PID-Controlled Constrained RL for Hub-based Joint Pricing, Dispatching, and Routing with Service Guarantees

Pengfei Du[1] , **Yucen Gao**[2*] , **Bin Wang**[1] and **Xiaochun Yang**[1*]
[1]School of Computer Science and Engineering, Northeastern University, China
[2]Software College, Northeastern University, China
{dupengfei, gaoyucen, yangxc, binwang}@mail.neu.edu.cn

Abstract

Joint optimization of pricing, dispatching, and routing is critical for hub-based mobility services but challenging due to complex decision couplings and strict service guarantees, such as Order Response Rate (ORR). Conventional constrained reinforcement learning often struggles in this mixed continuous–combinatorial action space, suffering from oscillatory behavior in Lagrangian dual variables and unstable constraint satisfaction. To address this, we propose PID-SACA, a unified framework that integrates an entropy-regularized actor–critic policy for continuous pricing and dispatching assisted by an embedded routing solver for execution-aware feedback. Crucially, we adapt the PID control mechanism to the Lagrangian dual update process. This approach leverages proportional, integral, and derivative feedback to dampen oscillations caused by stochastic gradient variance, ensuring robust long-term constraint enforcement. We provide theoretical analysis on the boundedness of dual variables, and experiments on publicly available large-scale mobility datasets demonstrate that PID-SACA significantly outperforms baselines, achieving high revenue with stable service compliance. Code: https://github.com/jerry0375/PID-SACA

1 Introduction

Hub-based mobility-on-demand (MoD) systems are integral to modern urban transit and paratransit services [Pavia *et al.*, 2024], concentrating demand and supply within tight spatio-temporal windows [Gao *et al.*, 2021]. Efficiently managing these systems requires the synergistic optimization of three coupled decisions: destination-level pricing, fleet dispatching, and vehicle routing [Sun *et al.*, 2022; Yuan *et al.*, 2023; Sun *et al.*, 2024; Zhou *et al.*, 2019]. As illustrated in Fig. 1, these decisions are intrinsically linked in a closed loop. Pricing reshapes the spatial distribution of demand by leveraging user sensitivity [Xin *et al.*, 2025], dispatching allocates scarce capacity to specific regions using advanced control policies [Han *et al.*, 2025], and routing determines the feasibility and operational cost of executing these commitments.

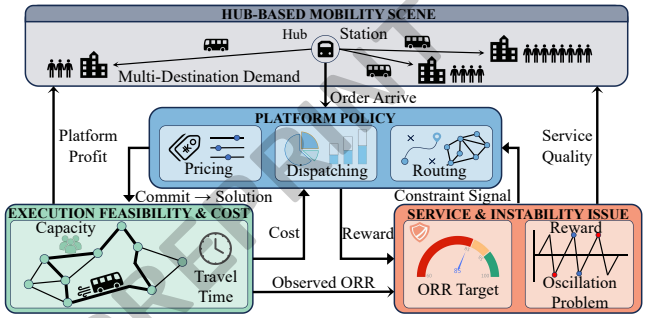


Figure 1: Schematic view of the Hub-based MoD system showing the interaction between Pricing, Dispatching, and Routing module.

Decoupling these components or employing myopic optimization strategies often yields suboptimal or execution-infeasible solutions. For instance, seemingly optimal dispatching decisions may become infeasible when grounded in actual road network constraints and travel times, a gap that recent neural routing solvers aim to bridge [Meng *et al.*, 2025; Xiao *et al.*, 2025]. Beyond profit maximization, platforms must strictly adhere to service-level requirements to ensure user retention and regulatory compliance [Pujari and Panda, 2025]. A critical metric in this domain is ORR, which measures the proportion of requests successfully served within a time slot. Neglecting this constraint in favor of pure profit maximization incentivizes reinforcement learning (RL) agents to "cherry-pick" high-value demand hotspots while systematically underserving remote or low-profit areas, violating fairness principles [Jiang *et al.*, 2025]. This behavior not only degrades user experience but also violates long-term operational mandates. Consequently, the problem evolves into a Constrained Markov Decision Process (CMDP) characterized by a hierarchical action space: continuous pricing and dispatching that dictate the inputs for combinatorial routing.

Solving this CMDP poses significant algorithmic challenges. Standard approaches, such as Lagrangian relaxation, often struggle with stability in this context [Wachi *et al.*, 2024; Gao *et al.*, 2023]. The interplay between profit maximization and service constraints often induces dual-variable oscillations, as widely reported in constrained RL [Tessler *et al.*, 2019]. The system alternates between aggressively overserving difficult regions to reduce accumulated penalties and

reverting to greedy, profit-seeking behavior when penalties subside. Furthermore, standard primal-dual updates are sensitive to the high variance typical of MoD environments, preventing convergence to a stable policy.

To address these challenges, this paper proposes PID-SACA, a unified framework for the Constrained Joint Dispatching, Routing, and Pricing (C-JDRP) problem. We model the system as a CMDP and employ an entropy-regularized actor–critic policy to jointly generate pricing and dispatching actions. To bridge the gap between high-level decisions and physical execution, we embed an attention-based routing solver into the learning loop to construct feasible routes and provide feedback on realized costs, enabling routing-aware policies. The core of our approach is the application of PID control theory to stabilize the Lagrangian multiplier updates, inspired by control-theoretic adaptations in deep RL [Stooke *et al.*, 2020]. By incorporating proportional, integral, and derivative feedback, this mechanism effectively dampens dual variable oscillations. Specifically, the derivative term helps anticipate error trends, smoothing the learning dynamics against stochastic noise and enabling the system to satisfy long-term ORR guarantees.

Our main contributions are:

- We formalize a hub-based joint pricing, dispatching, and routing problem with an explicit long-term ORR service requirement, capturing realistic operational coupling between learning decisions and routing feasibility.
- We demonstrate that adapting PID control to the Lagrangian dual update significantly improves training stability in high-variance MoD environments compared to standard primal-dual methods.
- We validate our approach on public large-scale real-world mobility benchmarks, showing consistent gains in revenue while meeting ORR targets, supported by ablations that isolate the benefits of the routing integration and the PID mechanism.

2 Related Work

This work lies at the intersection of joint decision-making in MoD systems and constrained reinforcement learning. We review prior efforts on coupled optimization of pricing, dispatching, and routing, and then highlight the missing service-aware perspective that motivates our CMDP formulation.

2.1 Joint Decision-Making for MoD

Due to the complexity of jointly optimizing pricing, dispatching, and routing, most studies focus on coupled subsets of decisions. In pricing–dispatching, [Zheng *et al.*, 2018] studied price-aware ridesharing dispatch with profit objectives and detour constraints, while [Tong *et al.*, 2018a] proposed matching-based dynamic pricing for spatial crowdsourcing to balance supply and demand. These methods provide economic control but often simplify route feasibility and network-level execution. In dispatching–routing, [Tong *et al.*, 2018b] developed a unified route-planning framework with flexible insertion operators, and [Jin *et al.*, 2019] proposed CoRide, a hierarchical multi-agent RL approach integrating dispatching with fleet repositioning. However, execution-focused methods typically assume exogenous demand and thus cannot shape demand through pricing.

Recent progress moves toward fully unified JDRP. [Gao *et al.*, 2021] introduced SACA, an end-to-end framework that uses a Soft Actor-Critic agent for pricing and dispatching together with an attention-based route planner, allowing high-level decisions to leverage low-level routing feedback. Despite strong performance, existing unified approaches largely pursue unconstrained revenue maximization and do not explicitly encode service-level requirements such as ORR. This can induce systematic preference for dense, high-value regions and chronic underservice of remote areas, conflicting with long-term service mandates. Our work addresses this gap by formulating JDRP as a CMDP, enabling stability-aware optimization under explicit service guarantees.

2.2 Constrained Reinforcement Learning for MoD

Enforcing strict service guarantees transforms the problem into a CMDP. While standard Lagrangian relaxation is the de facto method for solving CMDPs [Wachi *et al.*, 2024], it faces severe stability issues in MoD environments characterized by high variance and sparsity. Standard primal-dual updates often suffer from "overshoot," where the dual variable oscillates violently, causing the policy to alternate between ignoring constraints and over-conserving resources.

Recent works have attempted to mitigate this. [Hazra *et al.*, 2025] proposed *IP3O*, which uses adaptive incentives as soft barriers, and [Zhang *et al.*, 2025] introduced *SCRL* to handle situational constraints via probabilistic selection. However, these methods often introduce sensitive hyperparameters or fail to dampen the momentum of error accumulation in long-horizon tasks. Distinct from these approaches, we adapt PID control theory to the Lagrangian update [Stooke *et al.*, 2020]. By incorporating Integral (accumulated error) and Derivative (rate of change) feedback, our method explicitly dampens the oscillation of the multiplier, ensuring robust convergence to the target service level without manual reward shaping.

3 Problem Formulation

3.1 Preliminaries

We consider a hub-based bus service scenario where all buses depart from a central depot to deliver passengers to various destinations across the city and then return. The system operates in a round-based online manner.

Definition 1 (Time Slot)**.** *The operation time horizon is divided into discrete time slots $t = 1, 2, \ldots, T$ with a fixed interval Δt. In each slot t, the platform collects orders, updates vehicle states, and executes dispatching decisions.*

Definition 2 (Order Request)**.** *Let $\mathcal{D} = \{1, \ldots, K\}$ be the set of discrete destination regions. An order request o_i^t arriving at slot t is denoted as a tuple:*

$$o_i^t = \langle l_j, p_i \rangle, \quad j \in \mathcal{D},$$

where l_j represents the geographical location of destination j, and $p_i \in [p_{\min}, p_{\max}]$ denotes the trajectory-based price

charged to the passenger. The aggregate demand for destination k at slot t is collected in the set $\mathcal{O}_k^t$, and the total order set is $\mathcal{O}^t = \bigcup_{k\in\mathcal{D}} \mathcal{O}_k^t$.

Definition 3 (Bus Service)**.** *Consider a fleet of buses $\mathcal{B} = \{b_1, \ldots, b_M\}$. At the beginning of slot t, the state of bus b_i is defined as a tuple:*

$$b_i^t = \langle C_i,\ v_i,\ \zeta_i^t,\ \psi_i^t \rangle,$$

where C_i is the maximum passenger capacity, v_i is the average travel speed, ζ_i^t represents the planned route, and $\psi_i^t \in \mathbb{R}_{\geq 0}$ denotes the remaining service time*. A bus is available if and only if $\psi_i^t = 0$. The transition of the remaining service time follows:*

$$\psi_i^{t+1} = \max\{0,\ \psi_i^t - \Delta t\} + \mathbf{1}\{\zeta_i^{t+1} \neq \emptyset\} \cdot \mathrm{Time}(\zeta_i^{t+1}), \quad (1)$$

where $\mathrm{Time}(\cdot)$ estimates the travel time for a newly assigned route, and the indicator function $\mathbf{1}\{\cdot\}$ ensures the service time is updated only when a valid route is assigned ($\zeta_i^{t+1} \neq \emptyset$).

3.2 Supply and Demand Modeling

Demand Modeling. Let $\mathcal{O}^t$ be the set of order requests in slot t. We denote the location of destination region k as l_k. The *potential demand* N_k^t is the count of users wishing to travel to l_k, i.e., $N_k^t = |\{o_i^t \in \mathcal{O}^t \mid \mathrm{dest}(o_i^t) = l_k\}|$.

We define a unified pricing strategy p_k^t for destination k. The realized demand D_k^t is modeled by a conversion rate function $F_k^d(\cdot)$, which captures the user's sensitivity to both price and travel distance. Let d_k denote the distance from the hub to location l_k. The demand function is formulated as:

$$D_k^t = N_k^t \cdot F_k^d(p_k^t, d_k), \quad (2)$$

$$F_k^d(p_k^t, d_k) = 1 - \frac{d_{\max} + d_{\min} - d_k}{d_{\max}} \cdot \left(\frac{e^{p_k^t} - 1}{e - 1} \right), \quad (3)$$

where $d_{\min}$ and $d_{\max}$ represent the minimum and maximum distances in the service area, respectively. This formulation reflects that users traveling longer distances are less sensitive to price increases compared to those with shorter trips.

Supply Modeling. The platform controls the fleet and allocates capacity explicitly. The total available capacity at slot t is $S_{total}^t = \sum_{b_i\in\mathcal{B}} C_i \cdot \mathbf{1}\{\psi_i^t = 0\}$. Let $a_k^t \in [0,1]$ be the dispatching ratio for region k, subject to $\sum_k a_k^t \leq 1$. The effective supply for destination k is $S_k^t = S_{total}^t \cdot a_k^t$. Consequently, the number of fulfilled orders O_k^t is the minimum of demand and supply:

$$O_k^t = \min\{D_k^t, S_k^t\}.$$

3.3 Objective and Constraints

Definition 4 (Profit)**.** *The platform's profit R^t in slot t is the profit from fulfilled orders minus operational costs:*

$$R^t = \sum_{k\in\mathcal{D}} p_k^t O_k^t - \sum_{b_i\in\mathcal{B}} \mathrm{cost}(\zeta_i^t) \cdot \mathbf{1}\{\psi_i^t = 0\} \cdot \mathbf{1}\{\psi_i^{t+1} > 0\},$$

The routing process is formulated to find the optimal trajectory ζ_i^t that minimizes the total travel distance given the assigned destination set. Consequently, the operational cost is defined as $\mathrm{cost}(\zeta_i^t) = \beta_d \,\mathrm{dist}(\zeta_i^t)$, where β_d is a distance coefficient and $\mathrm{dist}(\zeta_i^t)$ is the length of this distance-optimized route. The indicator terms charge this cost only when a bus is newly dispatched.

The routing process is formulated as finding the optimal trajectory ζ_i^t that minimizes travel distance/time given the assigned destination set, serving as the physical constraint for the dispatching layer.

Definition 5 (Order Response Rate)**.** *The ORR quantifies the service level at slot t:*

$$ORR^t = \frac{\sum_{k\in\mathcal{D}} O_k^t}{\sum_{k\in\mathcal{D}} N_k^t}.$$

Definition 6 (C-JDRP)**.** *The C-JDRP problem aims to sequentially determine pricing and dispatching decisions over a horizon T, while* explicitly enforcing *a minimum service-level requirement measured by ORR. Formally, we solve:*

$$\max \quad \sum_{t=1}^{T} R^t \quad \textit{s.t.}\ \frac{1}{T}\sum_{t=1}^{T} ORR^t \geq \delta, \quad (4)$$

where δ denotes the required minimum long-term average ORR over the time horizon.

4 Methodology

As illustrated in Fig. 2, our framework employs an alternating optimization scheme: the primal policy jointly generates pricing and dispatching, while the dual variable is regulated by a PID controller to stabilize constraint satisfaction.

4.1 CMDP Formulation

We first map the JDRP elements to a standard MDP tuple $\langle \mathcal{S}, \mathcal{A}, \gamma \rangle$.

State Space $\mathcal{S}$. The state s_t captures the global demand distribution and fleet status. We define:

$$s_t = \left\langle \mathbf{N}^t,\ \mathbf{B}_{idle}^t,\ \mathcal{G}^t \right\rangle,$$

where $\mathbf{N}^t = \{N_k^t\}_{k\in\mathcal{D}}$ represents the potential demand across all regions, $\mathbf{B}_{idle}^t = \{b_i^t \in \mathcal{B} \mid \psi_i^t = 0\}$ represents the set of available buses, and $\mathcal{G}^t$ encapsulates global context features such as time of day and traffic conditions.

Action Space $\mathcal{A}$. The action $a_t = \langle \mathbf{P}^t, \mathbf{A}^t \rangle$ corresponds to the decision variables in the JDRP formulation:

- **Pricing $\mathbf{P}^t$:** A continuous vector determining the price $p_k^t \in [p_{\min}, p_{\max}]$ for each destination $k \in \mathcal{D}$.
- **Dispatching $\mathbf{A}^t$:** A continuous vector representing the dispatching ratio $a_k^t \in [0,1]$ for each region, normalized such that $\sum a_k^t \leq 1$.

Reward Function. We consider R^t in Definition 4 as the immediate reward function.

With the state, action, and reward defined above, we now cast C-JDRP (Definition 6) as a CMDP. Specifically, we maximize the discounted cumulative profit while enforcing the long-term ORR constraint, which will be handled via Lagrangian relaxation in the next subsection.

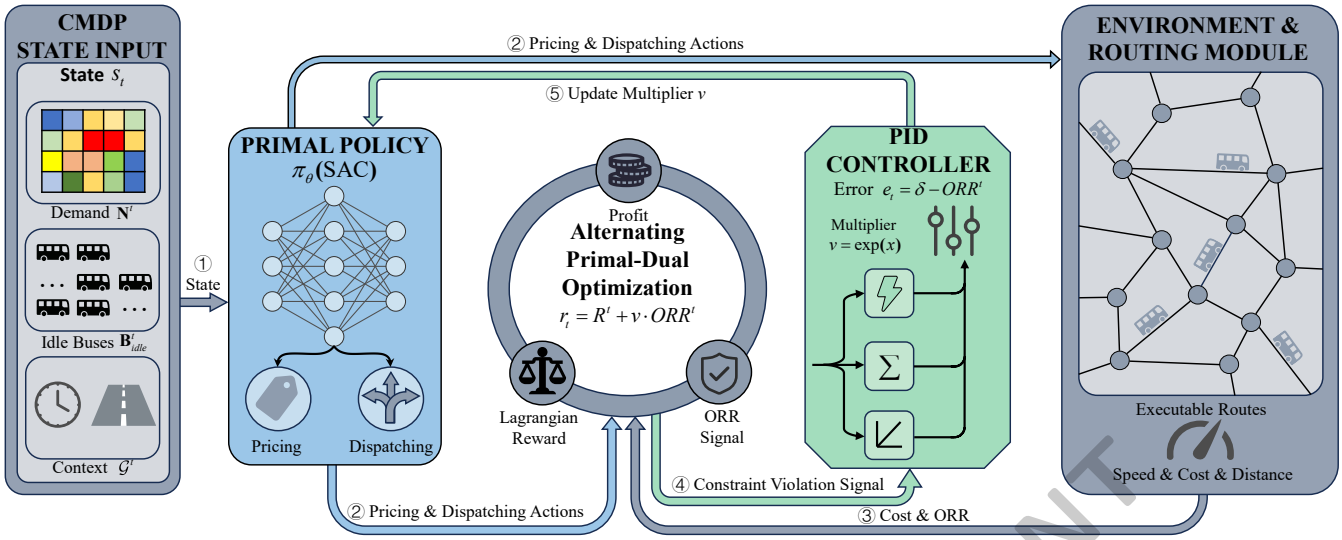


Figure 2: Overview of the proposed PID-SACA framework. The problem is formulated as a CMDP. The primal policy is optimized via SAC, while a PID-controlled Lagrangian multiplier enforces ORR constraint through alternating updates.

$$\begin{aligned} \max_{\pi_\theta} \quad & J_{\text{Pro}}(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T}\gamma^t\, R^t(s_t,a_t)\right] \\ \text{s.t.} \quad & J_{\text{ORR}}(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\frac{1}{T}\sum_{t=0}^{T}\text{ORR}^t\right] \geq \delta\,. \end{aligned} \tag{5}$$

4.2 Execution Layer: Attention-Based Routing

To ensure high-level dispatching decisions are physically feasible and service-oriented, we integrate a differentiable routing solver into the learning loop. Unlike heuristic estimates, this solver explicitly constructs routes to minimize the total travel distance, providing accurate feedback on operational costs and improving real-world execution reliability.

We formulate the routing sub-problem via the pointer-network paradigm [Vinyals *et al.*, 2015]. For a bus b_i assigned a set of destinations, the decoder sequentially selects the next station to visit based on the encoder embeddings of all pending stations. The probability of visiting station k_m at step m is given by:

$$\pi_\phi(k_m \mid k_{<m}, s_t) = \text{softmax}\big(\text{Attn}(h^m_{\text{dec}}, \{h_k\}_{k\in\mathcal{D}})\big), \tag{6}$$

where h^m_{dec} is the decoder state and $\{h_k\}$ are station embeddings. This routing module effectively acts as the environment's transition function component, determining the realization terms $\text{Time}(\tau_i^{t+1})$ and $\text{cost}(\tau_i^t)$ defined in Eq. (1) and Definition 4. By closing the loop, the upper-level pricing and dispatching policy π_θ learns to avoid decisions that would incur excessive routing latency or costs.

4.3 PID Lagrangian Relaxation and Alternating Optimization

Since the direct optimization of the constrained objective in a high-dimensional continuous action space is intractable, we employ the Lagrangian multiplier method to convert the CMDP into an unconstrained dual problem.

We construct the Lagrangian function $\mathcal{L}(\theta,\nu)$ by relaxing the constraint into the objective:

$$\mathcal{L}(\theta,\nu) = J_{\text{Pro}}(\pi_\theta) + \nu\cdot\left(J_{\text{ORR}}(\pi_\theta) - \delta\right), \tag{7}$$

where $\nu \geq 0$ is the learnable Lagrange multiplier acting as a penalty coefficient for constraint violation. This formulation establishes a min-max game (primal-dual problem):

$$\begin{aligned} &\min_{\nu\geq 0}\max_{\theta}\mathcal{L}(\theta,\nu) = \\ &\min_{\nu\geq 0}\max_{\theta}\mathbb{E}\left[\sum_{t=0}^{T}\gamma^t R^t + \nu(ORR^t - \delta)\right]. \end{aligned} \tag{8}$$

To solve this minimax problem within the SAC framework, we perform alternating updates during the training phase.

Primal Update (Policy Optimization). Fixing ν, we update the policy parameters θ via the standard SAC gradient estimator to maximize [Haarnoja *et al.*, 2018] the Lagrangian reward r^t_{lag}:

$$r^t_{\text{lag}} = R^t + \nu\cdot ORR^t. \tag{9}$$

This formulation dynamically shifts the agent's focus between profit maximization and service quality (ORR^t) based on the magnitude of ν.

Dual Update (PID Multiplier Optimization). We update ν while fixing π_θ. To strictly enforce non-negativity, we parameterize $\nu = \exp(x)$. Since standard gradient descent acts as pure integral control and is prone to oscillation in dynamic environments, we employ a Proportional-Integral-Derivative (PID) controller on x to ensure robust convergence.

We define the constraint violation error at step t as:

$$e_t := \delta - ORR^t. \tag{10}$$

Based on this error, the PID update magnitude Δx_t is computed as:

$$\Delta x_t = K_P e_t + K_I \sum_{\tau=0}^{t} e_\tau + K_D(e_t - e_{t-1}), \tag{11}$$

where $K_P, K_I, K_D \geq 0$ are the hyperparameters for the proportional, integral, and derivative terms, respectively. The log-multiplier x is then updated via gradient ascent with a learning rate η:

$$x_{t+1} \leftarrow x_t + \eta \cdot \Delta x_t. \tag{12}$$

This mechanism offers superior robustness over standard Lagrangian methods by integrating three terms: the Proportional (K_P) addresses immediate violations (e_t); the Integral (K_I) eliminates steady-state bias by accumulating past errors over long horizons; and the Derivative (K_D) dampens oscillations by penalizing the error rate ($e_t - e_{t-1}$), collectively ensuring training stability and strict constraint satisfaction.

4.4 Theoretical Analysis

We establish the convergence and long-term constraint satisfaction of our PID-based Lagrangian relaxation using Lyapunov optimization theory. By constructing a Lyapunov function based on Kullback-Leibler divergence—tailored to the exponential updates of log-parameterization—we prove that the dual multiplier remains bounded and the service constraint is satisfied over the long run.

Definition 7 (Lyapunov Function)**.** *Let ν^* be the optimal Lagrange multiplier. We define the potential function $V(\nu_t)$ as:*

$$V(\nu_t) = \nu^* \log\left(\frac{\nu^*}{\nu_t}\right) + \nu_t - \nu^*. \tag{13}$$

$V(\nu_t)$ is strictly convex and non-negative, with $V(\nu_t) = 0$ if and only if $\nu_t = \nu^*$.

Theorem 1 (Boundedness)**.** *Define the violation $e_t \triangleq \delta - \mathrm{ORR}_t$ and integral $I_t \triangleq \sum_{\tau=0}^{t} e_\tau$. Assume: (1) Slater's condition: $\exists \bar{\pi}, \epsilon > 0$ s.t. $\mathbb{E}[e_t|\bar{\pi}] \leq -\epsilon$; (2) Boundedness: $e_t \in [-1, 1]$ and rewards are bounded; (3) PID parameters $K_P, K_D \geq 0, K_I > 0$; (4) Update rule $\nu_{t+1} = \nu_t \exp(\eta \Delta x_t)$ with effective step $\Delta x_t = K_P e_t + K_I I_t + K_D(e_t - e_{t-1})$ satisfying $|\eta \Delta x_t| \leq 1$ (via clipping). Then, the multiplier sequence $\{\nu_t\}$ is bounded.*

Proof. **Decomposition.** Rewrite the PID signal as $\Delta x_t = a e_t + b_t$, where $a \triangleq K_P + K_I + K_D$ aggregates current feedback, and $b_t \triangleq K_I I_{t-1} - K_D e_{t-1}$ captures history. Note that the derivative component in b_t is bounded since $e_t \in [-1, 1]$.

Drift. Using $e^z \leq 1 + z + z^2$ for $|z| \leq 1$, the drift $\Delta V_t \triangleq V(\nu_{t+1}) - V(\nu_t)$ satisfies:

$$\begin{aligned}\frac{1}{\eta}\Delta V_t - R_t &\leq (\nu_t - \nu^*)\Delta x_t + \eta \nu_t \Delta x_t^2 - R_t \\ &\leq \underbrace{[a\nu_t e_t - R_t]}_{\text{Actor Objective}} - a\nu^* e_t + (\nu_t - \nu^*) b_t + \eta \nu_t \Delta x_t^2. \end{aligned} \tag{14}$$

The actor minimizes the Lagrangian-like term $a\nu_t e_t - R_t$.

Boundedness. By Slater's condition, for sufficiently large $\nu_t \geq \nu_{\mathrm{th}}$, the optimal policy forces $\mathbb{E}[e_t|\mathcal{F}_t] \leq -\epsilon' < 0$. Consequently, the integral state I_t acquires a negative drift: $\mathbb{E}[I_t|\mathcal{F}_t] \leq I_{t-1} - \epsilon'$. If ν_t were to diverge to $+\infty$, it would persistently exceed ν_{th}, causing I_{t-1} (and thus b_t) to decrease indefinitely. Since $K_I > 0$, a sufficiently negative I_{t-1} dominates the bounded terms ae_t and $-K_D e_{t-1}$, forcing $\Delta x_t < 0$ and $\nu_{t+1} < \nu_t$. Therefore, ν_t cannot diverge. □

Theorem 2 (Long-Term Constraint Satisfaction)**.** *Under the assumptions of Theorem 1, the time-averaged constraint violation is non-positive:*

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} (\delta - \mathrm{ORR}_t) \leq 0.$$

Proof. Let $e_t = \delta - \mathrm{ORR}_t$ and $I_t = \sum_{\tau=0}^{t} e_\tau$ as in Theorem 1. By Theorem 1, ν_t is bounded above, hence $x_t = \log \nu_t$ is bounded above as well.

We prove by contradiction. Suppose there exists $\bar{\epsilon} > 0$ such that

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} e_t \geq \bar{\epsilon}.$$

Then for infinitely many T large enough, implying

$$I_{T-1} = \sum_{t=0}^{T-1} e_t \geq \frac{\bar{\epsilon}}{2} T \xrightarrow[T\to\infty]{} +\infty. \tag{15}$$

From the PID law and the bound $|e_t - e_{t-1}| \leq 2E_{\max}$,

$$\begin{aligned}\Delta x_t &= K_P e_t + K_I I_t + K_D(e_t - e_{t-1}) \\ &\geq K_I I_t - (K_P + 2K_D) E_{\max}. \end{aligned} \tag{16}$$

Combining with (15) shows that Δx_t becomes unboundedly positive along an infinite subsequence, and thus

$$x_{t+1} - x_t = \eta \Delta x_t \to +\infty \Rightarrow x_t \to +\infty \Rightarrow \nu_t = e^{x_t} \to +\infty,$$

contradicting the boundedness of $\{\nu_t\}$ established in Theorem 1. Hence the assumed positive long-run average violation cannot occur, and we must have

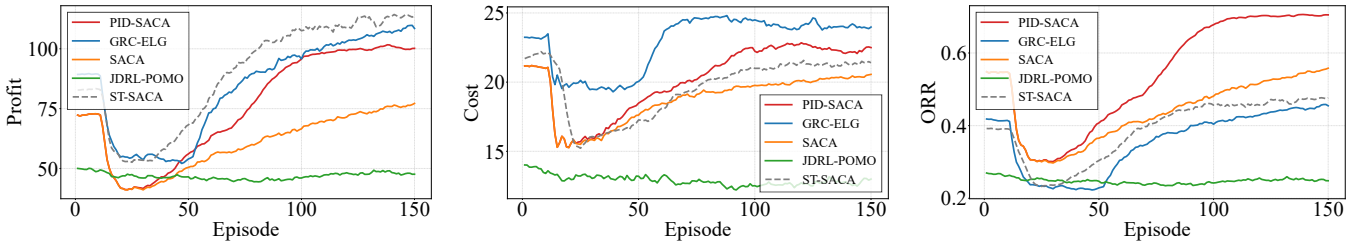
$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} e_t \leq 0,$$

i.e., the service constraint is satisfied in the long run. □

5 Experiments

In this section, we present a comprehensive evaluation of our proposed framework. We aim to answer the following research questions:

- **RQ1:** Can our method maximize the platform's profit while strictly satisfying ORR constraint compared to state-of-the-art baselines?
- **RQ2:** How does the proposed PID-based Lagrangian mechanism contribute to training stability and convergence speed?
- **RQ3:** What is the impact of individual modules and hyperparameters on the final performance?

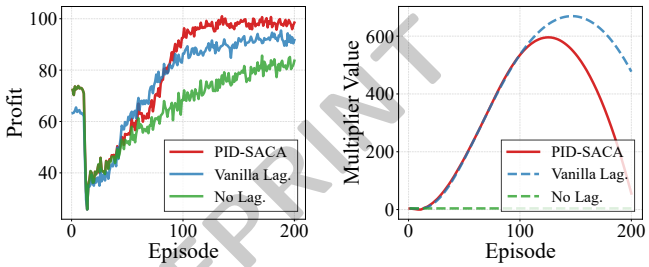


(a) Profit Convergence (b) Operational Cost (c) Order Response Rate

Figure 3: Training curves of different methods over 150 episodes under an ORR constraint threshold of $\delta = 0.7$. (a) Profit trends; (b) Operational Cost trends; (c) ORR trends. Our PID-SACA (red) achieves stable convergence while maintaining the target service level.

Method	Profit	Cost	ORR	Dist.
PID-SACA	100.14	22.46	**0.70**	150.50
GRC-ELG	<u>106.12</u>	24.00	0.44	160.02
SACA	74.49	<u>20.27</u>	<u>0.54</u>	<u>135.15</u>
JDRL-POMO	48.02	**12.77**	0.25	**85.12**
ST-SACA	**111.99**	21.35	0.47	142.30

Table 1: Quantitative comparison of different methods. Bold indicates the best result, and underline indicates the second best.



(a) Profit Convergence (b) Lagrange Multiplier

Figure 4: Analysis of training stability and convergence.

5.1 Experimental Settings

Dataset. We evaluate our method on a large-scale real-world ride-hailing order dataset released by the Didi Chuxing GAIA Initiative [DiDi Chuxing GAIA Initiative, 2016]. The dataset contains orders from Chengdu, China, spanning November 1–30, 2016 (30 days) and includes 7,065,937 records. Each record provides fields such as a unique ID, billing time, and GPS coordinates.

Baselines. We compare our method against the following representative state-of-the-art baselines:

- **JDRL-POMO:** A robust baseline that combines the JDRL dispatching policy [Sun *et al.*, 2022] with POMO [Kwon *et al.*, 2020], a routing solver leveraging instance augmentation to improve routing quality.
- **GRC-ELG:** A hybrid baseline integrating goal-conditioned RL for order dispatching [Lei and Ukkusuri, 2023] and ELG attention for route planning [Yoon *et al.*, 2024]. It jointly optimizes pricing and routing to improve long-term fleet revenue and operational efficiency.
- **SACA:** A two-module baseline with a SAC-based policy for joint pricing and dispatching, coupled with an attention-based routing policy [Kool *et al.*, 2019] to construct routes for assigned demand [Gao *et al.*, 2021].
- **ST-SACA:** An enhanced SACA variant that injects order coordinates during encoding, strengthening station-level spatio-temporal demand representation.

5.2 Destination Selection

Data Denoising. We apply DBSCAN [Ester *et al.*, 1996] to filter raw GPS noise and outliers, retaining a high-density destination set that characterizes core travel patterns.

Destination Construction. The filtered points are clustered using Partitioning Around Medoids (PAM) [Kaufman and Rousseeuw, 1990] to ensure robustness, yielding $K = 30$ destination medoids $\mathcal{D} = \{1, \ldots, 30\}$.

Integration with Dispatching. Given pricing and dispatching decisions, this module constructs feasible routes ζ_i^t to determine realized travel times and operational costs.

5.3 Results and Analysis

Comparison. We evaluate the performance of our proposed PID-SACA against four baselines: JDRL-POMO, GRC-ELG, ST-SACA, and the vanilla SACA. The training curves for Profit, Operational Cost, and ORR are illustrated in Fig. 3, and the quantitative metrics are summarized in Table 1.

As shown in Fig. 3 and Table 1, while ST-SACA and GRC-ELG maximize profit via "cherry-picking," they suffer from poor coverage (ORR ≈ 0.47). In contrast, PID-SACA achieves a dominant ORR of 0.70 ($\sim$30% improvement) with only a marginal profit trade-off, successfully balancing global service guarantees. Regarding cost (Fig. 3(b)), PID-SACA maintains consistently high operational efficiency comparable to ST-SACA. In contrast, JDRL-POMO's lower cost is merely an artifact of fleet idleness.

Ablation Study And Training Stability. To investigate the impact of the proposed mechanism on learning dynamics, we analyze the convergence behavior and stability of the Lagrange multiplier. The unconstrained baseline w/o Lagrange yields the lowest profit of 77.18, indicating that without service constraints acting as regularization, the agent adopts a myopic policy failing to sustain long-term revenue. Among

Method	Capacity = 40				Capacity = 60				Capacity = 80			
	Profit	ORR	Cost	Dist.	Profit	ORR	Cost	Dist.	Profit	ORR	Cost	Dist.
PID-SACA	48.31	**0.39**	21.57	143.83	84.49	**0.61**	21.99	146.63	100.14	**0.70**	22.46	150.50
GRC-ELG	**52.19**	0.25	22.50	149.97	81.29	0.35	23.44	156.28	106.12	0.44	24.00	160.02
JDRL-POMO	27.45	0.12	8.49	56.63	33.39	0.15	8.80	58.70	48.02	0.25	12.77	85.12
SACA	43.88	0.34	19.51	130.09	62.18	0.46	20.30	135.36	74.49	0.54	20.27	135.15
ST-SACA	47.93	0.21	16.52	110.15	**84.77**	0.37	20.78	138.52	**111.99**	0.47	21.35	142.30

Table 2: Performance comparison under varying vehicle capacities $C \in \{40, 60, 80\}$.

Method	Profit	ORR	Dist.	Cost	Std(ν)
PID-SACA	**100.14**	**0.70**	150.50	22.46	208.11
w/o PID	90.18	0.63	136.50	20.48	241.68
w/o Lagrange	77.18	0.56	136.45	20.47	\

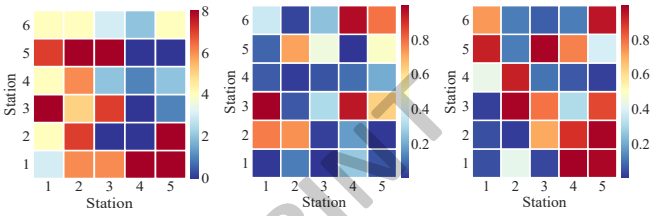
Table 3: Analysis of the impact of the PID and Lagrangian relaxation on training stability and convergence performance.

Target δ	0.10	0.25	0.40	0.55	0.70	0.80	0.90
Profit	81.93	81.90	87.35	99.56	**100.46**	96.13	89.88
ORR	0.59	0.59	0.63	0.70	0.71	0.81	**0.85**

Table 4: Sensitivity analysis of the constraint threshold δ.

constrained methods, the w/o PID variant exhibits instability characterized by a higher multiplier standard deviation of 241.68. This oscillation implies the dual variable overshoots and disrupts policy updates, resulting in suboptimal convergence with a profit of 90.18. In contrast, PID-SACA effectively utilizes derivative feedback to dampen fluctuations and reduces the multiplier standard deviation to 208.11. This stabilized dual update process enables the agent to reach a superior equilibrium, achieving the highest profit of 100.14 while satisfying the ORR requirement of 0.70.

Sensitivity Analysis. Table 4 and Table 2 investigate the impact of the constraint threshold δ and vehicle capacity C, respectively. (1) Impact of Threshold δ: Table 4 illustrates a non-trivial trade-off between revenue and service quality (ORR). Interestingly, at moderate levels ($\delta = 0.70$), the system achieves a *synergy* where improved coverage boosts revenue to a peak of 100.46. However, as we tighten the constraint ($\delta \geq 0.80$), the PID mechanism effectively prioritizes strict service guarantees (ORR rises to 0.85) over peak profitability. This confirms PID-SACA functions as a flexible tuner, navigating the Pareto frontier between operational efficiency and social responsibility. (2) Scalability on Capacity C: Table 2 further validates the robustness of our method against varying capacities. While baselines like ST-SACA aggressively pursue profit at the cost of service fairness (low ORR), PID-SACA consistently maintains the highest ORR across all capacities (e.g., 0.70 vs. 0.47 at $C = 80$) while delivering competitive revenue growth. This demonstrates that our approach ensures reliable service coverage regardless of resource abundance or scarcity.



(a) Demand Distribution (b) PID-SACA (Ours) (c) Vanilla (Base)

Figure 5: Spatial heatmaps at peak snapshot ($t = 12$). **(a)** Demand distribution. **(b)** PID-SACA aligns prices with demand. **(c)** Vanilla baseline shows misaligned pricing. Red indicates high values.

Case Study. Fig. 5 visualizes the pricing policy during a peak snapshot. As demand concentrates in the CBD and Transport Hub (Fig. 5(a)), PID-SACA (Fig. 5(b)) demonstrates topology-aware regulation: it applies surge pricing in hotspots while lowering prices in sparse western regions. This confirms that the PID controller guides the agent to stimulate conversion in underserved areas to satisfy global ORR constraints without sacrificing core profit. In contrast, the Vanilla Lagrangian baseline (Fig. 5(c)) counter-productively maintains high prices in low-demand zones, suppressing order volume. This highlights the necessity of the PID mechanism in stabilizing the trade-off between local profit maximization and global service mandates.

6 Conclusion

This paper investigates the constrained joint pricing, dispatching, and routing problem (C-JDRP) for mobility-on-demand systems, emphasizing the critical trade-off between profit maximization and service level guarantees. We propose PID-SACA, a unified framework that couples an entropy-regularized actor–critic policy with an attention-based routing module to ensure operational feasibility. To address the instability of standard Lagrangian methods in mixed continuous-combinatorial action spaces, we introduce a PID-controlled dual update mechanism. Theoretical analysis confirms that this approach ensures bounded dual variables and long-term constraint satisfaction. Extensive experiments on real-world datasets demonstrate that PID-SACA significantly outperforms competitive baselines, achieving high profitability while robustly maintaining the target Order Response Rate through stable learning dynamics. This work provides a scalable and theoretically grounded paradigm for sustainable urban fleet management.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions. This work was partially supported by the National Key Research and Development Program of China (No. 2024YFF0617702); the National Natural Science Foundation of China (Nos. U22A2025, 62232007, and U23A20309); the 111 Project (No. B16009); the Ant Group Research Program (No. 2025021900003); the Fundamental Research Funds for the Central Universities (No. N2417007); and the CCF-DiDi GAIA Collaborative Research Funds for Young Scholars.

References

[DiDi Chuxing GAIA Initiative, 2016] DiDi Chuxing GAIA Initiative. DiDi Chuxing GAIA Initiative. https://outreach.didichuxing.com/, 2016. Accessed: 7 July 2022.

[Ester *et al.*, 1996] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996.

[Gao *et al.*, 2021] Yucen Gao, Yuanning Gao, Yuhao Li, Xiaofeng Gao, Xiang Li, and Guihai Chen. An attention-based bi-gru for route planning and order dispatch of bus-booking platform. In *Proceedings of the International Conference on Database Systems for Advanced Applications*, pages 609–624, 2021.

[Gao *et al.*, 2023] Yucen Gao, Yulong Song, Xikai Wei, Xiaofeng Gao, and Guihai Chen. Saca: An end-to-end method for dispatching, routing, and pricing of online bus-booking. In *International Conference on Database Systems for Advanced Applications*, pages 303–313, 2023.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 1856–1865. PMLR, 2018.

[Han *et al.*, 2025] Xiao Han, Zijian Zhang, Xiangyu Zhao, Yuanshao Zhu, Guojiang Shen, Xiangjie Kong, Xuetao Wei, Liqiang Nie, and Jieping Ye. GARLIC: gpt-augmented reinforcement learning with intelligent control for vehicle dispatching. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, pages 255–263. AAAI Press, 2025.

[Hazra *et al.*, 2025] Somnath Hazra, Pallab Dasgupta, and Soumyajit Dey. Incentivizing safer actions in policy optimization for constrained reinforcement learning. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, pages 5318–5326, 2025.

[Jiang *et al.*, 2025] Lin Jiang, Yu Yang, and Guang Wang. Hcride: Harmonizing passenger fairness and driver preference for human-centered ride-hailing. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, pages 10289–10297, 2025.

[Jin *et al.*, 2019] Jiarui Jin, Ming Zhou, Weinan Zhang, Minne Li, Zilong Guo, Zhiwei Qin, Yan Jiao, Xiaocheng Tang, Chenxi Wang, Jun Wang, et al. Coride: Joint order dispatching and fleet management for multi-scale ride-hailing platforms. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1983–1992, 2019.

[Kaufman and Rousseeuw, 1990] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley, 1990.

[Kool *et al.*, 2019] Wouter Kool, Herke van Hoof, and Max Welling. Attention, learn to solve routing problems! In *Proceedings of the International Conference on Learning Representations*, 2019.

[Kwon *et al.*, 2020] Yeong-Dae Kwon, Jinho Choo, Byoungjip Kim, Iljoo Yoon, Youngjune Gwon, and Seungjai Min. POMO: Policy Optimization with Multiple Optima for Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21188–21198, 2020.

[Lei and Ukkusuri, 2023] Zengxiang Lei and Satish V. Ukkusuri. Scalable reinforcement learning approaches for dynamic pricing in ride-hailing systems. *Transportation Research Part B: Methodological*, 178:102848, 2023.

[Meng *et al.*, 2025] Dian Meng, Zhiguang Cao, Yaoxin Wu, Yaqing Hou, Hongwei Ge, and Qiang Zhang. Eformer: An effective edge-based transformer for vehicle routing problems. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, pages 8582–8590, 2025.

[Pavia *et al.*, 2024] Sophie Pavia, David Rogers, Amutheezan Sivagnanam, Michael Wilbur, Danushka Edirimanna, Youngseo Kim, Philip Pugliese, Samitha Samaranayake, Aron Laszka, Ayan Mukhopadhyay, and Abhishek Dubey. Deploying mobility-on-demand for all by optimizing paratransit services. In *Proceedings of the 33th International Joint Conference on Artificial Intelligence*, pages 7430–7437, 2024.

[Pujari and Panda, 2025] Siva Kumar Pujari and Sangita Panda. Impact of dynamic pricing and driver behavior on service quality in ride-hailing operations: a study of bangalore's urban dynamics. *The TQM Journal*, pages 1–19, 2025.

[Stooke *et al.*, 2020] Adam Stooke, Joshua Achiam, and Pieter Abbeel. Responsive safety in reinforcement learning by PID lagrangian methods. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9133–9143. PMLR, 2020.

[Sun *et al.*, 2022] Jiahui Sun, Haiming Jin, Zhaoxing Yang, Lu Su, and Xinbing Wang. Optimizing Long-Term Efficiency and Fairness in Ride-Hailing via Joint Order Dispatching and Driver Repositioning. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3950–3960, 2022.

[Sun *et al.*, 2024] Jiahui Sun, Haiming Jin, Zhaoxing Yang, and Lu Su. Optimizing long-term efficiency and fairness

in ride-hailing under budget constraint via joint order dispatching and driver repositioning. *IEEE Trans. Knowl. Data Eng.*, 36(7):3348–3362, 2024.

[Tessler *et al.*, 2019] Chen Tessler, Daniel J. Mankowitz, and Shie Mannor. Reward constrained policy optimization. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

[Tong *et al.*, 2018a] Yongxin Tong, Libin Wang, Zimu Zhou, Lei Chen, Bowen Du, and Jieping Ye. Dynamic pricing in spatial crowdsourcing: a matching-based approach. In *Proceedings of the 2018 International Conference on Management of Data*, pages 773–788, 2018.

[Tong *et al.*, 2018b] Yongxin Tong, Yuxiang Zeng, Zimu Zhou, Lei Chen, Jieping Ye, and Ke Xu. A unified approach to route planning for shared mobility. *Proceedings of the VLDB Endowment*, 11(11):1633–1646, 2018.

[Vinyals *et al.*, 2015] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700, 2015.

[Wachi *et al.*, 2024] Akifumi Wachi, Xun Shen, and Yanan Sui. A survey of constraint formulations in safe reinforcement learning. In *Proceedings of the 33th International Joint Conference on Artificial Intelligence*, pages 8262–8271, 2024.

[Xiao *et al.*, 2025] Yubin Xiao, Yuesong Wu, Rui Cao, Di Wang, Zhiguang Cao, Xuan Wu, Peng Zhao, Yuanshu Li, You Zhou, and Yuan Jiang. DGL: dynamic global-local information aggregation for scalable VRP generalization with self-improvement learning. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, pages 8669–8677, 2025.

[Xin *et al.*, 2025] Yun Xin, Jianfeng Lu, Shuqin Cao, Gang Li, Haozhao Wang, and Guanghui Wen. Daringfed: A dynamic bayesian persuasion pricing for online federated learning under two-sided incomplete information. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, pages 6687–6695, 2025.

[Yoon *et al.*, 2024] Jee Seok Yoon, Junghyo Sohn, Wootaek Jeong, and Heung-Il Suk. ExpertDiff: Head-less Model Reprogramming with Diffusion Classifiers for Out-of-Distribution Generalization. In *Proceedings of the 33th International Joint Conference on Artificial Intelligence*, pages 6866–6874, 2024.

[Yuan *et al.*, 2023] Enpeng Yuan, Wenbo Chen, and Pascal Van Hentenryck. Reinforcement learning from optimization proxy for ride-hailing vehicle relocation (extended abstract). In *Proceedings of the 32th International Joint Conference on Artificial Intelligence*, pages 6990–6994, 2023.

[Zhang *et al.*, 2025] Libo Zhang, Yang Chen, Toru Takisaka, Kaiqi Zhao, Weidong Li, and Jiamou Liu. Situational-constrained sequential resources allocation via reinforcement learning. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence*, pages 9121–9129, 2025.

[Zheng *et al.*, 2018] Libin Zheng, Lei Chen, and Jieping Ye. Order dispatch in price-aware ridesharing. *Proceedings of the VLDB Endowment*, 11(8):853–865, 2018.

[Zhou *et al.*, 2019] Hao Zhou, Yucen Gao, Xiaofeng Gao, and Guihai Chen. Real-time route planning and online order dispatch for bus-booking platforms. In *International Conference on Database Systems for Advanced Applications*, pages 748–763. Springer, 2019.