Online Multi-modal Evacuation during Passenger Flow Outburst in Urban Transit System: A Heterogeneous Multi-agent Reinforcement Learning Framework

Enze Liu[a], Shuguang Zhan[b,*], Yongqiu Zhu[c], Zhiyuan Lin[d], Dian Wang[b]

[a] *School of Mechanical Engineering, Hefei University of Technology, Hefei, China*
[b] *School of Automotive and Transportation Engineering, Hefei University of Technology, Hefei, China*
[c] *Department of Transport and Planning, Delft University of Technology, Delft, The Netherlands*
[d] *Institute for Transport Studies, University of Leeds, Leeds, United Kingdom*

Abstract

With growing demand straining urban transit systems' resilience in managing outburst passenger flows, existing approaches focused on offline and single-modal evacuations remain limited. This study proposes an online multi-modal evacuation framework that coordinates on-duty taxis, buses, and metros while minimizing impact on their regular services. We develop a data-driven agent-based environment to update multi-modal transit data and stranded passenger information in real time. Two coordination strategies are introduced: (1) an independent strategy using a distributed training and distributed execution algorithm, and (2) a collaborative strategy using a hybrid centralized training and distributed execution algorithm. To dynamically assess evacuation effectiveness, we design a resilience framework with three metrics: robustness, rapidity, and resourcefulness. These metrics are transformed into demand-responsive feedback at each time step, enabling agents to proactively generate resilient evacuation plans. In a real-world case study triggered by a railway disruption, our approach outperforms genetic algorithms and multi-agent deep deterministic policy gradient algorithms in computation time and solution quality under offline conditions. Simulated new environments further validate its online applicability, demonstrating its potential for real-world deployment.

Keywords: Urban transit, multi-modal evacuation, online, resilience, multi-agent reinforcement learning

1. Introduction

Outburst passenger flow (OPF), triggered by unexpected transit disruptions or large-scale gathering events, poses significant challenges to the resilience of urban transit systems, namely the adaptability and recoverability under abnormal conditions (Zhou et al., 2019). For example, a disruption on Singapore's urban rail line in September 2024 affected 358,000 commuters on the first day, many of whom sought alternative routes (Land Transport Guru, 2024). During the July 2021 Henan flood in China, all trains along the Longhai and Jingguang mainland railway lines were urgently canceled for an entire day, leaving thousands of passengers attempting to leave overcrowded railway stations (Hu et al., 2024). As sudden surges of passengers overwhelm transit services, the risk of overcrowding is heightened, causing potential accidents for stranded passengers (Xu et al., 2021). For passengers unable to evacuate independently (such as those without vehicles, carrying luggage, or traveling long distances), transit-based evacuation strategies are essential, which involve the dispatch of additional transit capacities to relocate stranded passengers safely and efficiently (Matherly et al., 2015). Once an unexpected OPF occurs, transit-based evacuation must rely on the on-duty services, such as unoccupied taxis, buses and metro capacities near the affected area, due to the absence of reserved back-up resources (Zhang et al., 2025). However, over-reliance on any single mode may not only constrain overall evacuation capacity but also compromise its serviceability for regular passengers outside the OPF area. Therefore, coordinated multi-modal evacuation is essential by strategically allocating passenger demand across modes based on the respective characteristics and capacity of each mode.

Due to the unpredictability of passenger accumulation and capacity availability, it is often infeasible to evacuate all passengers in a single planning cycle. Online evacuation planning is necessary, which continuously adapts evacuation plans to real-world conditions. To support such dynamic planning, real-time perception of both passenger mobility and transit capacity is essential. Agent-based models have been studied to simulate the dynamic interactions between passengers and vehicles during evacuations

*Corresponding author
Email address: shuguangzhan@hfut.edu.cn (Shuguang Zhan)

(Chang et al., 2024). However, existing models often rely on static simulation software and rule-based logic, limiting their responsiveness to real-world uncertainties. A data-driven agent-based model that can update environmental conditions in real time is therefore required to provide accurate inputs for effective evacuation planning. To adapt evacuation plans, a decision-making tool must be capable of generating solutions based on up-to-date information. Reinforcement learning, a machine learning paradigm for adaptive decision-making, offers a promising solution (Sutton and Barto, 2018). Pretrained agents can be directly deployed in the environment, without depending on predefined scenarios or computationally intensive optimization. This paradigm significantly enhances the flexibility of emergency response and reduces the computational time, enabling real-time emergency management in various domains, e.g., metro inflow control (Jiang et al., 2018) and train rescheduling (Ying et al., 2020). Therefore, this paper proposes a novel data-driven agent-based environment coupled with a reinforcement learning framework to enable effective and scalable online multi-modal evacuation.

To comprehensively evaluate the effectiveness of online evacuation planning during OPF, a dynamic indicator is needed to provide real-time feedback on the effectiveness of plans at each time step. While traditional indicators, such as network clearance time, risk exposure time, and the number of evacuated passengers (Jiang et al., 2022; Li et al., 2023), focus on the overall post-evacuation effectiveness, resilience, which reflects the propagation of system performance during abnormal events, has not been fully explored. Resilience can be captured by the "resilience triangle" and the "4R" metrics (i.e., redundancy, robustness, rapidity, and resourcefulness) (Bruneau et al., 2003). Therefore, the resilience of OPF evacuation, as a dynamic evaluation framework, requires further development.

This paper aims to propose an online multi-modal evacuation planning approach throughout the period of regional OPF. Specifically, this approach is designed to dynamically generate evacuation plans by coordinately dispatching the redundant capacity of each mode, while explicitly accounting for the impact of evacuation on the regular service. To achieve this, a multi-agent reinforcement learning (MARL) framework is developed that seamlessly integrates dispatching strategies for taxi, bus, and metro systems. Key strategies include: (a) deploying idle taxis to an OPF area; (b) dynamically reallocating bus fleets across emergency routes; and (c) regulating metro passenger inflow volumes to mitigate congestion. Considering the real-world mobility of multi-modal transit vehicles and passengers, a data-driven agent-based environment is constructed by leveraging multi-source datasets—including the taxi GPS traces, bus smart card transactions, bus automated vehicle location (AVL) logs, time-dependent metro origin-destination (OD) demand matrices, and anonymized mobile dataset. The framework is validated in a multi-modal hub in Xi'an, China, where a railway disruption causes thousands of passengers stranded, necessitating urgent evacuation. Our approach demonstrates superior computational efficiency compared to genetic algorithms (GA) and multi-agent deep deterministic policy gradient (MADDPG) algorithms in offline conditions. Furthermore, its online applicability is proved by transferring pre-trained agents to a series of new scenarios. The contributions of this work are threefold:

- We propose an online multi-modal evacuation planning approach. Unlike offline evacuation planning, our approach continuously generates evacuation plans in real time for each mode at each time step until the OPF subsides. In contrast to existing studies that focus on OPF within a single system, our approach targets the OPF in a geographic area, enabling a more practical and comprehensive analysis of overall passenger demand. Different from single-mode evacuation, we leverage the coordination of multiple modes with each mode having advantageous transit characteristics in terms of capacity and efficiency. By distributing passengers across multiple modes, our approach only relies on redundant capacities from on-duty services, thereby improving the generality of application by eliminating the need for backup capacities.
- We develop a novel MARL framework that coordinates capacity dispatch across multiple modes. Unlike MARL frameworks with homogeneous agents, our framework adopts heterogeneous agents, where each agent controls a specific mode, each with a distinct dispatching strategy and operational characteristics. To dynamically evaluate evacuation plans and provide feedback to agents, we establish a set of demand-responsive feedback functions based on a customized resilience framework.
- We introduce two MARL algorithms to train heterogeneous agents for coordinating multi-modal evacuations. A distributed training and distributed execution (DTDE) algorithm trains agents to independently control each mode and evacuate their respective demands. A hybrid centralized training and distributed execution (H-CTDE) algorithm trains agents to collaboratively manage multi-modal capacities, accounting for passengers' mode shifting. Compared with the traditional

centralized training and distributed execution (CTDE) algorithm, the H-CTDE algorithm incorporates a central critic that evaluates the overall effectiveness of evacuation plans, and a local critic that distinguishes the contribution of each mode and impact on its respective regular service. This algorithm ensures balanced utilization across modes, preventing over-reliance on any single mode and mitigating lazy or conservative behaviors among agents.

This paper is structured as follows. Section 2 reviews the related literature. Section 3 defines our strategy and the resilience framework. Section 4 formulates the mathematical model. Section 5 builds the MARL framework. Section 6 introduces the DTDE and H-CTDE training algorithms. Section 7 validates our approach through real-world case studies. Finally, the paper is concluded in Section 8.

2. Literature review

This paper focuses on the online multi-modal evacuation problem for improving the system's resilience under OPF. Therefore, related studies are reviewed from two aspects: transit-based evacuation planning for the OPF in Section 2.1, and dynamic evaluation for evacuation planning in Section 2.2.

2.1. Transit-based evacuation planning for outburst passenger flows

Studies on transit-based evacuation planning for OPF can be classified into offline and online approaches based on their reliance on the information, as described in Section 1. Studies on the offline and online evacuation planning are reviewed in Sections 2.1.1 and 2.1.2, respectively.

2.1.1. Offline evacuation planning

Traditional offline evacuation planning has focused on addressing a fixed number of evacuees using predetermined capacities. Studies have primarily focused on the bus system, with strategies including fleet dispatching, routing, and shelter (or destination) location optimization. Goerigk et al. (2015) first proposed an integer linear programming model that framed bus evacuation planning as a vehicle routing problem with multiple pick-up and shelter locations. The objective was to minimize the overall evacuation time. Teichmann et al. (2021) addressed large-scale evacuation planning under the nuclear leakage scenario. Evacuees were concentrated around the nuclear facility and needed to be transported away from the risky zone. Different from studies like Goerigk et al. (2015), the bus routing was assumed to be predefined, while the bus dispatching was carefully optimized to determine which fleet was assigned to which route. The primary objective was to minimize the clearance time of the risky area. However, overly relying on buses could restrict the evacuation capacity. Therefore, Yang et al. (2018) proposed a mixed-integer linear programming model that integrated taxis, buses and metros for collaborative evacuation. Different from single-modal evacuation, multi-modal capacities were coordinately dispatched. Passengers were distributed across multiple modes by a systematic optimization model, assuming that all passengers followed the assignment and were fully evacuated. Jiang et al. (2025) integrally dispatched conventional fixed-route buses and demand-responsive flexible-route buses to optimize passenger travel time and the number of evacuated passengers through vehicle routing and passenger assignment strategies. However, the assumption that evacuees and capacities are fixed is not applicable in many non-noticed emergencies where the dynamics of passenger growth and capacity availability increase the uncertainty in planning.

Some studies have incorporated dynamic models into offline evacuation planning to simulate evacuees' mobility. Chang et al. (2024) proposed a two-stage agent-based model to simulate pedestrian movement and bus routing process under a toxic gas leak scenario. An agent-based model was developed to dynamically simulate the mobility of passenger elements during bus evacuation (Note: To distinguish the term "agent" in the MARL and agent-based modeling context, the agents in the agent-based modeling method will be referred to as "elements" hereafter). However, the number of bridging buses was assumed to be sufficient to fully accommodate all passengers, which overlooked the propagation of stranded passengers. Hao et al. (2024) proposed a train dispatching strategy to address the OPF scenario in a metro system. The strategy focused on adding operating trains to enhance evacuation capacity. Wang and Jin (2025) proposed a train rescheduling and passenger assignment approach to evacuate stranded passengers following a rail disruption. The method incorporated non-disrupted train lines to construct alternative evacuation routes. Dynamic passenger inflow was considered, and the evacuation of all passengers was not mandatory. Instead, they aimed to maximize the number of evacuated passengers and minimize the number of additional trains. Ma et al. (2024) addressed another OPF scenario caused by significant airline cancellations, where stranded passengers at the airport required transit to return. Passenger demand was treated as dynamically arriving, as these passengers were originally heading for flights. Although

the above approaches considered the dynamic arrival of evacuees, the resources, including available buses and trains, were treated as statically given in the background. This assumption neglects the real-world capacity availability, which often depends on the redundancy of transit systems.

By using the dynamic redundant capacities in the urban transit system, Jiang et al. (2022) proposed a multi-modal evacuation strategy that involved taxis, buses, and metros. The availability of capacities was dynamically driven by taxi GPS, bus and metro smart card datasets. However, similar to Yang et al. (2018), both studies assumed evacuees as statically predetermined and ultimately fully evacuated, overlooking the passenger growth and their mode choice behavior in real-world scenarios.

2.1.2. Online evacuation planning

To enable online applications for OPF evacuation, several studies have explored inflow control strategies within metro systems. Zhang et al. (2021) addressed coordinated passenger flow control for multiple stations in a metro system using dynamic programming. A heuristic decomposition algorithm was proposed to manage flow control dynamically, relying only on real-time information. Liang et al. (2023), which also employed dynamic programming, developed a passenger flow control strategy for stations along the oversaturated metro lines, effectively addressing uncertainties in passenger demand through an online forward algorithm for real-time control. Moreover, Jiang et al. (2019) leveraged Q-learning to implement online inflow control and train stop-skipping strategies. Recently, Zhang et al. (2025) have explored bus evacuation for stranded passengers during rail disruptions. They integrated bus bridging and dispatching approach based on the rolling horizon approach, aiming to dynamically dispatch bridging buses while minimizing the impact on regular passengers. However, both their inflow control and bus evacuation components relied on a single mode of transport for evacuation, which limited overall capacity and operational efficiency.

In the multi-modal system, the metro and bus were coordinately dispatched for large-scale evacuation by Abdelgawad and Abdulhai (2010). They integrated demand estimation and routing simulation into a rolling horizon framework that generated evacuation plans for each time step. Wang et al. (2024) leveraged a rolling horizon framework, dispatching taxis while considering the surrounding regular bus services to address OPF in the taxi system. Passengers' mode choice was modeled based on travel time and fare costs. However, rolling horizon methods lack the Bellman equation structure to balance immediate and future benefits (Powell, 2011), thereby limiting the ability to proactively generate resilient evacuation plans during the OPF period. Additionally, Su et al. (2024, 2025) proposed a data-driven agent-based evacuation model that integrated taxis and buses for OPF evacuation. However, this method focused on the macroscopic evacuation rate, which did not assess precise plans of vehicle dispatching or passenger distribution. Therefore, current online multi-modal evacuation research has yet to fully integrate taxi, bus, and metro systems while accounting for the impact on their respective regular services. Moreover, as evacuation systems become increasingly complex, enhancing computational efficiency is essential to ensure real-time applicability.

2.1.3. Summary

Existing transit-based evacuation studies for OPF are summarized in Table 1, with their application scenarios and methodologies systematically classified. The abbreviations used in the table are defined as follows. The "Onl." column indicates whether the evacuation strategy is applicable online. "Eva." and "Res." columns denote, respectively, whether evacuees and resources are treated as static (S) or dynamic (D). "Mode" column refers to the transit modes considered, including taxi (T), bus (B), and metro (M). "Imp." column indicates whether the impact of evacuation on regular services is taken into account. "Cho." column represents the mode choice modeling approach, categorized as systematic optimization (Sys), independent choice (Ind), or mode shift behavior (Mod). "Model" column specifies the modeling methods used, such as integer linear programming (ILP), mixed-integer linear programming (MILP), mixed-integer nonlinear programming (MINLP), dynamic programming (DP), robust optimization (RO), simulation (Sim), rolling horizon approaches (RH), reinforcement learning (RL), and multi-agent reinforcement learning (MARL). "Strategy" column denotes the adopted evacuation strategies, including vehicle routing (VR), emergency vehicle dispatching (ED), destination relocating (DL), inflow control (IC), train rescheduling (TR), and passenger distribution (PD). "Objective" column refers to the components considered as optimization goals, including waiting time (WT), in-vehicle time (VT), operating cost (OC), and evacuated passengers number (EN).

According to the table, although many studies have focused on evacuation planning in offline settings,

online evacuation strategies have primarily matured within metro and bus systems (see Columns "Onl." and "Mode"). Moreover, few studies have simultaneously considered the dynamic nature of both evacuees and resources (see Columns "Eva." and "Res."). Additionally, the impact of evacuation on regular services has not been clearly addressed in the taxi system (see Column "Imp."). In the studies of multi-modal evacuation planning, passengers' mode choice behavior has not been addressed, except for Wang et al. (2024) and Su et al. (2024, 2025) (see Column "Cho."). Passengers' mode choice in these two studies is completely independent based on the trip cost of each mode (i.e., travel time and fare cost), regardless of their original mode choice while waiting for evacuation. However, as studies on mode-shifting behavior have shown, passengers typically prefer to remain with their original mode; only when the original mode is unavailable are they willing to shift to alternative modes (Li et al., 2020a,b; Gu and Chen, 2023). According to Column "Model", online multi-modal evacuation planning studies have primarily adopted the rolling horizon approach, which cannot proactively generate resilient evacuation plans by balancing immediate and future benefits. Consequently, developing an online multi-modal evacuation approach that accounts for passengers' mode-shifting behavior and the impact of evacuation on regular services requires further exploration.

To ensure generality, the proposed framework incorporates widely adopted evacuation strategies, including emergency dispatching (for taxis and buses), bus destination relocating, metro inflow control, and passenger distribution (see Column "Strategy"). The objectives include minimizing passenger waiting time, operational cost (measured by the impact on regular services), and maximizing the evacuated passenger number (see Column "Objective").

Table 1: Summary of transit-based evacuation literature

	Applicable scenario					Methodology			
Publication	Onl.	Eva.	Res.	Mode	Imp.	Cho.	Model	Strategy	Objective
Abdelgawad and Abdulhai (2010)	✓	D	S	B, M	×	Sys	Sim, RH	ED, PD	WT, VT, OC, EN
Goerigk et al. (2015)	×	S	S	B	×	-	ILP	VR	VT
Yang et al. (2018)	×	S	S	T, B, M	×	Sys	ILP	ED, PD	WT, VT
Jiang et al. (2018)	✓	D	S	M	✓	-	RL	IC, TR	WT
Teichmann et al. (2021)	×	S	S	B	×	-	MILP	ED, DL	VT, OC
Zhang et al. (2021)	✓	D	S	M	✓	-	DP	IC	EN
Jiang et al. (2022)	×	S	D	T, B, M	×	Sys	ILP	PD	WT
Liang et al. (2023)	✓	D	S	M	✓	-	DP	IC	WT, EN
Chang et al. (2024)	×	S	S	T, B	×	-	Sim	VR, ED, DL	WT, VT
Ma et al. (2024)	×	D	S	B	×	-	MINLP	VR, ED, DL	WT, VT, OC
Hao et al. (2024)	×	D	S	M	×	-	RO	IC, ED, TR	OC, EN
Wang et al. (2024)	✓	D	S	T, B	×	Ind	RO, RH	ED, DL, PD	WT, VT, OC, EN
Su et al. (2024, 2025)	✓	D	D	T, B	×	Ind	Sim	ED, PD	WT, VT, OC
Jiang et al. (2025)	×	D	S	B	×	-	ILP	VR, ED, PD	VT, EN
Wang and Jin (2025)	×	D	S	M	×	-	MINLP	TR	VT, EN
Zhang et al. (2025)	✓	D	D	B	✓	-	MILP, RH	ED, PD	WT, OC
This paper	✓	D	D	T, B, M	✓	Mod	MARL	ED, IC, PD	WT, OC, EN

2.2. Dynamic evaluation for evacuation planning

Traditional evacuation indicators often focus on the overall evacuation effectiveness. In our proposed online evacuation problem, which updates evacuation plans dynamically at each time step, resilience is required to evaluate not only the overall effectiveness but also the contribution of each plan at each time step. Quantification of resilience under various interventions and abnormal conditions has been studied. Studies such as Jin et al. (2014) and Tang et al. (2021) quantified the resilience of metro network under bus bridging strategies using the difference between satisfied demand and total demand. Bešinović et al. (2022) quantified the resilience of railway network under infrastructure maintenance and train rescheduling strategies using the demand satisfaction as an indicator as well. Though the "resilience triangle" has been well demonstrated by these quantification methods, the "4R" metrics have not been explicitly analyzed in their frameworks. The metrics of resilience have been addressed separately in recent studies. Li et al. (2023) addressed the system's robustness by focusing on the peak crowdedness in the station during rail disruptions. It was indicated by the maximum number of stranded passengers. However, this indicator emphasized only the worst-case scenario conditions, overlooking scenarios where crowdedness remained at a moderately high level for an extended period, which could also pose significant

challenges. In the bus system, Wang et al. (2025) addressed the system's rapidity of OPF by focusing on passengers' extra waiting time. However, the risk of overcrowding, as addressed by Li et al. (2023), was not considered.

Resilience, as a multi-metric framework, comprehensively reflects the severity of the worst-case condition (robustness), the speed of recovery (rapidity), and the extent of recovery (resourcefulness) (Bruneau et al., 2003). The evacuation mode, evacuation indicators, and resilience metrics used in the above studies are summarized in Table 2. Overall, existing evaluation methods rarely integrate the resilience triangle and metrics within a unified framework for evaluating dynamic interventions (refer to evacuation in our study) in OPF scenarios

Table 2: Summary of dynamic evaluation for evacuation planning

Paper	Mode	Evacuation indicator	Resilience metrics
Jin et al. (2014)	M	Demand satisfaction	Resilience triangle
Tang et al. (2021)	M	Demand satisfaction	Resilience triangle
Bešinović et al. (2022)	R	Demand satisfaction	Resilience triangle
Li et al. (2023)	M	Maximum stranded passengers	Robustness
Wang et al. (2025)	B	Passengers' waiting time	Rapidity
This paper	T, B, M	Demand satisfaction, Passengers' waiting time, Maximum stranded passengers	Resilience triangle, robustness, rapidity, resourcefulness

Notation: T-taxi, B-bus, M-metro, R-railway

3. Problem statement

The online multi-modal evacuation strategy is explained in Section 3.1, and the framework of resilience for OPF evacuation is defined in Section 3.2.

3.1. Online multi-modal evacuation for outburst passenger flows

The OPF evacuation is addressed in online settings by multi-modal coordination. The time horizon is discretized into time steps where an evacuation plan is provided at each interval Δt (Note that Δt should be relatively short, e.g., 5 minutes, to maintain synchronization with the real-world scenario). Passengers stranded in the OPF area require transit to various destinations scattered across the city. To handle this scattered passenger demand pattern, multi-modal transits are dispatched coordinately with each mode having a specific dispatching strategy and operational characteristics. Since our strategy aims to use redundant capacities from on-duty services, the impact of evacuation on regular services is explicitly considered. Two coordination dispatching strategies are proposed to facilitate multi-modal evacuation. The primary objective is to enhance the resilience under OPF and minimize the impact on regular services. To dynamically evaluate evacuation plans throughout the OPF period, a resilience framework is defined in Section 3.2.

3.1.1. Dispatching strategies and operational characteristics for multi-modal evacuation

To accommodate different passenger demands, the dispatching strategy and operational characteristics of each mode are demonstrated in Fig. 1 and described as follows:

- Taxis, with limited capacity and flexible routing options, are ideal for short-distance door-to-door service. As illustrated in Fig. 1, each taxi directly transits passengers (orange lines) to their destinations (black circles). It is needed to determine the number of taxis to be dispatched for evacuation.
- Buses, with moderate capacity, provide adaptable route-based coverage. As shown in Fig. 1, passengers, whose destinations (black circles) are located in a nearby geographic area, need to be aggregated onto one emergency route. These emergence routes (blue lines) directly transport passengers to a predefined terminal station (blue points) with each terminal station serving a specific geographic area (blue dashed circles). This strategy aligns with established methodologies, such as the spatial cluttering method described in Ma et al. (2024). The last-mile trip costs (black dashed lines), measured by the distance between bus terminal station and passengers' destinations, are considered in our model to capture passengers' disutility due to transfers. As buses are assigned to the emergency route, the buses that are stationed at the depot before they commence the next operation are deployed to avoid leaving on-board regular passengers stranded midway. Decisions are made in two steps: a) the number of buses to be dispatched; b) the bus assignment among emergency routes. This requires two key decisions: first determining how many buses to dispatch, then assigning them among emergency routes. To maintain proper bus headways, at most one bus is assigned to each route at each time step, as the evacuation plan is updated every Δt.

- Metro systems serve as high-capacity backbones for rapid evacuation but are constrained by fixed infrastructure, including rail lines, stations, and train frequency (shown as green lines and points). Metro passengers are assumed to disembark at the station nearest to their destinations. Their last-mile travel costs (black dashed lines) are calculated based on the distance between their disembarkation stations (green points) and final destinations (black circles). Passenger inflow at the metro station within the OPF area is the focus of this paper.

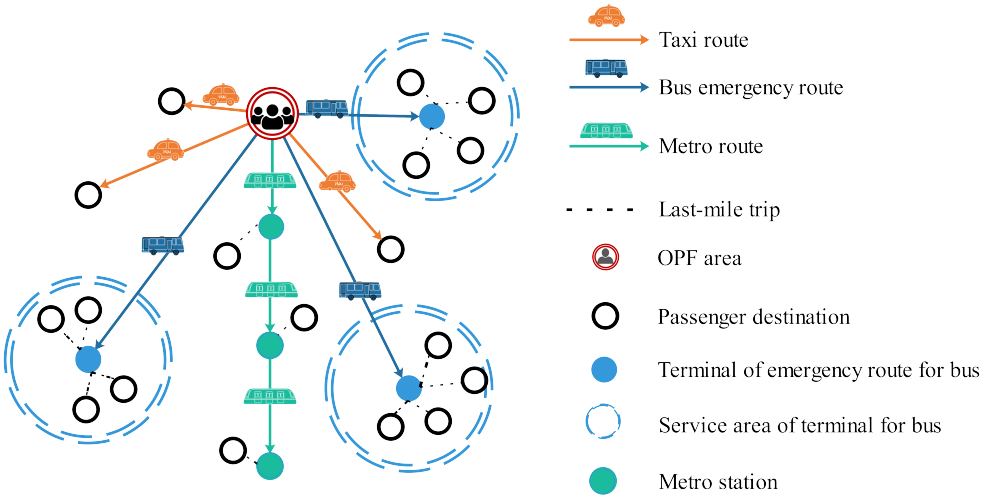


Figure 1: Characteristics of multi-modal evacuation

While this paper does not optimize bus routing and stopping patterns, our approach can accommodate flexible routing schemes by incorporating predefined candidate emergency routes. Such routes could be determined by solving the bus routing problem, as shown in existing studies like Zhu et al. (2024).

3.1.2. Impact of evacuation on regular services

To explicitly evaluate the impact of evacuation on regular services, the consequence of dispatching on-duty vehicles (for taxis and buses) and additional metro inflow is reflected by the number of passengers who are denied boarding (Ma et al., 2019), which is referred to as passenger abandonment hereafter. This kind of abandonment is induced differently for the three modes. The considerations are as follows, and the corresponding quantitative methods will be described in Section 5.2.1.

- Unoccupied taxis are identified within a certain area around the OPF, similar to the definition of searching area in Su et al. (2024). Dispatching taxis from a certain area induces a lack of vehicles for other passengers within this area. Regular passengers located outside the OPF area, whose taxis are dispatched for evacuation, are considered as abandoned.
- Buses are dispatched from their original routes. Thus, regular passengers heading to the dispatched buses at all stations along the bus route are deemed as abandoned.
- Metro trains become highly occupied due to the additional inflow at the station within the OPF area. Regular passengers at the downstream stations will be denied boarding if the arriving train is already at full capacity. Those passengers are deemed as abandoned in the metro system.

3.1.3. Coordination strategies

To coordinate multi-modal capacity dispatch, two strategies are developed to address passengers' mode-choice dynamics. While evacuees naturally exhibit mode preferences during waiting periods (their original mode choice), these preferences can be identified in real time through mobile data analysis within each mode's designated waiting zones (Zhong et al., 2017). The following coordination strategies are proposed based on distinct treatments of these mode-choice patterns:

- Independent strategy: Each mode handles its respective demand, with passengers adhering to their original mode choices and not allowed to shift modes while waiting. This restriction is similar to the assumptions in single-modal evacuation, as reviewed in Section 2.1, where passenger demand is fully managed within each mode without considering mode-shifting behavior. For practical implementation, this strategy makes the demand for each mode more predictable. However, it may lead to longer waiting times for passengers in the mode with significant imbalance between capacity and demand. This strategy will be facilitated by an MARL framework, as described in Section 5.3.1, and trained with a DTDE algorithm, as outlined in Section 6.1.

- Collaborative strategy: Each mode dispatches capacity based on the overall passenger demand within the OPF area. Passengers can dynamically shift modes depending on their original mode's available capacity at each time step and the mode-specific trip costs for each mode. If the passenger's original mode is fully occupied and alternative modes still have residual capacity at the time step, he/she is allowed to perform a mode shift based on a probability model, which considers the trip cost of each mode. If no residual capacity exists in any modes, passengers continue to wait for evacuation at the next time step (only Δt minutes later). For practical implementation, when any mode faces high demand, passengers can shift to less busy modes to reduce the stranding duration. This strategy will be facilitated by another MARL framework, as described in Section 5.3.2, and trained with an H-CTDE algorithm, as outlined in Section 6.2.

3.1.4. Assumptions

Five critical assumptions underpin our problem formulation.

A1. Self-evacuation, such as by active transportation and regular buses, is not considered. Taxi, bus, and metro systems are considered major modes for large-scale evacuation in urban areas.

A2. All stranded passengers queue at the waiting zone of each mode within the OPF area. Passengers are served following the first-come-first-served principle.

A3. Dispatched buses and taxis operate on a one-way route rather than a round trip.

A4. Passengers are relieved from overcrowding once they are assigned to a suitable transit mode.

A5. To minimize the impact of evacuation on regular services, train rescheduling and inflow control in other stations outside the OPF area are not considered. Running time and dwell time of trains are fixed following the timetable.

Assumption A1 defines the scope of stranded passengers to be evacuated. This paper excludes self-evacuation, meaning that our approach is responsible for evacuating all stranded passengers who have entered the OPF area. This assumption isolates the demand in the OPF area from the regular transit system, which mitigates the uncertainty of passenger dynamics. Taxi, bus, and metro systems are considered major modes due to their accessibility, flexibility, and complementary strengths in balancing capacity and coverage. Assumption A2 assumes that passengers follow their order of arrival, consistent with assumptions commonly made in transit-based evacuation studies (Ma et al., 2024; Su et al., 2024). In some practical emergencies, vulnerable populations may require prioritization. This can be addressed by modifying the queuing principle in the model to incorporate prioritization based on factors such as age, gender, or other relevant criteria. The model is described in Section 5.3.1, and an example illustrating how to adjust the queuing principle is provided in the Appendix C. Assumption A3 is valid for large-scale evacuations, similar to the studies like Liu et al. (2022) and Jiang et al. (2022). In scenarios where passengers' destinations are dispersed throughout the city, most routes involve long-distance travel. Unlike short-distance bridging services such as Zhang et al. (2025), the circulation of buses and taxis is not considered in this study. Assumption A4 excludes passengers' boarding time from the model. This assumption is reasonable when boarding efficiency is high. As stated in Assumption A2, passengers are already queuing in the designated waiting zones of each mode, and the mixing of passenger flows is neglected, further supporting the validity of this simplification. Similar assumptions have also been adopted in other studies, such as Zhang et al. (2021) and Zhu et al. (2024). Assumption A5 limits the scope of emergency response strategies in the metro system, which is commonly adopted in metro passenger inflow control studies, like Zhang et al. (2021); Liang et al. (2023) and Jiang et al. (2022). Train operations are typically considered robust enough to handle the oversaturated passenger flows without rescheduling.

3.2. Resilience framework for outburst passenger flows evacuation

To dynamically evaluate the effectiveness of evacuation at each time step, a resilience framework is defined for OPF evacuation. To quantify the OPF within this resilience framework, "capacity deficiency" is defined as the result of the mismatch between capacity (represented by satisfied demand) and demand (represented by total passenger demand). To diagram the "resilience triangle", the capacity deficiency is calculated as a negative value, shown in Eq. (1).

$$\text{Capacity deficiency} = \text{Satisfied demand} - \text{Total demand} \tag{1}$$

The absolute value of this deficiency indicates the number of stranded passengers who remain unsatisfied in the crowded area.

To demonstrate the resilience framework indicated by capacity deficiency, two example curves under OPF are shown in Fig 2. The grey curve represents the case without evacuation, serving as a benchmark, while the blue curve represents the case with a well-designed evacuation. An overcrowding threshold (red dashed line) is outlined, representing the safety limit for the number of stranded passengers. It serves as a trigger for initiating evacuation planning, which is also known as an alarm mechanism in other emergency management studies (Abdelgawad and Abdulhai, 2010). Under normal conditions (black curve), capacity deficiency is near zero, indicating an equilibrium state with only a few stranded passengers. When OPF occurs at t_0, the capacity deficiency intensifies, leading to a rise in the number of stranded passengers. The evacuation measures are activated at t_{start} when the number of stranded passengers surpasses the threshold. By comparison of the two curves, a well-designed evacuation approach (blue curve) can effectively reduce the maximum number of stranded passengers, shorten the duration of overcrowding, and accelerate the restoration of capacity-demand equilibrium, thereby enhancing resilience under OPF.

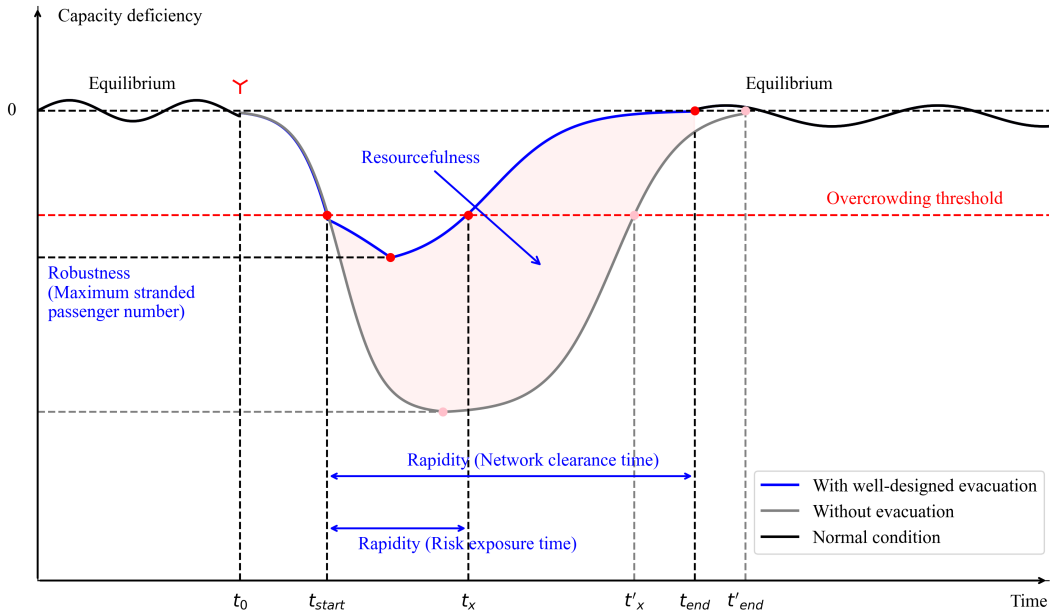


Figure 2: Resilience framework of OPF evacuation

To develop the resilience framework for evacuation, the "4R" metrics need to be defined in alignment with existing evacuation metrics (including maximum stranded passenger number, network clearance time and risk exposure time). Note that as redundancy is typically enhanced during pre-event stage (Xu et al., 2021), robustness, rapidity, and resourcefulness are emphasized in this paper which aims to enhance the resilience during the evacuation process. For the definition of resilience metrics, please refer to prominent review articles (Zhou et al., 2019; Gu et al., 2020). The labels of resilience metrics in Fig. 2 follow the concepts provided in these reviews. Overall, the definitions of resilience metrics used to evaluate the evacuation process are as follows.

- Robustness captures performance under worst-case conditions, measured by the maximum number of stranded passengers. To ensure passenger safety, this metric should remain near or below the overcrowding threshold. Enhancing robustness involves minimizing the maximum number of stranded passengers during disruptions.
- Rapidity reflects the total duration of the abnormal condition caused by the OPF. Two time spans require attention: (1) The time spans from the onset time t_{start} to the time of equilibrium restoration (t_{end} for the case with well-designed evacuation and t'_{end} for the case without evacuation). It aligns with the concept of network clearance time as an evacuation indicator; (2) the time spans from the onset time to the dissipation of overcrowding (t_x for the case with a well-designed evacuation and t'_x for the case without evacuation). This corresponds to the concept of risk exposure time. Both time spans can be shortened by minimizing the stranded duration for each passenger during evacuation planning.
- Resourcefulness measures the effectiveness of evacuation strategies in enhancing system resilience,

quantified by the performance gap between scenarios with and without evacuation (represented by the area between their respective curves). This metric can be improved by simultaneously maximizing demand satisfaction and minimizing stranded passenger durations during evacuation.

4. Mathematical model

To enhance the resilience under OPF and minimize the impact of evacuation on regular services, an optimization model is required to consider both objectives during OPF evacuation. The decision variables are the dispatched capacity of each mode $m \in M = \{\text{taxi}, \text{bus}, \text{metro}\}$ at each time step $t \in T$, indicated by a_m^t. Among them, a_{taxi}^t and a_{bus}^t represent the number of dispatched taxis and buses, receptively, while a_{metro}^t represents the dispatched number for passenger inflow at the metro station within the OPF area (indicating the capacity of the metro system to accommodate stranded passengers). Thus, the objective function of the entire period can be generally expressed as Eq. (2),

$$\max R = \sum_{t \in T} \sum_{m \in M} \left(R_{\text{resilience}}^{m,t}(a_m^t) - R_{\text{abandon}}^{m,t}(a_m^t) \right), \tag{2}$$

where R indicates the overall objective of the three modes throughout the OPF period. $R_{\text{resilience}}^{m,t}(a_m^t)$ is a general representation for the objective of resilience enhancement, which is transformed into demand-responsive feedback for each mode at each time step in Section 5.4. $R_{\text{abandon}}^{m,t}(a_m^t)$ is a general representation for the objective of the impact mitigation of evacuation on regular services, captured by the number of regular passengers abandoned by each mode at each time step. Abandoned regular passengers are variables related to the dispatched capacities, which will be described in Section 5.2.1.

To formulate the constraints related to capacity dispatching, several parameters need to be defined. Let n_m^t indicate the number of dispatchable capacities of mode m at time step t. Let c_m indicate the capacity limit of the vehicle in mode m (taxi, bus or train). To formulate the constraints related to passenger distribution, several parameters and auxiliary variables need to be defined. For practical considerations, some individual passengers actually travel together, such as family members. Therefore, let p index a passenger group who shares the same entry time, leaving time, mode choice and destination. Recall from Section 3.1.3 that passengers have mode choices while waiting for evacuation, where the derivation of their original mode choices will be described in Section 5.2.2 and their mode-shifting mechanism will be described in Section 5.3.2. Let P_m^t denote the set of passenger groups that are evacuated by mode m at time step t. Let $q(p)$ indicate the number of passengers in group p where $q(p) \geq 1$. In the case of a single passenger with a specific entry time, leaving time and destination, $q(p) = 1$. Recall that, in the bus system, passengers are distributed across different emergency routes. Let r index a specific route, γ indicate the set of all emergency routes, and γ^t indicate the set of emergency routes with a bus assigned at time step t. Let $P_{\text{bus},r}^t$ indicate the set of passenger groups which are distributed to route r at time step t. Values of parameters and variables are derived from the data-driven agent-based environment, as described in Section 5.2. Thus, the constraints can be expressed as follows:

$$0 \leq a_m^t \leq n_m^t, \quad \forall m \in M, t \in T, \tag{3}$$

$$\sum_{p \in P_{\text{taxi}}^t} q(p) \leq a_{\text{taxi}}^t c_{\text{taxi}}, \quad \forall t \in T, \tag{4}$$

$$\sum_{p \in P_{\text{bus},r}^t} q(p) \leq c_{\text{bus}}, \quad \forall t \in T, r \in \gamma^t, \tag{5}$$

$$|\gamma^t| = a_{\text{bus}}^t, \quad \forall t \in T, \tag{6}$$

$$\sum_{p \in P_{\text{metro}}^t} q(p) \leq a_{\text{metro}}^t, \quad \forall t \in T, \tag{7}$$

$$a_m^t \in \mathbb{N}, \quad \forall m \in M, t \in T. \tag{8}$$

Constraint (3) ensures that the dispatched capacity of each mode at each time step cannot exceed the dispatchable capacity. Constraint (4) ensures that the number of passengers evacuated by taxis cannot exceed the dispatched taxi capacity. Constraint (5) ensures that the number of passengers distributed to each bus route cannot exceed the bus capacity. Recall that each bus route dispatches at most one bus at

time step t. Constraint (6) indicates that the number of routes with a bus assigned equals the number of buses dispatched. Constraint (7) ensures that the number of inflows at the metro station cannot exceed the dispatched number. Constraint (8) indicates that the decision variable of dispatched capacity per mode must be a natural number.

Proposition 1. If the number of routes with assigned buses equals the total number of dispatched buses, then each route must be assigned exactly one bus.

Proof. Let $r(v) \in \gamma$ denote the route assigned to the dispatched bus v. The set of dispatched buses is denoted by $\bar{V}^t_{bus} = \{v_1, \ldots, v_n\}$, where $|\bar{V}^t_{bus}| = a^t_{bus}$. The set of routes with buses assigned is denoted by $\gamma^t = \{r(v_1), \ldots, r(v_n)\}$, where $|\gamma^t| \leq a^t_{bus}$ because there could be duplicate routes within the set γ^t. Only if each route is assigned exactly one bus, then $|\gamma^t| = a^t_{bus}$; otherwise, $|\gamma^t| < a^t_{bus}$.

The formulated problem is a dynamic combinatorial resource allocation problem with discrete decision variables and multi-stage temporal dependencies. These characteristics render the problem NP-hard and intractable for exact algorithms under large-scale instances. Especially, each passenger group's entry time, leaving time and destination are incorporated into the model, and track the passenger growth and capacity availability based on real-time datasets. Therefore, a reinforcement learning-based approach is developed to effectively adapt evacuation plans over time (Sutton and Barto, 2018). Decision variables a^t_m will be determined by our MARL agents with either independent or collaborative strategy in Sections 5.3.1 and 5.3.2, respectively.

All notations used in this paper are organized into five appendix tables. Decision variables are listed in Table A1, followed by objectives and feedback in Table A2, indices and sets in Table A3, modeling parameters in Table A4, and training parameters in Table A5.

5. Multi-agent reinforcement learning formulation

First, our MARL framework for online multi-modal evacuation is introduced in Section 5.1. Next, the establishment of our data-driven agent-based environment is introduced in Section 5.2. Then, two agents are formulated to respectively implement the independent and the collaborative strategies in Section 5.3. Finally, the objectives are transformed into demand-responsive feedback in Section 5.4.

5.1. Multi-agent reinforcement learning framework

The framework of our MARL is given in Fig. 3. A data-driven agent-based environment (blue box) consists of three dynamic transit environments (taxi, bus, metro) and a dynamic passenger environment. During an OPF event, real-time datasets—including taxi GPS, bus AVL, smart card transactions and metro time-dependent OD demand datasets—are processed to generate dynamic transit environments. These environments continuously track: (1) dispatchable capacities, and (2) regular passenger number for each mode. The mobile dataset is processed to generate the dynamic passenger environment. This environment tracks the conditions of stranded passengers within the OPF area. The MARL agents (white box) receive these capacity and demand metrics as state variables, enabling coordinated dispatch of taxis, buses, and metro inflows while optimizing passenger distribution across all modes.

Given agents' actions in capacity dispatch and passenger distribution (orange arrows), the environment updates accordingly (orange loop) and provides feedback to agents (green arrows). The feedback includes the resilience enhancement and impact of evacuation on regular services (green boxes), corresponding to the objective function in Eq. (2). As analyzed in Section 3.2, resilience metrics, including robustness, rapidity and resourcefulness, can be enhanced by different evacuation indicators. The framework of transformation relation between objectives and feedback at each time step is shown in Fig. 4. Since the number and stranded duration of evacuated passenger groups both depend on the results of passenger distribution of evacuation plans, an integrated feedback, referred to as demand satisfaction, is proposed. The resilience enhancement is driven by two demand-responsive feedback mechanisms: (1) overcrowding monitoring and (2) demand satisfaction evaluation, while the impact of evacuation on regular services is quantified through the abandoned passenger number.

5.2. Data-driven agent-based environment

Data-driven agent-based environments aim to track the evolving condition of capacity availability and passenger mobility, and evaluate the consequences of the agents' actions on the environments. Three dynamic transit environments are designed with the elements of vehicles and regular passengers in taxi, bus and metro systems, respectively, while a dynamic passenger environment is created with the elements of stranded passengers.

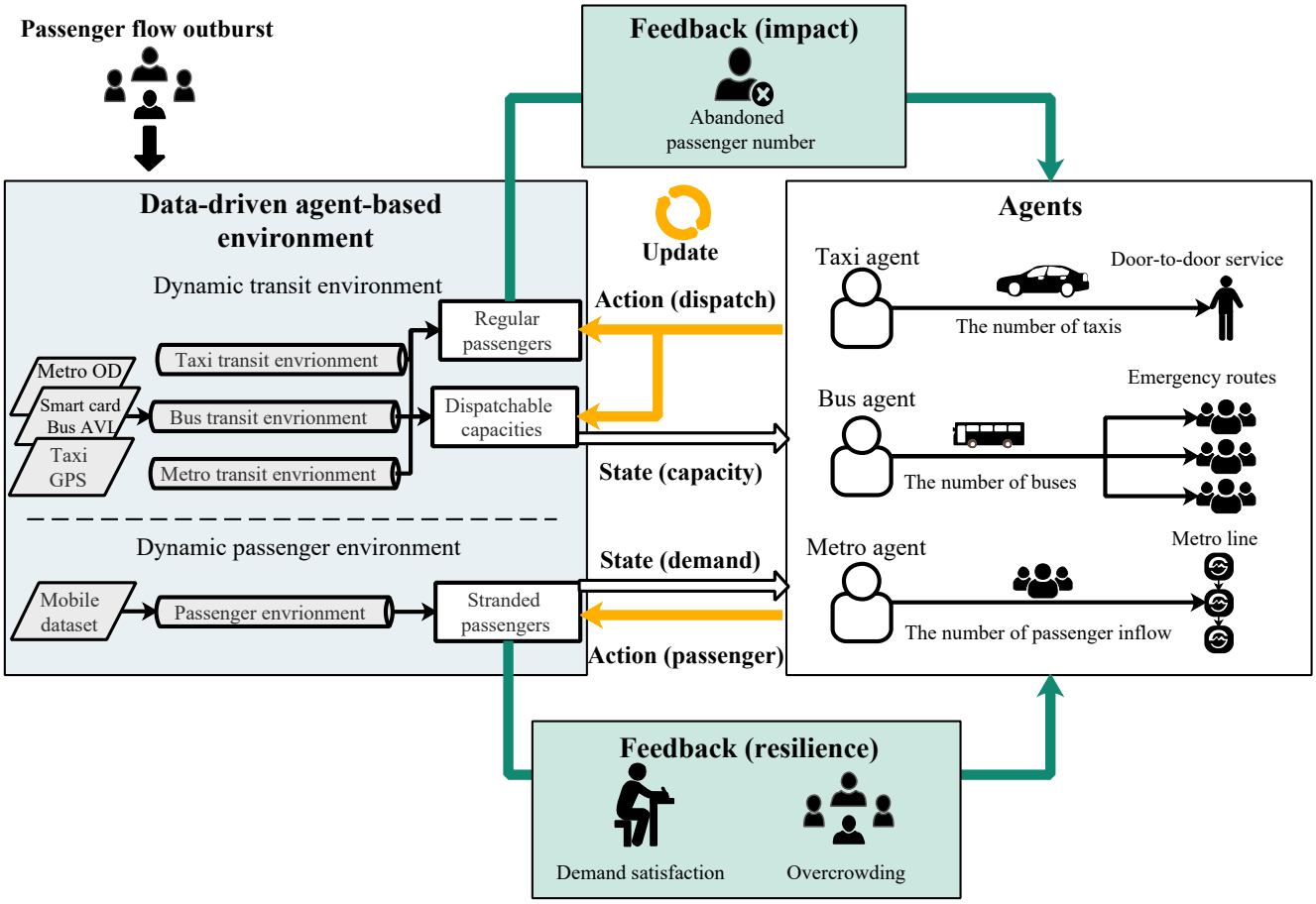


Figure 3: Framework of MARL

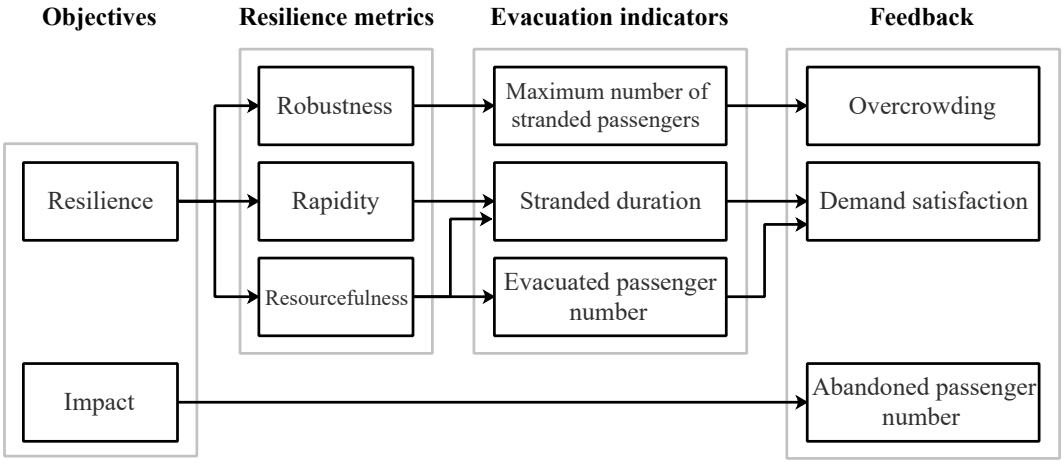


Figure 4: Framework of the feedback

The environment in this section refers to the offline training environment, constructed using historical datasets with complete information on capacities and passenger demand. However, passengers leave following the agent-determined evacuation plan rather than their original travel patterns. This setup simulates real-world feedback while training agents to iteratively optimize their decision-making. To ensure a seamless online application, agents are restricted to observable real-time states during the operation.

5.2.1. Dynamic transit environment

Agent-based models are established using datasets from multi-modal systems. To update the positioning and occupancy conditions of taxis, buses and trains with real-time information, the dynamic transit environment for each mode is established as follows.

(1) Taxi transit environment

The taxi transit environment is formulated based on the taxi GPS dataset. An example, including

elements of taxis and regular passengers, is illustrated in Fig. 5a. For a given searching area and time step, taxis located within the searching area, which are unoccupied at the beginning of the time step (white taxis in Fig. 5a), are deemed to be dispatchable. Let $v \in V^t_{\text{taxi}}$ indicate a specific taxi, where V^t_{taxi} indicates the set of dispatchable taxis at time step t. A detailed description of the taxi GPS dataset, and the identification of dispatchable taxis is presented in Appendix B.1. Then, the number of dispatchable taxis at time step t can be calculated as $n^t_{\text{taxi}} = |V^t_{\text{taxi}}|$, which serves as the capacity limit in Constraint (3).

For each taxi v, the GPS dataset records its ID, location, timestamp (a time instant, denoted by $\tilde{t}$, that records the moment a data point is generated and the timestamp interval, denoted by $\Delta\tilde{t}$, is significantly smaller than the time step interval, i.e., $\Delta\tilde{t} \ll \Delta t$) and occupancy (a binary indicator, denoted by $\hat{\rho}_{\text{taxi}}(v, \tilde{t})$, representing whether the taxi is occupied at each timestamp). A taxi that is originally dispatched to serve regular passengers can be revealed in the GPS dataset by checking the occupancy status $\hat{\rho}_{\text{taxi}}(v, \tilde{t})$ of the taxi v within the time step $[t, t + \Delta t]$. If a taxi's occupancy status becomes occupied within the searching area at subsequent time steps, it means that the taxi is heading to pick up a regular passenger (dashed line in Fig. 5a) in practice. When such a taxi is dispatched to evacuate stranded passengers in the OPF area, the associated regular passenger must be abandoned. To capture this, a binary indicator, $\rho_{\text{taxi}}(v)$, is introduced to indicate whether a regular passenger would be abandoned by taxi v if it is dispatched for evacuation. Specifically, the abandonment of a regular passenger by taxi v is calculated by Eq. (9).

$$\rho_{\text{taxi}}(v) = \begin{cases} 1 & \exists \hat{\rho}_{\text{taxi}}(v, \tilde{t}) = 1 \\ 0 & \forall \hat{\rho}_{\text{taxi}}(v, \tilde{t}) = 0 \end{cases}, \quad \forall t \in T, v \in V^t_{\text{taxi}}, \tilde{t} \in [t, t + \Delta t] \tag{9}$$

(2) Bus transit environment

Bus transit environment is formulated based on bus smart card transactions and the AVL dataset. An example, including elements of bus and regular passengers, is illustrated in Fig. 5b. Let $v \in V^t_{\text{bus}}$ indicate a specific bus, where V^t_{bus} indicates the set of dispatchable buses at time step t. The bus smart card dataset records each regular passenger's ID (denoted by $\hat{p}$), station code, bus ID (denoted by $v(\hat{p})$) and timestamp, and the bus AVL dataset records buses' IDs, location and timestamps. Recall that our strategy only dispatches buses from the depot, as described in Section 3.1.1. For a given depot (located in the upper-left corner of Fig. 5b) and a time step, the bus, which is located at the depot, can be identified via the AVL data. A detailed description of the bus smart card dataset, and the identification of dispatchable buses is presented in Appendix B.2. The number of dispatchable buses at time step t can be calculated by $n^t_{\text{bus}} = |V^t_{\text{bus}}|$, which serves as the capacity limit in Constraint (3).

The impact of dispatching an on-duty bus for evacuation can be revealed in the smart card dataset by checking the bus ID and calculating the number of regular passengers associated with the bus (dashed lines in Fig. 5b). Let $\rho_{\text{bus}}(v) \in \mathbb{N}$ indicate the number of abandoned regular passengers associated with a bus when it is dispatched. It is calculated by Eq. (10).

$$\rho_{\text{bus}}(v) = |\{\hat{p} | v(\hat{p}) = v\}|, \quad \forall t \in T, v \in V^t_{\text{bus}} \tag{10}$$

(3) Metro transit environment

The Metro transit environment is formulated based on a time-dependent OD demand matrix. An example, including elements of trains and regular passengers, is illustrated in Fig. 5c. Time-dependent OD demand matrix records the regular passenger demand between any two stations within a time interval. A detailed description of the time-dependent OD demand dataset is presented in Appendix B.3. Given a station, denoted by i^*, within the OPF area (the station within the white circle in Fig. 5c), let $v_t \in V_{\text{metro}}$ denote the train arriving at station i^* at time step t, where V_{metro} denotes the set of trains. Under the fourth assumption in Section 3.1.4, where the train's running and dwell time are fixed, the demand for each train at a given time step can be derived (Liang et al., 2023). Then, let $w(v_t, i, j)$ denote the demand on train v_t from station i to j, where $i, j \in I$. I is the set of metro stations. Specifically, let $\underline{i}, \bar{i}$ denote the origin and terminal stations of the train, respectively. For a given time step t, the available capacity

on train v_t arriving at station i, denoted by $n^{t,i}_{\text{metro}}$, is determined by Eq. (11),

$$n^{t,i}_{\text{metro}} = c_{\text{metro}} - \sum_{i'=\underline{i}}^{i-1} \sum_{j=i+1}^{\bar{i}} w(v_t, i', j), \quad \forall t \in T, i \in I, \tag{11}$$

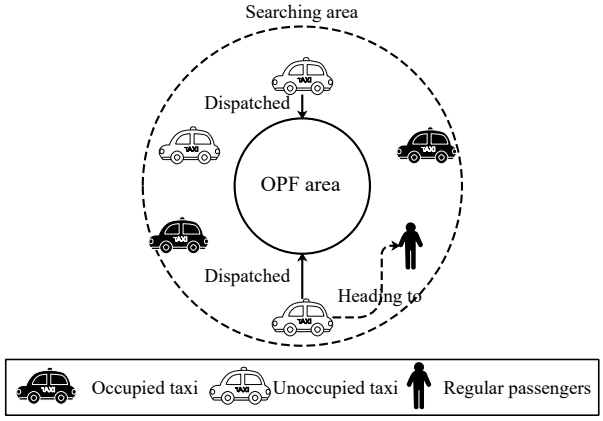
where c_{metro} is the capacity of an empty train. The summation calculates the number of passengers who board upstream of station i and disembark at downstream stations.

The available capacity at the OPF station i^* can also be obtained by Eq. (11). To maintain the generality with the mode of taxi and bus, let n^t_{metro} to indicate the dispatchable capacity at station i^*, which serves as the capacity limit in Constraint (3).

When additional passenger flows enter station i^*, the train may become fully occupied, resulting in regular passengers at downstream stations being denied boarding. For any downstream station i, the number of abandoned passengers is denoted by $\rho_{\text{metro}}(v_t, i)$, which is determined by Eq. (12).

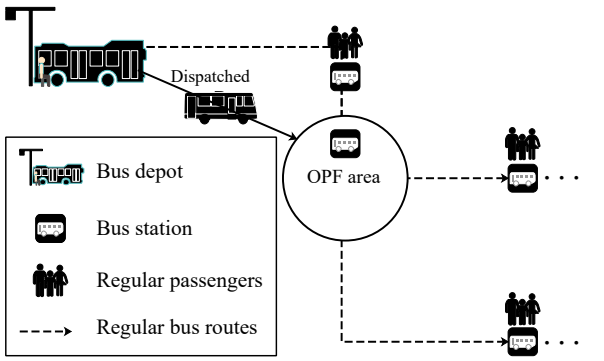
$$\rho_{\text{metro}}(v_t, i) = \max\left(\sum_{j=i+1}^{\bar{i}} w(v_t, i, j) - n^{t,i}_{\text{metro}}, 0\right), \quad \forall t \in T, i \in \{I | i^* < i < \bar{i}\} \tag{12}$$

Within the maximizing function, the first summation calculates the number of regular passengers at a downstream station i, while the second term represents the available capacity $n^{t,i}_{\text{metro}}$ of train v_t arriving at this station. The maximizing function ensures that this indicator is non-negative.



Taxi information = [ID, location, timestamp, occupancy]

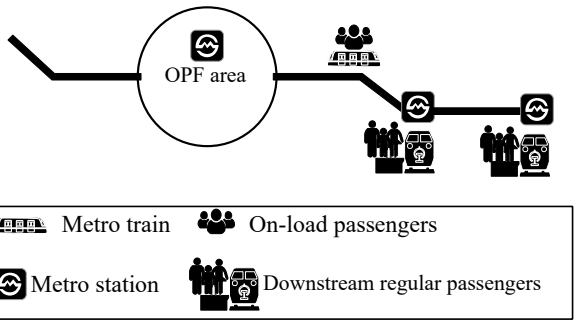
(a) Taxi transit environment



Bus information = [bus, location, timestamp]

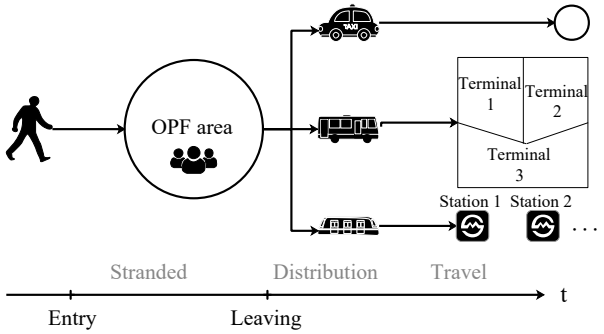
Regular passenger information = [passenger, station, bus, timestamp]

(b) Bus transit environment



Metro OD information = [OD demand on each train]

(c) Metro transit environment



Time information = [entry time, leaving time]

Passenger information = [ID, destination, distance, mode]

Preprocessed information = [passenger number, metro station]

(d) Passenger environment

Figure 5: Examples of data-driven agent-based environments

5.2.2. Dynamic passenger environment

An agent-based model is established using the mobile dataset, with each passenger as an element. The application of mobile datasets to model human mobility patterns during emergencies has been validated in several recent studies (Wang et al., 2021; Diaz et al., 2023). An example illustrating passenger elements is shown in Fig. 5d. Time and passenger information are given by the mobile dataset. The time information

includes each passenger's original entry and leaving time within the OPF area. The passenger information contains each passenger's ID, destination, travel distance, and original mode choice. A detailed description of the mobile dataset is presented in Appendix B.4.

However, two additional types of information are further required to apply our strategy in practice:

- As described in Section 4, passengers are classified based on their original entry time, leaving time, mode choices and destinations to evacuate them as a passenger group. The number of passengers within each passenger group is referred to as $q(p)$.
- To calculate the occupancy level of a metro train, it is essential to know the passengers' disembarkation station if they take the metro train. Based on each passenger's trajectory (namely, the destination and travel distance here), the disembarkation station can be estimated by matching the trajectory with the geographic locations of metro stations. This approach is similar to the method used in Huang et al. (2024), where passengers are mapped to specific stations based on their holding locations along the metro line. In this case, let $i(p)$ to denote the estimated disembarkation station for a passenger group p if it takes the metro train.

For a given time step t, passengers, whose entry time falls within the time step, are considered as entering and becoming stranded in the OPF area. Let $\bar{P}^t$ indicate the set of stranded passengers at time step t, which collects the stranded passenger groups $p \in \bar{P}^t$. The method for identifying stranded passengers is presented in Appendix B.4. Additionally, for each passenger group p, let $e(p)$ indicate its entry time, $\tilde{l}(p)$ indicate its original leaving time, $d(p)$ indicate its destination, $k(p)$ indicate its travel distance, $\tilde{m}(p)$ indicate its original mode choice. These parameters are essential for passenger distribution in Section 5.3.1 and evacuation evaluation in Section 5.4.

5.3. Agents

To facilitate the online multi-modal evacuation with the corresponding strategies described in Section 3.1.3, the formulation of agents under the independent strategy is introduced in Section 5.3.1 and that under collaborative strategy is introduced in Section 5.3.2.

5.3.1. Independent strategy: Demand splitting and independent dispatching

In our first coordination strategy, it is assumed that passengers always follow their original mode choice without shifting to alternative modes. Then, the stranded passengers can be split by their mode choice as: $\bar{P}^t_m = \{p | p \in \bar{P}^t, \tilde{m}(p) = m\}$, where $\bar{P}^t_m$ indicates the set of stranded passenger groups in mode m at time step t. In this setup, each agent can independently dispatch capacities to evacuate the stranded passengers within the respective system. The construction of the agents is designed as follows.

For each mode $m \in M$, an agent dispatches capacities based on its local observable state at each time step t. The state variables include three key components as follows:

- Dispatchable capacity n^t_m: This is derived from the dynamic transit environment of each mode, which quantifies the available capability to evacuate passengers.
- Stranded passenger number δ^t_m: This is calculated based on the set of stranded passenger groups $p \in \bar{P}^t_m$ by $\delta^t_m = \sum_{p \in \bar{P}^t_m} q(p)$, which reflects the urgency of the evacuation at current time step t.
- Current time step t: By including the current time as a component, the agent is expected to balance the dispatching decisions over time.

Therefore, the state variables for each mode m at time step t are formulated as $S^t_m = [n^t_m, \delta^t_m, t]$.

The agent's action, denoted as a^t_m, follows constraints of the mathematical model in Section 4, and updates the environment following the mechanism outlined below.

- For the taxi agent, its action a^t_{taxi} determines the number of taxis to be dispatched at time step t, respecting Constraint (3). Under the second assumption in Section 3.1.4, taxis operate on the one-way route. Thus, the dispatched taxis are collected in a set $\bar{V}^t_{\text{taxi}}$, which are prevented from being dispatched in subsequent time steps.
- For the bus agent, its action a^t_{bus} determines the number of buses to be dispatched at time step t, respecting Constraint (3). The dispatched buses at time step t are collected in a set $\bar{V}^t_{\text{bus}}$, which are prevented from being dispatched in subsequent time steps. Each route at most receives one bus, subject to Constraint (6). When the number of available buses is insufficient to serve all emergency routes, routes with the highest passenger demand are prioritized.
- For the metro agent, its action a^t_{metro} determines the number of passenger inflow at time step t, respecting Constraints (3) and (7).

Given the agent's action for each mode, stranded passengers are distributed to dispatched capacities based on their original mode choices. The set of evacuated passenger groups, denoted by P_m^t, collects passenger groups who are evacuated by mode m at time step t. The passenger distribution follows the mechanisms outlined below.

- For the taxi mode, stranded passengers can be assigned to the dispatched taxis until the capacity limit is reached, as specified by Constraint (4). Recall the second assumption that passengers follow the first-come-first-served principle, and remaining passengers wait for the evacuation plan in subsequent time steps. The set of passenger groups evacuated by taxi is represented by Eq. (13). Specifically, the selection aims to minimize the latest entry time among all evacuated passenger groups, i.e., the maximum $e(p)$ within the selected subset P under the Constraint (4).

$$P_{\text{taxi}}^t = \arg\min_{P \subseteq \bar{P}_{\text{taxi}}^t} \left\{ \max_{p \in P} e(p) | \sum_{p \in P} q(p) \leq a_{\text{taxi}}^t c_{\text{taxi}} \right\}, \quad \forall t \in T \tag{13}$$

- For the bus mode, stranded passengers whose destinations are served by a bus can be evacuated until the capacity limit is reached, as specified by Constraint (5). Remaining passengers wait for the evacuation plan in subsequent time steps. The set of passenger groups on bus emergency route r is represented by Eq. (14),

$$P_{\text{bus},r}^t = \arg\min_{P \subseteq \bar{P}_{\text{bus}}^t} \left\{ \max_{p \in P} e(p) | \sum_{\substack{p \in P \\ d(p) \in D_r^t}} q(p) \leq c_{\text{bus}} \right\}, \quad \forall t \in T, r \in \gamma^t, \tag{14}$$

  and the set of passenger groups evacuated by bus at time step t is $P_{\text{bus}}^t = \bigcup_{\forall r \in \gamma^t} P_{\text{bus},r}^t$.
- For the metro mode, the set of passenger groups evacuated by metro, which follows Constraint (7), is represented by Eq. (15).

$$P_{\text{metro}}^t = \arg\min_{P \subseteq \bar{P}_{\text{metro}}^t} \left\{ \max_{p \in P} e(p) | \sum_{p \in P} q(p) \leq a_{\text{metro}}^t \right\}, \quad \forall t \in T \tag{15}$$

Although the model is established based on the first-come-first-served principle, alternative queuing rules for vulnerable populations can be incorporated when additional information about stranded passengers is available. The implementation is provided in Appendix C.

Finally, the evacuated passenger groups in the set P_m^t are removed from the set of stranded passenger groups, representing their relief from the dynamic passenger environment, i.e., $\bar{P}_m^{t+1} \leftarrow \bar{P}_m^t - P_m^t$.

Note that some residual capacities exist in some modes after passenger distribution when the number of stranded passengers is smaller than dispatched capacities at time step t. In the independent strategy, these residual capacities are released back into the dynamic transit environments, as passengers in other modes are not considered for using these capacities.

5.3.2. Collaborative strategy: Passengers' mode shifting and collaborative dispatching

In our second coordination strategy, the assumption that passengers will only take their original modes as restricted in the independent strategy is relaxed. Instead, a passenger mode-shifting mechanism is proposed that considers both their original mode choice and mode-shifting behavior. The residual capacities released in the independent strategy are utilized in the collaborative strategy. Accordingly, agents are further designed as follows.

For each mode $m \in M$, an agent's state variables include the dispatchable capacity of each mode n_m^t, the total stranded passenger number δ^t where $\delta^t = \sum_{p \in \bar{P}^t} q(p)$, and the current time step t. Therefore, the state variables at time step t are formulated as $S_m'^t = [n_m^t, \delta^t, t]$. Note that the total stranded passenger number δ^t provides global observation, stimulating agents with higher dispatchable capacities to allocate more capacities rather than solely handling their own demands.

Action for each agent a_m^t remains the same as that in the independent strategy, which is to determine the number of capacities to be dispatched for each mode m at time step t.

Given agent's action for each mode, stranded passengers are initially distributed according to their original mode choice, as outlined in Section 5.3.1. Let a set $\hat{P}_m^t$ denote the set of initial evacuated passenger groups that collects passenger groups that are distributed with respect to their original mode choice. If the dispatched capacities exceed the number of passengers after the initial distribution, the residual capacity is then allocated to passengers whose mode-shifting choice aligns with this mode. The residual capacity for each mode can be checked as follows:

- For taxis, the residual capacity is calculated as the difference between the dispatched taxi capacities and the number of passengers in the set of initial evacuated passenger groups. The residual capacity of taxi system is calculated by Eq. (16).

$$\hat{n}_{\text{taxi}}^t = a_{\text{taxi}}^t c_{\text{taxi}} - \sum_{p \in \hat{P}_{\text{taxi}}^t} q(p), \quad \forall t \in T \tag{16}$$

- For buses, the residual capacity on each emergency route is calculated as the difference between the maximum bus capacity and the number of onboard passengers on that route in the set of initial evacuated passenger groups. The residual capacity of bus system is calculated by Eq. (17).

$$\hat{n}_{\text{bus},r}^t = c_{\text{bus}} - \sum_{p \in \hat{P}_{\text{bus},r}^t} q(p), \quad \forall t \in T, r \in \gamma^t \tag{17}$$

- For metro trains, the residual capacity is updated based on the inflow number. The residual capacity of metro system is calculated by Eq. (18).

$$\hat{n}_{\text{metro}}^t = n_{\text{metro}}^t - a_{\text{metro}}^t, \quad \forall t \in T \tag{18}$$

Remaining passengers, whose original mode choices have reached its capacity limit at the current time step, are allowed to shift to a more efficient alternative mode with residual capacity.

As studied by Li et al. (2020a,b), passengers prefer to choose a mode with less travel time, fare cost, and inconvenience of transfer. Additionally, passengers are inclined to remain with their original mode choice, even when an efficient alternative mode exists, based on their perception of the abnormal condition. To capture this, let $u_m(p)$ indicate the trip cost for passenger group p when traveling by mode m. The cost factors include the travel time, fare and last-mile trip (capturing the inconvenience of transfer). Additionally, a mode loyalty factor, denoted by ϵ_0^m, is introduced, which reflects the passengers' willingness to remain with their original mode. Thus, the general trip cost is calculated by Eq. (19),

$$u_m(p) = \epsilon_0^m(\epsilon_1^m k(p) + \epsilon_2^m k(p) + \epsilon_3^m \hat{k}_m(p)), \quad \forall m \in M, t \in T, p \in \bar{P}_m^t, \tag{19}$$

where $\epsilon_0^m \in (0,1)$ discounts the perceived trip cost of passengers' original mode choice, while $\epsilon_0^m = 1$ is applied to modes other than the original mode. The first two terms in the parentheses represent the travel time cost and fare cost, respectively, both of which are related to passengers' travel distance $k(p)$. The third term represents the last-mile trip cost. $\epsilon_1^m, \epsilon_2^m, \epsilon_3^m$ are the coefficients that capture the unit costs on travel time, fare and last-mile trip, respectively. The last-mile trip terms, $\hat{k}_m(p)$, are formulated differently based on the characteristics of modes:

- For taxi trips, last-mile costs are not incurred, as taxis provide door-to-door service, $\hat{k}_{\text{taxi}}(p) = 0$.
- For bus trips, it should be recalled that a terminal is designated for each service area. The last-mile term $\hat{k}_{\text{bus}}(p)$ represents the distance from a passenger group's disembarking terminal station to its destination.
- For metro trips, given a disembarking metro station of a passenger group, the last-mile term $\hat{k}_{\text{metro}}(p)$ represents the distance from passenger group's disembarking metro station to its destination.

A Logit model, commonly used to represent passengers' probability in mode choice decisions (Wang et al., 2024; Su et al., 2024), is established to formulate each passenger's mode-shifting behavior. The probability is calculated by Eq. (20).

$$\mathbb{L}_m(p) = \frac{exp(1/u_m(p))}{\sum_{m' \in M} exp(1/u_{m'}(p))}, \quad \forall m \in M, t \in T, p \in \bar{P}_m^t \tag{20}$$

When a passenger group's original mode is fully occupied and alternative modes have residual capacities, the passenger group is allowed to shift modes. In our agent-based model, the mode-shifting choice is implemented by letting each passenger group perform a random sampling based on the Logit probability model. Passengers are then continuously distributed to the residual capacities following the methodology outlined in Section 5.3.1, until the capacity limit is reached or no passengers remain.

5.4. Feedback

According to the objective function in Eq. (2), the agents aim to maximize the resilience while minimizing the impact on regular service. As analysis in Section 5.1, the feedback function of resilience enhancement, denoted by $R_{resilience}^{m,t}$, for each mode m at each time step t is transformed into the combination of demand satisfaction, denoted by $R_{\text{demand}}^{m,t}$, and overcrowding, denoted by $R_{\text{crowd}}^{m,t}$. The feedback function of resilience enhancement is calculated by Eq. (21),

$$R_{\text{resilience}}^{m,t} = \alpha_1 R_{\text{demand}}^{m,t} - \alpha_2 R_{\text{crowd}}^{m,t}, \quad \forall m \in M, t \in T, \tag{21}$$

where α_1 and α_2 are the coefficients to balance the feedback of demand satisfaction and overcrowding. The overall feedback function, denoted by R_m^t, for each mode m at each time step t is the weighted sum of resilience feedback and the feedback of impact on regular services, denoted by $R_{\text{abandon}}^{m,t}$. The overall feedback function is calculated by Eq. (22),

$$R_m^t = \alpha_1 R_{\text{demand}}^{m,t} - \alpha_2 R_{\text{crowd}}^{m,t} - \alpha_3 R_{\text{abandon}}^{m,t}, \quad \forall m \in M, t \in T, \tag{22}$$

where α_3 is the coefficient to balance the feedback of impact on regular services.

5.4.1. Feedback 1: Resilience enhancement

Resilience metrics, as posterior evaluators, are not sufficient to guide agents during the OPF. Two demand-responsive feedback functions, namely penalty of overcrowding and reward of demand satisfaction, are proposed as follows.

Penalty of overcrowding is evaluated by comparing the total number of stranded passengers with a predefined overcrowding threshold η. Above this threshold, a penalty will be imposed to incentivize agents to evacuate more passengers. Since overcrowding results from the total number of stranded passengers, each mode shares the responsibility for the occurrence. To ensure fairness and prevent over-penalizing any single mode, the overall penalty is proportionally distributed based on the dispatchable capacity of each mode m at time step t, reflecting the greater responsibility of higher-capacity modes. The penalty of overcrowding for mode m at time step t is expressed by Eq. (23).

$$R_{\text{crowd}}^{m,t} = \frac{n_m^t}{\sum_{m' \in M} n_{m'}^t} \max\Big(\sum_{p \in \bar{P}^t} q(p) - \eta, 0\Big), \quad \forall m \in M, t \in T \tag{23}$$

Reward of demand satisfaction focuses on two aspects: the number of evacuated passengers and their stranded duration, respectively. The number of evacuated passengers is captured by $q(p)$ for each passenger group $p \in P_m^t$. Passengers' stranded duration is measured by the time before they are evacuated. Let $h(p,t)$ indicate the stranded duration of passenger group p which is evacuated at time step t. Stranded duration of each passenger group is calculated by the time from the occurrence of the OPF, t_{start}, (or the entry time, $e(p)$, for passengers who arrive during the OPF) to the time step t that the passengers are evacuated. It is calculated by Eq. (24).

$$h(p,t) = t - \max(t_{start}, e(p)), \quad \forall t \in T, m \in M, p \in P_m^t \tag{24}$$

Passengers' stranded duration is a negative consequence associated with the evacuation plan that agents are trained to reduce. However, agents are more effectively stimulated through positive feedback, as suggested by the reward hypothesis in Sutton and Barto (2018). Thus, a stranded duration factor, denoted by $H(p,t)$, is introduced to be nonnegative, as depicted in Fig. 6. Since the strand is inevitable, the agent is supposed to reduce passenger groups' evacuated stranded duration compared to their original stranded duration. For each passenger group p, let $h(p, \tilde{l}(p))$ be its original stranded duration where $\tilde{l}(p)$ is its original leaving time. When the evacuated stranded duration is longer than the original stranded duration, a discount should be applied to the reward for demand satisfaction, reducing the value of the

evacuation plan that evacuates a large number of passengers at the cost of additional stranded duration. Conversely, when the evacuated stranded duration is shorter than the original stranded duration, a leverage should be applied to increase the value of the evacuation plan. Then, the stranded duration factor is formulated through a negative exponential function, calculated by Eq. (25),

$$H(p,t) = e^{1-\frac{h(p,t)}{h(p,\tilde{l}(p))+\xi}}, \quad \forall t \in T, m \in M, p \in P_m^t, \tag{25}$$

where the denominator contains an additional term ξ, which is a relatively small number, to prevent the equation from becoming invalid when $h(p,\tilde{l}(p)) = 0$, while the influence is negligible for $h(p,\tilde{l}(p)) \gg \xi$. This function ensures a smooth and continuous evaluation with $H(p,t) > 0$ even when $h(p,\tilde{l}(p)) = 0$. When the evacuated stranded duration is close to the original stranded duration, where $h(p,t) \approx h(p,\tilde{l}(p))$, $H(p,t) \approx 1$, indicating neutrality. If the evacuated stranded duration is longer, where $h(p,t) > h(p,\tilde{l}(p))$, $H(p,t) < 1$, indicating a discount. Conversely, if the evacuated stranded duration is shorter, where $h(p,t) < h(p,\tilde{l}(p))$, then $H(p,t) > 1$, indicating a leverage.

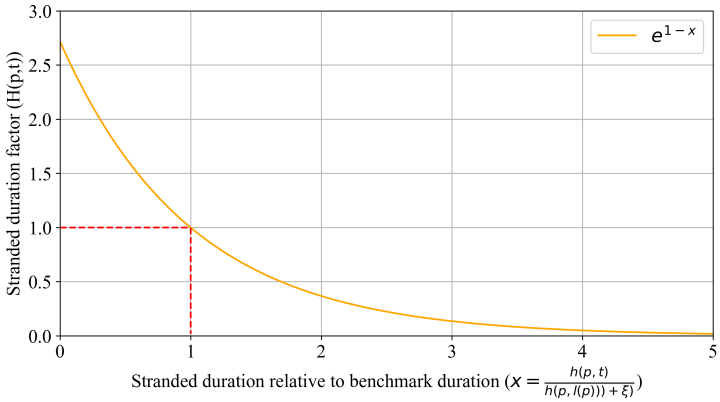


Figure 6: Trend of the stranded duration factor

Therefore, the reward for demand satisfaction for mode m at time step t is calculated as a sum that combines the stranded duration factor and the number of evacuated passengers by Eq. (26).

$$R_{\text{demand}}^{m,t} = \sum_{p \in P_m^t} H(p,t)q(p), \quad \forall m \in M, t \in T \tag{26}$$

5.4.2. Feedback 2: Impact on regular services

The impact on regular services is reflected by the number of passengers abandoned at time step t, which is formulated differently for each mode as follows:

- For the taxi mode, the regular passengers associated with the dispatched taxis $v \in \bar{V}_{\text{taxi}}^t$ are captured by $\rho_{\text{taxi}}(v)$ in Eq. (9). The number of abandoned regular passengers at time step t is then calculated by Eq. (27).

$$R_{\text{abandon}}^{\text{taxi},t} = \sum_{v \in \bar{V}_{\text{taxi}}^t} \rho_{\text{taxi}}(v), \quad \forall t \in T \tag{27}$$

- For the bus mode, the regular passengers associated with the dispatched buses $v \in \bar{V}_{\text{bus}}^t$ are captured by $\rho_{\text{bus}}(v)$ in Eq. (10). The number of abandoned regular passengers at time step t is then calculated by Eq. (28).

$$R_{\text{abandon}}^{\text{bus},t} = \sum_{v \in \bar{V}_{\text{bus}}^t} \rho_{\text{bus}}(v), \quad \forall t \in T \tag{28}$$

- For the metro mode, passengers at downstream stations may be denied boarding if the train is fully utilized when arriving. Given the set of evacuated passenger P_{metro}^t and their disembarkation station $i(p)$, the metro OD demand from station i^* (station in the OPF area) to each of the downstream stations j can be represented by Eq. (29).

$$w(v_t, i^*, j) = \sum_{\{p \in P_{\text{metro}}^t | i(p)=j\}} q(p), \quad \forall t \in T \tag{29}$$

Given the metro OD demand $w(v_t, i^*, j)$ from station i^*, the residual capacity on train v_t after loading passengers at station i^* can be updated by Eq. (11). The number of regular passengers abandoned in each downstream station, $\rho_{\text{metro}}(v_t, i)$, is derived from Eq. (12). Thus, the number of abandoned regular passengers at time step t is calculated by Eq. (30).

$$R_{\text{abandon}}^{\text{metro},t} = \sum_{i=i^*+1}^{\bar{i}} \rho_{\text{metro}}(v_t, i), \quad \forall t \in T \tag{30}$$

6. Training algorithm

Two training algorithms are introduced in this section. A DTDE structure is introduced for our independent strategy in Section 6.1. An H-CTDE structure is customized for our collaborative strategy in Section 6.2.

6.1. DTDE framework

To train the agents in the independent strategy in Section 5.3.1, a DTDE training method is proposed, where each agent has its own actor-critic network, trained by a deep deterministic policy gradient algorithm. Within the DTDE framework, each agent learns its policy based on the state variables while executing independently (Gronauer and Diepold, 2022). DTDE is particularly effective for multiple agents that require only their own history of observations during both training and inference, as demonstrated in previous studies (Wang et al., 2023; Yu et al., 2023). It is applicable for our independent strategy because separate agents control different modes, avoiding confusion among heterogeneous agents with distinct dispatching strategies and operational characteristics.

Within the DTDE framework, each agent is equipped with an actor network that decides the number of dispatched capacities based on its state variables, and a critic network that evaluates the value of the proposed actions. The actor and critic networks used for each agent are described as follows:

Actor Networks: An agent of each mode $m \in M$ has its own actor network π_m, which maps its state variables S_m^t to an action a_m^t. However, because each mode has a unique dispatchable capacity as a constraint at each time step t, a general intermediate action is required. Let $\hat{a}_t^m \in [0, 1]$ represent the proportion of dispatched capacities, which is derived from $\hat{a}_m^t = \pi_m(S_m^t \mid \theta_m^\pi)$ where θ_m^π are the parameters of the actor network. A target actor network, $\pi'_m(S_m^t \mid \theta_m^{\pi'})$, is applied as well to estimate the target action for the next state S_m^{t+1}. To encourage exploration during the training process, a noise factor is added to the actions, which helps to explore different actions in the early stage, preventing premature convergence to suboptimal policies. The noise is gradually reduced to ensure that the agents focus more on exploiting learned behaviors. The noise is modeled following the Gaussian distribution in Eq. (31), which decays with the number of episodes.

$$\epsilon^{\text{noise}} \mathcal{N}(0, (\underline{\epsilon}^{\text{noise}})^{\frac{e}{E}}), \tag{31}$$

where e is the current episode and E is the total training episodes. $\underline{\epsilon}^{\text{noise}}$ represents the minimum noise value. In each time step t, the proportion is modified by adding the noise, and clip the modified proportion into $[0, 1]$. The proportion is calculated by Eq. (32).

$$\hat{a}_m^t = \text{clip}(\pi_m(S_m^t \mid \theta_m^\pi) + \epsilon^{\text{noise}}, 0, 1), \quad \forall m \in M, t \in T \tag{32}$$

Incorporation of constraints: The basic reinforcement learning framework does not explicitly consider the constraints imposed on the decision space (Ying et al., 2020). For taxi and metro dispatch, Constraint (3) should be satisfied, where the dispatched capacity of each mode m at each time step t cannot exceed the dispatchable capacity. $\hat{a}_m^t$ is multiplied by the dispatchable capacity n_m^t for each mode m at a time step t. Considering Constraint (8), the result is rounded to the nearest integer. Agents' actions are calculated by Eq. (33).

$$a_m^t = \text{round}(\hat{a}_m^t n_m^t), \quad \forall m \in M = \{\text{taxi, metro}\}, t \in T \tag{33}$$

For bus dispatch, Constraint (6) should be additionally considered to ensure that each route is assigned at most one bus at each time step t.

Proposition 2. When the number of dispatched buses a_{bus}^t exceeds the total number of emergency

routes $|\gamma|$, at least one route must be assigned more than one bus.

Proof. The set of dispatched buses is denoted by $\bar{V}^t_{bus} = \{v_1, \ldots, v_n\}$, where $|\bar{V}^t_{bus}| = a^t_{bus}$. The set of routes assigned to buses is denoted by $\gamma^t = \{r_1, \ldots, r_n\}$, where $\gamma^t \subset \gamma$ and $|\gamma^t| \leq |\gamma|$. Since Proposition 1 has shown that each route can have at most one bus assigned when $|\gamma^t| = a^t_{bus}$, it follows that $a^t_{bus} \leq |\gamma|$ must hold before buses are assigned to specific routes.

Therefore, $\hat{a}^t_{\text{bus}}$ is multiplied by the dispatchable capacity of bus n^t_{bus} at a time step t, and round the result to the nearest integer. Then, the rounded integer is clipped into $[0, |\gamma|]$ to ensure that the number of dispatched buses cannot exceed the total number of emergency routes. Agent's action is by Eq. (34).

$$a^t_{\text{bus}} = \text{clip}(\text{round}(\hat{a}^t_{\text{bus}} n^t_{\text{bus}}), 0, |\gamma|), \quad \forall t \in T \tag{34}$$

Then, Constraint (6) is enforced by in-built rules for bus assignment in the agent-based model, as described in Section 5.3.1.

Critic Networks: Each agent has a critic network, $Q_m(S^t_m, a^t_m \mid \theta^Q_m)$, that estimates the Q-value of taking action a^t_m given the state variables S^t_m, where θ^Q_m are the parameters of the critic network. A target critic network, $Q'_m(S^t_m, a^t_m \mid \theta^{Q'}_m)$, is applied to estimate the target Q-value of the next state.

The training process begins by initializing the actor and critic networks along with their corresponding target networks. During each time step t, a tuple of experience $(S^t_m, a^t_m, R^t_m, S^{t+1}_m)$ is collected and stored in a replay buffer, denoted as RB. A mini-batch of experience tuples of size $B \in \mathbb{N}$ is then sampled from the replay buffer to update the networks. These samples are used to calculate target Q-values and train both the actor and critic networks.

The critic network is trained by minimizing the loss function $L(\theta^Q_m)$, defined as Eq. (35).

$$\min_{\theta^Q_m} L(\theta^Q_m) = \mathbb{E}_{(S^t_m, a^t_m, R^t_m, S^{t+1}_m) \sim RB} \left[\left(Q_m(S^t_m, a^t_m \mid \theta^Q_m) - R'^t_m \right)^2 \right] \tag{35}$$

The target Q-value R'^t_m is calculated by Eq. (36),

$$R'^t_m = R^t_m + \alpha_d Q'_m(S^{t+1}_m, a^{t+1}_m \mid \theta^{Q'}_m) \mid_{a^{t+1}_m = \pi'_m(S^{t+1}_m)}, \tag{36}$$

where α_d is the discount factor in the Bellman equation.

The actor network aims to maximize the expected Q-value over the policy's action, which serves as its objective function, calculated by Eq. (37).

$$\max_{\theta^\pi_m} J(\theta^\pi_m) = \mathbb{E}_{(S^t_m, a^t_m) \sim RB}[Q_m(S^t_m, a^t_m \mid \theta^Q_m)] \tag{37}$$

Let $\nabla L(\theta^Q_m)$ and $\nabla J(\theta^\pi_m)$ denote the gradients used to update the critic and actor networks, respectively. The weight updates are performed using learning rates α^m_Q and α^m_π, calculated by Eqs. (38) and (39).

$$\theta^Q_m \leftarrow \theta^Q_m + \alpha^m_Q \nabla L(\theta^Q_m), \tag{38}$$

$$\theta^\pi_m \leftarrow \theta^\pi_m + \alpha^m_\pi \nabla J(\theta^\pi_m) \tag{39}$$

Target networks Q'_m and π'_m are used to stabilize training by providing consistent targets. These networks are updated using a soft update mechanism with a given update rate α_s, which is calculated by Eqs. (40) and (41).

$$\theta^{Q'}_m \leftarrow \alpha_s \theta^Q_m + (1 - \alpha_s)\theta^{Q'}_m, \tag{40}$$

$$\theta^{\pi'}_m \leftarrow \alpha_s \theta^\pi_m + (1 - \alpha_s)\theta^{\pi'}_m \tag{41}$$

Algorithm 1 illustrates the framework of DTDE training process.

6.2. H-CTDE framework

Under the collaborative strategy, heterogeneous agents share the rewards and penalties of OPF evacuation based on passenger mode-shifting mechanism. Since each mode has unique capacity, modes with larger capacities have more significant contributions to the reward, while the contribution by modes with smaller capacities may be overlooked. Additionally, due to the different regular passenger distribution across modes, modes with more regular passengers are likely to face higher penalties. As a result, agents

Algorithm 1 DTDE Training Algorithm

Initialize parameters θ_m^{π} and θ_m^{Q} for actor and critic networks, respectively
Initialize target networks: $\theta_m^{\pi'} \leftarrow \theta_m^{\pi}$, $\theta_m^{Q'} \leftarrow \theta_m^{Q}$
Initialize replay buffer RB into an empty set
while $e \leq E$ **do**
  for each time step $t \in T$ **do**
    Observe state S_m^t for each agent $m \in M$
    Select the proportion of dispatched capacity $\hat{a}_m^t$ for each agent $m \in M$
    Calculate the action a_m^t by Eqs. (33) and (34) for each agent $m \in M$
    Execute actions a_m^t and observe the reward R_m^t and the next state S_m^{t+1}
    Store the tuple of experiences $(S_m^t, a_m^t, R_m^t, S_m^{t+1})$ in replay buffer RB
  end for
  for each agent $m \in M$ **do**
    Sample a batch of tuples $(S_m^t, a_m^t, R_m^t, S_m^{t+1})$ with the batch size of B from replay buffer RB
    Calculate target Q-values by Eq. (36)
    Update critic using the sampled gradient by Eqs. (35) and (38)
    Update actor using the sampled gradient by Eqs. (37) and (39)
    Update target networks by Eqs. (40) and (41)
  end for
  Update the episode number: $e \leftarrow e + 1$
  Update the noise factor: $\epsilon^{\text{noise}} \mathcal{N}(0, (\underline{\epsilon}^{\text{noise}})^{\frac{e}{E}})$
end while

with limited capacity may become lazy, relying on higher-capacity modes to shoulder the burden. This aligns with the lazy behavior in MARL, where certain agents fail to actively contribute to the system's overall performance (Liu et al., 2023). Furthermore, agents managing modes with higher regular demand may become conservative, avoiding dispatching residual capacities to minimize penalties. This phenomenon has been observed in MARL training (Shao et al., 2019).

A novel CTDE structure is customized to capture the complex interdependencies among multiple heterogeneous agents. Fig. 7 shows the multi-agent actor-critic architecture with H-CTDE. Each agent has a distributed actor network (black components) that references its own local states (blue components) and controls its specific dispatch actions (yellow components). In addition, each agent is equipped with a local critic (light green components), taking its local state and action as input, that evaluates the demand satisfaction and passenger abandonment within its respective mode. By emphasizing local metrics, the local critic ensures that each agent is aware of its individual contribution and impact of its action. Then a central critic network (dark green component) is further developed, which takes the joint state and action as input, to account for the overall objective. The combination of local critics and central critic allows agents to optimize performance for their specific modes while aligning with broader system objectives. The neural networks are formulated as follows:

Actor Networks: The agent of each mode $m \in M$ has an actor network, denoted by $\pi_m(S_m'^t \mid \theta_m^{\pi})$, which takes local state $S_m'^t$ as input and outputs an action a_m^t. A target actor network, $\pi_m'(S_m'^t \mid \theta_m^{\pi'})$, is applied. The proportion of dispatched capacity of each mode at each time step $\hat{a}_m^t$ is calculated by Eq. (32), and the value of action is determined by Eqs. (33) and (34).

Central Critic Network: The central critic network is responsible for evaluating the joint action-value function, taking into account the global reward. Let $S^t = [\delta^t, t, n_{\text{taxi}}^t, n_{\text{bus}}^t, n_{\text{metro}}^t]$ denote the joint states, $a^t = [a_{\text{taxi}}^t, a_{\text{bus}}^t, a_{\text{metro}}^t]$ denote the joint actions. It takes the joint state and action as input and outputs a global Q-value, denoted as $Q(S^t, a^t \mid \theta^Q)$. A target central critic network, $Q'(S^t, a^t \mid \theta^{Q'})$, is applied as well. The overall feedback at each time step t is formulated as: $R^t = \sum_{m \in M} R_m^t$.

Local Critic Network: Each agent also has a local critic network $Q_m(S_m'^t, a_m^t \mid \theta_m^Q)$ that evaluates its actions based on local feedback for mode m at time step t, denoted by $R_{\text{local}}^{m,t}$, which combines the rewards for demand satisfaction and penalties for passenger abandonment, where $R_{\text{local}}^{m,t} = \alpha_1 R_{\text{demand}}^{m,t} - \alpha_3 R_{\text{abandon}}^{m,t}$. A target local critic network, $Q_m'(S_m'^t, a_m^t \mid \theta_m^{Q'})$, is applied as well.

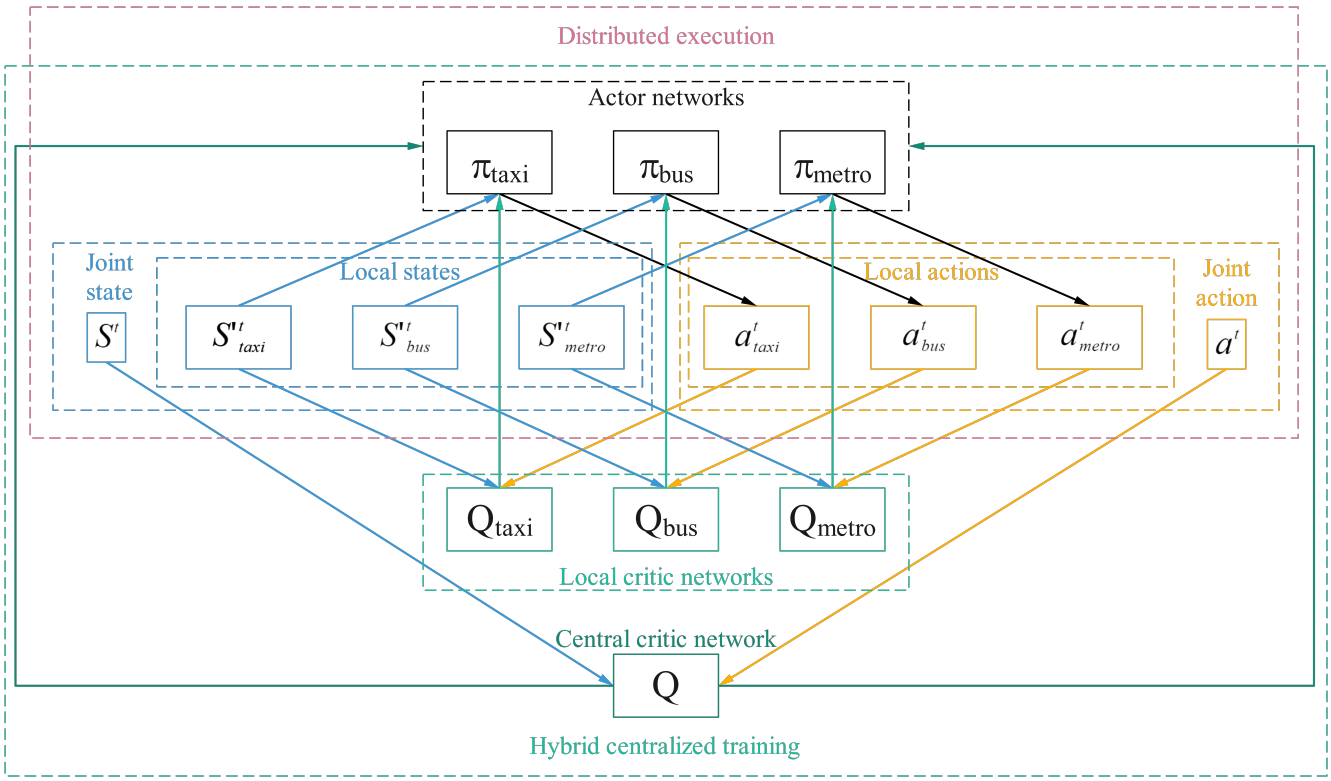


Figure 7: Multi-agent actor-critic architecture with H-CTDE

The target Q-value of local critic network, denoted by $R'^{m,t}_{\text{local}}$, is calculated by Eq. (42).

$$R'^{m,t}_{\text{local}} = R^{m,t}_{\text{local}} + \alpha_d Q'_m(S'^{t+1}_m, a^{t+1}_m \mid \theta^{Q'}_m)\mid_{a^{t+1}_m=\pi'_m(S'^{t+1}_m)} \tag{42}$$

The loss function of local critic network, denoted by $L(\theta^Q_m)$, is calculated by Eq. (43).

$$\min_{\theta^Q_m} L(\theta^Q_m) = \mathbb{E}_{(S^t_m,a^t_m,R^{m,t}_{\text{local}},S^{t+1}_m)\sim RB}\left[\left(Q_m(S^t_m,a^t_m \mid \theta^Q_m) - R'^{m,t}_{\text{local}}\right)^2\right] \tag{43}$$

The target Q-value of central critic network, denoted by R'^t, is calculated by Eq. (44).

$$R'^t = R^t + \alpha Q'(S'^{t+1}, a^{t+1} \mid \theta^{Q'})\mid_{a^{t+1}=[\pi'_{\text{taxi}}(S'^{t+1}_{\text{taxi}}),\pi'_{\text{bus}}(S'^{t+1}_{\text{bus}}),\pi'_{\text{metro}}(S'^{t+1}_{\text{metro}})]} \tag{44}$$

The loss function of central critic network, denoted by $L(\theta^Q)$, is calculated by Eq. (45).

$$\min_{\theta^Q} L(\theta^Q) = \mathbb{E}_{(S^t,a^t,R^t,S^{t+1})\sim RB}\left[\left(Q(S^t,a^t \mid \theta^Q) - R'^t\right)^2\right] \tag{45}$$

The objective function of actor network, denoted by $J(\theta^\pi_m)$, is calculated by Eq. (46).

$$\max_{\theta^\pi_m} J(\theta^\pi_m) = \mathbb{E}_{(S^t_m,a^t_m,S^t,a^t)\sim RB}[Q_m(S^t_m,a^t_m \mid \theta^Q_m) + Q(S^t,a^t \mid \theta^Q)] \tag{46}$$

The local critic and actor networks are updated by Eqs. (38) and (39), and the central critic network is updated by $\theta^Q \leftarrow \theta^Q + \alpha_Q \nabla L(\theta^Q)$. The target local critic and actor networks are updated by Eqs. (40) and (41), and the target central network is updated by $\theta^{Q'} \leftarrow \alpha_s \theta^Q + (1-\alpha_s)\theta^{Q'}$. Algorithm 2 illustrates the framework of H-CTDE training process.

7. Case study

Our online multi-modal evacuation approach is evaluated in a real-world OPF scenario. In Section 7.1, the study area and parameter settings for the case are presented as well as the OPF peak-hour and full-term period are identified. The computational efficiency of our MARL algorithms is demonstrated in Section 7.2. The effectiveness of our online multi-modal evacuation in enhancing resilience and impact

Algorithm 2 H-CTDE Training Algorithm

Initialize parameters of actor θ_m^{π}, local critic θ_m^{Q} for each agent m, and central critic θ^{Q}
Initialize target networks with $\theta_m^{\pi'} \leftarrow \theta_m^{\pi}$, $\theta^{Q'} \leftarrow \theta^{Q}$ and $\theta_m^{Q'} \leftarrow \theta_m^{Q}$
Initialize replay buffer RB into an empty set
while $e \leq E$ **do**
  for each time step $t \in T$ **do**
    Observe state $S_m^{'t}$ for each agent $m \in M$
    Select the proportion of dispatched capacity $\hat{a}_m^t$ for each agent $m \in M$
    Calculate the action a_m^t by Eqs. (33) and (34) for each agent $m \in M$
    Execute actions a_m^t and observe global feedback R^t, local reward $R_{\text{local}}^{m,t}$ and the next state $S_m^{'t+1}$
    Store the tuple $(S^t, S_{\text{taxi}}^{'t}, S_{\text{bus}}^{'t}, S_{\text{metro}}^{'t}, a^t, a_{\text{taxi}}^t, a_{\text{bus}}^t, a_{\text{metro}}^t, R^t, R_{\text{local}}^{m,t}, S^{t+1}, S_{\text{taxi}}^{'t+1}, S_{\text{bus}}^{'t+1}, S_{\text{metro}}^{'t+1})$ into RB
  end for
  for each agent $m \in M$ **do**
    Sample a batch of tuples $(S^t, S_{\text{taxi}}^{'t}, S_{\text{bus}}^{'t}, S_{\text{metro}}^{'t}, a^t, a_{\text{taxi}}^t, a_{\text{bus}}^t, a_{\text{metro}}^t, R^t, R_{\text{local}}^{m,t}, S^{t+1}, S_{\text{taxi}}^{'t+1}, S_{\text{bus}}^{'t+1}, S_{\text{metro}}^{'t+1})$ with the batch size of B
    Calculate target local and central Q-values by Eqs. (42) and (44), respectively
    Calculate the loss of local critic, central critic, and actor networks by Eqs. (43), (45) and (46), respectively
    Update the local critic and actor networks by Eqs. (38) and (39), respectively. Update the central critic network by: $\theta^{Q} \leftarrow \theta^{Q} + \alpha_Q \nabla L(\theta^{Q})$
    Update target local critic networks and target actor network by Eqs. (40) and (41), respectively. Update target central critic networks as: $\theta^{Q'} \leftarrow \alpha_s \theta^{Q} + (1-\alpha_s)\theta^{Q'}$
  end for
  Update the episode number: $e \leftarrow e+1$
  Update the noise factor: $\epsilon^{\text{noise}} \mathcal{N}(0, (\underline{\epsilon}^{\text{noise}})^{\frac{e}{E}})$
end while

mitigation is shown in Section 7.3. A series of agents trained in Section 7.2 are applied to new environments for online decision-making, verifying the transferability of our MARL approaches in Section 7.4. Finally, the practical, theoretical implications, and managerial insights are summarized in Section 7.5.

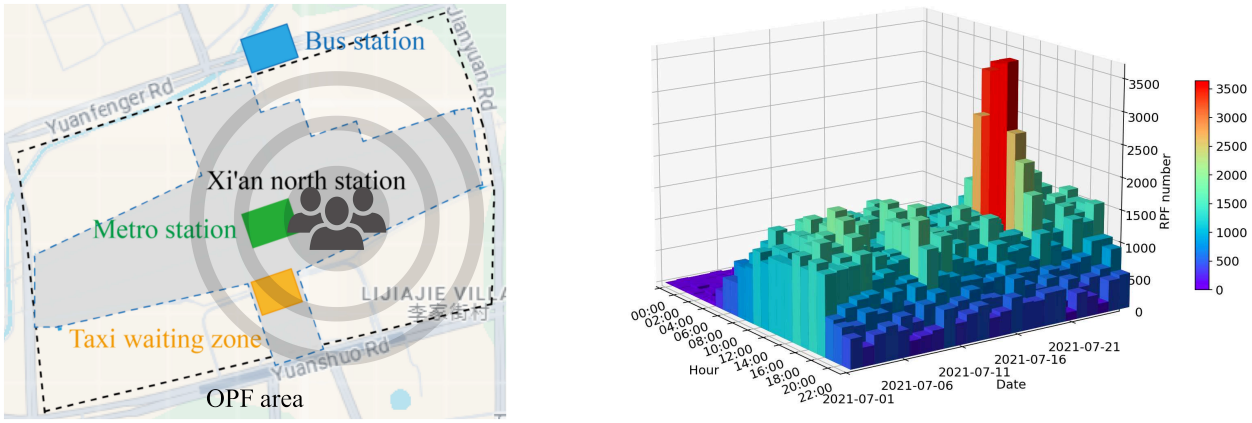
7.1. Experiment description

This case study is based on a severe railway disruption that occurred on July 20, 2021, when heavy rainfall caused a severe malfunction at the central mainland railway hub in Zhengzhou, China (Hu et al., 2024). The study area is focused on Xi'an North Station, a high-speed railway station as well as a multimodal urban transit hub, with a taxi waiting zone (orange area), a bus station (blue area), and a metro station (green area), as shown in Fig. 8a. Following the cancellation of all trains, railway passengers waiting in the station are forced to return, leading to the OPF overwhelming the local urban transit system. Unnoticed passengers continue to arrive at the station, worsening the situation.

Fig. 8b demonstrates the distribution of passengers leaving Xi'an North Station across the days in July 2021 through a heat map. The horizontal axes represent different days and hours, respectively, while the vertical axis represents the number of passengers leaving the station in each hour. On July 20, there is an anomalously large number of passengers leaving in the middle of the day, as indicated by the red bars between 12:00 and 16:00. The OPF on July 20 forms the basis of our case study. An overcrowding threshold is set by referring to the 70th percentile of passenger numbers leaving the area throughout July 2021 (Ma et al., 2024). Therefore, based on the historical records of passenger numbers leaving the railway station, the overcrowding threshold is set as 5,000.

Fig. 9 shows the number of passengers stranded at the OPF area throughout the day of our case study. Any time when the number of stranded passengers exceeds the overcrowding threshold, an evacuation is required. Therefore, the full term of OPF period is determined as from 10:00 to 18:40. As the peak occurs around 14:00, four hours centered on this peak are selected as the peak-hour period, which is from 12:00 to 16:00. The varying starting times and time spans create distinct scenarios, each with different numbers of stranded passengers and time steps. These differences lead to agents developing unique decision-making policies during training. Scenarios with varying time spans are applied to train MARL agents, allowing for the evaluation of their problem-solving effectiveness and the testing of their ability to transfer solutions in online applications.

The taxi GPS, bus smart card transactions, AVL, and time-dependent metro OD demand are synchronized with mobile data during the OPF on July 20 in Xi'an. The time step Δt is set to 5 minutes.



(a) Study area

(b) Distribution of passengers leaving the station

Figure 8: Study area and passenger distribution during OPF

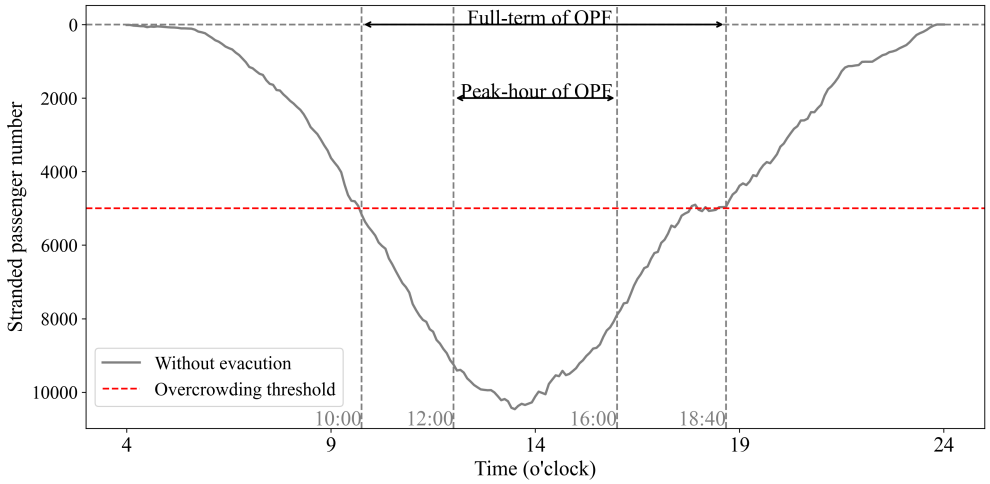


Figure 9: Identification of OPF period

Based on the administrative divisions of passengers' destinations, the city is divided into 13 service areas, as illustrated in Fig. 10. Each area is assigned a terminal station (black point) for the emergency bus routes based on passengers' average travel distance. Typically, the maximum capacity under emergency conditions is 150% of the normal seating capacity (Abdelgawad and Abdulhai, 2012). Hence, the bus capacity c_{bus} is set to 100 psg/veh, while the metro capacity c_{metro} is set to 3000 psg/veh. The loyalty factor of passengers' original mode choice ϵ_0^m is set as 0.9. Based on practical implementation, the weights for trip cost $[\epsilon_1^m, \epsilon_2^m, \epsilon_3^m]$ are set as [0.5, 3.5, 0] for taxi, [2, 1, 1] for bus, and [1.5, 1.5, 1] for metro. Feedback weights $[\alpha_1, \alpha_2, \alpha_3]$ are set as [1, 0.5, 0.5] to balance the objectives of the evacuation process. The first weight rewards the agent for demand satisfaction. The second and third weights (0.5 for both) penalize overcrowding and passenger abandonment.

The hyperparameters for MARL training are systematically adjusted to identify the best-performing configuration. The actor, local critic and central critic networks for each agent all have two hidden layers with 200 and 100 units, respectively. The activation functions of the hidden layers are ReLU. The activation functions of the output layers are set as a sigmoid function to ensure that the output values remain within the range $[0, 1]$. The minimum action noise is set as $\underline{\epsilon}^{\text{noise}} = 0.1$. The discount factor for the target Q-value is set to $\alpha_d = 0.9$. The total number of training episodes is set to $E = 5,000$. The sizes of the reply buffer and batch vary with the size of the episode (i.e., time span) $|T| = (t_{end} - t_{start})/\Delta t$, which are set as $100 * |T|$ and $20 * |T|$, respectively. The target network soft update rate is $\alpha_s = 0.8$, which updates every 20 episodes. The learning rates are set as 10^7 for central critic, 10^5 for taxi actor, 10^8 for the taxi critic, 10^5 for bus actor, 10^6 for the bus critic, 10^5 for the metro actor, and 10^7 for the metro critic.

7.2. Computational efficiency and policy learning analysis

This section focuses on illustrating the computational efficiency of our MARLs in addressing the complexity of dynamic multi-modal evacuation, and demonstrating the detailed policy learning process

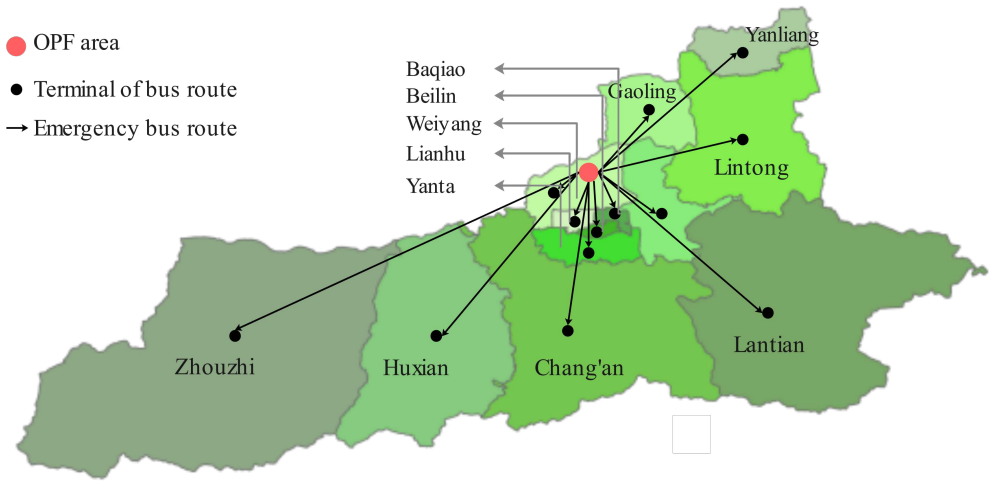


Figure 10: Classification of service areas and terminal stations of emergency bus routes

of agents under both independent and collaborative strategies. First of all, the computational efficiency is demonstrated under a series of environments with different time spans in Section 7.2.1. Then, the convergence processes of our customized MARLs are analyzed by comparing them to GAs and MADDPG algorithms in Section 7.2.2. Finally, different policies learned under independent and collaborative strategies are compared in Section 7.2.3.

Note that both MARL and GA can solve the offline problem with the assumption of complete knowledge of the entire period. However, only MARL is applicable to online settings. This section solely focuses on illustrating the computational efficiency of our customized MARLs for the offline problem. Two other algorithms are used for comparison: GA, a representative search-based heuristic algorithm, and MADDPG, a classic MARL algorithm with a conventional CTDE framework. These two algorithms are commonly employed as benchmarks in offline cases (Li and Ni, 2022; Ying et al., 2022).

7.2.1. Computational efficiency: our MARL vs benchmarks

To demonstrate the computational efficiency of our MARL, a series of environments with different time spans are built. As the full-term period ranges from 10:00-18:40, the start time is fixed at 10:00. Then, the end times are set as 12:00, 14:00, 16:00, 18:00, and 20:00, respectively. This creates environments with time spans of 2, 4, 6, 8, and 10 hours, which sufficiently cover the OPF period. These environments are used to evaluate our MARL approaches and the benchmark algorithms (i.e., GA and MADDPG). The GA is configured with a population size of 50 and a generation of 100, resulting in a total of 5,000 attempts, which matches the number of training episodes in our MARL approaches. The MADDPG uses the same parameters as our MARL approaches.

The computational efficiency, including the maximum rewards, final average rewards (average reward over the last 100 episodes) and computational time, are demonstrated in Table 3. GA-I refers to the GA under independent strategy, while GA-C refers to that under collaborative strategy. Let the results of our H-CTDE algorithm as the baseline, with the gaps of the other algorithms calculated relative to it.

First, both the maximum and final average rewards for our MARL approaches under the collaborative strategy are 12-24% higher than those under the independent strategy, demonstrating the effectiveness of the passenger mode-shifting mechanism in the collaborative strategy. The detailed policy learning process under these two strategies will be analyzed in Section 7.2.3.

Second, the maximum rewards achieved by GAs are slightly (1-3%) higher than those of our approach under both the independent and collaborative strategies. This is due to GAs' higher randomness in their search mechanism. However, GAs require significantly more computation time, approximately 22-48% higher than that of our approaches. Additionally, the final average rewards, which reflect the stable performance after training, obtained by our MARL approaches are 2-4% higher than those of GAs under both strategies. These gaps tend to increase with the length of time span. This result indicates that our MARL has higher stability, which is crucial for online applications that require continuous provision of reliable results based on updated information.

Finally, our H-CTDE approach achieves 2-6% higher maximum rewards and 7-11% higher final average reward compared to those of MADDPG. This demonstrates the effectiveness of our H-CTDE approach

in solving multi-modal evacuation problems using hybrid critic networks, especially when dealing with heterogeneous agents that have different dispatching strategies and operational characteristics.

The detailed convergence processes of the algorithms are provided in Fig. D1 in the Appendix D. Each subfigure compares different algorithms under the same evacuation strategy and time span.

Two special scenarios are tested: one representing the peak hour of the OPF and the other representing the full-term duration of the OPF. As analyzed in Section 7.1, the peak hour spans from 12:00 to 16:00, while the full-term duration extends from 10:00 to 18:40. The computational efficiencies are listed in Table 4. The convergence process of the peak-hour scenario and the maximum-reward solutions of the full-term scenario will be analyzed in detail in Sections 7.2.2 and 7.3, respectively.

Table 3: Comparison of the offline computational efficiency

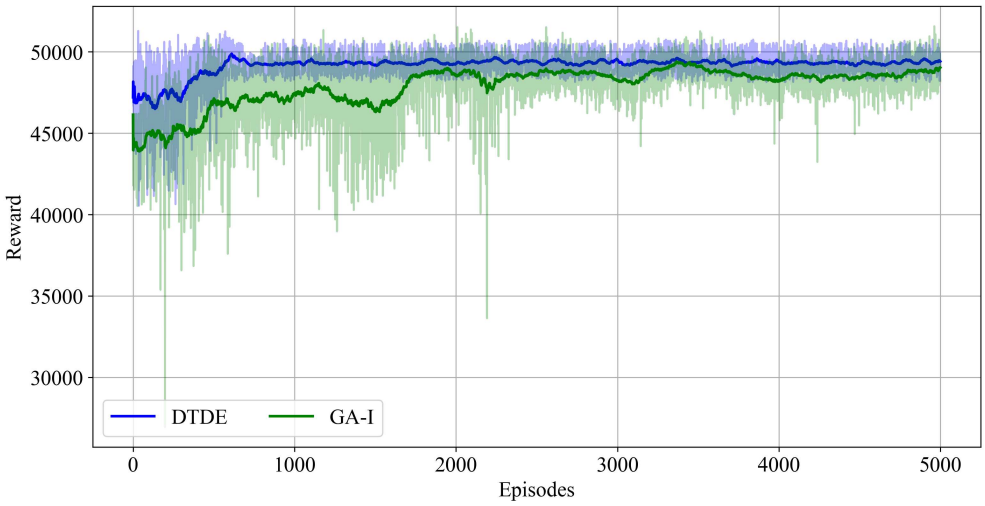
Time span	Strategy	Algorithm	Maximum reward		Final average reward		Time	
			Value	Gap (%)	Value	Gap (%)	Value (s)	Gap%
10:00-12:00	Independent	**DTDE**	30100.17	-23	28524.98	-24	7966	-12
		GA-I	30719.30	-21	27578.90	-26	10294	+14
	Collaborative	**H-CTDE**	38962.82	-	37369.52	-	9007	-
		GA-C	38750.43	-0.5	36116.55	-3	11824	+31
		MADDPG	37379.58	-4	33272.04	-11	6906	-23
10:00-14:00	Independent	**DTDE**	47640.52	-14	46480.61	-15	14185	-20
		GA-I	48819.09	-12	45713.33	-19	19541	+10
	Collaborative	**H-CTDE**	55596.37	-	54512.70	-	17791	-
		GA-C	56358.92	+1	53283.90	-2	26291	+48
		MADDPG	54379.58	-2	50272.04	-8	12298	-31
10:00-16:00	Independent	**DTDE**	60524.69	-14	59179.18	-14	20965	-19
		GA-I	62161.38	-12	56777.54	-18	27666	+6
	Collaborative	**H-CTDE**	70181.90	-	68819.78	-	26016	-
		GA-C	71796.90	+2	67396.71	-2	35654	+37
		MADDPG	67379.58	-4	63272.04	-8	18176	-30
10:00-18:00	Independent	**DTDE**	70665.16	-12	69621.00	-12	28233	-17
		GA-I	72577.52	-9	67702.55	-15	35832	+6
	Collaborative	**H-CTDE**	80126.83	-	78975.00	-	33832	-
		GA-C	81477.27	+1	77470.19	-2	41300	+22
		MADDPG	77379.58	-4	73272.04	-7	24478	-28
10:00-20:00	Independent	**DTDE**	81041.88	-13	80005.75	-13	34395	-21
		GA-I	81724.83	-12	74961.22	-19	45358	+4
	Collaborative	**H-CTDE**	93292.66	-	92264.47	-	43533	-
		GA-C	93683.06	+0.4	88963.22	-4	55033	+26
		MADDPG	87379.58	-6	83272.04	-10	29820	-32

Table 4: Computational efficiency of peak-hour and full-term scenarios

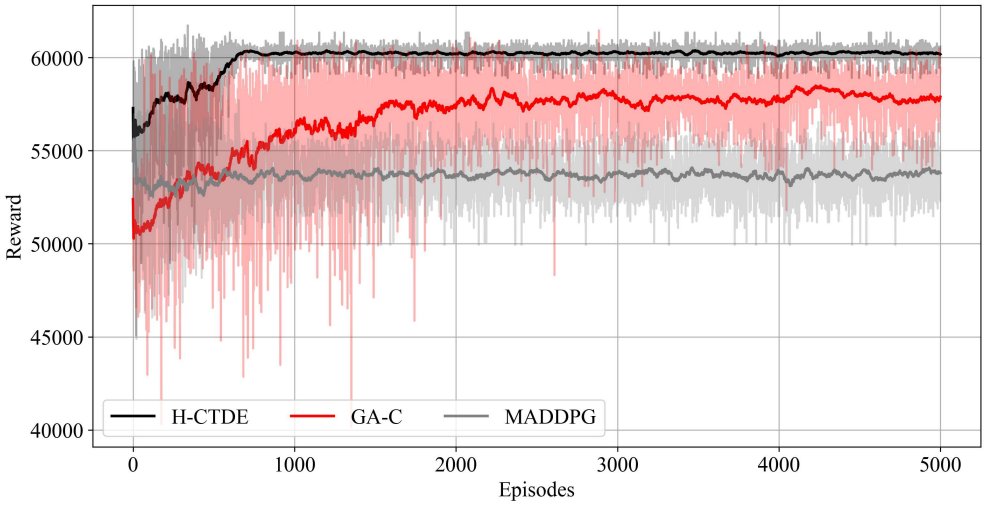
Time span	Strategy	Algorithm	Maximum reward		Final average reward		Time	
			Value	Gap (%)	Value	Gap (%)	Value (s)	Gap (%)
12:00-16:00 (Peak hour)	Independent	**DTDE**	51284.11	-20	49434.07	-22	9133	-0.2
		GA-I	51565.14	-20	48907.59	-23	11565	+21
	Collaborative	**H-CTDE**	61729.92	-	60129.79	-	9154	-
		GA-C	61470.01	-0.4	57870.45	-4	12165	+25
		MADDPG	57995.29	-6	53794.49	-12	7926	-15
10:00-18:40 (Full term)	Independent	**DTDE**	84643.59	-26	78611.27	-32	30589	-24
		GA-I	80411.41	-33	76575.23	-36	38967	+3
	Collaborative	**H-CTDE**	107009.84	-	104113.92	-	37911	-
		GA-C	104869.64	-2	98075.31	-6	44913	+16
		MADDPG	96308.65	-11	91492.60	-14	26619	-42

7.2.2. Convergence process: our MARL vs benchmarks

To analyze the convergence process, the training processes under the peak-hour environment with a four-hour time span from 12:00 to 16:00 is shown in Section 7.2.2. The convergence processes of independent and collaborative strategies are presented in Fig. 11a and Fig. 11b, respectively. Both Fig. 11a and Fig. 11b illustrate that the MARLs have a clear advantage in computational efficiency, which converge within approximately 800 episodes, whereas GAs require over 3000 episodes to reach convergence. The maximum rewards, final average rewards and computation time have been presented in Table 4.



(a) Independent strategy



(b) Collaborative strategy

Figure 11: Convergence of GA and MARL

According to Table 4, under the independent strategy, DTDE achieves lower maximum reward (51284.11 vs. 51565.14) but higher final average reward after the algorithm convergence (49434.07 vs. 48907.59) compared to GA-I. This difference suggests that while GA-I can stochastically achieve individual high rewards, the DTDE approach can stably achieve higher rewards after training. Under the collaborative strategy, the gap between H-CTDE and GA-C widens further. H-CTDE achieves the highest performance with a maximum reward of 61,729.92 and a final average reward of 60,129.79. In contrast, GA-C has a final average reward of 57,870.45 (4% lower than H-CTDE) and a slightly lower maximum reward of 61,470.01. This result highlights the superior adaptability and efficiency of H-CTDE in managing the complexity of collaborative dispatch and passengers' mode-shifting behavior. This is because the reward structure requires balancing competing objectives, such as maximizing demand satisfaction and minimizing passenger abandonment. The transition between states depends on the actions taken, making the reward highly interactive with the environment. The learning-based algorithm effectively optimizes these trade-offs by iteratively refining policies based on feedback, whereas searching-based algorithms like GA struggle because they rely on predefined heuristic strategies and limited initial search spaces. A similar phenomenon is also reported by Ying et al. (2022).

By comparing our proposed H-CTDE with the MADDPG training framework, the improvements become more significant. The maximum reward achieved by the MADDPG is 57,995.29 (6% lower than H-CTDE), with a final average reward of 53,794.49 (12% lower than H-CTDE). Another phenomenon can be found in Fig. 11b that the reward per episode achieved by MADDPG (grey line) drops during the first 100 training episodes and remains stable thereafter, indicating an unfavorable learning direction where agents are stuck in suboptimal policies. In the MADDPG approach, the central critic evaluates the policies of all agents collaboratively, which can inadvertently lead to certain agents becoming "lazy".

This issue is particularly evident in heterogeneous agent frameworks, such as ours, where the capacities of different modes (e.g., $c_{\text{metro}} = 3000$ vs. $c_{\text{bus}} = 100$) vary significantly. Agents controlling higher-capacity modes, like metro, can dominate the optimization process by achieving higher demand satisfaction, while lower-capacity modes contribute less visibly to the global reward. Furthermore, the global penalty in the MADDPG framework fails to adequately capture the impact of evacuation on regular services. Regular passenger demand varies across modes, times, and areas, as proved by the results in Figs. 17 and 18, which will be thoroughly discussed in Section 7.3.3. Therefore, a global penalty is counterproductive, potentially encouraging overly conservative behavior. These phenomena of lazy and conservative behavior align with challenges frequently highlighted in previous studies (Shao et al., 2019; Liu et al., 2023). In contrast, H-CTDE addresses these challenges by enabling better coordination and individualized learning for each mode, ensuring that all agents are effectively stimulated to contribute to the evacuation.

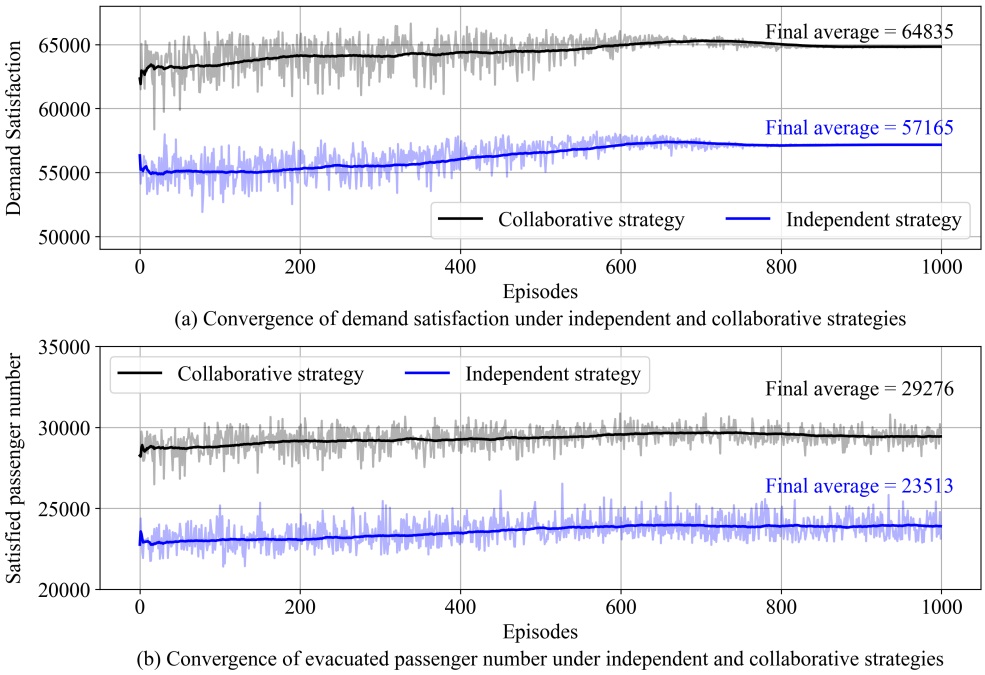
7.2.3. Policy learning analysis: independent vs. collaborative strategies

An obvious result, supported by Tables 4, shows that the rewards achieved by the collaborative strategy are generally higher than those achieved by the independent strategy. The final average reward of collaborative strategy is 22% higher than that of independent strategy (60129.79 vs. 49434.07). This result shows different policies learned by agents under the two evacuation strategies. To better understand agents' policy learning process, this section analyzes the convergence of the three feedback components: demand satisfaction, overcrowding and passenger abandonment, through the training processes of their corresponding algorithms.

Fig. 12 demonstrates the convergence process of demand satisfaction and evacuated passenger number through independent (blue) and collaborative (black) strategies. Subplot Fig. 12a presents the reward associated with demand satisfaction, incorporating both the number of evacuated passengers and their corresponding stranded durations, as defined in Eq. (26). Subplot Fig. 12b focuses solely on the evacuated passenger number. The light lines represent the values for each episode, while the dark lines indicate the moving averages. As shown in the Fig. 12a, the collaborative strategy consistently achieves higher demand satisfaction compared to the independent strategy. The collaborative strategy yields a reward of 64,835, compared to 57,165 for the independent strategy. In Fig. 12b, the collaborative strategy consistently evacuated 29,276 passengers after training, compared to 23,513 passengers evacuated by the independent strategy. Although the collaborative strategy satisfies only 5,763 more passengers than the independent strategy, the demand satisfaction reward improves by 7,670. This improvement highlights the effectiveness of the passenger mode-shifting mechanism in not only increasing the number of satisfied passengers but also reducing stranded duration. By effectively using the residual capacity of underutilized modes, the collaborative approach ensures that more passengers are accommodated, particularly alleviating the pressure on heavily relied-upon modes. The policies of multi-modal collaboration learned by agents will be analyzed in detail in Section 7.3.

Fig. 13 demonstrates the convergence of overcrowding penalties in both strategies (blue representing the independent strategy and black representing the collaborative strategy). The light lines represent the values for each episode, while the dark lines indicate the moving averages. The overcrowding is inevitable since the algorithm begins with the number of stranded passengers exceeding the threshold. During the first 100 training episodes, the collaborative strategy results in higher overcrowding penalties than the independent strategy. However, after training, the collaborative strategy incurs an average overcrowding penalty of only 899, compared to 1,263 for the independent strategy, representing a 44.8% increase. The collaborative strategy coordinates passengers' mode shifting across the multi-modal system, preventing bottlenecks and avoiding excessive loading on high-demand routes or transit modes. This system-wide passenger mode-shifting mechanism ensures that no single mode is overwhelmed, maintaining service stability. In contrast, the independent strategy focuses on individual mode optimization, which can lead to localized overcrowding.

Fig. 14 shows a notable difference in passenger abandonment between the two strategies. The collaborative strategy allows the residual capacities of other modes to be effectively utilized, reducing the burden on heavily used modes and ensuring that fewer regular passengers are displaced. Conversely, the independent strategy attempts to increase demand satisfaction by dispatching additional capacities for evacuation, often at the expense of abandoning more regular passengers. The behavior of sacrificing the services of regular passengers indicates that the agents recognize the reward of demand satisfaction as being higher than the penalty of passenger abandonment. As a result, agents intentionally prioritize de-



(a) Convergence of demand satisfaction under independent and collaborative strategies

(b) Convergence of evacuated passenger number under independent and collaborative strategies

Figure 12: Convergence of demand satisfaction and evacuated passenger number under independent and collaborative strategies

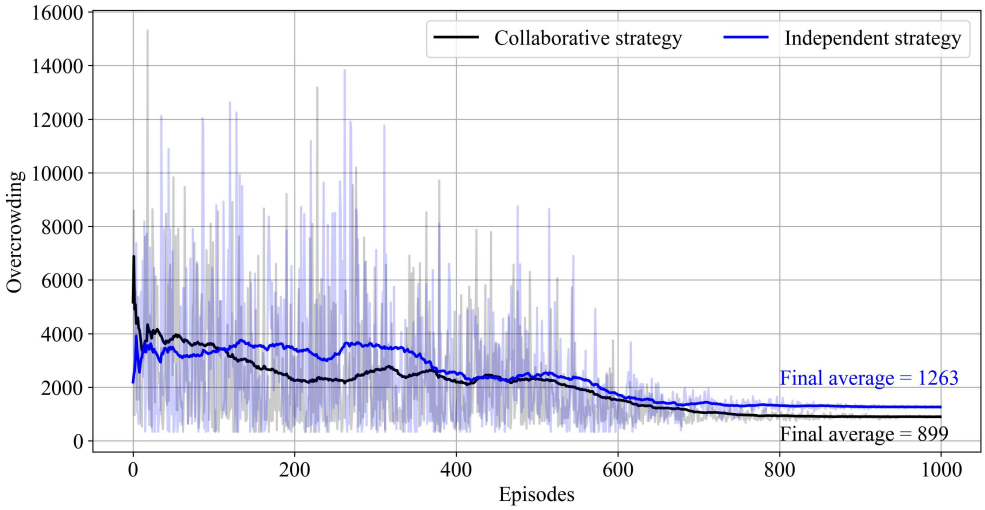


Figure 13: Convergence of overcrowding under independent and collaborative strategies

mand satisfaction over the potential penalty of passenger abandonment. After training, the collaborative strategy abandons 67% less passengers than the independent strategy (8511 vs. 14199).

7.3. Analysis for resilience enhancement and impact mitigation

This section takes the solution with the maximum reward obtained under the full-term environment from 10:00 to 18:40 to illustrate agents' policies on resilience enhancement and impact mitigation. Training hyperparameters and results are given in Section 7.3.1. Section 7.3.2 demonstrates the effectiveness of evacuation strategies in enhancing the resilience under OPF, specifically focusing on the robustness, rapidity and resourcefulness. Section 7.3.3 presents detailed passenger distribution and abandonment at each time step during dynamic multi-modal evacuation.

7.3.1. Training parameters

To analyze agents' capability of resilience enhancement and impact mitigation, the full-term OPF period (10:00-18:40) is taken as the offline training environment. To obtain higher-quality solutions, extensive parameter tuning is conducted through multiple training iterations. Both the learning rate and the number of training episodes are systematically adjusted to identify the best-performing configuration. Specifically, the number of episodes is set as $E = 10,000$, The learning rates are set as 10^9 for central critic, 10^6 for taxi actor, 10^9 for the taxi critic, 10^6 for bus actor, 10^7 for the bus critic, 10^6 for metro actor,

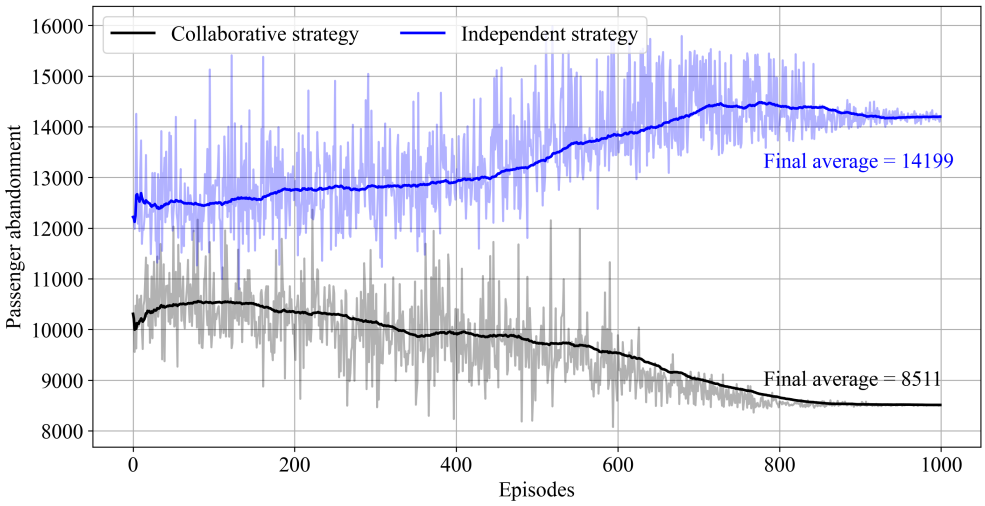


Figure 14: Convergence of passenger abandonment under independent and collaborative strategies

and 10^8 for the metro critic. The results of the maximum reward, final average reward and computation time are shown in Table 4.

7.3.2. Resilience analysis

Based on the solutions with maximum rewards of DTDE and H-CTDE, the resilience performance with independent, collaborative and without evacuation is presented in Fig. 15. The tuples {time, stranded passenger number} of some specific points are labeled above the curves. At the beginning, 5,590 passengers are stranded at 10:00, exceeding the overcrowding threshold. The risk exposure time lasts 8 hours and 45 minutes until 18:40 in the real-world case. The blue line represents the independent evacuation process, while the black line indicates the collaborative evacuation process. By comparing with the real-world condition without evacuation (grey line), both evacuation strategies can enhance the resilience, while the indicators, including the Robustness, Rapidity and resourcefulness, perform differently, which is summarized in Table 5.

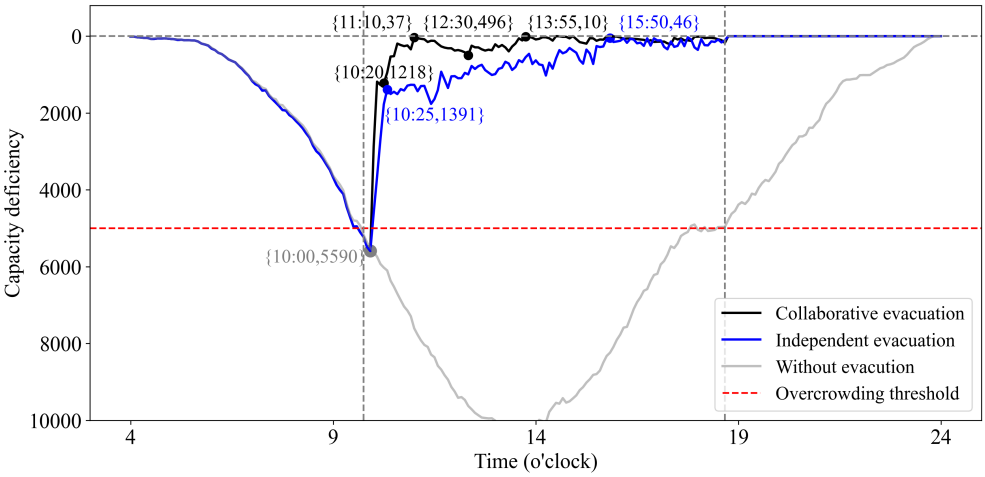


Figure 15: Resilience performance of our approaches and the case without evacuation

Table 5: Resilience metrics of our approaches and the case without evacuation

Strategies	Robustness (psg)	Rapidity-1 (min)	Rapidity-2 (min)	Resourcefulness (psg*min)
Independent	5590	350	25	4400750
Collaborative	5590	70	20	4671790
Without	10,677	780	520	-

Note: Rapidity-1-network clearance time; Rapidity-2-risk exposure time; psg-passenger; min-minute

In the collaborative evacuation process, only 1218 stranded passengers are left within the first 20 minutes (4 time steps), demonstrating strong robustness and short risk exposure time. The system achieves equilibrium with only 37 passengers remaining after 70 minutes (14 time steps), which highlights the rapidity by clearing the network with only 9% of the time compared to the case without evacuation.

While there are some passengers accumulating between 11:10 and 13:55, during the peak OPF period, the maximum accumulation reaches only 496 passengers, which remains within an acceptable range. According to Table 5, the resourcefulness of collaborative strategy is 4,671,790, which is 6% higher than that of independent strategy. Subsequently, agents' main focus is changed to managing arriving passenger flow and preventing overcrowding, which results in consistently maintaining a low level of stranded passengers.

Conversely, in the independent evacuation process, the restoration duration is significantly extended. Although only 1391 stranded passengers are left within the first 25 minutes (5 time steps), demonstrating comparable rapidity (risk exposure time), equilibrium is reached with 46 passengers remaining at 15:50. This process takes 350 minutes, five times longer than the collaborative strategy, highlighting a drawback in network clearance time. The slower restoration curve, indicating weaker resourcefulness, underscores the capacity limitations of individual modes. This limitation forces stranded passengers to wait for subsequent services, thereby prolonging the overall evacuation time span. Although the time to reach equilibrium with the independent strategy is longer than that with the collaborative strategy, the independent strategy still uses only 45% of the time compared to the case without evacuation.

7.3.3. Demand distribution and passenger abandonment across modes

The total demand, satisfied demand and passenger abandonment by each mode at each time step are depicted in Figs. 16, 17 and 18 for the case without evacuation, with independent strategy and with collaborative strategy, respectively. In these figures, the red points represent the total demand. The cumulative bars above the x-axis illustrate the satisfied demand by each mode: green for metro, blue for bus, and orange for taxi. Meanwhile, the bars below the x-axis represent the number of abandoned passengers. This visualization demonstrates the passenger distribution and abandonment across modes, highlighting the dynamic performance of the evacuation strategies in balancing emergency dispatch and impact of evacuation on regular passengers.

Fig 16 illustrates the real-world case without evacuation. The red points represent the total passenger demand at each time step, corresponding to the left axis, while the stacked bars represent the satisfied demand by each mode—green for metro, blue for bus, and yellow for taxi—corresponding to the right axis. Due to the limited capacities, the transit for stranded passengers remains inadequate under OPF. The number of stranded passengers (passenger demand) increases sharply, from 5,590 at 10:00 to 10,677 by 13:30. Although demand decreases after the peak hour, the decline lasts 5 hours until 18:40 with about 5,000 passengers remaining. During this period, demands are evenly distributed across modes, consistently averaging about 104 for metro, 65 for taxis, and 68 for buses at each time step. However, the capacity utilization across modes is not balanced, as metro train capacity is significantly larger than that of buses and taxis. This discrepancy highlights the intense use of bus and taxi services to accommodate the OPF, while the metro system retains some residual capacity. It reflects passengers' real-world mode choice in the absence of proactive multi-modal evacuation strategies. The increasing number of stranded passengers highlights the urgent need for effective evacuation and guidance.

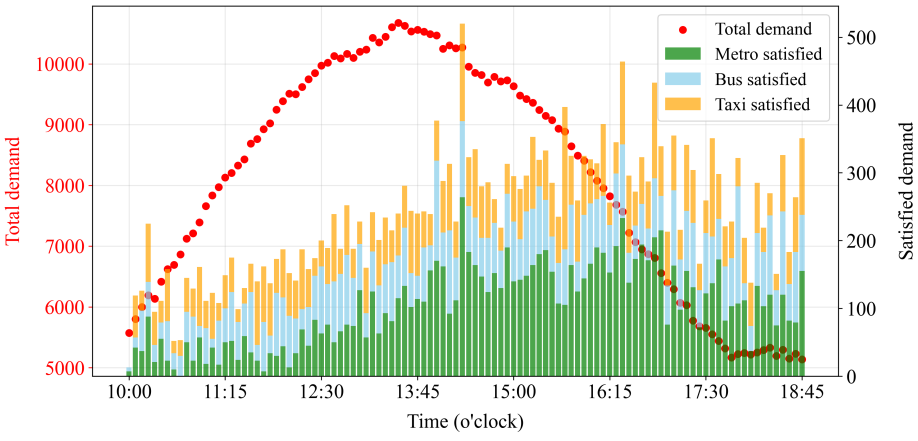


Figure 16: Total passenger demand and satisfied demand without evacuation

Fig. 17 illustrates the dynamic evacuation under independent strategy. The total demand depicts a gradual decrease from 10:00 to 15:30. The metro, with higher capacity, totally accommodates 2,089 passengers within the first two steps. The efficient utilization of metro capacity enables passengers to

depart earlier. This mode also proves effective during the rapid demand increase between 12:00 and 13:00, evacuating an average of 185 passengers per time step. However, the metro's advantage of high capacity diminishes later, which only evacuates an average of 112 passengers per time step. This is because most stranded passengers, who are willing to take the metro, have already left, and newly arriving passengers after that time are less likely to choose the metro. In contrast, the bus and taxi systems are not extensively dispatched at the first two steps, which evacuate the majority of stranded passengers from the third to fifth steps, with buses evacuating 947 passengers and taxis evacuating 2,197 passengers. However, due to the limited available capacity within the bus and taxi systems, the evacuation process is gradual and prolonged. Between 12:00 and 13:00, buses and taxis evacuate an average of 97 and 86 passengers per time step, respectively. However, after 13:00, their average demand satisfaction drops to 89 and 70 passengers per time step. Therefore, the independent strategy effectively groups passengers by mode and allocates additional capacity at specific time steps to high-demand modes, thereby reducing stranded duration. However, since passengers are restricted to their original mode choice, the overall evacuation speed is limited by the lower-capacity modes, like bus and taxi.

According to the bars below the x-axis, in the metro system, the minimal abandonment occurs before 12:00 due to low regular demand. However, 1,391 regular passengers are abandoned between 12:00 and 13:00, when the OPF rapidly increases, coinciding with the metro's peak service. For the bus system, passenger abandonment averages 150 per step due to continuous dispatching, which rises to 200 during high regular demand (12:00–14:00). Taxis experience notable abandonment only in the first five steps when capacity is heavily dispatched. This imbalance highlights the strain on regular services during heavy dispatch periods, emphasizing the need for better coordination of dispatching across modes.

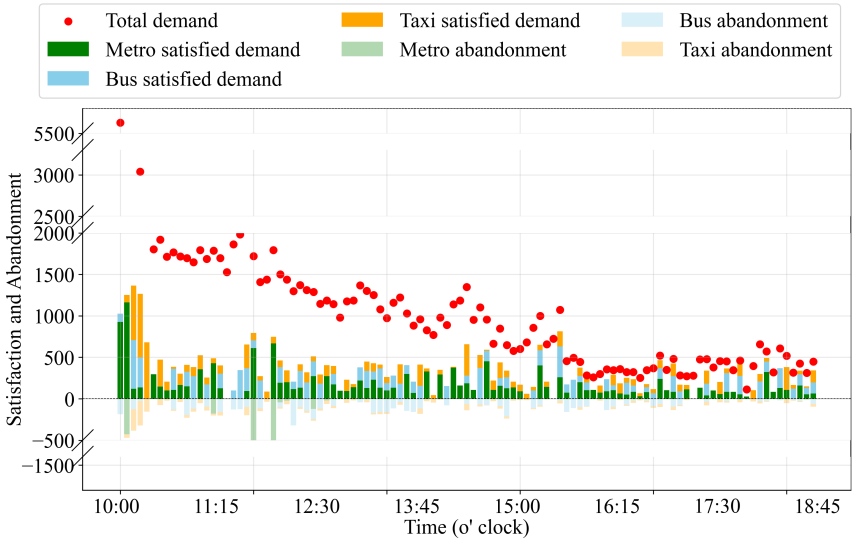


Figure 17: Demand satisfaction and passenger abandonment with independent strategy at each time step

Fig. 18 illustrates the dynamic evacuation under collaborative strategy. Agents swiftly dispatch sufficient capacities at the first time step, evacuating 2,761 passengers in the first time step, with 2572 by metro, 100 by bus and 89 by taxi. Taxis, throughout the period, evacuate an average of 130 passengers per time step. Buses are strategically dispatched only during specific periods, such as regular services off-peak hours (10:00–11:00) and the OPF peak hour (13:00–14:00). During these time periods, buses evacuate an average of 100 stranded passengers but only abandon an average of 79 regular passengers per time step. Buses are not heavily relied upon, possibly due to the higher penalty of passenger abandonment, compared to the average abandonment of 17 for metro and 41 for taxi per time step.

According to the bars below the x-axis, passenger abandonment is clearly visible at the first time step, as agents dispatch extensive capacities to evacuate overwhelming OPF. 1,522 regular metro passengers are abandoned and have to wait for the next train. These passengers are all accommodated in the next time step, just $\Delta t = 5$ minutes later. This is in stark contrast to the independent framework, where abandonment remains higher throughout the period due to passengers' rigid mode choice. The collaborative evacuation's ability to use residual capacities across modes results in minimal ongoing abandonment after the first time step, which effectively mitigates the negative impact on regular passengers.

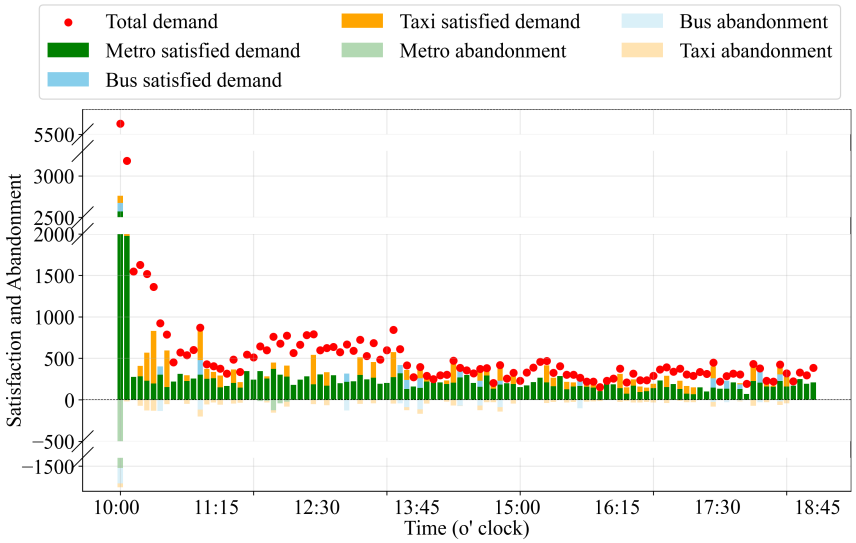


Figure 18: Demand satisfaction and passenger abandonment with collaborative strategy at each time step

7.4. Transferability

One key advantage of reinforcement learning approach lies in its transferability when adapting the pre-trained agents to new but similar scenarios. Based on the result in Section 7.3, the majority of stranded passengers are evacuated within the first several time steps, while the rest of time is focused on maintaining the crowdedness with arriving passengers. This creates a dilemma in reinforcement learning, where increasing the complexity and duration of the training environment may not always benefit online performance (Zhang et al., 2024). To demonstrate the transferability of the MARL in online cases and further illustrate the impact of the length of training datasets, several representative pre-trained MARL agents will be employed in a series of new environments in this section.

The generation of new environments used to evaluate the transferability is introduced in Section 7.4.1. Then, the efficiency of agents pre-trained by peak-hour and full-term environments is compared in Section 7.4.2. Finally, a sensitivity analysis is presented to investigate how various factors, such as the length of training environment, the length of new environment, and the variance between training and new environments, influence the agents' online performance in Section 7.4.3.

7.4.1. Environment generation

To create new environments, a series of variant factors is added to the number of passengers in each passenger group, as referred to $q(p)$ in the agent-based environment, in the training dataset. The variant factors follow a Gaussian distribution with a mean of 0 and standard deviations of 3, 6 and 10. The standard deviations correspond to the 30th, 60th and 90th percentiles of the passenger numbers in the dataset, ensuring that the new environments differ from the training environment at multiple levels. Any generated negative passenger numbers are eliminated from the dataset. The propagation of stranded passengers in these new environments is shown in Fig. 19. With a variant factor of 3, the OPF environment spans from 9:40 to 18:40, with the maximum number of stranded passengers reaching 11,076. With a variant factor of 6, the OPF environment spans from 9:10 to 20:10, with the maximum number of stranded passengers reaching 13,829. With a variant factor of 10, the OPF environment spans from 8:50 to 20:50, with the maximum number of stranded passengers reaching 17,894.

The transferability of a set of representative agents is evaluated in three new environments with different time spans. All environments start at the starting time of the OPF, while their time span are set as 2, 8 and 12 hours, receptively. Longer time spans imply more diverse conditions, which in turn require greater robustness from the agents. The pre-trained agents are evaluated over 100 episodes with a fixed $\epsilon^{\text{noise}} = 0.2$ without updating the weights of neural networks. The best rewards are derived after 3,000 episodes of training as benchmarks.

7.4.2. Transferability efficiency

Firstly, two pre-trained agents are transferred to the new environment with a variant factor of 10, which spans from 8:50-20:50. One agent is trained by the peak-hour environment discussed in Section 7.2.2, while the other is trained by the full-term environment in Section 7.3. Both the independent and collaborative strategies are tested as follows.

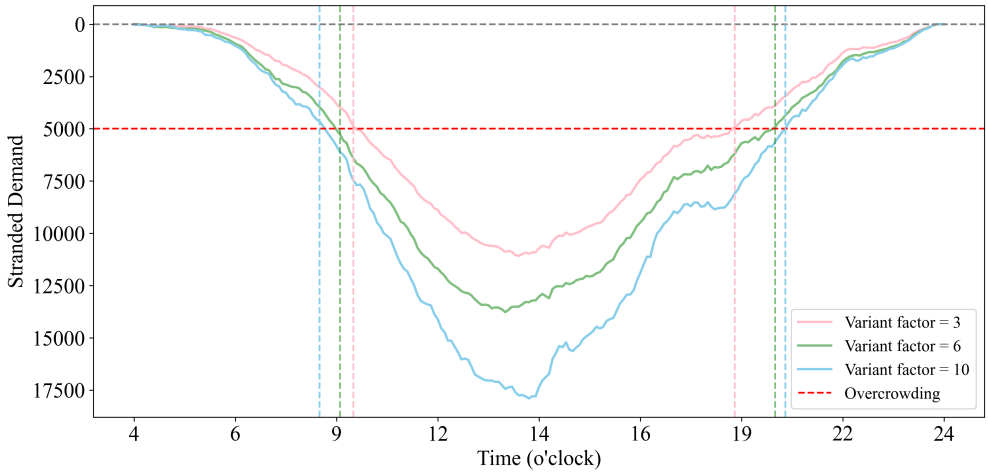


Figure 19: Demand propagation in new environments for online applications

The best, maximum and median rewards, the total computational time per episode and the average computational time per time step of the online application under independent strategy, are listed in Table 6. For agents trained in peak-hour environments, the average gap between the maximum rewards and the best rewards is 4.84%, while the median rewards show an average gap of 7.47%. For agents trained in full-term environments, the average gap between the maximum rewards and the best rewards is 5.75%, while the median rewards show an average gap of 10.23%. The average computational time per time step is similar for both agents trained by peak-hour and full-term environments (0.30s vs 0.29s).

Table 6: Comparison of online application efficiency for independent strategy

Ending	Best	Trained by Peak Hour						Trained by Full Term					
		Reward				Time		Reward				Time	
		Maximum	Gap	Median	Gap	Total	AVG	Maximum	Gap	Median	Gap	Total	AVG
10:50	37781.31	36741.52	2.75%	35657.72	5.63%	6.7s	0.28s	36054.62	4.57%	34375.35	9.03%	7.2s	0.30s
12:50	64953.95	61457.89	5.37%	59730.62	8.04%	14.8s	0.31s	60051.46	7.55%	55628.65	14.36%	13.4s	0.28s
14:50	89675.31	85430.48	4.73%	82946.49	7.51%	20.7s	0.29s	83780.58	6.57%	80219.03	10.55%	20.5s	0.28s
16:50	107709.79	103382.17	4.02%	100469.79	6.72%	29.0s	0.30s	102000.14	5.30%	97342.81	9.57%	29.2s	0.30s
18:50	128752.85	118621.94	7.86%	115289.53	10.46%	36.2s	0.30s	119602.64	7.10%	115400.13	10.35%	34.8s	0.29s
20:50	134583.20	128777.60	4.32%	125900.96	6.45%	44.3s	0.31s	129966.88	3.43%	124440.20	7.54%	44.7s	0.31s
Average	-	-	**4.84%**	-	**7.47%**	-	**0.30s**	-	**5.75%**	-	**10.23%**	-	**0.29s**

Note: Total-total computational time per episode; AVG-average computational time per time step

Similarly, the results of the online application under collaborative strategy are presented in Table 7. For agents trained in peak-hour environments, the average gap between the maximum rewards and the best rewards is 0.91%, while the median rewards show an average gap of 1.55%. For agents trained in full-term environments, the average gap between the maximum rewards and the best rewards is 1.22%, while the median rewards show an average gap of 2.68%. The average computational time per time step is similar for both agents trained by peak-hour and full-term environments (0.39s vs 0.40s).

Table 7: Comparison of online application efficiency for collaborative strategy

Ending	Best	Trained by Peak Hour						Trained by Full Term					
		Reward				Time		Reward				Time	
		Maximum	Gap	Median	Gap	Total	AVG	Maximum	Gap	Median	Gap	Total	AVG
10:50	50111.82	49737.75	0.74%	49298.18	1.63%	9.5s	0.40s	48967.99	2.28%	47674.32	4.86%	9.5s	0.40s
12:50	86881.88	85908.08	1.12%	85314.62	1.80%	19.4s	0.41s	85235.48	1.89%	83721.22	3.64%	18.8s	0.39s
14:50	118211.07	116678.19	1.30%	116035.22	1.84%	27.8s	0.39s	116788.61	1.20%	115261.26	2.49%	26.5s	0.37s
16:50	141379.29	140475.31	0.64%	139592.98	1.26%	37.4s	0.39s	140477.14	0.64%	139266.16	1.49%	38.6s	0.40s
18:50	165541.90	164405.30	0.69%	163464.10	1.25%	49.0s	0.41s	164807.19	0.44%	162698.90	1.72%	47.7s	0.40s
20:50	181650.60	179888.19	0.97%	178839.26	1.55%	52.9s	0.37s	180066.13	0.87%	178242.33	1.88%	53.6s	0.37s
Average	-	-	**0.91%**	-	**1.55%**	-	**0.39s**	-	**1.22%**	-	**2.68%**	-	**0.40s**

Note: Total-total computational time per episode; AVG-average computational time per time step

Agents trained in peak-hour environments (refer to peak-hour-trained agents) exhibit smaller gaps, compared to those trained in full-term environments (refer to full-term-trained agents), when the evacuation time span is no more than 6 hours. According to the results of independent strategy in Table 6, the median gap is 5.63% for peak-hour-trained agents versus 9.03% for full-term-trained agents in the 2-hour environment (end at 10:50). In the 4-hour environment (end at 12:50), the median gaps are 8.04% vs 14.36%, and in the 6-hour environment (end at 14:50), the gaps are 7.51% vs 10.55%. Similar results can be found in the collaborative strategy in Table 7. In the 2-hour environment, the median gap is 1.63%

for peak-hour-trained agents versus 4.86% for full-term-trained agents. In the 4-hour environment, the median gaps are 1.80% vs 3.64%, and in the 6-hour environment, the gaps are 1.84% vs 2.49%. These results show that peak-hour-trained agents are better for handling short-term OPF events, focusing on rapid evacuation and quick restoration of equilibrium.

In contrast, full-term-trained agents are better for longer time spans. When the evacuation time span is within 6 hours, the average median gap is 11.31% (i.e., $(9.03\% + 14.36\% + 10.55\%)/3$) for independent strategies in Table 6 and 3.66% (i.e., (4.86%+3.64%+2.49%)/3) for collaborative strategies in Table 7. For time spans over 6 hours, these gaps decrease to 9.15% (i.e., $(9.57\% + 10.35\% + 7.54\%)/3$) for independent strategies and 1.70% (i.e., $(1.49\% + 1.72\% + 1.88\%)/3$) for collaborative strategies. These results show that full-term-trained agents are better at managing long-term OPF by adapting to the dynamic regular passengers with redundant capacities. The distribution of results from online applications, shown in Fig. 20, supports these findings as well. Peak-hour-trained agents (blue bars) produce more concentrated results, reflecting their focus on rapid responses and efficient handling of the surge in OPF. In contrast, full-term-trained agents (red bars) exhibit broader solution ranges, demonstrating greater adaptability to diverse conditions over a longer period.

The results also demonstrate that the collaborative strategy achieves smaller gaps compared to the independent strategy, highlighting its effectiveness in producing high-quality solutions even in new environments. In addition, the average computational time per time step under collaborative strategy is longer than that under independent strategy (0.40s vs 0.30s). By considering passengers' mode shifting across systems, the collaborative strategy ensures adaptability and robustness, addressing unexpected fluctuations and efficiently balancing demand.

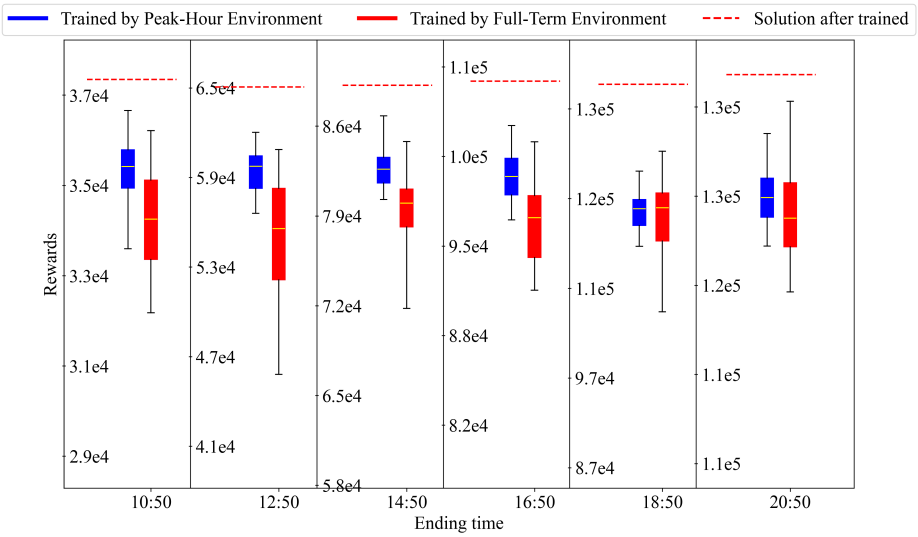
By transferring the trained agents to new environments, no additional training time is required. The average computational time per time step under both independent and collaborative strategies is less than 1 second, though the computational time under collaborative strategy is approximately 0.1 second longer than that under independent strategy. The computational time results demonstrate that our algorithm has ample potential for online application.

7.4.3. Sensitivity analysis

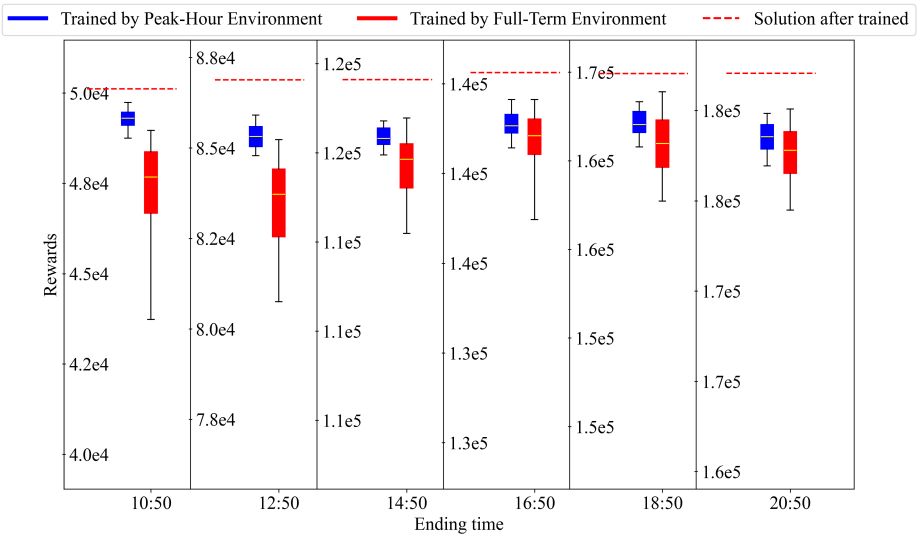
To demonstrate the transferability of our MARL approaches, a sensitivity analysis is conducted under a set of new environments with different time spans and variant factors. Three representative agents are selected from Section 7.2.1. Each one is trained using an environment with distinct time spans: the first two hours before the peak (10:00–12:00), the first four hours covering the peak hours (10:00–14:00), and the full ten-hour period covering the entire OPF period (10:00–20:00). Hereafter, the agents are referred to as the *2h-agent*, *4h-agent* and *10h-agent* for descriptive purposes. These time spans expose the agents to different environmental characteristics. In the first two hours, the environment experiences a rapid increase in passenger demand, allowing the 2h-agent to frequently interact with incoming passengers. The four-hour environment, which starts with a large number of stranded passengers, enables the 4h-agent to focus on evacuating stranded passengers during the initial time steps. Lastly, the ten-hour environment, with its more complex capacity and passenger dynamics, trains the 10h-agent to flexibly adapt its evacuation plans over the long term. By testing these agents in new environments, the impact of the training datasets on their transferability can be demonstrated.

The distributions of rewards across different scenarios for agents under the independent and collaborative strategies are shown in Fig. 21 and Fig. 22, respectively. Each subfigure corresponds to a specific scenario, where the variant factors increase from left to right, and the time span increases from top to bottom. The red line indicates the best reward achieved after training, while the blue line represents the median reward obtained by agents without training, which are referred to as *naive-agent* hereafter. Specifically, the detailed results, including the maximum and median rewards obtained by the agents, benchmark rewards with and without training, as well as the gaps and improvements, are summarized in Table E1 in the Appendix E.

First of all, the phenomena reported in Section 7.4.2 also hold true in the set of new scenarios. The gaps between the best rewards and the median rewards obtained by pre-trained agents are larger under the independent strategy (average of 5.96%) than under the collaborative strategy (average of 2.26%). The advantage of training over a longer time span becomes more evident when the time span of the new environment is extended. Results of agents' transferability for scenarios where the time spans of training environments are longer than those of new environments are summarized in Table E2, and results for



(a) Performance of independent strategy in online applications



(b) Performance of collaborative strategy in online applications

Figure 20: Distribution of results in online applications

scenarios where the time spans of training environments are shorter than those of new environments are summarized in Table E3. When the training time span is longer than that of new environments, for example, the 10h-agent under 2-hour or 8-hour environments, the average gaps are 5.33% under the independent strategy and 1.89% under the collaborative strategy. In contrast, the average gaps are 6.47% under the independent strategy and 2.55% under the collaborative strategy when the time spans of training environments are shorter than those of new environments.

By comparing vertically across the scenarios with different time spans, the 2h-agent shows less adaptability than the other two pre-trained agents, especially as the time spans increase in subfigures from top to bottom. Particularly, when the time span is 12h and variant factor is 3 under independent strategy, the median reward achieved by the 2h-agent is 2.73% lower than that achieved by naive-agent, as shown in Table E1. It demonstrates that 2h-agent's policy becomes overly conservative during longer time spans, leading to suboptimal performance. Results of 4h-agent are summarized in Table E4, and those of *10h-agents* are summarized in Table E5. Under the independent strategy, the 4h-agent shows comparable capability to the 10h-agent with the increase of time span. The average gaps between the median and best rewards are 5.24% for the 4h-agent and 5.27% for the 10h-agent. This is likely because the interactions between environments and agents are isolated among different modes, making the environment simpler under the independent strategy. As a result, the 4h-agent is sufficiently trained, and 10h-agent shows less obvious advantage in long-term environments. Under the collaborative strategy, the 4h-agent is only comparable to 10h-agent under 2-hour environment with average gaps of 2.23% (i.e., (2.32%+2.45%+1.93%)/3) and 2.36% (i.e., (2.80%+2.05%+2.23%)/3), respectively. However, the

average gap of all scenarios by 4h-agent is 2.82%, while that by 10h-agent is 1.77%. This result indicates that the transferability of 4h-agent and 10h-agent diverges significantly as the time span of the new environment increases. Therefore, a longer environment is more effective in training agents, enabling them to adapt better to new environments with a larger number of time steps, particularly under the collaborative strategy.

By comparing horizontally across the scenarios with different variant factors, the improvement of three pre-trained agents becomes more evident compared to the naive-agent, especially as the variant factor increases. Results of agents' transferability under scenarios with variance factors of 3, 6, and 10 are summarized in Tables E6, E7 and E8, respectively. Under the independent strategy, the average improvement for scenarios with variant factors of 3, 6, 10 are 2.58%, 7.09%, and 7.20%, respectively. Under the collaborative strategy, the average improvements are 1.38%, 5.36%, and 4.45%, respectively. It proves the transferability of MARL in online cases to new environments after offline training. Notably, under the collaborative strategy, the superiority of the 10h-agent becomes more pronounced with the increase of variant factor. This indicates that a longer training environment is more effective in training agents to be better adaptable to new environments that are significantly different from the training one.

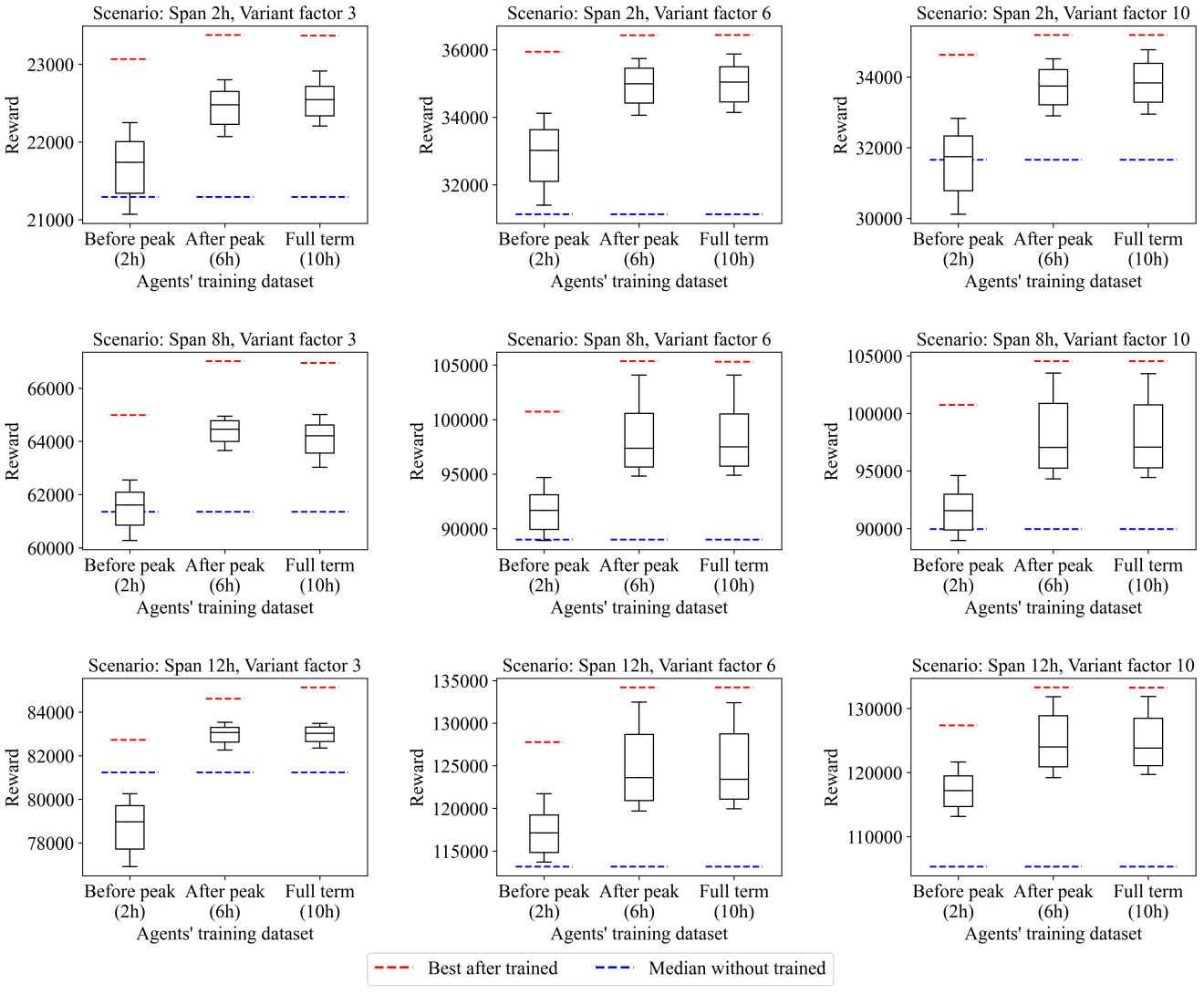


Figure 21: Sensitivity analysis of agents' transferability under independent strategy

The total computational time for each episode and the average computational time per time step under different scenarios and strategies are shown in Table E9. Comparing different scenarios under the same strategy (i.e., collaborative or independent), the average computational time per time step does not differ significantly. Specifically, the average computational time per time step is approximately 0.40s under the collaborative strategy and around 0.30s under the independent strategy. The computational time is primarily influenced by the evacuation strategy, with little impact from the length of the new scenario and the variance between the new scenario and the training environment. This result proves

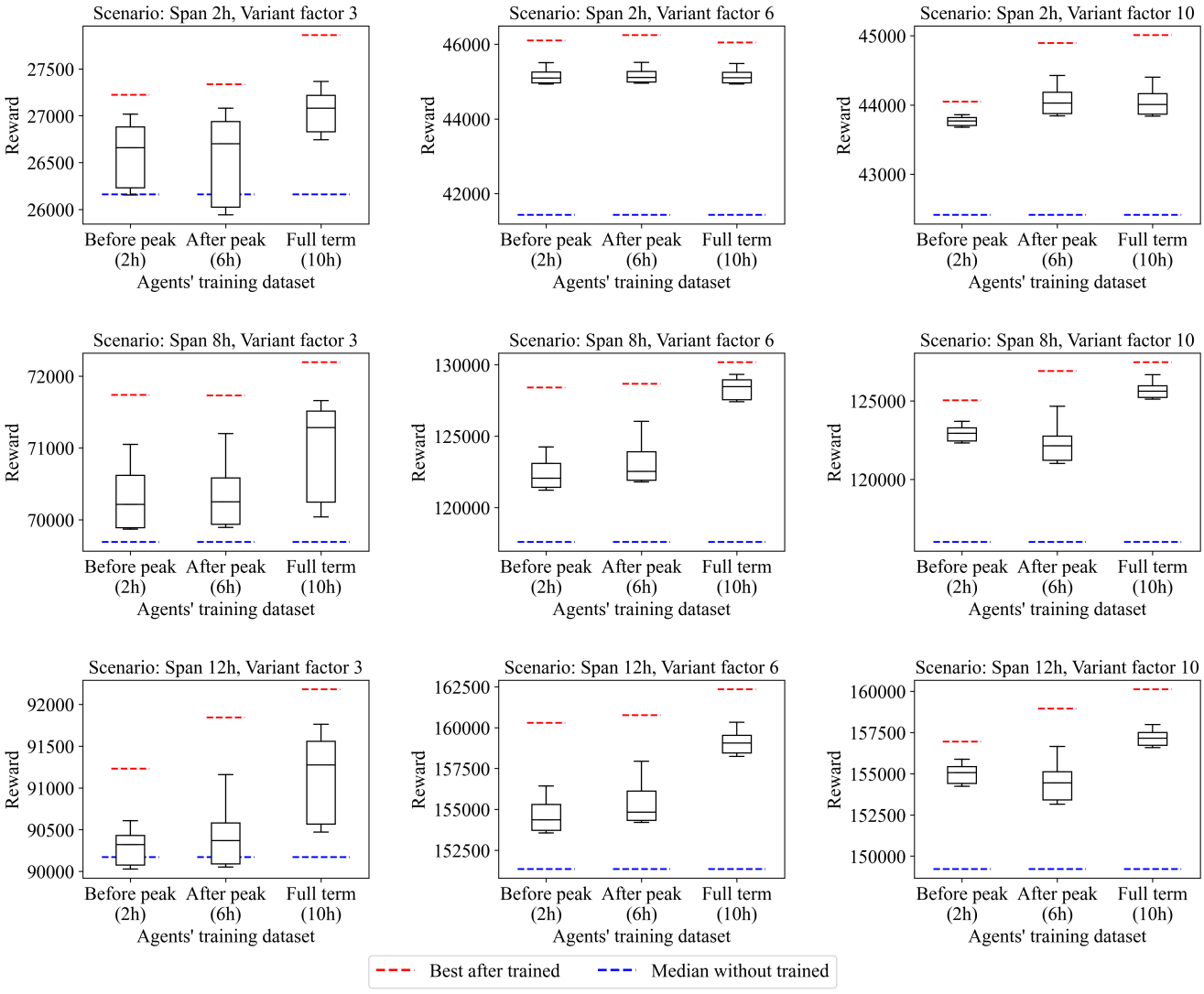


Figure 22: Sensitivity analysis of agents' transferability under collaborative strategy

that our approach is generalizable and can be transferred to a wide range of new online environments.

7.5. Practical, theoretical implications and managerial insights

Practical implications. The proposed online approach can be seamlessly integrated into the emergency management module of a smart city platform, enabling continuous monitoring of passenger accumulation and surrounding transit capacity during mass gatherings or peak periods in rail transit systems. Once passenger volume exceeds a predefined safety threshold, the system can automatically generate dispatching plans in less than one second. For passengers, the rapid responsiveness of our approach ensures timely interventions, which are critical for maintaining safety during emergencies. For transport authorities, such intelligent evacuation mechanisms contribute to the broader objectives of smart cities—ensuring safe, efficient, and adaptive mobility in real time. From an emergency management perspective, the system mitigates the risks of stampedes, severe delays, and secondary disruptions by preventing excessive overcrowding, thereby improving the overall resilience of urban transit systems.

Theoretical implications. Our study bridges the gap between multi-modal evacuation and data-driven decision-making, contributing to the growing body of literature on evacuation strategies, a topic recently identified as a research hotspot (Bergantino et al., 2024). The use of online decision-making effectively addresses uncertainties regarding the onset and duration of OPF, which have traditionally constrained evacuation planning. Furthermore, our approach redefines resilience as a real-time interactive metric rather than a retrospective one, addressing a key theoretical gap in resilience research highlighted by Wei et al. (2024). This framework also serves as a foundation for data-driven resilience evaluation, in line with recent advancements by Dui et al. (2023) and Knoester et al. (2024).

Managerial insights. This study highlights the critical role of multi-modal collaboration in evacuation management, emphasizing the importance of both flexibility and operational efficiency. The distinct

characteristics of each transport mode, as well as the potential impact of evacuation activities on regular services, must not be overlooked. Based on the analysis in Section 7.3, the following managerial insights are derived:

- Evacuation should be continuous. Effective evacuation requires ongoing monitoring of passenger accumulation and capacity availability until the end of the peak period. Evacuating only the initially accumulated passengers is insufficient, as overcrowding may still occur if new passengers continue to arrive and are not accounted for.
- Proactive passenger redistribution is essential. Evacuation managers should proactively guide passengers to different modes based on real-time capacity availability and travel efficiency. If passengers persist in their original mode choices, clearance of the OPF will be constrained by bottleneck modes with limited capacity, limiting the overall improvement in evacuation efficiency.
- Consider the impact on regular services. The impact of evacuation on regular services should be carefully assessed. Each mode exhibits different levels of residual capacity throughout the day—buses, taxis, and metros vary in the number of abandoned regular passengers when their capacities are dispatched for evacuation. To minimize impact, evacuation resources should be dispatched during off-peak periods or prioritized from modes experiencing lower regular passenger demand.

These insights support the development of more adaptive and balanced evacuation strategies that align with real-world operational constraints and passenger behavior.

8. Conclusion

This paper proposed an online multi-modal evacuation framework based on heterogeneous MARL, targeting dynamic OPF scenarios. A novel data-driven agent-based environment was developed to capture real-time interactions between passenger growth and capacity availability. Two coordination strategies were implemented: an independent strategy under the DTDE framework and a collaborative strategy under a customized H-CTDE algorithm. Resilience metrics—robustness, rapidity, and resourcefulness—were transformed into demand-responsive feedback mechanisms to guide proactive evacuation planning.

Comparative results demonstrated that the proposed MARL algorithms outperformed both the GA and MADDPG algorithms in terms of computation time and solution quality. H-CTDE achieved 7–11% and 2–4% higher rewards than MADDPG and GA, respectively, while reducing computation time by 22–48% compared to GA. The collaborative and independent strategies restored equilibrium using only 9% and 55% of the time required in the case without evacuation, respectively. In online settings, pre-trained agents consistently maintained solution gaps within 10% across most new environments, with computational times ranging from 0.3 to 0.4 seconds per time step. These results validated our approaches for online resilient evacuation planning across heterogeneous transit modes.

Future work can extend the proposed framework by incorporating train rescheduling and flexible bus routing to further reduce passengers' en-route delays. Addressing demand uncertainty caused by self-evacuation is also essential. To systematically balance the competing priorities of resilience enhancement (e.g., minimizing stranded passengers through proactive capacity allocation) and impact mitigation (e.g., preserving regular service quality by limiting evacuation-driven disruptions), multi-objective reinforcement learning could be employed to optimize Pareto-optimal solutions. This approach would enable adaptive policy development by dynamically weighting objectives based on real-time system states and predefined priority rules. Additionally, the H-CTDE framework may be developed as a generalized paradigm to validate its broader applicability in heterogeneous multi-agent settings. Integrating the data-driven agent-based environment with microscopic simulation platforms such as SUMO or AnyLogic would provide sufficient training scenarios for our approach. Lastly, improving data quality remains critical where current limitations, such as the lack of precise regular passenger counts in taxi GPS data, may lead to underestimated system impacts.

References

Abdelgawad, H., Abdulhai, B., 2010. Managing large-scale multimodal emergency evacuations. Journal of Transportation Safety & Security 2, 122–151.

Abdelgawad, H., Abdulhai, B., 2012. Large-scale evacuation using subway and bus transit: Approach and application in city of toronto. Journal of Transportation Engineering 138, 1215–1232.

Bergantino, A.S., Gardelli, A., Rotaris, L., 2024. Assessing transport network resilience: empirical insights from real-world data studies. Transport Reviews , 1–24.

Bešinović, N., Nassar, R., Szymula, C., 2022. Resilience assessment of railway networks: Combining infrastructure restoration and transport management. Reliability Engineering & System Safety 224, 108538.

Bruneau, M., Chang, S., Eguchi, R., Lee, G., O'Rourke, T., Reinhorn, A., Shinozuka, M., Tierney, K., Wallace, W., von Winterfeldt, D., 2003. A framework to quantitatively assess and enhance the seismic resilience of communities. Earthquake Spectra 19, 733–752.

Chang, K., Hsu, C., Su, W., 2024. An agent-based simulation framework for emergency evacuations from toxic gas incidents and an empirical study in Taiwan. Computers & Operations Research 167, 106645.

Diaz, R., Acero, B., Behr, J., Hutton, N., 2023. Impacts of household vulnerability on hurricane logistics evacuation under COVID-19: The case of U.S. Hampton Roads. Transportation Research Part E: Logistics and Transportation Review 176, 103179.

Dui, H., Liu, K., Wu, S., 2023. Data-driven reliability and resilience measure of transportation systems considering disaster levels. Annals of Operations Research .

Goerigk, M., Deghdak, K., T'Kindt, V., 2015. A two-stage robustness approach to evacuation planning with buses. Transportation Research Part B: Methodological 78, 66–82.

Gronauer, S., Diepold, K., 2022. Multi-agent deep reinforcement learning: a survey. Artificial Intelligence Review 55, 895–943.

Gu, Y., Chen, A., 2023. Modeling mode choice of customized bus services with loyalty subscription schemes in multi-modal transportation networks. Transportation Research Part C: Emerging Technologies 147, 104012.

Gu, Y., Fu, X., Liu, Z., Xu, X., Chen, A., 2020. Performance of transportation network under perturbations: Reliability, vulnerability, and resilience. Transportation Research Part E: Logistics and Transportation Review 133, 101809.

Hao, S., Song, R., He, S., 2024. Robust optimization modelling of passenger evacuation control in urban rail transit for uncertain and sudden passenger surge. International Journal of Rail Transportation 13, 151–170.

Hu, X., Huang, J., Long, X., Wang, J., 2024. Service vulnerability assessment of China's high-speed train network: A simulation approach. Reliability Engineering & System Safety 245, 109971.

Huang, D., Peng, X., Liu, Z., Chen, J., Liu, P., 2024. Understanding the route choice behavior of metro passenger using the smartphone applications. Travel Behaviour and Society 36, 100804.

Jiang, J., Tu, W., Kong, H., Zeng, W., Zhang, R., Konecny, M., 2022. Large-scale urban multiple-modal transport evacuation model for mass gathering events considering pedestrian and public transit system. IEEE Transactions on Intelligent Transportation Systems 23, 23059–23069.

Jiang, X., Long, S., Liu, Y., Meng, F., Luan, X., 2025. Integrated optimization of vehicle scheduling and passenger assignment of demand-responsive transit and conventional buses under urban rail transit disruptions. Journal of Transportation Engineering, Part A: Systems 151, 04025037.

Jiang, Z., Fan, W., Liu, W., Zhu, B., Gu, J., 2018. Reinforcement learning approach for coordinated passenger inflow control of urban rail transit in peak hours. Transportation Research Part C: Emerging Technologies 88, 1–16.

Jiang, Z., Gu, J., Fan, W., Liu, W., Zhu, B., 2019. Q-learning approach to coordinated optimization of passenger inflow control with train skip-stopping on a urban rail transit line. Computers & Industrial Engineering 127, 1131–1142.

Jin, J., Tang, L., Sun, L., Lee, D., 2014. Enhancing metro network resilience via localized integration with bus services. Transportation Research Part E: Logistics and Transportation Review 63, 17–30.

Knoester, M.J., Bešinović, N., Afghari, A.P., Goverde, R.M., Van Egmond, J., 2024. A data-driven approach for quantifying the resilience of railway networks. Transportation Research Part A: Policy and Practice 179, 103913.

Land Transport Guru, 2024. 2024 East-West MRT Line disruption (25–30 Sep). https://landtransportguru.net/2024-ewl-disruption/.

Li, B., Yao, E., Yamamoto, T., Huan, N., Liu, S., 2020a. Passenger travel behavior analysis under unplanned metro service disruption: Using stated preference data in Guangzhou, China. Journal of Transportation Engineering, Part A: Systems 146, 04019069.

Li, B., Yao, E., Yamamoto, T., Tang, Y., Liu, S., 2020b. Exploring behavioral heterogeneities of metro passenger's travel plan choice under unplanned service disruption with uncertainty. Transportation Research Part A: Policy and Practice 141, 294–306.

Li, W., Ni, S., 2022. Train timetabling with the general learning environment and multi-agent deep reinforcement learning. Transportation Research Part B: Methodological 157, 230–251.

Li, Z., Yin, J., Chai, S., Tang, T., Yang, L., 2023. Optimization of system resilience in urban rail systems: Train rescheduling considering congestions of stations. Computers & Industrial Engineering 185, 109657.

Liang, J., Zang, G., Liu, H., Zheng, J., Gao, Z., 2023. Reducing passenger waiting time in oversaturated metro lines with passenger flow control policy. Omega 117, 102845.

Liu, B., Pu, Z., Pan, Y., Yi, J., Liang, Y., Zhang, D., 2023. Lazy agents: a new perspective on solving sparse reward problem in multi-agent reinforcement learning, in: Proceedings of the 40th International Conference on Machine Learning, pp. 21937–21950.

Liu, E., Barker, K., Chen, H., 2022. A multi-modal evacuation-based response strategy for mitigating disruption in an intercity railway system. Reliability Engineering & System Safety 223, 108515.

Ma, R., Huang, A., Jiang, Z., Wang, Z., Luo, Q., Zhang, X., 2024. A data-driven optimal method for massive passenger flow evacuation at airports under large-scale flight delays. Reliability Engineering & System Safety 245, 109988.

Ma, Z., Koutsopoulos, H., Chen, Y., Wilson, N., 2019. Estimation of denied boarding in urban rail systems: Alternative formulations and comparative analysis. Transportation Research Record 2673, 771–778.

Matherly, D., Murray-Tuite, P., Wolshon, B., 2015. Traffic Management for Planned, Unplanned, and Emergency Events. John Wiley & Sons, Ltd. chapter 16. pp. 599–636.

Powell, W., 2011. Approximate dynamic programming. 2nd ed., John Wiley & Sons, Inc., New Jersey, USA.

Shao, K., Zhu, Y., Zhao, D., 2019. StarCraft micromanagement with reinforcement learning and curriculum transfer learning.

IEEE Transactions on Emerging Topics in Computational Intelligence 3, 73–84.
Su, H., Wang, X., Liu, W., Zhang, X., Xu, M., 2024. Modelling and accelerating the passenger evacuation process in a bi-modal system with bus and e-hailing modes after mass gathering events. Travel Behaviour and Society 35, 100713.
Su, H., Wang, X., Xu, M., Zhang, X., 2025. Boarding space design for passenger evacuation with bus and e-hailing services under a surge in traffic demand. Travel Behaviour and Society 40, 101021.
Sutton, R., Barto, A., 2018. Reinforcement learning: An introduction. 2nd ed., MIT Press, Cambridge, MA, USA.
Tang, J., Xu, L., Luo, C., Ng, T., 2021. Multi-disruption resilience assessment of rail transit systems with optimized commuter flows. Reliability Engineering & System Safety 214, 107715.
Teichmann, D., Dorda, M., Sousek, R., 2021. Creation of preventive mass evacuation plan with the use of public transport. Reliability Engineering & System Safety 210, 107437.
Transportation Research Board, 2008. The Role of Transit in Emergency Evacuation: Special Report 294. The National Academies Press, Washington, DC.
Turner, D., Evans, W., Kumlachew, M., Wolshon, B., Dixit, V., Sisiopiku, V., Islam, S., Anderson, M., 2010. Issues, practices, and needs for communicating evacuation information to vulnerable populations. Transportation Research Record: Journal of the Transportation Research Board 2196, 159–167.
Wang, J., Cai, J., Yue, X., Suresh, N., 2021. Pre-positioning and real-time disaster response operations: Optimization with mobile phone location data. Transportation Research Part E: Logistics and Transportation Review 150, 102344.
Wang, L., Jin, J.G., 2025. Utilizing and optimizing non-disrupted lines for evacuating passengers in urban rail transit networks during disruptions. Transportmetrica B: Transport Dynamics 13, 2440588.
Wang, X., Chen, X., Xie, C., Cheong, T., 2024. Coordinative dispatching of shared and public transportation under passenger flow outburst. Transportation Research Part E: Logistics and Transportation Review 189, 103655.
Wang, X., D'Ariano, A., Su, S., Tang, T., 2023. Cooperative train control during the power supply shortage in metro system: A multi-agent reinforcement learning approach. Transportation Research Part B: Methodological 170, 244–278.
Wang, Y., Mo, P., Chen, J., Liu, Z., 2025. Managing oversaturation in BRT corridors: A new approach of timetabling for resilience enhancement using a tailored integer L-shaped algorithm. European Journal of Operational Research 320, 219–238.
Wei, Y., Yang, X., Xiao, X., Ma, Z., Zhu, T., Dou, F., Wu, J., Chen, A., Gao, Z., 2024. Understanding the resilience of urban rail transit: Concepts, reviews and trends. Engineering , S2095809924001280.
Xu, X., Chen, A., Xu, G., Yang, C., Lam, W.H., 2021. Enhancing network resilience by adding redundancy to road networks. Transportation Research Part E: Logistics and Transportation Review 154, 102448.
Yang, X., Ban, X.J., Mitchell, J., 2018. Modeling multimodal transportation network emergency evacuation considering evacuees' cooperative behavior. Transportation Research Part A: Policy and Practice 114, 380–397.
Ying, C., Chow, A., Chin, K., 2020. An actor-critic deep reinforcement learning approach for metro train scheduling with rolling stock circulation under stochastic demand. Transportation Research Part B: Methodological 140, 210–235.
Ying, C., Chow, A., Nguyen, H., Chin, K., 2022. Multi-agent deep reinforcement learning for adaptive coordinated metro service operations with flexible train composition. Transportation Research Part B: Methodological 161, 36–59.
Yu, J., Laharotte, P., Han, Y., Leclercq, L., 2023. Decentralized signal control for multi-modal traffic network: A deep reinforcement learning approach. Transportation Research Part C: Emerging Technologies 154, 104281.
Zhang, L., Meng, Q., Wang, H., Yu, B., 2025. Joint bus dispatching and bus bridging timetabling for mass rapid transit disruption management. Transportation Research Part B: Methodological 196, 103215.
Zhang, P., Sun, H., Qu, Y., Yin, H., Jin, J., Wu, J., 2021. Model and algorithm of coordinated flow controlling with station-based constraints in a metro system. Transportation Research Part E: Logistics and Transportation Review 148, 102274.
Zhang, Z., Chen, Y., Lee, J., Du, S., 2024. Settling the sample complexity of online reinforcement learning, in: Proceedings of 37th Conference on Learning Theory, pp. 5213–5219.
Zhong, G., Wan, X., Zhang, J., Yin, T., Ran, B., 2017. Characterizing passenger flow for a transportation hub based on mobile phone data. IEEE Transactions on Intelligent Transportation Systems 18, 1507–1518.
Zhou, Y., Wang, J., Yang, H., 2019. Resilience of transportation systems: Concepts and comprehensive review. IEEE Transactions on Intelligent Transportation Systems 20, 4262–4276.
Zhu, Y., Jin, J.G., Wang, H., 2024. Path-choice-constrained bus bridging design under urban rail transit disruptions. Transportation Research Part E: Logistics and Transportation Review 188, 103637.

Appendix A. Table of notations

Table A1: Notations of decision variables

Symbol	Explanation
a^t_{taxi}	The number of taxis dispatched for evacuation at time step t.
a^t_{bus}	The number of buses dispatched for evacuation at time step t.
a^t_{metro}	The number of dispatched passenger inflow at the metro station at time step t.
a^t_m	A general indicator of dispatched capacity of mode m at time step t.

Table A2: Notations of objectives

Symbol	Explanation
R	Overall objective.
$R^{m,t}_{\text{resilience}}$	Objective of resilience enhancement for mode m at time step t.
$R^{m,t}_{\text{abandon}}$	Objective of impact mitigation on regular services for mode m at time step t.
$R^{m,t}_{\text{demand}}$	Reward of demand satisfaction for mode m at time step t.
$R^{m,t}_{\text{crowd}}$	Penalty of overcrowding for mode m at time step t.
R^t_m	Objective value for mode m at time step t.

Table A3: Notations of indices and sets

Symbol	Explanation
t	Index of time step.
m	Index of mode.
p	Index of passenger group.
r	Index of emergency bus route.
v	Index of vehicle in taxi and bus system.
i, i', j, j'	Indices of metro stations.
$\underline{i}, \bar{i}$	Indices of original and terminal station, respectively.
i^*	Index of the station within the OPF area.
v_t	Index of train which arrives at the metro station i^* at time step t.
T	Set of time steps.
M	Set of modes, where $M = \{\text{taxi}, \text{bus}, \text{metro}\}$.
P^t_m	Set of evacuated passenger groups of mode m at time step t.
$\bar{P}^t_m$	Set of stranded passenger groups for mode m at time step t.
$\bar{P}^t$	Set of total stranded passenger groups at time step t.
γ	Set of emergency bus routes.
γ^t	Set of emergency bus routes with buses assigned at time step t.
D^t_r	Set of destinations on route r at time step t.
$P^t_{\text{bus},r}$	Set of evacuated passenger groups on emergency bus route r at time step t.
V^t_{taxi}	Set of dispatchable taxis until time step t.
V^t_{bus}	Set of dispatchable buses until time step t.
$\bar{V}^t_{\text{taxi}}, \bar{V}^t_{\text{bus}}$	Sets of dispatched taxis and buses at time step t, respectively.
V_{metro}	Set of metro trains.
I	Set of metro stations.
$\hat{P}^t_m$	Set of initial evacuated passenger groups.

Table A4: Notations of modeling parameters

Symbol	Explanation
t_{start}, t_{end}	Starting and ending time of the online multi-modal evacuation, respectively.
$\Delta t, \Delta \tilde{t}$	Length of a time step and an interval between timestamps.
c_m	Vehicle capacity of mode m.
n_m^t	The number of dispatchable capacities of mode m at time step t.
$\tilde{t}$	The timestamp of a taxi GPS data.
$\hat{\rho}_{\text{taxi}}(v, \tilde{t})$	A binary indicator indicating whether the taxi v is occupied at each timestamp $\tilde{t}$.
$\rho_{\text{taxi}}(v)$	A binary indicator indicating whether a regular passenger is abandoned when taxi v is dispatched.
$\hat{p}$	Index of a regular bus passenger.
$v(\hat{p})$	The ID of the bus boarded by regular passenger $\hat{p}$ takes.
$\rho_{\text{bus}}(v)$	The number of regular passengers abandoned when bus v is dispatched.
$w(v_t, i, j)$	Passenger demand on train v_t from station i to station j.
$n_{\text{metro}}^{t,i}$	Available capacity of train v_t arriving at station i.
$\rho_{\text{metro}}(v_t, i)$	The number of regular passengers that are abandoned on train v_t at station i.
$q(p)$	Passenger number of a passenger group p.
$i(p)$	Estimated disembarkation station of passenger group p if it takes metro.
$e(p)$	Entry time of passenger group p.
$\tilde{l}(p)$	Original leaving time of passenger group p.
$d(p)$	Destination of passenger group p.
$k(p)$	Total distance for the entire journey of passenger group p.
$\tilde{m}(p)$	Original mode choice of passenger group p.
$S_m^t, S_m'^t$	State variables for agents under independent and collaborative strategies, respectively.
δ_m^t, δ^t	Single-modal and total stranded passenger number at time step t, respectively.
$h(p, t)$	Stranded duration of passenger group p which is evacuated at time step t.
$H(p, t)$	Stranded duration factor of passenger group p which is evacuated at time step t.
ξ	A relatively small number.
$u_m(p)$	Trip cost of passenger group p when traveling by mode m.
ϵ_0^m	Mode loyalty factor for passengers' willingness to remain with their original mode.
$\hat{k}(p)$	Last-mile trip term of passenger group p.
$\mathbb{L}_m(p)$	Probability model for choosing mode m by passenger group p for evacuation.
$\hat{n}_m^t$	Residual capacity after passengers are distributed based on their original mode choice for mode m at time step t.
$\hat{n}_{\text{bus},r}^t$	Residual capacity on emergency bus route r at time step t.
$\alpha_1, \alpha_2, \alpha_3$	Coefficients in the feedback functions.
$\epsilon_1^m, \epsilon_2^m, \epsilon_3^m$	Coefficients of unit costs.

Table A5: Notations of training parameters

Symbol	Explanation
π_m, π_m'	Actor and target actor networks for mode m, respectively.
Q_m, Q_m'	Local critic and target local critic networks for mode m, respectively.
Q, Q'	Central critic and target central critic networks, respectively.
$\theta_m^\pi, \theta_m^{\pi'}$	Parameters for actor and target actor networks, respectively.
$\theta_m^Q, \theta_m^{Q'}$	Parameters for local critic and target local critic networks, respectively.
$\theta^Q, \theta^{Q'}$	Parameters for central critic and target central critic networks, respectively.
$\hat{a}_m^t$	Proportion of dispatched capacities of mode m at time step t.
S^t	Vector of combined states as the input of central critic network at time step t.
a^t	Vector of combined actions as the input of central critic network at time step t.
R^t, R'^t	Global reward and target global Q-value at time step t, respectively.
$R_{\text{local}}^{m,t}, R_{\text{local}}^{'m,t}$	Local reward and target local Q-value at time step t, respectively.
e, E	Current and total number of training episodes, respectively.
$\epsilon^{\text{noise}}, \underline{\epsilon}^{\text{noise}}$	Noise factor and minimum noise added to agents' actions, respectively.
RB, B	Reply buffer and the size of a batch of tuples, respectively.
L	Loss function for training the critic networks.
J	Objective function for training the actor networks.
$\alpha_d, \alpha_Q^m, \alpha_Q, \alpha_\pi^m, \alpha_s$	Coefficients for discount factor, learning rates for local critic, central critic and actor networks, and soft updating rate for target networks, respectively.

Appendix B. Processing of multi-source data for environment formulation

This section introduces the processing method for the multi-source dataset used to formulate a data-driven transit and passenger environment for MARL training. The environment on the transit side is based on taxi GPS, bus smart card, bus AVL, and metro OD demand datasets, and the environment on the passenger side is based on the mobile dataset.

Appendix B.1. Taxi GPS dataset

The taxi transit environment is driven by taxi GPS data to capture the positioning and occupancy conditions of taxis. A sample of the taxi GPS dataset is shown in Table B6. Let V_{taxi} denote the overall GPS dataset throughout the OPF period. Each taxi has a specific taxi ID v, listed in the first column. The location is recorded with longitude $Lon(v,\tilde{t})$ and latitude $Lat(v,\tilde{t})$. A column of timestamp $\tilde{t}$ in the dataset records the time instance of each record. In practice, records are updated every 30 seconds, i.e., $\Delta\tilde{t} = 30s$. The occupancy status $\hat{\rho}(v,\tilde{t})$ is recorded in the last column, where 1 means that the taxi is occupied and 0 means that it is empty.

Table B6: Sample of taxi GPS data

Taxi ID	Timestamp	Longitude	Latitude	Occupancy
30000XXXX	12:18:04	108.9339	34.37231	0
30000XXXX	12:18:34	108.937215	34.37301	1
30000XXXX	12:19:04	108.938705	34.3734	1
30000XXXX	12:19:34	108.93822	34.37523	1
30000XXXX	12:20:04	108.936007	34.375725	1

By setting a searching area around the center of the OPF region, denoted by a coordinate (Lon^*, Lat^*), only unoccupied taxis (i.e., occupancy=0) within this designated searching area are considered dispatchable. The real-time distance of each taxi v to the center at timestamp $\tilde{t}$ is $k(v,\tilde{t})$, which is calculated by the Haversine formula [1] as Eq. (B1).

$$k(v,\tilde{t}) = 2\theta^R \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{Lon(v,\tilde{t}) - Lon^*}{2}\right) + \cos(Lat^*)\cdot\cos(Lat(v,\tilde{t}))\cdot\sin^2\left(\frac{Lat(v,\tilde{t}) - Lat^*}{2}\right)}\right), \quad \text{(B1)}$$

where $\theta^R = 6371$km denotes the radius of earth. The set of dispatchable taxis at time step t can be expressed as Eq. (B2).

$$V^t_{taxi} = \{v \in V_{taxi} | k(v,\tilde{t}) \le \theta^B, \hat{\rho}(v,\tilde{t}) = 0, \tilde{t} \in [t, t+\Delta t]\}, \quad \text{(B2)}$$

where θ^B denotes the radius of the searching area. In practice, the searching radius could be 3km or more for ride-hailing(Su et al., 2024). Thus, the value of searching radius is set as 3km in our study.

Appendix B.2. Bus smart card dataset

The bus transit environment is driven by the AVL dataset to capture bus positioning and smart card data to capture the occupancy conditions.

The AVL dataset records the longitude, latitude, and timestamp of each bus, similar to the taxi GPS dataset. By substituting the coordinates and area of the bus depot into Eqs. (B1) and (B2), the set of dispatchable buses V^t_{bus} at time step t can be determined.

The bus smart card dataset records the occupancy of passengers on each bus. A sample of the bus smart card dataset is shown in Table B7. Each row records a passenger's ID $\hat{p}$, ID $v(\hat{p})$ for the bus taken by passenger $\hat{p}$, timestamp (i.e., boarding time), and the station where passenger $\hat{p}$ boards bus $v(\hat{p})$. The regular passengers associated with the dispatched bus can be revealed by Eq. (10).

Table B7: Sample of bus smart card data that will be updated dynamically

Passenger ID	Bus ID	Timestamp	Station
125XXX	151XXX	06:04:16	Beishaomen
351XXX	151XXX	06:04:20	Beishaomen
498XXX	151XXX	06:09:08	Fangxincun
165XXX	151XXX	06:10:21	Yahehuayuan
794XXX	151XXX	06:24:16	Wenjinglukou

Appendix B.3. Metro OD dataset

The metro transit environment is driven by time-dependent OD demand data to capture the occupancy of trains. A sample of the OD demand matrix at time step t is shown in Table B8 , where the columns denote the

[1] *https://en.wikipedia.org/wiki/Haversine_formula*

origin stations and rows denote the destination stations. Each number denotes the demand volume $w(v_t, i, j)$ from metro station i to j on train v_t that arrives at the OPF station at time step t. The values in the matrix are updated with time step t. The dataset is used to calculate Eqs. (11) and (12).

Table B8: Sample of metro OD demand

	Houweizhai	Sanqiao	Zaohe	Zaoyuan	Hanchenglu	...
Beikezhan	58	23	22	9	86	...
Beiyuan	2	3	2	1	7	...
Yundonggongyuan	41	14	9	6	26	...
Xingzhengzhongxin	53	5	15	5	13	...
Fengchengwulu	50	16	13	7	21	...
Shitushuguan	78	12	14	4	20	...
...	...	...	...	...	...	...

Appendix B.4. Mobile dataset

The mobile dataset captures individual passenger mobility. A sample of mobile dataset is presented in Table B9. Each passenger group p has a record of its entry time $e(p)$, original leaving time $\tilde{l}(p)$, total distance for the entire journey $k(p)$, destination $d(p)$, original mode choice $\tilde{m}(p)$, passenger number $q(p)$, and debarking station $i(p)$.

Table B9: Sample of mobile dataset

ID	Entry time	Leave time	Destination	Distance	Mode	Number	Disembarkation station	Age	Gender
1	10:00:37	11:59:30	Weiyang	3824m	taxi	5	-	1	M
2	10:00:37	13:59:30	Lianhu	16807m	bus	5	-	3	F
3	10:00:47	17:51:30	Weiyang	4123m	metro	1	Beiyuan	5	M

The passengers who are stranded in the OPF area during the time interval $[t, t+\Delta t]$ are deemed stranded. Passengers who belong to the set of stranded passengers at time step t are presented by Eq. (B3).

$$\bar{P}^t = \{p | t \le e(p) \le t + \Delta t\}. \tag{B3}$$

Appendix C. Implementation of priority queuing for vulnerable populations

In emergencies, vulnerable populations should be proactively informed and receive special care (Transportation Research Board, 2008). Such populations can be feasibly identified in our problem, since they are concentrated in the waiting zone and can be recognized through targeted surveys (Turner et al., 2010). Our model framework allows for the incorporation of various queuing principles beyond the default first-come-first-served rule. To demonstrate the flexibility of the model, we provide an example that prioritizes vulnerable populations. Specifically, the vulnerable populations we focus on are those at higher risk during overcrowding, including the elderly, children, individuals with disabilities, and pregnant women. Assuming that each passenger group's vulnerability level $v^*(p)$ is assessed by the number of vulnerable populations it contains, which can be defined as a sum of multiple indicators:

$$v^*(p) = a^*(p) + b^*(p) + p^*(p) \tag{C1}$$

where $a^*(p) \in \{0, 1\}$ denotes whether the passenger group has an elder or child (1 if it has, 0 otherwise); $b^*(p) \in \{0, 1\}$ denotes whether the passenger group has a disabled passenger (1 if it has, 0 otherwise); $p^*(p) \in \{0, 1\}$ denotes whether the passenger group has a pregnant passenger (1 if it has, 0 otherwise). The sets of evacuated vulnerable passenger groups for the taxi, bus, and metro modes can be reformulated as follows:

$$P'^t_{\text{taxi}} = \arg \max_{P \subseteq \bar{P}^t_{\text{taxi}}} \left\{ \min_{p \in P} v^*(p) | \sum_{p \in P} q(p) \le a^t_{\text{taxi}} c_{\text{taxi}} \right\}, \quad \forall t \in T, \tag{C2}$$

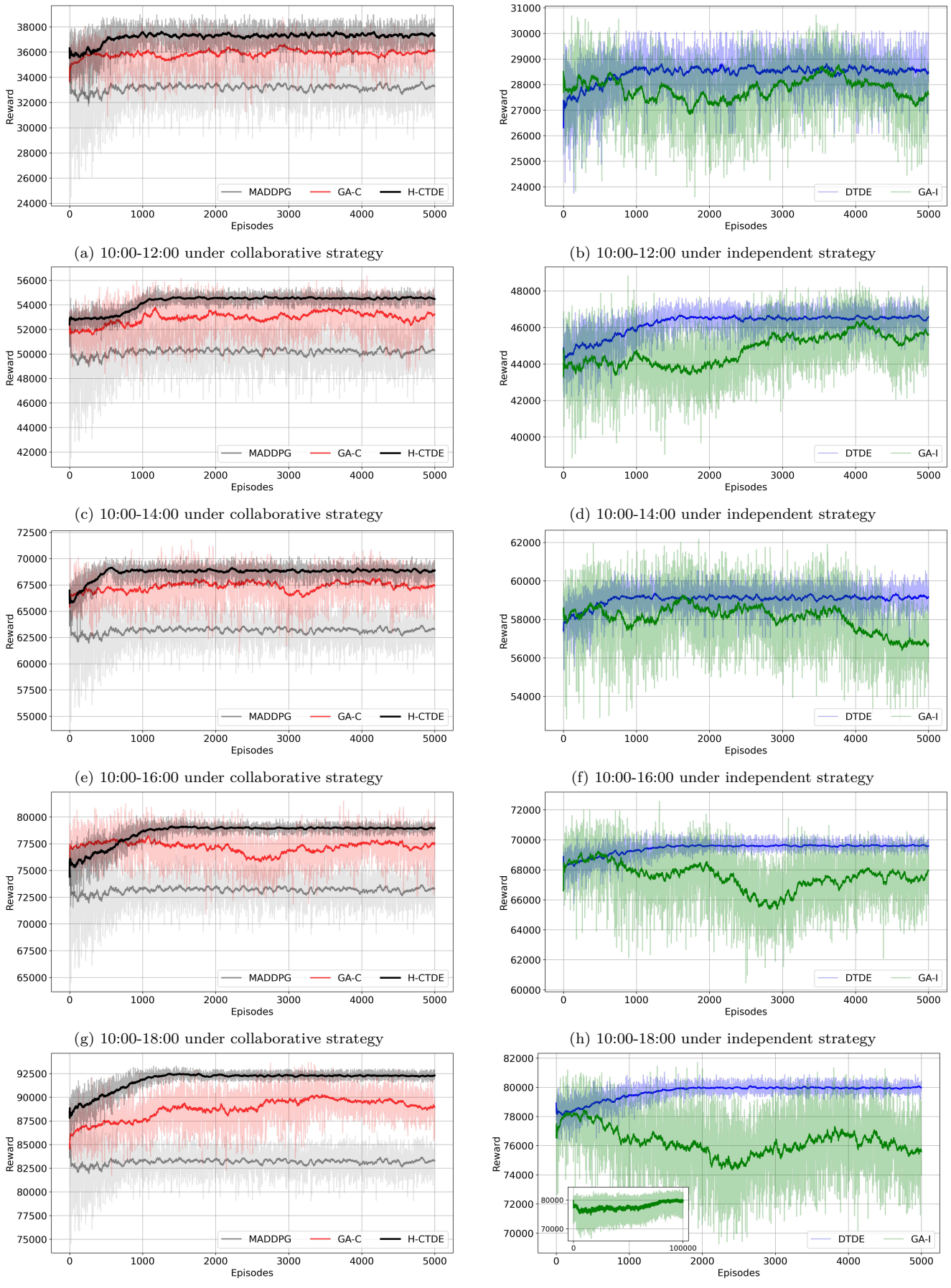
$$P'^t_{\text{bus},r} = \arg \max_{P \subseteq \bar{P}^t_{\text{bus}}} \left\{ \min_{p \in P} v^*(p) | \sum_{\substack{p \in P \\ d(p) \in D^t_r}} q(p) \le c_{\text{bus}} \right\}, \quad \forall t \in T, r \in \gamma^t, \tag{C3}$$

$$P'^t_{\text{metro}} = \arg \max_{P \subseteq \bar{P}^t_{\text{metro}}} \left\{ \min_{p \in P} v^*(p) | \sum_{p \in P} q(p) \le a^t_{\text{metro}} \right\}, \quad \forall t \in T. \tag{C4}$$

Vulnerable populations should be prioritized, with the sets of stranded passenger groups updated by $P^t_m \leftarrow P^t_m - P'^t_m, \forall t \in T, m \in M$. The remaining passengers are then selected based on the first-come-first-served principle as defined by Eqs. (13)-(15).

Appendix D. Sensitivity analysis of agents' computational efficiency

The convergence processes of agents under environments described in Section 7.2.1 are listed in Fig. D1. Note that for the GA in Fig. D1j, which does not converge within 5000 episodes, the population size and the number of generations are increased, resulting in a total of 100000 episodes, to achieve convergence, which is shown in the bottom-left corner.



(a) 10:00-12:00 under collaborative strategy

(b) 10:00-12:00 under independent strategy

(c) 10:00-14:00 under collaborative strategy

(d) 10:00-14:00 under independent strategy

(e) 10:00-16:00 under collaborative strategy

(f) 10:00-16:00 under independent strategy

(g) 10:00-18:00 under collaborative strategy

(h) 10:00-18:00 under independent strategy

(i) 10:00-20:00 under collaborative strategy

(j) 10:00-20:00 under independent strategy

Figure D1: Convergence of the algorithms under different environments

Appendix E. Sensitivity analysis of agents' transferability

The results of the sensitivity analysis, as presented in Section 7.4.3, are shown in Table E1. Several comparable results are presented in Tables E2-E8. Note that within the tables: "S" represents the time span of the new environment, "A" represents the time span of the training environment, "V" represents the variant factor, "After" represents the maximum reward after training, "Non-" represents the median reward achieved by agent without training, "Max" represents the maximum reward achieved by the pre-trained agents, "Med" represents the mediant reward, "Gap" represents the difference between the maximum reward achieved by the pre-trained agents and the maximum reward achieved after training, and "Imp" represents the improvement achieved by the pre-trained agents compared to the median reward of non-trained agents.

Table E1: Results of transferability

Scenario			Independent strategy									Collaborative strategy								
			Benchmarks			Results						Benchmarks			Results					
S	A	V	After-	Non-	Gap	Max	Gap	Imp	Med	Gap	Imp	After-	Non-	Gap	Max	Gap	Imp	Med	Gap	Imp
2h	2h	3	23063	21291	7.68%	22246	3.54%	4.14%	21736	5.75%	1.93%	27222	26161	3.90%	27020	0.74%	3.15%	26658	2.07%	1.83%
2h	2h	6	35939	31126	13.39%	34118	5.07%	8.33%	33015	8.14%	5.26%	46100	41432	10.13%	45501	1.30%	8.83%	45096	2.18%	7.95%
2h	2h	9	34628	31651	8.60%	32824	5.21%	3.39%	31745	8.33%	0.27%	44046	42413	3.71%	43857	0.43%	3.28%	43767	0.63%	3.07%
2h	4h	3	23375	21291	8.92%	22798	2.47%	6.45%	22478	3.84%	5.08%	27334	26161	4.29%	27080	0.93%	3.36%	26700	2.32%	1.97%
2h	4h	6	36428	31126	14.55%	35733	1.91%	12.65%	34992	3.94%	10.61%	46242	41432	10.40%	45514	1.58%	8.83%	45110	2.45%	7.95%
2h	4h	9	35186	31651	10.05%	34513	1.91%	8.13%	33746	4.09%	5.95%	44892	42413	5.52%	44425	1.04%	4.48%	44026	1.93%	3.59%
2h	10h	3	23369	21291	8.89%	22911	1.96%	6.93%	22544	3.53%	5.36%	27860	26161	6.10%	27533	1.17%	4.92%	27080	2.80%	3.30%
2h	10h	6	36434	31126	14.57%	35872	1.54%	13.03%	35042	3.82%	10.75%	46041	41432	10.01%	45484	1.21%	8.80%	45097	2.05%	7.96%
2h	10h	9	35189	31651	10.06%	34775	1.18%	8.88%	33828	3.87%	6.19%	45009	42413	5.77%	44401	1.35%	4.42%	44007	2.23%	3.54%
8h	2h	3	64986	61346	5.60%	62545	3.76%	1.84%	61604	5.20%	0.40%	71737	69690	2.85%	71048	0.96%	1.89%	70211	2.13%	0.73%
8h	2h	6	100720	89009	11.63%	94697	5.98%	5.65%	91671	8.98%	2.64%	128405	117593	8.42%	124246	3.24%	5.18%	122042	4.96%	3.46%
8h	2h	9	100721	89952	10.69%	94615	6.06%	4.63%	91552	9.10%	1.59%	125036	116025	7.21%	123696	1.07%	6.14%	122934	1.68%	5.53%
8h	4h	3	67012	61346	8.45%	64945	3.09%	5.37%	64457	3.81%	4.64%	71729	69690	2.84%	71205	0.73%	2.11%	70247	2.07%	0.78%
8h	4h	6	105352	89009	15.51%	104088	1.20%	14.31%	97377	7.57%	7.94%	128652	117593	8.60%	126026	2.04%	6.55%	122525	4.76%	3.83%
8h	4h	9	104510	89952	13.93%	103484	0.98%	12.95%	97036	7.15%	6.78%	126899	116025	8.57%	124656	1.77%	6.80%	122146	3.75%	4.82%
8h	10h	3	66944	61346	8.36%	65003	2.90%	5.46%	64204	4.09%	4.27%	72194	69690	3.47%	71660	0.74%	2.73%	71281	1.26%	2.20%
8h	10h	6	105295	89009	15.47%	104076	1.16%	14.31%	97491	7.41%	8.05%	130174	117593	9.66%	129319	0.66%	9.01%	128451	1.32%	8.34%
8h	10h	9	104522	89952	13.94%	103446	1.03%	12.91%	97060	7.14%	6.80%	127463	116025	8.97%	126660	0.63%	8.34%	125623	1.44%	7.53%
12h	2h	3	82721	81224	1.81%	80247	2.99%	-1.18%	78966	4.54%	-2.73%	91230	90169	1.16%	90606	0.68%	0.48%	90321	1.00%	0.17%
12h	2h	6	127765	113163	11.43%	121713	4.74%	6.69%	117127	8.33%	3.10%	160296	151355	5.58%	156433	2.41%	3.17%	154360	3.70%	1.87%
12h	2h	9	127380	105284	17.35%	121686	4.47%	12.88%	117180	8.01%	9.34%	156943	149216	4.92%	155871	0.68%	4.24%	155065	1.20%	3.73%
12h	4h	3	84617	81224	4.01%	83530	1.28%	2.73%	83068	1.83%	2.18%	91843	90169	1.82%	91164	0.74%	1.08%	90368	1.61%	0.22%
12h	4h	6	134175	113163	15.66%	132480	1.26%	14.40%	123602	7.88%	7.78%	160763	151355	5.85%	157947	1.75%	4.10%	154831	3.69%	2.16%
12h	4h	9	133303	105284	21.02%	131842	1.10%	19.92%	123970	7.00%	14.02%	158939	149216	6.12%	156649	1.44%	4.68%	154434	2.83%	3.28%
12h	10h	3	85132	81224	4.59%	83479	1.94%	2.65%	83028	2.47%	2.12%	92181	90169	2.18%	91763	0.45%	1.73%	91274	0.98%	1.20%
12h	10h	6	134193	113163	15.67%	132409	1.33%	14.34%	123425	8.02%	7.65%	162343	151355	6.77%	160325	1.24%	5.53%	159062	2.02%	4.75%
12h	10h	9	133266	105284	21.00%	131876	1.04%	19.95%	123798	7.10%	13.89%	160119	149216	6.81%	157986	1.33%	5.48%	157158	1.85%	4.96%
Average			-	-	11.59%	-	2.63%	8.95%	-	5.96%	5.62%	-	-	5.99%	-	1.20%	4.79%	-	2.26%	3.73%

Table E2: Results of transferability
(training time span > new time span)

Scenario			Independent strategy			Collaborative strategy		
S	A	V	After-	Med	Gap	After-	Med	Gap
2h	2h	3	23063	21736	5.75%	27222	26658	2.07%
2h	2h	6	35939	33015	8.14%	46100	45096	2.18%
2h	2h	9	34628	31745	8.33%	44046	43767	0.63%
2h	4h	3	23375	22478	3.84%	27334	26700	2.32%
2h	4h	6	36428	34992	3.94%	46242	45110	2.45%
2h	4h	9	35186	33746	4.09%	44892	44026	1.93%
2h	10h	3	23369	22544	3.53%	27860	27080	2.80%
8h	10h	3	66944	64204	4.09%	72194	71281	1.26%
2h	10h	6	36434	35042	3.82%	46041	45097	2.05%
8h	10h	6	105295	97491	7.41%	130174	128451	1.32%
2h	10h	9	35189	33828	3.87%	45009	44007	2.23%
8h	10h	9	104522	97060	7.14%	127463	125623	1.44%
Average					5.33%			1.89%

Table E3: Results of transferability
(training time span < new time span)

Scenario			Independent strategy			Collaborative strategy		
S	A	V	After-	Med	Gap	After-	Med	Gap
12h	10h	3	85132	83028	2.47%	92181	91274	0.98%
12h	10h	6	134193	123425	8.02%	162343	159062	2.02%
12h	10h	9	133266	123798	7.10%	160119	157158	1.85%
8h	2h	3	64986	61604	5.20%	71737	70211	2.13%
12h	2h	3	82721	78966	4.54%	91230	90321	1.00%
8h	2h	6	100720	91671	8.98%	128405	122042	4.96%
12h	2h	6	127765	117127	8.33%	160296	154360	3.70%
8h	2h	9	100721	91552	9.10%	125036	122934	1.68%
12h	2h	9	127380	117180	8.01%	156943	155065	1.20%
8h	4h	3	67012	64457	3.81%	71729	70247	2.07%
12h	4h	3	84617	83068	1.83%	91843	90368	1.61%
8h	4h	6	105352	97377	7.57%	128652	122525	4.76%
12h	4h	6	134175	123602	7.88%	160763	154831	3.69%
8h	4h	9	104510	97036	7.15%	126899	122146	3.75%
12h	4h	9	133303	123970	7.00%	158939	154434	2.83%
Average					6.47%			2.55%

Table E4: Results of transferability (*4h-agents*)

Scenario		Independent strategy			Collaborative strategy		
S	V	After-	Med	Gap	After-	Med	Gap
2h	3	23375	22478	3.84%	27334	26700	2.32%
8h	3	67012	64457	3.81%	71729	70247	2.07%
12h	3	84617	83068	1.83%	91843	90368	1.61%
2h	6	36428	34992	3.94%	46242	45110	2.45%
8h	6	105352	97377	7.57%	128652	122525	4.76%
12h	6	134175	123602	7.88%	160763	154831	3.69%
2h	9	35186	33746	4.09%	44892	44026	1.93%
8h	9	104510	97036	7.15%	126899	122146	3.75%
12h	9	133303	123970	7.00%	158939	154434	2.83%
Average				5.24%			2.82%

Table E5: Results of transferability (*10h-agents*)

Scenario		Independent strategy			Collaborative strategy		
S	V	After-	Med	Gap	After-	Med	Gap
2h	3	23369	22544	3.53%	27860	27080	2.80%
8h	3	66944	64204	4.09%	72194	71281	1.26%
12h	3	85132	83028	2.47%	92181	91274	0.98%
2h	6	36434	35042	3.82%	46041	45097	2.05%
8h	6	105295	97491	7.41%	130174	128451	1.32%
12h	6	134193	123425	8.02%	162343	159062	2.02%
2h	9	35189	33828	3.87%	45009	44007	2.23%
8h	9	104522	97060	7.14%	127463	125623	1.44%
12h	9	133266	123798	7.10%	160119	157158	1.85%
Average				5.27%			1.77%

Table E6: Results of transferability (Variance factor = 3)

Scenario			Independent strategy						Collaborative strategy					
			Benchmarks			Results			Benchmarks			Results		
S	A	V	After-	Non-	Gap	Med	Gap	Imp	After-	Non-	Gap	Med	Gap	Imp
2h	2h	3	23063	21291	7.68%	21736	5.75%	1.93%	27222	26161	3.90%	26658	2.07%	1.83%
2h	4h	3	23375	21291	8.92%	22478	3.84%	5.08%	27334	26161	4.29%	26700	2.32%	1.97%
2h	10h	3	23369	21291	8.89%	22544	3.53%	5.36%	27860	26161	6.10%	27080	2.80%	3.30%
8h	2h	3	64986	61346	5.60%	61604	5.20%	0.40%	71737	69690	2.85%	70211	2.13%	0.73%
8h	4h	3	67012	61346	8.45%	64457	3.81%	4.64%	71729	69690	2.84%	70247	2.07%	0.78%
8h	10h	3	66944	61346	8.36%	64204	4.09%	4.27%	72194	69690	3.47%	71281	1.26%	2.20%
12h	2h	3	82721	81224	1.81%	78966	4.54%	-2.73%	91230	90169	1.16%	90321	1.00%	0.17%
12h	4h	3	84617	81224	4.01%	83068	1.83%	2.18%	91843	90169	1.82%	90368	1.61%	0.22%
12h	10h	3	85132	81224	4.59%	83028	2.47%	2.12%	92181	90169	2.18%	91274	0.98%	1.20%
Average								2.58%						1.38%

Table E7: Results of transferability (Variance factor = 6)

Scenario			Independent strategy						Collaborative strategy					
			Benchmarks			Results			Benchmarks			Results		
S	A	V	After-	Non-	Gap	Med	Gap	Imp	After-	Non-	Gap	Med	Gap	Imp
2h	2h	6	35939	31126	13.39%	33015	8.14%	5.26%	46100	41432	10.13%	45096	2.18%	7.95%
2h	4h	6	36428	31126	14.55%	34992	3.94%	10.61%	46242	41432	10.40%	45110	2.45%	7.95%
2h	10h	6	36434	31126	14.57%	35042	3.82%	10.75%	46041	41432	10.01%	45097	2.05%	7.96%
8h	2h	6	100720	89009	11.63%	91671	8.98%	2.64%	128405	117593	8.42%	122042	4.96%	3.46%
8h	4h	6	105352	89009	15.51%	97377	7.57%	7.94%	128652	117593	8.60%	122525	4.76%	3.83%
8h	10h	6	105295	89009	15.47%	97491	7.41%	8.05%	130174	117593	9.66%	128451	1.32%	8.34%
12h	2h	6	127765	113163	11.43%	117127	8.33%	3.10%	160296	151355	5.58%	154360	3.70%	1.87%
12h	4h	6	134175	113163	15.66%	123602	7.88%	7.78%	160763	151355	5.85%	154831	3.69%	2.16%
12h	10h	6	134193	113163	15.67%	123425	8.02%	7.65%	162343	151355	6.77%	159062	2.02%	4.75%
Average								7.09%						5.36%

Table E8: Results of transferability (Variance factor = 10)

Scenario			Independent strategy						Collaborative strategy					
			Benchmarks			Results			Benchmarks			Results		
S	A	V	After-	Non-	Gap	Med	Gap	Imp	After-	Non-	Gap	Med	Gap	Imp
2h	2h	9	34628	31651	8.60%	31745	8.33%	0.27%	44046	42413	3.71%	43767	0.63%	3.07%
2h	4h	9	35186	31651	10.05%	33746	4.09%	5.95%	44892	42413	5.52%	44026	1.93%	3.59%
2h	10h	9	35189	31651	10.06%	33828	3.87%	6.19%	45009	42413	5.77%	44007	2.23%	3.54%
8h	2h	9	100721	89952	10.69%	91552	9.10%	1.59%	125036	116025	7.21%	122934	1.68%	5.53%
8h	4h	9	104510	89952	13.93%	97036	7.15%	6.78%	126899	116025	8.57%	122146	3.75%	4.82%
8h	10h	9	104522	89952	13.94%	97060	7.14%	6.80%	127463	116025	8.97%	125623	1.44%	7.53%
12h	2h	9	127380	105284	17.35%	117180	8.01%	9.34%	156943	149216	4.92%	155065	1.20%	3.73%
12h	4h	9	133303	105284	21.02%	123970	7.00%	14.02%	158939	149216	6.12%	154434	2.83%	3.28%
12h	10h	9	133266	105284	21.00%	123798	7.10%	13.89%	160119	149216	6.81%	157158	1.85%	4.96%
Average								7.20%						4.45%

Table E9: Computational time during online application

	Collaborative strategy						Independent strategy					
Scenario	Variance = 3		Variance = 6		Variance = 9		Variance = 3		Variance = 6		Variance = 9	
	Total	AVG	Total	AVG	Total	AVG	Total	AVG	Total	AVG	Total	AVG
2h	9.4s	0.39s	9.9s	0.41s	9.7s	0.40s	7.0s	0.29s	7.5s	0.31s	7.9s	0.33s
8h	39.5s	0.41s	39.8s	0.41s	40.0s	0.42s	27.5s	0.29s	27.6s	0.29s	27.9s	0.29s
12h	57.5s	0.40s	57.6s	0.40s	58.1s	0.40s	40.5s	0.28s	40.9s	0.28s	41.1s	0.29s

Note: Total-total computational time for each episode; AVG-average computational time per time step