Equity-Oriented Inflow Control in Congested Urban Rail Systems

Michael Sun[a], Baichuan Mo[b,*], Ruoyun Ma[c]

[a]*Appleby College, Oakville, Ontario, L6K 3P1, Canada*
[b]*Department of Civil Engineering, Tsinghua University, Beijing, China, 100084*
[c]*Department of Management Science and Engineering, Stanford University, Stanford, CA 94305*

Abstract

Rail transit systems under capacity constraints often exhibit boarding inequities, where downstream passengers experience excessive waiting due to first-come-first-served boarding. This study investigates inflow control policies to improve equity by minimizing the system-wide maximum number of trains a passenger must wait before boarding (left-behind times). We develop an event-based modeling framework that captures the discrete interaction between passenger queues and train capacity, and formulate an equity-oriented control problem over platform-level inflow decisions. To address the resulting non-smooth and non-convex optimization problem, we propose a physical gradient solving algorithm that constructs descent directions directly from passenger flow dynamics. The algorithm iteratively identifies the worst-off passenger and traces the minimum upstream capacity adjustments required to enable earlier boarding, yielding a physically interpretable and computationally efficient solution approach. Theoretical analysis establishes finite termination and local optimality at physical termination under physically feasible perturbations. Numerical results on a synthetic metro network inspired by the Shanghai Metro demonstrate an approximately 57% reduction in maximum left-behind time and clear improvement over heuristic and black-box methods. Additional experiments show that the method remains robust under varying network sizes and demand levels.

Keywords: Urban rail transit; Boarding control; Equity; Event-based modeling; Discrete optimization

1. Introduction

Rail transit congestion remains a challenge in metropolitan regions (Mo et al., 2023), particularly on suburban-to-urban corridors that serve both long-distance commuters and dense inner-city demand. During peak periods, the mismatch between passenger demand and train capacity leads to persistent overcrowding, increased dwell times, and cascading operational delays across the network (Anupriya et al., 2023; Zhang et al., 2023).

On many such lines, trains accumulate passengers at upstream stations earlier along the line, and even under relatively uniform demand, first-come-first-served boarding gives upstream stations a structural priority. This results in systematically lower waiting times upstream and severe capacity imbalances downstream. As a consequence, passengers at upstream stations are often able to board near-empty trains, while those at downstream, high-demand stations may be forced to wait through multiple full trains. This phenomenon has been documented on several major systems, including Shanghai Metro Line 9 and Beijing's Batong Line, where downstream stations experience persistent pass-ups and excessive waiting under peak demand (Shi

*Corresponding author

et al., 2018; Gong et al., 2020). Similar spatial inequities in crowding and waiting have also been observed more broadly in urban transit systems, where uneven exposure to crowding significantly affects passenger experience and system fairness (Lin et al., 2023).

Existing approaches to rail congestion mitigation primarily focus on supply-side operational control strategies, such as service frequency adjustment, short-turning, skip-stop or express operations, and timetable rescheduling under capacity constraints (Shi et al., 2018; Gong et al., 2020; Zhu and Goverde, 2019; Cao et al., 2023; Wang et al., 2024; Li et al., 2025; Salode and Ramamoorthy, 2024). Other studies further explore flexible capacity allocation across stations and integrated optimization of train services and passenger control to better match spatially heterogeneous demand (Shi et al., 2022; Li et al., 2023). While effective in some contexts, these strategies often require substantial operational coordination and do not directly regulate how train capacity is consumed along a line.

In contrast, demand-side interventions such as passenger inflow control and path or route recommendation offer operationally deployable alternatives. Inflow regulation has been shown to play a critical role in preventing system breakdown near capacity and maintaining stable operating regimes (Anupriya et al., 2023). A growing body of work studies coordinated and dynamic passenger flow control across multiple stations, incorporating network-wide interactions and transfer flows (Xu et al., 2024; Yang et al., 2022). However, in practice, inflow control has largely been applied reactively for safety and crowd management purposes, rather than systematically as a tool for improving boarding equity across stations (Jiang et al., 2018; Meng et al., 2022).

In this study, we investigate inflow control policies aimed at improving boarding equity along congested rail corridors. We focus on one specific dimension of equity: limiting the worst boarding delay caused by repeated pass-ups. Equity is therefore formalized as the minimization of the maximum number of trains a passenger must wait for before boarding (i.e., left-behind times), in contrast to most inflow-control studies that focus on safety, crowding reduction, or aggregate demand regulation. This perspective is closely related to transportation equity and transport justice research, which emphasizes that fairness analysis should explicitly identify the distributional outcome of concern and examine how transport benefits and burdens are allocated across users and locations (Golub and Martens, 2014; Martens, 2016; Pereira et al., 2017; Lucas et al., 2016; Bruno et al., 2025). At the same time, this objective should be interpreted as a boarding-equity metric rather than a complete welfare measure. Reducing the worst left-behind experience may require additional restrictions at upstream stations, and thus involves trade-offs with average waiting time, total delay, and the distribution of inconvenience across passenger groups.

To address this problem, we develop an event-based modeling and optimization framework that captures the discrete interaction between passenger queues and train capacity through train arrival and departure events. Based on this representation, we formulate an equity-oriented control problem that determines platform-level inflow control to minimize the worst-case left-behind time.

Due to the discrete and non-convex nature of the problem, conventional gradient-based and black-box methods are ineffective. We therefore propose a *physical gradient solving algorithm*, which constructs descent directions directly from passenger flow dynamics. The algorithm identifies the worst-off passenger and traces the minimum upstream capacity adjustments required to enable earlier boarding, effectively redistributing congestion from downstream to upstream stations in a structured and priority-preserving manner. Theoretical analysis shows that the algorithm terminates in finite steps and achieves local optimality at physical termination under physically feasible perturbations. We evaluate the method on a large-scale synthetic network inspired by selected Shanghai Metro corridors. Starting from severe congestion, the

algorithm reduces the maximum left-behind time by around 57% within a limited number of iterations, significantly outperforming heuristic and black-box optimization approaches. The results remain robust under varying network sizes and demand levels.

The remainder of the paper is organized as follows. Section 2 reviews related literature on rail transit congestion management, passenger flow control, and equity in public transport systems. Section 3 presents the proposed event-based modeling framework, the equity-oriented control formulation, and the physical gradient solving algorithm. Section 4 provides theoretical analysis, including convergence and optimality properties. Section 5 reports results from the synthetic case study and benchmark comparisons. Section 6 concludes the paper.

2. Literature review

2.1. Inflow control

Early work on congestion mitigation in rail transit emphasizes passenger inflow control, i.e., regulating station entry to prevent platform overcrowding and system breakdown. For example, Jiang et al. (2018) develop a reinforcement-learning-based controller that dynamically meters passenger entry to reduce safety risks and stranded passengers. More recent studies extend inflow control to coordinated and network-level settings. Deng et al. (2024) propose a flexible inflow control strategy incorporating heterogeneous passenger routes and adaptive control intervals, while Pan et al. (2022) formulate a corridor-level model predictive control framework that jointly optimizes inflow control and train operations, achieving improvements in waiting time and operational efficiency. Similarly, Yang et al. (2022) and Xu et al. (2024) study multi-station coordinated inflow control by explicitly modeling spatial–temporal interactions and transfer passenger flows across the network.

A related stream of literature focuses on the fundamental role of inflow control in maintaining system stability under near-capacity conditions. Using macroscopic fundamental diagram (MFD) theory, Anupriya et al. (2023) show that regulating passenger inflow is essential to prevent throughput collapse and sustain efficient operating regimes. These studies highlight that inflow control is not only a safety mechanism but also a critical tool for managing congestion dynamics in urban rail systems.

Another important direction is the integration of inflow control with train operations and capacity allocation. Shi et al. (2018) formulate a joint optimization of train timetables and station-level boarding control using integer programming, minimizing total passenger waiting time under capacity constraints. Extensions of this framework include integrated models with short-turning strategies (Xue et al., 2022) and flexible capacity allocation across stations (Shi et al., 2022; Li et al., 2023). While these studies demonstrate the effectiveness of coordinated supply–demand control, they primarily focus on improving system efficiency rather than explicitly addressing service equity.

2.2. Equity-oriented control

A growing body of research explicitly considers equity in congested transit systems, motivated by the systematic disadvantage faced by downstream passengers under first-come-first-served boarding. Gong et al. (2020) propose an integrated optimization framework that combines train scheduling and passenger flow control to reduce service imbalance across stations. Their results demonstrate that coordinated control can partially mitigate spatial inequities, although the analysis is limited to single-line systems.

More broadly, recent studies emphasize fairness as a key performance dimension in public transport systems. For example, Lin et al. (2023) show that crowding exposure is unevenly distributed across passengers

and locations, with important implications for service design and policy evaluation. Similarly, Bruno et al. (2025) review equity considerations in transit systems and highlight the need to balance efficiency with fairness across users. The broader transport justice literature further argues that equity evaluation should be explicit about the ethical basis of the analysis, the population or users being compared, and the service outcome being distributed (Golub and Martens, 2014; Martens, 2016; Pereira et al., 2017; Lucas et al., 2016). In this paper, the relevant users are passengers competing for scarce train capacity at congested platforms, and the distributed service outcome is the number of trains missed before boarding.

Some works incorporate fairness into passenger flow control or operational models. Liang et al. (2023) develop real-time control strategies that balance efficiency and fairness across origin–destination pairs, while Meng et al. (2022) propose a stochastic optimization framework that accounts for demand uncertainty and robustness. In the broader transportation and optimization literature, fairness has also been studied through min–max, max-min, or worst-case formulations, which aim to protect the worst-off user rather than only optimizing aggregate performance (Rawls, 1999; Bertsimas et al., 2011; Ogryczak et al., 2014). This provides a theoretical motivation for using the maximum left-behind time as a targeted boarding-equity indicator in a capacity-constrained rail system.

Despite these advances, existing approaches typically treat equity as a secondary objective or embed it within efficiency-driven formulations. Moreover, most models rely on continuous approximations or aggregate flow representations, which do not fully capture the discrete boarding process and its implications for individual passenger experience.

2.3. Summary

Overall, the literature reveals two key gaps. First, inflow-control models are primarily designed for safety, stability, or efficiency, and rarely consider equity as a primary objective. Second, existing equity-oriented approaches are often supply-side dominated or rely on formulations that are difficult to scale to large, event-driven systems with discrete passenger dynamics. These limitations motivate a demand-side formulation that directly regulates passenger inflow while explicitly targeting worst-case service outcomes. By minimizing the maximum left-behind time across all passengers and leveraging an event-based representation of boarding dynamics, our approach provides a scalable and physically interpretable framework for equity-oriented control in congested rail systems.

3. Methodology

3.1. Event-based urban rail transit system dynamics

We model the urban rail transit system using an event-based simulation framework, in which system states are updated only at discrete events corresponding to train arrivals and departures. This approach captures the asynchronous nature of train movements and passenger interactions while avoiding unnecessary time discretization. Let p index platforms and t index events. A platform is defined as a specific combination of transit line, station, and travel direction, such that all passengers waiting on the same platform are served by the same set of trains. Events are generated from the train timetable and correspond to the arrival or departure of a scheduled train at a given platform. Each event t is therefore uniquely associated with a platform and an event type (arrival or departure). The system state at each event is characterized by train loads and platform passenger queues.

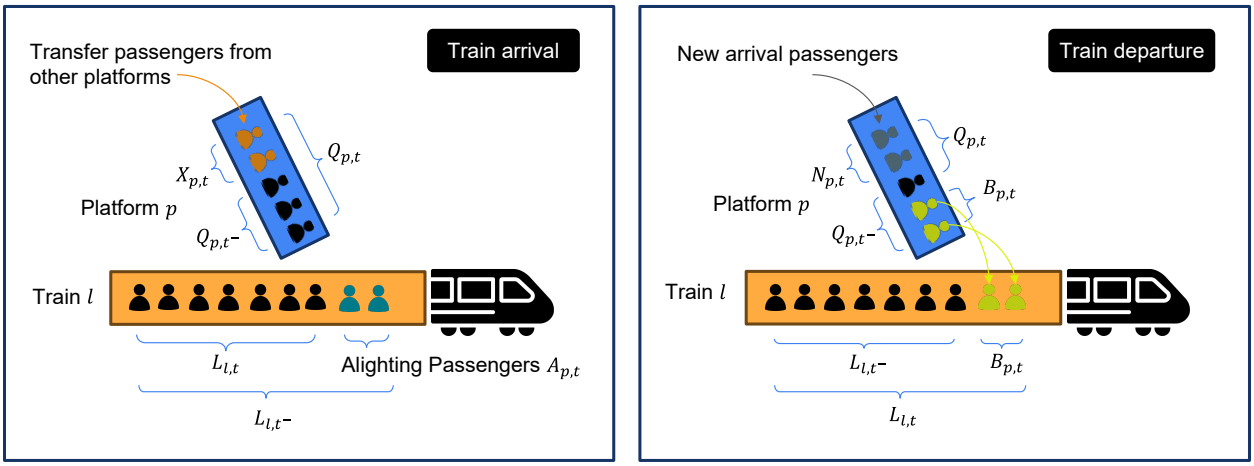


Figure 1: Illustration of the event-based simulation model

Let $Q_{p,t}$ denote the number of passengers waiting on platform p immediately after event t, and let $L_{l,t}$ denote the passenger load of train l immediately after event t. System states are updated differently depending on whether event t corresponds to a train arrival or departure at platform p as shown in Figure 1.

Train arrival events. When a train arrives at platform p, passengers alight from the train according to their destinations. Let $A_{p,t}$ denote the number of alighting passengers at platform p at event t. A subset of these alighting passengers may transfer to other lines or directions and thus enter the platform queue; let $X_{p,t}$ denote the number of such transfer passengers from other platforms to platform p since the previous event of t. The train load and platform queue are updated as

$$L_{l,t} = L_{l,t^-} - A_{p,t}, \quad Q_{p,t} = Q_{p,t^-} + X_{p,t}, \quad \forall l \in \mathcal{L}, p \in \mathcal{P}, t \in \mathcal{T}^{\text{Arr}}, \tag{1}$$

where t^- denotes the system state immediately before the arrival event t. $\mathcal{L}$, $\mathcal{P}$, and $\mathcal{T}^{\text{Arr}}$ are the sets of all trains, platforms, and arrival events, respectively.

Train departure events. When a train departs from platform p, new passengers may arrive at the platform according to their travel demand. Let $N_{p,t}$ denote the number of newly arriving passengers since the previous event at the platform. These passengers are added to the platform queue following their intended routes. We assume that all trains serving platform p travel toward passengers' destinations; therefore, no route-based filtering occurs at the queue formation stage.

Passenger boarding is constrained by both passenger availability and train capacity. Let $a_{p,t} \in [0, 1]$ denote the proportion of waiting passengers at platform p who are allowed to board the departing train at event t, reflecting our inflow control strategies (described later in Section 3.2). Let $R_{p,t}$ denote the remaining available capacity of the train upon departure. The number of boarding passengers is given by

$$B_{p,t} = \min \left\{ Q_{p,t^-} \cdot a_{p,t},\ R_{p,t} \right\}, \quad p \in \mathcal{P}, t \in \mathcal{T}^{\text{Dep}}. \tag{2}$$

where $\mathcal{T}^{\text{Dep}}$ is the set of departure events. The platform queue and train load are updated as

$$Q_{p,t} = Q_{p,t^-} + N_{p,t} - B_{p,t}, \quad L_{l,t} = L_{l,t^-} + B_{p,t}, \quad p \in \mathcal{P}, t \in \mathcal{T}^{\text{Dep}}. \tag{3}$$

This event-based formulation enables a consistent representation of passenger–train interactions, capacity constraints, and transfer dynamics, and serves as the foundation for subsequent control and optimization analyses.

3.2. Equity-oriented control problem description

We consider an inflow control problem formulated on the discrete-event transit system described above. Control decisions are made only at train departure events, as boarding eligibility is determined immediately before a train departs. For each platform p and departure event $t \in \mathcal{T}^{\text{Dep}}$, we introduce a control variable $a_{p,t} \in [0, 1]$, which represents the proportion of passengers waiting on platform p who are allowed to board the departing train at event t. This control captures inflow regulation mechanisms such as platform access control, gate restrictions, or holding policies, and directly affects the number of boarding passengers through the boarding process defined previously.

Let $\mathcal{N}$ denote the set of passengers in the system, and let $\mathcal{B}_n$ denote the set of platforms visited by passenger n during their journey. For each passenger $n \in \mathcal{N}$ and platform $p \in \mathcal{B}_n$, let $W_{n,p}$ denote the left-behind time experienced by passenger n at platform p. For example, if the passenger misses the first arriving train after joining the queue and boards the second one, the left-behind time is 1.

Our objective is to achieve equity-oriented control by minimizing the worst-case left-behind time across all passengers and platforms. Formally, the control problem is given by

$$\min_{\mathbf{a}} \quad \max_{n \in \mathcal{N},\, p \in \mathcal{B}_n} W_{n,p}, \tag{4}$$

where $\mathbf{a} = \{a_{p,t} \in [0, 1] : p \in \mathcal{P},\, t \in \mathcal{T}^{\text{Dep}}\}$ denotes the collection of inflow control decisions over all platforms and departure events.

This min–max formulation explicitly emphasizes fairness by limiting the maximum left-behind time experienced by any individual passenger, rather than optimizing average system performance. Such an equity-oriented objective is particularly relevant under capacity constraints and demand surges, where uncontrolled boarding may result in excessive waiting times for a subset of passengers, especially those at downstream stations. The rationale is similar to worst-case service guarantees and max-min fairness: the objective prevents a small group of passengers from bearing an extreme share of the congestion burden simply because of their position along the line (Rawls, 1999; Bertsimas et al., 2011; Ogryczak et al., 2014).

We note, however, that the objective is intentionally narrow. It measures equity in terms of the maximum number of pass-ups at the passenger-platform level, and does not by itself represent total passenger welfare, full-trip generalized cost, crowding discomfort, transfer reliability, accessibility, or socioeconomic equity. This distinction is important because transportation equity is multidimensional and can be evaluated using different ethical principles and outcome measures (Martens, 2016; Pereira et al., 2017; Lucas et al., 2016). Moreover, improving the worst case generally requires reallocating scarce train capacity from upstream passengers to downstream passengers. The proposed formulation is therefore best understood as a tool for studying the trade-off between worst-case boarding equity and system-level efficiency, rather than as a claim that minimizing the maximum left-behind time is always the socially optimal policy.

3.3. Physical gradient solving algorithm

The equity-oriented inflow control problem is challenging due to the discrete and non-smooth dependence of passenger left-behind times on boarding control variables. To address this, we propose a *physical gradient solving algorithm* that leverages the event-based passenger dynamics to construct a physically interpretable adjustment direction for the worst-case left-behind time. The key idea is to identify a worst-off passenger, determine the additional train capacity required for this passenger to board one train earlier, and then trace this required capacity backward to upstream boarding events.

For a given control vector $\mathbf{a}$, let the event-based simulation return the passenger-level left-behind values $W_{n,p}(\mathbf{a})$, the realized boarding counts $B_{p,t}(\mathbf{a})$, train loads, and ordered platform queues. Define the set of passenger-platform pairs attaining the maximum left-behind time as

$$\mathcal{C}(\mathbf{a}) = \{(n,p) \mid W_{n,p}(\mathbf{a}) = \max_{n' \in \mathcal{N},\, p' \in \mathcal{B}_{n'}} W_{n',p'}(\mathbf{a})\}. \tag{5}$$

Among these candidates, the critical passenger-platform pair is selected by the lexicographic rule

$$(n^*, p^*) = \arg \min_{(n,p) \in \mathcal{C}(\mathbf{a})} \left(l_{n,p},\ -s_{n,p}, -T^{\mathrm{Arr}}_{n,p} \right), \tag{6}$$

where $l_{n,p}$ denotes the index of the train that passenger n eventually boards at platform p (smaller value indicates earlier departure time), $s_{n,p}$ denotes the platform sequence index along the line (smaller values indicate upstream), and $T^{\mathrm{Arr}}_{n,p}$ denotes the arrival time of n at platform p. This rule selects, among the worst-off passengers, the one associated with the earliest boarded train; ties are resolved in favor of downstream platforms and then later arrivals at the same platform. The first two ordering components ensure a consistent target passenger for the subsequent capacity adjustment and prevent cycling behavior in the iterative procedure (see Proposition 1). The final ordering component implies that when (n^*, p^*) is improved, earlier-arriving passengers at the same platform and train are also protected by the first-come-first-served boarding rule.

Let l_{n^*,p^*} denote the train on which passenger n^* eventually boards at platform p^*, and let $l^{-1}_{n^*,p^*}$ be the train immediately preceding l_{n^*,p^*}, on which n^* was left behind. Let t^* be the departure event of train $l^{-1}_{n^*,p^*}$ at platform p^*. A diagram for the physical gradient solving algorithm is shown in Figure 2.

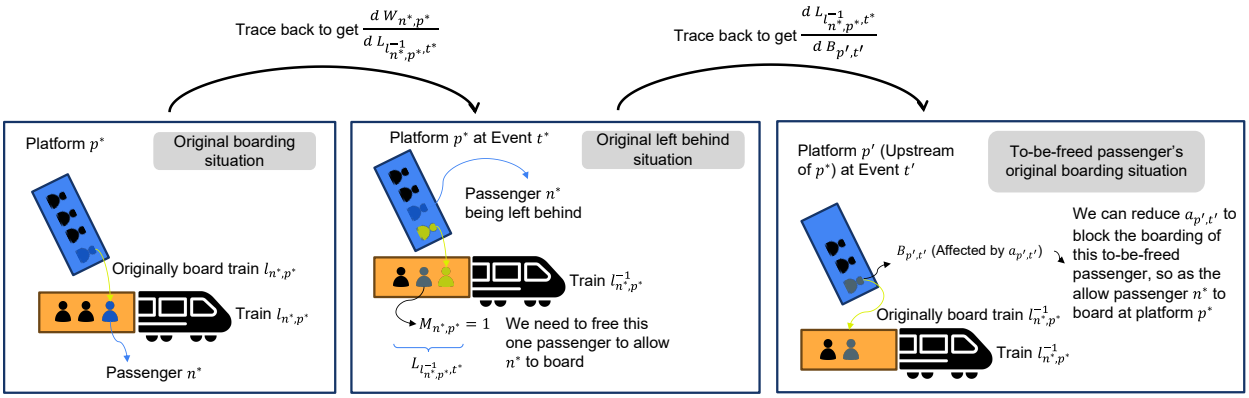


Figure 2: Diagram of the physical gradient solving algorithm

To reduce W_{n^*,p^*} by one unit, the algorithm attempts to let n^* board train $l^{-1}_{n^*,p^*}$ instead of l_{n^*,p^*}. Let $\mathrm{rank}_{p^*,t^*}(n^*)$ be the position of passenger n^* in the first-come-first-served queue at platform p^* immediately before the departure event t^*. The number of additional boarding positions required at (p^*, t^*) is

$$M_{n^*,p^*} = \max\left\{0,\ \mathrm{rank}_{p^*,t^*}(n^*) - B_{p^*,t^*}(\mathbf{a})\right\}. \tag{7}$$

Thus, M_{n^*,p^*} denotes the additional capacity required for passenger n^* to board train $l^{-1}_{n^*,p^*}$ at platform p^* under the current boarding target. If a_{p^*,t^*} already reserves capacity for downstream stations, this definition preserves that reserved capacity and only computes the extra seats needed to advance n^* by one train.

The required capacity is created by reducing upstream boarding on train $l^{-1}_{n^*,p^*}$. Define the set of

removable passengers as

$$\mathcal{U}_{n^*,p^*}(\mathbf{a}) = \{m \in \mathcal{N} : m \text{ is onboard train } l^{-1}_{n^*,p^*} \text{ immediately before boarding at } p^*,\ s_{p_m} < s_{p^*}\}, \quad (8)$$

where p_m is the platform at which passenger m boarded train $l^{-1}_{n^*,p^*}$. Passengers in $\mathcal{U}_{n^*,p^*}(\mathbf{a})$ are ordered by the last-board-first-remove rule: passengers who boarded train $l^{-1}_{n^*,p^*}$ later are removed first. This ordering protects the reserved capacity created for passenger n^*: if an earlier-boarded passenger were removed while later-boarded downstream passengers remained on the train, those downstream passengers could occupy the released capacity before train $l^{-1}_{n^*,p^*}$ reaches p^*. By removing the latest-boarded passengers first, releasing M_{n^*,p^*} passenger spaces guarantees that the required capacity is available at p^* for n^* to board. Let

$$\mathcal{S}_{n^*,p^*}(\mathbf{a}) = \mathrm{First}_{M_{n^*,p^*}}\left(\mathcal{U}_{n^*,p^*}(\mathbf{a}); \text{last-board-first-remove}\right) \quad (9)$$

be the first M_{n^*,p^*} removable passengers under this ordering. If $|\mathcal{U}_{n^*,p^*}(\mathbf{a})| < M_{n^*,p^*}$, then the required capacity cannot be released from upstream boarding events and the current adjustment is infeasible. This means that even if all removable upstream passengers are prevented from boarding train $l^{-1}_{n^*,p^*}$, the available capacity when the train arrives at p^* is still insufficient for passenger n^* to board, for example because the queue ahead of n^* remains too long. In this case, the algorithm stops and returns the current best maximum left-behind value.

For each upstream departure event (p,t) contributing passengers to train $l^{-1}_{n^*,p^*}$, define the event-specific removal count

$$\Delta B_{p,t} = \left|\{m \in \mathcal{S}_{n^*,p^*}(\mathbf{a}) : m \text{ boarded train } l^{-1}_{n^*,p^*} \text{ at } (p,t)\}\right|, \quad \forall (p,t) \in \mathcal{T}^{\mathrm{Up}}_{l^{-1}_{n^*,p^*},p^*}. \quad (10)$$

The target boarding counts after the adjustment are then

$$\widehat{B}_{p,t} = B_{p,t}(\mathbf{a}) - \Delta B_{p,t}, \quad \forall (p,t) \in \mathcal{T}^{\mathrm{Up}}_{l^{-1}_{n^*,p^*},p^*}, \quad (11)$$

$$\widehat{B}_{p^*,t^*} = B_{p^*,t^*}(\mathbf{a}) + M_{n^*,p^*}, \quad (12)$$

where $\mathcal{T}^{\mathrm{Up}}_{l^{-1}_{n^*,p^*},p^*}$ denotes the set of upstream departure events at which passengers board train $l^{-1}_{n^*,p^*}$ before it reaches p^*. The corresponding control update is obtained by converting these target boarding counts back to boarding proportions:

$$a^+_{p,t} = \Pi_{[0,1]}\left(\frac{\widehat{B}_{p,t}}{Q_{p,t^-}}\right), \quad \forall (p,t) \in \mathcal{T}^{\mathrm{Up}}_{l^{-1}_{n^*,p^*},p^*} \cup \{(p^*,t^*)\}, \quad (13)$$

where $\Pi_{[0,1]}(x) = \min\{1, \max\{0,x\}\}$ is the projection onto the feasible interval. All other control variables remain unchanged.

The physical meaning of the update in Eq. 13 can be understood from the sensitivity of the critical passenger's left-behind time to the affected control variables. For an upstream control variable, the sensitivity follows the physical chain

$$\frac{\mathrm{d}W_{n^*,p^*}}{\mathrm{d}a_{p,t}} = \frac{\mathrm{d}W_{n^*,p^*}}{\mathrm{d}L_{l^{-1}_{n^*,p^*},t^*}} \cdot \frac{\mathrm{d}L_{l^{-1}_{n^*,p^*},t^*}}{\mathrm{d}B_{p,t}} \cdot \frac{\mathrm{d}B_{p,t}}{\mathrm{d}a_{p,t}}, \quad \forall (p,t) \in \mathcal{T}^{\mathrm{Up}}_{l^{-1}_{n^*,p^*},p^*}. \quad (14)$$

The three components in this expression correspond to the construction above. The term $\mathrm{d}B_{p,t}/\mathrm{d}a_{p,t}$ links the control variable to the realized boarding count through the boarding rule in Eq. 2. In the finite update, this link is implemented by first reducing the target boarding count from $B_{p,t}(\mathbf{a})$ to $\widehat{B}_{p,t} = B_{p,t}(\mathbf{a}) - \Delta B_{p,t}$

and then converting $\widehat{B}_{p,t}$ back to the updated control $a^+_{p,t}$ in Eq. 13. The term $\mathrm{d}L_{l^{-1}_{n^*,p^*},t^*}/\mathrm{d}B_{p,t}$ captures how reducing boarding at upstream event (p,t) decreases the load of train $l^{-1}_{n^*,p^*}$ when it reaches p^*; this effect is computed by the removable-passenger set $\mathcal{U}_{n^*,p^*}(\mathbf{a})$ and the event-specific removal count $\Delta B_{p,t}$. Finally, $\mathrm{d}W_{n^*,p^*}/\mathrm{d}L_{l^{-1}_{n^*,p^*},t^*}$ represents the effect of the released train capacity on the critical passenger's left-behind time. This effect is quantified by the required additional capacity M_{n^*,p^*} in Eq. 7: once enough upstream capacity is released, passenger n^* can board train $l^{-1}_{n^*,p^*}$ and W_{n^*,p^*} decreases by one. For the critical event (p^*,t^*), the sensitivity is direct:

$$\frac{\mathrm{d}W_{n^*,p^*}}{\mathrm{d}a_{p^*,t^*}} = \frac{\mathrm{d}W_{n^*,p^*}}{\mathrm{d}B_{p^*,t^*}} \frac{\mathrm{d}B_{p^*,t^*}}{\mathrm{d}a_{p^*,t^*}} \leq 0. \tag{15}$$

Increasing a_{p^*,t^*} increases the boarding target $\widehat{B}_{p^*,t^*} = B_{p^*,t^*}(\mathbf{a}) + M_{n^*,p^*}$, which creates the required capacity for n^* to board train $l^{-1}_{n^*,p^*}$.

Therefore, the finite change in each affected control variable is

$$\begin{aligned}\Delta a_{p,t} &= a^+_{p,t} - a_{p,t} \\ &= \Pi_{[0,1]}\left(\frac{\widehat{B}_{p,t}}{Q_{p,t^-}}\right) - a_{p,t} \quad \forall (p,t) \in \mathcal{T}^{\mathrm{Up}}_{l^{-1}_{n^*,p^*},p^*} \cup \{(p^*,t^*)\}.\end{aligned} \tag{16}$$

For upstream events, $\Delta a_{p,t} \leq 0$ because $\widehat{B}_{p,t} = B_{p,t}(\mathbf{a}) - \Delta B_{p,t}$; for the critical event, $\Delta a_{p^*,t^*} \geq 0$ because $\widehat{B}_{p^*,t^*} = B_{p^*,t^*}(\mathbf{a}) + M_{n^*,p^*}$. These two signs are consistent with the corresponding physical sensitivities: increasing upstream boarding tends to increase W_{n^*,p^*}, while increasing the critical-event boarding target tends to decrease W_{n^*,p^*}. Hence, the finite update satisfies the generalized descent relation

$$\Delta a_{p,t} \frac{\mathrm{d}W_{n^*,p^*}}{\mathrm{d}a_{p,t}} \leq 0, \quad \forall (p,t) \in \mathcal{T}^{\mathrm{Up}}_{l^{-1}_{n^*,p^*},p^*} \cup \{(p^*,t^*)\}. \tag{17}$$

This relation connects the derivative expression to the actual computation of $a^+_{p,t}$: the physical gradient identifies which controls should be decreased upstream and which control should be increased at (p^*,t^*), while Eq. 13 gives the finite control values that realize these changes in the discrete event-based system.

Therefore, reducing upstream boarding through $a_{p,t}$ reduces $B_{p,t}$, decreases the load of train $l^{-1}_{n^*,p^*}$ when it reaches p^*, and creates additional capacity for the critical passenger. Unlike an analytical gradient in a smooth optimization problem, this physical gradient is constructed from the discrete boarding order and passenger-flow propagation.

The whole algorithm procedure is summarized in Algorithm 1. Starting from an initial no-control strategy, the algorithm iteratively simulates the system, identifies the worst-off passenger, computes the capacity deficit required to advance that passenger by one train, traces removable passengers upstream, and updates inflow controls along the physical adjustment direction. The procedure stops when the maximum number of iterations is reached or when no feasible one-step improvement can be constructed.

There are two cases in which the current adjustment becomes infeasible. The first is when $W^{(i)}_{n^*,p^*} \leq 1$. In this case, any further left-behind reduction for passenger (n^*,p^*) would require restricting upstream passengers to create additional capacity on train $l^{-1}_{n^*,p^*}$. Such restrictions would increase those passengers' left-behind values by at least one, so the system-wide maximum left-behind value cannot be reduced further by this physical adjustment. The second is when $|\mathcal{U}_{n^*,p^*}(\mathbf{a}^{(i)})| < M_{n^*,p^*}$. This means that the number of removable upstream passengers on train $l^{-1}_{n^*,p^*}$ is insufficient to create the required capacity for passenger n^* to board earlier.

Algorithm 1 Physical gradient solving algorithm for equity-oriented inflow control

Input: Initialize boarding control $\mathbf{a}^{(0)} \leftarrow \mathbf{1}$, iteration number $i \leftarrow 0$, best maximum left-behind time $W^*_{\text{Max}} \leftarrow \infty$, best control strategy $\mathbf{a}^* \leftarrow \mathbf{a}^{(0)}$.
repeat
  Simulate event-based system with current $\mathbf{a}^{(i)}$ and get left-behind times $W^{(i)}_{n,p}$ for all n, p
  Identify critical passenger-platform (n^*, p^*) according to the lexicographic rule
  if $W^{(i)}_{n^*,p^*} \leq W^*_{\text{Max}}$ **then**
    $W^*_{\text{Max}} \leftarrow W^{(i)}_{n^*,p^*}$, $\mathbf{a}^* \leftarrow \mathbf{a}^{(i)}$ ▷ Update best results so far
  if $W^{(i)}_{n^*,p^*} \leq 1$ **then**
    break ▷ Cannot further reduce maximum left-behind time; terminate
  Determine the train l_{n^*,p^*}, the previous train $l^{-1}_{n^*,p^*}$, and the departure event t^* at p^*
  Compute the required additional capacity M_{n^*,p^*} using Eq. 7
  Construct the removable passenger set $\mathcal{U}_{n^*,p^*}(\mathbf{a}^{(i)})$
  if $|\mathcal{U}_{n^*,p^*}(\mathbf{a}^{(i)})| < M_{n^*,p^*}$ **then**
    break ▷ Cannot free enough capacity; terminate
  Select $\mathcal{S}_{n^*,p^*}(\mathbf{a}^{(i)})$ according to the last-board-first-remove rule
  Compute event-specific removal counts $\Delta B_{p,t}$
  Compute target boarding counts $\widehat{B}_{p,t}$ and update controls by Eq. 13
  Assign the new control strategy to $\mathbf{a}^{(i+1)}$
  $i \leftarrow i + 1$
until maximum number of iterations
Output: W^*_{Max}, $\mathbf{a}^*$

In practical implementations, if $\mathcal{C}(\tilde{\mathbf{a}})$ involves passengers left behind on different lines, the physical gradient can be applied in parallel to trains on different lines to accelerate the process, as they do not interfere with each other.

4. Theoretical analysis

In this section we analyze the convergence properties of the proposed physical gradient solving algorithm. We first show that the algorithm does not exhibit cycling behavior due to the specific ordering rule used to select the critical passenger. We then establish finite termination and local optimality at physical termination under the physically feasible perturbations considered by the algorithm.

4.1. Absence of cycling

To demonstrate that the algorithm converges, we must show it does not enter an infinite loop (cycling) during the adjustment process. We define a specific sequence of iterations as a *congestion propagation phase* and analyze the behavior of the algorithm within it.

Definition 1 (Congestion propagation phase)**.** A congestion propagation phase starts when the algorithm selects a critical passenger (n^*, p^*) at iteration i and attempts to reduce their left-behind time $W^{(i)}_{n^*,p^*}$. The phase consists of all subsequent iterations for which the system-wide maximum left-behind time remains greater than or equal to the initial value $W^{(i)}_{n^*,p^*}$. Mathematically, the congestion propagation phase for

(n^*, p^*) is defined as

$$\mathcal{I}_{n^*,p^*} = \{j \in \mathbb{N} : i \le j \le I^{\text{Tem}},\ W^{(k)}_{\text{Max}} \ge W^{(i)}_{n^*,p^*} \text{ for all } k = i, \dots, j\}, \tag{18}$$

where I^{Tem} is the iteration index when the algorithm terminates.

During this phase, the algorithm applies the finite update in Eq. 13: it restricts upstream boarding to free capacity on the previous train and increases the boarding target at the critical event. The upstream restriction can cause "congestion propagation", where upstream passengers experience increased left-behind times and potentially become new critical passengers, keeping $W^{(j)}_{\text{Max}}$ elevated until the congestion is smoothed across upstream platforms.

Proposition 1. *(Monotonicity of lexicographic selection) During a congestion propagation phase initiated by passenger (n^*, p^*), no passenger (n, p) with a lexicographically smaller triple $(l_{n,p}, -s_{n,p}, -T^{Arr}_{n,p})$ than (n^*, p^*) can be selected as the critical passenger.*

Proof. Let (n^*, p^*) be the passenger selected at the start of the phase. The algorithm attempts to reduce W_{n^*,p^*} by restricting boarding volumes at upstream platforms and increasing the boarding target at (p^*, t^*). We categorize all passengers (n, p) with a smaller lexicographic order into two groups to show that their left-behind times $W_{n,p}$ remain unchanged and below the initial W_{n^*,p^*} during the phase.

1. **Case 1: Downstream or later arrival at the same platform.** These are passengers for whom $l_{n,p} = l_{n^*,p^*}$ and either $s_{n,p} > s_{n^*,p^*}$, or $s_{n,p} = s_{n^*,p^*}$ with $T^{\text{Arr}}_{n,p} > T^{\text{Arr}}_{n^*,p^*}$. Upstream inflow control does not restrict capacity for downstream platforms and can only release additional capacity before the train reaches them. Furthermore, under the first-come-first-served boarding rule, passengers who arrive later than n^* at the same platform are behind n^* in the queue, so advancing the boarding of n^* does not reduce their boarding priority. Thus, their left-behind times will not be increased during the congestion propagation phase.
2. **Case 2: Earlier trains.** These are passengers for whom $l_{n,p} < l_{n^*,p^*}$. Since the algorithm first restricts boarding for train $l^{-1}_{n^*,p^*}$, restricting this train may propagate congestion to trains coming later than it. However, the occupancy of earlier trains is unaffected. Consequently, their left-behind times do not increase.

At the start of the phase with iteration index i, for all passenger-platform pairs with smaller lexicographic order, we have $W^{(i)}_{n,p} < W^{(i)}_{n^*,p^*}$ (by the definition of (n^*, p^*) being the worst-off). Throughout the subsequent iterations of the phase $j \in \mathcal{I}_{n^*,p^*}$, $W^{(j)}_{n,p}$ does not increase. Thus, they cannot be selected as critical passengers until the phase ends and $W^{(j)}_{\text{Max}}$ drops below the initial threshold. □

4.2. Finite termination

We now show that the proposed algorithm terminates in a finite number of iterations, even if the maximum allowable number of iterations is unbounded.

Proposition 2 (Finite termination)**.** *The proposed physical gradient solving algorithm terminates after a finite number of iterations.*

Proof. The proof proceeds in two steps.

Step 1: Decomposition into congestion propagation phases and finiteness of phases. By Definition 1, every iteration of the algorithm belongs to a congestion propagation phase initiated by some critical passenger

(n^*, p^*). Moreover, the entire sequence of iterations can be partitioned into such phases, each starting when a new critical passenger is selected after the previous phase ends. From the definition of a congestion propagation phase, a phase ends only when the system-wide maximum left-behind time strictly decreases below the value at the start of that phase. Since left-behind times are integer-valued and nonnegative, the system-wide maximum at the initial iteration $W_{\text{Max}}^{(0)}$ can only decrease a finite number of times. Therefore, the number of congestion propagation phases is finite.

Step 2: Finiteness of iterations within each phase. Consider any congestion propagation phase initiated by (n^*, p^*). By Proposition 1, during this phase the algorithm will never select a passenger with a lexicographically smaller triple $(l_{n,p}, -s_{n,p}, -T_{n,p}^{\text{Arr}})$ than (n^*, p^*). Hence, all critical passengers selected during this phase must come from a subset of passengers whose lexicographic order is no smaller than that of (n^*, p^*). Observe that the set of all passenger-platform pairs is finite. Therefore, the number of distinct lexicographic triples is also finite. If the same critical passenger-platform pair is selected again, the algorithm either constructs a feasible update that reduces its left-behind time by one train, or it declares the current adjustment infeasible and terminates. Thus, an infinite cycle within a single phase cannot occur, and the number of iterations within each congestion propagation phase must be finite.

In summary, since the total number of congestion propagation phases is finite (Step 1), and each phase contains a finite number of iterations (Step 2), the total number of iterations executed by the algorithm is finite. □

Remark 1. Although Proposition 2 guarantees finite termination, the number of iterations required in practice may be very large under severe congestion. This is because each iteration of the physical gradient solving algorithm propagates congestion relief in a highly localized manner—typically from a single platform at a single departure event upstream to one train. When congestion spans multiple stations and persists over consecutive time periods, the worst-case left-behind passengers may be associated with different trains and platforms across iterations. In such cases, capacity adjustments must be propagated sequentially across stations and time slices, leading to slow convergence. This behavior is inherent to the discrete and event-driven nature of the problem: relieving congestion at one location-time pair does not immediately resolve downstream or future congestion. As a result, although the algorithm converges in finite time, the convergence speed may be slow in highly congested networks.

4.3. Local optimality

We now characterize the optimality of a control strategy at which the proposed physical gradient solving algorithm stops because no feasible physical adjustment can be constructed. If the algorithm is stopped only because an external maximum iteration limit is reached, the returned strategy should be interpreted as the best solution found within the computational budget, rather than as an optimality-certified solution. Due to the discrete, event-driven nature of the problem, classical notions of differentiability do not apply. Instead, we define a notion of *local optimality* based on physically feasible perturbations of the control variables.

Definition 2 (Local optimality under physical perturbations)**.** A feasible control strategy $\tilde{\mathbf{a}}$ is said to be *locally optimal* if there exists no admissible perturbation $\Delta\mathbf{a}$, corresponding to a finite sequence of physical updates of the form in Eq. 13, such that

$$\max_{n,p} W_{n,p}(\tilde{\mathbf{a}} + \Delta\mathbf{a}) < \max_{n,p} W_{n,p}(\tilde{\mathbf{a}}). \tag{19}$$

This definition captures the essential physical constraint of the system: any improvement in the worst-case left-behind time must be achieved through feasible capacity reallocation, consisting of upstream boarding restrictions and the corresponding critical-event boarding-target increase, while respecting the priority structure induced by the lexicographic rule.

Proposition 3 (Local optimality at physical termination)**.** *Suppose the algorithm terminates at a control strategy* $\tilde{\mathbf{a}}$ *because either* $W_{n^*,p^*} \leq 1$ *or* $|\mathcal{U}_{n^*,p^*}(\tilde{\mathbf{a}})| < M_{n^*,p^*}$ *holds for the current critical passenger* (n^*, p^*)*. Then* $\tilde{\mathbf{a}}$ *is locally optimal under physical perturbations.*

Proof. Let (n^*, p^*) denote the critical passenger at the stopping iteration under $\tilde{\mathbf{a}}$, with maximum left-behind time $W_{n^*,p^*}(\tilde{\mathbf{a}}) = \max_{n,p} W_{n,p}(\tilde{\mathbf{a}})$. We prove that no admissible physical perturbation can strictly reduce this maximum value.

Step 1: Necessity of upstream restriction. To reduce W_{n^*,p^*}, passenger n^* must be able to board an earlier train $l^{-1}_{n^*,p^*}$. Since train capacity is fixed, this requires freeing space on that train before it reaches p^* and increasing the boarding target at (p^*, t^*) so that the released capacity can be used by n^*. Therefore, any admissible perturbation that reduces W_{n^*,p^*} must follow the same physical structure as Eq. 13: upstream boarding restrictions paired with a critical-event boarding-target increase.

Step 2: Exhaustion of feasible upstream adjustments. At physical termination, one of the following conditions holds:

(i) $W_{n^*,p^*} \leq 1$, in which case any further attempt to reduce W_{n^*,p^*} would require removing at least one upstream passenger, thereby increasing their left-behind time by at least one and preserving the system-wide maximum;

(ii) $|\mathcal{U}_{n^*,p^*}(\tilde{\mathbf{a}})| < M_{n^*,p^*}$, meaning that the number of removable upstream passengers on train $l^{-1}_{n^*,p^*}$ is insufficient to create enough capacity for n^* to board earlier.

In both cases, no feasible physical update of the form in Eq. 13 exists that can reduce W_{n^*,p^*} without preserving or increasing the system-wide maximum.

Step 3: Conclusion. Combining the above arguments, any feasible perturbation that attempts to reduce the current maximum left-behind time must either:

(i) violate physical feasibility (insufficient releasable capacity), or

(ii) worsen the left-behind time of some passengers, thereby maintaining or increasing the system-wide maximum.

Therefore, no admissible perturbation exists that strictly improves the objective. Hence, $\tilde{\mathbf{a}}$ is locally optimal under physical perturbations. □

Remark 2. The above notion of local optimality is inherently tied to the discrete and physical structure of the problem. Unlike classical smooth optimization, where local optimality is defined via infinitesimal perturbations, here it is defined with respect to *finite, physically realizable adjustments* along passenger-flow propagation.

Moreover, the local optimality result aligns with the algorithmic design: each iteration performs an effective feasible adjustment to reduce the worst-case left-behind time, while preserving priority order. As a result, once no such adjustment exists, the terminal control strategy reaches a state where all remaining improvements within this physical perturbation class would require unacceptable trade-offs. If the procedure stops earlier because of a user-specified iteration limit, this local optimality certificate is not claimed.

4.4. Scope of the optimality result

The above analysis should be interpreted as an algorithm-specific optimality result rather than a guarantee over the full non-convex control problem. The local optimality definition is restricted to finite capacity adjustments that follow the physical perturbation structure used by the proposed algorithm. This scope is consistent with the purpose of the method: it constructs interpretable, passenger-flow-based adjustments that improve the worst left-behind outcome when such adjustments are available.

Because the feasible control space is discrete, highly nonlinear, and affected by passenger ordering, other control strategies outside this perturbation class may exist. Establishing optimality over all admissible boarding-control policies is therefore beyond the scope of the present study. Instead, the theoretical results provide finite termination and local optimality at physical termination for the proposed physical adjustment procedure, while the numerical experiments evaluate its empirical performance against benchmark methods.

5. Case study

5.1. Synthetic network and data

The test network is a synthetic metro network inspired by selected corridors of the Shanghai Metro system, rather than a calibrated reproduction of actual operations. This synthetic design is used intentionally. It preserves key structural features of a large metro system, such as radial line geometry, transfer stations, peak-direction demand, and upstream–downstream capacity competition, while allowing the experimental factors to be controlled transparently. In particular, it makes it easier to validate how the proposed algorithm responds to changes in network size and demand level, because the topology, OD demand, event list, and passenger paths can be regenerated consistently under each test scenario. The base topology uses the approximate ordering, transfer structure, and spatial arrangement of Lines 1, 2, 3, and 9 to create a realistic multi-line test bed with upstream–downstream capacity interactions. These corridors are chosen because they provide both radial passenger-flow structure and transfer connectivity, allowing congestion propagation and capacity interactions across lines to be represented in the synthetic setting.

For each line, approximately one-third of the stations are selected. The selected stations are assigned synthetic residential and employment attraction ratings and are spatially distributed along the line to preserve the longitudinal structure of passenger flows. This sampling strategy balances computational tractability with the need to represent realistic congestion patterns and inter-station dependencies. The resulting synthetic network topology and station layout are illustrated in Figure 3.

Origin–destination (OD) passenger demand is synthetically generated using a gravity-based model. The attraction and production strengths of stations are represented by synthetic population and employment ratings assigned to each station. The gravity model captures both spatial interaction intensity and distance decay effects, producing plausible OD flow patterns across the selected network. Temporal demand profiles are imposed to reflect typical morning and evening peak travel behavior in large metropolitan transit systems. The resulting demand exhibits pronounced peak periods with elevated arrival rates, as well as lower off-peak demand.

The train timetable used in the simulation is also synthetic. It is specified using typical metro service patterns, including shorter headways during peak periods and longer headways during off-peak periods, with values inspired by publicly available Shanghai Metro timetable information. The resulting case study should therefore be interpreted as a Shanghai-inspired synthetic benchmark rather than a real-world empirical validation.

Figure 3: Synthetic study network inspired by selected Shanghai Metro corridors

5.2. *Benchmark models*

5.2.1. *Marginal load reduction heuristic*

The marginal load reduction (MLR) heuristic is a rule-based, gradient-like benchmark that mimics the directional adjustment logic of the proposed method but without explicitly tracing discrete passenger boarding order or computing the minimal required capacity shift.

Starting from an initial feasible control vector $\tilde{\mathbf{a}}$, the system is first simulated to obtain all passenger left-behind times $\tilde{W}_{n,p}$. Let $(n^*, p^*) = \arg\max_{n,p} \tilde{W}_{n,p}$ denote the passenger-platform pair experiencing the maximum left-behind time. The key idea of this heuristic is to marginally reduce the boarding proportions of all upstream departure events that contribute passengers to train $l^{-1}_{n^*,p^*}$. Specifically, for every upstream platform $p' \in \mathcal{P}^{\text{Up}}_{p^*}$ and departure event t' such that passengers boarding at (p', t') load onto train $l^{-1}_{n^*,p^*}$, the control variable is updated as

$$a_{p',t'} \leftarrow \max\left\{0,\ a_{p',t'} - \Delta\right\}, \tag{20}$$

where $\Delta = 0.05$ is a fixed marginal reduction step size.

After updating the control vector, the system is re-simulated and the process is repeated until no further reduction in the maximum left-behind time can be achieved or a maximum number of iterations is reached.

5.2.2. *Black-box optimization benchmarks*

To evaluate the performance of the proposed physical gradient-based solution algorithm, we compare it with several benchmark optimization methods that treat the problem as a black-box optimization task (Mo et al., 2021). Specifically, we define a black-box function $\texttt{Simulation}(\mathbf{a}) : \mathbb{R}^{|\mathbf{a}|} \to \mathbb{R}$, which runs the simulation model and returns the maximum left-behind time in the system under control vector $\mathbf{a}$. The

optimal control problem can therefore be formulated as

$$\min_{\mathbf{a}} \quad W_{\text{Max}}(\mathbf{a}), \tag{21}$$

$$\text{s.t.} \quad W_{\text{Max}}(\mathbf{a}) = \texttt{Simulation}(\mathbf{a}), \tag{22}$$

$$0 \le a_{p,t} \le 1, \quad \forall p \in \mathcal{P},\ t \in \mathcal{T}. \tag{23}$$

where $W^{\text{Max}}(\mathbf{a}) = \max_{n,p} W_{n,p}(\mathbf{a})$ is the maximum left-behind time under control strategy $\mathbf{a}$.

To facilitate convergence of black-box optimization methods and reduce the search space dimension, we restrict the control vector $\mathbf{a}$ to platforms that are upstream of those experiencing left-behind passengers (including the affected platforms themselves). Let t^{Min} and t^{Max} denote the time interval during which left-behind events occur in the uncontrolled system. We further restrict the control horizon to the interval $[t^{\text{Min}} - 20 \text{ min}, t^{\text{Max}}]$, since controls outside this range do not affect the maximum left-behind time. These restrictions significantly reduce the dimensionality of the decision space and accelerate the search of the benchmark models.

We consider the following black-box optimization methods: **(1) Bayesian optimization (BYO).** Bayesian optimization is a sample-efficient global optimization framework designed for expensive black-box functions (Mockus, 1994; Shahriari et al., 2015). It constructs a probabilistic surrogate model (commonly a Gaussian process) to approximate the objective function and uses an acquisition function (e.g., expected improvement) to balance exploration and exploitation when selecting new evaluation points. BO is particularly suitable when each function evaluation (here, a simulation run) is computationally expensive. **(2) Differential evolution (DE).** Differential evolution is a population-based evolutionary algorithm for continuous global optimization (Storn and Price, 1997). It iteratively improves a population of candidate solutions through mutation, crossover, and selection operators. DE is derivative-free and robust to nonconvex, nonlinear, and noisy objective functions, making it a strong baseline for simulation-based optimization problems.

All methods are evaluated under the same simulation budget of 100 iterations.

5.3. Results

5.3.1. Model convergence

Figure 4 illustrates the convergence trajectory of the proposed physical gradient solving algorithm. Starting from the uncontrolled system with $W_{\text{Max}}^{(0)} = 7$, the algorithm rapidly reduces the maximum left-behind time to 3 within approximately 37 iterations. After this point, no further reduction in maximum left-behind time is observed.

The early-stage rapid decrease demonstrates the effectiveness of the physically constructed descent direction. By explicitly tracing the worst-off passenger and freeing the minimum required upstream capacity, the algorithm directly targets the structural bottleneck responsible for the maximum delay. The plateau at $W_{\text{Max}} = 3$ is caused by inherent capacity limitations in the network. At this stage, further reduction would require removing more passengers than the available onboard load of the critical train, which is infeasible.

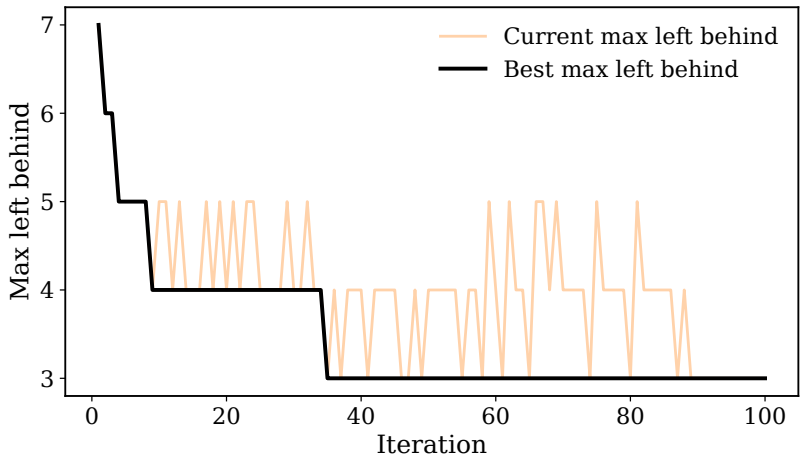


Figure 4: Convergence of the proposed physical gradient algorithm

It is worth noting that the *current* maximum left-behind curve is not monotonically decreasing. In several iterations, the current W_{Max} temporarily increases or remains unchanged. This phenomenon reflects congestion propagation dynamics. To reduce the global maximum, the algorithm may deliberately restrict upstream inflow, which can temporarily worsen waiting conditions at certain platforms. Such controlled redistribution of congestion is necessary to shift capacity toward the most disadvantaged passenger. Only after sufficient upstream adjustments does the global maximum decrease. This non-monotonic behavior highlights the discrete and network-coupled nature of the problem and further demonstrates that simple monotonic descent heuristics are insufficient.

Figure 5 shows the change in number of passengers with left-behind times equal to W_{Max} when W_{Max} decreases from 7 to 3. As the algorithm iterates, the number of passengers experiencing the worst service level (i.e., left-behind times equal to the current W_{Max}) consistently decreases within each phase. This pattern reflects the mechanism of the proposed algorithm: it explicitly targets the most disadvantaged passengers and incrementally reallocates capacity to improve their boarding opportunities.

Importantly, even in the final phase where $W_{\text{Max}} = 3$ remains unchanged for a substantial number of iterations, the algorithm continues to reduce the number of passengers subjected to this highest level of delay. This indicates that, although further reduction in W_{Max} is constrained by system capacity, the algorithm is still actively improving equity by shrinking the group of worst-off passengers. In other words, the system is progressively moving toward a more balanced state, where fewer individuals experience extreme delays, even when the global worst-case metric cannot be further improved.

Overall, the figure demonstrates that the algorithm not only reduces the maximum left-behind time across phases but also continuously enhances equity within each phase by redistributing congestion and alleviating the burden on the most affected passengers.

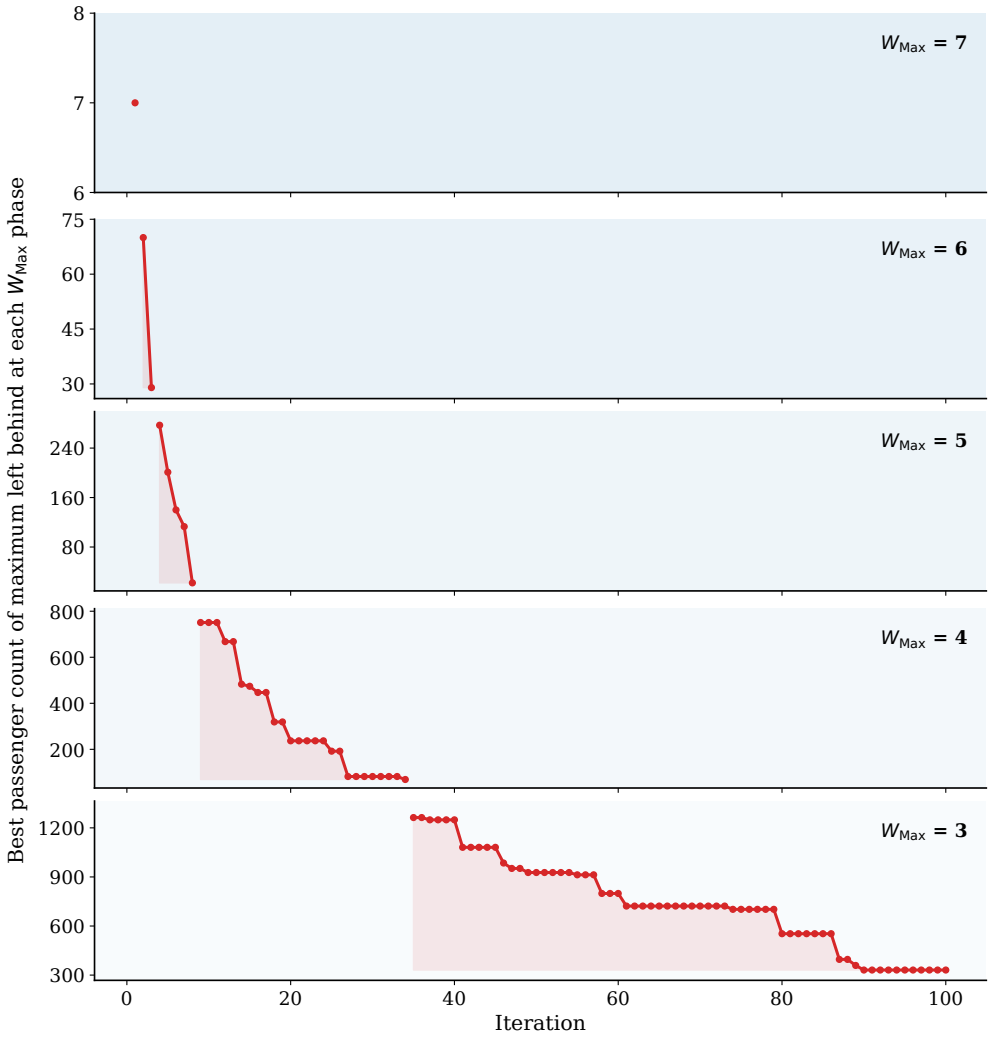


Figure 5: Number of passengers with left-behind times equal to W_{Max}

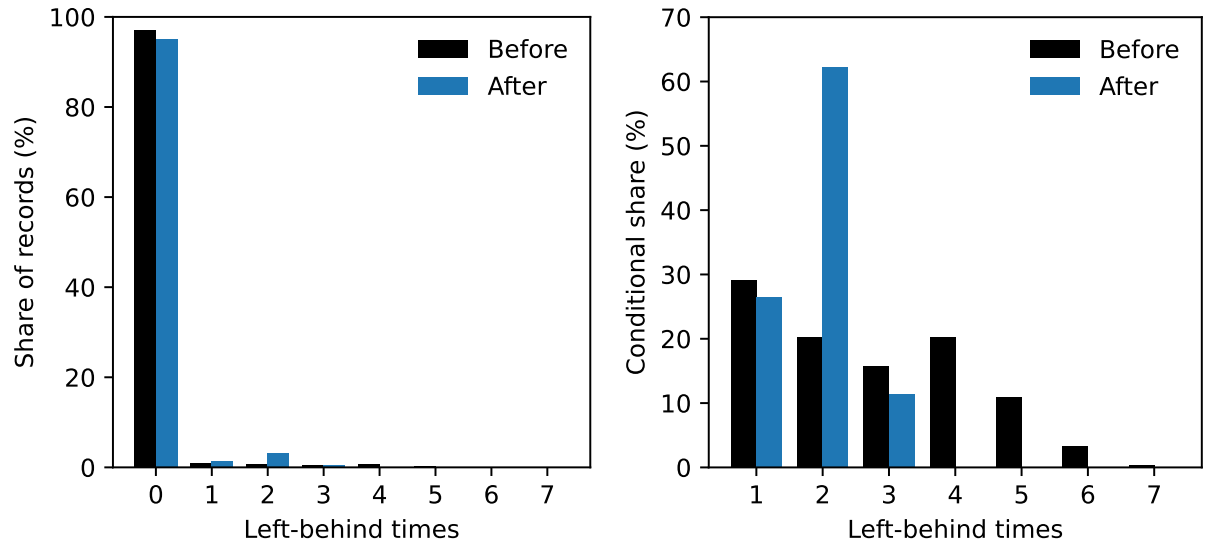
5.3.2. Distribution of left-behind times

To complement the maximum-based convergence analysis, Figure 6 compares the distribution of left-behind times before and after applying the proposed control strategy. Panel (a) reports the distribution over all passenger-platform records, while Panel (b) focuses only on records with positive left-behind times. In both panels, the before and after distributions are shown together, making it possible to compare both the full population-level pattern and the conditional distribution among affected records.

Before control, the distribution has a long right tail: although most records have zero left-behind times, 658 passenger-platform records experience four or more left-behind times, and the maximum reaches 7. After control, this extreme tail disappears entirely, with no record exceeding three left-behind times. This confirms that the algorithm directly reduces the most severe boarding delays rather than only changing the location of the maximum value.

The distribution also reveals the trade-off created by the control policy. The number of records with positive left-behind times increases from 1,886 to 3,074, and the average left-behind value increases slightly from 0.083 to 0.091. This indicates that the algorithm converts a small number of extreme delays into a

broader set of lower-intensity delays. In other words, the improvement is mainly distributional: the policy reduces worst-case inequity by spreading part of the congestion burden upstream, while keeping the resulting left-behind values bounded at a substantially lower level.



(a) All records

(b) Left-behind times ≥ 1

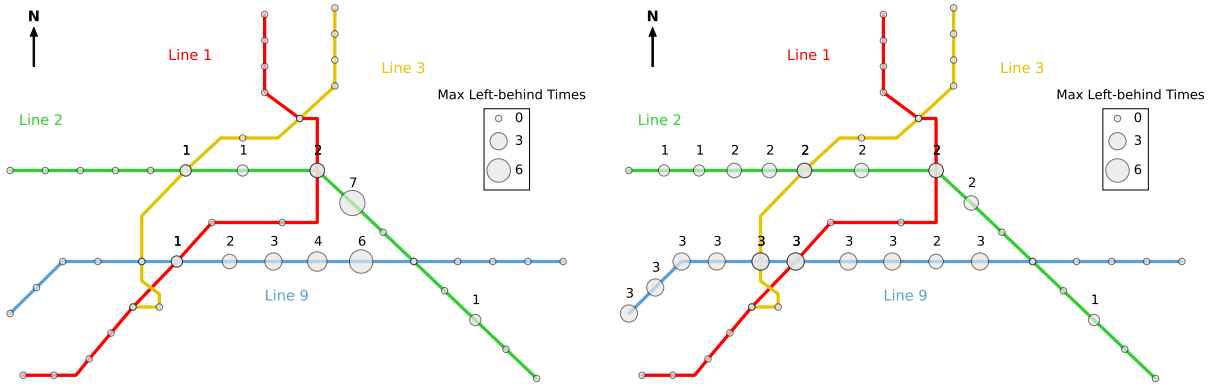
Figure 6: Distribution of left-behind times before and after applying the proposed control strategy

5.3.3. Spatial distribution

Figure 7 illustrates the spatial distribution of left-behind (LB) times across the network before and after applying the proposed control strategy. A clear improvement in equity can be observed from the reduction in both the magnitude and concentration of extreme LB values.

In the initial ("Before") state, severe congestion is concentrated around the central interchange and downstream segments, where several stations exhibit high values (e.g., up to 6 to 7). This pattern is consistent with the morning peak demand direction (west to center), where large passenger inflows accumulate and exceed local boarding capacity. As a result, congestion propagates downstream, creating a cluster of stations with high left-behind times near the bottleneck.

After applying the algorithm ("After"), the left-behind values are significantly reduced and become more spatially balanced. The most notable change is that extreme congestion at the central interchange is alleviated, and high values are no longer concentrated at a single location. Instead, a moderate level of congestion is redistributed to upstream stations. This reflects the intended mechanism of the algorithm: by restricting inflow at upstream platforms, capacity is reserved for downstream passengers who would otherwise experience excessive delays. Such redistribution should not be interpreted as cost-free. It improves the worst-case boarding-equity metric, but it may impose additional waiting on some upstream passengers. The acceptability of this trade-off depends on the operator's policy objective and on complementary performance measures such as average waiting time, total passenger delay, station crowding, and the number of passengers made worse off.



(a) Before (b) After

Figure 7: Spatial distribution of left-behind times before and after the algorithm

5.3.4. Model comparison

Figure 8 compares the best-so-far convergence curves of all benchmark models under the same simulation budget. Overall, all benchmark methods perform substantially worse than the proposed algorithm. The black-box optimization approaches reduce the maximum left-behind time only marginally, reaching $W_{\text{Max}} = 5$ after 100 iterations. This limited improvement is mainly due to the structural properties of the objective function. The maximum left-behind time is piecewise-constant and non-smooth with respect to the control variables $\mathbf{a}$. Small perturbations of $\mathbf{a}$ often do not change the identity of the worst-off passenger, resulting in flat objective regions. Consequently, numerical gradient estimation becomes unreliable and surrogate modeling is ineffective. Without explicit knowledge of the discrete boarding order and capacity propagation mechanism, black-box optimizers struggle to identify meaningful descent directions.

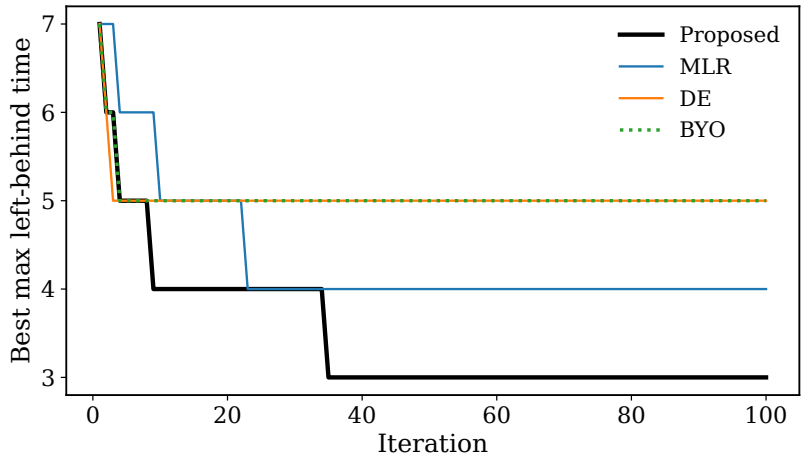


Figure 8: Comparison of different methods

The MLR heuristic performs better than purely black-box methods and converges to $W^{\text{Max}} = 4$. This improvement is expected, as MLR incorporates partial physical insight by reducing upstream boarding proportions of the critical train. However, unlike the proposed method, MLR does not compute the exact

number of passengers that must be removed to benefit the worst-off passenger. Instead, it applies a uniform marginal reduction with fixed step size. As a result, MLR fails to correctly execute congestion propagation. It may either under-adjust, leaving insufficient capacity to shift the boarding event of the critical passenger, or over-adjust, wasting capacity without improving the global maximum.

5.4. Impact of network size

To examine whether the proposed algorithm remains effective as the synthetic network expands, we construct three Shanghai-inspired subnetworks with different sizes. The small case includes Lines 1 and 2, the medium case adds Line 3, and the full case further adds the Line 9 corridor used in the reference experiment. For each network, the station set, OD demand, event list, and passenger paths are regenerated consistently within the selected subnetwork, so that no demand is assigned to stations outside the tested topology. In Tables 1 and 2, "Iter." denotes the first iteration at which the proposed method reaches its reported best maximum left-behind time.

Table 1: Performance under different synthetic network sizes

Network	**Lines**	**Stations**	**Platforms**	**Initial**	**Max left-behind time**				**Iter.**
					Proposed	MLR	DE	BYO	
Small	1+2	28	58	7	2	2	4	4	51
Medium	1+2+3	35	78	7	2	3	3	4	39
Full	1+2+3+9	48	110	7	3	4	5	5	34

Table 1 shows that the proposed method consistently reduces the worst left-behind time across all tested network sizes and performs favorably against the benchmark models. In the small two-line subnetwork, the proposed method and MLR both reduce the maximum left-behind time from 7 to 2, while DE and BYO remain at 4. In the medium and full networks, the proposed method obtains the lowest maximum left-behind time. The full network remains more difficult because it contains more transfer interactions and a larger set of platform-level control decisions. Nevertheless, the proposed algorithm still achieves a substantial reduction in the worst-case boarding delay, indicating that the physical-gradient mechanism is not limited to the reference topology.

5.5. Impact of demand

To further examine the robustness of the proposed method under different congestion levels, we evaluate three demand scenarios: a reduced-demand case (90% of the reference demand), the reference scenario (100%), and an increased-demand case (110%). Table 2 summarizes the maximum left-behind time obtained by different methods.

Table 2: Performance under different demand levels

Scenarios	**Initial**	**Max left-behind time**				**Iter.**
		Proposed	MLR	DE	BYO	
90% demand	4	2	3	4	4	15
100% demand	7	3	4	5	5	34
110% demand	11	4	4	6	6	71

The initial maximum left-behind time increases rapidly with demand. When demand increases to 110%, the uncontrolled maximum left-behind time rises sharply to 11, compared to 7 under the reference

scenario and 4 under reduced demand. This reflects the highly nonlinear amplification of congestion in capacity-constrained rail systems.

The proposed method consistently achieves the best performance across all demand levels. In the low-demand scenario, the algorithm reduces the maximum left-behind time from 4 to 2. Under the reference demand, it achieves a reduction from 7 to 3, and in the high-demand case, it further reduces the maximum left-behind time from 11 to 4. These results demonstrate the robustness of the proposed approach across varying congestion intensities.

Compared with benchmark methods, the proposed method maintains a clear advantage in the 90% and 100% demand scenarios. The MLR heuristic achieves moderate improvements, reducing the maximum left-behind time to 3, 4, and 4 under the three scenarios, respectively. Under 110% demand, MLR reaches the same maximum left-behind value as the proposed method, but it does so with a broader congestion burden: at the first iteration where the best maximum value is achieved, the proposed method has a lower mean left-behind time (0.162 versus 0.181) and fewer records with positive left-behind time (3,985 versus 5,677). In contrast, the black-box optimizers (DE and BYO) show more limited effectiveness, particularly under medium and high demand, where they converge to higher maximum left-behind times (5 and 6 in the 100% and 110% scenarios).

6. Conclusion and discussion

This study investigates inflow control policies for improving boarding equity in congested urban rail systems. We formulate boarding equity as the minimization of the maximum left-behind time, thereby shifting the focus from average system efficiency to worst-case service guarantees. This objective is intended to capture one important aspect of passenger experience–repeated inability to board due to insufficient residual capacity–rather than a complete measure of social welfare. To support this objective, we develop an event-based modeling framework that captures the discrete interaction between passenger queues and train capacity, and propose a physical gradient solving algorithm that constructs descent directions directly from passenger flow dynamics.

The proposed method offers several key contributions. Methodologically, it introduces a physically interpretable optimization approach that avoids the limitations of conventional gradient-based and black-box methods in discrete, non-smooth systems. By explicitly tracing the worst-off passenger and reallocating upstream capacity in a minimal and structured manner, the algorithm provides a novel way to solve equity-oriented control problems in transit systems. Theoretical analysis establishes finite termination and local optimality at physical termination under the physically feasible perturbations considered by the algorithm.

The numerical results demonstrate both the effectiveness and the structural advantages of the proposed approach. In the reference scenario, the algorithm reduces the maximum left-behind time from 7 to 3 (over 55%) within a limited number of iterations, and consistently outperforms all benchmark methods. The network-size experiment indicates that the same mechanism remains effective across two-line, three-line, and four-line synthetic networks. The demand sensitivity analysis further shows that, as demand increases, the uncontrolled system exhibits highly nonlinear growth in left-behind times.

From a practical perspective, the results suggest that inflow control can be used not only for safety and crowd management, but also as a proactive tool for improving service equity. The proposed framework provides actionable guidance on where and when to apply upstream restrictions to balance service across stations. In particular, it highlights the importance of coordinated, system-wide control rather than isolated station-level interventions.

Several limitations and future research directions remain. First, the current model assumes deterministic demand and fixed passenger routes; extending the framework to incorporate stochastic demand, adaptive passenger behavior, and real-time information feedback would improve realism. Second, while the proposed algorithm is computationally efficient relative to black-box methods, scalability to very large networks with high demand or real-time deployment needs further investigation. Third, the present objective focuses on a worst-case boarding-equity metric and does not explicitly optimize average waiting time, total passenger delay, crowding, station holding capacity, or passenger welfare. Integrating equity objectives with broader system goals–such as total travel time, waiting time, energy consumption, or operational cost–would enable a more comprehensive multi-objective framework and would help quantify the trade-offs created by upstream inflow restrictions. Finally, empirical validation using real-world smart card and train movement data would provide further insights into behavioral responses and implementation feasibility.

References

Anupriya, Graham, D.J., Bansal, P., Hörcher, D., Anderson, R.J., 2023. Optimal congestion control strategies for near-capacity urban metros: Informing intervention via fundamental diagrams. Physica A: Statistical Mechanics and its Applications 609, 128390. doi:10.1016/j.physa.2022.128390.

Bertsimas, D., Farias, V.F., Trichakis, N., 2011. The price of fairness. Operations Research 59, 17–31. doi:10.1287/opre.1100.0865.

Bruno, M., Kouwenberg, M., van Oort, N., 2025. Equity in transit fare policy: a literature review. Transport Reviews , 1–22doi:10.1080/01441647.2025.2601689.

Cao, Z., Wang, Y., Yang, Z., Chen, C., Zhang, S., 2023. Timetable rescheduling using skip-stop strategy for sustainable urban rail transit. Sustainability 15, 14511. doi:10.3390/su151914511.

Deng, Y., Liu, W., Jiang, Z., Xu, R., Liu, J., 2024. Method for coordinated passenger inflow control strategy of metro with flexible control time step considering long–short route pattern. Journal of Transportation Engineering, Part A: Systems 150. doi:10.1061/JTEPBS.TEENG-8535.

Golub, A., Martens, K., 2014. Using principles of justice to assess the modal equity of regional transportation plans. Journal of Transport Geography 41, 10–20. doi:10.1016/j.jtrangeo.2014.07.014.

Gong, C., Mao, B., Wang, M., Zhang, T., 2020. Equity-oriented train timetabling with collaborative passenger flow control: A spatial rebalance of service on an oversaturated urban rail transit line. Journal of Advanced Transportation 2020, 8867404. doi:10.1155/2020/8867404.

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. doi:10.1016/j.trc.2018.01.008.

Li, F., Zhang, Y., Guo, X., Chen, T., 2025. Timetable optimization for urban rail transit via flexible train formation and skip-stop strategies. Transportation Planning and Technology , 1–25doi:10.1080/03081060.2025.2559234.

Li, J.j., Bai, Y., Ho, T.k., Jia, W.z., Li, T., 2023. Integrated optimization of train services plan and passenger flow control on an oversaturated suburban metro line. Journal of Central South University 30, 625–641. doi:10.1007/s11771-023-5247-2.

Liang, J., Lyu, G., Teo, C.P., Gao, Z., 2023. Online passenger flow control in metro lines. Operations Research 71, 768–775. doi:10.1287/opre.2022.2417.

Lin, J.Y.J., Jenelius, E., Cebecauer, M., Rubensson, I., Chen, C., 2023. The equity of public transport

crowding exposure. Journal of Transport Geography 110, 103631. doi:10.1016/j.jtrangeo.2023.103631.

Lucas, K., van Wee, B., Maat, K., 2016. A method to evaluate equitable accessibility: combining ethical theories and accessibility-based approaches. Transportation 43, 473–490. doi:10.1007/s11116-015-9585-2.

Martens, K., 2016. Transport Justice: Designing Fair Transportation Systems. Routledge, New York. doi:10.4324/9781315746852.

Meng, F., Yang, L., Shi, J., Jiang, Z.Z., Gao, Z., 2022. Collaborative passenger flow control for oversaturated metro lines: a stochastic optimization method. Transportmetrica A: Transport Science 18, 619–658. doi:10.1080/23249935.2021.1886195.

Mo, B., Ma, Z., Koutsopoulos, H.N., Zhao, J., 2021. Calibrating path choices and train capacities for urban rail transit simulation models using smart card and train movement data. Journal of Advanced Transportation 2021, 5597130. doi:10.1155/2021/5597130.

Mo, B., Ma, Z., Koutsopoulos, H.N., Zhao, J., 2023. Ex post path choice estimation for urban rail systems using smart card data: An aggregated time-space hypernetwork approach. Transportation Science 57, 313–335. doi:10.1287/trsc.2022.1177.

Mockus, J., 1994. Application of bayesian approach to numerical methods of global and stochastic optimization. Journal of Global Optimization 4, 347–365. doi:10.1007/BF01099263.

Ogryczak, W., Luss, H., Pióro, M., Nace, D., Tomaszewski, A., 2014. Fair optimization and networks: A survey. Journal of Applied Mathematics 2014, 612018. doi:10.1155/2014/612018.

Pan, H., Liu, Y., Liu, Z., Hu, H., 2022. Passenger inflow control along rail corridor for recurrent heavy passenger flow mitigation. Transportation Research Record: Journal of the Transportation Research Board 2676, 59–75. doi:10.1177/03611981221085527.

Pereira, R.H.M., Schwanen, T., Banister, D., 2017. Distributive justice and equity in transportation. Transport Reviews 37, 170–191. doi:10.1080/01441647.2016.1257660.

Rawls, J., 1999. A Theory of Justice. Revised edition ed., Belknap Press of Harvard University Press, Cambridge, MA.

Salode, C.K., Ramamoorthy, P., 2024. Train timetabling with rolling stock assignment, short-turning and skip-stop strategy for a bidirectional metro line. arXiv preprint arXiv:2404.04233 doi:10.48550/arXiv.2404.04233.

Shahriari, B., Swersky, K., Wang, Z., Adams, R.P., De Freitas, N., 2015. Taking the human out of the loop: A review of bayesian optimization. Proceedings of the IEEE 104, 148–175. doi:10.1109/JPROC.2015.2494218.

Shi, J., Qin, T., Yang, L., Xiao, X., Guo, J., Shen, Y., Zhou, H., 2022. Flexible train capacity allocation for an overcrowded metro line: A new passenger flow control approach. Transportation Research Part C: Emerging Technologies 140, 103676. doi:10.1016/j.trc.2022.103676.

Shi, J., Yang, L., Yang, J., Gao, Z., 2018. Service-oriented train timetabling with collaborative passenger flow control on an oversaturated metro line: An integer linear optimization approach. Transportation Research Part B: Methodological 110, 26–59. doi:10.1016/j.trb.2018.02.003.

Storn, R., Price, K., 1997. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization 11, 341–359. doi:10.1023/A:1008202821328.

Wang, Y., Zheng, C., Shen, J., Chen, Z., Ma, J., Shen, Y., 2024. Metro stop-skipping rescheduling optimization considering passenger waiting time fairness. Transportation Letters 16, 1115–1125. doi:10.

1080/19427867.2023.2270243.
Xu, L., Lu, J., Zhang, S., Ren, G., He, K., 2024. Subway multi-station coordinated dynamic control method considering transfer inbound passenger flow. Sustainability 16, 11292. doi:10.3390/su162411292.
Xue, H., Jia, L., Li, J., Guo, J., 2022. Jointly optimized demand-oriented train timetable and passenger flow control strategy for a congested subway line under a short-turning operation pattern. Physica A: Statistical Mechanics and its Applications 593, 126957. doi:10.1016/j.physa.2022.126957.
Yang, R., Zhou, W., Han, B., Li, D., Zheng, B., Wang, F., 2022. Research on coordinated passenger inflow control for the urban rail transit network based on the station-to-line spatial-temporal relationship. Journal of Advanced Transportation 2022, 8895935. doi:10.1155/2022/8895935.
Zhang, J., Wada, K., Oguchi, T., 2023. Morning commute in congested urban rail transit system: a macroscopic model for equilibrium distribution of passenger arrivals. Transportmetrica B: Transport Dynamics 11, 2195582. doi:10.1080/21680566.2023.2195582.
Zhu, Y., Goverde, R.M., 2019. Railway timetable rescheduling with flexible stopping and flexible short-turning during disruptions. Transportation Research Part B: Methodological 123, 149–181. doi:10.1016/j.trb.2019.02.015.