Real-time coordinated estimation of passenger demand of urban rail transit and network-wide passenger flow dynamics in a computational-graph framework

Siyu Zhuo*[1], Pan Shang[2], Zhengke Liu[3], and Feixiong Liao[4]

[1]Ph.D Candidate, School of Traffic and Transportation, Beijing Jiaotong University, Beijing, 100044, China
[2]Associate Professor, School of Traffic and Transportation, Beijing Jiaotong University, Beijing, 100044, China
[3]Ph.D Candidate, School of Transportation Science and Engineering, Beihang University, Beijing, 100191, China
[4]Assistant Professor, Urban Planning and Transportation Group, Eindhoven University of Technology, POB 513, NL-5600 MB, Eindhoven, Netherlands

Short summary

The real-time estimation of origin-destination (OD) matrices and passenger flow dynamics in urban rail transit is essential for efficient operations but is hindered by challenges such as data scarcity, high dimensionality, and spatiotemporal dynamics. This study proposes a feedback-enhanced iterative estimation framework that leverages multi-source data to achieve coordinated estimation of OD demand and link-level passenger flows. The framework incorporates a path-to-link temporal correlation matrix to model spatial-temporal dependencies and a time-extended, four-layer computational graph to address high-dimensional complexity. A temporal feedback mechanism embedded in the backward propagation process iteratively refines earlier estimates by calibrating them against errors observed in later intervals, enabling the framework to adapt to dynamic transit environments. Validation on the Sioux-Falls and Beijing metro networks demonstrates the framework's scalability, robustness, and high accuracy, achieving an average estimation accuracy of 98.86%. The experimental results show its potential to enhance operational decision-making in urban rail transit.
Keywords: Real-time estimation; OD matrix; Network-wide passenger flow dynamics; Feedback-enhanced calibration; Time-extended computational graph

1 Introduction

Urban rail transit (URT) systems are vital for urban mobility, providing high-capacity, efficient, and sustain-able solutions. As passenger volumes grow and networks expand, real-time monitoring of passenger flow dynamics becomes crucial for effective management during disruptions or peak periods, enabling timely interventions like train rescheduling and entry restrictions, which outperform static plans in handling fluctuating demand.

Despite advancements in OD estimation and passenger flow modeling, significant challenges remain. Traditional approaches, such as gravity models and survey-based methods, rely on static datasets and fail to ac-count for temporal variations and dynamic passenger behaviors (Munizaga et al., 2014). Data-driven techniques leveraging multi-source data, including Automatic Fare Collection (AFC) records, mobile signaling, and sensor data, have improved modeling accuracy (Pelletier et al., 2011). Bayesian methods (Castillo et al., 2013), Kalman filters (Xiong et al., 2020), and machine learning models (Xu et al., 2024) dynamically integrate real-time data, capturing spatiotemporal patterns and nonlinearities. However, few studies address OD estimation and passenger flow dynamics simultaneously, meaning that their interdependences are often neglected.

Passenger flow dynamics estimation has similarly evolved. Assignment-based models, such as equilibrium assignment and Markov chain methods, allocate passengers based on predefined criteria but lack the flexibility to adapt to real-time demand fluctuations (Szeto and Jiang, 2014). Recent advancements in machine learning, such as graph neural networks and spatiotemporal convolutional

networks, excel at modeling complex spatial and temporal dependencies (Huang et al., 2024). However, the scalability and real-time adaptability of these methods remain limited, particularly for large-scale networks with high-dimensional data. In addition, three critical challenges persist in real-time estimation: (1) spatiotemporal heterogeneities in passenger demand, (2) data sparsity in multi-source datasets, and (3) the computational complexity of high-dimensional coordinated estimation across OD pairs and paths.

To address these challenges, this study aims to develop a robust framework that utilizes sparse data to simultaneously estimate OD matrices and link-level passenger flow dynamics in real time. The key contributions are as follows:

(1) A novel real-time iterative framework integrates a computational graph and feedback mechanism, enabling dynamic calibration of OD distributions and passenger flow states using limited multi-source data to ensure accurate and timely decision-making.
(2) A path-to-link temporal correlation matrix models temporal interdependencies to minimize discrepancies between observed and estimated data for precise flow dynamics representation.
(3) A time-extended four-layer computational graph integrates hierarchical relationships (Origin $\rightarrow$ OD $\rightarrow$ Path $\rightarrow$ Link) to address high-dimensional challenges and maximize multi-source data utilization for enhanced accuracy.
(4) A feedback-based iterative algorithm ensures convergence and superior accuracy, validated through experiments on varying scales, including the large-scale Beijing URT network.

The remainder of this paper is organized as follows. Section 2 discusses the formulations of the real-time coordinated estimation model and the design of the feedback-enhanced iterative algorithm. Section 3 dis-cusses the experimental results. Section 4 concludes with key findings and future research directions.

2 Methodology

2.1 Real-time Coordinated Estimation Model

2.1.1 Construction of space-time network and path-to-link temporal correlation matrix

To illustrate the URT network structure and its temporal characteristics, a simplified two-OD, four-station network (Stations A, B, C, and D) is illustrated as Fig. 1. Station A serves as the origin, with Stations C and D as destinations. The network is structured with nodes and links, where OD pairs $(1 \rightarrow 8)$ and $(1 \rightarrow 9)$ represent trips between the origin and destinations. Physical paths are defined by link sequences, such as: $(1,2) \rightarrow (2,3) \rightarrow (3,4) \rightarrow (4,7) \rightarrow (7,8)$.

To capture time-related dynamics (departure, transfer, and travel times), a space-time network $G' = (V, E)$ is constructed. Each node i is expanded into space-time vertices (i,t), and links are represented as arcs (i,j,t,t'), where $(t'-t)$ reflects travel time. This transforms physical paths into space-time paths with detailed temporal sequences. For instance, space-time path p_1 is expressed as: $(1,2,0) \rightarrow (2,3,1) \rightarrow (3,4,3) \rightarrow (4,7,4) \rightarrow (7,8,8)$.

This framework enables the construction of path-to-link temporal correlation matrices, capturing the two-dimensional relationships between paths and links across time intervals, as shown in Fig. 2. These matrices provide a structured view of temporal dependencies, highlighting how paths map to links over time for addressing inconsistencies in link-specific travel times.

2.1.2 Model formulation

In this section, we present the mathematical model for the coordinated estimation problem under real-time applications.

The objective function of the model aims to minimize the squared deviations between the estimated results (the flow volume $f_{i,j}(t)$, the OD distribution proportions $\rho_{o-w}(t)$, and the path choice proportions $\rho_{w-p}(t)$) and the observed parameters (the link volumes $\hat{f}_{i,j}(t)$, the historical

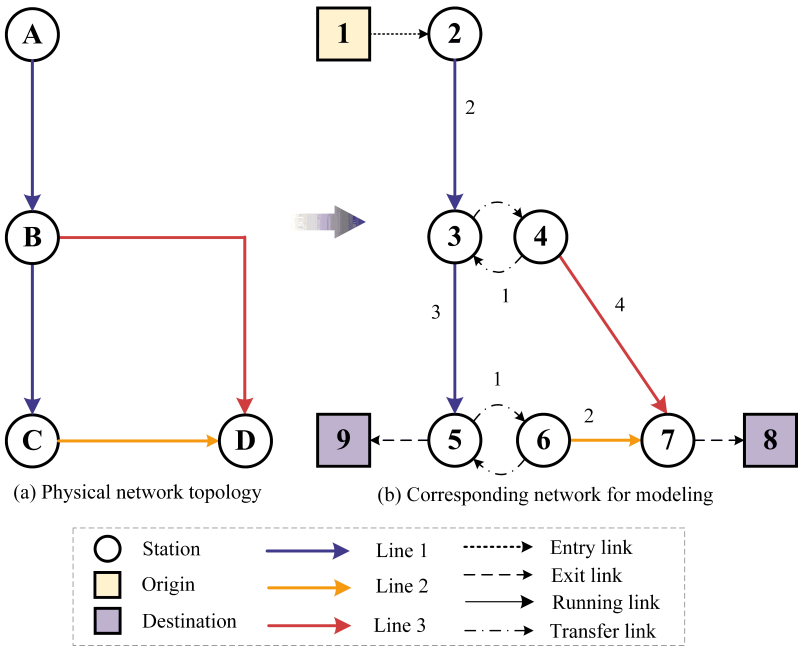


Figure 1: A two-OD four-station urban rail transit network.

proportions of OD distributions $\hat{\rho}_{o-w}^{t^{\text{his}}}$, and the historical path choices $\hat{\rho}_{w-p}^{t^{\text{his}}}$).

$$\min Z_{\text{error}}(t) = \sum_{\forall (i,j) \in L_p^{\text{obs}}} \left[f_{i,j}(t) - \hat{f}_{i,j}(t) \right]^2 + \sum_{\forall p \in P_w^{\text{obs}}} \left[\rho_{w-p}(t) - \hat{\rho}_{w-p}^{t^{\text{his}}} \right]^2 + \sum_{\forall w \in W_o^{\text{obs}}} \left[\rho_{o-w}(t) - \hat{\rho}_{o-w}^{t^{\text{his}}} \right]^2$$

The constraints of the model ensure consistency and feasibility. Origin demand $\hat{f}_o(t)$ is allocated to OD flows $f_w(t)$, distributed to path flows $f_p(t)$, and mapped to link volumes $f_{i,j}(t')$, with temporal dependencies captured by the path-to-link temporal correlation matrix $\delta_{i,j,t'}^{p,t}$. Proportional variables $\rho_{o-w}(t)$ and $\rho_{w-p}(t)$ are defined to represent the relationships between origins, OD pairs, and paths, linking flows across layers. The flows are subject to non-negativity constraints ($f_w(t) \geq 0, \quad f_p(t) \geq 0, \quad f_{i,j}(t) \geq 0$) and conservation constraints ($\sum_{\forall w \in W_o} \rho_{o-w}(t) = 1, \quad \sum_{\forall p \in P_w} \rho_{w-p}(t) = 1$), ensuring balanced distributions.

2.2 Feedback-Enhanced Iterative Algorithm

2.2.1 Feedback-enhanced iterative mechanism based on rolling horizons

The proposed real-time estimation framework uses a rolling mechanism to link multi-source data across time intervals, as shown in Fig. 3. At each time interval t (estimation horizon), OD distributions and path choices are estimated to determine network-wide link flows. Contributions from prior intervals (calibration horizon) and ongoing trips into future intervals (prediction horizon) are incorporated. The framework integrates four data types: real-time observations (e.g., AFC data, partial link flows), historical data (e.g., OD and path choice proportions), predicted data from previous intervals, and calibrated data from subsequent intervals. By propagating predictions forward and calibrations backward, the framework iteratively refines estimates, improving accuracy over time.

2.2.2 Time-extended computational graph and model reformulation

For real-time coordinated estimation of OD distributions and path choices, a time-extended, four-layer computational graph is developed as Fig. 4, comprising the Origin, OD, Path, and Link layers. Incorporating a temporal dimension ensures accurate representation of travel times, particularly in the transition from Path to Link layers, where a time-shift mechanism aligns path flows with link flows based on a path-to-link temporal correlation matrix introduced in Section 2.1.1, addressing temporal dependencies effectively.

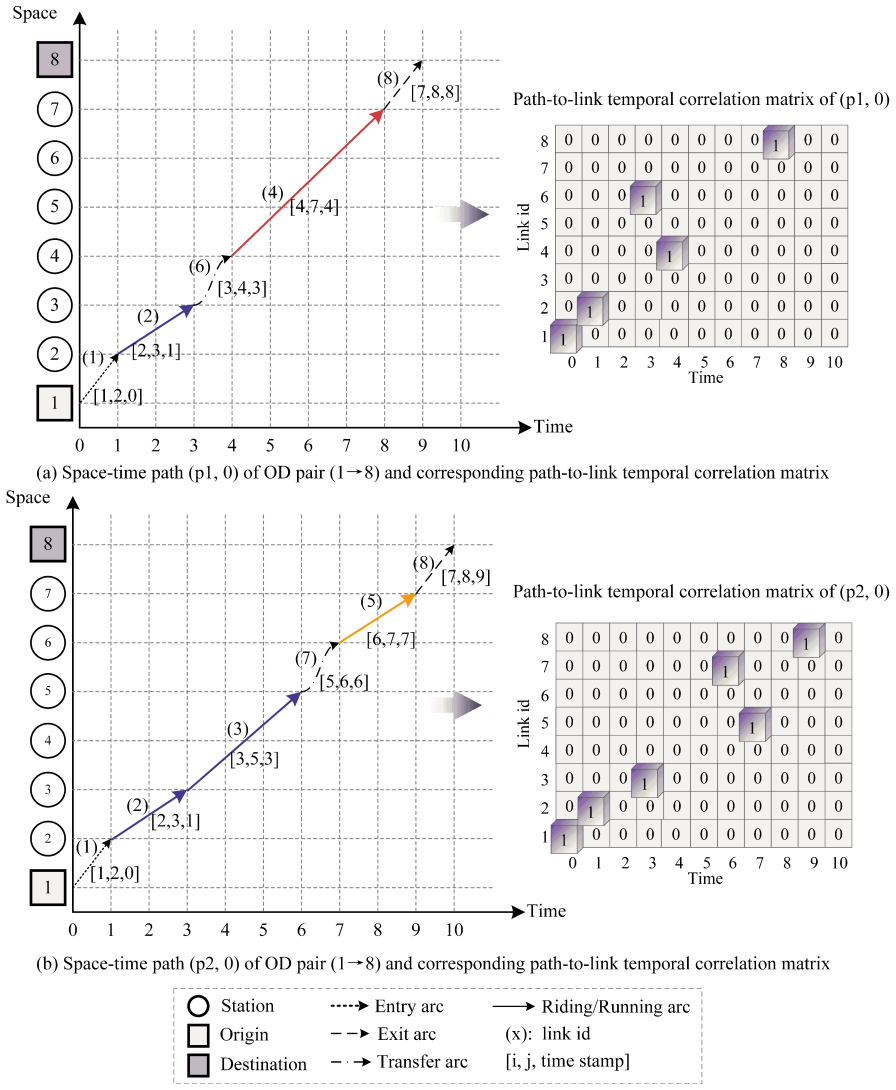


Figure 2: Two space-time paths with path-to-link temporal correlation matrices.

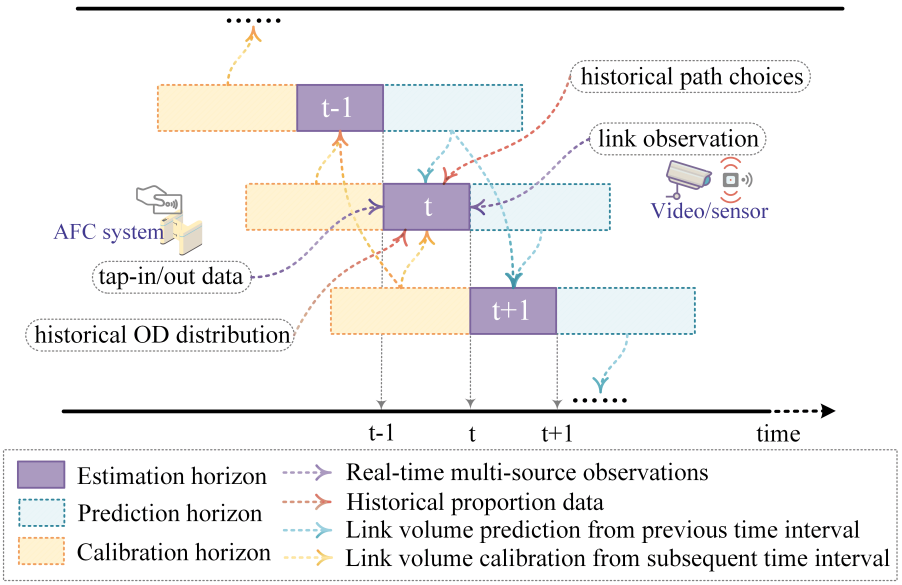


Figure 3: Illustration of feedback-enhanced iterative mechanism based on rolling horizons

To address the high-dimensional model established in Section 2.1.2, we reformulate it within the

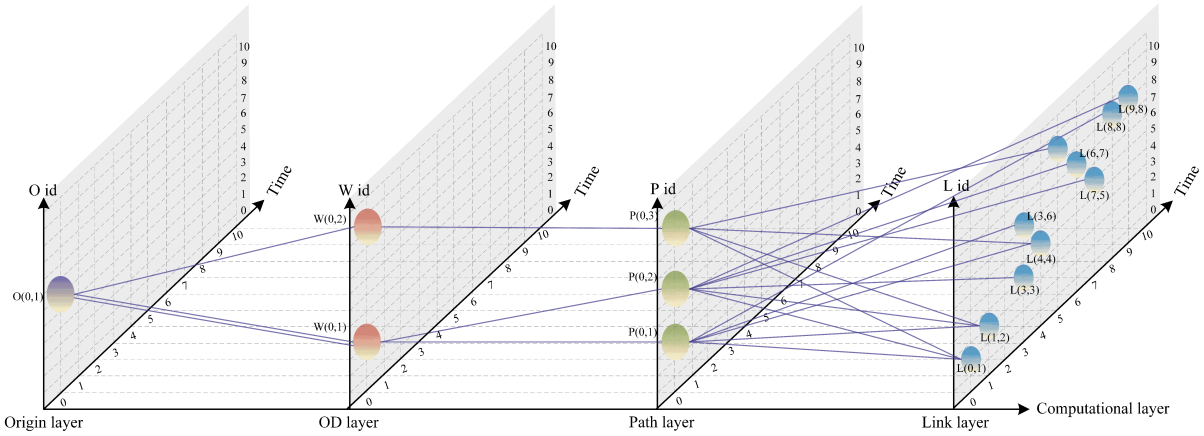


Figure 4: Illustration of the time-extended four-layer computational graph

four-layer computational graph by introducing two flow proportion variables $\rho_{o-w}(t)$ and $\rho_{w-p}(t)$. Edges in the computational graph correspond to these variables, linking elements across layers. For instance, $\rho_{o-w}(t)$ defines edges from the Origin to OD layers, and $\rho_{w-p}(t)$ connects OD to Path layers.

This reformulation simplifies decision variables $f_w(t)$, $f_p(t)$, and $f_{i,j}(t)$ into edge weights that are iteratively updated during forward and backward propagation, enabling efficient optimization within the computational graph framework.

2.2.3 Operations for training the computational graph

(1) Forward passing
The *Forward passing* sub-algorithm calculates time-dependent flow distributions by sequentially propagating flows through the computational graph's four layers. As shown in Fig. 5, flows are allocated from origins to OD pairs using OD weights, distributed across paths based on path weights, and mapped to links via the path-to-link temporal correlation matrix, ensuring accurate network-wide flow estimation.

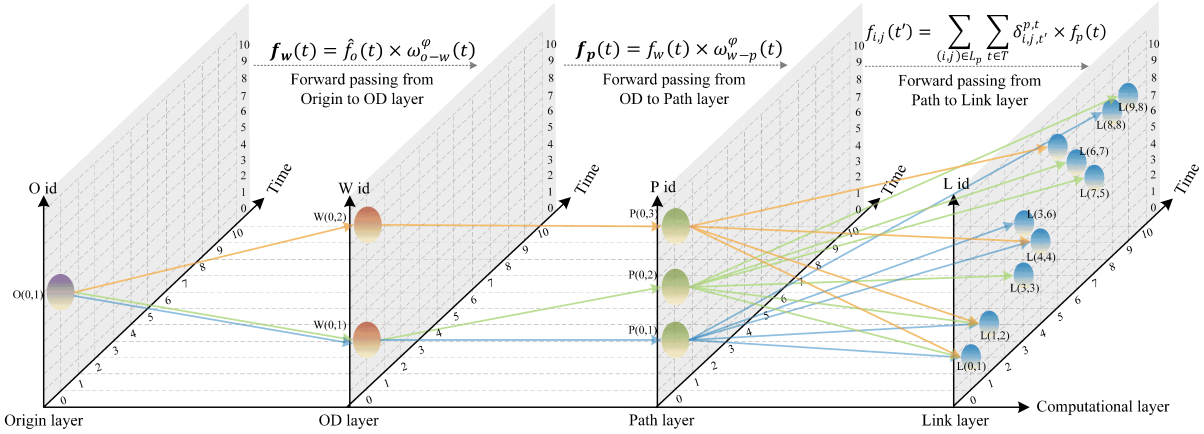


Figure 5: Illustration of the Forward passing in the four-layer computational graph

(2) Backward propagation
The *Backward propagation* sub-algorithm serves as a corrective mechanism, incorporating temporal feedback to refine flow distributions across the computational graph. While *Forward passing* distributes flows from the Origin layer to the Link layer, *Backward propagation* adjusts initial distributions by propagating errors from observed and estimated flows. As illustrated in Fig. 6, starting at the Link layer, deviations are calculated and back-propagated through the Path, OD, and Origin layers, spanning time intervals. Unlike traditional methods, this process leverages temporal feedback, allowing errors from future intervals to refine earlier estimates. By capturing temporal and spatial dependencies, this iterative recalibration improves estimation accuracy and

ensures consistency across layers and time.

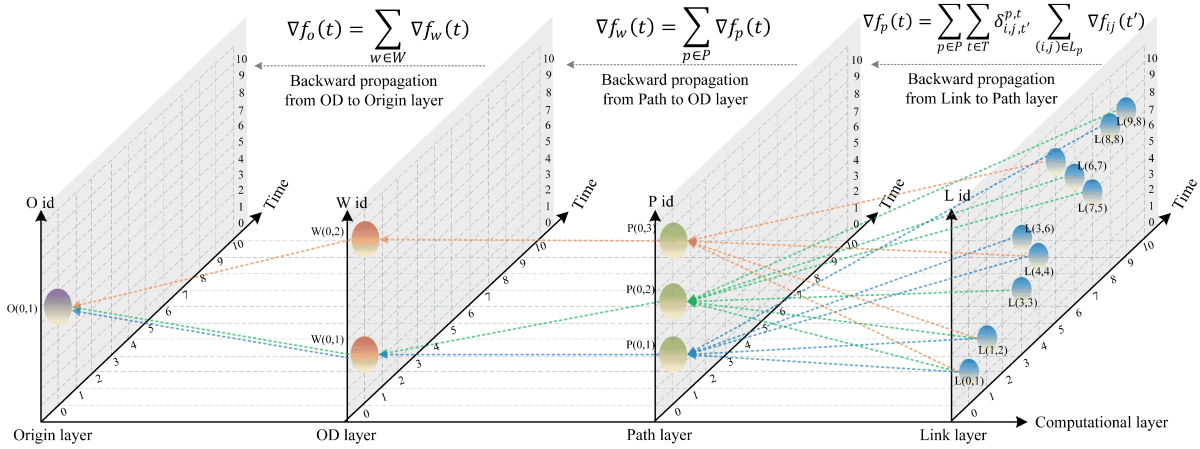


Figure 6: Illustration of the Backward Propagation in the four-layer computational graph

(3) Weight update & Regulation
The *Weight update and regulation* sub-algorithm refines flow weights between layers using gradient descent to minimize residual errors while maintaining stability. Iterative adjustments ensure optimal flow distribution, with a regulation step enforcing weights to sum to one, preserving logical consistency, stability, and convergence across OD pairs and paths.

2.2.4 Framework of the feedback-enhanced iterative algorithm

We present a flowchart of the feedback-enhanced iterative framework in Fig. 7. The framework refines flow estimation across a four-layer computational graph using *Forward passing*, *Backward propagation*, and *Weight update & regulation*. *Forward passing* distributes flows layer-by-layer, generating initial estimates, while *Backward propagation* leverages temporal feedback to adjust earlier estimates based on errors observed in later intervals. Weights are iteratively updated to ensure spatial-temporal consistency and global optimization. The algorithm iterates until the estimation error converges to a predefined threshold ε or the maximum iterations $\phi^{\max}$ are reached, outputting dynamically calibrated spatial-temporal flow distributions. This scalable, modular framework progressively enhances accuracy, adapts to temporal variations, and offers robust solutions for dynamic transportation network optimization.

3 Results and discussion

3.1 Numerical Validation on Sioux-Falls Network

3.1.1 Network Overview

A simplified Sioux-Falls network (Fig. 8) with 3 origins, 2 destinations, 17 stations, and 34 links was constructed to analyze passenger flow. With 700 passengers and 30 time intervals, initial OD and path proportions were estimated using a prior framework (Shang et al., 2024). The solution converged in 30 iterations, reducing deviations to zero, and completed in 1 second, demonstrating the framework's efficiency and real-time applicability.

3.1.2 Convergence under detector noisy and demand fluctuation

The framework's robustness and efficiency were evaluated through two tests. First, 15 perturbed initial solutions demonstrated consistent convergence within 30 iterations, aligning with ground truth despite observational noise (Fig. 9). Second, six scenarios with passenger demand levels from 280 to 2100 showed rapid convergence to zero within 15 iterations (Fig. 10), highlighting the framework's reliability in handling dynamic demand fluctuations. Notably, solution speed depended on network complexity, not demand levels.

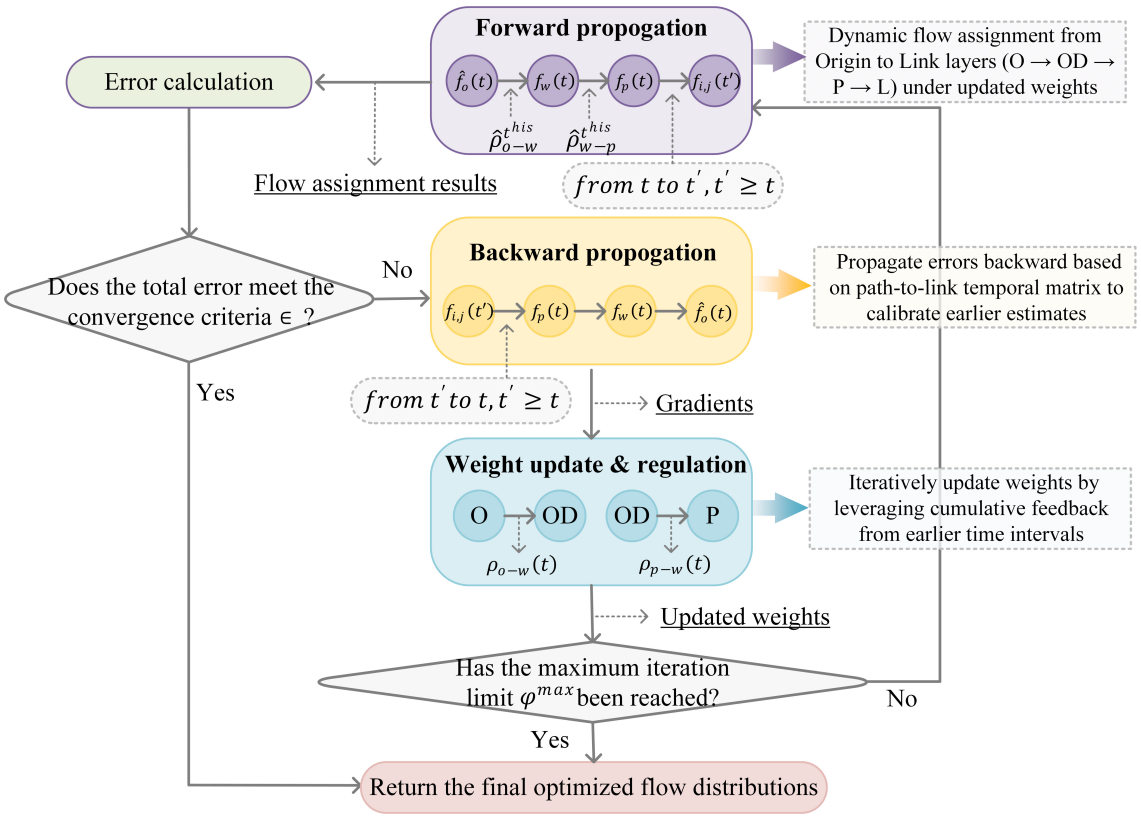
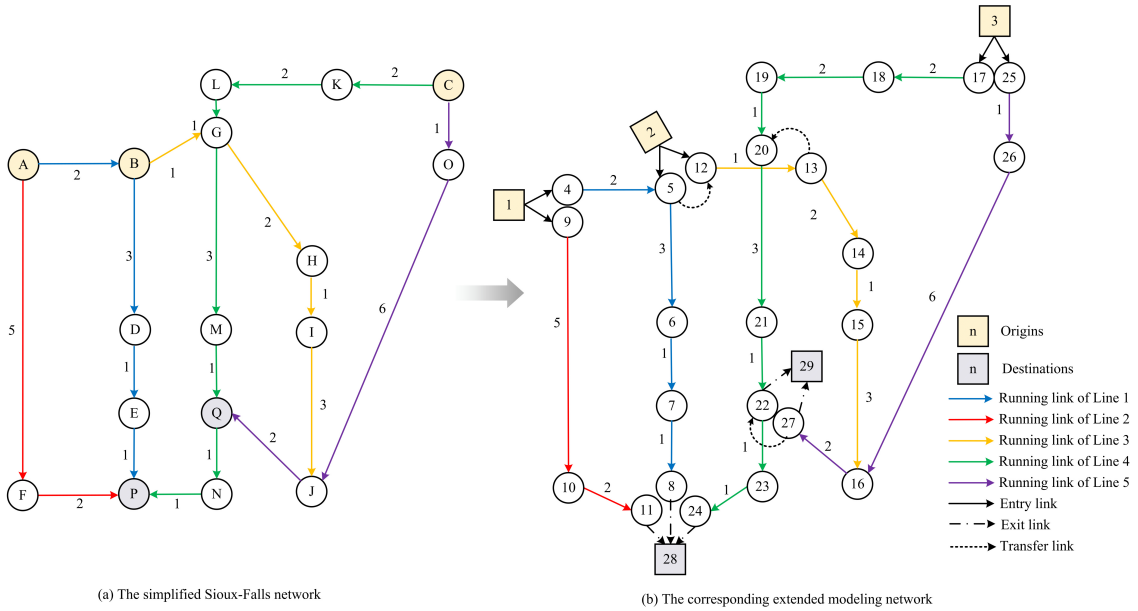


Figure 7: A flowchart for the feedback-enhanced iterative algorithm



(a) The simplified Sioux-Falls network

(b) The corresponding extended modeling network

Figure 8: The simplified Sioux-Falls network and corresponding modeling network

3.1.3 Parameter calibration of learning rates

Parameter calibration is crucial in the proposed estimation framework, particularly for the learning rates of OD distribution and path choice proportions. As shown in Fig. 11, testing 49 learning rate combinations (0.00005 to 0.02) revealed that moderate rates improve convergence accuracy, while excessively large rates (>0.02) cause failures. Path choice learning rates (0.0005–0.02) significantly impact accuracy, whereas OD learning rates have minimal influence. These findings emphasize prioritizing accurate path choice estimation for reliable URT modeling.

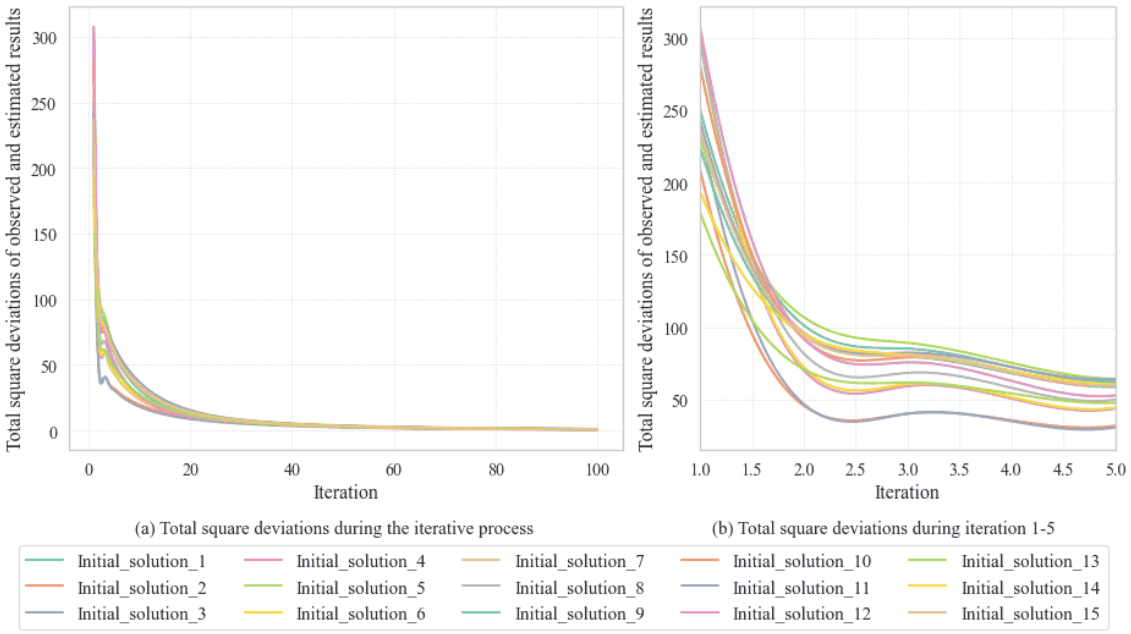


Figure 9: Iterative processes starting from 15 different initial solutions with random perturbations

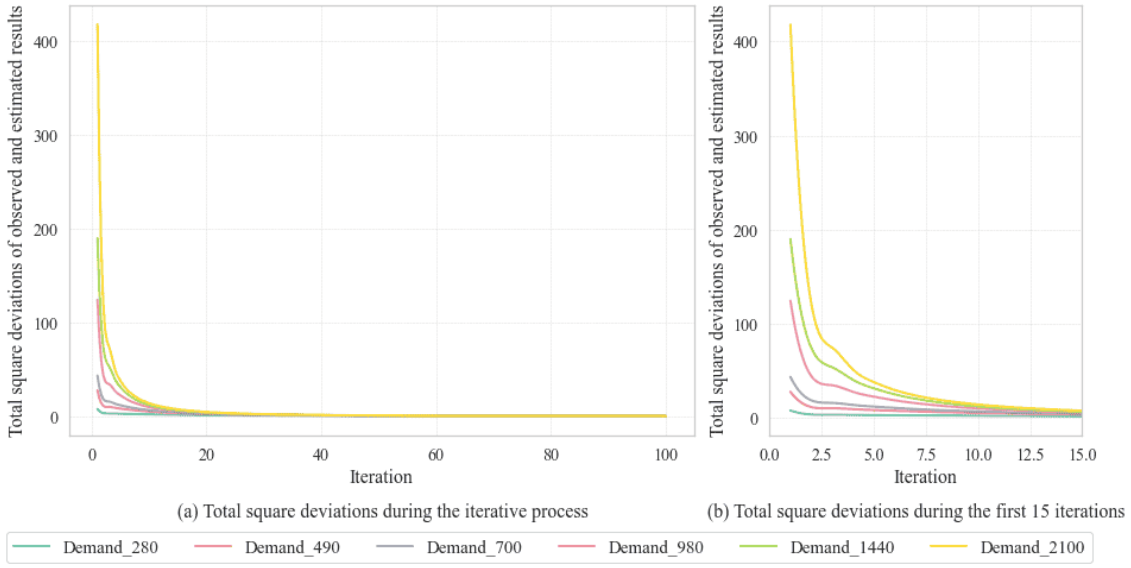


Figure 10: Iterative processes under different passenger demand scenarios

3.1.4 Method comparison with three baselines

To evaluate the proposed feedback-enhanced estimation framework ("Ours"), three alternative estimation methods were tested on the Sioux-Falls network: (1) Average observation method, which estimates flows by averaging historical observations, disregarding temporal variations; (2) Logit model method, which uses a probabilistic path choice model based on static OD distributions; and (3) Framework without feedback, which estimates flows sequentially without leveraging future observations for calibration.

Results in Fig. 12 show that "Ours" consistently outperformed all alternatives. It achieved the lowest convergence values across varying demand levels, remaining robust even in high-demand scenarios, while other methods showed poor adaptability, particularly the average observation (25.57% error) and logit-based (12.68% error) methods.
The feedback mechanism dynamically calibrates earlier estimates using later-time observations, ensuring global temporal-spatial consistency and preventing error accumulation. This makes the

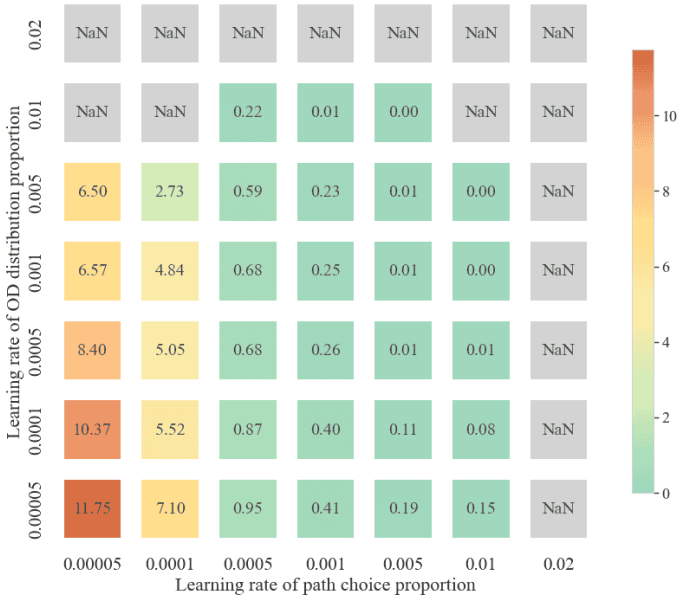
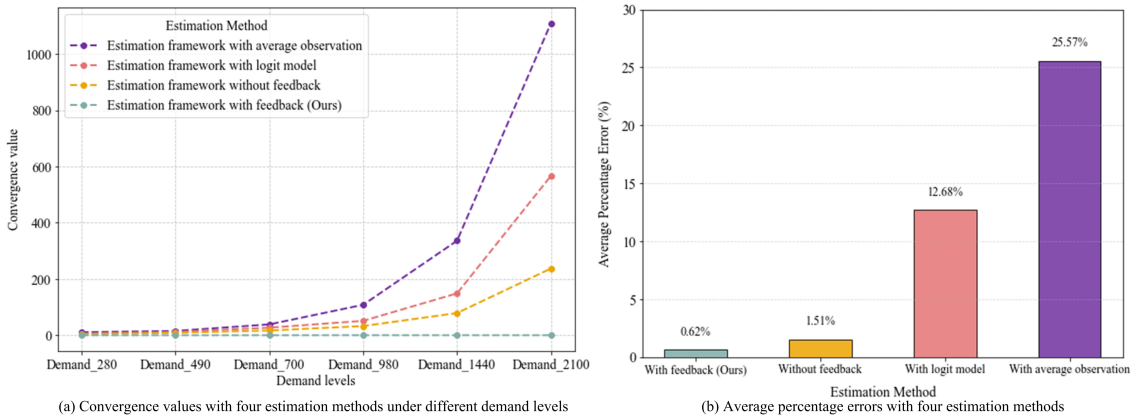


Figure 11: Convergences with different learning rates of OD distribution and path choice proportions.



(a) Convergence values with four estimation methods under different demand levels

(b) Average percentage errors with four estimation methods

Figure 12: Comparisons among different estimation frameworks.

framework particularly effective for dynamic or high-demand scenarios, achieving superior accuracy, robustness, and efficiency.

3.2 Real-World Application: Beijing Metro Network

3.2.1 Network Overview

The Beijing Subway features 27 lines, 490 stations, and spans 836 km, in Fig. 13. This study models the network with 1,740 nodes and 3,643 links, analyzing 1.96 million AFC records and link flow observations for 980 sections during the morning peak (6:00–10:00). The framework estimates 152,100 OD pairs and 381,990 paths, retaining 2-3 candidate paths per pair for efficient analysis.

3.2.2 Estimation results

The framework for the Beijing metro network completes 20 iterations in approximately 40 minutes, estimating time-dependent OD distributions for 152,100 OD pairs, path choices for 1.96 million passengers, and link flows for 980 sections.

(1) OD distributions
Fig. 14 visualizes the OD distributions across the entire Beijing URT network, where node size represents origin demand, and line thickness indicates flow magnitude. As shown, during the

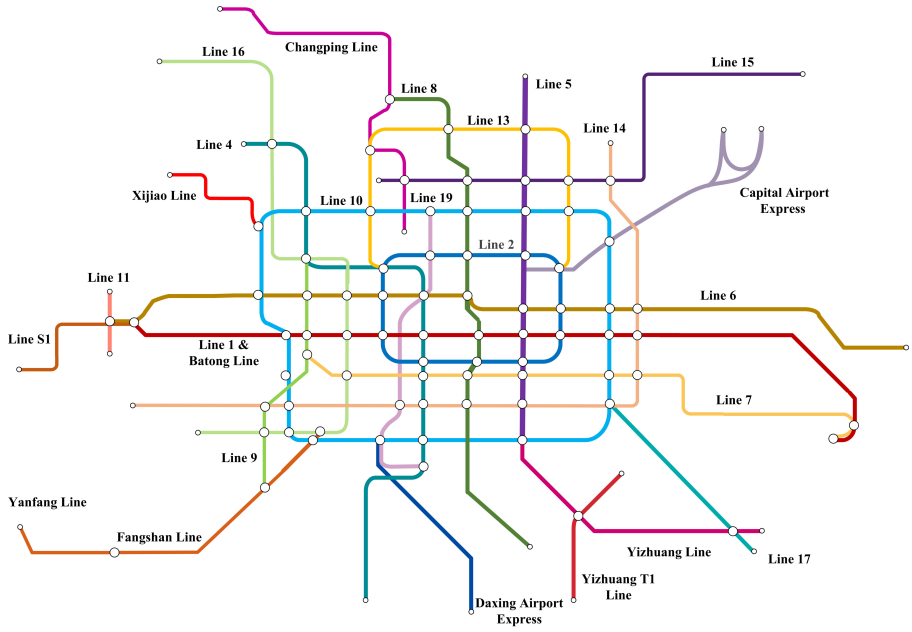


Figure 13: Layout of the Beijing metro network

study period (6:00–9:00), OD demand is concentrated primarily during the morning peak hours (7:15–9:30). Notably, several high-demand nodes are located in the northern region, with the overall demand pattern exhibiting a radial flow from suburban areas (primarily in the north) toward the city center.

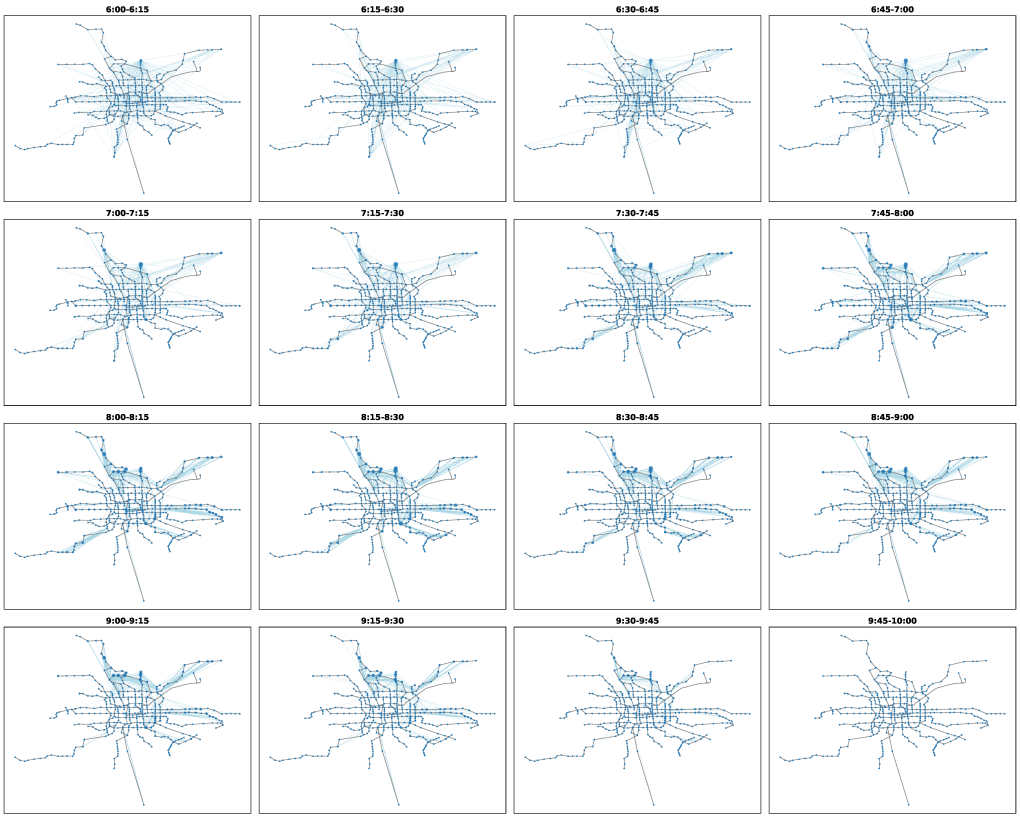


Figure 14: The estimated OD distribution among the Beijing metro network

(2) Link flows

The estimated link flows are illustrated in Fig. 15. Colored lines represent sections with varying passenger volumes, where green denotes lower volumes and red denotes higher volumes. As ob-

served, congestion is more severe between 7:45 and 9:00, primarily concentrated in the city center.

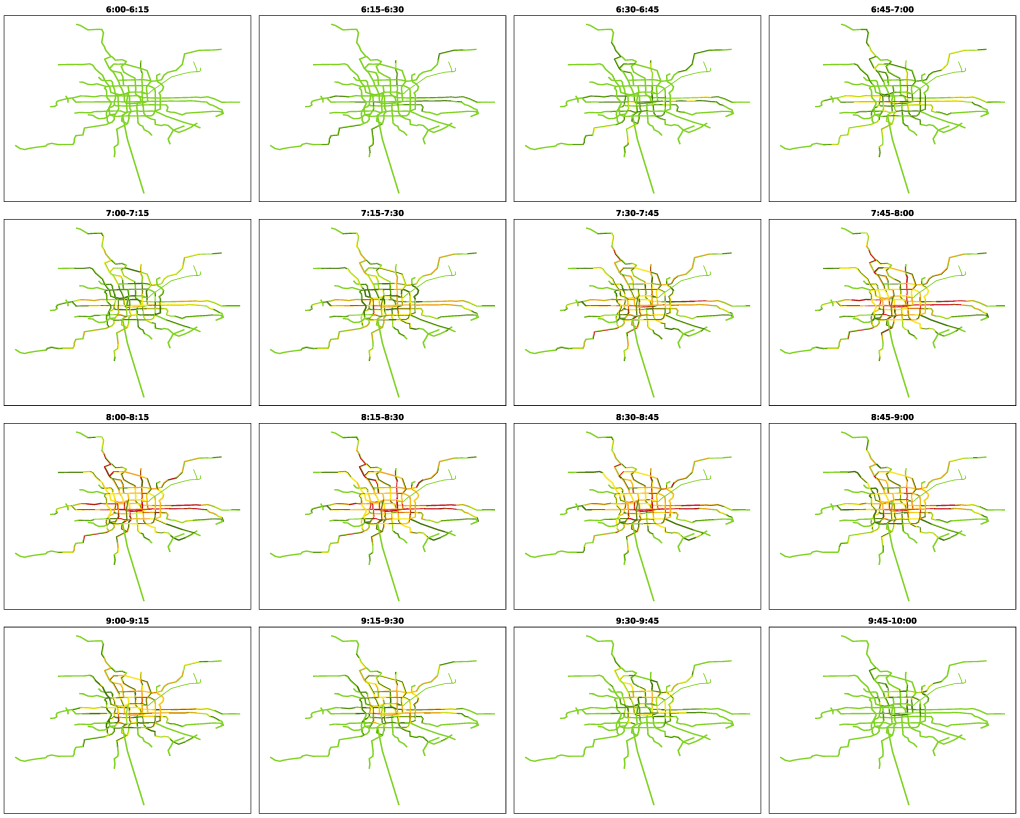


Figure 15: The estimated link flow results among the Beijing metro network

3.2.3 Estimation accuracy

The framework estimated 94,718 section flows with an average accuracy of 98.86%, with 94.28% of sections showing errors below 2%. Across 27 lines, all had errors under 4%, and 24 achieved errors below 2% (Fig. 16). These results demonstrate the framework's accuracy in handling the complexity of the Beijing metro network.

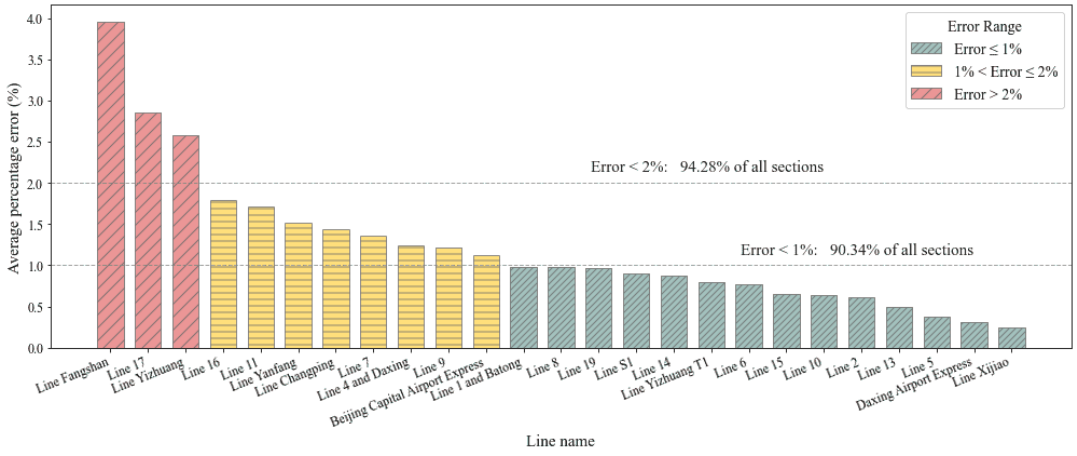
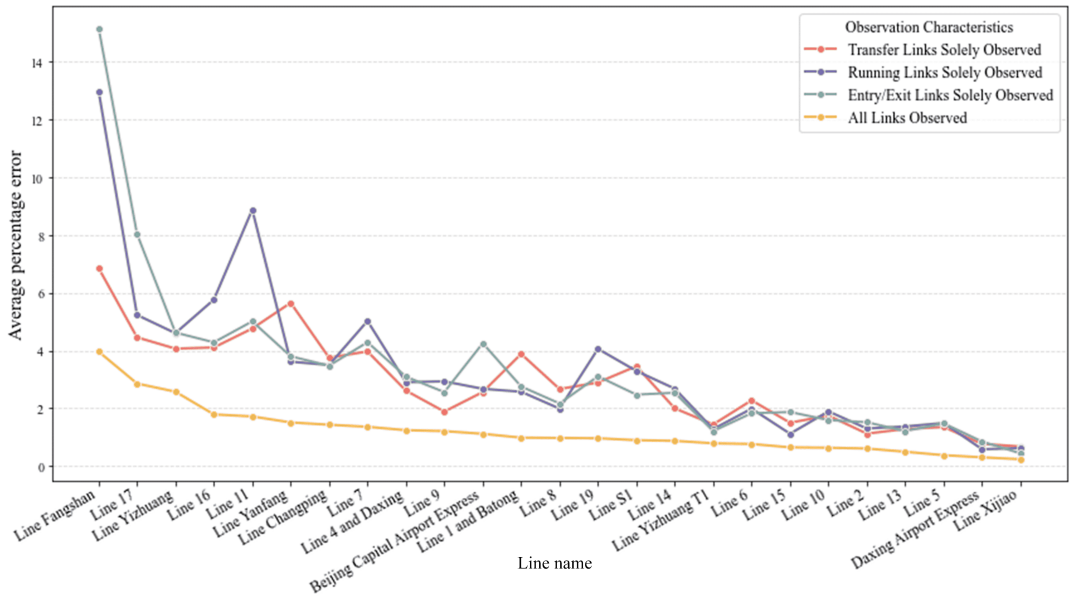


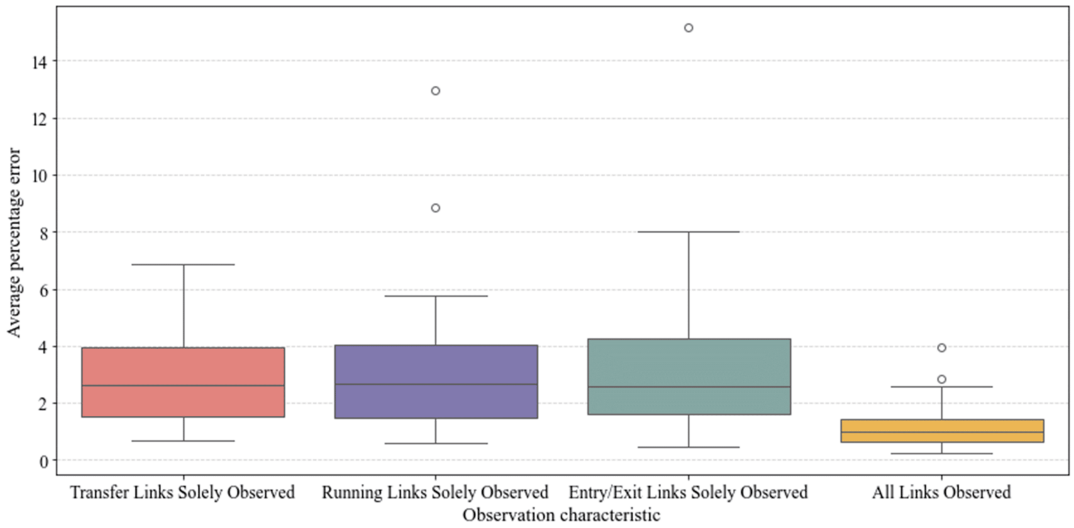
Figure 16: Average percentage error of all sections for each line

3.2.4 Effect of sensor locations

Data sparsity challenges URT systems, making optimal sensor placement critical. Three sensor types—AFC systems (entry/exit links), weight sensors (running links), and cameras (transfer links)—were tested across four deployment scenarios. As shown in Fig. 17, full observation yielded the lowest errors (<2%) and high-est stability but is resource-intensive. Observing only transfer links, while requiring fewer sensors, offered comparable accuracy and stability. These findings guide detector placement, balancing accuracy and re-sources in urban rail transit monitoring.



(a) Average absolute errors of all sections for each line under different observation characteristics (Line Plot).



(b) Average absolute errors of all sections for each line under different observation characteristics (Box Plot).

Figure 17: The average absolute errors for each line under different sensor deployments

4 Conclusions

This study proposes a feedback-enhanced iterative estimation framework for real-time passenger flow estimation in URT networks. Utilizing a time-extended, four-layer computational graph, it captures hierarchical relationships and integrates a path-to-link temporal correlation matrix to address temporal misalignments be-tween path and related links. A feedback mechanism with time rolling horizons iteratively calibrates OD distribution proportions and path choice weights, progressively improving accuracy. Validation on the Sioux-Falls and Beijing metro networks demonstrated the framework's robustness against detector errors and fluctuating demands. Comparative experiments confirmed its superior performance over alternative methods. Prioritizing detectors on

transfer arcs proved effective for resource-constrained scenarios. Future work should incorporate advanced behavioral models, deep learning techniques, and distributed computing frameworks to improve computational efficiency and scalability for large-scale networks.

References

Munizaga, M., Devillaine, F., Navarrete, C., & Silva, D. (2014). Validating travel behavior estimated from smartcard data. *Transportation Research Part C: Emerging Technologies*, *44*, 70–79. https://doi.org/10.1016/j.trc.2014.03.008

Pelletier, M.-P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, *19*, 557–568. https://doi.org/10.1016/j.trc.2010.12.003

Castillo, E., Jiménez, P., Menéndez, J. M., & Nogal, M. (2013). A bayesian method for estimating traffic flows based on plate scanning. *Transportation*, *40*, 173–201. https://doi.org/10.1007/s11116-012-9443-4

Xiong, X., Ozbay, K., Jin, L., & Feng, C. (2020). Dynamic origin–destination matrix prediction with line graph neural networks and kalman filter. *Transportation Research Record: Journal of the Transportation Research Board*, *2674*, 491–503. https://doi.org/10.1177/0361198120919399

Xu, H., Chen, Y., Li, C., & Chen, X. M. (2024). Space-time adaptive network for origin-destination passenger demand prediction. *Transportation Research Part C: Emerging Technologies*, *167*, 104842. https://doi.org/10.1016/j.trc.2024.104842

Szeto, W. Y., & Jiang, Y. (2014). Transit assignment: Approach-based formulation, extragradient method, and paradox. *Transportation Research Part B: Methodological*, *62*, 51–76. https://doi.org/10.1016/j.trb.2014.01.010

Huang, H., Mao, J., Kang, L., Lu, W., Zhang, S., & Liu, L. (2024). A dynamic graph deep learning model with multivariate empirical mode decomposition for network-wide metro passenger flow prediction. *Computer-Aided Civil and Infrastructure Engineering*, *39*, 2596–2618. https://doi.org/10.1111/mice.13214

Shang, P., Xiong, Y., Guo, J., Xian, K., Yu, Y., & Xu, H. (2024). A modeling framework to integrate frequency- and schedule-based passenger assignment approaches for coordinated path choice and space-time trajectory estimation based on multi-source observations. *Transportation Research Part B: Methodological*, *183*, 102945. https://doi.org/10.1016/j.trb.2024.102945