https://doi.org/10.1038/s44333-025-00072-2

Passenger flow distribution forecasting at integrated transport hub via group evolution mechanism and multimodal data

Zhicheng Dai[1,2], Dewei Li[1,3] ✉, Soora Rasouli[2], Yan Feng[4], Hua Li[5], Linhan Zou[1] & Ruonan Zhang[1]

Integrated transport hubs require reliable, fine-grained forecasts of crowd distribution to safeguard operations and sustainable urban mobility. We present Group Evolution Mechanism Embedded Network (GEME-Net), a passenger flow distribution forecasting architecture that fuses multimodal data, including video-derived counts, digital twin-based mobility chains, and railway/metro operations information, via multi-graph spatial representations and event-aware temporal modules, with a distilled lightweight student model for deployment. In a real-world case at Shanghai Hongqiao, GEME-Net consistently outperforms statistical, convolutional, recurrent, graph-based and Transformer baselines across MAE, RMSE and WMAPE, while retaining inference latency compatible with near-real-time use. Ablations indicate that schedule encoding and event-driven frequency enhancement, together with learned long-range and community graphs, are principal contributors to accuracy. By coupling operational signals with spatial semantics, our approach improves hub-scale situation awareness and short-horizon decision support, offering a practical route to resilient crowd management without asserting broader societal or policy impacts.

Amid the intensifying global warming and environmental pollution, the pursuit of sustainable transportation solutions has become an urgent priority. Multimodal public transportation systems, integrating buses, subways, railways and other modes, offer a viable path for reducing carbon emissions, alleviating urban congestion and promoting sustainable urban development[1]. Compared with traditional single-mode systems, multimodal transport significantly enhances travel efficiency[2], reduces environmental burdens[3], and plays a vital role in addressing climate change and advancing sustainability. Integrated transport hubs serve as the core nodes of multimodal transportation networks. These hubs connect various transport modes within and between urban regions, ensuring the seamless daily mobility of residents and travelers[4]. As crucial anchors of urban transport systems, they greatly facilitate commuter, tourist, and business travel by optimizing transfer experiences and improving service accessibility[5], thereby encouraging greater public transit utilization.

As illustrated in Fig. 1, an integrated transport hub is defined as a comprehensive infrastructure connecting multiple transport services, typically encompassing railway, metro, bus, and airport nodes. By providing a unified platform for intermodal connectivity, such hubs enable efficient and convenient transfers for daily travelers. However, the operation of these hubs faces several significant challenges:

(1) Crowd and safety: With the convergence of multiple transport modes, integrated transport hubs are often required to handle enormous volumes of passenger traffic. During peak hours, concentrated inflows may lead to severe crowding, exceeding the hub's designed capacity and posing serious challenges to operational efficiency and passenger safety.
(2) Complex passenger composition and spatiotemporal distribution. Passenger flows within hubs are composed of heterogeneous groups, including commuters, tourists, and business travelers. At the same time, hubs typically integrate a variety of non-transportation services, such as retail, dining, and leisure facilities, leading to increased heterogeneity in passengers' spatial travel behavior. The overlap of different travel purposes and transportation modes results in complex passenger flow generation patterns and spatial-temporal distributions. This diversity increases the difficulty of accurate prediction and real-time management, presenting significant operational challenges.
(3) There is no clear quantitative method yet for assessing the impact of public transport operations on passenger flow fluctuations. Despite the heterogeneity of travel within a given space, passenger behavior within a hub is still largely constrained by the operational dynamics of public transport as departure time approaches. Therefore, it is necessary to

[1]School of Traffic and Transportation, Beijing Jiaotong University, Beijing, China. [2]Department of Built Environment, Eindhoven University of Technology, Eindhoven, The Netherlands. [3]Frontiers Science Center for Smart High-speed Railway System, Beijing Jiaotong University, Beijing, China. [4]Transport & Planning Department, Delft University of Technology, Delft, The Netherlands. [5]China Railway Shanghai Group Co., Ltd., Shanghai, China. ✉e-mail: lidw@bjtu.edu.cn

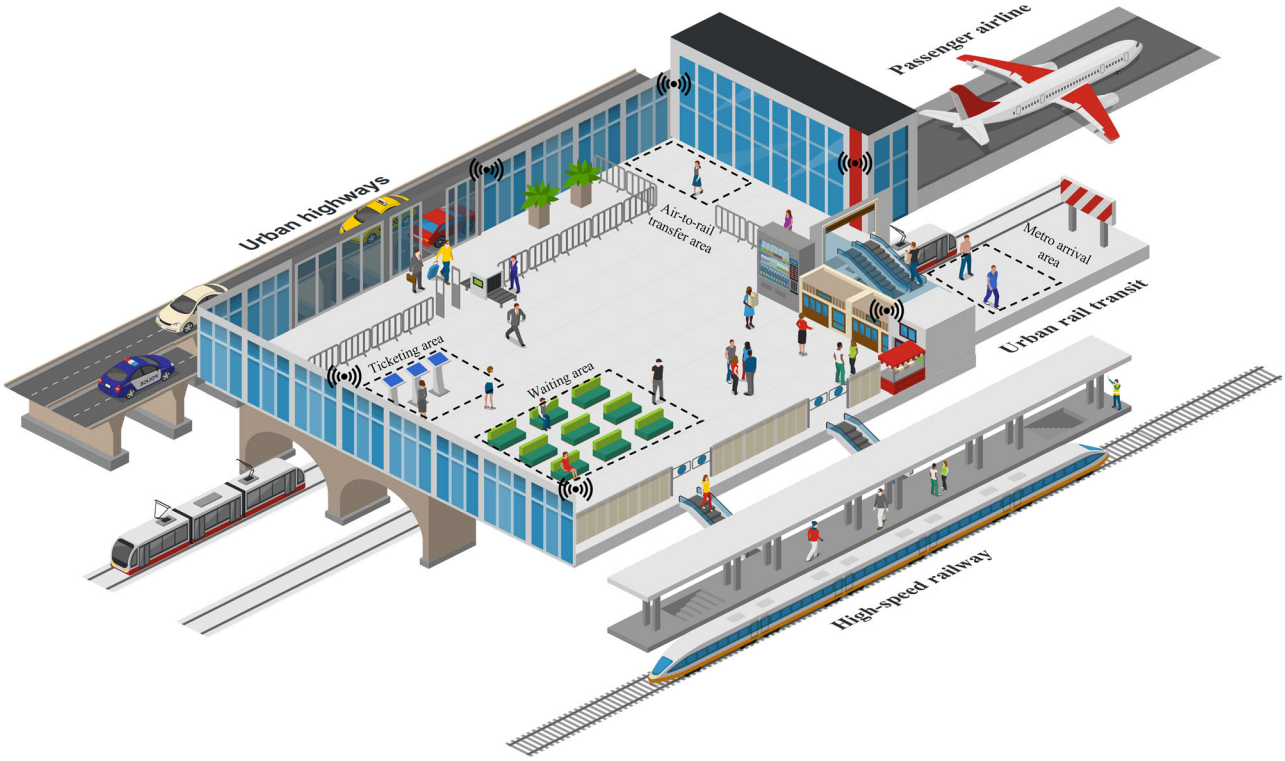


Fig. 1 | A real-world integrated public transportation infrastructure system. Travelers from multi-modal transport (airline, urban rail, road) converge in the space. Activities within the space are driven by the heterogeneous needs (ticketing, shopping and waiting) and travel purpose.

propose expressions that can quantify regional passenger flow fluctuations and multi-modal transport operational dynamics. Effectively modeling and accurately predicting these evolving patterns remain major technical and practical challenges.

To address the aforementioned challenges, it is essential to gain a deeper understanding of individual passenger behaviors within the hub's spatial environment. At the same time, integrating multimodal data, such as multi-angle surveillance video within the infrastructure, ticketing records, and public transportation operational data, enables the construction of more accurate passenger flow prediction models. Given the need for real-time demand forecasting in order to optimize hub operations, it is also crucial to develop lightweight predictive models that can deliver fast and reliable forecasts at low computational cost, thereby supporting dynamic and responsive management.

Passenger flow forecasting, as a classic time-series prediction problem, can generally be categorized into traditional statistical methods and machine learning-based approaches. Early studies often relied on parametric models, such as Autoregressive Integrated Moving Average (ARIMA) models[6–8] and historical average models (HA)[9]. However, these methods face substantial limitations when dealing with the complex and nonlinear dynamics of passenger flows, restricting their applicability in real-world scenarios[10]. During the same period, some researchers turned to computer-based simulation approaches. These include macroscopic dynamic models[11,12] and microscopic individual-based behavioral models[13–15], which simulate the dynamic evolution of crowds over time by scheduling agents' movements within virtual environments. While such simulations help capture basic structural patterns of flow and provide short-term demand estimations, they face challenges similar to those of parametric models. Essentially, simulation methods rely on parameterized assumptions about individual passenger behavior[16] or macroscopic flow dynamics. As a result, they struggle to accurately reflect the complex relationships between traveler movement, spatial layout of facilities, and external events in transport hubs, often leading to significant discrepancies between predicted and actual flow patterns.

The application of machine learning has introduced new perspectives to the task of accurate passenger flow forecasting. Algorithms such as Support Vector Machines (SVM)[17,18], k-Nearest Neighbors (KNN)[19,20], and Kalman Filtering[21] have been employed to predict passenger volumes, achieving improvements in forecasting accuracy. However, these models are limited in their ability to capture spatial features and long-term temporal dependencies inherent in time-series data, making them unsuitable for simultaneous forecasting across multiple spatial regions. Emerging deep learning techniques have become powerful tools for improving forecasting performance. In 2015, the Long Short-Term Memory (LSTM) network was first introduced into the field of traffic flow prediction[22]. Since then, a wide range of deep learning models based on architectures such as Convolutional Neural Networks (CNN)[23,24], Gated Recurrent Units (GRU)[25,26], and LSTM[27,28] have been proposed for passenger flow forecasting. Subsequent research recognized that regions within a transport network are not independent entities. Accordingly, many studies incorporated traffic network topologies into their model inputs and employed graph-based neural network modules, such as Graph Convolutional Networks (GCN)[29,30], Graph Attention Networks (GAT)[31,32], and Graph Transformer models[33] to jointly model temporal and spatial dependencies, thereby further enhancing predictive accuracy. However, several limitations remain in current research on passenger flow distribution forecasting in transport hubs. First, most existing models focus on coarse-grained metrics such as total inflow and outflow volumes at stations[34–36], with limited attention paid to how the intra-station spatial distribution of flows affects overall operational efficiency. Second, although spatial information has been widely incorporated into predictive models, the modeling of spatial networks is often based solely on the physical layout of the facility[37,38] or pre-defined activity routes[39]. This approach neglects the spatiotemporal correlations between non-adjacent areas and fails to account for actual passenger behavioral patterns, limiting the model's ability to capture the complex, dynamic relationships across functional areas within the hub. Considering external factors influencing

passenger flow fluctuations is also indispensable. Although existing prediction models incorporate and quantify the impact of weather[27], social media[10], and other information on passenger flows, thereby enhancing prediction accuracy, they lack deep exploration and quantification of external data, such as transportation operational information, which undoubtedly drives individual passenger travel behaviors. Consequently, these models have not yet established a modeling framework that captures the association mechanisms between passenger flow fluctuations and public transportation operational information.

In recent years, the introduction of the Transformer architecture[40] has brought a new paradigm to the design of passenger flow forecasting models. By leveraging multi-head attention mechanisms, Transformer networks are capable of adaptively capturing multi-scale temporal dependencies in time-series data while simultaneously modeling spatial–temporal correlations. These capabilities have established Transformer-based models as the state-of-the-art in the forecasting domain. Transformer-based approaches[41,42] have effectively addressed key limitations of traditional deep learning models, such as the difficulty in integrating spatial and temporal features and modeling long sequences. Specifically, the global self-attention mechanism at the core of Transformers inevitably incurs substantial computational overhead, which limits their suitability for real-time applications where fast inference is essential. As predictive models are increasingly expanded and deployed in real-time public transportation systems, keeping a balance between model complexity and predictive performance remains a critical challenge.

Building upon the limitations identified in existing studies, we position our work as a methodological and empirical contribution to hub-scale passenger flow forecasting. First, we design a regression-informed **T**rain-**S**chedule **E**ffect **E**ncoding and an **E**vent-**D**riven **F**requency-**E**nhanced **M**odule (TSEE and EDFEM) that translates timetable-driven influence into sparse attention masks and frequency-enhanced features, improving the predictive model's responsiveness and learning capacity regarding the short-term impact of event-induced volatility. Second, utilizing mobility chain data from VR behavioral experiments as input, we propose a graph construction strategy driven by topological features. This approach generates three spatial correlation graphs, integrating diverse spatial association patterns derived from the mobility network's topological structure. This enables a fine-grained representation of the non-local correlations and evolution mechanisms underlying spatial passenger flows. Third, leveraging the fusion of operational data of the integrated transportation hub and spatial correlation graphs constructed above, we propose a robust and generalizable passenger flow distribution forecasting model, named **G**roup **E**volution **M**echanism **E**mbedded **Net**work (GEME-Net). The model accounts for multiple dimensions of external factors—including passenger behavior, facility layout, and operational events—and embeds spatial-temporal features to enhance its adaptability to heterogeneous passenger groups and diverse spatial settings. GEME-Net addresses the shortcomings of traditional approaches in accurately predicting passenger dynamics under complex and variable conditions.

In terms of practical value for improving the operational efficiency of public transportation and enhancing passenger travel experience, our research offers several key contributions. Specifically, we employ knowledge distillation techniques to transfer knowledge from a complex deep learning model (teacher model) to a lightweight model (student model). This approach reduces computational costs while maintaining stable and reliable predictive performance. As a result, the distilled model could meet the practical requirements of real-time management and dynamic control in integrated transport hubs, where high efficiency and fast inference are essential for responsive operational decision-making.

Methods

Data sources

For empirical validation of our prediction framework, this study selects the Hongqiao Station located in Minhang District, Shanghai, China, as the case study site. The Shanghai Hongqiao Integrated Transport Hub spans a total area of over 1.3 million square meters, with the main waiting hall covering ~11,340 square meters and capable of accommodating up to 10,000 passengers simultaneously. The hub is seamlessly connected to Shanghai Hongqiao International Airport, as well as Shanghai Metro Lines 2, 10, and 17, and the city's road transportation network. The floor plan and functional area of Hongqiao Station's layout scheme are illustrated in Fig. 2.

The multimodal input data used in this study consist of three main branches: (1) time-series data of passenger flow distributions, (2) passenger mobility chain datasets, and (3) public transport operational information linked to railway infrastructure.

These datasets are sourced respectively from: (1) monitoring video data distributed throughout the hub infrastructure; (2) behavioral experiments conducted within a digital twin representation of the hub; (3) the Electronic Ticketing Management System (ETM).

First, monitoring video data was extracted from the station monitoring system for August 10, 2024, covering the period from 7:00 to 20:00, the operational hours during which the infrastructure is open to passengers. The video data were collected from 18 predefined areas located across the waiting hall level and the commercial level. To process this data, we developed an Automated Passenger Counting (APC) system. The system segments input surveillance videos from each zone into 10-second intervals, and then utilizes the Baidu AI Platform's automatic passenger counting API (https://ai.baidu.com/tech/body/num) to detect and count individuals within predefined spatial regions. Finally, time-aligned passenger counts from all zones are aggregated to generate the spatiotemporal passenger flow distribution dataset X^p_{N,T_h} used in this study.

$$X^p_{N,T_h} = \begin{pmatrix} x^p_{1,t-T_h+1} & \cdots & x^p_{N,t-T_h+1} \\ x^p_{1,t-T_h+2} & \cdots & x^p_{N,t-T_h+2} \\ \vdots & \ddots & \vdots \\ x^p_{1,t} & \cdots & x^p_{N,t} \end{pmatrix} \tag{1}$$

where $X^p_{N,T_h} \in R^{T_h \times N}$ represents the collection of passenger flow data for N functional areas over T_h historical time steps. $x^p_{N,t}$ represent the passenger flow data for area N at time t.

Second, to capture the complete mobility chains of passengers within the hub space, we constructed a high-fidelity digital replica of Shanghai Hongqiao Station, including all vertical levels and the connecting corridors to metro station exits and the airport transfer passages. Behavioral experiments conducted within digital representations of public spaces have been proven effective in studying pedestrian wayfinding strategies[43]. The digital environment was developed following a three-layer framework: Physical Layer → Input/Output Layer → Digital Layer, as illustrated in the "Digital Scenario Layer" module of Fig. 3.

In October 2024, we recruited 60 graduate students (35 male, 25 female; aged 21–30, average age 24.3) to take part in a behavioral experiment based on the digital environment. Prior to the experiment, each participant completed a questionnaire that collected demographic information (e.g., gender, age, previous railway travel experience) and habitual behaviors (e.g., whether they typically collect tickets or shop during waiting periods). Participants who reported such non-mandatory behaviors were assigned optional task nodes (e.g., ticketing, shopping) in the digital scenario, allowing their behavior in the virtual environment to realistically reflect their real-world travel habits. In the main experiment, each participant was asked to complete six independent wayfinding tasks. At the beginning of each task, the system randomly assigned an entry gate and a target boarding gate corresponding to a specific train, simulating realistic multimodal transfer scenarios. Each trial ended once the participant reached the assigned boarding gate.

During the experiment, participants' movements were recorded at a frequency of 1 second, resulting in a total of 420 valid trajectory sequences.

We treated the transfer counts between regional pairs as count variables and examined their distributional characteristics to estimate the

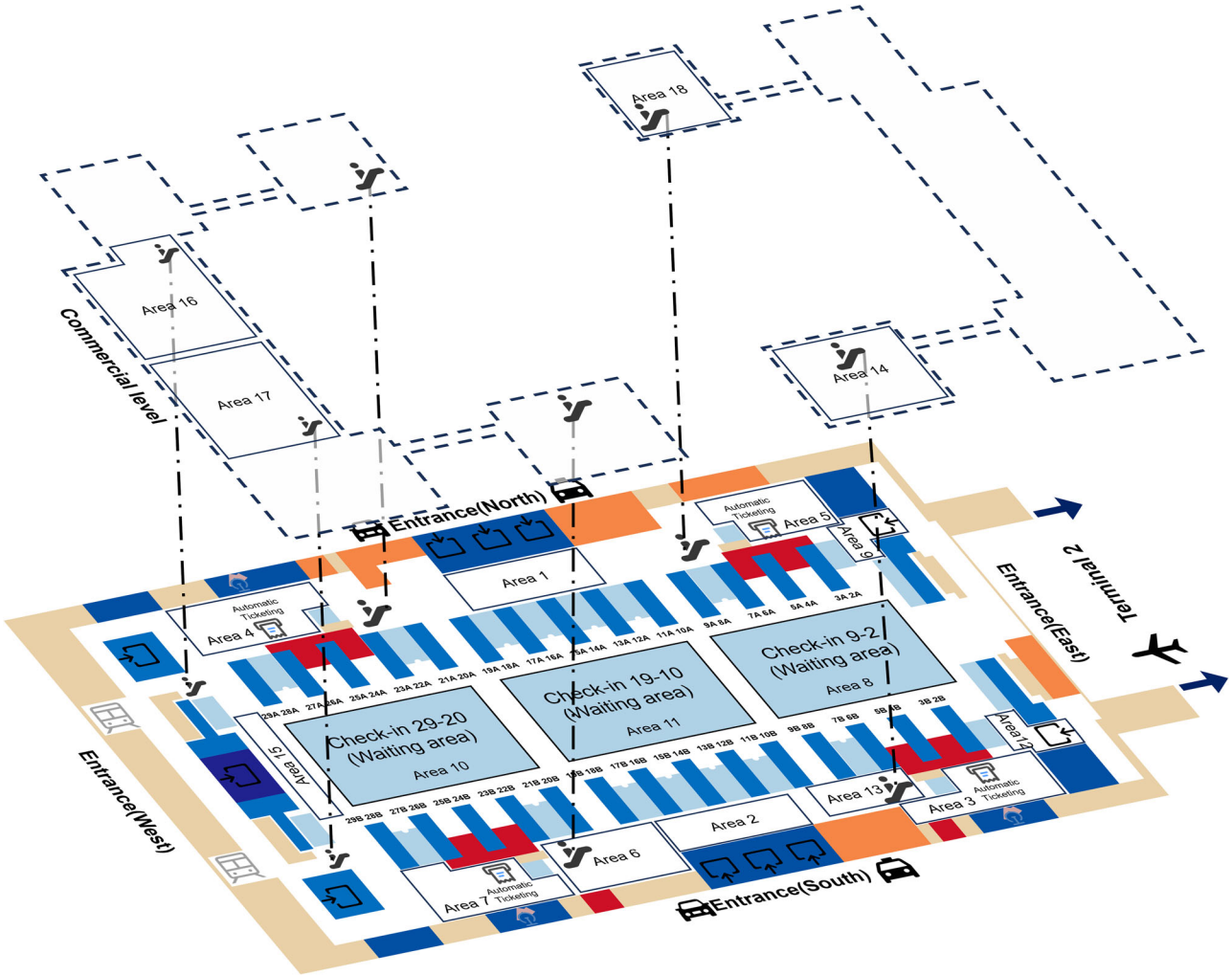


Fig. 2 | The layout and area division plan of the Shanghai Hongqiao integrated transport hub. Entrance area: 1,2,9,12,15. Ticketing area: 3,4,5,7. Waiting area: 8,10,11. Commercial area: 14,16,17,18. Transition area: 6,13.

minimum sample size required for the study. The overall transfer frequency histogram (Fig. 4) shows a pronounced right-skewed distribution. The Kolmogorov–Smirnov test (statistic = 0.457, $p < 0.001$) rejected the Poisson distribution. We therefore compared Poisson, mixed Poisson, and negative binomial models, using the minimum AIC criterion for model selection, and found that the negative binomial model (AIC = 1626.6) provided a significantly better fit. Accordingly, the negative binomial distribution was adopted to characterize the transfer counts. This result confirms the heavy-tailed property, where a small number of high-frequency transfers dominate the overall transfer pattern. Based on this, we assumed that the typical transfer probability corresponds to the contribution ratio of highly correlated regional pairs and used the top 10% of high-transfer pairs as the estimation benchmark. The theoretical minimum sample size was then calculated using the standard single-population proportion formula[44] $\bar{n} = Z^2_{\alpha/2} \cdot p \cdot (1-p)/E^2$, with a 95% confidence level ($Z_{\alpha/2} = 1.96$) and an error margin $E = \pm 5\%$. The resulting minimum sample size was 322, indicating that the 420 activity-chain records obtained from behavioral experiments in this study meet the required threshold.

Each trajectory sequence contains the 3D spatial coordinates and corresponding timestamps t' of a participant i's position $p_i(t') = \{x_i(t'), y_i(t'), z_i(t')\}$ in the digital environment. Subsequently, we applied a region-matching function $\delta_m^{i,n}$ to identify and record the sequence and timing t' of area n's entries during each participant's activity. These entries were mapped to a predefined set of bounded spatial regions R_m^n, $\mathcal{R}_m\{R_m^n | k = 1, 2, \ldots, A_m\}$, $\mathcal{R}_m$, denoted as a collection of functional zones within the station, each defined by a number of A_m known set of vertex coordinates in the facility layout.

$$\delta_m^{i,n}(t') = \begin{cases} 1, & p_i(t') \in R_m^n \\ 0, & otherwise \end{cases} \tag{2}$$

As illustrated in Fig. 5, each participant's activity sequence was encoded based on the order in which they entered functional areas, starting from their initial entry into the station environment. By sequentially labeling the identified regions according to the first time of entry, we constructed a set of mobility chains $\mathcal{L}_T$ that represent the ordered spatial trajectories of passengers within the hub.

The public transportation operational data used in this study include: (1) railway timetables connected to the hub, (2) urban rail transit (URT) timetables, and (3) ticketing records extracted from the Railway ETM during the study period. All three datasets were resampled to align with the temporal resolution of the passenger flow data, enabling seamless integration into the forecasting model.

Exploratory characterization of Hongqiao hub's mobility network

To study the topological characteristics of passenger collective mobility networks from a macroscopic perspective, this research integrates the set of mobility chains obtained from prior behavioral experiments in a digitalized hub scenario. We transform the mobility chains into a directed weighted mobility network, where the weight on edge i to j equals the number of times

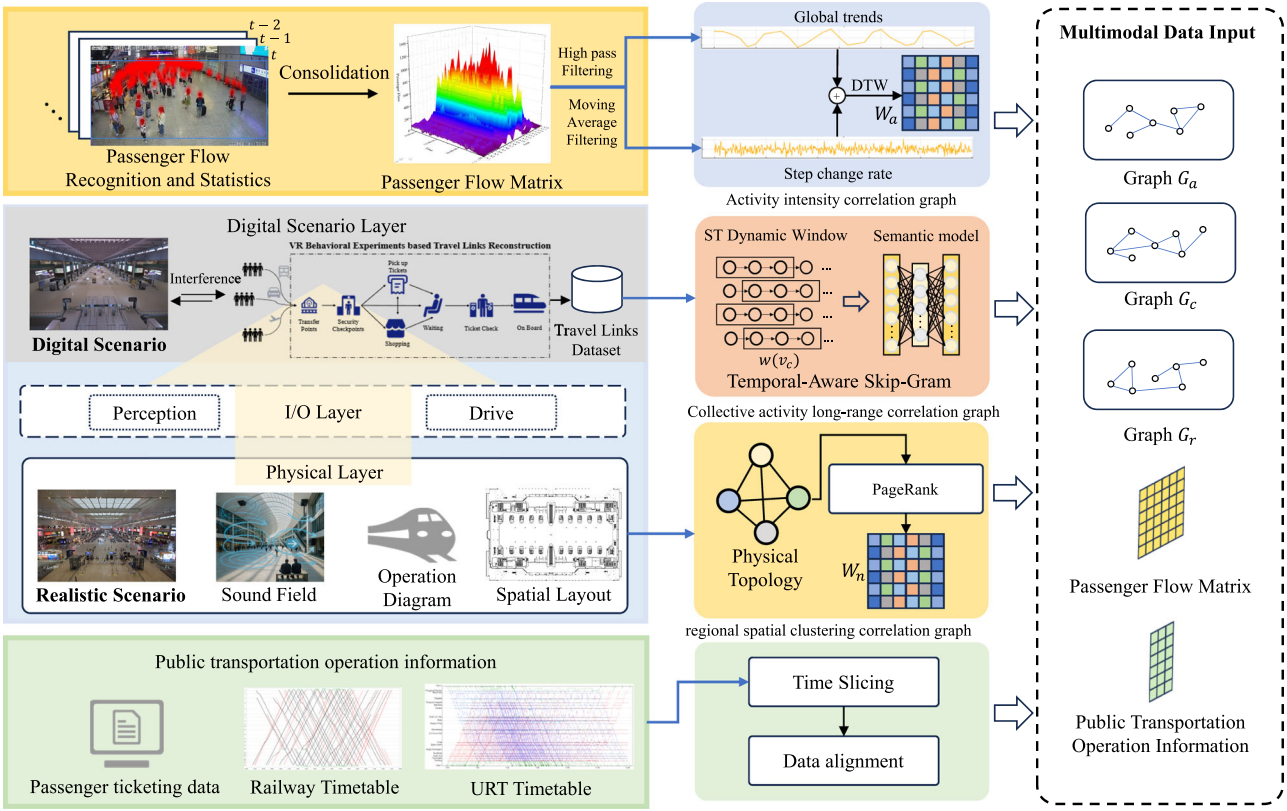
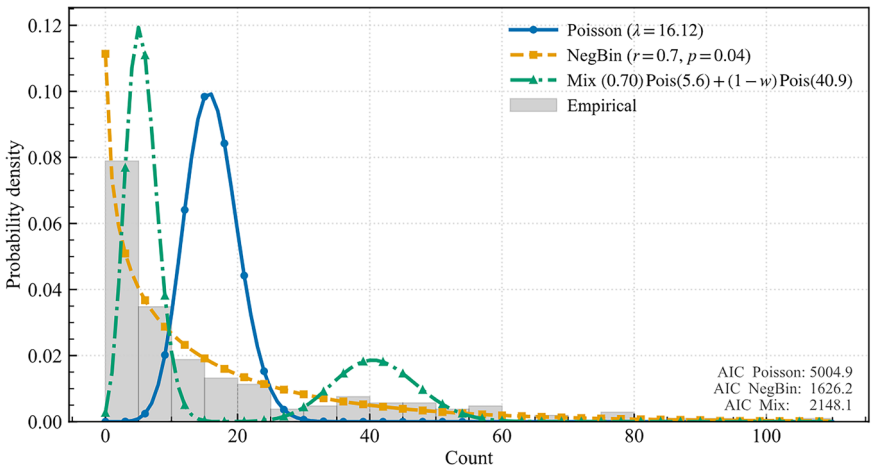


Fig. 3 | Preprocessing and input flow of spatially relevant multimodal data for an integrated transport hub.

Fig. 4 | Distribution fit of passenger area transition frequencies: Poisson, Negative Binomial, and Mixed Poisson.



region j follows i in the sequences. The average clustering coefficient and the average shortest path length are calculated as key indicators to characterize the network's "small-world" properties, aiming to explore the spatial aggregation of regions induced by collective passenger activities. A randomized reference network is constructed as a baseline for comparison, and based on the small-world model proposed by Watts and Strogatz[45], the overall small-world characteristics of the mobility network are quantitatively evaluated. The results show that the network's average clustering coefficient (for each node in the network, the ratio of the actual number of connections between neighboring nodes to the possible number of connections) is $C = 0.63$, and short average path length is $L = 2.58$ (which is a topological distance considering one unit for each edge connecting two nodes), the global small world coefficient (the clustering degree and average shortest path length of the network are compared to the corresponding values in the random graph of the same size. The larger the value, the stronger the small-world property of the network is $\sigma_g = 1.88$, yielding a high global small-world coefficient. This indicates that the network exhibits significant small-world properties, with strong local clustering among functional areas. Furthermore, for the undirected version of the activity network, node-level metrics including degree centrality, betweenness centrality and closeness centrality are calculated. The top 10 functional areas ranked by degree centrality are listed in Table 1. Among them are areas 8, 10, 11, 15, and 20, which include waiting areas and commercial zones near metro transfer corridors, exhibiting high centrality and playing crucial roles in both connectivity and mediating flows across the network. These areas are thus key to hub operations and passenger flow guidance.

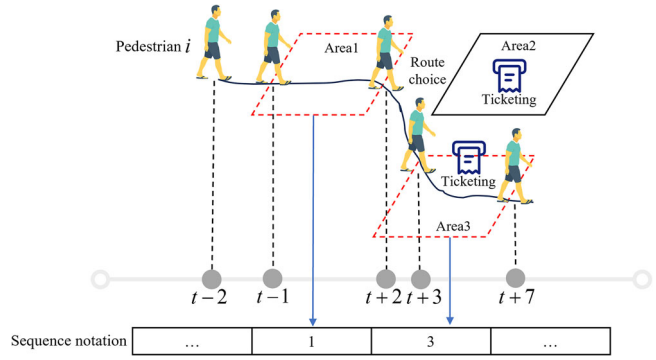


Fig. 5 | A reconstruction and coding method for travelers' mobility links based on behavioral experiments.

Table 1 | Feature metrics of the top 10 high-centrality regional nodes

Node	Degree centrality	Betweenness centrality	Closeness centrality	Eigenvector centrality
8	0.421	0.306	0.527	0.391
11	0.368	0.454	0.593	0.435
10	0.316	0.227	0.501	0.311
20	0.263	0.084	0.463	0.306
15	0.263	0.125	0.380	0.168
13	0.263	0.159	0.487	0.256
19	0.263	0.127	0.463	0.271
6	0.211	0.079	0.452	0.221
4	0.158	0.007	0.373	0.165
1	0.157	0.017	0.432	0.222

Degree centrality represents the proportion of a node's neighbors to all possible connections. Betweenness centrality measures how often a node serves as the "shortest-path bridge" between pairs of nodes in the network. Closeness centrality reflects the average length of the shortest paths from a node to every other node in the network. Eigenvector centrality is derived from the leading eigenvector of the network's adjacency matrix and evaluates centrality by considering both the number of a node's connections and the importance of the nodes it connects to.

The study further applies the Greedy Modularity Algorithm[46] for community detection within the activity network. The algorithm first treats each node as an individual community. Then repeatedly selects the pair of nodes whose merger maximally increases the network's modularity and immediately merges them into the same community. Each identified community subnetwork is then analyzed separately for small-world characteristics, enabling the identification of localized small-world structures. As shown in Fig. 6, the visualization of highly cohesive subnetworks reveals a clear pattern of "transport mode homogeneity" in community divisions. That is, nodes representing functional areas within the same community (indicated by the same color) tend to serve passengers arriving via the same mode of transportation. As observed in Table 2, areas within a single community, such as ticketing zones and commercial areas near metro transfer corridors—form subnetworks with high transition probabilities and significantly higher internal edge densities compared to inter-community connections. This structural pattern unveils the functional stratification of station interior space, where passengers associated with different transport modes self-organize into topologically distinct subsystems within the network.

This process, at the macroscopic level, reveals that the mobility network of passengers arriving via multi-modal transportation within the hub exhibits a significant spatial clustering effect. Simultaneously, at the microscopic level, it provides deeper insights into the spatial interaction characteristics among specific functional areas within the hub.

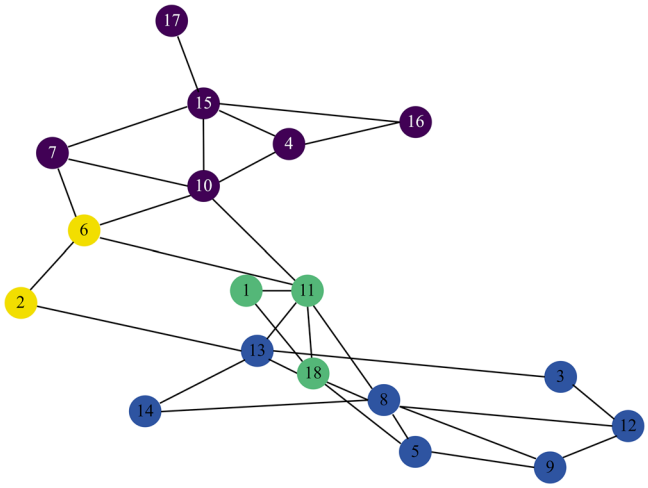


Fig. 6 | Visualization of the distribution of community sub-maps in spatial functional areas. The node numbering in the figure follows the same definition as previously described. Nodes within the same community are represented by the same color.

Causal association between operational information and flow fluctuations

To verify and quantify the strength of correlation between passenger flow fluctuations within different functional areas and the occurrence of public transportation operational events, this study employs Granger causality tests[47] to assess the statistical significance of the relationship between operational events and regional passenger flow dynamics. Furthermore, a multiple linear regression model is constructed to establish causal regression relationships. By recording the occurrence of operational events at time step t, the regression coefficients are used to capture the direct contribution of each event type to passenger volume changes in each area. These coefficients serve as indicators of the causal association strength between operational events and passenger flow variations.

$$X_{n,t}^{p} = \sum_{d=1}^{l} \beta_{1,d,n} \cdot E_{t-d} + \sum_{d=1}^{l} \beta_{2,d,n} \cdot U_{t-d} + \sum_{d=1}^{l} \gamma_{d,n} \cdot CV_{n,t-d} + \varepsilon_{n,t} \tag{3}$$

where, β, γ denote the regression coefficient measuring the direct contribution of an operational event (railway train departure E_t, metro train arrival U_t) to passenger volume in a given area, and $\varepsilon_{n,t}$ represent the random error term. If $\beta_{e,d,n} > 0$(when the coefficient is less than 0, it is uniformly treated as 0), the operational event is considered to have a Granger causal relationship with passenger flow in area n, indicating that the event can statistically explain future fluctuations in that area's flow. Additionally, $CV_{n,t-d} = \frac{\sigma_s(X^p_{n,t-d:t})}{\mu_s(X^p_{n,t-d:t})+\varepsilon}$ denotes the coefficient of variation of passenger volume in area n, calculated over a fixed lag window of length l. It is defined as the ratio of the standard deviation σ_s to the mean μ_s within the window and is used to capture local fluctuation characteristics of passenger flow in the area.

The results indicate that the average regression coefficient of railway train departure events on station-wide passenger flow fluctuations is 62.5% higher than that of metro train arrival events. As shown in Fig. 6a, b, railway departures exert a stronger and more sustained impact on passenger flow variations in different areas of the station compared to metro arrivals. The scatter plot in Fig. 6c further demonstrates that the causal strength remains positively correlated with network distance, suggesting that the influence of railway departures extends broadly across spatial regions. Furthermore, as the departure time approaches, the further away from the ticket gate, the greater the intensity of passenger flow fluctuations. In contrast, the

Table 2 | Subnetwork node composition and characteristics

Community	Size	σ_s	Path length	Aggregation factor	Nodes
1	6	1.24	1.29	0.64	4, 7, 10, 15, 16, 17
2	3	1.01	1.33	0.58	1, 11, 18
3	7	1.13	1.57	0.52	3, 5, 8, 9, 12, 13, 14

(Size represents the number of nodes contained in different communities). Path length is the average distance between nodes within a community. The calculation method for σ_s is the same as that for σ_g, but the research subjects are various communities. Only subnets with a small-world coefficient σ_s greater than 1 are retained.

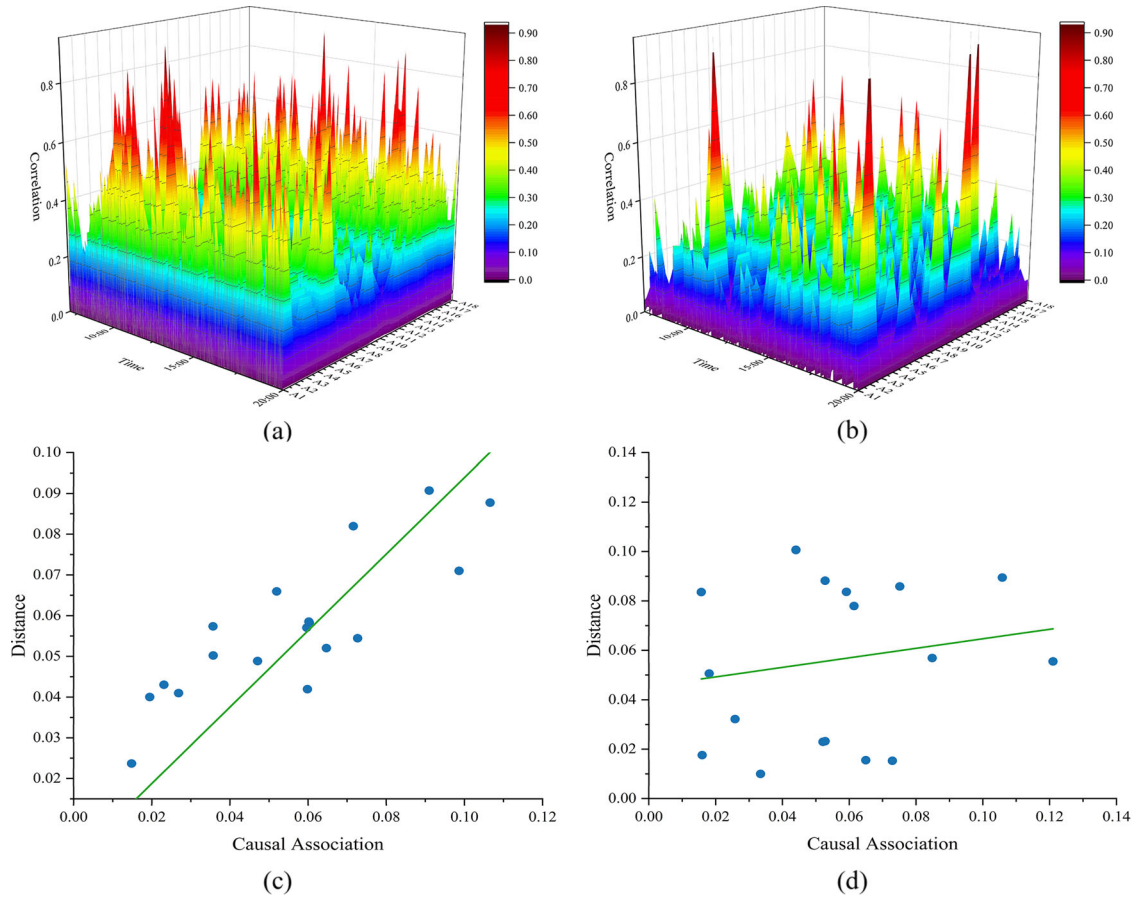


Fig. 7 | Visualization results of the association between two public transport operational events and regional passenger flow dynamics. a shows the railway departure event, and **b** shows the metro train arrival event. Scatterplot of correlation strength of passenger flow fluctuations versus network distance from the event area. **c**, **d** show the distance and strength of association of the region with railway train ticket gates and metro exit lanes, respectively. Distance and correlation data are normalized separately in order to keep the indicators on the same scale.

association map for metro arrivals (Fig. 6d) shows no significant trend, indicating that their impact is highly localized. We attribute this phenomenon to the differences in the spatiotemporal propagation patterns between the two types of operational events:

Railway train departures and metro arrivals differ markedly in the intensity, synchronization, and spatial propagation of passenger flow shocks within integrated transport hubs. Railway departures trigger high-intensity shocks strongly synchronized with collective passenger behaviors, manifesting as densely clustered crowds near waiting halls and ticket gates as departure times approach, with single-event inflows ranging from 10^2 to 10^3 passengers. These intense and synchronized flows generate chain-like fluctuations propagating widely along highly coupled pathways identified through small-world network analysis, often reaching distant, functionally distinct areas such as commercial and dining zones. In contrast, metro arrivals occur more frequently but involve smaller passenger volumes (typically between 10^1 to 10^2) and induce rapid passenger dispersal due to varied individual travel routes. Consequently, their impacts remain spatially localized within subnetworks such as concourses and transfer corridors, without significant long-range interactions, consistent with the small-world functional network's characteristic of forming spatially distinct communities with limited inter-community coupling.

To further verify the relationship between the strength of regional passenger flow fluctuations and the distance from the event occurrence area, defined as the metro exit or the ticket gate associated with the departing railway train at the current time, we conducted correlation tests (Pearson test) and linear fitting based on the scatter plots in Fig. 7c, d. The results reveal a significant positive correlation between passenger flow fluctuations and railway train departure events, with a Pearson correlation coefficient of 0.8395 and a high goodness-of-fit in the linear regression ($p = 0.0059 < 0.05, R^2 = 0.9405$). This indicates that the impact of railway departures on the intensity of regional passenger flow fluctuations increases with increasing spatial distance. In contrast, for metro train arrival events, the correlation between passenger flow fluctuations and distance is not statistically significant ($p = 0.33411 > 0.05$), and the linear regression

Table 3 | Results of fitting the relationship between the correlation strength of passenger flow fluctuation and the distance from the event area

Scenario	Railway		URT	
Formulas	$y_r = a_1 x_r + b_1$		$y_u = a_2 x_u + b_2$	
Parameters	a_1	b_1	a_1	b_2
Value	0.94	−0.017	0.54	0.02

x_r and x_u represent the distance from the target area to the ticket gate and the subway entrance. y_r and y_u represent the correlation strength between two operational events and passenger flow fluctuations.

exhibits a low goodness-of-fit ($R^2 = 0.4892$). This supports the conclusion that metro-induced flow variations are localized and do not exhibit a clear spatial decay pattern. Table 3 summarizes the linear regression results for both event types.

Optimization strategies for passenger flow forecasting modeling

Based on the previous analysis, we propose two modeling optimizations for passenger flow prediction: graph construction and transit event encoding. Previous studies have shown that considering multiple spatial dependencies in passenger flow prediction tasks can help improve prediction accuracy[10]. Consequently, we constructed three types of spatial correlation graphs, namely the collective activity long-range correlation graph G_c, the regional spatial clustering correlation graph G_r and the activity intensity correlation graph G_a. These graphs are used to model the long-range coupling characteristics between regions in the activity network (the passenger-activity network's average clustering coefficient is markedly higher than that of a random graph of the same size, while its average shortest-path length is comparable to the random baseline. As a result, functional zones that appear far apart can typically be reached through only two or three transfers, forming cross-area "long-range coupling" shortcuts), the strong correlation between passenger flows in adjacent regions within a community (community testing results reveal the spatial activity homogeneity of passengers traveling by the same mode of transport, indicating a high degree of correlation between passenger flow dynamics between adjacent nodes within the same community), and the similarity in passenger flow fluctuations between non-adjacent regions with similar service functions (within each community, the nodes generally include a mix of functional areas such as ticketing zones, commercial areas, and waiting halls, rather than clusters of homogeneous functions. This reflects a high degree of similarity in the spatiotemporal utilization of hub resources, leading to similar intensities of passenger activity over time).

For the construction of graph G_c, we propose a Temporal-Aware Skip-Gram semantic model, which extends the traditional Skip-Gram[48] by introducing a direction-sensitive, dynamic sampling window. This allows the model to capture unidirectional transitions and temporal constraints in passenger movements within the hub via learning semantic associations derived from passenger mobility chains. By inputting the set of passenger mobility chains $\mathcal{L}_T$, edges are constructed between nodes with high co-occurrence frequency, enabling the model to learn functional semantic associations and long-distance spatial dependencies beyond the physical topology. The size of the dynamic window is jointly determined by a base size (to prevent overfitting from an overly small window) and a function modulated by a distance-sensitivity coefficient.

$$w_t(v_{r,c}) = \left\lfloor w_{base} + \frac{\eta_t}{\min_{j>c} D_{c,j}} \right\rfloor \tag{4}$$

where $w_t(v_{r,c})$ indicates the sliding window size with $v_{r,c}$ as the target node. w_{base} denotes the base window size(set its value to the average length of the mobility chain), $\min_{j>c} D_{c,j}$ is the minimum Euclidean distance between the target node $v_{r,c}$ and a future node $v_{r,j}, j > c$ (appearing after central node $v_{r,c}$, $c < m$, r is the mobility chain number) in the mobility chain, and η_T is the distance sensitivity coefficient expressed as an exponential function. By treating the future node within the sliding window as the context node, the model samples temporal co-occurrence relationships between the target node and context nodes. Only future nodes are sampled to respect directionality. To achieve this goal, the d_v dimensional feature vector for each node's spatial characteristics is learned by minimizing the negative log-likelihood loss L_{SG} between the target node and context node. Finally, the cosine similarity between these feature vectors is used to quantify the spatial correlation between areas.

$$L_{SG} = -\sum_{c=1}^{r} \sum_{kc=1}^{w_t(v_{r,c})} \log p(v_{r,c+kc} | v_{r,c}) \tag{5}$$

In constructing G_r, the study employs PageRank centrality[49] and Dijkstra-based distance decay to reconstruct the influence range of physically neighboring functional areas in the mobility network. This approach can identify node pairs that are both physically close and have strong flow interaction, leading to the construction of the regional spatial clustering correlation graph G_r.

$$PR_i = \sum_{j=1}^{N} \frac{E_r(i,j) \times PR_j}{k_d(j)} \tag{6}$$

$$W_r(i,j) = \frac{PR_i \times PR_j}{D_{ij}} \tag{7}$$

where PR_i represents the PageRank score of node i. The computational process can be described as a damped random walk on a directed graph, where each step involves a uniform transition along outgoing edges with a certain probability, and a random jump to any node with the remaining probability. The steady-state visitation probability at each node ultimately constitutes its PageRank score. $k_d(j)$ denotes the degree of node j. D_{ij} is the shortest route length calculated by Dijkstra's algorithm between nodes i and j. Our objective is to assign greater weight to links connecting nodes with higher PageRank (those connected to more nodes) and their neighboring nodes (those with shorter distances), thereby emphasizing high-throughput neighboring pairings. The edge set (binary variable) E_r is constructed based on the actual physical topology. We consider area i and j to be topologically directly connected if there is no need to reach region j from area i through other areas.

Graph G_a is constructed based on two key temporal features of each node, stepwise rate of passenger flow change (the first-order difference of passenger flow between successive time steps) and global trend of variation. These are derived from the original passenger flow time series for each region. A moving average filter[50] is applied to extract trend components, and a high-pass filter is used to isolate high-frequency fluctuations. To eliminate sequence-similarity errors caused by phase shifts arising from sequence fluctuations, thereby accommodating passenger-flow sequences from different regions that share the same shape but are time-misaligned, the dynamic time warping (DTW) algorithm[51] is employed to compute the combined similarity (weighted sum of similarity between two trends across areas) between areas. The similarity is then used to assign edge weights between nodes in G_a.

To explicitly embed the impact of public transportation operational events into the passenger flow prediction model, this study designs an Event-Driven Spatio-Temporal Focus Module (EDSFM) including Train-Schedule Effect Encoding (TSEE) and Event Driven Sparse Attention. Based on previously estimated multivariate regression coefficients quantifying the influence of operational events on regional flow fluctuations, TSEE models the shock effects of both railway and urban transit (metro) train arrivals/departures as well as railway ticketing information. This module enables the model to incorporate the dynamic impact of operational schedules on passenger flow variations across functional areas.

$$I_{t,n} = \alpha_{1,n}(0.94 \cdot d_n) E_t \left(\frac{P_t}{\overline{P}}\right) e^{-\frac{t'-t}{\tau_1}} + \alpha_{2,n} U_t e^{-\frac{t-t'}{\tau_2}} \tag{8}$$

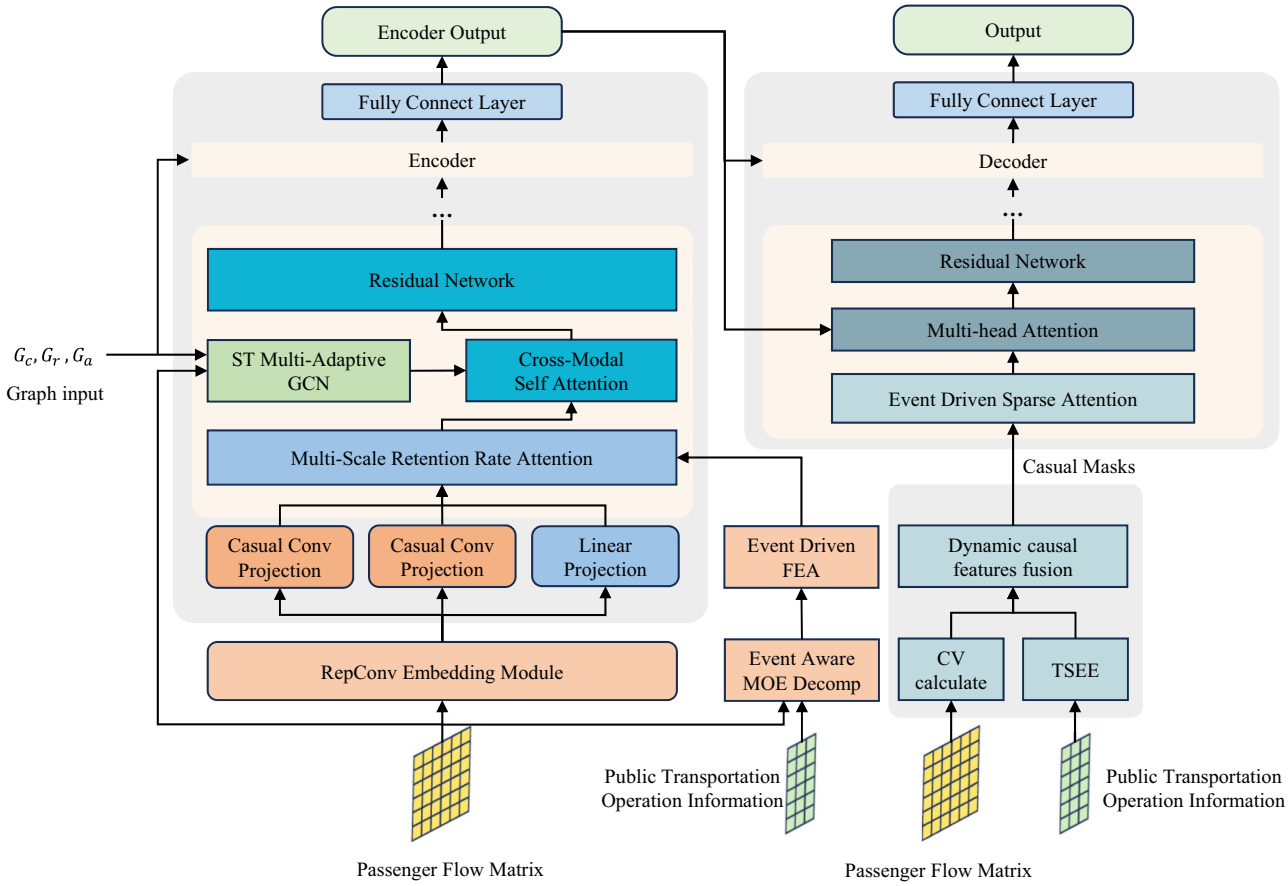


Fig. 8 | Overall architecture of GEME-Net. The overall framework adopts an encoder-decoder architecture, where both the encoder and the decoder consist of multiple identical sub-layers. The figure illustrates the structure of a single submodule within the encoder and decoder, while the ellipsis "…" indicates the sequential stacking of multiple identical modules.

where $\alpha_e, e \in \{1, 2\}$ denotes the influence weights of two types of events on different areas. E_t for the number of railway train departures and U_t for the number of urban rail arrivals (discrete variable). These weights are the mean values of the full-period regression coefficients obtained by Granger causal analysis for the corresponding events in the n^{th} area. $\bar{P}$ represents the average number of tickets sold per train on a given day, used for normalization. $\tau_e, e \in \{1, 2\}$ is the duration of the event's influence, these values are set as 30 minutes and 15 minutes according to empirical observations of transit operations. t' is the timestamp of the event, and d_n denotes the topological distance between area n and the event location. An exponential decay term is introduced in the formula to model the temporal attenuation of the event's impact.

Forecasting problem definition

We define the passenger flow prediction problem addressed in this study before introducing the prediction architecture. The task can be stated as follows: given the passenger flow matrix $X^p_{N,T_h} \in R^{T_h \times N}$ over T_h historical periods, the multi-type spatial association graphs $\mathcal{G}$ including G_c, G_r and G_a the public transportation operation data $PT^{T_h \times 3}$, the goal is to learn a mapping function $f(\cdot)$ to predict the passenger flow vector X^p_{N,T_h+tf} in each area for the future time period tf.

$$X^p_{n,ts+tf} = f(X^p_{n,ts}, G_c, G_r, G_a, PT) \tag{9}$$

Re-parameterized convolution embedding

The architecture of the proposed GEME-Net is illustrated in Fig. 8, adopting an encoder-decoder framework composed of multiple sub-layers. At the initial stage, a re-parameterized convolutional embedding layer[52] is applied to transform the raw temporal passenger flow data from multiple regions into dense vector representations. This step captures fine-grained temporal fluctuations and spatiotemporal dependencies, serving as the input features to the encoder. During the re-parameterized convolutional embedding process, the passenger flow distribution data $X^p_{N,T_h} \in \mathbb{R}^{T_h \times N}$ is mapped into a 3D tensor $X' \in \mathbb{R}^{1 \times T_h \times N}$ and the original convolutional kernel weights $W_{rc} \in \mathbb{R}^{C_{out} \times C_{in}/g_r \times k_{rc} \times k_{rc}}$ are initialized as an all-zero tensor. Here, C_{in} and C_{out} denote the number of input and output channels, g_r is the number of groups, and k_{rc} is the kernel size. These weights are learned via gradient updates during training. Next, kernel domain unfolding is applied: the convolutional kernels are flattened into 2D local convolutional feature maps $W^{flatten}_{rc} \in \mathbb{R}^{1 \times N_k \times k_{rc} \times k_{rc}}, N_k = C_{out} \cdot C_{in}$ for each spatial group. A kernel-space convolution $k_m \times k_m$ is then performed over these maps to learn intra-kernel structures in Eq. (10). To reduce noise and improve stability, depthwise separable convolution is used, enabling adaptive smoothing of the raw weights while reducing $O(N \cdot k_{rc})$ computational complexity. In the weight reassembly stage, the original kernel $W_{rc} \in \mathbb{R}^{C_{out} \times C_{in} \times k_{rc} \times k_{rc}}$ is fused with the learned structural kernel $\hat{W}_{rc}$ via element-wise addition to produce the final kernel $W^{final}_{rc} \in \mathbb{R}^{C_{out} \times (C_{in}/g_r) \times k_{rc} \times k_{rc}}$ in Eq. (11). This final kernel is then used in standard convolution to extract temporal dynamic features from the input, as shown in in Eq. (12).

$$\hat{W}_{rc} = \text{Conv2D}^{k_m \times k_m}(W^{flatten}_{rc}, \theta_{rc}) \tag{10}$$

$$W^{final}_{rc} = W_{rc} + \text{reshape}(\hat{W}_{rc}) \tag{11}$$

$$Z_{rc} = \text{Conv2D}^{k_{rc} \times k_{rc}}(X', W^{final}_{rc}) \tag{12}$$

where $\theta_{rc} \in \mathbb{R}^{(C_{out} \cdot C_{in}) \times 1 \times k_{rc} \times k_{rc}}$ denotes the kernel-space mapping weights. The original input $X^{p}_{N,T_h} \in \mathbb{R}^{T_h \times N \times 1}$ is ultimately embedded into a high-order spatial representation $Z_{rc} \in \mathbb{R}^{C_{out} \times T_h \times N}$. This efficient re-parameterized convolution process reduces parameter overhead while enhancing the model's ability to decouple multi-scale temporal features, thereby providing a robust foundational representation for the subsequent encoder's feature learning.

Encoder

The encoder consists of a projection layer, composed of a causal temporal convolution and a linear transformation, followed by stacked residual-connected encoder layers. Each encoder layer integrates four core components: the Multi-Scale Retention Rate Attention module explicitly models temporal dependencies across multiple time scales, focusing on local patterns while suppressing noise from distant time steps; the Event-Driven Frequency-Enhanced Module emphasizes learning from specific passenger flow frequency fluctuations induced by public transportation events, enhancing the model's responsiveness to short-term event-driven changes; the Spatial-Temporal Adaptive Multi-Graph Convolution Network (ST-AMGCN) dynamically constructs adaptive adjacency matrices to capture complex and evolving spatial dependencies, enabling flexible spatio-temporal representation; and the Cross-Modal Self-Attention (CMSA) module performs deep alignment and cross-attention between spatial and multi-scale temporal features, facilitating efficient integration of multimodal information. The outputs from all modules are fused through CMSA and passed through a fully connected layer to generate the encoder's final output, which is then fed into the decoder.

Event-driven frequency-enhanced module

The Event-Driven Frequency-Enhanced Module consists of two sequential submodules: Event-Aware MOE-Decomposition (EA-MOE) and Event-Driven Frequency-Enhanced Attention (ED-FEA). EA-MOE performs explicit temporal decomposition, stabilizing trend extraction and mitigating overfitting caused by abrupt event-induced fluctuations. ED-FEA then employs spectral gating masks to enhance periodic signals associated with transit events, enabling downstream attention mechanisms to focus directly on short-term, high-amplitude variations driven by those events. EA-MOE takes as input the historical passenger flow data $X^{p}_{N,T_h} \in R^{T_h \times N}$ and public transit event data PT on the same time scale. A multi-scale sliding average pooling is applied along the temporal axis to extract trend features $T^{(m)}_{E}$ at different scales, producing five expert-specific trend representations in Eq. (13). Simultaneously, PT is embedded via a fully connected layer into a low-dimensional event feature vector, which is then processed by another fully connected layer to generate gating weights for each expert in Eqs. (14) and (15). These weights are used to fuse expert trends into a unified, event-aware trend representation in Eq. (16). This is then concatenated with the original input X' and passed through a 1×1 convolution to integrate features. Finally, a residual connection between the fused output and the original input is applied to produce the final output Y_{EA} of the EA-MOE module, as shown in Eq. (17)

$$\{T^{(m)}_{E}\}_{m=1}^{M_e} = Avgpool^{k_e \times k_e}(X^{p}_{N,T_h}), k_e \in \mathrm{K}_e \tag{13}$$

$$z_e = GELU(PT \cdot W_1 + b_1),\ s_e = z_e \cdot W_2 + b_2 \tag{14}$$

$$\alpha_{MoE} = \mathrm{softmax}(s_e) \in [0,1],\ \sum_{m=1}^{M_e} \alpha_{MoE,m} = 1 \tag{15}$$

$$T_{MoE} = \sum_{m=1}^{M} \alpha_{b,m} T^{(m)}_{E} \tag{16}$$

$$Y_{EA} = \mathrm{Conv}^{1\times 1}[X' || T_{MoE}] + X' \tag{17}$$

where m represents the sliding average indices at five time scales, used to capture smoothed trends over 5 min, 10 min, 20 min, 40 min, and 60 min intervals, thereby identifying event pulse patterns of varying widths. K_e denotes the expert averaging pool size, and M_e is the number of experts, set to 5. W_1, W_2 are learnable parameter matrix. s_e is an intermediate variable z_e generated by a fully connected layer from the event embedding, representing the initial weight scores for each expert. Through this process, the module explicitly separates routine trends from event-driven passenger flow fluctuations in the temporal domain.

ED-FEA further refines event-related periodic features. The feature Y_{EA} is first flattened along temporal and spatial dimensions into a sequence $X_{flat} \in \mathbb{R}^{(T_h \times N) \times C_{emb}}$, and each sequence is transformed via Real Fast Fourier Transform[53] (RFFT) to obtain its frequency-domain representation X_f. In the frequency domain, the module computes the magnitude of each frequency component to capture underlying periodic patterns in passenger flow data. It then concatenates this frequency amplitude vector with the event embedding vector s_e and feeds the result into a fully connected network to generate a frequency-band gating mask M_{FEA}. This mask dynamically controls which frequency components are amplified or suppressed, allowing the model to selectively enhance event-relevant frequencies while attenuating irrelevant ones. The enhanced frequency representation X^{enh}_{f} is then transformed back into the time domain using inverse RFFT (IRFFT), yielding the event-enhanced spatiotemporal feature sequence $\tilde{X}_f \in \mathbb{R}^{(T_h \times N) \times C_{emb}}$.

$$\tilde{X}_f = \mathrm{F}_{IRFFT}(X^{enh}_{f}) = (\mathrm{Re}X_f) \odot M_{FEA} + j_e(\mathrm{Im}X_f) \tag{18}$$

where $\mathrm{Re}X_f$ and $\mathrm{Im}X_f$ denote the real and imaginary parts of the complex frequency spectrum tensor X_f, respectively, with j_e as the imaginary unit. Following frequency enhancement, the feature $\tilde{X}_f$ is projected into query Q_{FEA}, key K_{FEA}, and value V_{FEA} representations for the multi-head self-attention mechanism.

$$\begin{aligned} &\textbf{Multihead}\,\mathrm{A}_{FEA}(Q_{FEA}, K_{FEA}, V_{FEA}) = \textbf{Concat}(head_1, head_2, \ldots, head_{H_F}) \\ &where\, head_i = \mathrm{A}_{FEA}(Q^{(i)}_{FEA}, K^{(i)}_{FEA}, V^{(i)}_{FEA}) = \textbf{softmax}\left(\frac{\mathrm{Q}^{(i)}_{\mathrm{FEA}}\mathrm{K}^{(i)}_{\mathrm{FEA}}}{\sqrt{\mathrm{d_f}}}\right) V^{(i)}_{FEA} \end{aligned} \tag{19}$$

where H_F represents the number of attention heads, and $\sqrt{d_f} = \frac{C_{emb}}{H_F}$ denotes the scaled dot-product. Through this attention mechanism, the model explicitly captures fine-grained spatiotemporal dependencies among event-enhanced passenger flow features. The output is then passed through a linear projection layer, resulting in the final module output $Y_{FEA} \in \mathbb{R}^{T_h \times N \times C_{emb}}$.

Multi-scale retention rate attention

Given that passenger flow data is a typical time series, its short-term fluctuations, such as peak periods or unexpected events, often exhibit strong local temporal correlations. While traditional linear projections in attention mechanisms can capture global patterns, they are limited in modeling localized temporal structures. In Multi-Scale Retention Rate Attention (MSR-Atte), instead of applying standard linear projections for generating the query $Q_{MSR} \in \mathbb{R}^{T_h \times N \times C_p}$ and key $K_{MSR} \in \mathbb{R}^{T_h \times N \times C_p}$ the model uses causal convolutional projections along the channel dimension with output size C_p. Using a fixed kernel size k_{msr} causal convolution slides along the temporal axis to explicitly capture retentive correlations with the previous $k_{msr} - 1$ time steps, while preserving temporal causality. Meanwhile, the value tensor is still derived via linear projection, retaining the global temporal feature information. This design ensures that, after computing attention

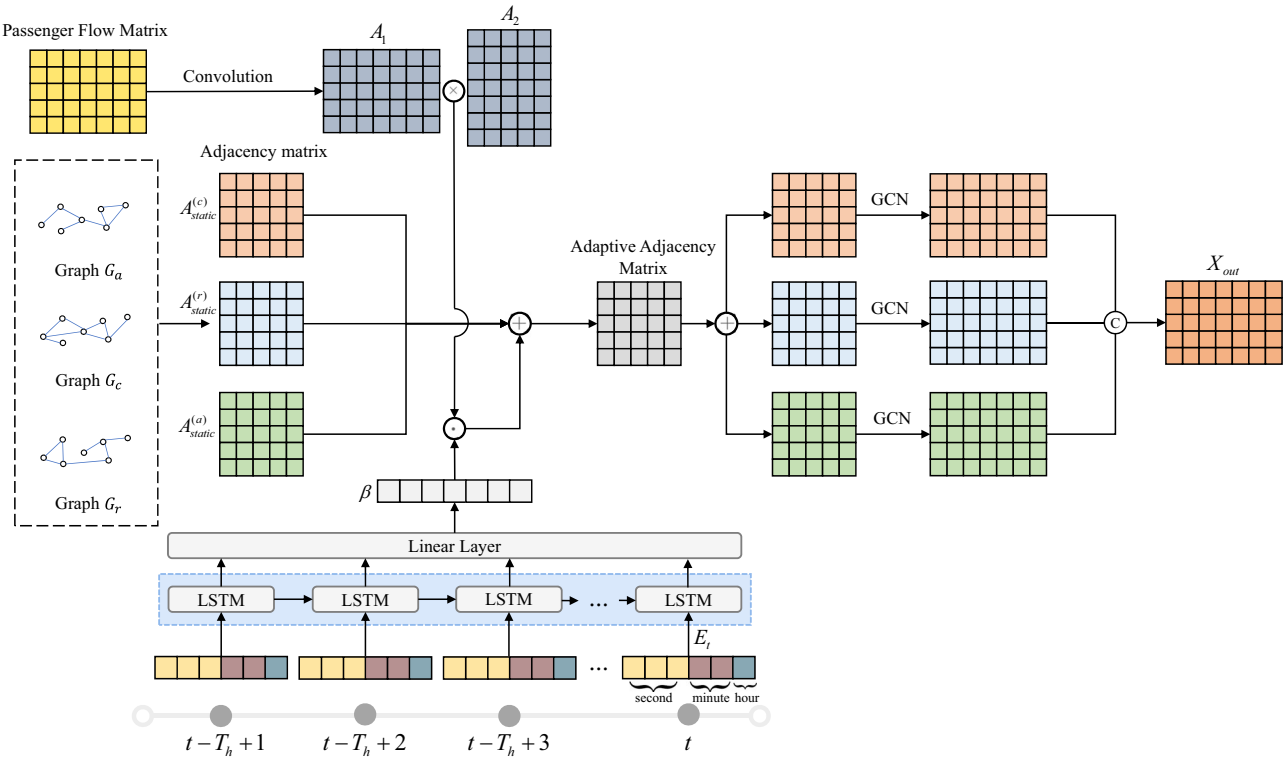


Fig. 9 | The overview of ST-AMGCN.

weights, the mechanism integrates both short-term dependencies and long-term trends effectively.

$$Q^{(t)}_{MSR} = \sum_{k_1=0}^{k_{msr}-1} W^{(k_1)}_{MSR,Q} \cdot Z^{(t-k_1)}_r \tag{20}$$

$$K^{(t)}_{MSR} = \sum_{k_1=0}^{k_{msr}-1} Q^{(k_1)}_{MSR,K} \cdot Q^{(t-k_1)}_r \tag{21}$$

where $W^{(k_1)}_{MSR,Q}$ and $W^{(k_1)}_{MSR,K}$ is the learnable parameter of temporal convolution kernel. $Z_r = \sigma(Conv^{(1\times1)}([Z_{rc}||Y_{FEA}]))$ is the passenger flow matrix higher-order embedding tensor and frequency domain enhanced temporal features obtained after splicing by a one-dimensional point-by-point convolution operation. is the time convolution kernel parameter.

Inspired by the retention mechanism in RetNet[54], we introduce a logarithmically spaced decay rate for each attention head h_m, controlling the retention strength of historical dependencies at different temporal scales. This design enables the model to simultaneously learn multi-scale temporal patterns ranging from minute-level to hour-level cycles within a single forward pass. Specifically, given the query Q_{MSR} and key K_{MSR} representations, MSR-Atte maps the feature dimension C_p into H_{MSR} multi-head subspaces. Each attention head is assigned a distinct temporal memory window, determined by a log-uniform partition of the time axis in the logarithmic domain, ensuring diversity in temporal focus. Finally, the outputs from all heads are concatenated to form a composite temporal representation, capturing multi-period temporal dependencies with fine-to-coarse semantic granularity. This process can be formulated as follows:

$$\gamma_{MSR,h_m} = 1 - \exp[\ln\left(\frac{1}{32}\right) + u_{h_m}\left(\ln(\frac{1}{512}) - \ln\left(\frac{1}{32}\right)\right)] \tag{22}$$

$$u_{h_m} = \frac{h_m - 1}{H_{MSR} - 1} \tag{23}$$

$$D^{h_m}_{MSR}(n,m) = \begin{cases} \gamma^{n-m}_{MSR,h_m}, & n \geq m, \\ 0, & otherwise, \end{cases} \tag{24}$$

$$MH_{MSR} = Concat^{H_{MSR}}_{h_m=1}[((\tilde{Q}_{h_m} \cdot \tilde{K}^{\top}_{h_m}) \odot D^{h_m}_{MSR}) \cdot V^{h_m}_{MSR}] \tag{25}$$

where u_{h_m} is a normalization coefficient used to position the k^{th} attention head within a linear interval, facilitating subsequent interpolation in the logarithmic domain to generate decay rates, thereby enabling the effective memory window of each head to expand geometrically. (n, m) is the integer index along the sequence dimension. When $n \geq m$, the query time is either after or exactly at the key time, ensuring attention values are retained and subjected to exponential decay to enforce strict temporal causality. $\tilde{Q}$ and $\tilde{K}$ represent the tensor forms of the original query and key, which help achieve smoother extrapolation over long sequences. These are element-wise multiplied (Hadamard product) with the decay matrix to compute decayed attention scores, which are then matrix-multiplied with the values V_{MSR} to obtain the causal influence matrix across time steps within the respective scale window for each head. Finally, the outputs from all heads are concatenated to form the composite temporal semantic tensor $MH_{MSR} \in \mathbb{R}^{C_p \times T_h \times N}$, capturing multi-scale periodic temporal dependencies.

Spatial-temporal adaptive multi-graph convolution network

The study proposes a spatial-temporal adaptive multi-graph convolutional network (ST-AMGCN) to capture the complex, dynamic spatiotemporal relationships among passenger flows across multiple regions within a transport facility. The architecture is illustrated in Fig. 9. Unlike conventional adaptive graph convolution methods, ST-AMGCN integrates a temporal dynamic weighting module, where an LSTM encodes both the temporal embedding E_t and recent historical passenger flow data to generate time-dependent weights β_t. This allows the model to adaptively adjust

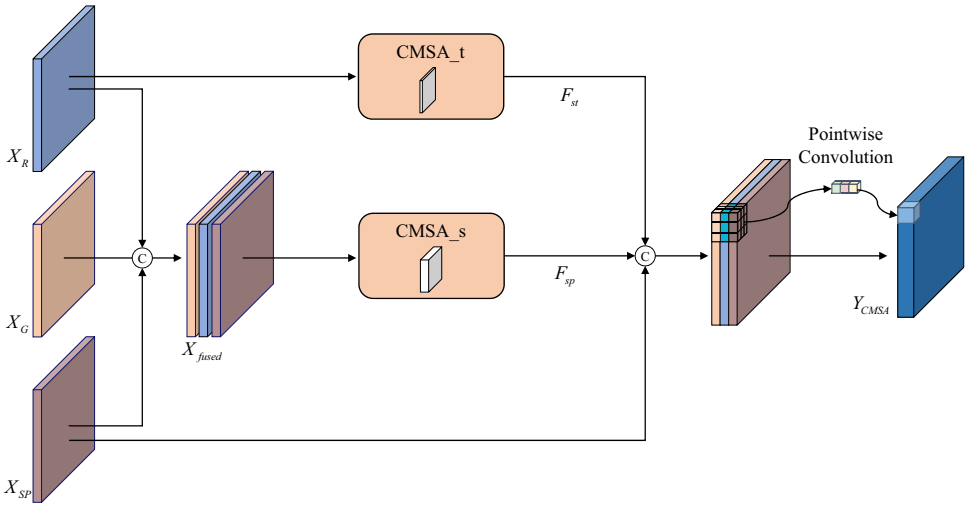


Fig. 10 | The overview of CMSA. The three feature tensors are concatenated along the channel dimension to form an integrated cross-modal feature representation X_{fused}, after which the output features F_{sp} and F_{st} from the two submodules are fused through concatenation and convolution operations.

the spatial dependencies between nodes as time progresses. The input passenger flow matrix X' is first projected into an intermediate feature space via 2D convolution, producing A_1, A_2. These projected features are used to compute a learned dynamic adjacency matrix A_{adapt}. This matrix is modulated by the dynamic weight and a set of static adjacency matrices $A^{(i)}_{static} \in M_{subsets} = \{A^{(c)}_{static}, A^{(r)}_{static}, A^{(a)}_{static}\}$ derived from predefined spatial correlation graphs, resulting in the final dynamic spatiotemporal adjacency matrix A_{st}. This matrix is then element-wise combined with the node features (passenger flows) and passed through multi-layer graph convolutional operations. The outputs from different subgraphs are concatenated to produce X_{out}, capturing complex dynamic interactions and spatial dependencies among regions.

$$h_t, c_t = LSTM(E_t) \tag{26}$$

$$\beta_t = \sigma([h_{te} \| X'_{t-T_h-1,t}] \cdot W_{te} + b_{te}) \tag{27}$$

$$A_1 = \mathbf{Conv2D}(X'), A_2 = \mathbf{Conv2D}(X') \tag{28}$$

$$A_{adapt} = \tanh\left(\frac{A_1 \cdot (A_2)^\top}{T_h}\right) \tag{29}$$

$$A^{(i)}_{st,t} = A^{(i)}_{static} + \beta_t \cdot A_{adapt} \tag{30}$$

$$S^l_{(i)} = GCN(S^{l-1}) = \sigma((A^{(i)}_{static} \oplus A^{(i)}_{st}) S^{l-1}_{(i)} W^{l-1}_g) \tag{31}$$

$$X_{out} = Linear([S^l_{(c)} \| S^l_{(n)} \| S^l_{(a)}]) \in \mathbb{R}^{C_{g,out} \times T_h \times N} \tag{32}$$

In the equation, the temporal embedding feature E_t is obtained by embedding one-hot vectors of the original temporal attributes—specifically second-level (the finest statistical granularity), minute-level, and hour-level features via a linear layer. And h_t, c_t denote the encoding of sequence information at the current time step and the long-term memory unit, respectively. W_{te}, b_{te} are the learnable weight matrix and bias of the fully connected layer. (i) represents the subgraph index, and S^{l-1} denotes the input feature matrix for the $(l-1)^{\text{th}}$ input. W^{l-1}_g is the learnable parameter matrix of that layer.

Cross-modal self-attention

The spatio-temporal features obtained from preceding modules are fused in this component to fully exploit the complementary spatial and temporal information embedded in passenger flow data. As shown in Fig. 10, the architecture of this module is designed to efficiently integrate multi-modal information. Compared to existing cross-modal attention networks[55], CMSA performs fusion on a shared spatial grid, integrating deep features from space X_G, time X_R, and position within a unified attention framework. This design allows parallel attention modeling strictly on the spatial plane, significantly reducing computational overhead. The CMSA_s submodule models cross-location spatial attention, explicitly capturing dynamic inter-regional dependencies. In parallel, CMSA_t focuses on temporal features extracted by MSR-Atte, learning interactions across different time scales. Finally, the module incorporates spatial position embeddings X_{SP}, and concatenates the outputs of CMSA_s and CMSA_t. A point-wise 1×1 convolution is applied to fuse spatial, temporal, and positional information, yielding the final cross-modal fused representation Y_{CMSA}.

$$\begin{aligned} F_{sp} &= CMSA_s([X_{fused}]) = CMSA_s([X_G \| X_R \| X_{SP}]) \\ F_{sp} &= \mathbf{Conv3D}\left(\mathbf{softmax}\left(\frac{Q_s \cdot K_s}{\sqrt{C_{CMSA}}}\right) V_s\right) \end{aligned} \tag{33}$$

Decoder

The decoder module is composed of multiple identical decoder layers stacked via residual connections, and concludes with two fully connected layers to generate short-term passenger flow forecasts for all regions. The decoder's core function is to build upon the spatiotemporal features extracted by the encoder, further incorporating event-driven signals and spatiotemporal fluctuations of regional flows to establish causal relationships. These causal dependencies are used to generate attention masks, enabling a sparse attention mechanism that enhances the precision of flow prediction for target regions. A multi-head cross-attention mechanism then integrates the encoder's spatiotemporal representations into the decoder. Unlike the Event-Driven Frequency-Enhanced Module, which embeds transit operations into the underlying frequency structure of the sequence via expert gating and band selection, the Event-Driven Spatial-Temporal Focusing Mechanism directly imposes event impact in the spatial domain. This mechanism modulates attention weights in the decoder based on causal influence, enabling local incremental adjustments to better capture region-specific flow variations.

In the decoder module, the event impact tensor I_t calculated from formula (8), which quantifies the influence of events at each time step, is first passed through a linear projection to generate a normalized attention mask matrix $\tilde{E}_{h_e}$. Subsequently, a sparse attention mechanism is constructed,

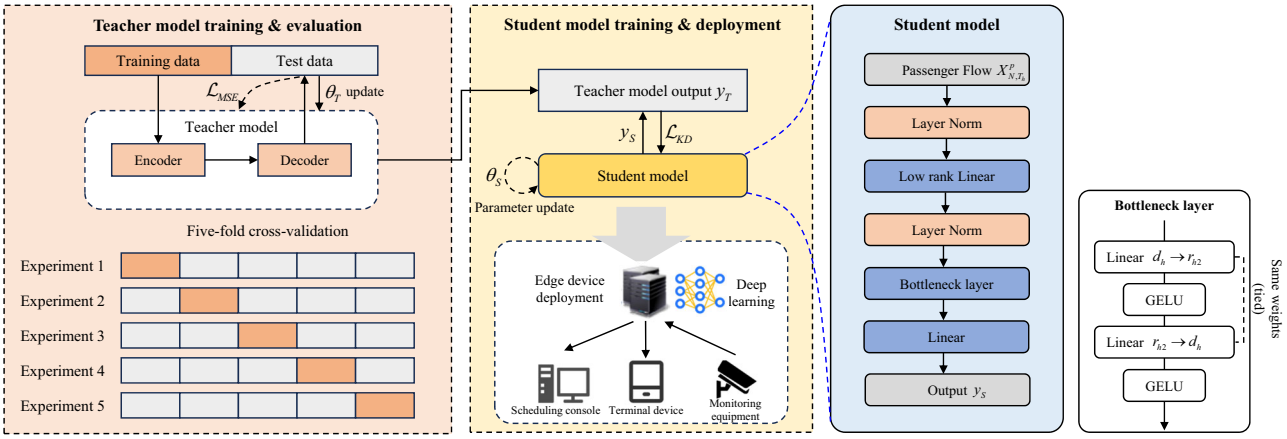


Fig. 11 | Schematic diagram of the GEME-Net network teacher model training, knowledge distillation, student model training and edge device deployment processes.

allowing the event-induced passenger flow response patterns to be explicitly injected into the spatiotemporal feature aggregation process.

$$\tilde{E}_{h_e,t} = \sigma(W_e \cdot I_t + b_e) \in (0,1) \tag{34}$$

$$\textbf{Multihead}\ \mathrm{A}_{ED}(Q_{ED}, K_{ED}, V_{ED}) = \textbf{Concat}(head_1, head_2, \ldots, head_{H_e})$$
$$where\ head_{h_e} = \mathrm{A}_{ED}(Q^{ED}, K^{ED}, V^{ED}) = \textbf{softmax}\left(\frac{\mathrm{Q}^{\mathrm{ED}}_{\mathrm{h_e},i}\mathrm{K}^{\mathrm{EDT}}_{\mathrm{h_e},i}}{\sqrt{\mathrm{d_e}}} \cdot \tilde{\mathrm{E}}_{\mathrm{h_e}} + (1-\tilde{\mathrm{E}}_{\mathrm{h_e}})(-\infty)\right) V^{ED}_{h_e,j} \tag{35}$$

where W_e, b_e denote the learnable weight matrix and bias vector, respectively. During the attention computation, when an element in the event-based mask $\tilde{E}_{h_e}$ approaches 0, the corresponding row i in the attention score matrix is assigned as $-\infty$, effectively forcing its softmax output to approach zero, thereby preventing attention allocation. Additionally, for each token, a causal mask $M_{gc}(i,e), i = t \cdot N + n$ is used to determine whether the token's associated region has any significant causal events. If not, the entire row in the attention matrix is masked, ensuring that no attention is distributed to non-event-related areas.

Next, the Multi-Head Attention mechanism adopts a standard multi-head self-attention structure, where the queries are derived from the output of the preceding Event-Driven Sparse Attention module, while the keys and values come from the encoder's spatiotemporal representations. The high-dimensional features output from the stacked decoder layers are then flattened and passed through two fully connected layers, yielding the predicted passenger flow $X^p_{n,ts+tf}$ across N areas for the next tf time steps. During training, the model optimizes the Mean Squared Error (MSE) loss, measuring the accuracy of short-term passenger flow distribution forecasts within the facility space.

$$\min \mathrm{L}_{MSE} = \frac{1}{N \cdot tf} \sum_{n=1}^{N} \sum_{t=1}^{tf} (\hat{y}_{n,t} - y_{n,t})^2 \tag{36}$$

Model knowledge distillation

Although GEME-Net, as the teacher model, achieves accurate predictions by integrating multiple modules to capture spatiotemporal passenger flow features and public transport operation events, it also introduces a large parameter scale and substantial computational overhead, which hinders low-latency deployment on resource-constrained edge devices. To address this, we adopt knowledge distillation by using the pre-trained GEME-Net as the teacher model to construct and train a student model with reduced complexity and latency, thereby lowering parameter size and computational cost while retaining as much predictive accuracy as possible.

The student model is implemented as a lightweight MLP, in which low-rank linear layers are combined with a shared bottleneck layer to capture the associations between key priors—such as event-driven dynamics and multi-scale temporal patterns—and passenger flow fluctuations distilled from the teacher model. Its training process and network architecture are illustrated in Fig. 11. Specifically, the historical passenger flow tensor $X^p_{N,T_h} \in \mathbb{R}^{T_h \times N \times 1}$ is first flattened and normalized through a LayerNorm layer, and then embedded into a d_h dimensional feature tensor $W_S \in \mathbb{R}^{(T_h \times N) \times d_h}$ via two low-rank fully connected layers (first projected into a lower-dimensional intermediate space of size r_{h1} and then mapped to a d_h dimensional feature). The embedded features are subsequently passed through a shared-weight bottleneck layer, where the same set of low-rank linear weights is invoked twice. This design approximates the passenger flow features W_S as a low-rank factorization $U_S V_S, U_S \in \mathbb{R}^{(T_h \times N) \times r_{h2}}, V_S \in \mathbb{R}^{r_{h2} \times d_h}$, reducing the parameter size from $T_h N d_h$ to $T_h N r_{h2} + r_{h2} d_h$, while the shared bottleneck further lowers computational complexity without sacrificing nonlinear expressiveness. The final features are projected through a linear layer to generate the student model's multi-region passenger flow predictions y_S. During training, the student model is optimized against the teacher's outputs $y_T = X^p_{n,ts+tf}$ using a composite regression distillation loss (smooth L1 loss and cosine similarity loss) so as to align both the magnitude of passenger flows and the proportional distribution across regions with the teacher model, thereby updating the student parameters.

$$\min \mathrm{L}_{KD} = \alpha_d \cdot SmoothL1(y_T, y_S) + (1-\alpha_d) \cdot cosine(y_T, y_S) \tag{37}$$

where α_d denotes the weighting coefficient of the two loss terms in the overall distillation loss, which is set to 0.5 in this study to balance their contributions to the final results.

Results

Model configurations and evaluation metrics

During the experiments, the dataset was split into training, validation, and test sets in a 7:2:1 ratio. The batch size was set to 16. All models were implemented using PyTorch 1.13.1, Tensorflow 2.6.0 and Python 3.8.0, and trained/evaluated on a GeForce RTX 3060 Ti GPU. The Adam optimizer was used, with a learning rate of 0.0001 for the teacher model and 0.001 for the student model. A regularization coefficient of 1×10^{-4} was applied to reduce overfitting. Model generalization was evaluated using 5-fold cross-validation, with 200 epochs per fold. The mean squared error (MSE) loss function was used for training. In terms of evaluation metrics, this study adopts a set of widely recognized indicators for assessing passenger flow prediction performance, including Root Mean Squared Error (RMSE),

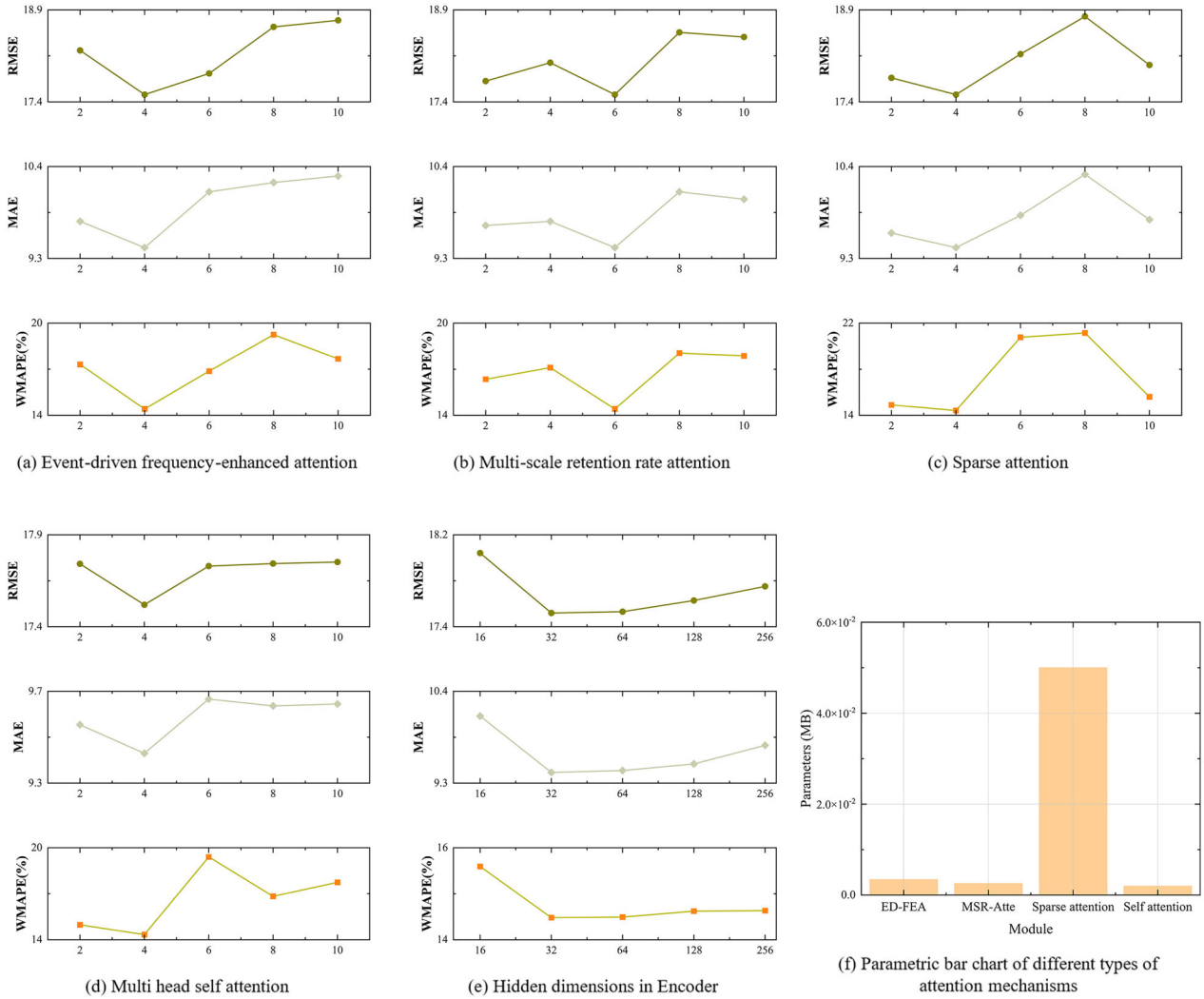


Fig. 12 | The impact of varying head counts in four multi-head attention mechanisms and hidden dimensions on the predictive performance of GEME-Net. a The effect of event-driven frequency-enhanced attention head counts. **b** The effect of multi-scale retention rate attention head counts. **c** The effect of sparse attention head counts. **d** The effect of multi head self attention head counts. **e** The effect of hidden dimensions in encoder. **f** Parameter scales of different types of attention mechanisms.

Mean Absolute Error (MAE) and Weighted Mean Absolute Percentage Error (WMAPE). These metrics are used to comprehensively evaluate the model's prediction accuracy and stability across different scenarios.

$$RMSE = \sqrt{\frac{1}{N \cdot tf}\sum_{n=1}^{N}\sum_{t=1}^{tf}(\hat{y}_{n,t} - y_{n,t})^2} \tag{38}$$

$$MAE = \frac{1}{N \cdot tf}\sum_{n=1}^{N}\sum_{t=1}^{tf}|\hat{y}_{n,t} - y_{n,t}| \tag{39}$$

$$WMAPE = \frac{\sum_{n=1}^{N}\sum_{t=1}^{tf}|\hat{y}_{n,t} - y_{n,t}|}{\sum_{n=1}^{N}\sum_{t=1}^{tf}|y_{n,t}|} \tag{40}$$

where $\hat{y}_{n,t}$ and $y_{n,t}$ denote the predicted and actual passenger flow values for area n at time step t.

Since our model incorporates improved multi-head attention mechanisms to learn and fuse multimodal data within the hub, we evaluated the impact of the number of attention heads across different attention types on predictive performance, as shown in Fig. 12. The results indicate that the number of heads is a critical factor influencing both accuracy and complexity, with the largest effect observed for the sparse attention used to integrate public transport operation times. Event-driven frequency-enhanced attention in Fig. 12a. shows the next most pronounced effect, suggesting that modeling exogenous factors has a greater impact on capturing passenger flow fluctuations than further refining temporal autocorrelation learning. Moreover, the stronger variability in prediction metrics caused by sparse attention compared with event-driven frequency-enhanced attention demonstrates that hard-masked exogenous variable modeling is more effective than soft modulation via frequency gating. However, the drawback of hard coding lies in increased computational complexity, as illustrated in Fig. 12f. We also examined the effect of embedding dimension on prediction performance, where Fig. 12e shows that the best accuracy is achieved at a dimension of 32, which is therefore adopted throughout the experiments. The list of model hyperparameters is shown in Table 4.

Comparison and analysis of prediction results

We next provide a detailed comparison of the prediction error metrics of the proposed model against those of established baseline models. In our experiments, we selected seven representative baseline models, including:

Table 4 | List of hyperparameter values for passenger flow prediction networks

	Parameters	Module name	Value
GEME-Net (Teacher model)	C_{in}	Re-parameterized convolutional embedding layer	18
	C_{out}	Re-parameterized convolutional embedding layer	12
	C_p	Encoder module	32
	k_m	Re-parameterized convolutional embedding layer	3
	k_{rc}	Re-parameterized convolutional embedding layer	5
	M_e	Event-driven frequency-enhanced module	5
	Head number of event-driven frequency-enhanced attention	Event-driven frequency-enhanced module	4
	k_{msr}	Multi-scale retention rate attention	32
	H_{MSR}	Multi-scale retention rate attention	6
	I	Spatial-temporal adaptive multi-graph convolutional network	3
	H_e	Sparse attention mechanism	4
	H	Multi-head self-attention	4
GEME-Net (Student model)	d_h	-	128
	r_{h1}	-	16
	r_{h2}	-	8
	Dropout ratio	-	0.1
	α_d	-	0.5

Table 5 | Comparison of prediction errors between GEME-Net and baseline models for passenger flow data

Time granularity	10 s			20 s			30 s			Sig.
Model	RMSE	MAE	WMAPE (%)	RMSE	MAE	WMAPE (%)	RMSE	MAE	WMAPE (%)	
ARIMA	31.36	16.33	28.47	62.50	32.28	29.38	107.93	55.82	30.72	*
LSTM	20.05	10.38	17.89	33.42	17.60	15.45	46.22	25.40	16.29	*
ConvLSTM	19.25	9.96	17.15	33.31	17.59	14.31	49.47	26.18	16.35	*
2D CNN	20.32	12.60	18.24	33.43	17.66	19.05	46.42	24.79	19.98	*
ST-ResNet	19.10	11.92	17.10	35.69	19.04	17.04	46.81	24.60	18.53	*
Transformer	18.76	11.05	16.83	31.83	18.03	14.64	67.88	39.10	19.89	*
ST-GCN	18.54	10.43	16.15	32.26	17.43	13.27	46.52	24.64	15.04	*
Informer	18.14	10.97	15.52	31.83	18.03	13.73	45.70	22.20	13.05	*
GEME-Net (Teacher model)	17.52	9.43	14.41	30.24	16.40	13.48	41.99	22.04	13.21	-
GEME-Net (Student model)	18.96	11.03	17.24	31.72	17.85	14.71	46.89	26.79	17.93	-

Models marked with * indicate statistical significance compared to the Teacher model of GEME-Net (T test with p value < 0.01).

- ARIMA: a statistical model widely used for time-series forecasting. In our experiments, the lag order, differencing order, and moving average order were set to 2, 1, and 1 respectively.
- Long Short-Term Memory (LSTM): This network consists of two hidden layers and one fully connected layer. Each hidden layer contains 128 neurons.
- Convolutional LSTM (ConvLSTM): The prediction model consists of three hidden layers and one fully connected layer, with each layer containing 8 filters and performing convolution operations using 3×3 convolution kernels.
- Convolutional Neural Network (2D CNN): Comprised of two convolutional layers and one fully connected layer, with 32 and 64 filters respectively and a kernel size of 3×3.
- ST-ResNet[56]: A model that uses residual convolution to capture spatiotemporal features of passenger flow. In our experiments, we used only three residual convolution branches and omitted the original module for extracting weather features. Other network parameters remained consistent with the original paper.
- Transformer[40]: The traditional Transformer model consists of three encoder and decoder layers. Each layer uses 8-head multi-head attention, and the feature embedding dimension is set to 64.
- Spatio-Temporal Graph Convolutional Networks (ST-GCN)[57]: A network that models spatiotemporal features of passenger flow using graph convolution and temporal gated causal convolution layers. Its structure comprises nine spatiotemporal convolutional units, using convolution kernels of size 3×3.
- Informer[41]: An improved Transformer network with the same attention mechanism parameters and layer configuration as the standard Transformer.

To ensure a fair and reproducible comparison, all models share the same input, train/validation/test split and pre-processing pipeline (Z-score scaling fitted on the training dataset). Exogenous features are provided to architectures that can accept them (Graph input: ConvLSTM, ST-ResNet and ST-GCN, public transport operational information: Transformer and Informer. The self-attention in the baseline model was redesigned as a sparse attention module to ensure the incorporation of exogenous variables. While for univariate statistical baselines (ARIMA and LSTM), exogenous inputs are disabled by design, and we report multi-step recursive forecasts with an identical train process.

Table 5 presents the prediction accuracy comparison across all models. The results demonstrate that GEME-Net significantly outperforms all baseline models in terms of MAE, RMSE, and MAPE. Specifically,

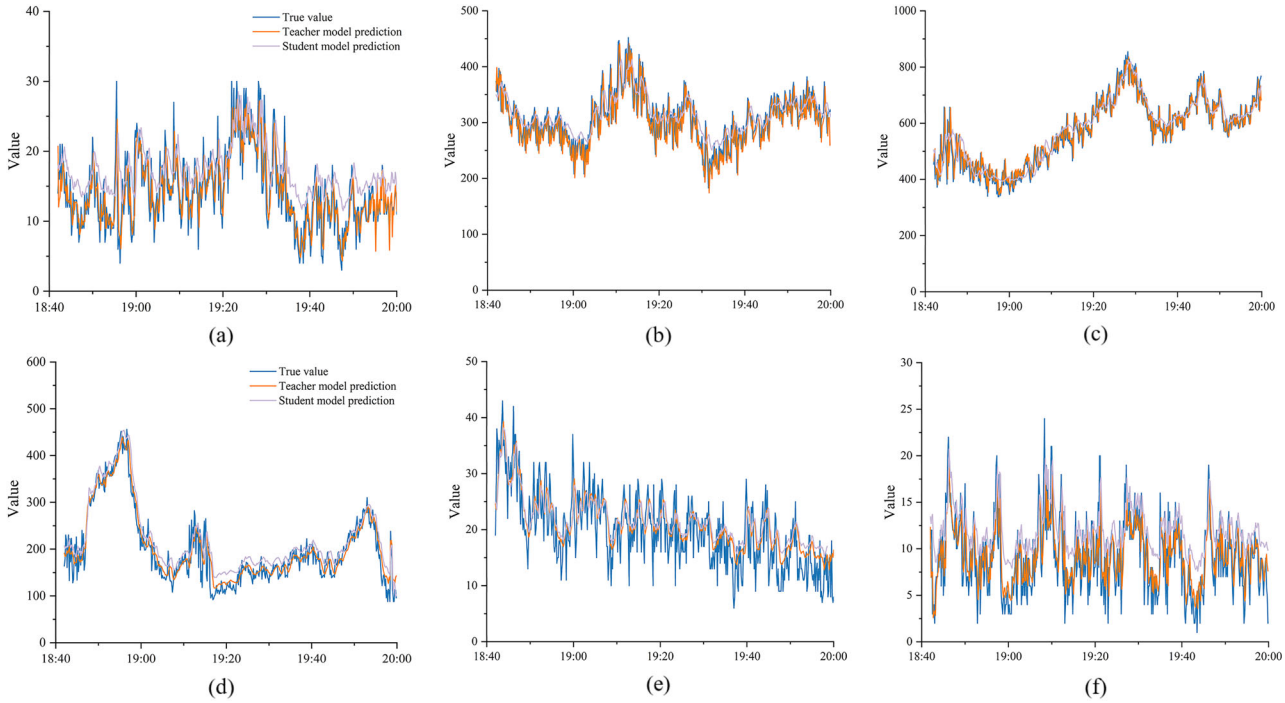


Fig. 13 | Visualization of the model's prediction performance for passenger flows in different functional areas. a Passenger flow in ticketing area (area 5). **b** Passenger flow in waiting area (area 10). **c** Passenger flow in check-in area (area 11). **d** Passenger flow at the airport arrival entrance (area 12). **e** Passenger flow in the eastern commercial area (area 14). **f** Passenger flow in the western commercial area (area 17).

compared to the traditional time-series model LSTM, GEME-Net achieves improvements of over 12.6% (MAE), 9.1% (RMSE), and 19.45% (MAPE). When compared to the Transformer-based state-of-the-art model Informer, GEME-Net still yields performance gains of over 6.6%, 14.6%, and 14.4% improvement in the same metrics. When the time granularity of the passenger flow data increases from 10 s to 30 s, the performance gap between GEME-Net and the baseline models widens. Furthermore, the comparison between the teacher and student models indicates that, although the student model experiences a minor decrease in prediction accuracy, knowledge distillation enables a substantial compression of the model size, reducing the parameter storage from 6.86 MB in the teacher model to only 0.16 MB in the student model. This substantially lowers computational overhead. Moreover, the student model built on multi-layer perceptron and multi-layer convolutional structures achieves accuracy comparable to Transformer-based models (with the same hyperparameters, such as the number of attention heads and embedding dimensions as the teacher model, the parameter size of Transformer-based models is 3.51 MB). These findings confirm that GEME-Net, through knowledge distillation, effectively balances predictive performance and computational efficiency, making it well-suited for deployment in real-world resource-constrained environments.

To more intuitively demonstrate the predictive performance of our proposed model, we extract the actual and predicted passenger flow values for different functional areas, as illustrated in Fig. 13. The results show that GEME-Net achieves a high degree of alignment between predicted and actual values, with particularly strong performance during peak flow periods. Visual inspection of Fig. 8 reveals that GEME-Net can accurately capture peak flows across both high- and low-traffic areas, maintaining low prediction error rates at peak times. Notably, in high-demand regions, GEME-Net exhibits superior predictive accuracy, with significantly reduced errors, further confirming its effectiveness in handling large-scale passenger flow variations.

The comparative analysis between the teacher and student models indicates that the teacher model consistently achieves lower prediction errors and more accurately captures subtle variations and peak flows. This reflects its stronger capacity to learn complex patterns and data dependencies. In contrast, although the student model exhibits slightly higher prediction errors, it still effectively captures the overall trends and major fluctuations in passenger flow. More importantly, it offers a significant advantage in computational efficiency, with a parameter size comparable to Transformer models, making it well-suited for lightweight deployment scenarios.

Table 6 | Comparison of the results of ablation studies with different graph inputs and attention mechanisms

Model	RMSE	MAE	WMAPE (%)
GEME-Net no G_c	18.12	10.18	15.06
GEME-Net no G_r	18.06	9.62	14.53
GEME-Net with G_p	17.82	9.64	14.70
GEME-Net with GCN	18.04	9.73	14.79
GEME-Net no EDSFM	18.98	10.37	15.02
GEME-Net no MSR-Atte	18.31	9.81	15.11
GEME-Net (teacher model)	17.57*	9.48*	14.30*

Numbers marked with * indicate statistical significance compared to the ablation model (T test with p value < 0.01).

We conducted an ablation study to examine the sensitivity of passenger flow prediction to different input graphs and optimization modules. As shown in Table 6, removing the graph G_c led to increases of 3.13% (RMSE), 7.38% (MAE), and 5.31% (WMAPE). Excluding the regional spatial clustering graph G_r resulted in smaller error increases of 2.79%, 1.58%, and 1.61%, respectively. This indicates that G_c has a greater impact on model performance, with the MAE increasing most significantly. Moreover, replacing G_c with the physical topology graph (GEME-Net with G_p) caused a further drop in prediction accuracy. This highlights the importance of capturing semantic and temporal information embedded in long-range mobility chains, which are more effective than raw physical topology for

learning spatial dependencies. Among all modules, removing the EDSFM had the largest negative impact, increasing errors by 8.03% (RMSE), 9.39% (MAE), and 5.03% (WMAPE). These results underscore the effectiveness of incorporating public transit operational information and confirm the central role of EDSFM in the prediction framework. Specifically, by dynamically integrating causal features of transit events, EDSFM significantly enhances short-term prediction accuracy under event-driven flow fluctuations.

Discussion

This study explores the evolution mechanism of passenger flow distribution within integrated transportation hubs. Based on multimodal data fusion and the reconstruction of passenger mobility chains, we developed a novel passenger flow prediction model named GEME-Net. Specifically, this research uncovers how passenger flow fluctuations in different areas within an integrated multimodal infrastructure are significantly impacted by public transportation and reveals the topological characteristics of passenger mobility networks. Additionally, we propose a deep learning prediction framework that incorporates spatial semantic relationships and public transportation event encoding. Overall, the study integrates real-time multimodal transport data (railway and metro) with spatiotemporal passenger behavior features, and comprehensively applies techniques from behavioral analysis, spatial modeling, and deep learning to effectively uncover dynamic patterns of passenger movement under multimodal transport events. These insights provide methodological support for operators to implement adaptive real-time management strategies—such as targeted crowd dispersion and flexible scheduling, thereby enhancing the urban transport system's responsiveness to dynamic mobility demands.

The detailed application process can be conceived as follows: firstly, high-precision monitoring sensor systems deployed in various functional areas within the hub dynamically capture real-time passenger flow and spatial distribution data, integrating operational information from multiple transportation modes, including railways, subways, buses, and flight schedules. Subsequently, this real-time data is transmitted to a central database, cleaned and integrated through data pre-processing modules, and then used as input for the GEME-Net prediction model. The model quickly processes this input to generate real-time predictions of future passenger flow distributions at a regional level (prediction time of teacher model: 6.87 s, student model: 1.41 s, Transformer: 4.97 s on validation set). These results are directly interfaced with the hub's operational management platform, enabling instant visualization and alerting. When significant fluctuations or crowding trends are predicted, the system automatically triggers an early-warning mechanism, sending real-time alerts to relevant departments. Based on these insights, operational managers can implement precise passenger guidance, dynamically adjust facility resources (such as temporarily opening additional entrances or adjusting the operating hours of commercial facilities), and redeploy staff, significantly enhancing passenger flow management efficiency and safety assurance within the hub.

The application of knowledge distillation provides significant advantages to our prediction model. On one hand, knowledge distillation effectively reduces the number of model parameters and computational load, facilitating rapid, lightweight predictive responses during real-world deployments, thereby lowering reliance on computational resources. On the other hand, transferring knowledge from the teacher model to the student model allows the student model to maintain relatively high accuracy while significantly enhancing computational efficiency. The practical significance of this method is particularly pronounced in emergency response scenarios, allowing rapid deployment on terminals or edge devices, supporting managers in making efficient real-time decisions, thus greatly enhancing the agility and practicality of predictive responses.

Compared to traditional methods based on physical topology for spatial correlation, this research employs digital hub scenarios and behavioral experiments to reconstruct passenger spatial mobility chains, offering significant theoretical advantages. Traditional methods often struggle to capture nonlinear and diverse actual passenger activity paths, whereas digital scenarios combined with behavioral experiments can more finely characterize individual choices and collective interaction behaviors within space. Although the proposed method requires additional behavioral experiment data collection through digital twin environments, resulting in higher data acquisition costs compared to baseline models. This additional effort significantly enhances model performance. We also acknowledge that the behavioral dataset underpinning our mobility-chain reconstruction comprises 420 trajectories, sufficient for the present analyses yet still modest and future work will develop more efficient, scalable protocols to collect and reconstruct large-scale hub travel processes. Moreover, the value of the digital twin approach extends beyond improved predictive accuracy; it also supports long-term strategic functions such as spatial planning, operational optimization, and emergency response simulation, thereby playing a vital role in sustainable transport management. For example, establishing a digital hub could preemptively simulate and evaluate the impacts of temporary spatial layout adjustments on passenger behaviors. The method proposed in this study allows managers to anticipate potential impacts of layout changes, thereby promoting refined and advanced management of public transportation.

However, several limitations remain in this study. First, the current prediction framework has not yet fully incorporated real-time data from urban road traffic and flight operations, which may constrain its accuracy in forecasting overall flows across the integrated transport system. Future work will aim to include these data sources to enhance the model's holistic forecasting capabilities. Second, this study characterizes the Hongqiao hub's activity network and observes a small-world–like organization. We emphasize that this finding is case-specific: network structure in transport hubs is shaped by local layout, functional area, and operational practices. Future work should replicate the analysis across multiple hubs of varying sizes and service mixes, and across temporal regimes.

Data availability

The data generated and/or analyzed during the current study are not publicly available for legal/ethical reasons.

Code availability

The code developed for this study can be made available upon request to the corresponding author.

Received: 4 July 2025; Accepted: 25 November 2025;
Published online: 27 January 2026

References

1. Mohan K. M., Timme, M. & Schröder, M. Efficient self-organization of informal public transport networks. *Nat. Commun*. 15, 4910 (2024).
2. Yu, C. et al. Multi-layer regional railway network and equitable economic development of megaregions. *npj. Sustain. Mobil. Transp.* **2**, 3 (2025).
3. Ma, C., Peñasco, C. & Anadón, L. D. Technology innovation and environmental outcomes of road transportation policy instruments. *Nat. Commun.* **16**, 4467 (2025).
4. Wen, X., Si, B., Xu, M., Zhao, F. & Jiang, R. A passenger flow spatial–temporal distribution model for a passenger transit hub considering node queuing. *Transport. Res. Part C. Emerg. Technol.* **163**, 104640–104640 (2024).
5. Auad-Perez, R. & Van Hentenryck, P. Ridesharing and fleet sizing for on-demand multimodal transit systems. *Transport. Res. Part C: Emerg. Technol.* **138**, 103594 (2022).
6. Van Der Voort, M., Dougherty, M. & Watson, S. Combining kohonen maps with arima time series models to forecast traffic flow. *Transport. Res. Part C: Emerg. Technol.* **4**, 307–318 (1996).
7. Kumar, S. V. & Vanajakshi, L. Short-term traffic flow prediction using seasonal ARIMA model with limited input data. *Eur. Transp. Res. Rev.* **7**, 21 (2015).

8. Yang, H. et al. A network traffic forecasting method based on SA optimized ARIMA–BP neural network. *Comput. Netw.* **193**, 108102 (2021).
9. Yang, H.-F., Dillon, T. S., Chang, E. & Phoebe Chen, Y.-P. Optimized configuration of exponential smoothing and extreme learning machine for traffic flow forecasting. *IEEE Trans. Ind. Inform.* **15**, 23–34 (2019).
10. Zhang, S., Zhang, J., Yang, L., Wang, C. & Gao, Z. COV-STFormer for short-term passenger flow prediction during COVID-19 in urban rail transit systems. *IEEE Trans. Intell. Transport. Syst.* **25**, 3793–3811 (2023).
11. Hoogendoorn, S. P., van Wageningen-Kessels, F., Daamen, W., Duives, D. C. & Sarvi, M. Continuum theory for pedestrian traffic flow: local route choice modelling and its implications. *Transport. Res. Proc.* **7**, 381–397 (2015).
12. Zanariah, A. R., Mustafa, M., Ashaari, Y. & Sadullah, A. F. M. Modeling pedestrian behavior in rail transit terminal. *Appl. Mech. Mater.* **567**, 742–748 (2014).
13. Helbing, D. & Molnár, P. Social force model for pedestrian dynamics. *Phys. Rev. E* **51**, 4282–4286 (1995).
14. Xiao, Y., Lv, Y. & Zhu, Z. Understanding pedestrian route choice behavior in the continuous space: diversity and equilibrium. *Transport. Res. Part C: Emerg. Technol.* **156**, 104336 (2023).
15. Berceanu, C., Banu, I., Husebo, B. S. & Patrascu, M. Predictive agent-based crowd model design using decentralized control systems. *IEEE Syst. J.* **17**, 1383–1394 (2022).
16. Leite, D. & Bacco, C. D. Similarity and economy of scale in urban transportation networks and optimal transport-based infrastructures. *Nat. Commun.* **15**, 7981 (2024).
17. Sun, Y., Leng, B. & Guan, W. A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system. *Neurocomputing* **166**, 109–121 (2015).
18. Castro-Neto, M., Jeong, Y.-S., Jeong, M.-K. & Han, L. D. Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. *Expert Syst. Appl.* **36**, 6164–6173 (2009).
19. Xu, D. et al. Real-time road traffic state prediction based on kernel-KNN. Transportmetrica. *A Transp. Sci.* **16**, 104–118 (2018).
20. Tak, S., Kim, S., Jang, K. & Yeo, H. Real-time travel time prediction using multi-level k-nearest neighbor algorithm and data fusion method. *Comput. Civil Build. Eng.* 1861–1868 (2014).
21. Guo, J., Huang, W. & Williams, B. M. Adaptive Kalman filter approach for stochastic short-term traffic flow rate prediction and uncertainty quantification. *Transport. Res. Part C: Emerg. Technol.* **43**, 50–64 (2014).
22. Ma, X., Tao, Z., Wang, Y., Yu, H. & Wang, Y. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transport. Res. Part C: Emerg. Technol.* **54**, 187–197 (2015).
23. Ma, X. et al. Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction. *Sensors* **17**, 818 (2017).
24. He, Y., Zhao, Y., Luo, Q. & Tsui, K.-L. Forecasting nationwide passenger flows at city-level via a spatiotemporal deep learning approach. *Phys. A Stat. Mech. Appl.* **589**, 126603–126603 (2021).
25. Fu, R., Zhang, Z. & Li, L. Using LSTM and GRU neural network methods for traffic flow prediction. In *2016 31st Youth Academic Annual Conference of Chinese Association of Automation* (YAC) (2016).
26. Zhao, L. et al. T-GCN: a temporal graph convolutional network for traffic prediction. *IEEE Trans. Intell. Transport. Syst.* **21**, 3848–3858 (2020).
27. Zhang, J., Chen, F., Cui, Z., Guo, Y. & Zhu, Y. Deep learning architecture for short-term passenger flow forecasting in urban rail transit. *IEEE Trans. Intell. Transport. Syst.* **22**, 7004–7014 (2021).
28. Dai, Z., Li, D. & Feng, S. Attention mechanism with spatial-temporal joint deep learning model for the forecasting of short-term passenger flow distribution at the railway station. *J. Adv. Transport*. **2024**, 7985408 (2024)
29. Liu, Y. M., Rasouli, S., Wong, M., Yan, H. & Huang, T. RT-GCN: Gaussian-based spatiotemporal graph convolutional network for robust traffic prediction. *Inf. Fusion* **102**, 102078–102078 (2024).
30. Chen, C. et al. Gated Residual Recurrent Graph Neural Networks for Traffic Prediction. *Proc. AAAI Conf. Artif. Intell.* **33**, 485–492 (2019).
31. Chen, Y. et al. Graph attention network with spatial-temporal clustering for traffic flow forecasting in intelligent transportation system. *IEEE Trans. Intell. Transport. Syst.* **24**, 8727–8737 (2022).
32. Wang, Y., Jing, C., Xu, S. & Guo, T. Attention based spatiotemporal graph attention networks for traffic flow forecasting. *Inf. Sci.* **607**, 869–883 (2022).
33. Li, G. et al. Towards integrated and fine-grained traffic forecasting: a spatio-temporal heterogeneous graph transformer approach. *Inf. Fusion* **102**, 102063–102063 (2023).
34. Tan, Y., Liu, H., Pu, Y., Wu, X. & Jiao, Y. Passenger flow prediction of integrated passenger terminal based on K-means–GRNN. *J. Adv. Transport.* **2021**, 1–14 (2021).
35. Toqué, F., Khouadjia, M., Come, E., Trepanier, M. & Oukhellou, L. Short & long term forecasting of multimodal transport passenger flows with machine learning methods. *IEEE Xplore* 560–566 (2017).
36. Guo, X., Grushka-Cockayne, Y. & De Reyck, B. Forecasting airport transfer passenger flow using real-time data and machine learning. *Manuf. Serv. Oper. Manag.* **24**, 2797–3306 (2021).
37. Zheng, W. & Mou, R. A dynamic network loading model for hub station pedestrian flow collection and distribution. *Mathematics* **11**, 3654–3654 (2023).
38. Li, Y. & Zhang, M. Resilience assessment for streamline network of transportation hub complex based on service performance. *J. Infrastruct. Syst.* **29**, 2 (2023).
39. Yue, H., Zhang, M., Duan, Y. & Yang, B. Layout of guidance signs in passenger hubs based on passenger activity line. *IOP Conf. Ser. Mater. Sci. Eng.* **688**, 044061–044061 (2019).
40. Vaswani A. et al. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* **30**, 1–15 (2017).
41. Zhou, H. et al. Informer: beyond efficient transformer for long sequence time-series forecasting. *arxiv* https://arxiv.org/abs/2012.07436 (2020).
42. Chen, C., Liu, Y., Chen, L. & Zhang, C. Bidirectional spatial-temporal adaptive transformer for urban traffic flow forecasting. In *IEEE Transactions on Neural Networks and Learning Systems* 1–13 (2022).
43. Feng, Y., Duives, D. C. & Hoogendoorn, S. P. Development and evaluation of a VR research tool to study wayfinding behaviour in a multi-story building. *Saf. Sci.* **147**, 105573 (2022).
44. Emi, M. Y. et al. Prevalence and associated risk factors of suicidal behaviors among cancer patients in a tertiary care hospital in Bangladesh. *Sci. Rep.* **15**, 7055 (2025).
45. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
46. Rustamaji, H. C. et al. Community detection with greedy modularity disassembly strategy. *Sci. Rep.* **14**, 4694 (2024).
47. Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–438 (1969).
48. Perozzi, B., Al-Rfou, R. & Skiena, S. DeepWalk: online learning of social representations. *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD* **14**, 701–710 (2014).
49. Page, L., Brin, S., Motwani, R. & Winograd, T. The PageRank citation ranking: bringing order to the web. *Web Conf.* **98**, 161–172 (1999).

50. Alvarez-Ramirez, J., Rodriguez, E. & Carlos Echeverría, J. Detrending fluctuation analysis based on moving average filtering. *Phys. A: Stat. Mech. Appl.* **354**, 199–219 (2005).
51. Bringmann, K., van Fischer, N., Evangelos K., Tomasz K., Rotenberg, E. Dynamic time warping. In *information retrieval for music and motion* 69–84 (2007).
52. Ding, X., Chen, H., Zhang, X., Han, J. & Ding, G. RepMLPNet: hierarchical vision MLP with re-parameterized locality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 568–577 (2022).
53. Heideman, M., Johnson, D. & Burrus, C. Gauss and the history of the fast fourier transform. *IEEE ASSP Mag.* **1**, 14–21 (1984).
54. Sun, Y. et al. Retentive network: a successor to transformer for large language models. *arxiv* https://arxiv.org/pdf/2307.08621 (2023).
55. Ye, L., Rochan M., Liu, Z. & Wang, Y. Cross-modal self-attention network for referring image segmentation. *arxiv.1904.04745* (2019).
56. Zhang, J., Zheng, Y. & Qi, D. Deep spatio-temporal residual networks for citywide crowd flows prediction. *arxiv* 1655–1661 (2016).
57. Yu, B., Yin, H. & Zhu, Z. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In *Proceedings twenty-seventh international joint conference on artificial intelligence* 3634–3640 (2018).

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities (Grant no. 2025JBZX077, 2022JBQY006 and 2024YJS081) and the National Natural Science Foundation of China (72471023). We also thank the China Scholarship Council for its financial contribution.

Author contributions

Z.D. conceived and designed the study. Z.D., L.Z., H.L. and R.Z. collect the data. Z.D. analyzes and interprets the results. Z.D. and R.Z. Make a draft manuscript preparation. D.L., S.R. and Y.F. supervised this study. All authors reviewed the results and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Dewei Li.