

Short-Term Metro Passenger Flow Forecasting Using a Context-Aware Graph Transformer

Ruilin Li¹, Honghong Li^{1,*}, Jianhua Wu², Yongfu He¹

¹Gongqing College of Nanchang University, Jiujiang, Jiangxi, China

²School of Electronic and Information Engineering, Nanchang University, Nanchang, Jiangxi, China

*Corresponding Author

Abstract: Metro operators require dependable station-level forecasts over the next hour to align service capacity with demand and intervene before local crowding develops. Such forecasts are difficult because recent flow trajectories interact with inter-station movement and with disruptions caused by weather, holidays, and events. This paper develops a Context-Aware Graph Transformer with Adaptive Context Fusion (CaGT). The spatial block learns unequal influence among connected stations, while the temporal block represents the evolution of the network over successive 15-min intervals. Instead of appending every external variable directly to the historical series, a learned gate controls how much contextual information enters the prediction at each state. Tests on operational records from all 94 stations of Nanchang Metro produced an RMSE of 1.76, an MAE of 1.28, and a MAPE of 6.9. Relative to the Transformer baseline, RMSE and MAE decreased by 20.4% and 21.0%, respectively. The ablation results attribute these gains to both attention-weighted station interaction and state-dependent context use. CaGT therefore offers a practical means of representing short-horizon metro demand when routine temporal regularity is interrupted by changing operating conditions.

Keywords: Short-Term Ridership Prediction; Metro Networks; Graph Transformer; Graph Attention; Adaptive Context Fusion; Spatio-Temporal Modelling

1. Introduction

Reliable forecasts for the next few operating intervals allow metro managers to adjust capacity, prepare station staff, and respond before crowding becomes critical. The task differs from an ordinary univariate forecast:

demand develops across a connected rail system and changes with both the timetable and the wider travel environment.

A station's current inflow is partly inherited from its own recent history, but it is also shaped by transfers, route choice, and accessibility elsewhere in the network. Weather, public holidays, and major events can further displace the timing or location of demand. A useful model must therefore distinguish regular temporal structure from spatial propagation and scenario-specific disturbance.

Deep learning has widened the range of dependencies that can be represented in transport data. For example, DCRNN embeds diffusion over a network within a recurrent forecasting architecture [1]. The Transformer takes a different route, using self-attention to connect observations at separated positions in a sequence [2]. Together, these developments provide the spatial and temporal foundations for network-wide demand prediction.

Neither foundation, however, fully resolves the conditions encountered in metro operation. Three issues are particularly relevant to short-horizon station forecasting. First, a physical link does not imply a stable or equal passenger relationship. Adjacent stations can exchange very different volumes because transfer status, surrounding land use, and time of day alter the strength and direction of movement. A graph fixed only by infrastructure may consequently miss operational changes in station coupling.

Second, external variables are commonly appended to historical observations as if they were uniformly informative. In practice, light weather variation may add little during a routine commute, whereas severe conditions or a local event can become decisive. The model should determine the relevance of context rather than assign it a constant role.

Third, irregular periods contain shifts that are weakly represented in recent history. Event-related surges, holiday travel, or unusual weather can move demand away from its normal time and station, reducing the reliability of pattern matching alone.

Dynamic graphs, Graph Transformers, and multi-source models have begun to address these problems by enlarging the set of spatial relations and explanatory inputs. Less attention has been paid to controlling when those inputs should affect the forecast. This distinction is important: collecting context and using it selectively are not the same modelling decision.

To address this gap, the proposed CaGT architecture couples graph-attention-based station learning with Transformer temporal encoding. Its adaptive fusion gate conditions the contribution of weather, calendar, and event variables on the current passenger-flow representation, rather than merging them indiscriminately.

The paper makes three contributions:

- (1) It formulates short-term metro demand as a joint network and sequence problem, allowing station interaction and flow evolution to be learned in one forecasting pipeline.
- (2) It introduces a state-dependent context gate that can suppress weak external signals during routine periods and strengthen them when operating conditions depart from the historical pattern.
- (3) It evaluates the architecture on 15-min records from 94 Nanchang Metro stations through baseline comparisons, component ablation, and scenario-specific error analysis.

2. Related Work

2.1 Graph-Based Spatio-Temporal Forecasting

Graph-based forecasting treats locations as nodes and operational relationships as edges. This representation allows a station estimate to draw on observations beyond its own sequence, but model behaviour depends heavily on how the graph is constructed and updated.

Recent work has extended graph reasoning beyond a single fixed adjacency structure. Hu et al. embedded a Graph Transformer in a multi-gate mixture-of-experts model for urban rail demand [3], while PDFormer represented delayed propagation over long spatial ranges [4]. These designs enlarge the receptive field,

although they also make the choice of spatial relations more consequential.

STGCN established an earlier benchmark by coupling graph convolution with temporal convolution [5]. Removing recurrent updates improved parallel computation, but the spatial operation still inherited the assumptions encoded in a predefined adjacency matrix. For a metro system, those assumptions may be too rigid when transfer intensity changes across periods.

GAT relaxes uniform neighbourhood aggregation by learning a separate attention weight for each connected node [6]. This mechanism is attractive for metro data because two physically adjacent stations need not exert comparable predictive influence.

Subsequent models have made the spatial structure increasingly adaptive. ASTGCN applies attention in both space and time [7]; Graph WaveNet learns latent adjacency in addition to the supplied graph [8]; and GMAN coordinates several attention mechanisms to represent network interactions [9]. The common direction is a shift from fixed averaging towards data-dependent station coupling.

Most of these architectures were designed for general traffic networks. Metro demand adds transfer behaviour, station function, and service conditions to physical connectivity, so a road-oriented graph cannot be assumed to capture all relevant passenger relationships.

2.2 Transformer-Based Spatio-Temporal Forecasting

The temporal component of metro demand combines sharp commuting peaks, recurring daily structure, and short local disturbances. It must preserve recent continuity without losing dependencies that extend beyond adjacent intervals.

Self-attention in the Transformer permits direct comparison between separated positions and avoids the strictly sequential updates of a recurrent network. This property is useful when a current station count depends on several earlier parts of the input window rather than only the most recent observation.

Transformer moved towards unified time-space prediction by allowing temporal and spatial representations to interact within the same attention architecture [10]. Such integration can capture broad dependence, but it does not by itself specify how heterogeneous metro context should enter the representation.

STKGformer further introduced knowledge-graph information into spatio-temporal Transformer forecasting [11]. The study shows the value of relational information that lies beyond a single infrastructure graph, particularly when station connections have different semantic meanings.

Graph Transformer models therefore provide a flexible bridge between network structure and sequence learning. Their advantage is the ability to compare distant or heterogeneous nodes; their cost is greater computational demand and a stronger need to control irrelevant relations.

For short-horizon metro operation, additional complexity is justified only when it improves the representation of changing station influence or irregular demand. A large attention model that treats every input as equally useful can still amplify noise.

This leaves a specific design problem: temporal attention and graph attention can learn when observations are related, but external context also requires a mechanism that decides whether a weather, calendar, or event signal is informative for the current state.

2.3 Metro Passenger-Flow Forecasting

Metro forecasting differs from road-traffic prediction because the network is fixed, entry and exit counts are station based, and transfers propagate demand along scheduled routes. At the same time, station roles and travel purposes create heterogeneity that is not visible from topology alone.

Zhan et al. addressed network-wide ridership with parallel multi-graph convolution and gated recurrent units [12], demonstrating that distinct graphs can contribute complementary spatial evidence.

Other studies combine graph learning with sequence encoders. Chen et al. stacked graph convolution with bidirectional LSTM [13], whereas Bao et al. weighted several graph views before sequence learning [14]. Both approaches recognise that a single physical adjacency matrix provides an incomplete account of station interaction.

Multi-source information has also entered metro forecasting. Fu et al. fused spatial, temporal, and auxiliary data [15], and He et al. used a multi-graph convolutional-recurrent architecture to represent different passenger-flow relations [16]. These studies improve input coverage but largely fuse sources according to a preset

architecture.

Semantic and dynamic relationships provide another line of development. Zeng and Tang incorporated a knowledge graph [17]; Huang et al. learned changing network-wide propagation patterns [18]; and IG-Net represented interaction between stations rather than limiting the graph to direct physical links [19].

Wu et al. connected prediction with the interpretation of irregular travel patterns [20]. This emphasis is operationally important because the largest errors often occur when routine commuting regularity no longer dominates.

Taken together, the literature has progressed from single-station history to multi-graph, semantic, and context-rich representations. Weather, holidays, and events are now common inputs, but they are still frequently encoded as static additions. The conditional value of those inputs remains less developed.

2.4 Research Gap

The preceding work reveals three connected limitations rather than three independent modelling problems.

First, even adaptive graph models usually begin from a fixed or narrowly defined set of station relations. Their learned weights can vary, but potentially useful changes in passenger interaction may remain outside the initial graph.

Second, strong temporal encoders do not guarantee that spatial and contextual evidence will be used coherently. A model may capture a long sequence while still admitting external variables at the wrong time or station.

Third, direct concatenation assumes that context has stable predictive value. Metro demand contradicts that assumption: a holiday label, weather observation, or event indicator may be decisive in one state and almost redundant in another.

The required framework must therefore learn station influence, retain temporal continuity, and condition the flow of external information on the state being forecast. Context should be selected within the model, not merely appended to the input.

CaGT is designed around this requirement. It combines graph attention, Transformer encoding, and a learned context gate so that the three information sources can contribute at different strengths.

3. Methodology

3.1 Problem Formulation

Let the metro system be encoded as a graph whose nodes correspond to stations and whose edges represent operational connections. The task is to infer a future sequence of station-level counts from recent graph-indexed observations and their accompanying context:

$$G=(V,E,A) \quad (1)$$

Here, V is the station set, E contains the links between stations, and A is the adjacency matrix used to describe network topology.

For N stations, the network-wide passenger-flow vector observed at interval t is written as:

$$X_t=[x_t^1, x_t^2, \dots, x_t^N] \quad (2)$$

Each element records the flow observed at one station i during interval t .

Each observation is paired with variables that describe the operating environment. Their values at interval t form the contextual vector:

$$C_t=[c_t^1, c_t^2, \dots, c_t^M] \quad (3)$$

M denotes the number of contextual variables. In this application they encode calendar position, weather, holidays, and event-related conditions. Given the recent flow sequence, the corresponding contextual sequence, and the metro graph, the forecasting mapping is:

$$\hat{Y}_{t+1:t+T} = f_{\theta}(F_{t-T+1:t}, C_{t-T+1:t}, A) \quad (4)$$

In this notation, L is the historical window, T' is the output horizon, and the function f_{θ} denotes the mapping learned by CaGT. Records are sampled every 15 min. Each sample uses the preceding 12 intervals (3 h) to estimate the next four intervals, giving a 1-h prediction horizon.

A 3-h input window retains the immediate build-up and decay of station demand while limiting the amount of stale information passed to the model. Longer windows were avoided because additional history increases computation and may mix the current regime with earlier travel conditions.

3.2 Overall Framework of CaGT

CaGT is organised around the three information decisions identified above: which stations matter, which earlier intervals matter, and how strongly external context should enter the current estimate.

1. learn unequal dependencies among connected stations;
2. encode the evolution of those station representations over time; and
3. regulate contextual information in response to

the current flow state.

These decisions are implemented by three linked components:

- (1) a graph-attention spatial encoder;
- (2) a Transformer temporal encoder; and
- (3) an adaptive context-fusion gate.

The first component produces a representation for each station from its neighbourhood. The second tracks how that representation changes across the input window, after which the gate combines it with the projected context vector.

Figure 1 shows the complete data path from historical inputs to multi-interval forecasts.

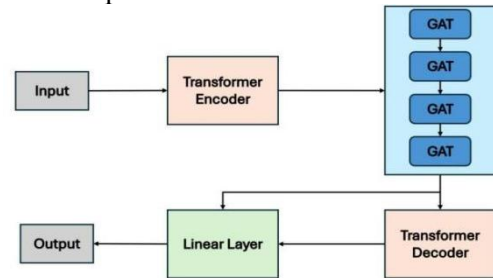


Figure 1. Architecture of the Context-Aware Graph Transformer with Adaptive Context Fusion (CaGT)

The fusion block does not assign a fixed coefficient to weather, calendar, or event variables. Its coefficient is learned jointly from the spatio-temporal state and the contextual embedding.

Consequently, the model can rely mainly on recent passenger flow during routine operation and admit more contextual evidence when the observed state departs from that routine.

3.3 Adaptive Spatial Dependency Learning

Rail infrastructure determines which stations are directly connected, but it does not determine a constant strength of interaction. The volume propagated over a link can change with transfer activity, station function, and time of day.

Recent multi-source spatio-temporal work likewise indicates that origin-destination and contextual features can alter the apparent relationship between stations [21]. This motivates a spatial representation that is reweighted from the observed features.

CaGT therefore uses a Graph Attention Network (GAT). Instead of averaging all neighbours uniformly, GAT assigns a learned coefficient to each connected station before aggregation.

For a connected pair of stations i and j , the unnormalised attention score is:

$$e_{ij,t}^{(m)} = \text{LeakyReLU} \left[\left(\mathbf{a}^{(m)} \right)^T \left(\mathbf{W}^{(m)} \mathbf{h}_{i,t} \| \mathbf{W}^{(m)} \mathbf{h}_{j,t} \right) \right] \quad (5)$$

For each attention head m , W is the learnable transformation, a is the attention vector, and $\parallel$ denotes concatenation.

Scores over the neighbourhood of station i are normalised as:

$$\alpha_{ij,t}^{(m)} = \frac{\exp(e_{ij,t}^{(m)})}{\sum_{k \in N_i} \exp(e_{ik,t}^{(m)})} \quad (6)$$

The neighbourhood set contains the stations connected to station i .

The spatial state is then obtained by attention-weighted aggregation:

$$H_i^s = \sigma\left(\sum_{j \in N_i} \alpha_{ij} W h_j\right) \quad (7)$$

The resulting vector is the spatial representation produced for station i .

Because the coefficients are feature dependent, the same infrastructure edge can carry different predictive weight under different operating states.

3.4 Temporal Dependency Learning

The station representations from each interval are ordered into a sequence before temporal encoding.

The sequence must preserve commuting peaks, recurring daily structure, rapid local fluctuations, and demand responses that appear after a short delay.

A Transformer encoder is used because self-attention can compare all positions in the 12-interval window directly. This choice is consistent with recent joint self-attention and multi-graph metro forecasting [22]. For one attention head, the scaled dot-product operation is:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V} \quad (8)$$

$\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote the query, key, and value matrices, respectively.

Applying the Transformer to the spatial sequence gives:

$$H^t = \text{Transformer}(H^s) \quad (9)$$

The output is the resulting spatio-temporal representation.

The encoder contains two layers, each using eight attention heads; the hidden representation has 256 dimensions.

3.5 Adaptive Context Fusion

The distinctive element of CaGT is not the presence of contextual data, but the way their contribution is controlled.

A direct concatenation scheme exposes every forecast to the same set of external inputs. That

design can transmit irrelevant variation when recent flow already provides a stable signal.

The usefulness of context is scenario dependent: severe weather can change mode choice, holidays can redistribute commuting demand, and an event can create a local surge around a small group of stations.

CaGT represents this conditional role with a gate learned from both the spatio-temporal state and the contextual embedding.

The raw context vector is first mapped into the model feature space:

$$\tilde{C} = W_c C + b_c \quad (10)$$

The contextual projection is parameterised by a weight matrix and a bias term.

The gate is then computed from the two information sources:

$$g = \sigma(W_g [H^s \parallel \tilde{C}] + b_g) \quad (11)$$

The vector g assigns an element-wise share to the passenger-flow representation and the projected context.

The gated representation used by the prediction layer is:

$$H^f = g \odot H^s + (1-g) \odot \tilde{C} \quad (12)$$

The circled-dot operator in Eq. (12) denotes element-wise multiplication.

Values of g near one preserve the learned passenger-flow state; smaller values admit a larger contextual correction. The balance is estimated separately for the features rather than fixed in advance.

This state-dependent balance is the principal difference between CaGT and a model that simply appends context to its input.

3.6 Prediction Layer and Optimisation

A fully connected output layer maps the fused state to station-level forecasts for the four target intervals.

Parameters are estimated by minimising mean squared error:

$$L = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (13)$$

The loss in Eq. (13) compares each observed passenger-flow value with its estimate.

Adam W is used for parameter updates; Section 4.2 reports the remaining training controls.

4. Experiments and Results

4.1 Dataset Description

The empirical study uses operational passenger-flow records from Nanchang Metro, a four-line system comprising 94 stations. Its lines

carried noticeably different annual volumes in 2024 (Figure 2), providing a heterogeneous setting in which to test whether the model can distinguish strong and weak spatial demand relationships.

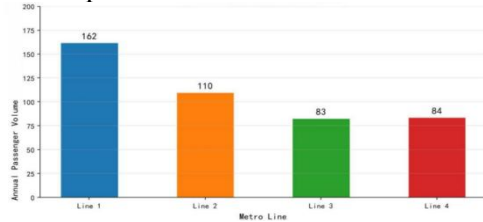
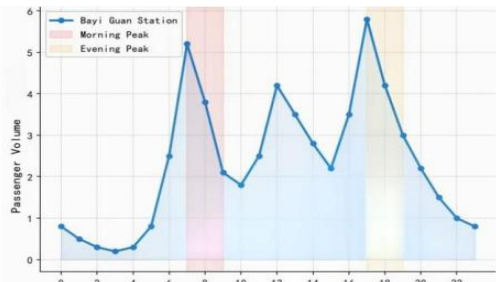
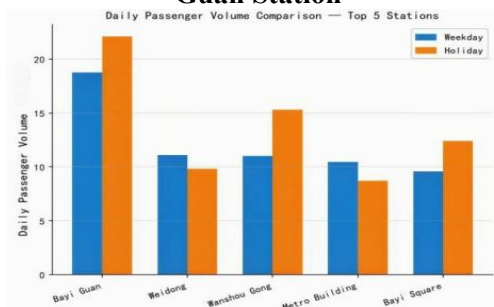


Figure 2. Annual Passenger Volume by Nanchang Metro Line, 2024

Variation is also visible within the day and across operating scenarios. Figure 3(a) shows the weekday profile at Bayi Guan Station, where demand rises sharply around the morning and evening commutes. Figure 3(b) contrasts weekday and holiday totals for the five busiest stations. The unequal changes across stations indicate that a calendar effect cannot be represented adequately by applying one common adjustment to the whole network.



(a) Hourly Passenger-Flow Profile at Bayi Guan Station



(b) Weekday and Holiday Passenger-Volume Comparison

Figure 3. Temporal and Contextual Variation in Station Passenger Flow

Counts were aggregated into 15-min intervals. Each of the 94 stations forms one graph node, and direct rail links provide the initial edge set; attention weights are subsequently learned from the station features.

Three groups of contextual variables accompany the flow records: time of day and calendar

position (including weekday and holiday status), observed weather, and indicators for special events that may produce localised demand.

These variables are included to distinguish deviations from a routine commuting profile. Their value is expected to be limited on an ordinary working day but greater when weather, a holiday, or an event changes where and when passengers travel. Samples were ordered by time. The earliest 70% were used for model fitting, the following 15% for validation, and the final 15% for out-of-sample testing. Preserving this order prevents later observations from leaking into the estimation period.

A rolling window converted the series into supervised samples: 12 observed intervals form the input and the immediately following four intervals form the target. Thus, every prediction uses 3 h of history to estimate the next hour.

4.2 Experimental Settings

CaGT was implemented in PyTorch and trained on a workstation with an NVIDIA RTX 3090 GPU.

AdamW was configured with an initial learning rate of 1×10^{-4} and a weight-decay coefficient of 1×10^{-4} . Each update used 64 samples, and training was capped at 150 epochs.

Validation loss was monitored after each epoch. Training stopped if no improvement occurred for 15 successive epochs, and all runs used random seed 42.

Table 1 consolidates the settings used in the reported experiments.

The two-layer design was chosen to retain short-horizon structure without adding depth that contributed little new information and increased computation.

Table 1. CaGT Hyperparameter Settings

Parameter	Setting
Input sequence length	12 intervals
Forecasting horizon	4 intervals
GAT layers	2
Transformer layers	2
Attention heads	8
Hidden dimension	256
Dropout rate	0.1
Batch size	64
Optimiser	AdamW
Learning rate	1×10^{-4}
Maximum epochs	150
Early stopping patience	15
Random seed	42

4.3 Baseline Models

Six baselines were selected to separate gains due to non-linearity, graph structure, temporal-spatial coupling, and self-attention. The comparison includes the following model families:

ARIMA, a statistical reference based on autoregressive and moving-average terms;

LSTM, a recurrent encoder for non-linear sequence dependence;

GCN, which aggregates information over the station graph;

GAT, which learns unequal weights for connected stations;

STGCN, which couples graph and temporal convolution; and

Transformer, which models the input window through self-attention.

Together these baselines provide a progression from a single-series statistical model to network-aware and attention-based predictors.

4.4 Evaluation Metrics

Forecast quality is reported from three complementary perspectives: average absolute deviation, sensitivity to large errors, and relative percentage deviation.

Mean Absolute Error (MAE)

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (14)$$

The metrics compare each observed passenger-flow value with its corresponding estimate.

Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (15)$$

Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{|S|} \sum_{i \in S} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (16)$$

For MAPE, the evaluation set contains only observations whose actual passenger flow is non-zero.

Zero-flow intervals are omitted only from the MAPE denominator to prevent division by zero; they remain available to the absolute-error measures.

4.5 Comparison with Baseline Models

Table 2 reports performance on the held-out chronological test set.

Table 2. Test-Set Performance of the Forecasting Models

Model	RMSE	MAE	MAPE
-------	------	-----	------

ARIMA	4.82	3.58	18.6
LSTM	3.65	2.71	14.2
GCN	3.21	2.38	12.5
GAT	2.89	2.08	11.1
STGCN	2.54	1.85	9.8
Transformer	2.21	1.62	8.5
CaGT	1.76	1.28	6.9

The pattern across models is more informative than any single error value.

ARIMA records the highest errors (RMSE 4.82; MAE 3.58), showing that a separate linear series for each station is inadequate for the non-linear, network-coupled behaviour in these data.

LSTM lowers RMSE to 3.65 by learning non-linear temporal structure, but it does not exchange information between stations. The remaining gap to the graph models is therefore consistent with missing spatial evidence.

GCN and GAT reduce RMSE to 3.21 and 2.89, respectively. The additional gain from GAT indicates that treating all connected stations alike is too restrictive; learned neighbour weights provide a better description of passenger interaction.

STGCN reaches an RMSE of 2.54, while the Transformer reaches 2.21. Their relative performance suggests that the 3-h window contains temporal relationships that benefit from direct self-attention across intervals.

CaGT gives the lowest error on every metric: RMSE 1.76, MAE 1.28, and MAPE 6.9. Against the Transformer, the reductions are 20.4% in RMSE and 21.0% in MAE. Because the temporal encoder is shared conceptually with that baseline, the remaining improvement is associated with station reweighting and controlled use of context.

4.6 Ablation Study

Ablation was performed with three nested configurations: Transformer alone, GAT plus Transformer, and the complete CaGT model.

Their test errors are reported in Table 3.

Adding GAT to the Transformer lowers RMSE from 2.21 to 2.02 and MAE from 1.62 to 1.47. Thus, information arriving from connected stations remains useful after the temporal sequence has been encoded.

Table 3. Component Ablation for CaGT

Configuration	RMSE	MAE	MAPE
Transformer	2.21	1.62	8.5
GAT+Transformer	2.02	1.47	7.8
CaGT	1.76	1.28	6.9

Introducing the context gate produces a second

and larger reduction, taking RMSE to 1.76 and MAPE from 7.8 to 6.9. The result supports selective context use rather than unconditional feature concatenation. The two steps contribute different information: graph attention refines where demand is related, whereas the gate controls when external conditions should modify the learned flow pattern.

4.7 Analysis under Different Operating Conditions

Aggregate accuracy can conceal scenario-dependent failure. Figure 4 therefore separates four operating regimes: ordinary working days, holidays, adverse weather, and large-scale events.

Errors are lowest on ordinary working days, when departure times and station use follow a familiar commuting rhythm. In this regime, the recent flow window supplies most of the information required by the forecast.



Figure 4. Forecasting Error across four Operating Conditions

Holiday errors are higher because leisure and tourism trips redistribute demand across both time and station. A regular weekday profile is consequently a weaker reference. Adverse weather adds a different form of uncertainty. Residential stations, commercial areas, and transport hubs do not respond uniformly, so a network-wide weather correction would obscure local effects.

Large events are the most difficult regime. Their effect is concentrated near particular venues and intervals, and an event indicator alone cannot describe attendance, arrival waves, or dispersal behaviour.

CaGT remains more robust when routine patterns are disturbed, but the residual error under events also shows where richer real-time mobility and origin-destination information would be valuable.

5. Discussion

5.1 Interpretation of Model Performance

The test results support a network-conditioned

interpretation of short-term metro demand. Temporal history remains central, but its predictive value increases when the model can trace how demand is related across stations and recognise when the operating context has changed.

The progression from LSTM to GCN and GAT isolates the spatial contribution. LSTM models each sequence without an explicit station graph; GCN introduces that graph; and GAT allows the strength of each neighbour to vary. The monotonic reduction in error across these models is consistent with increasingly realistic station interaction.

This result does not imply that physical topology is unimportant. Topology provides a defensible initial neighbourhood, while attention determines how much evidence should pass over each connection for the observed state.

The distinction matters operationally. A transfer link can dominate during the commute yet become less informative later in the day, whereas a commercial destination may show the opposite pattern on weekends or holidays. Fixed graph aggregation cannot express this change as directly.

5.2 Effectiveness of Adaptive Context Fusion

The context gate accounts for the improvement from GAT plus Transformer to the full CaGT model.

External information has no uniform baseline value. Weather may be weak during routine commuting, a holiday may replace rather than merely scale the weekday profile, and an event may affect only a few stations. Direct concatenation gives all of these signals a permanent route into the predictor, even when they add little.

CaGT instead computes the contextual share from the current spatio-temporal state. The ablation result shows that this conditional route adds information beyond spatial and temporal encoding alone.

The practical implication is that richer data are useful only if the model can also suppress irrelevant inputs. Future context-aware systems should evaluate both the coverage of external variables and the mechanism by which those variables enter a forecast.

5.3 Practical Implications for Metro Operation

At a one-hour horizon, the forecasts are most

relevant to short-cycle operating decisions rather than long-term planning.

Network-wide estimates can identify where a demand increase is likely to emerge before it becomes visible in platform crowding. Operators may then adjust train frequency or prepare capacity at the affected part of the line.

Station-level outputs also support targeted crowd management, including staff positioning, passenger guidance, and the allocation of temporary barriers or access controls.

The context gate is particularly relevant when normal travel behaviour is disrupted, because it allows the same forecasting system to respond differently to a holiday, weather episode, or event.

Deployment would nevertheless require an interface with live passenger counting, service control, and station-management systems. The present experiments evaluate predictive accuracy; they do not yet demonstrate a closed-loop operational intervention.

5.4 Limitations and Future Research

Three limitations define the boundary of the reported results.

First, the contextual inputs are structured variables. They do not include live crowd density, detailed origin-destination movement, or a continuous measure of surrounding activity. Those sources may be especially important for the event regime, where the largest residual errors occur.

Second, attention reweights an initial graph derived mainly from infrastructure. It cannot create every missing relation. A dynamic multi-view graph that combines topology, transfer flows, similarity, and station function may represent changing interaction more completely.

Third, all experiments use one metro system. Network scale, line layout, passenger composition, and service practice vary by city, so cross-city tests are necessary before generalisation can be claimed.

6. Conclusion

This paper examined whether short-horizon metro forecasts improve when external context is regulated rather than appended uniformly. The resulting CaGT model combines attention-weighted station aggregation, Transformer temporal encoding, and a gate conditioned on the current passenger-flow state.

On 15-min records from 94 Nanchang Metro stations, CaGT achieved an RMSE of 1.76, an MAE of 1.28, and a MAPE of 6.9. Compared with the Transformer, RMSE fell by 20.4% and MAE by 21.0%.

The ablation sequence clarifies the source of this improvement. Graph attention first reduces error by distinguishing strong from weak station influence; adaptive fusion then reduces it further by controlling the contextual correction. The gain therefore does not arise from adding inputs alone.

The results indicate that metro context is most valuable when interpreted relative to the current network state. This conclusion is strongest for the observed Nanchang system and should not yet be treated as evidence of cross-city transferability.

Future work will test dynamic multi-view graphs, incorporate real-time origin-destination and crowd information, and evaluate the model on metro networks with different scales and operating patterns.

Acknowledgments

This work was supported by:

- (1) Jiujiang City Natural Science Foundation: Optimization of Subway Short-Term Passenger Flow Prediction Model Using Context-Aware Graph Transformer (CaGT) (2025_000504).
- (2) Jiangxi Provincial Higher Education Institutions' Humanities and Social Sciences Research Project (Red Culture Special Project): Research on Jiangxi's Approach to Integrating Red Culture into Engineering Education Using Generative Artificial Intelligence (HSWH25025).
- (3) Jiangxi Provincial Department of Education: Research on Reforming Electrical Engineering Talent Training under the "Government-Enterprise-University-Association" Collaboration Mechanism (BKJG-2026-31-11).

References

- [1] Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2018). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. International Conference on Learning Representations (ICLR).
- [2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

- [3] Hu, S., Chen, J., Zhang, W., Liu, G., & Chang, X. (2024). Graph transformer embedded deep learning for short-term passenger flow prediction in urban rail transit systems: A multi-gate mixture-of-experts model. *Information Sciences*, 679, 121095.
- [4] Jiang, J., Han, C., Zhao, W.X., & Wang, J. (2023). PDFormer: Propagation Delay-Aware Dynamic Long-Range Transformer for Traffic Flow Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4), 4365–4373.
- [5] Yu, B., Yin, H., & Zhu, Z. (2018). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 3634–3640.
- [6] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph Attention Networks. *International Conference on Learning Representations (ICLR)*.
- [7] Guo, S., Lin, Y., Feng, N., Song, C., & Wan, H. (2019). Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 922–929.
- [8] Wu, Z., Pan, S., Long, G., Jiang, J., & Zhang, C. (2019). Graph WaveNet for deep spatial-temporal graph modelling. *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 1907–1913.
- [9] Zheng, C., Fan, X., Wang, C., & Qi, J. (2020). GMAN: A Graph Multi-Attention Network for Traffic Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(1), 1234–1241.
- [10] Jin, D., Shi, J., Wang, R., Li, Y., Huang, Y., & Yang, Y. (2023). Trafformer: Unify Time and Space in Traffic Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(7), 8114–8122.
- [11] Hu, Z., Chang, S., Zhang, W., et al. (2026). STKGformer: A spatio-temporal knowledge graph enhanced transformer for passenger flow prediction. *Information Sciences*, 735, 123055.
- [12] Zhan, S., Cai, Y., Xiu, C., Zuo, D., Wang, D., & Wong, S.C. (2024). Parallel framework of a multi-graph convolutional network and gated recurrent unit for spatial-temporal metro passenger flow prediction. *Expert Systems with Applications*, 251, 123982.
- [13] Chen, P., Fu, X., & Wang, X. (2022). A graph convolutional stacked bidirectional LSTM neural network for metro ridership prediction. *IEEE Transactions on Intelligent Transportation Systems*, 23(7), 6950–6962.
- [14] Bao, J., Kang, J., Yang, Z., & Chen, X. (2022). Forecasting network-wide multi-step metro ridership with an attention-weighted multi-view graph to sequence learning approach. *Expert Systems with Applications*, 210, 118475.
- [15] Fu, X., Zuo, Y., Wu, J., Yuan, Y., & Wang, S. (2022). Short-term prediction of metro passenger flow with multi-source data: A neural network model fusing spatial and temporal features. *Tunnelling and Underground Space Technology*, 124, 104486.
- [16] He, Y., Li, L., Zhu, X., & Tsui, K.L. (2022). Multi-Graph Convolutional-Recurrent Neural Network for short-term forecasting of transit passenger flow. *IEEE Transactions on Intelligent Transportation Systems*, 23(10), 18155–18174.
- [17] Zeng, J., & Tang, J. (2023). Combining knowledge graph into metro passenger flow prediction: A split-attention relational graph convolutional network. *Expert Systems with Applications*, 213, 118790.
- [18] Huang, H., Mao, J., Kang, L., Lu, W., Zhang, S., & Liu, L. (2024). A dynamic graph deep learning model with multivariate empirical mode decomposition for network-wide metro passenger flow prediction. *Computer-Aided Civil and Infrastructure Engineering*, 39, 2596–2618.
- [19] Li, P., Wang, S., Zhao, H., Yu, J., Hu, L., Yin, H., & Liu, Z. (2023). IG-Net: An interaction graph network model for metro passenger flow forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 24(4), 4147–4157.
- [20] Wu, F., Zheng, C., Zhou, S., Lu, Y., Wu, Z., & Zheng, S. (2025). An interpretable approach to passenger flow prediction and irregular passenger travel pattern understanding in metro systems. *Expert Systems with Applications*, 265, 125991.
- [21] Ren, Y., Tang, J., Ying, C., & Dong, Q. (2026). A multi-source feature-integrated spatiotemporal graph transformer network

- for short-term origin–destination passenger flow prediction in urban rail transit. *Journal of Intelligent Transportation Systems*, 1–19.2669875
- [22] Yang, Q., Yu, J., Wang, Z., Hu, X., Chen, Y., & Li, F. (2026). Short-term metro flow forecasting via spatiotem- poral joint self-attention and multi- graph weighted fusion. *International Journal of Transportation Science and Technology*. Volume 159,106491,