MF^3: Efficient Multimodal Federated Learning with Dual-Path Mamba-Transformer for Metro Flow Prediction

Anonymous Author(s)
Submission Id: 2104

ABSTRACT

Metro flow prediction is a critical application in smart city and Web of Things infrastructures, essential for optimizing urban mobility. However, building such predictive systems faces three key challenges: (1) the fragmentation of multimodal spatiotemporal data, (2) the inefficiency of existing models in capturing long-range dependencies, and (3) the data silos and privacy concerns inherent in distributed station infrastructures. To address these challenges, a multimodal federated learning framework named **MF^3** (**M**amba-Trans**f**ormer-**F**ederated Metro **F**low Prediction) is proposed. **First**, a multimodal alignment (MA) module is designed, where cross-modal alignment attention bridges visual and spatiotemporal features, thereby enhancing feature complementarity and alignment. **Second**, a dual-path Mamba-Transformer (DMT) module is designed, in which Mamba's linear long-range memory and the Transformer's global perception operate in parallel, reducing information loss. **Third**, a blockchain-based federated reputation (BFR) module is established to perform personalized federated learning, thereby enhancing privacy protection. Finally, extensive experiments on real metro datasets from Hangzhou and Shanghai demonstrate that MF^3 achieves superior performance in terms of prediction accuracy. In summary, the proposed MF^3 framework provides a new feasible paradigm for metro flow prediction, supporting urban traffic optimization, metro operation and scheduling, and the development of smart city and Web of Things infrastructures.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; *Distributed computing methodologies*; • **Information systems** → **Spatial-temporal systems**.

KEYWORDS

Metro flow prediction, Multimodal alignment, Dual-path Mamba-Transformer, Federated learning

1 INTRODUCTION

Urban rail transit systems, especially metro networks, are a critical component of modern metropolitan infrastructures and play an essential role in Web of Things (WoT)-enabled smart city ecosystems [25, 42]. Unlike traditional traffic flow prediction that primarily deals with road networks characterized by irregular topology and vehicle flow, metro flow prediction focuses on structured, station-based passenger movements constrained by fixed routes and schedules, making it inherently multimodal and spatiotemporally dependent. With the widespread deployment of sensing devices, mobile applications, and automated fare collection systems, massive amounts of multimodal passenger flow data are continuously generated and connected through the web-based infrastructure of urban mobility services [23, 26]. These data streams provide unprecedented opportunities to enhance operational efficiency, support real-time passenger services, and enable the seamless integration of metro systems into broader WoT infrastructures [22]. Consequently, accurate metro passenger flow prediction is not only a core task in intelligent transportation management but also a fundamental requirement for building reliable systems and infrastructures that optimize train dispatch intervals, reduce passenger waiting times, and alleviate congestion in WoT-driven urban environments [14, 38, 41]. Nevertheless, reliable prediction remains challenging due to multimodal and fragmented data distributions, long-term temporal dependencies coupled with local dynamics, and decentralized station-level data storage that complicates both privacy protection and federated collaboration.

Early studies primarily rely on statistical methods, such as regression analysis [9] and time-series modeling [31], to perform metro flow prediction. Although these approaches are simple and interpretable, they often fail to capture nonlinear dynamics and complex spatial dependencies inherent in metro systems. To overcome these limitations, machine learning techniques such as decision trees and support vector machines are gradually introduced, offering stronger predictive power [37]. However, these models still lack the ability to effectively represent sequential correlations across time. With the rise of deep learning, recurrent neural networks (RNNs) [11], long short-term memory (LSTM) networks [27], and gated recurrent units (GRUs) [40] become dominant, as they are able to capture temporal dependencies more comprehensively. Nevertheless, their sequential nature limits scalability and makes it difficult to model long-range dependencies. Furthermore, to address spatial complexity, graph neural networks (GNNs) [7, 13] are proposed to handle non-Euclidean structures and uncover inter-station relationships. However, their effectiveness often diminishes when dealing with multimodal and heterogeneous inputs. More recently, Transformers demonstrate strong capabilities in modeling long-range temporal dependencies through attention mechanisms [14], while the emerging Mamba model opens new opportunities for efficient long-sequence processing with linear complexity and enhanced local feature extraction [10, 15]. Despite these advancements, existing approaches still fall short in WoT-driven urban infrastructures, where metro flow prediction must be accurate, robust to multimodal inputs, and efficient under decentralized and privacy-sensitive environments. Consequently, these gaps call for new frameworks that can not only enhance predictive performance but also align with the requirements of Web, mobile, and WoT infrastructures, thereby motivating the design of the proposed MF^3 framework.

Building upon the recent developments and their limitations, it becomes evident that existing models still struggle to meet the demands of WoT-integrated metro systems. Specifically, metro passenger flow prediction in WoT-enabled infrastructures faces three key challenges:

(1) Heterogeneous modalities pose a significant challenge to effective fusion. Multimodal data from time series, spatial topologies, and external factors are heterogeneous and fragmented, making cross-modal fusion in WoT systems difficult.

(2) Global attention mechanisms suppress the modeling of local dynamics. Transformers struggle with long-sequence efficiency, where global attention often suppresses local dynamics, limiting accurate modeling of both short- and long-term patterns.

(3) Decentralized data silos hinder effective collaborative modeling. Metro stations operate as decentralized data silos, and in WoT infrastructures, privacy protection and communication overhead hinder effective collaborative modeling.

To address these challenges, a privacy-preserving multimodal federated learning (FL) framework, MF^3 (Multimodal FL with Dual-path Mamba-Transformer), is proposed. MF^3 consists of three core modules: MA (Multimodal Alignment), DMT (Dual-Path Mamba-Transformer), and BFR (Blockchain-based Federated Reputation), which correspond to the key functional components of the framework. Specifically, MA enhances cross-modal feature fusion, DMT improves long-sequence pattern capture, and BFR enables privacy-aware collaborative modeling across decentralized stations. Figure 1 provides an overview of the proposed MF^3 framework, illustrating the integration of metro flow prediction within the selected WoT-driven infrastructure and highlighting the interactions among its three major modules.

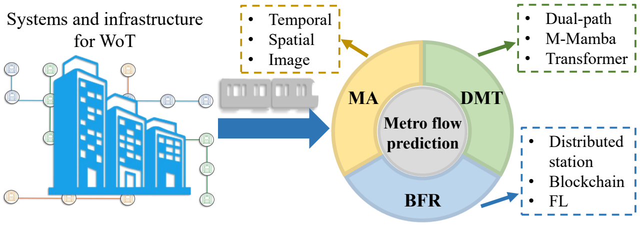

Figure 1: The overall architecture of the proposed MF^3 framework. It supports the construction of reliable systems and infrastructures for Web, mobile, and WoT environments through metro flow prediction.

First, the MA module is designed to mitigate the heterogeneity across modalities. In particular, a self-generated heatmap method based on passenger flow data is introduced: for each metro station (viewed as a client in FL), four heatmaps are generated to represent weekday inflow, weekday outflow, weekend inflow, and weekend outflow. Moreover, by incorporating both the station itself and its neighboring stations, this visual modality provides an intuitive representation of the dynamic distribution of passenger flows. Consequently, imaging-based encoding enhances the robustness of cross-modal fusion and lays a solid foundation for unified modeling, which is essential for accurate metro flow prediction in WoT-enabled urban infrastructures.

Second, the DMT module with a fusion gate is designed to capture both local fluctuations and long-term dependencies. In the temporal path, the DMT adopts a dual-channel embedding: one channel inputs fine-grained time series data to represent continuous variations in passenger flows, while the other introduces weekday/weekend-based periodic features to model rhythmic travel patterns. In the spatial path, the model leverages self-generated heatmaps as visual inputs, combined with station adjacency information, to jointly learn spatial–visual dependencies. Through efficient long-sequence modeling, the DMT module enhances predictive robustness and supports accurate metro flow prediction within WoT-enabled smart transportation networks.

Third, to address privacy and communication challenges in decentralized metro systems, the BFR module is designed. This framework dynamically evaluates each station's contribution through multi-dimensional metrics including local performance, historical consistency, and model similarity. The reputation-based aggregation prioritizes reliable clients while suppressing noisy participants. Additionally, personalized model delivery through similarity clustering and adaptive blending ensures customized solutions for each station's operational patterns. This approach significantly reduces communication overhead while enhancing model fairness and robustness, making it particularly suitable for WoT-enabled urban infrastructures with strict privacy requirements.

The main contributions of this paper can be summarized as follows:

(1) MA for unified metro flow prediction. An effective multimodal embedding and alignment mechanism is designed to integrate temporal, spatial, and visual modalities. In particular, self-generated metro flow heatmaps are introduced as visual features, which directly reflect passenger dynamics and enhance cross-modal fusion robustness.

(2) DMT for efficient spatiotemporal modeling. The Mamba-Former model is proposed, incorporating a dual-path fusion gate that combines Mamba's linear-complexity local feature extraction with Transformer's global dependency modeling. This design captures complex spatiotemporal dependencies efficiently and effectively.

(3) BFR for robust and fair federated aggregation. A dynamic reputation mechanism built on blockchain is developed to support adaptive model updates. This approach significantly reduces communication overhead, improves system robustness against unreliable participants, and strengthens privacy protection in distributed metro environments.

(4) Comprehensive empirical validation. Extensive experiments are conducted on real-world metro datasets. The results show that MF^3 achieves significant performance improvements compared with the strongest available baselines, with up to **55.10%** improvement for MAE, **62.52%** for MAPE, and **47.32%** for RMSE on the Shanghai dataset, demonstrating its superior accuracy, efficiency, and privacy preservation.

Through these contributions, MF^3 offers a scalable and secure framework for metro passenger flow prediction, effectively integrating multimodal learning, efficient long-sequence modeling, and trustworthy FL, while addressing the challenges of decentralized data and privacy constraints inherent in WoT-enabled urban infrastructures.

2 PROBLEM AND METHOD

In this section, the metro flow prediction problem is introduced by clarifying the basic terminology and providing a formal definition.

Based on this formulation, the proposed MF^3 framework is then presented, followed by a detailed explanation of its methodological components. Finally, the overall architecture of MF^3 is illustrated in Figure 2, highlighting how the framework supports accurate and scalable metro flow prediction within WoT-enabled urban infrastructures and distributed smart transportation systems.

2.1 Problem Statement

Before delving into the details of the MF^3 framework, it is helpful to provide an overview of the metro flow prediction problem and its key concepts, facilitating a clearer understanding of the task addressed by the model.

Let the metro system be represented by a directed graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}), \tag{1}$$

where each vertex $v_i \in \mathcal{V}$ is a station and each edge $(v_i, v_j) \in \mathcal{E}$ is a physical track segment.

The metro flow data are represented using three complementary components:

- **Time-series tensor:**

$$\mathbf{X} = \left\{ X_i^{(t)} \in \mathbb{R}^d \mid v_i \in \mathcal{V},\ t \in \mathcal{T} \right\}, \tag{2}$$

  where $X_i^{(t)}$ stacks d-dimensional measurements, including entries, exits, and transfers, for the station v_i in the time slot t. This tensor captures the temporal dynamics of passenger flows across all stations.
- **Spatial adjacency:**

$$\mathbf{A} \in \{0,1\}^{|\mathcal{V}|\times|\mathcal{V}|}, \tag{3}$$

  where represents the connectivity of the metro network, where $\mathbf{A}_{ij} = 1$ if stations v_i and v_j are directly connected, and $\mathbf{A}_{ij} = 0$ otherwise. This adjacency matrix encodes the spatial topology of the rail network.
- **Flow-density heatmap:**

$$\mathbf{H}_i^{(t)} \in \mathbb{R}^{h\times w}, \tag{4}$$

  where provides a pixel-level representation of passenger density on the platform of the station v_i at the time slot t. This visual modality captures spatial distributions of passenger flows within each station.

Given the historical observations $\mathcal{D}_{\text{hist}} = \{\mathbf{X}, \mathbf{A}, \mathbf{H}\}$ up to the current time-stamp T, the goal is to learn a predictive mapping:

$$\mathcal{F}_\theta : \left(\mathbf{X}^{(T-\tau+1:T]}, \mathbf{A}, \mathbf{H}^{(T-\tau+1:T]}\right) \longrightarrow \hat{\mathbf{Y}}^{(T+1:T+p]}, \tag{5}$$

where $\hat{\mathbf{Y}}^{(T+1:T+p]}$ denotes the predicted passenger flows for each station v_i over the future horizon $k = T+1, \dots, T+p$. This formulation unifies temporal, spatial, and visual information to support accurate multi-step flow prediction.

The resulting predicts are injected into the automatic train-regulation engine to optimize headways Δt_{head} in real time, minimise the average platform dwell time $\bar{w}_{\text{plat}}$, and generate early crowd-guidance alerts.

2.2 Data preprocessing

Before model training, the raw metro data must be preprocessed to ensure consistency across modalities. Proper preprocessing helps reduce noise, align heterogeneous data sources, and enhance feature representation quality, thereby laying a solid foundation for multimodal fusion.

Flow data normalization. To address the long-tailed distribution of metro flow data and mitigate the impact of extreme values, a logarithmic transformation followed by min-max normalisation is applied to the inflow and outflow time series. Specifically, for each station i and time step t, the raw flow value $x_i(t)$ is transformed as:

$$x_{\text{norm}} = \frac{\log(1 + \max(0, x_i(t)))}{\max(\log(1+X))}, \tag{6}$$

where X represents all flow values in the training set. This approach effectively handles the sparse nature of metro flow data, particularly during off-peak hours.

Temporal feature engineering. Beyond the flow values themselves, rich temporal features are extracted to capture periodic patterns in metro usage:

- **Circular time encoding**: Hour of day is encoded using sine and cosine transformations to preserve cyclical continuity:

$$\text{hour\_sin} = \sin\left(\frac{2\pi \cdot \text{hour}}{24}\right), \quad \text{hour\_cos} = \cos\left(\frac{2\pi \cdot \text{hour}}{24}\right). \tag{7}$$

- **Day type indicators**: Binary flags for weekends/holidays and peak hours (7-9 AM, 5-7 PM)
- **Sequential temporal context**: The immediate historical flow patterns are incorporated through the station's ego-graph neighborhood

Self-generated heat-maps. To provide spatial-visual context of crowd dynamics around each station, four daily heat-maps are generated per station. Formally, for the station v_i with its neighborhood $\mathcal{N}(i) = \{j \mid (v_i, v_j) \in \mathcal{E}\} \cup \{i\}$, flow patterns are aggregated by:

$$\mathbf{H}_i^{d,f} \in \mathbb{R}^{96\times|\mathcal{N}(i)|}, \qquad \mathbf{H}_i^{d,f}(t,n) = \frac{1}{|D_d|}\sum_{\delta\in D_d} x_n^f(t,\delta), \tag{8}$$

where $t = 0, 1, \dots, 95$ indexes 15-minute intervals, $n \in \mathcal{N}(i)$, D_d contains all dates of type $d \in \{\text{weekday}, \text{weekend}\}$, and $x_n^f(t,\delta)$ is the raw flow. Each heat-map row is min-max normalized to $[0,1]$ and converted to RGB images, providing visual priors for the multimodal learning framework.

2.3 Multimodal Alignment

To fully exploit the complementary information from heterogeneous data sources, it is crucial to align their representations before fusion. In metro systems, temporal sequences, spatial structures, and visual heatmaps differ in characteristics and scale, making direct integration suboptimal. To address this, the MA module is constructed, which aligns these modalities into a unified latent space and facilitates effective cross-modal feature interaction.

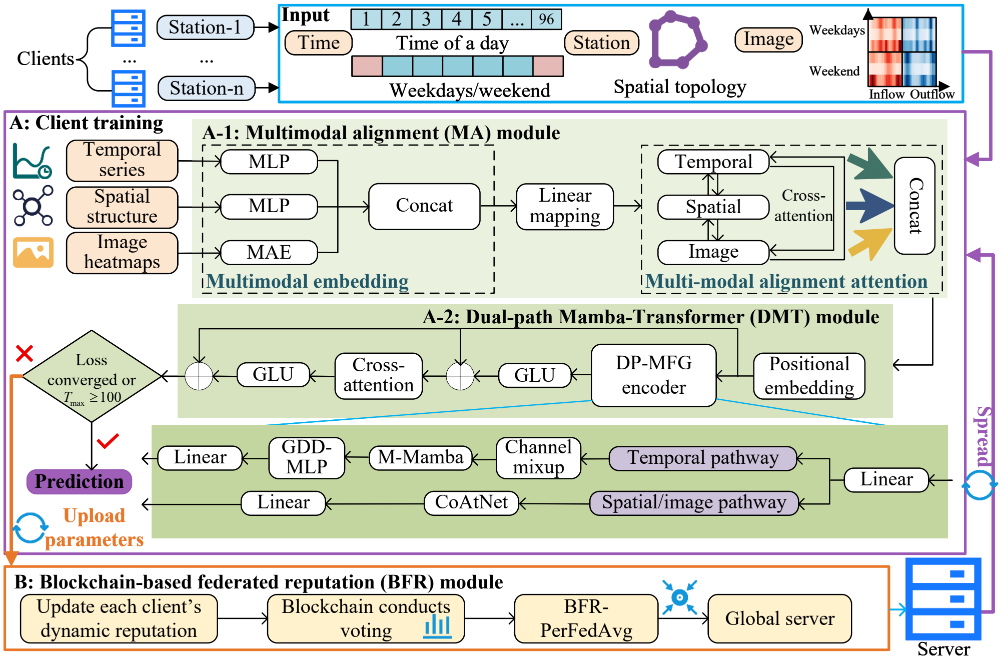

Figure 2: The Workflow of MF^3 for metro flow prediction. It consists of three modules: MA for multimodal alignment, DMT for efficient spatiotemporal modeling, and BFR for privacy-preserving federated aggregation.

Specifically, the MA module transforms each modality into modality-specific embeddings before alignment. Temporal sequences $X_t \in \mathbb{R}^{T \times d_t}$ and spatial structures $X_s \in \mathbb{R}^{S \times d_s}$ are first projected into a unified latent space through dedicated multi-layer perceptrons (MLPs):

$$\begin{aligned} Z_t &= \mathrm{MLP}_{\mathrm{temporal}}(X_t) = \mathrm{ReLU}(X_t W_t^{(1)} + b_t^{(1)}) W_t^{(2)} + b_t^{(2)}, \\ Z_s &= \mathrm{MLP}_{\mathrm{spatial}}(X_s) = \mathrm{ReLU}(X_s W_s^{(1)} + b_s^{(1)}) W_s^{(2)} + b_s^{(2)}. \end{aligned} \tag{9}$$

Concurrently, station heatmap images $X_i \in \mathbb{R}^{H \times W \times 3}$ are processed through a pre-trained Masked Autoencoder (MAE) to extract rich visual representations:

$$Z_i = \mathrm{MAE}_{\mathrm{visual}}(X_i). \tag{10}$$

The three modality embeddings are subsequently concatenated and linearly projected to ensure dimensional consistency:

$$Z_{\mathrm{concat}} = \mathrm{Linear}([Z_t \| Z_s \| Z_i]), \tag{11}$$

where $\|$ denotes the concatenation operation and the linear projection adjusts the combined dimensionality to d_{hidden}.

The core innovation of the MA module lies in its comprehensive cross-modal attention mechanism. For each of the six possible directional pairs among the three modalities, bidirectional cross-attention is performed:

$$\begin{aligned} \mathrm{CrossAttn}_{m \to n} &= \mathrm{Softmax}\left(\frac{(Z_m W_m^Q)(Z_n W_n^K)^\top}{\sqrt{d_k}} \right) (Z_n W_n^V), \\ \mathrm{CrossAttn}_{n \to m} &= \mathrm{Softmax}\left(\frac{(Z_n W_n^Q)(Z_m W_m^K)^\top}{\sqrt{d_k}} \right) (Z_m W_m^V), \end{aligned} \tag{12}$$

where $m, n \in \{\text{temporal}, \text{spatial}, \text{visual}\}$ represent distinct modalities. This results in six cross-attention outputs that capture all pairwise modal interactions.

Finally, these cross-modal representations are concatenated to form the comprehensive multimodal embedding:

$$\begin{aligned} Z_{\mathrm{ma}} = \mathrm{Concat}\Big(&\mathrm{CrossAttn}_{t \to s}, \ \mathrm{CrossAttn}_{s \to t}, \ \mathrm{CrossAttn}_{t \to i}, \\ &\mathrm{CrossAttn}_{i \to t}, \ \mathrm{CrossAttn}_{s \to i}, \ \mathrm{CrossAttn}_{i \to s}\Big). \end{aligned} \tag{13}$$

This meticulously designed alignment strategy ensures that temporal patterns are contextualized with spatial constraints and visual crowd dynamics, while spatial configurations are informed by temporal evolution and visual context. The output Z_{ma} serves as the input to the subsequent DMT module, carrying rich, cross-modally aligned representations for further processing.

2.4 Dual-Path Mamba-Transformer

The DMT module represents a sophisticated fusion of Mamba's selective state space mechanisms with Transformer's attention capabilities, designed to capture multi-scale spatiotemporal dependencies in metro flow data. The processing pipeline begins with positional encoding of the multimodal-aligned input:

$$Z_{\text{in}} = Z_{\text{ma}} + \text{PE}(t), \tag{14}$$

where $\text{PE}(t)$ denotes the sinusoidal positional embeddings that preserve temporal ordering information.

The encoded input is then processed by the DMT encoder, which operates through two specialized pathways. The temporal pathway is further divided into dual channels to capture complementary temporal dynamics:

$$\begin{aligned} Z_{\text{fine}} &= \text{MLP}_{\text{fine}}(Z_{\text{in}}), \\ Z_{\text{periodic}} &= \text{MLP}_{\text{periodic}}(Z_{\text{in}}). \end{aligned} \tag{15}$$

The fine-grained channel processes high-resolution temporal data to model continuous passenger flow variations, while the periodic channel focuses on rhythmic patterns based on weekly cycles. Both temporal representations are subsequently integrated and processed through the M-Mamba module [34], a modified variant of Mamba proposed in CMamba [34], which enhances the original state space model by introducing data-dependent mechanisms and improved cross-time dependency modeling tailored for multivariate time series prediction:

$$Z_{\text{temporal}} = \text{M-Mamba}(\text{Concat}[Z_{\text{fine}}, Z_{\text{periodic}}]). \tag{16}$$

Followed by gated dilated deep MLP (GDD-MLP) for multi-scale feature extraction:

$$Z_{\text{temp\_out}} = \text{Linear}(\text{GDD-MLP}(Z_{\text{temporal}})). \tag{17}$$

Concurrently, the spatial/image pathway processes self-generated heatmaps alongside station spatial adjacency information to model spatiotemporal dependencies at both local and global scales. This pathway employs a convolutional attention network (CoAtNet) for joint representation learning:

$$Z_{\text{spatial}} = \text{Linear}(\text{CoAtNet}(Z_{\text{in}})). \tag{18}$$

Upon exiting the DP-MFG encoder, outputs from both pathways undergo gated fusion. First, a gated linear unit (GLU) is applied to the combined representations:

$$Z_{\text{glu}} = \text{GLU}(\text{Concat}[Z_{\text{temp\_out}}, Z_{\text{spatial}}]). \tag{19}$$

Followed by residual connection with the originally position-encoded input:

$$Z_{\text{res1}} = Z_{\text{glu}} + Z_{\text{in}}. \tag{20}$$

The refined representations then proceed through cross-attention mechanisms that facilitate interaction between temporal and spatial modalities:

$$\begin{aligned} &Q = Z_{\text{res1}} W^Q, \quad K = Z_{\text{res1}} W^K, \quad V = Z_{\text{res1}} W^V, \\ &Z_{\text{attn}} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V. \end{aligned} \tag{21}$$

Another GLU operation is applied to enhance feature selectivity:

$$Z_{\text{glu2}} = \text{GLU}(Z_{\text{attn}}). \tag{22}$$

Followed by a final residual connection with the position-encoded input:

$$Z_{\text{out}} = Z_{\text{glu2}} + Z_{\text{in}}. \tag{23}$$

This meticulously designed architecture ensures comprehensive modeling of both transient flow variations and persistent spatial constraints, while the hierarchical fusion strategy enables effective information exchange across temporal and spatial domains. The output Z_{out} carries rich spatiotemporal representations suitable for final flow prediction tasks.

2.5 Blockchain-based Federated Reputation

To address the challenges of client data heterogeneity and model quality variations in FL environments, the BFR module is designed. This mechanism enhances global model performance while preserving data privacy through distributed reputation assessment and weighted aggregation strategies.

Client Model Upload and Reputation Initialization. Each client c_i uploads a triple to the server after local training:

$$\mathcal{U}_i = \left\langle i, \Theta_i, \mathcal{L}_i^{\text{val}} \right\rangle, \tag{24}$$

where Θ_i represents the model parameters of the client i, and $\mathcal{L}_i^{\text{val}}$ denotes the validation loss. The federated system initializes the reputation book:

$$\mathcal{R}^0 = \{r_i^0 = r_{\text{init}} \mid \forall i \in [N]\}, \tag{25}$$

where r_{init} is the initial reputation value, and N is the total number of clients.

Dynamic reputation update mechanism. Based on client-reported validation performance, reputation scores are updated via exponential moving average:

$$r_i^{t+1} = \alpha_{\text{ema}} \cdot r_i^t + (1 - \alpha_{\text{ema}}) \cdot \phi(\mathcal{L}_i^{\text{val}}), \tag{26}$$

where $\alpha_{\text{ema}} \in [0, 1]$ is a smoothing factor, and the transformation function $\phi(\mathcal{L}) = \frac{1}{1+\mathcal{L}}$ maps losses to reputation scores in the $[0, 1]$ interval.

Multi-dimensional voting weight calculation. The improved blockchain voting mechanism computes aggregation weights by integrating three dimensions: reputation, performance metrics, and model similarity:

$$w_i = \beta_{\text{rep}} \cdot \tilde{r}_i + \beta_{\text{metric}} \cdot \tilde{m}_i + \beta_{\text{sim}} \cdot \tilde{s}_i, \tag{27}$$

where $\tilde{r}_i$ is the normalized reputation score: $\tilde{r}_i = \frac{r_i}{\sum_{j=1}^N r_j}$, $\tilde{m}_i$ is the normalized performance score: $\tilde{m}_i = \frac{1/(1+\mathcal{L}_i^{\text{val}})}{\sum_{j=1}^N 1/(1+\mathcal{L}_j^{\text{val}})}$ and $\tilde{s}_i$ is the average model similarity: $\tilde{s}_i = \frac{1}{N-1} \sum_{j \neq i} \text{cosine}(\Theta_i, \Theta_j)$. The weight coefficients satisfy $\beta_{\text{rep}} + \beta_{\text{metric}} + \beta_{\text{sim}} = 1$, with typical settings of 0.5, 0.3, 0.2.

Personalized model aggregation strategy. For the target client c_t, top-k similar clients are selected based on data feature similarity:

$$\mathcal{S}_t = \text{topk}\left(\{\text{cosine}(\mathbf{f}_t, \mathbf{f}_i) \mid i \neq t\}\right) \cup \{t\}, \tag{28}$$

where $\mathbf{f}_i$ is the data feature vector of the client i. Personalized aggregation weights are computed as:

$$w_{i\to t} = \frac{w_i \cdot \text{cosine}(\mathbf{f}_t, \mathbf{f}_i) \cdot \gamma_i}{\sum_{j\in\mathcal{S}_t} w_j \cdot \text{cosine}(\mathbf{f}_t, \mathbf{f}_j) \cdot \gamma_j}, \tag{29}$$

where $\gamma_i = 1/(1 + \mathcal{L}_i^{\text{val}})$ is the performance adjustment factor. The aggregated model for the target client is:

$$\Theta_t^{\text{global}} = \sum_{i\in\mathcal{S}_t} w_{i\to t} \cdot \Theta_i. \tag{30}$$

Adaptive personalized blending.

The final client model is obtained through linear interpolation between global aggregation results and local models:

$$\Theta_t^{\text{final}} = (1 - \alpha_t) \cdot \Theta_t^{\text{global}} + \alpha_t \cdot \Theta_t^{\text{local}}. \tag{31}$$

The personalization coefficient α_t is dynamically adjusted based on client reputation:

$$\alpha_t = \begin{cases} \min(\alpha_t + \Delta_\alpha, \alpha_{\max}) & \text{if } \mathcal{L}_t^{\text{val}} < \text{median}(\{\mathcal{L}_i^{\text{val}}\}) \\ \max(\alpha_t - \Delta_\alpha, \alpha_{\min}) & \text{otherwise} \end{cases}, \tag{32}$$

where Δ_α is the adjustment step, and $\alpha_{\max}$ and $\alpha_{\min}$ are the upper and lower bounds for personalization, respectively.

This mechanism ensures that high-performing clients retain more personalized characteristics, while low-performing clients tend to rely more on global knowledge.

Through this blockchain-based reputation FL framework, the system can effectively screen high-quality client contributions while suppressing the influence of low-quality or malicious clients, achieving robust and personalized metro flow prediction model training while preserving data privacy.

2.6 Loss Function Design

The training objective for metro flow prediction is formulated as a multi-task regression problem, where both inflow and outflow predictions are simultaneously optimized. To address the dual challenges of typical flow variations and anomalous passenger surges in metro systems, a carefully designed loss framework is proposed.

Huber loss for robust optimization. The Huber loss is employed as the core component to balance efficient learning during normal conditions with robustness during anomalous events. This hybrid approach provides quadratic convergence for typical flow variations while maintaining linear robustness for sudden passenger surges, preventing the model from being overly sensitive to outliers while ensuring efficient gradient propagation:

$$L_\delta(a, b) = \begin{cases} \frac{1}{2}(a - b)^2 & \text{for } |a - b| \le \delta \\ \delta(|a - b| - \frac{\delta}{2}) & \text{otherwise} \end{cases}. \tag{33}$$

Composite loss formulation. Building upon the robust foundation of Huber loss, a composite loss function is designed to leverage the complementary strengths of multiple loss components:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{mse}} \cdot \mathcal{L}_{\text{mse}} + \lambda_{\text{mae}} \cdot \mathcal{L}_{\text{mae}} + \lambda_{\text{huber}} \cdot \mathcal{L}_{\text{huber}}, \tag{34}$$

where the Huber loss component is computed as:

$$\mathcal{L}_{\text{huber}} = \frac{1}{2B} \sum_{b=1}^{B} \left[L_\delta(y_b^{\text{in}}, \hat{y}_b^{\text{in}}) + L_\delta(y_b^{\text{out}}, \hat{y}_b^{\text{out}}) \right]. \tag{35}$$

Table 1: The summary statistic of two metro flow datasets.

Datasets	Stations	Connections	Steps	Intervals
Shanghai	288	958	6,716	15 min
Hangzhou	80	248	1,825	15 min

This multi-component strategy ensures balanced optimization across different operational scenarios in metro systems.

Optimization framework. To ensure stable training across heterogeneous clients in the FL environment, gradient clipping is applied during local training:

$$\text{clip}(\mathbf{g}, c) = \begin{cases} \mathbf{g} & \text{if } \|\mathbf{g}\| \le c \\ c \cdot \frac{\mathbf{g}}{\|\mathbf{g}\|} & \text{otherwise} \end{cases}. \tag{36}$$

The Adam optimizer with learning rate scheduling provides adaptive optimization:

$$\theta_{t+1} = \theta_t - \eta_t \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}. \tag{37}$$

This hierarchical loss framework progresses from robust individual components to an integrated optimization strategy, ensuring both efficient convergence during normal conditions and resilient performance during anomalous events in metro systems.

3 EXPERIMENTS

This section presents a series of experiments to comprehensively evaluate the MF^3 model, including comparative analysis, ablation studies, performance analysis, and few-shot experiment. The source code and experimental settings are available at https://anonymous.4open.science/r/MF3-code-308B

3.1 Datasets and Evaluation Metrics

Two real-world metro flow datasets, **Shanghai metro** and **Hangzhou metro** [20], covering large networks with numerous stations and extensive historical passenger flow records, are used for evaluation. Both datasets provide detailed spatiotemporal information as directed graphs with stations as nodes and tracks as edges, and time is discretized into 15-minute intervals. Table 1 summarizes key statistics, showing Shanghai Metro is larger with longer sequences, while Hangzhou Metro, though smaller, exhibits rich temporal variation, supporting robust assessment of model generalization across network scales.

In the prediction setting, each dataset is partitioned into training, validation, and testing sets with a ratio of 70%, 10%, and 20%, respectively. Three widely used evaluation metrics are adopted to assess model performance: mean absolute error (MAE), root mean square error (RMSE), and mean absolute percentage error (MAPE). The detailed computation formulas for these metrics are provided in the Appendix C.1.

3.2 Comparative Analysis

To comprehensively evaluate the performance of the proposed method, a series of comparison experiments are designed. The baselines can be categorized into two groups: the LSTM and the SVR are classical time series prediction methods, serving to verify

Table 2: Performance comparison on Shanghai and Hangzhou datasets.

	Shanghai											
	15min			30min			45min			60min		
Models	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
LSTM	33.76	34.37	60.48	49.16	41.32	58.91	45.42	45.13	68.56	47.25	49.06	70.93
SVR	36.61	42.58	50.69	41.82	46.62	57.38	48.35	48.62	57.44	52.57	58.13	60.48
ReDyNet	21.22	40.58	17.78	21.78	42.17	18.03	22.46	44.09	18.39	23.21	46.28	18.89
PDFormer	22.25	38.58	18.68	22.54	39.43	18.94	23.19	42.49	19.52	23.83	42.92	20.84
FGNN	24.48	46.61	21.23	25.58	48.64	21.91	27.06	51.54	22.65	28.79	55.08	24.07
STDGRL	22.02	42.28	18.73	22.52	43.16	19.23	23.64	45.61	20.12	24.58	47.91	23.52
AGCRN	22.37	45.92	21.17	23.69	48.01	24.29	24.66	49.51	23.26	26.15	49.51	29.16
STNNs	26.87	51.20	21.90	27.56	51.78	25.35	27.37	52.33	23.09	29.38	55.13	33.19
LLGformer	23.47	26.59	23.47	29.44	32.20	28.44	23.19	26.71	23.19	24.60	27.86	24.61
S-DGNN	12.94	15.68	29.28	12.03	14.83	26.25	11.06	13.28	24.62	9.81	12.96	22.23
MF^3	**5.81**	**8.26**	**10.97**	**8.75**	**11.00**	**14.28**	**8.29**	**10.61**	**14.19**	**7.13**	**9.72**	**12.59**
Improvement (%)	55.10	47.32	38.29	27.27	25.83	20.80	25.05	20.11	22.84	27.32	25.00	33.35
	Hangzhou											
	15min			30min			45min			60min		
Models	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
LSTM	34.02	36.56	60.32	38.31	40.95	68.76	45.23	44.62	69.35	46.73	48.62	70.64
SVR	36.84	42.78	50.23	41.52	47.91	57.23	48.09	48.61	58.19	53.62	59.93	60.77
ReDyNet	21.22	37.51	18.65	22.18	38.28	19.12	22.67	38.93	19.53	23.25	40.00	19.82
PDFormer	22.35	37.45	21.63	21.98	38.83	18.88	22.34	40.78	19.34	23.02	42.80	19.97
FGNN	27.36	48.33	27.19	28.86	51.83	27.49	30.64	56.26	29.10	32.57	60.88	30.74
STDGRL	23.62	47.17	24.97	24.09	51.67	26.09	25.06	51.67	25.06	25.73	51.67	23.03
AGCRN	23.16	41.38	22.78	24.61	45.23	24.32	26.19	49.84	26.09	27.56	53.88	26.58
STNNs	29.93	50.47	26.91	28.45	52.12	25.40	29.36	54.76	25.91	30.12	56.63	26.31
LLGformer	23.68	27.88	23.68	27.08	32.49	27.08	36.07	29.39	26.07	29.74	32.19	29.74
S-DGNN	11.49	14.80	27.14	11.26	13.86	24.29	10.71	13.82	22.40	8.91	12.20	20.83
MF^3	**6.29**	**7.86**	**17.31**	**5.22**	**7.12**	**15.80**	**6.61**	**9.50**	**18.70**	**6.29**	**7.86**	**17.31**
Improvement (%)	45.26	46.89	7.18	53.64	48.63	16.31	38.28	31.26	3.31	29.41	35.57	12.66

the basic effectiveness of the model; the subsequent methods are state-of-the-art approaches proposed in recent years, which have demonstrated strong capability in modeling complex spatiotemporal dependencies and dynamic variations in metro or traffic flow prediction.

The comparison results of all baseline methods and the proposed approach are summarized in Table 2, where **bold** and underline denote the optimal and suboptimal values, respectively. As shown, the proposed method consistently outperforms the baselines across multiple evaluation metrics and time intervals. The calculated improvement percentages quantitatively demonstrate its superior effectiveness. Specifically, it achieves a remarkable performance gain over the best baseline by reducing the error metrics by up to 55.10% in MAE, 62.52% in MAPE, and 47.32% in RMSE on the Shanghai dataset. Similar substantial improvements are observed on the Hangzhou dataset, confirming the model's robustness and exceptional capability in capturing complex spatiotemporal dependencies for metro flow prediction.

3.3 Ablation Studies

To validate the effectiveness of each module in the proposed MF^3 framework, ablation experiments were conducted by removing different components and comparing the prediction performance. The results are illustrated in Figure 3. It can be observed that removing the MA module leads to a noticeable degradation across all three metrics, indicating its importance in integrating heterogeneous modal information. Furthermore, eliminating the DMT results in further performance decline, which highlights the necessity of this module for capturing complex spatiotemporal dependencies. In contrast, the removal of the BFR mechanism causes a dramatic deterioration, where MAE, RMSE, and MAPE increase to 18.47, 21.12, and 27.04, respectively, far worse than the complete model. This demonstrates that BFR plays a crucial role in ensuring robustness and trustworthiness during federated collaboration. Overall, the three modules complement each other and collectively guarantee the superior performance of MF^3 in metro flow prediction.

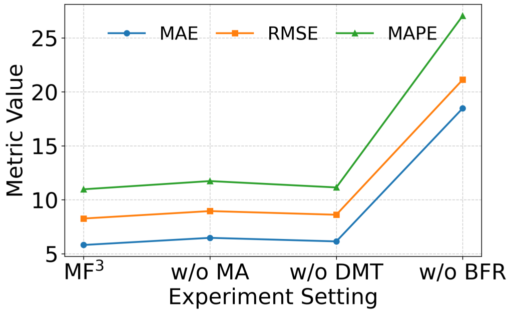

Figure 3: Ablation experiments on the Shanghai dataset.

3.4 Performance Analysis

This subsection systematically analyzes the prediction performance of the proposed MF^3 model in Shanghai and Hangzhou. By examining the error trends over training rounds for different time intervals as well as the client performance distribution, the model's convergence characteristics, prediction accuracy, and fairness in FL can be comprehensively evaluated.

Figure 4 illustrates the MAE and RMSE over training rounds for Shanghai and Hangzhou under different time intervals. It can be observed that as the number of training rounds increases, the MAE generally decreases and gradually converges, indicating stable prediction performance and good convergence in both cities. In Shanghai, the convergence speed and final accuracy vary among different time intervals, with 15-minute and 30-minute intervals achieving the best performance and the lowest final MAE. In comparison, Hangzhou exhibits slightly higher MAE values overall, but the trend is similar to that of Shanghai, demonstrating the model's robustness across cities and reflecting the impact of geographic characteristics on prediction performance.

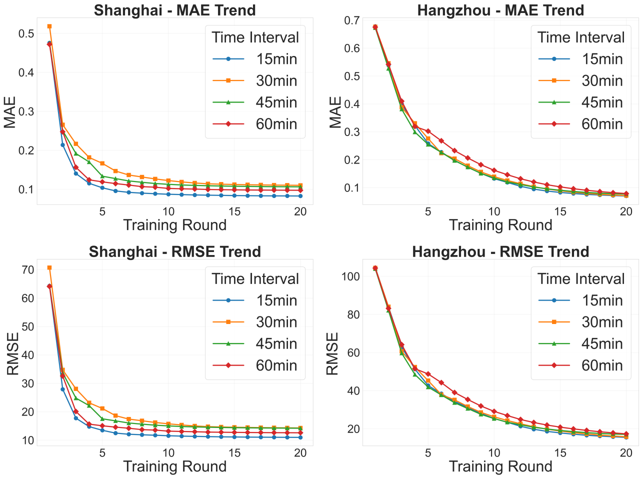

Figure 4: MAE and RMSE of MF^3 over training rounds in Shanghai and Hangzhou.

Moreover, RMSE is more sensitive to outliers, and its trends reflect the stability of predictions. In Shanghai, the RMSE decreases smoothly with minimal differences across intervals, indicating stable handling of outliers. Hangzhou's RMSE is slightly higher overall but follows a similar trend, providing a basis for cross-city comparison.

Additional experimental results, including hangzhou metro loss curves and client performance under MF^3 in FL, are provided in the Appendix C.3.

In summary, the experimental results demonstrate that the MF^3 achieves good convergence, stability, and fairness across different cities and time intervals, effectively supporting multi-city, multi-modal metro flow prediction tasks.

3.5 Few-Shot Experiment

To evaluate the performance of MF^3 under limited training data, a few-shot experiment is conducted. In this setting, the training, validation, and test sets are split in a ratio of 1:2:7, ensuring that effective feature representations can be learned even with scarce training data. The model is trained and tested on data from different time intervals and cities to examine its convergence and prediction accuracy under low-data conditions.

Table 3 presents the few-shot learning performance on Shanghai dataset across different time intervals:

Table 3: The few-shot learning performance on shanghai dataset.

Intervals	MAE	RMSE	MAPE
15min	14.38	16.65	22.38
30min	19.51	21.68	28.74
45min	20.38	22.67	29.61
60min	10.79	13.66	17.42

The experimental results demonstrate that even under the extreme condition of using only 10% of the training data, the MF^3 model maintains commendable prediction performance. Compared with the state-of-the-art baseline methods, the proposed few-shot scheme delivers significantly superior performance with a markedly lower MAE, thereby conclusively proving its exceptional data efficiency.

4 CONCLUSION

This paper proposes MF^3, a privacy-preserving multimodal FL framework for metro flow prediction within WoT-enabled smart city infrastructures. It integrates the MA module for cross-modal fusion, the DMT module for spatiotemporal modeling, and the BFR module for fair aggregation. Experiments on Shanghai and Hangzhou datasets show that MF^3 consistently outperforms existing methods in accuracy, stability, and few-shot performance, demonstrating its effectiveness in real-world WoT scenarios.

Future work will incorporate few-shot and zero-shot learning to improve adaptability under limited or unseen data and leverage transfer learning across cities and time domains to enhance generalization and scalability, making the framework more flexible and intelligent. This will further strengthen the model's practical applicability in real-world metro systems.

REFERENCES

[1] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. 2020. Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in Neural Information Processing Systems* 33 (2020), 17804–17815.

[2] Liyuan Chang and Bin Cao. 2025. Towards efficient network traffic classifications via multi-modal learning. *IEEE Internet of Things Journal* (2025). https://doi.org/10.1109/JIOT.2025.3563987

[3] Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. 2025. Visionts: Visual masked autoencoders are free-lunch zero-shot time series forecasters. (2025). arXiv:2408.17253

[4] Dawei Cui, Zewei Zhang, and Tongfeng Sun. 2026. Prediction and warning method for large passenger flow in metro transfer stations based on spatial and temporal characteristics of personnel trajectories. *Expert Systems with Applications* 296 (2026), 129193.

[5] Satvik Dixit, Laurie M Heller, and Chris Donahue. 2024. Vision language models are few-shot audio spectrogram classifiers. (2024). arXiv:2411.12058

[6] Harris Drucker, Christopher J Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. 1996. Support vector regression machines. *Advances in neural information processing systems* 9 (1996), 155–161.

[7] Mingjiang Duan, Tongya Zheng, Yang Gao, Gang Wang, Zunlei Feng, and Xinyu Wang. 2024. DGA-GNN: Dynamic grouping aggregation gnn for fraud detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 11820–11828.

[8] Qiang Gao, Zizheng Wang, Li Huang, Goce Trajcevski, Guisong Liu, and Xueqin Chen. 2025. Responsive dynamic graph disentanglement for metro flow forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 11690–11698.

[9] Yongshun Gong, Zhibin Li, Jian Zhang, Wei Liu, and Yu Zheng. 2022. Online spatio-temporal crowd flow distribution prediction for complex metro system. *IEEE Transactions on Knowledge and Data Engineering* 34, 2 (2022), 865–880.

[10] Sicheng He, Junzhong Ji, and Minglong Lei. 2025. Decomposed spatio-temporal mamba for long-term traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 11772–11780.

[11] Yuxin He, Lishuai Li, Xinting Zhu, and Kwok Leung Tsui. 2022. Multi-graph convolutional-recurrent neural network (MGC-RNN) for short-term forecasting of transit passenger flow. *IEEE Transactions on Intelligent Transportation Systems* 23, 10 (2022), 18155–18174.

[12] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[13] Yayao Hong, Liyue Chen, Leye Wang, Xiuhuai Xie, Guofech Luo, Cheng Wang, and Longbiao Chen. 2025. STKOpt: Automated spatio-temporal knowledge optimization for traffic prediction. In *Proceedings of the ACM on Web Conference 2025*. 2238–2249.

[14] Songhua Hu, Jianhua Chen, Wei Zhang, Guanhua Liu, and Ximing Chang. 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 (2024), 121095.

[15] Yizhou Huang, Yihua Cheng, and Kezhi Wang. 2025. Trajectory mamba: Efficient attention-mamba forecasting model based on selective ssm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 12058–12067.

[16] Jiawei Jiang, Chengkai Han, Wayne Xin Zhao, and Jingyuan Wang. 2023. PDFormer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 4365–4373.

[17] Di Jin, Cuiying Huo, Jiayi Shi, Dongxiao He, Jianguo Wei, and Philip S Yu. 2025. LLGformer: Learnable long-range graph transformer for traffic flow prediction. In *Proceedings of the ACM on Web Conference 2025*. 2860–2871.

[18] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2018. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. (2018). arXiv:1707.01926

[19] Zhihong Li, Xiaoyu Wang, Hua Cai, and Han Xu. 2024. Novel hybrid spatiotemporal convolution neural network model for short-term passenger flow prediction in a large-scale metro system. *Journal of Transportation Engineering, Part A: Systems* 150, 5 (2024), 04024016.

[20] Lingbo Liu, Jingwen Chen, Hefeng Wu, Jiajie Zhen, Guanbin Li, and Liang Lin. 2020. Physical-virtual collaboration modeling for intra-and inter-station metro ridership prediction. *IEEE Transactions on Intelligent Transportation Systems* 23, 4 (2020), 3377–3391.

[21] Qingxiang Liu, Sheng Sun, Yuxuan Liang, Min Liu, and Jingjing Xue. 2025. Personalized federated learning for spatio-temporal forecasting: A dual semantic alignment-based contrastive approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 12192–12200.

[22] Qinzhi Lv, Lijuan Liu, Ruotong Yang, and Yan Wang. 2025. Multimodal urban traffic flow prediction based on multi-scale time series imaging. *Pattern Recognition* 164 (2025), 111499.

[23] Jingchao Ni, Ziming Zhao, ChengAo Shen, Hanghang Tong, Dongjin Song, Wei Cheng, Dongsheng Luo, and Haifeng Chen. 2025. Harnessing vision models for time series analysis: A survey. (2025). arXiv:2502.08869

[24] Zhiqi Shao, Ze Wang, Xusheng Yao, Michael GH Bell, and Junbin Gao. 2025. ST-MambaSync: Complement the power of Mamba and Transformer fusion for less computational cost in spatial–temporal traffic forecasting. *Information Fusion* 117 (2025), 102872.

[25] Hongyuan Su, Yu Zheng, Jingtao Ding, Depeng Jin, and Yong Li. 2024. MetroGNN: Metro network expansion with reinforcement learning. In *Companion Proceedings of the ACM Web Conference 2024*. 650–653.

[26] Chaoqi Sun, Xiangmin Yang, Yongfeng Zhen, Yunhai Bai, and Lei Peng. 2024. Research on multimodal fusion indoor positioning under high-throughput passenger flow: A case study of metro station. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*. 2214–2220.

[27] Tao Wang, Jianjun Song, Jing Zhang, Junfang Tian, Jianjun Wu, and Jianfeng Zheng. 2025. Short-term metro passenger flow prediction based on hybrid spatiotemporal extraction and multi-feature fusion. *Tunnelling and Underground Space Technology* 159 (2025), 106491.

[28] Yichen Wang and Chengcheng Yu. 2024. CSP-AIT-Net: A contrastive learning-enhanced spatiotemporal graph attention framework for short-term metro OD flow prediction with asynchronous inflow tracking. (2024). arXiv:2412.01419

[29] Peng Xie, Minbo Ma, Tianrui Li, Shenggong Ji, Shengdong Du, Zeng Yu, and Junbo Zhang. 2023. Spatio-temporal dynamic graph relation learning for urban metro flow prediction. *IEEE Transactions on Knowledge and Data Engineering* 35, 10 (2023), 9973–9984.

[30] Mingxing Xu, Wenrui Dai, Chunmiao Liu, Xing Gao, Weiyao Lin, Guo-Jun Qi, and Hongkai Xiong. 2020. Spatial-temporal transformer networks for traffic flow forecasting. (2020). arXiv:2001.02908

[31] Rui Xue, Daniel Sun, and Shukai Chen. 2015. Short-term bus passenger demand prediction based on time series model and interactive multiple model approach. *Discrete Dynamics in Nature and Society* 2015, 1 (2015), 682390.

[32] Kun Yi, Qi Zhang, Wei Fan, Hui He, Liang Hu, Pengyang Wang, Ning An, Longbing Cao, and Zhendong Niu. 2023. FourierGNN: Rethinking multivariate time series forecasting from a pure graph perspective. *Advances in Neural Information Processing Systems* 36 (2023), 69638–69660.

[33] Xiaoming Yuan, Zhenyu Luo, Ning Zhang, Ge Guo, Lin Wang, Changle Li, and Dusit Niyato. 2025. Federated transfer learning for privacy-preserved cross-city traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems* 26, 4 (2025), 4418–4431.

[34] Chaolv Zeng, Zhanyu Liu, Guanjie Zheng, and Linghe Kong. 2024. CMamba: Channel correlation enhanced state space models for multivariate time series forecasting. (2024). arXiv:2406.05316

[35] Dongran Zhang, Jiangnan Yan, Kemal Polat, Adi Alhudhaif, and Jun Li. 2024. Multimodal joint prediction of traffic spatial-temporal data with graph sparse attention mechanism and bidirectional temporal convolutional network. *Advanced Engineering Informatics* 62 (2024), 102533.

[36] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31. 1655–1661.

[37] Lizong Zhang, Nawaf R Alharbe, Guangchun Luo, Zhiyuan Yao, and Ying Li. 2018. A hybrid forecasting framework based on support vector regression with a modified genetic algorithm and a random forest for traffic flow prediction. *Tsinghua Science and Technology* 23, 4 (2018), 479–492.

[38] Tianlong Zhang, Xiaoxi He, Yuxiang Wang, Yi Xu, Rendi Wu, Zhifei Wang, and Yongxin Tong. 2025. FedMetro: Efficient metro passenger flow prediction via federated graph learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25)*. Association for Computing Machinery, 5215–5224.

[39] Weiwen Zhang, Shuo Yang, and Yifeng Jiang. 2025. A two-phase client selection strategy for cost-optimal federated learning in traffic flow prediction. *IEEE Transactions on Consumer Electronics* 71, 2 (2025), 2955–2964.

[40] Yang Zhang, Yanling Chen, Ziliang Wang, and Dongrong Xin. 2023. TMFO-AGGRU: A graph convolutional gated recurrent network for metro passenger flow forecasting. *IEEE Transactions on Intelligent Transportation Systems* 25, 3 (2023), 2893–2907.

[41] Zhuangzhuang Zhao, Di Yang, Peng Wang, and Eryan Li. 2024. Metro passenger flow prediction: A double-stage decomposition combined with enhanced-BiGRU model considering multiple factors. In *Proceedings of the 4th International Conference on Artificial Intelligence and Computer Engineering (ICAICE '23)*. Association for Computing Machinery, 957–962.

[42] Li Zhu, Cheng Chen, Hongwei Wang, F. Richard Yu, and Tao Tang. 2024. Machine learning in urban rail transit systems: A survey. *IEEE Transactions on Intelligent Transportation Systems* 25, 3 (2024), 2182–2207.

A RELATED WORK

This section summarizes related work on metro flow prediction from three perspectives: multimodal learning for traffic flow prediction, deep learning-based metro flow prediction, and federated learning in traffic flow prediction.

A.1 Multimodal Learning for Traffic Flow Prediction

Metro flow prediction naturally extends to the broader field of traffic flow prediction, where data are inherently multimodal. Each modality, such as temporal, spatial, visual, and contextual, provides unique and complementary information about complex urban dynamics. Compared with unimodal methods, multimodal learning integrates heterogeneous signals, enhancing robustness, interpretability, and generalization across diverse scenarios [23]. Visual modalities demonstrate particular advantages due to their ability to encode high-level spatial correlations and contextual semantics. Unlike traditional tabular data, visual representations preserve structural continuity and latent relationships within traffic environments. For instance, VisionTS [3] converts temporal sequences into image-like matrices by segmenting and stacking them into heatmaps, allowing visual models to perceive periodic patterns. Similarly, VisionSTFT [5] transforms time series into spectrograms using short-time Fourier transform, enabling visual networks to capture frequency-domain dependencies that complement temporal features. In the domain of traffic flow prediction, CNN-based models extract local spatial patterns from grid-based traffic maps, while RNNs and GNNs capture temporal evolution and inter-station dependencies. For instance, DeepST [36] partitions cities into regular grids and employs CNN layers to learn spatial correlations among adjacent regions. DiffusionLSTM [18] combines LSTMs with graph convolutions to model both temporal variations and diffusion-based spatial dependencies. More recent frameworks adopt attention-based fusion strategies to align multimodal representations and emphasize the most informative components [2, 35]. For instance, MMGAT [35] integrates graph attention mechanisms for adaptive multimodal traffic prediction, capturing cross-modal spatiotemporal dependencies efficiently. Nevertheless, multimodal traffic flow prediction remains a challenging task. Discrepancies in data resolution, sampling frequency, and semantic meaning hinder modality alignment, and incorporating image-based representations often increases computational overhead in large-scale metro networks. Therefore, designing efficient and scalable frameworks capable of flexibly fusing temporal, spatial, and visual modalities while maintaining robustness continues to be an open research problem.

A.2 Deep Learning-Based Metro Flow Prediction

Compared with general traffic flow prediction, metro flow prediction exhibits unique characteristics. Metro routes and operation schedules are largely fixed, and passenger flow patterns show strong periodicity and spatial regularity. These characteristics constrain model flexibility while emphasizing structured dependencies among stations and time intervals. Deep learning significantly advances metro flow prediction by capturing nonlinear dependencies and long-term temporal correlations. RNNs and their variants such as LSTM and GRU model sequential patterns in metro flow data. For instance, At-STGCN-BiLSTM [19] combines graph convolution, bidirectional LSTM, and attention mechanisms to enhance short-term prediction performance in large-scale metro systems. GNNs capture spatial correlations and dynamic interactions among stations, leveraging the fixed non-Euclidean topology of metro networks. For instance, ReDyNet [8] adapts to variations in metro flow and external environments through a responsive mechanism to construct an appropriate dynamic graph. Transformer-based architectures model long-range dependencies and extract global patterns through attention mechanisms. CSP-AIT-Net [28] applies spatiotemporal graph attention with asynchronous inflow tracking and semantic embeddings for origin–destination flow prediction, demonstrating the effectiveness of attention for metro OD prediction. Emerging architectures such as Mamba provide linear computational complexity and strong local feature extraction capability for long sequence modeling. For instance, ST-MambaSync [24] embeds Mamba blocks into a spatial–temporal synchronization network, capturing lane-level speed surges within long multi-sensor sequences and enabling minute-by-minute traffic prediction for urban expressways. Despite these advances, metro-specific challenges remain. Fixed routes and timetables impose rigid spatial-temporal patterns, amplifying the impact of peak-hour fluctuations and service disruptions. The strong periodicity and dense station topology increase the difficulty of generalizing across days or lines. Furthermore, real-time metro operation requires models that balance global dependency modeling and computational efficiency. Achieving such a balance between prediction accuracy, scalability, and responsiveness continues to be an open research problem.

A.3 Federated Learning in Traffic Flow Prediction

Traffic flow prediction often faces challenges related to privacy, data heterogeneity, and decentralization. Raw data cannot always be shared across regions or cities due to privacy concerns, and centralized modeling may lead to biased or insecure predictions. FL addresses these issues by allowing local model training at each station or client while aggregating updates at a central server, thus mitigating data silos and privacy leakage [38]. Prior studies demonstrate the effectiveness of FL in traffic flow prediction while preserving privacy. For instance, FedTPCS [39] introduces a two-phase client selection strategy to minimize energy consumption and latency by addressing device and data heterogeneity. FUELS [21] employs dual semantic alignment-based contrastive learning to capture spatio-temporal heterogeneity. Additionally, 2MGTCN [33] combines graph convolutional networks and temporal convolutional networks with federated transfer learning to enhance cross-city prediction accuracy and privacy preservation. Nevertheless, designing FL frameworks that handle station heterogeneity, real-time metro flow fluctuations, and privacy-preserving multi-station collaboration remains challenging.

Despite significant advances in multimodal learning, deep learning-based metro flow prediction, and federated learning, several overarching challenges remain. Efficiently integrating heterogeneous modalities while preserving spatial-temporal dependencies, capturing both global patterns and local dynamics in highly periodic and dynamic metro networks, and ensuring privacy-preserving, fair, and scalable collaborative learning across decentralized stations continue to pose significant difficulties for practical metro flow prediction.

B ALGORITHM PSEUDOCODE AND COMPLEXITY ANALYSIS

The MF^3 algorithm framework integrates multimodal data processing with privacy-preserving FL for metro flow prediction. As shown in Algorithm 1, the framework operates through iterative client-server interactions while maintaining data privacy through local model training.

Time complexity analysis. Let N be the number of clients, E the local epochs, B the batch size, T the sequence length, and S the number of stations. The computational complexity is primarily determined by four components: the MA module requires $O(T \cdot S \cdot d^2)$ operations for cross-modal attention mechanisms; the DMT module exhibits $O(T \cdot S \cdot d + T \cdot d^2)$ complexity by combining the linear scaling of Mamba with attention mechanisms; client training dominates the overall cost with $O(N \times E \times (|D_i|/B) \times (T \cdot S \cdot d^2))$ where d represents the model dimension and $|D_i|$ denotes client i's data volume; and the BFR module contributes $O(N^2 \cdot |\Theta|)$ for similarity computations and model averaging during server aggregation. The framework's design ensures scalability for large-scale metro networks while maintaining the privacy guarantees of FL through efficient distributed computation.

C EXPERIMENT

This section provides supplementary details for the experiments presented in the main text.

C.1 Evaluation Metrics

To quantitatively assess the prediction performance of different models, three commonly used evaluation metrics are employed: MAE, RMSE, and MAPE. Their mathematical formulations are provided below:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|, \tag{38}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}, \tag{39}$$

$$\mathrm{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i + \epsilon}\right|, \tag{40}$$

where y_i and $\hat{y}_i$ represent the ground truth and predicted values, respectively, N denotes the total number of samples, and ϵ is a small constant to prevent division by zero. Note that for all three metrics, lower values indicate better prediction performance, as they reflect smaller deviations between the predicted and true values.

C.2 Baseline Methods

To ensure a comprehensive and fair evaluation, the baseline models are grouped into classical methods, graph-based models, and transformer-based or advanced spatiotemporal deep learning methods. These methods cover diverse paradigms such as traditional regression, deep spatiotemporal learning, multimodal fusion, and graph-based modeling, with the majority of baselines proposed in the past three years, reflecting recent advances in metro flow prediction. The main baselines and their representative significance are summarized as follows:

Algorithm 1 The MF^3 framework for metro flow prediction

```
Require: Historical metro data $\mathbf{X}$, spatial graph $\mathbf{A}$, heatmaps $\mathbf{H}$
Ensure: Predicted passenger flows $\hat{\mathbf{Y}}$
Initialize: Global model $\Theta_G$, client reputations $\mathcal{R}$
for round = 1 to $R$ do
    // Client Local Training
    for each client $c_i$ in parallel do
        Load local data $\mathcal{D}_i$
        Initialize local model $\Theta_i \leftarrow \Theta_G$
        for epoch = 1 to $E$ do
            Process data through the MA and the DMT modules
            Compute composite loss $\mathcal{L}_{\text{total}}$
            Update $\Theta_i$ with gradient clipping
        end for
        Compute validation loss $\mathcal{L}_i^{\text{val}}$
        Upload $\langle \Theta_i, \mathcal{L}_i^{\text{val}} \rangle$ to server
    end for
    // Server Aggregation
    for each client $i$ do
        Update reputation: $r_i \leftarrow \alpha \cdot r_i + (1 - \alpha) \cdot \phi(\mathcal{L}_i^{\text{val}})$
    end for
    Compute aggregation weights based on reputation and performance
    Identify similar clients for each target client
    for each target client $t$ do
        Compute personalized weights $w_{i \to t}$
        Aggregate: $\Theta_t^{\text{global}} \leftarrow \sum w_{i \to t} \cdot \Theta_i$
        Adjust personalization coefficient $\alpha_t$
        Blend: $\Theta_t^{\text{final}} \leftarrow (1 - \alpha_t)\Theta_t^{\text{global}} + \alpha_t \Theta_t^{\text{local}}$
    end for
    Update global model $\Theta_G$
    Distribute updated models to clients
end for
// Prediction Phase
function PREDICT(historical data, current time $T$)
    Preprocess: Normalize flows, extract features, generate heatmaps
    Multimodal Alignment: Fuse temporal, spatial, visual data
    DMT Processing:
        Temporal path: Fine-grained + periodic patterns
        Spatial path: Convolutional feature extraction
        Cross-attention fusion between pathways
    Generate predictions $\hat{\mathbf{Y}}$ for $T + 1$ to $T + p$
    return $\hat{\mathbf{Y}}$
end function
Output: Trained models for all clients, prediction function
```

Classical methods:

LSTM [12]: A classical recurrent neural network capable of capturing long-term dependencies in historical flow data, widely used in traffic and metro flow prediction. LSTM serves as a baseline to evaluate the benefits of more advanced graph or transformer-based architectures over standard sequence modeling.

SVR [6]: A traditional regression method suitable for small-sample or nonlinear time series prediction. SVR represents classical statistical approaches, providing a baseline for assessing the value added by deep learning and multimodal fusion techniques.

Graph-based models:

ReDyNet [8]: A responsive dynamic graph neural network designed for metro flow. ReDyNet constructs dynamic graphs and disentangles redundant signals to achieve accurate spatiotemporal modeling, representing state-of-the-art graph-based approaches for metro networks with complex station interactions.

FourierGNN [32]: Reformulates multivariate time series as hypervariate graphs and applies graph convolutions in Fourier space to efficiently capture spatiotemporal dynamics. FourierGNN exemplifies spectral-domain GNN methods for high-dimensional traffic data modeling.

STDGRL [29]: Utilizes spatiotemporal node embeddings and dynamic graph relationship learning combined with a transformer module for long-term flow prediction, highlighting the integration of graph learning with sequential modeling.

AGCRN [1]: Incorporates node-adaptive and data-adaptive modules to automatically capture fine-grained spatiotemporal correlations without requiring a predefined graph, representing flexible GNNs for heterogeneous urban traffic data.

S-DGNN [4]: Models metro transfer stations as nodes in a directed graph, leveraging temporal and spatial correlations with GRU to predict short-term transfer passenger flow, demonstrating practical utility for operational decision support.

Transformer-based and advanced spatiotemporal models:

PDFormer [16]: A propagation delay-aware long-range transformer that captures short- and long-range spatial dependencies while modeling the time delay of information propagation, illustrating the advantage of transformers in handling long-range spatiotemporal patterns.

STNNs [30]: Employs spatial and temporal transformers with multi-head attention to jointly model directed spatial dependencies and long-range temporal dependencies, providing a representative baseline for scalable long-term prediction.

LLGformer [17]: Constructs a learnable long-range spatiotemporal graph and applies attention mechanisms to capture complex patterns across time and space, improving traffic and metro flow prediction efficiency. This method is representative of recent advances in combining graph structures with attention mechanisms for metro flow prediction.

C.3 Performance Analysis

Figure 5 shows the training and validation loss curves over training rounds for different time intervals in Hangzhou.

It can be seen that the model continuously converges on the training set, and the 60-minute interval exhibits more stable validation performance, indicating that coarser-grained predictions achieve more reliable generalization.

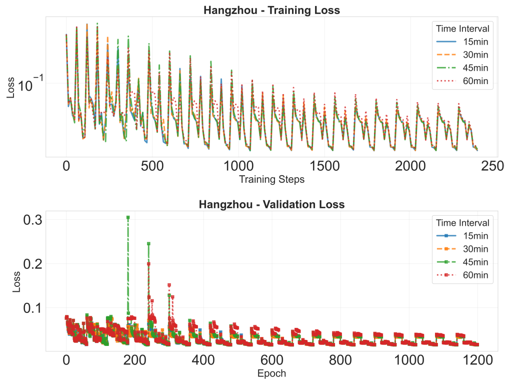

Figure 5: Hangzhou metro loss curves.

Figure 6 presents the client performance distribution using boxplots. In the figure, the box represents the interquartile range (middle 50% of clients), the line inside the box denotes the median, the whiskers indicate the normal range of the data, and the outliers reflect clients with exceptional performance.

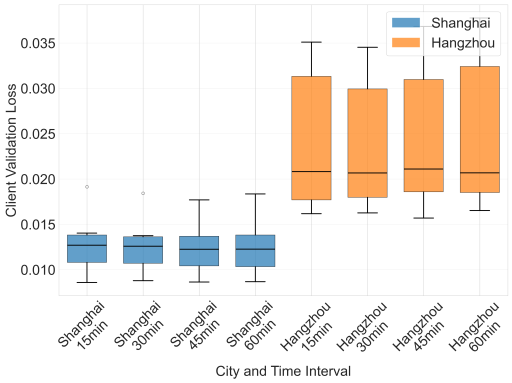

Figure 6: The client performance under MF^3 in FL.

It can be observed that client performance is generally stable across different configurations, with medians concentrated, indicating good fairness of MF^3 in FL and limited performance variation among clients.