Learning Fine-Grained Urban Mobility Dynamics through Large Model-Enhanced Multimodal Representations

Tianhong Zhao[a], Jianbin Li[a], Jinzhou Cao[a,*], Wei Tu[b], Filip Biljecki[c], Shengao Yi[d], Zhilu Yuan[b]

[a]*Spatio-temporal Intelligence Center & School of Artificial Intelligence, Shenzhen Technology University, Shenzhen 518118, China*
[b]*School of Architecture and Urban Planning, Shenzhen University, Shenzhen 518060, China*
[c]*Department of Architecture, National University of Singapore, Singapore 117356*
[d]*Department of City and Regional Planning, University of Pennsylvania, Philadelphia, PA 19104 USA*

Abstract

Accurately predicting fine-grained urban mobility is essential for optimizing transportation, accessibility, and urban management. However, existing approaches often depend on dynamic data such as trajectories or signaling records, which are sparsely available across cities, thereby limiting their applicability and generalizability to new urban contexts. To address these limitations, this study proposes a Large Model Enhanced Multimodal Representations (LMEMR) framework to learn hourly grid-level mobility dynamics solely from static geospatial data—including remote sensing imagery, building data, street view imagery, and points of interest—which are widely accessible. Large vision–language models are employed to generate natural-language descriptions of each modality, enriching the data with human-understandable semantics. A dual-level contrastive learning strategy aligns raw and textual features both within and across modalities, mitigating semantic gaps and enhancing multimodal consistency. Spatial dependencies are modeled through a graph attention network, and temporal dynamics are captured via a transformer encoder to produce 24-hour mobility sequences.

*Corresponding author
Email address: caojinzhou@sztu.edu.cn (Jinzhou Cao)

Results from Shenzhen demonstrate that LMEMR outperforms the baseline CLIP model, achieving an R^2 of 0.856 and an 18.07% reduction in MAE. Ablation experiments confirm the effectiveness of semantic enhancement, spatial graph reasoning, and cross-modal fusion. Overall, this research reveals the potential of static multimodal data for dynamic mobility inference, offering a scalable, interpretable, and privacy-friendly solution for smart city planning and management.

Keywords: Urban mobility prediction, Multimodal learning, Vision–language models, Contrastive learning, Graph attention networks

1. Introduction

Cities are inherently dynamic systems shaped by continuous human movement, interaction, and adaptation. This dynamic nature is most directly reflected in urban mobility patterns, which describe how people travel through the city throughout the day. In this study, urban mobility patterns are defined as the temporal sequence of outbound flows within grids over a 24-hour period [1, 2]. Understanding and predicting these patterns is fundamental for effective transportation management, equitable allocation of public services, and resilient urban planning [3]. Accurate modeling of urban mobility provides essential information for optimizing transit operations, mitigating congestion, and designing cities that can adapt to daily fluctuations and long-term changes in human activity [4].

Despite extensive research on urban activity modeling, most existing approaches remain limited to activity prediction or functional classification, without systematically characterizing the fine-grained temporal evolution of urban mobility [5, 6, 7]. Large-scale geographic datasets commonly available for urban analysis—such as remote sensing imagery (RSI), street view imagery (SVI), building data, and points of interest (POI)—are essentially static. In contrast to trajectory or signaling data that track real-time urban mobility dynamics, static spatial data provide distinct benefits in terms of wide availability, spatial coverage, and cross-temporal transferability [8]. These datasets record the physical form and functional attributes of the built environment, but do not directly capture the temporal dynamics of residents' travel demand [9]. To bridge this gap, this study aims to develop a framework that integrates multi-source static geographic data—which is widely accessible across cities—to accurately infer hourly, grid-level outbound flows.

Previous research has shown that static geographic data have the potential to reveal certain dynamic patterns of human activity. For example, POI data, which reflect the spatial distribution of urban functional areas, have been widely used to predict human mobility and regional activity levels [10]. Jiang et al. proposed a transfer learning method based on POI embeddings and deep learning, which achieved significant improvements in the prediction of human mobility in cities [11]. Meanwhile, RSI provides objective information on the physical environment, such as building density, green space, and road networks. These features have the potential to infer human activity patterns from urban morphology [12, 13]. Complementarily, SVI provide finer-grained representations of the city from a human perspective [14]. SVI captures details such as amenities, building façades, and cultural styles, revealing urban functions and reflecting mobility patterns on streets with similar functions [15, 16]. Zhang et al. showed that combining high-level semantic features from SVI via deep convolutional neural networks with taxi mobility data can effectively forecast hourly taxi demand [17].

Nevertheless, three major challenges remain to be addressed. *First, semantic representation is insufficient.* Previous methods mainly depend on low-level features, such as pixel values, lacking high-level semantic understanding of regional functions and behavioral implications. For example, high-density built-up areas can correspond to residential neighborhoods or industrial parks, exhibiting drastically different mobility patterns—distinctions that raw visual or structural features cannot capture [18]. *Second, fusing heterogeneous modalities remains challenging.* RSI, SVI, building, and POI data differ fundamentally in form and semantics. A naive concatenation method often introduces noise and semantic shifts, affecting the robustness and generalization of the model [19, 20]. *Third, spatial neighborhood dependencies are often underexplored.* Urban mobility patterns are influenced not only by the intrinsic attributes of a grid but also by the functions and transport facilities of its surrounding context. For example, residential grids adjacent to metro stations experience substantial demand surges during morning peaks—yet such non-local dependencies are difficult to capture without explicit spatial graph structures [21].

To address these challenges, we propose the *Large Model Enhanced Multimodal Representations (LMEMR)* framework, which transforms static multimodal geospatial data into dynamic representations of human mobility. LMEMR consists of three key modules that correspond to the identified challenges: (i) To enhance semantic representation, a *multimodal semantic*

enhancement module leverages large vision–language models (VLMs/LLMs) to enrich the high-level semantics of RSI, building, SVI, and POI data; (*ii*) To improve multimodal fusion, a *cross-modal fusion* module aligns modality-specific embeddings through contrastive learning and attention-based integration for hourly mobility prediction; and (*iii*) To capture spatial dependencies, a *spatial-context representation learning* module employs a graph attention network (GAT) to model local and non-local neighborhood interactions. By jointly encoding semantic, spatial and multimodal features, LMEMR enables an interpretable and accurate inference of fine-grained urban mobility patterns.

2. Related Work

2.1. Urban Mobility Pattern Prediction

Driven by urban big data, urban mobility prediction has evolved from aggregate estimation to fine-grained spatiotemporal forecasting [22, 23]. Existing research generally falls into three categories: short-term demand, origin–destination (OD) flow, and mobility pattern prediction. Short-term demand prediction focuses on immediate variations, where CNN–LSTM architectures have proven effective in capturing nonlinear dependencies for applications like metro and ride-hailing forecasting [24, 25]. OD flow prediction models dynamic interzonal interactions. While traditional gravity models laid the foundation, recent graph-based and attention-driven networks have significantly improved large-scale estimation [26, 27].

Mobility pattern prediction aims to reconstruct fine-grained temporal rhythms. Arenas et al. identified distinct patterns for functional zones using clustering [28], while recent works employ Transformer and GCNs to model spatiotemporal dependencies [29]. To explain behavioral differences, semantic data (POIs, SVI) have been integrated via knowledge graphs [30] or geographic-semantic GCNs [31]. Recent advances also explore LLM-powered semantic synthesis for dynamic mobility prediction [32] and deep learning approaches for predicting mobility flows directly from satellite imagery [33]. This research line shifts focus from aggregate demand to understanding structural mobility regularities linking the built environment with human dynamics; however, most existing methods still rely on dynamic data, leaving static-only mobility inference largely unexplored.

2.2. Multimodal Fusion for Geospatial Data

Multimodal fusion of geospatial data (RSI, Building, SVI, POI) has become central to urban computing [34, 35]. Research has shifted from single-modal classification [36] to integrating multiple sources for tasks like functional zone identification [37]. Recent advances in GNNs and contrastive learning facilitate joint modeling of heterogeneous data, enhancing regional representations [38] and mitigating label scarcity [39]. Current progress focuses on cross-modal semantic alignment [40], multi-level spatial graph modeling [41], and contrastive strategies for consistency [42].

Existing fusion methods are generally categorized as feature-based, alignment-based, or contrast-based. Feature-based concatenation often lacks semantic alignment and generalization [43]. Alignment-based methods map modalities to shared spaces but can be sensitive to noise [44]. Contrast-based frameworks explicitly enforce cross-modal consistency, offering robust generalization across heterogeneous data [45, 46]. Recent stochastic multimodal fusion approaches [47] and multi-scale contrastive learning frameworks [48] further enhance robust spatial representations across diverse urban tasks. These efforts mark a shift toward semantically aligned and structurally informed integration, yet few adopt a dual-level (intra- and inter-modal) contrastive strategy.

2.3. Semantic Representation of Geospatial Data

Despite data growth, urban models struggle with the semantic gap—the disconnect between low-level features and high-level human concepts [49, 50]. Traditional representations (e.g., spectral indices) lack behavioral meaning [51]. While POI embeddings [52] and land-use classification [53] offer partial solutions, they are limited by fixed taxonomies and lack contextual flexibility. Consequently, models often describe where activities occur without explaining why.

To bridge this gap, recent studies integrate natural-language priors. Large multimodal models (e.g., CLIP, Qwen-VL) enable vision–language alignment, reshaping spatial representation [54, 55, 56]. This "GeoAI + LLM" trend facilitates tasks like image annotation and knowledge extraction [57]. Benchmarks for evaluating VLMs on urban scene perception [58] reveal that models achieve stronger alignment on objective properties than subjective appraisals. Contrastive learning further aligns embeddings in a unified semantic space [59]. Building on these developments, our work employs VLMs

to generate textual descriptions from geospatial inputs and aligns them with spatial features via contrastive learning for mobility modeling.

3. Methodology

3.1. Problem Formulation and Notations

This study addresses the task of inferring fine-grained urban mobility dynamics from static geospatial data. Let $\mathcal{G} = \{g_1, g_2, \ldots, g_N\}$ denote a set of N non-overlapping urban grids partitioning the study area, where each grid corresponds to a spatial unit of size $500\,\mathrm{m} \times 500\,\mathrm{m}$. For each grid g_i, the input comprises multimodal static geospatial data from four modalities:

$$\mathcal{X}_i = \left\{ \mathbf{x}_i^{\mathrm{rsi}}, \mathbf{x}_i^{\mathrm{bld}}, \mathbf{x}_i^{\mathrm{svi}}, \mathbf{x}_i^{\mathrm{poi}} \right\}, \tag{1}$$

where $\mathbf{x}_i^{\mathrm{rsi}}$, $\mathbf{x}_i^{\mathrm{bld}}$, $\mathbf{x}_i^{\mathrm{svi}}$, and $\mathbf{x}_i^{\mathrm{poi}}$ denote remote sensing imagery, building height maps, street view images, and POI feature vectors, respectively. The model predicts a 24-hour mobility sequence $\hat{\mathbf{y}}_i = [\hat{y}_{i,1}, \hat{y}_{i,2}, \ldots, \hat{y}_{i,24}] \in \mathbb{R}^{24}$ for each grid, where $\hat{y}_{i,t}$ denotes the predicted outbound flow volume at hour t. Given a training set $\{(\mathcal{X}_i, \mathbf{y}_i)\}_{i=1}^{N}$, the model learns a mapping function $f_\theta : \mathcal{X} \to \hat{\mathbf{y}}$ by minimizing the prediction error:

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{t=1}^{24} \ell\left(\hat{y}_{i,t}, y_{i,t}\right), \tag{2}$$

where $\ell(\cdot, \cdot)$ denotes a suitable loss function.

Table 1 summarizes the key notations used throughout the methodology.

3.2. Overview of LMEMR

The proposed LMEMR establishes a framework that transforms static multimodal geospatial data into dynamic representations of human mobility, as shown in Fig. 1. The LMEMR comprises three key components. (*i*) The *Multimodal Semantic Enhancement* module uses VLMs/LLMs to enhance the semantic representation of RSI, building, SVI, and POI data. (*ii*) The *Graph-Enhanced Spatial-Context Learning* module models spatial relationships through GATs to capture both local and non-local neighborhood dependencies. (*iii*) The *Cross-Modal Fusion and Temporal Modeling* module integrates multimodal features via attention-based fusion and predicts 24-hour grid-level mobility sequences using a Transformer encoder. By

Table 1: Summary of main notations.

Symbol	Definition
$\mathcal{M} = \{\text{rsi}, \text{bld}, \text{svi}, \text{poi}\}$	Set of data modalities.
$\mathbf{x}_i^m, s_i^m$	Raw input and textual description for grid i and modality m.
$E_{\text{v}}^m(\cdot), E_{\text{t}}^m(\cdot)$	Spatial and semantic encoders for modality m.
$\mathbf{H}^m, \widetilde{\mathbf{H}}^m$	Intra-modally aligned and graph-enhanced feature matrices for modality m.
$\mathbf{z}_i$	Cross-modally fused embedding for grid i.
$\hat{y}_{i,t}$	Predicted outbound flow for grid i at hour t.
$\mathcal{L}_{\text{cla}}, \mathcal{L}_{\text{reg}}, \mathcal{L}_{\text{CL}}$	Classification, regression, and contrastive losses.

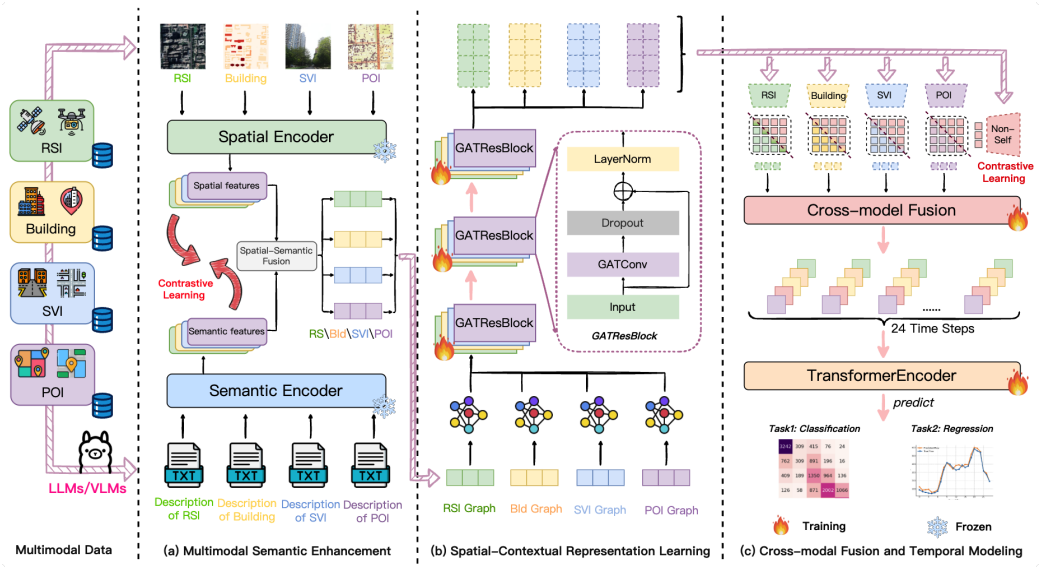

Figure 1: Overview of the proposed LMEMR. The framework consists of three modules: (a) multimodal semantic enhancement, (b) spatial–contextual representation learning, and (c) cross-modal fusion with temporal modeling.

combining semantic alignment, graph-based spatial reasoning, and temporal sequence modeling, LMEMR enables interpretable high-fidelity inference of fine-grained urban mobility dynamics from static multimodal data.

3.3. Multimodal Semantic Enhancement

This module enriches the semantic expressiveness of static geospatial data by aligning raw spatial features with text-derived semantics. The process consists of two stages: (*i*) *Semantic enrichment*, utilizing VLMs to generate descriptive text for each modality; and (*ii*) *Intra-modal alignment*, bridging the gap between spatial and semantic representations to capture behavior-related urban patterns.

3.3.1. Semantic Enrichment of Multimodal Geospatial Data

As shown in Fig. 2, each 500 m × 500 m grid is characterized by four behaviorally informative modalities. *RSI* captures large-scale morphological cues (e.g., layouts, vegetation), while *Building* height maps reflect vertical urban intensity and trip-generation potential. *SVI* offers human-scale visual attributes, sampled via 20 random images per grid. For grids with sparse road networks containing fewer than 20 available images, oversampling with replacement is applied to maintain consistent input dimensions while preserving realistic visual features. *POI* vectors describe functional composition and accessibility. Although rich in structural data, these modalities lack explicit semantic abstraction, limiting the model's ability to interpret high-level concepts like "commercial corridors."

To address this, we leverage VLMs to inject interpretable semantic priors into the spatial representation [60]. Specifically, Qwen-VL processes visual inputs (RSI, Building, SVI) for its strong visual understanding, while DeepSeek-R1 handles structured POI data via chain-of-thought reasoning that first analyzes category distributions and then synthesizes functional descriptions. These models serve as plug-and-play components that can be replaced by alternatives without architectural changes (see Table 8 for a systematic comparison). A unified three-level instruction prompt guides the generation of multi-dimensional urban semantics: *(i) land features and functional zoning; (ii) estimated activity intensity (0–100); and (iii) predicted 24-hour departure trends (0–1).* The resulting descriptions are encoded via Qwen3-Embedding for subsequent alignment.

3.3.2. Intra-modal Representation Alignment

This phase aligns VLM-derived semantics with original spatial features to ensure consistency. It proceeds through dual-branch encoding, contrastive alignment, and feature fusion.

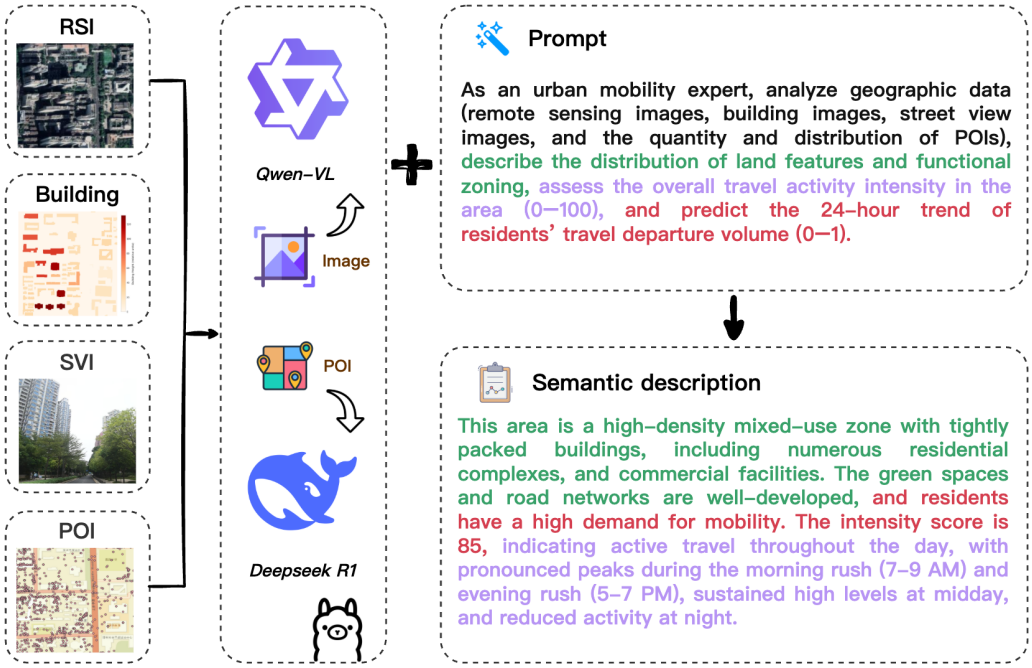

Figure 2: Illustration of text generation using LLMs/VLMs. Multiple models are employed to generate natural-language descriptions from multimodal geospatial data inputs.

Dual-branch Encoding. For each modality $m \in \{\text{rsi}, \text{bld}, \text{svi}, \text{poi}\}$, we extract parallel representations from the spatial input $\mathbf{x}_i^m$ and textual description s_i^m. The spatial encoder $E_{\text{v}}^m(\cdot)$ maps $\mathbf{x}_i^m$ to latent vector $\mathbf{f}_i^m \in \mathbb{R}^{d_f}$. Specifically, we employ a pre-trained ResNet-152 [61] for image-based modalities and an MLP for POI vectors. Simultaneously, the semantic encoder $E_{\text{t}}^m(\cdot)$ (Qwen3-Embedding [62]) processes s_i^m to produce $\mathbf{g}_i^m \in \mathbb{R}^{d_f}$. The pair $(\mathbf{f}_i^m, \mathbf{g}_i^m)$ thus encodes the grid from both physical and semantic perspectives.

Intra-modal Contrastive Alignment. To unify these branches, we employ a bidirectional contrastive framework [63] (Fig. 3). For all N grids, the pair $(\mathbf{f}_i^m, \mathbf{g}_i^m)$ constitutes a positive sample, while mismatched pairs $(\mathbf{f}_i^m, \mathbf{g}_j^m)_{i \neq j}$ serve as negatives. We define directional similarity distributions as:

$$
\begin{aligned}
p_{i \to j}^m &= \frac{\exp\big(\cos\big(\mathbf{f}_i^m,\, \mathbf{g}_j^m\big)/\tau\big)}{\sum_{k=1}^{N} \exp\big(\cos\big(\mathbf{f}_i^m,\, \mathbf{g}_k^m\big)/\tau\big)}, \\
q_{i \to j}^m &= \frac{\exp\big(\cos\big(\mathbf{g}_i^m,\, \mathbf{f}_j^m\big)/\tau\big)}{\sum_{k=1}^{N} \exp\big(\cos\big(\mathbf{g}_i^m,\, \mathbf{f}_k^m\big)/\tau\big)},
\end{aligned}
\tag{3}
$$

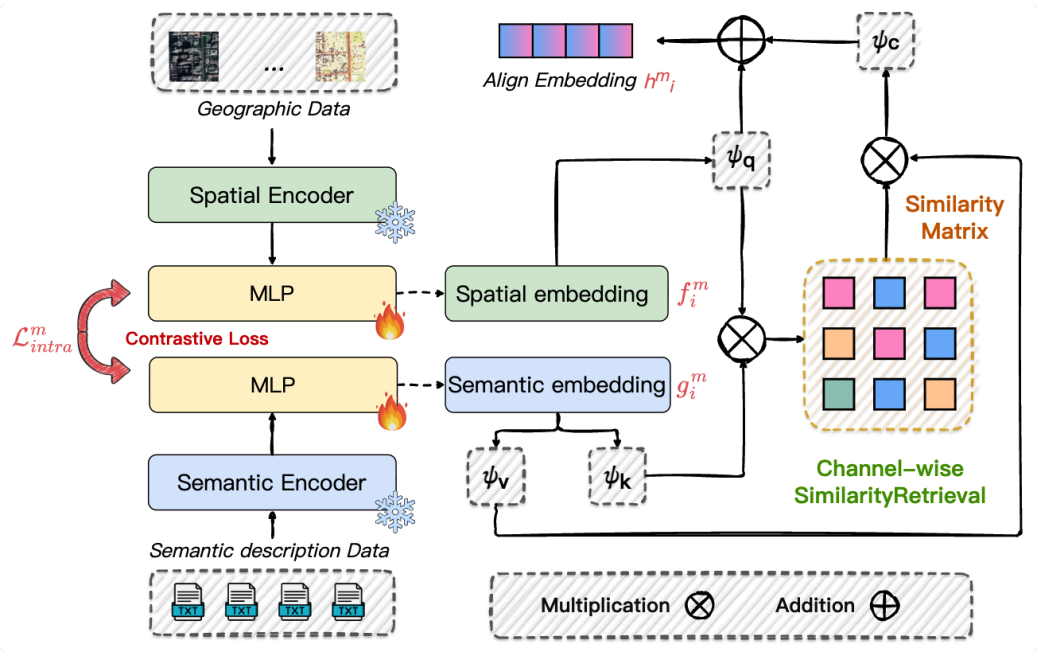

Figure 3: Architecture of the intra-modal contrastive alignment and feature fusion module.

where τ is the temperature coefficient. The bidirectional InfoNCE loss enforces mutual alignment:

$$\mathcal{L}^m_{\text{intra}} = -\frac{1}{2N}\sum_{i=1}^{N}\Big[\log p^m_{i\to i} + \log q^m_{i\to i}\Big]. \tag{4}$$

This optimization creates a coherent latent space where spatial and textual features mutually reinforce each other.

Intra-modal Feature Fusion. Post-alignment, we integrate the branches using channel-wise similarity retrieval [60]. Let $\mathbf{F}^m$ and $\mathbf{G}^m$ denote the stacked spatial and semantic embeddings. We project these into subspaces via ψ_q, ψ_k, ψ_v to compute a similarity matrix:

$$\mathbf{S}^m = \text{rowsoftmax}\big((\psi_q(\mathbf{F}^m))^\top \psi_k(\mathbf{G}^m)\big)\,. \tag{5}$$

The aligned representation $\mathbf{H}^m$ is obtained via residual fusion:

$$\mathbf{H}^m = \omega^c(\psi_v(\mathbf{G}^m)\,\mathbf{S}^m) \oplus \mathbf{F}^m, \tag{6}$$

where ω^c is a learnable transformation. The resulting $\mathbf{H}^m$ preserves geometric structure while incorporating high-level semantics, preparing robust inputs for cross-modal fusion.

3.4. Spatial-Contextual Representation Learning

Following intra-modal alignment, we address the spatial dependencies inherent in urban environments. Nearby grids often exhibit strong correlations in morphology and function due to the continuity of the built environment and socioeconomic interactions [64]. To capture these contextual relationships, we employ a graph-based neural network to model spatial dependencies among grids, as illustrated in Fig. 1. The process consists of two key stages: (*i*) *Graph construction and initialization*, which establishes a shared spatial graph topology with modality-specific node features; and (*ii*) *Attention-based Spatial Aggregation*, which leverages GAT to adaptively propagate and aggregate contextual features from local neighborhoods.

3.4.1. Graph construction and initialization

To explicitly model such spatial relations, we construct a spatial graph $\mathcal{G}^m = (\mathcal{V}^m, \mathcal{E}^m)$ for each data modality $m \in \{\text{rsi}, \text{bld}, \text{svi}, \text{poi}\}$, sharing the same adjacency topology across modalities while initializing nodes with modality-specific features. Each node $v_i^m \in \mathcal{V}^m$ corresponds to an urban grid, and each edge $(v_i^m, v_j^m) \in \mathcal{E}^m$ encodes the spatial adjacency between two grids. The node v_i^m is initialized with the intra-modally aligned embedding $\mathbf{h}_i^m$ obtained from the previous stage. The node feature matrix for the modality m is expressed as

$$\mathbf{H}^{(0,m)} = [\mathbf{h}_1^m, \mathbf{h}_2^m, \ldots, \mathbf{h}_N^m]^\top \in \mathbb{R}^{N \times d_f}, \tag{7}$$

where N is the number of urban grids and d_f is the embedding dimension. Spatial relationships are encoded in an adjacency matrix $\mathbf{A}^m \in \mathbb{R}^{N \times N}$, where $A_{ij}^m = 1$ if the grids i and j are spatially adjacent (e.g., share a boundary) and $A_{ij}^m = 0$ otherwise. This shared graph topology, combined with modality-specific node features, provides the foundation for capturing spatial context and non-local dependencies in the subsequent graph-based representation learning. We denote the updated graph-enhanced representation as $\widetilde{\mathbf{H}}^m = [\widetilde{\mathbf{h}}_1^m, \widetilde{\mathbf{h}}_2^m, \ldots, \widetilde{\mathbf{h}}_N^m]^\top$.

3.4.2. Attention-based Spatial Aggregation

For each modality, we adopt a GAT [65] to adaptively propagate contextual information among neighboring grids. As shown in Fig. 1, three

stacked *GATResBlocks* are employed to deepen the receptive field and capture multi-scale spatial dependencies. At the l-th layer, the hidden state is first normalized and then passed through a graph attention operator:

$$\mathbf{h}_i^{(l+1,m)} = \mathrm{GELU}\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^m \, \mathbf{W}^m \mathbf{h}_j^{(l,m)}\right), \tag{8}$$

where $\mathcal{N}(i) = \{j \mid A_{ij}^m = 1\}$, $\mathbf{W}^m$ is a learnable weight matrix, and the attention coefficients α_{ij}^m are normalized using a softmax function:

$$\begin{aligned}\alpha_{ij}^m = \mathrm{softmax}_{j\in\mathcal{N}(i)}\Big(&\mathrm{LeakyReLU}\big((\mathbf{a}^m)^\top[\mathbf{W}^m\mathbf{h}_i^{(l,m)}\| \\ &\mathbf{W}^m\mathbf{h}_j^{(l,m)}]\big)\Big),\end{aligned} \tag{9}$$

with $\mathbf{a}^m$ being a learnable attention vector and $\|$ denoting the concatenation of features. Dropout and residual connections are applied to stabilize training and enhance convergence.

$$\widetilde{\mathbf{h}}_i^{(m)} = \mathbf{h}_i^{(l,m)} + \mathrm{Dropout}(\mathbf{h}_i^{(l+1,m)}). \tag{10}$$

This attention-based residual design (denoted as *GATResBlock*) allows the network to dynamically assign importance weights to neighboring nodes, effectively capturing multi-scale spatial dependencies across different modalities.

3.5. Cross-modal Fusion and Temporal Modeling

After obtaining semantically aligned and spatially contextualized embeddings for each modality, the next step is to perform cross-modal fusion and learn the temporal dynamics of human mobility. This stage aims to (i) achieve inter-modal semantic consistency through contrastive learning, (ii) integrate complementary multimodal features via attention-based fusion, and (iii) model 24-hour temporal variations using a Transformer-based sequence encoder.

3.5.1. Inter-modal Contrastive Learning

Although intra-modal alignment ensures semantic consistency within each individual modality, the embeddings of different modalities (e.g., RSI, building, SVI, and POI) may still reside in heterogeneous latent spaces. To unify

these representations, we design an *inter-modal contrastive learning* framework that encourages cross-modal correspondence among semantically related regions.

Specifically, for any pair of modalities (m, n), we regard the graph-enhanced embeddings $\widetilde{\mathbf{h}}_i^m$ and $\widetilde{\mathbf{h}}_i^n$ corresponding to the same spatial location i as a *positive pair*, while embeddings from different locations $(\widetilde{\mathbf{h}}_i^m, \widetilde{\mathbf{h}}_j^n)$ $(i \neq j)$ serve as *negative pairs*. Following the contrastive paradigm introduced in Section 3.3.2, we compute pairwise similarities between modality-specific embeddings using cosine similarity and apply softmax normalization across all N grids. The bidirectional InfoNCE objective is then formulated as:

$$\mathcal{L}_{\text{inter}}^{(m,n)} = -\frac{1}{2N} \sum_{i=1}^{N} \Big[\log p_{i \to i}^{(m,n)} + \log q_{i \to i}^{(n,m)} \Big], \tag{11}$$

where $p_{i \to i}^{(m,n)}$ and $q_{i \to i}^{(n,m)}$ denote the normalized similarity probabilities from the modality m to n and n to m, respectively, and τ is the shared temperature coefficient that controls the sharpness of the distribution.

3.5.2. Multimodal Fusion

Let $\{\widetilde{\mathbf{H}}^{\text{rsi}}, \widetilde{\mathbf{H}}^{\text{bld}}, \widetilde{\mathbf{H}}^{\text{svi}}, \widetilde{\mathbf{H}}^{\text{poi}}\}$ denote the modality-specific embedding matrices from Section 3.4, where each $\widetilde{\mathbf{H}}^m = [\widetilde{\mathbf{h}}_1^m, \widetilde{\mathbf{h}}_2^m, \ldots, \widetilde{\mathbf{h}}_N^m]^\top \in \mathbb{R}^{N \times d_f}$ represents the features of all N urban grids under modality m, and $\widetilde{\mathbf{h}}_i^m \in \mathbb{R}^{d_f}$ denotes the feature vector of grid i. For each grid i, we construct a modality token matrix $\mathbf{T}_i = [\widetilde{\mathbf{h}}_i^{\text{rsi}}, \widetilde{\mathbf{h}}_i^{\text{bld}}, \widetilde{\mathbf{h}}_i^{\text{svi}}, \widetilde{\mathbf{h}}_i^{\text{poi}}]^\top \in \mathbb{R}^{4 \times d_f}$ and apply a lightweight Transformer to learn inter-modal dependencies:

$$\mathbf{z}_i = \text{AttnPool}\big(\text{Transformer}_{\text{modal}}(\mathbf{T}_i)\big), \tag{12}$$

where $\text{AttnPool}(\cdot)$ aggregates modality-aware features into a unified representation. The fused feature matrix is then expressed as $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_N]^\top \in \mathbb{R}^{N \times d_f}$, where each $\mathbf{z}_i$ encodes the integrated semantic and structural characteristics of grid i.

3.5.3. Temporal Modeling

To capture the diurnal variation of human mobility, the fused embeddings $\mathbf{Z}$ are temporally expanded into a 24-step sequence for each spatial unit. Let $\{\mathbf{e}_t\}_{t=1}^{24}$ denote learnable hour embeddings. We form per-hour contexts by conditioning the fused feature on $\mathbf{e}_t$:

$$\mathbf{z}_i^{(t)} = \phi([\mathbf{z}_i \| \mathbf{e}_t]) \in \mathbb{R}^{d_f}. \tag{13}$$

This yields $\mathbf{Z}_{\text{seq}} \in \mathbb{R}^{N\times 24\times d_f}$, where each spatial unit is represented by a 24-step temporal feature sequence. A Transformer encoder [66] is then applied to model long-range dependencies across hours, producing updated temporal representations $\mathbf{U} = \text{TransformerEncoder}(\mathbf{Z}_{\text{seq}}) \in \mathbb{R}^{N\times 24\times d_u}$.

For each grid i and hour t, the model outputs a probability distribution over C magnitude bins through a classification head:

$$\hat{\mathbf{p}}_{i,t} = \text{softmax}(\text{MLP}_{\text{cla}}(\mathbf{U}_{i,t})), \quad \hat{\mathbf{p}}_{i,t} \in \mathbb{R}^C. \tag{14}$$

Following the classification-then-regression strategy, the continuous outbound flow $\hat{y}_{i,t}$ is calculated as the expectation of the predicted bin distribution. Let the n-th bin correspond to the interval $[b_n, b_{n+1}]$ with midpoint $m_n = (b_n + b_{n+1})/2$. The final predicted mobility intensity is obtained by:

$$\hat{y}_{i,t} = \sum_{n=1}^{C} \hat{p}_{i,t}^{(n)} \cdot m_n, \tag{15}$$

where $\hat{p}_{i,t}^{(n)}$ is the predicted probability of bin n. This expectation-based formulation allows the model to produce smooth and physically meaningful continuous predictions while benefiting from the stability of a classification objective.

3.6. Loss Functions and Optimization

To jointly optimize multimodal, spatial, and temporal representations, the proposed framework employs two task-specific objectives—classification and regression—along with a contrastive alignment loss introduced in Sections 3.3 and 3.5. The overall training objective ensures that the model simultaneously learns to predict mobility intensity levels, capture fine-grained temporal variations, and maintain semantic consistency across modalities.

(i) Classification Loss: For each urban grid i and hour t, the model predicts a categorical travel intensity level from C discrete bins. If the ground-truth label $y_{i,t}$ is represented as a single class index (*hard labels*), the classification loss adopts a standard cross-entropy form:

$$\mathcal{L}_{\text{cla}} = -\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \log \hat{p}_{i,t}[y_{i,t}], \tag{16}$$

where $\hat{p}_{i,t}[y_{i,t}]$ denotes the predicted probability for the true class of grid i at hour t, and N and T represent the total number of grids and time steps ($T = 24$).

(ii) Regression Loss: To capture fine-grained variations in hourly travel intensity, a regression objective is introduced to directly estimate the continuous outbound flow volume. It is formulated as a mean squared error (MSE) between the predicted and ground-truth sequences:

$$\mathcal{L}_{\text{reg}} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(\hat{y}_{i,t} - y_{i,t} \right)^2, \tag{17}$$

where $\hat{y}_{i,t}$ and $y_{i,t}$ denote the predicted and observed hourly flow values for grid i at hour t.

(iii) Contrastive Alignment Loss: To ensure consistent multimodal semantics, the total contrastive loss $\mathcal{L}_{\text{CL}}$ combines both the intra- and inter-modal contrastive objectives defined in previous sections:

$$\mathcal{L}_{\text{CL}} = \sum_{m \in \mathcal{M}} \mathcal{L}_{\text{intra}}^{m} + \frac{1}{\binom{|\mathcal{M}|}{2}} \sum_{m<n} \mathcal{L}_{\text{inter}}^{(m,n)}, \tag{18}$$

where $\mathcal{L}_{\text{intra}}^{m}$ aligns spatial and semantic features within modality m, and $\mathcal{L}_{\text{inter}}^{(m,n)}$ enforces consistency between different modalities.

(iv) Overall Optimization: The final loss function jointly optimizes the classification, regression, and contrastive objectives:

$$\mathcal{L}_{\text{total}} = \alpha \, \mathcal{L}_{\text{cla}} + \beta \, \mathcal{L}_{\text{reg}} + \lambda \, \mathcal{L}_{\text{CL}}, \tag{19}$$

where α, β, and λ are weighting coefficients controlling the relative importance of the three components. This multi-objective formulation enables the model to simultaneously enhance prediction accuracy, temporal smoothness, and multimodal semantic coherence across the entire training process.

4. Experiments and Results

4.1. Datasets and Experimental Setup

4.1.1. Datasets

This study selected Shenzhen, China as the study area, dividing the built-up region into 5,263 grid tiles of 500 m × 500 m (Fig. 4). We integrated four types of static geospatial data: *RSI* from Google Earth (2.15 m resolution); *Building* height data from Gaode Map; *SVI* from Baidu Map (≈78k raw images, 1024 × 768, oversampled to 20 per grid); and *POI* data (≈1.5 million)

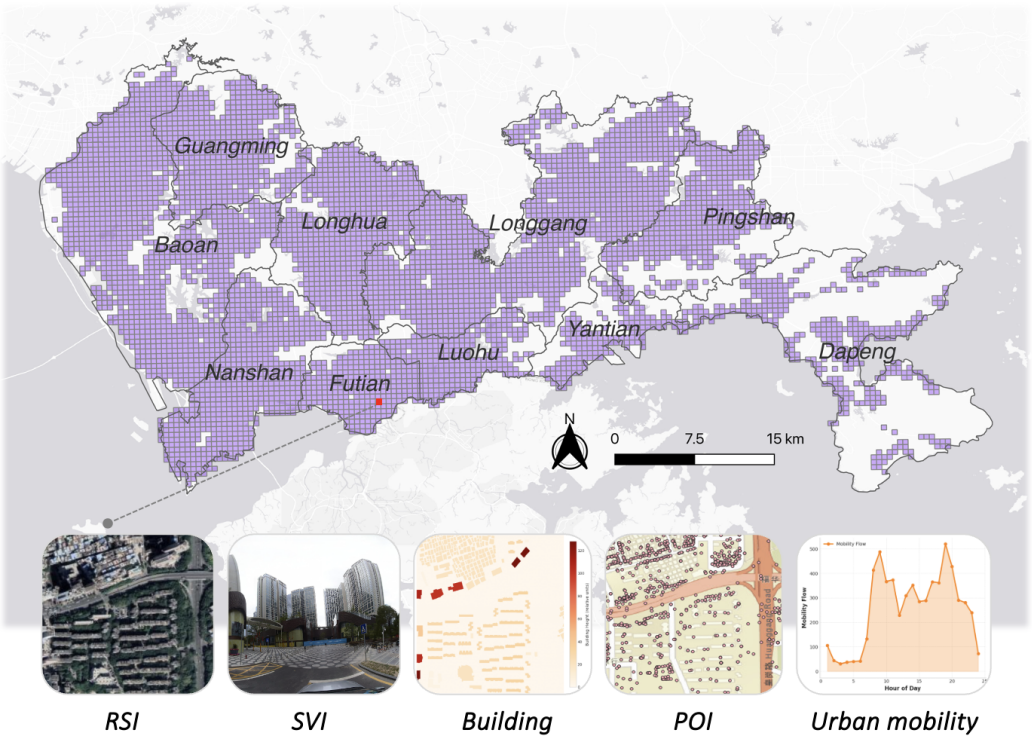

Figure 4: Study area and multimodal geospatial data overview. Shenzhen comprises 10 administrative districts, and its built-up region is partitioned into 5,263 grids of 500 m × 500 m, with four static data modalities illustrated.

from Gaode Map. The prediction target is the hourly outbound mobility derived from anonymized mobile signaling data. We utilized weekday records from June 18 to June 25, 2019. The mobility flows were log-transformed, Z-score normalized, and discretized into 10 bins. All modalities were collected within a one-year window centered on June 2019, minimizing temporal misalignment. The dataset was divided into training, validation, and test sets with a ratio of 6:2:2 using spatially random masks applied at the grid level.

4.1.2. Experimental Setup

We adopt *Overall Accuracy (OA)* for classification and *Coefficient of Determination* (R^2), *Mean Absolute Error* (MAE), and *Mean Absolute Percentage Error* (MAPE) for regression. The framework is implemented in PyTorch and trained on an NVIDIA RTX A6000 (48GB) for 300 epochs (Adam optimizer, lr= 1×10^{-4} with cosine annealing). Key hyperparameters: $d_f = 256$,

$\tau = 0.1$, α=0.5, β=0.5, λ=0.1.

To improve efficiency and transferability, the framework separates components into *frozen* and *trainable* groups. Frozen modules—VLM encoders (Qwen-VL, DeepSeek-R1) for offline semantic generation, spatial encoders (ResNet-152) for image features, and the text encoder (Qwen3-Embedding)—leverage large-scale pretraining without fine-tuning. Trainable modules—GAT layers, cross-modal Transformer fusion, temporal Transformer encoder, and prediction heads—are optimized via backpropagation. Fig. 5 illustrates the training process with stable convergence.

4.2. Results

This section presents the experimental results of the proposed LMEMR for grid-level mobility prediction. We comprehensively evaluate the model in terms of quantitative performance and interpretability. Comparisons with baseline models and ablation variants are provided to demonstrate the advantages of the proposed modules.

4.2.1. Overall Performance Comparison with Baselines

To evaluate LMEMR, we compare it against representative machine learning, deep learning baselines, and specialized spatiotemporal models. Except for CLIP, all baselines utilize the same 256-dimensional raw spatial encodings $\mathbf{f}_i^m$ (Section 3.3.2) without textual semantics or contrastive learning. The baselines are:

- *Linear Regression (LR).* Maps grid embeddings to hourly mobility intensity via ordinary least squares [67].
- *Random Forest (RF).* An ensemble of 500 decision trees capturing nonlinear feature–mobility relationships [68].
- *CNN.* A 2D convolutional model capturing local spatial correlations, lacking explicit temporal reasoning [69].
- *LSTM.* A two-layer RNN modeling the 24-hour sequence of grid embeddings to predict mobility dynamics [70].
- *Transformer.* A temporal attention model learning long-range dependencies across hourly time steps [66].

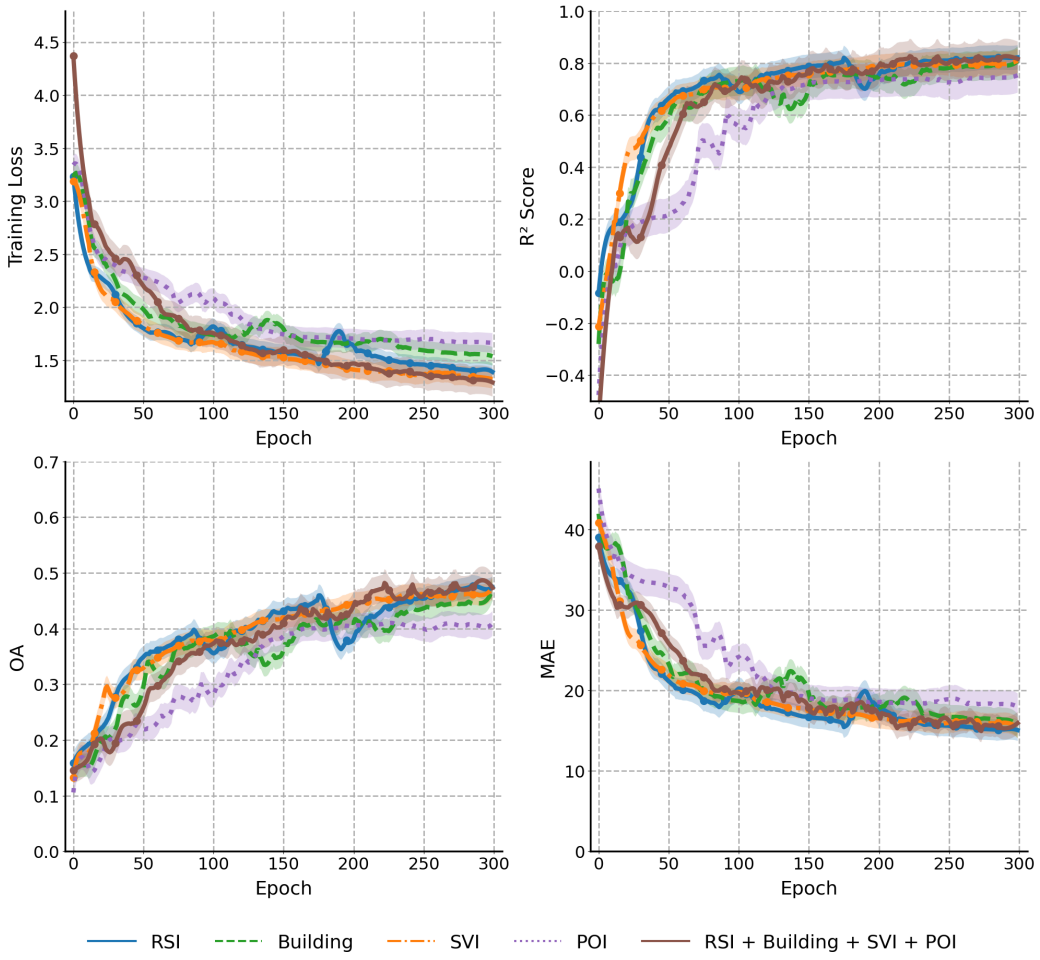

Figure 5: Training and validation curves of the proposed model. The plots illustrate the evolution of Training Loss, OA, R^2, and MAE over 300 epochs for different modality configurations.

- *CLIP.* A multimodal baseline using pretrained CLIP ViT-L/14 for visual encoding (RSI, Building, SVI) and CLIP Text Encoder for POI, retaining the same spatial and temporal modules as LMEMR [54].

As presented in Table 2, LMEMR outperforms all baselines. Traditional models (LR, RF) show limited accuracy due to insufficient spatial-behavioral modeling. Deep learning approaches (CNN, LSTM) improve performance by capturing local spatial or temporal dependencies, with the Transformer further enhancing results via attention mechanisms. The multimodal baseline

(CLIP) leverages pretrained vision-language embeddings but achieves suboptimal results ($R^2 = 0.761$), indicating that generic multimodal representations lack domain-specific spatial-behavioral knowledge. LMEMR significantly surpasses all baselines with inference time of 17.2s and training time of 2,327s, confirming that jointly modeling spatial, semantic, and temporal dimensions offers a superior understanding of urban mobility with reasonable computational cost.

Table 2: Overall predictive performance comparison with baselines. Best in **bold**, second best underlined.

Model	OA ↑	R^2 ↑	MAE ↓	MAPE ↓	Training (s)	Inference (s) ↓
LR	0.237	0.452	26.794	0.405	**135**	**10.3**
RF	0.226	0.521	24.916	0.374	271	**10.3**
CNN	0.411	0.734	18.904	0.259	1283	11.9
LSTM	0.423	0.733	17.882	0.264	1225	12.3
Transformer	0.425	0.755	17.525	0.249	1451	14.9
CLIP	0.430	0.761	16.769	0.245	1542	14.9
LMEMR	**0.521**	**0.856**	**13.738**	**0.187**	2327	17.2

4.2.2. Performance in Temporal Dynamics

Fig. 6 demonstrates that LMEMR accurately reproduces diurnal mobility dynamics across diverse urban environments. The model successfully captures the *morning-peak* in dense residential areas (Baoan urban village), the *evening-peak* in office clusters (Futian CBD), the *stable flow* in transport hubs (Longhua railway station), and the *dual-peak* pattern in mixed-use districts (Luohu). The predictions align closely with observed trends, showing high fidelity in both magnitude and temporal variation, with only minor deviations in highly dynamic zones.

4.2.3. Performance in Mobility Intensity Classification

Fig. 7 compares the classification performance across ten mobility intensity bins. LMEMR exhibits the strongest diagonal concentration (OA=0.521), significantly outperforming Transformer (OA=0.425) and RF (OA=0.226). Specifically, LMEMR reduces misclassification in the dominant low-to-mid intensity bins (0–5) and maintains robustness in high-activity bins (6–8), effectively mitigating the over-smoothing and underestimation issues observed

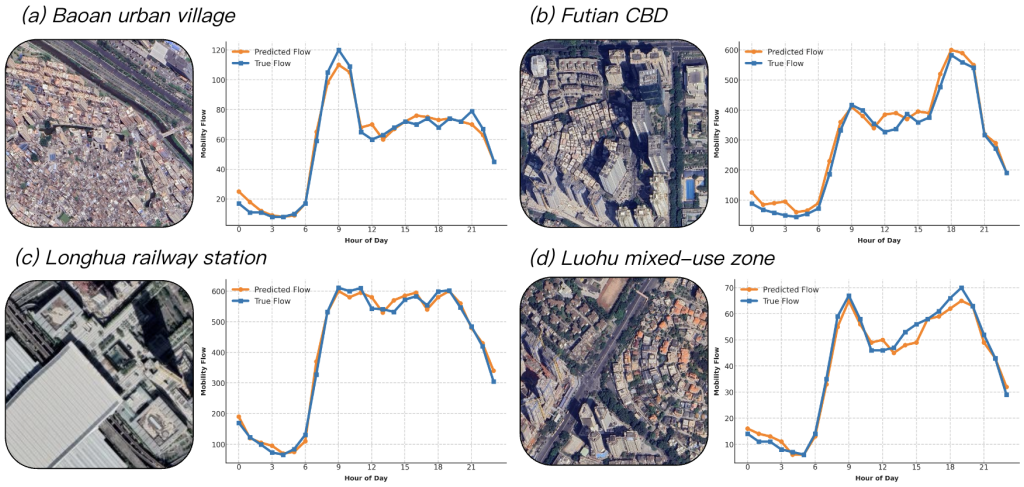

Figure 6: Predicted and observed hourly mobility in representative urban areas. Each pair shows a high-resolution remote sensing image (left) and its corresponding flow curve (right): (a) Baoan urban village (morning-peak), (b) Futian CBD (evening-peak), (c) Longhua railway station (stable flow), and (d) Luohu mixed-use zone (dual peaks).

in baselines. This confirms that integrating multimodal semantics and spatial reasoning enhances the model's precision in distinguishing fine-grained mobility states.

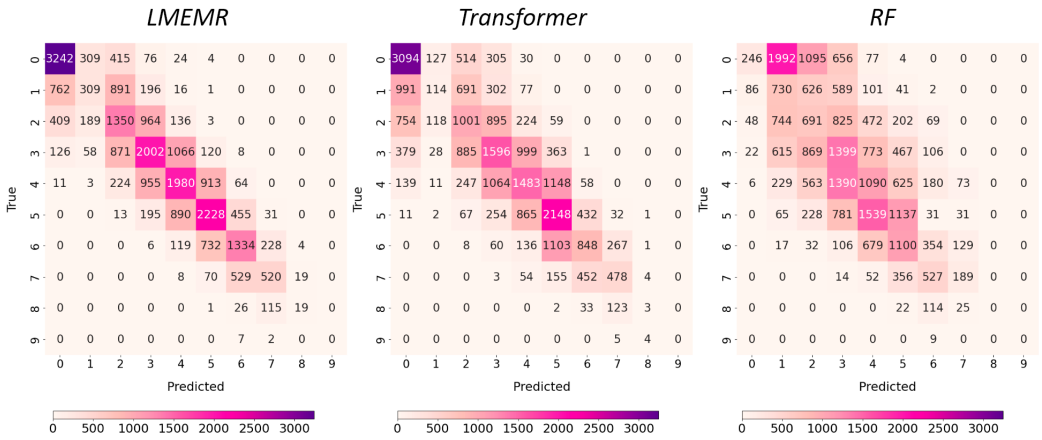

Figure 7: Confusion matrices of LMEMR, Transformer, and RF across ten mobility intensity bins (0–9). LMEMR achieves the strongest diagonal concentration and highest accuracy (OA=0.521), clearly outperforming Transformer (0.425) and RF (0.226).

4.2.4. Cross-City Generalization Analysis

To validate the generalizability of LMEMR across diverse urban contexts, we conducted independent experiments on three additional cities in the Greater Bay Area: Guangzhou (mega city, 8,054 grids), Dongguan (large industrial city, 5,022 grids), and Zhuhai (medium-sized tourism city, 2,173 grids), each trained and evaluated independently using identical protocols (6:2:2 split). As shown in Table 3, LMEMR consistently outperforms the CLIP baseline across all four cities, achieving high predictive accuracy with R^2 ranging from 0.734 to 0.863 across cities of varying scales and economic structures. Guangzhou achieves the highest R^2 (0.863), surpassing even Shenzhen (0.856), likely influenced by its larger sample size. Dongguan shows the largest relative improvement in R^2 over CLIP, suggesting that VLM-generated semantics offer notable benefits for cities with different urban structures. Zhuhai maintains competitive performance ($R^2 = 0.810$) despite the smallest sample size. These results confirm that LMEMR generalizes effectively across diverse urban contexts.

Table 3: Cross-city generalization validation across four cities in the Greater Bay Area.

City	**Model**	**OA ↑**	**R^2 ↑**	**MAE ↓**	**MAPE ↓**
Shenzhen	LMEMR	0.521	0.856	13.738	0.187
	CLIP	0.430	0.761	16.769	0.245
Guangzhou	LMEMR	0.578	0.863	10.943	0.238
	CLIP	0.513	0.777	12.934	0.283
Dongguan	LMEMR	0.421	0.734	5.967	0.273
	CLIP	0.374	0.644	6.339	0.337
Zhuhai	LMEMR	0.596	0.810	3.909	0.244
	CLIP	0.543	0.733	4.138	0.370

4.3. Ablation Studies

To isolate the contribution of each proposed component, all ablation variants are independently retrained from scratch using the same training configuration (300 epochs, Adam optimizer, lr= 1×10^{-4}) and identical data splits. For each variant, only the targeted module is modified while all other components remain unchanged, ensuring fair and controlled comparisons.

4.3.1. Effectiveness of the Semantic Enhancement Module

To evaluate the impact of VLM-derived semantics and contrastive alignment, four variants are developed: *(i) w/o Semantics*, which removes VLM-generated textual descriptions and uses only raw visual features from the spatial encoder E_{v}^{m}; *(ii) w/o Intra-CL*, which removes the intra-modal contrastive loss $\mathcal{L}_{\mathrm{intra}}^{m}$ while keeping all other components; *(iii) Prompt-Lite*, which uses simplified prompts without behavioral cues; and *(iv) Prompt-Shuffle*, which employs randomized captions to test robustness.

As summarized in Table 4, LMEMR achieves superior performance across all metrics. The removal of textual semantics leads to a significant degradation, indicating that high-level semantic cues are essential for predictive accuracy. Similarly, the performance drop in *w/o Intra-CL* confirms the necessity of explicitly aligning visual and textual representations. The intermediate results of *Prompt-Lite* and the poor performance of *Prompt-Shuffle* further validate that behavior-aware prompting and accurate text–location correspondence are critical for capturing travel dynamics. Qualitative analysis in Fig. 8 further illustrates this contribution. LMEMR accurately reproduces diurnal peaks, whereas the variant without semantics underestimates amplitudes and misaligns timing. The generated descriptions explicitly link built-environment characteristics (e.g., "mixed-use," "office clusters") with mobility behaviors, serving as semantic priors that guide the model to better associate spatial structures with travel rhythms.

Table 4: Ablation of the semantic enhancement module. Best in **bold**, second best underlined.

Variant	**OA ↑**	**R^2 ↑**	**MAE ↓**	**MAPE ↓**
w/o Semantics	0.491	0.825	14.833	0.213
w/o Intra-CL	0.505	0.837	14.285	0.209
Prompt-Lite	<u>0.514</u>	<u>0.850</u>	<u>13.783</u>	<u>0.199</u>
Prompt-Shuffle	0.475	0.811	15.426	0.228
LMEMR	**0.521**	**0.856**	**13.738**	**0.187**

4.3.2. Contribution of Spatial-Contextual Representation Learning

To quantify the role of spatial dependencies, we compare against two variants: *(i) w/o Spatial*, where grid embeddings are fed directly into the

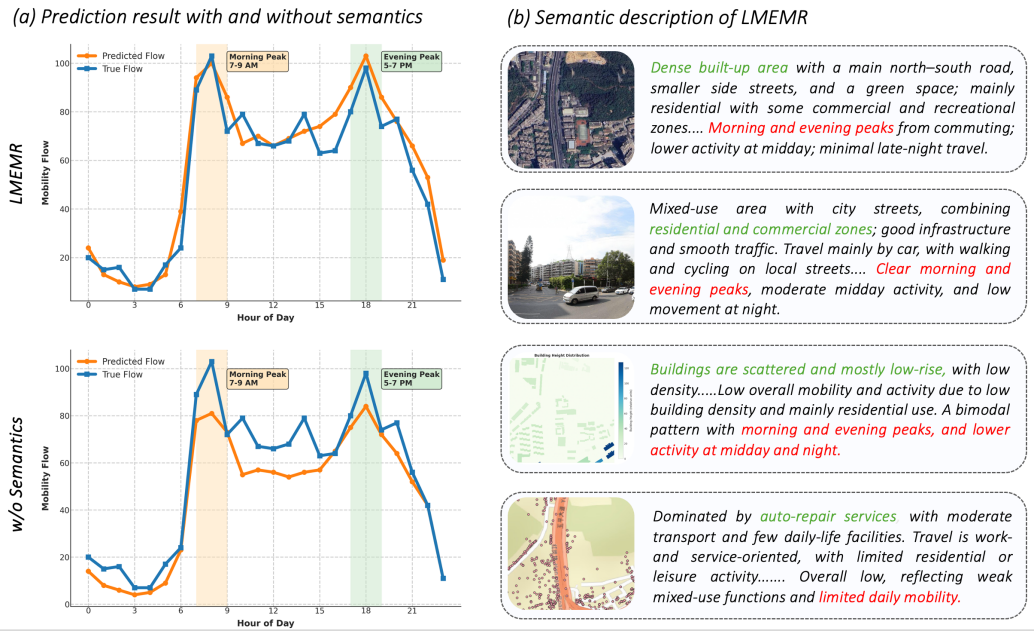

Figure 8: Comparison of mobility flow predictions and semantic interpretation. (a) LMEMR captures morning and evening peaks more accurately than the version without semantics. (b) VLM-generated descriptions provide explanatory context for the predicted patterns.

temporal encoder without spatial aggregation (GAT removed); and *(ii) GCN*, which replaces the attention-based GAT with fixed-weight mean aggregation using standard graph convolution [71].

Table 5 reports the results. The complete LMEMR framework yields the highest accuracy, while removing the spatial module causes a substantial drop in R^2 (from 0.856 to 0.764), underscoring that spatial context is indispensable for grid-level prediction. The GCN variant improves upon the non-spatial baseline but remains inferior to the GAT-based approach, demonstrating the advantage of adaptive attention in capturing non-uniform spatial interactions. Fig. 9 visualizes the learned attention weights. The spatial distribution reveals a clear urban hierarchy, with high-weight areas concentrated in core commercial corridors and transit hubs (e.g., Luohu–Futian–Nanshan). These zones exert strong spatial influences on adjacent regions, justifying the higher contextual weights assigned by the model. Furthermore, modality-specific maps show distinct hotspots, reflecting the complementary roles of structural (RSI/Building) and functional (SVI/POI) features in shaping ur-

ban mobility.

Table 5: Ablation of spatial-contextual representation learning. Best in **bold**, second best underlined.

Variant	OA ↑	R^2 ↑	MAE ↓	MAPE ↓
w/o Spatial	0.434	0.764	17.434	0.247
GCN	0.488	0.820	15.061	0.217
LMEMR	**0.521**	**0.856**	**13.738**	**0.187**

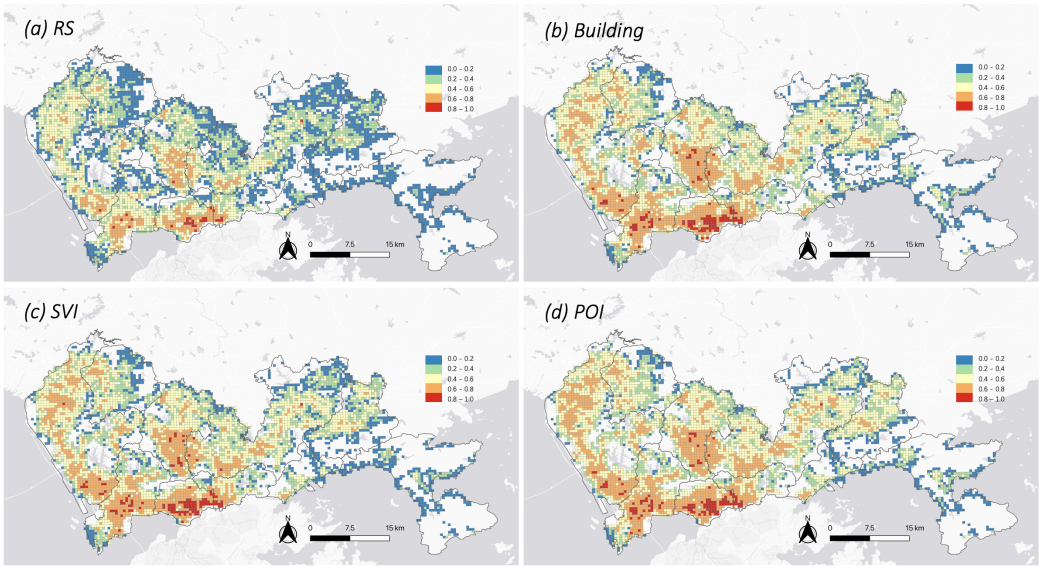

Figure 9: Spatial distribution of learned attention weights. High coefficients (red) align with core urban districts (e.g., Luohu, Futian), indicating strong spatial dependencies in dense functional zones.

4.3.3. Role of Cross-modal Semantic Fusion Module

This section investigates the contribution of cross-modal mechanisms through four sets of variants: *(i) w/o Inter-CL*, which disables inter-modal contrastive loss $\mathcal{L}_{\text{inter}}$ while retaining intra-modal alignment and the Transformer fusion; *(ii) Late-Concat*, which replaces the Transformer-based fusion with simple feature concatenation followed by MLP; *(iii) Gated-Fusion*, which uses gated summation instead of attention-based fusion; and *(iv)*

Single-Mod Drop, which retains only individual modalities while keeping the GAT and temporal modules.

Results in Table 6 demonstrate that the full LMEMR model achieves the highest accuracy. The degradation observed in *w/o Inter-CL* confirms that explicit cross-modal alignment fosters consistent representations among heterogeneous data. Furthermore, the attention-based fusion outperforms both concatenation and gating, indicating that the multi-head mechanism better captures nonlinear cross-modal dependencies. Among single-modality variants, RSI and SVI perform relatively well, suggesting that morphology and visual perception are primary determinants of mobility, whereas POI data alone is insufficient. Overall, the findings validate the necessity of the proposed contrastive–Transformer fusion design for robust multimodal integration.

Table 6: Ablation of cross-modal semantic fusion. Best in **bold**, second best underlined.

Variant	**OA ↑**	**R^2 ↑**	**MAE ↓**	**MAPE ↓**
w/o Inter-CL	0.492	0.828	14.879	0.219
Late-Concat	0.507	0.831	14.377	0.215
Gated-Fusion	0.495	0.827	14.787	0.213
RSI-only	0.482	0.820	14.924	0.222
Building-only	0.478	0.810	15.381	0.218
SVI-only	0.478	0.815	15.016	0.214
POI-only	0.435	0.764	17.571	0.243
LMEMR	**0.521**	**0.856**	**13.738**	**0.187**

4.3.4. Effect of Prediction Head Design

To evaluate the contribution of the classification-then-regression strategy, we compare three loss configurations: *(i) Regression only*, which uses MSE loss directly on continuous flow values; *(ii) Classification only*, which discretizes flow into intensity bins and uses cross-entropy loss without regression refinement; *(iii) Classification + Regression (LMEMR)*, the full design combining both losses.

As shown in Table 7, the combined strategy ($R^2 = 0.856$) outperforms both regression-only ($R^2 = 0.845$) and classification-only ($R^2 = 0.828$) approaches. Regression-only already achieves competitive performance, but the

Table 7: Ablation of prediction head design. Best in **bold**, second best underlined.

Variant	OA ↑	R^2 ↑	MAE ↓	MAPE ↓
Regression-Only	0.497	0.845	14.102	0.203
Classification-Only	0.492	0.828	14.696	0.215
LMEMR	**0.521**	**0.856**	**13.738**	**0.187**

distributional approach further improves by handling the long-tail mobility distribution through probability bins, mitigating extreme value influence during training. Classification-only performs worst, indicating that discretization alone loses important magnitude information. These results confirm that the two components are complementary.

5. Discussion

5.1. Error Analysis across Different Hours

As illustrated in Fig. 10, model performance exhibits a distinct diurnal pattern linked to the regularity of human mobility. Quantitative metrics indicate that prediction precision is highest during stable, low-mobility periods (e.g., early morning $R^2 > 0.84$) and decreases during transient phases such as morning (07:00–09:00) and evening (18:00–19:00) peaks. Specifically, MAPE peaks at 04:00 and 06:00, reflecting sensitivity to rapid shifts in travel intensity. This temporal variation aligns with previous findings that static features effectively capture structural determinants (e.g., land use) but struggle with volatile congestion dynamics [72]. Nevertheless, the consistently high R^2 and low average MAPE confirm that multimodal semantics and spatial–graph reasoning substantially enhance temporal robustness.

5.2. Semantic Analysis of Mobility Patterns

Hierarchical clustering of 24-hour mobility vectors reveals spatially distinct travel rhythms rooted in urban functionality. As shown in Fig. 11 (a), increasing clusters to K=7 progressively refines the urban structure from a simple built/non-built dichotomy to a polycentric hierarchy consistent with functional zones [73]. Note that K=7 is not claimed as an optimal value but serves as an exploratory setting that balances granularity and interpretability; the hierarchical evolution from K=2 to K=7 demonstrates that the discovered patterns remain coherent across resolutions. At this resolution,

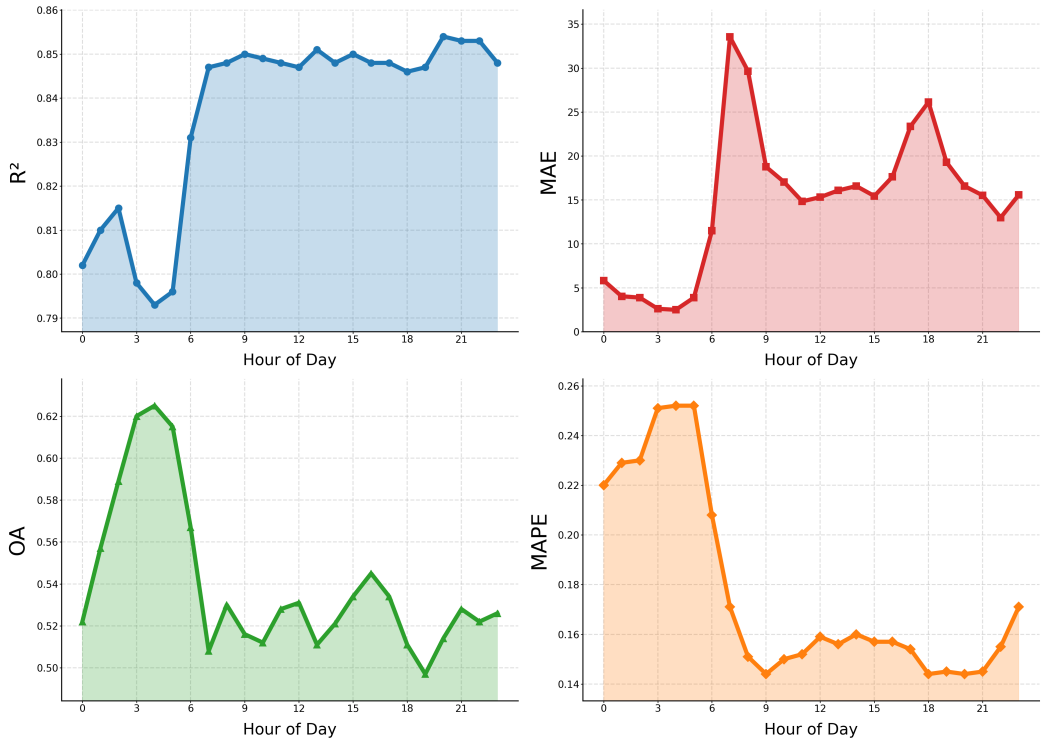

Figure 10: Hourly error analysis. Model performance across 24 hours evaluated using R^2, OA, MAE, and MAPE.

four representative patterns emerge from the seven clusters, supported by VLM-generated semantics (Fig. 11 (b)): (i) *Cluster 0 (Morning-peak)*, characterized by 7:00–9:00 AM surges and terms like "*commuting*" and "*residential*," indicating outbound flow from housing areas; (ii) *Cluster 1 (Evening-peak)*, associated with "*commercial district*" and "*night life*," reflecting post-work activity in CBDs; (iii) *Cluster 2 (Dual-peak)*, featuring bidirectional flows typical of mixed-use "*transportation nodes*"; and (iv) *Cluster 3 (Low-activity)*, corresponding to "*ecological zones*" with minimal human presence. These results demonstrate that integrating temporal signatures with textual semantics enables a coherent, explainable interpretation of urban dynamics.

5.3. Impact of Contrastive Learning

Fig. 12 visualizes the evolution of multimodal embedding distributions via t-SNE [74]. In the absence of contrastive learning (CL), embeddings remain scattered and modality-dependent, indicating weak semantic correspondence.

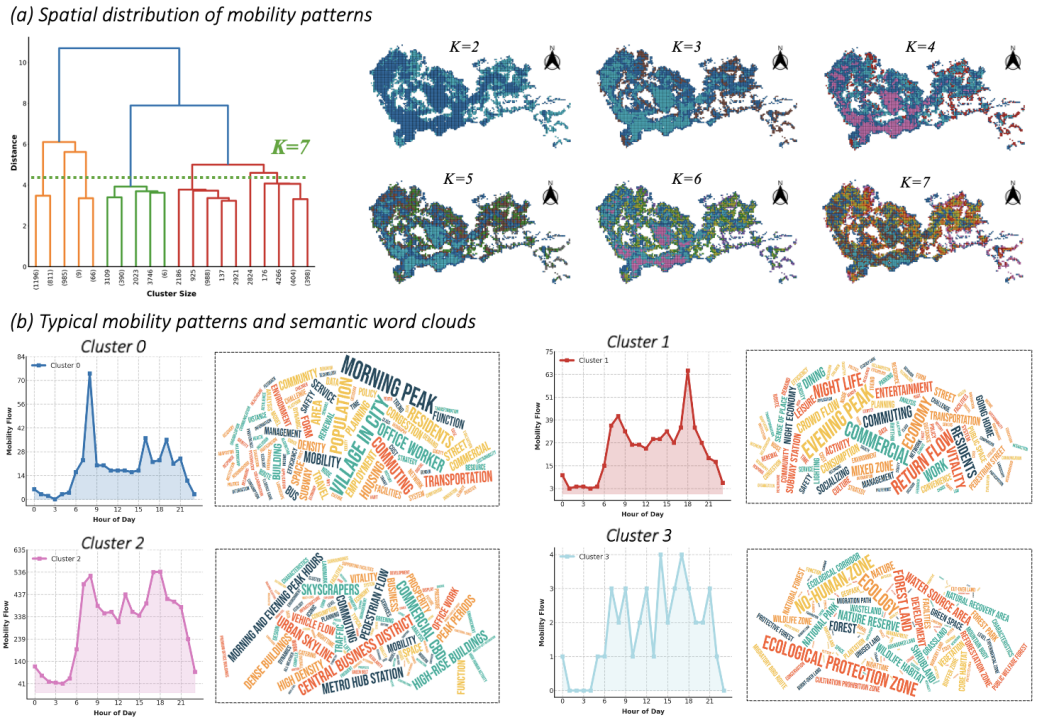

Figure 11: Spatiotemporal analysis of mobility patterns. (a) *Hierarchical evolution*: Clusters from K=2 to K=7 illustrate the structural refinement from a binary built/non-built dichotomy to a polycentric urban hierarchy. (b) *Semantic-temporal patterns*: Representative diurnal curves and VLM word clouds at K=7. The identified patterns include: *Cluster 0 (Morning-peak)*: residential zones driven by early commuting; *Cluster 1 (Evening-peak)*: commercial CBDs with nightlife activity; *Cluster 2 (Dual-peak)*: mixed-use areas and transport hubs; and *Cluster 3 (Low-activity)*: ecological protection zones. Word clouds link quantitative mobility rhythms to qualitative urban functions.

Conversely, the inclusion of CL drives the convergence of heterogeneous features into compact, semantically coherent clusters based on shared urban functions (e.g., residential vs. commercial). This explicit optimization of intra-class compactness and inter-modality similarity yields a significantly more discriminative embedding space than traditional fusion methods that rely solely on concatenation.

5.4. Impact of VLM Model Size

Table 8 presents a controlled comparison of six VLMs with varying parameter scales. Performance generally correlates with model capacity; the largest

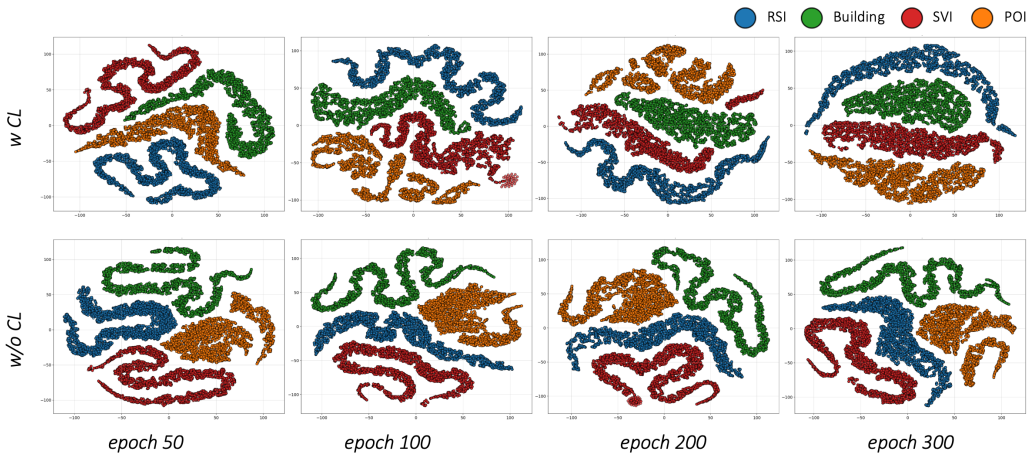

Figure 12: t-SNE visualization of multimodal embeddings with and without contrastive learning (CL). Colors denote modalities (RSI, Building, SVI, POI).

model, Qwen-VL-Max, achieves the best accuracy (R^2 = 0.856, MAE= 13.738), outperforming the lightweight Gemini-Flash-Lite by a clear margin. This suggests that larger models provide richer semantic reasoning for text–image alignment. However, gains diminish beyond 70B parameters, with mid-sized models (e.g., Qwen3-VL-32B, GPT-4o-mini) offering a favorable balance between accuracy ($R^2 > 0.83$) and computational efficiency (≈27s/image). Consequently, models with approximately 30B parameters represent the most practical trade-off for large-scale urban semantic applications.

Table 8: Performance comparison across different VLM model sizes. Time denotes the per-image semantic generation time. Best in **bold**, second best underlined.

Model	OA ↑	R^2 ↑	MAE ↓	MAPE ↓	Time (s/image)
Qwen-VL-Max	**0.521**	**0.856**	**13.738**	**0.187**	28.6
Qwen2.5-VL-72B	0.501	0.844	14.103	0.209	25.5
Qwen3-VL-32B	0.495	0.834	14.422	0.216	27.4
Qwen3-VL-8B	0.491	0.834	14.514	0.217	20.6
GPT-4o-mini	0.499	0.843	14.148	0.211	28.0
Gemini-Flash-Lite	0.499	0.838	14.285	0.214	**18.0**

5.5. Practical Applications

The proposed framework enables several practical applications in urban management. For urban planning, the predicted 24-hour mobility profiles allow planners to evaluate how land-use configurations shape daily activity patterns, informing zoning and development decisions without requiring dynamic trajectory data [75]. For transit operations, the hourly temporal resolution supports demand-responsive scheduling, enabling agencies to adjust service frequency and routes based on predicted ridership [76]. More broadly, because the framework relies solely on static geospatial data—which is widely available through open platforms (e.g., OpenStreetMap, Google Earth)—it can be readily generalized to cities worldwide that lack proprietary mobility datasets, as evidenced by consistent performance across four cities in the Greater Bay Area (Table 3), offering a scalable and privacy-preserving solution for global urban mobility modeling [77].

6. Conclusion

This study proposes LMEMR, a framework that predicts hourly grid-level mobility solely from static geospatial data by combining VLM-based semantic enhancement, dual-level contrastive learning, and graph-attention spatiotemporal modeling. Experiments on four cities confirm consistent generalizability; on Shenzhen, LMEMR achieves an 18.07% MAE reduction (R^2=0.856) over the best baseline (CLIP), validating the potential of widely accessible static data for scalable and privacy-friendly mobility inference. The main contributions are: (i) a unified multimodal semantic graph framework for hourly mobility prediction; (ii) a dual-level contrastive strategy aligning raw and semantic features across modalities; (iii) an attention-based intermodal fusion and spatiotemporal module; and (iv) extensive multi-city experiments confirming superiority, interpretability, and generalizability.

Despite these advances, several limitations remain. (i) Static features cannot capture event-driven perturbations (e.g., extreme weather, public events) [78]; integrating weakly dynamic priors may mitigate this gap. (ii) All four evaluated cities lie within the Greater Bay Area of China and share similar urban morphologies and data infrastructures. Generalizability to cities with fundamentally different forms—such as North American sprawl or informal settlements in developing countries—requires further validation. (iii) The 500 m grid resolution introduces scale sensitivity, and the Modifiable Areal Unit Problem (MAUP) implies that spatial conclusions may vary with

the chosen zoning scheme. Future work should examine multi-scale effects and adaptive spatial partitioning strategies.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 42401553, in part by the Natural Science Foundation of Top Talent of Shenzhen Technology University under Grant GDRC202415, in part by the Shenzhen Science and Technology Program under Grants JCYJ20240813113300001 and 20231127180406001, and in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2026A1515011875.

References

[1] E. Chen, Y. Liu, M. Yang, Revealing senior mobility patterns and activities in urban transit systems, IEEE Transactions on Intelligent Transportation Systems 24 (2023) 11424–11437. doi:10.1109/TITS.2023.3275389.

[2] M. Chen, Q. Yuan, C. Yang, Y. Zhang, Decoding urban mobility: Application of natural language processing and machine learning to activity pattern recognition, prediction, and temporal transferability examination, IEEE Transactions on Intelligent Transportation Systems 25 (2024) 7151–7173. doi:10.1109/TITS.2023.3339772.

[3] L. G. Azolin, A. N. R. da Silva, N. Pinto, Incorporating public transport in a methodology for assessing resilience in urban mobility, Transportation Research Part D-transport and Environment 85 (2020) 102386. doi:10.1016/j.trd.2020.102386.

[4] E. Suryani, R. A. Hendrawan, P. F. E. Adipraja, R. Indraswari, System dynamics simulation model for urban transportation planning: a case study, International Journal of Simulation Modelling (2020). doi:10.2507/ijsimm19-1-493.

[5] K. Zhang, Q. Jin, K. Pelechrinis, T. Lappas, On the importance of temporal dynamics in modeling urban activity, in: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, 2013, pp. 7:1–7:8. doi:10.1145/2505821.2505825.

[6] K. Qin, Y. Xu, C. Kang, S. Sobolevsky, M. Kwan, Modeling spatio-temporal evolution of urban crowd flows, ISPRS Int. J. Geo Inf. 8 (2019) 570. doi:10.3390/ijgi8120570.

[7] C. Liu, L. Chen, Q. Yuan, H. Wu, W. Huang, Revealing dynamic spatial structures of urban mobility networks and the underlying evolutionary patterns, ISPRS Int. J. Geo Inf. 11 (2022) 237. doi:10.3390/ijgi11040237.

[8] L. B. Santos, L. Carvalho, W. Seron, F. Coelho, E. Macau, M. G. Quiles, A. Monteiro, How do urban mobility (geo)graph's topological properties fill a map?, Applied Network Science 4 (2019) 1–14. doi:10.1007/s41109-019-0211-7.

[9] Z. Ye, Construction and application of urban mobility diagnosis model, Frontiers of Urban and Rural Planning 1 (2023). doi:10.1007/s44243-023-00016-9.

[10] X. Shi, F. Lv, D. Seng, B. Xing, B. Chen, Visual exploration of mobility dynamics based on multi-source mobility datasets and poi information, Journal of Visualization 22 (2019) 1209–1223. doi:10.1007/s12650-019-00594-1.

[11] R. Jiang, X. Song, Z. Fan, T. Xia, Z. Wang, Q. Chen, Z. Cai, R. Shibasaki, Transfer urban human mobility via POI embedding over multiple cities, ACM Transactions on Data Science 2 (2021) 1–26. doi:10.1145/3416914.

[12] M. Pastorino, F. Gallo, A. D. Febbraro, G. Moser, N. Sacco, S. Serpico, Multimodal fusion of mobility demand data and remote sensing imagery for urban land-use and land-cover mapping, Remote. Sens. 14 (2022) 3370. doi:10.3390/rs14143370.

[13] W. Tu, Z. Hu, L. Li, J. Cao, J. Jiang, Q. Li, Q. Li, Portraying urban functional zones by coupling remote sensing imagery and human sensing data, Remote. Sens. 10 (2018) 141. doi:10.3390/rs10010141.

[14] M. Quintana, Y. Gu, X. Liang, Y. Hou, K. Ito, Y. Zhu, M. Abdelrahman, F. Biljecki, Global urban visual perception varies across demographics and personalities, Nature Cities 2 (11) (2025) 1092–1106. doi:10.1038/s44284-025-00330-x.

[15] C. Ye, F. Zhang, L. Mu, Y. Gao, Y. Liu, Urban function recognition by integrating social media and street-level imagery, Environment and Planning B: Urban Analytics and City Science 48 (2020) 1430–1444. doi:10.1177/2399808320935467.

[16] T. Zhao, X. Liang, W. Tu, Z. Huang, F. Biljecki, Sensing urban soundscapes from street view imagery, Comput. Environ. Urban Syst. 99 (2023) 101915. doi:10.1016/j.compenvurbsys.2022.101915.

[17] F. Zhang, L. Wu, D. Zhu, Y. Liu, Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns, ISPRS journal of photogrammetry and remote sensing 153 (2019) 48–58. doi:10.1016/j.isprsjprs.2019.04.017.

[18] Y. Zhang, P. Liu, F. Biljecki, Knowledge and topology: A two layer spatially dependent graph neural networks to identify urban functions with time-series street view image, ISPRS Journal of Photogrammetry and Remote Sensing (2023). doi:10.1016/j.isprsjprs.2023.03.008.

[19] Y. Liu, X. Zhang, J. Ding, Y. Xi, Y. Li, Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction, Proceedings of the ACM Web Conference 2023 (2023). doi:10.1145/3543507.3583876.

[20] Z. Wang, H. Li, R. Rajagopal, Urban2vec: Incorporating street view imagery and pois for multi-modal urban neighborhood embedding, ArXiv abs/2001.11101 (2020). doi:10.1609/AAAI.V34I01.5450.

[21] M. Chen, Z. Li, H. Jia, X. Shao, J. Zhao, Q. Gao, M. Yang, Y. Yin, MGRL4RE: A multi-graph representation learning approach for urban region embedding, ACM Transactions on Intelligent Systems and Technology 16 (2025) 1–23. doi:10.1145/3712698.

[22] X. Wang, J. Cao, T. Zhao, B. Zhang, G. Chen, Z. Li, H. Chen, W. Tu, Q. Li, St-camba: A decoupled-free spatiotemporal graph fusion state space model with linear complexity for efficient traffic forecasting, Information Fusion (2025) 103495doi:10.2139/ssrn.5206097.

[23] X. Wang, T. Zhao, W. Tu, B. Zhang, G. Chen, J. Cao, Sat2flow: A structure-aware diffusion framework for human flow generation from

satellite imagery, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 40, 2026, pp. 15886–15894.

[24] Y. Kang, Short-term passenger flow prediction in urban rail transit based on hybrid deep learning models, Highlights in Science, Engineering and Technology (2024). doi:10.54097/nqxs4628.

[25] Y. Sun, Prediction of short-term passenger flow in the metro station with cnn-lstm model, 2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI) (2023) 1218–1222doi:10.1109/icetci57876.2023.10176978.

[26] L. Liu, Z. Qiu, G. Li, Q. Wang, W. Ouyang, L. Lin, Contextualized spatial–temporal network for taxi origin-destination demand prediction, IEEE Transactions on Intelligent Transportation Systems 20 (10) (2019) 3875–3887. doi:10.1109/tits.2019.2915525.

[27] J. Zhang, H. Che, F. Chen, W. Ma, Z. He, Short-term origin-destination demand prediction in urban rail transit systems: A channel-wise attentive split-convolutional neural network method, Transportation Research Part C: Emerging Technologies 124 (2021) 102928. doi:10.1016/j.trc.2020.102928.

[28] I. D. P. Arenas, H. Alatrista-Salas, M. N. del Prado Cortez, Discovery of urban mobility patterns, Advances in Data Science and Information Engineering (2021). doi:10.1007/978-3-030-71704-9_33.

[29] J. Cheng, K. Li, Y. Liang, L. Sun, J. Yan, Y. Wu, Rethinking urban mobility prediction: A multivariate time series forecasting approach, IEEE Transactions on Intelligent Transportation Systems 26 (2025) 2543–2557. doi:10.1109/tits.2024.3498054.

[30] Q. Pan, Y. Chen, G. Shen, Y. Yang, X. Kong, Spatio-temporal knowledge embedding via circular correlation: insights into functional urban area travel pattern mining, Neural Comput. Appl. 36 (2024) 19075–19095. doi:10.1007/s00521-024-10167-5.

[31] T. Li, Q. Feng, B. Niu, B. Chen, F. Yan, J. Gong, J. Liu, Mapping urban villages based on point-of-interest data and a deep learning approach, Cities (2025). doi:10.1016/j.cities.2024.105549.

[32] R. Chen, S. Jiang, W. Huang, Semob: Semantic synthesis for dynamic urban mobility prediction, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 2025, pp. 15346–15366.

[33] Y. Xu, S. Gao, Q. Huang, A. Göçmen, Q. Zhu, F. Zhang, Predicting human mobility flows in cities using deep learning on satellite imagery, Nature Communications 16 (1) (2025) 10372.

[34] J. Cao, X. Wang, G. Chen, W. Tu, X. Shen, T. Zhao, J. Chen, Q. Li, Disentangling the hourly dynamics of mixed urban function: A multimodal fusion perspective using dynamic graphs, Information Fusion 117 (2025) 102832. doi:10.1016/j.inffus.2024.102832.

[35] F. Zhang, B. Zhou, L. Liu, Y. Liu, H. H. Fung, H. Lin, C. Ratti, Measuring human perceptions of a large-scale urban region using machine learning, Landscape and Urban Planning 180 (2018) 148–160. doi:10.1016/j.landurbplan.2018.08.020.

[36] M. Wu, Q. Huang, S. Gao, Z. Zhang, Mixed land use measurement and mapping with street view images and spatial context-aware prompts via zero-shot multimodal learning, International Journal of Applied Earth Observation and Geoinformation 125 (2023) 103591. doi:10.1016/j.jag.2023.103591.

[37] C. Su, X. Hu, Q. Meng, L. Zhang, W. Shi, M. Zhao, A multimodal fusion framework for urban scene understanding and functional identification using geospatial data, International Journal of Applied Earth Observation and Geoinformation 127 (2024) 103696. doi:10.1016/j.jag.2024.103696.

[38] Y. Zhao, K. Zhao, Z. Tang, X. Lu, Y. Zhang, Y. Du, GraphJCL: A dual-perspective graph-based framework for urban region representation via joint contrastive learning, in: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Springer, 2025. doi:10.1007/978-3-032-06066-2_3.

[39] Y. Chen, W. Huang, K. Zhao, Y. Jiang, G. Cong, Self-supervised representation learning for geospatial objects: A survey, Information Fusion (2025) 103265doi:10.2139/ssrn.5173896.

[40] Y. Li, W. Huang, G. Cong, H. Wang, Z. Wang, Urban region representation learning with openstreetmap building footprints, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023, pp. 1363–1373. doi:10.1145/3580305.3599538.

[41] Y. Tao, W. Liu, J. Chen, J. Gao, R. Li, X. Wang, Y. Zhang, J. Ren, S. Yin, X. Zhu, et al., A graph-based multimodal data fusion framework for identifying urban functional zone, International Journal of Applied Earth Observation and Geoinformation 136 (2025) 104353. doi:10.1016/j.jag.2024.104353.

[42] R. Yang, Y. Zhong, Y. Su, Self-supervised joint representation learning for urban land-use classification with multisource geographic data, IEEE Transactions on Geoscience and Remote Sensing 63 (2025) 1–21.

[43] W. Xu, J. Wang, Y. Wu, Multi-dimension geospatial feature learning for urban region function recognition, in: IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2022, pp. 5832–5835. doi:10.1109/igarss46834.2022.9884450.

[44] X. Sun, J. Gao, Y. Yuan, Alignment and fusion using distinct sensor data for multimodal aerial scene classification, IEEE Transactions on Geoscience and Remote Sensing 62 (2024) 1–11.

[45] Y. Feng, J. Jin, Y. Yin, C. Song, X. Wang, Mcft: Multimodal contrastive fusion transformer for classification of hyperspectral image and lidar data, IEEE Transactions on Geoscience and Remote Sensing 62 (2024) 1–17.

[46] J. Cao, X. Wang, J. Chen, W. Tu, Z. Li, X. Yang, T. Zhao, Q. Li, Urban representation learning for fine-grained economic mapping: A semi-supervised graph-based approach, ISPRS Journal of Photogrammetry and Remote Sensing 226 (2025) 317–331. doi:10.1016/j.isprsjprs.2025.05.007.

[47] D. J. Mühlematter, L. Che, Y. Hong, M. Raubal, N. Wiedemann, Urbanfusion: Stochastic multimodal fusion for contrastive learning of robust spatial representations, arXiv preprint arXiv:2510.13774 (2025).

[48] J. Cao, J. Chen, X. Wang, W. Huang, D. Chen, T. Zhao, W. Tu, Q. Li, Urbanmmcl: Urban region representations via multi-modal and multi-graph self-supervised contrastive learning, ISPRS Journal of Photogrammetry and Remote Sensing 232 (2026) 75–93.

[49] Y. Lu, L. Zhang, J. Liu, Q. Tian, Constructing concept lexica with small semantic gaps, IEEE Transactions on Multimedia 12 (2010) 288–299. doi:10.1109/tmm.2010.2046292.

[50] J. Zhang, T. Li, X. Lu, Z. Cheng, Semantic classification of high-resolution remote-sensing images based on mid-level features, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9 (2016) 2343–2353. doi:10.1109/jstars.2016.2536943.

[51] J. Wang, C. Gao, M.-L. Wang, Y. Zhang, Identification of urban functional areas and urban spatial structure analysis by fusing multi-source data features: A case study of zhengzhou, china, Sustainability (2023). doi:10.3390/su15086505.

[52] W. Huang, L. zhen Cui, M. Chen, D. Zhang, Y. Yao, Estimating urban functional distributions with semantics preserved poi embedding, International Journal of Geographical Information Science 36 (2022) 1905–1930. doi:10.1080/13658816.2022.2040510.

[53] Q. Qin, S. Xu, M. Du, S. Li, Identifying urban functional zones by capturing multi-spatial distribution patterns of points of interest, International Journal of Digital Earth 15 (2022) 2468–2494. doi:10.1080/17538947.2022.2160841.

[54] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[55] J. Li, D. Li, C. Xiong, S. Hoi, BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International conference on machine learning, PMLR, 2022, pp. 12888–12900.

[56] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, et al., Qwen2-VL: Enhancing vision-language

model's perception of the world at any resolution, arXiv preprint arXiv:2409.12191 (2024).

[57] G. Mai, W. Huang, J. Sun, S. Song, D. R. Mishra, N. Liu, S. Gao, T. Liu, G. Cong, Y. Hu, C. Cundy, Z. Li, R. Zhu, N. Lao, On the opportunities and challenges of foundation models for geoai (vision paper), ACM Transactions on Spatial Algorithms and Systems 10 (2024) 1–46. doi:10.1145/3653070.

[58] R. Mushkani, Do vision-language models see urban scenes as people do? an urban perception benchmark, arXiv preprint arXiv:2509.14574 (2025).

[59] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PMLR, 2020, pp. 1597–1607. arXiv:2002.05709.

[60] C. Liu, Q. Xu, H. Miao, S. Yang, L. Zhang, C. Long, Z. Li, R. Zhao, Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39, 2025, pp. 18780–18788. doi:10.1609/aaai.v39i18.34067.

[61] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778. doi:10.1109/cvpr.2016.90.

[62] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, et al., Qwen3 embedding: Advancing text embedding and reranking through foundation models, arXiv preprint arXiv:2506.05176 (2025). arXiv:2506.05176.

[63] Y. Yan, H. Wen, S. Zhong, W. Chen, H. Chen, Q. Wen, R. Zimmermann, Y. Liang, UrbanCLIP: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web, Proceedings of the ACM Web Conference 2024 (2023). doi:10.1145/3589334.3645378.

[64] T. Zhao, Z. Huang, W. Tu, F. Biljecki, L. Chen, Developing a multiview spatiotemporal model based on deep graph neural networks to predict the travel demand by bus, International Journal of Geographical Information Science 37 (7) (2023) 1555–1581. doi:10.1080/13658816.2023.2203218.

[65] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, arXiv preprint arXiv:1710.10903 (2017).

[66] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[67] D. Maulud, A. M. Abdulazeez, A review on linear regression comprehensive in machine learning, Journal of applied science and technology trends 1 (2) (2020) 140–147. doi:10.38094/jastt1457.

[68] A. Liaw, M. Wiener, et al., Classification and regression by randomforest, R news 2 (3) (2002) 18–22.

[69] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: analysis, applications, and prospects, IEEE transactions on neural networks and learning systems 33 (12) (2021) 6999–7019. doi:10.1109/tnnls.2021.3084827.

[70] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[71] B. Yu, H. Yin, Z. Zhu, Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting, arXiv preprint arXiv:1709.04875 (2017). doi:10.24963/ijcai.2018/505.

[72] S. Yang, S. Qian, Understanding and predicting travel time with spatio-temporal features of network traffic flow, weather and incidents, IEEE Intelligent Transportation Systems Magazine 11 (2019) 12–28. doi:10.1109/mits.2019.2919615.

[73] Y. Chen, Z. Zhang, T. Liang, Assessing urban travel patterns: An analysis of traffic analysis zone-based mobility patterns, Sustainability (2019). doi:10.3390/su11195452.

[74] A. Chatzimparmpas, R. M. Martins, A. Kerren, T-viSNE: Interactive assessment and interpretation of t-SNE projections, IEEE transactions on visualization and computer graphics 26 (8) (2020) 2696–2714. doi:10.1109/tvcg.2020.2986996.

[75] J. Zhang, Y. Zheng, D. Qi, R. Li, X. Yi, Predicting citywide crowd flows using deep spatio-temporal residual networks, Artificial Intelligence 259 (2018) 147–166. doi:10.1016/j.artint.2018.03.002.

[76] W. Wang, F. Zong, B. Yao, A proactive real-time control strategy based on data-driven transit demand prediction, IEEE Transactions on Intelligent Transportation Systems 22 (4) (2021) 2404–2416. doi:10.1109/TITS.2020.3028415.

[77] C. Rong, X. Zhang, Y. Xi, H. Sui, J. Ding, Y. Li, Satellites reveal mobility: A commuting origin-destination flow generator for global cities, arXiv preprint arXiv:2505.15870 (2025).

[78] T. Fontes, R. Correia, J. Ribeiro, J. L. Borges, A deep learning approach for predicting bus passenger demand based on weather conditions, Transport and Telecommunication 21 (4) (2020) 255–264. doi:10.2478/ttj-2020-0020.