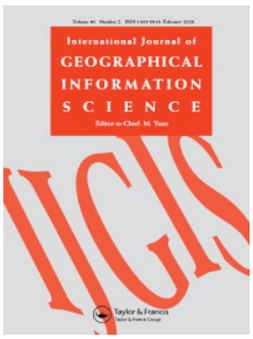

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: www.tandfonline.com/journals/tgis20

MGIM: a masked modeling framework for land parcel-level Geo-Inference

Xiang Zhang, Yao Yao, Chenglong Yu, Zhihui Hu, Geyuan Zhu, Mariko Shibasaki, Liangyang Dai, Yanduo Guo, Qingfeng Guan & Ryosuke Shibasaki

To cite this article: Xiang Zhang, Yao Yao, Chenglong Yu, Zhihui Hu, Geyuan Zhu, Mariko Shibasaki, Liangyang Dai, Yanduo Guo, Qingfeng Guan & Ryosuke Shibasaki (18 Feb 2026): MGIM: a masked modeling framework for land parcel-level Geo-Inference, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2026.2630403

To link to this article: https://doi.org/10.1080/13658816.2026.2630403

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2026.2630403

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

MGIM: a masked modeling framework for land parcel-level Geo-Inference

Xiang Zhang[a,b], Yao Yao[a,c], Chenglong Yu[a,b], Zhihui Hu[a], Geyuan Zhu[a,b], Mariko Shibasaki[b], Liangyang Dai[a], Yanduo Guo[a,b,d], Qingfeng Guan[a] and Ryosuke Shibasaki[b,e,f]

[a]UrbanComp Lab, School of Geography and Information Engineering, China University of Geosciences, Wuhan, Hubei, China; [b]LocationMind Institution, LocationMind Inc, Chiyoda, Tokyo, Japan; [c]Hitotsubashi Institute for Advanced Study, Hitotsubashi University, Kunitachi, Tokyo, Japan; [d]School of Computer Science, China University of Geosciences, Wuhan, China; [e]Faculty of Engineering, Reitaku University, Kashiwa, Chiba, Japan; [f]Interfaculty Initiative in Information Studies & Graduate School of Interdisciplinary Information Studies, The University of Tokyo, Tokyo, Japan

ABSTRACT

Effective modeling of spatio-temporal contexts to support geographic reasoning is essential for advancing Geospatial Artificial Intelligence. Inspired by masked language models, this paper introduces the Masked Geographical Information Model (MGIM), a novel self-supervised framework for learning context-aware representations from multi-source spatio-temporal data. The framework's core innovations include a parcel-scale method for multi-source data fusion and a custom self-supervised masking strategy for diverse geographic elements. This integrated modeling approach enables the model to capture complex spatio-temporal relationships and achieve consistently strong performance across diverse geographic reasoning tasks, such as trajectory inference, people flow inference, event identification, and land parcel function analysis. MGIM accurately reasons from spatio-temporal contexts and dynamically adjusts inferences according to contextual changes. The visualization of attention mechanisms further illustrates MGIM's capacity to construct contextually-aware representations and task-specific attention patterns analogous to natural language processing models. This study presents a new paradigm for general-purpose spatio-temporal modeling in real-world geographic scenarios, offering significant theoretical and practical value, and promising an effective solution for building a geographic foundation model.

ARTICLE HISTORY
Received 22 June 2025
Accepted 8 February 2026

KEYWORDS
GeoAI; multi-source data fusion; spatio-temporal contextual reasoning; Geo-Inference

1. Introduction

Cities represent complex and dynamic systems formed by interdependent interactions among human mobility, land structure, economic activity, and social processes (Wang and Biljecki, 2022). Gaining a deep understanding of these spatio-temporal

CONTACT Yao Yao yaoy@cug.edu.cn

Supplemental data for this article can be accessed online at https://doi.org/10.1080/13658816.2026.2630403.

interdependencies and reasoning about their interactions are crucial for applications such as smart city management, urban planning, and emergency response (Cao *et al.* 2025). For example, urban planners need to anticipate how new development projects may affect surrounding commercial vitality and residents' travel behaviors, while emergency managers must identify potential risk areas and respond to unexpected incidents based on real-time population movement, traffic conditions, and event information. These reasoning tasks require models capable of integrated, context-sensitive reasoning to interpret the dynamic behavior of urban systems.

Geographic reasoning refers to the process of using available spatio-temporal information to determine what to believe or what actions to take in a specific geographic context (Hooghuis *et al.* 2014). This reasoning process is analogous to the human cognitive process of perceiving the environment, analyzing situations, and solving problems (Ishikawa, 2013, Du *et al.* 2002), and it constitutes a fundamental basis for developing intelligent models capable of understanding and simulating complex geographic systems.

In existing GeoAI approaches for geographic inference, mainstream research paradigms typically decompose complex urban environments into a series of independent prediction tasks. Although such models often achieve high accuracy on specific tasks, their task-specific nature overlooks the ubiquitous dynamic interdependencies among different geographic phenomena. This task-centric simplification of the real world constrains the models' holistic understanding of cities as complex systems, thereby limiting their generalization and applicability in dynamic real-world scenarios. In the context of contemporary GeoAI and large models, geographic reasoning increasingly emphasizes a cross-element, cross-task, and transferable general reasoning capability, which fundamentally relies on a holistic understanding of the intertwined relationships among multiple elements within complex spatio-temporal contexts, going beyond inference targeting individual geographic phenomena.

Moving beyond single-target prediction toward general geographic reasoning models represents a crucial step for advancing GeoAI, as it enables the integration of multiple spatial and temporal dependencies (Janowicz *et al.* 2025). At the core of such models is the ability to transcend single-phenomenon analysis and model multi-element spatio-temporal contexts (Zhao *et al.* 2023). This necessitates a novel framework that can autonomously learn complex intrinsic relationships from multi-source data. Such a framework must build generalizable, adaptable representations of spatio-temporal contexts for diverse reasoning tasks, providing a foundation for simulating real-world environments.

1.1. Spatio-temporal modeling in GeoAI: from single-task prediction to multi-source fusion

In recent years, artificial intelligence technologies, particularly deep learning, have been widely applied in GeoAI (Janowicz *et al.* 2020). Researchers have developed specialized deep learning models for various geographic tasks, achieving notable results in trajectory prediction (Musleh *et al.* 2022), human and traffic flow forecasting (Ali *et al.* 2022), urban anomaly detection (Sharif *et al.* 2025), and urban functional zone

identification (Hu *et al.* 2023). However, these studies generally follow a paradigm that decomposes complex geographic problems into well-defined single-target supervised learning tasks. The limitation is that models are typically optimized for specific tasks. They struggle to capture the frequent interactions among geographic phenomena, and this constrains their holistic understanding of urban systems.

To overcome the limitations of single data sources, researchers have begun exploring the integration of multi-source heterogeneous data to enhance model performance. For instance, in urban functional zone identification, incorporating economic factors (Tu *et al.* 2024), transportation elements (Kanyepe *et al.* 2021), and human activity data can significantly improve the accuracy of land-use classification. Similarly, in human mobility prediction, integrating geographic semantic information has been shown to enhance the precision of next-location forecasting (Yao *et al.* 2023). These studies collectively highlight the value of multi-source data fusion in GeoAI applications.

Nevertheless, most existing data fusion approaches mainly rely on feature concatenation or treat additional data as auxiliary variables to improve the prediction accuracy of a predefined target. This approach is essentially a prediction-oriented, unidirectional form of data fusion and does not explicitly model the intrinsic, bidirectional, and complex spatio-temporal dependencies among different geographic phenomena (Choudhury *et al.* 2024). In reality, the spatio-temporal context represents an integrated system of geographic processes in which human activities, land use, and economic dynamics interact and evolve together (Li *et al.* 2022). Therefore, new approaches are needed that go beyond single-target prediction frameworks, enabling models to learn and represent complex spatio-temporal contexts.

1.2. Insights from NLP: masked self-supervised learning and contextual understanding

In recent years, natural language processing (NLP) technologies, particularly large language models, have achieved revolutionary breakthroughs, demonstrating strong capabilities in understanding complex textual contexts. The success of NLP has largely been attributed to the masked self-supervised learning paradigm (Devlin *et al.* 2019). By randomly masking portions of text in massive corpora and training models to predict the masked content based on context, these models can learn the semantic and syntactic relationships among words (Zhu *et al.* 2021).

The masked self-supervised learning paradigm provides valuable insights for developing geospatial models with spatio-temporal contextual understanding. Prior studies have demonstrated structural and semantic similarities between geospatial data and natural language text (Huang *et al.* 2025). For instance, urban regions can be conceptualized as documents, different land-use types as topics, and the spatial distribution of points of interest as word sequences within a document (Yao *et al.* 2017). Individual mobility trajectories can be regarded as sequences of spatio-temporal tokens (Musleh *et al.* 2022). These studies support the feasibility of drawing analogies between geospatial structures and linguistic constructs, opening new avenues for developing models capable of understanding complex spatio-temporal contexts.

1.3. Existing attempts and limitations of geospatial masked models

Inspired by advances in NLP, some studies have begun exploring the application of masked self-supervised learning to geospatial modeling. For example, Zhang *et al*. (2025) proposed a masked learning framework for region-focused learning in traffic flow prediction, while Yang *et al*.(2025a) introduced a masked reconstruction approach for the completion of DEM data. The UniST model (Yuan *et al*. 2024) also employs a mask-based spatio-temporal contextual modeling mechanism, demonstrating generalization across multiple urban spatio-temporal prediction tasks. These studies suggest that self-supervised learning is an effective approach to enhancing the reasoning capabilities of geospatial models. However, their task objectives are primarily focused on single numerical regression tasks, limiting applicability to broader geographic reasoning tasks, such as functional zone identification or event inference. Moreover, they typically rely on irregular grids that may fragment geographic entities and adopt masking strategies along a single dimension, overlooking the intrinsic relationships among multiple attributes present in real-world scenarios.

Although GeoAI has made preliminary progress in multi-source data fusion and self-supervised learning, there remains an urgent need for a general spatio-temporal contextual model that can overcome existing limitations. Current studies exhibit shortcomings in the semantic representation of spatial units, task generalization, and multi-dimensional masking strategies. These limitations constrain the models' capacity for comprehensive understanding and reasoning within complex geographic systems. The proposed MGIM framework aims to fill this gap. The main contributions of this work are as follows:

1. A novel framework for general geographic reasoning: We design and implement MGIM, a new self-supervised framework that operationalizes the masked-modeling paradigm for GeoAI. Its core innovations are a parcel-scale fusion method that aligns diverse, multi-source spatio-temporal data and a custom multi-element masking strategy specifically designed to capture the complex interdependencies among these geographic elements.
2. Validation of generalization across diverse tasks: We validate the effectiveness and high generalization of MGIM. A single pre-trained MGIM achieves strong performance across multiple geographic reasoning tasks, including trajectory location, people flow, urban event, and land parcel function inference.
3. Evidence of spatio-temporal contextual understanding: We provide evidence that MGIM moves beyond simple pattern fitting to acquire a functional understanding of spatio-temporal contexts. Through task-adaptive attention analysis and dynamic reasoning experiments, we show that the model can learn the functional implications of context on geographic processes, suggesting a semantic reasoning capability analogous to that of language models.

2. Study area and data

This study focuses on the 23 special wards of Tokyo, Japan, a region spanning approximately 627.5 km^2 with a population of around 9.7 million. As the political,

economic, and cultural center of Japan and a quintessential global metropolis, the 23 wards of Tokyo are characterized by a vast scale, functional complexity, and high population density. These attributes make the area a representative study case of large-scale urban development, ideal for investigating spatial distribution patterns and the functional evolution of major urban centers.

The following datasets support our analysis. The POI data is derived from the Telepoint Pack DB (Yellow Pages), which provides multiple attributes including the POI name, telephone number, address, postal code, category and other information. As shown in Table S1, several key attributes are selected for display. The POIs are categorized into 25 types, such as education, shopping, food service, among others. Within the study area, a total of 407,020 POIs are included, and their spatial distribution is illustrated in Figure 1a.

The fundamental spatial unit for analysis is the land parcel, which was generated by segmenting the area based on the OpenStreetMap (OSM) road network. After removing parcels that did not contain any POIs, a total of 22,455 parcels were retained for the study. The distribution of parcel areas after division is shown in Figure 1b, with a median value of 6,036 m^2.

The trajectory data were obtained from the 'Konzatsu-Tokei (R)' dataset, consisting of mobile device location information for 1,271,557 users within the study area throughout May 2024 (Li *et al.* 2025). To focus on macroscopic movements between land parcels and protect individual privacy, the trajectory data were preprocessed. Duplicate and abnormal points were first removed, and then each trajectory point was matched to its corresponding parcel. For consecutive points falling within the same parcel, only the entry and exit records were retained. This procedure simplifies intra-parcel activities, eliminates positioning errors and redundancies (Lin *et al.* 2025), and extracts key inter-parcel travel chains to support the analysis of spatial interactions at the city scale.

By spatially matching the trajectory data with the land parcel data, the hourly people flow dynamics for each parcel throughout May were calculated. As depicted in Figure 1c, the hourly people flow data exhibit a pronounced long-tail distribution, indicating that a small minority of parcels accommodate most of the human activity.

3. Methodology

In Figure 2, the Masked Geographic Information Model (MGIM) proposed in this study comprises five key modules: (1) Spatio-temporal Alignment of Multi-source Data: At the parcel scale, multi-source spatio-temporal data are aligned to construct trajectory point sequences that integrate diverse information sources. (2) Spatio-temporal Masking Strategy: A variety of masking methods are designed to target the multi-source information associated with each trajectory point, thereby generating diverse self-supervised learning tasks. (3) Spatio-temporal Element Encoding: The input spatio-temporal elements are uniformly encoded to obtain their corresponding embedding representations. (4) Feature Fusion and Trajectory Reconstruction: The encoded features are fused, and a Transformer architecture is employed to infer the masked information, thereby capturing complex spatio-temporal contextual

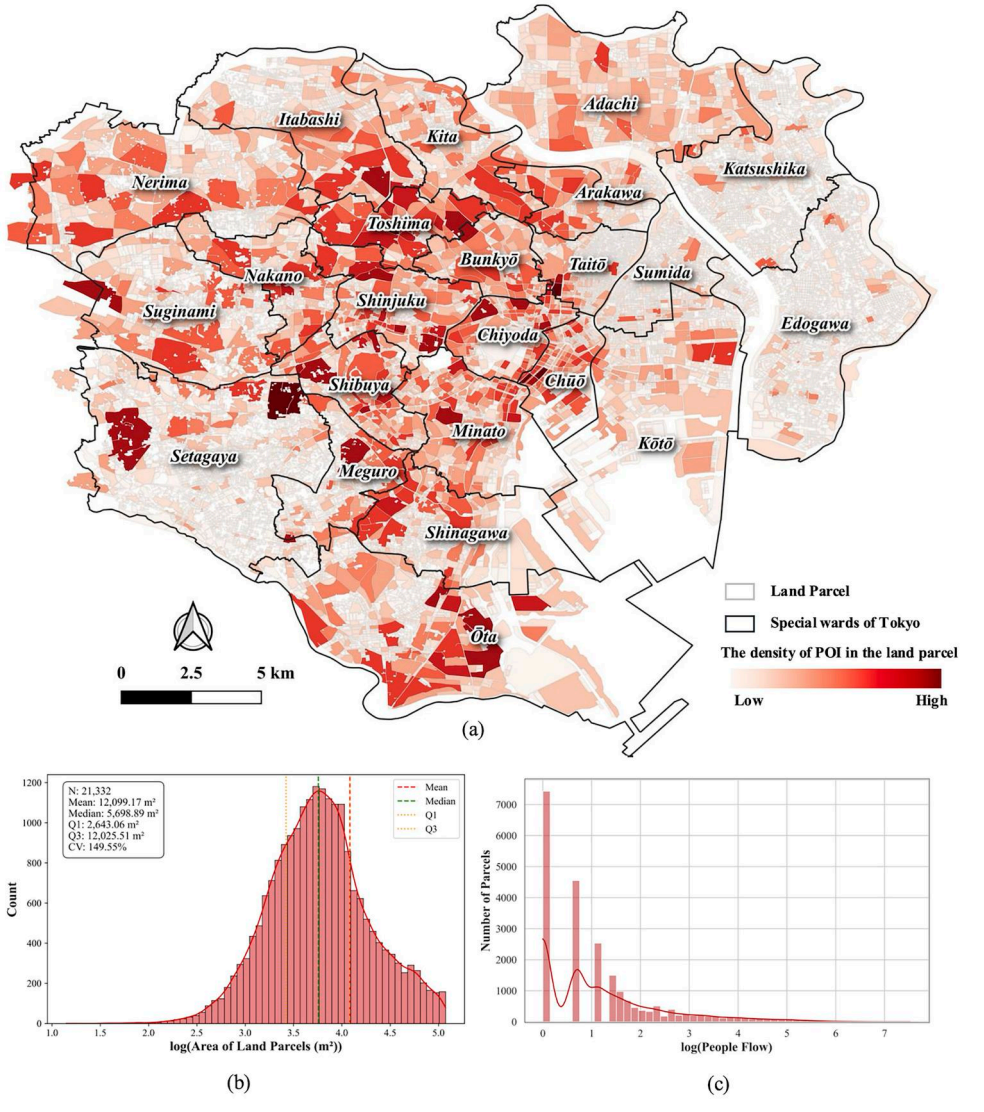

Figure 1. (a) The Study Area: Tokyo's 23 Special Wards and the Density of POI Distribution at the Parcel Scale. (b) Statistical Distribution of Land Parcel Areas. (c) Distribution of Parcel-Level People Flow in the Study Area.

relationships. (5) Multi-source Information Decoding: Multiple decoders are utilized to reconstruct the various types of spatio-temporal information from the reconstructed trajectory vectors.

3.1. Multi-source spatio-temporal data alignment

To align the multi-source, heterogeneous data, this study employs land parcels as the fundamental spatial units for analysis. This parcel-based approach is chosen over traditional grid-based partitioning to avoid the arbitrary splitting of real-world geographic features.

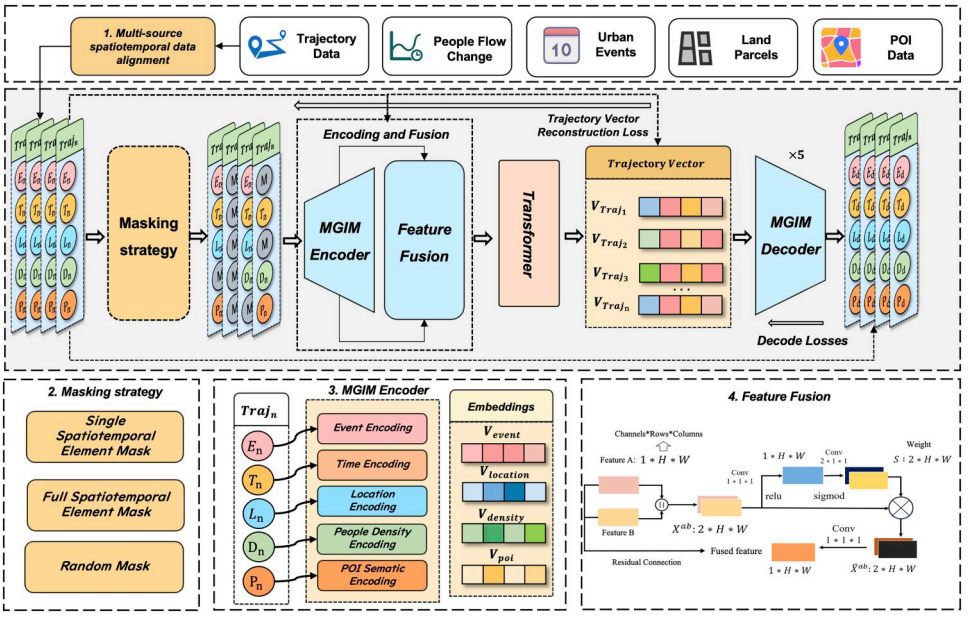

Figure 2. Framework of MGIM.

As illustrated in Figure 3, to construct structured inputs for embedding and relationship modeling, we first align the multi-source spatio-temporal data. A raw trajectory sequence is consequently transformed into an augmented trajectory sequence.

We define the augmented trajectory sequence as $S = \langle p_1, p_2, \cdots, p_n \rangle$, where n is the sequence length. Each p_n in this sequence is an augmented trajectory point, representing a multi-dimensional feature representation of the contextual information associated with the original point, generated by aligning it to its corresponding land parcel. This augmented trajectory point p_n is defined as: $p_n = (E_n, T_n, L_n, D_n, P_n)$, where the components are defined as follows:

- E_n denotes the urban event information associated with the trajectory point. In this study, urban events are simplified as a binary variable distinguishing between workdays and holidays.
- T_n represents the discrete timestamp of the trajectory point.
- L_n corresponds to the coordinates of the matched parcel center.
- D_n refers to the population density within matched parcel during the time T_n.
- P_n denotes the set of POIs within the matched parcel.

The p_n is designed to encode not only time (T_n) and location (L_n), but also the event context (E_n), population density (D_n), and functional attributes (P_n). The augmented trajectory sequence S serves as the input for our subsequent models.

3.2. Masking strategy for multiple spatio-temporal elements

The model is based on a masked self-supervised learning approach, analogous to that of Masked Language Models (MLMs). However, a fundamental challenge arises from

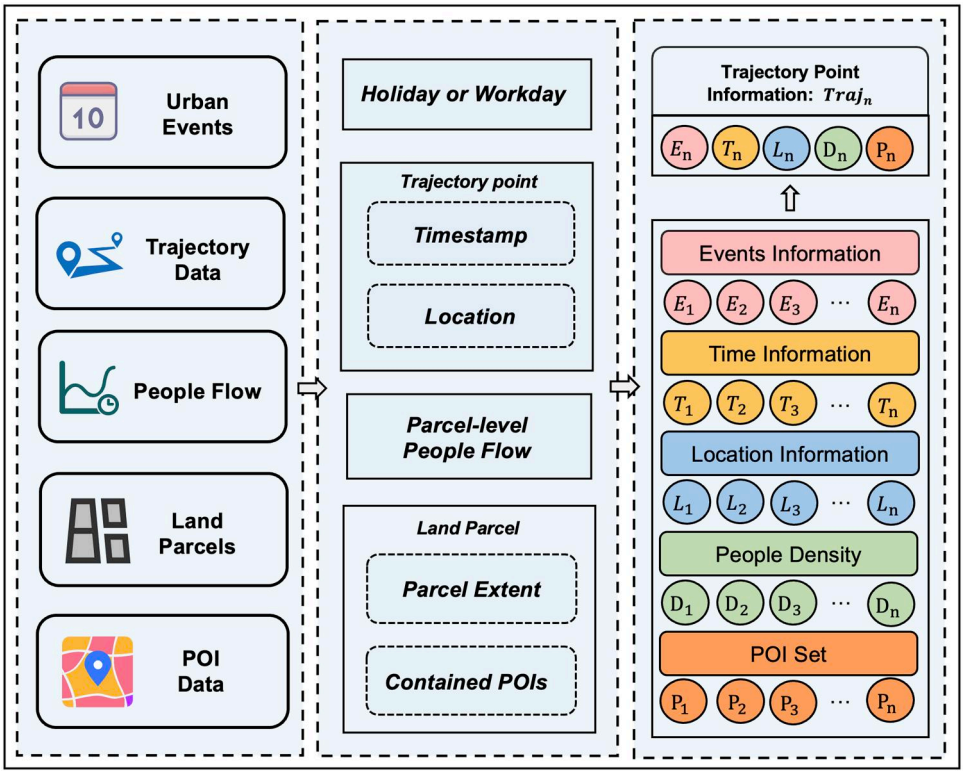

Figure 3. Spatio-temporal alignment of multi-source data at the parcel scale.

the data's structure: while MLMs mask individual tokens in a sequence, each trajectory point in this work comprises a set of heterogeneous spatio-temporal attributes. Consequently, a naive strategy of masking an entire trajectory point as a single unit would impede the model's ability to learn the complex interdependencies among these intra-point attributes (Choudhury *et al.* 2024). To overcome this limitation, this paper proposes a set of fine-grained masking strategies.

Three masking strategies are used: (1) Single Element Spatio-temporal Masking: The POI, urban event, people flow, time, and location information associated with a trajectory point are each masked independently. This allows for an examination of how the absence of a single element impacts the model's predictive capabilities. (2) Full Element Spatio-temporal Masking: All spatio-temporal attributes contained within a trajectory point are masked together as a single unit. This simulates the scenario of a completely missing trajectory point and enhances the model's ability to recover and predict missing points. (3) Random Masking: Within a sequence of trajectory points, either single-element masking or full-element masking is randomly applied. This strategy improves the model's robustness against complex patterns of missing data. The masking process is mathematically expressed as:

$$\text{Mask}\ (S) = (1 - M) \odot S + M \odot (-1) \tag{1}$$

$$S = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix} = \begin{pmatrix} E_1 & T_1 & L_1 & D_1 & P_1 \\ E_2 & T_2 & L_2 & D_2 & P_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ E_n & T_n & L_n & D_n & P_n \end{pmatrix} \tag{2}$$

$$M = \begin{pmatrix} m_{E_1} & m_{T_1} & m_{L_1} & m_{D_1} & m_{P_1} \\ m_{E_2} & m_{T_2} & m_{L_2} & m_{D_2} & m_{P_2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ m_{E_n} & m_{T_n} & m_{L_n} & m_{D_n} & m_{P_n} \end{pmatrix} \tag{3}$$

m is a binary mask indicator where a value of 1 signifies a masked element and 0 an unmasked element. The masked elements are subsequently replaced with a special value whereas unmasked elements retain their original information.

3.3. Multi-Spatio-temporal element encoding

An appropriate encoding and embedding methodology is crucial for enhancing a deep learning model's ability to perceive the heterogeneity and spatio-temporal dependencies inherent in multi-source data. This is required for the subsequent phases of mask-based feature selection and interaction modeling. In this study, distinct encoding methods are employed for different types of spatio-temporal elements, which are detailed in the following four subsections. After encoding, each of the features is processed by its respective linear layer to be projected into a common dimensionality, after which they undergo feature fusion.

3.3.1. Parcel functional attribute encoding

The Semantic2Vec method (Huang *et al.* 2022) is used to vectorize POI categories, thereby obtaining an embedding vector representation for each category. To create a parcel-level representation, the embedding vectors of all POIs located within the boundaries of a given parcel are aggregated by summation. The resulting vector is then L2-normalized to generate the final functional attribute encoding for that parcel. The specific calculation method is:

$$V_{poi} = \frac{\sum_{i=1}^{n} e_i}{\left|\left|\sum_{i=1}^{n} e_i\right|\right|} \tag{4}$$

e_i represents the embedding vector of the i-th POI within the parcel, n is the total number of POIs in the parcel, and V_{poi} is the normalized functional attribute vector for the parcel.

3.3.2. Urban event and temporal encoding

The cyclical nature of hourly information is captured using sine and cosine functions, which map discrete hours into a continuous vector space. The temporal encoding vector V_{time} is mathematically expressed as:

$$V_{time} = \left[\sin\left(\frac{2\pi \cdot hour}{24}\right), \cos\left(\frac{2\pi \cdot hour}{24}\right)\right] \tag{5}$$

hour denotes the hour of the day (ranging from 0 to 23) for the current trajectory point's timestamp. This cyclic encoding preserves temporal continuity and prevents discontinuities caused by treating time as a scalar variable. Concurrently, to differentiate between holidays and workdays, an event flag vector V_{event} is defined as follows:

$$V_{event} = \begin{cases} 1 \ if \ \ event \neq holiday \\ 2 \ if \ \ event = holiday \end{cases} \tag{6}$$

By multiplying the temporal encoding vector with the event flag vector, a joint encoding result is obtained. This result is then passed through a linear transformation function $f(*)$ to compute the temporal embedding for the event:

$$E = f(V_{time} \odot V_{event}) \tag{7}$$

The output of this encoding module, E is a fixed-dimension feature representation. This encoding captures temporal periodicity and increases model sensitivity to urban events, thereby providing a robust feature foundation for subsequent multi-source fusion and prediction tasks.

3.3.3. People flow encoding

To characterize people flow patterns, this study constructs a time-series sequence $(d_1, d_2 \ldots d_{24})$ based on the people flow data for the parcel associated with the current trajectory point over the preceding 24 hours. Here, d_t represents the number of people in the parcel during the t-th hour. The LSTM is used to model temporal patterns in people flow. This produces $V_{density}$, an encoding vector representing the parcel's people flow feature. Its mathematical expression is as:

$$V_{density} = LSTM(d_1, d_2 \ldots d_{24}) \tag{8}$$

The LSTM effectively captures the dynamic evolutionary patterns of people flow, generating stable time series embedding representations (Yu *et al.* 2019). The $V_{density}$ provides information about human mobility for the MGIM.

3.3.4. Location information encoding

To ensure numerical stability and capture fine-grained spatial variations, land parcel locations are represented using projected offsets rather than raw coordinates. All trajectory points are projected onto the UTM Zone 54 N system (EPSG:32654), which is appropriate for Tokyo. To minimize issues related to numerical precision, coordinates are stored as offsets relative to the southwest corner of the study area's bounding box:

$$V_{\text{pos}} = (X_i - X_{\min},\ \mathrm{Y_i} - \mathrm{Y_{min}}) \tag{9}$$

$(X_i,\ \mathrm{Y_i})$ represents the planar coordinates of a trajectory point i in the projected coordinate system. $(X_{\min},\ \mathrm{Y_{min}})$ represents the coordinates of the southwest corner of the study area's bounding box. V_{pos} is the resulting relative positional encoding for the trajectory point.

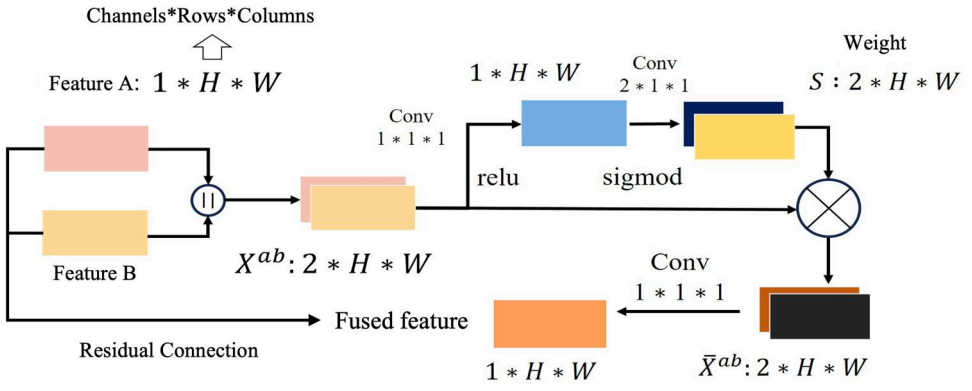

Figure 4. Feature fusion module.

3.4. Feature fusion module

The feature fusion module is designed to integrate complementary information from multi-source data and mitigate the limitations of single-feature representations. Feature aggregation is first performed through channel-wise concatenation, followed by convolutional layers and activation functions that learn adaptive feature weights. These weights modulate the fused representation, enhancing the model's expressive capacity and generalization. A residual connection is incorporated to preserve original information and prevent feature degradation. The module architecture is illustrated in Figure 4.

3.5. Loss functions definition

A multi-task loss function is designed to jointly optimize diverse objectives, including trajectory vector completion and multiple decoding tasks, arising from our multi-spatio-temporal masking strategy. The trajectory vector reconstruction loss $\mathcal{L}_{mgm}$ primarily updates the model modules responsible for completing missing information, including the feature encoding, feature fusion, and Transformer modules. It is mathematically defined as the Euclidean distance between the predicted and ground-truth trajectory vectors:

$$\mathcal{L}_{mgm} = \frac{1}{N}\sum |P_{traj_v} - L_{traj_v}|_2 \tag{10}$$

The P_{traj_v} represents the model's output for the reconstructed trajectory vector. The L_{traj_v} is the ground-truth trajectory vector. N is the total number of samples.

To further optimize the inference performance for different types of information, this study applies several decoding loss functions, each targeting its corresponding encoding module. For temporal information prediction, the decoder loss uses the Mean Absolute Percentage Error (MAPE), which is defined as:

$$\mathcal{L}_{time} = \frac{\sum Mape(P_{time}, L_{time})}{N} \tag{11}$$

P_{time} is the predicted time value and L_{time} is the ground-truth time value.

The loss for the location information decoding module is measured by the distance between the predicted and ground-truth locations:

$$\mathcal{L}_{loc} = \frac{\sum \sqrt{(P_x - l_x)^2 + (P_y - l_y)^2}}{N} \tag{12}$$

P_x, P_y are the predicted location coordinates output by the model, while l_x, l_y are the ground-truth location coordinates.

Given the prominent long-tail distribution of people flow data, using Mean Squared Error (MSE) can cause the model to be dominated by a small number of high-value samples, thereby degrading generalization performance. Therefore, we employ the Huber loss. The Huber loss function provides the fine-grained optimization of a squared loss for small errors while using a linear loss for large errors to reduce the impact of high-value outliers. This approach improves the model's robustness to data imbalance, and is defined as follows:

$$\mathcal{L}_{peo} = \begin{cases} \dfrac{1}{2(P_{peo} - L_{peo})^2} & if \left|P_{peo} - L_{peo}\right| \leq \delta \\ \delta \cdot \left(\left|P_{peo} - L_{peo}\right| - \dfrac{\delta}{2}\right) if \left|P_{peo} - L_{peo}\right| > \delta \end{cases} \tag{13}$$

P_{peo} represents the predicted people flow value. L_{peo} is the actual observed people flow value. δ is a hyperparameter used to control the robustness of the loss function.

The urban event decoding module involves a classic classification task. Therefore, this study uses the Cross-Entropy Loss function to classify the event types:

$$\mathcal{L}_{event} = \frac{\sum CrossEntropy(P_{event}, L_{event})}{N} \tag{14}$$

P_{event} is predicted probability distribution over the event classes. L_{event} is the ground-truth class label.

The land parcel functional attribute decoding module focuses on learning continuous semantic representations rather than discrete classes. Therefore, this study employs the Cosine Similarity Loss to measure the consistency between the predicted and ground-truth functional attribute vectors:

$$\mathcal{L}_{func} = 1 - \frac{\sum CosineSim(P_{func}, L_{func})}{N} \tag{15}$$

where P_{func} denotes the predicted functional attribute vector, and L_{func} represents the ground-truth vector.

3.6. Experimental setup and evaluation strategy

To evaluate the effectiveness of the MGIM framework, we randomly partitioned the trajectory sequences into training and testing sets at a 9:1 ratio. To prevent data leakage, we ensured that all trajectories belonging to a single user were restricted to only one of the sets. For the model configuration, a pre-training masking probability of 0.2 was selected. The empirical justification for this choice is detailed in Appendix B. Furthermore, to qualitatively demonstrate the model's performance across diverse

tasks, we extracted a representative three-day trajectory sequence from the test set for case studies and visualization. This sample covers both weekdays and holidays and includes distinct residential and outdoor activities, providing a robust basis for interpretative analysis. Detailed data for this sample is available in Table S2.

To evaluate the model's performance across diverse geographic reasoning tasks, we employ a suite of task-specific metrics. For numerical regression tasks, including trajectory vector reconstruction, location inference (m), and time inference (h), we measure performance using the Euclidean distance or Mean Absolute Error to quantify physical deviations. For people flow inference, both Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) are reported to account for the data's long-tail distribution. And for semantic tasks, we use classification accuracy (%) for urban events and cosine similarity for land parcel functions.

4. Results

The results are organized in two main parts. Section 4.1 provides a quantitative evaluation of model performance, analyzing the pre-training convergence and the impact of varying masking ratios on inference tasks. Section 4.2 presents case studies and visualization, utilizing representative trajectory sequences to illustrate specific inference outcomes and attention patterns.

4.1. Model training and quantitative performance evaluation

4.1.1. Pre-training results under the Random masking strategy

Figure 5 illustrates the pre-training convergence using a random masking strategy. As training progresses, the loss curves for all six tasks consistently decrease and trend towards convergence. These results indicate the model's convergence and stability.

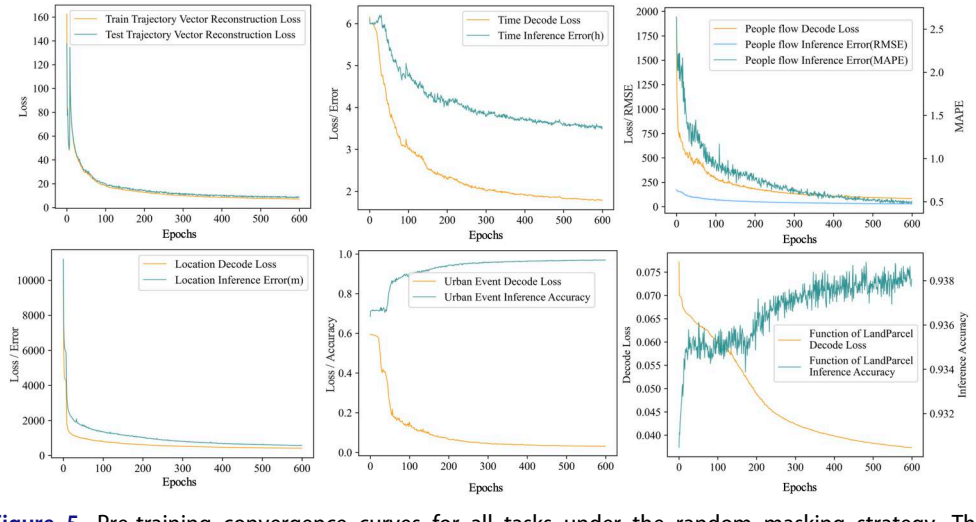

Figure 5. Pre-training convergence curves for all tasks under the random masking strategy. The plots show training loss versus test error/accuracy for trajectory vector reconstruction and the five spatio-temporal element inference tasks.

Table 1. Final loss and performance of the pre-trained model.

Task type	Train loss	Performance
Trajectory vector reconstruction ($\mathcal{L}_{mgm}$)	7.11	7.15
People flow	85.79	MAPE: 0.48 RMSE: 30.4 people
Trajectory point location	415.21	Mean:575.38 m
Time	1.79	3.55 h
Urban event	0.03	96.98%
Function of land parcel	0.04	0.94

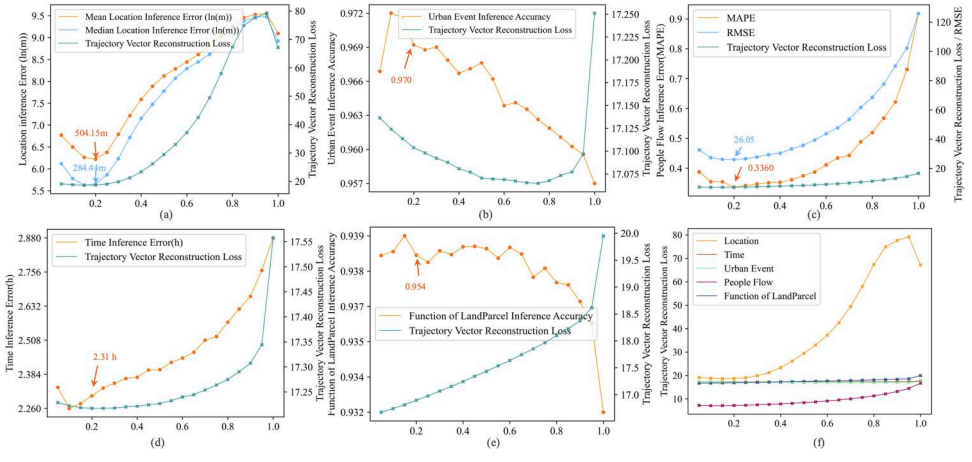

Figure 6. Model performance on spatio-temporal inference tasks and the corresponding change in trajectory vector reconstruction loss across different masking probabilities. The x-axis represents the masking probability.

Table 1 presents the model's final loss values and performance metrics on the test set after pre-training. The model demonstrates robust performance across all objectives, achieving low inference errors in continuous spatio-temporal tasks while maintaining high fidelity in semantic inference tasks. The results validate the effectiveness of the proposed method and the MGIM's multi-task inference capability.

4.1.2. Evaluation of model performance for inference tasks

Based on the pre-trained model, we conducted few-shot fine-tuning tests across five inference tasks. Compared to the pre-training baseline, the fine-tuned model exhibited consistent performance improvements across all metrics. As presented in Figure 6 and Table 2, the model demonstrated robust performance at a 0.2 masking probability, achieving high accuracy in semantic tasks and low error in numerical spatio-temporal inferences. Notably, the significant disparity between the mean and median location inference errors suggests that the overall metric is disproportionately skewed by a small number of high-error outliers, a phenomenon further examined in the discussion section.

Figure 6 illustrates a consistent performance degradation across all tasks as the masking probability increases. Inference errors for location (a), time (d), and people flow (c) reach their minimum at a masking probability of 0.2–0.3 before rising sharply. In the case of urban events (b) and parcel functions (e), accuracies peak within the

Table 2. Results across multiple geographic inference tasks at 0.2 masking probability.

Inference tasks	Performance
People flow	MAPE: 0.336 RMSE: 26.05 people
Trajectory point location	Mean: 504.15 m Median: 284.44 m
Time	2.31h
Urban event	97.00%
Function of parcel	0.95

0.1–0.25 range and subsequently follow a fluctuating downward trend. These results show that MGIM retains robust reasoning capabilities until the context becomes excessively sparse. The model maintains notable resilience in tasks (b) and (e), demonstrating the ability to infer semantic information even under high masking ratios.

The relative importance of each spatio-temporal element is determined by comparing their impact on trajectory vector reconstruction loss under various masking proportions. Figure 6f highlights the hierarchy of influence among elements. Location information has the most significant impact, followed by time, urban events, and parcel functional attributes. In contrast, people flow exerts a less significant influence, although its associated loss consistently increases with the masking probability.

4.2. Case studies and visualization across different tasks

To further demonstrate the model's performance across different tasks, we randomly selected a real trajectory sequence of a user from the preprocessed test set for case study and visualization. The selected trajectory spans three days and contains a total of 44 trajectory points, covering both weekdays and a holiday. It includes clearly identifiable residential locations and outdoor activities, making it representative and suitable for interpretative analysis. All results presented below are based on this sample. Detailed information about the trajectory sequence can be found in the appendix Table S2.

4.2.1. Inference result for trajectory location

With the masking probability set to 0.2, the model was tasked with inferring the masked locations. Figure 7 shows that 87.5% of the masked points were correctly restored to their original parcels. Only a single point (T-id 33) exhibited a minor deviation, landing adjacent to its true parcel. The result demonstrates that the proposed model can accurately infer masked trajectory locations.

Visualizing the Transformer's attention scores (Figure 8) reveals that when reconstructing a trajectory vector, the model does not just focus on the current point but also considers information from other key points in the sequence. In this location inference example, the model paid significant attention to trajectory points with IDs 8, 13, 19, 24, and 28. An analysis based on the raw data shows that points 8, 13, 30, and 28 are in areas with sparse crowds, indicating the model can identify features in an individual's trajectory that deviate from mainstream patterns. At the same time, the point (ID 24) with the highest attention score is in a subway station area, which suggests the model is highly sensitive to important geographical locations in a person's daily travel.

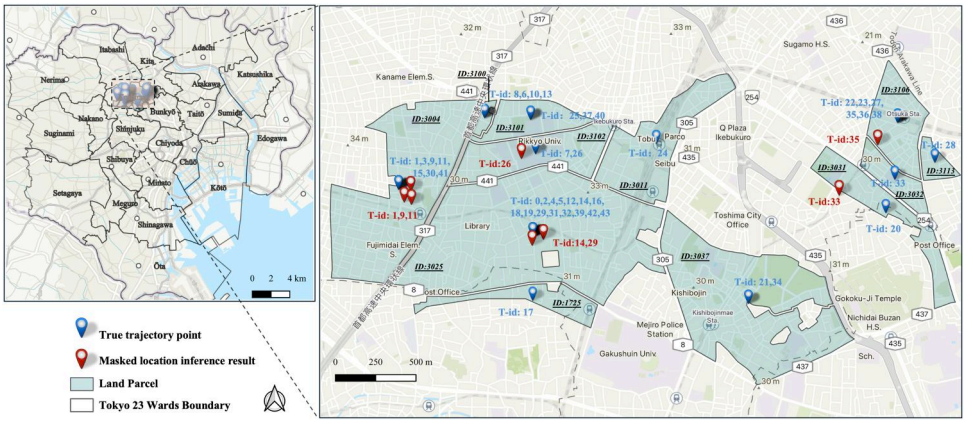

Figure 7. Visualization of Trajectory Location Inference Result. Blue markers represent true trajectory points, while red markers represent the model's inferred results after the true values were masked. The ground-truth trajectory data is detailed in Appendix Table S2.

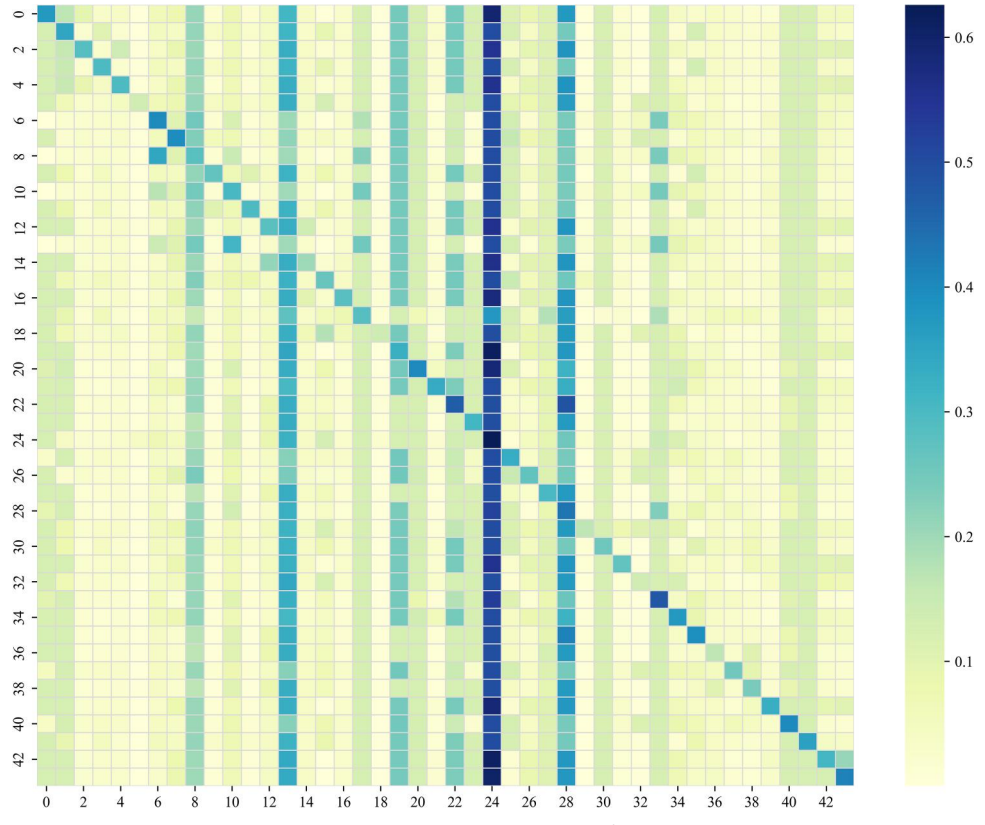

Figure 8. Attention scores in the trajectory point location inference task. The horizontal and vertical axes represent trajectory point IDs. See Appendix Table S2 for trajectory point details.

4.2.2. Inference result for parcel people flow

The model's performance on people flow inference (Figure 9) was evaluated on the same sample. For masked parcels, the inferred values align closely with the ground-

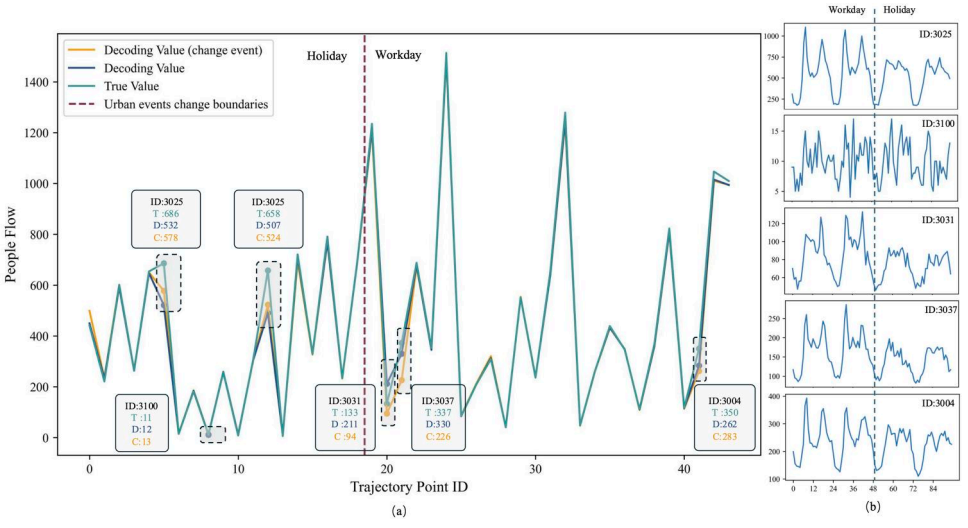

Figure 9. Inference results for parcel people flow: (a) People flow of the parcel visited at each specific time point of the trajectory. (b) Historical people flow values for masked parcels, showing patterns for both holidays and weekdays.

truth data, demonstrating the model's proficiency in people flow imputation. For unmasked parcels, the reconstructed values showed high fidelity to the original data. This result suggests that the model's information encoding and feature fusion components effectively capture and preserve the dynamics of people flow.

To further assess the model's understanding of urban event contexts, the input urban event type attribute was altered to observe the impact on people flow inferences. The model's inferences correctly reflected real-world patterns: changing a holiday to a weekday increased the predicted people flow for relevant parcels (e.g. ID 3025 from 532 to 578), while the reverse change decreased flow for locations typically quieter on holidays (e.g. IDs 3031, 3037, 3004). Additionally, for parcels with historically stable people flow regardless of event type (e.g. ID 3100), the predicted value differed from the observation by only one person.

The attention scores for people flow inference reveal the model's reasoning process (Figure 10). Similar to the location inference task, the model looks beyond the current point, allocating high attention scores to areas of significant human activity, such as train stations. Concurrently, the model also focuses on infrequent visit areas for the user (e.g. trajectory point ID 13), which reflects its ability to capture unique individual behavioral patterns. A key difference from the location completion task is the model's added focus on trajectory point ID 32 which the source data identifies as the user's residence. Since residential areas are often key origin/destination points and exhibit regular temporal patterns, this focus highlights MGIM's ability to model the spatio-temporal dependencies of core behavioral nodes.

4.2.3. Inference result for urban events

The urban event inference task was conducted on the same test sample, with the results presented in Table 3. The model achieved 100% accuracy under two distinct

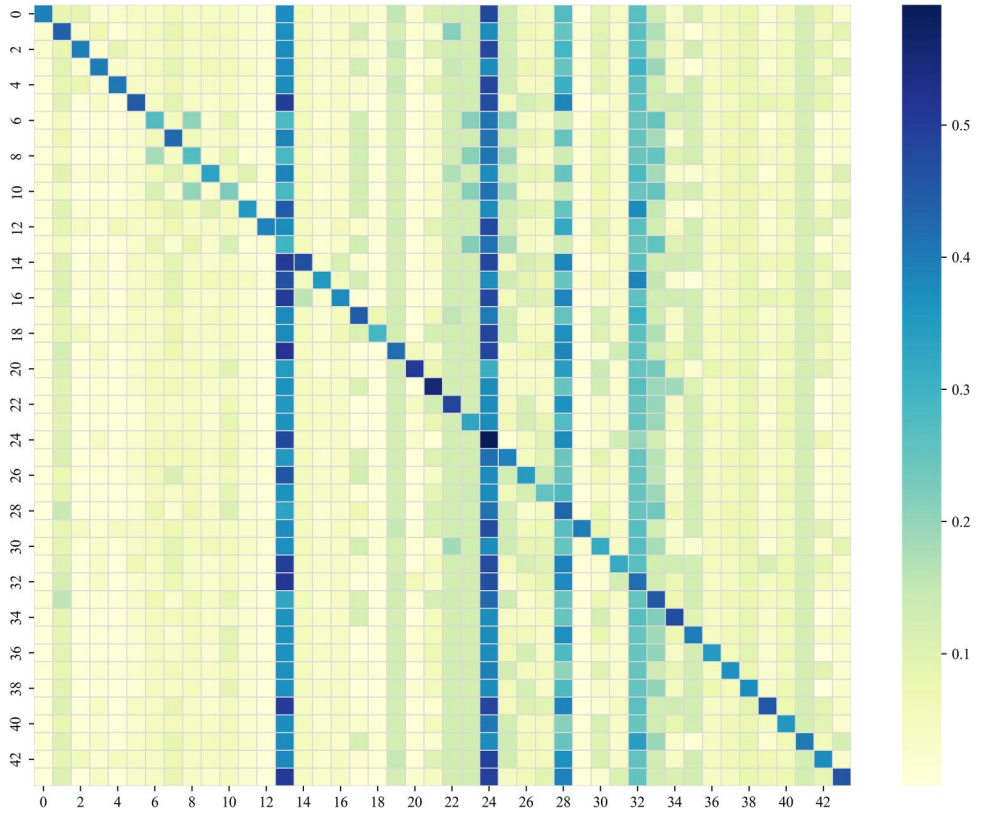

Figure 10. Attention scores in the parcel people flow inference task.

Table 3. Inference results for urban events.

Sample	Trajectory points during holidays	Trajectory points during workday
Ground truth	id: 0-18	id: 19-43
Inference results at 20% masking probability	id: 0-18	id: 19-43
Inference results at 100% masking probability	id: 0-18	id: 19-43

masking probabilities: 0.2 and 1.0. This accuracy, particularly under complete urban event information masking, demonstrates that urban events can be reliably inferred from other available spatio-temporal information, such as trajectory location and parcel people flow.

Observing the attention scores reveals a key difference in this task. In contrast to other tasks that prioritize spatial information, this task demonstrates the model's heightened sensitivity to temporal dynamics. As illustrated in Figure 11, the model focuses most on trajectory point IDs 19–22. The original data confirms that these points are located precisely at the transition between a holiday and a weekday. This targeted attention suggests that the model has learned to identify critical turning points within urban events.

4.2.4. Inference result for time information

In the time information inference task, the model's output demonstrates a high degree of accuracy. As shown in Figure 12, the decoded values for unmasked time points

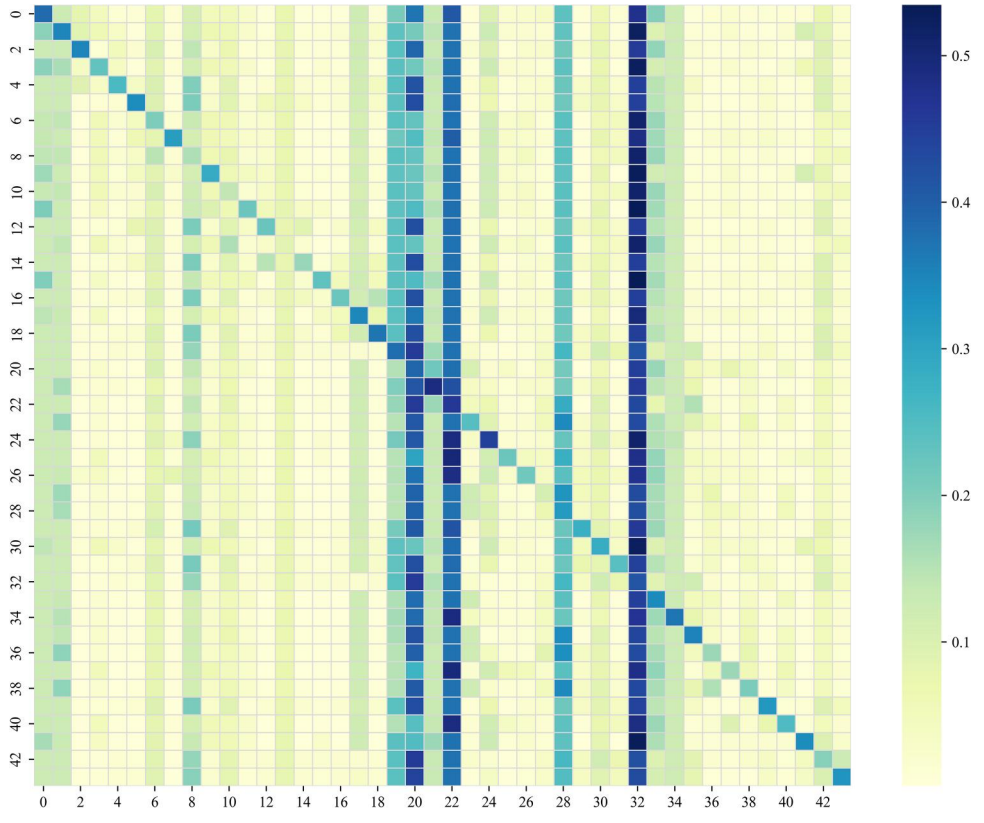

Figure 11. Attention scores in the urban event inference task.

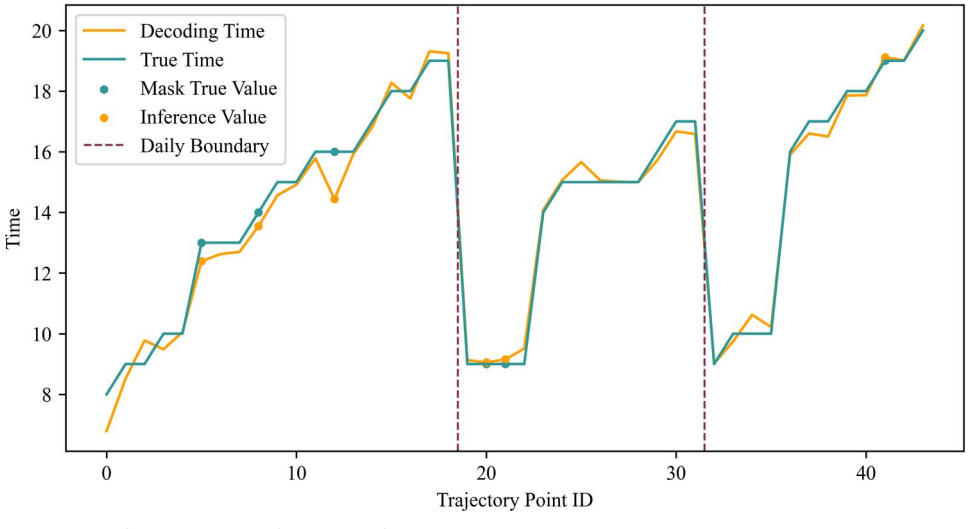

Figure 12. Inference results for time information.

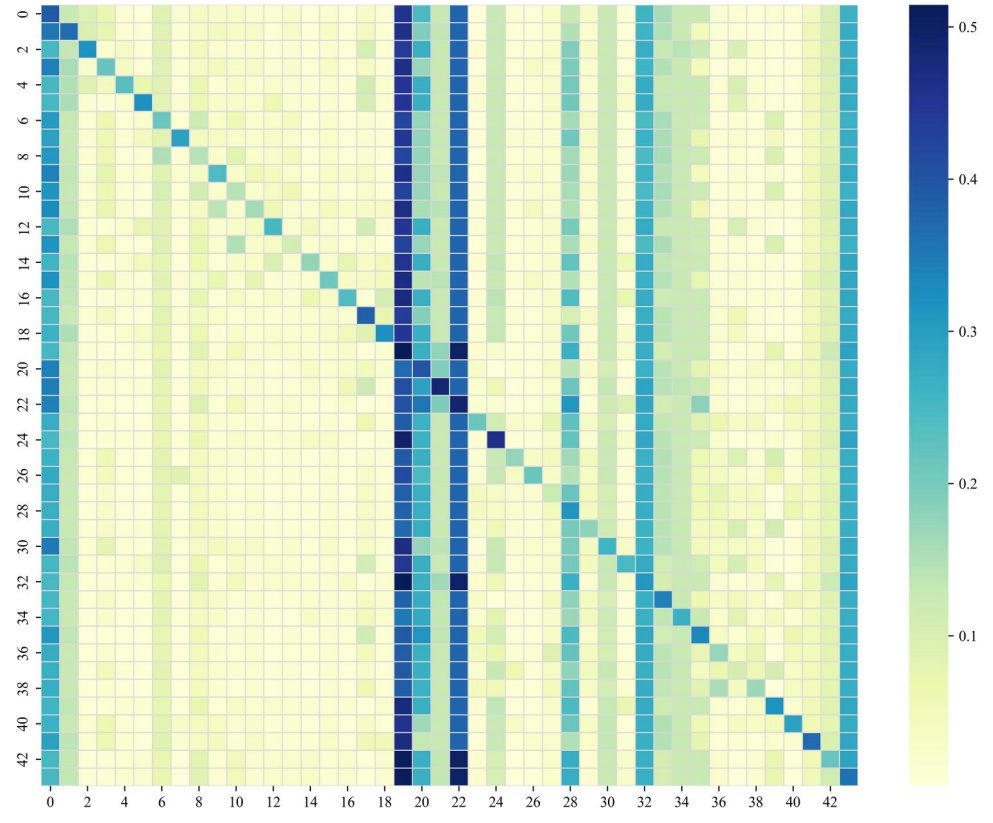

Figure 13. Attention scores in the time information inference task.

closely align with the ground-truth values. For the masked time points, the inferred results also exhibit minimal deviation from the ground-truth.

The attention score results reveal a unique focus for the time information inference task. In addition to the current trajectory point, the model assigns high attention weights to the first point in the sequence and to points located at day transitions (such as points 19–22 and 32 in this case, which mark the beginning of the second and third days). The result indicates that when performing temporal inference, the model references trajectory points that have distinct temporal stage characteristics, reflecting how the model leverages the sequence's temporal structure (Figure 13).

5. Discussion

To address the challenges of spatio-temporal contextual understanding in geographic reasoning, this study introduces the Masked Geo-Information Model (MGIM), inspired by the masked self-supervised learning paradigm in natural language processing. Experimental results demonstrate that MGIM exhibits stable training dynamics, well-converged loss functions, and strong performance across multiple geographic reasoning tasks, validating the effectiveness of the proposed framework. However, the contribution of MGIM extends beyond serving as a superior general learning framework compared to single-task models. More importantly, the findings of this study

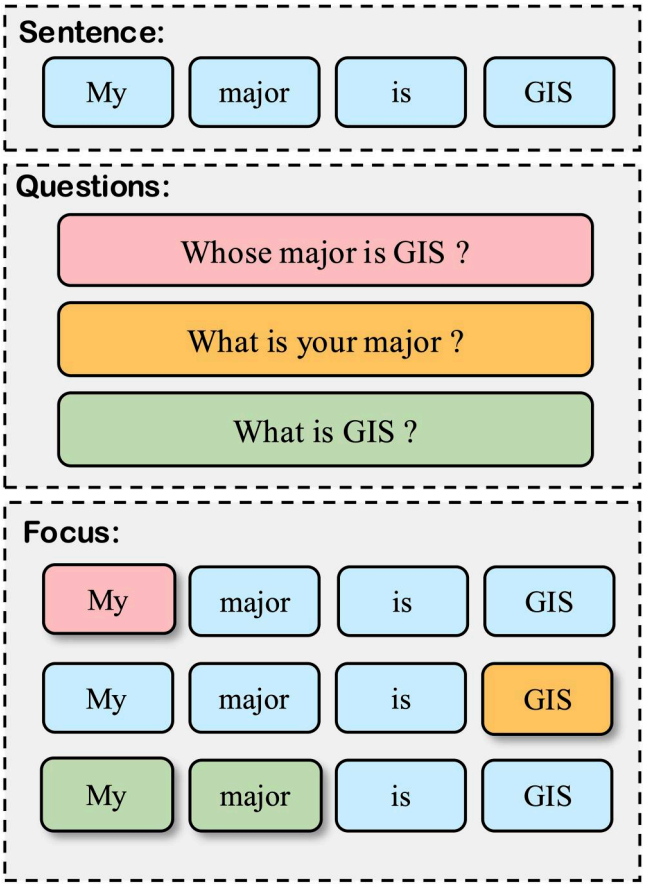

Figure 14. The process of contextual understanding of natural language texts across different scenarios.

reveal a possibility: that a model can move beyond mere pattern fitting to develop a deeper semantic understanding of spatio-temporal contexts.

This semantic understanding is most clearly reflected in the analogy between MGIM and language models. As illustrated in Figure 14, natural language understanding requires identifying relevant contextual components in response to different queries to produce appropriate answers (Warschauer and Healey, 1998, Karanikolas *et al.* 2023). MGIM constructs spatio-temporal contexts from multi-source data. It further exhibits task-specific attention patterns that vary with the inference objective. This ability to flexibly adapt its internal representations according to task requirements provides evidence that the model goes beyond static feature perception toward a semantic understanding of geographic processes.

The effectiveness of MGIM across multiple geographic reasoning tasks provides a foundation for analyzing the model's spatio-temporal contextual understanding. As shown in Figure 6f, location information contributes most to trajectory vector reconstruction, highlighting the role of spatial position as a core anchor in geographic reasoning (Mai *et al.* 2022). Therefore, we focus on the location inference task to further examine MGIM's performance. The observation that the mean error exceeds the

median error indicates that prediction deviations are primarily caused by a few highly fluctuating, non-regular scenarios, reflecting the dual nature of human mobility: high-frequency periodic behaviors (e.g. commuting) and low-frequency, sporadic activities (Wang *et al.* 2015).

For periodic behaviors, MGIM achieves high accuracy and stability, as evidenced by the low median error, and can effectively support macroscopic analyses of urban population dynamics (Xu *et al.* 2025, Yao *et al.* 2023). For sporadic behaviors, despite the greater prediction difficulty (Yang *et al.* 2025b), the model can still produce reasonable inferences based on multi-factor spatio-temporal context. For instance, in Figure 7, the only point with a notable deviation (T-id 33) corresponds to a sporadic behavior, yet the predicted location remains within adjacent parcels to the true position. Through its multi-factor masking mechanism, MGIM explicitly learns semantic correlations among spatial, temporal, and event features, enabling robust reasoning even for non-periodic behaviors. MGIM consistently learns high-generalization representations from noisy and heterogeneous spatio-temporal data, accurately reflecting the inherent uncertainty of human behavior rather than merely reproducing periodic patterns, thereby demonstrating the effectiveness of the model framework.

MGIM achieves deep representations of multi-source spatio-temporal data, providing a robust foundation for modeling spatio-temporal contexts. Its high-fidelity reconstruction demonstrates the ability to capture intrinsic relationships among spatio-temporal elements. Unlike task-specific models that optimize performance for a single objective at the expense of input integrity (Hu *et al.* 2025), MGIM preserves and reconstructs masked features with high accuracy through its decoder. This indicates that the learned trajectory vectors are not merely compressed representations but integrated embeddings that retain complex cross-dimensional dependencies, supporting higher-level reasoning beyond single-task prediction.

The dynamic contextual reasoning results further demonstrate MGIM's contextual adaptability. As shown in Figure 9, when the contextual input of the same land parcel is modified, the model dynamically predicts corresponding variations in people flow, with trends closely aligned with real-world patterns. This suggests that MGIM's reasoning extends beyond isolated elements, enabling it to evaluate how contextual changes influence human activity patterns and other spatio-temporal components, thereby supporting generalized context-aware modeling.

Finally, the task-adaptive attention mechanism provides insight into MGIM's semantic understanding. As illustrated in Figures 8, 10, 11, and 13, attention distributions vary systematically across tasks: trajectory inference emphasizes critical spatial nodes, people flow inference focuses on origin–destination structures, and event or time inference highlights temporal boundaries. This task-dependent reallocation of attention demonstrates MGIM's ability to dynamically reinterpret spatio-temporal contexts and identify the most relevant dependencies, reflecting its generalization capability across diverse reasoning tasks.

The semantic understanding and dynamic reasoning capabilities demonstrated by the MGIM framework hold significant potential for addressing real-world challenges in geographic reasoning. These capabilities suggest that MGIM could serve as a foundational approach to support more anticipatory and context-aware decision-

making in urban management. For instance, the model's dynamic contextual reasoning capability enables what-if scenario simulations (Kishita *et al.* 2020). By modifying contextual inputs to simulate corresponding shifts in people flow patterns. This could allow urban analysts to explore the potential impacts of hypothetical events, such as the closure of major transport hubs or large public gatherings, and to identify high-risk areas in advance (Yazdani and Haghani, 2023). In addition, the model's robust performance in trajectory inference may help planners better evaluate how new infrastructure developments, including metro stations, might influence human mobility patterns (Tsunoda *et al.* 2020). While further validation and integration into operational decision-support systems are still required, these findings highlight MGIM's potential as a step toward more fine-grained, adaptive, and predictive approaches to urban governance.

Although the study has made notable progress, several limitations remain, and future work can advance in the following directions. First, the current modeling of parcel functionality, based on aggregated POI categories, could be enhanced with more detailed representation schemes to enable finer-grained, interpretable quantitative analysis. Second, the model's generalization, currently limited by data availability to a single region and basic event types, requires validation across more diverse geographic and event contexts. Employing transfer learning techniques for such validation would be crucial for establishing the model's broader applicability and practical value.

6. Conclusion

This paper introduced MGIM, demonstrating the effectiveness of a self-supervised paradigm for geographic reasoning. By integrating a custom masking strategy with multi-source data fusion, MGIM learns deep contextual relationships, achieving high accuracy and reconstruction fidelity across diverse tasks. The model's effectiveness was validated through extensive experiments, where it achieved high accuracy on a diverse suite of downstream tasks while maintaining high fidelity in feature reconstruction. Another key capability of the MGIM is contextual adaptability, a feature that enables the model to dynamically modify its inferences based on evolving spatio-temporal conditions. The adaptive attention mechanism of MGIM further demonstrates its strong capability to comprehend spatio-temporal contexts, exhibiting a semantic understanding ability analogous to that of language models.

By adapting the masked-modeling paradigm to the heterogeneity of geospatial data, MGIM provides a robust, domain-specific foundation model capable of capturing complex spatio-temporal dependencies without relying on generic language models. Beyond methodological contributions, the framework offers researchers and urban managers a unified way to learn unified spatio-temporal representations that support multiple downstream analyses, such as scenario-based exploration of human mobility responses to urban events or infrastructure changes and trajectory-based assessment of mobility impacts, without retraining separate models for each task. The successful implementation confirms the feasibility of building powerful, context-aware geographic foundation models and establishes the proposed MGIM as a novel approach

for future research in general-purpose GeoAI. Future work will focus on expanding the model's generalization capabilities across more diverse urban environments and exploring its potential in fine-grained POI semantic analysis.

Acknowledgements

The authors would like to thank the editor and the anonymous reviewers for their constructive comments and insightful suggestions, which have significantly improved the quality of this work.

During writing this paper, the authors used Google Gemini 3 to assist with language polishing and grammar checking. This tool was employed to enhance the stylistic flow and clarity of the text. Following the use of this tool, the authors manually reviewed, edited, and verified all AI-suggested modifications to ensure they accurately reflect the research findings and intended meaning. The authors accept full responsibility for the content and integrity of the final published work.

Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by the National Natural Science Foundation of China (42471491, 42171466), the National Key Research and Development Program of China (2023YFB3906803). This work was partly supported by the Project Grant from the Co-creation Center for Disaster Resilience, IRIDeS, Tohoku University (ID: 2-QR001).

Notes on contributors

Xiang Zhang is a graduate student at China University of Geosciences (Wuhan), China. His research interests include GeoAI and human mobility.

Yao Yao is a Professor at China University of Geosciences (Wuhan) and Hitotsubashi University. His research interests include spatiotemporal big data mining, social geographic computing, and urban geographic information systems.

Chenglong Yu is a graduate student at China University of Geosciences (Wuhan), China and an intern student at LocationMind Institute, LocationMind Inc., Japan. His research interests include GeoAI and Large Language Model.

Zhihui Hu is a graduate student at China University of Geosciences (Wuhan), China. His research interests are geospatial big data mining and geospatial foundation modelling.

Geyuan Zhu is a graduate student at China University of Geosciences (Wuhan), China and an intern student at LocationMind Institute, LocationMind Inc., Japan. His research interests are intelligent agriculture, and large language model.

Mariko Shibasaki is a Consultant at LocationMind Institute, LocationMind Inc., Japan. She has received the master degree from the Graduate School of Frontier Sciences, the University of Tokyo. Her interest is application geospatial foundation models to sustainable and inclusive development involved with human society and the natural environment.

Liangyang Dai is a graduate student at China University of Geosciences (Wuhan). His research interests are geospatial big data mining and health geography.

Yanduo Guo is an undergraduate student at China University of Geosciences(Wuhan), China and an intern student at LocationMind Institute, LocationMind Inc., Japan. His research interests include GeoAI and retrieval-augmented large language models.

Qingfeng Guan is a Professor at China University of Geosciences (Wuhan). His research interests include high-performance spatial intelligence computation and urban computing.

Ryosuke Shibasaki is a Project Professor at the School of Interdisciplinary Information Studies at the University of Tokyo, Japan. His research interests cover mobile big data analysis, satellite/aerial imagery and sensor data analysis, including automated mapping with deep learning, human behavior modeling/simulation, and data assimilation of discrete moving objects.

Data and codes availability statement

The code and data supporting the reproducibility of this study are available at https://doi.org/10.6084/m9.figshare.29364092. We have released the full implementation of our proposed model, along with related test datasets, including parcel-level human mobility data, parcel function representations, and processed trajectory data packages. A detailed user guide is also provided to facilitate the reproduction of the experiments described in the paper. Due to concerns regarding personal privacy and commercial sensitivity, the original trajectory data cannot be made publicly available. Researchers interested in accessing the complete dataset may apply for purchase and use through https://www.blogwatcher.co.jp/terms.

References

Ali, A., Zhu, Y., and Zakarya, M., 2022. Exploiting dynamic spatio-temporal graph convolutional neural networks for citywide traffic flows prediction. *Neural Networks: The Official Journal of the International Neural Network Society*, 145, 233–247.

Cao, X., *et al.*, 2025. U-RNN high-resolution spatio-temporal nowcasting of urban flooding. *Journal of Hydrology*, 659, 133117.

Choudhury, S., *et al.*, 2024. Towards a trajectory-powered foundation model of mobility. Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Spatial Big Data and AI for Industrial Applications, 1–4.

Devlin, J., *et al.*, 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186.

Du, Y.-Y., *et al.*, 2002. Theoretic and application research of geo-case based reasoning. *Acta Geographica Sinica*, 57, 151–158.

Hooghuis, F., *et al.*, 2014. The adoption of thinking through geography strategies and their impact on teaching geographical reasoning in Dutch secondary schools. *International Research in Geographical and Environmental Education*, 23 (3), 242–258.

Hu, D., *et al.*, 2025. An information-theoretic multi-task representation learning framework for natural language understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39 (16), 17276–17286.

Hu, J., *et al.*, 2023. Recognizing mixed urban functions from human activities using representation learning methods. *International Journal of Digital Earth*, 16 (1), 289–307.

Huang, F., *et al.*, 2025. SPOK: Tokenizing geographic space for enhanced spatial reasoning in GeoAI. *International Journal of Geographical Information Science*, 39 (12), 2768–2808.

Huang, W., *et al.*, 2022. Estimating urban functional distributions with semantics preserved POI embedding. *International Journal of Geographical Information Science*, 36 (10), 1905–1930.

Ishikawa, T., 2013. Geospatial thinking and spatial ability: An empirical examination of knowledge and reasoning in geographical science. *The Professional Geographer*, 65 (4), 636–646.

Janowicz, K., *et al.*, 2020. GeoAI: Spatially explicit artificial intelligence techniques for geographic knowledge discovery and beyond. *International Journal of Geographical Information Science*, 34 (4), 625–636.

Janowicz, K., *et al.*, 2025. GeoFM: How will geo-foundation models reshape spatial data science and GeoAI? *International Journal of Geographical Information Science*, 39 (9), 1849–1865.

Kanyepe, J., Tukuta, M., and Chirisa, I., 2021. Urban land-use and traffic congestion: Mapping the interaction. *Journal of Contemporary Urban Affairs*, 5 (1), 77–84.

Karanikolas, N., *et al.*, 2023. Large language models versus natural language understanding and generation. Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics, 278–290.

Kishita, Y., *et al.*, 2020. Scenario structuring methodology for computer-aided scenario design: An application to envisioning sustainable futures. *Technological Forecasting and Social Change*, 160, 120207.

Li, K., *et al.*, 2022. Uniformer: Unified transformer for efficient spatio-temporal representation learning. arXiv Preprint arXiv:2201.04676.

Li, P., *et al.*, 2025. GeoAvatar: A big mobile phone positioning data-driven method for individualized pseudo personal mobility data generation. *Computers, Environment and Urban Systems*, 119, 102252.

Lin, L., *et al.*, 2025. Robust and Efficient Human Mobility Data Processing through the Lens of Topological Persistence. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems, SIGSPATIAL '25*. Association for Computing Machinery, New York, NY, USA, pp. 696–705.

Mai, G.C., *et al.*, 2022. A review of location encoding for GeoAI: Methods and applications. *International Journal of Geographical Information Science*, 36 (4), 639–673.

Musleh, M., Mokbel, M.F., and Abbar, S., 2022. Let's speak trajectories. Proceedings of the 30th International Conference on Advances in Geographic Information Systems, 1–4.

Sharif, M.H., Jiao, L., and Omlin, C.W., 2025. Deep crowd anomaly detection: State-of-the-art, challenges, and future research directions. *Artificial Intelligence Review*, 58 (5), 139.

Tsunoda, K., Hata, T., and Obana, K., 2020. Predicting people flow for supporting facility management. *Proceedings of the 4th International Conference on Software and E-Business*, 57–63.

Tu, W., *et al.*, 2024. Spatial cooperative simulation of land use–population–economy in the Greater Bay Area, China. *International Journal of Geographical Information Science*, 38 (2), 381–406.

Wang, J., and Biljecki, F., 2022. Unsupervised machine learning in urban studies: A systematic review of applications. *Cities*, 129, 103925.

Wang, Y., *et al.*, 2015. Regularity and conformity: Location prediction using heterogeneous mobility data. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1275–1284.

Warschauer, M., and Healey, D., 1998. Computers and language learning: An overview. *Language Teaching*, 31 (2), 57–71.

Xu, F., *et al.*, 2025. Using human mobility data to quantify experienced urban inequalities. *Nature Human Behaviour*, 9 (4), 654–664.

Yang, J., *et al.*, 2025a. GeomorPM: A geomorphic pretrained model integrating convolution and transformer architectures based on DEM data. *International Journal of Geographical Information Science*, 39 (2), 422–451.

Yang, X., *et al.*, 2025b. Causalmob: Causal human mobility prediction with LLMs-derived human intentions toward public events. *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1773–1784.

Yao, Y., *et al.*, 2017. Sensing spatial distribution of urban land use by integrating points-of-interest and Google Word2Vec model. *International Journal of Geographical Information Science*, 31 (4), 825–848.

Yao, Y., *et al.*, 2023. Predicting mobile users' next location using the semantically enriched geo-embedding model and the multilayer attention mechanism. *Computers, Environment and Urban Systems*, 104, 102009.

Yazdani, M., and Haghani, M., 2023. Elderly people evacuation planning in response to extreme flood events using optimisation-based decision-making systems: A case study in western Sydney, Australia. *Knowledge-Based Systems*, 274, 110629.

Yu, Y., *et al.*, 2019. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation*, 31 (7), 1235–1270.

Yuan, Y., *et al.*, 2024. Unist: A prompt-empowered universal model for urban spatio-temporal prediction. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4095–4106.

Zhang, Y., *et al.*, 2025. Focus on hard areas of reconstruction: A fine-grained urban flow inference framework. *Transactions in GIS*, 29 (3), e70039.

Zhao, Y., *et al.*, 2023. Generative causal interpretation model for spatio-temporal representation learning. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3537–3548.

Zhu, Q., *et al.*, 2021. When does further pre-training MLM help? An empirical study on task-oriented dialog pre-training. *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, 54–61.