Enhancing Ride-Hailing Forecasting at DiDi with Multi-View Geospatial Representation Learning from the Web

Xixuan Hao[†]
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
xhao390@connect.hkust-gz.edu.cn

Guicheng Li[†]
China University of Geoscience (Wuhan)
Wuhan, China
liguicheng@cug.edu.cn

Daiqiang Wu
Didichuxing Co. Ltd
Beijing, China
wudaiqiang@didiglobal.com

Xusen Guo
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
xguo796@connect.hkust-gz.edu.cn

Yumeng Zhu
Didichuxing Co. Ltd
Beijing, China
zhuyumeng@didiglobal.com

Zhichao Zou
Didichuxing Co. Ltd
Beijing, China
zouzhichao@didiglobal.com

Peng Zhen
Didichuxing Co. Ltd
Beijing, China
zhenpeng@didiglobal.com

Yao Yao
China University of Geoscience (Wuhan)
Wuhan, China
yaoy@cug.edu.cn

Yuxuan Liang[‡]
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
yuxliang@outlook.com

Abstract

The proliferation of ride-hailing services has fundamentally transformed urban mobility patterns, making accurate ride-hailing forecasting crucial for optimizing passenger experience and urban transportation efficiency. However, ride-hailing forecasting faces significant challenges due to geospatial heterogeneity and high susceptibility to external events. This paper proposes `MVGR-Net` (Multi-View Geospatial Representation Learning), a novel framework that addresses these challenges through a two-stage approach. In the pre-training stage, we learn comprehensive geospatial representations by integrating Points-of-Interest and temporal mobility patterns to capture regional characteristics from both semantic attribute and temporal mobility pattern views. The forecasting stage leverages these representations through a prompt-empowered framework that fine-tunes Large Language Models while incorporating external events. Extensive experiments on DiDi's real-world datasets demonstrate the state-of-the-art performance.

CCS Concepts

• **Mathematics of computing → Time series analysis**.

Keywords

Web Mining; Ride-Hailing Forecasting; Geospatial Representation; LLMs; Prompt Learning; Heterogeneity

[†]Equal Contribution.
[‡]Corresponding author. Email: yuxliang@outlook.com

WWW '26, Dubai, United Arab Emirates.

ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3792789

ACM Reference Format:
Xixuan Hao, Guicheng Li, Daiqiang Wu, Xusen Guo, Yumeng Zhu, Zhichao Zou, Peng Zhen, Yao Yao, and Yuxuan Liang. 2026. Enhancing Ride-Hailing Forecasting at DiDi with Multi-View Geospatial Representation Learning from the Web. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3774904.3792789

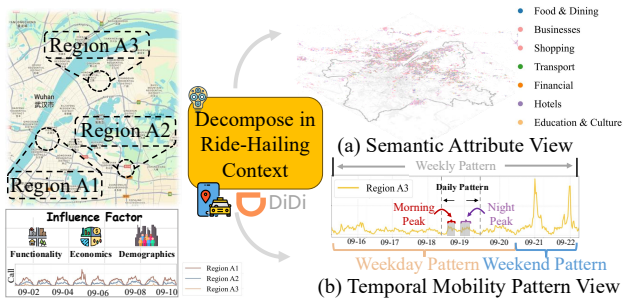


Figure 1: Geospatial Heterogeneity and its two principal view in the context of ride-hailing services.

1 Introduction

With the rapid expansion of the World Wide Web, interconnectivity has induced a profound transformation in the way humans live. The subsequent emergence and rapid development of mobile Internet have given rise to a multitude of online services, among which ride-hailing platforms serve as prominent representatives. As an exemplary web-data-driven service, the proliferation of ride-hailing has fundamentally reshaped urban mobility patterns. It has significantly enhanced transportation accessibility and flexibility, while simultaneously catalyzing innovations in intelligent transportation systems (ITS) and urban governance. Ride-hailing forecasting [15, 41, 42, 62] endeavors to predict temporal dynamics

of key supply-demand indicators, such as Calls and Total Supply Hours (TSH), by leveraging historical transactional data in conjunction with exogenous variables including seasonal holidays and meteorological conditions. Accurate ride-hailing forecasting enables optimization of passenger experience, enhancement of platform operational efficiency, mitigation of traffic congestion, and advancement of urban intelligent transportation ecosystems.

However, ride-hailing indicators, due to their close relationship with urban mobility patterns, exhibit unique inherent characteristics such as significant geospatial heterogeneity, and high susceptibility to external events [42, 50, 68]. These properties render the task of achieving accurate ride-hailing forecasting a formidable challenge. Existing studies [26, 36, 39, 67] commonly formulate ride-hailing forecasting as a spatio-temporal prediction problem [27], influenced by both historical time-series patterns and spatial dependencies across regions. However, our empirical analysis of DiDi's operational data reveals that temporal dependencies exert a dominant role, driven by human mobility patterns across daily, weekly, and seasonal cycles. Conversely, given that the spatial hierarchy of ride-hailing forecasting extends down to the county level, the geographic distance between counties and variations in their internal transportation conditions lead to a relatively limited impact on ride-hailing indicators of adjacent counties. Consequently, *we formulate ride-hailing forecasting as a time series forecasting problem.*

As a specific application of time series forecasting, ride-hailing forecasting possesses a unique and pronounced characteristic: ***Geospatial Heterogeneity***. As illustrated in the left panel of Figure 1, three regions with diverse urban contexts exhibit varying Call patterns that are correlated with regional factors encompassing functionality, economics, and demographics. From our industrial practice, a region's identity can be characterized by two fundamental aspects, as shown in the right panel of Figure 1: **(1) Semantic Attribute View**. The static identity of a region is characterized by its semantic attributes, exemplified by the distribution of Points-of-Interest (POIs), which encapsulate its functional characteristics and operational purpose. **(2) Temporal Mobility Pattern View**. The dynamic essence of a region is captured through temporal mobility patterns, which delineate its operational rhythm across various temporal cycles, from intra-day patterns (e.g., morning / evening peaks) to weekly trends (e.g., weekday / weekend shifts). Effectively synergizing these static and dynamic dimensions is crucial for achieving accurate ride-hailing forecasting.

In recent years, the advent of Large Language Models (LLMs) [2, 12, 40] has begun to shift the paradigm in time series forecasting [30]. Characterized by their profound generalization abilities [5] and world-scale knowledge bases [45], LLMs exhibit a distinct advantage in this domain [30]. Nevertheless, in specialized domains with a scarcity of open-source data, such as ride-hailing forecasting, the direct application of these models faces significant challenges due to their inherent deficiency in specialized domain knowledge. Therefore, to enhance the adaptability of LLMs for ride-hailing forecasting tasks, we propose augmenting them with the following two types of priors: **(1) Geospatial Heterogeneity Prior**. As previously analyzed, geospatial heterogeneity, reflecting intrinsic disparities in urban functionality, economic profiles, and demographic composition, underpins the diverse demand patterns across a region. Consequently, explicitly introducing this factor is crucial for developing more precise and context-aware forecasting models. **(2) External Event Prior**. Ride-hailing supply-demand dynamics are shaped not only by the platform's own operational cycles but also by exogenous events including weather, holidays, and special activities (e.g., concerts, sporting events) [42, 50, 68]. The sporadic and often unpredictable nature of such events makes them notoriously difficult to model using methods that rely solely on historical pattern mining. Therefore, incorporating an External Event Prior stays essential for enhancing the responsiveness to external events of ride-hailing forecasting systems.

In this paper, we propose a Multi-View Geospatial Representation Learning (`MVGR-Net`) framework for Ride-Hailing Forecasting. The proposed framework comprises two sequential stages: a pretraining stage designed to address the challenge of geospatial heterogeneity by learning comprehensive geospatial representations, and a subsequent forecasting stage that leverages these representations and further incorporates external events and textual descriptions to enhance ride-hailing forecast performance. In the pretraining stage, we incorporate two complementary data modalities — POIs and temporal mobility patterns — to capture geospatial heterogeneity across different regions from both semantic attribute and temporal mobility pattern views. A dual cross-attention mechanism, coupled with an attentional pooling module, produces the final comprehensive geospatial representations. In the subsequent ride-hailing forecasting stage, through the integration of multi-view geospatial representation, an elaborated prompt generation network effectively identifies underlying and shared regional properties, and subsequently learns to adaptively utilize these properties to generate informative prompt features that enhance predictive performance. Additionally, contextual factors are captured by integrating three key external variables: rainfall, holidays, and special events.

In summary, our contributions lie in the following aspects:

- *Multi-View Geospatial Representation Learning*. To the best of our knowledge, this is the first attempt to conduct pre-trained modeling from both semantic attribute view and temporal mobility pattern view to enhance ride-hailing forecasting.
- *Prompt-Empowered Ride-Hailing Forecasting*. We propose a prompt-empowered framework for fine-tuning LLMs on the task of ride-hailing forecasting, which simultaneously incorporates both external factors and textual descriptions.
- *Extensive empirical studies*. We conduct extensive experiments over DiDi's real-world datasets. The results demonstrate that our model achieves state-of-the-art performance with an average improvement of 1.8% for Call and 1.5% for TSH prediction.
- *Practical deployment*. Our proposed method has been successfully deployed on the DiDi platform. We demonstrate the system's user interface, geospatial embedding vector library, and an Intelligent Subsidy Allocation experiment to showcase our practicality.

2 Preliminary

Problem Setting. In ride-hailing forecasting, we are given a historical ride-hailing series $X = \{x_{t-L+1:t}\} \in \mathbb{R}^{L \times N_c}$, where L is the look-back window size, N_c denotes the number of regions. Our goal is to learn a forecasting model $f(\cdot)$, which predicts the future T time steps of the series, $\hat{X} = \{x_{t+1:t+T}\} \in \mathbb{R}^{T \times N_c}$, based on historical observations X.

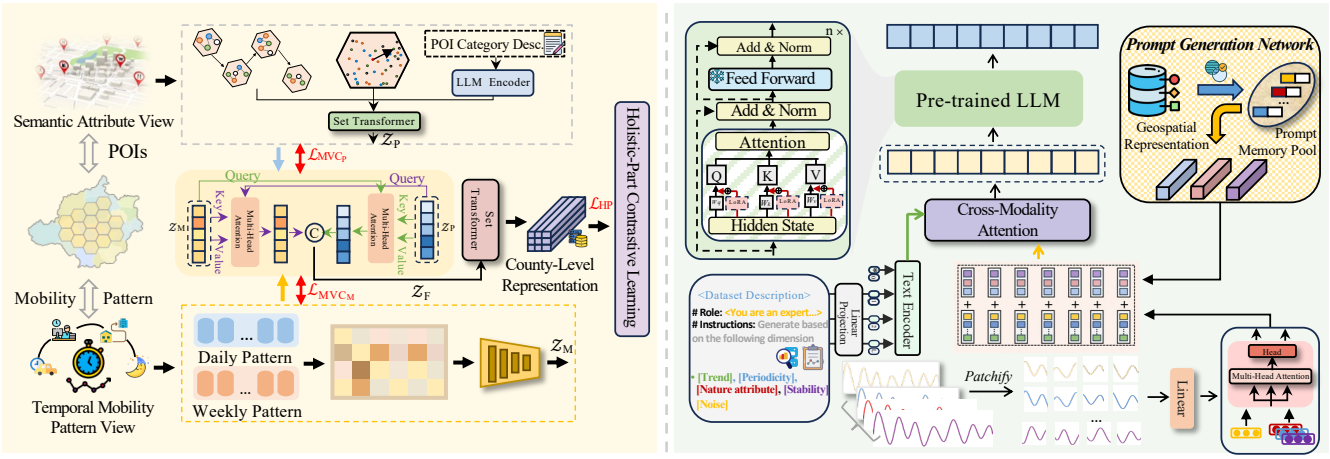


Figure 2: The Overall Framework of MVGR-Net.

Ride-hailing indicators can be classified into two categories: supply-side and demand-side.

- **Call**: The total number of passenger-initiated ride requests within a specified time frame, *acting as the core indicator of user **demand***.
- **Total Supply Hour (TSH)**: The aggregate in-service duration of all drivers on the platform. *It serves as a core indicator for measuring transport capacity **supply***.

We also incorporate three external variables identified through operational practice as most impactful on ride-hailing forecasting: **Rainfall**, **Holidays**, and **Special Events** (e.g., concerts and national examinations). More details can be found in Appendix D.

3 Methodology

In this section, we present details of our proposed MVGR-Net in Figure 2, which consists of two main stages:

- **Stage 1**: Our proposed Multi-View Geospatial Representation Learning framework comprises two complementary branches that model geospatial heterogeneity from semantic attribute view and temporal mobility pattern view, respectively. The framework then facilitates cross-view adaptive interaction through a dual attention mechanism coupled with gated fusion operations.
- **Stage 2**: In the forecasting phase, the time series data initially interacts with external variables and is then augmented with prompt-empowered geospatial representations from Stage 1. Subsequently, this enhanced representation undergoes cross-modal interaction with domain-specific text features. Finally, the resulting feature is fed into a pre-trained LLM, which is fine-tuned using Low-Rank Adaptation (LoRA) to generate predictions.

In industrial applications, the DiDi platform's ride-hailing forecasts currently operate fundamentally at the county level. However, China's county-level administrative units' size disparities (hundreds to thousands of square kilometers [1]) hinder fine-grained geospatial heterogeneity modeling. As a result, we commence our representation learning at a finer-grained grid level (hexagonal cells with 600m edge length) and subsequently aggregate incrementally upward to the county level.

3.1 Multi-View Geospatial Representation Learning

Semantic Attribute View Modeling. The distribution of POIs reflects a region's functional characteristics and operational purpose [56]. To comprehensively model the semantic information behind POIs, we construct a multi-faceted representation by capturing their features from three diverse perspectives. ❶ **Spatial Proximity** [59]. Leveraging a K-Nearest Neighbors (KNN) approach, we capture the distributional characteristics of POIs within their local spatial context. For each POI p_i^s, its k nearest neighbors $p_j^s \in \mathcal{N}_k(p_i^s)$ are retrieved based on spatial distance. c_i^s is the corresponding one-hot category vector associated with p_i^s, which are then passed through an encoder $F_s(\cdot)$ to obtain their feature representations. The encoder consists of an embedding layer followed by a multi-layer perceptron (MLP). We denote the final spatial proximity representation as $\mathcal{Z}_{P_1} = F_s(c_i^s)$. p_j^s primarily contributes during training, with details provided in Appendix C.1. ❷ **Hierarchical Category Semantics** [24]. To capture hierarchical semantic relationships between POI categories, we construct a POI graph where nodes are POIs and edges are spatially weighted. Random walks [33] are used to sample spatial co-occurring sequences. For each sequence, the first node is the target POI p_i^h, and the rest $p_j^h \in \mathcal{N}_k(p_i^h)$ form its context. Each POI is associated with its secondary category one-hot vector c_i^h, and encoded by $F_h(\cdot)$ to obtain feature representations. We denote the final hierarchical category semantics representation as $\mathcal{Z}_{P_2} = F_h(c_i^h)$. p_j^h primarily contributes during training, with details provided in Appendix C.2. ❸ **Textual Semantics.** To extract the semantic characteristics of POI categories from a natural language perspective, we developed a specialized prompt that exploits the inherent knowledge embedded within LLMs. The prompt (Appendix A) is specifically designed to generate comprehensive descriptions that precisely capture the distinct urban functional roles of individual POI categories. These generated POI descriptions are subsequently encoded by the LLM [49] to obtain the feature representation $\mathcal{Z}_{P_3}$. An attentional pooling mechanism [34, 63] is then utilized to capture complex inter-feature dependencies and

execute feature aggregation, ultimately yielding the final semantic attribute representation of grid-level region with d-dimension $\mathcal{Z}_P \in \mathbb{R}^{N_r \times d}$, where N_r denotes the number of regions:

$$\mathcal{Z}_P = \text{AttentionalPooling}(\text{Concat}(\mathcal{Z}_{P_1}, \mathcal{Z}_{P_2}, \mathcal{Z}_{P_3})). \tag{1}$$

Temporal Mobility Pattern View Modeling. Given that daily cycles constitute the fundamental rhythm of urban activities and that weekdays exhibit distinct characteristics from weekends, these temporal factors significantly influence the fluctuation patterns in ride-hailing indicators. To capture these patterns, we compress the time series data of ride-hailing indicators within each region over a given period by computing averages along two temporal dimensions. ❶ Aggregating into 24-hour daily cycles to form a 24-dimensional vector; ❷ Aggregating into weekly cycles to form a 7-dimensional vector. We subsequently construct a joint hour-day matrix, with each element denoting the average indicator value for a specific hour within a particular day of the week. This methodology enables the extraction of fine-grained temporal mobility patterns, such as the potential differences between Monday morning peak hours and Sunday morning peak hours. These temporal patterns are subsequently encoded into feature representations $\mathcal{Z}_M \in \mathbb{R}^{N_r \times d}$ through an encoder consist of MLP layers, followed by Multi-Head Attention [53] to adaptively capture time-dependent patterns.

Cross View Adaptive Interaction & Fusion. Subsequently, we fuse the semantic attribute features $\mathcal{Z}_P$ and temporal pattern features $\mathcal{Z}_M$ via a dual cross-attention mechanism [8, 72] to establish dynamic interactions between complementary domain features, yielding unified multi-view geospatial representations.

$$\mathcal{M}_{att} = \text{Softmax}\left(\frac{(W_Q\mathcal{Z}_M)(W_K\mathcal{Z}_P)^T}{\sqrt{d}}\right)(W_V\mathcal{Z}_P), \tag{2}$$

$$\mathcal{I}_{att} = \text{Softmax}\left(\frac{(W_Q\mathcal{Z}_P)(W_K\mathcal{Z}_M)^T}{\sqrt{d}}\right)(W_V\mathcal{Z}_M), \tag{3}$$

$$\mathcal{Z}_F = \text{Concat}(I_{att}, M_{att}), \tag{4}$$

where $W_Q \in \mathbb{R}^{d\times d}, W_K \in \mathbb{R}^{d\times d}, W_V \in \mathbb{R}^{d\times d}$ are learnable matrices. A multi-view consistency loss is utlized to enforce semantic consistency between $\mathcal{Z}_F$ and the single-view features $\mathcal{Z}_P$ and $\mathcal{Z}_M$.

$$\mathcal{L}_{\text{MVC}_\text{P}} = -\log \frac{\exp(\text{sim}(\mathcal{Z}_{Fi}, \mathcal{Z}_{P+})/\tau)}{\exp(\text{sim}(\mathcal{Z}_{Fi}, \mathcal{Z}_{P+}) + \sum_{j=0}^{N_L-1}\exp(\text{sim}(\mathcal{Z}_{Fi}, \mathcal{Z}_{P-})/\tau)}, \tag{5}$$

where $\mathcal{Z}_{P+}$ denotes the $\mathcal{Z}_P$ representation that resides in the same region as $\mathcal{Z}_{Fi}$, while $\mathcal{Z}_{P-}$ denotes the representation located in different regions. $\text{sim}(\cdot)$ denotes cosine similarity. N_L stands for the number of negative samples selected from the batch. τ is the temperature parameter.

$$\mathcal{L}_{\text{MVC}_\text{M}} = -\log \frac{\exp(\text{sim}(\mathcal{Z}_{Fi}, \mathcal{Z}_{M+})/\tau)}{\exp(\text{sim}(\mathcal{Z}_{Fi}, \mathcal{Z}_{M+}) + \sum_{j=0}^{N_L-1}\exp(\text{sim}(\mathcal{Z}_{Fi}, \mathcal{Z}_{M-})/\tau)}, \tag{6}$$

where $\mathcal{Z}_{M+}$ denotes the $\mathcal{Z}_M$ representation that resides in the same region as $\mathcal{Z}_{Fi}$, while $\mathcal{Z}_{M-}$ denotes the representation located in different regions.

After obtaining the grid-level representations $\mathcal{Z}_F$, we derive the county-level representations through the attentional pooling fusion mechanism:

$$\mathcal{H} = \text{AttentionalPooling}(\mathcal{Z}_F) \tag{7}$$

where $\mathcal{H} \in \mathbb{R}^{N_c\times d}$, N_c denotes the number of county. The Holistic-Part Loss is designed to align a county's feature representation with those of its constituent grid cells, while simultaneously distinguishing it from the representations of grid cells in other counties:

$$\mathcal{L}_{HP} = -\log \frac{\exp(\text{sim}(\mathcal{H}_i, \mathcal{Z}_{Fj})/\tau)}{\exp(\text{sim}(\mathcal{H}_i, \mathcal{Z}_{Fj}) + \sum_{k=0}^{N_L-1}\exp(\text{sim}(\mathcal{H}_i, \mathcal{Z}_{Fk})/\tau)}, \tag{8}$$

where $\mathcal{Z}_{Fj}$ denotes the $\mathcal{Z}_F$ representation that resides in $\mathcal{H}_i$ area, while $\mathcal{Z}_{Fk}$ denotes the grid representation located in different counties, N_L denotes the number of negative samples from the batch.

3.2 Prompt-Empowered Forecasting

In Stage 2, we integrate the comprehensive geospatial representation learned from Stage 1 into the forecasting process, incorporating county-level heterogeneity prior.

3.2.1 Ride-Hailing Forecasting with Exogenous Factors. In the context of ride-hailing practice, the supply-demand dynamics are governed by an interplay of endogenous factors, such as the platform's operational cycles, and exogenous variables like weather, holidays, and special events (e.g., concerts, sporting events). Accurately forecasting ride-hailing indicators thus necessitates a model capable of capturing the complex relationships between these two types of influences. To this end, we leverage the semantic understanding capabilities of LLMs [23] to interpret and encode the impact of these endogenous and exogenous factors on ride-hailing patterns.

As illustrated in the right panel of Figure 2, for the ride-hailing indicator with input length L, $X \in \mathbb{R}^{N_c\times L}$, where N_c denotes the number of counties, together with three exogenous variables — rainfall $X_r \in \mathbb{R}^{N_c\times L}$, holiday $X_h \in \mathbb{R}^{L}$, and special events $X_e \in \mathbb{R}^{N_c\times L}$ — we first split the time series into non-overlapping patches, and then performs self-attention interactions through a [EOS] token that aggregates global information.

$$\hat{\mathbf{x}} = \pi_N(\text{MLP}(\text{MSA}(\mathbf{x} + p))). \tag{9}$$

where $\text{MSA}(\cdot)$ denotes multi-head attention applied to time series, p represents the position embedding, $\pi_N(\cdot)$ denotes the projection operation for selecting the last patch, $\mathbf{x} \in \{X, X_r, X_h, X_e\}$, $\hat{\mathbf{x}} \in \{\hat{X}, \hat{X}_r, \hat{X}_h, \hat{X}_e\} \in \mathbb{R}^D$, where D denotes feature dimension.

After acquiring time series features $\hat{\mathbf{x}}$, we proceed to interact them with textual features and fuse the endogenous ride-hailing features with geospatial representation priors, then utilizing a pre-trained LLM [30] to encode the combined representations. The textual descriptions of the time series data represent a critical component in exploiting the prior knowledge embedded within LLMs. We utilize text generation prompts from [23] to provide comprehensive descriptions of both ride-hailing indicators and external variables from five perspectives: nature attribute, trend, periodicity, stability, and noise. The generated textual descriptions are encoded using an LLM encoder [23, 49] to obtain textual features $\mathcal{T}$.

3.2.2 Heterogeneity-Informed Prompt Learning. As depicted in the right panel of Figure 2, to integrate the learned geospatial representations into the ride-hailing forecasting model, we employ a

Prompt Generation Network to maintain a globally shared memory pool, which consists of M Key-Value pairs:

$$\mathrm{MP} = \{(k_0, v_0), (k_1, v_1), ..., (k_{M-1}, v_{M-1})\}, \tag{10}$$

where $\mathrm{MP} \in \mathbb{R}^{M \times d}$, (k_i, v_i) are all learnable parameters. These are continuously optimized throughout the entire training process to store universal geospatial patterns. The keys will ultimately learn to become various distinct prototypes, with each key representing a type of urban area that shares similar urban characteristics and intrinsic rhythms. The values store a set of mobility behavioral biases tailored to specific spatio-temporal prototypes.

Within this network, the learned geospatial representation functions as the query, which is used to match all Keys in the prompt memory pool, selecting the k_p highest-scoring candidates, yielding a set of attention weights. The computed attention weights are then used to perform a weighted summation of all corresponding Values in the memory pool. Specifically, for the geospatial representation $\mathcal{H}_r$ corresponding to region r, we have:

$$\mathcal{P}_r = \sum_{j=0}^{k_p-1} \alpha_j \, v_j, \ \{\alpha_i\}_{i=0}^{k_p-1} = \underset{l \subseteq [0,\, M-1]}{\arg\max} \ \gamma(\mathcal{H}_r, \, k_l), \tag{11}$$

where $\gamma(\mathcal{H}, \, k)$ calculates cosine similarity. Hereby, we establish a dynamic mapping from geospatial representations to model mobility behavioral heterogeneity, enabling the model to flexibly adapt to the ever-changing urban spatio-temporal scenarios, thus achieving exceptional generalization capability.

Finally, we concatenate the obtained prompt features $\mathcal{P} \in \mathbb{R}^{N_c \times d}$ and ride-hailing time series features to get $\mathcal{U} \in \mathbb{R}^{N_c \times d}$, and then interact them with text features.

$$F = \mathrm{Softmax}\left(\frac{(W_Q\mathcal{T})(W_K\mathcal{U})^T}{\sqrt{d}}\right)(W_V\mathcal{U}), \tag{12}$$

where W_Q,W_K,W_V are learnable matrices. Since the fusion of pre-trained geospatial representations with time series features differs from pure temporal feature distributions and semantic structures, unlike [23, 29], we adopt LoRA [21] to perform efficient fine-tuning in a parameter-efficient fine-tuning manner, thereby better leveraging the introduced spatiotemporal prior knowledge to enhance prediction performance. LoRA facilitates the adaptation of Large Models (LMs) by injecting trainable, low-rank matrices into their Transformer layers to approximate weight updates. For a given pre-trained weight matrix $\mathbf{W} \in \mathbb{R}^{d \times k}$, its update ΔW is represented by a low-rank factorization: $\mathbf{W} + \Delta W = \mathbf{W} + \mathbf{W}_{\text{down}}\mathbf{W}_{\text{up}}$. Here, $\mathbf{W}_{\text{down}} \in \mathbb{R}^{d \times r}$ and $\mathbf{W}_{\text{up}} \in \mathbb{R}^{r \times k}$ are the trainable low-rank matrices, with $r \ll \min(d, k)$. This method is specifically applied to the query (W_Q), key (W_K) and value (W_V) projection matrices in the multi-head attention sub-layer, modifying the output projection h for any given input x.

$$h = h + \varsigma \cdot xW_{\text{down}}W_{\text{up}}, \tag{13}$$

where $\varsigma \geq 1$ is a tunable scalar hyperparameter.

4 Experiments

In our experiments, we aim to address the following research questions (RQs):

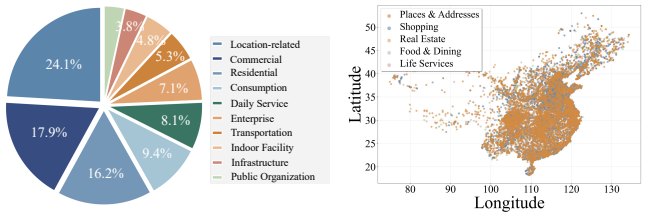
- **RQ1**: Can `MVGR-Net` outperform prior approaches under DiDi's real-world oper- ational datasets? ⇒ **Sec. 4.2**.
- **RQ2**: What are the individual contributions of the various components of `MVGR-Net` to its overall effectiveness?⇒ **Sec. 4.3**.
- **RQ3**: What does qualitative analysis reveal about the performance and interpretability of `MVGR-Net`? ⇒ **Sec. 4.4**.
- **RQ4**: How is the practical application of `MVGR-Net` in real-world business? ⇒ **Sec. 4.5**.

4.1 Experimental Setup

4.1.1 Datasets. We study our problem on DiDi's real-world operational datasets encompassing two ride-hailing indicators: Call and TSH. We focus on two core service categories: C1 (Regular Express) and C3 (Economy Express). The data is collected from 392 key counties of business interest across China, with records taken at 30-minute intervals. To ensure broad temporal span coverage, we select data spanning from 2023, 2024 and 2025. The division of the dataset into train, val, and test sets is shown in Table 2. The effective time window for evaluation is from 6:00 to 22:30 each day. We also incorporate external variables including rainfall, holidays, and special events, with temporal coverage consistent with the ride-hailing indicators in the dataset. Specifically, rainfall and special events are county-specific, while holidays are nationally uniform.

The dataset employed for pre-training consists of 93,441,589 POI entries with nationwide coverage across China. These entries are classified according to a two-tiered categorical system, which includes 18 primary and 218 secondary categories. Figure 3 demonstrates the proportional distribution and geographic distribution of POI primary categories. More details can be found in Appendix B.

Ride-hailing forecasting faces an inherent data insufficiency problem. Given that COVID-19's unprecedented impact [14] induced highly irregular consumption patterns during 2020-2022, data from this pandemic period proves unsuitable for operational forecasting applications. Additionally, earlier data from before 2019 represent outdated economic environments incompatible with present-day conditions. Consequently, the effective volume of applicable historical data remains severely constrained.



(a) (b)

Figure 3: The proportional distribution of top-10 and geographic distribution of top-5 POI primary categories.

4.1.2 Baselines. We compare `MVGR-Net` with the following baselines on our dataset: deep learning based models like PatchTST [46], DLinear [66], CrossFormer [71] and LLM-based model ExoLLM [23]. In addition, we also compared the traditional statistical learning method: XGB [10], ARIMA [51] and Weekly Counterpart. Weekly Counterpart utilizes the value from the corresponding day of the preceding week, serving as a common baseline in industrial applications, as ride-hailing time series data characteristically exhibit

Table 1: Ride-hailing indicators prediction results. The best results are in bold and the second-best results are underlined.

Methods			Ours		ExoLLM		iTransformer		PatchTST		DLinear		CrossFormer		XGB		Weekly Counterpart		ARIMA	
Metric			WMAPE	MAE	WMAPE	MAE	WMAPE	MAE	WMAPE	MAE	WMAPE	MAE	WMAPE	MAE	WMAPE	MAE	WMAPE	MAE	WMAPE	MAE
2025	Call	C1	**0.114**	**48.158**	0.141	59.377	0.192	74.621	0.210	81.321	0.213	81.389	0.178	69.042	0.165	70.155	0.293	124.262	0.191	81.085
		C3	**0.089**	**50.762**	0.105	60.244	0.127	67.520	0.139	73.903	0.139	73.751	0.122	64.676	0.127	78.937	0.170	97.263	0.125	71.359
	TSH	C1	**0.027**	**14.419**	0.043	22.664	0.061	29.447	0.070	34.509	0.070	33.966	0.062	29.976	0.057	29.908	0.075	39.583	0.059	30.869
		C3	**0.039**	**15.414**	0.055	21.773	0.089	32.592	0.092	33.804	0.091	33.434	0.085	31.390	0.089	34.879	0.124	48.844	0.090	35.239
2024	Call	C1	**0.104**	**42.762**	0.122	50.083	0.153	55.924	0.166	60.587	0.170	60.680	0.137	49.932	0.190	80.643	0.194	82.507	0.193	81.462
		C3	**0.078**	**47.662**	0.089	54.230	0.111	61.956	0.119	66.266	0.118	66.069	0.106	59.056	0.124	71.126	0.142	88.115	0.135	80.278
	TSH	C1	**0.028**	**14.196**	0.029	14.403	0.052	23.505	0.059	26.388	0.057	25.593	0.055	24.866	0.055	27.721	0.053	26.501	0.057	26.283
		C3	**0.032**	**13.122**	0.050	20.405	0.070	26.384	0.075	28.275	0.074	27.820	0.070	26.220	0.077	31.306	0.083	33.886	0.075	30.740
2023	Call	C1	**0.097**	**44.548**	0.116	53.154	0.157	63.842	0.168	68.462	0.167	68.333	0.151	61.462	0.178	76.432	0.217	100.558	0.183	83.225
		C3	**0.088**	**41.611**	0.103	48.673	0.136	58.143	0.145	62.200	0.147	62.112	0.128	54.916	0.145	64.902	0.175	82.709	0.158	71.324
	TSH	C1	**0.028**	**13.520**	0.049	23.847	0.076	33.250	0.083	36.357	0.082	35.887	0.074	32.418	0.084	37.514	0.090	44.740	0.087	41.256
		C3	**0.040**	**13.583**	0.057	19.395	0.097	29.342	0.099	30.101	0.098	29.814	0.092	27.984	0.102	33.070	0.131	44.168	0.114	36.251
1st Count			24		0		0		0		0		0		0		0		0	

strong daily and weekly seasonality. Building upon the original models, we also introduce external variables as supplementary input features.

Table 2: Dataset Division across training, validation, and test sets by time period.

Date	Train	Val	Test	Frequency	# Counties
2023	01/01-05/15	05/16-05/31	06/01-06/30	half-hour	392
2024	07/01-11/15	11/16-11/30	12/01-12/30	half-hour	392
2025	01/01-05/15	05/16-05/31	06/01-06/30	half-hour	392

4.1.3 Metrics and Implementation. To assess the prediction performance, we adopt two commonly used evaluation metrics: Weighted Mean Absolute Percentage Error (WMAPE) and Mean Absolute Error (MAE). The parameter initialization follows the setting from [23]. Adam [32] optimizer is chosen to minimize the training loss during parameter learning. All experiments are conducted on Tesla P40 and RTX A6000 GPUs. In our experiments, we set the batch size to 128, number of prompt pairs to 512 with k_p set to 128. We set the look-back window L to 336, corresponding to the past week, and the future timestep M to 48. We utilize DeepSeek-R1 [16] in POI category description generation.

4.2 RQ1: Overall Performance

To evaluate the effectiveness of our MVGR-Net framework, we conduct a comparison with existing state-of-the-art methods in Table 1. As can be seen, our model outperforms the other baselines across all years and in both categories. The performance gain is 1.7%,1.5%,2.2% for Call and 1.9%,1.0%,1.6% for TSH in 2023,2024,2025, respectively. Among the baseline methods, the LLM-based model (ExoLLM) demonstrates a clear advantage. Traditional deep learning methods (iTransformer, PatchTST, DLinear, CrossFormer) slightly outperform machine learning and statistical models (XGB, Weekly Counterpart, ARIMA). This progression indicates the evolutionary path and potential of model architectures for time series forecasting.

4.3 RQ2: Ablation Studies

As shown in Figure 4, we conduct ablation studies to examine each component in MVGR-Net on C1 category on our proposed dataset, including the Prompt Generation Network (PGN), LoRA adaptation, and External Variables (EV). The ablation study reveals that all three components are integral to the performance of MVGR-Net. Specifically, without the Prompt Generation Network (PGN), the integration of geospatial representations becomes rigid and static, failing to capture dynamic spatial contexts. Without LoRA, the LLM cannot adapt to the specific variations introduced by regional heterogeneity, thereby constraining its expressive capability for local patterns. Without external variables, the model is unable to respond to the impact of external events in a timely manner. Quantitatively, the inclusion of external variables consistently yields the most significant performance gain across all three years. Meanwhile, the relative contributions of the PGN and LoRA fluctuate, with their importance varying between different years.

We further investigate the impact of key hyperparameters, including the number of prompt pairs in the Prompt Generation Network (PGN) and the learning rate, with the results presented in Figure 5. The experimental findings reveal a positive correlation between the number of prompt pairs and model performance, with performance gains plateau at 512 pairs. Further increasing the number of prompt pairs to 1024 yields no additional performance improvement, suggesting that 512 pairs are sufficient to capture the essential patterns for our forecasting task. The learning rate experiments exhibit a similar trend. This analysis validates the effectiveness of our chosen hyperparameter settings.

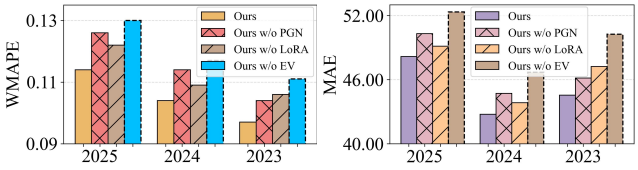


Figure 4: Results of ablation studies on both WMAPE and MAE metrics. PGN stands for Prompt Generation Network, EV denotes External Variables.

4.4 RQ3: Qualitative Analysis

4.4.1 Case Study for Predicted Results. To intuitively illustrate the performance of our proposed model, we present a visualization in Figure 6 that compares its predicted call volumes against those of

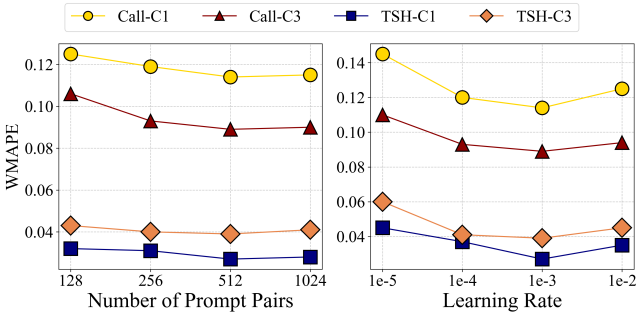


Figure 5: Performance comparison for different hyperparameter setting on 2025.

several baseline models. To aid in understanding the fluctuations in the time series, the plot also includes recorded precipitation and indicates weekend periods. As depicted, our method's predictions most closely align with the ground truth curve, accurately capturing both the underlying periodic trends (highlighted in **pink**) and the abrupt spikes (highlighted in **orange**). This result highlights the superior capability of our model to effectively integrate geospatial heterogeneity with external variables.

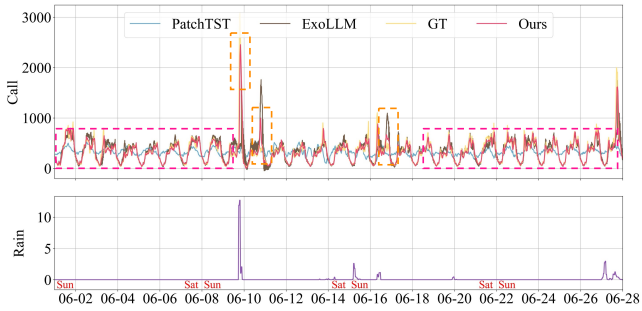


Figure 6: Results of Region A210, Category C1 in 2025. GT denotes Ground Truth.

4.4.2 Geospatial Representation Visualization. In this section, to intuitively demonstrate the effectiveness of the geospatial representation learned in `MVGR-Net`, we map the representation into two-dimensional space using the T-SNE algorithm [44] in Figure 8. Subsequently, we employed the K-Means algorithm [3] to partition the region representations into 10 clusters. As we can see, the results reveal a high degree of clustering coherence, demonstrating that counties with similar attributes are effectively grouped together. For an intuitive and illustrative analysis, we highlight four representative clusters in detail. These clusters correspond precisely to key geo-economic archetypes within China: (1) Primarily Agricultural Counties, (2) Underdeveloped and Ethnic Minority Inhabited Regions, (3) Remote, High-Altitude Cold Regions, and (4) Economically Strong Counties. It is noteworthy that the model successfully isolated the remote, high-altitude cold regions—areas empirically characterized by a significantly low volume of ride-hailing orders—as a unique and well-separated cluster. This indicates that the learned geospatial representations successfully encapsulate the intrinsic attributes of the regions, functioning effectively as robust digital profiles for regional analysis.

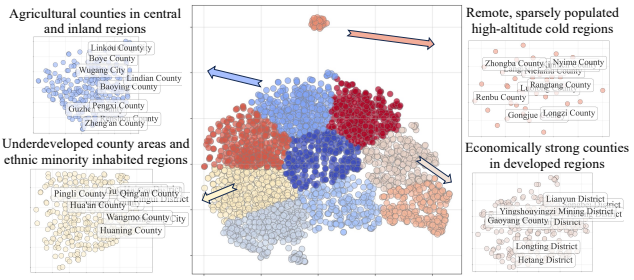


Figure 7: Geospatial Representation Visualization.

4.4.3 Visualization of Prompt Memory Pools. We visualize the features learned by the prompt memory pool in Stage 2 to gain an intuitive understanding of its functionality in Figure 7. We selected three regions, Region X, Y, and Z, where X and Y are economically similar, developed regions located in China's coastal provinces, while Z exhibits a relatively lower level of economic development and is situated in the central inland area. The upper right panel displays the Call volume trends for these regions. The bottom right figure illustrates the similarity scores between the geospatial representations of the corresponding regions and the 512 learnable Keys in the Prompt Generation Network, which we reshape into matrix form. As shown, the visualization matrices for Region X and Y exhibit similar distributions, which differ from the visualization matrix of Region Z. This indicates that the learnable keys adaptively capture distinguishable socioeconomic characteristics.

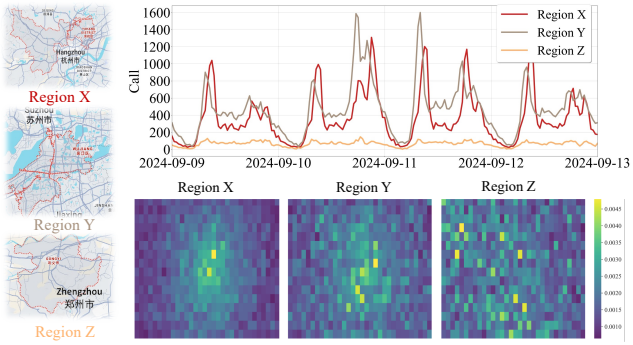


Figure 8: Visualization of Prompt Memory Pool.

4.5 RQ4: Practicality

4.5.1 Deployment. Our proposed `MVGR-Net` has been deployed in DiDi's operational environments, supporting multiple business scenarios such as demand-supply forecasting, resource allocation, smart subsidies, and dynamic pricing. To illustrate its practical use, we present a dashboard interface demonstration in Figure 11 in Appendix. The dashboard helps analysts and engineers visualize, monitor, and interactively analyze spatio-temporal mobility patterns. The data presented in the figure has been anonymized to protect sensitive information.

4.5.2 Geospatial Embedding Vector Library. The geospatial representation we developed has been operationalized within DiDi as a production-ready geospatial embedding vector library. To date, it

has undergone development across two generations, culminating in seven released versions. Figure 9 illustrates our internal web portal for the library, which provides key information such as the project overview, data structure, usage guidelines, and version history. The line chart in the bottom-left corner plots the growth in the number of visits that our embedding vector library has received since its release in April of this year.

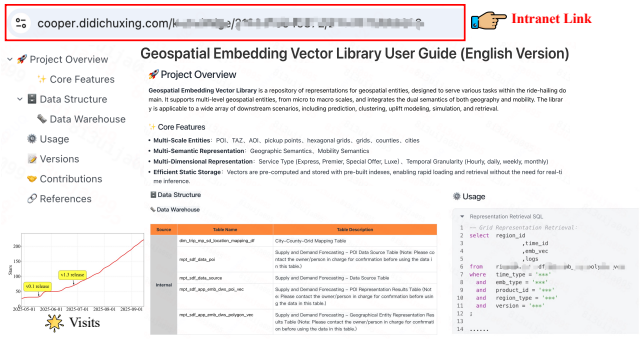


Figure 9: Demonstration of internal document for Geospatial Embedding Vector Library.

4.5.3 Online A/B Test. Our proposed MVGR-Net has been successfully deployed in live production environments to generate ride-hailing forecasts, thereby supporting downstream operational decision making processes. We follow [22, 47, 75] to conduct online A/B testing to validate the effectiveness of MVGR-Net. We report daily forecasting results conducted in August 2025. MVGR-Net achieves average WMAPE improvement of 3.6% and 2.3% in Call performance for C1 and C3 categories respectively; 2.7% and 2.8% in TSH performance for C1 and C3 category respectively.

To further validate the practical utility of our model, we report its performance in a significant downstream task: *Intelligent Subsidy Allocation*. Ride-hailing platforms commonly offer user subsidies to improve retention and operational efficiency. This task focuses on allocating subsidies based on demand forecasts to enhance overall efficiency and conversion performance. More details are provided in Appendix E.2.

5 Related Work

5.1 Urban Region Representation Learning.

Learning representations of urban regions [19, 77] targets the creation of highly transferable region embeddings via the incorporation of regions alongside their attributes. Existing literature has leveraged various data modalities to capture the characteristics of regions, including visual imagery [17, 18, 76], POI [57], human mobility [74], and knowledge graphs [43]. Pre-trained region features can be further adapted to various downstream tasks through fine-tuning, such as economic [9, 17, 58], environmental [11, 70], and demographic applications [69]. In this work, rooted in ride-hailing forecasting contexts, we combine POI data and temporal mobility patterns to model regional representations, infusing spatiotemporal heterogeneity into downstream applications.

5.2 Ride-Hailing Forecasting

In recent years, Ride-Hailing Forecasting [7, 15, 25, 26, 28, 36, 39, 42, 48, 54, 61, 62, 67] has received widespread attention due to its alignment with practical industry demands. [61] integrates CNNs, LSTMs, and graph embeddings to capture spatial, temporal, and semantic regional dynamics. Regarding multi-relational learning, Geng et al. [15] utilizes multi-graph convolution to represent inter-regional spatial, functional, and connectivity dependencies. For adaptive modeling, CCRNN [62] enables dynamic learning of location-specific adjacency matrices. To address noise mitigation, ADFormer [54] employs differential attention mechanisms at the architectural level for spatial correlation refinement. In this work, we formulate ride-hailing supply-demand forecasting as a time series forecasting problem, with detailed discussion in Section 1.

5.3 Prompt Learning

Prompt Learning [6, 13, 37, 38, 64, 65] suggests a methodology for efficiently adapting large-scale pre-trained models to downstream tasks by introducing a small number of trainable prompt parameters, and its influence has rapidly expanded from Natural Language Processing (NLP) [35] to Computer Vision (CV) [52, 73] and even multi-modal learning [31]. Furthermore, researchers have explored dynamic mechanisms that go beyond static prompts, such as leveraging a shared prompt pool to facilitate continual learning without rehearsal buffer [55]. HimNet [13] extracts spatio-temporal heterogeneity-informed prompt features through a query-pool paradigm. In this work, we deeply integrates dynamic prompts with the problem of spatio-temporal heterogeneity by proposing a novel, retrieval-based adaptive prompt framework.

6 Conclusion and Future Work

Accurate ride-hailing forecasting plays a pivotal role in the advancement of operational efficiency, the improvement of user experience, and the optimization of traffic management. To tackle the geospatial heterogeneity challenge inherent in ride-hailing demand forecasting, this work proposes learning comprehensive geospatial representations through two complementary viewpoints: semantic attributes and temporal mobility patterns. In subsequent ride-hailing forecasting, we integrate the geospatial representation into the time series forecasting through a dynamic prompt generation network, and combine it with external variables and domain-specific textual descriptions to enhance the prediction of ride-hailing indicators. Experiments on DiDi business data demonstrate the effectiveness of our method. Future directions include exploring the application of other types of data such as origin-destination (OD) in ride-hailing forecasting, as well as applying novel backbones such as TabPFN [20].

Acknowledgments

This work is supported by CCF-DiDi GAIA Collaborative Research Funds for Young Scholars, the National Natural Science Foundation of China (No.62402414), the Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011994), the Guangzhou Municipal Science and Technology Project (No. 2023A03J0011), the Guangzhou Industrial Information and Intelligent Key Laboratory Project (No. 2024A03J0628), and the Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No. 2023B1212010007).

References

[1] [n. d.]. http://xzqh.mca.gov.cn/map.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Mohiuddin Ahmed, Raihan Seraj, and Syed Mohammed Shamsul Islam. 2020. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics* 9, 8 (2020), 1295.

[4] Mouloud Belbahri, Alejandro Murua, Olivier Gandouet, and Vahid Partovi Nia. 2021. Qini-based uplift regression. *The Annals of Applied Statistics* 15, 3 (2021), 1247–1272.

[5] Mikhail Budnikov, Anna Bykova, and Ivan P Yamshchikov. 2025. Generalization potential of large language models. *Neural Computing and Applications* 37, 4 (2025), 1973–1997.

[6] Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. 2024. TEMPO: Prompt-based Generative Pre-trained Transformer for Time Series Forecasting. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=YH5w12OUuU

[7] Dun Cao, Kai Zeng, Jin Wang, Pradip Kumar Sharma, Xiaomin Ma, Yonghe Liu, and Siyuan Zhou. 2021. BERT-based deep spatial-temporal network for taxi demand prediction. *IEEE Transactions on Intelligent Transportation Systems* 23, 7 (2021), 9442–9454.

[8] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*. 357–366.

[9] Meng Chen, Zechen Li, Weiming Huang, Yongshun Gong, and Yilong Yin. 2024. Profiling urban streets: A semi-supervised prediction model based on street view imagery and spatial topology. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 319–328.

[10] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[11] Wei Chen, Xixuan Hao, Yuankai Wu, and Yuxuan Liang. 2024. Terra: A multimodal spatio-temporal dataset spanning the earth. *Advances in Neural Information Processing Systems* 37 (2024), 66329–66356.

[12] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[13] Zheng Dong, Renhe Jiang, Haotian Gao, Hangchen Liu, Jinliang Deng, Qingsong Wen, and Xuan Song. 2024. Heterogeneity-informed meta-parameter learning for spatiotemporal time series forecasting. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*. 631–641.

[14] Song Gao, Jinmeng Rao, Yuhao Kang, Yunlei Liang, and Jake Kruse. 2020. Mapping county-level mobility pattern changes in the United States in response to COVID-19. *SIGSpatial Special* 12, 1 (2020), 16–26.

[15] Xu Geng, Yaguang Li, Leye Wang, Lingyu Zhang, Qiang Yang, Jieping Ye, and Yan Liu. 2019. Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 3656–3663.

[16] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[17] Xixuan Hao, Wei Chen, Yibo Yan, Siru Zhong, Kun Wang, Qingsong Wen, and Yuxuan Liang. 2025. UrbanVLP: Multi-granularity vision-language pretraining for urban socioeconomic indicator prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 28061–28069.

[18] Xixuan Hao, Wei Chen, Xingchen Zou, and Yuxuan Liang. 2025. Nature makes no leaps: Building continuous location embeddings with satellite imagery from the web. In *Proceedings of the ACM on Web Conference 2025*. 2799–2812.

[19] Xixuan Hao, Yutian Jiang, Xingchen Zou, Jiabo Liu, Yifang Yin, and Yuxuan Liang. 2025. Unlocking Location Intelligence: A Survey from Deep Learning to The LLM Era. *arXiv preprint arXiv:2505.09651* (2025).

[20] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. 2025. Accurate predictions on small data with a tabular foundation model. *Nature* (09 01 2025). doi:10.1038/s41586-024-08328-6

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[22] Jun Hu, Wenwen Xia, Xiaolu Zhang, Chilin Fu, Weichang Wu, Zhaoxin Huan, Ang Li, Zuoli Tang, and Jun Zhou. 2024. Enhancing sequential recommendation via llm-based semantic embedding learning. In *Companion Proceedings of the ACM Web Conference 2024*. 103–111.

[23] Qihe Huang, Zhengyang Zhou, Kuo Yang, and Yang Wang. 2025. Exploiting Language Power for Time Series Forecasting with Exogenous Variables. In *Proceedings of the ACM on Web Conference 2025*. 4043–4052.

[24] Weiming Huang, Lizhen Cui, Meng Chen, Daokun Zhang, and Yao Yao. 2022. Estimating urban functional distributions with semantics preserved POI embedding. *International Journal of Geographical Information Science* 36, 10 (2022), 1905–1930.

[25] Ziheng Huang, Weihan Zhang, Dujuan Wang, and Yunqiang Yin. 2022. A GAN framework-based dynamic multi-graph convolutional network for origin–destination-based ride-hailing demand prediction. *Information Sciences* 601 (2022), 129–146.

[26] Guangyin Jin, Yan Cui, Liang Zeng, Hanbo Tang, Yanghe Feng, and Jincai Huang. 2020. Urban ride-hailing demand prediction with multiple spatio-temporal information fusion network. *Transportation Research Part C: Emerging Technologies* 117 (2020), 102665.

[27] Guangyin Jin, Yuxuan Liang, Yuchen Fang, Zezhi Shao, Jincai Huang, Junbo Zhang, and Yu Zheng. 2023. Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. *IEEE transactions on knowledge and data engineering* 36, 10 (2023), 5388–5408.

[28] Guangyin Jin, Zhexu Xi, Hengyu Sha, Yanghe Feng, and Jincai Huang. 2022. Deep multi-view graph-based network for citywide ride-hailing demand prediction. *Neurocomputing* 510 (2022), 79–94. doi:10.1016/j.neucom.2022.09.010

[29] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-LLM: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations (ICLR)*.

[30] Ming Jin, YiFan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. 2024. Position: What Can Large Language Models Tell Us about Time Series Analysis. In *Forty-first International Conference on Machine Learning*. https://openreview.net/forum?id=iroZNDxFJZ

[31] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. 2023. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 19113–19122.

[32] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[33] Gregory F Lawler and Vlada Limic. 2010. *Random walk: a modern introduction*. Vol. 123. Cambridge University Press.

[34] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. 2019. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*. PMLR, 3744–3753.

[35] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 3045–3059. doi:10.18653/v1/2021.emnlp-main.243

[36] Jianbo Li, Zhiqiang Lv, Zhaobin Ma, Xiaotong Wang, and Zhihao Xu. 2024. Optimization of spatial-temporal graph: A taxi demand forecasting model based on spatial-temporal tree. *Information Fusion* 104 (2024), 102178.

[37] Zhonghang Li, Long Xia, Lei Shi, Yong Xu, Dawei Yin, and Chao Huang. 2024. Opencity: Open spatio-temporal foundation models for traffic prediction. *arXiv preprint arXiv:2408.10269* (2024).

[38] Zhonghang Li, Lianghao Xia, Yong Xu, and Chao Huang. 2024. FlashST: A Simple and Universal Prompt-Tuning Framework for Traffic Prediction. In *Forty-first International Conference on Machine Learning*. https://openreview.net/forum?id=vye4OgLaTy

[39] Hongyi Lin, Yixu He, Yang Liu, Kun Gao, and Xiaobo Qu. 2023. Deep demand prediction: An enhanced conformer model with cold-start adaptation for origin–destination ride-hailing demand prediction. *IEEE Intelligent Transportation Systems Magazine* 16, 3 (2023), 111–124.

[40] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).

[41] Kai Liu, Zhiju Chen, and Yamamoto Liheng Tuo. 2023. Exploring the Impact of Spatiotemporal Granularity on the Demand Prediction of Dynamic Ride-Hailing. *IEEE transactions on intelligent transportation systems* 24, 1 (2023), 104–114.

[42] Mengjin Liu, Yuxin Zuo, Yang Luo, Daiqiang Wu, Peng Zhen, Jiecheng Guo, and Xiaofeng Gao. 2025. Weather-Conditioned Multi-graph Network for Ride-Hailing Demand Forecasting. In *Service-Oriented Computing. ICSOC 2025*.

[43] Yu Liu, Xin Zhang, Jingtao Ding, Yanxin Xi, and Yong Li. 2023. Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction. In *Proceedings of the ACM web conference 2023*. 4150–4160.

[44] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[45] Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. 2023. Geollm: Extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213* (2023).

[46] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers.

In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=Jbdc0vTOcol

[47] Wenjun Peng, Guiyang Li, Yue Jiang, Zilong Wang, Dan Ou, Xiaoyi Zeng, Derong Xu, Tong Xu, and Enhong Chen. 2024. Large language model based long-tail query rewriting in taobao search. In *Companion Proceedings of the ACM Web Conference 2024*. 20–28.

[48] Weiguo Pian, Yingbo Wu, Xiangmou Qu, Junpeng Cai, and Ziyi Kou. 2022. Spatial-Temporal Dynamic Graph Attention Networks for Ride-hailing Demand Prediction. arXiv:2006.05905 [cs.LG] https://arxiv.org/abs/2006.05905

[49] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[50] Jimeng Shi, Azam Shirali, and Giri Narasimhan. 2024. Boosting Time Series Prediction of Extreme Events by Reweighting and Fine-tuning. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 1450–1457.

[51] Robert H Shumway and David S Stoffer. 2017. ARIMA models. In *Time series analysis and its applications: with R examples*. Springer, 75–163.

[52] Kihyuk Sohn, Huiwen Chang, José Lezama, Luisa Polania, Han Zhang, Yuan Hao, Irfan Essa, and Lu Jiang. 2023. Visual prompt tuning for generative transfer learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 19840–19851.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[54] Haichen Wang, Liu Yang, Xinyuan Zhang, Haomin Yu, Ming Li, and Jilin Hu. 2025. ADFormer: Aggregation Differential Transformer for Passenger Demand Forecasting. In *Proceedings of International Joint Conference on Artificial Intelligence, IJCAI-25*.

[55] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. 2022. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 139–149.

[56] Yanxin Xi, Tong Li, Huandong Wang, Yong Li, Sasu Tarkoma, and Pan Hui. 2022. Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests. In *Proceedings of the ACM web conference 2022*. 3308–3316.

[57] Yanxin Xi, Tong Li, Huandong Wang, Yong Li, Sasu Tarkoma, and Pan Hui. 2022. Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests. In *Proceedings of the ACM web conference 2022*. 3308–3316.

[58] Congxi Xiao, Jingbo Zhou, Yixiong Xiao, Jizhou Huang, and Hui Xiong. 2024. Refound: Crafting a foundation model for urban region understanding upon language and visual foundations. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3527–3538.

[59] Bo Yan, Krzysztof Janowicz, Gengchen Mai, and Song Gao. 2017. From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems*. 1–10.

[60] Jiaqi Yang, Lexiao Chen, Zicheng Su, Wanjing Ma, Zhichao Zou, and Kun An. 2025. Decision-focused learning for optimal subsidy allocation in ride-hailing services. *Transportation Research Part C: Emerging Technologies* 180 (2025), 105301.

[61] Huaxiu Yao, Fei Wu, Jintao Ke, Xianfeng Tang, Yitian Jia, Siyu Lu, Pinghua Gong, Jieping Ye, and Zhenhui Li. 2018. Deep multi-view spatial-temporal network for taxi demand prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[62] Junchen Ye, Leilei Sun, Bowen Du, Yanjie Fu, and Hui Xiong. 2021. Coupled Layer-wise Graph Convolution for Transportation Demand Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 5 (May 2021), 4617–4625. doi:10.1609/aaai.v35i5.16591

[63] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* (2022).

[64] Yuan Yuan, Jingtao Ding, Jie Feng, Depeng Jin, and Yong Li. 2024. Unist: A prompt-empowered universal model for urban spatio-temporal prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4095–4106.

[65] Yuan Yuan, Chonghua Han, Jingtao Ding, Depeng Jin, and Yong Li. 2025. Urbandit: A foundation model for open-world urban spatio-temporal learning. *Advances in neural information processing systems* (2025).

[66] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. 2023. Are transformers effective for time series forecasting?. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 11121–11128.

[67] Chizhan Zhang, Fenghua Zhu, Yisheng Lv, Peijun Ye, and Fei-Yue Wang. 2021. MLRNN: Taxi demand prediction based on multi-level deep learning and regional heterogeneity analysis. *IEEE Transactions on Intelligent Transportation Systems* 23, 7 (2021), 8412–8422.

[68] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.

[69] Liang Zhang, Cheng Long, and Gao Cong. 2022. Region embedding with intra and inter-view contrastive learning. *IEEE Transactions on Knowledge and Data Engineering* 35, 9 (2022), 9031–9036.

[70] Ruixing Zhang, Bo Wang, Tongyu Zhu, Leilei Sun, and Weifeng Lv. 2025. Urban In-Context Learning: Bridging Pretraining and Inference through Masked Diffusion for Urban Profiling. *arXiv preprint arXiv:2508.03042* (2025).

[71] Yunhao Zhang and Junchi Yan. 2023. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*.

[72] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 11106–11115.

[73] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.

[74] Silin Zhou, Dan He, Lisi Chen, Shuo Shang, and Peng Han. 2023. Heterogeneous region embedding with prompt learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 4981–4989.

[75] Ding Zou, Wei Wei, Feida Zhu, Chuanyu Xu, Tao Zhang, and Chengfu Huo. 2024. Knowledge enhanced multi-intent transformer network for recommendation. In *Companion proceedings of the ACM web conference 2024*. 1–9.

[76] Xingchen Zou, Jiani Huang, Xixuan Hao, Yuhao Yang, Haomin Wen, Yibo Yan, Chao Huang, and Yuxuan Liang. 2024. Learning geospatial region embedding with heterogeneous graph. *arXiv e-prints* (2024), arXiv–2405.

[77] Xingchen Zou, Yibo Yan, Xixuan Hao, Yuehong Hu, Haomin Wen, Erdong Liu, Junbo Zhang, Yong Li, Tianrui Li, Yu Zheng, et al. 2025. Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook. *Information Fusion* 113 (2025), 102606.

Appendix

A POI Prompt

We provide a detailed POI Category Description generation prompt below, which comprehensively explores POI semantic features from five perspectives: Core Function, Hierarchy, Target Demographics, Temporal Pattern, and Spatial Context.

Prompt: POI Category Description

 # Role and Goal

You are an urban sociologist and data scientist. Your task is to write a descriptive paragraph for a given POI category.

 # Instructions

In your description, you must incorporate POI's category, location and administrative district, weave together the following aspects naturally in a paragraph format:

(1) Core Function: What is its primary purpose?
(2) Hierarchy: What broader category does it belong to? What are some specific examples?
(3) Target Demographics: Who are the primary customers or users?
(4) Temporal Pattern: When is it most active (e.g., daytime, nighttime, weekday, weekend)?
(5) Spatial Context: What other types of places is it often found near?

 # POI Information

[POI Category] [Location] [County&City Name]

Figure 10: Detailed prompt template for POI category description generation.

B More Details about Dataset

More information about POI secondary categories. For the POI data utilized in Stage 1: Multi-View Geospatial Representation Learning of our model, beyond the content introduced in the main text, we also introduce information about POI secondary categories here. POI secondary categories provide a finer-grained functional breakdown of primary POI categories. The top-10 POI secondary categories include: Door Number Information, Building Numbers, Administrative Places, Companies & Enterprises, Passage Facilities, Snacks & Fast Food, Furniture & Building Materials, Beauty & Hair Salons, Government Agencies, Chinese Restaurants.

Special events data types. Special events data encompasses 12 distinct types: *International Women's Day, Teacher Qualification Examination, Public Institution Recruitment Examination, Provincial Civil Service Examination, Self-taught Higher Education Examinations, Chinese Valentine's Day, Art College Entrance Examination, Concert, Sports Event, Marathon, Beer Festival and Large-scale Exhibition.* Some non-statutory holidays are also classified as special events.

C More Details about POI Representation

C.1 Spatial Proximity

Here we introduce the training process of POI representation learning from spatial proximity. For each POI p_i^s, and its k nearest neighbors $p_j^s \in \mathcal{N}_k(p_i^s)$ are retrieved based on spatial distance. c^s is the corresponding one-hot category vectors associated with p^s,

$$\mathbf{z}_{p_i^s} = F_s(c_i^s), \quad \mathbf{z}_{p_j^s} = F_s(c_j^s). \tag{14}$$

The objective is to maximize the similarity between the central POI and its neighbors:

$$\mathcal{L}_{\text{SP}} = -\sum_{p_j^s \in \mathcal{N}_k(p_i^s)} (\log \frac{\exp\left(\mathbf{z}_{p_j^s}^\top \mathbf{z}_{p_i^s}\right)}{\sum_{l=0}^{n_c-1} \exp\left(\mathbf{e}_l^\top \mathbf{z}_{p_i^s}\right)}), \tag{15}$$

where n_c is the total number of POI categories, e_l denotes the feature representation of the l-th POI category after encoding through F_s.

C.2 Hierarchical Category Semantics

Here we introduce the training process of POI representation learning from hierarchical category semantics. To capture hierarchical semantic relationships between POI categories, we construct a POI graph where nodes are POIs and edges are spatially weighted. Random walks [33] are used to sample spatial co-occurring sequences. For each sequence, the first node is the target POI p_i^h, and the rest $p_j^h \in \mathcal{N}_k(p_i^h)$ form its context. Each POI is associated with its secondary category one-hot vector c_i^h, and encoded by $F_h(\cdot)$ to obtain feature representations:

$$\mathbf{z}_{p_i^h} = F_h(c_i^h), \quad \mathbf{z}_{p_j^h} = F_h(c_j^h). \tag{16}$$

The hierarchical semantic representations are optimized via the following joint objective:

$$\begin{aligned}\mathcal{L}_{\text{HCS}} = &-\sum_{p_j^h \in \mathcal{N}_k(p_i^h)} \log \frac{\exp(\mathbf{z}_{p_j^h}^\top \mathbf{z}_{p_i^h})}{\sum_{l=0}^{n_c-1} \exp(\mathbf{e}_l^\top \mathbf{z}_{p_i^h})} \\ &+ \lambda \sum_{i,l} w_{il} \left\| \mathbf{z}_{p_i^h} - \mathbf{e}_l \right\|_2^2,\end{aligned} \tag{17}$$

while the first term models spatial co-occurrence between categories using a skip-gram objective, the second regularization term enforces embedding smoothness for POIs within the same primary category, n_c denotes the number of POI secondary categories, $w_{il} = 1$ if p_i^h and category l belong to the same primary category group. λ controls the trade-off between the two objectives.

D More Details about External Variables

- **Rainfall**. We incorporate rainfall data, measured as precipitation in millimeters (mm), as the sole weather variable due to its high impact. The forecasted rainfall over a prediction horizon of length S is represented by the vector $\mathrm{P} \in \mathbb{R}^{S\times 1}$ where $\mathrm{P} = (p_{t+1}, p_{t+2}, ..., p_{t+s})$. Data is sourced from the China Weather Network (https://www.weather.com.cn).
- **Holiday**. To capture the impact of public holidays on travel patterns, we incorporate holiday data as a categorical feature. We consider N_h distinct types of public holidays known to significantly influence travel demand. This information is encoded as a vector for the forecast period:

$$\mathbf{H} = (h_{t+1}, h_{t+2}, \ldots, h_{t+S}) \in \mathbb{Z}^{S\times 1}, h_i \in [0, N_h],$$

where S is the prediction horizon, N_h correspond to the number of holiday types. The data is based on official holiday schedules announced by the Chinese government.

- **Special Events**. Special events meta data accounts for non-periodic external events that can significantly impact travel demand. Examples include major concerts, sporting events, and national examinations.

$$\mathbf{E} = (e_{t+1}, e_{t+2}, \ldots, e_{t+S}) \in \mathbb{Z}^{S\times 1}, e_i \in [0, N_e],$$

where N_e represents the number of special events. Data is sourced from Damai (https://www.damai.cn/).

E More Details for Deployment

E.1 Deployment System Demonstration

As introduced in the main text, the interface allows flexible spatial and temporal exploration. Users can enter a grid ID in the top-left map to query the most similar grids from the embedding database. It also supports analysis of POI functions within the grid, helping to understand regional functionality and potential travel demand. For prediction, users can select any province-level, city-level, or county-level region to monitor future CALL, TSH and other relevant indicators such as ASP (Average Selling Price). This helps guide resource allocation, pricing strategies, and operational planning.

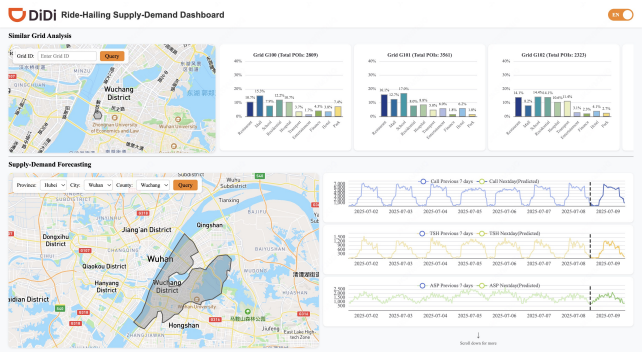


Figure 11: Deployment system demonstration of our MVGR-Net framework.

E.2 Intelligent Subsidy Allocation Experiment

To evaluate the transferability of the geospatial representations learned by MVGR-Net across multiple tasks, we additionally introduce an **Intelligent Subsidy Allocation Experiment** [60]. To enhance operational efficiency and economic benefits, ride-hailing platforms typically provide subsidies to users in order to improve user retention. Intelligent Subsidy Allocation aims to effectively allocate subsidies and improve overall operational efficiency and conversion performance. This task is formulated as a multi-treatment causal inference problem (*i.e.*, multiple subsidy strategies). The objective is to estimate the heterogeneous impact of different subsidy strategies on user conversion behavior. We adopt a neural network architecture with multiple output heads, which allows the model to predict potential outcomes under all treatment conditions within a unified framework.

Let the strategy set be $\mathcal{T} = \{ST_0, ST_1, \ldots, ST_{N_{st}-1}\}$, where N_{st} denotes the number of strategies such as 10% discount, fixed-amount reduction. ST_0 denotes no strategy. For input feature $\mathbf{x_{st}}$ of order and user interaction data, the model outputs the predicted conversion probability (the probability of a user taking a ride, conditional on receiving a subsidy) under each treatment: $\hat{Y}(ST_i \mid \mathbf{x_{st}})$. Each prediction is generated by a separate output head corresponding to one treatment. During training, the model learns from historical samples $\Omega = (\mathbf{x_{st}}_i, ST_i, \hat{Y}_i)$, where ST_i is the observed treatment and $\hat{Y}_i \in \{0, 1\}$ indicates whether the user converted. The cross-entropy loss function is used to optimize prediction accuracy under each treatment. During inference, the model computes the uplift effect of any treatment ST_i relative to the strategy $ST = 0$ as follows:

$$U(ST_i \mid \mathbf{x_{st}}) = \hat{Y}(ST_i \mid \mathbf{x_{st}}) - \hat{Y}(ST_0 \mid \mathbf{x_{st}}). \tag{18}$$

In business practice, user responses depend not only on order features but also on the spatial context, regional supply-demand conditions, and functional attributes. Therefore, we incorporate our proposed geospatial representations as features to capture these influences. We augment the feature vector for each order by concatenating it with the pre-trained geospatial representation of its originating region, yielding the final model input. The experiment results are demonstrated in Table 3. The online baseline model employs a neural network with multiple output heads. Based on this, the enhanced model Exp_1 utilized our city-level region representations, while Exp_2 utilized our county-level region representations. QINI coefficient [4] and WMAPE are utilized to measure how well the model identifies high-response users. Results demonstrate that the enhanced model almost significantly outperforms baselines on both QINI coefficient and WMAPE metrics. Notably, optimal performance varies by strategy, with Exp_1 excelling in some scenarios and Exp_2 in others. These results confirm the effectiveness and transferability of our learned geospatial representations.

Treatment	2025 / 05					
	QINI ↓			WMAPE ↓		
	Exp_1	Exp_2	Online	Exp_1	Exp_2	Online
85%-x	0.239	**0.220**	0.239	**0.167**	0.192	0.248
80%-x	0.233	0.232	**0.229**	**0.110**	0.152	0.307
75%-x	0.241	**0.238**	0.239	**0.063**	0.126	0.126
70%-x	**0.243**	0.252	0.249	0.115	**0.111**	0.142
60%-x	0.247	**0.245**	**0.245**	**0.038**	0.086	0.121

Table 3: Geospatial representation enhanced intelligent subsidy experiment results for May 2025. Treatment name's first number represents the discount percentage, x stands for a direct reduction discount of x yuan. We set x to 5 here.