# Retrieval-Augmented Universal Models for Spatio-temporal Data

Weilin Ruan*
The Hong Kong University of Science and Technology - Guangzhou Campus
Guangzhou, China
rwlinno@gmail.com

Yuxuan Liang
The Hong Kong University of Science and Technology - Guangzhou Campus
Guangzhou, China
yuxuanliang@outlook.com

## Abstract

Urban spatio-temporal prediction plays a critical role in domains such as traffic management, resource optimization, and smart city planning. However, existing methods are often optimized for specific scenarios, limiting their ability to generalize across different urban environments. Although pre-trained universal models have made significant progress in cross-region prediction, their performance in some tasks may still fall short of specialized models. To address the trade-off between performance and generalization, we propose a Retrieval-Augmented Spatio-temporal (RAST) universal model. By integrating pre-trained universal models with dynamic retrieval mechanisms, our approach efficiently handles diverse spatio-temporal tasks across multiple regions, reduces data redundancy, and improves prediction accuracy, thereby enhancing the generalization capabilities of spatio-temporal models. Our contributions include: (i) a scalable architecture that supports cross-region generalization, (ii) an effective pre-training strategy that captures complex spatio-temporal relationships, and (iii) a knowledge-guided retrieval mechanism that enhances model adaptability in few-shot and zero-shot scenarios. Extensive experiments in urban settings demonstrate that the RAST model significantly outperforms traditional methods, particularly in data-scarce environments. This research opens up a promising direction for developing more flexible and robust universal models for urban spatio-temporal prediction, especially in cross-region and cross-domain applications. The code implementation is released on https://github.com/RWLinno/RAST.

## CCS Concepts

• **Computing methodologies** → **Machine learning approaches**.

## Keywords

Pre-trained model, Retrieval-augmented, Spatio-temporal data, universal model

**ACM Reference Format:**
Weilin Ruan and Yuxuan Liang. 2018. Retrieval-Augmented Universal Models for Spatio-temporal Data. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

*Both authors contributed equally to this research.


*Conference acronym 'XX, June 03–05, 2018, Woodstock, NY*

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

With the rapid advancement of global urbanization and digitalization, cities are generating an immense amount of spatio-temporal data across multiple domains such as traffic management, human mobility, and environmental monitoring [14, 32]. This data is crucial for smart city applications, enabling urban planners and administrators to make more data-driven decisions in resource allocation and urban planning [7, 38, 39]. In recent years, spatio-temporal prediction tasks have increasingly become a research hotspot, aiming to leverage historical data to forecast future spatio-temporal dynamics. However, due to the complexity and diversity of spatio-temporal data [35], existing methods still face significant challenges in balancing generalization capability and prediction efficiency.

Early spatio-temporal prediction methods primarily relied on statistical models, such as Historical Average (HA) [30] and Auto-Regressive Integrated Moving Average (ARIMA) [2, 25]. These methods have certain advantages in capturing temporal dependencies but often fall short in handling complex spatial correlations. With the development of machine learning, models like Vector Auto-Regression (VAR) [28, 42] and Artificial Neural Networks (ANN) [12] were introduced to address non-linear relationships in the data. However, traditional machine learning models still exhibit limitations in scalability and effectively capturing complex spatial dependencies. The rise of deep learning has significantly enhanced the capability of spatio-temporal modeling. Convolutional Neural Networks (CNNs) [9] excel in capturing spatial dependencies, while Recurrent Neural Networks (RNNs) [36] and their variants, such as Long Short-Term Memory (LSTM) [11] networks, perform exceptionally well in modeling temporal dynamics [11, 20]. The combination of these deep learning approaches provides robust support for more complex spatio-temporal tasks. In recent years, Spatio-Temporal Graph Neural Networks (STGNNs) have further advanced the modeling of intricate spatio-temporal relationships by integrating graph structures with temporal modeling mechanisms, making them essential tools for traffic forecasting and human mobility prediction [1, 22, 33, 35].

The emergence of Self-Supervised Learning [4, 23, 27, 29] and Large Language Models (LLMs) [16, 17, 34, 40] has opened new opportunities for spatio-temporal data modeling. Large-scale pre-trained models have demonstrated substantial potential in enhancing model generalization and adaptability in Natural Language Processing (NLP) [5, 15] and Computer Vision (CV) [6, 10]. These models are trained on diverse datasets across multiple tasks and domains, enabling them to possess strong feature extraction and

transfer learning capabilities, thereby significantly improving cross-domain and cross-task prediction performance [24, 26, 37].

Despite the significant advantages brought by SSL and large-scale pre-trained models, numerous challenges remain in the field of spatio-temporal prediction:

- *Trade-off between Generalization Capability and Prediction Efficiency:* Different geographic regions exhibit significant differences in data distributions, traffic patterns, and urban infrastructures, limiting model generalization in new regions. Additionally, the large-scale and high-dimensional nature of spatio-temporal data increases the computational complexity of models, making it challenging to maintain high prediction efficiency while enhancing generalization capability.
- *Difficulty in Training and Dynamically Updating Pre-trained Spatio-temporal Models:* Large-scale pre-trained spatio-temporal models often require extensive labeled data and complex training processes, especially when dealing with multi-region data. Moreover, as urban environments dynamically change, efficiently updating and adapting spatio-temporal information for each region remains a pressing issue.

To address these challenges, researchers have begun exploring the application of Retrieval-Augmented Generation (RAG) models in spatio-temporal prediction. RAG models have achieved remarkable success in NLP and CV by integrating external retrieval mechanisms, allowing models to dynamically access relevant historical data or external knowledge bases during inference [15, 18, 21, 41]. This integration significantly enhances performance in knowledge-intensive tasks by not only improving model generalization but also increasing adaptability to evolving data and distribution shifts. However, the application of RAG models in the domain of spatio-temporal data remains underexplored. Spatio-temporal data involves complex multi-dimensional dependencies that require more refined and dynamic retrieval strategies to effectively integrate spatial and temporal information.

Motivated by this gap, we propose a novel framework for universal spatio-temporal models: the Retrieval-Augmented Spatio-Temporal Model (RAST). RAST combines pre-trained universal models with dynamic retrieval mechanisms, enabling the model to dynamically retrieve relevant historical data or external knowledge during inference. This integration reduces data redundancy and enhances both prediction performance and generalization capability across multiple regions and domains. Compared to existing methods, RAST offers several key contributions:

- *Scalable Architecture Supporting the Trade-off between Generalization Capability and Prediction Efficiency:* We introduce the Retrieval-Augmented Spatio-Temporal Model (RAST), a framework that enhances spatio-temporal generalization across multiple urban scenarios by dynamically retrieving relevant information during inference, thereby improving the model's adaptability to new tasks and unseen regions while optimizing prediction efficiency.
- *Efficient Pre-training Strategy for Capturing Complex Spatio-temporal Relationships:* Through pre-training on large and diverse spatio-temporal datasets, RAST effectively captures long-term patterns in the data, significantly boosting performance in few-shot and zero-shot settings. Additionally, RAST incorporates knowledge-guided retrieval mechanisms to enhance model adaptability, enabling dynamic access to relevant historical data or external knowledge based on current task requirements.
- *Comprehensive Evaluation on Real-world Spatio-temporal Datasets:* We conduct extensive experiments across multiple real-world urban scenarios, demonstrating that RAST significantly outperforms existing baseline models in cross-region generalization, few-shot learning, and zero-shot learning, particularly in data-scarce environments.

## 2 Related Work

***Retrieval-Augmented Generation for Large Language Models.*** Retrieval-Augmented Generation (RAG) has emerged as a powerful paradigm for improving the performance of large language models (LLMs), particularly in knowledge-intensive tasks [19]. Traditional LLMs, while highly capable in a variety of applications, often struggle with tasks requiring vast amounts of factual knowledge or domain-specific expertise. By integrating external retrieval mechanisms, RAG enables models to dynamically access relevant information from large knowledge bases during inference, significantly enhancing their ability to handle complex or specialized queries [8, 21]. Several works, such as RAG-Token and RAG-Sequence, have demonstrated the superiority of retrieval-augmented models in areas like open-domain question answering, few-shot learning, and fact-checking [13]. Rather than solely relying on the model's internalized knowledge, RAG dynamically retrieves the most relevant information from external sources, which can be particularly beneficial when dealing with data that is constantly evolving or when the model encounters out-of-distribution inputs. However, while RAG frameworks have been extensively explored in natural language processing (NLP), their application in spatio-temporal data domains remains underdeveloped. Spatio-temporal tasks often involve highly dynamic and context-dependent data, where both spatial and temporal correlations must be considered. In this paper, we extend the RAG framework to spatio-temporal prediction tasks, aiming to improve the generalization ability of models across different cities by dynamically retrieving relevant patterns from past data. Our approach leverages the strengths of retrieval-based methods to reduce data redundancy while enhancing model adaptability in diverse prediction scenarios.

***Pre-trained Models for Spatio-temporal Data.*** Pre-trained models have shown remarkable success in fields like NLP [3, 5, 31] and computer vision [10], and their potential in spatio-temporal data modeling has become increasingly recognized [29]. Spatio-temporal data, which includes both spatial and temporal dimensions (e.g., traffic flow, weather conditions, human mobility), requires models capable of capturing complex interactions between these two dimensions. Traditional approaches, such as convolutional neural networks (CNNs) [9], recurrent neural networks (RNNs) [36], and graph neural networks (GNNs) , have been applied to spatio-temporal tasks with some success, but they often require large amounts of labeled data and struggle to generalize across different regions. Recent research has explored the use of pre-trained models for spatio-temporal prediction, particularly in urban computing. Models like UniST [37] and UniTime [26] aim to develop universal

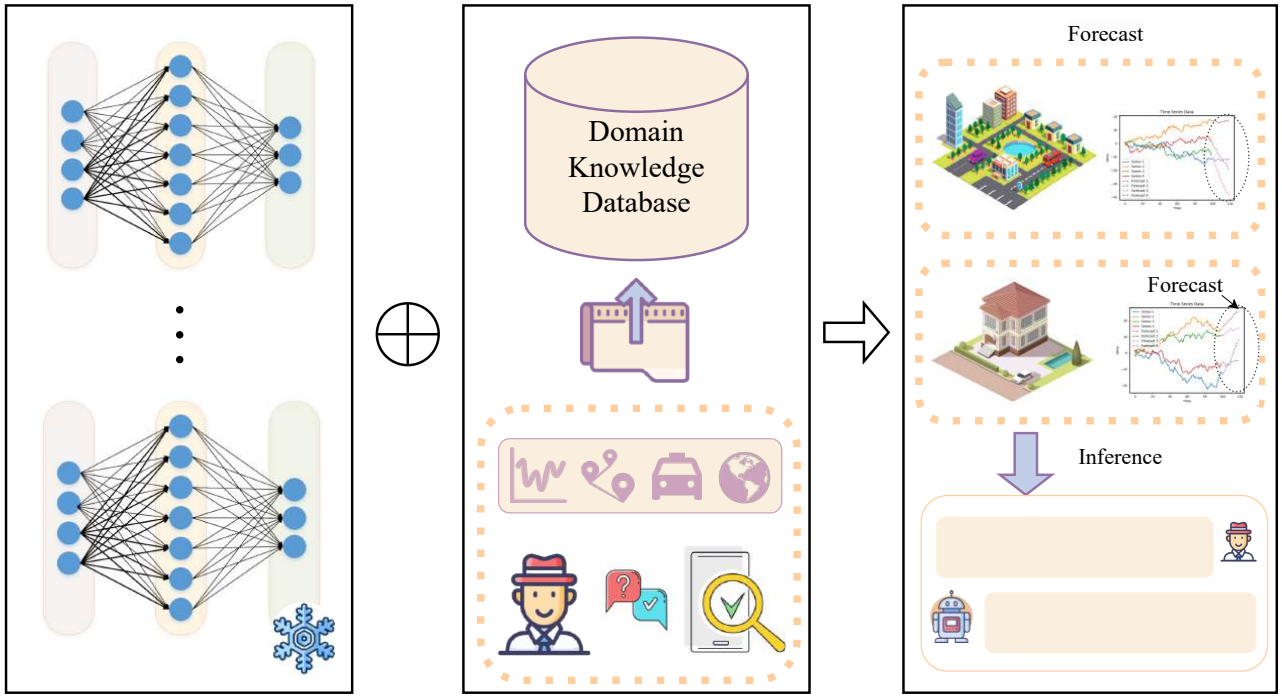


**Figure 1: The framework of our proposed method. (a) shows the cross-domain pre-training stage of the STGNNs. (b) is the retrieval-augmented stage for a specific domain. And (c) is the downstream application for our model, including forecasting and inference.**

frameworks that can generalize across regions by leveraging shared patterns in spatio-temporal data. These models are pre-trained on large datasets spanning multiple cities and domains, capturing underlying periodic patterns, spatial dependencies, and temporal correlations. However, pre-trained models still face challenges in terms of scalability and adaptability, especially in data-scarce environments or when dealing with out-of-distribution data. To address these limitations, our proposed RAST model combines the strengths of pre-training with retrieval-based techniques. By integrating a retrieval mechanism, we allow the model to dynamically access relevant historical data during inference, thereby improving both the accuracy and efficiency of spatio-temporal predictions. This approach not only enhances the model's generalization capabilities across different regions but also reduces the amount of redundant data processing, making it more efficient in real-world urban computing applications.

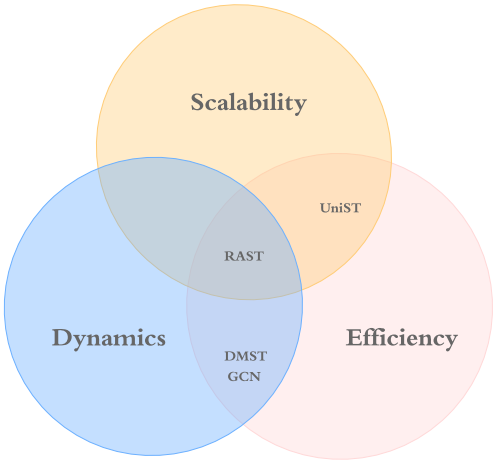


**Figure 2: The advantage of our proposed method.**

**Table 1: The Statistics Details of the Dataset.**

| Datasets | #Points | #Samples | #TimeSlices | Timespan |
|---|---|---|---|---|
| CA | 8600 | 301M | 35040 | 01/01/2019-12/31/2019 |
| GLA | 3834 | 134M | 35040 | 01/01/2019-12/31/2019 |
| GBA | 2352 | 82M | 35040 | 01/01/2019-12/31/2019 |
| SD | 716 | 25M | 35040 | 01/01/2019-12/31/2019 |

## 3 Methodology

## 4 Experiments

### 4.1 Setup.

### 4.2 Dataset.

### 4.3 Metrics.

### 4.4 Long-term Forecast

### 4.5 Generation

## 5 Conclusion

In this paper, we introduced RAST, a novel retrieval-augmented framework for spatio-temporal prediction. By integrating dynamic retrieval with pre-trained models, RAST addresses key challenges of generalization and adaptability in diverse urban environments. Our approach improves performance in data-scarce scenarios, reduces redundancy, and enhances cross-region generalization. Extensive experiments demonstrate the effectiveness of RAST, particularly in urban computing, where traditional models often fall short. For future work, we aim to improve the efficiency of the retrieval mechanism and explore applying RAST to a broader range of downstream tasks. Additionally, integrating more complex external knowledge sources and refining the framework for real-time applications will be key directions to further enhance its utility.

## 6 Acknowledgments

This work is mainly supported by the National Natural Science Foundation of China (No. 62402414). This work is also supported by the Guangzhou-HKUST(GZ) Joint Funding Program (No. 2024A03J0620), Guangzhou Municipal Science and Technology Project (No. 2023A03J0011), the Guangzhou Industrial Information and Intelligent Key Laboratory Project (No. 2024A03J0628), and a grant from State Key Laboratory of Resources and Environmental Information System, and Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No. 2023B1212010007).

## References

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

[2] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. 2015. *Time series analysis: forecasting and control*. John Wiley & Sons.

[3] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[4] Jiewen Deng, Renhe Jiang, Jiaqi Zhang, and Xuan Song. 2024. Multi-Modality Spatio-Temporal Forecasting via Self-Supervised Learning. *arXiv preprint arXiv:2405.03255* (2024).

[5] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[7] Rong Du, Paolo Santi, Ming Xiao, Athanasios V Vasilakos, and Carlo Fischione. 2018. The sensable city: A survey on the deployment and management for smart city monitoring. *IEEE Communications Surveys & Tutorials* 21, 2 (2018), 1533–1560.

[8] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* (2023).

[9] Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. 2018. Recent advances in convolutional neural networks. *Pattern recognition* 77 (2018), 354–377.

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.

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

[12] Wenhao Huang, Guojie Song, Haikun Hong, and Kunqing Xie. 2014. Deep architecture for traffic flow prediction: deep belief networks with multitask learning. *IEEE Transactions on Intelligent Transportation Systems* 15, 5 (2014), 2191–2201.

[13] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research* 24, 251 (2023), 1–43.

[14] 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* (2023).

[15] KyoHoon Jin, JeongA Wi, EunJu Lee, ShinJin Kang, SooKyun Kim, and YoungBin Kim. 2021. TrafficBERT: Pre-trained model with large-scale data for long-range traffic flow forecasting. *Expert Systems with Applications* 186 (2021), 115738.

[16] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728* (2023).

[17] 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*.

[18] Zhi Jing, Yongye Su, Yikun Han, Bo Yuan, Haiyun Xu, Chunjiang Liu, Kehai Chen, and Min Zhang. 2024. When large language models meet vector databases: a survey. *arXiv preprint arXiv:2402.01763* (2024).

[19] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*. PMLR, 15696–15707.

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).

[21] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[22] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2017. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926* (2017).

[23] Zhonghang Li, Chao Huang, Lianghao Xia, Yong Xu, and Jian Pei. 2022. Spatial-temporal hypergraph self-supervised learning for crime prediction. In *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 2984–2996.

[24] Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. 2024. Urbangpt: Spatio-temporal large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5351–5362.

[25] Marco Lippi, Matteo Bertini, and Paolo Frasconi. 2013. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. *IEEE Transactions on Intelligent Transportation Systems* 14, 2 (2013), 871–882.

[26] Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmermann. 2024. Unitime: A language-empowered unified model for cross-domain time series forecasting. In *Proceedings of the ACM on Web Conference 2024*. 4095–4106.

[27] Yixin Liu, Ming Jin, Shirui Pan, Chuan Zhou, Yu Zheng, Feng Xia, and S Yu Philip. 2022. Graph self-supervised learning: A survey. *IEEE transactions on knowledge and data engineering* 35, 6 (2022), 5879–5900.

[28] Helmut Lütkepohl. 2005. *New introduction to multiple time series analysis*. Springer Science & Business Media.

[29] Zezhi Shao, Zhao Zhang, Fei Wang, and Yongjun Xu. 2022. Pre-training enhanced spatial-temporal graph neural network for multivariate time series forecasting. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and*

*data mining*. 1567–1577.

[30] Brian L Smith and Michael J Demetsky. 1997. Traffic flow forecasting: comparison of modeling approaches. *Journal of transportation engineering* 123, 4 (1997), 261–266.

[31] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[32] Senzhang Wang, Jiannong Cao, and S Yu Philip. 2020. Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering* 34, 8 (2020), 3681–3700.

[33] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. 2019. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121* (2019).

[34] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM on Web Conference 2024*. 4006–4017.

[35] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2017. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875* (2017).

[36] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: LSTM cells and network architectures. *Neural computation* 31, 7 (2019), 1235–1270.

[37] 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.

[38] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Thirty-first AAAI conference on artificial intelligence*.

[39] Xinhu Zheng, Wei Chen, Pu Wang, Dayong Shen, Songhang Chen, Xiao Wang, Qingpeng Zhang, and Liuqing Yang. 2015. Big data for social transportation. *IEEE transactions on intelligent transportation systems* 17, 3 (2015), 620–630.

[40] Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. 2024. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems* 36 (2024).

[41] Kunlun Zhu, Yifan Luo, Dingling Xu, Ruobing Wang, Shi Yu, Shuo Wang, Yukun Yan, Zhenghao Liu, Xu Han, Zhiyuan Liu, et al. 2024. Rageval: Scenario specific rag evaluation dataset generation framework. *arXiv preprint arXiv:2408.01262* (2024).

[42] Eric Zivot and Jiahui Wang. 2006. Vector autoregressive models for multivariate time series. *Modeling financial time series with S-PLUS®* (2006), 385–429.