ST-MoE-BERT: A Spatial-Temporal Mixture-of-Experts Framework for Long-Term Cross-City Mobility Prediction

Haoyu He
Northeastern University
Boston, MA, USA
he.haoyu1@northeastern.edu

Haozheng Luo
Northwestern University
Evanston, IL, USA
robinluo2022@u.northwestern.edu

Qi R. Wang
Northeastern University
Boston, MA, USA
q.wang@northeastern.edu

Abstract

Predicting human mobility across multiple cities presents significant challenges due to the complex and diverse spatial-temporal dynamics inherent in different urban environments. In this study, we propose a robust approach to predict human mobility patterns called **ST-MoE-BERT**. Compared to existing methods, our approach frames the prediction task as a spatial-temporal classification problem. Our methodology integrates the Mixture-of-Experts architecture with BERT model to capture complex mobility dynamics and perform the downstream human mobility prediction task. Additionally, transfer learning is integrated to solve the challenge of data scarcity in cross-cities prediction. We demonstrate the effectiveness of the proposed model on GEO-BLEU and DTW, comparing it to several state-of-the-art methods. Notably, ST-MoE-BERT achieves an average improvement of 8.29%. [1]

CCS Concepts

• **Computing methodologies** → **Neural networks**; • **Information systems** → *Spatial-temporal systems.*

Keywords

human mobility prediction, spatial-temporal modeling, Mixture of Experts (MoE), BERT, cross-city forecasting

ACM Reference Format:
Haoyu He, Haozheng Luo, and Qi R. Wang. 2024. ST-MoE-BERT: A Spatial-Temporal Mixture-of-Experts Framework for Long-Term Cross-City Mobility Prediction . In *2nd ACM SIGSPATIAL International Workshop on the Human Mobility Prediction Challenge (HuMob'24), October 29-November 1, 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3681771.3699910

1 Introduction

Human mobility is a cornerstone of societal functioning, underpinning economic activities, urban development, and social interactions [34]. Understanding human travel patterns is essential for optimizing transportation systems, enhancing urban planning, and managing public health initiatives [1, 14]. In recent years, human mobility prediction has emerged as a critical component in the development of intelligent urban systems [23]. Accurate forecasting of individual and population movement patterns enables proactive decision-making, improves resource allocation, and enhances the responsiveness of services to dynamic urban environments.

However, predicting human mobility poses several significant challenges. Firstly, mobility data are often sparse and unevenly distributed spatially and temporarily, making it difficult to capture comprehensive movement patterns [39]. Secondly, human behaviors are inherently complex and influenced by a myriad of factors such as social interactions [41] and environmental changes [3], which complicates the modeling of their dependencies [32]. Lastly, transferring predictive models between different cities is challenging due to the heterogeneity in mobility patterns, urban infrastructure, and demographic characteristics [5].

To address these challenges, we propose **ST-MoE-BERT** (Spatial-Temporal Mixture-of-Experts with BERT), an innovative framework that integrates transformer-based architectures with a Mixture-of-Experts (MoE) layer [17]. The structure of our proposed framework is shown in Figure 1. ST-MoE-BERT leverages the sequence modeling capabilities of BERT [7] and the specialized expertise of MoE networks to capture both general and city-specific mobility patterns. Furthermore, our transfer learning strategy enables the model to adapt knowledge from data-rich cities to those with limited datasets, thereby enhancing prediction accuracy across multiple urban environments. Our key contributions are as follows.

- We introduce ST-MoE-BERT, a unique transformer-based method that combines BERT with an MoE layer to effectively address the long-term cross-city mobility prediction problem.
- We develop a transfer learning strategy that employs differential learning rates, thereby enhancing prediction accuracy in data-scarce environments and facilitating adaptation to diverse spatial distributions.
- Experimentally, we demonstrate that ST-MoE-BERT outperforms the state-of-the-art methods on the long-term cross-city prediction task with an average improvement of 8.29%.

2 ST-MoE-BERT

In this section, we present the proposed ST-MoE-BERT framework for long-term human mobility prediction.

Problem Setup. Given a sequence of historical mobility records $\mathcal{X} = \{x_1, x_2, \ldots, x_T\}$, where each $x_t \in \mathcal{L}$ represents the user's location at time step t, the goal is to predict the future mobility trajectory $\mathcal{Y} = \{y_{T+1}, y_{T+2}, \ldots, y_{T+H}\}$, where each $y_t \in \mathcal{L}$ denotes

[1] Code is available at **Github**.

HuMob'24, October 29-November 1, 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-1150-3/24/10
https://doi.org/10.1145/3681771.3699910

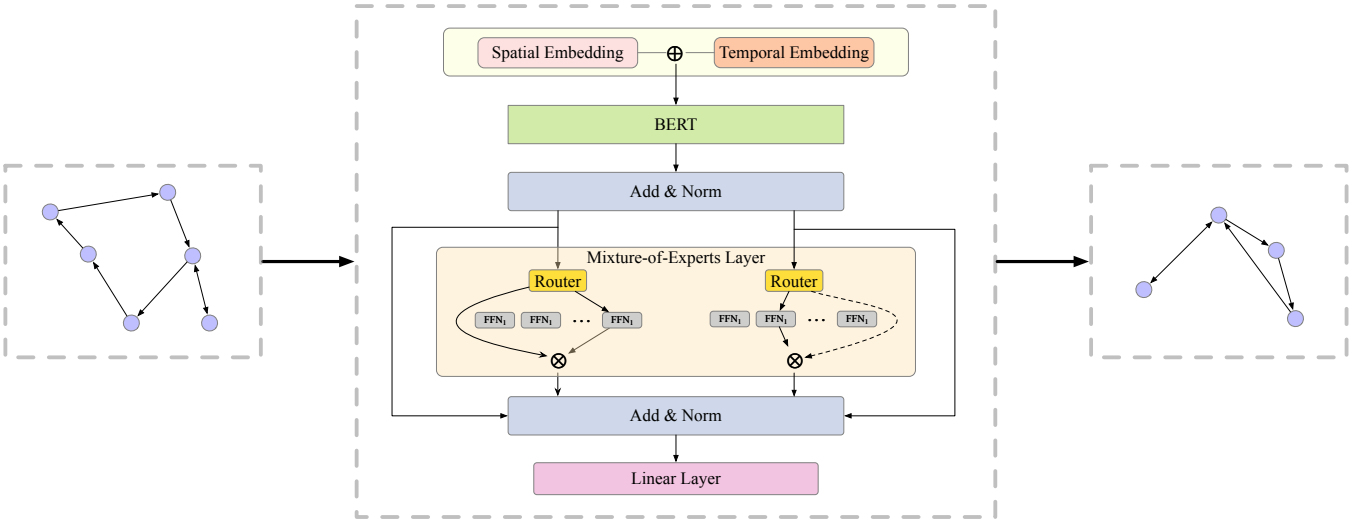

Figure 1: Overview of ST-MoE-BERT workflow. Input trajectories are first encoded with spatial and temporal embeddings, then combined and fed into the BERT model. The resulting representations are passed through a Mixture-of-Experts (MoE) layer consisting of eight feedforward network (FFN) experts. The MoE router selects the top two experts for each input, and their outputs are merged in a final linear layer to generate the predicted locations.

the user's location at time step t. Here, T is the length of the historical sequence, and H is the prediction horizon.

To achieve this, we aim to train a predictor $f : \mathcal{L}^T \rightarrow \mathcal{L}^H$ that minimizes the cross-entropy loss between the predicted and ground-truth trajectories on the training set of N trajectories. The loss function $\mathcal{J}$ is defined as:

$$\mathcal{J} = -\sum_{t=T+1}^{T+H} \sum_{i=1}^{N} \mathbb{I}(y_{t,i}) \log \hat{y}_{t,i},$$

where $\mathbb{I}(y_{t,i})$ equals 1 if the prediction for sample i at time t is correct, otherwise 0; and $\hat{y}_{t,i}$ represents the predicted probability distribution over all possible locations for sample i at time t.

In this classification set-up, each location is treated as a distinct class, and the predictor f outputs a probability distribution over the classes for each future time step.

Motivating Example. Recently, Liu et al. [22] propose modeling multivariate correlations by treating independent time series as tokens using self-attention mechanism. We use this observation as a starting point by considering our spatial-temporal foundation model with a combination embedding of weekday, day, time of day, and weekend. Since our problem is a long-term classification problem, we use the autoregressive model BERT [7] to model the temporal dependencies in the mobility data. To capture diverse and long-term mobility patterns, we incorporate a Mixture of Experts (MoE) layer [17] into our model architecture. The MoE layer consists of multiple specialized expert networks, each designed to model distinct aspects of human mobility. A gating network dynamically assigns the input data to these experts, allowing the model to adaptively focus on the relevant patterns for each prediction task.

2.1 Model Structure

In this section, we introduce our proposed ST-MoE-BERT framework. The primary architecture combines BERT with a Mixture of Experts (MoE) layer.

BERT.. To effectively capture the intricate temporal dependencies in human mobility data, we integrate the transformer architecture [31] into our model. Specifically, we employ BERT [7] as the backbone, leveraging its robust self-attention mechanisms to understand the relationship between a user's historical and future trajectories. Additionally, BERT excels at encoding contextual information from input sequences, and thus the model is able to discern complex movement patterns over time.

The self-attention mechanism, defined as:

$$\text{Attention}(X) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

computes attention weights by projecting the input sequence into query (Q), key (K), and value (V) matrices. This process allows the model to assign varying levels of importance to different time steps in historical sequences to focus on the most relevant locations when predicting future movements. The Softmax function ensures that the attention weights are normalized to effectively aggregate information across the sequence.

Furthermore, BERT utilizes a *[CLS]* token to capture global information from the entire input sequence. We leverage this *[CLS]* token as the input for the subsequent layer, ensuring that both local and global temporal contexts are effectively utilized.

Mixture of Experts. To effectively capture diverse and long-term mobility patterns across multiple cities, we incorporate a Mixture of Experts (MoE) layer [17] into our model architecture. The MoE layer comprises multiple specialized expert networks, each

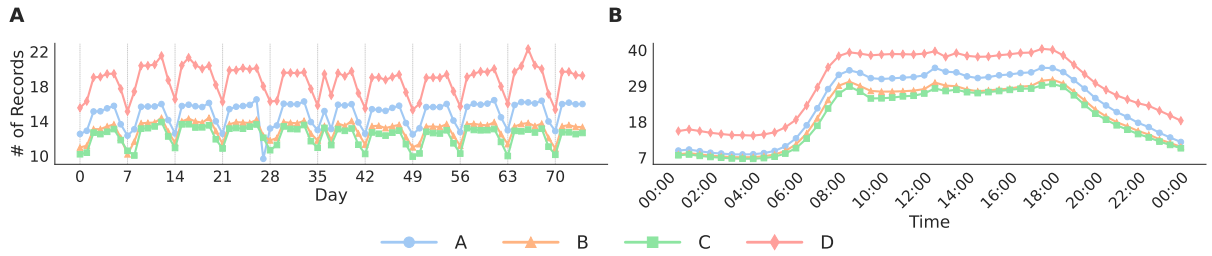

Figure 2: Average Mobility Records Per User Over Different Time Scales. A: Average number of records per user over a 75-day period in four cities, displaying a weekly cyclical pattern with higher records from Monday to Friday. **B:** Average half-hourly records per user for a single day, showing peak activity between 8 AM and 6 PM.

designed to model distinct aspects of mobility patterns, varying spatial regions, or temporal behaviors. A gating network dynamically assigns the input data to these experts, allowing the model to adaptively focus on the relevant patterns for each prediction task.

Mathematically, for an input feature vector $\mathbf{x} \in \mathbb{R}^d$, the MoE layer computes the output as:

$$\text{MoE}(\mathbf{x}) = \sum_{k=1}^{K} g_k(\mathbf{x}) f_k(\mathbf{x}),$$

where $f_k(\mathbf{x})$ represents the output of the k-th expert, and $g_k(\mathbf{x})$ is the gating network's probability for expert k, satisfying $\sum_{k=1}^{K} g_k(\mathbf{x}) = 1$ and $g_k(\mathbf{x}) \geq 0$. The gating probabilities $g_k(\mathbf{x})$ are derived using a softmax function:

$$g_k(\mathbf{x}) = \frac{\exp(\mathbf{w}_k^\top \mathbf{x} + b_k)}{\sum_{j=1}^{K} \exp(\mathbf{w}_j^\top \mathbf{x} + b_j)}.$$

In the context of human mobility prediction, the MoE layer allows the model to specialize different experts to distinct mobility patterns, such as varying user behaviors across different cities or temporal trends. This specialization enhances the model's ability to generalize across diverse mobility scenarios. Furthermore, the gating mechanism ensures that the model dynamically selects the most relevant experts based on the input data, thereby adapting to evolving mobility trends and improving prediction accuracy in multi-city settings.

2.2 Transfer Learning

Transfer learning [36] involves taking a pre-trained model on a large dataset and fine-tuning it for a different but related task. Fürst et al. [9] offers a theoretical analysis of how transfer learning operates within transformer architectures, highlighting the benefits of leveraging pre-trained models on related tasks. In human mobility prediction, transfer learning enables the adaptation of the model's learned representations to new urban environments, thereby leveraging knowledge from previously studied cities to enhance prediction accuracy in others.

Drawing inspiration from recent advancements in dense associative memory models [9, 16], where the attention mechanism in transformers is conceptualized as a form of associative memory, we propose a transfer learning strategy tailored for multi-city mobility prediction. Specifically, we pretrain our model on a large-scale mobility dataset from one city and subsequently fine-tune it on smaller datasets from other cities. This approach allows the model to capture both general and city-specific mobility patterns.

To effectively learn the distinct spatial information of each city, we employ different learning rates during fine-tuning. Specifically, we set the learning rate for the location embeddings to be ten times higher than the base learning rate used for the other model parameters. This higher learning rate enables the location embeddings to rapidly adapt and capture the unique spatial characteristics of the target city. Conversely, we apply a smaller learning rate to the rest of the model parameters to preserve the general knowledge acquired during pretraining. This strategy facilitates the transfer of knowledge from one city to another, improves prediction accuracy, and addresses uneven data distributions across cities.

3 Experiments

In this section, we conduct a series of experiments to demonstrate the performance and transferability of ST-MoE-BERT.

Models. In our experiments, we validate our approach using ST-MoE-BERT. We adopt the BERT model [2] with 110 million parameters. The input sequence length is set to 240, and the prediction horizon is 48. We pretrain the model using the masked language modeling (MLM) technique on mobility data from city A and then fine-tune it on mobility data from cities B, C, and D. The hyperparameters of ST-MoE-BERT are detailed in Table 2.

Datasets. We evaluate our method on human mobility data from four metropolitan areas, labeled as cities A, B, C, and D, as provided by Yabe et al. [40]. Each city is divided into a 200×200 grid, where each cell corresponds to a 500-meter by 500-meter area. The mobility datasets cover a 75-day period, with individual movements recorded at 30-minute intervals within these grid cells.

Evaluation Metrics. To evaluate the performance of predicting future mobility trajectories, we employ three metrics: Accuracy, GEO-BLEU [29], and Dynamic Time Warping (DTW) [25]. Accuracy quantifies the percentage of correctly predicted grid cells in future trajectories. GEO-BLEU is a metric that accounts for both precision and temporal alignment between the predicted and actual trajectories. DTW measures the alignment and distance between

[2]https://huggingface.co/google-bert/bert-base-uncased

Table 1: Comparison of ST-MoE-BERT with Benchmark Methods for Cross-city Prediction. We perform experiments on cross-city prediction tasks using two baseline models. Evaluation metrics include GEO-BLEU, DTW, and accuracy across four cities. 'ST-MoE-BERT w/o PT' refers to the model trained directly on each city without pretraining on city A. The best results are shown in bold. In the majority of configurations, ST-MoE-BERT consistently outperforms all baselines.

Method	A			B			C			D		
	GEO-BLEU ↑	DTW ↓	Acc. ↑	GEO-BLEU ↑	DTW ↓	Acc. ↑	GEO-BLEU ↑	DTW ↓	Acc. ↑	GEO-BLEU ↑	DTW ↓	Acc. ↑
HF	0.266	80.3	20.4%	0.265	56.4	21.0%	0.251	42.4	20.8%	0.295	80.0	21.0%
BERT	0.256	35.7	23.8%	0.284	**20.6**	27.0%	0.253	65.6	18.2%	0.253	65.6	18.2%
ST-MoE-BERT w/o PT	**0.286**	**30.2**	**27.9%**	0.286	28.2	27.5%	0.294	20.7	27.9%	0.250	67.6	21.4%
ST-MoE-BERT	-	-	-	**0.297**	29.3	**28.7%**	**0.297**	**19.7**	**28.9%**	**0.300**	**48.1**	**26.5%**

the predicted and true trajectories and is commonly used in long-term mobility prediction tasks [15]. Further details on GEO-BLEU and DTW are provided in subsection B.4.

Data Processing. We preprocess the human mobility data by categorizing the grid cells into 40,000 distinct classes. The dataset is then divided into training and testing sets, with the first 65 days allocated for training and the following 15 days for testing. To assess the percentage of missing data, we analyze the completeness of the data in each city, as shown in Figure 3. In our experiments, we focus our analysis exclusively on data where each time window has corresponding location records. Additionally, we examine the daily and hourly trends of data records across cities, as presented in Figure 2.

3.1 Cross-City Prediction with ST-MoE-BERT

To evaluate the efficiency of our method on cross-city prediction, we compare ST-MoE-BERT with baseline models on predictions for cities B, C, and D. Each evaluation is conducted three times using different random seeds, and we report the average performance for each metric.

Baselines. To evaluate the performance of our method, we use Historical Frequency (HF), naive BERT, and ST-MoE-BERT without pretraining to assess the efficiency of ST-MoE-BERT on cross-city prediction. HF predicts future locations using historical visit patterns based on time and weekday.

Results. As shown in Table 1, we observe an average improvement of 10.30% in GEO-BLEU, 15.50% in DTW, and 21.40% in accuracy across three cities. It is evident that ST-MoE-BERT outperforms all baseline models in predicting mobility for cities B, C, and D. HF scores high on GEO-BLEU by matching frequent past locations but may predict positions far from the true grid. In most configurations, ST-MoE-BERT without pretraining (w/o PT) outperforms the vanilla BERT model, highlighting the effectiveness of the MoE mechanism in capturing complex relationships across different cities. However, there are exceptions where the naive BERT model outperforms ST-MoE-BERT without pretraining, such as the GEO-BLEU score in city D. One possible reason is that the naive BERT model is more attuned to the distinct data distribution of city D, which differs from the other cities.

3.2 Ablation Study: Impact of Transfer Learning on Prediction Performance

We conduct experiments to assess the impact of transfer learning on the final prediction results in cross-city prediction tasks. As shown in Table 1, we observe an average improvement of 8.29% in GEO-BLEU, 9.90% in DTW, and 10.76% in accuracy across the three cities. It is evident that ST-MoE-BERT, when utilizing knowledge from other cities, enhances the robustness of cross-city predictions. The only exception, where ST-MoE-BERT without transfer learning outperforms the model with transfer learning, is the DTW score in city B.

4 Discussion

We introduce ST-MoE-BERT for predicting long-term human mobility patterns across cities. ST-MoE-BERT utilizes a transformer-based architecture with an MoE layer to capture complex spatio-temporal dependencies in mobility data. To enhance model robustness in target cities with limited data, we apply transfer learning by pretraining the model on city A. Our experiments show that ST-MoE-BERT improves prediction accuracy by an average of 8.29% compared to state-of-the-art models. These results highlight the significance of combining transformer architectures with MoE layers and a strong transfer learning strategy for better long-term cross-city human mobility prediction.

Limitation and Future Work. Despite its strengths, ST-MoE-BERT relies on a large-scale dataset for pre-training which may limit the model's applicability in scenarios where such comprehensive data is unavailable. Future work can address these limitations by exploring the integration of LLMs, which have shown promise in capturing intricate and multifaceted aspects of human behavior.

Acknowledgments

He and Wang acknowledge the support from the U.S. National Science Foundation (NSF) under Grant No. 2125326, 2114197, 2228533, and 2402438. The authors are grateful for the support of NSF. Any opinions, findings, conclusions, or recommendations expressed in the paper are those of the author and do not necessarily reflect the views of the funding agencies.

References

[1] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. 2018. Human mobility: Models and applications. *Physics Reports* 734 (2018), 1–74.

[2] Ciro Beneduce, Bruno Lepri, and Massimiliano Luca. 2024. Large Language Models are Zero-Shot Next Location Predictors. *arXiv preprint arXiv:2405.20962* (2024).

[3] Sebastiano Bontorin, Simone Centellegher, Riccardo Gallotti, Luca Pappalardo, Bruno Lepri, and Massimiliano Luca. 2024. Mixing Individual and Collective Behaviours to Predict Out-of-Routine Mobility. *arXiv preprint arXiv:2404.02740* (2024).

[4] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. 2006. The scaling laws of human travel. *Nature* 439, 7075 (2006), 462–465.

[5] Benjamin D Dalziel, Babak Pourbohloul, and Stephen P Ellner. 2013. Human mobility patterns predict divergent epidemic dynamics among cities. *Proceedings of the Royal Society B: Biological Sciences* 280, 1766 (2013), 20130763.

[6] Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2023. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688* (2023).

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. 2018. Deepmove: Predicting human mobility with attentional recurrent networks. In *Proceedings of the 2018 world wide web conference*. 1459–1468.

[9] Andreas Fürst, Elisabeth Rumetshofer, Johannes Lehner, Viet T Tran, Fei Tang, Hubert Ramsauer, David Kreil, Michael Kopp, Günter Klambauer, Angela Bitto, et al. 2022. Cloob: Modern hopfield networks with infoloob outperform clip. *Advances in neural information processing systems* 35 (2022), 20450–20468.

[10] Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723* (2020).

[11] Azul Garza and Max Mergenthaler-Canseco. 2023. TimeGPT-1. *arXiv preprint arXiv:2310.03589* (2023).

[12] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *nature* 453, 7196 (2008), 779–782.

[13] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2024. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems* 36 (2024).

[14] Haoyu He, Hengfang Deng, Qi Wang, and Jianxi Gao. 2022. Percolation of temporal hierarchical mobility networks during COVID-19. *Philosophical Transactions of the Royal Society A* 380, 2214 (2022), 20210116.

[15] Haoyu He, Xinhua Wu, and Qi Wang. 2023. Forecasting Urban Mobility using Sparse Data: A Gradient Boosted Fusion Tree Approach. In *Proceedings of the 1st International Workshop on the Human Mobility Prediction Challenge*. 41–46.

[16] Jerry Yao-Chieh Hu, Pei-Hsuan Chang, Robin Luo, Hong-Yu Chen, Weijian Li, Wei-Po Wang, and Han Liu. 2024. Outlier-Efficient Hopfield Layers for Large Transformer-Based Models. In *Forty-first International Conference on Machine Learning (ICML)*.

[17] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation* 3, 1 (1991), 79–87.

[18] 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).

[19] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691* (2021).

[20] Yuebing Liang, Yichao Liu, Xiaohan Wang, and Zhan Zhao. 2024. Exploring large language models for human mobility prediction under public events. *Computers, Environment and Urban Systems* 112 (2024), 102153.

[21] Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2016. Predicting the next location: A recurrent model with spatial and temporal contexts. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.

[22] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. 2023. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625* (2023).

[23] Massimiliano Luca, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. 2021. A survey on deep learning for human mobility. *ACM Computing Surveys (CSUR)* 55, 1 (2021), 1–44.

[24] Haozheng Luo, Jiahao Yu, Wenxin Zhang, Jialong Li, Jerry Yao-Chieh Hu, Xingyu Xin, and Han Liu. 2024. Decoupled Alignment for Robust Plug-and-Play Adaptation. *arXiv preprint arXiv:2406.01514* (2024).

[25] Meinard Müller. 2007. Dynamic time warping. *Information retrieval for music and motion* (2007), 69–84.

[26] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. 2024. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems* 36 (2024).

[27] Yuanyuan Qiao, Zhongwei Si, Yanting Zhang, Fehmi Ben Abdesslem, Xinyu Zhang, and Jie Yang. 2018. A hybrid Markov-based model for human mobility prediction. *Neurocomputing* 278 (2018), 99–109.

[28] Injong Rhee, Minsu Shin, Seongik Hong, Kyunghan Lee, Seong Joon Kim, and Song Chong. 2011. On the levy-walk nature of human mobility. *IEEE/ACM transactions on networking* 19, 3 (2011), 630–643.

[29] Toru Shimizu, Kota Tsubouchi, and Takahiro Yabe. 2022. GEO-BLEU: similarity measure for geospatial sequences. In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*. 1–4.

[30] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. 2010. Limits of predictability in human mobility. *Science* 327, 5968 (2010), 1018–1021.

[31] 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).

[32] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. 2011. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1100–1108.

[33] Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *arXiv preprint arXiv:2310.04793* (2023).

[34] Qi Wang, Nolan Edward Phillips, Mario L Small, and Robert J Sampson. 2018. Urban mobility and neighborhood isolation in America's 50 largest cities. *Proceedings of the National Academy of Sciences* 115, 30 (2018), 7735–7740.

[35] Yu Wang, Tongya Zheng, Yuxuan Liang, Shunyu Liu, and Mingli Song. 2024. Cola: Cross-city mobility transformer for human trajectory simulation. In *Proceedings of the ACM on Web Conference 2024*. 3509–3520.

[36] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data* 3 (2016), 1–40.

[37] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).

[38] Hao Xue, Flora Salim, Yongli Ren, and Nuria Oliver. 2021. MobTCast: Leveraging auxiliary trajectory forecasting for human mobility prediction. *Advances in Neural Information Processing Systems* 34 (2021), 30380–30391.

[39] Takahiro Yabe, Massimiliano Luca, Kota Tsubouchi, Bruno Lepri, Marta C Gonzalez, and Esteban Moro. 2024. Enhancing human mobility research with open and standardized datasets. *Nature Computational Science* 4, 7 (2024), 469–472.

[40] Takahiro Yabe, Kota Tsubouchi, Toru Shimizu, Yoshihide Sekimoto, Kaoru Sezaki, Esteban Moro, and Alex Pentland. 2024. YJMob100K: City-scale and longitudinal dataset of anonymized human mobility trajectories. *Scientific Data* 11, 1 (2024), 397.

[41] Xiao-Yong Yan, Wen-Xu Wang, Zi-You Gao, and Ying-Cheng Lai. 2017. Universal model of individual and population mobility on diverse spatial scales. *Nature communications* 8, 1 (2017), 1639.

[42] Jiahao Yu, Haozheng Luo, Jerry Yao-Chieh Hu, Wenbo Guo, Han Liu, and Xinyu Xing. 2024. Enhancing Jailbreak Attack Against Large Language Models through Silent Tokens. *arXiv preprint arXiv:2405.20653* (2024).

A Related Work

In this section, we present a concise overview of human mobility prediction and the application of foundation models for time series classification.

Human Mobility Prediction. Researchers have applied physics-based models to understand human mobility, depicting movements as scale-free Lévy flights with distances following a truncated power-law distribution [4, 28]. Patterns such as frequently revisited locations align with Zipf's Law [12], while the number of unique places visited grows sublinearly [30]. Although Markov models have been used to predict future locations based solely on the current state [27], they struggle to capture the complex spatio-temporal dependencies and varied travel behaviors in human mobility.

With the advancement of deep learning, more sophisticated models have been developed to address these limitations. Recurrent neural networks, such as STRNN [21], were applied to mobility

prediction, effectively capturing temporal dependencies in sequential mobility data. Building upon this, the attention mechanism was introduced in DeepMove [8], allowing the model to focus on the most relevant part of a user's trajectories. MobTCast [38] employed context-aware transformer to effectively model users' social and geographical interactions. More recently, cross-city prediction approaches [35] have highlighted the transferability of mobility knowledge across different urban areas by leveraging shared spatio-temporal patterns. A novel direction involves the exploration of large language models (LLMs) for mobility prediction [2, 20]. These studies utilize the vast capabilities of LLMs, along with prompt-based learning, to effectively capture latent mobility patterns and provide more accurate, context-specific predictions across diverse environments.

Foundation Models for Time Series Classification. The advent of Transformer-based architectures, initially introduced for language translation tasks [7, 31], has paved the way for the development of foundation models across diverse domains. These architectures are capable of capturing complex relationships within various types of data, thereby leading to significant advancements in fields, such as safety [24, 42] and prompting [10, 19]. Meanwhile, they have demonstrated exceptional performances across multiple scientific disciplines, including genomics [26], finance [33, 37] and many others. In the context of time series forecasting, foundation models have also been explored for time series forecasting, with approaches built on existing LLMs [11, 18] showing promising results on tasks like forecasting, imputation, and anomaly detection. However, current time-series foundation models [6, 13, 18] rely on patch-based embedding mechanisms, which are inadequate for capturing the intricate spatio-temporal dependencies in human mobility data. In contrast to TimeLLM [18], ST-MoE-BERT is specifically designed to model these complex spatio-temporal relationships, offering more accurate and context-aware predictions for human mobility.

B Experiment Setting

B.1 Hyperparameters of ST-MoE-BERT

The hyperparameters of ST-MoE-BERT are presented in Table 2.

B.2 Data Completeness Rate

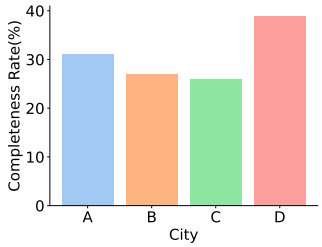

Figure 3: The completeness rate for each city.

Table 2: Hyperparameters of the Pretrained and Fine-Tuned Models

Pretrained Model	
Learning Rate	0.0003
Weight Decay	0.001
Number of Hidden Layers	12
Hidden Size	768
Number of Attention Heads	16
Number of Experts	8
Dropout	0.1
Day Embedding Size	64
Time Embedding Size	64
Day of Week Embedding Size	64
Weekday Embedding Size	32
Location Embedding Size	256
Fine-Tuned Model	
Learning Rate	0.00005
Location Embedding Learning Rate	0.0005

B.3 Distribution of Individuals Across Cities

Table 3: Number of users with movement data across cities A, B, C, and D.

City	A	B	C	D
Number of Users	100,000	25,000	20,000	6,000

B.4 Evaluation metrics

Given two sequences, $X = (x_1, x_2, \ldots, x_n)$ and $Y = (y_1, y_2, \ldots, y_m)$:

GEO-BLEU: GEO-BLEU extends the BLEU metric to geographic data. It evaluates n-gram similarities between sequences using the formula:

$$\text{GEO-BLEU} = BP * \exp\left(\sum_{n=1}^{N} w_n \log q_n\right)$$

where BP is the brevity penalty, w_n are the weights for each n-gram level, and q_n is the geometric mean of the n-gram similarities between the sequences.

DTW: DTW measures the similarity between X and Y by finding the optimal alignment that minimizes the sum of Euclidean distances between corresponding elements. It is defined as:

$$\text{DTW}(X, Y) = \min_{c}\left(\sum_{(i,j)\in c} \text{dist}(x_i, y_j)\right)$$

where c denotes the alignment path and $\text{dist}(x_i, y_j)$ is the Euclidean distance between x_i and y_j.