An LLM-Enhanced Conditional Diffusion Model for Mobile Traffic Prediction

Zhu Xiao, Rui Wang, Jing Bai, Tong Li, Shiyuan Zhang, Keqin Li, and Zhu Han

Abstract

Accurate traffic prediction is critical for ensuring the efficient operation of mobile networks and maintaining a seamless user experience. However, existing approaches face several significant challenges: high costs and incomplete environmental perception, difficulty capturing spatial correlations between base stations, and limited adaptability to irregular traffic fluctuations driven by user activities. This article proposes an LLM-enhanced Conditional Diffusion (LEC-Diff) model for cellular traffic prediction to address the abovementioned challenges. First, we utilize easily accessible aerial images to describe the static environmental context surrounding base stations. We further enhance this representation by leveraging comprehensive textual data from a Large Language Model (LLM) for these images, extracting the abundant inherent knowledge embedded within the LLM. Second, we employ a Graph Neural Network (GNN) to automatically model the spatial dependencies between base stations and enhance spatiotemporal information through mapping. Finally, we introduce a conditional diffusion model to capture complex traffic distributions by conditioning predictions on static environmental features and dynamic historical traffic features. Extensive experiments demonstrate that our proposed model surpasses state-of-the-art methods by over 5% in mobile traffic prediction.

Introduction

With the rapid evolution of wireless communications, diversified services such as short videos and live broadcasts have significantly accelerated mobile traffic growth. Rapidly developing these emerging mobile services has introduced significant challenges to cellular networks, including increasing pressure on network resources and issues like congestion and delays [1]. To overcome these challenges and deliver high-quality network services, accurate traffic prediction has become a critical capability for operators and infrastructure providers, enabling the optimization of resource allocation and proactive network control, ensuring the growing demand for reliable and efficient connectivity is met [2].

Numerous studies have sought to enhance the accuracy of mobile traffic prediction. Initially, researchers approached traffic prediction as a general time-series forecasting problem, using Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks [3] to capture the temporal dependencies of traffic at individual base stations. These purely temporal models fail to consider the spatial correlations between base stations. Thus, several studies have incorporated Graph Neural Networks (GNNs) into their frameworks, leveraging graph structures to model the spatial distributions of base stations. For example, Yu *et al.* [4] proposed an innovative architecture integrating the Graph Convolutional Network (GCN) with the gated Convolutional Neural Network (CNN). By extracting the topological structure of the graph network, the dynamic characteristics of mobile traffic are comprehensively analyzed, thereby effectively uncovering spatiotemporal variations in mobile traffic data. Beyond spatial and temporal dependencies, the contextual environmental information of base stations is also a critical factor in mobile traffic prediction. Gong *et al.* [5] were among the first to integrate environmental information into their models. Specifically, they constructed an urban knowledge graph to represent the spatial structure of a city and applied knowledge graph embedding techniques to capture environmental factors. However, building an urban knowledge graph for each city is not only time-intensive and labor-intensive but also requires access to sensitive data, such as citizens' trajectories, which raises privacy concerns. These issues severely limit the practicality of this approach in large-scale, real-world applications. Furthermore, most existing methods rely on deterministic predictions, which constrain their ability to model the distribution of traffic. As a result, these methods struggle to adapt to sudden traffic fluctuations caused by unexpected events. These limitations underscore the need for more robust and flexible tools to improve the performance of mobile traffic prediction, particularly in dynamic and complex environments.

In recent years, generative AI has achieved remarkable performance in domains such as image generation and Natural Language Processing (NLP) thanks to its high-quality generative capabilities, flexibility, and scalability [6]. Furthermore, generative AI models, such as diffusion models, excel

Zhu Xiao, Rui Wang, and Tong Li (corresponding author) are with Hunan University, China; Jing Bai is with Xidian University, China; Shiyuan Zhang is with Department of Electrical and Electronic Engineering, The University of Hong Kong, Pok Fu Lam, China; Keqin Li is with State University of New York, New Paltz, USA; Zhu Han is with the University of Houston, USA and Kyung Hee University, South Korea.

Digital Object Identifier: 10.1109/MCOM.001.2400779

0163-6804/25/$25.00 © 2025 IEEE

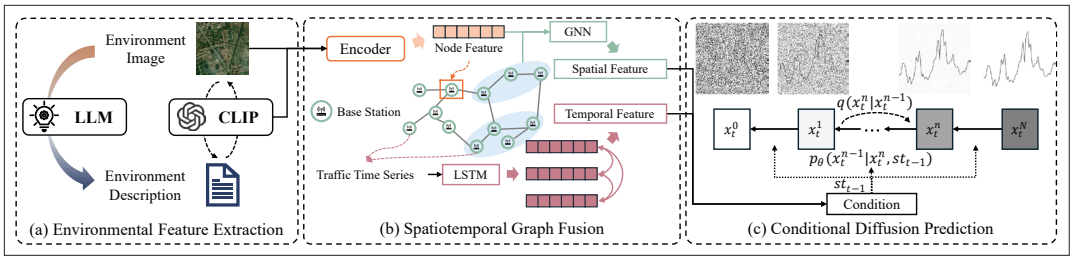

FIGURE 1. Overview of LEC-Diff. The Environmental Feature Extraction Module is responsible for extracting environmental features of base stations using image-text pairs. The Spatiotemporal Graph Fusion Module integrates environmental features with historical traffic temporal features across base stations by jointly modeling them through RNNs and GNNs. Finally, the Conditional Diffusion Prediction Module models future traffic distribution samples conditioned on the fused spatiotemporal features of each individual base station.

The GNN then fuses these node features through aggregation and propagation operations, providing a more comprehensive and accurate representation of spatiotemporal dependencies within the cellular network.

at effectively modeling underlying probability distributions and dependencies in data, showcasing their potential for time series prediction tasks that involve nonlinearity and non-stationarity [7]. Alternatively, Large Language Models (LLMs), as generative AI-based foundational models, are trained on vast corpora of natural language data [8, 9]. They have significant potential for leveraging the rich textual knowledge embedded in their vast training data to extract environmental features in urban settings, eliminating the need to construct urban knowledge graphs manually.

Despite the efforts, several challenging issues remain unresolved in the aforementioned works.

- **Environmental Feature Extraction for Base Stations.** Mobile users in similar environments exhibit similar network usage behaviors, leading to corresponding base stations sharing similar traffic patterns. This makes environmental contextual information critical and valuable for cellular traffic prediction [10, 11]. Existing studies have initially employed point-of-interest (POI) distributions to describe surrounding environments. POI data indicate locations of particular interest for specific purposes. However, such data are often inaccessible to operators and researchers as internet companies typically control them. Furthermore, POI datasets are not frequently updated, with updates often occurring over several years. Additionally, POI data provide only a limited view of the environment, neglecting rich regional textural features and urban structures such as building distributions and road layouts. Furthermore, some researchers have proposed building urban knowledge graphs to represent urban regional structures. In such graphs, elements like regions, business areas, and transportation hubs are represented as elements, with user behavior data used to establish relationships between these elements. However, constructing urban knowledge graphs is both time-consuming and labor-intensive. For new cities, this process requires manual data collection and graph construction, making it impractical in many scenarios. In summary, the effective and efficient extraction of environmental information remains a significant and unresolved challenge for cellular traffic prediction.
- **Traffic Uncertainty Fluctuations.** Mobile traffic is highly volatile, as it is influenced not only by spatiotemporal characteristics but also by numerous discrete potential factors. For instance, regional gatherings triggered by special events can cause sudden traffic surges. Traditional deterministic prediction methods often produce significant prediction errors in such cases [12]. In addition, abnormal traffic fluctuations caused by such behaviors in historical traffic data hinder model development, preventing the model from effectively capturing general patterns and reducing its overall accuracy.

In this article, we propose an LLM-Enhanced Conditional Diffusion model (LEC-Diff) for mobile traffic prediction. Figure 1 illustrates the framework of LEC-Diff, composed of three key modules, i.e., the LLM-enhanced environment feature extraction module, the spatiotemporal graph fusion module, and the conditional diffusion prediction module. In the LLM-enhanced environment feature extraction module, we leverage readily available aerial images, i.e., satellite images, to capture the urban environmental context surrounding base stations. Additionally, we harness the rich textual knowledge embedded in pre-trained LLMs to generate detailed textual descriptions for each aerial image, thereby enriching the contextual information. This process yields image-text pairs, which we utilize in a contrastive language-image pretraining [13] approach to integrate multimodal features effectively. This integration enables a more accurate and insightful understanding of environmental contexts. The spatiotemporal graph fusion module extracts spatiotemporal information by jointly modeling RNNs and GNNs. The GNN uses a graph structure to represent the spatial relationships among base stations, where nodes correspond to base stations and edges represent the distances between them. The environmental features extracted by the LLM-enhanced environmental feature extraction module are incorporated as the spatial features of the nodes. Simultaneously, the historical traffic time series of the base stations are processed by the RNN to generate the temporal features of the nodes. The GNN then fuses these node features through aggregation and propagation operations, providing a more comprehensive and accurate representation of spatiotemporal dependencies within the cellular network. Finally, in the conditional diffusion prediction module, the spatiotemporal enhancement information obtained from the spatiotemporal graph fusion module is used as a condition for the diffusion model. Through a Markov chain with Langevin sampling, white noise is progressively transformed into future traffic distribution samples under the guidance of conditional information, enabling accurate traffic predictions.

This enriched spatial context information is integrated with temporal features in an autoregressive model, providing accurate spatio-temporal guidance for the diffusion model to generate future traffic distributions.

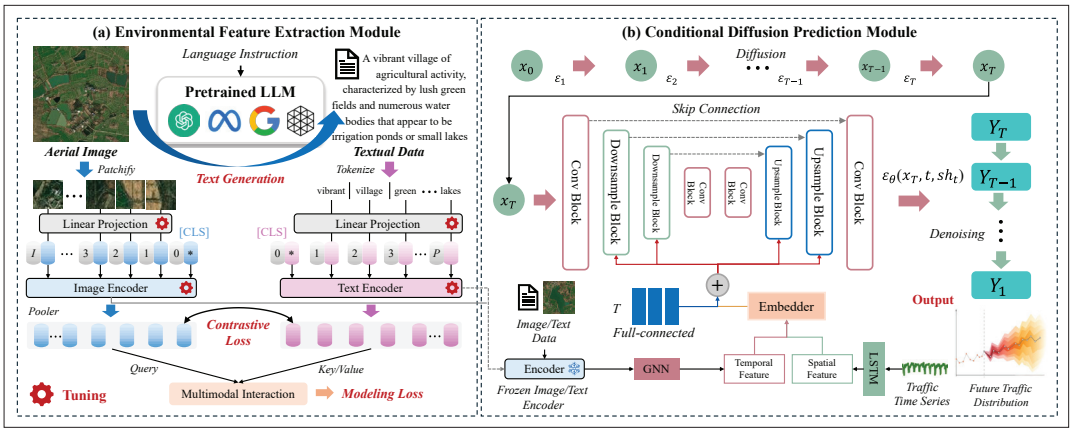

FIGURE 2. Architecture of LEC-Diff. The Environmental Feature Extraction Module leverages LLM to improve environmental perception and generate a comprehensive text modality, while the constructed image-text data pairs are aligned using CLIP. The conditional diffusion prediction module employs the diffusion model to predict future mobile traffic distributions using spatiotemporal conditions provided by the graph fusion module.

Our contributions are summarized as follows.

- We propose a generative AI-driven paradigm for mobile traffic prediction. The comprehensive text data generated by LLM is a critical supplement, aiding in the efficient and high-quality characterization of the spatial environment surrounding the base station. The conditional diffusion model can generate probability distribution predictions based on this enhanced spatial information and demonstrates greater robustness in adapting to uncertainties and dynamic variations in the cellular network environment.
- We propose a novel LEC-Diff model to enhance mobile traffic prediction. Specifically, the GNN effectively captures diverse relationships among base stations, environmental features, and inter-base station connections by utilizing context information enhanced by the LLM. This enriched spatial context information is integrated with temporal features in an autoregressive model, providing accurate spatiotemporal guidance for the diffusion model to generate future traffic distributions.
- We conduct extensive experiments on two real-world datasets. The results reveal that the proposed model achieved about 5% higher accuracy than the baselines, highlighting its superior accuracy and effectiveness. Furthermore, we analyze the roles of the modules within the network structure to confirm that the proposed method enhances the environmental perception and understanding capabilities for base station traffic prediction.

Solution

LEC-Diff comprises three key modules:

1. The environmental feature extraction module, which utilizes an LLM to analyze satellite images and extract spatial environmental information for base stations;
2. The spatiotemporal graph fusion engine, designed to capture the spatial distribution of base stations while integrating environmental and temporal latent information across spatiotemporal dimensions;
3. The autoregressive denoising diffusion traffic prediction engine, which employs an autoregressive diffusion model to generate accurate traffic predictions based on spatiotemporal information.

Environmental Feature Extraction Module

Given the complexity of factors affecting prediction, predicting the traffic of the base station is challenging if the environment in which the base station is located cannot be deeply analyzed. To better capture environmental information, as shown in Fig. 2a, we design the environmental feature extraction module. Specifically, we leverage the LLM to enhance the analysis of the environment around the base station and derive improved representations of base station environmental information through language-image comparison pre-training. As demonstrated in Fig. 3, empirical experiments with varying language instructions revealed that more detailed prompts — particularly those emphasizing specific aspects, such as urban infrastructure — can elicit the LLM's enhanced capability to generate high-quality summaries. Furthermore, given the well-documented hallucination issue, environmental descriptions generated by LLM frequently exhibit unrealistic or ambiguous information, which hinders the effective integration of LLM-based knowledge into the image encoder. To thoroughly enhance and produce high-quality environmental representations, it is essential to refine or rewrite textual content following established rules. Thus, we initially apply pre-configured regular filters to remove redundant and irrelevant textual information. Additionally, we integrate geographical and computational expertise to conduct secondary fact verification and devise a dual scoring mechanism to ensure the accuracy of this process.

The preprocessed images and texts are converted into slices using Pathify and Tokenize. After linear mapping, a unique token ([CLS]) is added at the beginning of the sequence to represent its overall information. The image and text data are then input into two unimodal encoders to encode the data into latent image and text representations. The LLM-enhanced semantic representation and the visual representation of the exact base station location are optimized to be as similar as possible. However, the inconsistent modal learning methods and the relationships between

different modalities may introduce ambiguity in the representations of base station environmental information. Therefore, we design a contrastive image-text loss to jointly optimize the image and text encoders by comparing the image-text pairs with other image-text pairs in the sampling batch. The contrastive loss function is formulated based on the InfoNCE (Information Noise Contrastive Estimation) framework. Its objective is to ensure that satellite images of the same urban area and their corresponding textual descriptions are closely aligned in the latent space. The similarity between images and text is quantified using a bidirectional loss function, ensuring alignment in both directions (image to text and text to image). Simultaneously, other samples within the batch serve as negative samples, enabling the model to learn more discriminative features and achieve improved generalization when applied to large-scale data. Finally, a cross-attention mechanism is incorporated within the multimodal interaction module to effectively learn a unified representation of image and text.

Admittedly, our proposed method for improving environmental representation using LLM leads to a higher computational cost during the training phase. However, we would like to emphasize that this design does not impose extra computational overhead during the inference phase, which is the most critical and frequently used stage after deploying our model in real-world networks. It is worth noting that environmental dynamics occur gradually. Once the representation vectors are generated using LLM during the training stage, they can be reutilized across different time periods and multiple queries within the inference stage. This means that although training computational costs may rise, the generated vectors can be continuously reused, thereby distributing computational expenses over time and yielding sustainable long-term advantages. In addition, our method can precisely and effectively model the spatial environment surrounding the base station. This comprehensive spatial contextual data, integrated with temporal features, provides robust and accurate spatiotemporal insights for future mobile traffic prediction, thereby effectively overcoming the limitations of existing approaches in environmental information extraction.

Spatiotemporal Graph Fusion Module

Base station traffic prediction tasks typically exhibit strong spatiotemporal dependencies. However, the nonlinear relationship between temporal and spatial information, as well as the conflict in learning dependencies across different dimensions, makes it challenging to integrate information effectively. We design the spatiotemporal graph fusion engine based on the unimodal encoder pre-trained earlier. As shown in Fig. 2b, we extract environmental features for each base station from satellite images using the pre-trained image encoder. We also extract adjacency relationships between base stations based on their spatial locations. Using the base-mentioned station-related data mentioned above, we model the spatial dependencies in mobile traffic across base stations by constructing a graph of a convolutional neural network. Finally, the spatial features output by the graph convolutional neural network are combined with the time series hidden features output by the recurrent neural network to enable an accurate spatiotemporal representation for base station traffic prediction. In the subsequent prediction process, the diffusion model can better capture the spatiotemporal distribution of base station traffic under complex spatial dependencies and dynamic changes.

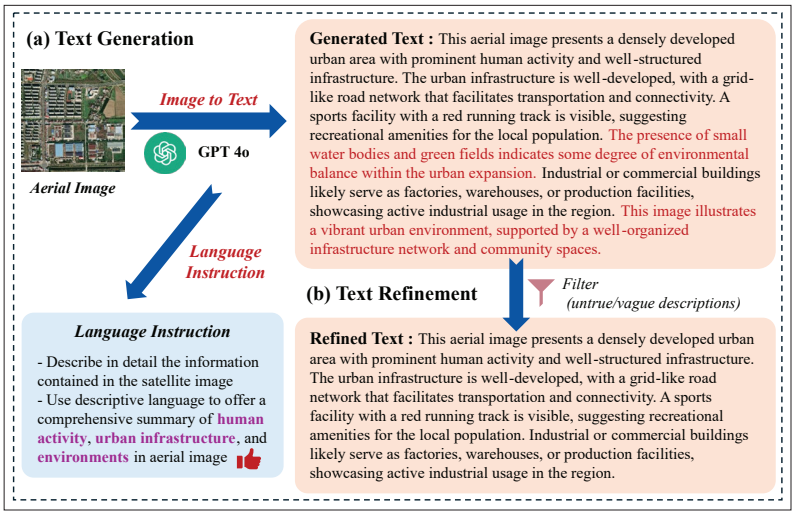

FIGURE 3. Text generation and refinement. Based on our carefully designed language instruction, we utilize a pre-trained LLM to generate detailed descriptions of the environment. To further eliminate unrealistic or ambiguous information within the text, we establish specific refinement rules to ensure accurate and high-quality representations of the environment.

Conditional Diffusion Prediction Module

Based on the spatiotemporal information obtained by the fusion network earlier, we construct LEC-Diff utilizing the autoregressive denoising diffusion model. As shown in Fig. 2b, the core of the diffusion model is the novel generative framework inspired by the diffusion process in non-equilibrium thermodynamics. The model consists of a forward diffusion process and a backward denoising process. At the same time, we formulate the base station traffic probability prediction task as modeling an autoregressive conditional probability distribution, precisely predicting the distribution of future values using past values and covariates as conditions.

RNN models the autoregressive process. A fixed distribution can represent the likelihood term, and a function can generate the parameters of the distribution. The information about past values is encoded into hidden states through the RNN sequential modeling process. Similar to the sequence-to-sequence (seq-to-seq) process in language models, this hidden state represents the output of the encoder. The prediction process involves obtaining an output from the past value encoder, fed into the decoder to generate the future value. The only difference between the encoder and the decoder is whether the ground truth is involved. The covariates, considered known conditions, consist of time-related features (e.g., the day of the week or the hour of the day), time-independent embeddings, and lagged features, which are determined by the frequency of the training dataset. We encode past period information and covariates into h using the RNN and feed h into the diffusion model to model the corresponding conditional probability distribu-

Analyzing spatial correlation based on spatial distribution and geographic patterns helps uncover the temporal patterns of change, thereby achieving higher performance in practical applications.

Dataset	Shanghai	Nanjing
Collection Duration	Aug. 1st–31st, 2014	Feb. 2nd–Mar. 31st, 2021
Time Interval	30 minutes	
Covered Users	≥ 150,000	≥ 1250,000
Covered BSs	4505	8000
Covered Area	6340	6587
Textual Description	11125	39400

TABLE 1. Statistics of the datasets used in our experiments.

tion. To better guide the diffusion model during prediction, we feed the RNN hidden state and perceived base station environmental information into the spatiotemporal graph fusion network to extract spatiotemporal dependency features of base station traffic, which serve as conditional information for the diffusion model.

Evaluation on Real-world base station

Datasets and Baselines

The datasets used in our experiments are derived from large-scale mobile cellular networks in two major Chinese cities, Shanghai and Nanjing [5]. Table 1 summarizes the statistics of the Shanghai and Nanjing datasets. The Shanghai dataset contains anonymous traffic data collected from 4,505 base stations at 30-minute intervals starting from August 2014. By spanning 6,340 regions, each data trace provides a comprehensive record of mobile data usage for over 150,000 users. It includes the anonymous device ID, the start time of the data connection, the base station location, and the amount of data used during the connection. The Nanjing dataset is larger than the Shanghai dataset, encompassing anonymous traffic data from 8,000 base stations collected at 30-minute intervals between February 2 and March 31, 2021, across 6,587 regions. This large-scale, fine-grained traffic data reinforces the credibility of our base station mobile traffic modeling and prediction. To achieve a more comprehensive understanding of the base station environment, we leverage the ArcGIS platform's map API (Application Programming Interface) to retrieve satellite images corresponding to the geographic locations of base stations in the two cities. For each satellite image, we employ the image-to-text model GPT-4o to generate 11,125 and 39,400 detailed textual descriptions for the base stations in the two cities, respectively.

To evaluate the performance of the proposed model, we compare the proposed model with several traditional spatiotemporal methods [3, 4] and up-to-date generative AI approaches [14, 15].

- **LSTM [3].** LSTM, as a specialized type of RNN, is designed to model sequential data while effectively capturing long-term dependencies. It addresses the vanishing gradient problem, enabling it to retain information across long time steps. LSTM employs a gating mechanism, consisting of input, forget, and output gates, to regulate the flow of information.
- **STGCN [4].** STGCN integrates GCN and gated CNN architectures to effectively capture spatiotemporal patterns in graph-structured data. It employs GCN to extract the graph's topological structure and gated CNN to analyze dynamic mobile traffic features.
- **WaveNet [14].** WaveNet, a generative model developed by DeepMind, is initially designed to generate raw audio data. Its core architecture, featuring causal convolutions and dilated convolutions, enables the model to effectively capture long-term dependencies.
- **TMAF [15].** TMAF is a generative model tailored to enhance probabilistic forecasting for multivariate time series. The model captures the dynamic characteristics of time series via the autoregressive structure and employs conditional normalization flow to model the intricate distribution of high-dimensional data, enabling it to more precisely capture the intricate relationships among variables and enhance forecasting performance.

In addition, we further analyze the effects of combining various modules in LEC-Diff on predicting base station traffic. Drop GEI entails removing the environmental feature extraction module from our proposed LEC-Diff. This modification allows us to evaluate the contribution of the text modality, provided by the LLM, in enriching environmental information. Drop SGF involves removing the spatiotemporal graph fusion module from our proposed LEC-Diff. This change helps us to verify the role of the GNN in enhancing spatial information within LEC-Diff.

Results and Discussion

All the experiments are conducted on Pytorch 2.0.1 based on Python 3.11.4 on the server equipped with Intel Xeon Silver 4310 with 2.1GHz and NVIDIA GeForce RTX 3090 with 24GB of memory. Based on this environment, we compare the prediction results of various models with the actual traffic data from two real-world datasets. We evaluate their performance using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Figure 4 presents not only a comparison of LEC-Diff with three baseline models, but also a validation of the effectiveness of our proposed module.

The proposed LEC-Diff model demonstrates outstanding performance on both datasets, outperforming all the compared algorithms. For instance, as shown in Fig. 4a, the proposed LEC-Diff model achieves over 5% reduction in MAE on the Shanghai dataset. Similarly, Fig. 4b highlights the superiority of our model on the Nanjing dataset, where it achieves more than a 2% reduction in RMSE. Compared with the LSTM baseline, temporal single-dimensional models are observed to perform poorly in the mobile traffic prediction task. This limitation stems from their inability to capture information across multiple dimensions. Analyzing spatial correlation based on spatial distribution and geographic patterns helps uncover the temporal patterns of change, thereby achieving higher performance in practical applications. STGCN is widely regarded as an effective method for spatiotemporal modeling, as it can effectively capture spatial features through the GNN structure. However, compared with STGCN, our proposed method has a performance advantage of about 5%. This is because the enhanced spatial information extracted by

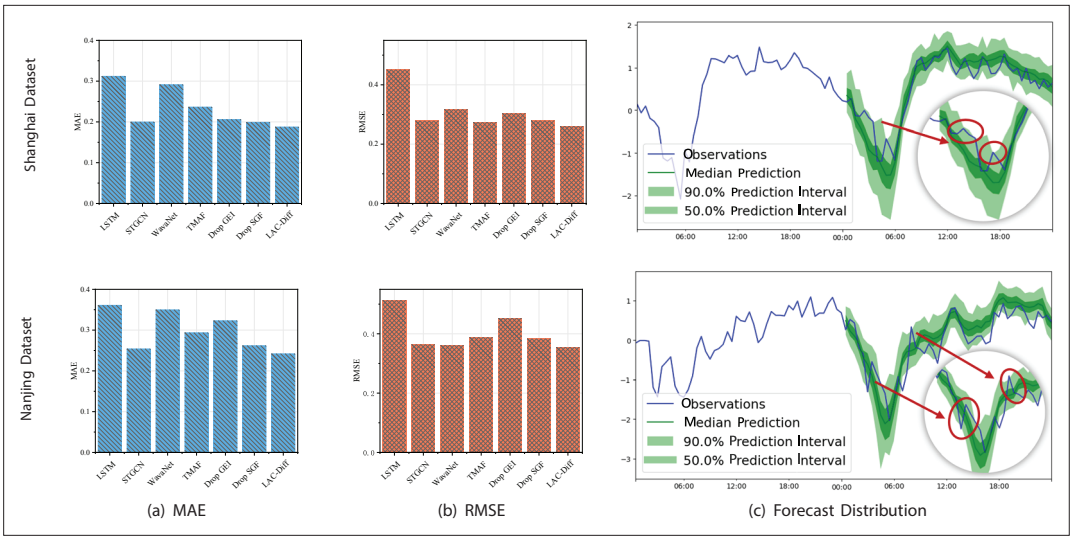

FIGURE 4. Prediction results. (a, b) Overall prediction performance of LEC-Diff in comparison with other algorithms on Shanghai and Nanjing datasets. The compared algorithms include not only baselines but also ablation studies; c) we select one base station from each of the two city datasets to analyze the impact of LEC-Diff on the predicted distribution. The uncertainty fluctuations in mobile traffic are highlighted with a red elliptical border.

In terms of the model capacity, the trained prediction model contains only 21.8M parameters, enabling flexible deployment on various lightweight nodes.

LLM incorporates a more comprehensive environmental context rather than relying exclusively on the distance matrix. Compared to the aforementioned deterministic prediction baseline methods, WaveNet and TMAF, which are generative AI-based prediction models, possess the capability to forecast uncertain probability distributions in non-stationary sequences. However, compared to LEC-Diff, these models exhibit notable limitations in environmental modeling. We employ the Continuous Ranked Probability Score (CRPS) to assess their performance in probability distribution forecasting. Using the Nanjing dataset as an example, compared to WaveNet (0.3025) and TMAF (0.3525), our method (0.2974) achieved an approximate performance improvement of 1.7%, demonstrating superior predictive ability for time series distribution. Overall, the proposed LEC-Diff model demonstrates substantial advantages over all existing spatiotemporal models.

To gain deeper insights into each component of our model, we conduct a series of ablation experiments. First, we remove the Environmental Feature Extraction Module, followed by the Spatiotemporal Graph Fusion module. The results of the ablation study, as shown by the "Drop GEI" results in Fig. 4, demonstrate that removing the Environmental Feature Extraction Module diminishes the model's environmental perception capability, impairs its understanding of environmental context, and significantly degrades prediction performance. Removing the Spatiotemporal Graph Fusion module (see "Drop SGF" results in Fig. 4) prevents the model from capturing spatial correlations between base stations through spatial modeling. Although it retains the ability to fully perceive environmental information, it fails to extract correlation patterns between base stations from the environmental information, leading to a notable reduction in prediction performance.

Moreover, to intuitively demonstrate the adaptability of LEC-Diff in addressing sudden or unexpected situations, Fig. 4c illustrates its capability to predict mobile traffic distribution. For this purpose, a base station from each city is selected for analysis as an example. As is shown in this figure, we mark some obvious traffic fluctuations during the observed time period. Notably, even when unexpected events lead to traffic fluctuations, our prediction method effectively captures such fluctuations within the 50% confidence interval. For general regular changes, our model successfully encapsulates regular changes within the broader 90% confidence interval. In summary, the proposed LEC-Diff model offers a robust and precise characterization of future changes in mobile cellular traffic.

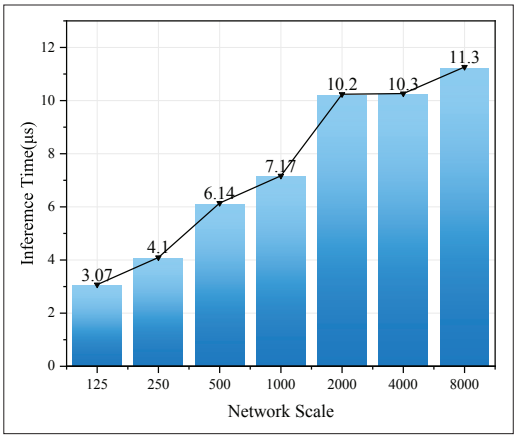

FIGURE 5. Computational Efficiency Analysis. Our proposed model has reduced inference time and enhanced adaptability in practical applications.

We conduct the efficiency analysis to evaluate the performance of LEC-Diff. In terms of the model capacity, the trained prediction model contains only 21.8M parameters, enabling flexible deployment on various lightweight nodes. Moreover, our model demonstrates high inference efficiency. As illustrated in Fig. 5, the inference time remains non-linear and does not rise proportionally as the network scale increases. Instead, the inference latency stabilizes progressively and consistently remains at a low level. For instance, in a network with 8,000 base station nodes, the model achieves an inference latency as low as 11.3 μs. Hence, the model demonstrates robust scalability,

supports flexible, lightweight deployment, and effectively meets real-time requirements in large-scale network applications.

Conclusion and Future Directions

In this article, we explore the potential of generative AI for addressing the problem of mobile cellular traffic prediction. To achieve that goal, we propose an LLM-enhanced conditional diffusion model named LEC-Diff, which adaptively generates environmentally enhanced information tailored to different cities and predicts future cellular traffic distribution by modeling spatiotemporal traffic patterns. Extensive experiments on large-scale mobile cellular traffic datasets demonstrate that LEC-Diff outperforms the baseline models.

Inspired by our work, the application of generative AI techniques in communications represents a promising research direction. In this context, the balance between its benefits and costs is a pivotal factor influencing the practical success of its application.

Benefits Analysis. Large generative AI models offer substantial advantages through extensive data training. LLMs utilize rich textual knowledge embedded within extensive training data to extract environmental features in urban environments, removing the need for manually constructing urban knowledge graphs. Moreover, the probabilistic modeling capabilities of generative AI significantly enhance its already exceptional potential. By modeling underlying probability distributions and dependencies within the data, generative AI can quantify the uncertainty of generated results and facilitate the evaluation of output reliability in time-series prediction tasks characterized by nonlinearity and non-stationarity.

Costs Analysis. Although generative AI offers considerable advantages and opportunities, it inevitably incurs higher short-term computational costs during training. Therefore, when deploying generative AI, it is essential to carefully assess performance gains against associated computational costs. In our study, environmental representation vectors produced by LLMs during training can be reused across various time instances and multiple requests during inference. This reuse strategy can effectively amortize overall computational costs while delivering sustainable long-term benefits.

Based on this, future efforts could focus on reducing the diffusion steps in the generation process, adopting a more lightweight neural network architecture, and leveraging hardware accelerations (such as GPU/TPU optimization) to speed up the inference process. In addition, exploring multi-task learning strategies, which involve simultaneous processing of related tasks (e.g., traffic prediction, congestion detection, and anomaly identification), could enhance computational efficiency through shared feature representations.

Acknowledgment

This research has been supported in part by the National Natural Science Foundation of China under Grant U24A20247 and Grant 62471277.

References

[1] B. Yu, X. Chen, and Y. Cai, "Age of Information for the Cellular Internet of Things: Challenges, key Techniques, and Future Trends," *IEEE Commun. Mag.*, vol. 60, no. 12, Dec. 2022, pp. 20–26.

[2] R. Blasco *et al.*, "Predictive Quality of Service in Cellular Networks: Challenges, Framework, and Application in Vehicular Communications," *IEEE Commun. Mag.*, vol. 61, no. 3, 2023, pp. 44–49.

[3] H. D. Trinh, L. Giupponi, and P. Dini, "Mobile Traffic Prediction from Raw Data Using LSTM Networks," *2018 IEEE 29th Annual Int'l. Symp. Personal, Indoor and Mobile Radio Communications (PIMRC)*, Bologna, Italy, Sept. 2018, pp. 1827–32.

[4] B. Yu, H. Yin, and Z. Zhu, "Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic forecasting," *Proc. 27th Int'l. Joint Conf. Artificial Intelligence*, IJCAI-18, Stockholm Sweden, July 2018, pp. 3634–40.

[5] J. Gong *et al.*, "Empowering Spatial Knowledge Graph for Mobile Traffic Prediction," *Proc. 31st ACM Int'l. Conf. Advances in Geographic Information Systems*, Hamburg, Germany, Dec. 2023, pp. 1–11.

[6] L. Bariah *et al.*, "Large Generative AI Models for Telecom: The Next Big Thing?," *IEEE Commun. Mag.*, vol. 62, no. 11, Nov. 2024, pp. 84–90.

[7] K. Rasul *et al.*, "Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting," *Proc. 38th Int'l. Conf. Machine Learning*, vol. 139, July 2021, pp. 8857–68.

[8] Y. Yan *et al.*, "Urbanclip: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web," *Proc. ACM on Web Conf. 2024*, New York, NY, USA, May 2024, pp. 4006–17.

[9] Y. Shen *et al.*, "Large Language Models Empowered Autonomous Edge AI for Connected Intelligence," *IEEE Commun. Mag.*, vol. 62, no. 10, Oct. 2024, pp. 140–46.

[10] J. Feng *et al.*, "Deeptp: An End-to-End Neural Network for Mobile Cellular Traffic Prediction," *IEEE Network*, vol. 32, no. 6, Nov./Dec. 2018, pp. 108–15.

[11] K. He *et al.*, "Graph Attention Spatialtemporal Network with Collaborative Global-Local Learning for Citywide Mobile Traffic Prediction," *IEEE Trans. Mobile Computing*, vol. 21, no. 4, Apr. 2022, pp. 1244–56.

[12] D. Salinas *et al.*, "Deepar: Probabilistic Forecasting with Autoregressive Recurrent Networks," *Int'l. J. Forecasting*, vol. 36, no. 3, July 2020, pp. 1181–91.

[13] A. Radford *et al.*, "Learning transferable Visual Models from Natural Language Supervision," *Proc. 38th Int'l. Conf. Machine Learning*, vol. 139, Jul. 2021, pp. 8748–63.

[14] A. van den Oord *et al.*, "Wavenet: A Generative Model for Raw Audio," *9th ISCA Wksp. Speech Synthesis*, Sunnyvale, CA, USA, Sep. 2016, p. 125.

[15] K. Rasul *et al.*, "Multivariate Probabilistic Time Series Forecasting via Conditioned Normalizing Flows," *Int'l. Conf. Learning Representations*, Vienna, Austria, May 2021.

Biographies

Zhu Xiao (zhxiao@hnu.edu.cn) received the M.S. and Ph.D. degrees in communication and information systems from Xidian University, Xi'an, China, in 2007 and 2009, respectively. From 2010 to 2012, he was a Research Fellow with the Department of Computer Science and Technology, University of Bedfordshire, Bedfordshire, U.K. He is currently a Full Professor with the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. His research interests include machine learning, mobile computing, and intelligent information processing.

Rui Wang (wray@hnu.edu.cn) received the B.S. degrees in Communication Engineering from Hunan University, Changsha, China, in 2015. He is currently pursuing the M.S. degree at the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. His research interests include signal processing and system security, mobile computing and task optimization.

Tong Li (t.li@connect.ust.hk) received the B.S. and M.S. degrees in Communication Engineering from Hunan University, Changsha, China, in 2014 and 2017, respectively, the Ph.D. in Computer Science and Engineering from the Hong Kong University of Science and Technology in 2021 and the Ph.D. in Computer Science from the University of Helsinki in 2022. He is currently a Full Professor at the College of Computer Science and Electronic Engineering, Hunan University, Changsha, China. His research interests include mobile computing, wireless network digital twins, network simulation and optimization, and data-driven networking.

Jing Bai (baijing@mail.xidian.edu.cn) received the B.S. degree in electronic and information engineering from Zhengzhou University, Zhengzhou, China, in 2004, and the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, Xi' an, China, in 2009. She is currently a Professor with Xidian University. Her research interests include Signal Modulation Recognition, machine learning, and intelligent information processing.

SHIYUAN ZHANG (zhangshi22@mails.tsinghua.edu.cn) received the B.S. degree in Automation from Harbin Institute of Technology, Harbin, China, in 2022, and the M.S. degree in Electronic Engineering from Tsinghua University, Beijing, China, in 2025. He is currently pursuing the Ph.D. degree in Electrical and Electronic Engineering at the University of Hong Kong. His research interests include mobile computing, wireless network digital twins, network simulation and optimization, and data-driven networking.

KEQIN LI [F] (lik@newpaltz.edu) received the Ph.D. degree in computer science from the University of Houston, Houston, Texas, USA, in 1990. He is currently a SUNY Distinguished Professor of Computer Science with the State University of New York, New Paltz, NY, USA. He has published over 620 journal articles, book chapters, and refereed conference papers. His current research interests include cloud computing, fog computing, mobileedge computing, energy-efficient computing and communication, embedded systems, cyber–physical systems, heterogeneous computing systems, big data computing, high-performance computing, CPU–GPU hybrid and cooperative computing, computer architectures and systems, computer networking, machine learning, and intelligent and soft computing. He received several best paper awards. He currently serves or has served on the editorial boards of *IEEE Transactions on Parallel and Distributed Systems*, *IEEE Transactions on Computers*, *IEEE Transactions on Cloud Computing*, *IEEE Transactions on Services Computing*, and *IEEE Transactions on Sustainable Computing*.

ZHU HAN [S'01, M'04, SM'09, F'14] (hanzhu22@gmail.com) received his Ph.D. degree in electrical and computer engineering from the University of Maryland, College Park. Currently, he is a professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department at the University of Houston, Texas.