

UoMo: A Universal Model of Mobile Traffic Forecasting for Wireless Network Optimization

Haoye Chai
Tsinghua University
Beijing, China

haoyechai@mail.tsinghua.edu.cn

Shiyuan Zhang
Tsinghua University
Beijing, China

zhangshi22@mails.tsinghua.edu.cn

Xiaoqian Qi
Tsinghua University
Beijing, China

qixiaoqian24@mails.tsinghua.edu.cn

Baohua Qiu
China Mobile
Beijing, China
qiubaohua@chinamobile.com

Yong Li
Tsinghua University
Beijing, China
liyong07@tsinghua.edu.cn

ABSTRACT

Mobile traffic forecasting allows operators to anticipate network dynamics and performance in advance, offering substantial potential for enhancing service quality and improving user experience. It involves multiple tasks, including long-term prediction, short-term prediction, and generation tasks that do not rely on historical data. By leveraging the different types of mobile network data generated from these tasks, operators can perform a variety of network optimizations and planning activities, such as base station (BS) deployment, resource allocation, energy optimization, *etc.* However, existing models are often designed for specific tasks and trained with specialized data, and there is a lack of universal models for traffic forecasting across different urban environments. In this paper, we propose a Universal model for Mobile traffic forecasting (UoMo), aiming to handle diverse forecasting tasks of short/long-term predictions and distribution generation across multiple cities to support network planning and optimization. UoMo combines diffusion models and transformers, where various spatio-temporal masks are proposed to enable UoMo to learn intrinsic features of different tasks, and a contrastive learning strategy is developed to capture the correlations between mobile traffic and urban contexts, thereby improving its transfer learning capability. Extensive evaluations on 9 real-world datasets demonstrate that UoMo outperforms current models in various forecasting tasks and zero/few-shot learning. It shows an average accuracy improvement of 27.85%, 25.80%, and 15.6% in long-term prediction, short-term prediction, and generation tasks, respectively, showcasing its strong forecasting capability. We deploy UoMo on China Mobile's JiuTian platform, using the predicted mobile data to optimize live networks. This includes BS deployment with a 25.3% increase in served users, BS sleep control reducing equipment depreciation by 40.1%, and resource block scheduling cutting transmission latency by 16.4%.

KEYWORDS

Mobile traffic forecasting; foundation model; diffusion model

1 INTRODUCTION

In recent years, foundation models [6, 42, 48] have made substantial strides in natural language processing and computer vision. These models are reshaping the AI ecosystem by harnessing their powerful data processing, generalization, and zero/few-shot learning

capabilities. An increasing number of specialized domains have developed foundational models tailored to their specific data and contextual demands, including healthcare, medicine, urban navigation, and beyond [4, 23, 66, 69]. Mobile networks encompass massive amounts of mobile traffic, user, and geographical data, providing inherent data support for building universal models. However, such dedicated models for mobile network domains have yet to be established. We hence aim to construct a universal model for mobile traffic forecasting, which can handle rich features of large-scale mobile data while retaining the generalization needed across multiple network applications [8, 55, 68].

Constructing such a universal model of mobile traffic forecasting is vital for mobile networks [14, 54, 76]. On the one hand, mobile traffic forecasting offers great potential for network planning and optimization. It enables operators to anticipate traffic dynamics, facilitating proactive perceptions of resource utilization and service quality, and allowing for the preemptive development of optimization strategies. On the other hand, mobile networks encompass a variety of optimization scenarios, including radio resource scheduling [31, 35], Base Station (BS) deployment [12, 13], and antenna configuration [28, 33], *etc.* These scenarios involve diverse objectives like throughput, coverage, and energy efficiency, which *impose distinct tasks on traffic forecasting*. For example, radio resource scheduling requires performing short-term traffic forecasting tasks, prioritizing traffic dynamics to improve user experience [9, 16], whereas BS deployment involves long-term traffic forecasting tasks, focusing on long-term traffic patterns within a region to align with network demands [52, 75]. For the planning and deployment of live wireless networks, it is essential to leverage the powerful data mining and robust generalization capabilities of universal models to simultaneously address a variety of optimization tasks.

Although numerous notable works have emerged in the area of mobile traffic forecasting [17, 24, 25, 39, 45, 59, 72], current methods typically employ one-to-one approaches: designing customized models by leveraging task-specific data [20, 27, 64, 76]. The complex, customized models deployed in live networks often require manual orchestration and scheduling, which will lead to the waste of computational and storage resources, increasing the overhead of model deployment. In addition, mobile traffic data is inherently heterogeneous of various collection granularity and scope. For example, Measure Report (MR) data primarily collects millisecond-level user traffic, while Performance Management (PM) data gathers cell-level

traffic statistics over 15-minute intervals [37, 53], leading to the absence of a unified representation akin to that found in natural language. Consequently, it is challenging to directly apply pre-trained models from the natural language/visual domains to mobile traffic data. Although some efforts have been made to reprogram mobile traffic data into a natural language format [18, 22], this approach heavily relies on the quality of manually crafted prompts, making it difficult to capture a universal representation of mobile traffic. Specifically, current mobile traffic forecasting models face two key limitations:

i) *Limited generalization*. Mobile traffic data is inherently shaped by the spatio-temporal dynamics of population distribution and communication demands. Due to variations in geographic environments, lifestyle habits, and urban layouts across different cities, mobile traffic can differ significantly [58, 67]. With relatively small parameters, current models struggle to capture the diverse spatio-temporal patterns inherent in large-scale data across multiple cities. Additionally, it is challenging to encapsulate the complex correlations between contextual factors and mobile traffic, resulting in poor transferability in multi-city scenarios.

ii) *Constrained task adaptability*. Mobile traffic forecasting is extensively applied across varying optimization scenarios. However, current models are often designed with specialized modules tailored to specific tasks. For instance, in short-term forecasting, models usually focus on capturing traffic fluctuations that employ autoregressive or event-driven methods. In contrast, long-term predictions emphasize the regular patterns of traffic and typically utilize time series decomposition techniques. These dedicated models increase design complexity and raise deployment costs when applied to diverse scenarios.

To tackle the limitations, we propose a Universal model for Mobile traffic forecasting (UoMo), which aims to learn universal features of mobile traffic data and to handle multiple tasks in mobile networks, thereby establishing a one-for-all forecasting model. **First**, inspired by Sora [5], UoMo adopts the transformer-based diffusion model as the backbone instead of the U-Net structure, to help the model understand the diverse features of massive mobile data. We propose a contrastive diffusion algorithm and adjust the variational lower bound by analyzing the cross-entropy between mobile traffic and contextual features. This helps the model better integrate environmental information, improving generalization and addressing the first limitation. **Second**, we adopt a task-oriented masking and self-supervised training paradigm, where we categorize traffic forecasting in mobile networks into three tasks: short-term prediction, long-term prediction, and generation. We design the corresponding masking strategies to enable the model to learn data features for various tasks and adapt to multiple tasks, thus addressing the second research challenge.

• To the best of our knowledge, it is the first universal model designed for mobile traffic forecasting. The proposed model enables various forecasting tasks in mobile networks across different urban environments via a unified framework, assisting network operators in achieving highly efficient network planning and optimization.

• We develop our universal model using a masked diffusion approach with spatio-temporal masking strategies tailored for diverse forecasting tasks, including short/long-term predictions and

distribution generation. To strengthen the correlation between contextual features and mobile traffic, we further propose a context-aware contrastive learning fine-tuning strategy, which can enhance forecasting and transfer learning capabilities.

• We conduct extensive evaluations with 9 real-world mobile traffic datasets. The results validate UoMo's superior generalization, multi-task capabilities, and robust few/zero-shot performance in unseen scenarios. We also deploy the UoMo model on China Mobile's Jiutian platform. Experiments on the live system data prove our model empowers multiple network optimization scenarios, including a 25.3% increase in served users for BS deployment, a 40.1% reduction in equipment depreciation for BS sleep control, and a 16.4% decrease in transmission latency for resource scheduling.

2 PROBLEM FORMULATION

Mobile Traffic refers to the volume of data transmitted over wireless channels between mobile devices and BSs over a period of time. We consider a discrete-time scenario $T = \{0, \dots, T\}$ with equal time intervals. For a single BS, the traffic variation over time T can be represented as $\{b_t\}_{t=0:T}$, where b_t denotes the aggregated traffic within the coverage area of BS at time t . To characterize the mobile traffic features across an urban region \mathcal{G} with contextual features $C_{\mathcal{G}}$, we define the geographical length and width of that region as H and V , respectively. The mobile traffic of region \mathcal{G} can then be defined as the sum of the aggregate traffic of all BSs located at \mathcal{G} : $S_{0:T} = \sum_b \{b_t\}_{t=0:T}$. Regarding diverse tasks in communication networks, such as base station deployment, user access control, and wireless resource allocation, different traffic forecasting tasks are often required. We categorize forecasting into 3 typical tasks:

• **Short-term prediction task** uses long historical data $0 \sim t$ to predict mobile traffic dynamics over a short future period, *i.e.*, $\{S_{0:t}, C_{\mathcal{G}}\} \rightarrow S_{t:T}$, where $t \gg (T - t)$. Based on the forecasts with fluctuations, operators can understand upcoming network demands and make optimizations such as resource allocation [34] and access control [30] to improve user experience in live networks.

• **Long-term prediction task** estimates future traffic patterns based on limited historical data, *i.e.*, $\{S_{0:t}, C_{\mathcal{G}}\} \rightarrow S_{t:T}$, where $t \ll T$. It explores inherent periodical patterns within the traffic. This forecasting enables operators to estimate and analyze network performance from a global perspective, thereby facilitating the formulation of network optimization planning strategies, such as cell dormancy [57] and network capacity expansion [78].

• **Generation task** focuses on identifying underlying network demand within a specific area without referring to historical data, *i.e.*, $\{C_{\mathcal{G}}\} \rightarrow S_{t=0:T}$. Traffic generation helps operators assess potential communication demands in new regions lacking historical data, allowing them to develop planning strategies such as BS deployment [33], network segmentation [61], and capacity planning [32], *etc.*

We aim to build a domain-universal model capable of achieving the above 3 forecast tasks. The problem can be defined as follows.

Problem Definition. Given arbitrary urban region \mathcal{G} , the goal is to use a model \mathcal{F} to forecast diverse mobile traffic sequence with short/long/generation tasks, conditioning on urban contextual factors $C_{\mathcal{G}}$, *i.e.*, $\mathcal{F}(S_{t=T_0:T_1} / S_{t=0:T_1}, C_{\mathcal{G}})$.

However, building such a universal model is not straightforward. Specifically, two key challenges arise: i). **What** strategies can be employed for the training process, ensuring that the model can handle the diverse forecasting tasks? ii). **How** to effectively integrate user dynamics and contextual features with mobile traffic?

3 DESIGN OF UOMO

3.1 Framework overview

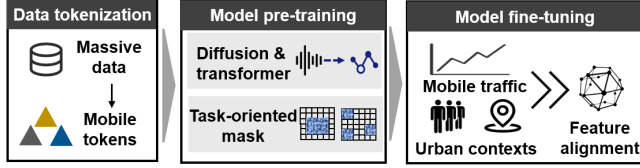


Figure 1: The flowchart of the UoMo framework

To tackle the challenges, we propose the UoMo framework incorporating three stages as illustrated in Figure 1.

i). *Data tokenization* reshapes mobile traffic data from various spatial-temporal spans across multiple cities into a unified mobile token for model training and capturing their diverse features.

ii). *Masked diffusion-based pre-training* tends to fully grasp the fundamental spatio-temporal features of mobile traffic across various forecasting tasks, where we design a diffusion-based backbone and task-oriented masks.

iii). *Urban context-aware fine-tuning* introduces a contrastive learning algorithm that integrates external factors closely associated with mobile traffic, including network user dynamics and urban POI distributions.

3.2 Masked diffusion-based pre-training

We propose a masked diffusion model with self-supervised training, where specific masks are tailored for the three forecasting tasks to enhance the model's understanding of various forecasting tasks and to capture the spatio-temporal correlations inherent in massive mobile data, as shown in Figure 2.

3.2.1 Mobile traffic data tokenization. We draw inspiration from NLP tokenization, where we decompose traffic data with varying sampling intervals and diverse spatial ranges into basic unit $h_0 \times v_0 \times t_0$. For traffic data S of length T within an urban region $H \times V$, the tokenization process breaks down S into multiple small mobile tokens X , which can be expressed as $S \in \mathbb{R}^{H \times V \times T} \rightarrow X \in \mathbb{R}^{(H' \times V' \times T') \times (h_0 \times v_0 \times t_0)}$, where $H' = H/h_0$, $V' = V/v_0$, and $T' = T/t_0$, the (h_0, t_0, v_0) of X represents the mobile token. Subsequently, we use an embedding layer $E_X(X)$ (e.g., pooling layer, convolutional layer, or fully connected layer) to map the mobile token with hidden features C , i.e., $E_X(X) \in \mathbb{R}^{(H' \times V' \times T') \times C}$.

3.2.2 Task-oriented mask. We develop 4 distinct masks: short-term, long-term, generation, and random masks, $m \in \mathbb{R}^{H' \times V' \times T'}$. The first three focus on specific forecasting tasks, while random masking explores spatio-temporal correlations to enhance generalization.

• **Short/Long-term masks.** The schemes mask the time dimension T' at a specific spatial location (h, v) to reconstruct the mobile traffic within the period $T' - t_0$, where $t_0 \in \{0, T'\}$. Depending on the ratio

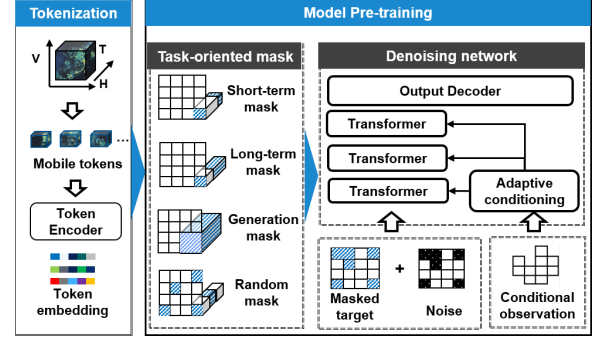


Figure 2: Masked diffusion-based pre-training network

of t_0 to T' , the schemes correspond to short/long-term predictions, respectively:

$$m_{h,v,t} = \{0, t_0 < t \leq T' \mid 1, 0 < t \leq t_0\}. \quad (1)$$

• **Generation mask.** It completely obscures the temporal dimension at a specific spatial location (h, v) , forcing the model to generate complete mobile traffic sequence within the region. Unlike prediction masks that rely on historical data, it captures spatio-temporal dependencies between the target region and its surroundings to generate underlying distributions:

$$m_{h,v,t} = \{0, 0 \leq t \leq T'\}. \quad (2)$$

• **Random mask.** It masks mobile traffic across both spatial and temporal dimensions, which aims to capture diverse correlations of mobile tokens, aiding the model in understanding the complex features of mobile data. Denote $\mathcal{R}(H', V', T')$ as randomly choosing items from H' , V' , and T' :

$$m_{h,v,t} = \{0, \mathcal{R}(H', V', T') \mid 1, \text{else}\}. \quad (3)$$

3.2.3 Self-supervised masked diffusion model. After completing the multi-task masking process, the original mobile token $E_X(X)$ is divided into two parts: the masked portion e requiring to be reconstructed, and the unmasked observation o :

$$e = E_X(X) \odot m, \quad o = E_X(X) \odot (1 - m), \quad (4)$$

where m corresponds to the four mask strategies, and \odot represents element-wise products. Subsequently, o is input as conditions into the denoising network, while e adds noise according to the forward process, which is given by

$$e_k = \sqrt{\hat{\alpha}_k} e + (1 - \hat{\alpha}_k) \epsilon, \quad \epsilon \sim N(0, 1). \quad (5)$$

Afterward, e_k is fed into the transformer-based denoising network for further feature extraction. To fully capture the dependencies between conditional observation and mobile traffic, we employ an adaptive conditioning method [40]. The method reshapes the scale and shift parameters of the layernorm of transformers by referring to the given conditions, which is proven to offer better effectiveness and computational efficiency [41]. It can be formulated as follows:

$$\alpha, \beta, \gamma = \mathcal{F}_\theta(o), \quad e_k \leftarrow e_k + \alpha \mathcal{A}_\theta(\beta e_k + \gamma), \quad (6)$$

where \mathcal{F}_θ and \mathcal{A}_θ denote linear layer and attention layer. α, β, γ are residual, scale, and shift parameters, respectively. The denoising network aims to fit the posterior distribution of the diffusion

process to predict the mean noise, ultimately reconstructing the final network traffic through an output decoder. Our objective thus emphasizes the reconstruction accuracy of the masked portion, which can be given as indicated in (5):

$$L_\theta = \min_{\theta} \mathbb{E}_{e \sim q(e)} \left\{ \|\epsilon - \epsilon_\theta(e, k|o)\|^2 \odot m \right\}. \quad (7)$$

3.3 Urban context-aware fine-tuning

Mobile traffic is not only a spatio-temporal sequence but also influenced by urban contexts. We thereby propose an urban context-aware fine-tuning scheme that integrates human dynamics and POIs into the UoMo, as shown in Fig 3.

3.3.1 Contextual data transformation. Mobile user refers to the number of users accessing the network, which can fully characterize the human dynamics in mobile networks. Similar to mobile traffic, it is inherently a spatio-temporal sequence and denoted as $U \in \mathbb{R}^{H \times V \times T}$. We apply the same processing method as traffic tokens where we perform tokenization on mobile users as $c^u \in \mathbb{R}^{(H' \times V' \times T') \times (h_0 \times v_0 \times t_0)}$, allowing this data to be directly input into the network for training.

POIs reflect the static distribution of urban layout that can be denoted as $P \in \mathbb{R}^{H \times V}$. Although the distribution of POIs is static, the impact of different categories of POIs on human behavior varies across different times, leading to corresponding variations in mobile traffic. For example, restaurant-type POIs typically show higher traffic during lunchtime and evening. In this regard, we design a dynamic POI transformation scheme. We first extract the intrinsic static features of POI distribution, which can be written as:

$$h_p^s = \sigma(W^s \cdot P + B^s), \quad (8)$$

where σ is the Sigmoid activation function, W^s and B^s are the weight and bias parameters of MLP network. We further utilize an MLP network $\tau(t)$ to project timestamp as temporal embeddings, then we fuse the static POI feature with the temporal indicators:

$$h_p^d = \sigma(W^l \cdot [h_p^s \oplus \tau(t)] + B^l), \quad (9)$$

where \oplus denotes vector concatenation, W^l and B^l are the learnable parameters. In this way, we can obtain spatio-temporal dynamic representations as $h_p^d \in \mathbb{R}^{H \times V \times T}$. The final features of POI can be calculated via the same tokenization method as mobile traffic data: $c^p \in \mathbb{R}^{(H' \times V' \times T') \times (h_0 \times v_0 \times t_0)}$. The ultimate contextual feature tokens can be denoted as $y = c^u + c^p$.

3.3.2 Context-aware alignment. To establish bridges between mobile traffic and contextual features, we propose a contrastive learning algorithm. We define positive samples as mobile traffic tokens and contextual feature tokens within the same spatio-temporal block, denoted as (e, y) ; while negative samples are defined as the two types of tokens from different spatio-temporal blocks, denoted (e', y) . Our goal is to maximize the mutual information between the traffic feature e and contextual feature y . According to previous research [49], a density ratio can be utilized while preserving the mutual information as $I(e, y) \propto \frac{p(e|y)}{p(e)}$, and the maximization problem is equivalent to minimizing the InfoNCE loss that yields:

$$\min_{e \in \mathbb{B}} \mathbb{E} \log \frac{-I(e, y)}{I(e, y) + \sum_{e'} I(e', y)} \leq \log(N) - I(e, y), \quad (10)$$

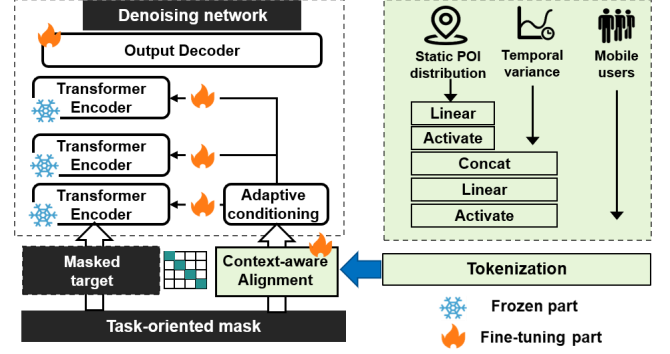


Figure 3: Context-aware fine-tuning process

where \mathbb{B} denotes the entire batch of samples.

We claim in Lemma 1 that training the diffusion model with positive and negative samples is equivalent to minimizing the InfoNCE loss in contrastive learning of Eq (10):

Lemma 1: By optimizing the Mean Squared Error (MSE) of positive and negative samples via Eq (11), we can achieve alignment between mobile traffic and contextual features.

$$L \approx \mathbb{E} \left\{ \left(\|\epsilon - \epsilon_\theta(e, k|y)\|^2 - \lambda \sum_{e'} \|\epsilon - \epsilon_\theta(e', k|y)\|^2 \right) \odot m \right\}. \quad (11)$$

The proof of the lemma is provided in the appendix A.4. During the fine-tuning process, we partially froze the main parameters of the pre-trained model, including the attention layer, linear layer, and MLP network, to preserve the model's ability to learn general spatio-temporal features of mobile traffic. We primarily update the parameters in the adaptive conditioning and output decoder layers. By partially updating these components, the time and computational cost of the fine-tuning process can be reduced.

4 EVALUATION

We perform evaluations on 9 real-world datasets to evaluate the UoMo with 13 baselines. The evaluations need to address the following 2 questions.

- **RQ1:** How does it perform in multi-task forecasting?
- **RQ2:** How does it perform in zero-shot and few-shot learning?

4.1 Evaluation setting from live mobile system

4.1.1 Datasets. Mobile traffic data. We collect mobile traffic data of live networks from 7 cities of varying scales in China, encompassing downlink traffic including 4G and 5G data. The time granularity of the data ranges from 15 minutes to 1 hour. Additionally, we utilize mobile traffic data from 2 other cities in China and Germany to validate UoMo's zero/few-shot capabilities.

Urban Contextual data. We collect mobile user data and mobile traffic data in each dataset. We crawl POI data from each city through public map services, including 15 categories related to living, entertainment, and other aspects.

4.1.2 Baselines. We select 13 baselines with 4 types. i). Statistical models. Historical moving average (**HA**) and **ARIMA** [60]. ii) NLP-based model. **Time-LLM** [22] and **Tempo** [7] describes time series features using natural language and uses these descriptions for forecasting. iii). Spatio-temporal based models. **CSDI** [47],

Table 1: Performance of short-term prediction task. Bold numbers denote the best results and underline numbers denote the second-best results. The proposed UoMo has the best prediction performance across 7 datasets.

Model	Beijing		Shanghai		Nanjing		Nanjing-4G		Nanchang		Nanchang-4G		Shandong	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
HA	0.1199	0.0697	0.1151	0.0576	0.0788	0.0353	0.0830	0.0371	0.0589	0.0266	0.0702	0.0339	0.1739	0.0578
ARIMA	0.2212	0.1333	0.1609	0.0819	0.1353	0.0622	0.1443	0.0668	0.1532	0.0666	0.1740	0.0789	0.1366	0.0531
SpectraGAN	0.2675	0.1228	0.2086	0.1226	0.2412	0.1186	0.2152	0.1151	0.2974	0.1467	0.1892	0.0935	0.2492	0.0814
keGAN	0.3307	0.2994	0.3456	0.2174	0.3586	0.3318	0.3579	0.3297	0.3123	0.1913	0.2521	0.2206	0.2662	0.2616
Adaptive	0.2779	0.2138	0.3007	0.2164	0.2606	0.1906	0.2219	0.1469	0.2305	0.1709	0.2572	0.1919	0.2688	0.1937
Open-Diff	0.1104	0.0899	0.1326	0.0981	0.1087	0.0823	0.1196	0.1005	0.1204	0.0801	0.1377	0.0906	0.1166	0.0799
Time-LLM	0.1511	0.1115	0.1388	0.0964	0.2351	0.1817	0.1754	0.1309	0.2039	0.1474	0.1770	0.1242	0.1571	0.0846
Tempo	0.1206	0.0873	0.0747	0.0455	0.0805	0.0625	0.0652	0.0498	0.0830	0.0638	0.0749	0.0550	0.0969	0.0763
CSDI	0.1752	0.1015	0.2060	0.1141	0.1722	0.0929	0.2299	0.1251	0.1797	0.0929	0.1587	0.0758	0.2131	0.0976
patchTST	0.1107	0.0686	0.1288	0.0872	0.0935	0.0616	0.0960	0.0631	0.1182	0.0635	0.1162	0.0638	0.1089	0.0703
TimeGPT	0.0598	0.0422	0.0866	0.0457	0.0646	0.0397	0.0657	0.0388	0.0502	0.0281	0.0576	0.0299	0.1219	0.0358
Lagllama	0.0501	0.0349	0.0853	0.0441	0.0529	0.0302	0.0530	0.0286	0.0505	0.0271	0.0625	0.0322	0.1272	0.0371
UniST	0.0332	0.0252	0.0658	0.0448	0.0623	0.0442	0.0608	0.0409	0.0433	0.0246	0.0852	0.0525	0.0766	0.0489
UoMo (our)	0.0284	0.0135	0.0588	0.0349	0.0442	0.0247	0.0439	0.0143	0.0360	0.0178	0.0408	0.0221	0.0609	0.0343
Improvement	14.45%	46.42%	10.63%	22.44%	16.44%	18.21%	16.60%	50.00%	16.85%	27.64%	29.16%	26.08%	20.49%	4.19%

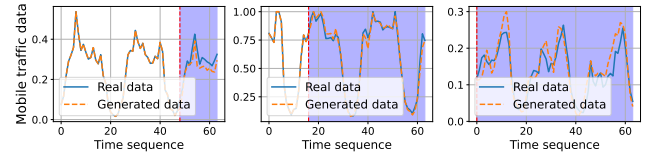
TimeGPT [15], Lagllama [43], PatchTST [38], and UniST [70] forecast mobile traffic as spatio-temporal series via autoregression, decomposition, and spatial convolution. iv). Dedicated models for mobile networks. SpectraGAN [59], KEGAN [21], ADAPTIVE [72], and Open-Diff [8] utilizes dedicated contextual data to generate mobile traffic. We describe the collected data, baselines, and metrics in the appendix A.1, A.2, and A.3.

4.2 Multitask forecasting (RQ1)

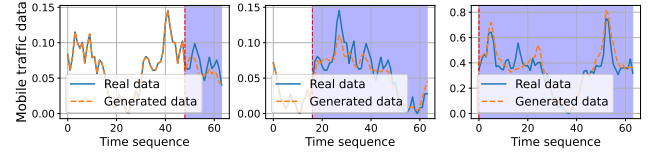
In our experiments, the temporal length is 64. For short-term prediction, the model forecasts 16 future points using the previous 48. For long-term prediction, the model forecasts 48 future points using the previous 16. For data generation, the model predicts all 64 points based on the current timestamp. We primarily use the 35M UoMo model (with 16 transformer layers and a hidden feature size of 256), and the other scaling evaluations are provided in appendix A.6.

- Short-term prediction. The results are presented in Table 1. Since sufficient historical data is available for reference, most baselines, leveraging their temporal feature extraction modules, effectively predict short-term changes. Overall, UoMo improved the average RMSE performance by 17.80% and the average MAE by 27.85% across 7 datasets, where UoMo can improve the RMSE metric by up to 29.1% (Shandong dataset) and the MAE metric by up to 50% (Nanjing-4G dataset), which exhibits stronger generalization capabilities compared to other models. Through the adaptive layer-norm module, the diffusion model iteratively integrates contextual features and leverages the transformer to capture long-term dependencies between mobile traffic and the environment. We believe this correlation can transfer across different cities, improving the model's generalization capability.

- Long-term prediction. The results are also shown in Table 2. For this task, the lack of sufficient historical observations often leads to performance degradation in some baselines. However, UoMo consistently achieves the best performance. It improves the average RMSE performance by 18.93% and the average MAE by 25.80% across 7 real-world datasets, with a maximum improvement of 31.36% in RMSE (Beijing dataset) and a maximum enhancement of 40.56% in MAE (Nanjing-4G dataset), which showcases its adaptability to various tasks.



(a) Forecasting performance of Beijing dataset.



(b) Forecasting performance of Nanchang dataset.

Figure 4: Visualization results. From left to right, it represents short-term/long-term/generation tasks.

- Mobile traffic generation. As shown in Table 3, the absence of historical observation of the generation task prevents some existing baselines from completing the task. Nevertheless, the UoMo still achieved strong generative results. It achieves a 6.95% improvement in the average JSD performance and a 15.6% improvement in the average MAE across all the datasets, with up to 21.19% improvement in the JSD metric (Shandong dataset), and a maximum enhancement of 29.04% in MAE (Nanjing-4G dataset). This is due to the contextual feature fusion module used during fine-tuning, which captures the correlation between contextual and mobile traffic features through contrastive learning. This allows the model to infer potential traffic distribution based on environmental changes, even without historical data.

4.2.1 Visualization. To intuitively demonstrate our model's universality for different tasks, we select two datasets as examples (Beijing and Nanchang) and plot the forecasting results in Figure 4. From left to right, it represents the tasks of short-term prediction → long-term prediction → traffic generation. The blue-shaded area indicates the model's predicted results, while the unshaded area represents historical observations. UoMo generates mobile traffic

Table 2: Performance of Long-term prediction task.

Model	Beijing		Shanghai		Nanjing		Nanjing-4G		Nanchang		Nanchang-4G		Shandong	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
HA	0.2945	0.1887	0.2214	0.1180	0.1808	0.0877	0.1914	0.0941	0.2011	0.0948	0.2285	0.1116	0.1331	0.0409
ARIMA	0.2023	0.1237	0.1560	0.0811	0.1269	0.0592	0.1340	0.0634	0.1533	0.0709	0.1751	0.0848	0.1224	0.0380
SpectraGAN	0.3880	0.3005	0.1962	0.1234	0.3621	0.2717	0.3212	0.2160	0.2432	0.1787	0.2352	0.1260	0.2438	0.0809
keGAN	0.3041	0.3716	0.2695	0.1837	0.2525	0.1809	0.2623	0.1917	0.2241	0.1770	0.2132	0.1837	0.1742	0.1315
Adaptive	0.2885	0.2234	0.3019	0.2197	0.2631	0.1876	0.2619	0.1907	0.1959	0.1419	0.2436	0.1752	0.1605	0.1144
Open-Diff	0.2801	0.1993	0.1562	0.1102	0.2042	0.1313	0.1809	0.1623	0.1822	0.1511	0.1907	0.1749	0.1496	0.1097
Time-LLM	0.1472	0.1099	0.1765	0.1124	0.2463	0.1843	0.2239	0.1602	0.2261	0.1751	0.2199	0.1621	0.1789	0.0868
Tempo	0.3514	0.2559	0.1518	0.0787	0.2896	0.1892	0.2780	0.1793	0.2380	0.1347	0.2365	0.1306	0.1020	0.0275
CSDI	0.3822	0.2836	0.2880	0.1856	0.4164	0.3034	0.3492	0.2520	0.3879	0.2913	0.3452	0.2347	0.2973	0.1705
patchTST	0.1512	0.1331	0.1627	0.0817	0.1521	0.1236	0.1644	0.0999	0.1430	0.0905	0.1789	0.1060	0.0985	0.0676
TimeGPT	0.3422	0.2433	0.1110	0.0766	0.2272	0.1391	0.2116	0.1345	0.1994	0.1001	0.1953	0.0919	0.0887	0.0253
Lagllama	0.2318	0.1879	0.1453	0.08742	0.0960	0.0683	0.1115	0.0959	0.1091	0.0788	0.1684	0.0889	0.1076	0.0439
UniST	0.1426	0.1014	0.1264	0.0803	0.1831	0.0845	0.1268	0.0869	0.1387	0.0835	0.1445	0.0763	0.0622	0.0337
UoMo (our)	0.1035	0.0696	0.0983	0.0679	0.0818	0.0532	0.0849	0.0570	0.0853	0.0576	0.1206	0.0563	0.0518	0.0197
Improvement	27.41%	31.36%	11.44%	11.35%	14.79%	22.10%	23.85%	40.56%	21.81%	26.90%	16.54%	26.21%	16.72%	22.13%

Table 3: Performance of generation task.

Model	Beijing		Shanghai		Nanjing		Nanjing-4G		Nanchang		Nanchang-4G		Shandong	
	JSD	MAE	JSD	MAE	JSD	MAE	JSD	MAE	JSD	MAE	JSD	MAE	JSD	MAE
SpectraGAN	0.3621	0.1584	0.3788	0.1284	0.3477	0.2888	0.3352	0.2494	0.2285	0.1288	0.3482	0.1762	0.3364	0.1794
keGAN	0.3435	0.2297	0.4909	0.2183	0.3071	0.1801	0.4862	0.1865	0.4032	0.1988	0.4792	0.2493	0.4007	0.2387
Adaptive	0.3044	0.2143	0.2751	0.1848	0.3040	0.1536	0.2587	0.2304	0.2730	0.2251	0.3201	0.1681	0.2806	0.1889
CSDI	0.3385	0.1431	0.2331	0.1025	0.4044	0.1844	0.3875	0.2516	0.3416	0.2163	0.2895	0.1734	0.2666	0.2170
Open-Diff	0.2155	0.1112	0.2299	0.1020	0.2114	0.1149	0.2275	0.1322	0.2296	0.1326	0.2624	0.1222	0.1977	0.1203
UoMo (our)	0.2013	0.0894	0.2259	0.1002	0.1971	0.0948	0.2164	0.0938	0.2226	0.1043	0.2494	0.1159	0.1558	0.0993
Improvement	6.58%	19.60%	1.74%	1.76%	6.76%	17.49%	4.87%	29.04%	2.58%	19.02%	4.95%	5.15%	21.19%	17.45%

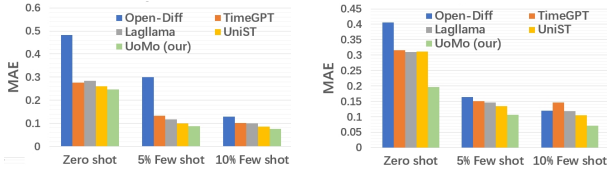
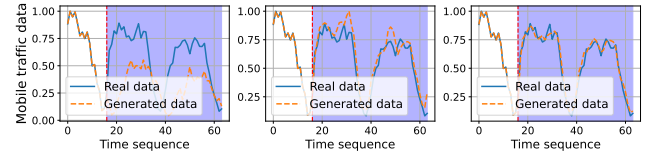


Figure 5: Zero/Few-shot across two distinct cities.

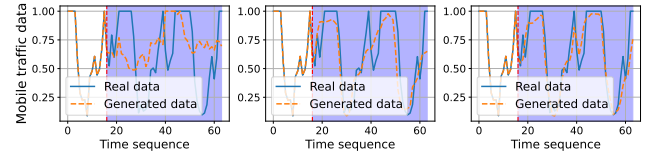
closely aligned with real values across all tasks, accurately predicting periodic trends and capturing fast dynamics, which shows that our UoMo model achieves forecasting across multiple cities and tasks, highlighting its generalization capability.

4.3 Zero/Few shot learning (RQ2)

To evaluate UoMo’s zero/ few-shot learning capabilities, we select two datasets that UoMo has not encountered during training: Hangzhou (China) and Munich (Germany). We choose 4 baselines that perform well in previous multitask forecasting: Open-Diff, TimeGPT, Lagllama, and UniST. The results are shown in Figure 5, where 5% few-shot and 10% few-shot represent the model training with a small amount of data (5% and 10%, respectively). It shows that UoMo exhibits good zero-shot performance, especially in the Munich dataset, where UoMo’s zero-shot performance even surpasses that of Open-Diff after small-scale training. After training with a small amount of data, all models show varying degrees of improvement. UoMo still demonstrates the best performance, indicating that UoMo can utilize the pre-trained model to quickly



(a) Long-term prediction results of zero/few-shots on Munica dataset.



(b) Long-term prediction results of zero/few-shots on Hangzhou dataset.

Figure 6: Visualization of zero/few-shot learning. From left to right: zero-shot → 5% few-shot → 10% few-shot.

capture general features within unseen mobile data. We also visualize the performance in zero/few-shot scenarios, as shown in Figure 6. We select a long-term forecasting task, with the results for zero-shot → 5% few-shot → 10% few-shot from left to right in the figure. It can be observed that UoMo can learn the general distribution characteristics of mobile traffic in the zero-shot phase, and after training with a small sample, the model realizes accurate traffic forecasting.

To give a comprehensive evaluation of the UoMo, We further conduct ablation and model scaling experiments in appendix A.5 and A.6.

5 DEPLOYMENT APPLICATIONS

Deployment. To validate UoMo's prediction effectiveness, we deploy the model on the Jiutian platform, an AI platform developed by China Mobile, featuring functions like scenario construction, network simulation, optimization strategy formulation, and performance evaluation. The Jiutian platform offers full-element network simulation capabilities, enabling efficient simulation of interactions between communication systems and user behavior. It also supports operators in developing customized algorithms and applications, deploying them into production environments, and performing product validation and testing with real network data. The platform has now been fully deployed within China Mobile, supporting network development across 31 provinces in China. We select predefined urban layout and human mobility data. UoMo is deployed in the mobile traffic module, with its predictions feeding into the optimization selection module. We focus on 3 optimization scenarios (highlighted in yellow in the figure) and evaluate the performance via network coverage, throughput, and energy consumption. The training of UoMo is conducted on 4 NVIDIA A100 GPUs (80GB each) using PyTorch 2.0.1. Table 4 summarizes the model's parameters and training/inference time per sample. We plot the framework of the Jiutian platform in Figure 12 in the appendix A.7.

Table 4: Model deployment efficiency on JiuTian.

Model	Layers	Hidden feature	Parameter scale	Training time	Inference time
Open-Diff	6	128	10M	0.09 min	0.021 min
UniST	12	512	30M	0.19 min	0.009 min
Lagllama	8	144	10M	0.21 min	0.011min
UoMo	12	128	5M	0.24 min	0.043 min
	16	256	35M	0.32 min	0.054 min
	20	768	200M	0.82 min	0.162 min

Optimization. Our optimization method on the Jiutian platform is shown in figure 7. First, we use UoMo or other mobile traffic forecasting algorithms to generate traffic data. The platform then uses this generated data to formulate and solve the problem of network optimization and planning. After obtaining the optimal network strategy, we input real live network traffic data into the platform to validate the optimization strategy and assess its performance.

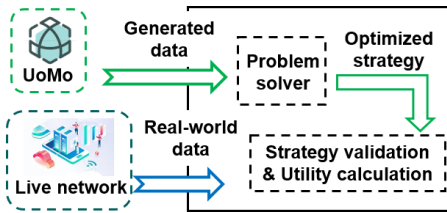


Figure 7: Optimization workflow via UoMo.

5.1 BS deployment

We examine a BS deployment planning at the grid level within a discrete-time $\{1, \dots, t\}$ framework. The target area consists of N grids, and M BSs need to be deployed across the grids. Each grid experiences varying mobile users at different times t , and each BS has

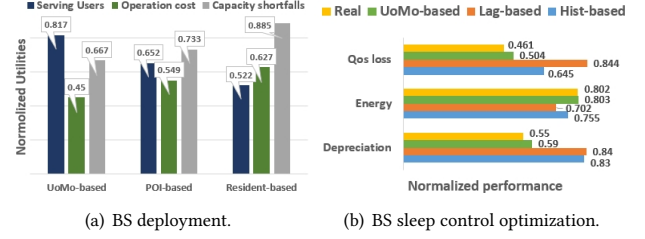


Figure 8: Optimization results with live system data.

a maximum capacity of C users at any given time. The deployment strategy's objective is to optimize the serving user count in each grid while minimizing operation cost (deploying fewer BSs in grids with fewer mobile users) and reducing capacity shortfalls (preventing situations where grids lack BSs and cannot accommodate excess users). Since there is no historical data on mobile user distribution for the target area, we rely on forecasting methods to estimate future mobile traffic demands. Three different estimation methods are utilized: the mobile traffic generation approach (UoMo-based) using the generation mask in (2), POI distribution (POI-based), and residential population distribution (Resident-based). Our focus is on addressing the BS deployment problem with a long-term perspective. The decision variables include x_i , which denotes the number of BSs deployed in grid i , and u_i^t , which represents the mobile traffic served by grid i at time t . Then the BS deployment problem yields

$$\begin{aligned} \max \quad & \sum_t \sum_i \left(y_i^t - \beta(\hat{U}_i^t - y_i^t)^+ - \alpha \sum_i x_i \right) \\ \text{s.t.} \quad & \sum_i x_i = M, \quad 0 \leq y_i^t \leq \min\{x_i C_0, \hat{F}_i^t\}, \end{aligned} \quad (12)$$

where $(y)^+ = \max(0, y)$ denote choosing the maximize value between y and 0, and \hat{U}_i^t is the mobile network demand that is estimated by UoMo, POI, and residential distribution. We employ the Pulp library to solve the optimization problem described above, and we can derive 3 distinct BS deployment strategies by referring to the three estimation methods. Subsequently, we test the optimization strategies using real user mobility trajectory data in Nanchang, streamed by the Jiutian platform, with the test results presented in Figure 8(a). The results demonstrate that the UoMo-based strategy maximizes the system's service revenue while significantly reducing both operational costs and capacity deficits compared to the other strategies, with an increase in the proportion of serving users by 25.3% (from 0.652 to 0.817), while reducing operational costs and capacity shortfall by 18.03% (from 0.549 to 0.45) and 9.00% (from 0.733 to 0.667), respectively. This success is attributed to the UoMo takes into account the time-varying nature of network usage across different regions. By accurately estimating the mobile traffic predictions with UoMo, we can effectively capture the dynamic patterns of human activities over time in a region. In contrast, the POI-based and resident-based strategies rely solely on static attributes to guide BS deployment. These methods are less effective in capturing the dynamic activity patterns of humans over time, and fall short when compared to those based on traffic.

5.2 BS sleep control

We consider the C-RAN scenario, where the BS achieves cell coverage by activating different numbers of RRUs [1]. BS sleep strategy involves controlling the RRU operational status (activated or sleep) based on the network load. We model the problem from the perspectives of service quality, depreciation cost, and energy consumption. We set N BSs, each of which has M cells to serve at time t . Define the traffic load that a single RRU can serve as c_0 , then for BS n , its Quality of Service (QoS) equals $Q(n) = \sum_t \sum_m \max(L_{m,t} - x_{m,t}c_0, 0)/L_{m,t}$, where $L_{m,t}$ is the actual cell load and $x_{m,t}$ is the activated RRUs. Moreover, frequent switching of RRUs can lead to a reduction in the lifespan of BSs, and the depreciation yields $W(n) = \sum_t \sum_m |x_{m,t} - x_{m,t-1}|$. BS energy consumption equals $E(n) = \sum_t \sum_m \mathcal{P}[\min(L_{m,t}, x_{m,t}c_0)]$ that is determined by the load at the RRU [73], where $\mathcal{P}[L] = \alpha L + \beta(L/c_0)$ is the energy consumption function. Therefore, the optimization objective yields:

$$\min \sum_n Y(x_{m,t}^{(n)} | L_{m,t}^{(n)}) = Q(n) + W(n) + E(n). \quad (13)$$

For the BS sleeping strategy, frequently adjusting RRUs over a short period is impractical. A more reasonable approach is to assess network dynamics over a longer period and develop a long-term adjustment strategy. Therefore, we leverage UoMo's long-term prediction capability to estimate $L_{m,t}$ via the long-term prediction mask in (1). The optimization results are shown in Figure 8(b), it can be seen that the UoMo-based strategy closely matches the real data-based strategy across different metrics, with a 21.9% improvement in QoS (from 0.645 to 0.504) and a maximum 40.7% reduction in BS depreciation (from 0.83 to 0.59) better than the other two. While energy consumption lags behind the comparison strategies, our algorithm aligns with that of the Real-based strategy, indicating that UoMo accurately predicts mobile traffic and effectively meets real user demands.

In addition to the above two deployment scenarios, we also develop a resource allocation application in appendix A.8.

6 RELATED WORK

Mobile traffic forecasting. It can be broadly categorized into two types: prediction and generation. Mobile traffic prediction involves estimating future values using historical data, while generation learns the underlying distribution of mobile traffic relying on external contextual information and samples new data from this distribution. Early forecasting used statistical approaches or simulation techniques [3, 11], but these methods typically struggled to capture complex traffic patterns. With the rise of machine learning, many studies used AI for mobile traffic forecasting. For mobile traffic prediction, LSTM models have been used to capture long-term dependencies in traffic patterns [10, 51, 77]. Some studies incorporated spatial attributes into traffic prediction, with Li *et al* [26] combining transformer and GCN to capture spatio-temporal correlations. Wu *et al* [56] combined GAN with GCN to capture spatial correlations across multiple cities. The MVSTGN model [65] divided urban spaces into multi-attribute graphs to capture mobile traffic features in latent space. For mobile traffic generation, early work used GANs to capture the overall distribution of mobile traffic [29, 44]. SpectraGAN [59] viewed cities as images, extracting POI and land use information via CNNs and incorporating it into traffic generation.

Sun *et al* [46] added user usage features to a GAN network, improving the accuracy of traffic generation. Hui *et al* [21] built a city knowledge graph incorporating extensive semantic features into traffic generation models. Open-Diff [8] used a diffusion model to generate grid-level traffic via open contextual data.

Universal and foundation models. These models show to excel in multitasking and zero/few-shot learning, and have been applied across various specialized domains. Yang *et al* [62] and Zhang *et al* [74] proposed foundation models aim to achieve various specialized tasks like investment, quantification, and urban navigation. Notably, many universal models for spatio-temporal forecasting have been proposed. Leveraging existing LLM, TEMPO [7] and Time-LLM [22] introduces a prompt mechanism in the pre-trained LLM for long-term forecasting by aligning features between mobile traffic and natural language tokens with a reprogramming approach. Some methods do not rely on existing language models but reconstruct spatio-temporal foundation models using transformer architectures. LagLLama [43] used lag indices to annotate multi-dimensional periodic features such as monthly, daily, and hourly periods. TimeGPT [15] replaced the Feedforward layer in the transformer with a CNN to enhance temporal correlations. UniST [70] achieved spatio-temporal prediction in urban contexts by employing a memory network.

While existing models offer advantages in forecasting mobile traffic, they are often designed for specific tasks like short-term prediction or generation. In real-world deployments, network optimization involves multiple forecasting tasks across cities, requiring frequent model switching and complex adaptations, which increases deployment costs. *A promising approach is to develop a universal model for mobile traffic forecasting.* The main challenge lies in mastering various forecasting tasks while integrating contextual features such as human dynamics and geographical traits. This integration is essential for building a robust model that captures the intrinsic correlations between the environment, users, and mobile traffic.

7 CONCLUSION

In this paper, we propose UoMo, a universal model with diffusion models for mobile traffic forecasting. To the best of our knowledge, it is the first universal model in mobile networks that simultaneously supports diverse forecasting tasks including short-term/long-term predictions and generation. By capturing the temporal, spatial, human dynamics, and geographical features related to mobile traffic, UoMo exhibits robust multi-task adaptability and zero/few-shot learning capability for diverse tasks across multiple cities, which exhibits good universality. Moreover, we identify the scaling properties of UoMo by examining the model performance with diverse parameter scales and data sizes. We deploy UoMo on the Jiutian platform, where it is used to optimize various aspects of network coverage, throughput, and energy consumption based on its accurate traffic forecasts. Currently, UoMo has been implemented in Nanning, Guangxi Province, China, for real-time traffic forecasting and data streaming. The model has the potential for large-scale deployment and can effectively assist operators in designing pricing strategies and network expansion, which improves both user experience and economic revenue.

REFERENCES

- [1] 2023. Artificial intelligence for reducing the carbon emissions of 5G networks in China. *Nature Sustainability* 6 (2023), 1522–1523. <https://api.semanticscholar.org/CorpusID:26097998>
- [2] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured Denoising Diffusion Models in Discrete State-Spaces. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 17981–17993. https://proceedings.neurips.cc/paper_files/paper/2021/file/958c530554f78bcd8e97125b70e6973d-Paper.pdf
- [3] Shruti Bothe, Haneya Naeem Qureshi, and Ali Imran. 2019. Which Statistical Distribution Best Characterizes Modern Cellular Traffic and What Factors Could Predict Its Spatiotemporal Variability? *IEEE Communications Letters* 23, 5 (2019), 810–813. <https://doi.org/10.1109/LCOMM.2019.2908370>
- [4] Vukasin Bozic, Abdelaziz Djelouah, Yang Zhang, Radu Timofte, Markus Gross, and Christopher Schroers. 2024. Versatile Vision Foundation Model for Image and Video Colorization. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) (*SIGGRAPH '24*). Association for Computing Machinery, New York, NY, USA, Article 94, 11 pages. <https://doi.org/10.1145/3641519.3657509>
- [5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators. (2024). <https://openai.com/research/video-generation-models-as-world-simulators>
- [6] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).
- [7] Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. 2024. TEMPO: Prompt-based Generative Pre-trained Transformer for Time Series Forecasting. *arXiv:2310.04948 [cs.LG]* <https://arxiv.org/abs/2310.04948>
- [8] Haoye Chai, Tao Jiang, and Li Yu. 2024. Diffusion Model-based Mobile Traffic Generation with Open Data for Network Planning and Optimization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Barcelona, Spain) (*KDD '24*). Association for Computing Machinery, New York, NY, USA, 4828–4838. <https://doi.org/10.1145/3637528.3671544>
- [9] Min Chen, Yiming Miao, Hamid Gharavi, Long Hu, and Iztok Humar. 2020. Intelligent Traffic Adaptive Resource Allocation for Edge Computing-Based 5G Networks. *IEEE Transactions on Cognitive Communications and Networking* 6, 2 (2020), 499–508. <https://doi.org/10.1109/TCCN.2019.2953061>
- [10] Anestis Dalgkitis, Malamati Louta, and George T. Karetos. 2018. Traffic forecasting in cellular networks using the LSTM RNN. In *Proceedings of the 22nd Pan-Hellenic Conference on Informatics* (Athens, Greece) (*PCI '18*). Association for Computing Machinery, New York, NY, USA, 28–33. <https://doi.org/10.1145/3291533.3291540>
- [11] Alvise De Biasio, Federico Chiariotti, Michele Polese, Andrea Zanella, and Michele Zorzi. 2019. A QUIC Implementation for Ns-3.
- [12] Miaomiao Dong, Taejoon Kim, Jingjin Wu, and Eric Wing-Ming Wong. 2020. Millimeter-Wave Base Station Deployment Using the Scenario Sampling Approach. *IEEE Transactions on Vehicular Technology* 69, 11 (2020), 14013–14018. <https://doi.org/10.1109/TVT.2020.3026216>
- [13] Ahmed Fahim and Yasser Gadallah. 2023. An Optimized LTE-Based Technique for Drone Base Station Dynamic 3D Placement and Resource Allocation in Delay-Sensitive M2M Networks. *IEEE Transactions on Mobile Computing* 22, 2 (2023), 732–743. <https://doi.org/10.1109/TMC.2021.3089329>
- [14] Zhiying Feng, Qiong Wu, and Xu Chen. 2024. Communication-efficient Multi-service Mobile Traffic Prediction by Leveraging Cross-service Correlations. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Barcelona, Spain) (*KDD '24*). Association for Computing Machinery, New York, NY, USA, 794–805. <https://doi.org/10.1145/3637528.3671730>
- [15] Azul Garza and Max Mergenthaler-Canseco. 2023. TimeGPT-1. *arXiv:2310.03589 [cs.LG]*
- [16] Abdennaceur Ghandri, Houssein Eddine Nouri, and Maher Ben Jemaa. 2024. Deep Learning for VBR Traffic Prediction-Based Proactive MBSFN Resource Allocation Approach. *IEEE Transactions on Network and Service Management* 21, 1 (2024), 463–476. <https://doi.org/10.1109/TNSM.2023.3311876>
- [17] Jiahui Gong, Yu Liu, Tong Li, Haoye Chai, Xing Wang, Junlan Feng, Chao Deng, Depeng Jin, and Yong Li. 2023. Empowering Spatial Knowledge Graph for Mobile Traffic Prediction. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems* (Hamburg, Germany) (*SIGSPATIAL '23*). Association for Computing Machinery, New York, NY, USA, Article 24, 11 pages. <https://doi.org/10.1145/3589132.3625569>
- [18] Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2023. Large Language Models Are Zero-Shot Time Series Forecasters. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 19622–19635. https://proceedings.neurips.cc/paper_files/paper/2023/file/3eb7ca52e8207697361b2c0fb3926511-Paper-Conference.pdf
- [19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) (*NIPS '20*). Curran Associates Inc., Red Hook, NY, USA, Article 574, 12 pages.
- [20] Yahui Hu, Yujiang Zhou, Junping Song, Luyang Xu, and Xu Zhou. 2023. Citywide Mobile Traffic Forecasting Using Spatial-Temporal Downsampling Transformer Neural Networks. *IEEE Transactions on Network and Service Management* 20, 1 (2023), 152–165. <https://doi.org/10.1109/TNSM.2022.3214483>
- [21] Shuodi Hui, Huandong Wang, Tong Li, Xinghao Yang, Xing Wang, Junlan Feng, Lin Zhu, Chao Deng, Pan Hui, Depeng Jin, and Yong Li. 2023. Large-scale Urban Cellular Traffic Generation via Knowledge-Enhanced GANs with Multi-Periodic Patterns. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Long Beach, CA, USA) (*KDD '23*). Association for Computing Machinery, New York, NY, USA, 4195–4206. <https://doi.org/10.1145/3580305.3599853>
- [22] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. *arXiv:2310.01728 [cs.LG]* <https://arxiv.org/abs/2310.01728>
- [23] Inwon Kang. 2024. Advancing Web Science through Foundation Model for Tabular Data. In *Companion Publication of the 16th ACM Web Science Conference* (Stuttgart, Germany) (*WebSci Companion '24*). Association for Computing Machinery, New York, NY, USA, 32–36. <https://doi.org/10.1145/3630744.3658614>
- [24] Migyeong Kang, Juho Jung, Minhan Cho, Daejin Choi, Eunil Park, Sangheon Pack, and Jinyoung Han. 2024. Poster: ISOML: Inter-Service Online Meta-Learning for Newly Emerging Network Traffic Prediction. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services* (Minato-ku, Tokyo, Japan) (*MOBISYS '24*). Association for Computing Machinery, New York, NY, USA, 718–719. <https://doi.org/10.1145/3643832.3661437>
- [25] Fatemeh Kavehmadavani, Van-Dinh Nguyen, Thang X. Vu, and Symeon Chatzinotas. 2024. Empowering Traffic Steering in 6G Open RAN with Deep Reinforcement Learning. *IEEE Transactions on Wireless Communications* (2024), 1–1. <https://doi.org/10.1109/TWC.2024.3396273>
- [26] He Li, Duo Jin, Xuejiao Li, Jianbin Huang, Xiaoke Ma, Jiangtao Cui, Deshuang Huang, Shaojie Qiao, and Jaesoo Yoo. 2023. DMGF-Net: An Efficient Dynamic Multi-Graph Fusion Network for Traffic Prediction. *ACM Trans. Knowl. Discov. Data* 17, 7, Article 97 (apr 2023), 19 pages. <https://doi.org/10.1145/3586164>
- [27] Tong Li, Shuodi Hui, Shiyuan Zhang, Huandong Wang, Yuheng Zhang, Pan Hui, Depeng Jin, and Yong Li. 2024. Mobile User Traffic Generation Via Multi-Scale Hierarchical GAN. *ACM Trans. Knowl. Discov. Data* 18, 8, Article 189 (jul 2024), 19 pages. <https://doi.org/10.1145/3664655>
- [28] Sheng-Hong Lin, Youyun Xu, Lei Wang, and Jin-Yuan Wang. 2022. Coverage Analysis and Chance-Constrained Optimization for HSR Communications With Carrier Aggregation. *IEEE Transactions on Intelligent Transportation Systems* 23, 9 (2022), 15107–15120. <https://doi.org/10.1109/ITITS.2021.3137030>
- [29] Zinan Lin, Alankar Jain, Chen Wang, Giulia Fanti, and Vyas Sekar. 2020. Using GANs for Sharing Networked Time Series Data: Challenges, Initial Promise, and Open Questions. In *Proceedings of the ACM Internet Measurement Conference* (Virtual Event, USA) (*IMC '20*). Association for Computing Machinery, New York, NY, USA, 464–483. <https://doi.org/10.1145/3419394.3423643>
- [30] Chunhui Liu, Minghui Xu, Hechuan Guo, Xiuzhen Cheng, Yin hao Xiao, Dongxiao Yu, Bei Gong, Arkady Yerukhimovich, Shengling Wang, and Weifeng Lyu. 2024. TBAC: A Token-Based Accountable Access Control Scheme for the Internet of Things. *IEEE Transactions on Mobile Computing* 23, 5 (2024), 6133–6148. <https://doi.org/10.1109/TMC.2023.3316622>
- [31] Sudha Lohani, Roya Arab Loodaricheh, Ekram Hossain, and Vijay K. Bhargava. 2016. On Multiuser Resource Allocation in Relay-Based Wireless-Powered Uplink Cellular Networks. *IEEE Transactions on Wireless Communications* 15, 3 (2016), 1851–1865. <https://doi.org/10.1109/TWC.2015.2496943>
- [32] Wencan Mao, Ozgur Umur Akgul, Byungjin Cho, Yu Xiao, and Antti Ylä-Jääski. 2023. On-Demand Vehicular Fog Computing for Beyond 5G Networks. *IEEE Transactions on Vehicular Technology* 72, 12 (2023), 15237–15253. <https://doi.org/10.1109/TVT.2023.3289862>
- [33] Weidong Mei and Rui Zhang. 2023. Joint Base Station and IRS Deployment for Enhancing Network Coverage: A Graph-Based Modeling and Optimization Approach. *IEEE Transactions on Wireless Communications* 22, 11 (2023), 8200–8213. <https://doi.org/10.1109/TWC.2023.3260805>
- [34] Erfan Meskar and Ben Liang. 2023. Fair Multi-Resource Allocation in Heterogeneous Servers With an External Resource Type. *IEEE/ACM Transactions on Networking* 31, 3 (2023), 1244–1262. <https://doi.org/10.1109/TNET.2022.3213426>
- [35] Jie Miao, Zheng Hu, Kun Yang, Canru Wang, and Hui Tian. 2012. Joint Power and Bandwidth Allocation Algorithm with QoS Support in Heterogeneous Wireless Networks. *IEEE Communications Letters* 16, 4 (2012), 479–481. <https://doi.org/10.1109/LCOMM.2012.030512.112304>
- [36] Soumik Mukhopadhyay, M. Gwilliam, Vatsal Agarwal, Namitha Padmanabhan, Archana Ashok Swaminathan, Srinidhi Hegde, Tianyi Zhou, and Abhinav Shrivastava. 2023. Diffusion Models Beat GANs on Image Classification. *ArXiv abs/2307.08702* (2023). <https://api.semanticscholar.org/CorpusID:259937835>

- [37] Dalia Naboulsi, Marco Fiore, Stephane Ribot, and Razvan Stanica. 2016. Large-Scale Mobile Traffic Analysis: A Survey. *IEEE Communications Surveys & Tutorials* 18, 1 (2016), 124–161. <https://doi.org/10.1109/COMST.2015.2491361>
- [38] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2023. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. *arXiv:2211.14730 [cs.LG]* <https://arxiv.org/abs/2211.14730>
- [39] Chandrasen Pandey, Vaibhav Tiwari, Joel J. P. C. Rodrigues, and Diptendu Sinha Roy. 2024. 5GT-GAN-NET: Internet Traffic Data Forecasting With Supervised Loss Based Synthetic Data Over 5 G. *IEEE Transactions on Mobile Computing* (2024), 1–12. <https://doi.org/10.1109/TMC.2024.3364655>
- [40] William Peebles and Saining Xie. 2023. Scalable Diffusion Models with Transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 4172–4182. <https://doi.org/10.1109/ICCV51070.2023.00387>
- [41] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence* (New Orleans, Louisiana, USA) (AAAI'18/AAAI'18/EAAI'18). AAAI Press, Article 483, 10 pages.
- [42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [43] Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadihir Hassen, Marin Biloš, Sahil Garg, Anderson Schneider, Nicolas Chapados, Alexandre Drouin, Valentina Zantedeschi, Yuriy Nevmyvaka, and Irina Rish. 2024. Lag-Llama: Towards Foundation Models for Probabilistic Time Series Forecasting. *arXiv:2310.08278 [cs.LG]* <https://arxiv.org/abs/2310.08278>
- [44] Markus Ring, Daniel Schlör, Dieter Landes, and Andreas Hotho. 2019. Flow-based network traffic generation using generative adversarial networks. *Computers & Security* 82 (2019), 156–172.
- [45] Zhi Sheng, Yuan Yuan, Jingtao Ding, and Yong Li. 2025. Unveiling the Power of Noise Priors: Enhancing Diffusion Models for Mobile Traffic Prediction. *arXiv:2501.13794 [cs.LG]* <https://arxiv.org/abs/2501.13794>
- [46] Chuanhao Sun, Kai Xu, Marco Fiore, Mahesh K. Marina, Yue Wang, and Cezary Ziemlicki. 2022. AppShot: A Conditional Deep Generative Model for Synthesizing Service-Level Mobile Traffic Snapshots at City Scale. *IEEE Transactions on Network and Service Management* 19, 4 (2022), 4136–4150. <https://doi.org/10.1109/TNSM.2022.3199458>
- [47] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. 2021. CSDI: Conditional Score-based Diffusion Models for Probabilistic Time Series Imputation. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 24804–24816. https://proceedings.neurips.cc/paper_files/paper/2021/file/cfe8504bda37b575c70e1a8276f3486-Paper.pdf
- [48] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [49] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748 [cs.LG]* <https://arxiv.org/abs/1807.03748>
- [50] Heng Wang, Yixuan Bai, and Xin Xie. 2024. Deep Reinforcement Learning Based Resource Allocation in Delay-Tolerance-Aware 5G Industrial IoT Systems. *IEEE Transactions on Communications* 72, 1 (2024), 209–221. <https://doi.org/10.1109/TCOMM.2023.3322736>
- [51] Xiaojie Wang, Laisen Nie, Zhaolong Ning, Lei Guo, Guoyin Wang, Xinbo Gao, and Neeraj Kumar. 2022. Deep Learning-Based Network Traffic Prediction for Secure Backbone Networks in Internet of Vehicles. *ACM Trans. Internet Technol.* 22, 4, Article 87 (nov 2022), 20 pages. <https://doi.org/10.1145/3433548>
- [52] Zi Wang, Jia Hu, Geyong Min, Zhiwei Zhao, and Jin Wang. 2021. Data-Augmentation-Based Cellular Traffic Prediction in Edge-Computing-Enabled Smart City. *IEEE Transactions on Industrial Informatics* 17, 6 (2021), 4179–4187. <https://doi.org/10.1109/TII.2020.3009159>
- [53] Zhen Wang, Guofu Wei, Yaling Zhan, and Yanhuan Sun. 2017. Big Data in Telecommunication Operators: Data, Platform and Practices. *Journal of Communications and Information Networks* 2, 3 (2017), 78–91. <https://doi.org/10.1007/s41650-017-0010-1>
- [54] Qiong Wu, Xu Chen, Zhi Zhou, Liang Chen, and Junshan Zhang. 2021. Deep Reinforcement Learning With Spatio-Temporal Traffic Forecasting for Data-Driven Base Station Sleep Control. *IEEE/ACM Trans. Netw.* 29, 2 (apr 2021), 935–948. <https://doi.org/10.1109/TNET.2021.3053771>
- [55] Qiong Wu, Kaiwen He, Xu Chen, Shuai Yu, and Junshan Zhang. 2021. Deep Transfer Learning Across Cities for Mobile Traffic Prediction. *IEEE/ACM Trans. Netw.* 30, 3 (dec 2021), 1255–1267. <https://doi.org/10.1109/TNET.2021.3136707>
- [56] Qiong Wu, Kaiwen He, Xu Chen, Shuai Yu, and Junshan Zhang. 2022. Deep Transfer Learning Across Cities for Mobile Traffic Prediction. *IEEE/ACM Transactions on Networking* 30, 3 (2022), 1255–1267. <https://doi.org/10.1109/TNET.2021.3136707>
- [57] Yi Wu, Pengcheng Jin, Yihan Zhang, Tiejiong Cai, and Ying Ji. 2023. Coverage Quality Optimization Strategy for Static Heterogeneous Wireless Sensor Networks. In *2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIE)*. 1–6. <https://doi.org/10.1109/AIKIE60097.2023.10390348>
- [58] Fengli Xu, Yong Li, Huandong Wang, Pengyu Zhang, and Depeng Jin. 2017. Understanding Mobile Traffic Patterns of Large Scale Cellular Towers in Urban Environment. *IEEE/ACM Trans. Netw.* 25, 2 (apr 2017), 1147–1161. <https://doi.org/10.1109/TNET.2016.2623950>
- [59] Kai Xu, Rajkarn Singh, Marco Fiore, Mahesh K. Marina, Hakan Bilen, Muhammad Usama, Howard Benn, and Cezary Ziemlicki. 2021. SpectraGAN: spectrum based generation of city scale spatiotemporal mobile network traffic data. In *Proceedings of the 17th International Conference on Emerging Networking EXperiments and Technologies* (Virtual Event, Germany) (CoNEXT '21). Association for Computing Machinery, New York, NY, USA, 243–258. <https://doi.org/10.1145/3485983.3494844>
- [60] Xiongqiao Xu, Xin Wang, Elkin Cruz-Camacho, Christopher D. Carothers, Kevin A. Brown, Robert B. Ross, Zhiling Lan, and Kai Shu. 2023. Machine Learning for Interconnect Network Traffic Forecasting: Investigation and Exploitation. In *Proceedings of the 2023 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation* (Orlando, FL, USA) (SIGSIM-PADS '23). Association for Computing Machinery, New York, NY, USA, 133–137. <https://doi.org/10.1145/3573900.3591123>
- [61] Awaneesh Kumar Yadav. 2024. SKAP-NS: A Symmetric Key-Based Authentication Protocol for 5G Network Slicing. *IEEE Transactions on Industrial Informatics* 20, 11 (2024), 13363–13372. <https://doi.org/10.1109/TII.2024.3435535>
- [62] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-Source Financial Large Language Models. *FinLLM Symposium at IJCAI 2023* (2023).
- [63] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Comput. Surv.* 56, 4, Article 105 (nov 2023), 39 pages. <https://doi.org/10.1145/3626235>
- [64] Shun-Ren Yang, Yu-Ju Su, Yao-Yuan Chang, and Hui-Nien Hung. 2019. Short-Term Traffic Prediction for Edge Computing-Enhanced Autonomous and Connected Cars. *IEEE Transactions on Vehicular Technology* 68, 4 (2019), 3140–3153. <https://doi.org/10.1109/TVT.2019.2899125>
- [65] Yang Yao, Bo Gu, Zhou Su, and Mohsen Guizani. 2023. MVSTGN: A Multi-View Spatial-Temporal Graph Network for Cellular Traffic Prediction. *IEEE Transactions on Mobile Computing* 22, 5 (2023), 2837–2849. <https://doi.org/10.1109/TMC.2021.3129796>
- [66] Chin-Chia Michael Yeh, Xin Dai, Huiyuan Chen, Yan Zheng, Yujie Fan, Audrey Der, Vivian Lai, Zhongfang Zhuang, Junpeng Wang, Liang Wang, and Wei Zhang. 2023. Toward a Foundation Model for Time Series Data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (Birmingham, United Kingdom) (CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 4400–4404. <https://doi.org/10.1145/3583780.3615155>
- [67] Jieli Yin, Yali Fan, Tong Xia, Yong Li, Xiang Chen, Zhi Zhou, and Xu Chen. 2020. Mobile App Usage Patterns Aware Smart Data Pricing. *IEEE Journal on Selected Areas in Communications* 38, 4 (2020), 645–654. <https://doi.org/10.1109/JSAC.2020.2971896>
- [68] Yucheng Yin, Zinan Lin, Minhao Jin, Giulia Fanti, and Vyas Sekar. 2022. Practical GAN-based synthetic IP header trace generation using NetShare. In *Proceedings of the ACM SIGCOMM 2022 Conference* (Amsterdam, Netherlands) (SIGCOMM '22). Association for Computing Machinery, New York, NY, USA, 458–472. <https://doi.org/10.1145/3544216.3544251>
- [69] Jinliang Yuan, Chen Yang, Dongqi Cai, Shihe Wang, Xin Yuan, Zeling Zhang, Xiang Li, Dingge Zhang, Hanzi Mei, Xianqing Jia, Shangguang Wang, and Mengwei Xu. 2024. Mobile Foundation Model as Firmware. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking* (Washington D.C., DC, USA) (ACM MobiCom '24). Association for Computing Machinery, New York, NY, USA, 279–295. <https://doi.org/10.1145/3636534.3649361>
- [70] 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* (Barcelona, Spain) (KDD '24). Association for Computing Machinery, New York, NY, USA, 4095–4106. <https://doi.org/10.1145/3637528.3671662>
- [71] Saniya Zafar, Sobia Jangsher, and Ararat Al-Dweik. 2022. Resource Allocation Using Deep Learning in Mobile Small Cell Networks. *IEEE Transactions on Green Communications and Networking* 6, 3 (2022), 1903–1915. <https://doi.org/10.1109/TGCN.2022.3146487>
- [72] Shiyuan Zhang, Tong Li, Shuodi Hui, Guangyu Li, Yanping Liang, Li Yu, Depeng Jin, and Yong Li. 2023. Deep Transfer Learning for City-scale Cellular Traffic Generation through Urban Knowledge Graph. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Long Beach, CA,

- USA) (*KDD '23*). Association for Computing Machinery, New York, NY, USA, 4842–4851. <https://doi.org/10.1145/3580305.3599801>
- [73] Sheng Zhang, Shenglin Zhao, Mingxuan Yuan, Jia Zeng, Jianguo Yao, Michael R. Lyu, and Irwin King. 2017. Traffic Prediction Based Power Saving in Cellular Networks: A Machine Learning Method. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (Redondo Beach, CA, USA) (*SIGSPATIAL '17*). Association for Computing Machinery, New York, NY, USA, Article 29, 10 pages. <https://doi.org/10.1145/3139958.3140053>
- [74] Weijiao Zhang, Jindong Han, Zhao Xu, Hang Ni, Hao Liu, and Hui Xiong. 2024. Towards Urban General Intelligence: A Review and Outlook of Urban Foundation Models. *ArXiv abs/2402.01749* (2024). <https://api.semanticscholar.org/CorpusID:267412420>
- [75] Zhongliang Zhao, Lucas Pacheco, Hugo Santos, Minghui Liu, Antonio Di Maio, Denis Rosário, Eduardo Cerqueira, Torsten Braun, and Xianbin Cao. 2021. Predictive UAV Base Station Deployment and Service Offloading With Distributed Edge Learning. *IEEE Transactions on Network and Service Management* 18, 4 (2021), 3955–3972. <https://doi.org/10.1109/TNSM.2021.3123216>
- [76] Shenghan Zhou, Chaofan Wei, Chaoferi Song, Xing Pan, Wenbing Chang, and Linchao Yang. 2023. Short-Term Traffic Flow Prediction of the Smart City Using 5G Internet of Vehicles Based on Edge Computing. *IEEE Transactions on Intelligent Transportation Systems* 24, 2 (2023), 2229–2238. <https://doi.org/10.1109/TITS.2022.3147845>
- [77] Feiyue Zhu, Lixiang Liu, and Teng Lin. 2020. An LSTM-Based Traffic Prediction Algorithm with Attention Mechanism for Satellite Network. In *Proceedings of the 2020 3rd International Conference on Artificial Intelligence and Pattern Recognition* (Xiamen, China) (*AIPR '20*). Association for Computing Machinery, New York, NY, USA, 205–209. <https://doi.org/10.1145/3430199.3430208>
- [78] Xiaomeng Zhu, Yi Li, Yuting Zheng, Rui Xia, Lexi Xu, Bei Li, Zixiang Di, Lu Zhi, and Xinzhou Cheng. 2022. Research on 5G Network Capacity and Expansion. In *2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*. 1473–1478. <https://doi.org/10.1109/TrustCom56396.2022.00209>

A APPENDIX

A.1 Description of dataset

We collected mobile traffic and user data from 9 different cities, covering over 30,000 base stations, with time granularities ranging from 15 minutes to 1 hour. POI data was crawled from the map service (available at <https://lbs.amap.com/>) and includes 15 categories, such as lifestyle, entertainment, work, and dining, as shown in Table 5.

Table 5: Description of datasets

Dataset	Usage	Data description	Mobile traffic	Mobile users	Time granularity
Beijing	Model training	5G data, October, 2021, 4000+ BSs	✓	✓	30 min
Shanghai		4G data, August, 2014, 5000+ BSs	✓	✓	30 min
Nanjing		5G data, February to March, 2021, 6000+ BSs	✓	✓	15 min
Nanjing-4G		4G data, February to March, 2021, 6000+ BSs	✓	✓	15 min
Nanchang		5G data, July, 2023, 5000+ BSs	✓	✓	30 min
Nanchang-4G		4G data, July, 2023, 7000+ BSs	✓	✓	30 min
Shandong	Zero/Few shot tests	5G data, February, 2024, 1000+ BSs	✓	✓	1 hour
Hangzhou		5G data, July, 2023, 1000+ BSs	✓	✓	1 hour
Munich		4G data, 2022, 2500+ grid-data	✓	—	1 hour
POI	Shopping, Enterprise, Restaurant, Local Living, Transportation, Public Health, Automobile, Physical facilities, Accommodation, Finance, Government organs, Education, Business, Public facilities, scenic spot.				

A.2 Description of baselines

We select a total of 13 baselines across 4 major types.

- Statistical models. Historical moving average method (**HA**) and **ARIMA** method that integrate autoregression with average moving [60].

- Natural language-based model. **Time-LLM** [22] describes time series features using natural language and uses these descriptions as prompts into a natural language pre-trained model (LLAMA-7B) for forecasting. **Tempo** [7] designs temporal prompts with trend and seasonal features for pre-trained models (GPT-2) to predict time series.

- Spatio-temporal-based models. They primarily forecast mobile traffic as spatio-temporal series via autoregression, decomposition, and spatial convolution. **TimeGPT** [15] replaces the Feedforward layer in the transformer with a CNN network and is trained on vast spatio-temporal data. **Lagllama** [43] uses a set of lag indices to capture different periodic correlations in the time series. **CSDI** [47] is a conditional diffusion model that uses a masking method for time-series data forecasting and imputation. **PatchTST** [38] decomposes time series into multiple segments and uses transformers for feature extraction. **UniST** [70] segments spatio-temporal data and fine-tunes the model using features like geographical proximity and temporal correlations.

- Dedicated models for mobile traffic forecasting. In addition to the general spatio-temporal forecasting models, we select 5 models dedicated to mobile networks. **SpectraGAN** [59] converts mobile traffic generation into an image generation problem and utilizes a CNN-based GAN network for traffic forecasting. **KEGAN** [21] is a hierarchical GAN that utilizes a self-constructed Urban Knowledge Graph (UKG) to explicitly incorporate urban features during the forecasting process. **ADAPTIVE** [72] leverages the UKG and a BS aligning scheme to transfer mobile traffic knowledge from one city to another. **Open-Diff** [8] utilizes open contextual data like satellite

images, residential count, and POI distribution to generate mobile traffic data.

A.3 Description of metrics

We choose 3 metrics to investigate the performance of the algorithms. **MAE** evaluates the similarity of generated values S and real values \hat{S} , which can be expressed as

$$m1 = Avg(|S - \hat{S}|). \quad (14)$$

RMSE measures the average magnitude of the errors between predicted values and actual observed values, which can be expressed as

$$m2 = \sqrt{Avg(|S - \hat{S}|^2)}. \quad (15)$$

Jensen-Shannon divergence (JSD) is a commonly used metric to measure the similarity between two distributions, which can be calculated by the mean of the Kullback-Leibler (KL) divergence:

$$m3 = \sqrt{\frac{KL(Pr(S||\hat{S})) + KL(Pr(\hat{S}||S))}{2}}. \quad (16)$$

A.4 Proof of Lemma 1

The underlying idea of DDPM is to use two Markov chains to characterize the transition from original data to noise data [36]. The forward chain gradually adds Gaussian noise $\epsilon \sim N(0, 1)$ to the original data that can be expressed as $q(x_k|x_{k-1}) = N(\sqrt{1-\beta_k}x_{k-1}, \beta_k\mathbf{I})$, $\{\beta_k \in (0, 1), k \in (1, K)\}$ is a set of scheduled noise weight. Based on the forward transition probability, the generated noised data in step k can be calculated by $x_k = \sqrt{\alpha_k}x_0 + (1 - \alpha_k)\epsilon$. The reversed chain utilizes a well-trained network p_θ to recurrently denoise x_K for recovering original data x_0 , which can be expressed as $p_\theta(x_{k-1}|x_k) = N(\mu_\theta(x_k^k, k), \sigma_\theta(x_k, k)\mathbf{I})$. The objective of the diffusion model is essentially to maximize the log-likelihood function of the denoising network p_θ for the initial data x_0 , i.e.,

$$L(\theta) = \mathbb{E}_{x_0 \sim q(x_0)} \{-\log p_\theta(x_0)\}. \quad (17)$$

Subsequently, this function is optimized using the Variational Lower Bound (VLB), which can be expressed as:

$$\begin{aligned} -\log p_\theta(x_0) &\leq -\log p_\theta(x_0) + D_{KL}(q(x_{1:T}|x_0)||p_\theta(x_{1:T}|x_0)) \\ &= \mathbb{E}_{q(x_{1:T}|x_0)} \left\{ \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T}|x_0)} \right\}. \end{aligned} \quad (18)$$

Taking the expectation on both sides of the above equation and applying Fubini's theorem [63], we can derive:

$$\begin{aligned} L(\theta) &= \mathbb{E}_{q(x_0)} \{-\log p_\theta(x_0)\} \leq \mathbb{E}_{q(x_0)} \left\{ \mathbb{E}_{q(x_{1:T}|x_0)} \left\{ \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T}|x_0)} \right\} \right\} \\ &= \mathbb{E}_{q(x_{0:T})} \left\{ \log \frac{q(x_{1:T}|x_0)}{p_\theta(x_{0:T}|x_0)} \right\} \triangleq L_{vb}(\theta). \end{aligned} \quad (19)$$

We can minimize the upper bound of $L(\theta)$ by minimizing L_{vb} , thereby maximizing the log-likelihood function of p_θ . Ho *et al.* [19] proved that $L_{vb}(\theta)$ can be further parameterized by $\mu_\theta(x_k, k) = \alpha_k^{-0.5}[x_k - \beta_k(1 - \hat{\alpha}_k)^{-0.5}\epsilon_\theta(x_k, k)]$, and σ_θ can be parameterized as $\sigma_\theta(x_k, k) = \sqrt{(1 - \hat{\alpha}_{k-1})/(1 - \hat{\alpha}_k)\beta_k}$. The network p_θ can then be optimized by the following objective:

$$\min_{\theta} L_{vb}(\theta) \approx \min_{\theta} \mathbb{E}_{x_0 \sim q(x_0), \epsilon \sim N(0, I)} [\|\epsilon - \epsilon_\theta(x_k, k)\|_2^2]. \quad (20)$$

The objective in equation (20) is fundamentally equivalent to that of InfoNCE in contrastive learning. We use p_θ to represent the probability in the mutual information as $I(e, y) = p_\theta(e_{0:K}|y)/p_\theta(e_{0:K})$. In this way, the origin InfoNCE loss can be rewritten as:

$$\begin{aligned} L &= \mathbb{E}_{e \in \mathbb{B}} -\log \frac{p_\theta(e_{0:K}|y)/p_\theta(e_{0:K})}{p_\theta(e_{0:K}|y)/p_\theta(e_{0:K}) + \sum_{e'} p_\theta(e'_{0:K}|y)/p_\theta(e'_{0:K})} \\ &= \mathbb{E}_{e \in \mathbb{B}} \log \left\{ 1 + \frac{p_\theta(e_{0:K})}{p_\theta(e_{0:K}|y)} \cdot N \mathbb{E}_{e'} \frac{p_\theta(e'_{0:K}|y)}{p_\theta(e'_{0:K})} \right\}, \end{aligned} \quad (21)$$

where e' denotes all negative samples. Referencing the parameterization in [2] where $p_\theta(x_{k-1}|x_k) = \sum_{x_0 \in \mathcal{Q}} q(x_{k-1}|x_k, x_0) p_\theta(x_0|x_k)$, the above loss can be further formulated as:

$$\begin{aligned} L &\approx \mathbb{E}_q \left\{ -\log \frac{p_\theta(e_{0:K}|y)}{q(e_{1:K}|e_0)} \right\} - \log N \mathbb{E}_{e'} \mathbb{E}_q \left\{ -\log \frac{p_\theta(e'_{0:K}|y)}{q(e'_{1:K}|e'_0)} \right\} \\ &= L_{ob}^e - \log N \sum_{e'} L_{ob}^{e'} \\ &\doteq \mathbb{E} \left\{ (\|\epsilon - \epsilon_\theta(e, k|y)\|^2 - \lambda \sum_{e'} \|\epsilon - \epsilon_\theta(e', k|y)\|^2) \odot m \right\}, \end{aligned} \quad (22)$$

where symbol \doteq denotes the loss function we used during the model training process and λ is a scaling parameter proportional to $\log N$.

A.5 Ablation study and comparison

To test the effectiveness of our proposed fine-tuning module, we conduct ablation experiments on UoMo, as shown in Table 6, with UoMo-user and UoMo-POI representing the incorporation of mobile users and POI distributions, respectively, during the fine-tuning process. It can be observed that adding these two contextual features at the fine-tuning stage enhances model performance to varying degrees. Moreover, the performance degradation of UoMo-POI is more significant, indicating that mobile users better reflect the dynamic characteristics of mobile traffic and are more critical for mobile traffic forecasting compared to POI distribution.

Table 6: Performance of ablation study. Δ represents the degradation after removing certain modules.

Model	Beijing		Shanghai		Nanchang	
	Prediction (RMSE)	Generation (JSD)	Prediction (RMSE)	Generation (JSD)	Prediction (RMSE)	Generation (JSD)
UoMo (our)	0.1035	0.2213	0.0983	0.2202	0.0360	0.2226
UoMo-user	0.1230	0.2294	0.1295	0.2264	0.0421	0.2260
Δ	-23.82%	-21.77%	-47.75%	-24.31%	-19.81%	-34.69%
UoMo-POI	0.1758	0.2464	0.1507	0.2363	0.0636	0.2301
Δ	-88.38%	-67.47%	-80.36%	-63.13%	-89.61%	-76.53%
Pre-train	0.1853	0.2585	0.1635	0.2457	0.0668	0.2324

We comprehensively summarize and compare the UoMo algorithm, UniST, Lagllama, and Open-Diff, as shown in Figure 9. In the figure, we highlight the performance of our model as well as the performance of the second-best model. UoMo consistently delivers the best performance across various tasks and multiple cities. For example, in few-shot tasks, UoMo achieves performance improvements of 10.6% and 32.1% in different cities, demonstrating the model's universality.

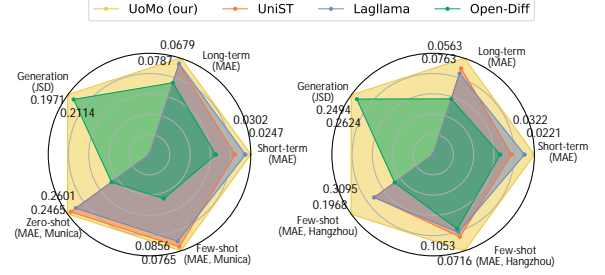


Figure 9: Comparison of forecasting models in Nanjing (left) and Nanchang (right) datasets.

A.6 Scaling performance of UoMo

Scaling performance reflects the relationship between model parameters, data size, and overall performance. Understanding the scaling performance of foundation models provides valuable guidance for parameter selection during model deployment, which optimizes computational and storage overheads for the entire system. We provide 5 different model settings based on various transformer layers and hidden features: UoMo with 5M, 35M, 100M, 150M, and 200M parameters.

We first examine the scaling performance between model parameters and training loss in Figure 10, where the loss is calculated via MSE (Equation 22). The results indicate that as the model size increases, the denoising network's capability improves, leading to enhanced learning capability of mobile traffic features. For smaller models, increasing the number of parameters significantly boosts learning capacity (as seen from 5M parameters \rightarrow 35M parameters in the figure). In contrast, for larger models, a marginal diminishing becomes apparent (as depicted of 100M parameters \rightarrow 200M parameters).

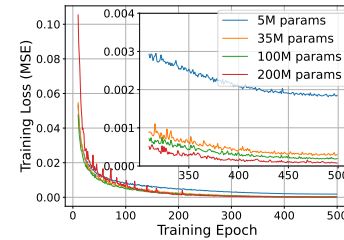


Figure 10: Training performance vs. model scales.

We explore the relationship between model size and performance across tasks, as shown in Figure 11(a). Smaller models improve quickly with more parameters, while larger models show diminishing returns. We attribute this to parameter redundancy in larger models relative to fixed training data, leading to overfitting and limiting performance improvements. To assess scaling performance, we evaluate the model with varying dataset sizes in Figure 11(b). Larger models degrade with small datasets, but as data volume increases, they leverage their extensive parameters to improve performance. In contrast, smaller models struggle with diverse features, leading to performance drops. From these observations, we identify a scaling regulation for UoMo: *Simply increasing model parameters*

does not guarantee better performance in mobile traffic forecasting. The optimal model size depends on the available data, prompting further investigation into the relationship between model scale and factors like urban size, population, and temporal granularity to enhance performance.

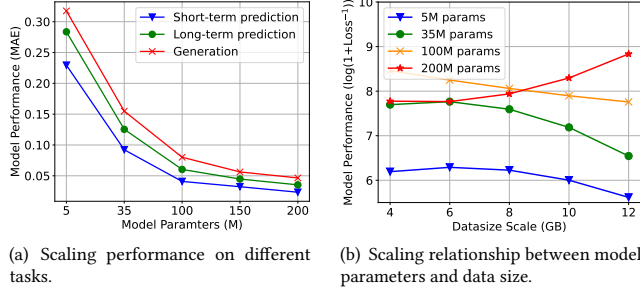


Figure 11: Scaling performance of UoMo.

A.7 The framework of Jiutian platform

The Platform is China Mobile's self-developed AI innovation platform, providing intelligent decision-making support for mobile networks. The framework of the Jiutian platform is illustrated in Figure 12. Our model acts as a key part of the platform to generate mobile traffic data (the green part in the figure), by which we implement optimization and planning applications and evaluate the network performance (the yellow parts in the figure). For different network optimization and planning scenarios, UoMo uses various masks to forecast mobile traffic. In this paper, we select three optimization scenarios:

In the base station deployment scenario, since there is no historical traffic data available, we use the generation mask to generate the mobile traffic distribution for the target area and estimate the region's potential network demand, which is formulated and optimized as described in 5.1.

In the base station sleep scenario, we use the long-term prediction mask to capture the long-term variations in mobile traffic patterns and optimize the overall network energy consumption from a long-term perspective as described in 5.2.

In the infinite resource allocation scenario, we use the short-term prediction mask to forecast traffic fluctuations over a short period and develop corresponding optimization strategies in advance. We provide a detailed formulation in A.8.

A.8 Resource allocation

We consider cellular networks involving a single cloud server and N cells, where the server allocates K PRBs to each cell [50, 71]. We aim to optimize the average delay of each cell by allocating a limited number of PRBs across multiple time slots. We assume that the average channel quality h_t of each cell remains constant at time slot t and varies independently between different slots. The average rate in cell n can be expressed as $r_n(t) = B \sum_k \rho_{k,n}(t) \log_2[1 + p_n(t)h_n(t)/\sigma^2]$, where B is the PRB bandwidth and $p_n(t)$ is the transmission power. $\rho_{k,n}(t)$ is a binary variable indicating PRB

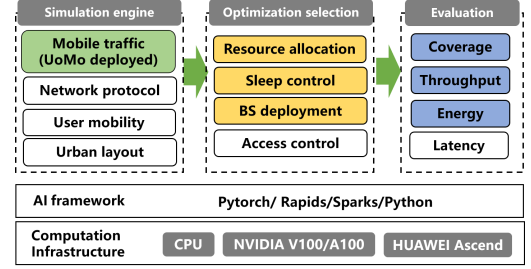


Figure 12: Optimization framework of Jiutian platform

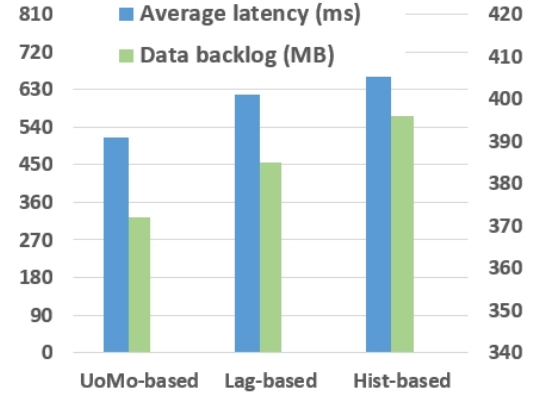


Figure 13: Wireless resource allocation.

allocation ($\rho_{k,n}(t) = 1$ represents PRB k is allocated to cell n , and vice versa). Based on the PRB selection, the end-to-end delay is expressed as $d_n(t) = U_n(t)/r_n(t)$, where $U_n(t)$ is the overall transmission tasks. When there are too many transmission tasks for a cell, it might not be possible to complete all transmissions within a single time slot, leading to data backlog. Therefore, the backlog data at time $t + 1$ for the cell can be expressed as $U_n(t + 1) = \max\{U_n(t) + P_n(t) - r_n(t), 0\}$, where $P_n(t)$ represents the newly arriving transmission data. Our ultimate goal is to minimize the accumulated backlog data and end-to-end delay for the cells, which yields:

$$\min \sum_t \sum_n U_n(t + 1) + d_n(t), \quad (23)$$

where $P_n(t)$ is provided by the UoMo. Compared to the BS sleep modes control, the PRB allocation primarily focuses on the user experience over a relatively short period. Therefore, we use the short-term prediction mask in (1) to generate network traffic for the near future. As shown in Figure 13, our UoMo-based strategy significantly reduces system latency and overall data backlog by 16.4% and 3.5%, respectively. This is because UoMo can more accurately characterize the cells' future transmission tasks, allowing us to make decisions based on the predicted data that are closer to reality, and improve the system performance.