UniST: A Prompt-Empowered Universal Model for Urban Spatio-Temporal Prediction

Yuan Yuan
Department of Electronic Engineering, BNRist, Tsinghua University, Beijing, China
y-yuan20@mails.tsinghua.edu.cn

Jingtao Ding
Department of Electronic Engineering, BNRist, Tsinghua University, Beijing, China
dingjt15@tsinghua.org.cn

Jie Feng
Department of Electronic Engineering, BNRist, Tsinghua University, Beijing, China
fengj12ee@hotmail.com

Depeng Jin
Department of Electronic Engineering, BNRist, Tsinghua University, Beijing, China
jindp@tsinghua.edu.cn

Yong Li
Department of Electronic Engineering, BNRist, Tsinghua University, Beijing, China
liyong07@tsinghua.edu.cn

ABSTRACT

Urban spatio-temporal prediction is crucial for informed decision-making, such as traffic management, resource optimization, and emergence response. Despite remarkable breakthroughs in pre-trained natural language models that enable one model to handle diverse tasks, a universal solution for spatio-temporal prediction remains challenging. Existing prediction approaches are typically tailored for specific spatio-temporal scenarios, requiring task-specific model designs and extensive domain-specific training data. In this study, we introduce UniST, a universal model designed for general urban spatio-temporal prediction across a wide range of scenarios. Inspired by large language models, UniST achieves success through: (i) utilizing diverse spatio-temporal data from different scenarios, (ii) effective pre-training to capture complex spatio-temporal dynamics, (iii) knowledge-guided prompts to enhance generalization capabilities. These designs together unlock the potential of building a universal model for various scenarios. Extensive experiments on more than 20 spatio-temporal scenarios demonstrate UniST's efficacy in advancing state-of-the-art performance, especially in few-shot and zero-shot prediction. The datasets and code implementation are released on https://github.com/tsinghua-fib-lab/UniST.

CCS CONCEPTS

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

KEYWORDS

Spatio-temporal prediction, prompt learning, universal model

ACM Reference Format:
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 (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671662

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08.
https://doi.org/10.1145/3637528.3671662

1 INTRODUCTION

Pre-trained foundation models have showcased remarkable success in Natural Language Processing (NLP) [3, 52], particularly excelling in few-shot and zero-shot settings [3, 27]. However, similar breakthroughs have not yet been achieved in the field of urban spatio-temporal prediction [17, 53, 73]. In this paper, our goal is to establish a foundation model for general urban spatio-temporal prediction — specifically, to develop a universal model that offers superior performance and powerful generalization capabilities across diverse spatio-temporal scenarios. This entails training a single model capable of effectively handling various urban contexts, encompassing various domains such as human mobility, traffic and communication networks across different cities.

The significance of such a universal model lies in its ability to address prevalent data scarcity issues in urban areas. The varying levels of digitalization across domains and cities often result in imbalanced and incomplete datasets. Despite notable advancements in existing spatio-temporal modeling approaches [1, 29, 35, 43, 67, 76], their effectiveness is typically confined to specific domains within a single city. The reliance on extensive training data further impedes the model's generalization potential. Consequently, current solutions are still far from "universality", and remain narrowly applicable.

A universal spatio-temporal model must possess two essential capabilities. *Firstly, it must be capable of leveraging abundant and rich data from different urban scenarios for training.* The training of the foundational model should ensure the acquisition of ample and rich information [2, 52, 58]. *Second, it should demonstrate robust generalization across different spatio-temporal scenarios.* Especially in scenarios with limited or no training data, the model can still work well without obvious performance degradation [14, 58].

However, realizing the aforementioned capabilities encounters significant challenges specific to spatio-temporal data, which impede the direct application of current foundation models developed for language and vision domains. The first challenge arises from the inherent *diverse formats* of spatio-temporal datasets. Unlike languages with a natural and unified sequential structure or images and videos adhering to standardized dimensions, spatio-temporal data collected from different sources exhibit highly varied features. These include variable dimensions, temporal durations, and spatial

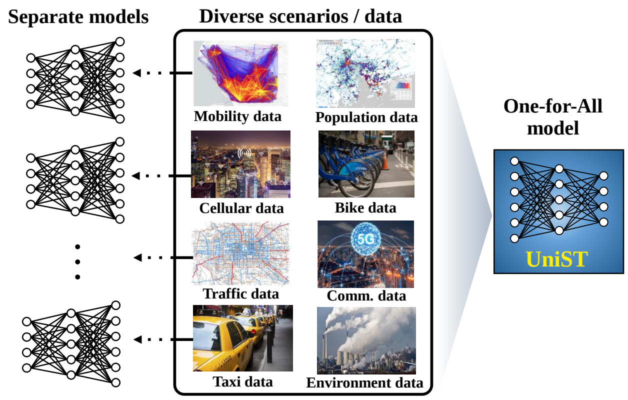


Figure 1: The transition from traditional separate deep learning models to a one-for-all universal model for urban spatio-temporal prediction.

coverages that differ significantly, posing difficulties in standardizing their structure. The second challenge arises from *high variations in data distributions across multiple scenarios*. Faced with highly distinct spatio-temporal patterns, the model may struggle to adapt to these differences. Unlike language, which benefits from a shared vocabulary, various scenarios of different domains and cities often operate on entirely different spatial and temporal scales, lacking common elements for effective training and generalization.

Although the displayed spatio-temporal patterns vary significantly, there are certain underlying laws that should be common among them. This principle arises from the intuition that human activity influences various spatio-temporal data generated in urban settings, leading to the existence of universal patterns. For example, traffic speed and communication networks exhibit distinct spatio-temporal patterns, yet both are influenced by human mobility and therefore adhere to similar underlying principles. Additionally, while temporal periodic patterns vary across domains, they share fundamental concept of repetition. Furthermore, city layouts vary considerably between different urban areas, but the relationships among various functional zones within cities may exhibit shared characteristics. Therefore, the key to building a one-for-all model is to capture, align and leverage these shared while underlying characteristics effectively.

To this end, we introduce ***UniST***, a **uni**versal solution for urban **s**patio-**t**emporal prediction through advanced pre-training and prompt learning. Notably, UniST achieves three essential capabilities of:

(1) Scalability across scenarios with diverse spatio-temporal data;
(2) Effective pre-training to capture complex spatio-temporal relationships;
(3) utilizing spatio-temporal prompts to align underlying shared patterns across scenarios.

UniST achieves the above capabilities through its holistic design driven by four key components: *data*, *architecture*, *pre-training*, and *prompt learning*. Firstly, we harness the rich diversity inherent in spatio-temporal scenarios by leveraging extensive *data* from various domains and cities. Secondly, we design spatio-temporal patching to unify diverse data into a sequential format, facilitating the utilization of the powerful Transformer *architecture*. Thirdly, drawing inspiration from large language and vision models [9, 18], UniST adopts the widely-used generative *pre-training* strategy – Masked Token Modeling (MTM). We further enhance the model's capability to capture complex spatio-temporal relationships by employing multiple masking strategies that comprehensively address multi-perspective correlations. Moreover, informed by the established domain knowledge in spatio-temporal modeling, we design an innovative prompt learning approach. The elaborated prompt network identifies underlying and shared spatio-temporal patterns, adapting dynamically to generate useful prompts. In this way, UniST aligns distinct data distributions of various datasets and advances towards developing a one-for-all universal model. We summarize our contributions as follows:

- To our best knowledge, this the first attempt to address universal spatio-temporal prediction by investigating the potential of a one-for-all model in diverse spatio-temporal scenarios.
- We propose UniST that harnesses data diversity and achieves universal spatio-temporal prediction through advanced pre-training and prompt learning. It has made a paradigm shift from traditional separate deep learning methods to a one-for-all model.
- Extensive experiments demonstrate the generality and universality of UniST. It achieves new state-of-the-art performance on various prediction tasks, particularly, superior few-shot and zero-shot capabilities.

2 RELATED WORK

Urban Spatio-Temporal Prediction. Urban spatio-temporal prediction [53, 73] aims to model and forecast the dynamic patterns of urban activities over space and time. Deep learning techniques has propelled significant advancements. A spectrum of models, including CNNs [30, 35, 67], RNNs [33, 55, 56], ResNets [67], MLPs [46, 69], GNNs [1, 15, 72], Transformers [6, 7, 21, 64], and diffusion models [65, 75], have been introduced to capture spatio-temporal patterns. Simultaneously, cutting-edge techniques like meta-learning [40, 63], contrastive learning [19, 68], and adversarial learning [42, 51] are also utilized. However, most approaches remain constrained by training separate models for each specific dataset. Some studies [25, 39, 40, 63] explore transfer learning between cities, however, a certain amount of data samples in the target city are still required. Current solutions are restrictive to specified spatio-temporal scenarios and require training data, while our model allows generalization across diverse scenarios and provides a one-for-all solution.

Foundation Models for Spatio-temporal Data and Time Series. Inspired by the remarkable strides in foundation models for NLP [3, 52] and CV [2, 45], foundation models for urban prediction have emerged recently. Some explorations unlock the potential of large language models (LLMs) in this context. Intelligent urban systems like CityGPT [10, 61], CityBench [11] and UrbanGPT [31] have demonstrated proficiency in addressing language-based tasks. Additionally, LLMs are utilized for describing urban-related images [62] to benefit downstream tasks and predict user activities [16]. Moreover, the application of LLMs extends to traffic signal control [28], showcasing their utility in tackling complex spatio-temporal problems beyond languages. Recently, there also has been great progress

Table 1: Comparison of UniST with other spatio-temporal models regarding important properties.

Model	Scalability[(1)]	Few-shot	Zero-shot	Efficiency
PromptST [70]	✗	✗	✗	✓
GPT-ST [32]	✗	✗	✗	✓
STEP [47]	✗	✗	✗	✓
ST-SSL [19]	✗	✗	✗	✓
TrafficBERT [22]	✓	✗	✗	✓
TFM [54]	✗	✗	✗	✓
UrbanGPT [31]	✓	✓[(2)]	✓[(2)]	✗
STG-LLM [34]	✗	✗	✗	✗
UniST	✓	✓	✓	✓

(1) Whether can leverage diverse datasets with diverse formats.
(2) Restricted in the same city.

in foundation models for time series [4, 23, 24, 74]. Unlike time series characterized by a straightforward sequential structure, spatio-temporal data presents a more intricate nature with intertwined dependencies across both spatial and temporal dimensions. While exploring the integration of LLMs is promising, it's important to recognize that spatio-temporal data is not inherently generated by language. Thus, developing foundation models specifically trained on pure spatio-temporal data is also an important direction. In Table 1, we compare the essential properties of UniST with other approaches employing pre-training, prompt learning, or LLMs. UniST encompasses all these essential capabilities, whereas other approaches have certain limitations.

Prompt Learning. Prompt learning has achieved superior performance in large models [20, 36, 44, 48], with the goal of enhancing the generalization capability of pretrained models on specific tasks or domains. Typically, language models usually use a limited number of demonstrations as prompts and vision models often employ a learnable prompt network to generate useful prompts, known as prompt learning. Our research aligns with prompt learning, where spatio-temporal prompts are adaptively generated based on spatio-temporal patterns through a prompt network.

3 METHODOLOGY

3.1 Preliminary

Spatial and Temporal Partitions. We use a grid system for spatial partitioning, dividing the city into equal, non-overlapping areas defined by longitude and latitude on an $H \times W$ map. For each area, the temporal dynamics are recorded at certain intervals.

Spatio-Temporal Data. A spatio-temporal data X is defined as a four-dimensional tensor with dimensions $T \times C \times H \times W$, where T represents time steps, C represents the number of variables, H and W represent spatial grids. T, C, H, and W can vary across different spatio-temporal scenarios.

Spatio-Temporal Prediction. For a specific dataset, given l_h historical observations for the grid map, we aim to predict the future k steps. The spatio-temporal prediction task can be formulated as learning a θ-parameterized model $\mathcal{F}$: $X_{[t:t+k]} = \mathcal{F}_\theta(X_{[t-l_h:t]})$.

Few-Shot and Zero-Shot Predictions. The model is trained on multiple source datasets and then adapted to a target dataset. In few-shot learning, it is fine-tuned with a small amount of target samples; in zero-shot learning, it makes predictions without any fine-tuning.

3.2 Pre-training and Prompt Learning

Universal spatio-temporal prediction aims to empower a single model to effectively handle diverse spatio-temporal scenarios, requiring the unification of varied spatio-temporal data within a cohesive model. This necessitates addressing significant distribution shifts across datasets of different scenarios. To achieve this goal, we propose a framework for pre-training and prompt learning, leading to a universal prediction model, UniST. Figure 2 shows the overview architecture, detailing UniST with two stages:

- **Stage 1: Large-scale spatio-temporal pre-training.** Different from existing methods limited to a single dataset, our approach utilizing extensive spatio-temporal data from a variety of domains and cities for pre-training.
- **Stage 2: Spatio-temporal knowledge-guided prompt learning.** We introduces a prompt network for in-context learning, where the generation of prompts is adaptively guided by well-developed spatio-temporal domain knowledge, such as spatial hierarchy and temporal periodicity.

3.3 Base Model

Our base model is a Transformer-based encoder-decoder architecture. Through spatio-temporal patching, it can handle diverse spatio-temporal data in a unified sequential format.

Spatio-Temporal Patching. The conventional Transformer architecture is designed for processing 1D sequential data. However, spatio-temporal data possesses a 4D structure. To accommodate this, we first split the data into channel-independent instances, which are 3D tensors. Then, we utilize spatio-temporal patching to transform the 3D tensor, denoted as $X \in \mathbb{R}^{L \times H \times W}$, into multiple smaller 3D tensors. If the original shape is $L \times H \times W$, and the patch size is (l, h, w), the resulting sequence is given by $E_x \in \mathbb{R}^{L' \times H' \times W'}$, $L' = \frac{L}{l}, H' = \frac{H}{h}, W' = \frac{W}{w}$.

This transformation involves a 3D convolutional layer with a kernel size and stride both set to (l, h, w). The process can be expressed as $E_x = \textsc{Conv}_{3d}(X)$, where E_x represents the converted 1D sequential data. The sequence length of E_x is $L' \times H' \times W'$.

Positional Encoding. As the original Transformer architecture does not consider the order of the sequence, we follow the common practice that incorporate positional encoding [9]. To enhance generalization, we choose sine and cosine functions rather than learnable parameters for positional encoding. This encoding is separately applied to the spatial and temporal dimensions.

Encoder-Decoder Structure. The base model utilizes an encoder-decoder framework inspired by Masked Autoencoder (MAE) [18]. It processes input patches with a certain masking ratio, where the encoder takes the unmasked patches and the decoder reconstructs the image using the encoder's output and the masked patches. Our focus is on capturing comprehensive spatio-temporal dependencies, including both high-level and low-level relationships, with the goal of accurately predicting values at specific time and space coordinates. Unlike MAE, which uses a lightweight decoder for pre-training, our model employs a full-sized decoder that plays a crucial role in both pre-training and fine-tuning. It can be formulated as:

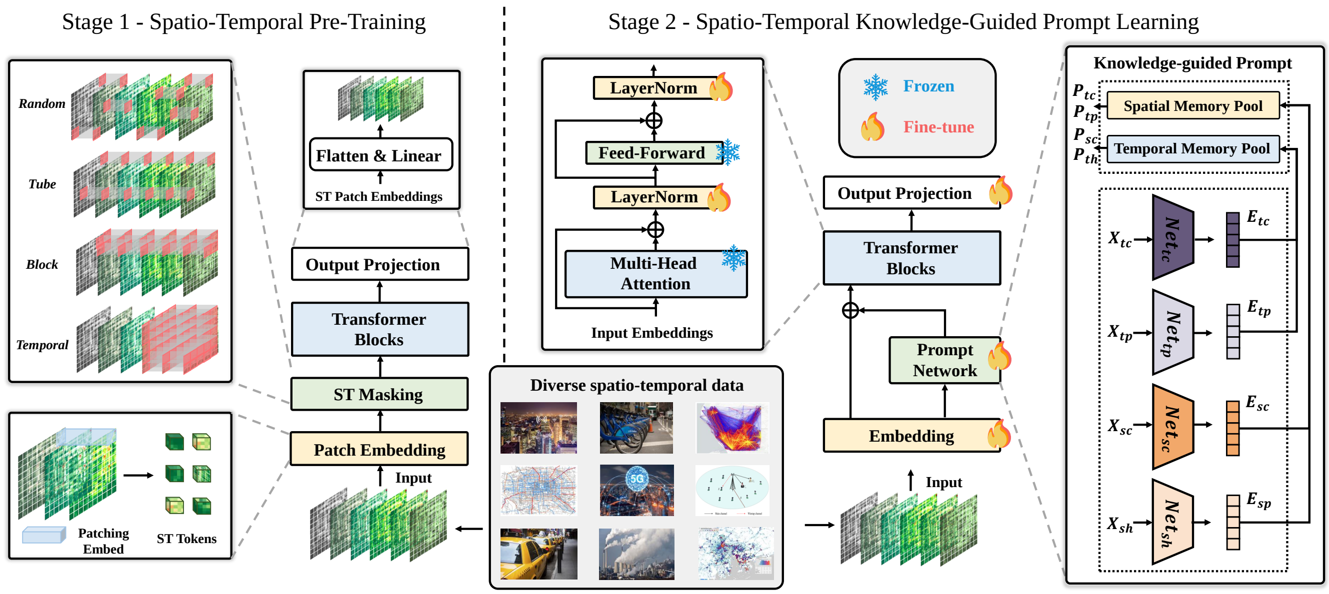


Figure 2: The overview architecture of UniST, which consists of two stages: (i) large-scale spatio-temporal pre-trianing, (ii) spatio-temporal knowledge-guided prompt learning.

$$E_{enc} = \textsc{Encoder}(E_x),\ Y_{dec} = \textsc{Decoder}(E_{enc}, E_{mask}),$$

where E_{mask} denotes the token embeddings for the masked patch.

3.4 Spatio-Temporal Self-Supervised Pre-train

In pretrained language models, the self-supervised learning task is either masking-reconstruction [9] or autoregressive prediction [3]. Similarly, in vision models, visual patches are randomly masked and the pre-training objective is to reconstruct the masked pixels. To further augment the model's capacity to capture intricate spatio-temporal relationships and intertwined dynamics, we introduce four distinct masking strategies during the pre-training phase, which are shown in the left box in the stage 1 of Figure 2. Suppose the masking percentage is r, we explain these strategies as follows:

- **Random masking.** This strategy is similar to the one used in MAE, where spatio-temporal patches are randomly masked. Its purpose is to capture fine-grained spatio-temporal relationships.
$$M \sim \mathbf{U}[0,1],\ E_x = E_x[M < 1-r],\ M \in \mathbb{R}^{L' \times H' \times W'}.$$
- **Tube masking.** This strategy simulates scenarios where data for certain spatial units is entirely missing across all instances in time, mirroring real-world situations where some sensors may be nonfunctional—a common occurrence. The goal is to improve spatial extrapolation competence.
$$M \sim \mathbf{U}[0,1],\ E_x = E_x[:, M < 1-r],\ M \in \mathbb{R}^{H' \times W'}.$$
- **Block masking.** A more challenging variant of tube masking, block masking involves the complete absence of an entire block of spatial units across all instances in time. The reconstruction task becomes more intricate due to limited context information, with the objective of enhancing spatial transferability.
$$M \sim \textbf{Uniform}(1,2),\ E_x = E_x[:, \frac{M-1}{2}H' : \frac{M}{2}H', \frac{M-1}{2}W' : \frac{M}{2}W'].$$
- **Temporal Masking.** In this approach, future data is masked, compelling the model to reconstruct the future based solely on historical information. The aim is to refine the model's capability to capture temporal dependencies from the past to the future.
$$M = \textsc{Concat}([\mathbf{1}_{(1-r)L' \times H' \times W'}, \mathbf{0}_{rL' \times H' \times W'}]),\ E_x = E_x[M = 1].$$

By employing these diverse masking strategies, the model can systematically enhance its modeling capabilities from a comprehensive perspective, simultaneously addressing spatio-temporal, spatial, and temporal relationships.

3.5 Spatio-Temporal Knowledge-Guided Prompt

Prompt learning plays a critical role in enhancing UniST's generalization ability. Before delving into the details of our prompt design, it is essential to discuss why pre-trained models can be applied to unseen scenarios.

3.5.1 ***Spatial-Temporal Generalization.*** In urban prediction tasks, the distributions of features and labels differ across domains and cities, denoted as $X_A \neq X_B, Y_A \neq Y_B$, where X and Y denote features and labels, while A and B represent different cities or domains. Taken A and B as a simple example, generalization involves leveraging knowledge acquired from the A dataset and adapt it to the B dataset. The key point lies in identifying and aligning "related" patterns between A and B datasets. While finding similar patterns for an entire dataset may be challenging, we claim that identifying and aligning fine-grained patterns is feasible. Specifically, we provide some assumptions that applies to prompt-empowered spatio-temporal generalization, which are expressed as follows:

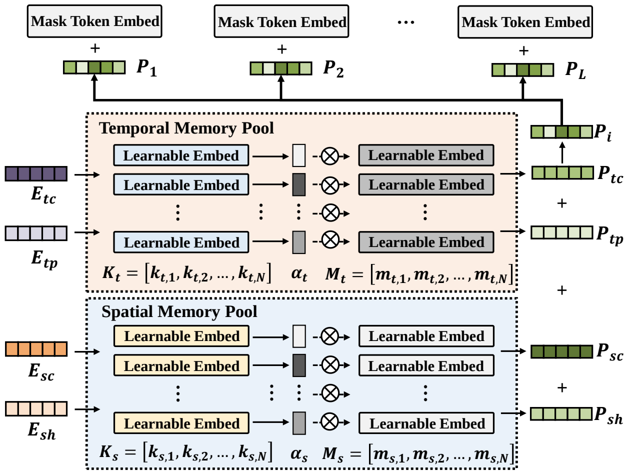


Figure 3: Illustration of the prompt generation process.

Assumption 1. *For a new dataset* $\mathbf{B}$*, it is possible to identify fine-grained patterns related to the training data* $\mathbf{A}$*.*

$$X_A \neq X_B,\ Y_A \neq Y_B,$$
$$\exists x_a \in X_A, y_a \in Y_A,\ \exists x_b \in X_B, y_b \in Y_B, : x_a \approx x_b, y_a \approx y_b.$$

Assumption 2. *Distinct spatio-temporal patterns correspond to customized prompts.*

$$P_i^* \neq P_j^* \quad \text{if} \quad D(x_i, x_j) > \epsilon,$$
$$D(P_i^*, P_j^*) > D(P_m^*, P_n^*) \quad \text{if} \quad D(x_i, x_j) > D(x_m, x_n),$$

where x_i *denotes the fine-grained spatio-temporal pattern,* P_i^* *represents the prompt of* x_i*, and* D *is the similarity between* x_i *and* x_j*.*

Assumption 3. *There exists* f_θ *that captures the mapping relationship from the spatio-temporal pattern* x_i *to prompt* P_i^**.*

$$P_i = f_\theta(x_i) \quad \text{where} \quad \theta = \underset{\theta}{\text{argmin}} \sum_i \text{DISTANCE}(P_i^*, f_\theta(x_i)).$$

Based on these assumptions, our core idea is that for different inputs with distinct spatio-temporal patterns, customized prompts should be generated adaptively.

3.5.2 ***Spatio-Temporal Domain Knowledge***. Given the aforementioned assumptions, a critical consideration is how to define the concept of "similarity" to identify and align shared spatio-temporal patterns. Here we leverage insights from well-established domain knowledge in spatio-temporal modeling [67, 73], encompassing properties related to both space and time. There are four aspects to consider when examining these properties:

- Spatial closeness: Nearby units may influence each other.
- Spatial hierarchy: The spatial hierarchical organization impacts the spatio-temporal dynamics, requiring a multi-level perception on the city structure.
- Temporal closeness: Recent dynamics affect future results, indicating a closeness dependence.
- Temporal period: Daily or weekly patterns exhibit similarities, displaying a certain periodicity.

For simplicity, we provide some straightforward implementations, which are shown in the four networks in Figure 2, *i.e.,* NET_{tc}, NET_{tp}, NET_{sc}, and NET_{sh}. For the spatial dimension, we first employ an attention mechanism to merge the temporal dimension into a representation termed E_s. Then, to capture spatial dependencies within close proximity, a two-dimensional convolutional neural network (CNN), *i.e.,* NET_{sc}, with a kernel size of 3 is employed. To capture spatial hierarchies, we utilize CNNs with larger kernel sizes, *i.e.,* NET_{sh}. These larger kernels enable the perception of spatial information on larger scales, which facilitate to construct a hierarchical perspective. As for the temporal dimension, we employ an attention network, *i.e.,* NET_{tc}, to aggregate the previous M steps denoted as X_c. Regarding the temporal period, we select corresponding time points from the previous N days, denoted as X_p. Subsequently, we employ another attention network, *i.e.,* NET_{tp}, to aggregate the periodical sequence, which captures long-term temporal patterns. The overall process is formulated as follows:

$$E_{sc} = \text{CONV}_{2D}[3](X_s),$$
$$E_{sh} = \{\text{CONV}_{2D}[2^i + 1](X_s)\}, i \in \{2, 3, 4\},$$
$$E_{tc} = \text{ATTENTION}(X_c),$$
$$E_{tp} = \text{ATTENTION}(X_p).$$

It is essential to emphasize that the learning of E_{sc}, E_{sh}, E_{tc}, and E_{tp} is not restricted by our practice. Practitioners have the flexibility to employ more complex designs to capture richer spatio-temporal properties. For example, Fourier-based approaches [38, 60] can be utilized to capture periodic patterns.

3.5.3 ***Spatio-Temporal Prompt Learner***. Given the representations of properties derived from spatio-temporal domain knowledge, the pivotal question is how to generate prompts—*how does spatio-temporal knowledge guide prompt generation?* Here we utilize prompt learning techniques. While prompt learning in computer vision [20] often train fixed prompts for specific tasks such as segmentation, detection, and classification. Due to the high-dimensional and complex nature of spatio-temporal patterns, training a fixed prompt for each case becomes impractical.

To tackle this issue, we draw inspirations from memory networks [49] and propose a novel approach that learns a spatial memory pool and a temporal memory pool. In the prompt learning process, these memory pools are optimized to store valuable information about spatio-temporal domain knowledge. As shown in Figure 3, the spatial and memory pools are defined as follows:

$$KM_s = \{(k_{s,0}, m_{s,0}), (k_{s,1}, m_{s,1}), ..., (k_{s,N-1}, m_{s,N-1})\},$$
$$KM_t = \{(k_{t,0}, m_{t,0}), (k_{t,1}, m_{t,1}), ..., (k_{t,N-1}, m_{t,N-1})\},$$

where $k_{s,i}, m_{s,i}, k_{t,i}, m_{t,i}, i \in \{0, 1, ..., N-1\}$ are all learnable parameters, and the memory is organized in a key-value structure following existing practice [49, 59].

Subsequently, useful prompts are generated based on these optimized memories. This involves using the representations of spatio-temporal properties as queries to extract valuable memory knowledge, *i.e.,* pertinent embeddings from the memory pool. Figure 3 illustrates the process, and it is formulated as follows:

Table 2: Performance comparison of short-term prediction on seven datasets in terms of MAE and RMSE. We use the average prediction errors over all prediction steps. Bold denotes the best results and underline denotes the second-best results.

	TaxiBJ		Crowd		Cellular		BikeNYC		TrafficJN		TDrive		TrafficSH	
Model	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
HA	53.77	29.82	17.80	6.79	72.94	27.57	11.41	3.43	1.38	0.690	150.2	74.5	1.24	0.771
ARIMA	56.70	39.53	21.87	10.23	81.31	40.22	12.37	3.86	1.20	0.651	211.3	108.5	1.17	0.769
STResNet	45.17	30.87	5.355	3.382	24.30	14.32	8.20	4.98	0.964	0.556	220.1	117.4	1.00	0.723
ACFM	37.77	21.59	4.17	2.34	22.79	12.00	3.93	1.67	0.920	0.559	98.1	51.9	0.833	0.566
STID	27.36	14.01	3.85	1.63	18.77	8.24	4.06	1.54	0.880	0.495	47.4	23.3	0.742	0.469
STNorm	29.37	15.71	4.44	2.09	19.77	8.19	4.45	1.66	0.961	0.532	54.3	47.9	0.871	0.579
STGSP	45.04	28.28	7.93	4.56	39.99	21.40	5.00	1.69	0.882	0.490	94.6	47.8	1.02	0.749
MC-STL	29.14	15.83	4.75	2.39	21.22	10.26	4.08	2.05	1.19	0.833	54.2	28.1	1.00	0.720
PromptST	27.44	14.54	3.52	1.54	15.74	7.20	4.36	1.57	0.953	0.490	47.5	22.8	0.811	0.523
MAU	38.14	20.13	4.94	2.35	39.09	18.73	5.22	2.06	1.28	0.697	48.8	22.1	1.37	0.991
PredRNN	27.50	14.29	5.13	2.36	24.15	10.44	5.00	1.74	0.852	0.463	54.9	25.2	0.748	0.469
MIM	28.62	14.77	5.66	2.27	21.38	9.37	4.40	1.62	1.17	0.650	51.4	22.7	0.760	0.505
SimVP	32.66	17.67	3.91	1.96	16.48	8.23	4.11	1.67	0.969	0.556	46.8	22.9	0.814	0.569
TAU	33.90	19.37	4.09	2.11	17.94	8.91	4.30	1.83	0.993	0.566	51.6	28.1	0.820	0.557
PatchTST	42.74	22.23	10.25	3.62	43.40	15.74	5.27	1.65	1.25	0.616	106.4	51.3	1.10	0.663
iTransformer	36.97	19.14	9.40	3.40	37.01	13.93	7.74	2.53	1.11	0.570	86.3	42.6	1.04	0.655
PatchTST(one-for-all)	43.66	23.16	13.51	5.00	56.80	20.56	9.97	3.05	1.30	0.645	127.0	59.26	1.13	0.679
UniST(one-for-all)	**26.84**	**13.95**	**3.00**	**1.38**	**14.29**	**6.50**	**3.50**	**1.27**	**0.843**	**0.430**	**44.97**	**19.67**	**0.665**	**0.405**

$$\alpha_{sc} = [k_{s,0}; k_{s,1}; ..., k_{s,N-1}]E_{sc}^T,\ P_{sc} = \sum_i \alpha_{sc,i} m_{s,i},$$
$$\alpha_{sh} = [k_{s,0}; k_{s,1}; ..., k_{s,N-1}]E_{sh}^T,\ P_{sh} = \sum_i \alpha_{sh,i} m_{s,i},$$
$$\alpha_{tc} = [k_{t,0}; k_{t,1}; ..., k_{t,N-1}]E_{tc}^T,\ P_{tc} = \sum_i \alpha_{tc,i} m_{t,i},$$
$$\alpha_{tp} = [k_{t,0}; k_{t,1}; ..., k_{t,N-1}]E_{tp}^T,\ P_{tp} = \sum_i \alpha_{tp,i} m_{t,i},$$

where $E_{sc}, E_{sh}, E_{tc}, E_{tp}$ represent four representations related to four types of spatio-temporal domain knowledge, and $P_{sc}, P_{sh}, P_{tc}, P_{tp}$ are the extracted prompts. This allows the model to adaptively select the most useful information for prediction. These prompts are then integrated into the input space of the Transformer architecture, which are displayed in the upper part of Figure 3.

4 PERFORMANCE EVALUATIONS

4.1 Experimental Setup

To evaluate the performance of UniST, we conducted extensive experiments on more than 20 spatio-temporal datasets. Due to the page limit, we select a few representative results on below, and a full benchmark can be found in the arXiv version[1].

Datasets. The datasets we used cover multiple cities, spanning various domains such as crowd flow, dynamic population, traffic speed, cellular network usage, taxi trips, and bike demand. Appendix Table 4 and Table 5 (arXiv) provide a summary of the datasets we used. These spatio-temporal datasets originate from distinct domains and cities, and have variations in the number of variables, sampling frequency, spatial scale, temporal duration, and data size.

Baselines. We compare UniST with a broad collection of state-of-the-art models for spatio-temporal prediction, which can be categorized into five groups:

- **Heuristic approaches.** History average (HA) and ARIMA.
- **Deep urban prediction approaches.** We consider state-of-the-art urban ST prediction models, including STResNet [67], ACFM [35], MC-STL [68], STGSP [71], STNorm [8], STID [46], and PromptST [70].
- **Video prediction approaches.** We compare with competitive video prediction models from the popular benchmark, including PredRNN [56], MAU [5], MIM [57], SimVP [13], and TAU [50].
- **Multivariate time series forecasting approaches.** We consider state-of-the-art multivariate time series forecasting models, including PatchTST [41] and iTransformer [37]. For a fair comparison, we also train PatchTST for all datasets, denoted as PatchTST(one-for-all).
- **Meta learning approaches.** To evaluate the generalization capability, we consider meta-learning approaches including MAML [12] and MetaST [63].

Metrics. We employed commonly used regression metrics, including Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). For more detailed information of the datasets, baselines, and metrics, please refer to Appendix A, B, and Appendix D (arXiv).

4.2 Short-Term Prediction

Setups. Following previous practices [23, 41], both the input step and prediction horizon are set as 6, i.e., 6 $\rightarrow$ 6. For baselines, we

[1] https://arxiv.org/abs/2402.11838

Table 3: Performance comparison of long-term prediction on three datasets in terms of MAE and RMSE. We use the average prediction errors over all prediction steps. Bold denotes the best results and underline denotes the second-best results.

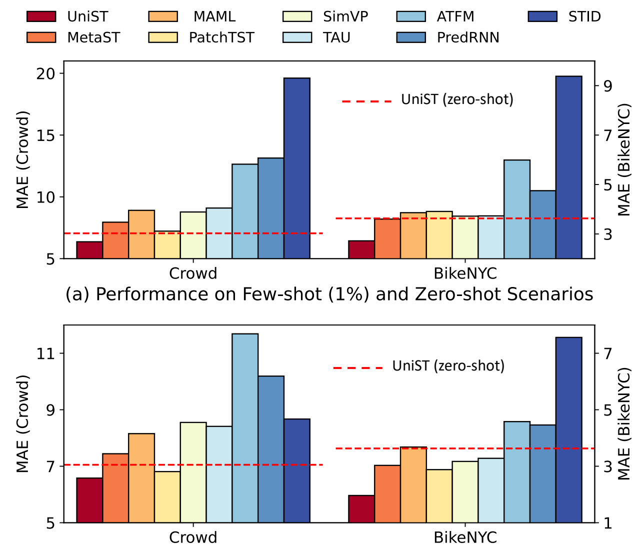
Model	TaxiNYC		Crowd		BikeNYC	
	RMSE	MAE	RMSE	MAE	RMSE	MAE
HA	61.03	21.33	19.57	8.49	11.00	3.66
ARIMA	68.0	28.66	21.34	8.93	11.59	3.98
STResNet	29.54	14.46	8.75	5.58	7.15	3.87
ACFM	32.91	13.72	6.16	3.35	4.56	1.86
STID	24.74	11.01	<u>4.91</u>	<u>2.63</u>	4.78	2.24
STNorm	31.81	11.99	9.62	4.30	6.45	2.18
STGSP	28.65	10.38	17.03	8.21	4.71	<u>1.54</u>
MC-STL	29.29	17.36	9.01	6.32	4.97	2.61
MAU	26.28	9.07	20.13	8.49	6.18	2.13
PredRNN	21.17	<u>7.31</u>	19.70	10.66	5.86	1.97
MIM	63.36	29.83	15.70	8.81	7.58	2.81
SimVP	<u>20.18</u>	9.78	5.50	3.13	4.10	1.71
TAU	24.97	10.93	5.31	2.81	<u>3.89</u>	1.73
PatchTST	30.64	17.49	5.25	2.83	5.27	1.65
iTransformer	33.81	11.48	6.94	2.63	6.00	2.02
PatchTST(one-for-all)	34.50	10.63	6.39	2.92	6.02	1.83
UniST (one-for-all)	**19.83**	**6.71**	**4.25**	**2.26**	**3.56**	**1.31**

train a dedicated model for each dataset, while UniST is evaluated across all datasets.

Results. Table 2 presents the short-term prediction results, with a selection of datasets due to space constraints. The complete results can be found in Table 11 and Table 12 in Appendix E (arXiv). As we can observe from Table 2, UniST consistently outperforms all baselines across all datasets. Compared with the best baseline of each dataset, it showcases a notable average improvement. Notably, time series approaches such as PatchTST and iTransformer exhibit inferior performance compared to spatio-temporal methods. This underscores the importance of incorporating spatial dependency as prior knowledge for spatio-temporal prediction tasks. Another observation is that PatchTST(one-for-all) performs worse than PatchTST dedicated for each dataset, suggesting that the model struggles to directly adept to these distinct data distributions. Moreover, baseline approaches exhibit inconsistent performance across diverse datasets, indicating their instability across scenarios. The consistent superior performance of UniST across all scenarios underscores the significant potential and benefits of a one-for-all model. Moreover, it demonstrates UniST's capability to orchestrate diverse data, where different datasets can benefit each other.

4.3 Long-Term Prediction

Setups. Here we extend the input step and prediction horizon to 64 following [23, 41]. This configuration accommodates prolonged temporal dependencies, allowing us to gauge the model's proficiency in capturing extended patterns over time. Similar to the short-term prediction, UniST is directly evaluated across all datasets, while specific models are individually trained for each baseline on respective datasets.



(a) Performance on Few-shot (1%) and Zero-shot Scenarios

(b) Performance on Few-shot (5%) and Zero-shot Scenarios

Figure 4: (a) Few-shot performance of UniST and baselines on Crowd and BikeNYC datasets using only 1% of the training data. (b) Few-shot performance of UniST and baselines using only 5% of the training data. The Dashed red lines denote the zero-shot performance of UniST.

Results. Table 3 shows the long-term prediction results. Even with a more extended prediction horizon, UniST still consistently outperforms all baseline approaches across all datasets. Compared with the best baseline of each dataset, it yields an average improvement of 10.1%. This highlights UniST's capability to comprehend temporal patterns effectively and its robustness in generalizing across extended durations. Table 13 in Appendix E (arXiv) illustrates the complete results.

4.4 Few-Shot Prediction

Setups. The hallmark of large foundation models lies in their exceptional generalization ability. The few-shot and zero-shot evaluations are commonly employed to characterize the ultimate tasks for universal time series forecasting [66, 74]. Likewise, the few-shot and zero-shot prediction capability is crucial for a universal spatio-temporal model. In this section, we assess the few-shot learning performance of UniST. Each dataset is partitioned into three segments: training data, validation data, and test data. In few-shot learning scenarios, when confronted with an unseen dataset during the training process, we utilized a restricted amount of training data, specifically, *1%, 5%, 10%* of the training data. We choose some baselines with relatively good performance for the few-shot setting evaluation, We also compare with meta-learning baselines, *i.e.,* MAML and MetaST, and pretraining and finetuning-based time series method, *i.e.,* PatchTST.

Results. Appendix Table 14 to 16 (arXiv) illustrate the overall few-shot results. Due to the space limit, Figure 4 only illustrates the 1% few-shot learning results on two datasets. In these cases, UniST still outperforms all baselines, it achieves a larger relative improvement

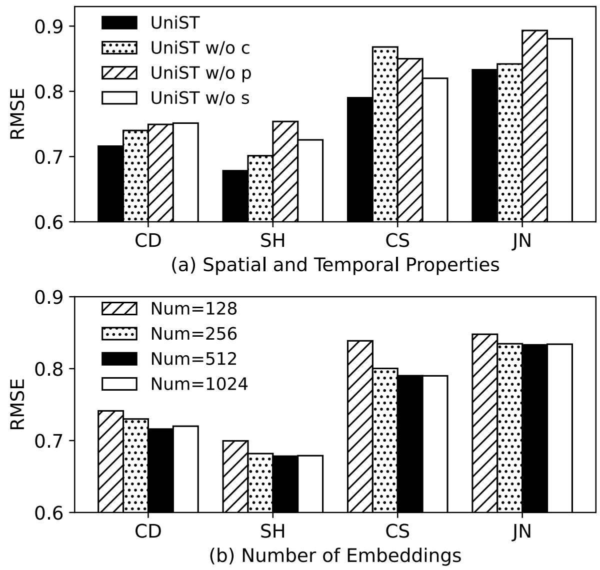


Figure 5: Ablation studies on four traffic speed datasets: Chengdu (CD), Shanghai (SH), Changsha (CS), and Jinan (JN). (a) illustrates the results of removing a prompt guided by one type of spatio-temporal knowledge. (b) presents the results of varying the number of learnable embeddings in the temporal and spatial memory pools.

over baselines compared to long-term and short-term predictions. The transferability can be attributed to successful knowledge transfer in our spatio-temporal prompt.

4.5 Zero-Shot Prediction

Setups. Zero-shot inference serves as the ultimate task for assessing a model's adaptation ability. In this context, after training on a diverse collection of datasets, we evaluate UniST on an entirely novel dataset—*i.e., without any prior training data from it.* The test data used in this scenario aligns with that of normal prediction and few-shot prediction.

Results. Figure 4 also compares the performance of UniST (zero-shot) and baselines (few-shot). As observed, UniST achieves remarkable zero-shot performance, even surpassing many baselines trained with training data that are highlighted by red dashed lines. We attribute these surprising results to the powerful spatio-temporal transfer capability. It suggests that for a completely new scenario, even when the displayed overall patterns are dissimilar to the data encountered during the training process, UniST can extract fine-grained similar patterns from our defined spatial and temporal properties. The few-shot and zero-shot results demonstrate the powerful generalization capability of UniST.

5 STUDY AND ANALYSIS ON UNIST

5.1 Ablation Study

The prompts play an essential role in our UniST model. Here we investigate whether the designed spatial and temporal properties contribute to the overall performance. We use 's' to denote spatial closeness and hierarchy, 'p' for temporal periodicity, and 'c' for temporal closeness. we compare the overall design that incorporates all three properties with three degraded versions that individually remove 's', 'p', or 'c'. Figure 5(a) shows the results on four traffic speed datasets. As we can observe, removing any property results in a performance decrease. The contributions of each spatial and temporal property vary across different datasets, highlighting the necessity of each property for the spatio-temporal design.

Additionally, we explore how the number of embeddings in the memory pools affects the final performance. As seen in Figure 5(b), increasing the number from 128 to 512 improves performance across the four datasets. When further increasing the number to 1024, the performance remains similar to 512, suggesting that 512 is the optimal choice.

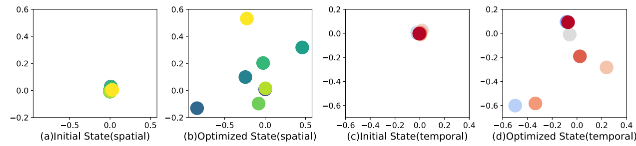


Figure 6: Embeddings visualization of spatial and temporal memory pools at the initial and final optimized states. The embeddings exhibit obvious divergence.

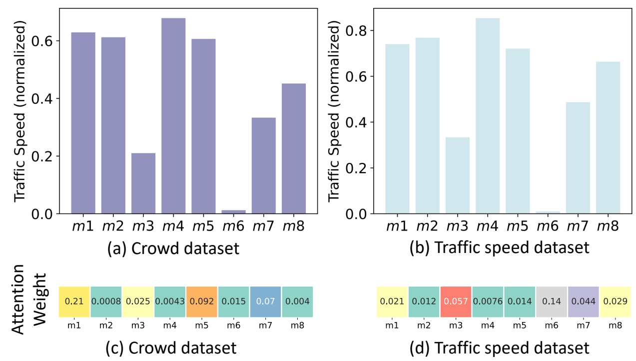


Figure 7: (a) and (b): Comparison of the mean value of inputs in each memory embedding, where the inputs assign the highest attention weight to the memory embedding. (c) and (d): Comparison of the attention weight on each memory embedding for two distinct datasets.

5.2 Prompt Learner

In this section, we conduct in-depth analyses of the prompt learner. To provide a clearer understanding, we leverage t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize the embeddings of both the spatial and temporal memory pools. Specifically, we plot the initial state and the optimized state in Figure 6. Notably, from the start state to the final optimized state, the embeddings gradually become diverged in different directions. This suggests that, throughout the optimization process, the memory pools progressively store and encapsulate personalized information.

Next, we delve into the memorized patterns of each embedding within the temporal memory pool. Specifically, we first select the inputs based on the attention weights. For each embedding, we aggregate the corresponding input spatio-temporal data with the

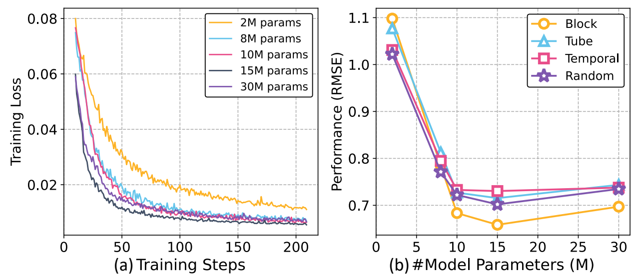


Figure 8: (a) Training loss across five models with varying parameter sizes. (b) Performance evaluation of masked patch reconstruction by increasing parameter sizes.

highest attention weight. Then, we calculate the mean value of the extracted spatio-temporal data. Figure 7(a) and Figure 7(b) illustrate the results for two datasets (Crowd and TrafficSH). As we can see, the memorized patterns revealed in the prompt tool exhibit remarkable consistency across different urban scenarios. This not only affirms that each embedding is meticulously optimized to memorize unique spatio-temporal patterns, but also underscores the robustness of the spatial and temporal memory pools across different scenarios.

Moreover, we examine the extracted spatio-temporal prompts for two distinct domains. Specifically, we calculate the mean attention weight for each embedding in the context of each dataset. Figure 7(c) and Figure 7(d) illustrate the comparison results. As we can observe, the depicted attention weight distributions for the two datasets manifest striking dissimilarities. The observed distinctiveness in attention weight distributions implies a dynamic and responsive nature in the model's ability to tailor its focus based on the characteristics of the input data. The ability to dynamically adjust the attention weights reinforces UniST's versatility and universality for diverse datasets.

5.3 Scalability

Scalability is a crucial characteristic for universal models, therefore, we explore the scaling behavior of our UniST model. Our investigation specifically concentrates on observing changes in training loss and prediction performance as we vary the model parameter size. Figure 8 depicts the training loss and testing RMSE of UniST with varying parameter sizes. Regarding training loss (left figure), several key observations emerge: (i) across different parameter sizes, the training loss consistently decreases and gradually converges with increasing training steps; (ii) increasing the parameter size accelerates the convergence of the training loss; (iii) there exist diminishing marginal returns, suggesting that reducing the training loss becomes progressively harder as parameter size increases. The right figure illustrates the reconstruction RMSE on the testing set, showing similar trends to the training loss.

These observations indicate that UniST has shown scalability behaviors,wherein larger models generally exhibit improved performance. However, unlike large language and vision models [2, 26], the scalability in spatio-temporal prediction shows diminishing marginal returns. This may stem from the relative lack of diversity in spatio-temporal data compared to language or visual datasets.

6 CONCLUSION

In this work, we address an important problem of building a universal model UniST for urban spatio-temporal prediction. By leveraging the diversity of spatio-temporal data from multiple sources, and discerning and aligning underlying shared spatio-temporal patterns across multiple scenarios, UniST demonstrates a powerful capability to predict across all scenarios, particularly in few-shot and zero-shot settings. A promising direction for future work entails the integration of various spatio-temporal data formats, such as grid, sequence, and graph data. Our study inspires future research in spatio-temporal modeling towards the universal direction.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China under grant 2020YFA0711403 and the National Natural Science Foundation of China under 62171260 and 62272260.

REFERENCES

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

[2] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. 2023. Sequential modeling enables scalable learning for large vision models. *arXiv preprint arXiv:2312.00785* (2023).

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[4] Defu Cao, Furong Jia, Sercan O Arik, Tomas Pfister, Yixiang Zheng, Wen Ye, and Yan Liu. 2023. Tempo: Prompt-based generative pre-trained transformer for time series forecasting. *arXiv preprint arXiv:2310.04948* (2023).

[5] Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Yan Ye, Xiang Xinguang, and Wen Gao. 2021. Mau: A motion-aware unit for video prediction and beyond. *Advances in Neural Information Processing Systems* 34 (2021), 26950–26962.

[6] Changlu Chen, Yanbin Liu, Ling Chen, and Chengqi Zhang. 2022. Bidirectional spatial-temporal adaptive transformer for Urban traffic flow forecasting. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[7] Weihuang Chen, Fangfang Wang, and Hongbin Sun. 2021. S2tnet: Spatio-temporal transformer networks for trajectory prediction in autonomous driving. In *Asian Conference on Machine Learning*. PMLR, 454–469.

[8] Jinliang Deng, Xiusi Chen, Renhe Jiang, Xuan Song, and Ivor W Tsang. 2021. St-norm: Spatial and temporal normalization for multi-variate time series forecasting. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*. 269–278.

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

[10] Jie Feng, Yuwei Du, Tianhui Liu, Siqi Guo, Yuming Lin, and Yong Li. 2024. CityGPT: Empowering Urban Spatial Cognition of Large Language Models. *arXiv preprint arXiv:2406.13948* (2024).

[11] Jie Feng, Jun Zhang, Junbo Yan, Xin Zhang, Tianjian Ouyang, Tianhui Liu, Yuwei Du, Siqi Guo, and Yong Li. 2024. CityBench: Evaluating the Capabilities of Large Language Model as World Model. *arXiv preprint arXiv:2406.13945* (2024).

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*. PMLR, 1126–1135.

[13] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. 2022. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3170–3180.

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

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

[16] Jiahui Gong, Jingtao Ding, Fanjin Meng, Guilong Chen, Hong Chen, Shen Zhao, Haisheng Lu, and Yong Li. 2024. A Population-to-individual Tuning Framework for Adapting Pretrained LM to On-device User Intent Prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3637528.3671984
[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*. 1–11.
[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16000–16009.
[19] Jiahao Ji, Jingyuan Wang, Chao Huang, Junjie Wu, Boren Xu, Zhenhe Wu, Zhang Junbo, and Yu Zheng. 2023. Spatio-Temporal Self-Supervised Learning for Traffic Flow Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 4 (2023), 4356–4364.
[20] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision*. Springer, 709–727.
[21] Jiawei Jiang, Chengkai Han, Wayne Xin Zhao, and Jingyuan Wang. 2023. PDFormer: Propagation Delay-aware Dynamic Long-range Transformer for Traffic Flow Prediction. *arXiv preprint arXiv:2301.07945* (2023).
[22] KyoHoon Jin, JeongA Wi, EunJu Lee, ShinJin Kang, SooKyun Kim, and YoungBin Kim. 2021. TrafficBERT: Pre-trained model with large-scale data for long-range traffic flow forecasting. *Expert Systems with Applications* 186 (2021), 115738.
[23] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728* (2023).
[24] Ming Jin, Qingsong Wen, Yuxuan Liang, Chaoli Zhang, Siqiao Xue, Xue Wang, James Zhang, Yi Wang, Haifeng Chen, Xiaoli Li, et al. 2023. Large models for time series and spatio-temporal data: A survey and outlook. *arXiv preprint arXiv:2310.10196* (2023).
[25] Yilun Jin, Kai Chen, and Qiang Yang. 2022. Selective cross-city transfer learning for traffic prediction via source city region re-weighting. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 731–741.
[26] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).
[27] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.
[28] Siqi Lai, Zhao Xu, Weijia Zhang, Hao Liu, and Hui Xiong. 2023. Large Language Models as Traffic Signal Control Agents: Capacity and Opportunity. *arXiv preprint arXiv:2312.16044* (2023).
[29] Ruikun Li, Huandong Wang, and Yong Li. 2023. Learning slow and fast system dynamics via automatic separation of time scales. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4380–4390.
[30] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2018. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *International Conference on Learning Representations*.
[31] Zhonghang Li, Lianghao Xia, Jiabin Tang, Yong Xu, Lei Shi, Long Xia, Dawei Yin, and Chao Huang. 2024. UrbanGPT: Spatio-Temporal Large Language Models. arXiv:2403.00813 [cs.CL]
[32] Zhonghang Li, Lianghao Xia, Yong Xu, and Chao Huang. 2023. Generative Pre-Training of Spatio-Temporal Graph Neural Networks. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=nMH5cUaSj8
[33] Zhihui Lin, Maomao Li, Zhuobin Zheng, Yangyang Cheng, and Chun Yuan. 2020. Self-attention convlstm for spatiotemporal prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 11531–11538.
[34] Lei Liu, Shuo Yu, Runze Wang, Zhenxun Ma, and Yanming Shen. 2024. How Can Large Language Models Understand Spatial-Temporal Data? arXiv:2401.14192 [cs.LG]
[35] Lingbo Liu, Ruimao Zhang, Jiefeng Peng, Guanbin Li, Bowen Du, and Liang Lin. 2018. Attentive crowd flow machines. In *Proceedings of the 26th ACM international conference on Multimedia*. 1553–1561.
[36] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.
[37] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. 2023. itransformer: Inverted transformers are effective for time series forecasting. *arXiv preprint arXiv:2310.06625* (2023).
[38] Yong Liu, Chenyu Li, Jianmin Wang, and Mingsheng Long. 2023. Koopa: Learning Non-stationary Time Series Dynamics with Koopman Predictors. *arXiv preprint arXiv:2305.18803* (2023).
[39] Zhanyu Liu, Guanjie Zheng, and Yanwei Yu. 2023. Cross-city Few-Shot Traffic Forecasting via Traffic Pattern Bank. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 1451–1460.
[40] Bin Lu, Xiaoying Gan, Weinan Zhang, Huaxiu Yao, Luoyi Fu, and Xinbing Wang. 2022. Spatio-Temporal Graph Few-Shot Learning with Cross-City Knowledge Transfer. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1162–1172.
[41] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730* (2022).
[42] Xiaocao Ouyang, Yan Yang, Wei Zhou, Yiling Zhang, Hao Wang, and Wei Huang. 2023. CityTrans: Domain-Adversarial Training with Knowledge Transfer for Spatio-Temporal Prediction across Cities. *IEEE Transactions on Knowledge and Data Engineering* (2023).
[43] Zheyi Pan, Yuxuan Liang, Weifeng Wang, Yong Yu, Yu Zheng, and Junbo Zhang. 2019. Urban traffic prediction from spatio-temporal data using deep meta learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 1720–1730.
[44] Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597* (2022).
[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
[46] Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. 2022. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4454–4458.
[47] Zezhi Shao, Zhao Zhang, Fei Wang, and Yongjun Xu. 2022. Pre-training Enhanced Spatial-temporal Graph Neural Network for Multivariate Time Series Forecasting. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*. ACM, 1567–1577.
[48] Sheng Shen, Shijia Yang, Tianjun Zhang, Bohan Zhai, Joseph E Gonzalez, Kurt Keutzer, and Trevor Darrell. 2024. Multitask vision-language prompt tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 5656–5667.
[49] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. *Advances in neural information processing systems* 28 (2015).
[50] Cheng Tan, Zhangyang Gao, Lirong Wu, Yongjie Xu, Jun Xia, Siyuan Li, and Stan Z Li. 2023. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18770–18782.
[51] Yihong Tang, Ao Qu, Andy HF Chow, William HK Lam, SC Wong, and Wei Ma. 2022. Domain adversarial spatial-temporal network: a transferable framework for short-term traffic forecasting across cities. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 1905–1915.
[52] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
[53] Senzhang Wang, Jiannong Cao, and S Yu Philip. 2020. Deep learning for spatio-temporal data mining: A survey. *IEEE transactions on knowledge and data engineering* 34, 8 (2020), 3681–3700.
[54] Xuhong Wang, Ding Wang, Liang Chen, and Yilun Lin. 2023. Building Transportation Foundation Model via Generative Graph Transformer. arXiv:2305.14826 [cs.LG]
[55] Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, and S Yu Philip. 2018. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *International Conference on Machine Learning*. PMLR, 5123–5132.
[56] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. 2017. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in neural information processing systems* 30 (2017).
[57] Yunbo Wang, Jianjin Zhang, Hongyu Zhu, Mingsheng Long, Jianmin Wang, and Philip S Yu. 2019. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9154–9162.
[58] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. 2023. Detecting everything in the open world: Towards universal object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11433–11443.
[59] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. 2022. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 139–149.
[60] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. 2022. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186* (2022).
[61] Fengli Xu, Jun Zhang, Chen Gao, Jie Feng, and Yong Li. 2023. Urban Generative Intelligence (UGI): A Foundational Platform for Agents in Embodied City

Environment. *arXiv preprint arXiv:2312.11813* (2023).
[62] Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2023. When Urban Region Profiling Meets Large Language Models. *arXiv preprint arXiv:2310.18340* (2023).
[63] Huaxiu Yao, Yiding Liu, Ying Wei, Xianfeng Tang, and Zhenhui Li. 2019. Learning from multiple cities: A meta-learning approach for spatial-temporal prediction. In *The world wide web conference*. 2181–2191.
[64] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. 2020. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 507–523.
[65] Yuan Yuan, Jingtao Ding, Chenyang Shao, Depeng Jin, and Yong Li. 2023. Spatio-temporal Diffusion Point Processes. *arXiv preprint arXiv:2305.12403* (2023).
[66] Yuan Yuan, Chenyang Shao, Jingtao Ding, Depeng Jin, and Yong Li. 2024. Spatio-Temporal Few-Shot Learning via Diffusive Neural Network Generation. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=QyFm3D3Tzi
[67] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.
[68] Xu Zhang, Yongshun Gong, Xinxin Zhang, Xiaoming Wu, Chengqi Zhang, and Xiangjun Dong. 2023. Mask-and Contrast-Enhanced Spatio-Temporal Learning for Urban Flow Prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 3298–3307.
[69] Zijian Zhang, Ze Huang, Zhiwei Hu, Xiangyu Zhao, Wanyu Wang, Zitao Liu, Junbo Zhang, S Joe Qin, and Hongwei Zhao. 2023. MLPST: MLP is All You Need for Spatio-Temporal Prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 3381–3390.
[70] Zijian Zhang, Xiangyu Zhao, Qidong Liu, Chunxu Zhang, Qian Ma, Wanyu Wang, Hongwei Zhao, Yiqi Wang, and Zitao Liu. 2023. PromptST: Prompt-Enhanced Spatio-Temporal Multi-Attribute Prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 3195–3205.
[71] Liang Zhao, Min Gao, and Zongwei Wang. 2022. St-gsp: Spatial-temporal global semantic representation learning for urban flow prediction. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1443–1451.
[72] Ling Zhao, Yujiao Song, Chao Zhang, Yu Liu, Pu Wang, Tao Lin, Min Deng, and Haifeng Li. 2019. T-gcn: A temporal graph convolutional network for traffic prediction. *IEEE transactions on intelligent transportation systems* 21, 9 (2019), 3848–3858.
[73] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang. 2014. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 3 (2014), 1–55.
[74] Tian Zhou, Peisong Niu, Xue Wang, Liang Sun, and Rong Jin. 2023. One Fits All: Power General Time Series Analysis by Pretrained LM. *arXiv preprint arXiv:2302.11939* (2023).
[75] Zhilun Zhou, Jingtao Ding, Yu Liu, Depeng Jin, and Yong Li. 2023. Towards Generative Modeling of Urban Flow through Knowledge-enhanced Denoising Diffusion. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*. 1–12.
[76] Zhengyang Zhou, Kuo Yang, Yuxuan Liang, Binwu Wang, Hongyang Chen, and Yang Wang. 2023. Predicting collective human mobility via countering spatiotemporal heterogeneity. *IEEE Transactions on Mobile Computing* (2023).

APPENDIX

A DATASETS

A.1 Basic Information

Here we provide more details of the used datasets in our study. We collect various spatio-temporal data from multiple cities and domains. Table 4 summarizes the basic information of the used datasets, and Table 5 (arXiv) reports the basic statistics. Specifically, values for Crowd and Cellular datasets in Table 2, Table 3, Table 13 (arXiv), Table 14 (arXiv) and Figure 4 should be scaled by a factor of 10^3.

A.2 Data Preprocessing

For each dataset, We split it into three non-overlapping periods: the first 70% of the period was used as the training set, the next 15% as the validation set, and the final 15% as the test set. To ensure no overlap between train/val/test sets, we removed intermediate sequences. We have normalized all datasets to the range $[-1, 1]$. The reported prediction results are denormalized results.

B BASELINES

- **HA**: History average uses the mean value of historical data for future predictions. Here we use historical data of corresponding periods in the past days.
- **ARIMA**: Auto-regressive Integrated Moving Average model a widely used statistical method for time series forecasting. It is a powerful tool for analyzing and predicting time series data, which are observations collected at regular intervals over time.
- **STResNet** [67]: It is a spatio-temporal model for crowd flow prediction, which utilizes residual neural networks to model the temporal closeness, period, and trend properties.
- **ACFM** [35]: Attentive Crowd Flow Machine model is proposed to predict the dynamics of the crowd flows. It learns the dynamics by leveraging an attention mechanism to adaptively aggregate the sequential patterns and the periodic patterns.
- **STGSP** [71]: This model propose that the global information and positional information in the temporal dimension are important for spatio-temporal prediction. To this end, it leverages a semantic flow encoder to model the temporal relative positional signals. Besides, it utilizes an attention mechanism to capture the multi-scale temporal dependencies.
- **MC-STL** [68]: It leverages an state-of-the-art training techniques for spatio-temporal predition, the mask-enhanced contrastive learning, which can effectively capture the relationships on the spatio-temporal dimension.
- **MAU** [5]: Motion-aware unit is a video prediction model. it broadens the temporal receptive fields of prediction units, which can facilitates to capture inter-frame motion correlations. It consists of an attention module and a fusion module.
- **PredRNN** [56]: PredRNN is a recurrent network-based model. In this model, the memory cells are explicitly decoupled, and they calculate in independent transition manners. Besides, different from the memory cell of LSTM, this network leverages zigzap memory flow, which facilitates to learn at distinct levels.
- **MIM** [57]: Memory utilize the differential information between adjacent recurrent states, which facilitates to model the non-stationary properties. Stacked multiple MIM blocks make it possible to model high-order non-stationarity.
- **SimVP** [13]: It is a simple yet very effective video prediction model. It is completely based on convolutional neural networks and uses MSE loss. It serves as a solid baseline in video prediction tasks.
- **TAU** [50]: Temporal Attention Unit is the state-of-the-art video prediction model. It decomposes the temporal attention into two parts: intra-frame attention and inter-frame attention, which are static and dynamical, respectively. Besides, it introduces a novel regularization, *i.e.,* differential divergence regularization, to consider the impact of inter-frame variations.
- **STID** [46]: It is a MLP-based spatio-temporal prediction model, which is simple yet effective. Its superior performance comes from the identification of the indistinguishability of samples in

Table 4: The basic information of the used datasets.

Dataset	Domain	City	Temporal Duration	Temporal interval	Spatial partition
TaxiBJ	Taxi GPS	Beijing, China	20130601-20131030 20140301-20140630 20150301-20150630 20151101-20160410	Half an hour	32×32
Cellular	Cellular usage	Nanjing, China	20201111-20210531	Half an hour	16 * 20
TaxiNYC-1	Taxi OD	New York City, USA	20160101-20160229	Half an hour	16 * 12
TaxiNYC-2	Taxi OD	New York City, USA	20150101-20150301	Half an hour	20 * 10
BikeNYC-1	Bike usage	New York City, USA	20160801-20160929	One hour	16 * 8
BikeNYC-2	Bike usage	New York City, USA	20160701-20160829	Half an hour	10 * 20
TDrive	Taxi trajectory	New York City, USA	20150201-20160602	One hour	32×32
Crowd	Crowd flow	Nanjing, China	20201111-20210531	Half an hour	16 * 20
TrafficCS	Traffic speed	Changsha, China	20220305-20220405	Five minutes	28×28
TrafficWH	Traffic speed	Wuhan, China	20220305-20220405	Five minutes	30×28
TrafficCD	Traffic speed	Chengdu, China	20220305-20220405	Five minutes	28×26
TrafficJN	Traffic speed	Jinan, China	20220305-20220405	Five minutes	32×18
TrafficNJ	Traffic speed	Nanjing, China	20220305-20220405	Five minutes	32×24
TrafficSH	Traffic speed	Shanghai, China	20220127-20220227	Five minutes	28×32
TrafficSZ	Traffic speed	Shenzhen, China	20220305-20220405	Five minutes	24×18
TrafficGZ	Traffic speed	Guangzhou, China	20220305-20220405	Five minutes	32×26
TrafficGY	Traffic speed	Guiyang, China	20220305-20220405	Five minutes	26×28
TrafficTJ	Traffic speed	Tianjin, China	20220305-20220405	Five minutes	24×30
TrafficHZ	Traffic speed	Hangzhou, China	20220305-20220405	Five minutes	28×24
TrafficZZ	Traffic speed	Zhengzhou, China	20220305-20220405	Five minutes	26×26
TrafficBJ	Traffic speed	Beijing, China	20220305-20220405	Five minutes	30×32

spatio-temporal dimensions. It demonstrates that it is promising to design efficient and effective models in spatio-temporal predictions.

- **STNorm** [8]: It proposed two types of normalization modules: spatial normalization and temporal normalization. These two normalization methods can separately consider high-frequency components and local components.
- **PatchTST** [41]: It first employed patching and self-supervised learning in multivariate time series forecasting. It has two essential designs: (i) segmenting the original time series into patches to capture long-term correlations, (ii) different channels are operated independently, which share the same network.
- **iTransformer** [37]: This is the state-of-the-art multivariate time series model. Different from other Transformer-based methods, it employs the attention and feed-forward operation on an inverted dimension, that is, the multivariate correlation.
- **MAML** [12]: Model-Agnostic Meta-Learning is an state-of-the-art meta learning technique. The main idea is to learn a good initialization from various tasks for the target task.
- **MetaST** [63]: It is an urban transfer learning approach, which utilizes long-period data from multiple cities for transfer learning. by employing a meta-learning approach, it learns a generalized network initialization adaptable to target cities. It also incorporates a pattern-based spatial-temporal memory to capture important patterns.
- **PromptST [70]**: It is the state-of-the-art pre-trianing and prompt-tuning approach for spatio-temporal prediction.