UrbanPG: An Efficient Framework with Personalized Context and General Backbone Interaction for Urban Spatio-Temporal Learning

Aoyu Liu and Yaying Zhang*

The Key Laboratory of Embedded System and Service Computing, Ministry of Education,
Tongji University, Shanghai, China
{liuaoyu, yaying.zhang}@tongji.edu.cn

Abstract

As urban data expands, existing spatio-temporal models encounter challenges such as high context dependency, poor cross-scenario generalization, and inefficient computational performance. To address these issues, we propose UrbanPG, an efficient and scalable framework for spatio-temporal learning. UrbanPG separates task-specific personalized patterns from general patterns, enabling unified spatio-temporal modeling and efficient knowledge generalization across scenarios. The key innovations of UrbanPG include: the development of a lightweight, context-independent general backbone utilizing linear spatio-temporal attention for scalable cross-scenario deployment; a personalized context prompt mechanism designed to model heterogeneity through spatio-temporal embeddings and random perturbation regularization, interacting with the backbone to enhance pattern differentiation; the proposal of multi spatio-temporal learning paradigms for rapid knowledge transfer and generalization to downstream tasks through fine-tuning personalized context prompts while freezing the backbone. Experimental results demonstrate that UrbanPG achieves state-of-the-art performance in large-scale forecasting, few-shot transfer, and continual learning tasks across eight real-world datasets, showcasing exceptional performance, strong generalization, and significant reductions in computational overhead.

Introduction

With the advancement of urban data mining technologies, the analysis of spatio-temporal data, such as traffic and meteorological data, has become essential for spatio-temporal forecasting (Yang et al. 2025; Liu et al. 2025c; Qiu et al. 2024). Accurate forecasting is crucial for the development of smart cities. Urban expansion has generated vast amounts of data, presenting both new opportunities and challenges for spatio-temporal forecasting (Kumar et al. 2024; Liang et al. 2025). While large datasets offer rich potential for model training, they also demand more efficient feature learning methods. Thus, achieving efficient representation learning in large-scale data has become a critical challenge.

In urban spatio-temporal forecasting tasks, deep learning-based spatio-temporal graph neural networks (STGNNs) have emerged as one of the most effective solutions (Dong et al. 2024; Qiu et al. 2025). These models capture temporal and spatial correlations by designing specialized temporal modules (for time periodicity and trends) and spatial modules (for dynamic spatial changes). However, as data scales grow, conventional STGNNs (Kong, Guo, and Liu 2024; Huang et al. 2025) struggle to efficiently handle large-scale spatio-temporal learning scenarios due to limited computational resources. In response, more efficient and lightweight models (Shao et al. 2022a; Fang et al. 2025) have emerged, performing exceptionally well in large-scale scenarios while effectively balancing computational overhead.

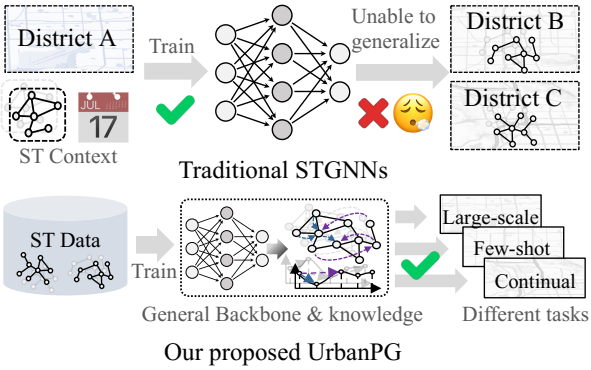

Figure 1: Traditional STGNNs vs. UrbanPG.

Although existing lightweight STGNNs have made significant progress, they are tend designed for specific scenarios. As shown in Fig. 1, different spatio-temporal learning scenarios correspond to distinct spatio-temporal graph structures, making the backbone parameters of existing models highly dependent on the temporal and spatial contexts of the current task, such as the time sampling period and the number of spatial nodes. With the widespread adoption of LLM techniques in various fields, pre-training on large-scale datasets and transferring learned knowledge to new tasks has become a common practice (Liu et al. 2025e; Li et al. 2024a; Liu et al. 2025a). However, the single-scenario modeling of existing methods limits the model's ability to transfer knowledge and generalize across different scenarios. While current spatio-temporal pre-training methods (Zhang et al. 2025; Liu et al. 2025f) can facilitate knowledge transfer and gener-

*Corresponding author.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

alization, they fail to effectively capture the diverse knowledge from large-scale data and struggle to distinguish between personalized and general patterns.

We believe that models trained on large-scale spatio-temporal data should not only perform well in their current scenario but also leverage the diverse knowledge acquired during pre-training to generalize effectively to similar tasks. Therefore, an ideal spatio-temporal learning framework for large-scale scenarios should have the following characteristics: ▷ Efficient and scalable, capable of extracting spatio-temporal correlations in large-scale data; ▷ Spatio-temporal backbone parameters independent of specific spatio-temporal contexts, supporting unified modeling across different scenarios; ▷ The ability to leverage diverse spatio-temporal knowledge from pre-training, distinguishing personalized from general patterns, and enabling effective generalization to downstream tasks.

In light of these challenges, we propose **UrbanPG**, an efficient and scalable urban spatio-temporal learning framework designed to address the complexities of large-scale spatio-temporal learning while enabling unified modeling and pattern generalization across diverse scenarios. UrbanPG combines personalized context prompts and a general backbone, with personalized and general patterns modeled separately yet interactively. Personalized context prompts consist of trainable parameters closely linked to the specific spatio-temporal context of the current task, effectively capturing task-specific patterns. The parameters of the general backbone, independent of the spatio-temporal context, focus on uncovering general patterns. The general backbone models spatio-temporal correlations through a linear context graph attention mechanism, addressing the computational challenges of large-scale spatio-temporal learning. Personalized context prompts interact with the general backbone via a prompting mechanism, integrating the patterns identified by both. To fully leverage the diverse knowledge derived from large-scale data, UrbanPG rapidly adapts to new tasks by freezing the general backbone and fine-tuning the personalized context prompts. Experimental results demonstrate that UrbanPG excels in various forecasting tasks with strong generalization capabilities.

The main contributions of this paper are as follows:

- We propose an interactive framework between personalized context prompts and an efficient, scalable backbone, which effectively reduces computational overhead in large-scale spatio-temporal learning and enables unified modeling across scenarios.
- We introduce paradigms that separate the modeling of personalized and general patterns, allowing the model to fine-tune only the personalized context prompts for spatio-temporal generalization tasks, thereby leveraging diverse spatio-temporal knowledge obtained from pre-training for rapid adaptation to downstream tasks.
- We evaluate UrbanPG on multiple real-world datasets, showing that it achieves low computational overhead, strong generalization capabilities, and state-of-the-art performance in large-scale, few-shot, and continual spatio-temporal forecasting tasks.

Related Work

Urban Spatio-Temporal Learning. Urban spatio-temporal learning focuses on uncovering complex spatio-temporal relationships within urban regions. Recently, STGNNs (Yu, Yin, and Zhu 2018; Liu et al. 2024a; Li et al. 2018) have become the dominant approach for modeling spatio-temporal dependencies and capturing urban dynamics. Early works, such as GWNet (Wu et al. 2019), integrated spatial dependencies and temporal modeling through adaptive adjacency matrices and dilated causal convolutions. STID (Shao et al. 2022a) enhanced accuracy with a lightweight MLP framework by introducing spatial identity and temporal periodicity embeddings, addressing indistinguishable spatio-temporal patterns. To handle non-stationary data, STWave (Fang et al. 2023) used wavelet transforms to decompose data, modeling stable trends and fluctuating events separately with a dual-channel network. In response to the computational challenges posed by massive urban datasets, BigST (Han et al. 2024) designed a long-sequence feature extractor and a linearized global spatial convolution network, reducing complexity to linear through matrix approximation for efficient predictions on large networks. PatchSTG (Fang et al. 2025) improved accuracy and interpretability in large-scale predictions by using a leaf-based KDTree for balanced partitioning of irregular nodes and capturing spatial dependencies with a deep-breadth attention mechanism.

Spatio-Temporal Pre-Training. Pre-training models (Liu and Zhang 2025; Liu et al. 2024b, 2025d) for spatio-temporal forecasting has significantly enhanced model generalization across tasks by learning diverse spatio-temporal patterns from external data. Research has focused on both single-domain optimization and cross-domain transfer. In the single-domain, STEP (Shao et al. 2022b) introduced pre-training in STGNNs with the TSFormer model, which learns segment-level representations from long sequences through masked autoencoding, alleviating computational pressure and improving predictions. Later, STD-MAE (Gao et al. 2024) refined the masking strategy with separate spatial and temporal masks, capturing spatio-temporal heterogeneity. In cross-domain, FlashST (Li et al. 2024b) employed transfer learning through a spatio-temporal prompt tuning framework, improving model adaptability by mitigating domain differences with a distribution mapping mechanism.

Continual Spatio-Temporal Learning. Continual spatio-temporal learning (Miao et al. 2025a; Zhou et al. 2025) focuses on dynamically adapting models to evolving spatio-temporal data and environments, with an emphasis on incremental learning strategies. The field began with the TrafficStream (Chen, Wang, and Xie 2021) framework, which combined spatio-temporal modeling with continual learning, processing streaming traffic data through historical data replay and parameter smoothing for accurate predictions. The STKEC (Wang et al. 2023) framework introduced influence-driven knowledge expansion and memory-augmented mechanisms, adapting to expanding road networks and mitigating catastrophic forgetting. More recently, the EAC (Chen and Liang 2025) framework used prompt

tuning and a dynamic expansion-compression mechanism to fine-tune a small number of parameters, adapting to new nodes while retaining historical knowledge, improving generalization and computational efficiency.

Preliminaries

Definition 1 (Spatio-Temporal Graph). A spatio-temporal graph is denoted as $G = (V, A)$, where $V = \{v_1, v_2, \ldots, v_N\}$ represents the set of spatio-temporal entities (e.g., specific city regions or sensors), and $N = |V|$ denotes the total number of nodes. The matrix $A \in \mathbb{R}^{N \times N}$ encodes the relationships or connectivity between nodes, such as distances or similarities.

Definition 2 (Spatio-Temporal Forecasting). Spatio-temporal forecasting on a graph G aims to predict future data $\mathbf{Y}_{T+1:T+T'} \in \mathbb{R}^{T' \times N \times D}$ for each spatiotemporal entity, using historical data $\mathbf{X}_{:T} \in \mathbb{R}^{T \times N \times D}$ over a specified period. Here, T represents the observation period, T' is the forecast horizon, and D is the feature dimension per node. The prediction can be expressed as:

$$\hat{\mathbf{Y}}_{T+1:T+T'} = \mathcal{F}_\theta(G, \mathbf{X}_{:T}), \tag{1}$$

where the model $\mathcal{F}_\theta$, parameterized by θ, is trained by minimizing:

$$\theta^* = \arg\min_\theta \mathbb{E}_{(G,\mathbf{X},\mathbf{Y})\sim\mathcal{D}} \left[\mathcal{L}\left(\mathcal{F}_\theta(G, \mathbf{X}), \mathbf{Y}\right)\right] \tag{2}$$

with $\mathcal{L}(\cdot, \cdot)$ representing the loss function, and $\mathcal{D}$ denoting the data distribution.

Methodology

As shown in Fig. 2, UrbanPG consists of two components: personalized context prompts and a general backbone. The personalized context prompts are designed to model the unique patterns of the current spatio-temporal scenario, enabling the distinction of spatio-temporal heterogeneity. In contrast, the general backbone serves as a lightweight general pattern extractor, with parameters that are independent of the spatio-temporal context, making it applicable to a wide range of spatio-temporal scenarios. During the training process, the personalized context prompts and the general backbone interact, jointly capturing spatio-temporal correlations. With lightweight architecture, UrbanPG efficiently captures diverse patterns in large-scale spatio-temporal scenarios, facilitating accurate predictions. Furthermore, by freezing the backbone and reconstructing or expanding the personalized context prompts for new scenarios, UrbanPG effectively generalizes to various spatio-temporal tasks, including few-shot learning and continual learning.

Personalized Context Prompts

In spatio-temporal forecasting tasks, factors such as the time sampling period, time position encoding, and the number of spatial nodes are critical for distinguishing different urban scenarios. These factors are essential for capturing the personalized patterns. Drawing inspiration from recent studies (Shao et al. 2022a; Liang et al. 2022), we adaptively mine the personalized context information in the current scenario by mapping trainable parameters to the spatio-temporal context. Specifically, we construct personalized context prompts, which include temporal context prompt $\mathbf{P}_t \in \mathbb{R}^{N \times d}$ and spatial context prompt $\mathbf{P}_s \in \mathbb{R}^{N \times d}$.

Temporal Context. For the time dimension, to preserve the periodicity and trend information in the sequence, we introduce two trainable time embedding parameters: $\mathbf{E}_{tod} \in \mathbb{R}^{f \times d}$ and $\mathbf{E}_{dow} \in \mathbb{R}^{w \times d}$, where f represents the sampling frequency within a day, w is the period (e.g., for a weekly period, $w = 7$), and d is the feature mapping dimension. As shown in Fig. 2, we use the position encoding information from the last time step in the input data $\mathbf{X} \in \mathbb{R}^{T \times N \times D}$ as an index to initialize $\mathbf{E}_{tod}$ and $\mathbf{E}_{dow}$. The temporal context prompt $\mathbf{P}_t \in \mathbb{R}^{N \times d}$ can be obtained by adding the two initialized time embeddings, which can be expressed as:

$$\mathbf{P}_t = \mathrm{index}(\mathbf{E}_{tod}, \mathbf{X}) + \mathrm{index}(\mathbf{E}_{dow}, \mathbf{X}), \tag{3}$$

where index denotes the indexing operation.

Spatial Context. Consistent with recent research (Shao et al. 2022a; Liu et al. 2023a), we believe that modeling spatial heterogeneity in spatio-temporal scenarios significantly enhances model performance. Different spatial nodes may exhibit similar or distinct patterns, and effectively distinguishing these patterns is crucial for improving prediction accuracy. To this end, we introduce spatial embedding $\mathbf{E}_s \in \mathbb{R}^{N \times d}$ to model spatial heterogeneity, enabling the adaptive recognition and distinction of spatial patterns through training. However, as the number of spatial nodes increases in large-scale forecasting tasks, the number of parameters in $\mathbf{E}_s$ grows significantly, which may lead to overfitting and negatively impact pattern distinction.

To mitigate this, we propose a random perturbation-based regularization method. Specifically, we introduce random noise $\mathbf{M}_d^{(n)} \in \mathbb{R}^d$ and trainable shared embedding $\mathbf{E}_d^{(n)} \in \mathbb{R}^d$, and perform random replacement of nodes in the spatial embedding $\mathbf{E}_s \in \mathbb{R}^{N \times d}$ during training. The replacement process employs a combination ratio of p and q, where both p and q are within the range $[0, 1)$, resulting in the spatial context prompt $\mathbf{P}_s \in \mathbb{R}^{N \times d}$. The replacement process is defined as:

$$\mathbf{P}_s^{(n)} = \begin{cases} \mathbf{E}_d^{(n)} & \text{if } 0 \le u_n < p \cdot q, \\ \mathbf{M}_d^{(n)} & \text{if } p \cdot q \le u_n < p, \text{ for } n = 1, 2, \ldots, N, \\ \mathbf{E}_s^{(n)} & \text{if } p \le u_n < 1. \end{cases} \tag{4}$$

where u_n is a random number uniformly distributed in the interval $[0, 1)$, which determines whether node n should be replaced. This random replacement operation effectively disrupts the original spatial pattern distinction, prompting the model to learn more robust spatial relationships.

General Spatio-Temporal Backbone

When processing spatio-temporal data across different scenarios, conventional STGNNs typically require redesigning for each specific scenario. To address this, we propose a general spatio-temporal backbone for modeling general spatio-temporal patterns. The general backbone is designed to be

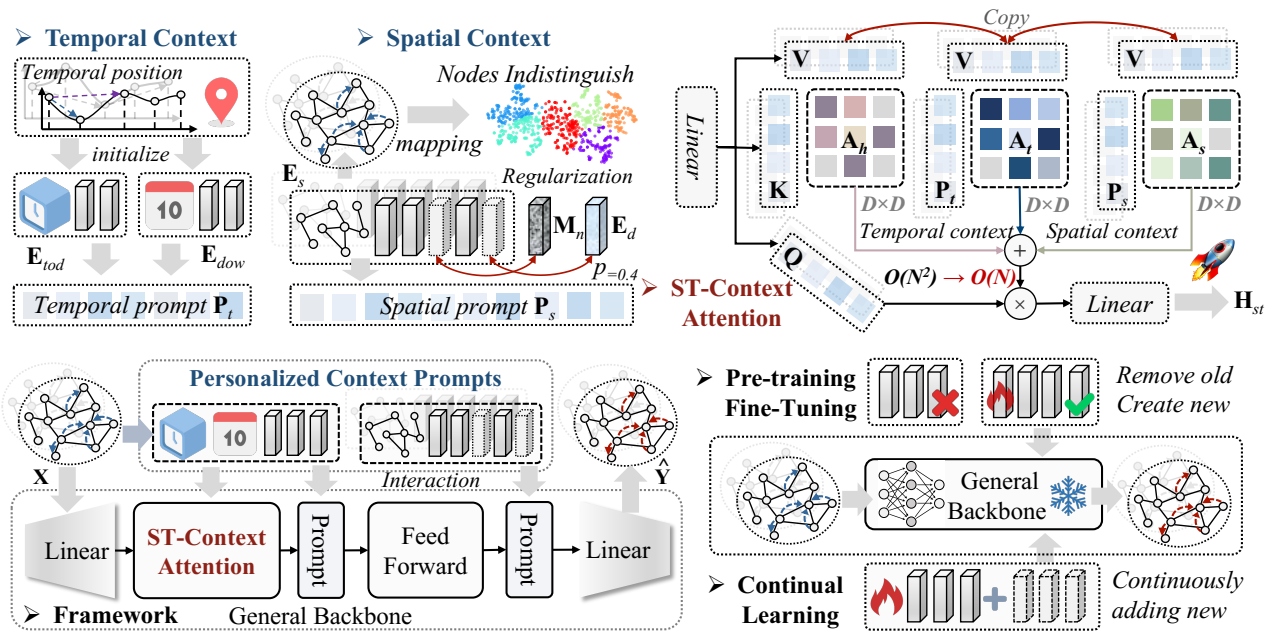

Figure 2: The overall framework of **UrbanPG** consists of **Personalized Context Prompts** and a **General Backbone**. ST refers to Spatio-Temporal. **Pre-training, Fine-Tuning**, and **Continual Learning** represent various learning paradigms extended from the current framework. UrbanPG is designed to handle large-scale, few-shot, and continual forecasting tasks.

lightweight, with parameters that are independent of the spatio-temporal context, allowing UrbanPG to seamlessly adapt to a wide range of spatio-temporal scenarios.

Spatio-Temporal Context Attention. As illustrated in Fig. 2, the input spatio-temporal data $\mathbf{X}$ is initially passed through a linear layer, which maps the temporal features to high-dimensional hidden representations $\mathbf{H} \in \mathbb{R}^{N \times d}$. To mitigate the computational overhead associated with scaling spatio-temporal dimensions, we propose a linear spatio-temporal context attention (STCA) module based on random feature mapping (Katharopoulos et al. 2020; Choromanski et al. 2020; Miao et al. 2024). The STCA module utilizes temporal and space context prompt information $\mathbf{P}_t$ and $\mathbf{P}_s$ as queries, and enhances the fusion and interaction of personalized and general patterns by performing cross-computation on spatio-temporal information. Specifically, the STCA module first maps $\mathbf{H}$ into attention features $\mathbf{Q} \in \mathbb{R}^{N \times d}$, $\mathbf{K} \in \mathbb{R}^{N \times d}$, and $\mathbf{V} \in \mathbb{R}^{N \times d}$, and, by altering the order of the attention computation, it implicitly uncovers dynamic correlations while maintaining linear complexity. The process can be expressed as:

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \text{Linear}(\mathbf{H}), \tag{5}$$

$$\begin{aligned}
\mathbf{H}_{st} &= \text{STCA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{P}_t, \mathbf{P}_s) \\
&= \text{Softmax}(\mathbf{Q}\mathbf{K}^\top + \mathbf{Q}(\mathbf{P}_t)^\top + \mathbf{Q}(\mathbf{P}_s)^\top)\mathbf{V}, \\
&\approx (\phi(\mathbf{Q})\phi(\mathbf{K})^\top + \phi(\mathbf{Q})\phi(\mathbf{P}_t)^\top + \phi(\mathbf{Q})\phi(\mathbf{P}_s)^\top)\mathbf{V} \\
&\approx \phi(\mathbf{Q})\big(\phi(\underbrace{\mathbf{K})^\top\mathbf{V}}_{\mathbf{A}_h \in \mathbb{R}^{d \times d}} + \underbrace{\phi(\mathbf{P}_t)^\top\mathbf{V}}_{\mathbf{A}_t \in \mathbb{R}^{d \times d}} + \underbrace{\phi(\mathbf{P}_s)^\top\mathbf{V}}_{\mathbf{A}_s \in \mathbb{R}^{d \times d}}\big),
\end{aligned} \tag{6}$$

where the function ϕ represents the random feature mapping function, and in practice, we use $\mathrm{Softmax}$ for approximation. STCA adjusts the computation order, transforming the original $N \times N$ query-key matrix into a $d \times d$ key-value matrix, reducing spatial computational complexity from quadratic to linear. In practical applications, d is significantly smaller than N, which makes STCA both efficient and highly scalable, while effectively integrating spatio-temporal context knowledge.

Next, the output representation $\mathbf{H}_{st}$ is further enhanced for non-linear expressiveness through a multilayer perceptron (MLP). Finally, the representation is restored to its original shape through a linear layer, and the prediction results $\hat{\mathbf{Y}} \in \mathbb{R}^{T' \times N \times D}$ are output.

Prompt Adjustment. In addition to participating in the cross-attention computation, $\mathbf{P}_t$ and $\mathbf{P}_s$ also function as auxiliary prompt information, contributing to the gated adjustment of the STCA module and the subsequent feedforward layer outputs. The adjustment process for the output of the STCA module can be expressed as:

$$\mathbf{H}^p_{st} = (\mathbf{H}_{st}(1 + \mathbf{P}_t) + \mathbf{P}_s)\mathbf{P}_t, \tag{7}$$

where $\mathbf{H}^p_{st} \in \mathbb{R}^{N \times d}$ represents the output after prompt adjustment. The personalized context prompt dynamically modulates the influence of spatio-temporal information through the gating mechanism, allowing the model to flexibly and effectively integrate both personalized and general spatio-temporal patterns across diverse urban scenarios.

Multi Spatio-Temporal Learning Paradigms

Given the complexity of spatio-temporal data, a single learning paradigm cannot address all tasks. Some tasks may face data scarcity (Lu et al. 2022; Hu et al. 2024; Liu et al. 2025b), while others require continuous updates as scenarios evolve (Miao et al. 2025b; Yi et al. 2025). Thus, a spatio-temporal learning framework must be flexible enough to adapt to these varying demands. To address this challenge, UrbanPG integrates two learning paradigms to leverage the diverse spatio-temporal knowledge acquired during pre-training: pre-training and fine-tuning (for tasks with limited urban samples, such as few-shot learning) and continual learning (to accommodate incremental learning from urban expansion). Let $\mathcal{M}$ represent the backbone trained on pre-trained datasets. In both paradigms, only the general backbone is frozen, while personalized context prompts are fine-tuned to adapt to downstream tasks.

Pre-Training and Fine-Tuning Paradigm. As illustrated in Fig. 2, in the pre-training and fine-tuning paradigm, we reconstruct the personalized context prompts, $\mathbf{P}'_t$ and $\mathbf{P}'_s$, and inject them into the frozen backbone. By training on downstream data and leveraging the general representations pre-learned by the backbone, the model can quickly adapt to specific tasks and achieve efficient generalization across tasks. The optimization process for this paradigm can be expressed as:

$$\begin{aligned}\psi_t^*, \psi_s^* = \arg\min_{\psi_t, \psi_s} \mathbb{E}_{(G,\mathbf{X},\mathbf{Y})\sim\mathcal{D}} \\ \left[\mathcal{L}\left(\mathcal{M}(G, \mathbf{X}, \mathcal{P}_{\psi_t,\psi_s}(\mathbf{X})), \mathbf{Y}\right)\right],\end{aligned} \tag{8}$$

where $\mathcal{M}$ represents the frozen backbone, and $\mathcal{P}_{\psi_t,\psi_s}(\mathbf{X})$ denotes the personalized context prompts. ψ_t^* and ψ_s^* represent the optimized parameters for reconstructing the personalized context prompts $\mathbf{P}'_t$ and $\mathbf{P}'_s$. This paradigm enables rapid knowledge transfer, especially in few-shot learning tasks, where the performance boost is more significant.

Continual Learning Paradigm. In continual spatio-temporal learning scenarios, urban development results in the continuous expansion of spatial scale, such as an increase in the number of nodes. After the first incremental training phase, as shown in Fig. 2, we freeze the general backbone and expand the spatial context prompts $\mathbf{P}_s^{(\tau)} \in \mathbb{R}^{N^{(\tau)}\times d}$ for the new incremental scenario:

$$\mathbf{P}_s^{(\tau)} = \mathbf{P}_s^{(\tau)-1} \parallel \Delta\mathbf{P}_s^{(\tau)}, \tau > 1, \tag{9}$$

where τ denotes the incremental period, and $\Delta\mathbf{P}_s^{(\tau)} \in \mathbb{R}^{(N^{(\tau)}-N^{(\tau-1)})\times d}$ represents the new parameters introduced in the current phase. The optimization process for continual learning is expressed as:

$$\begin{aligned}\varphi^{(\tau)*} = \arg\min_{\varphi^{(\tau)}} \mathbb{E}_{(G^{(\tau)},\mathbf{X}^{(\tau)},\mathbf{Y}^{(\tau)})\sim\mathcal{D}^{(\tau)}} \\ \left[\mathcal{L}\left(\mathcal{M}(G^{(\tau)}, \mathbf{X}^{(\tau)}, \mathcal{P}_{\varphi^{(\tau)}}(\mathbf{X}^{(\tau)})), \mathbf{Y}^{(\tau)}\right)\right],\end{aligned} \tag{10}$$

where $\varphi^{(\tau)*}$ represents the optimized parameters for reconstructing the personalized context prompts $\mathbf{P}_s^{(\tau)}$. Through this approach, UrbanPG is able to retain and distinguish patterns while incorporating patterns from new nodes, thereby effectively mitigating the catastrophic forgetting problem in incremental learning scenarios.

Experiments

Experimental Settings

Datasets. We evaluate the model's performance on three tasks: large-scale, few-shot, and continual spatio-temporal forecasting using eight datasets. All datasets were split into training, validation, and test sets in a 6:2:2 ratio, with predictions made for the next 12 time steps based on the previous 12. For large-scale spatio-temporal forecasting, we used four regional traffic flow datasets from the LargeST (Liu et al. 2023b): **SD** (716 nodes), **GBA** (2352 nodes), **GLA** (3834 nodes), and **CA** (8600 nodes) from 2019. In the few-shot task, we selected local regions **CA-D3** (480 nodes) and **CA-D5** (211 nodes) from the LargeST, using data from January to February 2019, with the training set consisting of only the first 10% of the data. For continual spatio-temporal forecasting, we used the **PEMS-Stream** traffic dataset (Chen et al. 2001) and the **AIR-Stream** meteorological dataset (Chen and Liang 2025). The PEMS-Stream dataset consists of seven incremental periods (2011-2017), with spatial nodes increasing from 655 to 871. The AIR-Stream dataset spans four incremental periods (2016-2019), with spatial nodes expanding from 1087 to 1202.

Baselines. To conduct a more comprehensive comparison, we selected four categories of baselines for evaluation: Conventional spatio-temporal models: **GWNet** (Wu et al. 2019), **STID** (Shao et al. 2022a), **STWave** (Fang et al. 2023); Large-scale spatio-temporal models: **BigST** (Han et al. 2024), **PatchSTG** (Fang et al. 2025); Spatio-temporal pre-training methods: **STEP** (Shao et al. 2022b), **STD-MAE** (Gao et al. 2024), **FlashST** (Li et al. 2024b); Spatio-temporal continual learning methods: **TrafficStream** (Chen, Wang, and Xie 2021), **STKEC** (Wang et al. 2023), **EAC** (Chen and Liang 2025). More details about each model are provided in the Related Work section.

Implementation Details. All experiments were conducted on machines with an Intel Xeon Gold 5220 CPU, an NVIDIA Tesla V100 GPU (32GB), running Ubuntu 22.04.2 and PyTorch 2.2.1. The batch size was set to 64, with 300 training epochs and early stopping applied. Following previous work (Fang et al. 2025; Chen and Liang 2025), we used three standard evaluation metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), with MAE as the loss function. For all tasks, we repeated the experiments five times and reported the average prediction metrics over 12 time steps. Additional implementation details and the **code** are available in the GitHub repository[1].

Performance Comparison

Large-Scale Forecasting Results. As shown in Table 1, UrbanPG outperforms all state-of-the-art models in large-

[1]https://github.com/Aoyu-Liu/UrbanPG

Model	SD			GBA			GLA			CA		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
GWNet	17.74	29.62	11.88%	20.91	33.41	17.66%	21.20	33.58	13.18%	21.72	34.20	17.40%
STID	17.86	31.00	11.94%	20.22	34.61	15.91%	19.76	34.56	12.41%	18.41	32.00	13.82%
STWave	18.22	30.12	12.20%	20.81	33.77	15.76%	20.96	33.48	12.70%	19.69	31.58	14.58%
BigST	18.80	31.73	12.91%	21.95	35.54	18.50%	22.08	36.00	14.57%	20.32	33.45	15.91%
PatchSTG	16.90	29.27	11.23%	19.50	33.16	14.64%	18.96	32.33	11.44%	17.35	29.79	12.79%
UrbanPG	**16.50**	**28.02**	**10.74%**	**19.04**	**31.98**	**14.61%**	**18.69**	**31.07**	**11.19%**	**17.23**	**29.08**	**12.49%**

Table 1: Comparison of large-scale spatio-temporal forecasting results. **Bold**: best, underline: second best.

Model	CA-D3			CA-D5		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE
STID	22.90	34.93	25.24%	15.08	23.57	22.64%
STWave	21.20	34.08	18.05%	14.42	23.63	18.41%
STEP	20.25	32.64	17.22%	14.04	22.60	18.17%
STD-MAE	20.09	32.37	17.28%	13.93	23.17	17.56%
FlashST	18.91	31.50	16.81%	13.47	22.44	17.29%
UrbanPG	**18.28**	**29.77**	**15.80%**	**12.70**	**21.07**	**16.41%**

Table 2: Comparison of few-shot spatio-temporal forecasting results. **Bold**: best, underline: second best.

Model	PEMS-Stream			AIR-Stream		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE
GWNet	16.62	27.07	25.40%	25.11	39.21	32.16%
STID	15.14	24.41	23.17%	23.27	36.38	27.64%
TrafficStream	14.06	22.90	19.48%	21.42	33.72	26.99%
STKEC	14.07	22.93	19.29%	21.85	34.22	27.35%
EAC	13.49	21.60	19.69%	20.77	32.94	26.77%
UrbanPG	**10.77**	**17.83**	**14.23%**	**19.67**	**31.55**	**24.68%**

Table 3: Comparison of continual spatio-temporal forecasting results. **Bold**: best, underline: second best.

scale spatio-temporal forecasting. Among conventional methods, STID surpasses GWNet and STWave, indicating that addressing spatial heterogeneity alone—without complex modeling—can yield decent performance, though with limited improvements. BigST uses low-rank methods in large-scale models, while PatchSTG employs irregular spatial partitioning and Transformers, which may lead to information loss. In contrast, UrbanPG addresses spatio-temporal heterogeneity with personalized context prompts and a scalable backbone architecture that facilitates interaction between general and personalized patterns. Unlike baseline methods, UrbanPG reduces information loss, directly handles large-scale data, eliminates the need for pre-training and partitioning, and is more user-friendly.

Few-Shot Forecasting Results. For the few-shot spatio-temporal forecasting task, we pre-trained UrbanPG using data from non-target domains of the LargeST dataset (from December 2018) and employed a pre-training and fine-tuning strategy for cross-domain generalization. As shown in Table 2, the pre-training methods, STD-MAE and FlashST, significantly outperform conventional models like STWave, as they effectively utilize knowledge gained from additional training. The results demonstrate that UrbanPG, by separately modeling personalized and general patterns, effectively leverages spatio-temporal knowledge from other regions to enhance downstream task generalization. Even when compared to similar models such as FlashST, UrbanPG shows notable performance improvements, underscoring the effectiveness and simplicity of its pre-training and fine-tuning methodology.

Continual Forecasting Results. Table 3 presents the average prediction results for each incremental period in the continual spatio-temporal forecasting task. Conventional methods, such as GWNet and STID, underperform compared to continual learning models like EAC, as they struggle to mitigate catastrophic forgetting during incremental learning. In contrast, UrbanPG demonstrates superior performance over existing continual learning methods. By freezing the backbone trained during the first incremental period and expanding the spatial context prompts alongside the scenario-parameter-independent backbone, UrbanPG effectively tackles the challenge of increasing spatial scale while alleviating catastrophic forgetting, resulting in superior prediction performance.

Ablation Study

To assess the effectiveness of each module in UrbanPG, we designed several variants for ablation experiments on the SD and CA datasets:

- **w/o TC**: Exclusion of temporal context prompts;
- **w/o SC**: Exclusion of spatial context prompts;
- **w/o RPR**: Exclusion of random perturbation regularization in spatial context prompts;
- **w/o STCA**: Exclusion of the STCA module.

The results of the ablation experiments, shown in Fig. 3, reveal that each module contributes positively to the model's performance. Specifically, the results for **w/o TC** and **w/o SC** demonstrate a significant performance drop when personalized context prompts, particularly spatial context prompt, are removed. This suggests that the model loses the

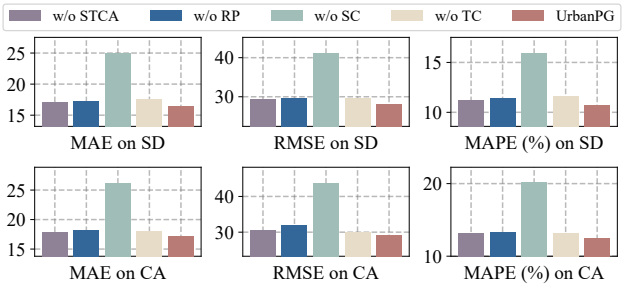

Figure 3: Results of ablation experiments.

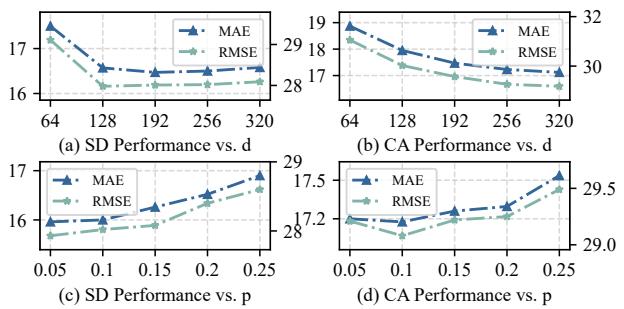

Figure 4: Results of hyperparameter experiments.

ability to distinguish different spatial patterns, making it less effective in handling spatial heterogeneity, a finding consistent with previous research (Shao et al. 2022a). Additionally, the **w/o RPR** results indicate that random perturbation plays a crucial role in reducing overfitting and addressing training issues in spatial context prompts. Finally, the **w/o STCA** results emphasize the importance of the general backbone design and its effective integration with context prompts in improving UrbanPG's performance.

Hyperparameter Study

To evaluate the effect of key hyperparameters on model performance, we investigated the feature mapping dimension d and the perturbation ratio p in the random perturbation regularization of UrbanPG (since q has little impact on model performance) on the SD and CA datasets. The results shown in Fig. 4 demonstrate that increasing d raises the model's parameter count, thereby enhancing its expressive capacity. When $d = 256$, the model achieves the optimal balance between performance and computational cost. Additionally, the experiments indicate that the best perturbation ratio is $p = 0.1$, as higher values of p may introduce excessive noise, impeding model training.

Efficiency Study

To assess the scalability of UrbanPG, we performed an efficiency comparison on large-scale datasets using the same experimental setup (batch size = 1) against baselines. The results shown in Fig. 5 reveal that, as the node scale increases, conventional spatio-temporal models such as GWNet and STWave experience significant increases in training time,

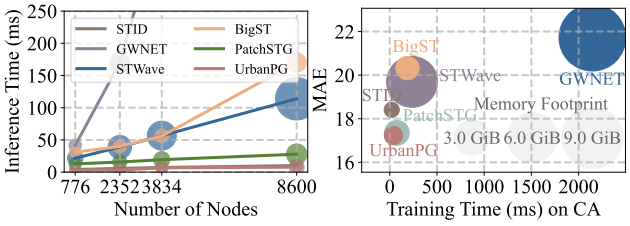

Figure 5: Efficiency comparison.

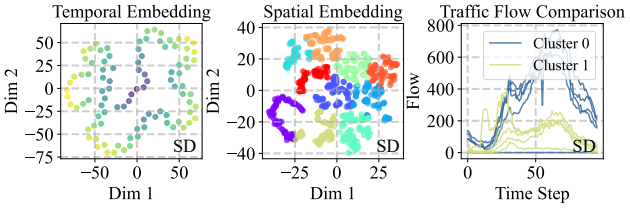

Figure 6: Interpretability of personalized context prompts.

inference time, and memory usage. In contrast, lightweight models like STID and large-scale spatio-temporal models such as BigST, PatchSTG, and UrbanPG effectively manage computational overhead. However, STID underperforms in large-scale spatio-temporal forecasting tasks. On the CA dataset, compared to the second-best model PatchSTG, UrbanPG reduces training time, inference time, and memory usage by 48.96%, 72.44%, and 45.72%, respectively.

Case Study

To enhance the interpretability of the personalized context prompts, we used the t-SNE method to visualize the time embeddings $\mathbf{E}_{tod}$ and spatial embeddings $\mathbf{E}_s$ trained on the SD dataset, as shown in Fig. 6. The visualization of $\mathbf{E}tod$ reveals that the data exhibits periodicity, with adjacent time periods sharing similar features. The visualization of $\mathbf{E}_s$ demonstrates its ability to categorize nodes based on pattern differences, grouping nodes with similar periods and trends into the same cluster. Overall, the personalized context prompts effectively address the issue of spatio-temporal heterogeneity.

Conclusion and Future Work

In this study, we introduced UrbanPG, a framework that effectively tackles key challenges in existing models, such as strong scenario dependency, poor generalization, and high computational overhead. By decoupling the design of personalized context prompts (for modeling spatio-temporal heterogeneity) from the general backbone (for extracting general patterns across scenarios), UrbanPG demonstrates exceptional scalability, robust generalization, and strong compatibility with multiple learning paradigms. However, the potential of UrbanPG to evolve into a spatio-temporal foundational model is limited by its inability to support multi-task parallel training. Moving forward, we plan to extend this work by training spatio-temporal foundational models on multi-source, large-scale spatio-temporal data.

Acknowledgments

This work was partly supported by the National Key Research and Development Program of China under Grant 2022YFB4501704, the National Natural Science Foundation of China under Grant 72342026, and Fundamental Research Funds for the Central Universities under Grant 2024-6-ZD-02.

References

Chen, C.; Petty, K.; Skabardonis, A.; Varaiya, P.; and Jia, Z. 2001. Freeway performance measurement system: mining loop detector data. *Transportation research record*, 1748(1): 96–102.

Chen, W.; and Liang, Y. 2025. Expand and Compress: Exploring Tuning Principles for Continual Spatio-Temporal Graph Forecasting. In *The Thirteenth International Conference on Learning Representations*.

Chen, X.; Wang, J.; and Xie, K. 2021. TrafficStream: A Streaming Traffic Flow Forecasting Framework Based on Graph Neural Networks and Continual Learning. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 3620–3626.

Choromanski, K. M.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J. Q.; Mohiuddin, A.; Kaiser, L.; et al. 2020. Rethinking Attention with Performers. In *International Conference on Learning Representations*.

Dong, Z.; Jiang, R.; Gao, H.; Liu, H.; Deng, J.; Wen, Q.; and Song, X. 2024. Heterogeneity-informed meta-parameter learning for spatiotemporal time series forecasting. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, 631–641.

Fang, Y.; Liang, Y.; Hui, B.; Shao, Z.; Deng, L.; Liu, X.; Jiang, X.; and Zheng, K. 2025. Efficient Large-Scale Traffic Forecasting with Transformers: A Spatial Data Management Perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 307–317.

Fang, Y.; Qin, Y.; Luo, H.; Zhao, F.; Xu, B.; Zeng, L.; and Wang, C. 2023. When spatio-temporal meet wavelets: Disentangled traffic forecasting via efficient spectral graph attention networks. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, 517–529. IEEE.

Gao, H.; Jiang, R.; Dong, Z.; Deng, J.; Ma, Y.; and Song, X. 2024. Spatial-Temporal-Decoupled Masked Pre-training for Spatiotemporal Forecasting. In Larson, K., ed., *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 3998–4006.

Han, J.; Zhang, W.; Liu, H.; Tao, T.; Tan, N.; and Xiong, H. 2024. Bigst: Linear complexity spatio-temporal graph neural network for traffic forecasting on large-scale road networks. *Proceedings of the VLDB Endowment*, 17(5): 1081–1090.

Hu, J.; Liu, X.; Fan, Z.; Yin, Y.; Xiang, S.; Ramasamy, S.; and Zimmermann, R. 2024. Prompt-Based Spatio-Temporal Graph Transfer Learning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, 890–899.

Huang, Y.; Mao, X.; Guo, S.; Chen, Y.; Shen, J.; Li, T.; Lin, Y.; and Wan, H. 2025. STD-PLM: Understanding Both Spatial and Temporal Properties of Spatial-Temporal Data with PLM. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(11): 11817–11825.

Katharopoulos, A.; Vyas, A.; Pappas, N.; and Fleuret, F. 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning (ICML)*, 5156–5165. PMLR.

Kong, W.; Guo, Z.; and Liu, Y. 2024. Spatio-Temporal Pivotal Graph Neural Networks for Traffic Flow Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8): 8627–8635.

Kumar, R.; Bhanu, M.; Mendes-Moreira, J.; and Chandra, J. 2024. Spatio-Temporal Predictive Modeling Techniques for Different Domains: a Survey. *ACM Computing Surveys*, 57(2): 1–42.

Li, Y.; Yu, R.; Shahabi, C.; and Liu, Y. 2018. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *International Conference on Learning Representations (ICLR)*.

Li, Z.; Xia, L.; Tang, J.; Xu, Y.; Shi, L.; Xia, L.; Yin, D.; and Huang, C. 2024a. Urbangpt: Spatio-temporal large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5351–5362.

Li, Z.; Xia, L.; Xu, Y.; and Huang, C. 2024b. FlashST: A simple and universal prompt-tuning framework for traffic prediction. In *International Conference on Machine Learning (ICML)*, ICML'24. JMLR.org.

Liang, Y.; Shao, Z.; Wang, F.; Zhang, Z.; Sun, T.; and Xu, Y. 2022. Basicts: An open source fair multivariate time series prediction benchmark. In *International symposium on benchmarking, measuring and optimization*, 87–101. Springer.

Liang, Y.; Wen, H.; Xia, Y.; Jin, M.; Yang, B.; Salim, F.; Wen, Q.; Pan, S.; and Cong, G. 2025. Foundation models for spatio-temporal data science: A tutorial and survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Liu, A.; and Zhang, Y. 2025. CrossST: An Efficient Pre-Training Framework for Cross-District Pattern Generalization in Urban Spatio-Temporal Forecasting . In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, 2935–2948.

Liu, C.; Hettige, K. H.; Xu, Q.; Long, C.; Xiang, S.; Cong, G.; Li, Z.; and Zhao, R. 2025a. ST-LLM+: Graph Enhanced Spatio-Temporal Large Language Models for Traffic Prediction. *IEEE Transactions on Knowledge and Data Engineering*.

Liu, C.; Miao, H.; Xu, Q.; Zhou, S.; Long, C.; Zhao, Y.; Li, Z.; and Zhao, R. 2025b. Efficient Multivariate Time Series Forecasting via Calibrated Language Models with Privileged Knowledge Distillation. In *2025 IEEE 41st Inter-*

national Conference on Data Engineering (ICDE), 3165–3178. IEEE Computer Society.

Liu, C.; Xiao, Z.; Long, C.; Wang, D.; Li, T.; and Jiang, H. 2024a. MVCAR: Multi-view collaborative graph network for private car carbon emission prediction. *IEEE Transactions on Intelligent Transportation Systems*.

Liu, C.; Xiao, Z.; Long, W.; Li, T.; Jiang, H.; and Li, K. 2025c. Vehicle trajectory data processing, analytics, and applications: A survey. *ACM Computing Surveys*, 57(9): 1–36.

Liu, C.; Xu, Q.; Miao, H.; Yang, S.; Zhang, L.; Long, C.; Li, Z.; and Zhao, R. 2025d. TimeCMA: Towards LLM-Empowered Multivariate Time Series Forecasting via Cross-Modality Alignment. volume 39, 18780–18788.

Liu, C.; Yang, S.; Xu, Q.; Li, Z.; Long, C.; Li, Z.; and Zhao, R. 2024b. Spatial-Temporal Large Language Model for Traffic Prediction. In *2024 25th IEEE International Conference on Mobile Data Management (MDM)*, 31–40.

Liu, C.; Zhou, S.; Xu, Q.; Miao, H.; Long, C.; Li, Z.; and Zhao, R. 2025e. Towards Cross-Modality Modeling for Time Series Analytics: A Survey in the LLM Era. In *Proceedings of the 34th International Joint Conference on Artificial Intelligence, IJCAI-25*.

Liu, H.; Dong, Z.; Jiang, R.; Deng, J.; Chen, Q.; and Song, X. 2023a. STAEformer: Spatio-Temporal Adaptive Embedding Makes Vanilla Transformers SOTA for Traffic Forecasting. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM)*, 21–25.

Liu, J.; Yi, Z.; Zhou, Z.; Huang, Q.; Yang, K.; Wang, X.; and Wang, Y. 2025f. SynEVO: A neuro-inspired spatiotemporal evolutional framework for cross-domain adaptation. In *Forty-second International Conference on Machine Learning*.

Liu, X.; Xia, Y.; Liang, Y.; Hu, J.; Wang, Y.; Bai, L.; Huang, C.; Liu, Z.; Hooi, B.; and Zimmermann, R. 2023b. LargeST: A Benchmark Dataset for Large-Scale Traffic Forecasting. In *Advances in Neural Information Processing Systems*.

Lu, B.; Gan, X.; Zhang, W.; Yao, H.; Fu, L.; and Wang, X. 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.

Miao, H.; Liu, Z.; Zhao, Y.; Guo, C.; Yang, B.; Zheng, K.; and Jensen, C. S. 2024. Less is more: Efficient time series dataset condensation via two-fold modal matching. *Proceedings of the VLDB Endowment*, 18(2): 226–238.

Miao, H.; Xu, R.; Zhao, Y.; Wang, S.; Wang, J.; Yu, P. S.; and Jensen, C. S. 2025a. A parameter-efficient federated framework for streaming time series anomaly detection via lightweight adaptation. *IEEE Transactions on Mobile Computing*.

Miao, H.; Zhao, Y.; Guo, C.; Yang, B.; Zheng, K.; and Jensen, C. S. 2025b. Spatio-Temporal Prediction on Streaming Data: A Unified Federated Continuous Learning Framework. *IEEE Transactions on Knowledge and Data Engineering*.

Qiu, X.; Hu, J.; Zhou, L.; Wu, X.; Du, J.; Zhang, B.; Guo, C.; Zhou, A.; Jensen, C. S.; Sheng, Z.; and Yang, B. 2024. TFB: Towards Comprehensive and Fair Benchmarking of Time Series Forecasting Methods. In *Proc. VLDB Endow.*, 2363–2377.

Qiu, X.; Wu, X.; Lin, Y.; Guo, C.; Hu, J.; and Yang, B. 2025. DUET: Dual Clustering Enhanced Multivariate Time Series Forecasting. In *SIGKDD*, 1185–1196.

Shao, Z.; Zhang, Z.; Wang, F.; Wei, W.; and Xu, Y. 2022a. 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.

Shao, Z.; Zhang, Z.; Wang, F.; and Xu, Y. 2022b. Pre-training enhanced spatial-temporal graph neural network for multivariate time series forecasting. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 1567–1577.

Wang, B.; Zhang, Y.; Shi, J.; Wang, P.; Wang, X.; Bai, L.; and Wang, Y. 2023. Knowledge expansion and consolidation for continual traffic prediction with expanding graphs. *IEEE Transactions on Intelligent Transportation Systems*, 24(7): 7190–7201.

Wu, Z.; Pan, S.; Long, G.; Jiang, J.; and Zhang, C. 2019. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 1907–1913.

Yang, B.; Liang, Y.; Guo, C.; and Jensen, C. S. 2025. Data Driven Decision Making with Time Series and Spatio-Temporal Data. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, 4517–4522. IEEE Computer Society.

Yi, Z.; Zhou, Z.; Huang, Q.; Chen, Y.; Yu, L.; Wang, X.; and Wang, Y. 2025. Get rid of isolation: a continuous multi-task spatio-temporal learning framework. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24. Red Hook, NY, USA: Curran Associates Inc. ISBN 9798331314385.

Yu, B.; Yin, H.; and Zhu, Z. 2018. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*.

Zhang, Y.; Wang, X.; Yu, X.; Sun, Z.; Wang, K.; and Wang, Y. 2025. Drawing Informative Gradients from Sources: A One-stage Transfer Learning Framework for Cross-city Spatiotemporal Forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(1): 1147–1155.

Zhou, Z.; Huang, Q.; Wang, B.; Hou, J.; Yang, K.; Liang, Y.; Zheng, Y.; and Wang, Y. 2025. Coms2t: A complementary spatiotemporal learning system for data-adaptive model evolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.