A GENERAL SPATIO-TEMPORAL BACKBONE WITH SCALABLE CONTEXTUAL PATTERN BANK FOR URBAN CONTINUAL FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

With the explosive growth of spatio-temporal data driven by IoT deployments and urban infrastructure expansion, accurate and efficient continual forecasting remains a critical challenge. Recent Spatio-Temporal Graph Neural Networks assume static graph topologies and temporal scales, making them ill-suited for dynamic real-world data streams. Meanwhile, existing continual learning methods often adopt simple backbones, limiting their ability to capture evolving dependencies and adapt to distributional drift. To address these limitations, we propose `STBP`, a novel framework for Continual Spatio-Temporal Forecasting that bridges the gap between STGNNs and continual learning. `STBP` integrates a general-purpose spatio-temporal backbone with a scalable contextual pattern bank. The backbone extracts stable spatio-temporal representations in the frequency domain and models dynamic spatial correlations using linear graph attention. To support continual adaptation and alleviate catastrophic forgetting, the contextual pattern bank is incrementally updated via parameter expansion, capturing evolving node-level heterogeneous patterns. During incremental training, the backbone remains frozen to preserve general knowledge, while the contextual pattern bank adapts to new scenarios and distributions. Extensive experiments show that `STBP` surpasses state-of-the-art baselines in both accuracy and scalability, underscoring its effectiveness for continual spatio-temporal forecasting. Code is available at Anonymous Github.

1 INTRODUCTION

With the rapid development of urban infrastructure (Kumar et al., 2024; Hu et al., 2023) and the widespread deployment of IoT sensing devices (Jin et al., 2024; Yang et al., 2025), spatio-temporal data—such as traffic flow (Shao et al., 2022b) and weather observations (Tian et al., 2025)—have grown explosively. Efficient and accurate forecasting of such large-scale, continuously evolving spatio-temporal data has become one of the key tasks in the development of smart cities.

However, urban spatio-temporal data inherently form a dynamic system: as the urban area expands, the spatial topology evolves, sensors are continuously added, and data distributions drift over time. These dynamic characteristics bring new challenges to recent spatio-temporal forecasting methods, such as Spatio-Temporal Graph Neural Networks (STGNNs) (Kong et al., 2024; Gao et al., 2024; Liu & Zhang, 2025), which have achieved significant progress in modeling spatio-temporal correlations. However, as illustrated in Figure 1, most existing methods are based on static assumptions—i.e., fixed temporal scales and static graph topologies—making them ill-suited for real-world data streams that evolve continuously. More critically, recent STGNNs rely on offline training; when encountering new data or topology changes, they often require retraining from scratch, which is impractical in resource-constrained or continuously growing environments.

To tackle these issues, Continual Spatio-Temporal Forecasting (CSTF) (Miao et al., 2024; Chen & Liang, 2025; Ma et al., 2025b) has emerged as a research hotspot. Its core goal is to achieve incremental learning and efficient forecasting on new data without retraining on old data. As shown in Figure 1, these methods typically construct a general spatio-temporal backbone and adopt strategies such as regularization, replay, and dynamic architectures to enhance adaptability and mitigate catastrophic forgetting. However, most existing methods mainly focus on retaining old knowledge and adopt

relatively simple spatio-temporal backbones, overlooking the ability to model dynamic spatio-temporal characteristics and adapt to distributional drift, thus limiting forecasting performance.

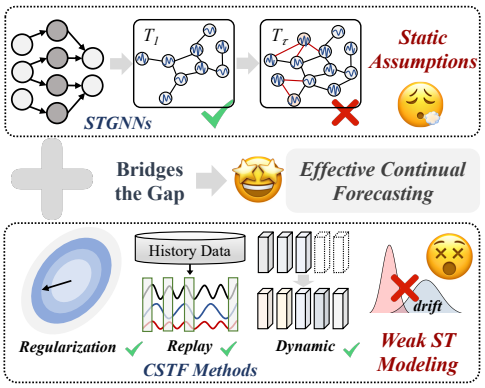

Figure 1: Limitations of existing studies.

An intuitive solution is to combine high-performing STGNNs with continual learning strategies to balance modeling capacity and adaptability. However, in practice, once the assumption of fixed topology is removed, the original spatio-temporal modeling ability of STGNNs degrades significantly (Shao et al., 2024; Ma et al., 2025a). Moreover, most STGNNs lack designs for incremental training, making them hard to scale efficiently and hindering the real-world deployment of continual learning strategies. Therefore, an ideal CSTF framework should simultaneously address the following four key challenges: ❶ *handling distributional drift*, ❷ *modeling dynamic spatio-temporal correlations*, ❸ *alleviating catastrophic forgetting*, and ❹ *supporting dynamic expansion of graph structures*.

To this end, we bridge the gap between STGNNs and continual learning by introducing a general-purpose **s**patio-**t**emporal **b**ackbone with scalable contextual **p**attern bank (`STBP`). Specifically, the backbone in `STBP` extracts stable spatio-temporal components in the frequency domain to mitigate distributional drift; meanwhile, a lightweight, scene-agnostic, data-driven linear graph attention is used to model dynamic spatial correlations with minimal computational overhead. To alleviate catastrophic forgetting and support the continual expansion of graph structures, the contextual pattern bank—composed of trainable parameters—is incrementally updated through parameter expansion to adapt to evolving scenarios. In this framework, the backbone models general and stable patterns, whereas the contextual pattern bank captures contextual and node-specific heterogeneous patterns that interact with the backbone to adapt to continuously evolving environments.

Our main contributions are summarized as follows: ❶ We propose a highly general and efficient backbone tailored for incremental forecasting. ❷ We introduce a contextual pattern-based optimization strategy that supports dynamic adaptation and mitigates catastrophic forgetting. ❸ Extensive experiments on multiple real-world benchmark datasets demonstrate that `STBP` significantly outperforms state-of-the-art baselines in terms of forecasting accuracy, adaptability, and scalability.

2 Related Work

Spatio-Temporal Forecasting. Early studies in spatio-temporal forecasting, including methods like STGCN (Yu et al., 2018) and DCRNN (Li et al., 2018), primarily focused on combining basic temporal and spatial elements for prediction tasks. These models typically depended on predefined geographic adjacency matrices, which limited their ability to capture the evolving nature of spatial correlations. In contrast, later advancements, such as GWNet (Wu et al., 2019), DGCRN (Li et al., 2023), and MegaCRN (Jiang et al., 2023b), addressed this limitation by incorporating adaptive adjacency matrices or learning spatial correlations directly from the data. This shift led to a notable improvement in forecasting accuracy. More recently, models like STID (Shao et al., 2022a), STAEformer (Liu et al., 2023a), and HimNet (Dong et al., 2024) have emphasized the significance of distinguishing spatial patterns to further enhance forecasting performance. These methods incorporate trainable components, including spatial embeddings, parameter pools, and contextual pattern bank, to more accurately capture spatial variations, boosting both prediction precision and model adaptability.

Continual Spatio-Temporal Forecasting. TrafficStream (Chen et al., 2021), one of the pioneering frameworks in CSTF, was instrumental in combining spatio-temporal modeling with continual learning. It utilized techniques such as historical data replay and parameter smoothing to effectively manage long-term streaming traffic data, delivering accurate traffic flow predictions. Following this, the STKEC (Wang et al., 2023a) introduced an influence-based knowledge expansion strategy along with a memory-augmented knowledge consolidation mechanism, which better supported the scaling of transportation networks while alleviating issues of catastrophic forgetting. The EAC (Chen & Liang, 2025) further advanced CSTF by incorporating prompt tuning, which enabled continual

spatio-temporal learning with a minimal number of trainable parameters. Its dynamic prompt pool, which allows for both "expansion" and "compression," enhances the model's adaptability to new nodes while preserving past knowledge, improving generalization and computational efficiency. Additionally, the UFCL (Miao et al., 2025) leveraged federated learning to protect data privacy and introduced a global replay buffer for synthetic spatio-temporal data, addressing challenges associated with distributed streaming environments.

3 PRELIMINARY

Definition 1 (Streaming Spatio-Temporal Graph). We define a streaming spatio-temporal graph as a sequence of evolving graphs $\mathbb{G} = \{G_\tau\}_{\tau=1}^{\mathcal{T}}$, where each graph $G_\tau = (V_\tau, E_\tau, A_\tau)$ represents the graph at incremental period τ. Here, V_τ denotes the node set, E_τ the edge set, and adjacency matrix $A_\tau \in \mathbb{R}^{N_\tau \times N_\tau}$ connections between nodes. The number of nodes at period τ is denoted by $N_\tau = |V_\tau|$. The graph evolves incrementally as $G_\tau = G_{\tau-1} + \Delta G_\tau$, where ΔG_τ captures structural or feature modifications between periods.

Definition 2 (Continual Spatio-Temporal Forecasting). Continual spatio-temporal forecasting aims to develop an optimal predictive model at each stage based on dynamic, streaming spatio-temporal graph data. At each incremental period τ, given the current graph G_τ and historical observations $\mathbf{X}_\tau \in \mathbb{R}^{N_\tau \times T_h}$, the goal is to predict future signals $\mathbf{Y}_\tau \in \mathbb{R}^{N_\tau \times T_f}$ as follows:

$$\hat{\mathbf{Y}}_\tau = f_\theta(G_\tau, \mathbf{X}_\tau), \tag{1}$$

where T_h is the length of the historical observation window, and T_f is the forecasting horizon. The model f_θ is parameterized by θ, and continually updated by minimizing:

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

where $\mathcal{L}(\cdot, \cdot)$ is a loss function, and $\mathcal{D}_\tau$ denotes the data distribution at period τ.

4 METHODOLOGY

4.1 OVERVIEW OF STBP

The workflow and architecture of STBP are shown in Figure 2. It consists of two core components: a general spatio-temporal backbone and a contextual pattern bank. The backbone, comprising temporal and spatial modules with a prediction layer, captures spatio-temporal correlations in evolving networks. The contextual pattern bank, made of trainable parameters, is dynamically expanded and fine-tuned as data evolves. While the backbone captures general, stable patterns, the contextual pattern bank adapts to environmental changes, focusing on context-specific patterns. Guided by prompts, both components collaborate to form an efficient and robust continual learning system.

In terms of workflow, streaming spatio-temporal data is sequentially fed into the STBP. During the initial incremental training phase, the backbone and contextual pattern bank are jointly trained to capture spatio-temporal correlations from current data. In later stages, the backbone is frozen (denoted by a snowflake) to retain knowledge learned from historical data, while the contextual pattern bank is updated (denoted by a flame) through expansion and fine-tuning. These updates serve as prompts, guiding the frozen backbone to adapt to new data distributions. This continual learning process, driven by the interplay between backbone and contextual pattern bank, enables the model to progressively enhance its representation power and adaptability while preserving core functionality. For detailed workflow steps, refer to Algorithm 1 in Appendix A.3.2.

4.2 CONTEXTUAL PATTERN BANK

Recent studies (Shao et al., 2022a; Dong et al., 2024; Chen & Liang, 2025) have shown that incorporating node-specific trainable parameters into STGNNs can significantly enhance forecasting performance. Following this insight, we propose an expandable `contextual pattern bank` $\mathbf{P}_\tau \in \mathbb{R}^{N\tau \times d}$, composed of trainable parameters, to consolidate historical spatio-temporal patterns and generalize to new ones, thereby mitigating *catastrophic forgetting* and continuously adapting to new incremental scenarios, where d denotes the feature dimension.

We posit that the model can utilize $\mathbf{P}_\tau$ to effectively distinguish both the *relevance* and *heterogeneity* of nodes, enabling a more nuanced understanding of the underlying data structures. Here,

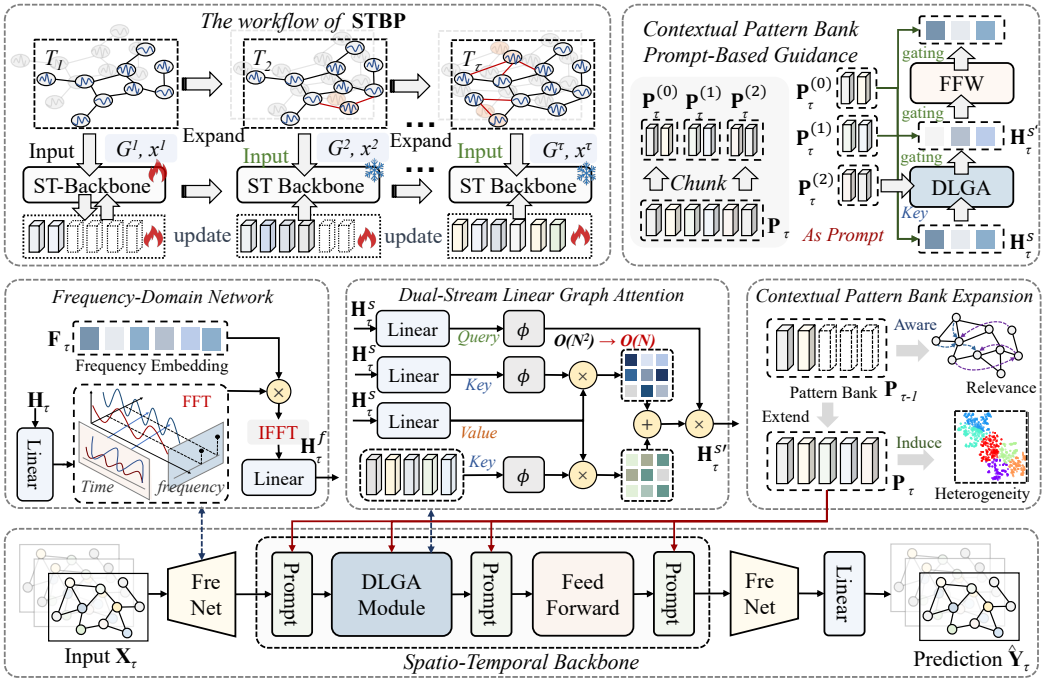

Figure 2: The overall workflow and architecture of STBP.

relevance refers to shared behavioral patterns among nodes—such as similar trends or periodic fluctuations—while *heterogeneity* captures differences arising from distinct node functions or external factors such as geography, policy, or events. To validate this hypothesis, we conduct a t-SNE-based analysis on $\mathbf{P}_\tau$ trained on spatio-temporal datasets (see Figure 3), which reveals meaningful clustering patterns. Each cluster exhibits distinct characteristics, corresponding to *heterogeneity*, while nodes within the same cluster display similar temporal dynamics, reflecting *relevance*.

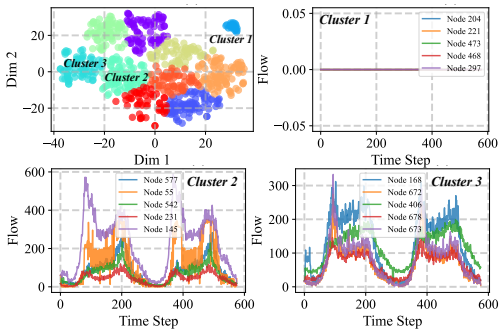

Figure 3: Contextual pattern bank visualization.

As shown in Figure 2, given an stream spatio-temporal input $\mathbf{X}_\tau \in \mathbb{R}^{N_\tau \times T_h}$, the backbone model $\mathcal{M}_\theta$, and contextual pattern bank $\mathbf{P}_\tau \in \mathbb{R}^{N_\tau \times d}$, the incremental learning process is formulated as:

$$\hat{\mathbf{Y}}_\tau = \mathcal{M}_\theta(\mathbf{X}_\tau, \mathbf{P}_\tau). \tag{3}$$

At the initial training stage ($\tau = 1$), both the backbone and contextual pattern bank are jointly trained (denoted with flame). For subsequent stages ($\tau > 1$), the backbone is frozen (denoted with snowflake), and only the contextual pattern bank is updated through expansion:

$$\mathbf{P}'_\tau = \mathbf{P}_{\tau-1} \parallel \Delta\mathbf{P}_\tau, \tag{4}$$

where $\Delta\mathbf{P}_\tau \in \mathbb{R}^{(N_\tau - N_{\tau-1}) \times d}$ represents newly introduced parameters for the current incremental period. Only the expanded contextual pattern bank $\mathbf{P}'_\tau \in \mathbb{R}^{N_\tau \times (d)}$ is fine-tuned during training. This strategy ensures that the backbone retains previously acquired knowledge, while the contextual pattern bank continually adapts to evolving distributions. It incrementally expands to represent an increasingly diverse set of environmental patterns, thereby avoiding the inadequacy exhibited by fixed models in novel scenarios.

Distinct from existing work (Wang et al., 2023a; Chen & Liang, 2025; Wang et al., 2023b), we introduce a *Prompt-Based Guidance*(Peebles & Xie, 2023; Zhang et al., 2023) mechanism to enhance $\mathbf{P}_\tau$'s capacity to model both node-level relevance and heterogeneity. Specifically, the contextual pattern bank comprises three groups of trainable parameters: $\mathbf{P}_\tau^{(i)} \in \mathbb{R}^{N_\tau \times d}$ for $i \in 0, 1, 2$. As

illustrated in Figure2, these components interact with the backbone's hidden representation $\mathbf{H}_\tau$ via the following prompt-based gating function:

$$\mathbf{H}'_\tau = \mathbf{P}_\tau^{(0)} \cdot h_\theta(\mathbf{H}_\tau \cdot \mathbf{P}_\tau^{(1)}), \tag{5}$$

where h_θ denotes an arbitrary submodule within the backbone. This gating mechanism enables adaptive modeling of node heterogeneity. Additionally, $\mathbf{P}_\tau^{(2)}$ acts as a key embedding in the attention module, guiding the backbone to generalize correlation-aware information under task constraints. Importantly, since the contextual pattern bank encodes high-level abstractions rather than raw historical data, our method supports knowledge retention without revisiting prior data—offering advantages in *privacy protection* and *storage efficiency*.

4.3 General Spatio-Temporal Backbone

While the contextual pattern bank mitigates catastrophic forgetting in continual learning, it lacks the ability to model dynamic spatio-temporal correlations and handle out-of-distribution generalization. To address this, we design a **general spatio-temporal backbone** aimed at handling distributional drift, spatio-temporal correlation modeling, and graph scalability during continual learning. The term *general* implies that the backbone is independent of the number of nodes and does not rely on any predefined adjacency matrix, making it adaptable to arbitrary spatio-temporal data structures.

As shown in Figure 2, the backbone operates as follows: input spatio-temporal data first pass through a **frequency-domain network** (`FreNet`), which maps it into high-dimensional temporal representations and extracts stable components via frequency domain analysis. A **dual-stream linear graph attention** (`DLGA`) module then captures dynamic spatial correlations, followed by a feedforward layer with a multilayer perceptron for enhanced nonlinear expressivity. Finally, the features are reconstructed to their original shape by another `FreNet` and passed through a prediction layer. We detail the `FreNet` and `DLGA` modules below.

Frequency-Domain Network. Spatio-temporal data in evolving environments often suffer from distributional drift (Wang et al., 2024; Ji et al., 2025; Zhou et al., 2023). Although the contextual pattern bank helps retain stable knowledge, we further address this issue through a dedicated frequency-domain analysis (Xia et al., 2023). `FreNet` is designed to capture temporal correlations while emphasizing stable components in the data, such as periodicity and trends, which are more resilient to distributional changes (Liu & Zhang, 2025). Specifically, `STBP` employs two `FreNets`—one at the beginning and one at the end of the backbone (Figure 2). The first maps input data $\mathbf{X}_\tau \in \mathbb{R}^{N_\tau \times T_h}$ through a linear layer into a high-dimensional representation $\mathbf{H}_\tau \in \mathbb{R}^{N_\tau \times d}$, which is then transformed to the frequency domain using a Fast Fourier Transform (FFT). A learnable frequency-domain embedding $\mathbf{F}_\tau \in \mathbb{C}^{(\frac{d}{2}+1)}$ adaptively highlights stable features. This process is formalized as:

$$\mathbf{H}_\tau^f = \text{IFFT}(\text{FFT}(\mathbf{H}_\tau) \odot \mathbf{F}_\tau), \tag{6}$$

where $\mathbf{H}_\tau^f \in \mathbb{R}^{N_\tau \times d}$ is further processed by a linear layer. The second `FreNet` performs an inverse operation, restoring the feature shape to $\mathbb{R}^{N_\tau \times T_h}$. Compared to traditional temporal modules like RNNs (Li et al., 2018; Bai et al., 2020) or TCNs (Zheng et al., 2023; Fang et al., 2023), `FreNet` offers higher computational efficiency and improved ability to extract stable components, thereby alleviating the impact of distributional drift.

Dual-Stream Linear Graph Attention. After obtaining stable components, it remains essential to capture complex spatial interactions and time-varying node correlations. An effective spatial module must adaptively learn node correlations in a data-driven manner, maintain computational efficiency, and scale to growing graphs. Graph attention mechanisms (Veličković et al., 2018) have emerged as promising solutions, enabling dynamic correlation modeling without relying on fixed adjacency matrices. However, conventional graph attention (Zheng et al., 2020; Jiang et al., 2023a; Liu et al., 2023a) incurs $O(N^2)$ complexity, limiting its scalability. To overcome this, we propose `DLGA` (Figure 2), which improves efficiency using a *random feature mapping*-based linear attention mechanism (Katharopoulos et al., 2020). Moreover, `DLGA` introduces a **dual-stream structure** by incorporating the contextual pattern bank $\mathbf{P}_\tau^{(2)} \in \mathbb{R}^{N_\tau \times d}$ as an additional *key*. This enables the model to assess the relationship between evolving input patterns and stored knowledge. Formally:

$$\mathbf{Q} = \mathbf{W}_q \mathbf{H}_\tau^s, \quad \mathbf{K} = \mathbf{W}_k \mathbf{H}_\tau^s, \quad \mathbf{V} = \mathbf{W}_v \mathbf{H}_\tau^s, \tag{7}$$

$$\begin{aligned}\mathbf{H}_\tau^{s'} &= \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{P}_\tau^{(2)}) \\ &= \text{Softmax}(\mathbf{Q}\mathbf{K}^\top + \mathbf{Q}(\mathbf{P}_\tau^{(2)})^\top)\mathbf{V},\end{aligned} \tag{8}$$

$$\begin{aligned}\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{P}_\tau^{(2)}) &\approx (\phi(\mathbf{Q})\phi(\mathbf{K})^\top + \phi(\mathbf{Q})\phi(\mathbf{P}_\tau^{(2)})^\top)\mathbf{V} \\ &= \phi(\mathbf{Q})\left(\phi(\mathbf{K})^\top\mathbf{V} + \phi(\mathbf{P}_\tau^{(2)})^\top\mathbf{V}\right).\end{aligned} \tag{9}$$

Here, $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ are trainable projection matrices. $\mathbf{H}_\tau^s$ and $\mathbf{H}_\tau^{s'} \in \mathbb{R}^{N_\tau \times d}$ denote the input and the spatially enriched representation passed to the feedforward layer of the DLGA module, respectively. The function $\phi(\cdot)$ denotes a random feature mapping, with Softmax used for approximation in our implementation. For further details on the approximation derivation, see Appendix A.3.1. Notably, the linear attention approximation does not explicitly construct an adjacency matrix. Instead, it implicitly models dynamic correlations by reordering operations in the attention computation. `DLGA` reduces computational complexity from quadratic to linear, while preserving dynamic spatial modeling and seamlessly integrating prompt-based knowledge from the contextual pattern bank.

5 Experiment

5.1 Experimental Settings

Datasets. We evaluate our model on three real-world streaming spatio-temporal datasets from the traffic and meteorology domains. The traffic datasets, **PEMS-Stream** (Chen et al., 2001) and **CA-Stream** (Liu et al., 2023b), consist of traffic flow measurements provided by the California Department of Transportation (CalTrans) [1], with a sampling interval of 5 minutes. The meteorological dataset, **Air-Stream** (Chen & Liang, 2025), is derived from urban air quality platform of the Chinese Environmental Monitoring Center [2], with hourly sampling intervals. To ensure fair evaluation, all datasets are split into training, validation, and test sets using a fixed ratio of 6:2:2. For each prediction task, the model is trained to forecast the next 12 time steps based on the previous 12 observations. Detailed dataset statistics are provided in Table 3 in Appendix A.4.1.

Baselines and Metrics. We select representative models from two categories as baselines: ▷ `Conventional spatio-temporal forecasting models`, including lightweight spatio-temporal architectures such as **GWNet** (Wu et al., 2019), **STID** (Shao et al., 2022a), and **iTransformer** (Liu et al., 2024). These models are adapted specifically for incremental training in our experiments. ▷ `Continual spatio-temporal forecasting models`, including **TrafficStream**, **STKEC** (Wang et al., 2023a), and **EAC** (Chen & Liang, 2025). The performance of all models is evaluated using the following metrics: Mean Absolute Error (**MAE**), Root Mean Squared Error (**RMSE**), and Mean Absolute Percentage Error (**MAPE**). More details on this are included in Appendix A.4.2.

5.2 Main Results

Overall results. The main experimental results are summarized in Table 1, which lists the average metrics for all incremental periods, including the averages for the 3rd, 6th, and 12th time steps, as well as the overall average. In terms of overall performance, our proposed `STBP` outperforms all other models. Compared to the best baseline, `STBP` improves the average MAE by **18.46%**, **16.69%**, and **5.1%** on the PEMS-Stream, CA-Stream, and Air-Stream datasets, respectively. This improvement is attributed to the bridge established between STGNNs and CSTF methods, where the carefully designed general spatio-temporal backbone and contextual pattern bank effectively capture dynamic spatio-temporal correlations, mitigating catastrophic forgetting and addressing distributional drift.

Results of conventional methods. STGNNs, including GWNet and STID, rely on static graph assumptions and are not designed for continual learning tasks. Consequently, a new spatio-temporal backbone is trained for each data stage, with each model trained only on current stage data for prediction. In contrast, iTransformer is trained on the complete node data of the current spatio-temporal graph and initializes with weights from the previous period, allowing fine-tuning across the entire model. As shown in Table 1, STGNNs trained from scratch show mediocre performance at each dataset. While these methods perform well under static assumptions, they fail to leverage past

[1] https://dot.ca.gov/programs/traffic-operations/mpr/pems-source

[2] https://air.cnemc.cn:18007/

Table 1: Main experimental results. **Bold**: best, underline: second best.

Dataset	Metric	Horizon	GWNet	STID	iTransformer	TrafficStream	STKEC	EAC	STBP
PEMS-Stream	MAE	3	$16.60_{\pm 0.24}$	$13.71_{\pm 0.17}$	$13.59_{\pm 0.23}$	$12.76_{\pm 0.12}$	$12.79_{\pm 0.05}$	$12.67_{\pm 0.12}$	$\mathbf{10.70}_{\pm 0.05}$
		6	$16.53_{\pm 0.16}$	$14.93_{\pm 0.10}$	$14.49_{\pm 0.38}$	$13.89_{\pm 0.10}$	$13.90_{\pm 0.04}$	$13.41_{\pm 0.15}$	$\mathbf{10.99}_{\pm 0.04}$
		12	$16.77_{\pm 0.25}$	$17.45_{\pm 0.12}$	$16.29_{\pm 0.70}$	$16.18_{\pm 0.13}$	$16.14_{\pm 0.05}$	$14.79_{\pm 0.20}$	$\mathbf{11.45}_{\pm 0.04}$
		Avg.	$16.62_{\pm 0.14}$	$15.14_{\pm 0.12}$	$14.62_{\pm 0.40}$	$14.06_{\pm 0.10}$	$14.07_{\pm 0.04}$	$13.49_{\pm 0.15}$	$\mathbf{11.00}_{\pm 0.04}$
	RMSE	3	$26.99_{\pm 0.39}$	$21.89_{\pm 0.24}$	$21.38_{\pm 0.41}$	$20.59_{\pm 0.16}$	$20.65_{\pm 0.05}$	$20.16_{\pm 0.13}$	$\mathbf{17.51}_{\pm 0.07}$
		6	$26.95_{\pm 0.21}$	$24.11_{\pm 0.18}$	$23.10_{\pm 0.73}$	$22.66_{\pm 0.18}$	$22.69_{\pm 0.06}$	$21.52_{\pm 0.19}$	$\mathbf{18.15}_{\pm 0.06}$
		12	$27.37_{\pm 0.35}$	$28.44_{\pm 0.30}$	$26.36_{\pm 1.35}$	$26.61_{\pm 0.25}$	$26.58_{\pm 0.06}$	$23.85_{\pm 0.29}$	$\mathbf{19.10}_{\pm 0.08}$
		Avg.	$27.07_{\pm 0.20}$	$24.41_{\pm 0.20}$	$23.29_{\pm 0.77}$	$22.90_{\pm 0.18}$	$22.93_{\pm 0.05}$	$21.60_{\pm 0.19}$	$\mathbf{18.15}_{\pm 0.06}$
	MAPE (%)	3	$25.46_{\pm 0.64}$	$21.41_{\pm 2.36}$	$30.88_{\pm 1.31}$	$17.57_{\pm 0.42}$	$17.70_{\pm 0.38}$	$18.67_{\pm 0.95}$	$\mathbf{14.22}_{\pm 0.08}$
		6	$25.22_{\pm 0.82}$	$22.83_{\pm 2.32}$	$32.22_{\pm 1.54}$	$19.14_{\pm 0.35}$	$18.97_{\pm 0.39}$	$19.58_{\pm 0.90}$	$\mathbf{14.50}_{\pm 0.11}$
		12	$25.63_{\pm 1.38}$	$26.12_{\pm 2.26}$	$35.87_{\pm 1.96}$	$22.67_{\pm 0.61}$	$22.00_{\pm 0.51}$	$21.27_{\pm 0.92}$	$\mathbf{14.98}_{\pm 0.10}$
		Avg.	$25.40_{\pm 0.87}$	$23.17_{\pm 2.32}$	$32.65_{\pm 1.53}$	$19.48_{\pm 0.41}$	$19.29_{\pm 0.42}$	$19.69_{\pm 0.90}$	$\mathbf{14.52}_{\pm 0.09}$
CA-Stream	MAE	3	$19.74_{\pm 1.05}$	$18.89_{\pm 0.34}$	$17.32_{\pm 0.44}$	$16.33_{\pm 0.14}$	$16.35_{\pm 0.17}$	$16.90_{\pm 0.30}$	$\mathbf{14.46}_{\pm 0.06}$
		6	$19.96_{\pm 0.62}$	$21.43_{\pm 0.53}$	$18.50_{\pm 0.44}$	$17.79_{\pm 0.12}$	$17.81_{\pm 0.17}$	$17.76_{\pm 0.21}$	$\mathbf{14.92}_{\pm 0.05}$
		12	$20.60_{\pm 0.31}$	$26.17_{\pm 0.73}$	$20.78_{\pm 0.72}$	$20.72_{\pm 0.12}$	$20.73_{\pm 0.17}$	$19.62_{\pm 0.22}$	$\mathbf{15.61}_{\pm 0.05}$
		Avg.	$20.05_{\pm 0.57}$	$21.74_{\pm 0.52}$	$18.65_{\pm 0.46}$	$18.01_{\pm 0.13}$	$18.03_{\pm 0.17}$	$17.91_{\pm 0.24}$	$\mathbf{14.92}_{\pm 0.05}$
	RMSE	3	$31.28_{\pm 1.93}$	$28.65_{\pm 0.37}$	$27.10_{\pm 0.64}$	$25.87_{\pm 0.16}$	$25.91_{\pm 0.27}$	$26.09_{\pm 0.32}$	$\mathbf{23.35}_{\pm 0.10}$
		6	$31.66_{\pm 1.20}$	$32.26_{\pm 0.60}$	$29.06_{\pm 0.69}$	$28.28_{\pm 0.13}$	$28.37_{\pm 0.30}$	$27.58_{\pm 0.22}$	$\mathbf{24.18}_{\pm 0.10}$
		12	$32.65_{\pm 0.48}$	$39.07_{\pm 0.88}$	$32.83_{\pm 1.20}$	$32.92_{\pm 0.15}$	$33.03_{\pm 0.31}$	$30.37_{\pm 0.26}$	$\mathbf{25.36}_{\pm 0.12}$
		Avg.	$31.78_{\pm 1.11}$	$32.72_{\pm 0.60}$	$29.31_{\pm 0.73}$	$28.59_{\pm 0.14}$	$28.67_{\pm 0.29}$	$27.73_{\pm 0.25}$	$\mathbf{24.17}_{\pm 0.09}$
	MAPE (%)	3	$20.03_{\pm 0.78}$	$19.56_{\pm 1.17}$	$18.07_{\pm 0.58}$	$15.77_{\pm 0.12}$	$15.69_{\pm 0.37}$	$16.63_{\pm 0.29}$	$\mathbf{14.05}_{\pm 0.47}$
		6	$20.24_{\pm 0.50}$	$21.85_{\pm 1.17}$	$19.21_{\pm 0.50}$	$16.96_{\pm 0.11}$	$16.90_{\pm 0.33}$	$17.30_{\pm 0.22}$	$\mathbf{14.39}_{\pm 0.40}$
		12	$20.87_{\pm 0.91}$	$26.68_{\pm 1.15}$	$21.74_{\pm 0.32}$	$19.60_{\pm 0.25}$	$19.51_{\pm 0.30}$	$19.06_{\pm 0.23}$	$\mathbf{14.96}_{\pm 0.39}$
		Avg.	$20.33_{\pm 0.49}$	$22.28_{\pm 1.16}$	$19.45_{\pm 0.45}$	$17.20_{\pm 0.13}$	$17.13_{\pm 0.33}$	$17.49_{\pm 0.24}$	$\mathbf{14.41}_{\pm 0.42}$
Air-Stream	MAE	3	$23.51_{\pm 0.88}$	$20.94_{\pm 1.31}$	$19.18_{\pm 0.43}$	$18.71_{\pm 0.46}$	$19.26_{\pm 0.32}$	$18.03_{\pm 0.38}$	$\mathbf{16.71}_{\pm 0.25}$
		6	$25.20_{\pm 0.60}$	$23.42_{\pm 0.97}$	$21.94_{\pm 0.33}$	$21.66_{\pm 0.45}$	$22.06_{\pm 0.37}$	$20.99_{\pm 0.24}$	$\mathbf{20.03}_{\pm 0.28}$
		12	$27.25_{\pm 0.39}$	$26.42_{\pm 0.77}$	$25.02_{\pm 0.26}$	$24.91_{\pm 0.42}$	$25.14_{\pm 0.44}$	$24.27_{\pm 0.22}$	$\mathbf{23.58}_{\pm 0.32}$
		Avg.	$25.11_{\pm 0.64}$	$23.27_{\pm 0.96}$	$21.71_{\pm 0.34}$	$21.42_{\pm 0.44}$	$21.85_{\pm 0.36}$	$20.77_{\pm 0.24}$	$\mathbf{19.71}_{\pm 0.24}$
	RMSE	3	$36.60_{\pm 1.22}$	$32.20_{\pm 1.86}$	$29.89_{\pm 0.71}$	$29.01_{\pm 0.69}$	$29.65_{\pm 0.56}$	$28.34_{\pm 0.59}$	$\mathbf{26.93}_{\pm 0.44}$
		6	$39.47_{\pm 0.82}$	$36.89_{\pm 1.44}$	$34.58_{\pm 0.49}$	$34.38_{\pm 0.64}$	$34.84_{\pm 0.70}$	$33.56_{\pm 0.33}$	$\mathbf{32.83}_{\pm 0.51}$
		12	$42.69_{\pm 0.49}$	$41.92_{\pm 1.21}$	$39.37_{\pm 0.34}$	$39.74_{\pm 0.56}$	$40.03_{\pm 0.91}$	$38.78_{\pm 0.42}$	$\mathbf{38.52}_{\pm 0.57}$
		Avg.	$39.21_{\pm 0.88}$	$36.38_{\pm 1.46}$	$34.01_{\pm 0.54}$	$33.72_{\pm 0.64}$	$34.22_{\pm 0.70}$	$32.94_{\pm 0.34}$	$\mathbf{32.02}_{\pm 0.47}$
	MAPE (%)	3	$29.72_{\pm 1.08}$	$24.57_{\pm 1.11}$	$24.37_{\pm 0.58}$	$23.21_{\pm 0.74}$	$23.73_{\pm 0.38}$	$22.75_{\pm 0.42}$	$\mathbf{21.10}_{\pm 0.17}$
		6	$32.22_{\pm 0.72}$	$27.75_{\pm 0.87}$	$28.37_{\pm 0.44}$	$27.09_{\pm 0.75}$	$27.41_{\pm 0.34}$	$26.93_{\pm 0.44}$	$\mathbf{25.13}_{\pm 0.23}$
		12	$35.46_{\pm 0.48}$	$31.80_{\pm 0.62}$	$33.39_{\pm 0.40}$	$31.95_{\pm 0.73}$	$32.06_{\pm 0.33}$	$32.02_{\pm 0.46}$	$\mathbf{29.85}_{\pm 0.32}$
		Avg.	$32.16_{\pm 0.78}$	$27.64_{\pm 0.84}$	$28.25_{\pm 0.48}$	$26.99_{\pm 0.73}$	$27.35_{\pm 0.34}$	$26.77_{\pm 0.39}$	$\mathbf{24.86}_{\pm 0.20}$

spatio-temporal knowledge, resulting in suboptimal performance. In contrast, iTransformer performs better by utilizing historical spatio-temporal knowledge through online training, though it still suffers from catastrophic forgetting, making it a less optimal solution.

Results of CSTF methods. The best-performing models are those that can effectively address catastrophic forgetting, such as CSTF models, TrafficStream, STKEC, and EAC. It is worth noting that despite EAC adopting parameter expansion and fine-tuning strategies, it performs poorly due to neglecting the specific design of the spatio-temporal backbone. In extreme incremental training scenarios, such as with the CA-Stream dataset, EAC's parameter expansion strategy is less effective than the regularization and memory replay strategies used by TrafficStream and STKEC. More detailed experimental results can be found in Appendix A.4.4.

5.3 Ablation Study & Parameter Sensitivity Analysis

Ablation Study Settings. To validate the core contributions of `STBP`, we designed the following variants for ablation experiments: ❶ **Retrain**: The contextual pattern bank is removed. Similar to GWNet and STID, a new backbone is trained for each incremental period using the spatio-temporal graph data of that period, with the corresponding model predicting the results for the current test set. ❷ **Online**: The contextual pattern bank is removed. Similar to iTransformer, the model is trained on the complete node data of the current spatio-temporal graph and initialized with the model from the previous period, allowing for adjustments across the entire model. ❸ **w/o Backbone**: The contextual pattern bank is retained, but the spatio-temporal backbone is replaced with the ones used in TrafficStream, STKEC, and EAC—i.e., replacing `FreNet` and `DLGA` with CNN and GCN. ❹ **w/o `DGLA`**: The `DLGA` module in the spatio-temporal backbone is ablated. ❺ **EAC**: We also included EAC, which follows a similar approach, for comparison in the ablation study.

Ablation findings. The ablation results are shown in Figure 4. The **Retrain** and **Online** results demonstrate that expanding contextual pattern bank parameters and distinguishing and prompting spatio-temporal patterns are crucial for mitigating catastrophic forgetting in continual learning. No-

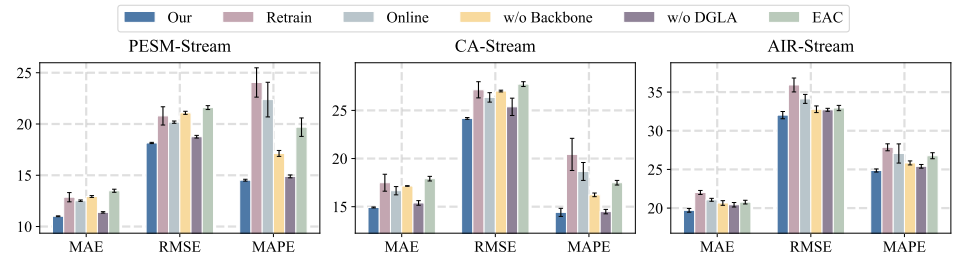

Figure 4: Results of ablation experiments.

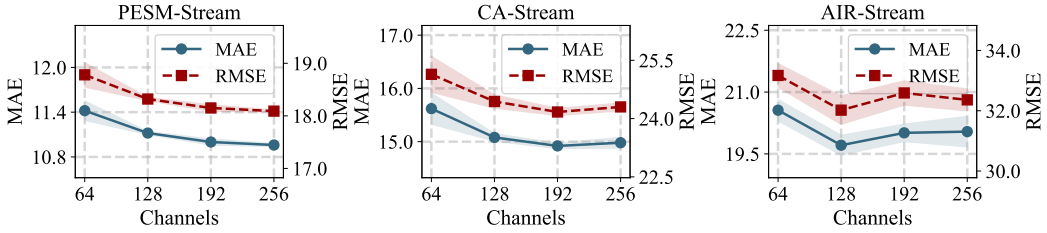

Figure 5: Results of parameter experiments.

tably, even without the contextual pattern bank, the spatio-temporal backbone achieves performance comparable to **EAC** through online training, highlighting the importance of real-time dynamic correlation modeling and mitigating temporal distributional drift in adapting to new incremental periods. The **w/o Backbone** and **w/o DGLA** variants further confirm the indispensability of both the general backbone and the contextual pattern bank. Removing the DLGA module significantly degrades performance, validating its role in capturing dynamic spatial correlations and integrating prompt-based knowledge. Additionally, the FreNet module in backbone improves computational efficiency and enhances the extraction of stable temporal components.

Parameter Sensitivity Analysis. Additionally, we performed a sensitivity analysis on the adjustable hyperparameter d in STBP. In STBP, d represents the feature dimension for each module's feature mapping, as well as the feature dimension of parameters in the contextual pattern bank. The analysis results are shown in Figure 5. Increasing d enhances the model's overall parameter count and improves its expressive power. However, the performance gains from increasing d do not grow indefinitely; after reaching a certain threshold, the performance gain stabilizes. Further increases in d not only fail to improve performance but may also lead to negative effects, causing parameter redundancy. More parameter sensitivity analysis can be found in Appendix A.4.5.

5.4 Case Study

To provide a more intuitive explanation of the contextual pattern bank's distinction and expandability in STBP, we perform dimensionality reduction and clustering analysis on $\mathbf{P}_\tau \in \mathbb{R}^{N_\tau \times d}$ using t-SNE on the PEMS-Stream dataset. Each scatter in Figure 6 represents a node in the spatio-temporal graph. The figure shows that the untrained contextual pattern bank exhibits a random, chaotic distribution, unable to effectively distinguish nodes with different patterns. After multiple incremental training periods, the contextual pattern bank parameters gradually form distinct clusters. By randomly selecting nodes from the same cluster and visualizing their real traffic data, we observe similar patterns, with shared periodic and trend-based characteristics.

In contrast, nodes from different clusters, such as Clusters 1, 2, and 3 in Figure 6, show significantly different patterns. For the 2011 PEMS-Stream dataset, which consists of 655 nodes, when the incremental training reaches 2017, Cluster 1 classifies newly emerging nodes into the current cluster. Nodes 693, 809, and 834, for example, are generalize into Cluster 1 after training on the 2017 data. This demonstrates that the contextual pattern bank, through the fine-tuning of trainable parameters, effectively distinguishes between different patterns and generalizes new ones, continuously adapting to changes. Additional case studies on other datasets can be found in Appendix A.4.6.

5.5 Efficiency Study

An effective CSTF method balances scalability, computational cost, and performance. We compare the efficiency of STBP with baselines under identical settings. Figure 7 shows the average computational

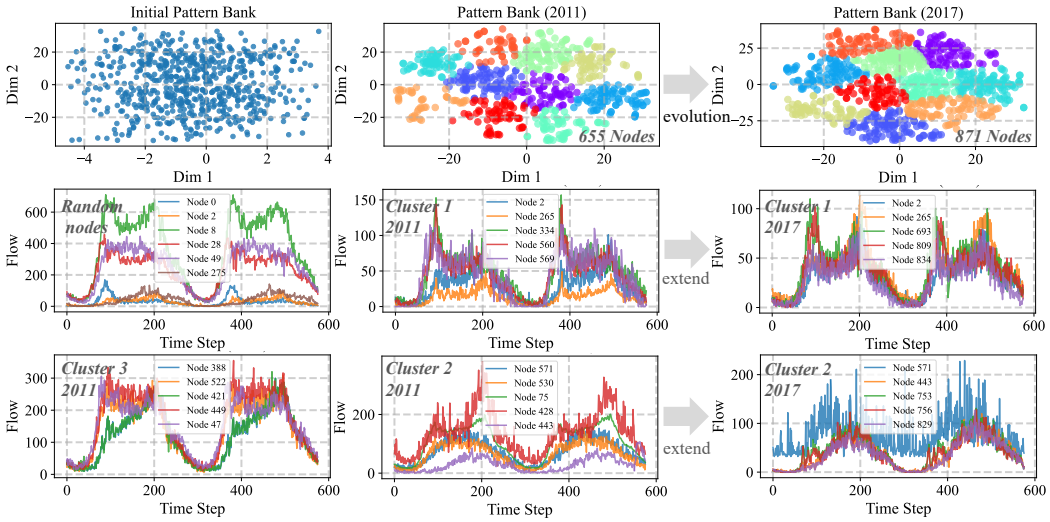

Figure 6: Case study on PEMS-Stream.

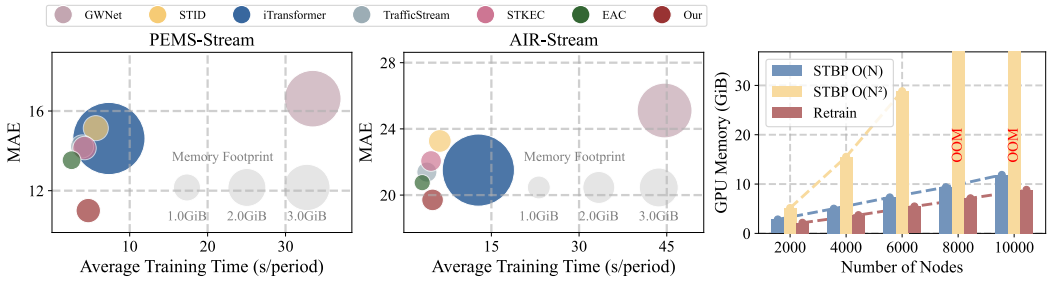

Figure 7: Efficiency comparison.

cost per period across PEMS-Stream and AIR-Stream, with scatter size indicating GPU memory usage. We also evaluate the impact of linear attention, full attention, and the contextual pattern bank's absence on model efficiency using synthetic datasets. Results on PEMS-Stream and AIR-Stream show that non-continual methods, such as GWNet, STID, and iTransformer, require global parameter adjustments at each phase, reducing efficiency. iTransformer, in particular, suffers from high memory consumption due to quadratic attention complexity. Even lightweight non-continual methods struggle with efficiency during incremental training.

In contrast, CSTF methods like EAC, TrafficStream, and STKEC are more efficient, thanks to lightweight backbones and non-global parameter fine-tuning. Despite its complex backbone, `STBP` incurs minimal cost compared to models like EAC, due to optimizations like frequency-domain processing and linear attention. `STBP` achieves substantial performance gains with minimal overhead. Synthetic dataset results further confirm that linear attention reduces cost, and as node count increases, the contextual pattern bank adds minimal burden, demonstrating scalability.

6 Conclusion

In this work, we propose `STBP`, a novel framework for continual spatio-temporal forecasting. By combining a general-purpose backbone with a scalable contextual pattern bank, `STBP` efficiently mitigates catastrophic forgetting while capturing dynamic spatio-temporal correlations. It adapts to evolving urban data without retraining from scratch, making it suitable for real-time applications. Validated on multiple datasets, `STBP` demonstrates strong continual learning capabilities. Nevertheless, `STBP` currently supports continual learning in a single-task setting. In the future, we plan to extend its application to cross-domain continual spatio-temporal forecasting, which will be a crucial step towards developing a foundational spatio-temporal model.

7 Ethics Statement

We affirm that this paper adheres to the ICLR 2026 Code of Ethics. The research presented does not involve human subjects, nor does it involve any harmful insights or methodologies that could negatively impact society. There are no conflicts of interest, sponsorship concerns, or issues with privacy or security. The dataset used is publicly available, and all experiments were conducted under ethical guidelines. The authors confirm that the study complies with legal and ethical standards in its domain, and no ethical issues have been encountered throughout the research process.

8 Reproducibility Statement

This paper is committed to ensuring the reproducibility of its findings. All experiments are conducted on publicly available, real-world datasets, and detailed descriptions of data processing, model architecture, and training procedures are provided in the main text. For reproducibility, the source code for our proposed method, STBP, has been made available at https://anonymous.4open.science/r/STBP/. This repository includes all necessary scripts for running the experiments and reproducing the results presented in the paper. For further clarity, the implementation details, model configurations, and hyperparameter settings are documented within the appendix sections.

References

Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. volume 33, pp. 17804–17815, 2020.

Lucas Caccia, Eugene Belilovsky, Massimo Caccia, and Joelle Pineau. Online learned continual compression with adaptive quantization modules. In *International conference on machine learning*, pp. 1240–1250. PMLR, 2020.

Chao Chen, Karl Petty, Alexander Skabardonis, Pravin Varaiya, and Zhanfeng Jia. Freeway performance measurement system: mining loop detector data. *Transportation research record*, 1748(1): 96–102, 2001.

Wei Chen and Yuxuan Liang. Expand and compress: Exploring tuning principles for continual spatio-temporal graph forecasting. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xu Chen, Junshan Wang, and Kunqing Xie. Trafficstream: A streaming traffic flow forecasting framework based on graph neural networks and continual learning. In Zhi-Hua Zhou (ed.), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 3620–3626, 8 2021. doi: 10.24963/ijcai.2021/498. URL https://doi.org/10.24963/ijcai.2021/498.

Zheng Dong, Renhe Jiang, Haotian Gao, Hangchen Liu, Jinliang Deng, Qingsong Wen, and Xuan Song. Heterogeneity-informed meta-parameter learning for spatiotemporal time series forecasting. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 631–641, 2024.

Yuchen Fang, Yanjun Qin, Haiyong Luo, Fang Zhao, and Kai Zheng. Stwave+: A multi-scale efficient spectral graph attention network with long-term trends for disentangled traffic flow forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Haotian Gao, Renhe Jiang, Zheng Dong, Jinliang Deng, Yuxin Ma, and Xuan Song. Spatial-temporal-decoupled masked pre-training for spatiotemporal forecasting. In Kate Larson (ed.), *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pp. 3998–4006. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/442. URL https://doi.org/10.24963/ijcai.2024/442. Main Track.

Danlei Hu, Lu Chen, Hanxi Fang, Ziquan Fang, Tianyi Li, and Yunjun Gao. Spatio-temporal trajectory similarity measures: A comprehensive survey and quantitative study. *IEEE Transactions on Knowledge and Data Engineering*, 36(5):2191–2212, 2023.

Jiahao Ji, Wentao Zhang, Jingyuan Wang, and Chao Huang. Seeing the unseen: Learning basis confounder representations for robust traffic prediction. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 577–588, 2025.

Jiawei Jiang, Chengkai Han, Wayne Xin Zhao, and Jingyuan Wang. Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023a.

Renhe Jiang, Zhaonan Wang, Jiawei Yong, Puneet Jeph, Quanjun Chen, Yasumasa Kobayashi, Xuan Song, Shintaro Fukushima, and Toyotaro Suzumura. Spatio-temporal meta-graph learning for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8078–8086, 2023b.

Ming Jin, Huan Yee Koh, Qingsong Wen, Daniele Zambon, Cesare Alippi, Geoffrey I Webb, Irwin King, and Shirui Pan. A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning (ICML)*, pp. 5156–5165. PMLR, 2020.

Weiyang Kong, Ziyu Guo, and Yubao Liu. Spatio-temporal pivotal graph neural networks for traffic flow forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8):8627–8635, Mar. 2024. doi: 10.1609/aaai.v38i8.28707. URL https://ojs.aaai.org/index.php/AAAI/article/view/28707.

Rahul Kumar, Manish Bhanu, João Mendes-Moreira, and Joydeep Chandra. Spatio-temporal predictive modeling techniques for different domains: a survey. *ACM Computing Surveys*, 57(2):1–42, 2024.

Sanghyun Lee and Chanyoung Park. Continual traffic forecasting via mixture of experts. *arXiv preprint arXiv:2406.03140*, 2024.

Fuxian Li, Jie Feng, Huan Yan, Guangyin Jin, Fan Yang, Funing Sun, Depeng Jin, and Yong Li. Dynamic graph convolutional recurrent network for traffic prediction: Benchmark and solution. *ACM Transactions on Knowledge Discovery from Data*, 17(1):1–21, 2023.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations (ICLR)*, 2018.

Aoyu Liu and Yaying Zhang. An efficient spatial-temporal transformer with temporal aggregation and spatial memory for traffic forecasting. *Expert Systems with Applications*, 250:123884, 2024a.

Aoyu Liu and Yaying Zhang. Spatial–temporal dynamic graph convolutional network with interactive learning for traffic forecasting. *IEEE Transactions on Intelligent Transportation Systems*, 2024b.

Aoyu Liu and Yaying Zhang. Crossst: An efficient pre-training framework for cross-district pattern generalization in urban spatio-temporal forecasting. In *41th IEEE International Conference on Data Engineering*, 2025.

Hangchen Liu, Zheng Dong, Renhe Jiang, Jiewen Deng, Q Chen, and X Song. 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)*, pp. 21–25, 2023a.

Xu Liu, Yutong Xia, Yuxuan Liang, Junfeng Hu, Yiwei Wang, Lei Bai, Chao Huang, Zhenguang Liu, Bryan Hooi, and Roger Zimmermann. Largest: A benchmark dataset for large-scale traffic forecasting. In *Advances in Neural Information Processing Systems*, 2023b.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2024.

Jiaming Ma, Bingwu Wang, Pengkun Wang, Zhengyang Zhou, Xu Wang, and Yang Wang. Robust spatio-temporal centralized interaction for ood learning. In *Proceedings of the Forty-Second International Conference on Machine Learning (ICML)*, 2025a.

Minbo Ma, Kai Tang, Huan Li, Fei Teng, Dalin Zhang, and Tianrui Li. Beyond fixed variables: Expanding-variate time series forecasting via flat scheme and spatio-temporal focal learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 2054–2065, 2025b.

Hao Miao, Yan Zhao, Chenjuan Guo, Bin Yang, Kai Zheng, Feiteng Huang, Jiandong Xie, and Christian S Jensen. A unified replay-based continuous learning framework for spatio-temporal prediction on streaming data. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pp. 1050–1062. IEEE, 2024.

Hao Miao, Yan Zhao, Chenjuan Guo, Bin Yang, Kai Zheng, and Christian S Jensen. Spatio-temporal prediction on streaming data: A unified federated continuous learning framework. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Zezhi Shao, Zhao Zhang, Fei Wang, Wei Wei, and Yongjun Xu. 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*, pp. 4454–4458, 2022a.

Zezhi Shao, Zhao Zhang, Wei Wei, Fei Wang, Yongjun Xu, Xin Cao, and Christian S Jensen. Decoupled dynamic spatial-temporal graph neural network for traffic forecasting. *arXiv preprint arXiv:2206.09112*, 2022b.

Zezhi Shao, Fei Wang, Yongjun Xu, Wei Wei, Chengqing Yu, Zhao Zhang, Di Yao, Tao Sun, Guangyin Jin, Xin Cao, et al. Exploring progress in multivariate time series forecasting: Comprehensive benchmarking and heterogeneity analysis. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 914–921, 2020.

Jindong Tian, Yuxuan Liang, Ronghui Xu, Peng Chen, Chenjuan Guo, Aoying Zhou, Lujia Pan, Zhongwen Rao, and Bin Yang. Air quality prediction with physics-guided dual neural odes in open systems. In *The Thirteenth International Conference on Learning Representations*, 2025.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Binwu Wang, Yudong Zhang, Jiahao Shi, Pengkun Wang, Xu Wang, Lei Bai, and Yang Wang. Knowledge expansion and consolidation for continual traffic prediction with expanding graphs. *IEEE Transactions on Intelligent Transportation Systems*, 24(7):7190–7201, 2023a.

Binwu Wang, Yudong Zhang, Xu Wang, Pengkun Wang, Zhengyang Zhou, Lei Bai, and Yang Wang. Pattern expansion and consolidation on evolving graphs for continual traffic prediction. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 2223–2232, New York, NY, USA, 2023b. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599463. URL https://doi.org/10.1145/3580305.3599463.

Binwu Wang, Jiaming Ma, Pengkun Wang, Xu Wang, Yudong Zhang, Zhengyang Zhou, and Yang Wang. Stone: A spatio-temporal ood learning framework kills both spatial and temporal shifts. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2948–2959, 2024.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1907–1913, 2019.

Yutong Xia, Yuxuan Liang, Haomin Wen, Xu Liu, Kun Wang, Zhengyang Zhou, and Roger Zimmermann. Deciphering spatio-temporal graph forecasting: A causal lens and treatment. *Advances in Neural Information Processing Systems*, 36:37068–37088, 2023.

Bin Yang, Yuxuan Liang, Chenjuan Guo, and Christian S Jensen. Data driven decision making with time series and spatio-temporal data. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pp. 4517–4522. IEEE Computer Society, 2025.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

Zijian Zhang, Xiangyu Zhao, Qidong Liu, Chunxu Zhang, Qian Ma, Wanyu Wang, Hongwei Zhao, Yiqi Wang, and Zitao Liu. Promptst: Prompt-enhanced spatio-temporal multi-attribute prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 3195–3205, 2023.

Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. Gman: A graph multi-attention network for traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 1234–1241, 2020.

Chuanpan Zheng, Xiaoliang Fan, Shirui Pan, Haibing Jin, Zhaopeng Peng, Zonghan Wu, Cheng Wang, and S Yu Philip. Spatio-temporal joint graph convolutional networks for traffic forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Zhengyang Zhou, Qihe Huang, Kuo Yang, Kun Wang, Xu Wang, Yudong Zhang, Yuxuan Liang, and Yang Wang. Maintaining the status quo: Capturing invariant relations for ood spatiotemporal learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 3603–3614, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599421. URL https://doi.org/10.1145/3580305.3599421.

A APPENDIX

A.1 NOTATIONS

Table 2 summarizes the notations frequently used throughout this manuscript.

Table 2: The notations that are commonly used in the manuscript.

Notation	Definition
$\mathbb{G} = \{G_\tau\}_{\tau=1}^{\mathcal{T}}$	Streaming spatio-temporal graph
$\mathbf{X}_\tau$	Inputs for the τ period
$\mathbf{Y}_\tau$	Prediction for the τ period
$\mathbf{P}_\tau$	contextual pattern bank for the τ period
$\mathbf{P'}_\tau$	Expanded contextual pattern bank
$\mathbf{H}_\tau$	Hidden representation for the τ period
$\mathcal{M}_\theta$	Spatio-Temporal backbone
$\mathbf{F}_\tau$	Frequency domain embedding
$\mathbf{H}_\tau^f$	Representation after frequency-domain processing
$\mathbf{W}_q$	Trainable parameter weights
$\mathbf{W}_k$	Trainable parameter weights
$\mathbf{W}_v$	Trainable parameter weights
$\phi(\cdot)$	Random mapping function
$\mathbf{H}_\tau^s$	Input of the `DLGA` module
$\mathbf{H}_\tau^{s'}$	Output of the `DLGA` module

A.2 RELATED WORK DETAILS

A.2.1 SPATIO-TEMPORAL FORECASTING

Spatio-temporal forecasting aims to support decision-making in critical domains such as intelligent transportation and smart cities by uncovering dynamic correlations embedded in spatio-temporal data. These data typically exhibit strong spatial-temporal correlations and pronounced heterogeneity. In recent years, deep learning-based STGNNs have emerged as effective tools for such forecasting tasks. STGNNs generally employ temporal modules (e.g., recurrent neural networks (RNNs) (Li et al., 2018; Jiang et al., 2023b; Shao et al., 2022b) and convolutional neural networks (CNNs)) (Yu et al., 2018; Liu & Zhang, 2024b;a) to capture temporal correlations, while leveraging spatial modules (e.g., graph neural networks (GNNs)) (Veličković et al., 2018; Song et al., 2020) to model spatial relationships.

Early STGNNs, such as STGCN (Yu et al., 2018) and DCRNN (Li et al., 2018), combined basic temporal and spatial components for forecasting tasks, often relying on predefined geographic adjacency matrices. However, these static assumptions hinder their ability to model dynamically changing spatial correlations in a data-driven manner. Subsequent works—such as GWNet (Wu et al., 2019), DGCRN (Li et al., 2023), and MegaCRN (Jiang et al., 2023b)—introduced adaptive adjacency matrices or learned spatial correlations directly from data, significantly improving prediction accuracy. More recent advances, including STID (Shao et al., 2022a), STAEformer (Liu et al., 2023a), and HimNet (Dong et al., 2024), have highlighted the importance of spatial pattern distinction in enhancing forecasting performance. These models incorporate trainable mechanisms such as spatial embeddings, parameter pools, and contextual pattern banks to distinguish spatial patterns more precisely, thereby improving both accuracy and adaptability.

Despite these advancements, most existing STGNNs are built on static assumptions and are not designed to operate in dynamic, continually evolving spatio-temporal environments—limiting their applicability in continual learning scenarios.

A.2.2 CONTINUAL SPATIO-TEMPORAL LEARNING

Early research in continual learning primarily focused on computer vision (Lee & Park, 2024) and natural language processing (Caccia et al., 2020). With the rapid development of IoT and intelligent

transportation systems, attention has increasingly shifted toward CSTL (Chen et al., 2021; Wang et al., 2023a; Chen & Liang, 2025; Wang et al., 2023b; Miao et al., 2025), which addresses the challenges of dynamically evolving and expanding spatio-temporal data. CSTL aims to enable models to continuously learn and adapt to new patterns and knowledge in changing environments, while minimizing forgetting of previously acquired information or performance degradation.

One of the earliest frameworks in this domain, TrafficStream (Chen et al., 2021), pioneered the integration of spatio-temporal modeling with continual learning. It employed strategies such as historical data replay and parameter smoothing to handle long-term streaming traffic data, achieving accurate traffic flow forecasting. Subsequently, the STKEC (Wang et al., 2023a) introduced an influence-based knowledge expansion strategy and a memory-augmented knowledge consolidation mechanism to better accommodate the growth of transportation networks while mitigating catastrophic forgetting. The EAC (Chen & Liang, 2025) further advanced the field by incorporating prompt tuning, enabling CSTL with a small number of trainable parameters. Its dynamic prompt pool, which supports both "expansion" and "compression," enhances adaptability to new nodes while preserving historical knowledge, improving both generalization and computational efficiency. In addition, the UFCL (Miao et al., 2025) leveraged federated learning to preserve data privacy and introduced a global replay buffer for synthetic spatio-temporal data, addressing the challenges of distributed streaming environments.

Despite these advancements, most existing methods primarily focus on alleviating knowledge forgetting, while overlooking the critical role of the spatio-temporal backbone in continual learning scenarios.

A.3 Further Methods Details

A.3.1 Approximation Derivation of Eq. 9

An approximate derivation of the attention mechanism in the dual-stream linear graph attention is presented below:

$$\begin{aligned}
\text{Attention}\left(\boldsymbol{q}_u,\, \boldsymbol{k}_v,\, \boldsymbol{v}_v,\, \boldsymbol{p}_v\right) &= \sum_{v=1}^{N} \frac{\exp\left(\boldsymbol{q}_u^\top \boldsymbol{k}_v\right) \boldsymbol{v}_v}{\sum_{w=1}^{N} \exp\left(\boldsymbol{q}_u^\top \boldsymbol{k}_w\right)} + \sum_{v=1}^{N} \frac{\exp\left(\boldsymbol{q}_u^\top \boldsymbol{p}_v\right) \boldsymbol{v}_v}{\sum_{w=1}^{N} \exp\left(\boldsymbol{q}_u^\top \boldsymbol{p}_w\right)} \\
&\approx \frac{\sum_{v=1}^{N} \phi\left(\boldsymbol{q}_u\right)^\top \phi\left(\boldsymbol{k}_v\right) \boldsymbol{v}_v}{\sum_{w=1}^{N} \phi\left(\boldsymbol{q}_u\right)^\top \phi\left(\boldsymbol{k}_w\right)} + \frac{\sum_{v=1}^{N} \phi\left(\boldsymbol{q}_u\right)^\top \phi\left(\boldsymbol{p}_v\right) \boldsymbol{v}_v}{\sum_{w=1}^{N} \phi\left(\boldsymbol{q}_u\right)^\top \phi\left(\boldsymbol{p}_w\right)} \\
&= \underbrace{\left[\frac{\phi\left(\boldsymbol{q}_u\right)^\top \sum_{v=1}^{N} \phi\left(\boldsymbol{k}_v\right) \boldsymbol{v}_v^\top}{\phi\left(\boldsymbol{q}_u\right)^\top \sum_{w=1}^{N} \phi\left(\boldsymbol{k}_w\right)}\right]}_{\text{Term 1: Representation-based aggregation}} + \underbrace{\left[\frac{\phi\left(\boldsymbol{q}_u\right)^\top \sum_{v=1}^{N} \phi\left(\boldsymbol{p}_v\right) \boldsymbol{v}_v^\top}{\phi\left(\boldsymbol{q}_u\right)^\top \sum_{w=1}^{N} \phi\left(\boldsymbol{p}_w\right)}\right]}_{\text{Term 2: Prompt-based aggregation}}
\end{aligned} \tag{10}$$

where $\boldsymbol{q}_u$ is the query tensor of node u; $\boldsymbol{k}_v$ and $\boldsymbol{v}_v$ are the key and value tensors of node v, respectively; and $\boldsymbol{p}_v$ represents the prompt information for node v.

A.3.2 Algorithm Workflow

The overall workflow of `STBP` for continual spatio-temporal forecasting is presented in a more intuitive manner in Algorithm 1.

A.4 Additional Experiment Details

A.4.1 Dataset Details

Table 3 and Table 4 jointly summarize the characteristics of the three continual spatio-temporal datasets used in this study: **PEMS-Stream**, **CA-Stream**, and **Air-Stream**. These datasets differ in domain (traffic vs. weather), temporal span, and topological evolution, collectively covering a broad spectrum of real-world non-stationary scenarios suitable for evaluating continual learning models. PEMS-Stream contains highway traffic sensor readings collected across California from July 2011 to September 2017. It spans seven periods with a gradual increase in the number of sensor nodes—from 655 to 871—resulting in a +33% relative growth. This dataset simulates realistic,

Algorithm 1 The workflow of `STBP` for continual spatio-temporal forecasting

Require: Spatio-temporal backbone $\mathcal{M}_\theta$, contextual pattern bank $\{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_\tau\}$, streaming train data $\{\mathbf{X}_1, \mathbf{X}_2 \ldots, \mathbf{X}_\tau\}$.
Ensure: Optimized backbone $\mathcal{M}_{\theta^*}$ and contextual pattern bank $\{\mathbf{P}_1^*, \ldots, \mathbf{P}_\tau^*\}$
Initialize:$\mathcal{M}_\theta \leftarrow \{\}$, $\mathbf{P}_1 \leftarrow \{\}$
for each period i in $\{1, 2, 3, \ldots, \tau\}$ **do**
  if $i == 1$ **then**
    ▷ *Initial training phase* ◁
    Construct the initial contextual pattern bank $\mathbf{P}_1$
    Optimize backbone and contextual pattern bank with initial data $\mathbf{X}_1$:
    $(\mathcal{M}_{\theta^*}, \mathbf{P}_1^*) \leftarrow \text{argmin}_\theta \mathcal{M}_\theta(\mathbf{X}_1, \mathbf{P}_1)$
  else
    ▷ *Streaming learning phase* ◁
    Expand contextual pattern bank $\mathbf{P}_i$: $\mathbf{P}_i \leftarrow \mathbf{P}_{i-1} \parallel \Delta\mathbf{P}_i$
    Inherit parameters: $(\mathcal{M}_\theta, \mathbf{P}_i) \leftarrow (\mathcal{M}_{\theta^*}, \mathbf{P}_{i-1}^*)$
    Freeze backbone parameters $\mathcal{M}_\theta$: $\theta \leftarrow \texttt{freeze}(\theta)$
    Fine-tune $\mathbf{P}_i$ with backbone $\mathcal{M}_\theta$ on $\mathbf{X}_\text{i}$:
    $\mathbf{P}_i^* \leftarrow \text{argmin}_\theta \mathcal{M}_\theta(\mathbf{X}_i, \mathbf{P}_i)$

Table 3: Overview of continual spatio-temporal forecasting datasets.

Dataset	Domain	Time Range	Period	Node Expansion	Frequency
PEMS-Stream	Traffic	07/10/2011 - 09/08/2017	7	655 → 715 → 786 → 822 → 834 → 850 → 871	5 min
CA-Stream	Traffic	01/01/2019 - 04/30/2019	4	480 → 691 → 1175 → 1698	5 min
Air-Stream	Weather	01/01/2016 - 12/31/2019	4	1087 → 1154 → 1193 → 1202	1 hour

long-term infrastructure expansion and serves as a benchmark for evaluating model adaptability under progressive and stable topological changes. CA-Stream, also in the traffic domain, covers a much shorter period (January to April 2019) but features a sharp and sudden node expansion—from 480 to 1,698—corresponding to a +254% relative increase. This explosive growth introduces significant distributional shifts, making CA-Stream a challenging testbed for assessing model robustness under rapidly evolving conditions. Air-Stream focuses on urban air quality and environmental measurements from 2016 to 2019. It exhibits modest but steady node growth—from 1,087 to 1,202 (+10%)—and represents a relatively stable expansion setting. Its distinct domain and smoother structural changes make it particularly suitable for evaluating cross-domain generalization and robustness to gradual environmental variation.

To further assess non-stationarity, we conducted Maximum Mean Discrepancy (MMD) tests across different periods, separately evaluating *original nodes* (present from the beginning) and *added nodes* (introduced during expansion), as shown in Table 5. A distribution shift is considered significant when MMD > 0.1 or $p < 0.05$. Across all datasets, added nodes consistently exhibit stronger distributional shifts, reflecting the spatial disruptions caused by topological expansion. For instance, CA-Stream shows a substantial shift for added nodes (MMD = 0.3361, $p = 0.0010$), consistent with its rapid growth. Interestingly, Air-Stream records the highest MMD among original nodes (0.3324, $p = 0.001$), despite minimal structural change—indicating notable temporal drift in environmental data. This highlights Air-Stream's importance for evaluating robustness to evolving distributions even under stable topology. By contrast, PEMS-Stream shows only moderate drift among original nodes (MMD = 0.0939), aligning with its smoother expansion. CA-Stream presents weaker drift in original nodes (MMD = 0.0792, $p = 0.1119$), likely due to its limited temporal span. These results underscore the dual challenge in continual spatio-temporal learning: managing both spatial shifts

Table 4: Topological dynamics and evaluation purposes of the datasets.

Dataset	Topology Change	△Nodes	Relative Change	Primary Purpose
PEMS-Stream	Gradual expansion	216	+33%	Realistic progressive growth
CA-Stream	Explosive expansion	1,218	+254%	Extreme incremental stress test
AIR-Stream	Stable expansion	115	+10%	Cross-domain validation

Table 5: Distribution shift analysis based on MMD tests.

Type	PEMS-Stream		AIR-Stream		CA-Stream	
	MMD	p	MMD	p	MMD	p
Original Node	0.0939	0.008	0.3324	0.001	0.0792	0.1119
Added Node	0.2958	0.001	0.2679	0.001	0.3361	0.0010

induced by node expansion and temporal non-stationarity inherent to dynamic environments, with their nature and intensity varying across domains.

A.4.2 Baselines and Metrics Details

In this paper, we provide a detailed comparison with two categories of representative models:

Conventional Spatio-Temporal Forecasting Models. ❶ **GWNet** (Wu et al., 2019): A STGNN model based on an adaptive adjacency matrix that can adaptively capture latent spatial dependencies. This model combines graph convolutional networks and temporal convolutions to effectively capture spatio-temporal correlations in the data. ❷ **STID** (Shao et al., 2022a): An efficient multilayer perceptron model that solves the problem of sample non-separability using trainable embeddings, showing outstanding performance in spatio-temporal forecasting tasks. ❸ **iTransformer** (Liu et al., 2024): A time-series model that does not rely on a static graph structure. By modeling the interactions between variables, it captures temporal features and is effectively applied to multivariate time series forecasting tasks.

Continual Spatio-Temporal Forecasting Models. These models are designed to handle time-varying data and are suitable for continual training tasks. Like `STBP`, they belong to the category of continual learning models. We selected the following three representative models for comparison: ❶ **TrafficStream** (Chen et al., 2021): The first model for CSTF, it employs a traffic pattern fusion approach, historical data replay, and parameter smoothing strategies to efficiently integrate and learn new spatio-temporal patterns in the continuously expanding and evolving traffic network. ❷ **STKEC** (Wang et al., 2023a): A traffic forecasting model based on the continual learning paradigm. Through an influence-based knowledge expansion strategy and a memory-augmented knowledge consolidation mechanism, STKEC helps the model effectively integrate new spatio-temporal traffic patterns in an ever-expanding road network while retaining previously learned spatio-temporal patterns. ❸ **EAC** (Chen & Liang, 2025): A CSTF based on prompt tuning. By integrating a base STGNN with a continual prompt pool, it efficiently addresses incremental learning and catastrophic forgetting in streaming data using lightweight trainable parameters.

The Excluded Models. Some baselines that might be considered relevant for comparison were excluded, and we provide explanations for their exclusion below. ❶ **STAEformer** (Liu et al., 2023a): a widely recognized baseline, was not included in our comparison due to non-convergence observed when applying the same experimental setting as used for GWNet and STID on the selected three datasets. To ensure fair and unambiguous evaluation, we excluded it from the results and have provided the corresponding training logs in the anonymous code repository. ❷ **UFCL** (Miao et al., 2025): The CSTF method UFCL is not included in the comparison due to differences in experimental settings, which prevent a fair evaluation.

Table 6: Comparison of prediction performance for each incremental period on PEMS-Stream. **Bold**: best, underline: second best.

Model	Metric	PEMS-Stream Period							
		2011	2012	2013	2014	2015	2016	2017	Avg.
GWNet	MAE	$16.48_{\pm 0.62}$	$16.59_{\pm 0.62}$	$15.23_{\pm 1.03}$	$15.50_{\pm 0.80}$	$17.43_{\pm 1.06}$	$15.30_{\pm 0.33}$	$19.77_{\pm 1.32}$	$16.62_{\pm 0.14}$
	RMSE	$25.82_{\pm 0.81}$	$26.01_{\pm 0.75}$	$25.44_{\pm 1.36}$	$24.81_{\pm 1.13}$	$28.31_{\pm 1.38}$	$26.69_{\pm 0.62}$	$32.40_{\pm 1.76}$	$27.07_{\pm 0.20}$
	MAPE (%)	$23.81_{\pm 1.38}$	$24.63_{\pm 1.99}$	$21.44_{\pm 1.34}$	$23.93_{\pm 1.54}$	$27.66_{\pm 3.19}$	$24.02_{\pm 4.99}$	$32.29_{\pm 1.03}$	$25.40_{\pm 0.87}$
STID	MAE	$16.26_{\pm 0.34}$	$15.89_{\pm 0.68}$	$14.54_{\pm 1.16}$	$14.87_{\pm 0.15}$	$14.44_{\pm 0.16}$	$14.73_{\pm 0.52}$	$15.24_{\pm 0.45}$	$15.14_{\pm 0.12}$
	RMSE	$24.26_{\pm 0.42}$	$24.99_{\pm 0.84}$	$23.82_{\pm 1.80}$	$24.13_{\pm 0.35}$	$23.77_{\pm 0.29}$	$25.76_{\pm 0.73}$	$24.12_{\pm 0.24}$	$24.41_{\pm 0.20}$
	MAPE (%)	$23.27_{\pm 1.96}$	$21.60_{\pm 1.68}$	$19.85_{\pm 1.90}$	$21.70_{\pm 1.40}$	$19.79_{\pm 1.26}$	$21.84_{\pm 5.47}$	$34.13_{\pm 8.59}$	$23.17_{\pm 2.32}$
iTransformer	MAE	$14.47_{\pm 0.19}$	$14.06_{\pm 0.46}$	$14.37_{\pm 0.41}$	$15.37_{\pm 0.62}$	$14.51_{\pm 0.39}$	$14.03_{\pm 0.47}$	$15.53_{\pm 0.66}$	$14.62_{\pm 0.40}$
	RMSE	$21.75_{\pm 0.41}$	$21.70_{\pm 0.83}$	$22.65_{\pm 0.72}$	$24.29_{\pm 1.27}$	$23.36_{\pm 0.77}$	$24.37_{\pm 0.72}$	$24.95_{\pm 1.03}$	$23.29_{\pm 0.77}$
	MAPE (%)	$30.14_{\pm 3.82}$	$30.18_{\pm 2.46}$	$36.45_{\pm 3.48}$	$34.89_{\pm 2.51}$	$32.19_{\pm 3.44}$	$31.94_{\pm 2.73}$	$32.79_{\pm 1.81}$	$32.65_{\pm 1.53}$
TrafficStream	MAE	$14.14_{\pm 0.16}$	$13.78_{\pm 0.19}$	$13.60_{\pm 0.10}$	$14.47_{\pm 0.09}$	$14.11_{\pm 0.13}$	$13.52_{\pm 0.11}$	$14.79_{\pm 0.07}$	$14.06_{\pm 0.10}$
	RMSE	$21.81_{\pm 0.22}$	$21.71_{\pm 0.28}$	$21.93_{\pm 0.21}$	$23.32_{\pm 0.13}$	$23.08_{\pm 0.22}$	$24.05_{\pm 0.16}$	$24.41_{\pm 0.15}$	$22.90_{\pm 0.18}$
	MAPE (%)	$19.14_{\pm 0.81}$	$19.48_{\pm 0.70}$	$19.55_{\pm 0.91}$	$20.30_{\pm 0.71}$	$20.07_{\pm 1.03}$	$18.16_{\pm 0.43}$	$19.66_{\pm 0.61}$	$19.48_{\pm 0.41}$
STKEC	MAE	$14.01_{\pm 0.10}$	$13.91_{\pm 0.23}$	$13.64_{\pm 0.08}$	$14.51_{\pm 0.11}$	$14.02_{\pm 0.05}$	$13.45_{\pm 0.06}$	$14.89_{\pm 0.11}$	$14.07_{\pm 0.04}$
	RMSE	$21.53_{\pm 0.21}$	$21.76_{\pm 0.33}$	$22.02_{\pm 0.11}$	$23.58_{\pm 0.23}$	$22.96_{\pm 0.08}$	$24.13_{\pm 0.22}$	$24.53_{\pm 0.10}$	$22.93_{\pm 0.05}$
	MAPE (%)	$18.54_{\pm 0.43}$	$18.90_{\pm 0.26}$	$19.64_{\pm 1.22}$	$19.39_{\pm 0.73}$	$20.31_{\pm 1.13}$	$18.20_{\pm 0.58}$	$20.05_{\pm 1.17}$	$19.29_{\pm 0.42}$
EAC	MAE	$13.26_{\pm 0.05}$	$12.97_{\pm 0.11}$	$12.95_{\pm 0.14}$	$13.91_{\pm 0.22}$	$13.56_{\pm 0.17}$	$13.01_{\pm 0.16}$	$14.72_{\pm 0.28}$	$13.49_{\pm 0.15}$
	RMSE	$20.15_{\pm 0.09}$	$20.06_{\pm 0.15}$	$20.61_{\pm 0.23}$	$22.06_{\pm 0.37}$	$21.78_{\pm 0.24}$	$23.00_{\pm 0.16}$	$23.52_{\pm 0.23}$	$21.60_{\pm 0.19}$
	MAPE (%)	$17.79_{\pm 0.52}$	$18.44_{\pm 0.69}$	$19.23_{\pm 0.82}$	$20.80_{\pm 0.96}$	$20.70_{\pm 1.05}$	$18.63_{\pm 1.06}$	$22.29_{\pm 1.85}$	$19.69_{\pm 0.90}$
STBP	MAE	**11.12**$_{\pm 0.07}$	**10.71**$_{\pm 0.09}$	**10.48**$_{\pm 0.08}$	**11.35**$_{\pm 0.08}$	**10.96**$_{\pm 0.06}$	**10.35**$_{\pm 0.05}$	**12.04**$_{\pm 0.05}$	**11.00**$_{\pm 0.04}$
	RMSE	**17.04**$_{\pm 0.07}$	**16.86**$_{\pm 0.11}$	**17.09**$_{\pm 0.16}$	**18.30**$_{\pm 0.12}$	**17.94**$_{\pm 0.08}$	**19.65**$_{\pm 0.09}$	**20.14**$_{\pm 0.04}$	**18.15**$_{\pm 0.06}$
	MAPE (%)	**14.44**$_{\pm 0.21}$	**14.05**$_{\pm 0.15}$	**13.75**$_{\pm 0.11}$	**14.87**$_{\pm 0.06}$	**14.49**$_{\pm 0.03}$	**13.17**$_{\pm 0.07}$	**16.88**$_{\pm 0.26}$	**14.52**$_{\pm 0.09}$

Metrics Details. Additionally, the performance metrics used in the experiments to evaluate the model, namely MAE, RMSE, and MAPE, are defined as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{11}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{12}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\% \tag{13}$$

where n represents the number of observed samples, y_i denotes the i-th real sample, and $\hat{y}_i$ is the corresponding predicted value.

A.4.3 Implementation Details

All experiments are conducted on a machine with an NVIDIA Tesla V100 GPU and 32 GB of memory. The Adam optimizer, with an initial learning rate of 0.01, is used to optimize the training process. The batch size is set to 64, the number of training epochs is set to 200, and an early stopping mechanism is implemented to ensure efficient convergence. The reported results for all baselines are the average of five repeated runs.

A.4.4 Experiment Results

Tables 6 and 7 present more detailed experimental results, including the prediction performance metrics for each incremental training period. The metrics for each period represent the average predictions over 12 time steps. The results demonstrate that, both in terms of overall CSTF performance across all stages and in each individual incremental period, STBP exhibits outstanding continual spatio-temporal learning capabilities. This advantage can be attributed to its well-designed spatio-temporal backbone structure and the effective support provided by the contextual pattern bank in consolidating and inductively incorporating both historical and new spatio-temporal knowledge.

To further evaluate the robustness of the proposed model under low-resource scenarios, we constructed a few-shot training setting and compared it against existing baselines. Specifically, we simulated a few-shot environment in which the sample size of the first incremental period was kept unchanged, while the training set size for subsequent periods was reduced to only 10% of the original. The test

Table 7: Comparison of prediction performance for each incremental period on CA-Stream and AIR-Stream. **Bold**: best, underline: second best.

Model	Metric	CA-Stream Period					AIR-Stream Period				
		Jan	Feb	Mar	Apr	Avg.	2016	2017	2018	2019	Avg.
GWNet	MAE	$22.39_{\pm 0.82}$	$21.08_{\pm 0.85}$	$17.96_{\pm 0.55}$	$18.78_{\pm 0.75}$	$20.05_{\pm 0.57}$	$26.82_{\pm 0.78}$	$27.10_{\pm 2.42}$	$25.29_{\pm 2.11}$	$21.22_{\pm 2.61}$	$25.11_{\pm 0.64}$
	RMSE	$35.73_{\pm 1.38}$	$33.61_{\pm 1.62}$	$28.74_{\pm 1.33}$	$29.04_{\pm 1.35}$	$31.78_{\pm 1.11}$	**$42.98_{\pm 1.18}$**	$40.86_{\pm 3.33}$	$41.06_{\pm 2.81}$	$31.95_{\pm 2.74}$	$39.21_{\pm 0.88}$
	MAPE (%)	$20.10_{\pm 1.10}$	$21.78_{\pm 1.79}$	$19.41_{\pm 0.47}$	$20.03_{\pm 0.31}$	$20.33_{\pm 0.49}$	$28.95_{\pm 0.50}$	$30.60_{\pm 3.13}$	$39.74_{\pm 1.76}$	$29.34_{\pm 2.62}$	$32.16_{\pm 0.78}$
STID	MAE	$24.20_{\pm 0.50}$	$25.87_{\pm 1.52}$	$19.98_{\pm 0.91}$	$16.93_{\pm 0.12}$	$21.74_{\pm 0.52}$	$29.09_{\pm 2.08}$	$26.23_{\pm 3.96}$	$18.50_{\pm 2.90}$	$19.24_{\pm 1.10}$	$23.27_{\pm 0.96}$
	RMSE	$36.41_{\pm 0.69}$	$38.24_{\pm 1.68}$	$30.39_{\pm 1.20}$	$25.86_{\pm 0.29}$	$32.72_{\pm 0.60}$	$45.10_{\pm 2.41}$	$38.67_{\pm 4.92}$	$32.28_{\pm 5.83}$	$29.47_{\pm 1.11}$	$36.38_{\pm 1.46}$
	MAPE (%)	$22.68_{\pm 1.36}$	$23.96_{\pm 3.28}$	$23.27_{\pm 0.83}$	$19.19_{\pm 1.74}$	$22.28_{\pm 1.16}$	$30.12_{\pm 1.96}$	$28.36_{\pm 3.57}$	$26.21_{\pm 4.22}$	$25.89_{\pm 1.40}$	$27.64_{\pm 0.84}$
iTransformer	MAE	$21.35_{\pm 0.88}$	$19.05_{\pm 0.37}$	$17.33_{\pm 0.49}$	$16.88_{\pm 0.38}$	$18.65_{\pm 0.46}$	$28.19_{\pm 1.17}$	$23.38_{\pm 0.77}$	$16.72_{\pm 0.34}$	$18.55_{\pm 0.43}$	$21.71_{\pm 0.34}$
	RMSE	$33.14_{\pm 1.29}$	$30.51_{\pm 0.55}$	$27.44_{\pm 0.81}$	$26.14_{\pm 0.59}$	$29.31_{\pm 0.73}$	$44.85_{\pm 1.58}$	$34.61_{\pm 1.11}$	$28.54_{\pm 0.54}$	$28.04_{\pm 0.31}$	$34.01_{\pm 0.54}$
	MAPE (%)	$20.92_{\pm 1.21}$	$19.80_{\pm 1.03}$	$18.90_{\pm 1.01}$	$18.16_{\pm 0.81}$	$19.45_{\pm 0.45}$	$31.54_{\pm 1.31}$	$28.23_{\pm 0.88}$	$26.12_{\pm 0.26}$	$27.10_{\pm 0.72}$	$28.25_{\pm 0.48}$
TrafficStream	MAE	$19.88_{\pm 0.22}$	$18.61_{\pm 0.23}$	$16.97_{\pm 0.19}$	$16.58_{\pm 0.05}$	$18.01_{\pm 0.13}$	$26.99_{\pm 0.76}$	$23.48_{\pm 1.08}$	$16.62_{\pm 0.50}$	$18.62_{\pm 0.94}$	$21.42_{\pm 0.44}$
	RMSE	$31.43_{\pm 0.24}$	$30.06_{\pm 0.28}$	$27.04_{\pm 0.26}$	$25.83_{\pm 0.06}$	$28.59_{\pm 0.14}$	$43.45_{\pm 1.19}$	$34.39_{\pm 1.36}$	$28.92_{\pm 0.71}$	$28.13_{\pm 1.00}$	$33.72_{\pm 0.64}$
	MAPE (%)	$17.12_{\pm 0.48}$	$17.93_{\pm 0.66}$	$17.05_{\pm 0.28}$	$16.70_{\pm 0.13}$	$17.20_{\pm 0.13}$	$28.81_{\pm 0.95}$	$27.61_{\pm 0.97}$	$25.68_{\pm 1.83}$	$25.89_{\pm 0.80}$	$26.99_{\pm 0.73}$
STKEC	MAE	$19.68_{\pm 0.26}$	$18.67_{\pm 0.32}$	$17.01_{\pm 0.13}$	$16.76_{\pm 0.11}$	$18.03_{\pm 0.17}$	$27.98_{\pm 1.23}$	$24.34_{\pm 0.78}$	$16.66_{\pm 0.62}$	$18.41_{\pm 0.68}$	$21.85_{\pm 0.36}$
	RMSE	$31.14_{\pm 0.28}$	$30.25_{\pm 0.60}$	$27.17_{\pm 0.26}$	$26.14_{\pm 0.15}$	$28.67_{\pm 0.29}$	$44.58_{\pm 1.53}$	$35.28_{\pm 1.08}$	$29.31_{\pm 1.45}$	$27.72_{\pm 0.42}$	$34.22_{\pm 0.70}$
	MAPE (%)	$17.34_{\pm 1.50}$	$17.58_{\pm 0.29}$	$17.28_{\pm 0.56}$	$16.32_{\pm 0.10}$	$17.13_{\pm 0.33}$	$29.15_{\pm 0.89}$	$28.78_{\pm 1.58}$	$25.08_{\pm 0.53}$	$26.38_{\pm 1.55}$	$27.35_{\pm 0.34}$
EAC	MAE	$19.18_{\pm 0.18}$	$17.83_{\pm 0.14}$	$17.19_{\pm 0.38}$	$17.45_{\pm 0.51}$	$17.91_{\pm 0.24}$	$28.13_{\pm 0.54}$	$21.68_{\pm 0.55}$	$16.03_{\pm 0.41}$	$17.24_{\pm 0.32}$	$20.77_{\pm 0.24}$
	RMSE	$30.06_{\pm 0.32}$	$28.39_{\pm 0.12}$	$26.38_{\pm 0.39}$	$26.07_{\pm 0.59}$	$27.73_{\pm 0.25}$	$45.21_{\pm 0.82}$	$32.60_{\pm 0.47}$	$27.35_{\pm 0.44}$	**$26.59_{\pm 0.27}$**	$32.94_{\pm 0.34}$
	MAPE (%)	$17.19_{\pm 0.54}$	$17.25_{\pm 0.26}$	$17.50_{\pm 0.41}$	$18.04_{\pm 0.54}$	$17.49_{\pm 0.24}$	$29.35_{\pm 0.40}$	$25.76_{\pm 0.76}$	$26.82_{\pm 1.41}$	$25.15_{\pm 0.52}$	$26.77_{\pm 0.39}$
STBP	MAE	**$16.72_{\pm 0.23}$**	**$15.03_{\pm 0.06}$**	**$13.99_{\pm 0.08}$**	**$13.94_{\pm 0.07}$**	**$14.92_{\pm 0.05}$**	**$26.52_{\pm 0.92}$**	**$20.49_{\pm 0.33}$**	**$14.95_{\pm 0.28}$**	**$16.89_{\pm 0.36}$**	**$19.71_{\pm 0.24}$**
	RMSE	**$27.20_{\pm 0.37}$**	**$25.16_{\pm 0.08}$**	**$22.57_{\pm 0.14}$**	**$21.73_{\pm 0.07}$**	**$24.17_{\pm 0.09}$**	$43.26_{\pm 1.74}$	**$31.76_{\pm 0.38}$**	**$26.41_{\pm 0.35}$**	$26.67_{\pm 0.41}$	**$32.02_{\pm 0.47}$**
	MAPE (%)	**$14.87_{\pm 0.84}$**	**$14.52_{\pm 0.55}$**	**$13.90_{\pm 0.21}$**	**$14.33_{\pm 0.17}$**	**$14.41_{\pm 0.42}$**	**$27.15_{\pm 0.55}$**	**$24.18_{\pm 0.37}$**	**$23.99_{\pm 0.48}$**	**$24.15_{\pm 0.47}$**	**$24.86_{\pm 0.20}$**

Table 8: Comparison of few-shot forecasting results. **Bold**: best, underline: second best.

Model	PEMS-Stream (10%)			CA-Stream (10%)			AIR-Stream (10%)		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
GWN	20.73	34.13	29.04%	28.69	44.85	30.72%	32.62	47.98	53.04%
STID	24.32	39.53	39.68%	31.94	48.90	34.72%	40.30	58.64	61.88%
iTransformer	19.22	30.66	43.26%	26.94	40.51	31.48%	31.38	45.43	55.33%
TrafficStream	14.09	22.81	19.73%	18.17	28.68	16.89%	28.54	42.38	43.97%
STKEC	14.14	23.01	19.20%	18.12	28.63	16.67%	23.73	36.61	32.71%
EAC	13.83	21.94	20.48%	19.12	29.15	19.10%	21.21	33.24	29.52%
STBP	**11.86**	**19.48**	**15.95%**	**17.17**	**27.17**	**16.32%**	**20.46**	**32.64**	**28.48%**

set size remained fixed throughout. As shown in Table 8, STBP consistently outperforms all other methods, highlighting its strong ability to extract meaningful patterns from limited data. Continual baselines such as TrafficStream, STKEC, and EAC are more resilient to low-resource conditions than conventional STGNNs (e.g., GWNet, STID), yet they still suffer from performance drops, especially on AIR-Stream and CA-Stream, which exhibit strong distributional drift (Table 5).

Existing continual spatio-temporal learning methods typically test the model immediately after training each incremental period, rather than conducting a unified evaluation on all historical periods once all incremental training has been completed. In other words, current practices do not directly assess the model's ability to retain historical knowledge. To address this, we reevaluated the model on the test sets from all historical periods after completing the full incremental training, in order to assess the extent of forgetting. The results in the Figure 8 below show that all continual learning methods showed varying degrees of performance degradation in this post-hoc evaluation, indicating catastrophic forgetting of old tasks as new nodes and data were introduced. Nevertheless, STBP achieved the best overall performance, demonstrating its relative advantage in mitigating forgetting.

A.4.5 Parameter Sensitivity Analysis

Beyond the feature dimension d, we further investigated the sensitivity of two key architectural hyperparameters: the number of DLGA layers and attention heads. As shown in Figure 9, increasing either parameter yields marginal gains at best, and in some cases, even leads to slight performance degradation. Overall, apart from the feature dimension, model performance remains relatively insensitive to these hyperparameter variations.

A.4.6 Additional Case Study

To maintain consistency with the case study on PEMS-Stream, we also conducted case studies on the CA-Stream and Air-Stream datasets to further validate the expansion and distinction capabilities of the contextual pattern bank in STBP. The experimental results for CA-Stream are shown in

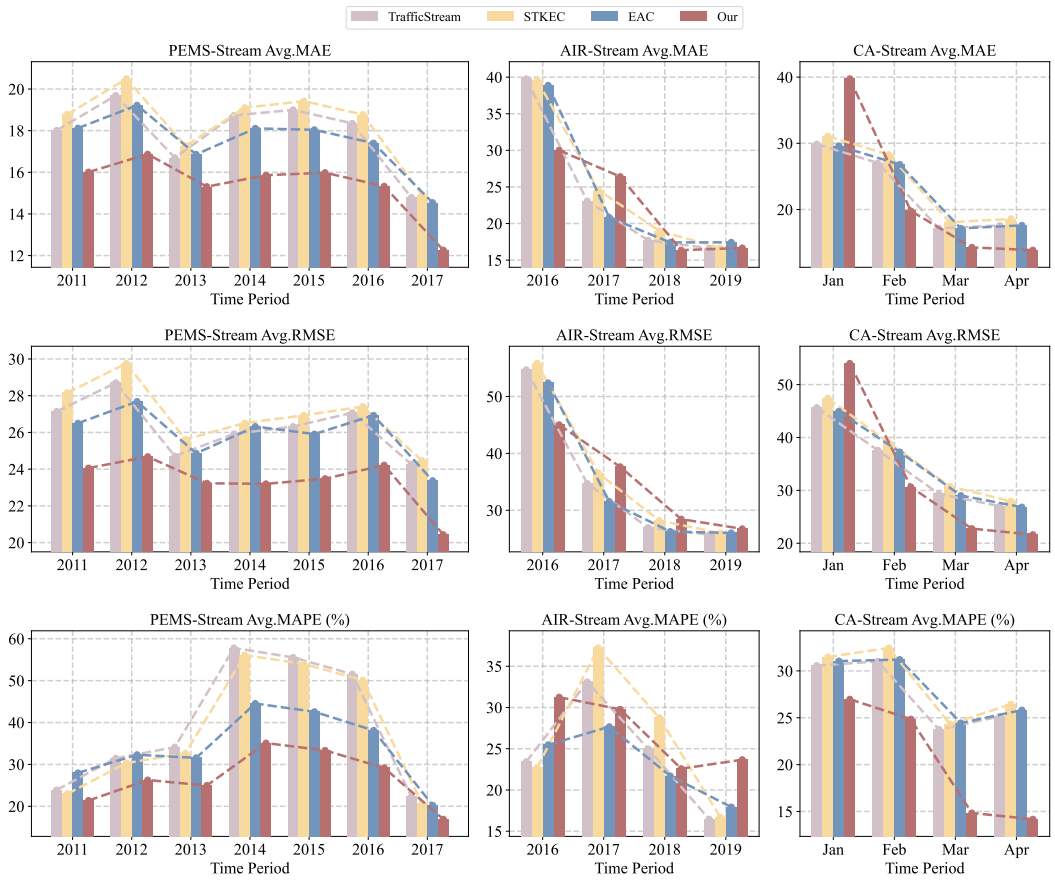

Figure 8: Comparison of historical knowledge forgetting at each incremental period.

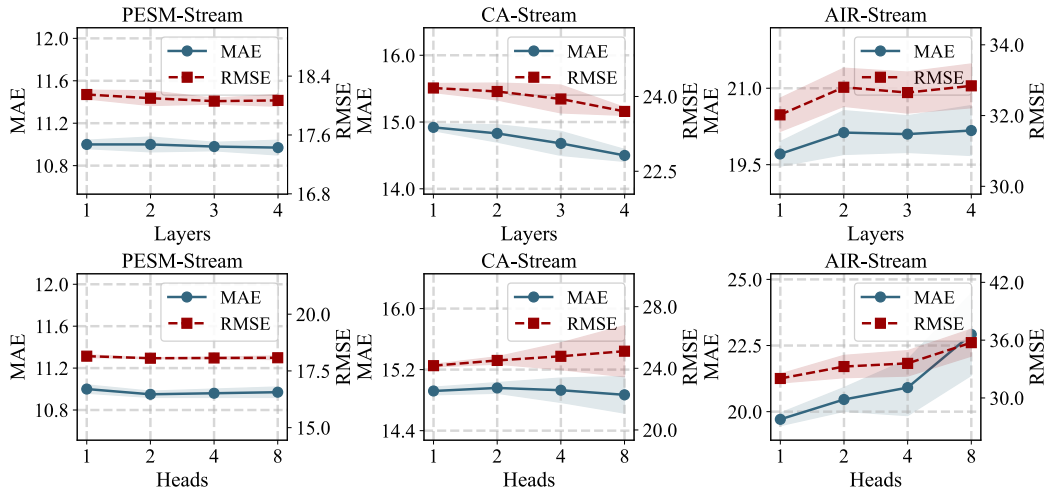

Figure 9: Additional Results of parameter experiments.

Figure 10. Even in the more challenging task of node increment, STBP 's contextual pattern bank effectively distinguishes and consolidates different spatio-temporal patterns, incorporating new patterns introduced by newly added nodes into the existing pattern clusters.

Figure 11 presents the results on Air-Stream. Compared to traffic flow data, the spatio-temporal patterns in this dataset exhibit more complex periodic and trend changes. Nonetheless, STBP continues to accurately differentiate and consolidate diverse patterns, indicating that its contextual pattern bank

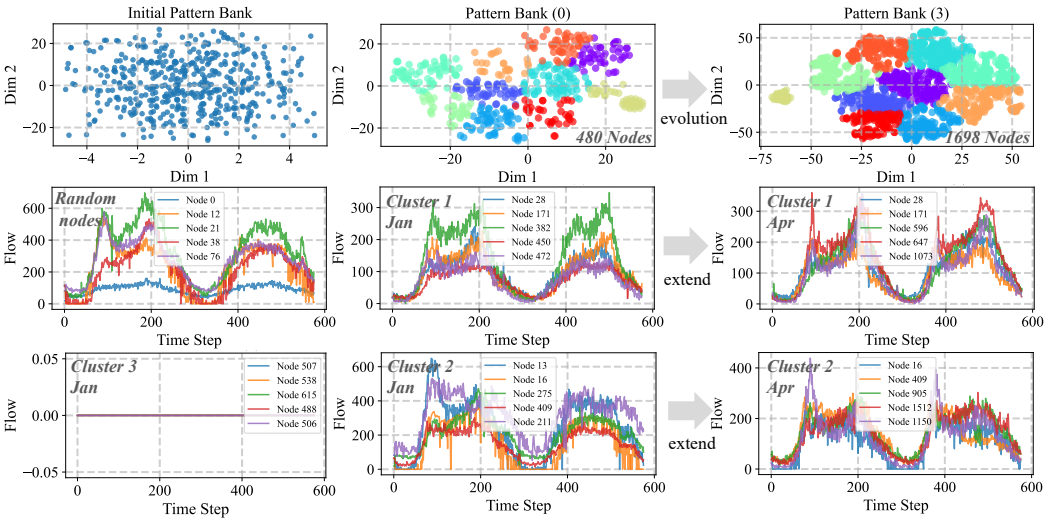

Figure 10: Case Study on CA-Stream.

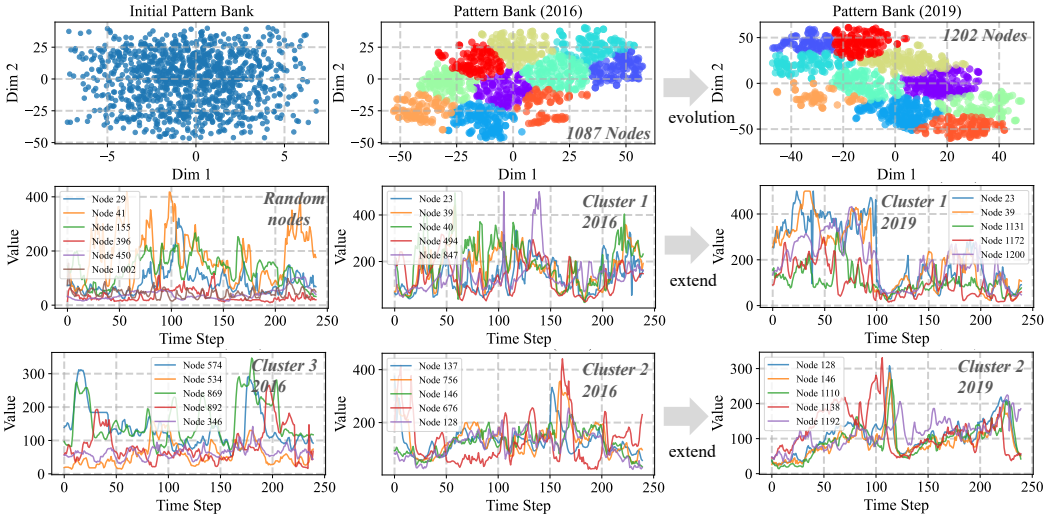

Figure 11: Case Study on AIR-Stream.

has adaptive inductive capabilities for various types of spatio-temporal patterns, independent of the specific dataset type. This mechanism enables STBP to exhibit greater flexibility and adaptability in CSTF tasks.

A.4.7 Efficiency Study

Figure 12 provides additional experiments assessing the efficiency and scalability of STBP. Overall, these results confirm that STBP achieves favorable scalability and efficiency, and that its linear-attention design and modular contextual pattern bank structure enable it to handle large-scale spatio-temporal graphs in continual learning settings.

A.5 Limitation

Despite the excellent performance of the STBP model on several benchmark datasets, there remain several theoretical and practical limitations that warrant further exploration. Firstly, current continual learning research, including this work, generally assumes an idealized scenario where all tasks processed during incremental learning come from the same or highly similar data domains. However, this assumption significantly deviates from the dynamic, complex, and diverse environments encountered in the real world. Cross-domain distribution shifts can introduce dual challenges when the model faces new tasks, including feature space mismatch and exacerbated catastrophic forgetting.

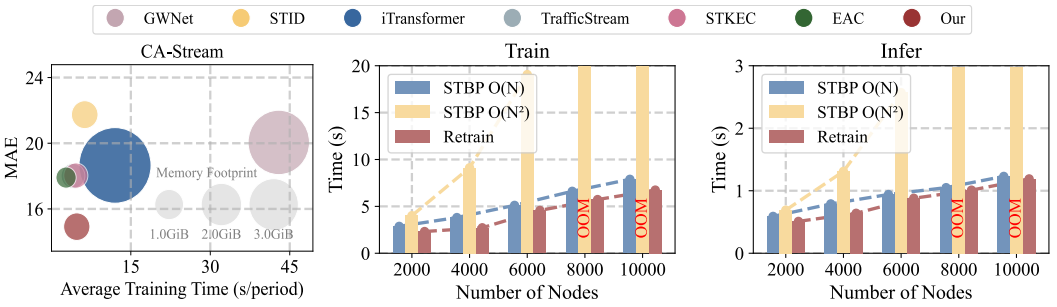

Figure 12: Additional efficiency comparison.

For instance, in intelligent transportation systems, when the model is applied to new urban traffic data with significant distributional differences, it must not only dynamically expand its spatio-temporal feature extraction capabilities but also develop effective representations of cross-domain invariant features. Although road network structures may vary across cities, certain microscopic traffic dynamics (e.g., traffic flow propagation speed, congestion formation mechanisms) could have inherent universality. How to construct a continual learning framework with domain adaptation capabilities, which can accurately distinguish domain-specific features from cross-domain shared features, will be a key breakthrough in improving the model's cross-domain generalization abilities.

A.6 Broader Impact

STBP, with its carefully designed general spatio-temporal backbone structure and contextual pattern bank expansion mechanism tailored for dynamic scenario changes, effectively achieves continual spatio-temporal forecasting. This approach demonstrates that the spatio-temporal backbone can serve as a stable infrastructure, consistently retaining the ability to model general spatio-temporal dependencies. When facing new or evolving scenarios, there is no need to retrain the backbone network. Instead, by introducing scalable parameters relevant to the current scenario, the model can rapidly adapt to new tasks.

Building on this concept, we aim to further explore the development of a spatio-temporal foundational model. This approach involves continuously training a unified backbone model with spatio-temporal data from multiple heterogeneous domains, thereby enhancing its spatio-temporal representational capacity. As data from various domains are continuously integrated and trained, the spatio-temporal foundational model will evolve, enabling efficient generalization and adaptation to entirely new scenarios or tasks by incorporating only a small number of additional parameters. Such a model holds the potential to benefit society by improving intelligent transportation through more accurate traffic forecasting and supporting climate resilience via advanced environmental modeling.

A.7 LLM Usage

In accordance with the ICLR 2026 policy on large language model (LLM) usage, we disclose that we used an LLM (ChatGPT) solely for the purpose of improving the grammar, clarity, and fluency of the manuscript. The content, structure, technical contributions, experiments, analysis, and all scientific writing were entirely conceived, drafted, and validated by the human authors. The LLM was not involved in research ideation, experimental design, data analysis, or any aspect of the scientific content creation. All outputs generated by the LLM were reviewed and edited by the authors to ensure accuracy and correctness. We confirm that no hidden prompts, prompt injections, or LLM-generated falsehoods were introduced in the manuscript, and all use of LLMs complies with the ICLR Code of Ethics.