Adaptive Frequency Pathways for Spatiotemporal Forecasting

Yanjun Qin[1*], Yuchen Fang[2*], Xinke Jiang[2], Hao Miao[3†], Xiaoming Tao[1,4]

[1]School of Computer Science and Technology, Research Center for Multimodal Information Perception and Intelligent Processing, Xinjiang University, Urumqi, China
[2]University of Electronic Science and Technology of China, Chengdu, China
[3]The Hong Kong Polytechnic University, Hongkong, China
[4]Tsinghua University, Beijing, China
qinyanjun@xju.edu.cn, fyclmiss@gmail.com, thinkerjiang@foxmail.com
hao.miao@polyu.edu.hk, taoxm@mail.tsinghua.edu.cn

Abstract

Spatiotemporal forecasting is a fundamental task in areas such as traffic flow prediction, environmental sensing, and urban planning. Recent advances have shown that decomposing temporal signals into multiple frequencies and modeling them jointly with spatial structures can significantly enhance forecasting performance. However, existing multi-frequency forecasting models still face two critical limitations. First, the importance of different temporal frequencies evolves over time, yet most models assume fixed or static frequency contributions. Second, spatial dependencies are inherently frequency-sensitive. For instance, low-frequency components often align with global spatial patterns, while high-frequency components tend to correspond to localized interactions. However, current approaches typically use a shared spatial information across all frequencies, introducing spatiotemporal inconsistency. To address these challenges, we propose a novel Adaptive Frequency Pathways (AdaFre) for spatiotemporal forecasting, which adaptively captures both dynamic frequency relevance and frequency-aligned spatial structures. AdaFre employs a multi-frequency routing mechanism to dynamically select and aggregate the most informative temporal frequency components, while associating each with its corresponding spatial representation derived from frequency-aware embeddings. Spatiotemporal backbones are then used to model each path independently before final aggregation. Extensive experiments on several real-world datasets demonstrate that AdaFre significantly outperforms state-of-the-art baselines.

1 Introduction

Spatiotemporal forecasting plays a vital role in real-world applications such as traffic management, environmental monitoring, and urban planning (Chen et al. 2021; Fang et al. 2025b; Miao et al. 2024). The core challenge lies in accurately modeling complex dependencies that evolve across both space and time (Liu et al. 2025a,b,c; Miao et al. 2025). Traditional approaches often adopt unified modeling paradigms, attempting to learn global spatiotemporal patterns from raw observations (Liu et al. 2023). However, such monolithic designs tend to overlook the inherent multi-scale nature of spatiotemporal dynamics, limiting their capacity to generalize across varying temporal rhythms and spatial granularities (Fang et al. 2023b; Liu et al. 2025d).

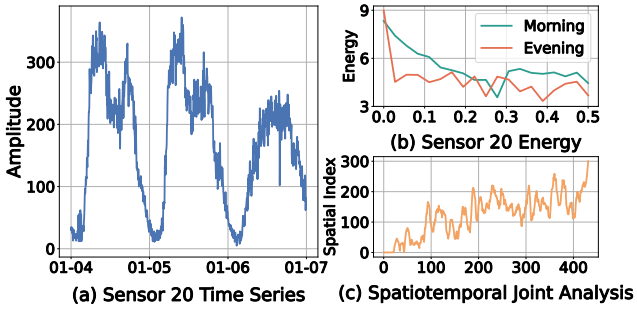


Figure 1: Spatio-temporal frequency analysis. (a) A representative sensor time series. (b) Varying frequency energy patterns across time demonstrate the dynamic evolution of temporal frequency importance. (c) Joint analysis of temporal and spatial frequency indices indicates that spatial dependencies are frequency-specific.

To overcome these limitations, recent studies have proposed decomposition-based multi-path models that break time series into components at different frequencies, and then jointly encode spatial dependencies under each temporal scale (Cao et al. 2025). This class of methods has demonstrated substantial performance gains by allowing models to capture spatiotemporal dependencies at multiple resolutions. For instance, extracting periodic patterns from low-frequency trends and detecting transient anomalies from high-frequency components. By explicitly disentangling frequency-specific information, these models have achieved more robust and interpretable forecasting results. Despite these improvements, existing frequency-aware spatiotemporal models still face two fundamental challenges.

Temporal Frequency Dynamics. A critical yet underexplored aspect of spatiotemporal forecasting is the dynamic importance of temporal frequency components. Unlike static signals, real-world time series often exhibit non-stationary frequency characteristics due to evolving external condi-

*These authors contributed equally.
†Corresponding author.

tions, human behaviors, or event-driven disruptions. For example, in traffic systems, low-frequency components (such as morning and evening commute patterns) dominate during rush hours, while high-frequency components (such as sudden traffic accidents or road closures) become more relevant during off-peak periods or special events. As illustrated in Figure 1 (b), the frequency energy distribution of a traffic sensor in the PeMSD4 dataset varies significantly across different times of day. Specifically, low-frequency components dominate during early morning hours (reflecting stable commuting trends), whereas high-frequency energy spikes in the evening, possibly reflecting irregular behaviors, congestion, or anomalies. This dynamic shift in frequency importance suggests that a fixed-weighted fusion of all frequency components, as adopted by most existing models, fails to capture the changing predictive relevance of each component over time. Instead, what is needed is a mechanism that can adaptively select and weight frequency components based on the current temporal context

Frequency-Specific Spatial Dependency. In addition to temporal dynamics, spatial dependencies are not uniform across temporal frequencies. Different temporal frequency components often correspond to qualitatively different spatial interaction patterns (Liu et al. 2025e). For instance, low-frequency trends—such as periodic weather changes or daily traffic flows—typically reflect broad, globally coherent phenomena, affecting most locations in a correlated manner. In contrast, high-frequency variations—such as local accidents, construction activity, or sensor malfunctions—are often spatially localized, impacting only small subsets of nodes or regions. This phenomenon is evident in Figure 1 (c), where we analyze the joint distribution of temporal and spatial frequency indices. The visualization reveals that temporal high-frequency components co-occur with higher spatial frequency indices, suggesting more fragmented or localized spatial influence. Conversely, low-frequency components exhibit smoother spatial propagation, reflected by their lower spatial frequency index. However, most existing spatiotemporal models rely on a shared or static spatial graph structure across all temporal components, implicitly assuming that spatial dependencies are independent of temporal frequency. This frequency-agnostic spatial prior introduces inconsistency: it may enforce global interactions even when only local, high-frequency patterns are relevant, or ignore long-range influences in low-frequency regimes.

To address these limitations, we propose a novel Adaptive Frequency Pathways (AdaFre) for spatiotemporal forecasting, which can model frequency-aware temporal dynamics and spatial structures in an adaptive and unified manner. AdaFre first employs a multi-frequency routing mechanism that adaptively selects the most relevant temporal frequency components, thereby mitigating the impact of dynamically irrelevant or noisy frequencies. Then, for each selected frequency, AdaFre further associates it with a tailored spatial embedding, capturing the corresponding spatial dependency at that frequency level and avoiding conflicts arising from frequency-agnostic spatial graphs. Finally, frequency-specific spatiotemporal backbones are employed to encode and predict each frequency component separately, followed by a learned aggregation across frequencies to produce the final prediction. Extensive experiments on several spatiotemporal datasets demonstrate that AdaFre consistently outperforms state-of-the-art methods.

In summary, this paper makes following contributions:

- We introduce a novel AdaFre with adaptive frequency pathways that dynamically selects most relevant temporal frequency components, allowing the model to adaptively focus on frequency-specific patterns over time.
- We align spatial dependencies with temporal frequencies, enabling the model to capture frequency-specific spatial structures and mitigate spatiotemporal inconsistency.
- Extensive experiments on multiple real-world spatiotemporal datasets demonstrate that our approach achieves substantial improvements over state-of-the-art baselines, especially in scenarios with strong frequency dynamics.

2 Preliminaries

2.1 Spatiotemporal Data

The spatiotemporal data is collected from a system of N spatial nodes (*e.g.*, traffic sensors, weather stations, or monitoring sites) over T discrete time steps. Let $X \in \mathbb{R}^{T\times N\times C}$ denote the historical observations, where T is the number of past time steps, N is the number of spatial nodes, C is the number of features per node (*e.g.*, traffic speed, temperature). Each slice $X_t \in \mathbb{R}^{N\times C}$ corresponds to the observations of all spatial nodes at time step t. To model spatial dependencies among the N nodes, we construct a spatial graph $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_N\}$ is the set of nodes, $E \subseteq V \times V$ is the set of edges representing spatial connections. The graph structure is encoded by an adjacency matrix $A \in \mathbb{R}^{N\times N}$, where A_{ij} indicates the strength of spatial dependency between node i and node j. The adjacency matrix can be defined based on physical distance, functional correlation, or learned adaptively from data (Li et al. 2018).

2.2 Spatiotemporal Forecasting

Given the past T observations $X \in \mathbb{R}^{T\times N\times C}$, the goal of spatiotemporal forecasting is to predict the future $\hat{T}$ time steps of all nodes:

$$Y = \{X_{T+1}, X_{T+2}, \ldots, X_{T+\hat{T}}\} \in \mathbb{R}^{\hat{T}\times N\times C}, \quad (1)$$

which can be predicted through a learnd function:

$$\hat{Y} = F_\theta(X, A), \quad (2)$$

where $F_\theta(\cdot)$ is a spatiotemporal model that captures both temporal dynamics and spatial correlations to produce accurate predictions for each node at each future time step.

3 Method

As shown in Figure 2, our Adaptive Frequency Pathway (AdaFre) is composed of two main components: a spatiotemporal multi-frequency decomposition and an adaptive pathway mechanism. The first part aims to capture multi-resolution temporal dynamics and corresponding spatial dependencies. It begins by decomposing the input time series into multiple frequency components using a temporal

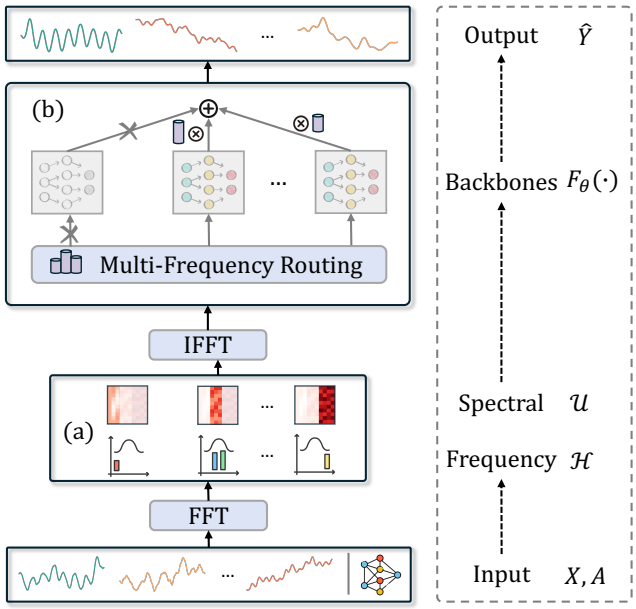


Figure 2: The overview of the proposed AdaFre

transform, followed by generating frequency-specific spectral spatial embeddings. The second part focuses on adaptively selecting and integrating frequency paths. A learnable routing module dynamically identifies the most informative frequencies for a given context, and a weighted fusion module aggregates their outputs.

3.1 Spatiotemporal Multi-Frequency Network

To capture temporal dynamics at multiple resolutions and their corresponding spatial structures, we first transform the spatiotemporal data into the frequency domain and extract frequency-specific representations along with matched spectral spatial embeddings. These are then jointly modeled through a spatiotemporal backbone.

Frequency Decomposition via Temporal Transform. Spatiotemporal data often exhibit behaviors at multiple temporal scales. These multi-scale temporal dynamics are not only informative for forecasting but also heterogeneous, *i.e.*, different time scales may require distinct modeling strategies. While some recent works attempt to model temporal hierarchies through sequential or multi-resolution designs, they often do so implicitly, without explicitly separating or exploiting frequency information. To address this, we adopt the Discrete Fourier Transform (DFT) to decompose temporal signals into constituent frequencies, which projects time-domain data into the frequency domain using a well-defined and invertible basis. Compared to wavelet or learned frequency decompositions, DFT separates different components into distinct frequency channels, which can be directly interpreted and manipulated. This design facilitates reduced interference between low- and high-frequency signals and specialized processing of different temporal patterns for downstream control (*e.g.*, selective modeling or fusion). Specifically, given an input spatiotemporal sequence, we first apply the Discrete Fourier Transform (DFT) along the temporal axis:

$$\hat{X} = \text{FFT}(X, \text{dim} = 0) \in \mathbb{C}^{T \times N \times C}. \tag{3}$$

The DFT produces a symmetric spectrum of complex-valued coefficients that encode temporal periodicity at each spatial node. Due to Hermitian symmetry for real signals, only the first $\lfloor \frac{T}{2} \rfloor + 1$ frequency bins are unique and meaningful. We denote this as the valid frequency range $\hat{X}_{[0:\lfloor \frac{T}{2} \rfloor + 1]}$. We divide this valid frequency range into P non-overlapping frequency bands and each frequency band I_p (for $p = 1, \ldots, P$) is defined as a contiguous index set:

$$I_p = \left\{ \left\lfloor \frac{(p-1)(\lfloor \frac{T}{2} \rfloor + 1)}{P} \right\rfloor, \ldots, \left\lfloor \frac{p(\lfloor \frac{T}{2} \rfloor + 1)}{P} \right\rfloor \right\}, \tag{4}$$

so that each band covers approximately $\left\lfloor \frac{\lfloor \frac{T}{2} \rfloor + 1}{P} \right\rfloor$ frequency components. This ensures that the entire frequency space is evenly divided, with band I_1 covering the lowest-frequency (i.e., long-term trend), band I_P capturing the highest-frequency dynamics (e.g., noise or rapid changes), and intermediate bands modeling medium-range temporal fluctuations. For each band p, we zero out the DFT coefficients outside I_p, and apply the Inverse FFT to project the frequency-limited signal back into the time domain:

$$H^{(p)} = \text{Re}\left(\text{IFFT}\left(\hat{X}_{I_p}\right)\right) \in \mathbb{R}^{T \times N \times C}. \tag{5}$$

Each resulting $H^{(p)}$ is a time-domain sequence that only preserves the temporal patterns within the frequency band I_p. Collectively, these form a set of band-limited temporal views of the same signal:

$$\mathcal{H} = \left\{ H^{(1)}, H^{(2)}, \ldots, H^{(P)} \right\}. \tag{6}$$

These frequency-specific representations provide diverse and disentangled temporal features, which are then passed to downstream adaptive pathway mechanism for dynamic frequency selection and fusion.

Frequency-Aware Graph Spectral Embeddings. Spatiotemporal data often reside on irregular spatial graphs, where node-wise correlations are best captured by the underlying graph topology. To encode this spatial structure into learning, graph spectral embeddings, typically derived from the Laplacian eigendecomposition, have been widely adopted in spatiotemporal forecasting tasks (Jiang et al. 2023). These embeddings act as frequency-like bases that reveal multi-scale diffusion patterns and spatial heterogeneity, greatly benefiting spatial relational modeling. While spectral embeddings are effective, existing approaches typically adopt a single set of graph spectral features for all temporal components, implicitly assuming that spatial patterns are invariant across time frequencies. This assumption, however, is often unrealistic. In real-world scenarios low-frequency temporal trends often reflect global, stable spatial interactions, while high-frequency components may correspond to

localized or rapidly changing spatial dependencies. Applying a shared spectral embedding across all temporal frequency bands introduces an inductive bias mismatch, *i.e.*, the spatial representation may amplify irrelevant correlations or suppress informative local signals and lead to representation inconsistency and noisy error propagation. To bridge this inconsistency, we propose to co-decompose the graph spectrum into P partitions, aligned with the temporal frequency bands. Concretely, given the normalized Laplacian matrix $L \in \mathbb{R}^{N\times N}$, we compute its eigendecomposition:

$$L = U\Lambda U^\top, \tag{7}$$

where $U \in \mathbb{R}^{N\times N}$ contains the eigenvectors (graph Fourier bases), and $\Lambda \in \mathbb{R}^{N\times N}$ is a diagonal matrix of eigenvalues. Then we partition the eigenvectors into P contiguous spectral groups:

$$U^{(p)} = U\left[:, \hat{I}_p\right], \quad p = 1, 2, \ldots, P, \tag{8}$$

where each index set $\hat{I}_p$ covers $\lfloor \frac{N}{P} \rfloor$ consecutive eigenvectors. The ordering follows the eigenvalue magnitudes (from low to high graph frequency), ensuring that $U^{(1)}$ captures global spatial modes, $U^{(P)}$ captures high-frequency local patterns, and intermediate $U^{(p)}$ model meso-scale spatial dependencies. This results in a set of graph spectral embeddings $\mathcal{U} = \{U^{(p)}\}_{p=1}^{P}$, where each is specialized for a particular temporal frequency band. Such pairing enables us to inject frequency-specific spatial priors into the model in a fully disentangled fashion.

3.2 Adaptive Pathways

Although the multi-frequency representations enable comprehensive temporal decomposition, the model so far treats all frequencies equally, regardless of the specific spatiotemporal context. However, in real-world systems, the importance of different frequency components often varies depending on the scenario, *e.g.*, rush hour traffic patterns or seasonal climate shifts. Thus, we propose an adaptive pathway to dynamically identify and utilize the most relevant frequency components for each input sequence, enhancing both efficiency and accuracy.

Multi-Frequency Routing. Let the input sequence be $X \in \mathbb{R}^{T\times N\times C}$. Given P extracted frequency-specific representations $\{H^{(p)}\}_{p=1}^{P}$, we compute their relevance scores of each node via a learned score function $g_\phi(\cdot)$, which maps input X into a soft frequency importance distribution:

$$R = g_\phi(X) \in \mathbb{R}^{N\times P}. \tag{9}$$

Then we apply softmax normalization with temperature scaling to derive probability. The process for node i can be formulated as:

$$\hat{r}_{ip} = \frac{\exp(r_{ip}/\tau)}{\sum_{j=1}^{K}\exp(r_{ij}/\tau)}, \tag{10}$$

and select the top-K frequency indices $\tilde{I} = \mathrm{TopK}(\hat{R}, K)$ on the frequency dim. This mechanism allows the model to prioritize frequencies most aligned with the current dynamics, while discarding irrelevant components.

Frequency-Specific Spatiotemporal Modeling. Each pair $(H^{(\tilde{I}_k)}, U^{(\tilde{I}_k)})$ in top-K frequencies captures temporal behavior and corresponding spatial structure at the $\tilde{I}_k$-th frequency. We feed each pair into $\tilde{I}_k$-th spatiotemporal backbone network $f_\theta^{(\tilde{I}_k)}$. In this paper, we adopt STID (Shao et al. 2022) to highlight the improvements introduced by our frequency adaptation mechanism, avoiding confounding effects from overly complex backbones:

$$\hat{Y}^{(\tilde{I}_k)} = f_\theta^{(\tilde{I}_k)}(H^{(\tilde{I}_k)}, U^{(\tilde{I}_k)}). \tag{11}$$

Frequency-Aware Fusion. Even among the selected frequency representations, not all contribute equally. For instance, low-frequency components may reflect long-term trends, while high-frequency ones carry short-term bursts. Therefore, we design a context-aware fusion mechanism that dynamically weighs selected frequency-specific predictions based on their utility. After obtaining the predictions $\hat{Y}^{(\tilde{I}_k)}$ from the spatiotemporal backbone for selected frequency indices $\tilde{I}_k$, we perform a weighted aggregation:

$$\hat{Y} = \sum_{k=1}^{|\tilde{I}|}\alpha_k \cdot \hat{Y}^{(\tilde{I}_k)}, \quad \sum_{k=1}^{|\tilde{I}|}\alpha_k = 1. \tag{12}$$

The fusion weights are computed via softmax normalization. This soft fusion step ensures that the model captures complementary patterns across frequencies while maintaining flexibility in how much each contributes.

3.3 Training Objective

Supervised Forecasting Loss To guide the backbone toward accurate frequency-specific modeling, we apply the mean absolute error (MAE) loss the final fused prediction:

$$\mathcal{L}_{\text{pred}} = \|\hat{Y} - Y\|_2, \tag{13}$$

where Y is the ground truth future observation. This loss encourages each frequency branch to learn effective spatiotemporal representations relevant to its spectral focus.

Frequency Balancing Loss Without additional constraints, the frequency router may overfit to a few dominant frequencies (*e.g.*, always selecting low-frequency trends), ignoring others even if they are useful in rare but important cases. To avoid this, we introduce a frequency usage regularizer to encourage balanced exploration and utilization of all frequency paths. Let $\tilde{r}_p$ be the average selection probability for the p-th frequency over a mini-batch:

$$\tilde{r}_p = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{N}\hat{r}_{jp}^{(i)}, \tag{14}$$

where B is the batch size. We define the balancing loss as:

$$\mathcal{L}_{\text{bal}} = \sum_{p=1}^{P}\left(\tilde{r}_p - \frac{1}{P}\right)^2. \tag{15}$$

This term penalizes deviation from uniform frequency usage across the training set, promoting diverse routing decisions.

Datasets	Nodes	Interval	Length	Time Span
PeMSD3	358	5 mins	26208	09/2018-11/2018
PeMSD4	307	5 mins	16992	01/2018-02/2018
PeMSD7	883	5 mins	28224	05/2017-08/2017
PeMSD8	170	5 mins	17856	07/2016-08/2016

Table 1: Dataset statistics.

Overall Objective The final objective combines the main spatiotemporal prediction loss with the balancing loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{bal}}. \tag{16}$$

4 Experiments

4.1 Experimental Setting

Datasets We evaluate our model on four widely used real-world spatiotemporal datasets: PeMSD3, PeMSD4, PeMSD7, and PeMSD8, all collected from the Caltrans Performance Measurement System (PeMS) in California. These datasets cover diverse urban and highway regions with varying numbers of sensor nodes and temporal spans, thereby offering a comprehensive benchmark for spatiotemporal forecasting. Following previous works (Song et al. 2020), we split each dataset into 60% training, 20% validation, and 20% testing based on chronological order. To stabilize model training, we apply Z-Score normalization using the mean and standard deviation computed from the training set. For spatial structure, the adjacency matrix is constructed based on pairwise physical distances between sensor stations, with a Gaussian kernel applied to define edge weights. This setting aligns with widely adopted practices in spatiotemporal GNN literature (Li et al. 2018). The detailed statistics of each dataset are provided in Table 1.

Baselines To comprehensively evaluate the effectiveness of our proposed AdaFre, we compare it against 22 baseline models spanning a wide range of modeling paradigms. We first include traditional non-deep learning approaches such as HA (Pan, Demiryurek, and Shahabi 2012), ARIMA (Kumar and Vanajakshi 2015), VAR (Chandra and Al-Deek 2009), and SVR (Castro-Neto et al. 2009). We then consider deep learning models that focus purely on temporal modeling, including LSTM (Zhao et al. 2017) and TCN (Bi et al. 2021). For static spatiotemporal deep models, we adopt STGCN (Yu, Yin, and Zhu 2018), DCRNN (Li et al. 2018), GWN (Wu et al. 2019), STSGCN (Song et al. 2020), AGCRN (Bai et al. 2020), STNorm (Deng et al. 2021), STFGNN (Li and Zhu 2021), STGODE (Fang et al. 2021), and STID (Shao et al. 2022). To capture dynamically changing spatiotemporal patterns, we also evaluate dynamic modeling baselines such as ASTGCN (Guo et al. 2019), PDFormer (Jiang et al. 2023), STAEFormer (Liu et al. 2023), and HimNet (Dong et al. 2024). In addition, to verify the contribution of frequency-aware modeling, we include decomposition-based models STWave (Fang et al. 2023b) and STDN (Cao et al. 2025).

Metrics To evaluate the predictive accuracy of all models, we adopt three widely used metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). These metrics provide a comprehensive assessment of both absolute and relative forecasting performance across different datasets.

Implementation Details All models are implemented using the PyTorch framework and trained with the Adam optimizer. Our model is trained for 100 epochs with an initial learning rate of 0.001. The learning rate is halved at the 2nd, 50th, and 80th epochs to improve convergence stability. Following previous works, we forecast the next 12 time steps (corresponding to 1 hour) but significantly extend the historical input to 288 steps (*i.e.*, 1 day) to better capture temporal patterns and highlight the role of frequency decomposition. For hyperparameter settings, we divide the input time series into $P = 4$ frequency bands and select the top $K = 2$ bands in the frequency routing stage, enabling the model to adaptively focus on the most informative frequency components.

4.2 Main Results

The main experimental results are summarized in Table 2, where our model, AdaFre, is compared against 22 baseline methods across four real-world spatiotemporal datasets. From the table, we derive the following key observations:

Superiority of Dynamic Modeling. Models that dynamically capture the evolving spatiotemporal dependencies (*e.g.*, STAEFormer, PDFormer, and HimNet) generally outperform static counterparts (*e.g.*, GWN and AGCRN). This confirms the necessity of dynamic modeling, as spatial relationships often drift over time, making static assumptions suboptimal for future predictions.

Effectiveness of Frequency Decomposition. Decomposition-based models (*e.g.*, STWave and STDN) outperform vanilla deep learning models, demonstrating that modeling different temporal components (*e.g.*, trend and seasonal) separately can reduce interference among them and yield more accurate predictions. This aligns with the intrinsic hierarchical nature of time series data.

Consistent Superiority of AdaFre. Our model outperforms all baselines across all metrics and datasets, showing not only strong accuracy but also robustness across diverse settings. We attribute this consistent advantage to the combination of both dynamic modeling and multi-frequency decomposition, as well as our novel design that incorporates spatial hierarchy through frequency-aligned spectral embeddings. This allows AdaFre to fully capture the complex and evolving spatiotemporal structures of the data.

4.3 Ablation Study

To assess the effectiveness of each major component in AdaFre, we conduct ablation studies by evaluating the following three variants: **w/o FD**: Removing frequency decomposition, where the input is routed directly without band separation. **w/o FSE**: Removing frequency-aware spectral embeddings, by sharing the same spectral embedding across all

Methods	PeMSD3			PeMSD4			PeMSD7			PeMSD8		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
HA	31.58	52.39	33.78%	38.03	59.24	27.88%	45.12	65.64	24.51%	34.86	59.24	27.88%
ARIMA	35.41	47.59	33.78%	33.73	48.80	24.18%	38.17	59.27	19.46%	31.09	44.32	22.73%
VAR	23.65	38.26	24.51%	24.54	38.61	17.24%	50.22	75.63	32.22%	19.19	29.81	13.10%
SVR	21.97	35.29	21.51%	28.70	44.56	19.20%	32.49	50.22	14.26%	23.25	36.16	14.64%
LSTM	21.33	35.11	23.33%	26.77	40.65	18.23%	29.98	45.94	13.20%	23.09	35.17	14.99%
TCN	19.32	33.55	19.93%	23.22	37.26	15.59%	32.72	42.23	14.26%	22.72	35.79	14.03%
STGCN	17.55	30.42	17.34%	21.16	34.89	13.83%	25.33	39.34	11.21%	17.50	27.09	11.29%
DCRNN	17.99	30.31	18.34%	21.22	33.44	14.17%	25.22	38.61	11.82%	16.82	26.36	10.92%
GWN	19.12	32.77	18.89%	24.89	39.66	17.29%	26.39	41.50	11.97%	18.28	30.05	12.15%
ASTGCN	17.34	29.56	17.21%	22.93	35.22	16.56%	24.01	37.87	10.73%	18.25	28.06	11.64%
STSGCN	17.48	29.21	16.78%	21.19	33.65	13.90%	24.26	39.03	10.21%	17.13	26.80	10.96%
AGCRN	15.98	28.25	15.23%	19.83	32.26	12.97%	22.37	36.55	9.12%	15.95	25.22	10.09%
STNorm	15.32	25.93	14.37%	18.96	30.98	12.69%	20.50	34.66	8.75%	15.41	24.77	9.76%
STFGNN	16.77	28.34	16.30%	20.48	32.51	16.77%	23.46	36.60	9.21%	16.94	26.25	10.60%
STGODE	16.50	27.84	16.69%	20.84	32.82	13.77%	22.59	37.54	10.14%	16.81	25.97	10.62%
STID	15.33	27.40	16.40%	18.38	29.95	12.04%	19.61	32.79	8.30%	14.21	23.28	9.27%
STWave	14.93	26.50	<u>15.05</u>%	18.50	30.39	12.43%	19.94	33.88	8.38%	<u>13.42</u>	23.40	8.90%
PDFormer	14.94	<u>25.39</u>	15.82%	18.36	30.03	12.00%	19.97	32.95	8.55%	13.58	23.41	9.05%
STAEFormer	15.35	27.55	15.18%	18.22	30.18	<u>11.98</u> %	<u>19.14</u>	<u>32.60</u>	<u>8.01</u>%	13.46	23.25	<u>8.88</u>%
HimNet	15.11	26.56	15.49%	<u>18.14</u>	<u>29.88</u>	12.00%	19.21	32.75	8.03%	13.57	<u>23.22</u>	8.98%
STDN	<u>14.89</u>	25.66	15.68%	18.40	30.41	12.21%	20.08	33.73	9.29%	13.67	23.49	9.22%
AdaFre	**14.27**	**24.22**	**15.04**%	**17.89**	**29.76**	**11.97**%	**18.58**	**31.55**	**7.85**%	**12.99**	**22.32**	**8.88**%
w/o FD	14.51	24.36	16.24%	18.35	30.13	12.18%	19.05	31.86	8.42%	13.50	22.46	9.45%
w/o FSE	14.44	24.34	16.19%	18.14	30.12	12.06%	18.78	31.68	7.96%	13.26	22.39	9.23%
w/o AP	14.37	24.31	15.55%	18.28	29.99	12.79%	18.70	31.58	7.93%	13.41	22.42	9.44%

Table 2: Performance comparison of AdaFre and baselines on four datasets. **Bold**: Best, <u>underline</u>: Second best.

routing branches. **w/o AP**: Removing adaptive path selection and importance-based weighting. The results in Table 2 reveal several key insights:

Benefit of Frequency Decomposition. Without decomposition, the model's performance degrades significantly. This suggests that improvements are not merely due to increased parameters, but from explicitly separating temporal components to reduce interference.

Importance of Spectral Coordination. Using shared spectral embeddings across all routes leads to a noticeable performance drop. This highlights the necessity of aligning temporal frequency content with spatial patterns, *i.e.*, when spectral embeddings do not reflect the localized temporal semantics of each band, the model fails to capture the joint spatiotemporal structure effectively.

Effectiveness of Adaptive Pathway. Simply aggregating all routing outputs without selective weighting results in suboptimal predictions. This demonstrates that the importance of different frequency bands is dynamically varying across both time and space, and a static aggregation cannot adapt to these changes. The adaptive selection mechanism is essential for exploiting the most relevant pathways at each timestep and location.

4.4 Hyperparameter Analysis

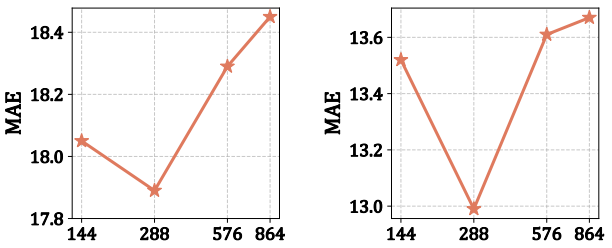
We conduct a systematic study on three key hyperparameters in AdaFreas shown in Figure 3: the historical input length T, the number of frequency bands P, and the number of selected frequency bands K in the adaptive routing phase.

Input length T. We evaluate four settings: 144 (half day), 288 (one day), 576 (two days), and 864 (three days). The results show that using 288 achieves the best performance. A shorter input fails to capture sufficient periodicity, while a longer input introduces redundant or noisy information that may not contribute to immediate prediction.

Number of frequency bands P. We vary P from 2 to 5 with $K = 2$. The best performance is achieved at $P = 4$, indicating that the temporal signals benefit from being decomposed into a moderate number of frequency bands. Too few bands result in coarse granularity and mixed semantics and too many lead to data fragmentation and insufficient information in each sub-band.

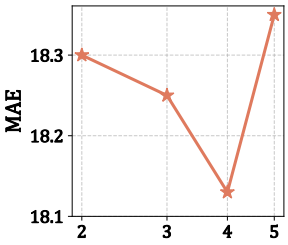
4.5 Efficiency Analysis

We compare the training time and GPU memory consumption of AdaFre with a wide range of baselines. As shown in Table 3, our method achieves a favorable balance between accuracy and efficiency. Specifically, AdaFre's training time and memory usage are only slightly higher than STID, a

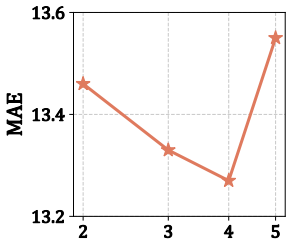


(a) T on PeMSD4

(b) T on PeMSD8



(c) P on PeMSD4



(d) P on PeMSD8

Figure 3: The influence of input length and frequency bands.

Methods	#Params	Time/Batch	Time/Epoch	Mem Usage
STID	123K	10ms	15s	1679MB
GWN	309K	48ms	69s	2046MB
AGCRN	763K	63ms	102s	2679MB
HimNet	11B	103ms	189s	6784MB
STDN	3B	73ms	118s	3533MB
AdaFre	337K	14ms	23s	2137MB

Table 3: Efficiency comparison on PeMSD4 dataset.

lightweight fully-connected model known for its simplicity. Compared to more sophisticated architectures such as STAEFormer, PDFFormer, and STFGNN, our method requires significantly less computational resources. This efficiency arises from the sparse selection strategy where only the most relevant K frequency paths (out of P) are activated per sample. This avoids unnecessary computation on uninformative branches and effectively reduces model overhead. Moreover, rather than using complex global spatiotemporal attention or full-graph propagation (which are both costly), we divide the problem into localized frequency subspaces.

5 Related Work

5.1 Spatiotemporal Forecasting

Early spatiotemporal forecasting relied heavily on statistical and classical machine learning techniques, including ARIMA (Kumar and Vanajakshi 2015), Kalman filters (Kumar 2017), and Gaussian processes (Zhao and Sun 2016). These models typically assume stationarity and linearity and are limited in capturing nonlinear temporal dynamics or intricate spatial interactions. Machine learning methods such as SVMs (Feng et al. 2018) and random forests (Evans, Waterson, and Hamilton 2019) improve on nonlinear modeling capabilities but require extensive feature engineering and generally fail to represent spatial relations explicitly. With the rise of deep learning, static GNN-based spatiotemporal frameworks became mainstream (Fang et al. 2021; Deng et al. 2021; Shao et al. 2022). These models represent sensors or spatial units as graph nodes and use graph convolution layers to capture spatial dependencies, combined with sequential temporal modules such as RNNs or 1D CNNs for time modeling (Li et al. 2018; Yu, Yin, and Zhu 2018; Wu et al. 2019; Bai et al. 2020). More recent works leverage spatiotemporal Transformer architectures to model long-range dependencies and dynamic interaction patterns (Guo et al. 2019; Jiang et al. 2023; Liu et al. 2023; Dong et al. 2024; Cao et al. 2025; Yi et al. 2024; Zhou et al. 2025; Fang et al. 2025a). These Transformer variants dynamically model interactions over time and space but typically assume fixed temporal decomposition.

5.2 Spatiotemporal Decomposition

To handle signals of different temporal scales, several works decompose time series into trend, seasonal, and residual components (via Fourier, wavelet, VMD or SSA methods) and model each separately. For example, TFD (Yang, Cheng, and Li 2022) extracts interpretable modes for downstream forecasting through variational mode decomposition. StemGNN (Cao et al. 2020) jointly transforms temporal and spatial signals into the spectral domain using DFT and Graph Fourier Transform to model temporal and spatial correlations in unified spectral space. Other systems such as STWave (Fang et al. 2023a) and DSTAGNN (Lan et al. 2022) apply wavelet-based multi-scale decomposition followed by graph-based attention modules to capture interactions across time scales. Independent of decomposition, some studies design multi-scale spatial graphs, either hierarchical or multiresolution, to capture local and global structural dependencies. These methods build multiple graphs ranging from local neighborhoods to global connectivity and fuse information across graph views to enrich spatial modeling (Wang et al. 2021, 2022; Fang et al. 2023b; Tian et al. 2023). Such multi-graph strategies enable richer spatial representation but often remain agnostic to temporal frequency components.

6 Conclusion

In this paper, we propose AdaFre, a novel spatiotemporal forecasting model that adaptively routes input signals across frequency-specific pathways. By explicitly decomposing time series into multiple frequency bands, AdaFre captures latent hierarchical structures across both temporal and spatial dimensions. The model further enhances flexibility by dynamically selecting the most relevant frequency components through a sparse routing mechanism, allowing it to adjust to the evolving importance of different signals over time and space. Extensive experiments on four real-world datasets demonstrate that AdaFre achieves consistent state-of-the-art performance across all benchmarks. We believe AdaFre offers a generalizable paradigm that can benefit a wide range of time-series forecasting tasks beyond spatiotemporal prediction in the future.

Acknowledgments

Supported by the National Natural Science Foundation of China Nos.62306164; "Tianchi Yingcai" Introduction Program; Basic Research Project of the Autonomous Region's Universities' Basic Research Operating Funds (XJEDU2025J001); Open Research Fund Program of Beijing National Research Center for Information Science and Technology; Key Research and Development Project of the Autonomous Region (2024B03028).

References

Bai, L.; Yao, L.; Li, C.; Wang, X.; and Wang, C. 2020. Adaptive graph convolutional recurrent network for traffic forecasting. In *NeurIPS*, 17804–17815.

Bi, J.; Zhang, X.; Yuan, H.; Zhang, J.; and Zhou, M. 2021. A hybrid prediction method for realistic network traffic with temporal convolutional network and LSTM. *IEEE Transactions on Automation Science and Engineering*, 1869–1879.

Cao, D.; Wang, Y.; Duan, J.; Zhang, C.; Zhu, X.; Huang, C.; Tong, Y.; Xu, B.; Bai, J.; Tong, J.; et al. 2020. Spectral temporal graph neural network for multivariate time-series forecasting. In *NeurIPS*, 17766–17778.

Cao, L.; Wang, B.; Jiang, G.; Yu, Y.; and Dong, J. 2025. Spatiotemporal-aware Trend-Seasonality Decomposition Network for Traffic Flow Forecasting. In *AAAI*, 11463–11471.

Castro-Neto, M.; Jeong, Y.-S.; Jeong, M.-K.; and Han, L. D. 2009. Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. *Expert systems with applications*, 6164–6173.

Chandra, S. R.; and Al-Deek, H. 2009. Predictions of freeway traffic speeds and volumes using vector autoregressive models. *Journal of Intelligent Transportation Systems*, 53–72.

Chen, Y.; Li, X.; Cong, G.; Bao, Z.; Long, C.; Liu, Y.; Chandran, A. K.; and Ellison, R. 2021. Robust road network representation learning: When traffic patterns meet traveling semantics. In *CIKM*, 211–220.

Deng, J.; Chen, X.; Jiang, R.; Song, X.; and Tsang, I. W. 2021. St-norm: Spatial and temporal normalization for multi-variate time series forecasting. In *SIGKDD*, 269–278.

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 *SIGKDD*, 631–641.

Evans, J.; Waterson, B.; and Hamilton, A. 2019. Forecasting road traffic conditions using a context-based random forest algorithm. *Transportation planning and technology*, 554–572.

Fang, Y.; Liang, Y.; Hui, B.; Shao, Z.; Deng, L.; Liu, X.; Jiang, X.; and Zheng, K. 2025a. Efficient large-scale traffic forecasting with transformers: A spatial data management perspective. In *SIGKDD*, 307–317.

Fang, Y.; Miao, H.; Liang, Y.; Deng, L.; Cui, Y.; Zeng, X.; Xia, Y.; Zhao, Y.; Pedersen, T. B.; Jensen, C. S.; et al. 2025b. Unraveling Spatio-Temporal Foundation Models via the Pipeline Lens: A Comprehensive Review. *arXiv preprint arXiv:2506.01364*.

Fang, Y.; Qin, Y.; Luo, H.; Zhao, F.; Xu, B.; Zeng, L.; and Wang, C. 2023a. When spatio-temporal meet wavelets: Disentangled traffic forecasting via efficient spectral graph attention networks. In *ICDE*, 517–529.

Fang, Y.; Qin, Y.; Luo, H.; Zhao, F.; and Zheng, K. 2023b. 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*, 2671–2685.

Fang, Z.; Long, Q.; Song, G.; and Xie, K. 2021. Spatial-temporal graph ode networks for traffic flow forecasting. In *SIGKDD*, 364–373.

Feng, X.; Ling, X.; Zheng, H.; Chen, Z.; and Xu, Y. 2018. Adaptive multi-kernel SVM with spatial–temporal correlation for short-term traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 2001–2013.

Guo, S.; Lin, Y.; Feng, N.; Song, C.; and Wan, H. 2019. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. In *AAAI*, 922–929.

Jiang, J.; Han, C.; Zhao, W. X.; and Wang, J. 2023. Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction. In *AAAI*, 4365–4373.

Kumar, S. V. 2017. Traffic flow prediction using Kalman filtering technique. *Procedia Engineering*, 582–587.

Kumar, S. V.; and Vanajakshi, L. 2015. Short-term traffic flow prediction using seasonal ARIMA model with limited input data. *European Transport Research Review*, 21.

Lan, S.; Ma, Y.; Huang, W.; Wang, W.; Yang, H.; and Li, P. 2022. Dstagnn: Dynamic spatial-temporal aware graph neural network for traffic flow forecasting. In *ICML*, 11906–11917.

Li, M.; and Zhu, Z. 2021. Spatial-temporal fusion graph neural networks for traffic flow forecasting. In *AAAI*, 4189–4196.

Li, Y.; Yu, R.; Shahabi, C.; and Liu, Y. 2018. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *ICLR*.

Liu, C.; Miao, H.; Xu, Q.; Zhou, S.; Long, C.; Zhao, Y.; Li, Z.; and Zhao, R. 2025a. Efficient Multivariate Time Series Forecasting via Calibrated Language Models with Privileged Knowledge Distillation. In *ICDE*, 3165–3178.

Liu, C.; Xu, Q.; Miao, H.; Yang, S.; Zhang, L.; Long, C.; Li, Z.; and Zhao, R. 2025b. Timecma: Towards llm-empowered multivariate time series forecasting via cross-modality alignment. In *AAAI*, 18780–18788.

Liu, C.; Zhou, S.; Xu, Q.; Miao, H.; Long, C.; Li, Z.; and Zhao, R. 2025c. Towards Cross-Modality Modeling for Time Series Analytics: A Survey in the LLM Era. In *IJCAI*.

Liu, H.; Dong, Z.; Jiang, R.; Deng, J.; Deng, J.; Chen, Q.; and Song, X. 2023. Spatio-temporal adaptive embedding makes vanilla transformer sota for traffic forecasting. In *CIKM*, 4125–4129.

Liu, J.; Yi, Z.; Zhou, Z.; Huang, Q.; Yang, K.; Wang, X.; and Wang, Y. 2025d. SynEVO: A neuro-inspired spatiotemporal evolutional framework for cross-domain adaptation. In *ICML*.

Liu, S.; Cao, N.; Chen, Y.; Jiang, Y.; and Cong, G. 2025e. Mixture-of-Experts for Personalized and Semantic-Aware Next Location Prediction. *arXiv preprint arXiv:2505.24597*.

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. *VLDB*, 226–238.

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

Pan, B.; Demiryurek, U.; and Shahabi, C. 2012. Utilizing real-world transportation data for accurate traffic prediction. In *ICDM*, 595–604.

Shao, Z.; Zhang, Z.; Wang, F.; Wei, W.; and Xu, Y. 2022. Spatial-temporal identity: A simple yet effective baseline for multivariate time series forecasting. In *CIKM*, 4454–4458.

Song, C.; Lin, Y.; Guo, S.; and Wan, H. 2020. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *AAAI*, 914–921.

Tian, R.; Wang, C.; Hu, J.; and Ma, Z. 2023. Multi-scale spatial-temporal aware transformer for traffic prediction. *Information Sciences*, 119557.

Wang, S.; Zhang, M.; Miao, H.; Peng, Z.; and Yu, P. S. 2022. Multivariate correlation-aware spatio-temporal graph convolutional networks for multi-scale traffic prediction. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 1–22.

Wang, S.; Zhang, M.; Miao, H.; and Yu, P. S. 2021. Mtstnets: Multi-task spatial-temporal networks for multi-scale traffic prediction. In *SDM*, 504–512.

Wu, Z.; Pan, S.; Long, G.; Jiang, J.; and Zhang, C. 2019. Graph wavenet for deep spatial-temporal graph modeling. In *IJCAI*, 1907–1913.

Yang, H.; Cheng, Y.; and Li, G. 2022. A new traffic flow prediction model based on cosine similarity variational mode decomposition, extreme learning machine and iterative error compensation strategy. *Engineering Applications of Artificial Intelligence*, 105234.

Yi, Z.; Zhou, Z.; Huang, Q.; Chen, Y.; Yu, L.; Wang, X.; and Wang, Y. 2024. Get rid of isolation: A continuous multi-task spatio-temporal learning framework. *NeurIPS*, 136701–136726.

Yu, B.; Yin, H.; and Zhu, Z. 2018. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. In *IJCAI*, 3634–3640.

Zhao, J.; and Sun, S. 2016. High-order Gaussian process dynamical models for traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 2014–2019.

Zhao, Z.; Chen, W.; Wu, X.; Chen, P. C.; and Liu, J. 2017. LSTM network: a deep learning approach for short-term traffic forecast. *IET intelligent transport systems*, 68–75.

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. *TPAMI*.