StaPerNet: Feature Stabilization and Period-Level Modeling for Long-Horizon Traffic Forecasting

Anonymous ICME submission

***Abstract*—Accurate traffic forecasting remains challenging due to the interplay of complex long-range dependencies and real-world data instability. To address these challenges, in this paper, we propose a novel framework based on adaptive feature stabilization and structured periodic sparse modeling. First, the framework employs a novel stability-aware mechanism that learns to decompose features adaptively and applies targeted regularization, thereby discriminating between robust and sensitive dimensions and improving generalization against distribution shifts and noise. Building upon these stabilized features, we propose a temporal modeling module that decomposes complex patterns into hierarchical periodic components through sparse coding. This process establishes a structured representation enabling efficient long-range dependency capture via direct cross-period mappings from historical to future periods, bypassing conventional sequential processing. Extensive experiments on five real-world benchmarks demonstrate that our approach achieves state-of-the-art performance with particularly significant improvements in long-term forecasting scenarios while maintaining computational efficiency. Our code is available at https://anonymous.4open.science/r/ICME_2026-648/.**

***Index Terms*—Spatio-temporal Time Series, Feature Stabilization, Periodic Modeling**

I. Introduction

In the pursuit of efficient Intelligent Transportation Systems (ITS), accurate traffic flow forecasting has become indispensable, powering applications from real-time routing to congestion management [1], [2]. Yet, constrained by the complex multi-modal traffic flow characteristics [3], modeling the dynamic and non-linear nature of traffic remains a formidable challenge, driving the need for models that can capture intricate spatiotemporal dynamics [4]. Within this context, two critical challenges stand out: the fragility of learned features in the face of data instability, and the inability to coherently model periodic patterns over extended horizons.

Firstly, regarding the issue of feature instability, existing approaches have directed their main efforts toward the continuous refinement of graph architectures, with the primary emphasis being placed on enhancing their capacity to model complex spatial dependencies [5]–[9]. Nevertheless, a growing body of research has recognized the non-stationary nature of traffic characteristics and attempted to address it within graph-based frameworks. Representative of this direction, PDFormer [10] employs masked spatial self-attention to capture dynamic dependencies and explicitly models non-stationary information propagation through a delay-aware mechanism. Similarly, MegaCRN [11] addresses non-stationarity through a meta-graph learner that dynamically constructs graph structures based on evolving traffic patterns. Yet, while these methods

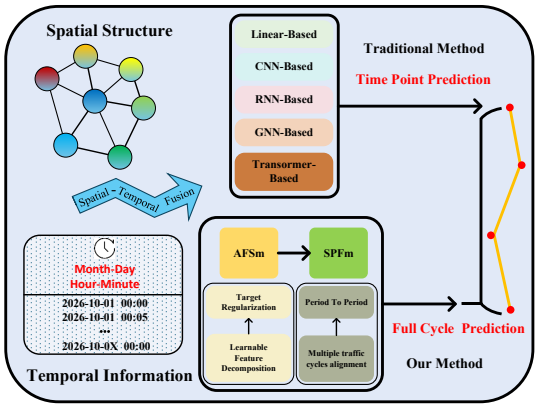

Fig. 1: Comparison between conventional step-wise forecasting and the proposed period-level modeling approach for traffic prediction.

effectively model dynamic graph structures, they process raw input features as homogeneous signals, lacking explicit mechanisms to distinguish stable from unstable components. This leaves them fundamentally vulnerable to distribution shifts, sensor noise, and anomalous events that corrupt the input feature stream, ultimately limiting their real-world robustness.

To bridge this gap, we introduce the Adaptive Feature Stabilization module (AFSm), which directly addresses feature-level non-stationarity through learnable feature decomposition and targeted regularization. Our approach automatically distinguishes between stable and unstable feature components, applying stability-enhancing techniques specifically to vulnerable elements while preserving robust features. This orthogonal stabilization strategy complements existing graph-based methods by providing more reliable input features, thereby enhancing overall model robustness without compromising their structural modeling capabilities.

Secondly, beyond the pursuit of feature stability, accurately modeling long-range temporal dependencies remains a critical yet persistently challenging task in traffic forecasting [12]. Conventional temporal approaches are fundamentally constrained in this regard. Models such as RNNs are plagued by error propagation across extended time horizons [13]–[15], while transformer-based architectures are hindered by prohibitive computational costs stemming from their quadratic attention complexity [10], [16]. Furthermore, as shown in Figure 1, most mainstream methods attempt to output predictions point by point after aggregating spatiotemporal features. This sequential prediction approach consumes substantial computa-

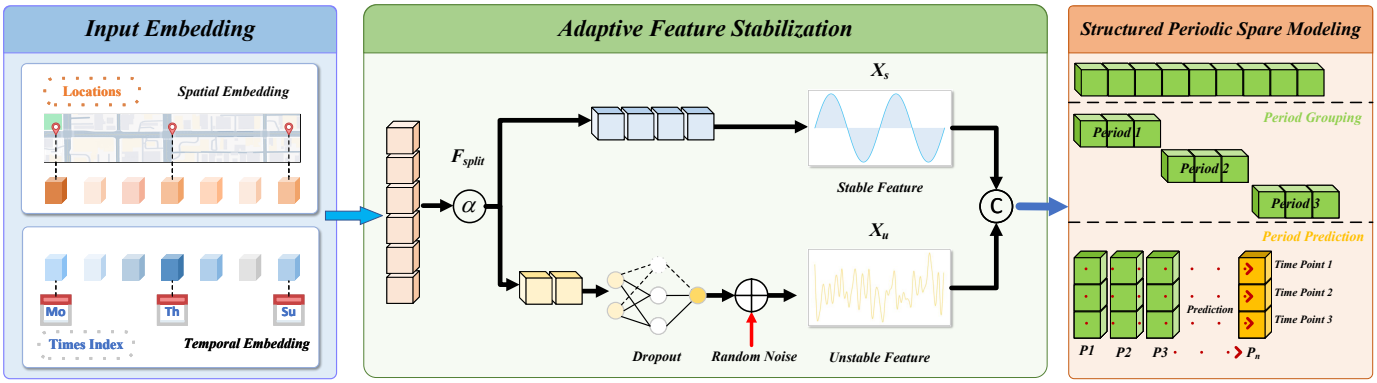

Fig. 2: **Pipeline of the proposed framework.** Input sequences are first mapped to spatiotemporal embeddings. The Adaptive Feature Stabilization module then separates latent channels into stable and unstable parts through a learnable ratio, perturbs only the unstable features with dropout and Gaussian noise, and fuses the two branches into a refined representation. The stabilized features are finally processed by the Structured Periodic Sparse Modeling module to produce future traffic predictions.

tional resources in temporal forecasting systems. Such fundamental limitation is their underlying assumption of temporal homogeneity, which leads to a failure in explicitly leveraging the strong periodic patterns, such as daily and weekly rhythms, that are inherent to traffic systems [17]. This oversight results in models that are not only computationally inefficient but also empirically suboptimal for capturing the structured dependencies essential for accurate long-term predictions. Consequently, these models often fail to generalize in real-world settings where periodic trends dominate long-term traffic behavior.

To address these temporal challenges, we propose the Structured Periodic Sparse Model (SPSm). This module explicitly decomposes the stabilized time series into periodic components, aligns multiple traffic cycles, and learns direct mappings between historical and future periods. The SPSm is designed as a period-to-period model that directly characterizes the underlying relationships across different temporal cycles. It thereby replaces conventional sequential processing with structured period-level modeling, which enables the efficient capture of long-range dependencies while maintaining computational efficiency. By establishing direct connections between corresponding periods across time, the module effectively leverages the inherent cyclical nature of traffic data, providing a principled solution for long-term forecasting.

Our main contributions can be summarized as follows:

- We introduce AFSm that employs learnable feature decomposition and targeted regularization mechanisms to enhance model robustness against distribution shifts, input perturbations, and noise.
- We introduce SPSm that models long-range dependencies via periodic pattern decomposition and cross-period mapping, thereby mitigating the computational constraints of sequential models.
- We develop an integrated framework where AFSm and SPSm work synergistically, achieving new state-of-the-art performance while maintaining computational efficiency across multiple real-world benchmarks.

II. Methodology

A. Preliminary

The traffic network is represented by an undirected graph $G = (V, E)$, where V is the set of N nodes, and E is the set of edges. The adjacency matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ is constructed based on the Euclidean distance between nodes.

Each node records traffic flow data as a graph signal $\mathbf{x}^t \in \mathbb{R}^N$ at time step t. Given historical S-step graph signals $\mathbf{X}^{1:S} = (\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^S) \in \mathbb{R}^{N \times S}$, the problem aims to predict the next T-step graph signals:

$$\mathbf{X}^{S+1:S+T} = (\mathbf{x}^{S+1}, \mathbf{x}^{S+2}, \ldots, \mathbf{x}^{S+T}) \in \mathbb{R}^{N \times T}, \quad (1)$$

this is formulated as finding a mapping function F such that:

$$(\mathbf{x}^{S+1}, \mathbf{x}^{S+2}, \ldots, \mathbf{x}^{S+T}) = F\left((\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^S)\right). \quad (2)$$

B. Adaptive Feature Stabilization

Spatiotemporal feature instability, arising from distribution shifts, sensor noise, and anomalous traffic events, poses a major challenge to model robustness in traffic forecasting. On top of the latent representations produced by the HimNet encoder [18], we design AFSm, which is a lightweight and learned module that explicitly decomposes spatiotemporal features into stable and unstable components and applies targeted stabilization exclusively to the latter.

AFSm is motivated in an information-theoretic objective: to maximize the mutual information $I(\mathbf{X}_s; \mathbf{Y})$ while minimizing the sensitivity $\|\nabla_{\mathbf{X}_u} \mathcal{L}\|$. Although not explicitly enforced via an auxiliary loss, this objective is implicitly optimized through the dynamic feature demultiplexer $\mathcal{F}_{\text{split}}$ and the stabilization pipeline. The gradient flow from $\mathcal{L}_{\text{forecast}}$ induces a self-organizing feature routing that naturally converges toward the information-theoretic optimum.

Our model adopts the feature encoding of HimNet [18], and the overall architecture is shown in Figure 2. HimNet first encodes raw data into spatiotemporal representations $\mathbf{X} \in \mathbb{R}^{B \times T \times N \times C}$. Upon these HimNet-derived features, our

proposed AFSm operates to enhance robustness by decomposing them into stable and unstable components. The core of our approach is $\mathcal{F}_{\text{split}}$, which adaptively decomposes the input features $\mathbf{X}$ into stable and unstable components:

$$\mathbf{X}_s, \mathbf{X}_u, d_s, d_u = \mathcal{F}_{\text{split}}(\mathbf{X}), \tag{3}$$

where $\mathbf{X}_s$ and $\mathbf{X}_u$ represent the stable and unstable feature components, with corresponding channel dimensions d_s and d_u. Unlike methods using pre-defined statistical metrics, $\mathcal{F}_{\text{split}}$ learns an *adaptive channel-wise partitioning* that designates a subset of channels as unstable. A central element is the decomposition ratio $\alpha = d_u/C$, which is governed by a learnable parameter θ_α as

$$\alpha = \sigma(\theta_\alpha) \cdot \gamma + \beta, \tag{4}$$

where the scaling and shifting parameters γ and β (set to 0.8 and 0.1 in our experiments) constrain $\alpha \in (\beta, \beta + \gamma)$ to prevent pathological cases where either branch is eliminated. The parameter θ_α is learnable. Consequently, features that are incompatible with the regularization applied to the unstable branch tend to migrate into $\mathbf{X}_s$, while those that remain functional under such perturbations constitute $\mathbf{X}_u$, which leads to a self-organizing and balanced decomposition.

To enhance the robustness of the unstable component $\mathbf{X}_u$, AFSm applies both dropout and additive Gaussian perturbations during training. Let $\mathcal{D}(\cdot; p)$ denote dropout with rate p. We first compute a global standard deviation over all entries of the unstable features

$$\sigma_{\mathbf{X}_u} = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}(x_i - \mu_{\mathbf{X}_u})^2}, \quad \text{where } M = |\mathbf{X}_u|, \tag{5}$$

which yields a scalar that reflects the overall variability of $\mathbf{X}_u$. We then draw a noise tensor $\boldsymbol{\epsilon}$ with the same shape as $\mathbf{X}_u$ from a standard normal distribution,

$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{6}$$

and scale it by a noise strength parameter σ_{noise} and the global deviation $\sigma_{\mathbf{X}_u}$. The resulting perturbation is

$$\boldsymbol{\Xi} = \sigma_{\text{noise}}\, \sigma_{\mathbf{X}_u}\, \boldsymbol{\epsilon}. \tag{7}$$

The unstable component is then updated as

$$\mathbf{X}'_u = \mathcal{D}(\mathbf{X}_u; p) + \boldsymbol{\Xi}. \tag{8}$$

This formulation injects isotropic Gaussian perturbations whose magnitude adapts to the overall variability of the unstable channels, thereby exposing them to semantically meaningful fluctuations during training.

The feature streams are synthesized through a stability-aware fusion:

$$\mathbf{X}' = \text{Concat}[\mathbf{X}_s, \mathbf{X}'_u], \quad \mathbf{H} = \mathbf{X}' \oplus \lambda \cdot \mathcal{G}(\mathbf{X}'), \tag{9}$$

where $\mathcal{G} : \mathbb{R}^C \to \mathbb{R}^{C/2} \to \mathbb{R}^C$ performs non-linear feature transformation. The composition operator $\oplus$ integrates transformed features while preserving the stabilized representation through adaptive scaling by λ.

In terms of computational overhead, the proposed components are highly efficient. The decomposition $\mathcal{F}_{\text{split}}$ is a tensor slicing operation with $\mathcal{O}(1)$ cost, while the enhancement network $\mathcal{G}$ uses a bottleneck architecture of complexity $\mathcal{O}(B \cdot T \cdot N \cdot C^2)$, comparable to a standard feed-forward layer. Dropout and covariance-aware noise injection are element-wise operations with negligible cost.

C. Structured Periodic Sparse Module

Traditional traffic forecasting models face dual challenges in capturing long-range dependencies: computational inefficiency and the overlooking of inherent periodicity. We address this with SPSm, which employs structured periodic decomposition. SPSm operates at the coarse-grained period level rather than at individual time steps, which allows it to efficiently capture long-range dependencies while explicitly encoding traffic cyclicality. Figure 2 illustrates the entire process of the SPSm module processing input features.

The core of the proposed approach is an adaptive multiple traffic cycle alignment mechanism that integrates domain knowledge with structural constraints. Given an expected period length $\hat{P}$, computational feasibility is ensured through an optimization that determines the optimal period P:

$$P = \arg \min_{P' \le \min(T, T_{pred})} \left(|T \bmod \hat{P}| + |T_{pred} \bmod \hat{P}| \right). \tag{10}$$

In implementation, the input data are preprocessed to ensure that T and T_{pred} are integer multiples of $\hat{P}$. This design generally results in $P = \hat{P}$ and prevents trivial solutions. The resulting factorization into $T = H \cdot P$ and $T_{pred} = F \cdot P$, enables efficient period-level modeling, where H and F denote the number of historical and future periods, respectively. Since each period corresponds to a complete and contiguous traffic cycle, this decomposition constitutes a lossless reorganization of the original time sequence.

Based on the period factorization established previously, the temporal dependencies are modeled through a structured linear transformation. Let $\mathbf{X} \in \mathbb{R}^{B \times N \times P \times (H \cdot D)}$ denote the input tensor representing H historical periods, where B is the batch size, N is the number of nodes, P is the period length, and D is the hidden dimension. The mapping to future periods is formulated as:

$$\mathbf{Y} = \mathbf{X}\mathbf{W}^\top, \tag{11}$$

where $\mathbf{W} \in \mathbb{R}^{(F \cdot D) \times (H \cdot D)}$ is a learnable weight matrix that projects historical period embeddings to future period embeddings, and F denotes the number of future periods. The resulting tensor $\mathbf{Y} \in \mathbb{R}^{B \times N \times P \times (F \cdot D)}$ is then reshaped to generate the future periodic feature $\mathbf{Y}_{\text{out}} \in \mathbb{R}^{B \times N \times T_{\text{pred}}}$, where $T_{\text{pred}} = P \times F$ corresponds to the prediction horizon. This formulation achieves computational efficiency by operating at the period granularity, reducing the parameter complexity from $\mathcal{O}(T^2)$ in sequential models to $\mathcal{O}(H \cdot F)$ while preserving temporal relationships through the periodic structure.

The term "structured" in this context emphasizes the organized decomposition of temporal patterns into periodic components. This period-level representation achieves substantial

computational efficiency compared to traditional sequential approaches. Specifically, the complexity is reduced from $\mathcal{O}(T^2 \cdot D^2)$ typical of attention mechanisms to $\mathcal{O}(H \cdot F \cdot D^2)$. Given that $H = T/P$ and $F = T_{\text{pred}}/P$, the effective complexity becomes $\mathcal{O}(\frac{T \cdot T_{\text{pred}}}{P^2} \cdot D^2)$, yielding approximately P^2-fold improvement in the quadratic term relative to sequence length compared to standard attention.

The SPSm exhibits strong synergy with the AFSm, forming a complementary framework that effectively addresses both temporal and feature-space challenges. While SPSm focuses on efficient temporal structure modeling through period-level transformations, AFSm ensures robustness in the feature representation space by stabilizing feature distributions. This integrated approach allows AFSm-processed stable features to serve as inputs to SPSm, therefore ensuring that the periodic decomposition operates on reliable representations. Through joint optimization, the framework thus simultaneously maximizes temporal structural efficiency and feature-space robustness, leading to comprehensive performance improvements across diverse real-world scenarios.

III. Experiments

A. Experimental Setup

Datasets. Our model is evaluated on five standard spatiotemporal forecasting benchmarks: **METR-LA** and **PEMS-BAY** [19], which record traffic speed from sensors in Los Angeles and the Bay Area, as well as **PEMS04**, **PEMS07**, and **PEMS08** [20], which contain traffic flow data from the Caltrans PeMS system. All datasets have a 5-minute sampling rate. We preprocess the raw data using Z-score normalization. All datasets are publicly available and used in accordance with standard ethical research practices. Comprehensive details are provided in table I.

TABLE I: Summary of datasets.

Dataset	Nodes	Timesteps	Training	Validation	Testing
METR-LA	207	34,272	70%	10%	20%
PEMSBAY	325	52,116	70%	10%	20%
PEMS04	307	16,992	60%	20%	20%
PEMS07	883	28,224	60%	20%	20%
PEMS08	170	17,856	60%	20%	20%

Settings. We maintain a consistent experimental setup across datasets unless specified otherwise. The model uses a single-layer encoder-decoder architecture. Key hyperparameters, including the hidden dimension h, temporal embedding size d_t, spatial embedding size d_s, and spatiotemporal embedding size d_{st}, are summarized in Table II.

TABLE II: Model configurations for different datasets.

Dataset	h	d_t	d_s	d_{st}	T, T_{pred}
METR-LA	64	16	16	16	12
PEMS-BAY	64	16	16	16	12
PEMS04	64	16	16	16	12
PEMS07	64	16	16	16	12
PEMS08	96	12	14	10	12

The optimization employs the Adam optimizer with an initial learning rate of 0.001, coupled with a learning rate scheduler. Models are trained with a batch size of 16 for a maximum of 200 epochs, incorporating early stopping (patience=20) to prevent overfitting. The loss function is Mean Absolute Error (MAE) for METR-LA and PEMS-BAY, and the more robust Huber loss for the PEMS04, PEMS07, and PEMS08 datasets. Model performance is evaluated using standard metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). All experiments are conducted on NVIDIA GeForce RTX 3090 GPUs.

Baselines. We compare our method against a comprehensive set of baselines spanning different architectural paradigms. Simple yet strong baselines include **HI** [21] and **GRU** [22]. Graph-based spatiotemporal models are represented by **STGCN** [8], **DCRNN** [23], and **GWNet** [6], which integrate graph convolutions with temporal modeling. Methods focusing on adaptive graph learning include **AGCRN** [7] and **GTS** [24], while **STNorm** [25] addresses distribution shifts through specialized normalization. Recent innovations include **STID** [26] and **ST-WA** [27] that leverage spatiotemporal identity embeddings, **PDFormer** [10] which targets propagation delay modeling, and meta-learning approaches like **MegaCRN** [11] and **HimNet** [18] designed to capture complex dependencies.

B. Performance Evaluation

The experimental results are presented in Figure III, with the best results highlighted in red. As clearly shown, our method achieves state-of-the-art performance across all five benchmarks, thus underscoring the efficacy of our feature stabilization paradigm. It attains the best results in 13 out of 15 metrics. Improvements are particularly striking on the METR-LA and PEMS-BAY datasets. Compared to the strongest baseline, our model achieves MAE values of **2.94** and **1.52**, corresponding to substantial improvements of **12.8%** and **17.4%**, respectively.

These consistent gains are evident across all evaluation metrics. On the METR-LA benchmark, our model attains an RMSE of **6.13** and a MAPE of **8.15%**, which correspond to improvements of **15.1%** and **16.8%** respectively. Similarly, on PEMS-BAY, it achieves an RMSE of **3.54** and a MAPE of **3.41%**, representing even greater improvements of **18.1%** and **21.2%**. Notably, this advantage extends robustly to the more challenging PEMS benchmarks, which feature more complex graph structures and relatively shorter temporal sequences. On PEMS04, it reduces MAE by **0.3%** and MAPE by **0.4%**. For the large-scale PEMS07 dataset with 883 nodes, it still improves MAE by **0.6%** and MAPE by **1.5%**. On PEMS08, our model achieves a competitive MAE of **13.57** while further lowering MAPE by **1.1%**.

In summary, the demonstrated performance advantage provides compelling validation for our core architectural strategy. This strategy focuses on learning stable, interpretable representations to model underlying periodic features, thereby effectively addressing the critical challenge of data non-stationarity.

TABLE III: Performance on METR-LA, PEMSBAY, PEMS04, PEMS07, and PEMS08 datasets

Dataset		Metric	HI	GRU	STGCN	DCRNN	GWNet	AGCRN	GTS	STNorm	STID	ST-WA	PDFormer	MegaCRN	HimNet	StaPerNet
METRLA	Average	MAE	6.80	4.88	3.60	3.54	3.51	3.59	3.59	3.57	3.55	3.68	3.62	3.51	3.37	**2.94**
		RMSE	14.21	9.75	7.43	7.47	7.28	7.45	7.44	7.51	7.55	7.59	7.47	7.39	7.22	**6.13**
		MAPE	16.71%	14.91%	10.35%	10.32%	9.96%	10.47%	10.25%	10.24%	10.95%	10.78%	10.91%	10.01%	9.79%	**8.15%**
PEMSBAY	Average	MAE	3.05	2.70	2.02	1.97	1.99	1.94	2.06	1.92	1.91	2.00	1.91	1.90	1.84	**1.52**
		RMSE	7.01	6.28	4.63	4.60	4.60	4.50	4.60	4.45	4.42	4.52	4.43	4.49	4.32	**3.54**
		MAPE	6.83%	6.72%	4.72%	4.68%	4.71%	4.55%	4.88%	4.46%	4.55%	4.63%	4.51%	4.53%	4.33%	**3.41%**
PEMS04	Average	MAE	42.35	25.55	19.57	19.63	18.53	19.38	20.96	18.96	18.38	19.06	18.36	18.72	18.14	**18.09**
		RMSE	61.66	39.71	31.38	31.26	29.92	31.25	32.95	30.98	29.95	31.02	30.03	30.53	29.88	**29.77**
		MAPE	29.92%	17.35%	13.44%	13.59%	12.89%	13.40%	14.66%	12.69%	12.04%	12.52%	12.00%	12.77%	12.00%	**11.96%**
PEMS07	Average	MAE	49.29	26.74	21.74	21.16	20.47	20.57	22.15	20.50	19.61	20.74	19.97	19.83	19.21	**19.10**
		RMSE	71.34	42.78	35.27	34.14	33.47	34.40	35.10	34.66	32.79	34.05	32.95	32.91	**32.75**	33.00
		MAPE	22.75%	11.58%	9.24%	9.02%	8.61%	8.74%	9.38%	8.75%	8.30%	8.77%	8.55%	8.36%	8.03%	**7.91%**
PEMS08	Average	MAE	34.66	19.36	16.08	15.22	14.40	15.32	16.49	15.41	14.21	15.41	13.58	14.75	**13.57**	**13.57**
		RMSE	50.45	31.20	25.39	24.17	23.39	24.41	26.08	24.77	23.28	24.62	23.41	23.73	**23.22**	23.27
		MAPE	21.63%	12.43%	10.60%	10.21%	9.21%	10.03%	10.54%	9.76%	9.27%	9.94%	9.05%	9.48%	8.98%	**8.88%**

C. *Ablation Study*

For the ablation study, we employ the METR-LA dataset, which offers a balanced scale in terms of node count and data volume. To isolate the contribution of each proposed module, we evaluate two ablated variants: (1) w/o AFSm: In this variant, the AFSm is replaced by a fully connected layer. (2) w/o SPSm: Similarly, the SPSm is substituted with an equivalent linear layer. The results are detailed in Fig 3.

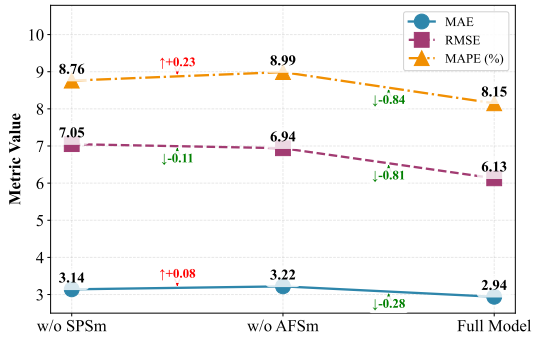

Fig. 3: Ablation study on the METR-LA dataset.

The ablation analysis reveals distinct functional roles for each module. Removing SPSm results in an MAE of 3.14 and MAPE of 8.76%, representing degradations of **6.8%** and **7.5%**, respectively. This consistent decline highlights the module's importance in capturing essential periodic patterns for urban traffic forecasting.

Strikingly, the removal of AFSm induces even more severe degradation, with MAE increasing to 3.22 and MAPE to 8.99%. This larger performance drop underscores the foundational role of feature stabilization in mitigating distribution shifts and non-stationarity.

D. *Efficiency Study*

A comparative analysis of computational efficiency between our model and the remaining benchmarks is conducted in this subsection. Figure 4 shows the number of parameters, per-batch time, and per-epoch time.

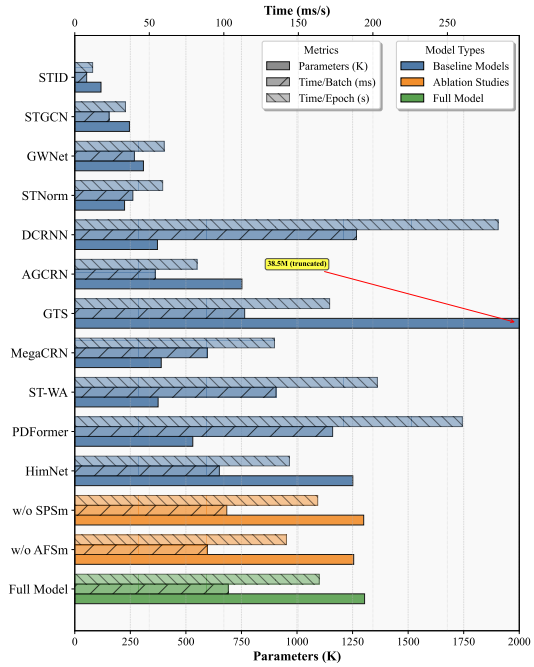

Fig. 4: Efficiency comparison on METR-LA dataset.

As shown in Figure 4, adding the SPSm module introduces only 4,000 additional parameters but yields a measurable acceleration in convergence. This combination of high parameter efficiency and faster convergence suggests that the module effectively captures underlying periodic patterns in the data, which in turn guides and streamlines the optimization process.

Furthermore, we analyze the model's sensitivity to the initial stable feature ratio. As shown in Fig. 5, performance improves as the ratio increases and peaks around 0.7, after which it gradually declines. Notably, AFSm consistently converges

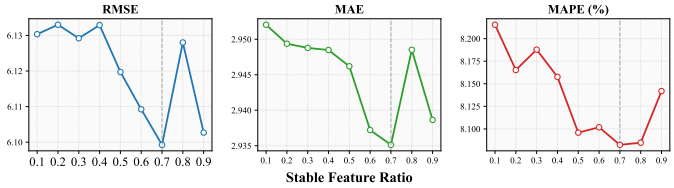

Fig. 5: Effect of the Stable Feature Ratio on Performance.

to near-optimal performance across a broad range of initial values, indicating that it adaptively learns an effective feature stabilization policy rather than relying on a sensitive hyperparameter.

IV. Conclusion

This paper introduces a novel traffic forecasting framework that addresses both feature instability and long-range temporal dependencies. Its core components, AFSm and SPSm, respectively enhance robustness against distribution shifts and capture long-range periodic patterns. Their synergy establishes a new state-of-the-art across five benchmarks, with particularly strong performance in long-term forecasting. By learning interpretable structures while maintaining resilience to data non-stationarity, this approach demonstrates significant potential for real-world intelligent transportation systems.

References

[1] Carlos Oliveira Cruz and Joaquim Miranda Sarmento, "Traffic forecast inaccuracy in transportation: a literature review of roads and railways projects," *Transportation*, pp. 1–36, 2020.

[2] Wei Li, Li Ying Sui, Min Zhou, and Hai-rong Dong, "Short-term passenger flow forecast for urban rail transit based on multi-source data," *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, 2021.

[3] Hao Niu, Yun Xiong, Xiaosu Wang, Biao Yang, and Yao Zhang, "How does textual information selection influence time series forecasting? A cross-modal perspective on financial volatility prediction," in *IEEE International Conference on Multimedia and Expo, ICME 2024*. 2024, pp. 1–6, IEEE.

[4] Shengna Guo, Youfang Lin, Huaiyu Wan, Xiucheng Li, and G. Cong, "Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, pp. 5415–5428, 2021.

[5] 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*. 2020, vol. 34, pp. 914–921, AAAI Press.

[6] 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*. 2019, pp. 1907–1913, IJCAI.

[7] Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang, "Adaptive graph convolutional recurrent network for traffic forecasting," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17804–17815, 2020.

[8] Bing Yu, Haoteng Yin, and Zhanxing Zhu, "Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. 2018, pp. 3634–3640, IJCAI.

[9] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang, "Connecting the dots: Multivariate time series forecasting with graph neural networks," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2020, pp. 753–763, ACM.

[10] Jiawei Jiang, Chenggkai Han, Wayne Xin Zhao, and Jingyuan Wang, "Pdformer: Propagation delay-aware dynamic long-range transformer for traffic flow prediction," in *AAAI*. 2023, AAAI Press.

[11] 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*. 2023, vol. 37, pp. 8078–8086, AAAI Press.

[12] Wentian Zhao, Yanyun Gao, Tingxiang Ji, Xili Wan, Feng Ye, and Guangwei Bai, "Deep temporal convolutional networks for short-term traffic flow forecasting," *IEEE Access*, vol. 7, pp. 114496–114507, 2019.

[13] Haiyang Yu, Zhihai Wu, Shuqin Wang, Yunpeng Wang, and Xiaolei Ma, "Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks," *Sensors (Basel, Switzerland)*, vol. 17, no. 10, 2017.

[14] Yang Jian, Jinhong Li, Lu Wei, Lei Gao, and Fuqi Mao, "Spatiotemporal deepwalk gated recurrent neural network: A deep learning framework for traffic learning and forecasting," *Journal of Advanced Transportation*, 2022.

[15] Junchen Ye, Leilei Sun, Bowen Du, Yanjie Fu, and Hui Xiong, "Coupled layer-wise graph convolution for transportation demand prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2021, vol. 35, pp. 4617–4625, AAAI Press.

[16] Peng Chen, Yingying Zhang, Yunyao Cheng, Yang Shu, Yihang Wang, Qingsong Wen, Bin Yang, and Chenjuan Guo, "Multi-scale transformers with adaptive pathways for time series forecasting," in *International Conference on Learning Representations*, 2024.

[17] Martín Saavedra, Alberto P. Muñuzuri, Monica Menendez, and José Balsa-Barreiro, "Analysing macroscopic traffic rhythms and city size in affluent cities: insights from a global panel data of 25 cities," *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, vol. 382, 2024.

[18] Zheng Dong, Renhe Jiang, and Haotian Gao, "Heterogeneity-informed meta-parameter learning for spatiotemporal time series forecasting," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2024, pp. 631–641, ACM.

[19] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," *arXiv: Learning*, 2017.

[20] Chao Song, Youfang Lin, S. Guo, and Huaiyu Wan, "Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting," in *AAAI Conference on Artificial Intelligence*, 2020.

[21] Yue Cui, Jiandong Xie, and Kai Zheng, "Historical inertia: A neglected but powerful baseline for long sequence time-series forecasting," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2021, pp. 2965–2969, ACM.

[22] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[23] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu, "Diffusion convolutional recurrent neural network: Data-driven traffic forecasting," in *International Conference on Learning Representations*, 2018.

[24] Chao Shang, Jie Chen, and Jinbo Bi, "Discrete graph structure learning for forecasting multiple time series," in *International Conference on Learning Representations*, 2021.

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

[26] 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*. 2022, pp. 4454–4458, ACM.

[27] Razvan-Gabriel Cirstea, Bin Yang, Chenjuan Guo, Tung Kieu, and Shirui Pan, "Towards spatio-temporal aware traffic time series forecasting," in *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. 5 2022, pp. 2900–2913, IEEE.