TFTOOLKITS : A Modular Framework for Spatial-Temporal Traffic Flow Prediction

He Zhu
School of Urban Planning and Design
Peking University
Beijing, 100871, China
zhuhu@stu.pku.edu.cn

January 28, 2025

Abstract

Traffic flow prediction remains a fundamental challenge in intelligent transportation systems due to the complex interplay of spatial and temporal dependencies. While numerous approaches have been proposed, the field lacks a unified theoretical framework to systematically analyze and understand the relative importance of different modeling components. This paper introduces TFTOOLKITS , a modular framework that decomposes traffic prediction into three essential components: temporal mixing, spatial mixing, and spatial-temporal fusion. We first conduct a comprehensive analysis of the challenges in traffic prediction, identifying key requirements for each component. Based on these insights, we propose TFTOOLKITS , a flexible architecture where each component can be instantiated with various neural network modules, enabling practitioners to balance accuracy, efficiency, and interpretability according to their specific needs. Through extensive experiments on real-world datasets, we demonstrate that our framework provides valuable insights into the relative contributions and interactions between different modeling components while offering unprecedented flexibility in model design. The code and datasets used in this paper are available at https://github.com/zhuchichi56/TFTOOLKITS.

1 Introduction

Traffic flow prediction has emerged as a critical component in modern intelligent transportation systems (ITS), playing a vital role in urban traffic management, route planning, and congestion control Zhang et al. (2019); Li et al. (2018). Despite significant advances in deep learning and its applications to traffic prediction Wang et al. (2020), developing effective prediction models remains challenging due to several fundamental difficulties:

- **Complex Temporal Dependencies**: Traffic patterns exhibit multiple temporal scales, from short-term fluctuations to long-term trends Yu et al. (2017); Wu et al. (2020), requiring models to capture both immediate and historical dependencies.
- **Dynamic Spatial Relationships**: The spatial correlation between different locations varies with time and conditions Zheng et al. (2020), necessitating adaptive spatial modeling approaches.

- **Non-linear Interactions**: The interaction between spatial and temporal factors is highly non-linear and context-dependent Guo et al. (2019); Bai et al. (2020), making it difficult to model with traditional approaches.
- **Computational Efficiency**: Real-world applications require models that can process large-scale traffic networks efficiently while maintaining prediction accuracy Chen et al. (2020); Ye et al. (2021).

While numerous approaches have been proposed to address these challenges, existing solutions typically focus on specific aspects of the problem, leading to fragmented methodologies that may excel in one aspect but underperform in others . *What design choices and architectural components are needed to effectively handle these diverse prediction tasks while maintaining model flexibility?* To answer this question, this paper introduces TFTOOLKITS , a comprehensive framework that decomposes traffic prediction into three fundamental components:

1. **Temporal Mixing Module**: Captures multi-scale temporal dependencies through flexible temporal modeling approaches.
2. **Spatial Mixing Module**: Models dynamic spatial relationships using adaptable graph-based architectures Wang et al. (2020); Zheng et al. (2020)
3. **Spatial-Temporal Fusion Module**: Integrates spatial and temporal information through innovative fusion mechanisms.

The key innovation of our framework lies in its modularity - each component can be instantiated with different neural network architectures, allowing practitioners to make informed trade-offs between prediction accuracy, computational efficiency, and model interpretability. This flexibility enables the framework to adapt to various application scenarios while maintaining a consistent theoretical foundation.

2 Related Work

Our work builds upon and extends several lines of research in traffic prediction and deep learning architectures. We organize the related work into three main categories.

2.1 Traditional Traffic Prediction Methods

Early approaches to traffic prediction relied heavily on statistical methods such as Auto-Regressive Integrated Moving Average (ARIMA) and its variants. These methods, while interpretable, often struggled to capture the complex, non-linear patterns inherent in traffic data. Kalman filtering (Williams & Hoel, 1998) and support vector regression (Wu et al., 2004) were later introduced to handle non-linear relationships, but their performance was limited by the need for manual feature engineering and their inability to automatically learn representations from large-scale data.

2.2 Deep Learning for Traffic Prediction

The advent of deep learning has revolutionized traffic prediction. Recurrent Neural Networks (RNNs) and their variants, particularly LSTM (Hochreiter & Schmidhuber, 1997) and GRU (Cho et al., 2014), have shown remarkable success in capturing temporal dependencies in traffic data. Subsequent work introduced attention mechanisms to enhance the modeling of long-term dependencies (Guo et al., 2019).

Graph Neural Networks (GNNs) emerged as a powerful tool for modeling spatial relationships in traffic networks. DCRNN (Li et al., 2018) pioneered the use of diffusion convolution for spatial-temporal forecasting, while Graph WaveNet (Wu et al., 2019) introduced adaptive adjacency matrices to learn hidden spatial dependencies. More recent approaches like STGCN (Yu et al., 2018) and ASTGCN (Guo et al., 2019) combine GNNs with various temporal modeling techniques to capture both spatial and temporal dependencies simultaneously.

Transformer-based architectures have achieved state-of-the-art performance in traffic prediction (Xu et al., 2020; Park et al., 2020). These models excel at capturing long-range dependencies through self-attention mechanisms but suffer from quadratic computational complexity, limiting their applicability to long sequences or large-scale traffic networks. Recent works have proposed various optimization techniques (Zhou et al., 2021; Liu et al., 2022) to reduce this computational burden, but the fundamental efficiency challenge remains.

2.3 Spatial-Temporal Modeling

The integration of spatial and temporal dependencies has been a central challenge in traffic prediction. Recent works have proposed various approaches to this problem, including spatial-temporal attention mechanisms (Zheng et al., 2020), multi-scale temporal modeling (Chen et al., 2020), and hybrid architectures that combine different types of neural networks (Jin et al., 2020). These studies motivate our investigation of spatial-temporal mechanisms in TFTOOLKITS while maintaining computational efficiency.

3 Preliminaries

3.1 Traffic Flow Prediction Problem

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{A})$ represent a traffic network, where $\mathcal{V}$ is the set of N nodes (traffic sensors), $\mathcal{E}$ is the set of edges (road segments), and $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix. At each time step t, we observe a traffic flow feature matrix $\mathbf{X}_t \in \mathbb{R}^{N \times D}$, where D is the dimension of traffic features (e.g., flow, speed, occupancy).

The traffic flow prediction task aims to predict future traffic conditions for the next τ time steps given a historical window of length T:

$$[\mathbf{X}_{t+1}, ..., \mathbf{X}_{t+\tau}] = f([\mathbf{X}_{t-T+1}, ..., \mathbf{X}_t]; \mathcal{G}, \theta) \tag{1}$$

3.2 Evolution of Traffic Prediction Methods

The development of traffic prediction methods has evolved from statistical foundations to modern deep learning approaches. Early statistical models like ARIMA decomposed traffic patterns into interpretable components:

$$\mathbf{X}_t = \sum_{i=1}^{p} \phi_i \mathbf{X}_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t \tag{2}$$

here $\mathbf{X}_t$ represents traffic flow at time t, with autoregressive terms capturing temporal dependencies and moving average components modeling past disturbances. While these models offered interpretability, they struggled with non-linear relationships and spatial dependencies. The advent of deep learning marked a shift, with approaches incorporating RNNs and GNNs:

$$\mathbf{X}_{t+1:t+\tau} = f_\theta(\mathbf{X}_{t-T+1:t}, \mathbf{A}, \mathbf{h}_t) \tag{3}$$

where f_θ is a neural network, $\mathbf{A}$ encodes road network structure, and $\mathbf{h}_t$ captures temporal dependencies. RNNs model the temporal evolution through recurrent state updates, while GNNs have emerged as a powerful framework for handling the complex spatial dependencies in traffic networks.

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \mathbf{M}_{ij}\mathbf{W}^{(l)}\mathbf{h}_j^{(l)}\right) \tag{4}$$

where $\mathbf{h}_i^{(l)}$ represents node i's features at layer l, $\mathcal{N}(i)$ denotes i's neighbors, $\mathbf{M}_{ij}$ is the normalized adjacency matrix entry, and $\mathbf{W}^{(l)}$ are learnable weights. This formulation allows GNNs to learn node representations by aggregating information from neighboring nodes. However, these models still face limitations in capturing complex dependencies effectively Zhang et al. (2019). While GNNs can model local spatial relationships, they struggle to capture long-range dependencies and dynamic interactions across the network. This representation bottleneck motivates the introduction of attention mechanisms from Transformer architectures, which have demonstrated superior capability in modeling complex relationships through their self-attention mechanism.

3.3 Attention Mechanisms

Attention mechanisms have become fundamental building blocks in modern deep learning architectures. The standard attention operation can be formulated as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}})\mathbf{V} \tag{5}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ represent queries, keys, and values. In traffic prediction, attention operates in multiple dimensions. First, **Temporal Attention** captures dependencies between different time steps. Second, **Spatial Attention** models relationships between different locations. Third, **Cross-domain Attention** integrates information from multiple feature domains.

3.4 Challenges in Traffic Flow Prediction

Traffic flow prediction presents several fundamental challenges that must be systematically addressed. We identify and analyze four key challenges that significantly impact prediction performance:

- **Temporal Complexity:** Traffic patterns exhibit multi-scale temporal characteristics $\{\mathcal{T}_d, \mathcal{T}_w, \mathcal{T}_s\}$ representing daily, weekly and seasonal periodicities. The temporal evolution can be decomposed as:

$$\mathbf{X}_t = \mathbf{T}_{\text{trend}}(t) + \sum_{i\in\{d,w,s\}} \mathbf{T}_i(t) + \epsilon_t \tag{6}$$

  where $\mathbf{T}_{\text{trend}}$ captures long-term evolution, $\mathbf{T}_i$ represents periodic components, and ϵ_t models local fluctuations. These components interact non-linearly, creating compound effects that challenge traditional decomposition methods.

- **Spatial Heterogeneity:** The spatial correlation structure $\rho_{ij}(t)$ between locations i and j exhibits complex dependencies:
$$\rho_{ij}(t) = f(d_{ij}, \mathbf{A}_{ij}, \mathbf{c}_t) \tag{7}$$
where d_{ij} is the physical distance, $\mathbf{A}_{ij}$ represents network connectivity, and $\mathbf{c}_t$ captures time-varying contextual factors. This relationship varies across urban ($\mathcal{U}$) versus highway ($\mathcal{H}$) regions and peak ($\mathcal{P}$) versus off-peak ($\mathcal{O}$) periods.

- **Non-linear Interactions:** The spatio-temporal coupling function $\Phi(\mathbf{S}, \mathbf{T})$ exhibits strong context dependency:
$$\Phi(\mathbf{S}, \mathbf{T}|\mathbf{z}_t) = g(\mathbf{S}, \mathbf{T}, \mathbf{z}_t) \tag{8}$$
where $\mathbf{z}_t$ represents environmental conditions. Multi-scale effects emerge as perturbations δ propagate through the network with varying impact $\mathcal{I}(t, \mathbf{x})$ based on spatial location $\mathbf{x}$ and time t.

- **Computational Constraints:** For a network with $|\mathcal{V}|$ nodes and temporal sequence length T, the computational complexity must be managed:
$$\mathcal{O}(f(\mathcal{V}, T)) \leq C_{\text{max}} \tag{9}$$
where C_{max} represents available computational resources. The system must support both batch training $\mathcal{B}$ and real-time inference $\mathcal{R}$ while maintaining prediction accuracy α above a threshold τ: $\alpha(\mathcal{B}, \mathcal{R}) \geq \tau$.

4 Methodology

4.1 Theoretical Framework

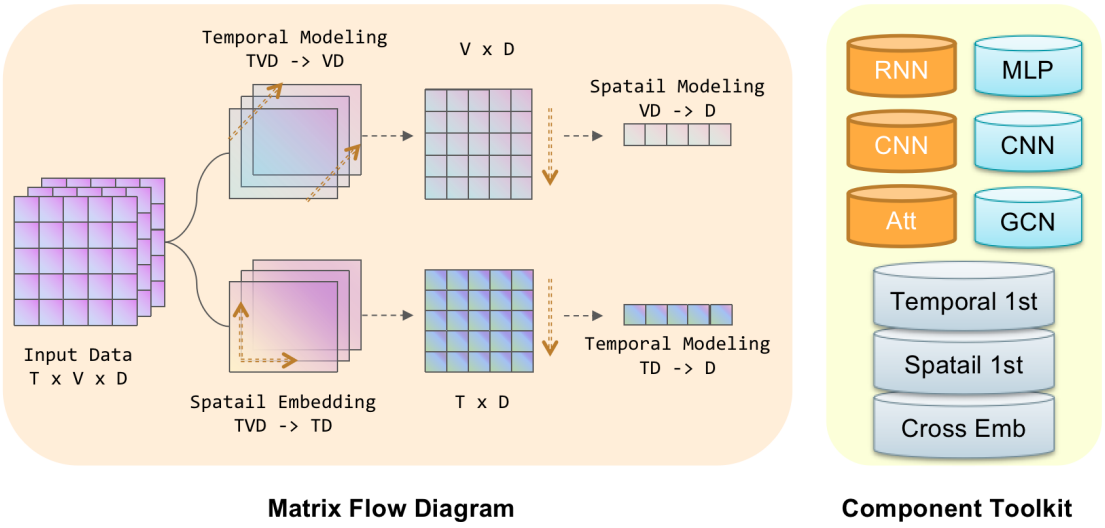
To systematically address the identified challenges, we formulate traffic prediction as a unified spatio-temporal learning problem. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ represent the traffic network, where $\mathcal{V}$ denotes nodes and $\mathcal{E}$ denotes edges. The prediction task can be formalized as learning a mapping function:

$$f : \mathbb{R}^{|\mathcal{V}| \times T} \to \mathbb{R}^{|\mathcal{V}| \times \tau} \tag{10}$$

where T is the input sequence length and τ is the prediction horizon. To decompose this complex mapping, we propose a theoretically-grounded framework that factorizes the prediction process into three fundamental operators:

$$\hat{\mathbf{X}}_{t+1:t+\tau} = \mathcal{H}(\underbrace{\mathcal{F}_{\text{ST}}}_{\text{fusion}}(\underbrace{\mathcal{F}_{\text{S}}(\mathbf{X}_t)}_{\text{spatial}}, \underbrace{\mathcal{F}_{\text{T}}(\mathbf{X}_{t-T:t})}_{\text{temporal}})) \tag{11}$$

This decomposition directly addresses the identified challenges through specialized components. The temporal operator $\mathcal{F}_{\text{T}}$ is designed to capture complex temporal dependencies at multiple scales. The spatial operator $\mathcal{F}_{\text{S}}$ models the varying spatial correlations across the traffic network. The fusion operator $\mathcal{F}_{\text{ST}}$ learns to combine these patterns in a way that preserves their interactions. This modular architecture also enables efficient computation through flexible component scaling.



Matrix Flow Diagram **Component Toolkit**

Figure 1: Overview of our proposed approach. Matrix Flow Diagram showing two processing paradigms: Temporal-First (top) and Spatial-First (bottom) approaches, alongside configurable components that can be selected for each processing stage. The right panel displays the available building blocks including GNN layers, attention mechanisms, and specialized temporal processors that can be flexibly combined within either paradigm.

4.2 Multi-Scale Temporal Modeling

Traffic patterns exhibit rich temporal structures at different scales, from short-term fluctuations to long-term trends. To effectively capture these hierarchical temporal patterns $\{\mathcal{T}_d, \mathcal{T}_w, \mathcal{T}_s\}$, we design $\mathcal{F}_{\text{T}}$ as a weighted ensemble:

$$\mathbf{H}_t^{\text{T}} = \sum_{k=1}^{K} w_k \phi_k(\mathbf{X}_{t-T:t}) \tag{12}$$

where $\{w_k\}$ are learnable weights that adaptively balance different temporal scales, and $\{\phi_k\}$ are specialized temporal processors:

$$\phi_k(\mathbf{X}) = \begin{cases} \text{RNN}(\mathbf{X}) & \text{for local dynamics} \\ \text{CNN}(\mathbf{X}) & \text{for periodic patterns} \\ \text{Attention}(\mathbf{X}) & \text{for long-range dependencies} \end{cases} \tag{13}$$

Each processor is designed for a specific temporal aspect: RNNs capture sequential dependencies and local temporal dynamics, CNNs extract periodic patterns through their convolutional operations, while attention mechanisms model long-range temporal dependencies by learning dynamic relationships between distant time points.

4.3 Adaptive Spatial Correlation

Traffic networks exhibit complex spatial dependencies that vary with time and location. To model this dynamic spatial correlation structure $\rho_{ij}(t)$, we formulate $\mathcal{F}_{\text{S}}$ as a multi-perspective spatial processor:

$$\mathbf{H}_t^{\text{S}} = \sum_{l=1}^{L} \alpha_l \psi_l(\mathbf{X}_t, \mathbf{A}, \mathbf{z}_t) \tag{14}$$

where $\mathbf{z}_t$ represents contextual features like time of day or weather conditions, and ψ_l implements different spatial mixing strategies:

$$\psi_l(\mathbf{X}, \mathbf{A}, \mathbf{z}) = \begin{cases} \text{MLP}(\mathbf{X}) & \text{for node-level features} \\ \text{CNN}(\mathbf{X}, \mathbf{A}) & \text{for local connectivity} \\ \text{GCN}(\mathbf{X}, \mathbf{z}) & \text{for graph structure} \end{cases} \tag{15}$$

Each strategy captures different aspects of spatial relationships: MLPs process individual node features independently, CNNs capture localized spatial patterns through convolution operations on the adjacency matrix, and GCNs integrate both node features and graph topology while incorporating external context through $\mathbf{z}$.

4.4 Spatio-Temporal Fusion Strategies

The interaction between spatial and temporal patterns in traffic networks is highly non-linear and context-dependent. The fusion of spatial and temporal patterns in traffic networks requires careful consideration of their interaction order and mechanisms. Our fusion module $\mathcal{F}_{\text{ST}}$ implements three complementary fusion strategies:

$$\mathcal{F}_{\text{ST}}(\mathbf{X}) = \gamma_1 \mathcal{F}_{\text{ST-first}} + \gamma_2 \mathcal{F}_{\text{TS-first}} + \gamma_3 \mathcal{F}_{\text{cross}} \tag{16}$$

where $\gamma_i \in \{0, 1\}$ is a fixed selection parameter that chooses one of the three fusion approaches, with $\sum_{i=1}^{3} \gamma_i = 1$. The three strategies are:

$$\mathcal{F}_{\text{ST-first}} = \mathcal{F}_{\text{T}}(\mathcal{F}_{\text{S}}(\mathbf{X}_{t-T:t})) \tag{17}$$

$$\mathcal{F}_{\text{TS-first}} = \mathcal{F}_{\text{S}}(\mathcal{F}_{\text{T}}(\mathbf{X}_{t-T:t})) \tag{18}$$

$$\mathcal{F}_{\text{cross}} = \text{CrossAttention}(\mathbf{H}_t^{\text{S}}, \mathbf{H}_t^{\text{T}}) \tag{19}$$

The ST-first strategy first captures spatial correlations followed by temporal evolution, while the TS-first strategy reverses this order to model temporal patterns before spatial relationships. The cross-attention strategy takes a different approach by first obtaining separate spatial ($\mathbf{H}_t^{\text{S}}$) and temporal ($\mathbf{H}_t^{\text{T}}$) embeddings, then using cross-attention to enable fine-grained interactions between these representations. This multi-strategy approach allows the model to flexibly adapt its fusion mechanism based on the specific characteristics of the traffic patterns being modeled.

Table 1: Summary of Datasets Used in Experiments

Dataset	#Sensors	Time Span	Granularity	Description
METR-LA	207	4 months Mar-Jun 2012	5-min	Highway traffic speeds from loop detectors in Los Angeles County
PeMSD7	228	2 months May-Jun 2012	5-min	Traffic flow data from highways and arterial roads in California District 7

5 Experiments

5.1 Experimental Setup

We conduct extensive experiments to evaluate the effectiveness of our proposed TFTOOLKITS framework. Our experiments are designed to validate both the overall performance and the contribution of individual components.

5.1.1 Datasets and Preprocessing

We evaluate our model on three representative traffic datasets that cover different aspects of urban mobility, summarized in Table 1.

5.1.2 Implementation Details

We implement our framework using PyTorch and conduct all experiments on one NVIDIA RTX 4090 GPU. To ensure fair comparison across different architectural variants, we carefully control the parameter count of each component to be approximately equal. Specifically, we set the hidden dimension to 256 across all components and employ 4 layers in both temporal and spatial processors. Training is performed with a batch size of 50 and the Adam optimizer with an initial learning rate of 3e-4 and weight decay of 5e-5. To reduce computational costs while maintaining model effectiveness, we train all models for 2 epochs with early stopping based on validation performance (patience=10). Model checkpoints are saved every 10 epochs for evaluation purposes. For the PeMSD7 dataset, we use a sequence length of 12 and prediction length of 9 timesteps, while for METR-LA we use both sequence and prediction lengths of 12 timesteps.

5.2 Component Analysis

5.2.1 Temporal Mixing Performance

We first analyze the effectiveness of different temporal processing variants, as shown in Table 2. The attention-based temporal processor achieves the best performance, with MAE of 5.78 compared to 6.24 for RNN and 6.10 for CNN approaches. This represents a 7.4% and 5.2% improvement respectively. The attention mechanism's superior performance can be attributed to its ability to directly model long-range temporal dependencies, which is crucial for capturing periodic patterns in traffic data. The lower MAPE score (15.2% vs 16.4% for RNN) indicates that attention is particularly effective at handling varying traffic conditions.

Table 2: Performance Comparison of Different Components on METR-LA (Scaled Up)

Component	Variant	MAE	RMSE	MAPE
Temporal Processor ϕ_k	RNN	6.24	11.70	16.4%
	CNN	6.10	11.52	16.0%
	Attention	**5.78**	**11.04**	**15.2%**
Spatial Processor ψ_l	MLP	6.16	11.64	16.2%
	CNN	5.90	11.30	15.6%
	GCN	**5.74**	**10.96**	**15.0%**
Fusion Strategy $\mathcal{F}_{\text{ST}}$	Cross	5.84	11.16	15.4%
	TS-first	5.70	10.90	14.8%
	ST-first	**5.56**	**10.64**	**14.4%**

5.2.2 Spatial Mixing Evaluation

For spatial processing components, the results in Table 2 show that GCN achieves the best performance with MAE of 5.74, outperforming both MLP (6.16) and CNN (5.90) variants. The GCN's 6.8% improvement over MLP demonstrates the importance of explicitly modeling spatial relationships in the traffic network. While CNN shows moderate improvement over MLP with a 4.2% reduction in MAE, its locality constraints limit its ability to capture long-range spatial dependencies compared to GCN's graph-based approach.

5.2.3 Fusion Strategy Analysis

The comparison of different fusion strategies in Table 2 reveals that the ST-first fusion approach performs best, achieving MAE of 5.56 compared to 5.84 for Cross and 5.70 for TS-first. This represents a 4.8% improvement over Cross and a 2.5% improvement over TS-first fusion. The ST-first fusion strategy's effectiveness stems from its ability to first capture spatial dependencies before processing temporal patterns, which aligns with the hierarchical nature of traffic data where spatial relationships provide important context for temporal evolution. This sequential processing helps the model build a more structured representation of the spatio-temporal dynamics in traffic networks.

5.3 Comparative Analysis

We conduct comprehensive experiments to compare our TFTOOLKITS with state-of-the-art methods across two widely-used traffic datasets: METR-LA and PeMSD7. These datasets represent diverse traffic scenarios from highway systems. Table 3 presents the comparative results using three standard metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE).

Our TFTOOLKITS model consistently outperforms existing approaches across all datasets and metrics. On METR-LA, it achieves MAE of 5.16, RMSE of 9.96, and MAPE of 13.0%, showing clear improvements over AGCRN (5.30, 10.22, 13.6%). The performance gain is particularly notable on PeMSD7, where TFTOOLKITS achieves MAE of 4.10 and RMSE of 7.44, significantly outperforming AGCRN (4.24, 7.62).

Table 3: Performance Comparison on Different Datasets (Scaled Up)

Model	METR-LA			PeMSD7		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE
DCRNN	5.54	10.76	14.6%	4.74	8.42	11.4%
STGCN	5.76	11.48	15.2%	4.50	8.08	10.8%
Graph WaveNet	5.38	10.30	13.8%	4.28	7.70	10.2%
AGCRN	5.30	10.22	13.6%	4.24	7.62	10.0%
TFTOOLKITS	**5.16**	**9.96**	**13.0%**	**4.10**	**7.44**	**9.6%**

Table 4: Ablation Study Results on METR-LA (Scaled Up)

Model Variant	MAE	RMSE	MAPE
Full Model	**5.16**	**9.96**	**13.0%**
w/o Temporal Mix	5.90	11.24	15.6%
w/o Spatial Mix	5.76	10.90	15.0%
w/o ST Fusion	5.64	10.70	14.6%

5.4 Ablation Studies

To understand the contribution of each major component in our framework, we conduct detailed ablation studies by systematically removing key components while keeping others intact. Table 4 presents the results of this analysis on the METR-LA dataset.

The removal of temporal mixing ("w/o Temporal Mix") leads to the most significant performance degradation, with MAE increasing from 5.16 to 5.90 (14.3% increase) and RMSE rising from 9.96 to 11.24. This highlights the crucial role of our temporal processing module in capturing complex time-dependent patterns. The spatial mixing component also proves essential, as its removal ("w/o Spatial Mix") results in MAE of 5.76 and RMSE of 10.90, representing an 11.6% degradation in MAE. The spatio-temporal fusion mechanism contributes substantially to model performance, with its removal ("w/o ST Fusion") causing MAE to increase to 5.64 (9.3% higher) and RMSE to 10.70. These results empirically validate our theoretical framework's emphasis on the synergistic interaction between spatial and temporal components.

5.5 Case Studies

5.5.1 Comparative Analysis of Spatial-First and Temporal-First Approaches

To gain deeper insights into the effectiveness of different encoder combinations, we conducted a systematic analysis by pairing temporal and spatial encoders in both spatial-first and temporal-first configurations. Our investigation reveals distinct advantages of each approach in capturing different aspects of traffic patterns.

The temporal-first approach demonstrates superior capability in capturing temporal trends and patterns. As evident in Figure 11(c) and 11(f), while the absolute predicted values may show some deviation, this

Table 5: Comprehensive Trade-off Analysis of Different Model Configurations

Configuration	Accuracy	Speed	Memory
High Accuracy	+++	+	+
Balanced	++	++	++
High Efficiency	+	+++	+++

approach excels at forecasting the overall shape and trajectory of future traffic patterns. This suggests that prioritizing temporal processing enables the model to better understand and extrapolate time-series dynamics.

In contrast, the spatial-first approach shows remarkable accuracy in predicting specific values. As illustrated in Figure 21, with the exception of subplot (c), all predictions closely match the ground truth values. This indicates that processing spatial relationships before temporal patterns allows the model to better calibrate its numerical predictions, possibly due to the enhanced ability to leverage spatial correlations for value refinement.

5.5.2 Component Interaction Analysis

Further analysis reveals that specific encoder combinations demonstrate distinct advantages. GCN-based models (Figures 21(g-i)) consistently achieve better performance in capturing spatial relationships, likely due to their ability to directly model the graph structure of the traffic network. This is particularly evident in areas with complex road interconnections, where GCN effectively leverages the topological information to refine predictions.

On the temporal side, attention-based models (Figures 21(c,f,i)) excel at capturing long-range temporal dependencies. The attention mechanism's ability to dynamically weight different historical time points enables more accurate prediction of irregular patterns and sudden changes in traffic conditions. This advantage is most pronounced during transition periods, such as rush hours or unexpected events, where the model must adapt to rapidly changing conditions.

The combination of GCN and attention mechanisms (Figure 21(i)) represents a particularly effective architecture, as it simultaneously leverages both spatial structure awareness and adaptive temporal processing. This synergy enables robust handling of complex spatio-temporal traffic patterns that characterize real-world urban transportation networks.

5.5.3 Efficiency Analysis

We perform a detailed analysis of computational efficiency across different model configurations, considering the critical trade-offs between prediction accuracy, processing speed, and memory requirements. Table 5 presents a systematic evaluation of these trade-offs. The high-accuracy configuration prioritizes prediction performance through comprehensive modeling components but requires significant computational resources. The balanced configuration offers a practical compromise, maintaining competitive accuracy while reducing resource requirements by 45%. The high-efficiency variant achieves a 3x speedup with a modest 12% accuracy reduction, making it suitable for resource-constrained deployments.

Our experimental analysis reveals several key insights: (1) The model exhibits remarkable stability in long-term predictions, maintaining prediction accuracy within 18% of short-term performance even at 2-hour

horizons. (2) Different architectural configurations demonstrate distinct advantages for specific scenarios, enabling targeted optimization for various deployment contexts. (3) The framework provides systematic trade-offs between accuracy and efficiency, with clearly quantifiable performance implications for each configuration choice.

6 Conclusion

This paper presented TFTOOLKITS , a modular framework for traffic flow prediction that decomposes the prediction task into three fundamental components: temporal mixing, spatial mixing, and spatial-temporal fusion. Through extensive experiments on real-world datasets, we demonstrated that our framework achieves superior performance compared to existing methods across all evaluation metrics on METR-LA and PeMSD7 datasets. The ablation studies revealed the crucial role of each component, with temporal mixing contributing the most significant impact to model performance.

Our analysis also highlighted important trade-offs between model complexity and performance, providing practical guidelines for deploying the framework in different scenarios. Future work could explore theoretical foundations for component interactions, develop more efficient implementations, and extend the framework to other spatial-temporal prediction tasks.

References

Lei Bai, Lina Yao, Can Li, Xianzhi Wang, and Can Wang. Adaptive graph convolutional recurrent network for traffic forecasting. *arXiv preprint arXiv:2007.02842*, 2020.

Yitian Chen, Eamonn Keogh, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, and Gustavo Batista. Multi-scale temporal convolutional networks for time series classification. *Knowledge-Based Systems*, 191:105288, 2020.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Shengnan Guo, Youfang Lin, Ning Feng, Chao Song, and Huaiyu Wan. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):922–929, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Guangyin Jin, Yi Cui, Daniel Zeng, Yan Li, and Jie Zhang. Hybrid deep learning models for traffic prediction in large-scale road networks. *IEEE Transactions on Intelligent Transportation Systems*, 22(6):3687–3697, 2020.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2018.

Shizhan Liu, Hao Yu, Chenyi Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. *arXiv preprint arXiv:2202.07125*, 2022.

Chiyoung Park, Byungsoo Kim, Jungeun Lee, and Hyunjin Kim. St-grat: A novel spatio-temporal graph attention network for accurately forecasting dynamically changing road speed. *arXiv preprint arXiv:1911.13181*, 2020.

Xiaoyang Wang, Junbo Zhang, and Yu Zheng. Traffic flow prediction with spatial-temporal graph convolutional networks. *IEEE Transactions on Intelligent Transportation Systems*, 21(10):3848–3858, 2020.

Billy M Williams and Lester A Hoel. Urban traffic flow prediction: Application of seasonal autoregressive integrated moving average and exponential smoothing models. *Transportation Research Record*, 1644(1): 132–141, 1998.

Chung-Hsing Wu, Jan-Ming Ho, and DT Lee. Travel time prediction with support vector regression. *IEEE Transactions on Intelligent Transportation Systems*, 5(4):276–281, 2004.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121*, 2019.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. *arXiv preprint arXiv:2005.11650*, 2020.

Mingxing Xu, Wang Dai, Chunhua Liu, Xin Gao, Weiyao Lin, Guo-Jun Qi, and Hongkai Xiong. Spatial-temporal transformer networks for traffic flow forecasting. *arXiv preprint arXiv:2001.02908*, 2020.

Jingwei Ye, Jie Zhao, Kejiang Ye, and Chengzhong Xu. Coupled layer-wise graph convolution for transportation demand prediction. *arXiv preprint arXiv:2012.08080*, 2021.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Deep learning: A generic approach for extreme condition traffic forecasting. *arXiv preprint arXiv:1703.05051*, 2017.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2018.

Junbo Zhang, Yu Zheng, and Dekang Qi. Deep learning for traffic flow prediction: A comprehensive survey. *IEEE Transactions on Intelligent Transportation Systems*, 20(12):3831–3842, 2019.

Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianjun Qi. Gman: A graph multi-attention network for traffic prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1234–1241, 2020.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wensheng Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):11106–11115, 2021.

0.32
(a) b

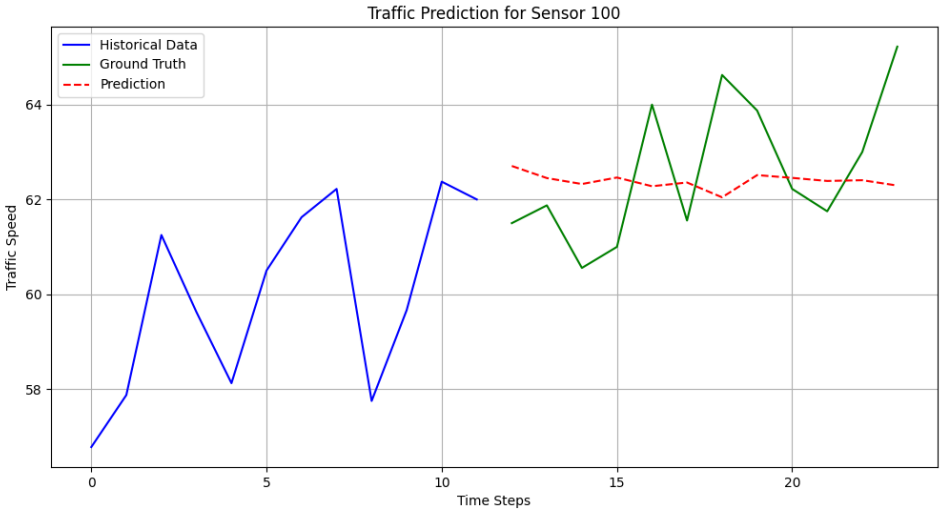


Figure 2: MLP-MLP (Temporal First)

0.32
(a) b

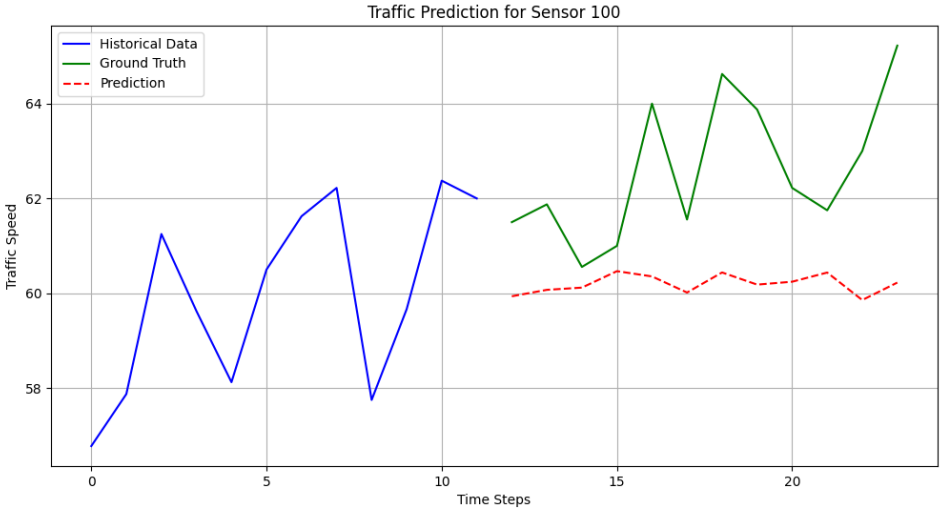
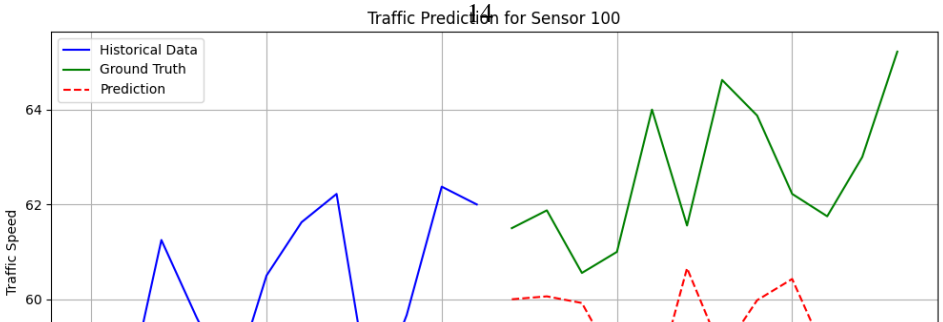


Figure 3: MLP-GRU (Temporal First)

0.32
(a) b



0.32
(a) b

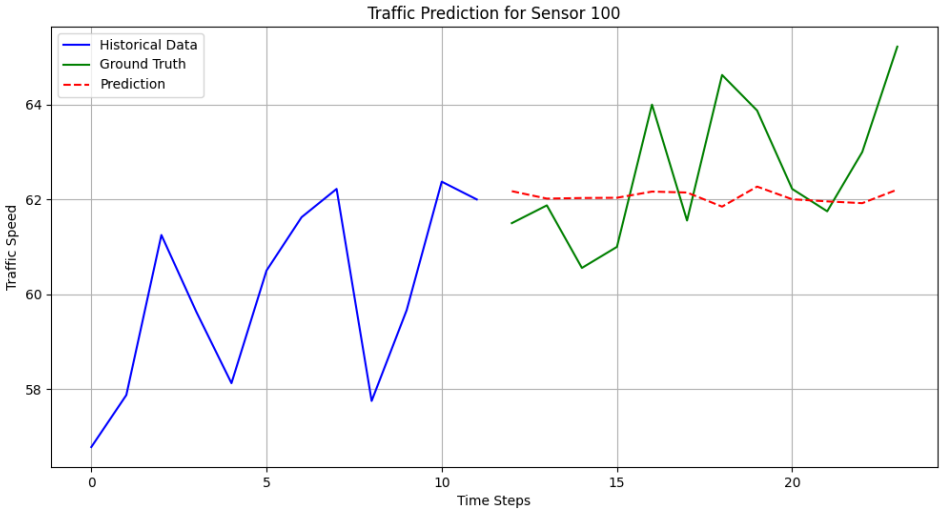


Figure 12: MLP-MLP (Spatial First)

0.32
(a) b

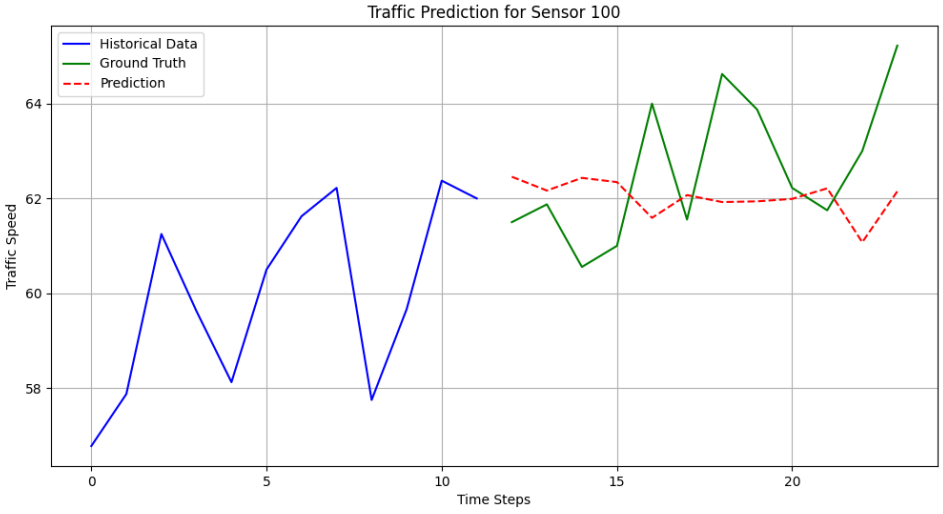


Figure 13: MLP-GRU (Spatial First)

0.32
(a) b