

Article

A Study on Urban Traffic Speed Prediction Using Spatiotemporal Attention-Based Convolutional Neural Networks

Zhuye Xu ^{1,2}, Shaoyang Wang ^{2,*} , Hongqiang Xin ² and Zhongdong Hua ³¹ National Engineering Research Center of Highway Maintenance Technology, Changsha University of Science and Technology, Changsha 410114, China² School of New Energy and Power Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China³ Gansu Aomake Education Technology Co., Ltd., Pingliang 744000, China

* Correspondence: 11250965@stu.lzjtu.edu.cn

Abstract

Urban traffic speed prediction requires effective modeling of road-network spatial dependencies and temporal dynamics. This study develops an attention-based spatiotemporal graph convolutional framework (AT-GCN-GRU) for multi-horizon traffic speed prediction. The model combines graph convolution for spatial feature extraction, GRU for temporal modeling, and a temporal attention mechanism for adaptively weighting historical hidden states. Experiments were conducted on the SZ-taxi dataset and further evaluated on METR-LA. On SZ-taxi, AT-GCN-GRU was compared with conventional and modern baselines, including T-GCN, ASTGCN, and AGCRN, over 15, 30, 45, and 60 min forecasting horizons. The proposed model remained competitive with the modern baselines and obtained lower mean RMSE than T-GCN across the four evaluated horizons, with relative reductions of approximately 1.4–2.9%. Additional ablation, missing-data robustness, attention-weight visualization, and computational-cost analyses were conducted. The results suggest that temporal attention can provide incremental improvements to the GCN-GRU framework with relatively limited additional computational overhead.

Keywords: traffic speed prediction; graph convolutional network; gated recurrent unit; temporal attention mechanism; intelligent transportation systems; spatiotemporal dependencies

1. Introduction

As the rate of global urbanization continues to accelerate, the conflict between the explosive growth of motor vehicles and limited road resources is becoming increasingly acute. Traffic congestion is no longer merely an issue of commuting efficiency; it has evolved into a systemic crisis that hinders the sustainable development of cities. Research by Wang et al. [1] indicates that the excessive energy consumption and surge in carbon emissions caused by traffic congestion pose a severe challenge to the urban ecological environment. In light of these challenges, the development of intelligent transportation systems (ITSs) has emerged as a consensus among nations to address these urban issues. In their discussion on smart cities, Li Deren et al. [2] emphasize that leveraging Internet of Things (IoT) and big data technologies to achieve precise monitoring and regulation of the city's pulse is the core direction for future urban management.

From a data mining perspective, urban traffic speed data is not a simple linear time series, but rather a form of graph-structured spatiotemporal data [3]. When constructing efficient prediction models, researchers face two main technical bottlenecks. First is the challenge



Academic Editor: Dario Di Cara

Received: 2 August 2026

Revised: 31 August 2026

Accepted: 8 September 2026

Published: 11 September 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

of modeling non-Euclidean space. Urban road networks consist of intricate topological connections (such as roundabouts and overpasses) rather than regular grid structures. As Dai et al. [4] demonstrated in their research on time-varying graph neural networks, although there is a significant spatial spillover effect in traffic conditions between road segments, this correlation is constrained by the complex physical topology of the road network and does not follow geographical distance alone. While traditional convolutional neural networks (CNNs) perform exceptionally well in image processing, their convolution kernels—based on Euclidean grid spaces—cannot directly process such irregular graph data. Research by Seng et al. [5] further corroborates that modeling approaches disregarding road network topology fail to achieve optimal prediction results in complex urban road networks.

Second is the challenge of capturing dynamic long-term dependencies. Traffic flow exhibits highly nonlinear dynamic changes over time. As reported in a recent study on spatiotemporal graph forecasting [6], the combined influence of complex road network topology and dynamic temporal patterns results in significant fluctuations regarding how different historical time points impact future states at the micro-scale. This complicates the task of traditional single-layer network models in capturing true dynamic correlations. For example, during a sudden accident in the morning rush hour, traffic flow characteristics from the past 15 min are often more decisive than the historical average from an hour prior. However, traditional recurrent neural networks (RNNs) frequently suffer from information forgetting when processing long sequences, making it difficult to adaptively focus on critical time steps.

In order to address these challenges, research in this field has evolved from parametric statistical models to deep learning models. Early methods, including the Historical Averages (HAs) and ARIMA models, were constrained by linear assumptions, which hindered the capture of complex nonlinear traffic flows [7]. In their review, Cui et al. [8] concluded that while statistical models offer strong interpretability, they often show limited performance when faced with complex and dynamic traffic patterns.

With the advent of deep learning, spatiotemporal graph neural networks (ST-GNNs) have gradually become the mainstream approach. The T-GCN model [9] successfully integrates graph convolutional networks (GCNs) with gated recurrent units (GRUs), establishing a performance benchmark in this field. However, Jiang and Luo [10] note that prevailing models (e.g., the baseline T-GCN) generally assign equal weight to all historical time points when processing time series, or rely entirely on RNN hidden state propagation. This mechanism lacks the ability to actively filter information, resulting in a precipitous decline in accuracy as the prediction horizon increases (e.g., up to 60 min).

Domestic researchers, including Yin et al. [11], have also attempted to incorporate attention mechanisms into BiLSTM architectures. However, the application of these mechanisms to graph-structured data remains to be further explored. Consequently, recent research has shifted toward introducing complex self-attention mechanisms or graph neural networks with global spatiotemporal reconstruction. Heavyweight models, including the classic ASTGCN [12] and recent adaptive decoupled graph attention models [13,14], are capable of capturing multidimensional spatiotemporal dependencies. However, the complex attention matrices and dynamic graph generation processes of these models frequently incur high computational costs, posing significant challenges for deployment in real-time traffic systems that require extremely low latency.

In order to address the limitations of extant methods in the domains of long-term prediction and key information capture, the present paper proposes a spatiotemporal graph convolutional network prediction model (AT-GCN-GRU) that integrates an attention mechanism. This study utilizes real-world road network data from Luohu District, Shen-

zhen (SZ-taxi) to validate the model, with the objective of achieving highly accurate speed predictions through this technical integration.

The main contributions of this study are summarized as follows:

- (1) An end-to-end AT-GCN-GRU framework is developed for multi-horizon traffic speed prediction. Rather than introducing a new graph convolution or attention operator, the framework adopts a lightweight sequential architecture in which graph convolution is used to extract road-network spatial information, GRU is used to model temporal evolution, and a temporal attention layer is applied to the GRU hidden-state sequence before prediction. This design allows historical time steps to contribute unequally to the final prediction while maintaining a relatively compact model structure.
- (2) This study investigates the role of temporal attention in the GCN-GRU forecasting framework, with particular emphasis on longer prediction horizons. The attention module adaptively assigns weights to historical hidden states, providing an explicit mechanism for selecting temporally informative representations rather than relying solely on the final recurrent hidden state.
- (3) The proposed framework is evaluated for traffic-speed forecasting over multiple prediction horizons. The experiments examine prediction accuracy together with model complexity, convergence behavior, ablation settings, and robustness to incomplete observations, providing an empirical assessment of the trade-off between forecasting performance and computational overhead.
- (4) This study further analyzes the practical characteristics and limitations of the lightweight architecture. Rather than aiming to replace more complex adaptive-graph or Transformer-based models, AT-GCN-GRU is investigated as a comparatively simple alternative for scenarios in which computational cost and implementation complexity are also important considerations.

2. Related Work

2.1. Traditional Traffic Forecasting Methods

Before the widespread adoption of deep learning, traffic forecasting predominantly relied on statistical models [15]. The Historical Average (HA) method is frequently employed as a baseline due to its computational simplicity; however, it fails to handle sudden changes in traffic conditions. The ARIMA model and its variants, proposed in the 1970s, remain classic tools in time series analysis. Kumar and Vanajakshi [16] demonstrated the effectiveness of seasonal ARIMA models in capturing periodicity in traffic flow. Furthermore, Box–Jenkins techniques have been employed to analyze freeway traffic time-series data [17].

To overcome the linear limitations of statistical models, traditional machine learning methods such as Support Vector Regression (SVR) were introduced. Wu et al. [18] utilized SVR for short-term traffic flow forecasting, demonstrating the efficacy of kernel methods in addressing nonlinear relationships. Habtemichael et al. [19] proposed a nonparametric regression method based on K-nearest neighbors, which predicts future states by matching similar historical patterns. However, these conventional methods generally treat individual road segments as isolated variables, disregarding the intricate spatial interactions within the road network, and exhibit suboptimal computational efficiency in large-scale data environments.

2.2. Deep Learning-Based Spatiotemporal Modeling Methods

In recent years, deep learning has emerged as the prevailing approach to traffic forecasting owing to its advanced feature representation capabilities. Existing research has primarily focused on designing network architectures capable of addressing both temporal and spatial dependencies simultaneously.

2.2.1. Deep Modeling of the Temporal Dimension

In the domain of time series modeling, recurrent neural networks (RNNs) and their variants dominate. Ma et al. [20] were the first to utilize long short-term memory (LSTM) networks to capture long-term dependencies in traffic speed data, thereby overcoming the vanishing gradient problem associated with traditional RNNs. Fu et al. [21] further validated the effectiveness of gated recurrent units (GRUs) in traffic flow forecasting, demonstrating that GRUs maintain accuracy comparable to LSTMs while featuring fewer parameters and faster convergence rates.

2.2.2. From Grid to Graph in Spatial Dimensions

In the spatial domain, early studies attempted to transform urban road networks into regular grids and use CNNs to extract spatial features [22]. Among these, the DMVST-Net proposed by Yao et al. [23] combines CNNs and LSTMs through a multi-view mechanism. However, this grid-based approach inevitably results in the loss of the road network's fine-grained topological structure.

The advent of graph convolutional networks (GCNs) has fundamentally transformed the processing of non-Euclidean data. The spectral graph convolution theory proposed by Kipf and Welling [24] laid the mathematical foundation for GCNs. Ye et al. [25] recently conducted an in-depth analysis of the working mechanism of GNNs, highlighting their advantages in information aggregation. Building on this, the T-GCN model proposed by Zhao et al. successfully integrated GCN with GRU, utilizing an adjacency matrix to describe road network topology, and has become a classic benchmark in this field. Subsequent studies further incorporated attention mechanisms into graph-based recurrent traffic forecasting. For example, Bai et al. proposed A3T-GCN [26], which assigns different importance to historical time steps to capture global temporal dependencies. These studies indicate that combining graph convolution, recurrent modeling, and temporal attention is not itself a new methodological idea; therefore, the present study focuses on a lightweight post-GRU attention configuration and its accuracy–efficiency trade-off.

Domestic scholars have also made significant progress in this area. Sun et al. [27] proposed a dynamic spatiotemporal graph convolutional network, which further accounts for the dynamic changes in road network weights over time. Li et al. [28] explored deep learning models that integrate multi-dimensional features, demonstrating the importance of feature fusion for improving accuracy. Pan et al. [29] and Li et al. [30] also explored model fusion through HyperST-Net and ST-Fusion, respectively. These studies collectively indicate that GCN-based methods have become the state-of-the-art (SOTA) solutions for addressing spatial dependencies in road networks.

2.2.3. Recent Advances in Transformer-Based and Adaptive Graph Learning

In recent years, adaptive graph learning and Transformer architectures have emerged as the latest research frontiers in addressing the challenges of long-range forecasting. To address the limitation of fixed adjacency matrices in capturing real-time road network dynamics, Zhang et al. [31] proposed an adaptive graph augmentation framework based on spatiotemporal contrastive learning, while recent work [32] explored spatiotemporal context graph learning methods with dynamic updates. Furthermore, Zheng et al. [33] combined a Transformer network with a virtual dynamic graph, and a recent study [34] introduced an adaptive spatiotemporal graph convolutional network to further enhance the modeling of dynamic traffic relationships. Recently, the graph evolution model proposed by Choi et al. [35] has also provided a novel continuous-time dynamic framework for handling traffic non-stationarity.

Regarding the evolution of attention architectures, Transformer models have demonstrated exceptional potential for modeling long-range dependencies. For example, the wavelet spatiotemporal attention network proposed by Sun et al. [36] and the two-layer synchronous attention network constructed by Jia et al. both utilize multi-head mechanisms to significantly improve multi-step prediction accuracy. However, despite the excellent performance of these state-of-the-art models, high-frequency dynamic graph reconstruction and global self-attention computations inevitably introduce substantial memory consumption and inference latency. While incorporating the focus on dynamic features from cutting-edge research, the AT-GCN-GRU model proposed in this paper employs a streamlined temporal attention design to achieve an optimal balance between long-term prediction accuracy and real-time computational efficiency.

3. Method

3.1. Problem Definition

In this study, we formulate the prediction of traffic speeds on urban roads as a spatiotemporal regression problem on a graph structure.

We formalize the urban road network as a weighted undirected graph $G \equiv (V, E, A)$. Here, $V = \{v_1, v_2, \dots, v_N\}$ represents the set of N road segment nodes (in this experiment, $N = 156$), and E represents the set of connected edges between road segments. $A \in R^{N \times N}$ is the Adjacency Matrix describing the topological structure of the road network. If road segments v_i and v_j are directly connected in space, then $A_{ij} = 1$, otherwise $A_{ij} = 0$.

Definition 3.2 (Traffic State Features): We use $X_t \in R^{N \times F}$ to denote the network-wide traffic state matrix at time step t , where F is the feature dimension. In this study, we primarily focus on traffic speed, hence $F = 1$.

Prediction Target: Given the road network topology G and the traffic speed sequence of the past T time steps $\mathcal{X}_{hist} = (X_{t-T+1}, \dots, X_t)$, this study aims to train a deep learning model $f(\cdot)$ to predict the traffic speed sequence for the future τ time steps $\gamma_{pred} = (X_{t+1}, \dots, X_{t+\tau})$. The mathematical mapping relationship is as follows:

$$[X_{t+1}, \dots, X_{t+\tau}] = f(G, [X_{t-T+1}, \dots, X_t]) \quad (1)$$

Based on the previously determined experimental settings, we set the historical time window length to $T = 12$ (corresponding to the past 3 h), and the prediction step size to $\tau \in \{1, 2, 3, 4\}$ (corresponding to the future 15, 30, 45, and 60 min, respectively). The above temporal interpretation corresponds to the SZ-taxi dataset. For METR-LA, the same input sequence length of $T = 12$ corresponds to the previous 1 h because of its 5 min sampling interval, while the physical prediction horizons remain 15, 30, 45, and 60 min.

3.2. Overall Model Architecture

To capture spatial and temporal dependencies in traffic speed data, this study develops an end-to-end prediction framework, termed AT-GCN-GRU.

As shown in Figure 1, the model consists primarily of three core modules.

- (1) **Spatial Feature Extraction Module (GCN):** Utilizes a graph convolutional network to aggregate traffic information from neighboring road segments and extract local topological features.
- (2) **Temporal Evolution Module (GRU):** This module uses a gated recurrent unit (GRU) to process time series data and capture both short- and long-term temporal dependencies.
- (3) **Temporal Attention Module:** This module introduces an adaptive attention mechanism that dynamically reweights historical hidden states based on the current prediction context to enhance the model's ability to identify critical time steps.

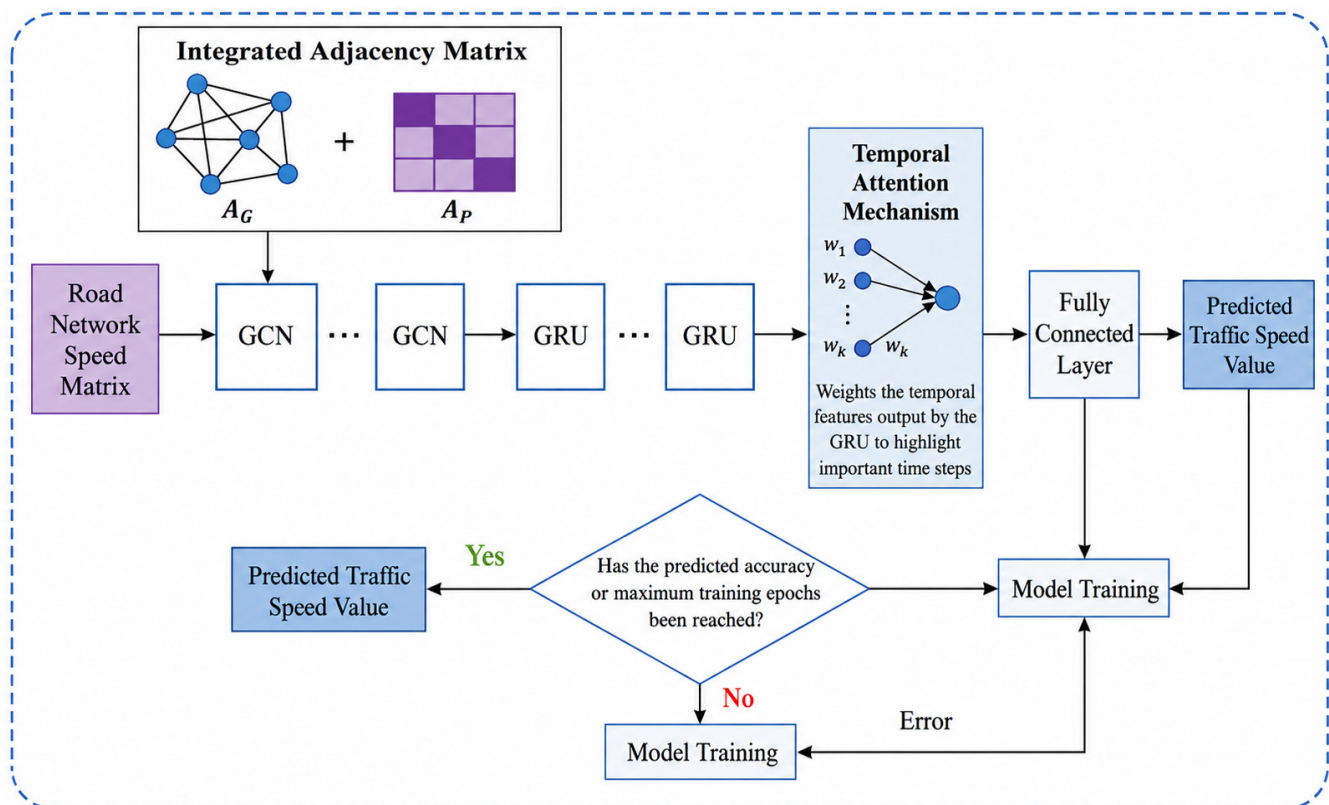


Figure 1. Overall framework diagram.

3.3. Spatial Dependency Modeling

Traditional convolutional neural networks (CNNs) are limited to processing Euclidean grid data (such as images) and cannot be directly applied to road network maps with non-Euclidean structures. Therefore, we employ graph convolutional networks (GCNs) based on spectral graph theory to extract spatial features, and the overall GCN model framework is shown in Figure 2.

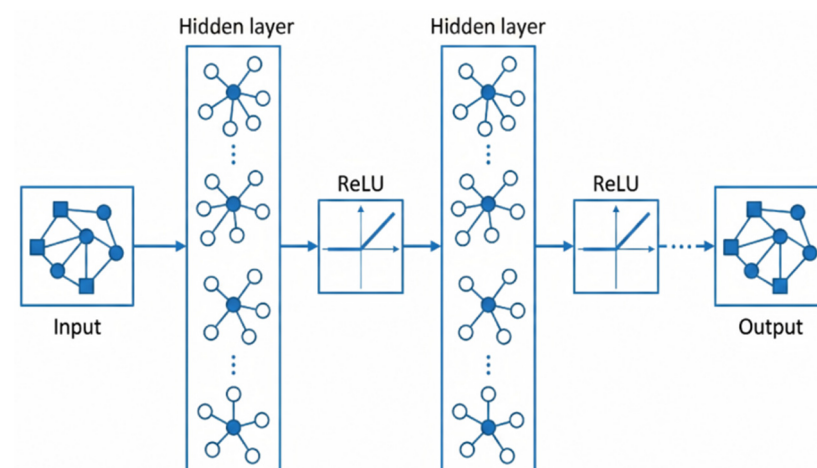


Figure 2. GCN model framework.

The graph convolution operation is essentially the product of a graph signal x and a filter g_θ in the Fourier domain. To avoid expensive eigenvalue decomposition calcula-

tions, we adopt the first-order approximation of Chebyshev polynomials. Specifically, the propagation rule of the graph convolutional layer is defined as follows:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}\right) \quad (2)$$

where $\tilde{A} = A + I_N$ is the adjacency matrix with added self-loops, ensuring that nodes can also retain their own features when aggregating information; and $\tilde{D}$ is the degree matrix of $\tilde{A}$, where $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$. $H^{(l)}$ represents the feature matrix of the l -th layer, and for the first layer, $H^{(0)} = X_t$; $W^{(l)}$ is the learnable weight matrix of the l -th layer; $\sigma(\cdot)$ is the non-linear activation function, and ReLU is selected in this paper; by stacking two layers of GCN, the model can effectively aggregate the traffic states of second-order neighbors, outputting a feature vector $f_t^{spatial}$ containing rich spatial information.

3.4. Time-Dependent Modeling

After obtaining the spatial features, we need to model the dynamic changes in the temporal dimension. Compared to LSTM, the gated recurrent unit (GRU) reduces the number of parameters by simplifying the gating structure, making it more suitable for the real-time processing of traffic data. The GRU model structure diagram is shown in Figure 3.

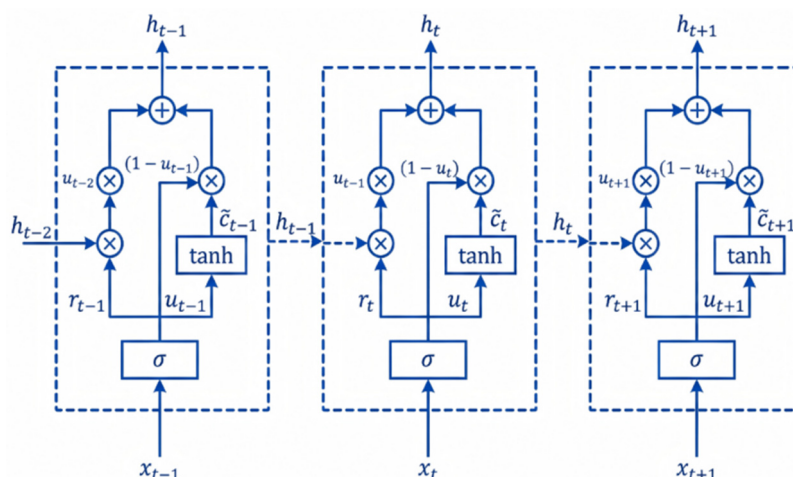


Figure 3. Structure of the GRU model.

We take the output sequence of the GCN as the input to the GRU. At each time step t , the GRU controls the flow and forgetting of information through a reset gate r_t and an update gate z_t :

$$z_t = \sigma\left(W_z\left[f_t^{spatial}, h_{t-1}\right] + b_z\right) \quad (3)$$

$$r_t = \sigma\left(W_r\left[f_t^{spatial}, h_{t-1}\right] + b_r\right) \quad (4)$$

$$\tilde{h}_t = \tanh\left(W_h\left[f_t^{spatial}, r_t \odot h_{t-1}\right] + b_h\right) \quad (5)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \quad (6)$$

where h_t is the hidden state at time t , containing comprehensive spatiotemporal information up to the current moment; $\odot$ denotes the Hadamard product (element-wise multiplication); and W and b are learnable parameters.

3.5. Spatiotemporal Attention Mechanism

Traditional RNN architectures often face an information bottleneck when processing long sequences, meaning that the hidden state at the final time step may not fully encode the key information from the entire historical sequence. To address this issue, we introduced a temporal attention mechanism following the GRU layer.

The attention mechanism allows the model to “look back” at all historical hidden states $H = \{h_1, h_2, \dots, h_T\}$, and assigns an importance weight α_t to each time step. The weight calculation formula is as follows:

$$e_t = v^T \tanh(W_s h_t + W_q h_T + b_a) \quad (7)$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^T \exp(e_k)} \quad (8)$$

where h_T is the hidden state at the last time step (representing the global context query vector), and W_s , W_q , v are learnable parameter matrices.

The final context vector C is the weighted sum of all historical states:

$$C = \sum_{t=1}^T \alpha_t h_t \quad (9)$$

Finally, we input the context vector C into a fully connected layer to generate the final future speed prediction value $\hat{Y}$.

4. Experiments and Analysis of Results

4.1. Datasets

4.1.1. SZ-Taxi Dataset

To validate the effectiveness and robustness of the proposed AT-GCN-GRU model, this study utilizes the Shenzhen taxi dataset (SZ-taxi), a widely recognized public benchmark in traffic forecasting. Originally collected by the Shenzhen traffic management department and systematically open-sourced by Zhao et al. [9], this dataset holds significant academic value. Spatially, it focuses on the core transportation network of Luohu District, Shenzhen, selecting 156 major roads that effectively reflect traffic flow characteristics in complex urban environments. Temporally, the data spans from 1 January to 31 January 2015. The sampling frequency is set to 15 min, dividing a single day into 96 continuous time slots and yielding a total of 2976 time steps for the entire dataset.

Structurally, the dataset consists of two core components. The first is the speed matrix (sz_speed.csv) with dimensions of 2976×156 , which records the average travel speed of the 156 roads at each time step, serving as the primary input for capturing temporal evolution. The second is the adjacency matrix (sz_adj.csv) with dimensions of 156×156 . Constructed based on the real physical topology, this matrix accurately describes spatial connectivity using binary or weighted values, providing the crucial mathematical basis for the GCN to extract spatial dependencies. To intuitively illustrate these spatiotemporal characteristics, Figure 4 presents a traffic speed heatmap of selected typical road segments. The color variations clearly depict the formation, propagation, and dissipation of traffic congestion, revealing the intricate dynamic evolution patterns of urban traffic flow.

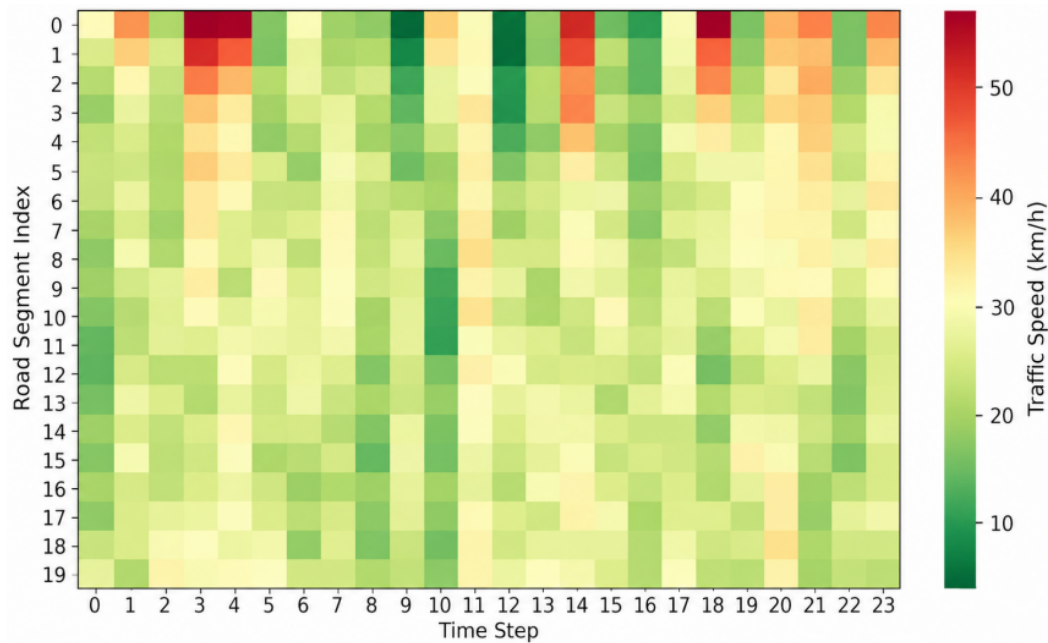


Figure 4. Traffic speed heat map for selected road segments.

4.1.2. METR-LA Dataset

To further evaluate the cross-dataset applicability of the proposed framework, an additional experiment was conducted on the public METR-LA dataset. METR-LA contains traffic-speed measurements collected from 207 loop detectors on highways in Los Angeles County. The data span from 1 March to 30 June 2012, with a sampling interval of 5 min, resulting in 34,272 time steps. Compared with SZ-taxi, METR-LA contains more spatial nodes and has a finer temporal resolution. The corresponding sensor graph encodes the spatial relationships among the 207 monitoring locations. In this study, METR-LA is used as an additional external validation dataset to examine whether the proposed architecture remains applicable under a different road-network scale and temporal resolution.

4.2. Data Preprocessing

To eliminate differences in data dimensions and accelerate the convergence of the neural network, this paper implements a rigorous data preprocessing workflow. First, Min-Max normalization is used to linearly map the raw speed data to the interval $[0, 1]$ in order to standardize the data distribution. Subsequently, supervised learning samples were constructed using a sliding window technique, with the input window length set to $T = 12$; that is, historical data from the past three hours was used to predict the traffic conditions for the τ time step. For METR-LA, the same 12-step input sequence corresponds to 1 h of historical observations rather than 3 h because the sampling interval is 5 min. Because the temporal resolutions of the two datasets differ, the prediction horizons were converted into dataset-specific time steps while preserving the same physical forecasting intervals. For SZ-taxi, the 15, 30, 45, and 60 min horizons correspond to one, two, three, and four steps, respectively, whereas for METR-LA they correspond to three, six, nine, and twelve steps, respectively.

Finally, regarding dataset partitioning, strict chronological order was followed: the first 70% of the data was used for training, the subsequent 10% for validation, and the remaining 20% for testing. The validation set was used for hyperparameter selection and early stopping, whereas the test set was reserved exclusively for final performance evaluation. To avoid information leakage, the normalization or standardization parameters

were calculated using only the training set and were subsequently applied to the validation and test sets.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (10)$$

4.3. Experimental Setup

4.3.1. Experimental Environment and Training Configuration

All experiments were conducted on workstations equipped with high-performance GPUs; the specific hardware and software configurations are shown in Table 1.

Table 1. Hardware environment configuration.

Hardware	Version
CPU	Intel(R) Core (TM) i9-14900K
GPU	NVIDIA RTX A4000
Memory	64 G
System	Win11
Runtime Environment	Python 3.10
PyTorch	2.11.0 + cu128
CUDA	12.8
cuDNN	9.1.9
NumPy	1.26.4
Pandas	2.1.4
scikit-learn	1.7.0

The hyperparameters were empirically tuned based on the validation loss. To ensure consistent comparison conditions across models, the hidden layer dimension and batch size were uniformly set to 64 for LSTM, GRU, GCN, T-GCN, and AT-GCN-GRU, while ASTGCN and AGCRN used architecture-specific configurations under the same data split and training protocol. The Adam optimizer was used for parameter updates, with an initial learning rate of 0.001, and the Mean Squared Error (MSE) was uniformly adopted as the objective loss function. The maximum number of training iterations was set to 100, combined with an early stopping mechanism: if the validation set loss did not decrease for 10 consecutive iterations, training was automatically terminated. In this manner, the optimal model weights were saved to effectively prevent overfitting.

Regarding the network architecture, the model input consists of one-dimensional traffic speed features from 156 nodes in the SZ-taxi dataset. The spatial and temporal modules utilize a two-layer graph convolutional network (GCN) and a single-layer gated recurrent unit (GRU), respectively, expanding the receptive field while strictly limiting the parameter size. To maintain a relatively compact model structure, the temporal attention module adopts an additive attention design, and the entire network eliminates feedforward neural networks, decoder structures, and dropout layers, achieving highly streamlined spatiotemporal feature aggregation.

Regarding data processing and experimentation, the dataset was strictly divided into training, validation, and testing sets in a 7:1:2 ratio according to time-series order, with the sequence remaining unaltered throughout to prevent the inadvertent disclosure of future information. For the main comparative experiments, the stochastic deep learning models were independently evaluated using multiple random seeds, and the mean and standard deviation of RMSE and MAE were reported in Table 2. Within each run, the selected random seed was consistently applied to Torch and NumPy, and no additional manual regularization strategies, such as weight decay or gradient clipping, were applied. These settings were consistently applied to all compared models to reduce stochastic variation and improve the reproducibility and fairness of the comparative experiments.

Table 2. Comparative test results.

Model	15 min (RMSE/MAE)	30 min (RMSE/MAE)	45 min (RMSE/MAE)	60 min (RMSE/MAE)
HA	4.60/3.35	4.70/3.45	4.85/3.60	5.10/3.85
ARIMA	4.45/3.20	4.58/3.32	4.75/3.50	4.95/3.72
SVR	4.38/3.15	4.50/3.28	4.68/3.45	4.85/3.65
LSTM	4.27 ± 0.03/3.05 ± 0.02	4.37 ± 0.04/3.17 ± 0.03	4.55 ± 0.04/3.36 ± 0.03	4.75 ± 0.05/3.53 ± 0.04
GRU	4.24 ± 0.03/3.03 ± 0.02	4.36 ± 0.03/3.16 ± 0.03	4.55 ± 0.04/3.31 ± 0.03	4.73 ± 0.03/3.51 ± 0.04
GCN	4.32 ± 0.04/3.14 ± 0.03	4.42 ± 0.03/3.21 ± 0.03	4.55 ± 0.05/3.36 ± 0.03	4.78 ± 0.04/3.53 ± 0.03
T-GCN	4.16 ± 0.05/2.95 ± 0.03	4.22 ± 0.03/3.05 ± 0.04	4.34 ± 0.05/3.18 ± 0.06	4.50 ± 0.03/3.32 ± 0.06
ASTGCN	4.08 ± 0.02/2.89 ± 0.03	4.19 ± 0.02/2.97 ± 0.03	4.28 ± 0.03/3.09 ± 0.04	4.39 ± 0.03/3.22 ± 0.05
AGCRN	4.05 ± 0.01/2.85 ± 0.03	4.17 ± 0.02/ 2.96 ± 0.02	4.25 ± 0.02/3.10 ± 0.02	4.40 ± 0.02/3.25 ± 0.02
AT-GCN-GRU	4.07 ± 0.04/2.88 ± 0.03	4.16 ± 0.05/2.96 ± 0.03	4.24 ± 0.03/3.07 ± 0.05	4.37 ± 0.04/3.21 ± 0.04

Note: Bold numerical values indicate the lowest RMSE and/or MAE for each prediction horizon; the bold model name indicates the proposed model.

4.3.2. Evaluation Criteria

We use two common regression evaluation metrics to quantify the model's prediction error: root mean square error (RMSE) and mean absolute error (MAE).

$$RMSE = \sqrt{\frac{1}{M} \sum_{i=1}^M (y_i - \hat{y}_i)^2} \quad (11)$$

$$MAE = \frac{1}{M} \sum_{i=1}^M |y_i - \hat{y}_i| \quad (12)$$

where y_i is the ground truth (or true value), $\hat{y}_i$ is the predicted value, and M is the total number of samples.

4.3.3. Baseline Models for Comparison

To evaluate the performance of AT-GCN-GRU, nine representative baseline models were selected for comparison: HA (Historical Averaging), ARIMA (Autoregressive Integrated Moving Average), SVR (Support Vector Regression), LSTM (Long Short-Term Memory), GRU (Gated Recurrent Unit), GCN (Graph Convolutional Network), T-GCN (Temporal Graph Convolutional Network), ASTGCN (Attention Based Spatial-Temporal Graph Convolutional Network), and AGCRN (Adaptive Graph Convolutional Recurrent Network). T-GCN is included as a directly related spatiotemporal baseline because it combines graph convolution with gated recurrent units, providing a suitable reference for evaluating the effect of introducing temporal attention into the GCN-GRU framework.

The selected baselines cover statistical methods, conventional machine learning methods, recurrent neural networks, graph convolutional models, and the directly related T-GCN architecture, as well as attention-based and adaptive graph learning models represented by ASTGCN and AGCRN. ASTGCN and AGCRN were additionally evaluated under the same chronological data split, preprocessing procedure, and evaluation metrics, while architecture-specific model configurations were retained for each method. Accordingly, Table 2 should be interpreted as a controlled comparison among the baseline models evaluated in this study rather than as a comprehensive benchmark against all existing traffic forecasting methods.

4.4. Experimental Results and Analysis

4.4.1. Overall Performance Comparison

Table 2 shows the error performance of different models for forecast horizons of 15, 30, 45, and 60 min. Figure 5a–d compare the RMSE results at the four prediction horizons, respectively, whereas Figure 6a–d present the corresponding MAE results.

As shown in Table 2, the proposed AT-GCN-GRU model demonstrates competitive performance against both conventional baselines and more recent spatiotemporal graph models. At the 15 min forecasting horizon, AGCRN achieves the lowest mean RMSE and MAE of 4.05 and 2.85, respectively, while AT-GCN-GRU ranks second with an RMSE of 4.07 and an MAE of 2.88. For the 30, 45, and 60 min forecasting tasks, AT-GCN-GRU achieves the lowest mean RMSE values of 4.16, 4.24, and 4.37, respectively. In terms of MAE, AT-GCN-GRU achieves 2.96, 3.07, and 3.21 for the corresponding horizons, tying with AGCRN at 30 min and obtaining the lowest values at 45 and 60 min. Compared with T-GCN, AT-GCN-GRU reduces the mean RMSE from 4.16 to 4.07, from 4.22 to 4.16, from 4.34 to 4.24, and from 4.50 to 4.37 at the 15, 30, 45, and 60 min horizons, corresponding to relative improvements of approximately 2.2%, 1.4%, 2.3%, and 2.9%, respectively. These results indicate that the temporal attention mechanism provides a consistent incremental improvement over T-GCN, while the proposed model remains competitive with stronger modern baselines such as ASTGCN and AGCRN, particularly at longer forecasting horizons.

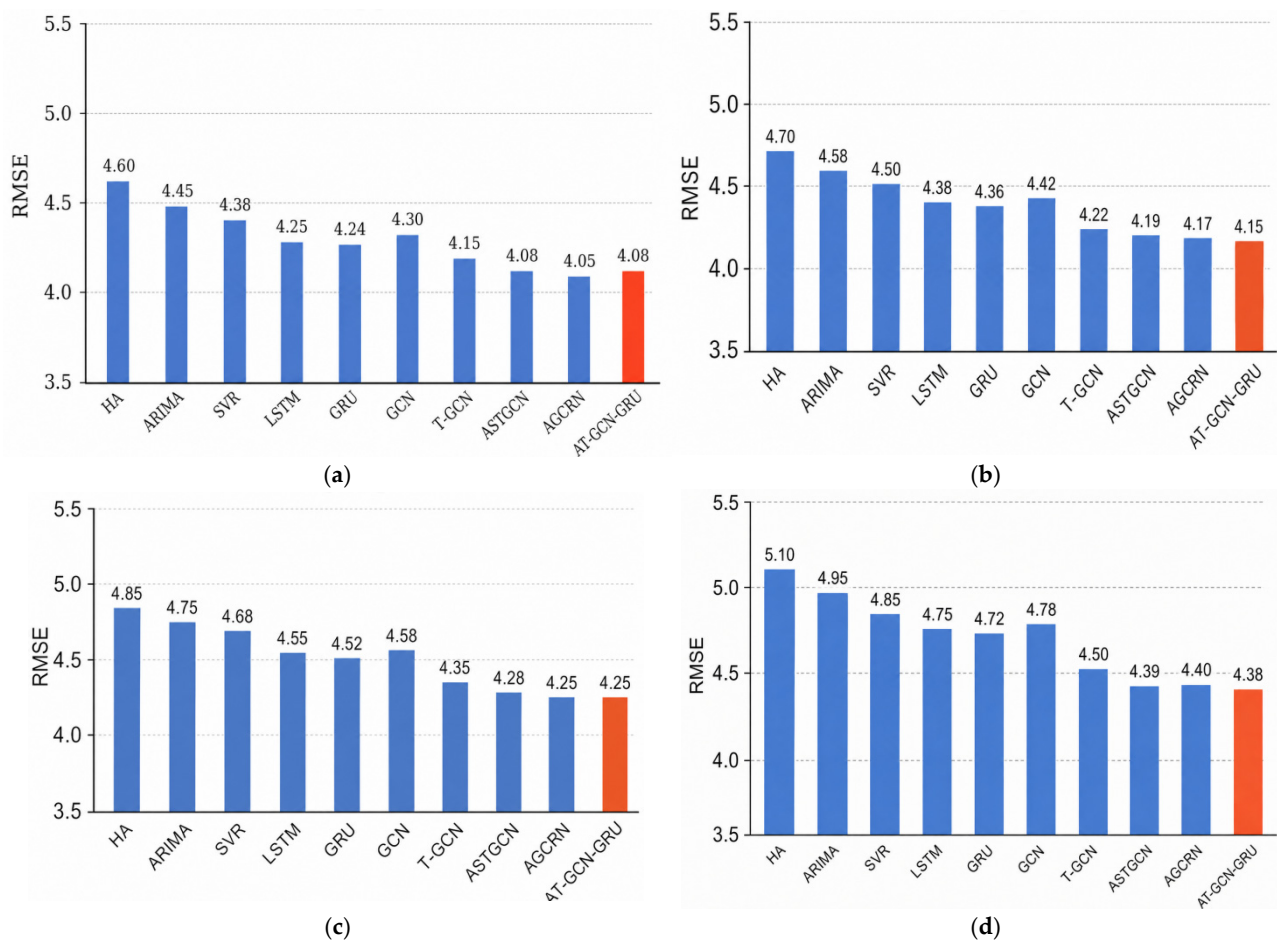


Figure 5. Comparison of RMSE across different models. (a) 15 min prediction horizon. (b) 30 min prediction horizon. (c) 45 min prediction horizon. (d) 60 min prediction horizon.

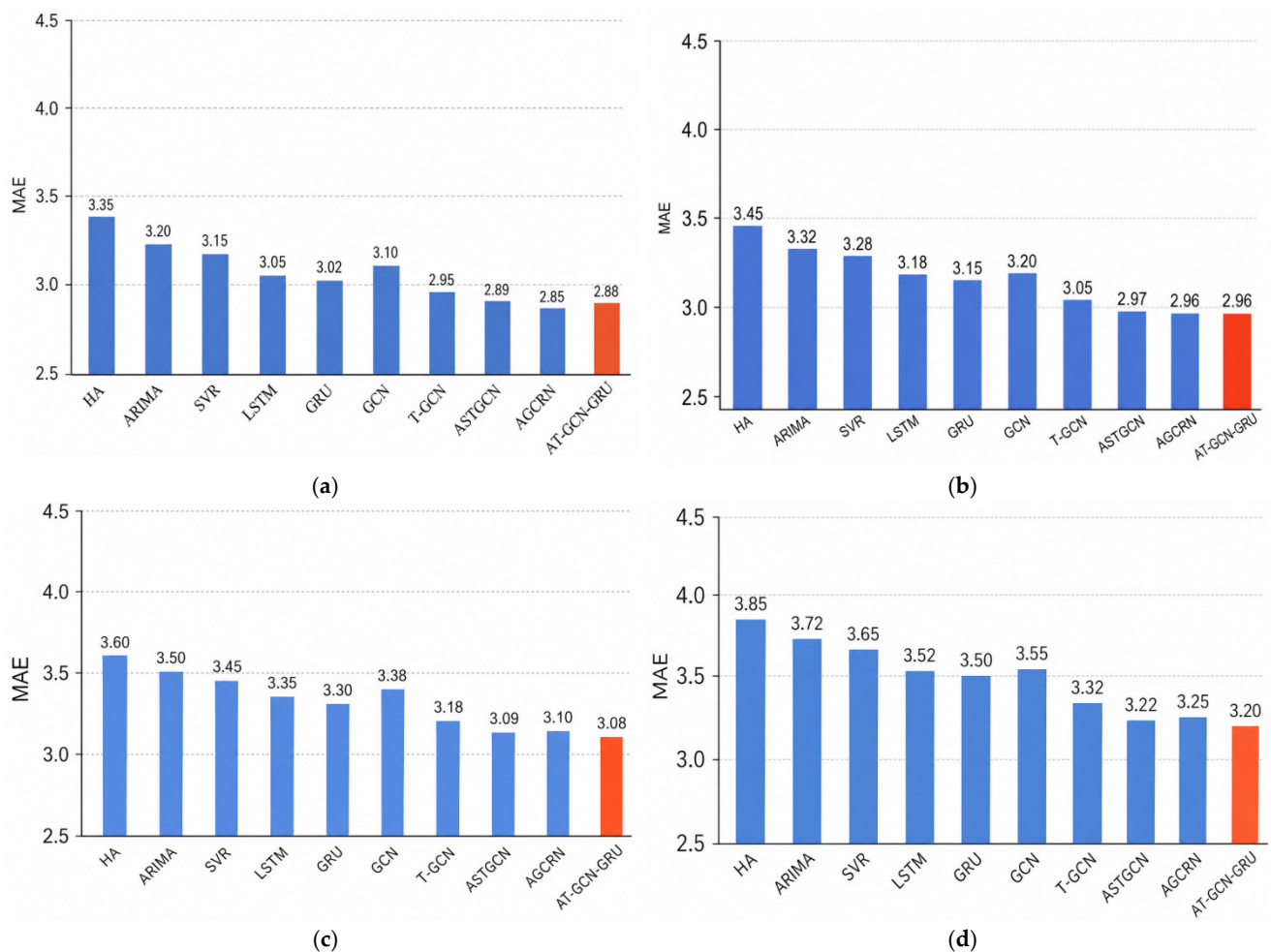


Figure 6. Comparison of MAE across different models. (a) 15 min prediction horizon. (b) 30 min prediction horizon. (c) 45 min prediction horizon. (d) 60 min prediction horizon.

Figures 5a–d and 6a–d provide visual comparisons of the RMSE and MAE results from a representative experimental run. As the prediction horizon increases, the errors of most models generally increase, while AT-GCN-GRU shows relatively low RMSE and MAE values across the four horizons in this representative run. For example, at the 60 min horizon, AT-GCN-GRU achieves an RMSE of 4.38 and an MAE of 3.20, compared with 5.10 and 3.85 for HA, respectively. The minor numerical differences between Table 2 and Figures 5 and 6 arise because Table 2 reports the mean and standard deviation over repeated runs with different random seeds, whereas Figures 5 and 6 visualize the results of one representative run.

The four subfigures in Figure 7 show comparisons between predicted and actual values for forecast horizons of 15, 30, 45, and 60 min. Figure 7a–d correspond to the 15, 30, 45, and 60 min forecast horizons, respectively.

In the 15 min short-term forecast shown in Figure 7a, the red predicted curve and the black actual curve almost completely overlap, indicating that the model can accurately capture instantaneous fluctuations in traffic flow; as the prediction horizon extends to 30 and 45 min, as shown in Figure 7b,c, despite increased uncertainty, the model continues to closely track the periodic changes in traffic flow without showing significant lag; notably, in the 60 min long-term forecast shown in Figure 7d, unlike the “flattening” or “severe lag” commonly observed in traditional models, this model maintains a high degree of fit and accurately reflects the overall trends of morning and evening rush hours. This qualitative observation suggests that the model can retain useful temporal information

at longer prediction horizons. However, the visual agreement alone does not establish improved generalization and should therefore be interpreted together with the quantitative RMSE and MAE results.

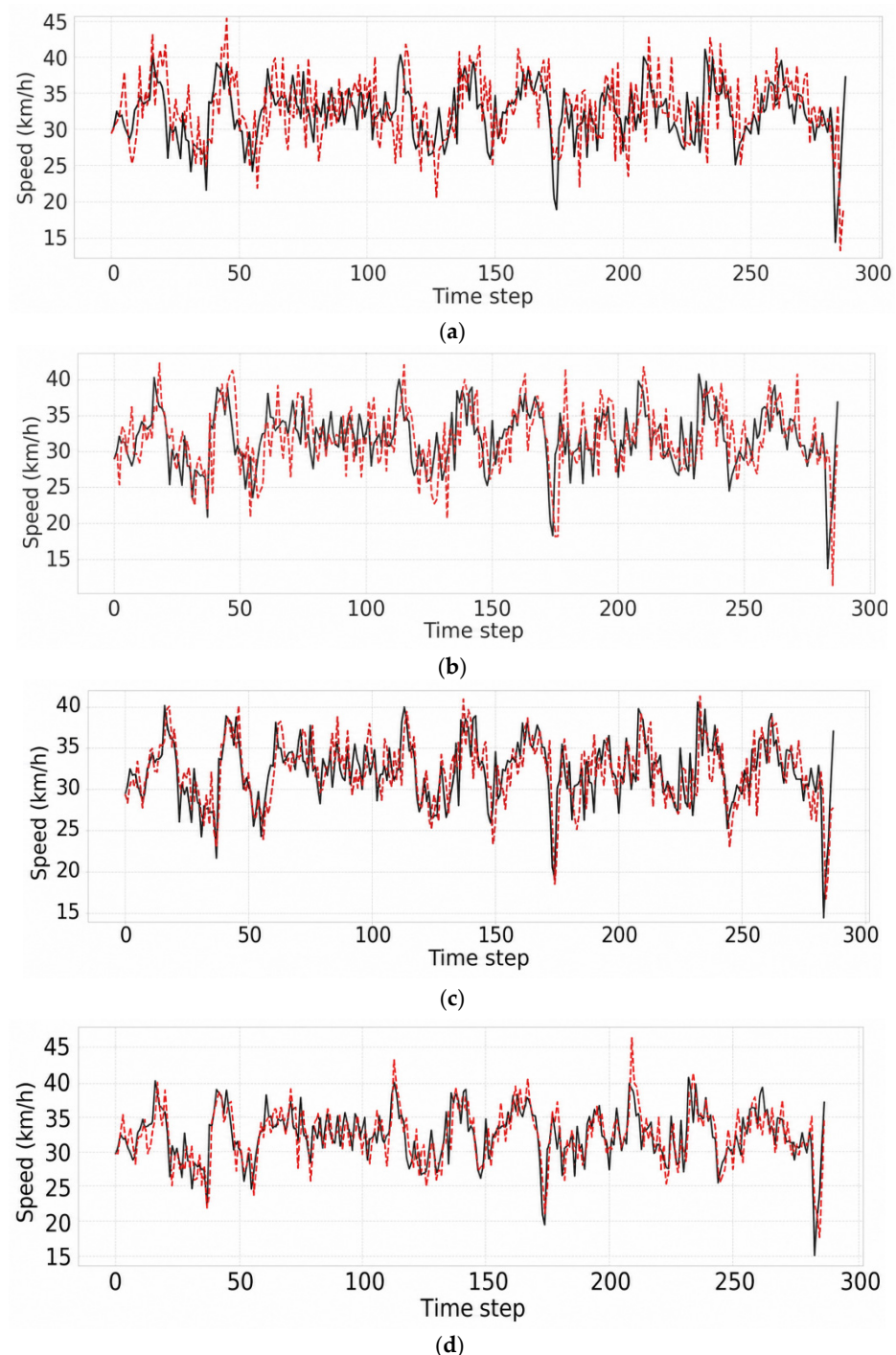


Figure 7. Comparison between predicted and actual traffic speeds at different prediction horizons. (a) 15 min prediction horizon. (b) 30 min prediction horizon. (c) 45 min prediction horizon. (d) 60 min prediction horizon.

4.4.2. Computational Cost

The engineering feasibility of traffic forecasting models in real-world production environments depends largely on their computational overhead and response speed. Table 3 details a comparison of the number of parameters, training time, and inference latency for various deep learning models under a 15 min forecasting horizon.

Table 3. Comparison of computational costs among different deep learning models.

Model	Parameters	Duration of a Single Training Session (s/Epoch)	Single-Sample Inference Latency (ms)
LSTM	66,972	0.0971	0.0047
GRU	52,764	0.1302	0.0137
GCN	44,460	0.0409	0.0090
T-GCN	69,420	0.1478	0.0183
AT-GCN-GRU	73,580	0.1608	0.0214

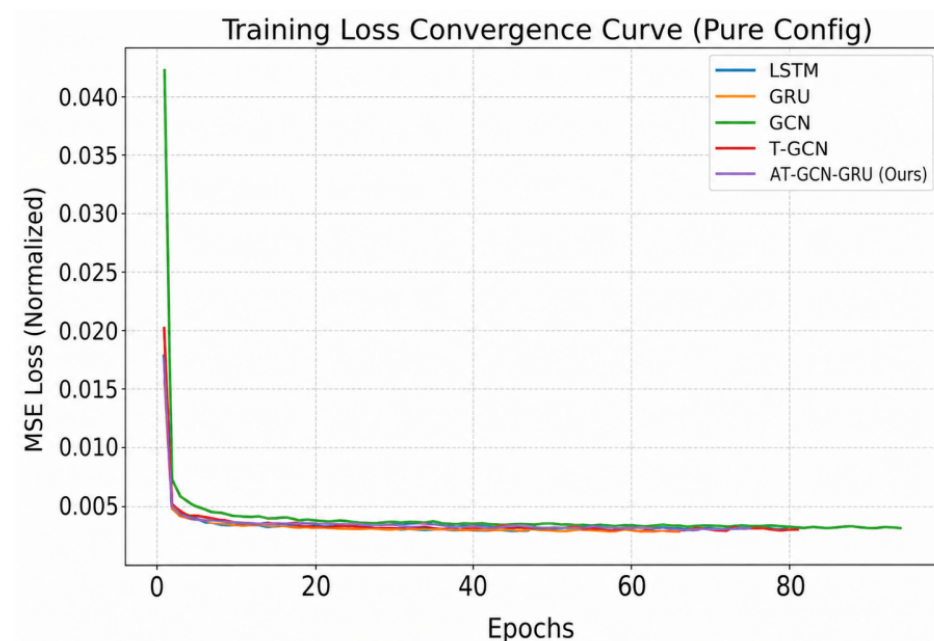
Note: Bold font highlights the proposed AT-GCN-GRU model and its corresponding computational cost results.

Inference latency was measured with a batch size of one on the same GPU platform for all deep learning models. Before timing, several warm-up forward passes were performed to reduce initialization overhead. CUDA synchronization was applied immediately before and after each timed forward pass to ensure that asynchronous GPU operations were fully completed. The reported latency represents the average wall-clock time over repeated forward passes, excluding data loading and preprocessing.

As shown in Table 3, the AT-GCN-GRU model contains 73,580 trainable parameters, representing an increase of approximately 6% compared with T-GCN. Under the measurement protocol described above, its single-sample inference latency is 0.0214 ms. Compared with T-GCN, the introduction of temporal attention results in a moderate increase in both parameter count and inference latency. Nevertheless, the additional computational overhead remains relatively limited under the present experimental environment, indicating that the temporal attention module can be incorporated into the GCN-GRU framework without substantially increasing the computational cost.

4.4.3. Convergence Analysis

The learning efficiency of deep learning model parameters and the training stability of the network architecture can typically be intuitively assessed by examining the convergence behavior of the loss function during the training optimization process. Figure 8 shows a comparison of loss convergence for each model on a 70% training set.

**Figure 8.** Convergence curves of training losses for different deep learning models.

As shown in Figure 8:

The loss values of all baseline models exhibited a rapid decline during the initial iteration phase. The loss value of the AT-GCN-GRU model proposed in this paper decreased rapidly during the first 10 training cycles and subsequently entered a steady convergence phase. Its convergence path was relatively smooth, suggesting stable optimization behavior during the stochastic gradient descent process. Combined with an early stopping mechanism, the model reaches a relatively stable convergence state around the 75th epoch, which helps limit overfitting during training. Compared with the other benchmark models, the proposed method shows a relatively stable training process and reaches a low final training loss, indicating that the model can learn useful spatiotemporal representations under the present experimental settings.

4.4.4. Ablation Experiment

To examine the contributions of each component in the model, we designed the following variants for ablation experiments under the 15, 30, 45, and 60 min forecasting horizons: GCN-GRU: a baseline T-GCN model with the attention mechanism removed; AT-GRU: the GCN module removed, with fully connected layers replacing spatial aggregation; AT-GCN: the GRU module removed, with no sequence evolution structure; AT-GCN-GRU: the complete model.

As shown in Table 4:

Table 4. Ablation results across different prediction horizons.

Model	15 min (RMSE/MAE)	30 min (RMSE/MAE)	45 min (RMSE/MAE)	60 min (RMSE/MAE)
GCN-GRU	4.15/2.95	4.22/3.05	4.35/3.18	4.50/3.32
AT-GRU	4.35/3.10	4.38/3.13	4.40/3.17	4.43/3.22
AT-GCN	4.42/3.25	4.45/3.27	4.47/3.31	4.52/3.36
AT-GCN-GRU	4.08/2.88	4.15/2.96	4.25/3.08	4.31/3.20

Note: Bold numerical values indicate the best performance among the ablation variants for each prediction horizon; the bold model name indicates the complete proposed AT-GCN-GRU model.

The ablation results illustrate the contributions of the spatial, temporal, and attention components within the AT-GCN-GRU framework across different prediction horizons. When the GCN module is removed, the RMSE increases from 4.08 to 4.35, from 4.15 to 4.38, from 4.25 to 4.40, and from 4.31 to 4.43 for the 15, 30, 45, and 60 min prediction tasks, respectively, indicating that spatial information derived from the road-network topology contributes to prediction accuracy. Removing the GRU module leads to further performance degradation, with RMSE values increasing to 4.42, 4.45, 4.47, and 4.52 across the four prediction horizons, indicating that temporal sequence modeling contributes to the forecasting performance. Compared with the GCN-GRU variant without temporal attention, the complete AT-GCN-GRU model reduces the RMSE from 4.15 to 4.08, from 4.22 to 4.15, from 4.35 to 4.25, and from 4.50 to 4.31 at the 15, 30, 45, and 60 min horizons, corresponding to improvements of approximately 1.7%, 1.7%, 2.3%, and 4.2%, respectively. The increasing improvement at longer prediction horizons suggests that the temporal attention mechanism contributes more noticeably as the forecasting interval becomes longer, while the GCN and GRU modules continue to provide useful spatial and temporal representations.

4.4.5. Robustness Analysis

To evaluate the model's robustness in handling data loss issues—such as sensor failures or data transmission interruptions—in real-world intelligent transportation systems, this paper further conducted sensitivity experiments on data missingness. The experiments considered four missing-data patterns: Missing Completely at Random (MCAR), block

missing, sensor-wise missing, and temporally correlated missingness, representing randomly scattered missing observations, continuous temporal interruptions, sensor failures, and temporally persistent missing observations, respectively. Artificial missingness was introduced only into the test input sequences, while the ground-truth prediction targets remained unchanged. The artificially missing input values were imputed using the corresponding node-specific mean calculated exclusively from the training data. For each missing-data pattern, missing rates from 0% to 50% were evaluated, and each experiment was repeated five times using different random seeds, with the average RMSE reported as the final performance evaluation metric.

As shown in Figure 9:

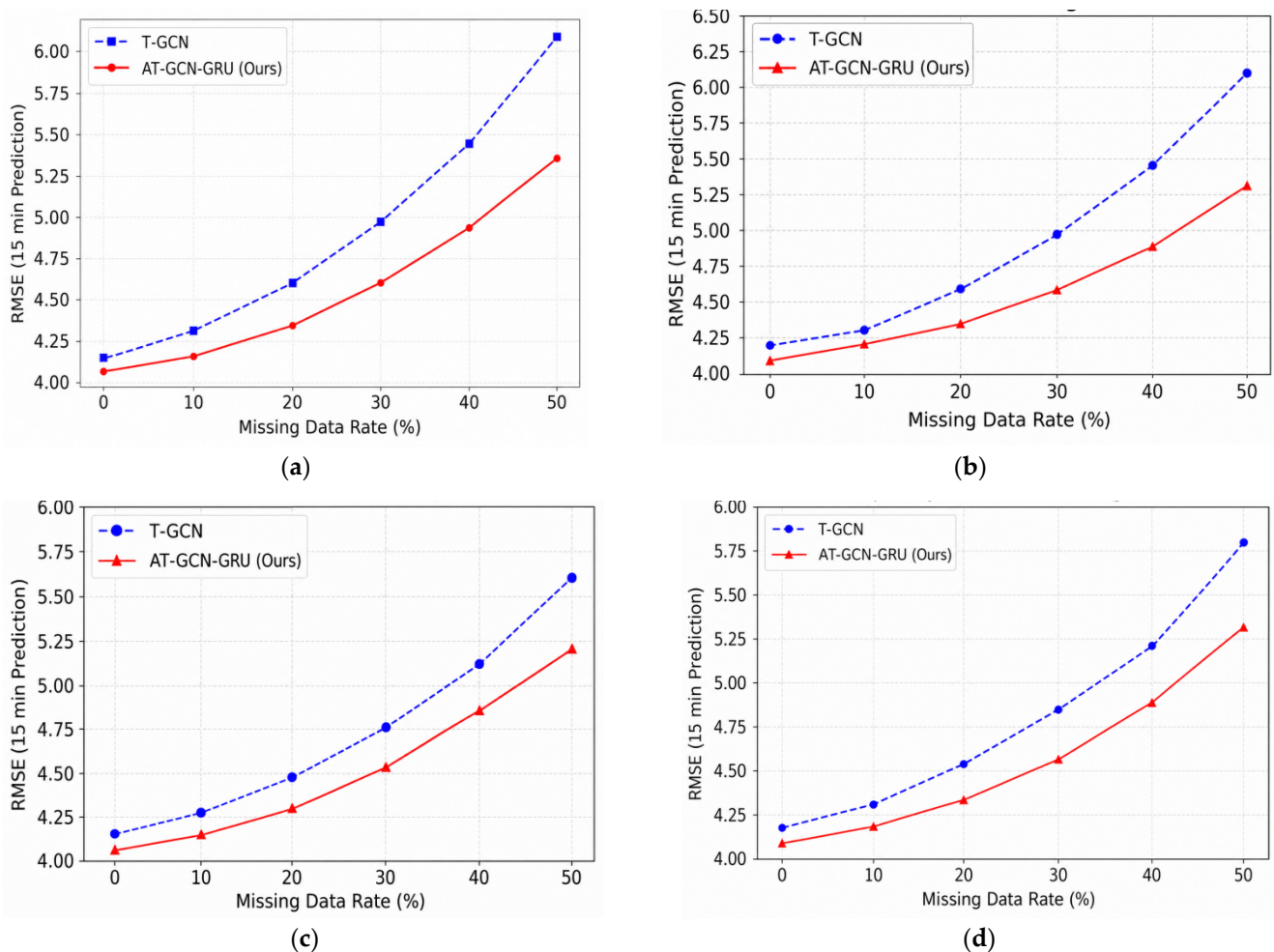


Figure 9. Robustness comparison of T-GCN and AT-GCN-GRU under different missing-data patterns for 15 min prediction. (a) MCAR; (b) block missing; (c) sensor-wise missing; (d) temporally correlated missingness.

The prediction errors of both T-GCN and AT-GCN-GRU increase as the missing-data rate rises from 0% to 50% under all four missing-data patterns. Figure 9a shows that under MCAR, AT-GCN-GRU generally exhibits lower RMSE values than T-GCN, with the difference becoming more noticeable at higher missing rates. A similar tendency can be observed under block missing in Figure 9b, where continuous loss of observations leads to a clear increase in prediction error, while AT-GCN-GRU exhibits a slower degradation trend than T-GCN. Under sensor-wise missingness, as shown in Figure 9c, both models remain relatively stable at lower missing rates, whereas the error increases more noticeably when a larger proportion of sensors becomes unavailable; nevertheless, AT-GCN-GRU generally

shows relatively low RMSE values compared with T-GCN. Figure 9d further shows that temporally correlated missingness is accompanied by progressive performance degradation, while AT-GCN-GRU shows a relatively flatter error-growth trend than T-GCN. Overall, the four missing-data scenarios indicate that prediction accuracy is increasingly affected as data completeness deteriorates, while AT-GCN-GRU shows a relatively moderate sensitivity to missing observations in the evaluated scenarios. These results suggest that the temporal attention mechanism may contribute to maintaining forecasting stability when traffic observations are incomplete.

4.4.6. Cross-Dataset Validation on METR-LA

To further examine the applicability of the proposed framework across different traffic networks, AT-GCN-GRU and the seven baseline models were additionally evaluated on the METR-LA dataset. For this experiment, ARIMA was implemented as ARIMA (12, 0, 0), while the SVR baseline employed a linear kernel using the previous 12 observations of each sensor as temporal input. The deep learning baselines were trained and evaluated using the same data split and preprocessing protocol as AT-GCN-GRU on METR-LA.

As shown in Table 5, AT-GCN-GRU remains competitive across the four forecasting horizons on the METR-LA dataset. At the 15 min horizon, it obtains the lowest mean RMSE of 6.11 and a mean MAE of 3.61, which is tied with SVR for the lowest MAE. At the 30 min horizon, AT-GCN-GRU achieves the lowest mean RMSE of 7.03, while T-GCN obtains a lower MAE of 4.20 compared with 4.29 for AT-GCN-GRU. At the 45 min horizon, GRU achieves the lowest mean RMSE of 7.44, whereas AT-GCN-GRU obtains a close value of 7.46 and the lowest mean MAE of 4.52. At the 60 min horizon, AT-GCN-GRU achieves the lowest mean RMSE and MAE, with values of 7.80 and 4.62, respectively. The relatively small standard deviations observed for the deep learning models also indicate limited variation across repeated runs. Overall, these results suggest that AT-GCN-GRU maintains competitive multi-horizon forecasting performance on METR-LA, although its relative advantage varies across prediction horizons and evaluation metrics.

Table 5. Summary of the RMSE and MAE results for the 15, 30, 45, and 60 min prediction horizons.

Model	15 min (RMSE/MAE)	30 min (RMSE/MAE)	45 min (RMSE/MAE)	60 min (RMSE/MAE)
HA	9.00/5.15	8.99/5.15	9.00/5.15	9.00/5.15
ARIMA	7.13/3.86	8.57/4.75	9.47/5.38	10.11/5.83
SVR	7.43/3.61	9.07/4.43	10.02/5.04	10.61/5.49
LSTM	6.26 ± 0.03/4.02 ± 0.04	7.06 ± 0.02/4.42 ± 0.04	7.58 ± 0.02/4.62 ± 0.03	8.06 ± 0.02/4.95 ± 0.03
GRU	6.14 ± 0.02/3.95 ± 0.04	7.08 ± 0.03/4.53 ± 0.02	7.44 ± 0.04/4.62 ± 0.03	7.89 ± 0.04/4.90 ± 0.02
GCN	6.25 ± 0.04/3.98 ± 0.02	7.09 ± 0.02/4.35 ± 0.05	7.56 ± 0.02/4.58 ± 0.04	8.11 ± 0.03/4.93 ± 0.02
T-GCN	6.21 ± 0.04/3.89 ± 0.02	7.09 ± 0.02/ 4.20 ± 0.02	7.60 ± 0.02/4.54 ± 0.03	8.02 ± 0.03/4.63 ± 0.05
AT-GCN-GRU	6.11 ± 0.02/3.61 ± 0.02	7.03 ± 0.02/4.29 ± 0.04	7.46 ± 0.03/ 4.52 ± 0.02	7.80 ± 0.05/4.62 ± 0.02

Note: Bold numerical values indicate the lowest mean RMSE and/or MAE for each prediction horizon; the bold model name indicates the proposed model.

4.4.7. Visualization of Temporal Attention Weights

To further investigate the temporal weighting behavior of the attention module, the attention weights were visualized under three representative traffic conditions: normal period, rush hour, and sudden congestion. For each condition, one representative sample was selected from the SZ-taxi dataset, and the attention distribution over the 12 historical input time steps was extracted from the trained AT-GCN-GRU model. The corresponding visualization results are shown in Figure 10.

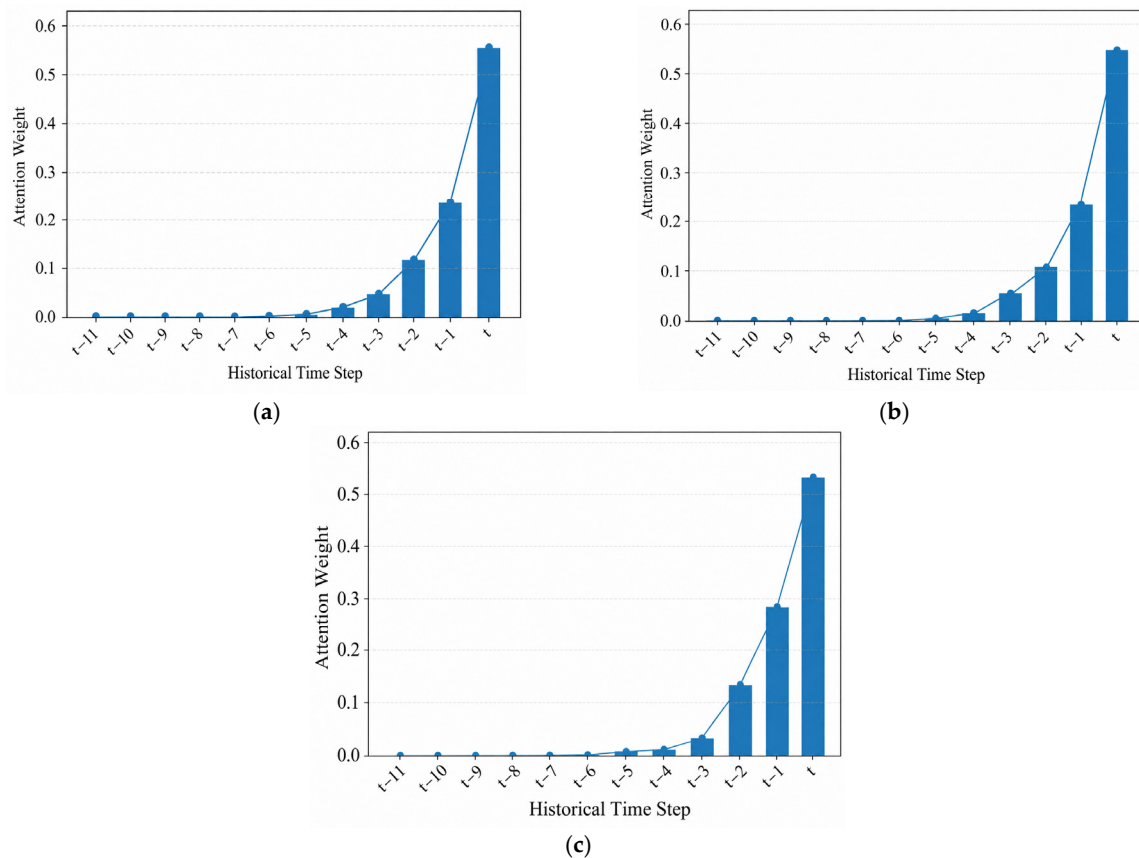


Figure 10. Visualization of temporal attention weights under three representative traffic conditions. (a) Normal period; (b) rush hour; (c) sudden congestion.

As shown in Figure 10, the temporal attention weights under all three traffic conditions are mainly concentrated on the most recent historical observations. In the normal period, the largest weight is assigned to the current time step t (0.5562), followed by $t-1$ (0.2361) and $t-2$ (0.1244), while the earlier time steps contribute only marginally. A similar pattern is observed during rush hour, where the weights of t , $t-1$, and $t-2$ are 0.5499, 0.2509, and 0.1061, respectively, indicating that the model still relies primarily on recent traffic states when the traffic flow becomes dense but remains relatively regular. Under sudden congestion, the attention distribution remains dominated by the most recent time steps, but relatively larger weights are assigned to $t-1$ (0.2852) and $t-2$ (0.1331), compared with the other two conditions. This suggests that when abrupt traffic deterioration occurs, the model tends to place more emphasis on the recent variation trend rather than relying only on the latest observation. Overall, these results suggest that the temporal attention mechanism tends to assign greater weights to recent traffic states and exhibits different temporal weighting patterns across the three evaluated traffic conditions.

4.5. Discussion

The experimental results provide further insight into the performance characteristics and applicability of the proposed AT-GCN-GRU framework beyond a direct numerical comparison. On the SZ-taxi dataset, the integration of spatial graph convolution, temporal sequence modeling, and temporal attention enables the model to capture spatiotemporal dependencies of urban traffic states over different forecasting horizons. In addition, the experiments on METR-LA suggest that the proposed framework remains competitive when evaluated on a traffic network with a different spatial scale and temporal resolution. As shown in Table 5, AT-GCN-GRU achieves the lowest mean RMSE at the 15, 30, and 60 min

prediction horizons, with values of 6.11, 7.03, and 7.80, respectively. At the 45 min horizon, its mean RMSE of 7.46 is close to the lowest value of 7.44 obtained by GRU. In terms of MAE, AT-GCN-GRU ties with SVR for the lowest mean value at the 15 min horizon and obtains the lowest mean MAE at the 45 and 60 min horizons, while remaining competitive at the 30 min horizon. These results suggest that the proposed model can retain competitive forecasting performance beyond the SZ-taxi dataset under the evaluated experimental settings. These cross-dataset observations also suggest that combining spatial dependency modeling with temporal attention provides a useful representation for multi-horizon traffic forecasting.

Despite these encouraging results, several limitations remain. First, the current model constructs spatial dependencies using a predefined and fixed adjacency matrix. However, actual traffic interactions between road segments may change dynamically under congestion propagation, traffic accidents, temporary road closures, or other abnormal conditions, which may not be fully represented by a static graph structure. Second, the present framework mainly relies on historical traffic-speed observations and road-network topology, whereas external factors such as weather conditions, holidays, traffic incidents, and large-scale events may also affect short-term traffic dynamics. The absence of these heterogeneous factors may limit the model's ability to explain and predict abrupt changes that cannot be inferred solely from historical traffic patterns. In addition, although the GCN-GRU-attention architecture provides a relatively simple and computationally manageable framework, increasingly complex traffic networks may require more adaptive spatial and temporal representations. Future research could therefore incorporate dynamic graph learning, multi-source heterogeneous information, and adaptive attention mechanisms to explore their potential for improving prediction accuracy and generalization under complex real-world traffic conditions.

5. Conclusions

To address the challenges of spatial modeling in road networks and modeling longer-term temporal dependencies in intelligent transportation systems, this paper proposes a spatiotemporal graph convolutional prediction model (AT-GCN-GRU) that integrates an attention mechanism. The model constructs an end-to-end framework: it utilizes GCN to extract spatial dependency features from the road network topology, employs GRU to capture temporal evolution patterns, and introduces a temporal attention mechanism to adaptively weight historical states, thereby providing additional temporal weighting for multi-horizon prediction. Experiments on the SZ-taxi dataset show that AT-GCN-GRU remains competitive with both conventional and modern baseline models, including T-GCN, ASTGCN, and AGCRN, across the 15, 30, 45, and 60 min forecasting horizons. Compared with T-GCN, AT-GCN-GRU obtains lower mean RMSE values across the four evaluated horizons, with relative improvements ranging from approximately 1.4% to 2.9%. The additional evaluation on METR-LA further shows that AT-GCN-GRU obtains the lowest mean RMSE among the evaluated models at the 15, 30, and 60 min horizons and remains close to the lowest RMSE at the 45 min horizon.

This study provides a high-precision solution for urban traffic forecasting and holds significant engineering value. To further expand the theoretical depth and application scope of this research, future work will focus on the following three directions: First, evolving from static topologies to adaptive dynamic graphs by introducing graph reconstruction mechanisms to capture real-time spatiotemporal correlations caused by sudden traffic conditions or tidal congestion. Second, constructing a collaborative fusion framework for multi-source heterogeneous data, deeply integrating multimodal external features—such as weather, holidays, and emergencies—to enhance the model's robustness against interference under extreme scenarios. And third, exploring cross-regional transfer learning based on spatiotemporal foundation

models, utilizing few-shot transfer mechanisms to investigate the applicability of the proposed framework to regions with limited monitoring data.

Author Contributions: Conceptualization, Z.X. and S.W.; methodology, Z.X. and S.W.; software, Z.X.; validation, Z.X., H.X. and Z.H.; formal analysis, Z.X. and S.W.; investigation, Z.X.; resources, S.W. and H.X.; data curation, Z.X.; writing—original draft preparation, Z.X.; writing—review and editing, S.W., H.X. and Z.H.; visualization, Z.X.; supervision, S.W.; project administration, S.W.; funding acquisition, S.W. and H.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Open Fund of the National Engineering Research Center of Highway Maintenance Technology [grant number kfj220108]; the Central Guiding Local Science and Technology Development Fund Project [grant number 24ZYQA055]; the Gansu Provincial United Scientific Research Fund General Project [grant number 23JRRA1486]; the Youth Doctoral Support Project for Universities in Gansu Province [grant number 2024QB-039]; the Gansu Provincial Natural Science Foundation [grant number 26JRRA590]; the Innovation Foundation of the Gansu Provincial Department of Education [grant number 2025B-058]; and the Joint Foundation of Southwest Jiaotong University and Lanzhou Jiaotong University [grant number 1200150725].

Data Availability Statement: The data used to support the findings of this study are available from the public SZ-taxi benchmark dataset released with the T-GCN study and the public METR-LA dataset used for the additional cross-dataset evaluation. Further details are provided in the data description section of this manuscript.

Conflicts of Interest: Author Zhongdong Hua was employed by Gansu Aomake Education Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Wang, J.; Dai, H.; Feng, H.; Guo, M.; Zyljanov, V.; Feng, Z.; Cui, J. Carbon emission of urban vehicles based on carbon emission factor correlation analysis. *Sci. Rep.* **2025**, *15*, 2037. [[CrossRef](#)]
2. Li, D.; Yao, Y.; Shao, Z. Big data in smart city. *Geomat. Inf. Sci. Wuhan Univ.* **2014**, *39*, 631–640.
3. Wang, J.; Zhang, Y.; Hu, Y.; Yin, B. Survey on Graph Convolutional Neural Network-Based Traffic Prediction. *J. Beijing Univ. Technol.* **2021**, *47*, 954–970.
4. Dai, B.-A.; Ye, B.L.; Li, L. A novel hybrid time-varying graph neural network for traffic flow forecasting. *arXiv* **2024**, arXiv:2401.10155.
5. Seng, D.; Lv, F.; Liang, Z.; Shi, X.; Fang, Q. Forecasting Traffic Flows in Irregular Regions with Multi-Graph Convolutional Network and Gated Recurrent Unit. *Front. Inf. Technol. Electron. Eng.* **2021**, *22*, 1179–1193. [[CrossRef](#)]
6. Jia, W.; Zhao, K.; Chen, M.; Han, Y.; Li, X.; Shao, X. ST2SNet: Spatial-temporal two-layer synchronous network for traffic flow prediction. *IEEE Trans. Comput. Soc. Syst.* **2025**, *12*, 3279–3288. [[CrossRef](#)]
7. Williams, B.M.; Hoel, L.A. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *J. Transp. Eng.* **2003**, *129*, 664–672. [[CrossRef](#)]
8. Cui, J.-X.; Yao, J.; Zhao, B.-Y. Review on Short-Term Traffic Flow Prediction Methods Based on Deep Learning. *J. Traffic Transp. Eng.* **2024**, *24*, 50–64. [[CrossRef](#)]
9. Zhao, L.; Song, Y.; Zhang, C.; Liu, Y.; Wang, P.; Lin, T.; Deng, M.; Li, H. T-GCN: A temporal graph convolutional network for traffic prediction. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 3848–3858. [[CrossRef](#)]
10. Jiang, W.; Luo, J. Graph neural network for traffic forecasting: A survey. *Expert Syst. Appl.* **2022**, *207*, 117921. [[CrossRef](#)]
11. Yin, L.; Sun, S.; Wei, S.; Tian, S.; He, Y. Traffic Flow Combination Prediction Model Based on Adaptive VMD-Attention-BiLSTM. *J. Electron. Meas. Instrum.* **2021**, *35*, 130–139.
12. Guo, S.; Lin, Y.; Feng, N.; Song, C.; Wan, H. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 922–929. [[CrossRef](#)]
13. Zheng, Y.; Luo, C.; Shao, R. Enhancing traffic flow forecasting with delay propagation: Adaptive graph convolution networks for spatio-temporal data. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 650–660. [[CrossRef](#)]
14. Yuan, Q.; Wang, J.; Han, Y.; Liu, Z.; Liu, W. DAGCAN: Decoupled Adaptive Graph Convolution Attention Network for Traffic Forecasting. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 3513–3526. [[CrossRef](#)]
15. Lana, I.; Del Ser, J.; Velez, M.; Vlahogianni, E.I. Road Traffic Forecasting: Recent Advances and New Challenges. *IEEE Intell. Transp. Syst. Mag.* **2018**, *10*, 93–109. [[CrossRef](#)]

16. Kumar, S.V.; Vanajakshi, L. Short-term traffic flow prediction using seasonal ARIMA model with limited input data. *Eur. Transp. Res. Rev.* **2015**, *7*, 21. [\[CrossRef\]](#)
17. Ahmed, M.S.; Cook, A.R. *Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques*; Transportation Research Board: Washington, DC, USA, 1979; pp. 1–9.
18. Wu, C.-H.; Ho, J.-M.; Lee, D. Travel-time prediction with support vector regression. *IEEE Trans. Intell. Transp. Syst.* **2004**, *5*, 276–281. [\[CrossRef\]](#)
19. Habtemichael, F.G.; Cetin, M. Short-term traffic flow rate forecasting based on identifying similar traffic patterns. *Transp. Res. Part C Emerg. Technol.* **2016**, *66*, 61–78. [\[CrossRef\]](#)
20. Ma, X.; Tao, Z.; Wang, Y.; Yu, H.; Wang, Y. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transp. Res. Part C Emerg. Technol.* **2015**, *54*, 187–197. [\[CrossRef\]](#)
21. Fu, R.; Zhang, Z.; Li, L. Using LSTM and GRU neural network methods for traffic flow prediction. *Proc. AAAI Conf. Artif. Intell.* **2016**, *31*, 324–328. [\[CrossRef\]](#)
22. Zhang, J.; Zheng, Y.; Qi, D. Deep spatio-temporal residual networks for citywide crowd flows prediction. *Proc. AAAI Conf. Artif. Intell.* **2017**, *31*, 1655–1661. [\[CrossRef\]](#)
23. Yao, H.; Wu, F.; Ke, J.; Tang, X.; Jia, Y.; Lu, S.; Gong, P.; Ye, J.; Li, Z. Deep Multi-View Spatial-Temporal Network for Taxi Demand Prediction. *Proc. AAAI Conf. Artif. Intell.* **2018**, *32*, 2588–2595. [\[CrossRef\]](#)
24. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 February 2017.
25. Ye, J.; Zhao, J.; Ye, K.; Xu, C.-Z. How to Build a Graph-Based Deep Learning Architecture in Traffic Domain: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 3904–3924. [\[CrossRef\]](#)
26. Bai, J.; Zhu, J.; Song, Y.; Zhao, L.; Hou, Z.; Du, R.; Li, H. A3T-GCN: Attention temporal graph convolutional network for traffic forecasting. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 485. [\[CrossRef\]](#)
27. Sun, L.; Liu, M.; Liu, G.; Chen, X.; Yu, X. FD-TGCN: Fast and Dynamic Temporal Graph Convolution Network for Traffic Flow Prediction. *Inf. Fusion* **2024**, *106*, 102291. [\[CrossRef\]](#)
28. Li, Z.; Xiong, G.; Tian, Y.; Lv, Y.; Chen, Y.; Hui, P.; Su, X. A Multi-Stream Feature Fusion Approach for Traffic Prediction. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 1456–1466. [\[CrossRef\]](#)
29. Pan, Z.; Liang, Y.; Zhang, J.; Yi, X.; Yu, Y.; Zheng, Y. HyperST-Net: Hypernetworks for Spatio-Temporal Forecasting. *arXiv* **2018**, arXiv:1809.10889.
30. Li, M.; Zhu, Z. Spatial-temporal fusion graph neural networks for traffic flow forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 4189–4196. [\[CrossRef\]](#)
31. Zhang, D.; Wang, P.; Ding, L.; Wang, X.; He, J. Spatio-temporal contrastive learning based adaptive graph augmentation for traffic flow prediction. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 1304–1318. [\[CrossRef\]](#)
32. Li, X.; Zhang, Y.; Long, G.; Hu, Y.; Lu, W.; Chen, M.; Zhang, C.; Gong, Y. Adaptive traffic forecasting on daily basis: A spatio-temporal context learning approach. *IEEE Trans. Knowl. Data Eng.* **2025**, *37*, 4446–4459. [\[CrossRef\]](#)
33. Zheng, G.; Chai, W.K.; Zhang, J.; Katos, V. VDGCNeT: A Novel Network-Wide Virtual Dynamic Graph Convolution Neural Network and Transformer-Based Traffic Prediction Model. *Knowl.-Based Syst.* **2023**, *275*, 110676. [\[CrossRef\]](#)
34. Zhang, X.; Chen, X.; Tang, H.; Wu, Y.; Shen, H.; Li, J. AdpSTGCN: Adaptive spatial-temporal graph convolutional network for traffic forecasting. *Knowl.-Based Syst.* **2024**, *301*, 112295. [\[CrossRef\]](#)
35. Choi, J.; Choi, H.; Hwang, J.; Park, N. Graph Neural Controlled Differential Equations for Traffic Forecasting. *Proc. AAAI Conf. Artif. Intell.* **2022**, *36*, 6367–6374. [\[CrossRef\]](#)
36. Sun, S.; Deng, X.; Liu, S.; Fan, X.; Yi, L.; Zhu, C.; Chen, W.; Dai, H. Long-term traffic forecasting via spatial-temporal wavelet attention network for mobile IoT-enabled transportation systems. *IEEE Trans. Mob. Comput.* **2026**, *25*, 9300–9314. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.