Applied Artificial Intelligence Research

INNOVIAIR

A Hybrid Spatiotemporal Framework with Memory and Diffusion Convolution for Traffic Flow Prediction

Xi Chen[a,*], Jiajia Chen[a]

[a]School of Intelligent Manufacturing, Chongqing Jianzhu College, Chongqing 400072,China

ARTICLEINFO

Keywords:
Traffic Flow Prediction
Spatiotemporal Forecasting
Memory-Augmented Transformer
Diffusion Convolution

ABSTRACT

Accurate traffic flow prediction is pivotal for intelligent transportation systems, yet it remains inherently challenging due to dynamic spatial correlations and long-range temporal dependencies. While existing forecasting paradigms predominantly rely on static, pre-defined graph structures, they often overlook the direct functional connections between non-adjacent time steps.This paper proposes a novel spatiotemporal forecasting framework, termed Memory-augmented Diffusion Convolutional LSTM network (MDC-LSTM), which integrates a MemBART-inspired memory mechanism, diffusion convolution, regional attention, and LSTM-based temporal modeling. By synthesizing these components with Long Short-Term Memory (LSTM) units, the proposed model effectively captures both localized spatial patterns and deep inter-temporal relationships. Empirical evaluations conducted on the METR-LA benchmark dataset demonstrate that our framework significantly enhances predictive accuracy across multiple metrics, consistently outperforming several state-of-the-art (SOTA) baselines and exhibiting strong robustness in long-term forecasting scenarios.

1.Introduction

With the rapid development of urbanization, intelligent transportation systems (ITS), connected vehicles, and large-scale sensing infrastructures, traffic flow prediction has become a fundamental task in transportation engineering, urban computing, and artificial intelligence. Recent studies have shown that urban traffic flow is affected by multiple heterogeneous factors, including temporal cycles, meteorological conditions, road-network structure, and unexpected events, which makes accurate forecasting highly challenging[1]. Modern traffic networks continuously generate massive spatiotemporal data from loop detectors, GPS trajectories, surveillance cameras, roadside units, and mobile sensing devices. These data provide essential support for traffic state estimation, congestion warning, route guidance, traffic signal control, emergency response, and smart city decision-making. Accurate traffic forecasting can help transportation authorities identify potential congestion in advance, optimize road network management, reduce travel delay, and improve the operational efficiency of urban transportation systems[2,3].

Despite its practical importance, traffic flow prediction remains a challenging problem because traffic systems are highly nonlinear, dynamic, and stochastic. Traffic states are influenced by multiple factors, including road topology, upstream and downstream propagation, commuting patterns, signal timing, accidents, weather conditions, holidays, and human travel behavior. These factors jointly lead to complex spatial-temporal dependencies. From the spatial perspective, the traffic state of one road segment is not only affected by geographically adjacent sensors but may also be influenced by distant yet functionally related roads, such as parallel corridors, alternative routes, and key intersections. From the temporal perspective, traffic sequences contain both short-term fluctuations and long-term periodic patterns, including morning peaks, evening peaks, daily regularities, and weekly periodicity. Therefore, an effective forecasting model should simultaneously capture local spatial diffusion, global spatial correlation, short-term temporal variation, and long-range temporal dependency.

Early traffic forecasting studies mainly relied on statistical and traditional machine learning models, such as Autoregressive Integrated Moving Average (ARIMA), Vector Autoregression (VAR), Kalman filtering, Support Vector Regression (SVR), and shallow neural networks[2,4]. These methods are computationally efficient and interpretable, but their ability to model nonlinear spatiotemporal patterns is limited. With the development of deep learning, recurrent

* Corresponding author.
E-mail address: jlinghu@163.com.
https://doi.org/10.65455/6rkv7629
Received 28 May 2026; Received in revised form 2 July 2026; Accepted 8 July 2026; Available 20 July 2026

neural networks, Long Short-Term Memory (LSTM), Gated Recurrent Units (GRUs), temporal convolutional networks, and sequence-to-sequence architectures have been widely applied to traffic forecasting[3,5]. These models improve temporal representation learning, but they often treat traffic sensors as independent or weakly connected sequences, making it difficult to explicitly capture the spatial dependency embedded in road networks.

To address this limitation, graph neural networks (GNNs) have become a dominant paradigm for traffic forecasting. By representing traffic sensors as graph nodes and road connections as graph edges, GNN-based models can naturally describe spatial correlations in transportation networks. Representative models such as Diffusion Convolutional Recurrent Neural Network (DCRNN)[6], Spatio-Temporal Graph Convolutional Network (STGCN)[7], Graph WaveNet[8], Attention Based Spatial-Temporal Graph Convolutional Network (ASTGCN)[9], Multivariate Time Series Graph Neural Network (MTGNN)[10], Adaptive Graph Convolutional Recurrent Network (AGCRN)[11], and Decoupled Dynamic Spatial-Temporal Graph Neural Network (D2STGNN)[12] have achieved remarkable progress. These studies demonstrate that graph-based spatial modeling is effective for capturing road-network dependencies and improving prediction accuracy. Nevertheless, many existing methods still rely heavily on predefined adjacency matrices or relatively simple adaptive graph structures. Such designs may not be sufficient for modeling the time-varying and context-dependent nature of traffic interactions.

In real-world traffic networks, spatial dependencies are not fixed. The influence between two road segments may change dynamically according to congestion levels, traffic incidents, weather conditions, and temporal context. For example, two road segments may be weakly correlated during normal periods but strongly correlated during peak hours or under incident-induced congestion. Similarly, downstream traffic states may be affected by upstream bottlenecks after a propagation delay rather than immediately. Although adaptive graph learning methods such as Graph WaveNet, MTGNN, AGCRN, and D2STGNN attempt to infer hidden or dynamic spatial relationships from data[8,10-12], the learned dynamic adjacency matrices may also become unstable when sensor data are noisy, sparse, or incomplete. Therefore, how to enhance dynamic spatial representation while maintaining stable and reliable graph learning remains an important challenge.

Long-range temporal dependency is another key limitation of many existing traffic forecasting models. RNN- and LSTM-based models are effective for sequential modeling, but they may suffer from gradient decay and error accumulation when the forecasting horizon becomes longer[5]. Convolution-based models can improve training efficiency, but their temporal receptive fields are still constrained by kernel size and network depth. Transformer-based models have recently attracted increasing attention because self-attention can directly capture global dependencies among different time steps[13]. Representative long-sequence forecasting models, such as Informer[14], Autoformer[15], and FEDformer[16], have shown the potential of attention mechanisms in modeling long-term temporal patterns. In the traffic forecasting field, Transformer-based models such as GMAN[17], PDFormer[18], and STAEformer[19] further demonstrate the effectiveness of attention mechanisms in capturing complex spatial-temporal dependencies. However, most Transformer architectures operate within a fixed input window and lack an explicit memory mechanism to preserve historical context beyond the current sequence segment. This limitation may weaken their ability to maintain temporal continuity in long-term traffic forecasting.

Memory-augmented architectures provide a promising direction for addressing this problem. By introducing memory states, memory-enhanced models can store, retrieve, and reuse historical information across sequence segments. Such mechanisms have been explored in natural language processing and long-context sequence modeling, such as Transformer-XL[20], Memorizing Transformers[21], and BART-style encoder-decoder architectures[22]. However, their integration with graph-based traffic forecasting remains insufficiently explored. Traffic forecasting requires not only long-range temporal memory but also dynamic spatial dependency learning over road networks. Therefore, integrating memory-augmented sequence modeling with graph diffusion and local spatial attention may provide a more comprehensive solution for spatiotemporal traffic prediction.

2.Related work

Traffic flow prediction has been extensively studied in transportation engineering, data mining, and artificial intelligence. Existing studies can be broadly divided into five categories: statistical and traditional machine learning methods, deep temporal models, graph neural network-based methods, Transformer-based spatiotemporal forecasting models, and memory-augmented sequence models. This section reviews representative advances in these areas and discusses their relevance to the proposed framework.

2.1.Statistical and traditional machine learning methods for traffic forecasting

Early traffic forecasting studies mainly relied on statistical time-series models. Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA, Vector Autoregression (VAR), Kalman filtering, and Gaussian process regression were widely used to model temporal changes in traffic speed, flow, and occupancy[2,4]. These methods are mathematically interpretable and computationally efficient, making them suitable for early traffic management systems with limited data scale. For example, ARIMA-based models can capture linear temporal autocorrelation, while VAR models can describe interactions among multiple traffic variables[4].

Traditional machine learning methods were later introduced to improve nonlinear modeling ability. Support Vector Regression, Random Forests, k-Nearest Neighbors, Gradient Boosting Regression Trees, and shallow artificial neural networks have been applied to short-term traffic prediction[2]. Compared with purely statistical models, these methods can better capture nonlinear relationships between input features and traffic states. Recent factor-aware studies

further indicate that multi-source factors and explainable learning strategies can help analyze the contribution of different variables to traffic flow forecasting[1]. Early deep learning-based traffic studies also demonstrated that neural networks could benefit from large-scale traffic data and improve prediction accuracy compared with conventional models[3].

2.2.Deep learning-based temporal models

With the rapid progress of deep learning, neural network models have become widely used for traffic forecasting. Recurrent Neural Networks, Long Short-Term Memory networks, and Gated Recurrent Units are representative temporal models. They are capable of learning sequential dependencies from historical traffic observations and can model nonlinear temporal evolution more effectively than traditional statistical methods[3,5]. LSTM models are particularly useful because their gating mechanisms can partially alleviate the vanishing gradient problem and preserve information over relatively long time intervals[5].

Sequence-to-sequence architectures further improved multi-step traffic prediction by mapping historical sequences into future sequences through encoder-decoder structures. In addition, Temporal Convolutional Networks and dilated causal convolutions have been introduced to capture temporal dependencies with higher computational efficiency than recurrent models. By stacking dilated convolutional layers, temporal convolution-based models can enlarge the receptive field and model longer historical patterns.

2.3.Graph neural networks for spatiotemporal traffic forecasting

Graph neural networks provide a natural solution for modeling road-network structures. In traffic forecasting, sensors or road segments can be represented as graph nodes, while physical road connections, distance-based relationships, or learned dependencies can be represented as graph edges. By propagating information over graph structures, GNN-based models can capture spatial correlations that are difficult to model using conventional temporal networks.

DCRNN is one of the representative works in this direction. It models traffic flow as a diffusion process on a directed graph and combines diffusion convolution with recurrent neural networks to capture both spatial and temporal dependencies[6]. STGCN replaces recurrent structures with graph convolution and temporal convolution, improving computational efficiency while modeling spatiotemporal features[7]. Graph WaveNet further introduces an adaptive adjacency matrix and dilated temporal convolution, allowing the model to learn hidden spatial dependencies even when the predefined graph is incomplete[8]. ASTGCN incorporates spatial-temporal attention mechanisms to capture recent, daily-periodic, and weekly-periodic dependencies[9]. MTGNN constructs graph structures for multivariate time series and applies mix-hop propagation to capture latent dependencies among variables[10]. AGCRN introduces node-adaptive parameter learning and data-adaptive graph generation, enabling the model to learn node-specific patterns without relying entirely on predefined graphs[11]. D2STGNN further decouples diffusion signals and inherent signals, providing a more detailed modeling strategy for dynamic traffic patterns[12].

2.4.Transformer-based models for long-range spatiotemporal forecasting

Transformer architectures have achieved strong performance in sequence modeling due to their self-attention mechanism, which can directly capture global dependencies across time steps[13]. Compared with recurrent models, Transformers can model long-range interactions more effectively and support parallel computation. As a result, Transformer-based methods have been increasingly applied to long-sequence time-series forecasting and traffic prediction.

Informer improves the efficiency of Transformer for long-sequence forecasting by introducing ProbSparse self-attention and a generative-style decoder[14]. Autoformer incorporates series decomposition and an auto-correlation mechanism to better capture periodic patterns and long-term dependencies[15]. FEDformer further introduces frequency-domain representation and seasonal-trend decomposition to reduce prediction error in long-term forecasting[16]. These models show that Transformer-based architectures can be effective for general long-sequence time-series prediction.

In the traffic forecasting field, attention-based and Transformer-based models have also been explored. GMAN applies spatial-temporal attention blocks and a transform attention layer to model relationships between historical and future time steps[17]. PDFormer introduces propagation delay-aware dynamic long-range attention to model delayed traffic influence and long-range spatial dependencies[18]. STAEformer shows that spatiotemporal adaptive embeddings can make vanilla Transformer architectures highly competitive for traffic forecasting[19].

2.5.Memory-augmented sequence modeling and research gap

Memory-augmented neural networks aim to improve sequence modeling by storing, retrieving, and reusing historical information. In natural language processing and long-context modeling, memory mechanisms have been used to extend the effective context length and preserve information across sequence segments. Representative models such as Transformer-XL, Memorizing Transformers, and BART-style encoder-decoder architectures show that explicit memory or context reuse can improve long-range dependency modeling[20–22].

Although memory-enhanced architectures have achieved progress in language modeling and general sequence tasks, their application to traffic forecasting remains relatively limited. Traffic data are inherently spatiotemporal: temporal patterns are coupled with road-network structure, and spatial dependencies evolve over time. Therefore, simply applying memory mechanisms to temporal sequences is insufficient. A traffic forecasting model should simultaneously address three issues: dynamic spatial dependency, localized traffic diffusion, and long-range temporal memory.

Existing studies have partially addressed these issues from different perspectives. GNN-based models focus on spatial dependency and graph diffusion[6–12]. Transformer-based models focus on long-range attention and global dependency[13–19]. However, few studies integrate memory-augmented temporal modeling, diffusion-based graph learning, and regional spatial attention into a unified framework. In particular, the interaction between dynamic adjacency learning and memory-enhanced Transformer modeling remains underexplored.

3.Methodologies

3.1.Datasets

We use the METR-LA dataset, a widely adopted benchmark dataset for traffic speed prediction. The dataset contains traffic speed records collected from 207 loop detectors on the Los Angeles highway network from March 1, 2012 to June 30, 2012. Following common practice, the raw traffic speed values are normalized using Z-score normalization before model training.

Data are of size $X \in R^{T*N*2}$, where T represents number of time stamps and N represents the number of nodes. For both datasets, Z-score normalization is applied on third dimension. The first value on the third dimension is the normalized value of speed, and the second value on the third dimension is the normalized value for time.

3.2.Model architecture

The proposed architecture integrates dynamic graph construction, diffusion-based spatial modeling, and memory-augmented sequence learning within a unified framework. Given the input representation $X \in R^{T*N*2}$, the model first constructs time-dependent spatial structures, followed by joint spatial-temporal encoding and sequence decoding. The proposed MDC-LSTM consists of three main components: a diffusion convolutional layer, a memory-augmented Transformer encoder, and a MemBART-style decoder. The RNN decoder is not used as the final architecture; it is only employed as an alternative decoder in the ablation study to evaluate the contribution of the MemBART-style decoder.

To capture dynamic spatial dependencies, we construct forward and backward dynamic adjacency matrices at each time step. Specifically, for timestamp t, the adjacency matrix is computed using a self-attention mechanism:

$$P_t = Softmax\left(\frac{Q_t K_t^T}{\sqrt{d_{model}}}\right) \odot A \quad (1)$$

where $X_t \in R^{N*F}$ denotes the node feature matrix at time step t, N is the number of traffic sensors, and F is the input feature dimension. In this study, F = 2, corresponding to the normalized traffic speed and temporal feature. The projection matrices W_Q and W_K are defined as W_Q, $W_K \in R^{F*dmodel,}$ where d_a is the attention projection dimension. The projected query and key matrices are $Q_t = X_t W_Q \in R^{N*dmodel}$ and $K_t = X_t W_K \in RR^{N*dmodel}$. The static adjacency mask $A \in \{0,1\}^{N*N}$ is derived from the predefined METR-LA road-network adjacency and is used to constrain attention to physically or topologically valid sensor relationships; self-connections are retained on the diagonal. The dynamic adjacency matrix is obtained as $P_t = softmax(Q_t K_t^T / sqrt(d_{model})) \odot A$, where $\odot$ denotes element-wise multiplication. W_Q and W_K are initialized using Xavier uniform initialization and are optimized jointly with the whole network. This formulation allows the connectivity between nodes to evolve over time while being regularized by the static road-network mask.

Based on the dynamic adjacency matrices, spatial dependencies are modeled via diffusion convolution. For an input signal $H^{(l)} \in R^{N*d\ model}$, the diffusion operation is defined as:

$$H^{(l+1)} = \sum_{k=0}^{K} \theta_k (P_t)^k H^l \quad (2)$$

where k is the maximum diffusion order rather than a projection matrix. In the experiments, k is set to 2, so information is propagated within two graph-diffusion steps. The parameters θ_k are learnable diffusion kernels, with $\theta_k \in R^{din \times dout}$ for the k-th diffusion step. They are initialized using Xavier uniform initialization, while bias terms are initialized to zero. The values of θ_k are not manually selected; they are learned by back-propagation from the training objective. The operator * used in the DCLSTM equations denotes the diffusion convolution defined above. Therefore, $\Theta * [X_t, h_{t-1}]$ means that the concatenated input and previous hidden state are first propagated over the forward and backward diffusion graphs and then linearly transformed by the corresponding learnable diffusion kernels. This operation enables multi-hop information propagation along directed graph structures. To incorporate temporal dynamics, we extend the standard LSTM by replacing linear transformations with diffusion convolution. At time step t, the DCLSTM cell is defined as:

$$i_t = \sigma(\Theta_i * [X_t, h_{t-1}]), f_t = \sigma(\Theta_f * [X_t, h_{t-1}]) \quad (3)$$

$$o_t = \sigma(\Theta_o * [X_t, h_{t-1}]), g_t = \tanh(\Theta g * [X_t, h_{t-1}]) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t, h_t = o_t \odot \tanh(c_t) \quad (5)$$

This formulation ensures that spatial interactions directly influence temporal state transitions.

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_{model}}}\right) \cdot \text{DiffConv}(V) \quad (6)$$

The output of the diffusion-recurrent layer is then fed into a memory-augmented Transformer encoder. The encoder operates on inputs of size $(M + N) \times 2d_{model}$, together with memory states $M_t \in R^{(M+N)*2\ d\ model}$. To integrate spatial information into attention, we modify the standard scaled dot-product attention: where the value matrix V is enhanced via diffusion convolution. This modification enables the attention mechanism to capture both global dependencies and graph structure.

In addition to global attention, a regional attention mechanism is introduced to emphasize local spatial interactions. For each node v, a neighborhood set $N(v)$ is constructed using graph traversal algorithms. The regional representation is computed as:

$$R(v) = \sum_{u \in N(v)} \alpha(v,u) F(u) \quad (7)$$

where $\alpha(v,u)$ denotes attention weights and $F(u)$ represents node features. This mechanism complements global attention by focusing on localized patterns.

The encoder maintains temporal continuity through memory states, which are propagated across input windows. For an input sequence of length T, the memory generated at

time t is reused at time $t+T$, enabling long-term dependency modeling beyond the fixed input horizon.

Finally, the decoder generates predictions for future time steps based on the encoded representations. A Transformer-based decoder is adopted, where adjacency information is further incorporated into attention by scaling value representations. The model outputs predictions $Y \in \mathrm{R}^{N * T}$ corresponding to the next T time steps.

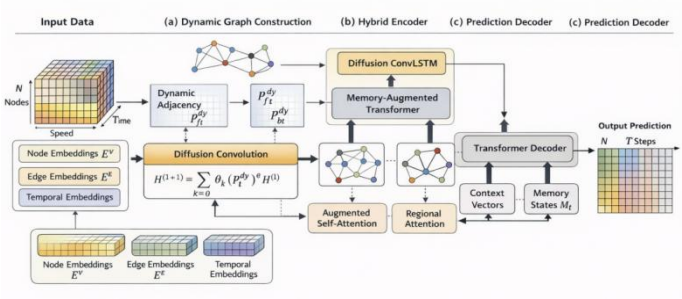

Fig 1. Architecture of the proposed MDC-LSTM model

3.3.Evaluation metrics

To evaluate the prediction performance, we adopt three widely used metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE).

The MAE measures the average magnitude of prediction errors:

$$\mathrm{MAE} = \frac{1}{\mathrm{NT}} \sum_{t=1}^{\mathrm{T}} \sum_{i=1}^{\mathrm{N}} |\mathrm{Y}_{t,i} - \widehat{\mathrm{Y}}_{t,i}| \tag{8}$$

They denote the ground truth and predicted value of node i at time step t, respectively.

The RMSE emphasizes larger errors by applying a quadratic penalty:

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{NT}} \sum_{t=1}^{\mathrm{T}} \sum_{i=1}^{\mathrm{N}} (\mathrm{Y}_{t,i} - \widehat{\mathrm{Y}}_{t,i})^2} \tag{9}$$

The MAPE evaluates the relative prediction error:

$$\mathrm{MAPE} = \frac{1}{\mathrm{NT}} \sum_{t=1}^{\mathrm{T}} \sum_{i=1}^{\mathrm{N}} \left| \frac{\mathrm{Y}_{t,i} - \widehat{\mathrm{Y}}_{t,i}}{\mathrm{Y}_{t,i} - \varepsilon} \right| \tag{10}$$

To avoid numerical instability when $\mathrm{Y}_{t,i}$ is close to zero, a small constant ε is added to the denominator in practice. In all experiments, ε is set to 1.0×10^{-5}. The same value is used for all models and all prediction horizons to ensure a fair comparison.

4.Experiments and results

4.1.Experimental setup

The dataset was divided into training, validation, and testing sets according to a 7:1:2 ratio. The historical input length was set to 12 time steps, corresponding to 60 minutes, and the model predicted future traffic states at 3, 6, and 12 steps, corresponding to 15, 30, and 60 minutes, respectively. The batch size was set to 64, the initial learning rate was set to 3×10^{-3}, and the model was trained using the Adam optimizer. Early stopping with a patience of 15 epochs was adopted based on the validation loss. To improve reproducibility, the main hyperparameters used in all experiments are summarized in Table 1.

Table 1. Main hyperparameter settings of MDC-LSTM.

Hyperparameter	Value / Setting
Dataset split	Train/validation/test = 7:1:2
Historical input length	12 steps (60 min)
Prediction horizons	3, 6, and 12 steps (15/30/60 min)
Batch size	64
Optimizer	Adam
Initial learning rate	3×10^{-3}
Loss function	MAE
Transformer dimension d_model	512
Attention heads	8
Dropout rate	0.1
Weight decay	1×10^{-4}
Gradient clipping threshold	5.0
Maximum training epochs	100
Early-stopping patience	15 epochs
Diffusion order K_d	2
Region size	96
Memory unit quantity M	64
Parameter initialization	Xavier uniform for trainable weights; zero for biases

The model was optimized using the Adam optimizer with Mean Absolute Error (MAE) as the primary objective function. A key technical modification involved enhancing the encoder with DCLSTM units and adapting the decoder's attention mechanism to scale based on the road network's adjacency matrix. To ensure robustness and prevent overfitting, we employed an early stopping strategy with a patience of 15 epochs. This configuration allows the model to effectively reconcile localized traffic perturbations with broader, long-term temporal trends.

4.2.Performance comparison and result analysis

To comprehensively evaluate the forecasting capability of the proposed MDC-LSTM model, we compare it with several representative baseline models, including DCRNN, STGCN, Graph WaveNet, ASTGCN, MTGNN, and AGCRN. The comparison is conducted on the METR-LA dataset under three commonly used prediction horizons: 15 minutes, 30 minutes, and 60 minutes, corresponding to Horizon 3, Horizon 6, and Horizon 12, respectively. The prediction performance is evaluated using MAE, RMSE, and MAPE. Lower values of these three metrics indicate better forecasting accuracy.

Table 2. Performance comparison of different models on the METR-LA dataset

	Methods	Horizon 3			Horizon 6			Horizon 12		
		MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
Datasets	DCRNN	2.77	5.38	7.30%	3.15	6.45	8.80%	3.6	7.6	10.50%
	STGCN	2.88	5.74	7.62%	3.47	7.24	9.57%	4.59	9.4	12.70%
	Graph WaveNet	2.69	5.15	6.90%	3.07	6.22	8.37%	3.53	7.37	10.01%
	ASTGCN	4.86	9.27	9.21%	5.43	10.61	10.13%	6.51	12.52	11.64%
	AGCRN	3.31	7.62	8.06%	4.13	9.77	10.29%	5.06	11.66	12.91%
	MTGNN	2.69	5.18	6.88%	3.05	6.17	8.19%	3.49	7.23	9.87%
	MDC-LSTM(ours)	2.62	5.01	6.63%	2.99	6.05	8.00%	3.44	7.19	9.73%

Table 2 presents the overall comparison results. It can be observed that MDC-LSTM achieves the best performance across all three forecasting horizons. Specifically, the proposed model obtains MAE values of 2.62, 2.99, and 3.44 for 15-minute, 30-minute, and 60-minute predictions, respectively. Compared with existing graph-based and attention-based baselines, MDC-LSTM consistently maintains lower prediction errors, indicating that the integration of diffusion convolution, dynamic graph learning, regional attention, and memory-enhanced temporal modeling is effective for spatiotemporal traffic forecasting.

4.2.1.Short-term forecasting analysis: 15-minute prediction

For the 15-minute forecasting horizon, all models generally achieve relatively low prediction errors because short-term traffic evolution is more strongly correlated with recent observations. In this setting, MDC-LSTM obtains the best performance, with an MAE of 2.62, an RMSE of 5.01, and a MAPE of 6.63%. Compared with the strongest baseline in terms of MAE, namely Graph WaveNet and MTGNN, both of which achieve an MAE of 2.69, MDC-LSTM reduces MAE by approximately 2.60%. In terms of RMSE, MDC-LSTM improves over Graph WaveNet from 5.15 to 5.01, corresponding to a relative reduction of approximately 2.72%. For MAPE, MDC-LSTM also outperforms MTGNN, reducing the value from 6.88% to 6.63%.

These results indicate that the proposed model has strong short-term forecasting capability. In short prediction intervals, traffic states are mainly affected by recent local variations and spatial propagation among neighboring road segments. The diffusion convolution module enables the model to capture directional traffic propagation over the road network, while the regional attention mechanism further strengthens local neighborhood representation. Therefore, MDC-LSTM can effectively model short-term traffic fluctuations and local spatial dependencies.

Compared with DCRNN, which also uses diffusion convolution, MDC-LSTM achieves lower MAE, RMSE, and MAPE. This suggests that diffusion convolution alone is not sufficient to fully capture complex traffic dependencies. The additional use of dynamic graph construction and regional attention provides more flexible spatial representation, allowing the model to better adapt to time-varying traffic conditions. Meanwhile, compared with STGCN, the proposed model achieves a substantial improvement, which indicates that combining recurrent temporal modeling with memory-enhanced attention is more effective than relying only on temporal convolution for short-term prediction.

4.2.2.Medium-term forecasting analysis: 30-minute prediction

When the prediction horizon increases to 30 minutes, the forecasting task becomes more difficult because uncertainty and error accumulation begin to increase. Under this setting, MDC-LSTM still achieves the best performance among all compared methods, with an MAE of 2.99, an RMSE of 6.05, and a MAPE of 8.00%. Compared with MTGNN, which is the strongest baseline at this horizon, MDC-LSTM reduces MAE from 3.05 to 2.99, RMSE from 6.17 to 6.05, and MAPE from 8.19% to 8.00%. These results correspond to relative improvements of approximately 1.97%, 1.94%, and 2.32%, respectively.

The medium-term results demonstrate that MDC-LSTM remains effective when the prediction interval becomes longer. Compared with short-term forecasting, 30-minute prediction requires the model to capture not only immediate traffic changes but also evolving temporal dependencies across a longer interval. The memory-augmented Transformer component contributes to this process by preserving historical context and enhancing temporal dependency modeling beyond local sequential patterns. At the same time, the LSTM-based temporal transition mechanism helps maintain sequential continuity, while diffusion convolution and regional attention jointly model the spatial propagation of traffic states.

The performance gap between MDC-LSTM and traditional graph-based models becomes more meaningful at this horizon. For example, compared with DCRNN, MDC-LSTM reduces MAE from 3.15 to 2.99 and RMSE from 6.45 to 6.05. This indicates that the proposed model has better capability in handling medium-term spatiotemporal evolution. Compared with Graph WaveNet and MTGNN, the improvements are relatively moderate but consistent across all metrics, suggesting that MDC-LSTM provides a more balanced modeling strategy by integrating adaptive spatial learning and memory-enhanced temporal modeling.

4.2.3.Long-term forecasting analysis: 60-minute prediction

The 60-minute forecasting horizon is the most challenging setting because long-term prediction is more vulnerable to accumulated errors, delayed traffic propagation, and dynamic changes in spatial dependencies. As shown in Table 2, MDC-LSTM achieves the best overall performance at this horizon, with an MAE of 3.44, an RMSE of 7.19, and a MAPE of 9.73%. Compared with MTGNN, which is the strongest baseline under the 60-minute setting, MDC-LSTM reduces MAE from 3.49 to 3.44, RMSE from 7.23 to 7.19, and MAPE from 9.87% to 9.73%. Although the relative improvement is not very large, the proposed model consistently obtains the lowest error values across all metrics.

The long-term results are particularly important because they reflect the model's ability to maintain predictive stability over extended forecasting horizons. In long-term traffic prediction, the influence of historical patterns may gradually weaken, while delayed spatial propagation and nonlinear traffic evolution become more significant. The memory-enhanced component of MDC-LSTM helps retain useful historical information across input windows, while the dynamic graph construction strategy allows the model to adjust spatial relationships according to changing traffic states. This combination enables the model to preserve competitive accuracy even when the prediction horizon is extended to 60 minutes.

Compared with STGCN, ASTGCN, and AGCRN, MDC-LSTM shows a clear advantage in long-term forecasting. For instance, the MAE of MDC-LSTM is 3.44, whereas STGCN, ASTGCN, and AGCRN report MAE values of 4.59, 6.51, and 5.06, respectively. These differences indicate that models relying on relatively limited temporal receptive fields or unstable attention structures may suffer more seriously from long-horizon error accumulation. In contrast, MDC-LSTM

benefits from the joint modeling of temporal memory, spatial diffusion, and local regional interactions.

It should also be noted that the performance improvement over the strongest baselines, such as MTGNN and Graph WaveNet, is relatively moderate at the 60-minute horizon. This suggests that although the proposed memory-augmented framework improves long-term prediction accuracy, long-horizon traffic forecasting remains difficult due to the intrinsic uncertainty of traffic systems. Therefore, further improvements in dynamic graph stability, noise robustness, and computational efficiency are still necessary for future research.

4.2.4.Overall discussion

Overall, the experimental results demonstrate that MDC-LSTM achieves the best forecasting performance across 15-minute, 30-minute, and 60-minute prediction horizons. The results verify the effectiveness of the proposed hybrid architecture from both spatial and temporal perspectives. For short-term forecasting, the advantage mainly comes from diffusion convolution and regional attention, which enhance the model's ability to capture local traffic propagation and fine-grained spatial interactions. For medium-term forecasting, the integration of dynamic graph learning and memory-enhanced temporal modeling helps the model maintain stable prediction accuracy as uncertainty increases. For long-term forecasting, the memory-augmented mechanism plays a key role in preserving historical context and alleviating the degradation of prediction performance over extended horizons.

Compared with classical spatiotemporal graph models, MDC-LSTM consistently achieves lower MAE, RMSE, and MAPE values. Compared with adaptive graph-based methods such as Graph WaveNet, MTGNN, and AGCRN, the proposed model shows competitive and generally superior performance, indicating that dynamic graph construction alone is not sufficient and should be combined with stronger temporal memory and local spatial attention. These findings support the effectiveness of combining diffusion convolution, regional attention, and memory-enhanced sequence modeling within a unified spatiotemporal forecasting framework.

Nevertheless, the results also reveal that the improvements over strong baselines become relatively small at the 60-minute horizon. This indicates that long-term traffic forecasting remains a challenging problem, and the proposed model still has room for improvement in terms of robustness and scalability. In particular, the stability of dynamic adjacency matrices and the computational cost introduced by regional attention should be further investigated in future work.

5.Ablation study

All experiments were performed under identical training settings unless otherwise stated. The learning rate, Transformer dimension, number of attention heads, dropout rate, weight decay, gradient clipping threshold, maximum training epoch, memory unit quantity, diffusion order, and region size are consistent with the unified hyperparameter settings reported in Table 1. This setting ensures that the ablation results are attributable to the removed or replaced module rather than to changes in optimization configuration.

5.1.Effect of regional attention

The Regional Attention (RA) mechanism was designed to explicitly capture localized spatial dependencies among neighboring traffic sensors. To evaluate its contribution, two model variants were compared: one incorporating Regional Attention and another without this component.

Table 3. Ablation study of the regional attention mechanism

Model Variant	MAE ↓	RMSE ↓	Epoch Time (min)
Without Regional Attention	3.58	7.51	8
With Regional Attention	3.44	7.19	37
Relative Improvement	3.91%	4.26%	-

As shown in Table 3, incorporating the Regional Attention mechanism significantly improves prediction performance. Specifically, MAE decreases from 3.58 to 3.44, while RMSE decreases from 7.51 to 7.19. These improvements indicate that local spatial interactions among neighboring traffic nodes contain valuable information that cannot be fully captured by global attention mechanisms alone.

The performance gain, however, comes at the cost of substantially increased computational complexity. The training time per epoch increases from 8 minutes to 37 minutes. This increase is mainly caused by regional attention rather than by diffusion-convolution-augmented self-attention. Regional attention requires explicit neighborhood construction, local feature aggregation, and attention-weight computation for each sensor region. Since the adopted region size is 96, the model performs a relatively large number of local attention operations at each training iteration. Therefore, the sharp increase in epoch time reflects the additional local-neighborhood modeling cost introduced by Regional Attention.

These findings demonstrate a trade-off between forecasting accuracy and computational efficiency. For applications where predictive performance is the primary objective, Regional Attention provides substantial benefits. Conversely, in resource-constrained environments, a simplified model without Regional Attention may offer a more practical solution.

5.2.Effect of diffusion-convolution-augmented self-attention

To investigate the effectiveness of incorporating graph diffusion information into the self-attention mechanism, a comparative experiment was conducted between the standard self-attention module and the proposed diffusion-convolution-augmented self-attention module.

The results presented in Table 4 demonstrate that augmenting self-attention with diffusion convolution leads to consistent improvements across all evaluation metrics when the regional attention setting is kept unchanged. Compared with the standard self-attention mechanism, the proposed approach reduces MAE by 1.43% and RMSE by 1.64%. This

comparison isolates the effect of diffusion-convolution augmentation under the same regional attention configuration.

Table 4. Ablation study of diffusion-convolution-augmented self-attention

Model Variant	MAE ↓	RMSE ↓	Epoch Time (min)
Standard Self-Attention + Regional Attention	3.49	7.31	37
Diffusion-Augmented Self-Attention + Regional Attention	3.44	7.19	37
Relative Improvement	1.43%	1.64%	-

This improvement can be attributed to the ability of diffusion convolution to explicitly encode graph structural information into the attention computation process. By incorporating topological relationships among traffic sensors, the attention mechanism becomes capable of modeling both temporal dependencies and spatial interactions simultaneously.

Notably, the inclusion of diffusion convolution introduces negligible additional computational overhead under the same regional attention setting. Both variants require approximately 37 minutes per training epoch. This does not contradict the result in Table 3: the increase from 8 minutes to 37 minutes is caused by adding Regional Attention, whereas Table 4 compares two attention variants after Regional Attention is already enabled. In other words, diffusion convolution mainly changes the representation operation inside attention, while the dominant computational cost comes from neighborhood search and local regional aggregation. Therefore, Diffusion-Convolution-Augmented Self-Attention provides an effective way to improve forecasting accuracy without becoming the primary source of additional training cost.

Therefore, Diffusion-Convolution-Augmented Self-Attention provides an effective and computationally efficient means of improving traffic forecasting accuracy.

5.3.Comparison of decoder architectures

To evaluate the impact of decoder design on forecasting performance, the proposed MemBART decoder was compared against a traditional RNN decoder equipped with an attention mechanism.

Table 5. Comparison of decoder architectures

Decoder Variant	MAE ↓	RMSE ↓	MAPE ↓	Epoch Time (min)
RNN Decoder with Attention	3.87	8.06	12.65%	37
MemBART Decoder	3.44	7.19	9.73%	37
Relative Improvement	11.13%	10.81%	23.08%	-

As shown in Table 5, the MemBART decoder consistently outperforms the RNN-based decoder across all evaluation metrics. Compared with the RNN decoder, the MemBART decoder achieves reductions of 11.13% in MAE, 10.81% in RMSE, and 23.08% in MAPE.

The superior performance of the MemBART decoder can be attributed to the stronger sequence modeling capability of Transformer-based architectures. Unlike recurrent neural networks, which process temporal information sequentially, self-attention mechanisms can directly capture long-range dependencies and complex interactions among different time steps. This capability enables the decoder to generate more accurate traffic flow predictions, particularly for longer forecasting horizons.

Interestingly, both decoder architectures require approximately the same training time per epoch. Therefore, the observed performance gains originate primarily from architectural improvements rather than increased computational complexity.

These results confirm that the MemBART decoder constitutes a more effective decoding strategy for traffic forecasting tasks.

5.4.Overall discussion

The largest performance gain originates from replacing the conventional RNN decoder with the MemBART decoder, highlighting the importance of advanced temporal modeling capabilities. The diffusion-convolution-augmented self-attention mechanism further improves prediction accuracy by effectively incorporating graph structural information into the attention process. Finally, Regional Attention enhances the model's ability to exploit localized spatial correlations, leading to additional performance improvements.

Collectively, these components form a complementary framework capable of simultaneously capturing long-range temporal dependencies, global spatial relationships, and local neighborhood interactions. The experimental results validate the effectiveness of the proposed MDC-LSTM architecture and demonstrate its superiority for traffic flow forecasting tasks.

6.Limitations and future work

Although the proposed MDC-LSTM model achieves promising performance in traffic flow prediction, several limitations should be acknowledged and further addressed in future work.

First, the computational cost of MDC-LSTM is relatively high compared with lightweight spatiotemporal forecasting models. This limitation mainly comes from the diffusion convolution operations, dynamic adjacency matrix construction, and regional attention mechanism. In particular, regional attention requires neighborhood search, local feature aggregation, and attention weight calculation for traffic sensors, which substantially increases training time and memory consumption. As observed in the ablation study, the inclusion of regional attention improves prediction accuracy but also leads to a clear increase in epoch time. Therefore, although the mechanism is beneficial for capturing fine-grained local spatial dependencies, its computational burden may limit the scalability of the model when it is applied to larger road networks or real-time traffic management systems.

Second, the dynamic adjacency matrix learned by the model may be unstable under noisy or abnormal traffic conditions. In real-world traffic systems, sensor failures,

missing values, accidents, weather changes, and sudden congestion may disturb the learned spatial correlations among sensors. Since the dynamic adjacency matrix is generated from input-dependent representations, its values may fluctuate when the input data contain noise or outliers. Such instability may affect the reliability of spatial dependency modeling and reduce the robustness of the model in complex traffic environments. Therefore, improving the stability and interpretability of dynamic graph learning remains an important research direction.

Third, the current model is evaluated mainly on the METR-LA benchmark dataset. Although this dataset is widely used in traffic forecasting studies, a single dataset cannot fully reflect the diversity of real-world transportation networks. Different cities may have different road structures, traffic regulations, commuting patterns, sensor densities, and congestion characteristics. Therefore, the generalization ability of MDC-LSTM should be further verified on additional datasets, such as PEMS-BAY, PEMS03, PEMS04, PEMS07, and PEMS08. Cross-dataset evaluation and transfer learning experiments would provide stronger evidence for the applicability of the proposed model in different traffic scenarios.

Fourth, MDC-LSTM contains multiple interconnected modules, including diffusion convolution, LSTM-based temporal transition, memory-augmented Transformer encoding, dynamic graph construction, and regional attention. Although this hybrid design improves representation capacity, it also introduces several hyperparameters, such as diffusion steps, region size, hidden dimension, memory length, attention heads, and Transformer dimension. The performance of the model may be sensitive to these settings, and extensive parameter tuning may be required when applying the model to new datasets. This may affect reproducibility and practical deployment efficiency.

To address these limitations, future work will focus on the following directions. First, computational optimization will be conducted to reduce the training and inference cost of MDC-LSTM. Sparse graph representations, efficient neighborhood sampling, low-rank approximation, attention pruning, and parallelized regional attention can be explored to reduce unnecessary computation. Model compression techniques, such as knowledge distillation, parameter pruning, and low-precision quantization, may also be introduced to construct a lightweight version of MDC-LSTM for real-time forecasting applications.

Second, more stable dynamic graph learning strategies will be investigated. Regularization constraints can be imposed on the learned adjacency matrices to reduce excessive fluctuation across time steps. Temporal smoothing, graph sparsity constraints, and topology consistency loss may help improve the robustness of dynamic graph construction. In addition, incorporating prior road-network topology into the adaptive graph learning process may prevent the learned adjacency matrix from deviating excessively from realistic traffic structures.

Third, future studies will improve the robustness of the model under noisy and incomplete traffic data. Frequency-domain denoising methods, such as Fourier Transform and Wavelet Transform, can be used to separate high-frequency noise from meaningful traffic patterns. Missing value imputation and anomaly detection modules may also be integrated into the forecasting framework to enhance reliability under real-world sensor faults or abnormal traffic events.

Fourth, the generalization ability of MDC-LSTM will be further evaluated on multiple benchmark datasets and real-world traffic networks. In addition to standard in-dataset testing, cross-city prediction, transfer learning, and domain adaptation experiments can be conducted to examine whether the learned spatial-temporal representations can be transferred across different road networks. Such experiments would provide more comprehensive evidence for the practical applicability of the proposed framework.

Finally, future research may extend MDC-LSTM from deterministic forecasting to probabilistic forecasting. Instead of only predicting a single future traffic value, probabilistic models can estimate prediction intervals or uncertainty distributions. This would be useful for risk-aware traffic management, congestion warning, and decision-making under uncertainty.

7.Conclusion

This study proposes MDC-LSTM, a memory-augmented diffusion convolutional LSTM network for traffic flow prediction. The proposed framework integrates dynamic graph construction, diffusion-based spatial modeling, regional attention, LSTM-based temporal transition, and memory-augmented Transformer encoding into a unified spatiotemporal forecasting architecture. By combining these components, MDC-LSTM is designed to capture local spatial diffusion, dynamic road-network dependencies, and long-range temporal correlations simultaneously.

The main motivation of this study is to address two key challenges in traffic forecasting: dynamic spatial dependency modeling and long-range temporal dependency learning. Traditional graph-based methods often rely on static or weakly adaptive adjacency matrices, which may fail to describe time-varying traffic interactions. Meanwhile, conventional recurrent or convolutional temporal models may suffer from limited long-range modeling ability when the prediction horizon increases. To overcome these limitations, MDC-LSTM introduces dynamic adjacency construction to adaptively learn spatial relationships among traffic sensors, diffusion convolution to model directional traffic propagation, regional attention to enhance local spatial representation, and a memory-augmented mechanism to preserve historical context across input windows.

Experimental results on the METR-LA dataset demonstrate that MDC-LSTM achieves competitive forecasting performance compared with representative baseline models. The horizon-wise analysis shows that the proposed model performs effectively across 15-minute, 30-minute, and 60-minute prediction horizons. In short-term forecasting, the advantage of MDC-LSTM mainly comes from its ability to capture local traffic propagation and neighborhood-level spatial interactions. In medium-term forecasting, the combination of dynamic graph learning and memory-enhanced temporal modeling helps maintain stable prediction

accuracy. In long-term forecasting, the memory-augmented mechanism contributes to preserving historical information and alleviating performance degradation over extended horizons.

The ablation studies further confirm the effectiveness of the key components of MDC-LSTM. The regional attention mechanism improves prediction accuracy by strengthening local spatial dependency modeling, although it increases computational cost. The diffusion-convolution-augmented self-attention mechanism enhances spatial-temporal representation learning without introducing substantial additional computation. The MemBART-style decoder outperforms the traditional RNN decoder, indicating that memory-enhanced Transformer-based decoding is more suitable for capturing long-range temporal dependencies in traffic forecasting tasks.

Despite these promising results, several limitations remain. The proposed model has relatively high computational complexity, mainly due to dynamic graph construction, diffusion convolution, and regional attention. This may restrict its scalability in large-scale road networks or real-time deployment scenarios. In addition, the learned dynamic adjacency matrix may become unstable under noisy, incomplete, or abnormal traffic conditions, which could affect the robustness of spatial dependency modeling. Furthermore, the current evaluation is mainly based on the METR-LA dataset, and additional experiments on multiple traffic datasets are needed to further verify the generalization ability of the model.

In future work, computationally efficient variants of MDC-LSTM will be explored through sparse graph learning, lightweight attention, model pruning, knowledge distillation, and quantization. More robust dynamic graph learning methods will also be investigated by introducing temporal smoothness constraints, graph regularization, and prior road-network topology. Moreover, the proposed framework will be evaluated on more traffic datasets and extended to cross-network transfer learning and probabilistic forecasting. These directions are expected to improve the scalability, robustness, and practical applicability of MDC-LSTM in real-world intelligent transportation systems.

Overall, this study provides a new perspective for combining graph diffusion, regional spatial attention, and memory-augmented temporal modeling in traffic flow prediction. The proposed MDC-LSTM framework contributes to the development of accurate and robust spatiotemporal forecasting models and has potential value for intelligent traffic management, congestion mitigation, and urban mobility optimization.

References

[1] WANG X. A Case Study of the Influence of Multifarious Factors on Traffic Flow Forecasting. Applied Artificial Intelligence Research, 2026, 2(2).

[2] VLAHOGIANNI E I, KARLAFTIS M G, GOLIAS J C. Short-term traffic forecasting: Where we are and where we are going. Transportation Research Part C: Emerging Technologies, 2014, 43: 3-19.

[3] LV Y, DUAN Y, KANG W, et al. Traffic flow prediction with big data: A deep learning approach. IEEE Transactions on Intelligent Transportation Systems, 2015, 16(2): 865-873.

[4] CHANDRA S R, AL-DEEK H. Predictions of freeway traffic speeds and volumes using vector autoregressive models. Journal of Intelligent Transportation Systems, 2009, 13(2): 53-72.

[5] HOCHREITER S, SCHMIDHUBER J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780.

[6] LI Y, YU R, SHAHABI C, et al. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting//Proceedings of the International Conference on Learning Representations, 2018.

[7] YU B, YIN H, ZHU Z. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting//Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018: 3634-3640.

[8] WU Z, PAN S, LONG G, et al. Graph WaveNet for deep spatial-temporal graph modeling//Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019: 1907-1913.

[9] GUO S, LIN Y, FENG N, et al. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting//Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 922-929.

[10] WU Z, PAN S, LONG G, et al. Connecting the dots: Multivariate time series forecasting with graph neural networks//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020: 753-763.

[11] BAI L, YAO L, LI C, et al. Adaptive graph convolutional recurrent network for traffic forecasting//Advances in Neural Information Processing Systems, 2020: 17804-17815.

[12] SHAO Z, ZHANG Z, WEI W, et al. Decoupled dynamic spatial-temporal graph neural network for traffic forecasting. Proceedings of the VLDB Endowment, 2022, 15(11): 2733-2746.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need//Advances in Neural Information Processing Systems, 2017: 5998-6008.

[14] ZHOU H, ZHANG S, PENG J, et al. Informer: Beyond efficient Transformer for long sequence time-series forecasting//Proceedings of the AAAI Conference on Artificial Intelligence, 2021: 11106-11115.

[15] WU H, XU J, WANG J, et al. Autoformer: Decomposition Transformers with auto-correlation for long-term series forecasting//Advances in Neural Information Processing Systems, 2021: 22419-22430.

[16] ZHOU T, MA Z, WEN Q, et al. FEDformer: Frequency enhanced decomposed Transformer for long-term series forecasting//Proceedings of the 39th International Conference on Machine Learning, PMLR, 2022: 27268-27286.

[17] ZHENG C, FAN X, WANG C, et al. GMAN: A graph multi-attention network for traffic prediction//Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 1234-1241.

[18] JIANG J, HAN C, ZHAO W X, et al. PDFormer: Propagation delay-aware dynamic long-range Transformer for traffic flow prediction// Proceedings of the AAAI Conference on Artificial Intelligence, 2023: 4365-4373.

[19] LIU H, DONG Z, JIANG R, et al. Spatio-temporal adaptive embedding makes vanilla Transformer SOTA for traffic forecasting//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023: 4125-4129.

[20] DAI Z, YANG Z, YANG Y, et al. Transformer-XL: Attentive language models beyond a fixed-length context//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 2978-2988.

[21] WU A, SZEGEDY C, MISRA D, et al. Memorizing Transformers// Proceedings of the International Conference on Learning Representations, 2022.

[22] LEWIS M, LIU Y, GOYAL N, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 7871-7880.