Research on STA-GRU Vehicle Trajectory Prediction Model Based on Spatio-Temporal Attention Mechanism

Shijie Liu [1, *], Shuobo Wang [2]

[1] Department of Mechatronics and Vehicle Engineering, Taiyuan University, Taiyuan, China, 030032

[2] School of Mechanical and Precision Instrument Engineering, Xi'an University of Technology, Xi'an, China, 710048

* Corresponding Author Email: shijie_liu1232025@163.com

Abstract. Autonomous driving needs to accurately predict the trajectory of surrounding vehicles and pedestrians to reduce the risk of accidents [1]. Traditional methods (such as physical models and machine learning) have limitations in complex scenarios and high-dimensional data processing, while existing deep learning models (such as LSTM, GRU) lack the ability to model long-term predictions and real-time interactions. Therefore, more advanced algorithms are needed to improve the accuracy and adaptability of trajectory prediction. To solve the above problems, this paper proposes a gated cyclic unit model (STA-GRU) that combines temporal and spatial characteristics with attention mechanism, aiming to improve the accuracy and robustness of trajectory prediction in complex traffic scenarios. In the simulation test of NGSIM data set, the mean RMSE of 1s is 0.32284. This model is capable of enhancing the accuracy of vehicle trajectory prediction. As a result, it furnishes crucial data support for traffic signal control, traffic flow optimization, and other related aspects. By doing so, it effectively improves traffic efficiency and mitigates congestion-related accidents. Furthermore, the incorporation of the spatio - temporal attention mechanism augments the interpretability of the model, facilitating a more in - depth understanding of its operational principles and prediction outcomes.

Keywords: Intelligent Connected Vehicles; Vehicle Trajectory Prediction; Spatio-temporal attention mechanism; GRU.

1. Introduction

With the continuous advancement of the automotive industry, autonomous driving technology has become a critical development direction in the field. Autonomous driving systems must replace a driver's predictive capabilities by forecasting the future trajectories of surrounding vehicles and pedestrians to reduce accident risks. Such systems require precise environmental perception and understanding, with trajectory prediction being a key component.

In the initial stages of autonomous vehicle trajectory prediction research, physics-based modeling approaches dominated the field. Zhang Chi-hao et al. formulated the control problem as a quadratic programming solution by developing longitudinal dynamic equations that accounted for various forces including tire friction, aerodynamic drag, and road gradient [2]. Similarly, Fei Xiansong et al. constructed a comprehensive vehicle dynamics model grounded in Newton's second law, integrating a nonlinear tire model to enhance trajectory prediction accuracy [3]. Vincent P, on the other hand, adopted a simplified approach by reducing the vehicle dynamics to a linear bicycle model for collision avoidance applications [4]. While these physics-based methods offer computational efficiency and straightforward implementation, they demonstrate limited adaptability when confronted with complex, real-world traffic scenarios.

With the advancement of machine learning techniques, conventional algorithms including Support Vector Machines (SVMs) and decision trees have been extensively employed in trajectory prediction tasks. Mandalia et al. implemented SVMs by projecting data into high-dimensional feature spaces to determine optimal hyperplanes for behavioral classification, specifically applying this methodology to lane-change maneuver detection. Subsequent developments introduced more sophisticated

approaches, such as Wang's graph convolutional interaction network that integrates lane topology constraints to enhance vehicle trajectory prediction accuracy [5]. Another significant contribution came from Shen Y. et al., who developed a context-aware framework combining multi-sensor perception with Enhanced Inverse Reinforcement Learning (EIRL) to improve decision-making in autonomous driving scenarios [6]. While these machine learning-based solutions demonstrated marked improvements in prediction precision over conventional physics-based models, they still face inherent challenges including heavy reliance on large-scale high-quality datasets, difficulties in effectively processing high-dimensional information, and limitations in capturing complex spatiotemporal correlations within dynamic environments.

Recent advances in artificial intelligence and big data have significantly enhanced deep learning's ability to handle complex pattern recognition and prediction tasks, opening new possibilities for trajectory prediction in autonomous driving. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs)—along with their variants such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs)—have been widely adopted for this purpose. CNNs excel at processing visual data (e.g., road images from onboard cameras) to extract spatial features like object position and shape, while RNNs specialize in modeling temporal dependencies in sequential trajectory data. To better capture spatiotemporal relationships, Sheng et al integrated Graph Neural Networks (GNNs) with CNNs for feature extraction and employed GRUs for trajectory decoding [7]. Li proposed a vehicle-lane aggregation model that encodes target and obstacle dynamics using GRUs and incorporates artificial potential fields for trajectory generation [8]. Gao Zhenhai Et Al. developed an intention-aware prediction framework, leveraging an encoder-decoder architecture to forecast trajectories up to 6 seconds ahead with high accuracy [9]. Meanwhile, Deo and Trivedi introduced a convolutional social pooling mechanism with LSTM to predict highway trajectories by considering surrounding vehicle interactions [10]. However, while LSTM-based methods effectively model temporal dependencies, they often overlook spatial interactions between agents, limiting their accuracy and robustness in complex, dynamic traffic environments.

In conclusion, this paper introduces the STA-GRU model, which integrates spatio-temporal features with an attention mechanism to address trajectory prediction. By modeling the temporal dependencies in historical vehicle trajectories and the spatial interactions between vehicles, the proposed framework enables precise forecasting of future trajectories. Additionally, the study visualizes the spatio-temporal attention weights, explicitly illustrating the critical distributions of time steps and spatial grid contributions, thereby enhancing the model's interpretability. This work highlights a data-driven approach that balances predictive accuracy with transparency, offering practical insights for intelligent transportation systems.

2. Introduction to the STA-GRU Model

This paper proposes a traffic trajectory prediction method based on the STA-GRU model. A social tensor is constructed using a lane-based grid definition. A 13×3 spatial grid is established around the predicted vehicle, where each column corresponds to a lane, and the distance between rows is approximately 15 feet (roughly equivalent to the length of a typical vehicle). Within this grid, all vehicles except the target vehicle are considered neighboring vehicles, with each neighbor assigned to a unique grid cell based on the position of its front bumper.

The input to the STA-GRU model includes the T-step historical trajectory data of all vehicles within the grid. The historical trajectory data of each vehicle is processed separately through its corresponding GRU model. The model's output is the predicted trajectory of the target vehicle for the next H steps. The STA-GRU model learns both temporal attention weights, which analyze the influence of the historical trajectories of the target vehicle and its neighbors on the prediction results, and spatial attention weights, which interpret the specific impact of neighboring vehicles on the target vehicle's trajectory prediction. By integrating spatiotemporal information and a dynamic attention mechanism, the model effectively handles complex interaction relationships in traffic scenarios. As

shown in Figure 1, the STA-GRU model adopts an encoder-decoder architecture. It incorporates a hierarchical attention mechanism to dynamically capture temporal dependencies and spatial interactions between vehicles while utilizing a social context grid to encode the positional information of neighboring vehicles, thereby optimizing the processing efficiency of sparse interactions.

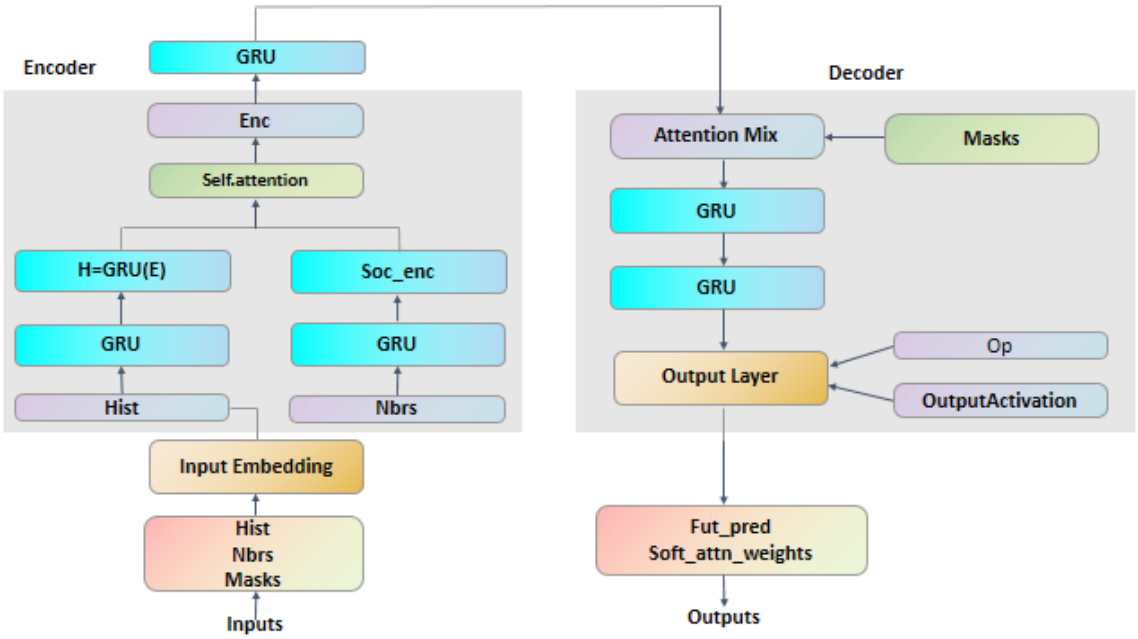

Figure 1. Model framework diagram.

2.1. GRU Encoder

In the STA-GRU model, the primary function of the encoder is to transform historical trajectory data into feature representations that capture vehicle motion patterns. The encoder first embeds the historical trajectory data (including position coordinates, velocity, etc.) through a linear layer, projecting it into a high-dimensional space. This process involves encoding both the ego vehicle's historical trajectory and neighboring vehicles' trajectories, while incorporating a temporal attention mechanism. Ultimately, the temporally-attended neighboring vehicle trajectory encodings are concatenated with the ego vehicle's historical trajectory encoding.

The embedding layer in the encoder is expressed by Equation (1), where X is the input historical trajectory data, W_ip and b_ip are the weight and bias of the embedding layer, respectively, and E represents the embedded features. The GRU, with its update gate and reset gate mechanisms[11], effectively controls information flow, enabling better handling of long sequential data. Thus, the embedded features are fed into the GRU layer to capture dynamic variations in the time series, as shown in Equation (2), where H denotes the GRU output, representing the hidden state of the vehicle's historical trajectory. The temporal attention mechanism dynamically adjusts the model's sensitivity to different time steps by computing weights for each time step, as formulated in Equations (3) and (4),α represents the attention weights, W_att is the weight matrix of the attention layer, and H_new is the hidden state adjusted by the attention mechanism.

$$E = Leaky\,\mathrm{Re}\,LU\left(W_{ip}X + b_{ip}\right) \tag{1}$$

$$H = GRU\left(E\right) \tag{2}$$

$$\alpha = Soft\max\left(W_{att}\tanh\left(H\right)\right) \tag{3}$$

$$H_{new} = \sum \alpha \square\ H \tag{4}$$

2.2. GRU Decoder

The decoder in the STA-GRU model is responsible for generating future trajectories. Its primary task is to progressively predict the vehicle's future trajectory based on the encoder's output and spatial attention weights.

The decoder's initial input is the encoder's output hidden state *H_new*, which contains spatiotemporal features of the historical trajectory and enhanced key information through the attention mechanism. The decoder employs GRU layers to sequentially generate future trajectory points. At each time step, the decoder's input consists of the output from the previous time step and the encoder's hidden state. The GRU's output can be expressed by Equation (5), where *D_t* represents the decoder's output at time step *t*. The decoder's output is then mapped to trajectory coordinates through a linear layer to generate future trajectory points, as shown in Equation (6), where *W_out* and *b_out* are the weight and bias of the output layer respectively, and Y_t denotes the predicted future trajectory point. Finally, the predicted trajectory points are processed through an activation function to ensure the output conforms to realistic physical constraints [12]. This complete process enables the STA-GRU model to generate accurate and physically plausible future trajectories while effectively capturing both temporal dynamics and spatial interactions in complex traffic scenarios.

$$Y_t = W_{out} D_t + b_{out} \tag{5}$$

$$D_t = GRU\left(H_{new}, D_{t-1}\right) \tag{6}$$

2.3. Spatio-temporal attention mechanism

The temporal and spatial attention mechanisms serve as key components in the STA-GRU model, dynamically allocating weights to enable focused attention on critical timesteps and spatial locations.

A feedforward network first calculates weights for each timestep using Equations (3) and (4), where tanh functions as the hyperbolic tangent activation. These attention weights then perform element-wise multiplication (denoted by $\odot$) with hidden states before weighted summation generates new hidden representations. The ReLU activation function subsequently applies nonlinear transformation to these states as expressed in Equation (7). This attention framework empowers the STA-GRU model to automatically emphasize prediction-relevant timesteps and spatial positions, thereby enhancing trajectory forecasting accuracy and robustness. The temporal attention mechanism employs Softmax layers to compute timestep importance weights that produce context vectors through weighted aggregation, while spatial attention concatenates encoded target and neighboring vehicle representations before applying quadratic attention to identify crucial interaction patterns - significantly improving the model's capacity to handle complex traffic environments.

$$H_{new} = \mathrm{Re}\,LU\left(H_{new}\right) \tag{7}$$

3. The establishment of the STA-GRU model

3.1. Analysis of single Feature and multiple feature input selection

The NGSIM (Next Generation Simulation) dataset, officially released by the Federal Highway Administration (FHWA) through rigorous and standardized procedures, demonstrates authoritative and cutting-edge leadership in the specialized field of vehicle trajectory prediction research. This study utilizes data collected from the I-80 freeway segment in Emeryville, California. Located at a critical transportation hub in the San Francisco Bay Area, this section exhibits highly dynamic traffic flow, diverse vehicle types, and complex participant behaviors, making the collected data particularly valuable for research. It provides researchers with highly referential traffic scenario samples, facilitating in-depth understanding of vehicle trajectory formation mechanisms and evolution patterns in complex traffic environments. The data from the I-80 freeway in Emeryville, California, was

annotated and processed to obtain 47 dimensional features as shown in Table 1. Dimensions 1-6 encode vehicle identification numbers, coordinates, lateral lane information, and other basic attributes.

The 7th dimension captures lateral lane-changing behaviors, including: left lane change, right lane change, and lane keeping. The determination method involves: for each vehicle at each frame, examining 40 preceding and 40 subsequent frames, then comparing current lane ID with past and future lane IDs. If the future lane ID is greater than the current one, or the current ID is greater than the past, it is labeled as a right lane change; if the future lane ID is smaller than the current one, or the current ID is smaller than the past, it is labeled as a left lane change; otherwise, it is labeled as lane keeping. The 8th dimension represents longitudinal speed change behaviors, including: deceleration and maintaining speed. The judgment method involves: for each vehicle at each frame, examining 50 preceding and 30 subsequent frames to calculate historical average speed (V_Hist) and future average speed (V_Fut). If the ratio V_Fut/V_Hist ≥ 0.8, it is labeled as maintaining speed; if < 0.8, labeled as deceleration. Dimensions 9-47 are constructed by establishing a 3×13 grid to identify neighboring vehicles for each vehicle at each frame. First, a spatial grid is defined with a longitudinal range of 180 meters (90 meters ahead and behind the target vehicle), divided into 15-meter intervals. Three lanes are identified: left adjacent lane, current lane, and right adjacent lane, with each lane containing 13 grid cells. For each target vehicle, the longitudinal distance (ΔY) to surrounding vehicles is calculated, and neighbor vehicle IDs are assigned to corresponding grid indices: left adjacent lane indices 9-21, current lane indices 22-34, and right adjacent lane indices 35-47. This results in a total of 39 dimensions (3 lanes × 13 grids) representing the spatial distribution of neighboring vehicles, as shown in Table 1.

Table 1. Dimensional index table.

Dimension Index	Description	Note
1	Scene	/
2	Vehicle Identification Number	Vehicle Number
3	Timestamp	/
4	Absolute X coordinate	Unit: Ft
5	Absolute Y coordinate	Unit: Ft
6	Horizontal lane coding	Encode from left to right in sequence
7	Horizontal Lane Behavior Coding	1- Change lanes to the left /2- Change lanes to the right /3- Go straight ahead.
8	Vertical behavioral coding	1- Slow down,2- Maintain
9-47	3*13 grid marking	/

3.2. Model training

The model employs a dual-layer GRU architecture, with inputs consisting of 1-second future trajectory prediction data and 3-second historical trajectory data. Experiments were conducted on a computer equipped with an Intel Core i7-13750HX processor, 16GB system memory, and an NVIDIA GeForce RTX 4050 GPU for training, utilizing the PyTorch deep learning framework.The network configuration sets both the encoder and decoder GRU layers to 2 layers, with a hidden layer size of 128 and dropout rate of 0.2. For training parameters, the model undergoes 80 epochs with a batch size of 1024, using the Adam optimizer and mean squared error as the loss function. The learning rate is set to 0.001 with a weight decay of 0.0001. The dataset was partitioned into training, validation, and test sets in a 7:2:1 ratio while ensuring no vehicle ID overlap across these subsets, and subsequently fed into the STA-GRU model for training.

3.3. Visualization of Model Spatiotemporal Attention

The spatial attention mechanism of the STA-GRU model is shown in Figure 2. The figure reveals that the ego vehicle assigns nearly zero attention weights to rear vehicles, indicating that trailing

vehicles have minimal impact on its behavior, which aligns with the common knowledge that drivers primarily focus on forward traffic conditions. When the ego vehicle is in the center lane, higher attention weights are allocated to the second and third forward grid cells, demonstrating that neighboring vehicles ahead significantly influence the ego vehicle's future trajectory on highway mainlines, as shown in Figure.2.

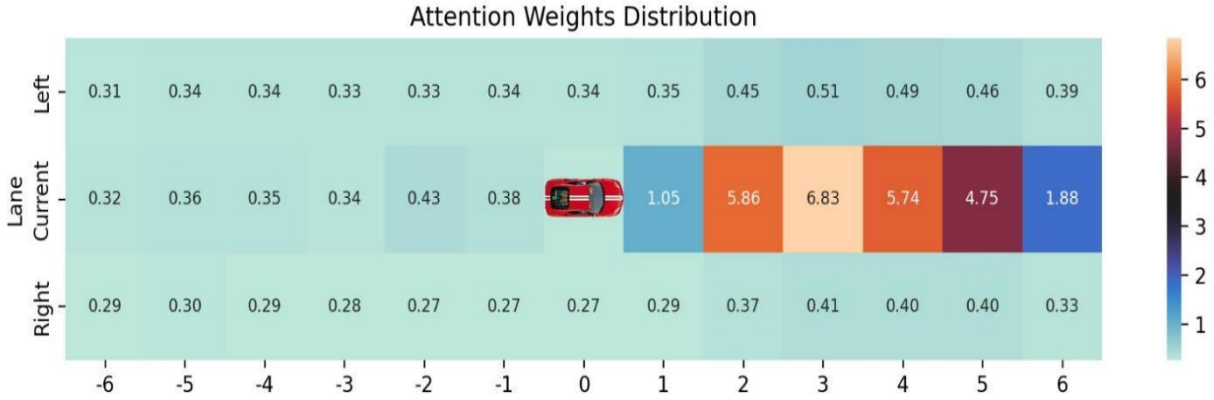

Figure 2. Spatial attention weight.

The temporal attention mechanism of the STA-GRU model is illustrated in Figure 3, which displays the attention weight distribution across different timesteps of the ego vehicle's historical trajectory. The horizontal axis represents the sequence of timesteps, while the vertical axis shows the normalized attention weight values.The results reveal a sub-peak at the initial timestep (t=0) and a prominent peak at timestep 15 (t=15), indicating that the model assigns significant importance to both the starting state of the trajectory and critical turning moments. This demonstrates the trajectory prediction model's sensitivity to initial motion conditions and key maneuver points. The peak values marked by red dashed lines clearly show the model's distinct temporal attention focusing phenomenon during decision-making.

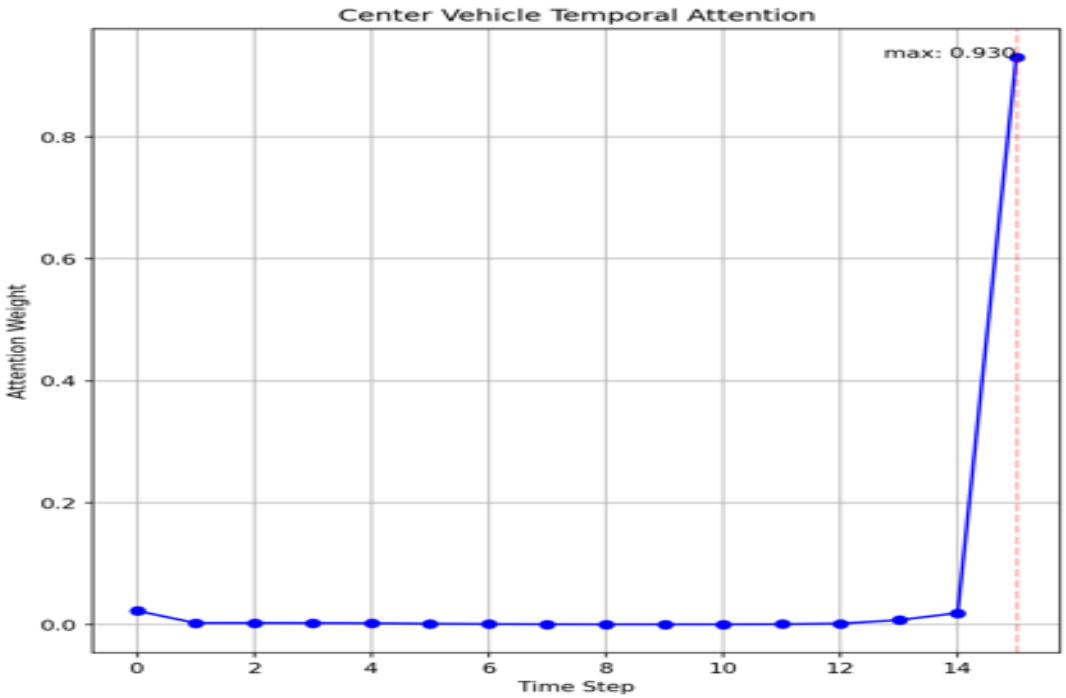

Figure 3. Time attention weight.

4. Analysis

4.1. Comparative Analysis of STA-GRU and GRU Models

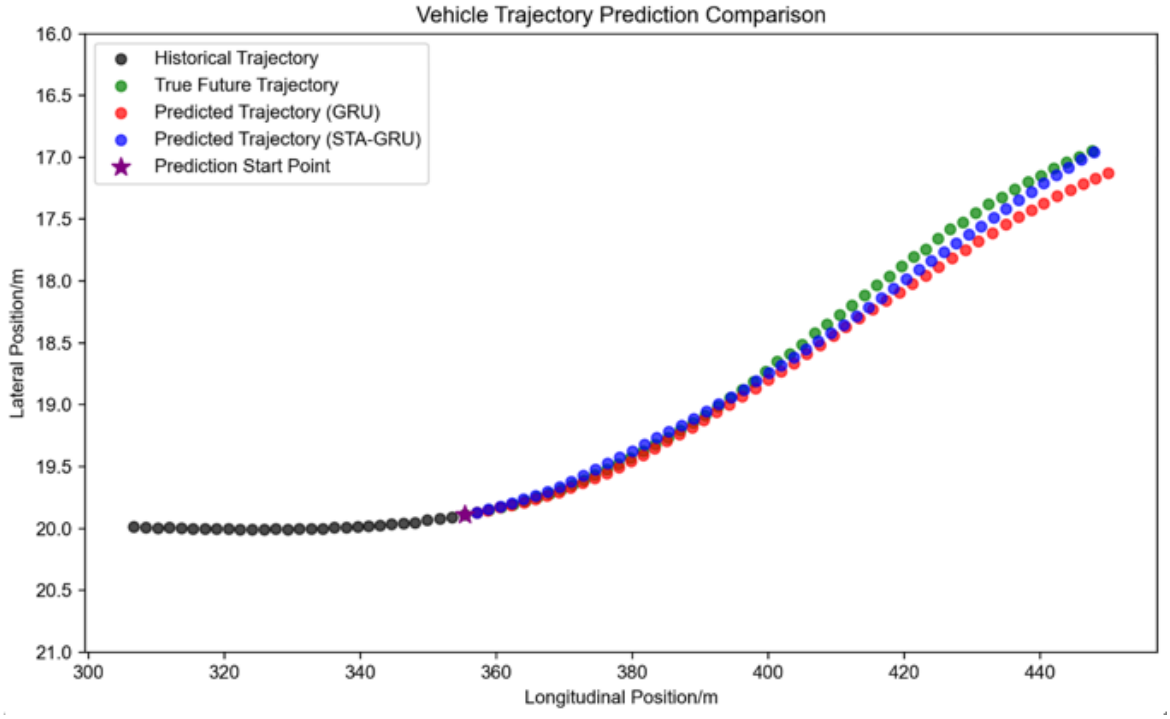

Figure 4. Comparison diagram of model trajectory prediction.

Figure 4 presents the vehicle trajectory prediction performance of different models. The horizontal and vertical axes represent lateral and longitudinal positions respectively, with the legend indicating various trajectory types and prediction starting points. The results demonstrate that STA-GRU's predicted trajectories align more closely with the ground truth compared to conventional GRU, which exhibits significantly larger prediction errors. This comparison visually validates the model's optimization effectiveness, showing STA-GRU's superior capability in long-term prediction accuracy.The improved performance confirms that by effectively integrating spatiotemporal information with dynamic attention mechanisms, the model successfully handles complex interaction patterns in traffic scenarios, leading to enhanced trajectory prediction precision. The visualization clearly illustrates how STA-GRU maintains better prediction stability over extended horizons where traditional methods tend to accumulate errors.

4.2. Performance comparison of different models

In order to further evaluate the performance of the proposed model, N-GRU, SA-GRU, CS-LSTM, GAIL-GRU were selected respectively to compare with the STA-GRU model proposed in this study, and the evaluation indicators were obtained as shown in Table 2 and Figure 5[13].

$$RMSE = \frac{1}{N}\sum_{i=1}^{N}\sqrt{(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2} \tag{8}$$

Table 2. Comparison result of different models.

Different models	Displacement Errors/m				
	1st	2st	3st	4st	5st
N-GRU	0.1345	0.2458	0.3642	0.4823	0.6527
SA-GRU	0.1024	0.2037	0.3187	0.4463	0.5741
CS-LSTM	0.1032	0.2029	0.3173	0.4459	0.5732
GAIL-GRU	0.1078	0.2105	0.3252	0.4476	0.5745
STA-GRU	0.0997	0.2011	0.3148	0.4363	0.5623

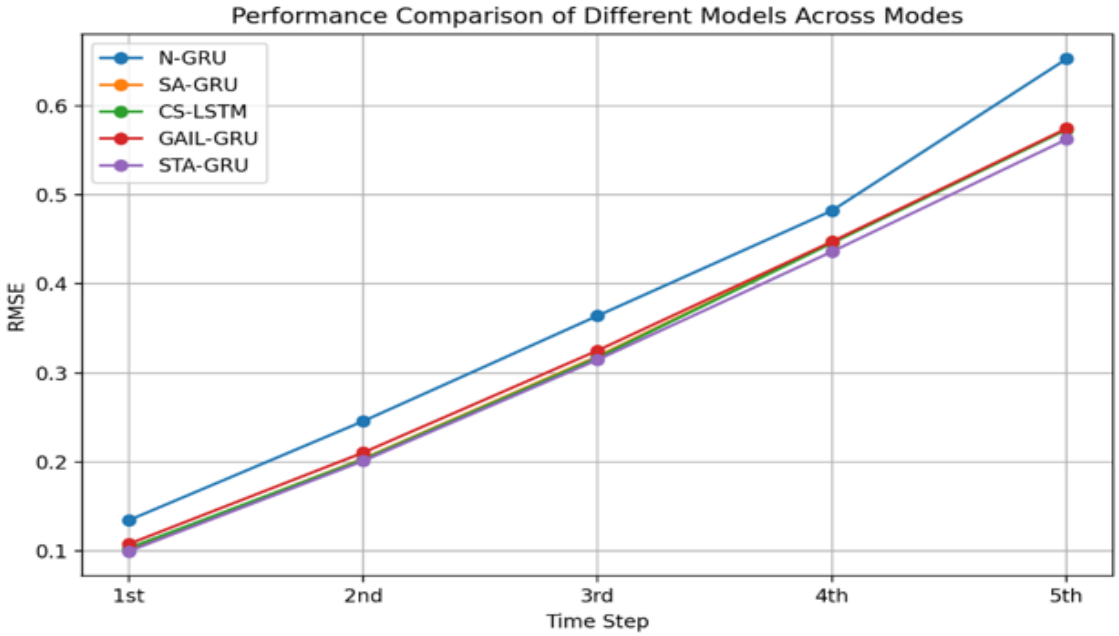

Figure 5. Result of different models.

The RMSE values of the predicted trajectories incorporating spatiotemporal features are lower than those of other existing state-of-the-art models. The results demonstrate that by integrating spatiotemporal information and dynamic attention mechanisms, the model can effectively handle complex interaction relationships in traffic scenarios, improving the accuracy and robustness of trajectory prediction.

5. Conclusions

This study proposes a hierarchical attention-based vehicle trajectory prediction model that simultaneously captures temporal motion dependencies and spatial inter-vehicle interactions through a dual-level attention architecture, effectively improving prediction accuracy in complex traffic scenarios. For social context encoding, a grid-based method structurally represents neighboring vehicles' positional information, combined with masked scatter operations to efficiently process sparse interaction data while significantly reducing computational complexity. The model employs an end-to-end training paradigm and innovatively incorporates a dual-mode output mechanism featuring both single-step decoding (for latency-sensitive applications) and iterative decoding (for enhanced long-term dependency modeling), providing flexible adaptation to diverse prediction requirements. Experimental results demonstrate that the proposed STA-GRU model outperforms existing benchmarks in highway trajectory prediction tasks.

However, current experimental validation is limited to structured road environments (such as controlled-access highways), and its applicability to unstructured traffic scenarios (including but not limited to intersections, crosswalks, roundabouts and other complex topological environments) has not been fully verified. Future research will focus on designing more adaptive attention mechanisms for driving behaviors specific to unstructured scenarios (such as yielding, obstacle avoidance, and traffic signal response), and exploring upgrades from the current grid-based spatial representation to graph neural network architectures to more flexibly handle dynamically changing traffic participant relationships.

References

[1] Yang Chenxi, Zhuang Xufei, Chen Junnan, et al. Research review of bus travel trajectory prediction based on Deep Learning [J]. Computer Engineering and Applications,2024,60(09):65-78

[2] Zhang Chihao, Shi Wei, Qin Yinan, et al. Research on Longitudinal Control of Intelligent Vehicles Based on Model Predictive Control [J]. Special Purpose Vehicle, 2023(5): 20-22.

[3] Fei Xiansong. Research on Vehicle Handling Stability with Combined Active steering and yaw Moment Control [D]. Nanjing University of Aeronautics and Astronautics, 2007.

[4] BengioY, Lamblin P, Popovici D ,et al.Greedy layer-wise training of deep networks[C]Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia Canada, December 4-7, 2006.DBLP, 2007.

[5] Wang Mengxi, Cai Yingfeng, Wang Hai, et al. Vehicle Trajectory Prediction Method Based on Graph Convolutional Interaction Network [J]. Automotive Engineering, 2024, 46(10): 1863-1872.

[6] Shen Y , Li W , Lin M .Autonomous Driving Via Context-aware Multi-sensor Perception and Enhanced Inverse Reinforcement Learning[J]. 2020. (5):20-22.

[7] SHENG Z,XU Y, XUE S, et al. Graph-based spatial-temporal convolutional network for vehicle trajectory prediction in autonomous driving [J]. IEEE Transactions on Intelligent Transportation System, 2022, 23 (10): 17654-17665.

[8] Li Jiufa, Zou Bowen, Ren Yue. Research on Vehicle Trajectory Prediction Algorithm Considering Vehicle-Road Interaction [J]. Journal of Mechanical Engineering, 2024, 60(10): 76-85.

[9] GAO Z H et al. The method of probabilistic multi-model expected trajectory prediction based on LSTM [J]. Automotive Engineering, 2023, 45(07): 1145-1152+1162.

[10] Deo N , Trivedi M M .Convolutional Social Pooling for Vehicle Trajectory Prediction[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).IEEE, 2018, pp. 1468–1476.

[11] Peng Ziran, Wang Shunhao, Xiao Shengping, et al. A Cyclic-Gated Model for Accurately Estimating the State of Charge (SOC) and State of Health (SOH) of Electric Vehicle Battery Cells [J]. Journal of Electronic Measurement and Instrumentation, 2024, 38(9): 11-23.

[12] Tao Yanyun, Shen Zhiwei, Wang Xiang, et al. A Multi-point Regression Prediction Model of Convolutional Neural Network for Traffic Flow Prediction: CN201810866657.4 [P]. CN108830430A [2025-04-13].

[13] Du Qian. Research on Autonomous Driving Decision-making Method Based on Risk Assessment and Deep Reinforcement Learning [D]. Qilu University of Technology, 2024.