Trajectory Prediction for Autonomous Driving of New Energy Vehicles based on Graph Transformer Networks

Chao Ye, Dongjing Chen*

AmazingX Academy, Shenzhen, China

*Corresponding author: sysucdj@amazingx.org

Abstract

Accurate trajectory prediction is crucial for ensuring the safe and efficient operation of autonomous vehicles, especially in complex and dynamic traffic environments. In this paper, we propose a novel approach for trajectory prediction in the context of new energy vehicles (NEVs) by combining Graph Neural Networks (GNNs) and Transformer models. Our model, Graph Transformer Networks (GTNet), effectively captures the spatial-temporal dependencies and interactions among multiple agents, including other vehicles, pedestrians, and traffic infrastructure, to predict future trajectories. Unlike traditional methods that focus only on local interactions, our approach integrates global context through the Transformer model to better handle long-range dependencies. The GNN component captures the underlying structure of the traffic environment, such as road networks and vehicle relationships, while the Transformer model leverages its attention mechanism to focus on relevant information across time and space. This combination allows for a more accurate prediction of future vehicle trajectories, especially in scenarios where the motion of surrounding agents is highly interdependent. We evaluate our approach on real-world datasets and demonstrate its superior performance compared to baseline methods, achieving improved prediction accuracy for both short- and long-term trajectory forecasts. Our results highlight the effectiveness of integrating Graph Neural Networks with Transformer models for enhancing trajectory prediction in autonomous driving, particularly for new energy vehicles navigating dynamic traffic scenarios.

Keywords

Deep Learning, new energy vehicles, autonomous driving, trajectory prediction.

1. Introduction

The rapid development of autonomous driving technologies has led to significant advancements in vehicle navigation, control, and decision-making systems. One of the most critical tasks in autonomous driving is trajectory prediction, which involves forecasting the future movement of surrounding vehicles, pedestrians, and other traffic agents. Accurate trajectory prediction is essential for ensuring the safety, efficiency, and comfort of autonomous vehicles, as it allows them to anticipate the behavior of other road users and make informed decisions. However, the task of predicting future trajectories is inherently challenging due to the dynamic, complex, and often unpredictable nature of real-world traffic environments.

In particular, the increasing adoption of new energy vehicles (NEVs)-including electric vehicles (EVs) and hybrid vehicles-presents unique challenges and opportunities for trajectory prediction. NEVs are characterized by distinct operational dynamics, such as energy consumption patterns and driving behavior, which can influence their interaction with other vehicles on the road. Moreover, the rise of NEVs is accompanied by the growing importance of optimizing the use of traffic resources, reducing energy consumption, and improving the overall

driving experience. Therefore, it is crucial to develop trajectory prediction models that not only account for traditional road users but also consider the specific characteristics of NEVs.

Conventional trajectory prediction approaches predominantly model traffic agents as isolated entities lacking spatial interdependencies, employing classical mathematical frameworks such as kinematic/dynamic models (Brännström et al.[1]) and Gaussian Processes (Rasmussen[2]). These methods exhibit constrained capabilities in interpreting complex scenarios and generating long-term predictions. The advent of deep neural networks has shifted research focus toward spatiotemporal feature extraction through neural architectures (Alahi et al.[3]; Huang et al.[4]; Ivanovic & Pavone[5]; Mohamed et al.[6]).

Existing trajectory prediction models largely rely on traditional deep learning approaches, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs), which focus on modeling the temporal dependencies of moving agents. However, these models often fail to capture the complex spatial and temporal relationships between multiple agents in dynamic traffic scenarios [7-13]. Moreover, they tend to overlook the crucial spatial structure of the road network and interaction mechanisms between agents. Some recent approaches have started integrating Graph Neural Networks (GNNs) to model the spatial dependencies between vehicles, pedestrians, and road networks [14]. However, these methods typically do not fully leverage the global context and long-range dependencies that are often critical for accurate trajectory prediction in complex traffic environments [15].

To address these limitations, we propose a novel framework for trajectory prediction based on Graph Transformer Networks (GTNet), which combines the strengths of both Graph Neural Networks and Transformer models [16-21]. Our model leverages the ability of GNNs to capture the spatial relationships between agents and traffic infrastructure, while the Transformer model enhances the learning of long-range temporal dependencies through its attention mechanism. Specifically, GTNet integrates spatial-temporal modeling to learn interactions between agents, road networks, and traffic rules, improving prediction accuracy in dynamic and heterogeneous traffic conditions.

The main contributions of this paper are as follows:

1) Graph Transformer Networks: We propose a novel trajectory prediction model that combines Graph Neural Networks and Transformer models. This integration allows for the effective modeling of spatial and temporal dependencies in a unified framework, enabling more accurate predictions of vehicle trajectories.

2) Modeling of Spatial-Temporal Interactions: We introduce a spatial-temporal attention mechanism that enables the model to learn and prioritize critical interactions between multiple agents in dynamic environments. This mechanism enhances the model's ability to predict future trajectories in scenarios involving complex agent interactions and traffic conditions.

3) Incorporation of New Energy Vehicle Dynamics: Our framework is specifically designed to handle the unique characteristics of new energy vehicles, considering their distinct driving behaviors and energy consumption patterns. This makes our approach well-suited for the growing fleet of NEVs on the road.

4) Superior Performance: Through extensive experiments on real-world trajectory datasets, we demonstrate that our approach outperforms existing state-of-the-art methods, achieving superior prediction accuracy for both short- and long-term trajectory forecasts. We also show that GTNet provides better generalization to various traffic scenarios compared to traditional approaches.

In summary, this paper introduces an innovative approach to trajectory prediction for autonomous driving, combining the power of Graph Neural Networks and Transformer models. Our model not only improves the accuracy and robustness of trajectory forecasting but also addresses the challenges specific to the emerging domain of new energy vehicles.

2. Problem and Methodology

2.1. Problem in Trajectory Prediction

A trajectory prediction task can be represented as a given in a scene Set a prediction time range T_{h}, the agent's historical trajectory X and Environmental information M, predicting its range in the future period of time T_f orbit Trace. The input to the model is generally the historical trajectory X, which is determined by the target vehicle S_0 and the physical state ($S_1, S_2, S_3, \ldots, S_n$) composition of surrounding vehicles. The physical state of the vehicle consists of speed, acceleration, steering angle, and the current environmental information M. Typically, it is assumed that there are N traffic participants (such as surrounding vehicles, cyclists, and pedestrians) around the target vehicle.

$$X = \{p^1, p^2, \ldots, p^{Th}\} \tag{1}$$

Here, $p^t (t \in 1, 2, \ldots, T_h)$ represents the historical trajectory within the time range t, $p = (x_t, y_t, M, S_n)$, where x_t and y_t denote the coordinates of the current vehicle, and p_{T_h} represents the state of the current target vehicle.

In the trajectory prediction task, the model's output is mostly the predicted trajectory Y for the nextT_f times steps:

$$Y = \{p^{T_h+1}, p^{T_h+2}, \ldots, p^{T_h+T_f}\} \tag{2}$$

2.2. Trajectory Prediction Networks

The main challenges in trajectory prediction stem from environmental uncertainty, the multimodality of generated trajectories, and the interpretability of the model. This section will discuss these issues in detail.

In trajectory prediction, the future behavior of surrounding traffic participants is difficult to predict, and their future trajectories are multimodal (as shown in Fig.1). Therefore, the predicted future trajectory is not unique; that is, under the same historical trajectory conditions, there are multiple possible future trajectories for the vehicle. As a result, the prediction model needs to account for uncertainty and predict multiple possible trajectories for other vehicles. Multimodal trajectory prediction involves the model assigning corresponding confidence levels to each predicted trajectory and generating multiple future trajectories for each traffic participant.

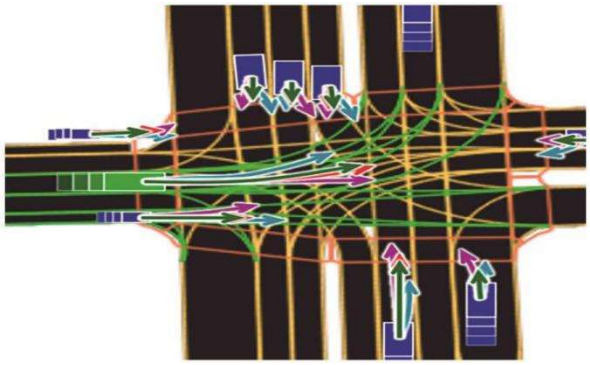

Fig. 1 Multimodal trajectory prediction visualization results.

Early autonomous driving trajectory prediction models were mainly unimodal, where the prediction algorithm would output a single, determined trajectory given a specified historical trajectory. However, unimodal trajectory prediction fails to capture the possibilities in complex scenarios and does not guarantee prediction accuracy or safety when integrated with the planning module. Deo et al.[9] used a Mixture Density Network (MDN) for highway trajectory prediction, where the model takes the historical trajectories of surrounding vehicles and the topology of the highway as input, and divides confidence values into 6 maneuver categories. Each driving action was defined as a prediction mode, and each mode was associated with a Gaussian component, thus outputting a multimodal distribution for future trajectories.

However, trajectory data is abundant and complex, and the predefined actions in reference cannot guarantee trajectory diversity. Therefore, in subsequent research, scholars have focused on generating multimodal trajectory predictions based on uncertainty, aiming to reflect real driving scenarios. This paper will focus on deep learning-based multimodal trajectory prediction models. The schematic diagram of the working mechanism of trajectory prediction is displayed in Fig. 2.

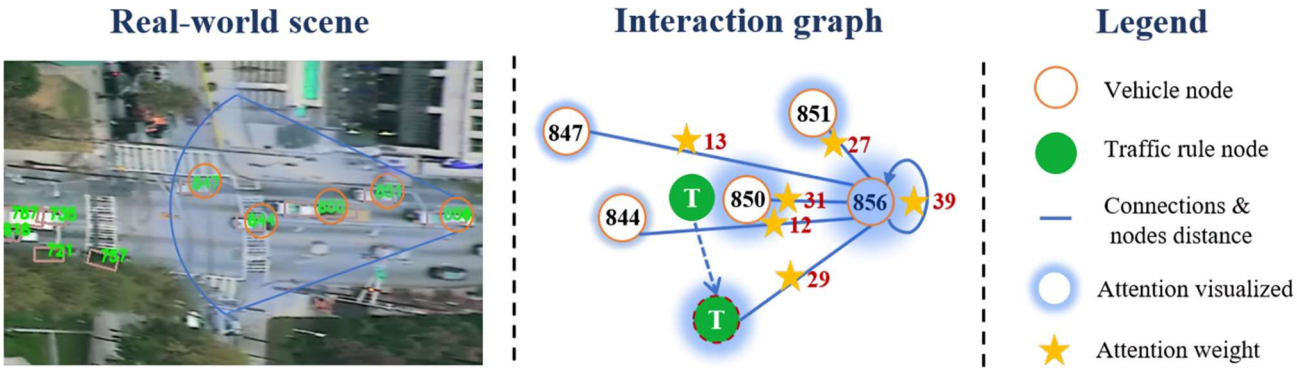

Fig. 2 The schematic diagram of the working mechanism of trajectory prediction.

2.3. Self-attention in Transformer

The architectural distinction between Transformers and LSTMs lies in their computational paradigms: Transformers leverage parallelized self-attention operations, whereas LSTMs sequentially aggregate temporal dependencies through recurrent gate mechanisms. The self-attention computation initiates with three trainable projections: query $q_i \in R^{d_q}$, *key* $k_i \in R^{d_k}$, key, and value $v_i \in R^{d_v}$ vectors derived from input embeddings $w_i \in R^e, i \in (1,n)$ (where n denotes sequence length).

The attention score is computed via $a_{ij} = q_i \cdot k_j^T, \forall i,j \in (1,n)$ (dot product of query and transposed key vectors), followed by softmax normalization to generate probabilistic weights a_{ij}. Contextualized representations are then synthesized by summing value vectors v_j scaled by their corresponding normalized weights. Formally, this multi-step process yields:

$$Attention(Q,K,V) = soft\max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

Where d_k is the dimension of each query. The division by $\sqrt{d_k}$ is used to increase gradients stability.

By adding multi-head attention mechanism, we can further improve the performance of self-attention. It gives multiple representation sub-spaces for self-attention and enables the model to jointly deal with information from varied sub-spaces at separate positions.

2.4. Overview of GTNet Model

As depicted in Fig. 3, the proposed framework adopts a hierarchical encoder-decoder structure integrating three core components: a spatio-temporal Transformer encoder, a temporal Transformer encoder, and a temporal Transformer decoder. To extract comprehensive motion patterns, historical trajectories of traffic agents are first projected into an expanded latent space via fully connected layers. Spatial interdependencies within individual frames are captured through self-attention mechanisms, while temporal coherence across frames is modeled using temporal convolutional networks (TCNs). The architecture innovatively couples spatial and temporal modeling by interleaving these two modules within unified spatio-temporal Transformer layers.

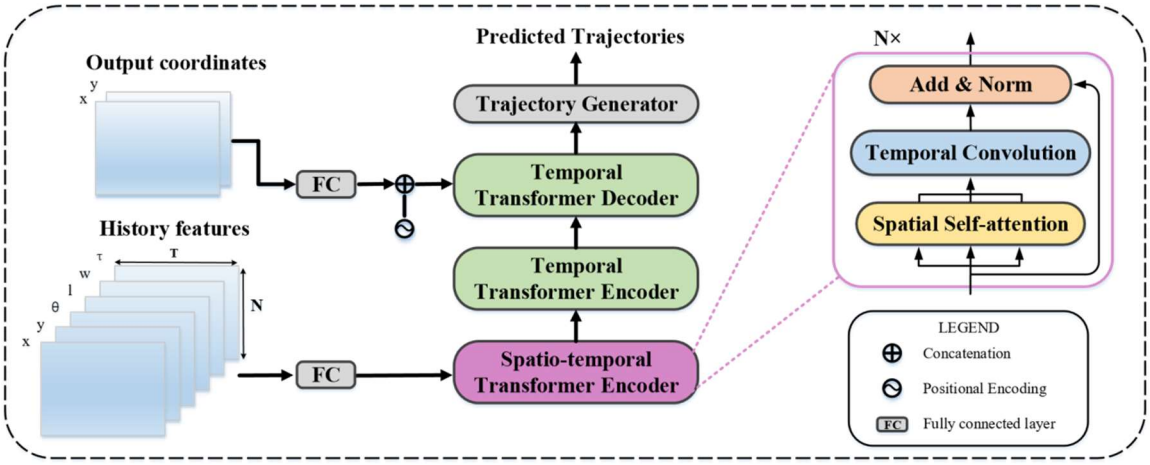

Fig. 3 Overview of GTNet.

2.5. Spatio-temporal Transformer

Reinterpreted through a graph-based paradigm, this component operates on spatial-edge representations within spatio-temporal graphs. Adopting a message-passing framework, it processes node relationships at each timestep t. For node i, query ($q_i^t \in R^{d_q}, keyk_i^t \in R^{d_k}, h_i^t \in R^C$) and value ($h_i^t \in R^C$) vectors are derived through linear projections:

$$h_i^t \in R^C \xrightarrow{W_q} q_i^t \in R^{d_q},\ key\ k_i^t \in R^{d_k},\ h_i^t \in R^C \xrightarrow{W_v} v_i^t \in R^{d_v} \tag{4}$$

Concurrently, temporal dependencies are captured through dilated convolutional operations along the temporal axis.

This dual-path architecture enables simultaneous learning of spatial correlations and temporal dynamics through alternating parameterized transformations. The hierarchical feature fusion process is mathematically expressed as:

$$q_i^t = W_q \cdot h_i^t,\ \ v_i^t = W_v \cdot h_i^t \tag{5}$$

Where $W_q \in R^{C\times d_q}$, $W_k \in R^{C\times d_k}$, $W_k \in R^{C\times d_v}$.Attention score between node i and $j \in (1, V)$ is then obtained by applying scaled dot-product between q_i^t and k_j^t representing the spatial-edge massage m_{ij}^t, j send from j to i.

$$m_{ij}^t = q_i^t \cdot k_j^{tT} \tag{6}$$

The framework aggregates interaction signals from neighboring nodes through normalized spatial edge weights. Each node's attention representation is generated by weighted summation of normalized messages across its spatial neighborhood, ensuring focused information propagation within localized contexts.

$$head_i^t = \sum soft\max\left(\frac{m_{ij}^t}{d_k}\right)v_j \tag{7}$$

A multi-branch architecture enables parallel feature extraction through h independent parameterizations. The resulting heterogeneous attention representations undergo concatenation followed by linear projection, achieving feature fusion and dimensionality reduction through trainable transformations.

$$MulltiHead_i^t = W_0 \cdot concat([head_{i0}^t, ..., head_{ih}^t]) \tag{8}$$

The temporal sub-module applies dilated convolutions along the temporal dimension to capture evolving patterns in trajectory sequences. Operating on spatiotemporal tensors, it employs specialized 1D kernels to model temporal dependencies while preserving spatial independence across nodes.

$$output_i = Conv_{K\times 1}(MultiHead_i^t) \tag{9}$$

The architecture incorporates residual connections and layer normalization modules after temporal operations. This design ensures stable gradient propagation during training while maintaining structural consistency in spatiotemporal graph processing through skip-connection enhanced feature refinement.

2.6. Temporal Transformer

Unlike its spatial counterpart that aggregates cross-node interactions, the temporal self-attention mechanism focuses on intra-node evolutionary patterns across sequential timesteps. For node i, this process isolates temporal dependencies through node-specific query-key-value computations.

The temporal self-attention for node i represented as:

$$MultiHead_i = W_u \cdot concat([head_{i0}, ..., head_{ih}]) \tag{10}$$

$$head_i = soft\max\left(\frac{Q_i \cdot K_i^T}{\sqrt{d_k}}\right)V_i \tag{11}$$

The temporal Transformer architecture employs query (Q_i), key (K_i), and value (V_i) matrices derived from node-specific input embeddings to establish localized temporal dependencies. Crucially, it replaces conventional dense feedforward layers with depthwise-separable convolutions in secondary sublayers, enhancing modeling precision through decoupled spatial-channel filtering. The decoder incorporates positional encoding mechanisms to embed sequential trajectory relationships, where absolute/relative position information is fused with output embeddings through additive transformations. This architectural innovation enables simultaneous preservation of temporal ordering semantics and adaptive feature refinement across prediction horizons, contrasting with standard Transformers' position-agnostic processing. The integrated framework achieves optimized parameter efficiency while maintaining temporal coherence through node-isolated attention computations and progressive trajectory conditioning.

$$PE(pos, 2i) = \sin\left(pos/10000^{2i/d_{model}}\right) \tag{12}$$

$$PE(pos, 2i) = \cos\left(pos/10000^{2i/d_{model}}\right) \tag{13}$$

Where pos is the position, i is the dimension and d_{model} the total dimensions of the output embeddings.

The decoder architecture introduces critical modifications to ensure causal temporal relationships: 1) A masked self-attention mechanism prevents temporal leakage by restricting attention to preceding timesteps, 2) Depthwise-separable convolutions maintain computational efficiency from encoder layers, and 3) A dedicated cross-attention sublayer bridges encoder-decoder information flow through multi-head attention over encoded representations. This triad of components enables autoregressive refinement while preserving temporal dependency integrity - the masked attention enforces prediction causality, the separable convolutions extend temporal receptive fields, and the cross-attention synthesizes global context from encoder states. The integrated design contrasts with encoder operations by implementing strict temporal masking and explicit encoder feature fusion, achieving coordinated trajectory prediction through constrained attention and multi-source feature integration.

3. Experiments

3.1. Dataset

Vehicle trajectories play a crucial role in traffic flow analysis, driving behavior analysis, decision model design, training, testing, and path planning. Therefore, choosing an appropriate dataset is essential for improving model performance. In the early NGSIM dataset, traffic participants were only vehicles in highway scenarios, and the data only contained historical trajectories. Using such datasets as input for trajectory prediction algorithms is clearly incomplete. The low environmental complexity made it difficult for early trajectory prediction methods to generate multimodal predicted trajectories. Later datasets included more environmental information, such as high-definition (HD) maps, multiple types of traffic participants (pedestrians, cyclists), and various scenarios (intersections, roundabouts), allowing recent models to achieve better prediction accuracy and generalization in complex, high-dynamic, and uncertain scenarios. This section will introduce the commonly used datasets in trajectory prediction tasks and list classic algorithms that use these datasets (It is displayed in Table 1).

Table 1. Dataset for autonomous driving trajectory prediction tasks.

Dataset	Year	Traffic participants	Sensor	Scene	Data Content	Algorithms based on this dataset
NuScenes[12]	2020	vehicles, pedestrians, bicycles	radar, cameras	city	Driving tracks, HD maps	MHA-JAM
Lyft[13]	2020	vehicles, pedestrians, bicycles	radar, cameras	city	Driving tracks, HD maps	Gatformer
Waymo[14]	2020	vehicles, pedestrians, bicycles	radar, cameras	city	Driving tracks, HD maps	Wayformer, DesneTNT
Argoverse[15][16]	2019	vehicles	radar, cameras	Intersection without signal lights	Driving tracks, HD maps	mmTransformer, HOME, Traj-Former, VertorNet
INTERACTION[17]	2019	vehicles, pedestrians	drones, cameras	highway, Intersection without signal lights, Intersection with signal lights, roundabout	Driving tracks, HD maps	GOHOME, THOMAS, TNT
HighD[18]	2018	vehicles	drones	city intersection	driving trajectory, lane	MHA-LSTM
Apolloscape[19]	2018	vehicles, pedestrians, bicycles	radar, cameras	city, highway	driving trajectory	AI-TP, S2TNet, GRIP++
KITTI[20]	2013	pedestrians, bicycles	radar, cameras	city, highway	images, point clouds	DESIRE
NGSIM	2006	vehicles	cameras	highway intersection	driving trajectory, lane	Maneuver based on LSTM, GRIP++

3.2. Evaluation Metrics

Assessment employs displacement-based metrics:

1) Average Displacement Error (ADE): Temporal mean of Euclidean distances between predicted and ground-truth positions

2) Final Displacement Error (FDE): Terminal timestep-specific displacement

3) Scale-Weighted Metrics (WSADE/WSFDE): Agent-type adjusted variants addressing motion magnitude disparities across traffic participants

This dual-focus metric suite quantifies both holistic trajectory adherence and critical endpoint precision, while accommodating inherent scale variations through dynamic weighting mechanisms.

$$WSADE = D_v \cdot ADE_p + D_b \cdot ADE_b \tag{14}$$

$$WSFDE = D_v \cdot FDE_v + D_b \cdot FDE_b \tag{15}$$

The weighting coefficients in evaluation metrics are derived from inverse average velocity measurements across distinct agent categories (vehicles, pedestrians, cyclists) within the dataset. This velocity-dependent scaling mechanism proportionally weights prediction errors according to characteristic motion patterns of different traffic participants, ensuring equitable performance assessment across agents with inherently divergent speed profiles.

Where D_v = 0.20, D_p = 0.58, and D_b = 0.22.

4. Results

4.1. Quantitative Results and Analyses

A critical insight emerges from the outperformance of the velocity-based CV model over advanced learning methods like STAR, suggesting limitations of homogeneous architectures in dense urban contexts. Conversely, our method excels in heterogeneous environments through adaptive spatiotemporal modeling, effectively addressing multi-agent interactions and motion diversity while maintaining computational efficiency. These results validate the framework's capability to balance precision and generalizability across complex traffic scenarios.

4.2. Qualitative Results and Analyses

Qualitative visualizations (Fig. 4) demonstrate three key capabilities of the framework:

1) Multi-Horizon Predictive Competence - Accurately forecasting 3s trajectories from 3s historical observations, particularly excelling in acute steering maneuvers (Fig. 4a-b) with superior error propagation characteristics compared to GRIP++ in extended horizons (Fig. 4c-d).

2) Interaction-Aware Modeling - Precisely resolving counterflow agent interactions (Fig. 4e-f upper section) where conventional methods exhibit trajectory divergence.

3) Stationary State Recognition - Correctly identifying near-zero velocity states during deceleration phases (Fig. 4e-f lower section), overcoming baseline limitations in motion cessation detection.

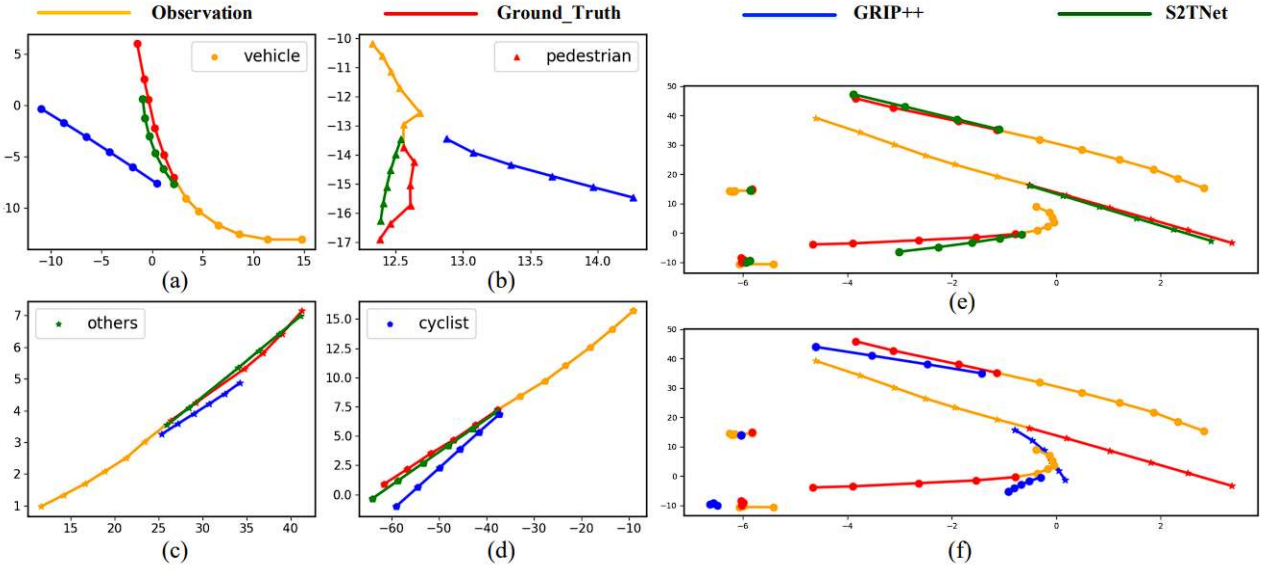

Fig. 4 Visualized Prediction Results in heterogeneous and dense traffic.

The visual evidence confirms the framework's capacity to balance long-term prediction fidelity with nuanced interaction modeling, particularly in scenarios requiring simultaneous handling

of kinematic constraints and multi-agent dynamics. Enhanced temporal coherence and interaction sensitivity are achieved through dedicated spatiotemporal attention mechanisms, enabling robust performance across heterogeneous motion patterns from free-flowing to static operational conditions.

5. Conclusion and Future Work

In this paper, we proposed Graph Transformer Networks, a novel framework for trajectory prediction in autonomous driving, with a particular focus on new energy vehicles (NEVs). Our approach integrates Graph Neural Networks and Transformer models to effectively capture the complex spatial-temporal dependencies between traffic agents, road infrastructure, and environmental factors. By leveraging the spatial modeling capabilities of GNNs and the long-range temporal dependency learning of Transformers, GTNet offers a robust solution to the challenge of predicting the future trajectories of vehicles in dynamic and heterogeneous traffic environments.

We demonstrated that GTNet outperforms traditional trajectory prediction methods, achieving superior performance in terms of both short- and long-term prediction accuracy. The model's ability to learn critical interactions between multiple agents and prioritize relevant information through its spatial-temporal attention mechanism has proven essential in improving prediction reliability, especially in complex urban driving scenarios. Furthermore, our framework's incorporation of NEV dynamics ensures that it addresses the specific characteristics and behaviors of new energy vehicles, a growing segment of the automotive market.

Future work could focus on several avenues as follows:

1) Integration of Multi-Modal Data: One potential direction is to incorporate additional data sources, such as LiDAR, radar, or camera input, to further enhance the model's ability to perceive the environment. Multi-modal data could provide richer information for trajectory prediction, especially in environments with limited visibility or adverse weather conditions.

2) Scalability and Real-Time Application: Another area for future work is optimizing the model for real-time prediction. Currently, the computational complexity of GNNs and Transformers can limit their scalability, particularly in large-scale traffic scenarios. Efficient model architectures, such as knowledge distillation or model pruning, could be explored to enhance real-time performance while maintaining accuracy.

3) Vehicle-to-Infrastructure (V2I) Communication: In the future, we plan to explore the integration of Vehicle-to-Infrastructure (V2I) communication into the trajectory prediction framework. By incorporating real-time traffic signals, road conditions, and other infrastructure-related information, we could further improve the accuracy of trajectory predictions, especially in highly dynamic environments such as intersections or urban streets.

4) Adversarial and Uncertainty Modeling: Autonomous vehicles often operate in uncertain environments, where interactions with other traffic agents can be unpredictable. Future research could focus on integrating adversarial models or uncertainty estimation techniques to better handle the inherent uncertainties in trajectory prediction. This would allow the model to not only predict a single trajectory but also account for multiple potential outcomes, providing more robust decision-making capabilities for autonomous vehicles.

5) Integration with Autonomous Control Systems: Another exciting direction for future work is the integration of GTNet with autonomous vehicle control systems. By linking trajectory prediction directly with motion planning and control tasks, we could develop end-to-end systems that seamlessly transition from prediction to action, improving the overall decision-making process and vehicle autonomy.

In summary, the proposed GTNet framework represents a significant step forward in trajectory prediction for autonomous driving, particularly in the context of new energy vehicles. By effectively combining Graph Neural Networks and Transformer models, we have developed a robust and accurate method for predicting the future movement of traffic agents. With further advancements, especially in the areas of real-time performance, multi-modal data integration, and uncertainty modeling, we believe that GTNet can play a key role in the future of autonomous driving and the broader landscape of intelligent transportation systems.

References

[1] Brännström M, Coelingh E, Sjöberg J. Model-based threat assessment for avoiding arbitrary vehicle collisions[J]. IEEE Transactions on Intelligent Transportation Systems, 2010, 11(3): 658-669.

[2] Rasmussen C E. Gaussian processes in machine learning[M]//Summer school on machine learning. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003: 63-71.

[3] Alahi A, Goel K, Ramanathan V, et al. Social lstm: Human trajectory prediction in crowded spaces[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 961-971.

[4] Huang Y, Bi H, Li Z, et al. Stgat: Modeling spatial-temporal interactions for human trajectory prediction[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6272-6281.

[5] Ivanovic B, Pavone M. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 2375-2384.

[6] Mohamed A, Qian K, Elhoseiny M, et al. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 14424-14432.

[7] Giuliari F, Hasan I, Cristani M, et al. Transformer networks for trajectory forecasting[C]//2020 25th international conference on pattern recognition (ICPR). IEEE, 2021: 10335-10342.

[8] Vaswani A. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017.

[9] Deo N, Trivedi M M. Convolutional social pooling for vehicle trajectory prediction[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018: 1468-1476.

[10] Li X, Ying X, Chuah M C. Grip: Graph-based interaction-aware trajectory prediction[C]//2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, 2019: 3960-3966.

[11] Yu C, Ma X, Ren J, et al. Spatio-temporal graph transformer networks for pedestrian trajectory prediction[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16. Springer International Publishing, 2020: 507-523.

[12] Caesar H, Bankiti V, Lang A H, et al. nuscenes: A multimodal dataset for autonomous driving[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11621-11631.

[13] Houston J, Zuidhof G, Bergamini L, et al. One thousand and one hours: Self-driving motion prediction dataset[C]//Conference on Robot Learning. PMLR, 2021: 409-418.

[14] Sun P, Kretzschmar H, Dotiwalla X, et al. Scalability in perception for autonomous driving: Waymo open dataset[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 2446-2454.

[15] Chang M F, Lambert J, Sangkloy P, et al. Argoverse: 3d tracking and forecasting with rich maps[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 8748-8757.

[16] Wilson B, Qi W, Agarwal T, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting[J]. arXiv preprint arXiv:2301.00493, 2023.

[17] Zhan W, Sun L, Wang D, et al. Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps[J]. arXiv preprint arXiv:1910.03088, 2019.

[18] Krajewski R, Bock J, Kloeker L, et al. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems[C]//2018 21st international conference on intelligent transportation systems (ITSC). IEEE, 2018: 2118-2125.

[19] Huang X, Cheng X, Geng Q, et al. The apolloscape dataset for autonomous driving[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018: 954-960.

[20] Geiger A, Lenz P, Stiller C, et al. Vision meets robotics: The kitti dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.

[21] Dosovitskiy A, Ros G, Codevilla F, et al. CARLA: An open urban driving simulator[C]//Conference on robot learning. PMLR, 2017: 1-16.