2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Human Observation-Inspired Trajectory Prediction for Autonomous Driving in Mixed-Autonomy Traffic Environments

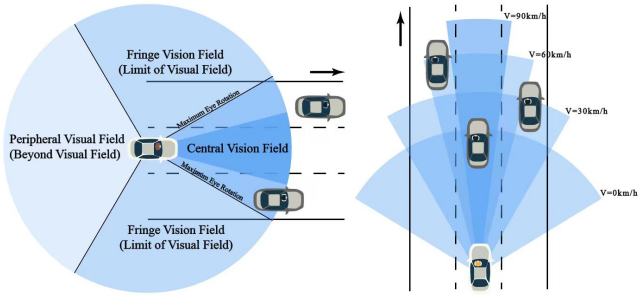
Haicheng Liao*[1], Shangqian Liu*[1], Yongkang Li[2], Zhenning Li†[3], Chengyue Wang[4], Yunjian Li[5], Shengbo Eben Li[6], and Chengzhong Xu[1]

***Abstract*— In the burgeoning field of autonomous vehicles (AVs), trajectory prediction remains a formidable challenge, especially in mixed autonomy environments. Traditional approaches often rely on computational methods such as time-series analysis. Our research diverges significantly by adopting an interdisciplinary approach that integrates principles of human cognition and observational behavior into trajectory prediction models for AVs. We introduce a novel "adaptive visual sector" mechanism that mimics the dynamic allocation of attention human drivers exhibit based on factors like spatial orientation, proximity, and driving speed. Additionally, we develop a "dynamic traffic graph" using Convolutional Neural Networks (CNN) and Graph Attention Networks (GAT) to capture spatio-temporal dependencies among agents. Benchmark tests on the NGSIM, HighD, and MoCAD datasets reveal that our model (GAVA) outperforms state-of-the-art baselines by at least 15.2%, 19.4%, and 12.0%, respectively. Our findings underscore the potential of leveraging human cognition principles to enhance the proficiency and adaptability of trajectory prediction algorithms in AVs.**

I. INTRODUCTION

In the pursuit of developing fully autonomous vehicles (AVs), one of the most daunting challenges is to accurately predict vehicle trajectories with the same level of proficiency as human drivers, especially in a mixed autonomy environment where AVs coexist with human-driven vehicles [1]. While existing research largely focuses on computational methods grounded in time-series analysis, our work takes a distinct direction. We adopt an interdisciplinary approach that seeks to integrate principles of human cognition and observational behavior.

To appreciate the relevance of this interdisciplinary approach, it's essential to understand the complexities of human driving behavior. These complexities are deeply rooted in a synergistic interplay between observational faculties and cognitive decision-making processes. It is worth noting that empirical evidence suggests that human drivers rely on visual information for approximately 90% of their driving decisions [2]. Given this, we are led to a pivotal research question: Can a trajectory prediction model for AVs be formulated that incorporates insights from the visual perception and attention allocation strategies exhibited by human drivers?



(a) Whole vision filed during driving (b) Central vison field

Fig. 1. Adaptive dynamic mechanism of the human driver's attention. **Left**: The human driver's visual field can be roughly divided into three parts: **central**, **fringe**, and **peripheral** vision fields. The central vision field receives the most attention, while visual information from the fringe and peripheral vision fields is prioritized during maneuvering changes, which are observed through side and rearview mirrors and receive comparatively less attention and observation time. **Right**: The coverage angle of the central vision field adjusts rapidly with velocity in real-time.

Answering this question necessitates an examination of the intrinsic limitations of human cognitive capacity and their measurable impacts on driving behavior. Research indicates that the human brain can effectively process information related to only a limited number of external agents, usually not exceeding four [3]. This limitation sets the stage for the priority schema that human drivers subconsciously follow. Specifically, when faced with multiple stimuli, drivers allocate attention based on factors like spatial orientation and proximity to the vehicle [4]. Additionally, given the increased risks of frontal collisions, a significant portion of cognitive resources is channeled towards the **central visual field** [5]. This focus enables quicker detection of and reaction to potential obstacles lying ahead [6]. In our study, we also pay close attention to a nuanced, yet vital, aspect of human driving behavior: the adaptability of the driver's 'visual sector,' which is defined by both its radius and angle. Notably, this sector isn't static but dynamically adapts in

*Both authors contributed equally to this research.

†Corresponding author.

[1]State Key Laboratory of Internet of Things for Smart City and Department of Computer and Information Science, University of Macau, Macau SAR, China. Email: {yc27979, mc25671, czxu}@um.edu.mo

[2]Department of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, China. Email: franklin1234560@163.com

[3]State Key Laboratory of Internet of Things for Smart City and Departments of Civil and Environmental Engineering and Computer and Information Science, University of Macau, Macau SAR, China. Email: zhenningli@um.edu.mo

[4]State Key Laboratory of Internet of Things for Smart City and Departments of Civil and Environmental Engineering, University of Macau, Macau SAR, China. Email: emailcyw@gmail.com

[5]Faculty of Innovation Engineering Macau University of Science and Technology, Macau SAR, China. Email: liyunjian@must.edu.mo

[6]School of Vehicle and Mobility, Tsinghua University, Beijing, China. Email: lishbo@tsinghua.edu.cn

real-time in relation to driving speed [7]. Counterintuitively, this adaptability manifests in a narrower visual sector at higher speeds and a wider one at lower speeds. The reason for this lies in human cognitive and perceptual adjustments. At higher speeds, the narrowed focus allows the driver to concentrate on the road directly ahead, ensuring a quicker response to sudden obstacles or changes. On the other hand, a broader focus at lower speeds enables the driver to absorb more peripheral information, such as roadside activity, pedestrians, and traffic signals, which becomes more relevant at slower paces. By incorporating an adaptive visual sector that adjusts its field-of-view based on speed, our model mimics this inherently human trait, providing a more intuitive and context-sensitive trajectory prediction.

Building on these insights, our model aims for a more human-like, adaptive trajectory prediction algorithm for AVs. The central contributions of our research are threefold:

- We introduce a sophisticated pooling mechanism that replicates human attention allocation with a novel **adaptive visual sector**. This mechanism dynamically adapts its focus in real-time, allowing the model to capture important perceptual cues from different scenes.
- We introduce a novel **dynamic traffic graph** to extract the interaction of agents using a unique **topology graph structure** constructed using Convolutional Neural Networks (CNN) and Graph Attention Networks (GAT). Enhanced by multi-head attention mechanisms, this architecture provides a robust means to model spatio-temporal dependencies and generate multiple hypothesis trajectories with corresponding credibility.
- In benchmark tests on the NGSIM, HighD, and MoCAD datasets, our model outperforms the state-of-the-art (SOTA) baselines by at least 15.2%, 19.4% and 12.0%, respectively, demonstrating its impressive accuracy and applicability in various traffic scenarios, including **highways** and **dense urban areas**.

II. RELATED WORKS

Deep Learning-based trajectory prediction. The rise of deep learning technologies has significantly propelled the field of trajectory prediction in autonomous vehicles (AVs). Long Short-Term Memory (LSTM) networks are renowned for their ability to model nonlinear time-series data, serving as a fundamental component in various trajectory prediction studies [9,10,11]. LSTMs have been particularly instrumental in predicting both longitudinal and lateral vehicle trajectories on highways [12]. Convolutional Neural Networks (CNNs) have also seen application in this domain, especially in modeling social interactions between vehicles through techniques like Social Pooling [13,14]. Hybrid CNN-LSTM models have also demonstrated promising results [15]. Further enriching the field are Graph Neural Networks (GNNs), which have been employed in Generative Adversarial Network (GAN)-based frameworks [16] and lightweight Graph Convolutional Network (GCN)-based models [17]. The introduction of multi-head attention mechanisms has led to advanced models such as HDGT, which combines Transformers [40] with Recurrent Neural Networks (RNNs) [18], and the spatio-temporal model STDAN [19], known for its high prediction accuracy.

Vision Aware Mechanisms in Traffic Behavior. One body of research explores the concept of the Visual Field, highlighting that a driver's field of view changes based on vehicle speed[6]. Specifically, slower speeds result in a broader focus on the surrounding environment, while higher speeds narrow the focus, emphasizing more distant views [20]. Another study builds upon this by showing that drivers can recognize more details and have a wider visual perspective at lower speeds. This research also delves into the implications of such visual mechanisms for accident rates [21]. A more recent study introduces the term "visual attention" to encompass these phenomena. Using eye-tracking technology, the study finds that drivers' focus shifts based on varying speeds and road types. The focus is generally divided into looking ahead for vehicle control and looking to the sides for lane-changing decisions, with these focus areas dynamically adapting based on the speed of the vehicle [22].

III. PROBLEM FORMULATION

This study aims to forecast the future trajectory of a target vehicle over a fixed time horizon T by leveraging historical states of both the target vehicle and its surrounding agents over a time duration of f.

We define x_t^i as the state of the target vehicle (superscript 0) and all observed surrounding vehicles (superscript 1 to N) at a given time t. This state includes position coordinates, velocity, acceleration, vehicle type, lateral and longitudinal behavior, and lane information. At each time step t, given the historical states of the spans T,

$$\boldsymbol{X} = \{x_{t-T:t}^0, x_{t-T:t}^1, x_{t-T:t}^2, \ldots, x_{t-T:t}^N\} \tag{1}$$

, which used for the inputs of the model.

The primary objective is to predict the trajectory of the target vehicle, where prediction spans from time step $T+1$ to $T+F$:

$$\boldsymbol{Y} = \{y_{T+1}^0, y_{T+2}^0, \ldots, y_{T+F}^0\}, \tag{2}$$

where F is the prediction horizon. Each y_{T+f}^0 (where $1 \leq f \leq F$) represents the future x and y coordinates of the target vehicle at time $T+f$.

In our model, the distribution of y_{T+f}^0 is captured using a bivariate Gaussian distribution. This is denoted as $P(\boldsymbol{Y}|\boldsymbol{X})$, which quantifies the uncertainty associated with each predicted trajectory point.

IV. PORPOSED MODEL

Figure 2 illustrates the architecture of our model, incorporating Context-aware, Interaction-aware, Vision-aware, and Priority-aware modules. These modules emulate the human observation process during driving and are trained together.

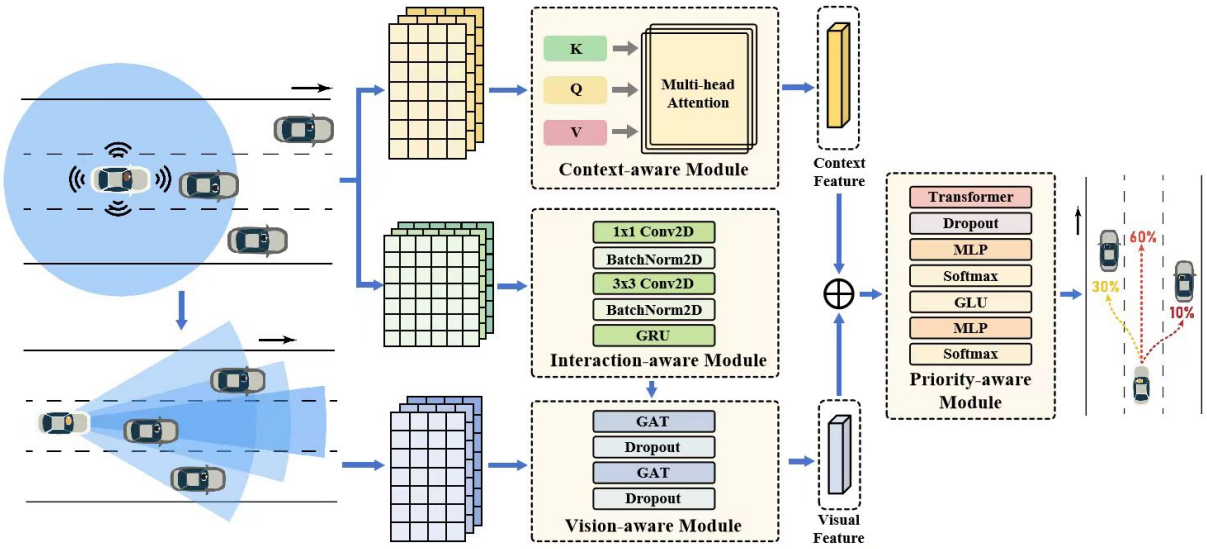


Fig. 2. A general overview of our proposed model. The original data undergoes processing through the Context-Aware Module to extract high-dimensional temporal information, resulting in the formation of the context feature. On the other hand, raw data is processed through the Visual-Aware and the Interaction-Aware Module, yielding Visual Features. Finally, the Priority-Aware Module is employed to generate multi-modal trajectory predictions based on the fusion of two features.

A. *Context-aware Module*

The Context-aware Module aims to efficiently capture temporal data by amalgamating the historical states of the target vehicle and its observed surrounding agents [19]. Specifically, a Gated Recurrent Unit (GRU) is used to process temporal data for both the target and the surrounding agents, encoding their trajectories over each observation interval and allowing the data to be amalgamated. A fully connected layer transforms the input motion state x_t^i into an embedded representation. This transformation uses a Multi-Layer Perceptron (MLP) with learnable parameters, activated by an Exponential Linear Unit (ELU). The ELU function is defined as

$$\mathrm{ELU}(x_t^i) = \begin{cases} (x_t^i, W_{\mathrm{E}})_{\mathrm{MLP}}, & \text{if } (x_t^i, W_{\mathrm{E}})_{\mathrm{MLP}} \geq 0 \\ \alpha(e^{(x_t^i, W_{\mathrm{E}})} - 1)_{\mathrm{MLP}}, & \text{if } (x_t^i, W_{\mathrm{E}})_{\mathrm{MLP}} < 0 \end{cases}. \tag{3}$$

Then GRU cells are used to encode each x_i^t for each vehicle i, $h_t^i = \mathrm{GRU}\left(x_t^i, h_{t-1}^i, W_G\right)$, where W_G are the learnable parameters common to all vehicles. The final GRU state for each vehicle can be expected to encode the motion state of that vehicle. Formally,

$$\alpha_t = \mathrm{softmax}\left(\frac{\langle q_t, K_t \rangle}{\sqrt{d_k}}\right) \tag{4}$$

where the α is computed by taking the dot product of the query(q) and key(K) vectors and applying a SoftMax operation. The features of the neighbor vehicles are aggregated using the equation:

$$Head_t = \sum_{i=1}^{N} \alpha_t^i v_t^i, \tag{5}$$

where the weighted sum is then computed using the value (v). Next, Gated Linear Units (GLU) are used to extract social dependencies and features from the target vehicle, resulting in $Head_t$. All multi-head information is processed and combined into the context feature: $H^0 = \left[h_1^0, h_2^0, \cdots, h_T^0\right]$. This output represents a collection of historical information processed by the attention mechanism, expressing the mapping of the temporal information of all neighboring vehicles to the temporal information of the target vehicle.

B. *Interaction-aware Module*

The Interaction-aware Module focuses on capturing spatial interactions between vehicles [39]. This is achieved by aggregating spatial data through a graph polymerization operation. Specifically, three matrices, denoted as A^{vel}, A^{acc}, and A^{man}, are used to represent the variations in velocity, acceleration, and behavioral consistency among vehicles within the same observation area at time t. These matrices are then merged into a 3D tensor D:

$$D = \{A^{vel}, A^{acc}, A^{man}\} \tag{6}$$

A 2D convolutional layer with a kernel size of 1×1 is applied to expand the number of channels in the tensor. Another convolutional layer with a kernel size of 3×3 is then applied to extract channel-specific features. Batch normalization and dropout techniques are used to improve the performance and generalization of the model. The processed tensor D is then fed into a GRU to capture the dynamic interactions between the target vehicle and its neighbors. The output from the aggregation layer is multiplied by the visual field matrix to simulate changes in focus. This results in a higher dimensional tensor, H_{conv}, which is stored in $\mathbb{R}^{T\times\mathrm{nbrs}}$. This output allows the model to capture a richer spatial context around the target vehicle in mixed-autonomy traffic environments.

C. Vision-aware Module

Traditional research often assumes that a driver's attention is evenly distributed throughout the environment [5]. However, it is important to recognize that human attention is limited. Traffic behavior studies have shown that drivers adjust their visual focus based on speed and other factors [6,14]. Specifically, a driver's central visual field (the frontal sector) of attention can be described as a "visual sector". For example, at relatively lower speeds, the driver's focus is concentrated within a visual sector defined by $r = 30$ meters and $\theta = 90°$. In contrast, at higher speeds, the focus shifts to a visual sector with $r = 90$ meters and $\theta = 45°$. This adaptation reflects a narrower but more distant field of view at higher speeds.

Given these findings, our model incorporates an adaptive visual sector concept to simulate the driver's changing visual field at different speeds. Based on the effectiveness confirmed by relevant research [7,14], we define speed thresholds at 30 km/h, 60 km/h, and 90 km/h to delineate different visual sectors. Different visual sectors are used to apply different weights to these regions. We use a visual matrix V to model the changes in the central visual field. We then perform element-wise multiplication between the interaction-aware output and the visual matrix:

$$N = H_{\text{conv}} \odot V, \tag{7}$$

where V is the visual matrix composed of varying values between 0 and 1 depending on the driver's speed. N denotes the node vectors to be fed into the graph neural network.

In mixed-autonomy traffic environments, where not all vehicles share the same control logic (including a significant number of human drivers), it becomes crucial to implement a mechanism for recognizing the different interaction patterns among vehicles in the scene. To address this issue, we used two layers of Graph Attention Network (GAT) with ELU activation functions, alternating with dropout layers, to analyze the interactions between targets. This approach allows us to simulate the different detection patterns of drivers towards different vehicles in mixed scenarios.

Considering the safety of vehicles in close proximity, we have defined a circular region with a certain safety distance as the " nearby area ". Within this boundary, we enhanced the weights of the GAT adjacency edges. At the same time, we construct a graph, denoted A^{edge}, containing all vehicles within the detection area. We compute similarity coefficients between each vertex N_i and its neighboring vehicles. Formally,

$$\alpha(i,j) = \frac{\text{LeakyReLU}\left(\mathbf{a}^{\boldsymbol{T}}\left[W * n_t(i) \| W * n_t(j)\right]\right)}{\sum_{k \in N_i} \text{LeakyReLU}\left(\mathbf{a}^{\boldsymbol{T}}\left[W * n_t(i) \| W * n_t(k)\right]\right)}, \tag{8}$$

where n denotes the features of a single node. To obtain the final features of each node at time t, we sum all the vectors connected to each node:

$$\hat{n}_t(i) = \sigma\left(\sum_{j \in N_i} \alpha(i,j) W * n_t(j)\right), \tag{9}$$

where σ is the nonlinear function ELU.

By incorporating the Vision-aware Module, we have not only improved the accuracy of the model but also enabled it to recognize the different driving patterns in mixed-autonomy traffic environments. The output of the module is called the visual feature.

D. Priority-aware Module

The Priority-aware Module uses a Transformer-based sequence-to-sequence model with an encoder-decoder architecture. Inspired by Transformer translation models, we use the visual feature as input for the encoder and the context feature as input for the decoder. We then perform multi-trajectory prediction based on different behavioral patterns and derive the final trajectories by mapping the results to a bivariate Gaussian distribution.

1) Encoder: The input tensor Z to the encoder is a visual feature that can be represented as $Z \in \mathbb{R}^{nbrs \times dim}$,. where $nbrs$ is the number of spatial details (e.g. position coordinates) in each trajectory, and dim is the dimension of the feature representation.

The encoder consists of L layers and uses multi-head attention to process three inputs: query (q), key (k), and value (v), all derived from the same input features $Z_n(t)$. Then, a dot product operation followed by a softmax operation is performed between Q and K to generate pairwise importance weights A. The weighted sum is then computed using the attention matrix A and the value V: $Z' = AV$. We define the attention function as $Z' = \text{Attn}(X)$. The multi-head attention mechanism in the encoder module is critical for capturing relevant spatial details. The results of multi-head attention are integrated using a linear transformation with weights $W_o \in \mathbb{R}^{n_h h_v \times h_v}$, where h_v is the dimension of the value V and n_h is the number of heads:

$$Z'' = \text{MHA}(Z) = \left[\text{Attn}_1(Z), \ldots, \text{Attn}_{n_h}(Z)\right] W_o \tag{10}$$

The feedforward module is a 2-layer MLP with activation $\delta(\cdot)$:

$$Z''' = \text{FFN}\left(Z''\right) = \delta\left(ZW_1 + b_1\right) W_2 + b_2 \tag{11}$$

2) Decoder: The Decoder module of the module receives two input sources: the encoded output Z''' from the encoder and the context feature $H_n(t)$ formed in Context-aware Module. The input trajectory tensor $H_n(t)$ is embedded using fully connected layers. The Transformer Decoder consists of L layers with multi-head self-attention and position-wise feedforward neural networks. The output of the Transformer denoted as $\tilde{H}^0 \in \mathbb{R}^{T \times dim}$, is generated using a linear transformation with Dropout and GLU:

$$\tilde{H}^0 = \left[\tilde{h}_1^0, \tilde{h}_2^0, \cdots, \tilde{h}_T^0\right] \tag{12}$$

The actual trajectory of a vehicle in real traffic scenes is often uncertain due to the complexity and diversity of possible driving maneuvers. To address this issue, we introduce the Generator, which is responsible for predicting the probability distribution of different driving maneuvers for the target vehicle[19].

Specifically, $\tilde{H}^0$ is transformed from historical data into future data using MLP layers, yielding the intention-specific feature $\tilde{Z} \in \mathbb{R}^{F \times dim}$, where F denotes the number of time steps the model is going to predict in the future. In order to predict the probability distribution of the future trajectory at each t time step based on different maneuvers, the intention-specific feature $\tilde{Z}$ and the prediction probability of maneuver classes $P(la)$ and $P(lo)$are concatenated and then fed into a fully connected layer with weight W_m as follows:

$$e_t^{la,lo} = \left(\tilde{Z} \oplus P(la) \oplus P(lo), W_m\right)_{\text{MLP}} \tag{13}$$

Additionally, to ensure temporal continuity in the predicted trajectory, the encoded representation $e_t^{la,lo}$ at each timestamp is sequentially fed into a decoder Long Short-Term Memory (LSTM) network. The LSTM outputs a bivariate Gaussian distribution for the vehicle's motion, represented as θ_t:

$$\theta_t = \left(\text{LSTM}\left(e_t^{la,lo}, W_{decoder}\right), W_d\right)_{\text{MLP}} \tag{14}$$

To estimate the probability distribution of the trajectory prediction Y, we apply the total probability theorem to incorporate the uncertainty captured by the bivariate Gaussian distribution. The parameters θ represent a collection of bivariate Gaussian distribution parameters for each time step in the future trajectory, denoted as $\theta = [\theta_{T+1}, \theta_{T+2}, \cdots, \theta_{T+F}]$. Each θ_t for $T+1 \leq t \leq T+F$ consists of the mean $\mu_{t,x}$ and $\mu_{t,y}$, the variances $\sigma_{t,x}$ and $\sigma_{t,y}$, and the correlation coefficient ρ_t for the vehicle's x and y positions. These parameters collectively represent the probability distribution of the predicted trajectory and account for the uncertainty inherent in the prediction process.

V. Experiments

In this paper, we use three datasets: Next Generation Simulation (NGSIM), Highway Drone (HighD), and Macao Connected Autonomous Driving (MoCAD) datasets, we use 3 seconds as trajectory history data for input, and the following 5 seconds for prediction. These datasets provide a wealth of information on real-world traffic scenarios. We use the Adam optimizer with an initial learning rate of 0.0005. The training and test experiments are conducted with a RTX3090 Graphics card.

A. Experimental Results

To verify the effect of the model proposed in this paper, we have compared the effect of existing reproducible models, and the results are shown in Table I.

We have evaluated the accuracy of the model proposed in this paper by comparing it to the aforementioned baseline using Root Mean Square Error (RMSE). Our analysis shows that the model presented in this paper consistently produces more accurate results across different time dimensions, regardless of the dataset used.

Previous research has shown that using information from surrounding vehicles leads to better results compared to not using such relevant information. In this paper, all baselines compared are based on interactions between the ego vehicle and other vehicles. Among the reproducible models selected in this study, we found that the performance of the S-LSTM, which uses only fully connected layers for prediction, and the trajectory prediction model based on GAN generation were the worst. The methods using social convolutional networks and the wave theory model using amplitude and phase performed better. The model based solely on the multi-head attention mechanism achieved the best performance prior to this study.

From the perspective of traffic behavior, we constructed the interaction part of the model and simulated the driver's visual focus changing with vehicle speed using a multi-head transformer model, which performed best. In the 5-second

TABLE I

Evaluation results for GaVa and the baselines in the three datasets. Note: RMSE (m) is the evaluation metric, where lower values indicate better performance, with some not specifying ('-'). Values in **bold** represent the best performance in each category.

Dataset	Model	Prediction Horizon (s)				
		1	2	3	4	5
NGSIM	S-LSTM [24]	0.65	1.31	2.16	3.25	4.55
	S-GAN [25]	0.57	1.32	2.22	3.26	4.40
	CS-LSTM [13]	0.61	1.27	2.09	3.10	4.37
	MATF-GAN [26]	0.66	1.34	2.08	2.97	4.13
	NLS-LSTM [27]	0.56	1.22	2.02	3.03	4.30
	M-LSTM [28]	0.58	1.26	2.12	3.24	4.66
	IMM-KF [29]	0.58	1.36	2.28	3.37	4.55
	GAIL-GRU [30]	0.69	1.51	2.55	3.65	4.71
	MFP [31]	0.54	1.16	1.89	2.75	3.78
	DRBP[32]	1.18	2.83	4.22	5.82	-
	DN-IRL [34]	0.54	1.02	1.91	2.43	3.76
	WSiP [8]	0.56	1.23	2.05	3.08	4.34
	CF-LSTM [35]	0.55	1.10	1.78	2.73	3.82
	MHA-LSTM [33]	0.41	1.01	1.74	2.67	3.83
	HMNet [36]	0.50	1.13	1.89	2.85	4.04
	TS-GAN [37]	0.60	1.24	1.95	2.78	3.72
	STDAN [19]	**0.39**	0.96	1.61	2.56	3.67
	GaVa	0.40	**0.94**	**1.52**	**2.24**	**3.13**
HighD	S-LSTM [24]	0.22	0.62	1.27	2.15	3.41
	S-GAN [25]	0.30	0.78	1.46	2.34	3.41
	WSiP [8]	0.20	0.60	1.21	2.07	3.14
	CS-LSTM(M) [13]	0.23	0.65	1.29	2.18	3.37
	CS-LSTM [13]	0.22	0.61	1.24	2.10	3.27
	MHA-LSTM [33]	0.19	0.55	1.10	1.84	2.78
	MHA-LSTM(+f) [33]	0.06	0.09	0.24	0.59	1.18
	NLS-LSTM [27]	0.20	0.57	1.14	1.90	2.91
	DRBP[32]	0.41	0.79	1.11	1.40	-
	EA-Net [38]	**0.15**	0.26	0.43	0.78	1.32
	CF-LSTM [35]	0.18	0.42	1.07	1.72	2.44
	STDAN [19]	0.19	0.27	0.48	0.91	1.66
	GaVa	0.17	**0.24**	**0.42**	**0.86**	**1.31**
MoCAD	S-LSTM [24]	1.73	2.46	3.39	4.01	4.93
	S-GAN [25]	1.69	2.25	3.30	3.89	4.69
	CS-LSTM(M) [13]	1.49	2.07	3.02	3.62	4.53
	CS-LSTM [13]	1.45	1.98	2.94	3.56	4.49
	MHA-LSTM [33]	1.25	1.48	2.57	3.22	4.20
	MHA-LSTM(+f) [33]	1.05	1.39	2.48	3.11	4.12
	NLS-LSTM [27]	0.96	1.27	2.08	2.86	3.93
	WSiP [8]	0.70	0.87	1.70	2.56	3.47
	CF-LSTM [35]	0.72	0.91	1.73	2.59	3.44
	STDAN [19]	0.62	0.85	1.62	2.51	3.32
	GaVa	**0.51**	**0.75**	**1.34**	**2.13**	**2.91**

prediction phases of the three data sets, it outperformed the former model by 15.2%, 19.4%, and 12.0%, respectively.

Based on the experimental results, we can draw the following conclusions: 1) Models using multimodal trajectories for loss propagation are superior because they consider multiple different states of the model predictions and determine the optimal solution through a comprehensive evaluation. 2) Convolutional networks are more effective compared to the simulation of physical factors, possibly because the computational approach of computers is better suited to convolutional spatial operations. 3) This experiment proves that the combination of convolutional networks and multi-head attention mechanisms can achieve excellent results, probably because their cooperation increases the robustness of the model.

B. Qualitative Results

In addition, we performed simulations to visualize predicted vehicle trajectories over a 5-second interval based on the experimental NGSIM and HighD datasets. As shown in Figure 3, the visual representation vividly illustrates the predictive capabilities of multimodal trajectories (shown as green solid lines) in complex traffic scenarios.

C. Ablation Studies

In this section, we perform an ablation study of the model to demonstrate the necessity and conciseness of the intermodularity of the GaVa model. Specifically, the model was altered into the following variants:

1) GVTA (-IaM): This variant directly removes the entire Interaction-aware module, i.e., it allows the model to make predictions relying solely on each vehicle's independent historical data without accessing the interaction characteristics of all vehicles through the graph neural network.
2) GVTA (-VM): This variant removes visual matrix, i.e., ignores the idea that the focus of the field of view varies with speed, and removes the corresponding module in the Vision-aware Module.
3) GVTA (+NVM): This variant uses not only the results of the Vision-aware Module outputs as node feature vectors but also adds the Interaction-aware Module results that have not been recognized by the visual matrix as node feature vectors.

TABLE II

PREDICTION ERROR OF ABLATION MODELS ON NGSIM DATASETS.

Horizon (s)	GaVa (-IaM)	GaVa (-VM)	GaVa (+NVM)	GaVa
1	0.62	0.45	0.43	**0.39**
2	1.43	1.16	0.96	**0.93**
3	2.45	2.13	1.59	**1.52**
4	3.89	3.35	2.37	**2.24**
5	5.14	4.26	3.36	**3.13**

We performed ablation experiments on the above three variants with the same set of other model parameters and training hyperparameters on NGSIM datasets, and the results are shown in Table II. From the results, it is not difficult

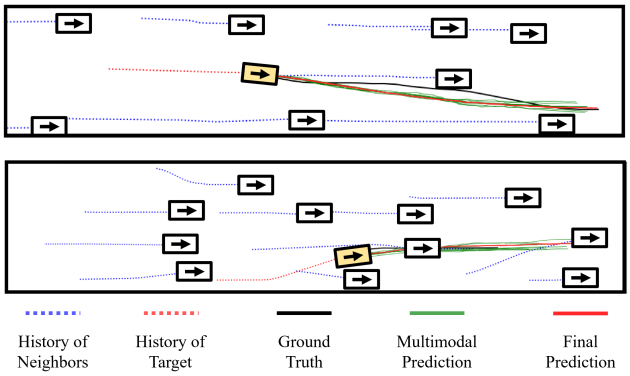


Fig. 3. Visualization of Predictions for Complex Traffic Situations in NGSIM (above) and highD (below) Datasets. The yellow rectangles denote the target vehicle, while other rectangles represent the neighboring vehicles.

to draw the following conclusions: First, the variant that ignores the composition of social network relationships performs the weakest, even worse than S-LSTM, confirming the importance of capturing social network interactions for trajectory prediction. Second, we find that the model variant that removes the visual field is not ideal, and its effect is comparable to that of S-GAN, suggesting that the visual field is critical to the overall role of the model. Third, instead of getting better, the results of the model with added convolutional data not detected by the visual field got relatively worse, with its RMSE increasing more and more as the number of seconds of prediction increased. This suggests that the necessary conciseness of the feature vectors of the graph neural network is beneficial for the final trajectory prediction.

The results confirm that extracting information about inter-vehicle interactions through graph neural networks is necessary for the machine, and in particular, Vision-aware Recognition validates the idea that the visual field of interest region of traffic behavior changes with speed. Both are essential for improving model performance.

VI. CONCLUSION

In this study, we have introduced a comprehensive trajectory prediction model called GaVa, which incorporates insights from traffic behavior studies. By incorporating temporal data, spatial data, and visual matrix, GaVa excels in accurately predicting future trajectories. The use of a multi-head attention mechanism and a novel graph neural network composition contribute to GaVa's superior trajectory prediction performance on three datasets, outperforming reproducible state-of-the-art results. Through an ablation study, we validate the importance of each component in the model and emphasize the need to integrate traffic behavioral science. Our work demonstrates the potential of combining domain knowledge from traffic behavior studies with advanced neural network architectures. We believe that this fusion of expertise opens promising avenues for future research in trajectory prediction and autonomous driving.

ACKNOWLEDGEMENTS

This research is supported by the Science and Technology Development Fund of Macau SAR (File no. 0021/2022/ITP, 0081/2022/A2, 001/2024/SKL), and University of Macau (SRG2023-00037-IOTSC). For correspondence related to this research, please contact with Zhenning Li (zhenningli@um.edu.mo) and Chengzhong Xu (czxu@um.edu.mo).

REFERENCES

[1] Huang, Yanjun, et al. "A survey on trajectory-prediction methods for autonomous driving." IEEE Transactions on Intelligent Vehicles 7.3 (2022): 652-674.

[2] Hills, B. L. (1980). Vision, visibility, and perception in driving. Perception, 9(2), 183-216.

[3] Louie, J. (2018). Working memory capacity and executive attention as predictors of distracted driving.

[4] Broadbent, David P., et al. "Cognitive load, working memory capacity and driving performance: A preliminary fNIRS and eye tracking study." Transportation research part F: Traffic psychology and behavior 92 (2023): 121-132.

[5] Hajime, I. T. O., ATSUMI, B., Hiroshi, U. N. O., and AKAMATSU, M. (2001). Visual distraction while driving: trends in research and standardization. IATSS Research, 25(2), 20-28.

[6] Lee, J. D. (2014). Dynamics of driver distraction: The process of engaging and disengaging. Annals of advances in automotive medicine, 58, 24.

[7] Tucker, A., and Marsh, K. L. (2021). Speeding through the pandemic: Perceptual and psychological factors associated with speeding during the COVID-19 stay-at-home period. Accident Analysis and Prevention, 159, 106225.

[8] Wang, Renzhi, et al. "WSiP: Wave Superposition Inspired Pooling for Dynamic Interactions-Aware Trajectory Prediction." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 4. 2023.

[9] Hao Xue, Du Q Huynh, and Mark Reynolds, "SS-LSTM: A hierarchical LSTM model for pedestrian trajectory prediction," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1186–1194, 2018, IEEE.

[10] Arsal Syed and Brendan Tran Morris, "SSeg-LSTM: semantic scene segmentation for trajectory prediction," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, pp. 2504–2509, 2019, IEEE.

[11] Zhenning Li, Haicheng Liao, Ruru Tang, Guofa Li, Yunjian Li, and Chengzhong Xu, "Mitigating the impact of outliers in traffic crash analysis: A robust Bayesian regression approach with application to tunnel crash data," *Accident Analysis & Prevention*, vol. 185, p. 107019, 2023, Elsevier.

[12] Florent Altché and Arnaud de La Fortelle, "An LSTM network for highway trajectory prediction," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 353–359, 2017, IEEE.

[13] Nachiket Deo and Mohan M Trivedi, "Convolutional social pooling for vehicle trajectory prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1468–1476, 2018.

[14] Zhenning Li, Zhiwei Chen, Yunjian Li, and Chengzhong Xu, "Context-aware trajectory prediction for autonomous driving in heterogeneous environments," *Computer-Aided Civil and Infrastructure Engineering*, 2023, Wiley Online Library.

[15] Guo Xie, Anqi Shangguan, Rong Fei, Wenjiang Ji, Weigang Ma, and Xinhong Hei, "Motion trajectory prediction based on a CNN-LSTM sequential model," *Science China Information Sciences*, vol. 63, pp. 1-21, 2020, Springer.

[16] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2255–2264, 2018.

[17] Hyeongseok Jeon, Junwon Choi, and Dongsuk Kum, "Scale-net: Scalable vehicle trajectory prediction network under random number of interacting vehicles via edge-enhanced graph convolutional neural network," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2095–2102, 2020, IEEE.

[18] Kai Gao, Xunhao Li, Bin Chen, Lin Hu, Jian Liu, Ronghua Du, and Yongfu Li, "Dual Transformer Based Prediction for Lane Change Intentions and Trajectories in Mixed Traffic Environment," *IEEE Transactions on Intelligent Transportation Systems*, 2023, IEEE.

[19] Xiaobo Chen, Huanjia Zhang, Feng Zhao, Yu Hu, Chenkai Tan, and Jian Yang, "Intention-aware vehicle trajectory prediction based on spatial-temporal dynamic attention network for internet of vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 19471–19483, 2022, IEEE.

[20] Hans Antonson, Selina Mårdh, Mats Wiklund, and Göran Blomqvist, "Effect of surrounding landscape on driving behavior: A driving simulator study," *Journal of Environmental Psychology*, vol. 29, no. 4, pp. 493–502, 2009, Elsevier.

[21] Shameen Randika Dharmasena and Edirisooriya Arachchige Tharanga Suresh, "A methodology to analyze road landscape in accident Black-Spots: the case of Southern Expressway, Sri Lanka," *ArchNet-IJAR: International Journal of Architectural Research*, vol. 12, no. 2, p. 347, 2018, Emerald Group Publishing Limited.

[22] Sophie Lemonnier, Lara Désiré, Roland Brémond, and Thierry Baccino, "Drivers' visual attention: A field study at intersections," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 69, pp. 206–221, 2020, Elsevier.

[23] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[24] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961-971, 2016.

[25] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2255-2264, 2018.

[26] Tianyang Zhao, Yifei Xu, Mathew Monfort, Wongun Choi, Chris Baker, Yibiao Zhao, Yizhou Wang, and Ying Nian Wu, "Multi-agent tensor fusion for contextual trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12126-12134, 2019.

[27] Messaoud, Amine, Martin Voigt, and Alois Knoll. "Non-linear trajectory prediction in interactive on-road scenarios using LSTM." In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6341-6347, 2019.

[28] Nachiket Deo and Mohan M Trivedi, "Multi-modal trajectory prediction of surrounding vehicles with maneuver based LSTMs," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1179-1184, 2018, IEEE.

[29] Lefkopoulos, Ioannis, et al. "Interaction-aware probabilistic trajectory prediction for autonomous vehicles." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11568-11577, 2020.

[30] Kuefler, Andreas, et al. "Imitating driver behavior with generative adversarial networks." In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, pp. 6235-6243, 2017.

[31] Tang, Shixiang, et al. "Multiple human tracking and trajectory prediction by using improved particle filter." In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5349-5355, 2019.

[32] Gao, Kai, Xunhao Li, Bin Chen, Lin Hu, Jian Liu, Ronghua Du, and Yongfu Li. "Dual Transformer Based Prediction for Lane Change Intentions and Trajectories in Mixed Traffic Environment." *IEEE Transactions on Intelligent Transportation Systems*, 2023, IEEE.

[33] Messaoud, Amine, Martin Voigt, and Alois Knoll. "Attention-based Trajectory Prediction for Autonomous Driving: A Survey." *IEEE Intelligent Transportation Systems Magazine*, 2021, IEEE.

[34] Fernando, Hasini B. H., et al. "Neighbourhood aware long short-term memory networks for urban traffic prediction." In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2394-2403, 2020.

[35] Xie, Guo, Anqi Shangguan, Rong Fei, Wenjiang Ji, Weigang Ma, and Xinhong Hei. "Congestion-aware trajectory prediction for autonomous driving using convolutional neural networks." In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4327-4333, 2021.

[36] Xue, Hao, Shihao Li, Zhongang Tang, and Du Q. Huynh. "Hierarchical LSTM with Spatio-Temporal Memory for Pedestrian Trajectory Prediction." In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3526-3532, 2021.

[37] Yu Wang, Shengjie Zhao, Rongqing Zhang, Xiang Cheng, and Liuqing Yang, "Multi-Vehicle Collaborative Learning for Trajectory Prediction With Spatio-Temporal Tensor Fusion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 236-248, 2022, IEEE.

[38] Yingfeng Cai, Zihao Wang, Hai Wang, Long Chen, Yicheng Li, Miguel Angel Sotelo, and Zhixiong Li, "Environment-attention network for vehicle trajectory prediction," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 11, pp. 11216-11227, 2021, IEEE.

[39] Xin Li, Xiaowen Ying, and Mooi Choo Chuah, "Grip++: Enhanced graph-based interaction-aware trajectory prediction for autonomous driving," *arXiv preprint arXiv:1907.07792*, 2019.

[40] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).