Regular Issue

IAtraj: Multi-Modal Trajectory Prediction Through Contextual Information Spatio-Temporal Interaction and Awareness

Xiaoliang Wang[1,2], Lian Zhou[1,2], Kuan-Ching Li[3*], Shiqi Zheng[1,2], Huijing Fan[1,2]

[1] School of Computer Science and Engineering, Hunan University of Science and Technology (China)
[2] Hunan Key Laboratory for Service Computing and Novel Software Technology (China)
[3] Dept of Computer Science and Information Engineering, Providence University (Taiwan)

* Corresponding author: kuancli@pu.edu.tw

Received 25 November 2023 | Accepted 25 July 2024 | Early Access 9 September 2024

Accurately and feasibly predicting the future trajectories of autonomous vehicles is a critically important task. However, this task faces significant challenges due to the variability of driving intentions and the complexity of social interactions. These challenges primarily arise from the need to understand one's driving behaviors and model the interaction information of the surrounding environment. A substantial amount of research has been focused on integrating interaction information from the surrounding environment, mainly using raster images or High-Definition maps (HD maps). However, the real-time update of environmental maps and the high computational cost associated with processing interaction information using compatible technologies such as vision have become limiting factors. Additionally, ineffective simulation and modeling of real driving scenarios, coupled with inadequate understanding of contextual environmental information, result in lower prediction accuracy. To overcome these challenges, we propose a multi-modal trajectory prediction model based on sequence modeling namely IAtraj, incorporating multiple attention mechanisms, focuses on the three critical elements in real traffic scenarios: the target agent's historical trajectory, effective interactions with neighboring vehicles, and lane supervision and retention strategies. To better model these elements, we design modules for Temporal Interaction (TI), Spatial Interaction (SI), and Lane Awareness (LA). Through extensive experiments conducted on the publicly available nuScenes dataset, IAtraj exhibits outstanding performance, successfully addressing the challenges of temporal dependencies in trajectory sequences and the representation of scene changes. Finally, comprehensive ablation experiments validate the effectiveness of each significant module, reinforcing the reliability and robustness of IAtraj in dealing with complex traffic scenarios.

Keywords

Attention Mechanisms, Contextual Information, Multi-Modal, Spatio-Temporal Interaction and Awareness, Trajectory Prediction.

DOI: 10.9781/ijimai.2024.09.001

I. Introduction

In multi-agent interactive prediction, accurate and feasible trajectory prediction of self-driving vehicles is an important prerequisite for safe and efficient vehicle operation. This requires a thorough comprehension and integration of the agents' historical trajectory sequences and the environmental information, encompassing traffic participants and lane details. However, fully understanding and modeling the historical trajectory information and the environmental information is a great challenge. In the actual driving scenarios, affected by various potential factors such as the driver's driving habits and the actual environmental conditions, there are many possibilities for the future driving trajectories of the target agent, so the future trajectories of the target agent should be extensively multi-modal. As shown in Fig. 1, according to the environment information, drivers' psychology, behavioral habits, or other potential factors, the target agent can choose to continue straight (vertical diversity) or steer (horizontal diversity) at different speeds presenting rich multi-modal future trajectory sequences. Earlier, due to the limited experimental conditions and equipment, researchers often explored trajectory prediction methods based on physical models. [1], [2], [3] all used Kalman filter-based physical models to predict the state of the agent (including the moving direction, traveling lane, speed, and acceleration, among others). Chen *et al.* [4] explored vehicle trajectory prediction using a deep Monte Carlo Tree Search (deep-MCTS) approach. Both [5], [6] used machine learning-based methods for path planning and prediction. However, traditional physical and machine learning models have lower prediction accuracy, robustness, and high latency.

As historical driving trajectories directly influence future driving behaviors, extracting features from chronological trajectory sequences

Please cite this article as: X. Wang, L. Zhou, K. C. Li, S. Zheng, H. Fan. IAtraj: Multi-Modal Trajectory Prediction Through Contextual Information Spatio-Temporal Interaction and Awareness, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 5, pp. 17-28, 2025, http://dx.doi.org/10.9781/ijimai.2024.09.001

Fig. 1. Multi-modal travel trajectory map of the target agent.

becomes crucial. This process involves time-series modeling of the trajectory data. Generally, the more historical information that can be provided and the better the model's ability to extract and model the trajectory information, the more accurate the trajectories we predict. Regrettably, the dataset imposes limitations on the available feature information for the model. Therefore, our focus lies in enhancing the model's capacity to comprehend historical information. Existing works such as [7], [8], [9], [10] fully consider the sequential nature of trajectory sequences and use Long Short-Term Memory (LSTM) Networks to extract the long-term temporal dependence of trajectory sequences. Additionally, [11], [12] utilize an attention mechanism to learn which time-step feature information needs more attention adaptively. Meanwhile, for generating future trajectory sequences that also have temporal features with long-term dependence, future prediction trajectories need to be generated with a complete understanding of the interaction information to generate plausible trajectory sequences that cover a wide range of future possibilities. Numerous studies, such as [8] and [12], correspond to the encoder using LSTM as a multi-modal trajectory decoder. PGP [13] and LAformer [14] introduce random noise to simulate longitudinal driving behaviors, such as acceleration and deceleration, and the introduction of random noise covers the diversity of changes in a variety of future driving trajectories. Nevertheless, this approach to some extent, heightens the model's uncertainty, thereby affecting the accuracy and realism of predictions.

The future travel paths of the vehicles are influenced not only by their historical trajectories but also by the surrounding environmental information in the scenes, and the neighboring agents may directly impact the current decisions of the target agent. Therefore, the model needs to consider the contextual scene information of the target agent and the environmental information comprehensively. In this regard, previous researchers have conducted numerous studies on extracting feature information, leading to a multitude of interaction-aware models. Early literature employed trajectory sequences and scene raster images as multi-modal inputs. Including raster images and videos facilitate interaction between multiple agents and the understanding of contextual information to a certain extent [15], [16], [17]. Although these methods can be implemented using visual techniques, they are often limited by the fact that they can't capture the dynamic information of the agents and scenes well, as well as require significant computational overhead. Additionally, the decoder struggles to decode spatio-temporal information accurately. In recent years, a large number of researchers have applied graphs to traffic prediction in traffic scenarios [18], [19], [20]. However, the incorporation of graphs can also simulate social interactions in traffic scenarios, and existing works use high-definition maps to vectorize and encode traffic scenario contextual information. Vectornet [21] vectorizes both historical trajectories and lane lines as folded segments and models them as global interaction graphs. PGP [13] divides lanes into nodes and models lane graphs. Similarly [22] and [23], Graph Neural Networks (GNNs) are used to realize the interaction and awareness of feature information among multiple agents. Meanwhile, numerous research efforts have explored the trajectory prediction based on the attention mechanisms [9], [12], [14], [24], [25], [26], [27]. They all use the improved attention mechanisms to calculate attention weights for realizing the effective interactions between multiple agents. Compared to the raster image methods, these approaches can fully comprehend environmental and scene information to predict the future trajectories of agents accurately.

Although studies have made significant progress in raster images, high-definition maps, and sequence modeling, there are still the following problems: (1) Lack of organic and unified modeling of the environment and awareness: most of the existing research only deals with and models one-sided factors, such as modeling interactions around the vehicles, but lacks monitoring and understanding of lane keeping aspects, (2) Insufficient understanding and modeling of feature information, as the availability of the vehicle history data is limited, the model can only increase the understanding and modeling of the vehicle history feature information, (3) Raster images can be compatible with advanced visual technologies but often face challenges of high computational costs and difficulty in effectively extracting and modeling subtle features, (4) More advanced models and predictive capabilities, providing accurate and efficient forecasting, can assist autonomous vehicles in making wiser and safer decisions.

To solve the abovementioned problems, we propose a multi-modal trajectory prediction model based on contextual spatio-temporal interaction and awareness: IAtraj. The model fully considers and analyzes the complex spatio-temporal interactions between the target agent and its neighboring agents, as well as its ability to perceive the road environment. Moreover, it establishes an integrated model that combines feature extraction and modeling, interaction awareness, and multi-modal decoding. We have the following key contributions:

- To study the problem of trajectory prediction based on historical sequences and environmental information, and construct a generalized prediction network framework that can be applied to efficient trajectory prediction for multiple agents, such as vehicles, pedestrians, and others.
- To propose an Interaction and Awareness Block (IAB) based on the attention mechanisms, from which intrinsic interaction and awareness features are extracted by fusing joint temporal, spatial, and lane features. The module takes into full consideration the three most crucial elements in actual driving scenarios (historical trajectories, neighboring vehicles, and road information). It establishes independent yet organically integrated processing strategies, thereby providing the target agent with accurate judgment and decision-making strategies.
- To achieve better performance on the nuScenes [28] dataset and verify the effectiveness of the module through extensive ablation studies.

The remainder of this article is organized as follows. Section II presents related work on sequence modeling and interaction awareness, and Section III presents the complete implementation of our proposed IAtraj model. Section IV conducts extensive comparisons and ablation experiments on the public dataset nuScenes, and finally, Section V gives concluding remarks and directions for future work.

II. Related Work

A. Time Sequence Modeling

Sequence modeling is a key component of the trajectory prediction, which determines whether the interaction and awareness block can effectively utilize trajectory features. In recent years, with the advancement of deep learning, researchers have extensively explored trajectory feature modeling. LSTMs are capable of modeling the long-term dependence within time sequence data and simultaneously addressing the issue of gradient explosion associated with Recurrent Neural Networks (RNNs). [9] uses LSTMs in the encoder stage to extract and model the target agent, neighboring agents, and lane information respectively, and similarly, in the decoding stage where the trajectory data is also time-sequential, [8] and [12] apply LSTMs to generate multi-modal future trajectory sequences.

With the emergence of attention mechanisms [29], a large number of attention mechanism variants have been produced, and the field of trajectory prediction has also used them for feature extraction and interaction. [12] uses a multi-head attention mechanism to adaptively assign weights to the agent's historical trajectory data, in which [30] combines LSTM with attention mechanisms for multi-dimensional extraction of the target agent's historical trajectories. However, attention mechanisms entail high computational complexity. [30] also designs a single-agent coding module for a multi-dimensional attention mechanism to improve the computational speed. For this purpose, we refer to the above works and design our Temporal Interaction (TI) module using LSTMs in combination with the AFT FULL module [31], the module eliminates matrix multiplication operation compared to a traditional attention mechanisms, which are used in a way that not only has better results but also has a lower computational complexity.

B. Spatial Interaction and Awareness

In real traffic scenarios, the target agent is not an isolated moving entity. It typically relies on real-time, efficient analysis, understanding, and response to the surrounding environment to make timely adjustments such as steering, acceleration, and deceleration. However, to achieve accurate judgments, we need to extract rich interaction information and feature representation from traffic scenes. Previous works have mainly transformed traffic scenes (e.g., lanes, pedestrian crossings, traffic signals, etc.) into bird's-eye views and performed feature extraction from the bird's-eye views by visual methods. Usually, the studies involve fusing the acquired feature information with the target agent's data processed through a temporal model, which is then utilized as input for subsequent predictions. Many works have leveraged visual technologies such as Convolutional Neural Networks (CNNs) to characterize the rich features of traffic scenes effectively. For example, H. Cui *et al.* [15] constructs an MTP model using MobileNet-v2 and ResNet models to represent traffic scene information as raster images. T. Phan-Minh *et al.* [32] improves the MTP model by approximating all possible motions through a set of trajectories and focusing on the multi-modal trajectory outputs. [33], [34], [35] adopt social pooling techniques to achieve effective interactions between the target agent and neighboring agents. However, raster images are susceptible to limitations of local awareness and tend to overlook important features in the global context and dynamic scenes, reducing the accuracy of trajectory prediction. The attention mechanisms also eliminate the need for raster images and can focus on more important information in the environment through adaptive weight allocation. [8] and [9] both employ LSTM to handle information from multi-agent, emphasizing the significance of neighboring agents and lanes in predicting the trajectories of the target agent through the use of attention mechanisms. [11] adopts a dual-attention mechanism to model intentional behaviors and trajectory prediction separately, which improves the accuracy of prediction. Therefore, to achieve effective interactions between multiple agents, we designed the spatial interactive attention module to perceive the interactions among multiple agents and accurately predict trajectories by utilizing effective spatial representation to the greatest extent.

III. Methods

Fig. 2 shows the proposed IAtraj model, in this section, we first introduce data preprocessing and problem formulation, followed by a detailed overview of the IAtraj model.

A. Preprocessing and Problem Formulation

Target agent history trajectory ($V_{Tar}^{(P)}$): Using the state information of the target agent in the past 2 seconds as input. $V_{Tar}^{(P)}$ represents the historical trajectory sequences of the target agent in the past T + 1 time steps. That is, $V_{Tar}^{(P)} = \left\{V_{T}^{(P)}, V_{T+1}^{(P)}, \dots, V_{0}^{(P)}\right\}$, each state information is denoted as $V_t^{(P)} = [\mathrm{x}_t, \mathrm{y}_t, \mathrm{v}_t, \mathrm{a}_t, \theta_t]$, where x_t, y_t denote the agent's transverse and longitudinal coordinate in the coordinate system in the t moment, and v_t, a_t and θ_t denote the agent's velocity, acceleration, and yaw angle information at the moment t.

Target agent future information ($V_{Tar}^{(F)}$): Generating predicted trajectory coordinates for the target agent in the next 6 seconds. $V_{Tar}^{(F)}$ represents the sequences of predicted trajectory states of the target agent at the future H time steps. That is, $V_{Tar}^{(F)} = \{V_f^{(1)}, V_f^{(2)}, \dots, V_f^{(H)}\}$, and each state information is denoted as $V_h^{(f)} = [\mathrm{x}_h, \mathrm{y}_h]$, where x_h, y_h denote the agent's transverse and longitudinal coordinate in the coordinate system at the h moment, respectively.

Lane information ($L^{(N)}$): Based on the target agent's centroid position, search for the nearest N lane segments within the surrounding threshold range. Subsequently, select the two preceding and following lane segments to ensure connectivity. Finally, resample their coordinates to have equal distances. Among them, the lane closest to the future trajectory in all lanes is labeled as the reference lane. $L^{(N)}$ represents the N lane information that the surrounding environment influences the target agent. That is, $L^{(N)} = \{L^{(1)}, L^{(2)}, \dots, L^{(N)}\}$.

Neighboring agent historical trajectory information $V_{Sur}^{(N)}$: Since the selected lanes are the closest to the target agent, it is only necessary to choose the state information of the closest agents within each of the selected paths. We consider the closest neighboring agents in the lanes to have the most significant impact on the target agent; therefore, there is no need to screen other neighboring agents. $V_{Sur}^{(N)}$ represents the trajectory sequences of the neighboring agents for the past T + 1 time steps. That is, $V_{Sur}^{(N)} = \left\{V_{-T}^{(n)}, V_{-T+1}^{(n)}, \dots, V_{0}^{(n)}\right\}$, and the specific state information is similar to the above target agent.

B. Detailed Overview of the Model

1. Vehicle-Lane Feature Encoder (VLFE)

The first step in trajectory prediction is to encode the trajectory sequence data and environment information, and the effective extraction of features determines whether the interaction and awareness block can fully understand and utilize the feature information. As shown in Fig. 2, the feature encoding module contains two key parts: the feature extraction and the information aggregation. Specifically, for the feature extraction module, to capture the feature information at different scales, a one-dimensional CNN (1D-CNN) is used to perform a sliding convolution operation on $\forall V \in \{V_{Sur}^{(1)}, V_{Sur}^{(2)}, \dots, V_{Sur}^{(N)}, V_{Tar}\}$ or $\forall L \in \{L^{(1)}, L^{(2)}, \dots, L^{(N)}\}$. In addition, introducing the LSTM helps to improve the model's understanding of sequence information, which in turn improves the ability of temporal modeling of contextual information. The feature extraction module can be expressed as Eq. (1)-(3):

$$\eta_L^i = LSTM(1D - CNN(L^i)) \tag{1}$$

$$\eta_S^i = LSTM(1D - CNN(V_{Sur}^i)) \tag{2}$$

$$\eta_T = LSTM(1D - CNN(V_{Tar})) \tag{3}$$

In the above section, L^i, V_{Sur}^i, and V_{Tar} denote the original sequence information of the lanes, neighboring agents, and the target agent,

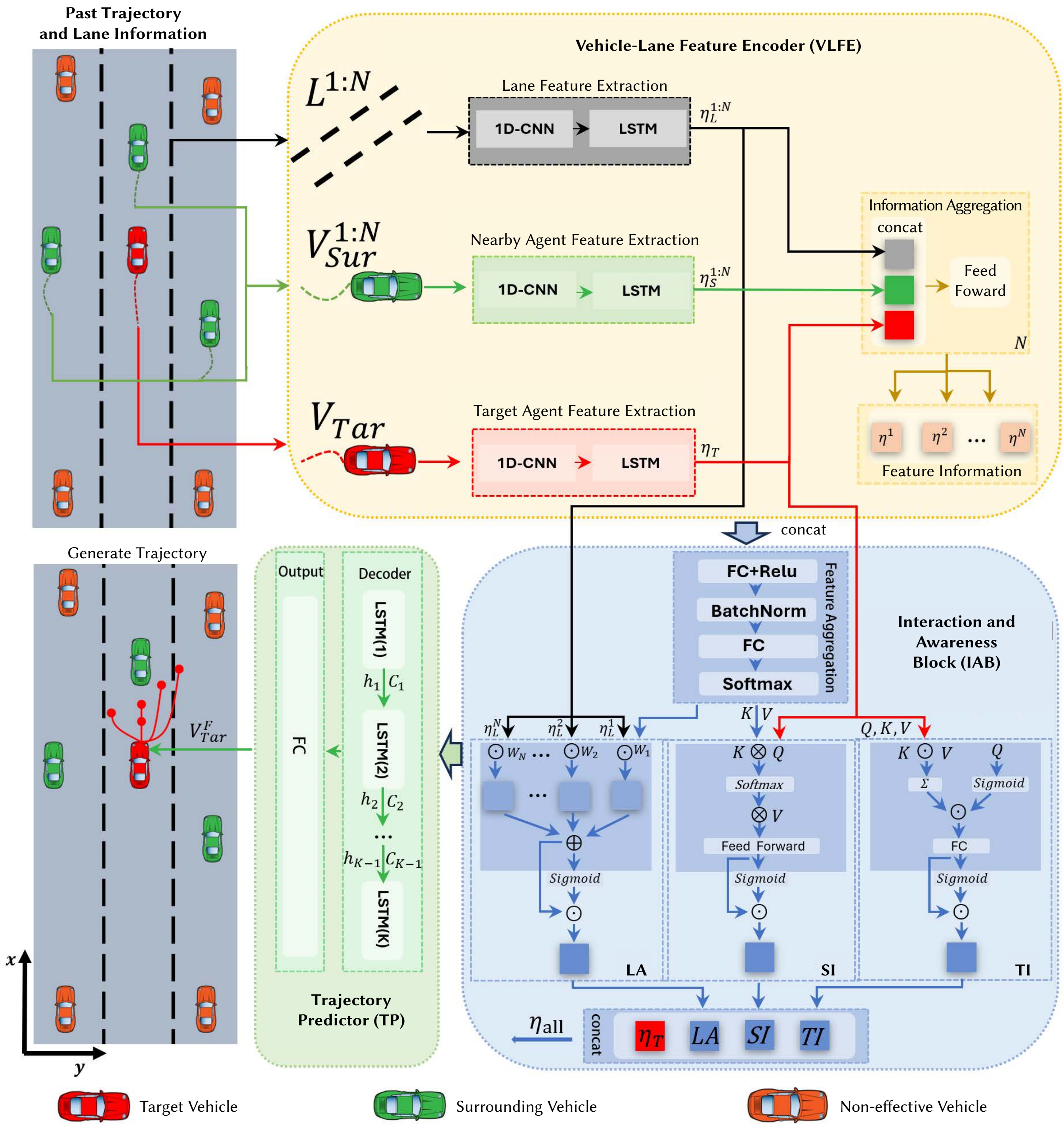


Fig. 2. A multi-modal trajectory prediction model framework using contextual information spatio-temporal interaction and awareness. The proposed model is divided into three main phases: **Vehicle-Lane Feature Encoder** (**VLFE**): processing and extracting features of the target agent (red), neighboring agents (green), and lanes (black); **Interaction and Awareness Block** (**IAB**): simulating effective spatio-temporal interactions of multiple agents and approximating lanes in case of lane deviation of the target agent; **Trajectory Predictor** (**TP**): generating multi-modal predicted trajectory sequences.

respectively. η_L^i, η_S^i, and η_T are the feature information generated by the lanes, neighboring agents, and the target agent after the feature extraction module, respectively. In the information aggregation module, we can combine the rich information $\eta_L^{1:N}$, $\eta_S^{1:N}$, and η_T in the trajectory sequences, and then provide more expressive features for the subsequent trajectory prediction tasks through the mapping of feedforward layers. Eq. (4) describes the process of information aggregation.

$$\eta^i = BatchNorm(\emptyset_2(Relu(\emptyset_1(concat(\eta_L^i, \eta_S^i, \eta_T))) \quad (4)$$

The $\emptyset_1$, $\emptyset_2$ layers are two fully connected layers, whose main function is to transform and map the features nonlinearly, and η^i is the output of the VLFE module, which contains the historical trajectories of the target agent, neighboring agents, and lane information.

2. Interaction and Awareness Block (IAB)

For multi-modal trajectory prediction, it is crucial to establish the spatio-temporal interactions between the target agent, neighboring agents, and the surrounding environment, so we use the IAB module to achieve the spatio-temporal interactions of the target agent and efficient awareness of the surrounding environment. Following the 1D-CNN with the LSTM, the temporal features of the target agent are initially extracted. To emphasize the significant spatio-temporal features, we predominantly employ the information passing through the VLFE module, emphasizing the significant temporal and spatial expressions of the target agent via the TI and SI modules. For the lane awareness module, the lane weights are adaptively assigned to select an appropriate driving lane.

Temporal Interaction Module (TI): The TI module enables the model to more selectively focus on and integrate information from different time steps in the sequences, and the features η_T generated by the target agent after the VLFE module are nonlinearly mapped into Q_{TI}, K_{TI}, and V_{TI}. The mapping relationship changes as shown in Eq. (5)-(7):

$$Q_{TI} = \rho_1(\eta_T, W_{Q_{TI}}) \tag{5}$$

$$K_{TI} = \rho_2(\eta_T, W_{K_{TI}}) \tag{6}$$

$$V_{TI} = \rho_3(\eta_T, W_{V_{TI}}) \tag{7}$$

Where, η_T denotes the feature information generated by the target agent after the VLFE module. $W_{Q_{TI}}$, $W_{K_{TI}}$ and $W_{V_{TI}}$ are the weight matrix. Additionally, ρ_1, ρ_2 and ρ_3 are the different linear transformation layers to compute the significant time expressions:

$$\eta_{TI} = \sigma(Q_{TI}) \odot (\frac{\sum exp(K_{TI} + W) \odot V_{TI}}{\sum exp(K_{TI} + W)}) \tag{8}$$

The temporal interaction attention values are calculated as described in Eq. (8). We calculate the time-series weighted average using K_{TI} and V_{TI}, then employ Q_{TI} for implicit attention calculations. This process allows us to acquire trajectory information based on these calculated weights. Consequently, it can emphasize or balance the significance of specific time steps, thereby obtaining more representative temporal features η_{TI}. Here, W is the weight matrix, and $\odot$ is the element-wise product, which makes the computational complexity of the AFT FULL much lower than that of other attentional mechanisms, as the element-wise operation replaces the matrix multiplication of traditional attentional mechanisms.

Spatial Interaction Module (SI): The SI module interacts with each element of the target agent's trajectory sequences (for example, features at each time step) with elements from other sequences through matrix multiplication. It generates weights based on their similarity to simulate the impact of neighboring agents on the ego vehicle's movement in real driving scenarios. The computation of the *Query*, *Key*, and *Value* in the multi-head attention mechanism is described in Eq. (9)-(11):

$$Q_{SI} = \rho_4(\eta_T, W_{Q_{SI}}) \tag{9}$$

$$K_{SI} = \rho_5(\eta, W_{K_{SI}}) \tag{10}$$

$$V_{SI} = \rho_6(\eta, W_{V_{SI}}) \tag{11}$$

In this case, the *Query* (Q_{SI}), *Key* (K_{SI}), and *Value* (V_{SI}) are obtained by nonlinear mapping of the target agent features η_T with the aggregated features η. The nonlinear mapping is shown in Equations (9)-(11). Each *head* ∈ 1, 2, ..., N_h of Q_{SI}, K_{SI}, and V_{SI} for attention computation can be defined as Eq. (12):

$$head_i = softmax(\frac{Q_{SI} \otimes transpose(K_{SI})}{\sqrt{d_k}}) \otimes V_{SI} \tag{12}$$

$$\eta_{SI} = Concat(head_1, head_2, \dots, head_{N_h})W + b \tag{13}$$

By utilizing the similarity relationship between the information of the target agent and that of neighboring agents, a weighted aggregation was conducted on the neighboring agents' information. Enable the final feature representation $head_i$ to more effectively capture interactions between the target agent and its neighboring agents. Where, Eq. (13) represents the value of attention for aggregating multiple heads, N_h represents the number of heads of multi-head attention, W represents the weight matrix, b represents the bias, $\otimes$ denotes the matrix multiplication, and η_{SI} represents the feature values generated by the spatial interaction module.

Lane Awareness Module (LA): Calculate the attention weights designed to decrease the extent of lane deviation for the target agent using lane-aware probabilities associated with neighboring lane features. The lane-aware feature results are obtained by summing the calculated probabilities. The specific formula expressions are shown in (14) - (15):

$$\omega_i = softmax(\emptyset_4(Relu(\emptyset_3(\eta^i)))) \tag{14}$$

$$\eta_{LA} = \sum_{i=1}^{N} \omega_i \, \eta_L^i \tag{15}$$

Where ω_i is derived from the feature aggregation module, mapping through softmax indicates the degree of attention to neighboring lanes in the form of probabilities. The weighted probability ω_i is multiplied by the feature values η_L to obtain lane features with weights η_{LA}. η_{LA} represents the feature value generated by the lane awareness module; $\emptyset_3$ and $\emptyset_4$ denote fully connected layers.

We employ a gated selection mechanism to compute and filter temporal, spatial, and lane-aware features concurrently, specifically through *Sigmoid* gated filtering to extract the effective feature information, multiply it with the original information, and then carry out the residual connection to input it into the layer normalization layers, as illustrated in Eq. (16)-(18):

$$TI = LayerNorm(\eta_{TI} + \eta_{TI} * Sigmoid(\eta_{TI})) \tag{16}$$

$$SI = LayerNorm(\eta_{SI} + \eta_{SI} * Sigmoid(\eta_{SI})) \tag{17}$$

$$LA = LayerNorm(\eta_{LA} + \eta_{LA} * Sigmoid(\eta_{LA})) \tag{18}$$

LayerNorm denotes the layer normalization, and TI, SI, and LA represent the outputs of the final temporal, spatial interaction, and lane awareness modules, respectively. The residual concatenation of feature values from the temporal, spatial interaction, and awareness modules, together with the temporal features of the original target agent, yields η_{all} as the output of the IAB module. Residual connectivity allows information to propagate more directly between different layers of the network and avoids model degradation caused by vanishing gradients [36]. Eq. 19 describes the feature aggregation process for IAB.

$$\eta_{all} = concat(\eta_T, LA, SI, TI) \tag{19}$$

3. Trajectory Predictor (TP)

The TP mainly consists of K LSTMs with a fully connected layer. Its main function is to aggregate the features of the target agent η_T with the features from the interaction and awareness block TI, SI, and LA, resulting in the combined features η_{all}, which is then inputted into the TP module. Using the feature information, the module generates multiple sequences of future trajectories V_{Tar}^f.

$$V_{Tar}^{(f,i)} = \emptyset_7(LSTM(\eta_{all})) \tag{20}$$

The TP module is described in Eq. (20). Among them, $\emptyset_7$ is composed of a multilayer linear network (*Linear*), a normalization layer (*BatchNorm*), and an activation function layer (*Relu*).

C. Realization Details

1. LOSS

We introduce the classification loss L_{class} to constrain behaviors such as lane change and mode choice, the regression loss $L_{regression}$ to constrain the degree of deviation of the predicted trajectories from the true trajectory, and the lane choice loss L_{lc} to encourage the model to choose an appropriate lane. Therefore, the total loss L_{total} function of IAtraj is shown in Eq. (21):

$$L_{total} = L_{class} + L_{regression} + L_{lc} \tag{21}$$

The class loss consists of trajectory modal classification loss L_{cls} and lane classification loss L_{lane_cls}. These two losses are implemented using cross-entropy loss, and the regression loss L_{reg} is implemented using Smooth L1. The class loss and regression loss are shown in Eq. (22) and (23), respectively.

$$\begin{aligned} L_{class} &= aL_{cls}^{K} + bL_{lane\_cls}^{L} \\ &= a\sum_{k=1}^{K} -\pi_{cls}^{k}\log(\hat{\pi}_{cls}^{k}) + b\sum_{l=1}^{L} -\pi_{lane\_cls}^{l}\log(\hat{\pi}_{lane\_cls}^{l}) \end{aligned} \tag{22}$$

$$L_{regression} = c\sum_{k=1}^{k} SmoothL1(\hat{V}_{t}^{k}, V_{t}^{gt}) \tag{23}$$

Where, K denotes the number of modalities, and L denotes the number of candidate lanes. $\pi_{cls/lane\_cls}^{(k/l)}$ denotes the target probability, $\hat{\pi}_{cls/lane\_cls}^{(k/l)}$ is the predicted probability. a, b and c are the weighting factors, balancing the overall impact of multiple factors on the model [37]. V_{t}^{gt} and $\hat{V}_{t}^{k}$ denote the real and predicted trajectories of the target agent, respectively.

$$L_{lc} = d\begin{cases} \frac{1}{H}\sum_{t=1}^{h} \delta(\hat{V}_{t}^{k}, L^{ref}), & if\ \delta(\hat{V}_{t}^{k}, L^{ref}) > \delta(V_{t}^{gt}, L^{ref}), \\ 0, & otherwise. \end{cases} \tag{24}$$

Lane choice loss L_{lc} is utilized to measure the deviation of the predicted trajectories from the lane, as illustrated in Eq. (24). This metric encourages the target agent to approach the reference lane to enhance prediction accuracy when the predicted trajectory's distance from the reference lane exceeds that of the true trajectory, as measured by $\delta(X, Y)$, denoting the distance difference between X and Y. Here, L^{ref} represents the reference lane.

2. Training

The training process for the IAtraj model is performed using the NAdam optimizer and end-to-end training for 33 epochs on NVIDIA RTX 3090 GPUs, taking approximately 4 hours. We use the PyTorch framework to implement the proposed model. To provide a better understanding and implementation, we provide a pseudo-code form algorithm for the entire model, refer to Algorithm 1 for details.

IV. Experiment

A. Dataset

nuScenes: We evaluated the IAtraj model on the large-scale public trajectory prediction dataset nuScenes, which is an automated driving dataset created by nuTonomy, Inc. The dataset comprises 1,000 different scenarios occurring at various times of the day and under different weather conditions, encompassing settings like city streets, highways, parking lots, and more. Each sample in the dataset includes multiple sensor data points and annotated information about associated vehicles, pedestrians, bicycles, and other objects. The maximum length of the dataset is based on the agent's historical trajectory data from the past 2 seconds to predict the target agent's motion trajectories for the next 6 seconds.

Algorithm 1. Trajectory Prediction through Contextual Information Spatio-Temporal Interaction and Awareness

```
Input: Historical trajectory sequences and lane information,
        $X_i, i \in \{V_{Tar}, V_{Sur}, L\}$
Output: Future trajectory sequences
procedure VLFE($X_i, i \in \{V_{Tar}, V_{Sur}, L\}$
    for each $i$ do
        $e_{1D}^{i}$ = Calculate the 1D-CNN embeddings.
        $\eta^{i}$ = Calculate the LSTM outputs.
        $\eta_{agg}^{i}$ = Calculate the aggregation information.
    end for
    return $\eta_{agg}^{i}, i \in \{V_{Tar}, V_{Sur}, L\}$
end procedure
procedure IAB $(\eta^{i}, \eta_{agg}^{i}, i \in \{V_{Tar}, V_{Sur}, L\})$
    $\eta$ = Calculate the feature aggregation output.
    $TI$ = Calculate the Time Interaction Module output.
    $SI$ = Calculate the Spatial Interaction Module output.
    $LA$ = Calculate the Lane Awareness output.
    Concatenate $\eta_T$, $TI, SI, LA$ to generate $\eta_{all}$.
    return $\eta_{all}$
end procedure
procedure TP($\eta_{all}$)
    $V_{Tar}^{F}$ = Calculate decode outputs and generate initial future
    trajectories
    return $V_{Tar}^{F}$
end procedure
if then train == True:
    for each epoch do
        $\eta_{agg}^{i}$ ← VLFE ($X_i, i \in \{V_{Tar}, V_{Sur}, L\}$)
        $\eta_{all}$ ← IAB ($\eta^{i}, \eta_{agg}^{i}, i \in \{V_{Tar}, V_{Sur}, L\}$)
        $V_{Tar}^{F}$ ← TP($\eta_{all}$)
        Calculate Loss and update backward.
    end for
end if
for test dataset do
    $\eta_{agg}^{i}$ ← VLFE ($X_i, i \in \{V_{Tar}, V_{Sur}, L\}$)
    $\eta_{all}$ ← IAB ($\eta^{i}, \eta_{agg}^{i}, i \in \{V_{Tar}, V_{Sur}, L\}$)
    $V_{Tar}^{F}$ ← TP($\eta_{all}$)
    return All future trajectories $V_{Tar}^{F}$.
end for
```

B. Performance Evaluation

In this section, we will introduce two common evaluation metrics for trajectory prediction, **Average Displacement Error** (*ADE*) and **Final Displacement Error** (*FDE*), and use these two evaluation metrics to assess the performance of the proposed model.

Average Displacement Error (*ADE*): The ADE is computed by calculating the average Euclidean distance difference between the true trajectory and the corresponding moment in the predicted trajectories for each moment, reflecting the overall level of prediction effectiveness. In the equation, K denotes the predicted modal number, H denotes the future time step, V_{Tar}^{f}, and V_{Tar} denote the predicted trajectory positions

TABLE I. Comparison Results With Existing State-of-the-art Methods in the NuScenes Test Set

Network	ADE_K			FDE_K		
	$K = 1$	$K = 5$	$K = 10$	$K = 1$	$K = 5$	$K = 10$
AME	-	1.99	1.53	-	4.23	3.08
CoverNet	3.87	1.96	1.48	10.16	-	-
GATraj	-	1.87	1.46	-	4.08	2.97
AgentFormer	-	1.86	1.45	-	3.89	2.86
SGNet	-	1.85	1.32	-	3.87	2.50
ContextVAE	3.54	1.59	-	8.24	3.28	-
Lapred	3.51	1.53	**1.12**	8.12	3.37	2.39
IAtraj(Ours)	**3.27**	**1.48**	1.21	**7.59**	**2.90**	**2.11**

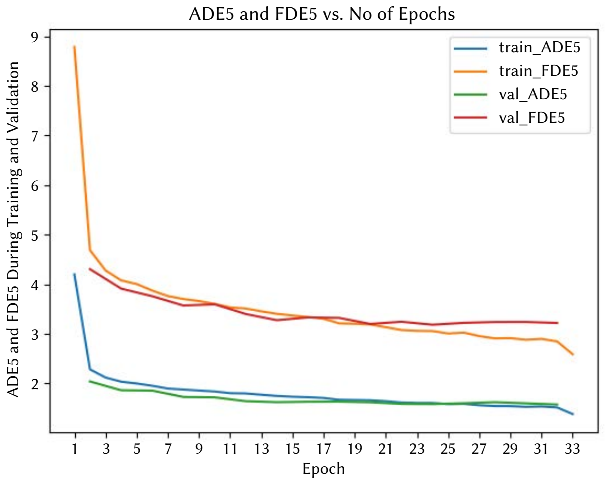


Fig. 3. Change curve of evaluation metrics for training and validation process.

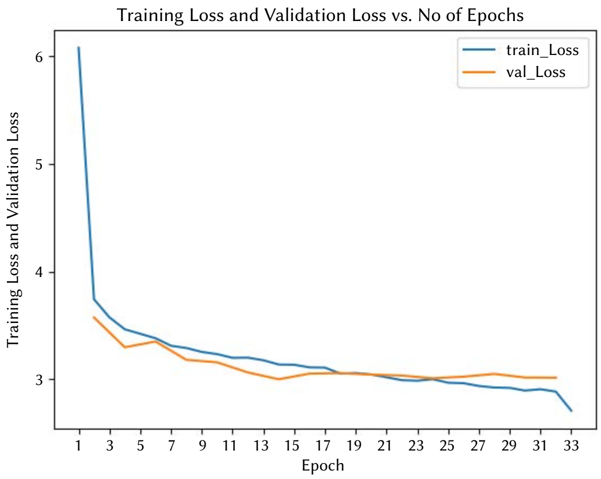


Fig. 4. Loss variation curve for training and validation process.

and true trajectory position of the target agent at time t, respectively. The ADE is calculated as shown in Eq. (25).

$$ADE_K = \frac{1}{K * H} \sum_{k=1}^{K} \sum_{t=1}^{H} || V_{Tar}^{(f)} - V_{Tar} ||_2 \tag{25}$$

Final Displacement Error (*FDE*): The FDE is computed by calculating the Euclidean distance difference between the predicted trajectories and the position of the endpoint of the corresponding true trajectory. Where, K denotes the predicted modal number, $V_{Tar(F)}^{f}$ and $V_{Tar(F)}$ denote the final predicted trajectory positions corresponding to the target agent and the final true trajectory position, respectively. The FDE is calculated as shown in Eq. (26).

$$FDE_K = \frac{1}{K} \sum_{k=1}^{K} || V_{Tar(F)}^{(f)} - V_{Tar(F)} ||_2 \tag{26}$$

C. Comparison

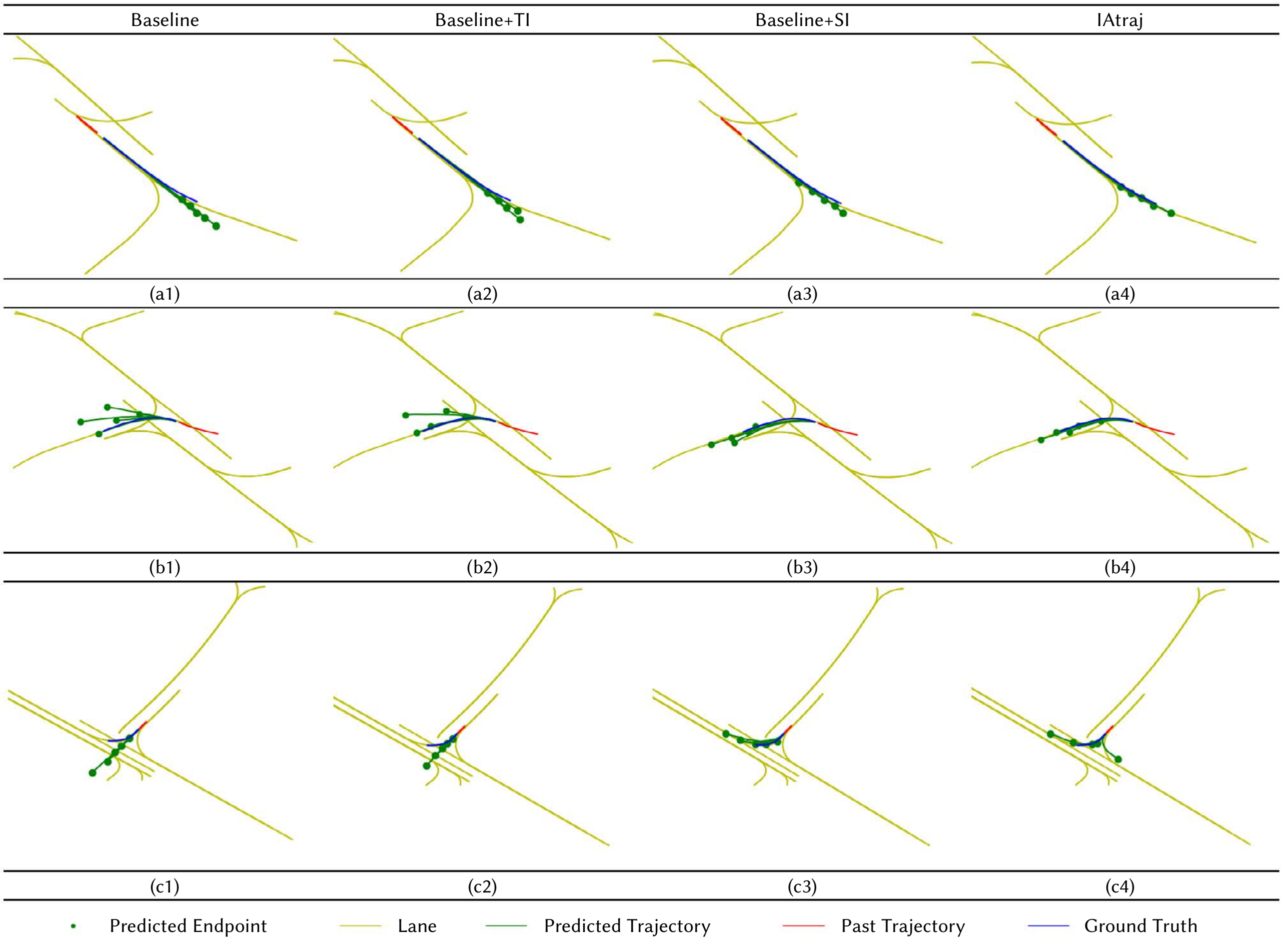
To fully assess the overall performance of our model, we compared the IAtraj model with mainstream models and methods in recent years.

- AME [38]: Relies on bird's eye view (BEV) local perception maps to supplant the need for high-definition maps, avoiding dependency on HD maps and enabling accurate predictions of practical significance.
- CoverNet [32]: Constructs the trajectory prediction problem as a prediction method for classifying different sets of trajectories using trajectory state sequences and raster images as inputs.
- GATraj [10]: Applies Graph Convolutional Networks (GCNs) to simulate interactions between multiple agents, and incorporates the attention mechanism to model the spatio-temporal dynamics of the agents. The method demonstrates excellent prediction speed and efficiency while maintaining prediction accuracy, presenting a graphical model based on the attention mechanism.
- AgentFormer [26]: Joints modeling of temporal and social dimensions mainly using the attention mechanism, through which effective interactions between traffic participants can be achieved, allowing the social behaviors of traffic participants to influence the model of other participants.
- SGNet [39]: The method considers that the agent's movement will change with time, so the designed model mainly performs step-by-step goal estimation and application on multiple time scales for use in predicting successive goals in the future.
- ContextVAE [40]: An environment-aware vehicle trajectory prediction model in real-time is developed utilizing a temporal VAE architecture and map encoding module, generating high-fidelity and effective trajectories corresponding to the given map.
- Lapred [9]: By selecting the target agent along with the potential lanes around it and applying attention computation, the model enhances the effective interactions between the target agent and the lanes, thereby improving the accuracy of prediction.

We evaluated the IAtraj model on the nuScenes dataset and compared it with existing research, whose results are presented in Table I. It can be observed that our proposed method outperforms the other prediction methods in most of the metrics for modal numbers K of 1, 5, and 10. Although Lapred [9] slightly outperforms our method in the ADE_{10} metric, it achieves significant improvement in all other metrics, especially the FDE indicators. We attribute the excellent performance in terms of evaluation metrics to the exquisite and complete design of IAtraj. It differs from models with raster images attempting to extract contextual environment information using image techniques and also differs from models focusing on processing local

TABLE II. Ablation Experiments on the NuScenes Test Set

Ablations	Modules					Metrics			
	Time		Spatial	Awareness	Gate	ADE_K		FDE_K	
	LSTM	TI	SI	LA		$K = 1$	$K = 5$	$K = 1$	$K = 5$
Baseline	✓					4.29	2.08	10.36	4.57
Method A	✓	✓			✓	4.11	1.99	10.06	4.39
Method B	✓		✓	✓	✓	3.39	1.53	7.89	3.06
Method C		✓	✓	✓	✓	3.38	1.51	7.88	2.98
Method D	✓		✓		✓	3.33	**1.47**	7.86	2.93
Method E	✓	✓	✓	✓		3.37	1.50	7.78	2.95
Method F	✓	✓	✓	✓	✓	**3.27**	1.48	**7.59**	**2.90**



(a1) (a2) (a3) (a4)

(b1) (b2) (b3) (b4)

(c1) (c2) (c3) (c4)

Fig. 5. Qualitative analysis of the module ablation study on the nuScenes dataset. Horizontal represents being in the same scenes (respectively straight, left turn, right turn), while vertical means being in the same ablation research models. The past trajectories are shown in red, the ground-truth trajectories are shown in blue, and the predicted trajectories are shown in green. The same applies to the following image.

information. IAtraj focuses on three key elements in traffic scenarios: its historical trajectory, neighboring vehicle behaviors, and lane-keeping. By cleverly modeling these factors, it can effectively simulate variations in real driving scenarios, establish temporal dependencies, and generate plausible prediction results.

D. Ablation Studies

1. Quantitative Analysis

In order to comprehensively study and evaluate the model's overall performance, focusing particularly on the effectiveness of the Time Interaction (TI), Spatial Interaction (SI), and Lane Awareness (LA) modules, we take a step-by-step approach to add modules to verify the predicted performance of the model. The baseline model uses a fully connected layer to replace the entire interaction and awareness block. Based on this, Method A introduces the Temporal Interaction (TI) module in the baseline model to consider only the role of temporal information on the target agent; Methods B and C consider the ability of the LSTM component and the TI component to extract temporal information, respectively; Method D retains the Spatial Interaction (SI) module to assess the degree of influence of the surrounding environment on the target agent; Method E lacks the gating selection mechanism, which reduces the screening and filtering ability of feature information; and Method F includes the complete interaction and awareness block with the best prediction performance. The specific performance evaluation is presented in Table II.

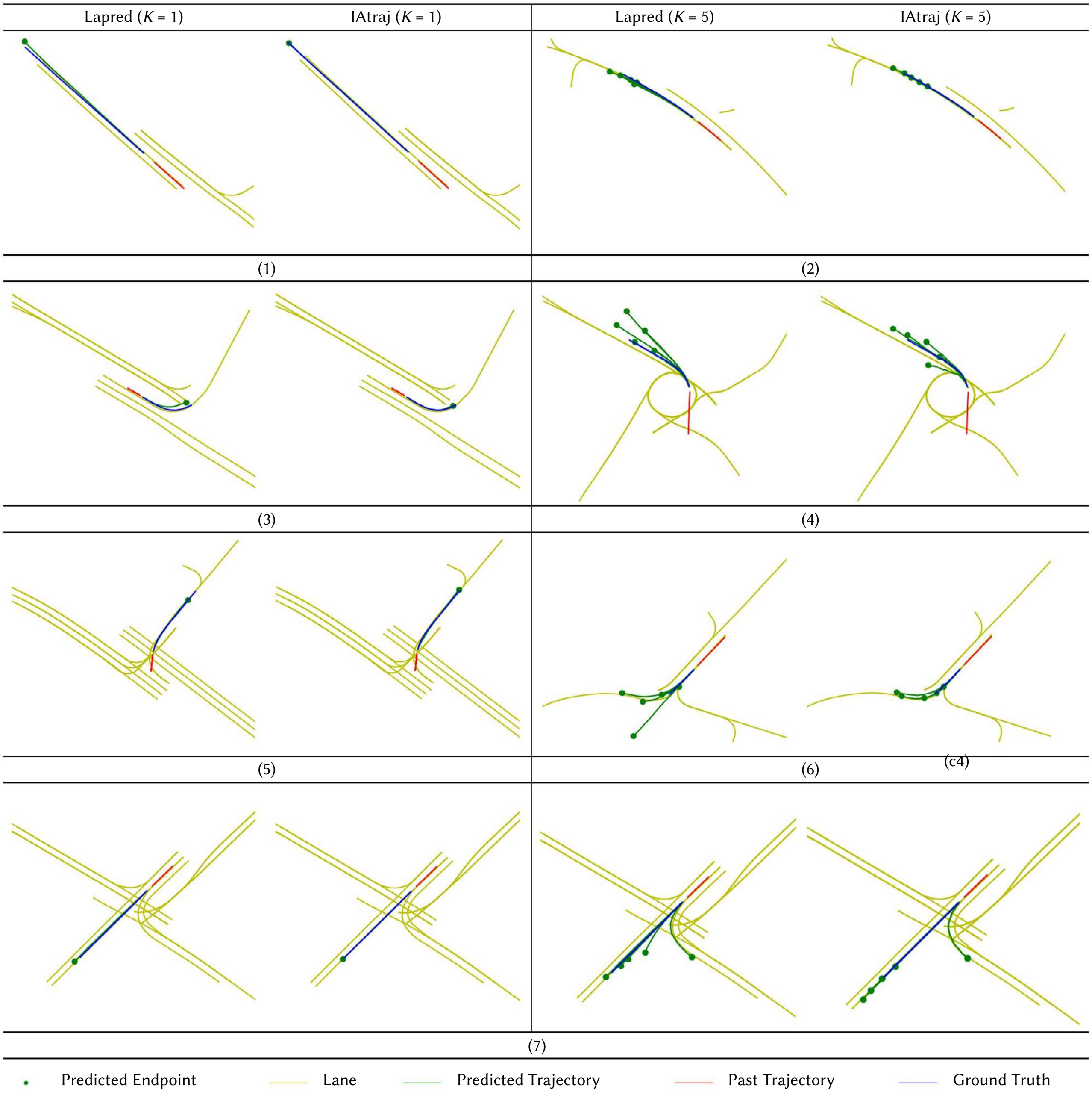


Fig. 6. Qualitative analysis of IAtraj and Lapred on the nuScenes dataset, with the left and right columns representing scenarios with modal numbers K=1 and K=5, respectively. Horizontal means being in the same scenes (respectively straight, left turn, right turn, multimodal variation), while vertical represents being in the same models.

From the data in Table II, it is evident that the baseline model's performance is notably inferior to that of the other models due to its lack of adequate interaction awareness. However, all the improved methods have shown significant enhancements when compared with the baseline method. This suggests the effectiveness of our proposed interaction and awareness block in the trajectory prediction tasks. Notably, Method D, which simply incorporates the Spatial Interaction (SI) module, demonstrates the most substantial performance improvement over the baseline model. This highlights the paramount importance of comprehensively considering and understanding spatio-temporal interaction behaviors around the target agent's driving path in trajectory prediction tasks, aligning with real-world driving situations.

To have a more comprehensive understanding of changes in the IAtraj model training process, we experimented with various configurations. By evaluating the curves of metric changes and loss changes, we observed that the training loss and validation loss almost converged around the 33rd epoch. Therefore, we decided to halt the training at epoch=33. The curves depicting changes in evaluation metrics and loss during the training and validation processes related to the model are displayed in Fig.3, 4.

E. Qualitative Analysis

Fig. 5 visualizes the scenarios of baseline, temporal interaction, spatial interaction, and IAtraj model, respectively. It mainly explores the influence of different modules with varying critical information on the prediction accuracy of specific scenarios. Horizontally observing different ablation research models, a comprehensive understanding of feature information, including time, space, and lanes, proves advantageous in enhancing the accuracy of predictions.

The detailed comparative analysis is shown in Fig. 5. Trajectories predicted by the baseline model are approximate extensions of historical trajectories, suggesting that the target agent relies solely on historical trajectories to continue moving along potentially possible directions. The introduction of the Temporal Interaction (TI) module brings the predicted trajectories closer to the actual trajectory, thereby improving the understanding of historical sequences beyond being mere extensions of historical trajectories. Upon adding the Spatial Interaction (SI) module, the predicted trajectories generally align with actual driving trajectories. However, the judgment of possible neighboring driving lanes and predicted endpoints remains insufficient, particularly in Fig. 5 (b3) where the predicted endpoint location judgment is poor. Interestingly, the comparison between Fig. 5 (c3) and Fig. 5 (c4) illustrates that thorough consideration and comprehension of lane information can generate potential left turn trajectory sequences, thus expanding the richness of modalities.

Fig. 6 compares predictions between IAtraj and Lapred [9] in specific scenarios on the nuScenes large-scale dataset. The left and right columns showcase the prediction scenarios for modalities $K = 1$ and $K = 5$. The overall comparison indicates that our proposed IAtraj model outperforms the Lapred model in predicting trajectories and endpoint positions. Further detailed comparisons reveal that in Scenario (2), the IAtraj model effectively predicts potential variations in agent speed, demonstrating its advantage in longitudinal richness. However, in Scenario (6), regrettably, the IAtraj model's richness is inferior to that of the Lapred model, as it fails to predict potential driving possibilities other than right turns. Interestingly, in Scenario (7), despite the accurate prediction generated in the case of modality $K = 1$, the model still comprehensively understands lane information, proposing potential modes for left turns, with most modes aligning with the real trajectory. This scenario reflects the diverse and rich prediction possibilities in real driving situations influenced by various potential factors.

Based on the qualitative analysis mentioned earlier, our IAtraj model has demonstrated significant improvements over baseline models in terms of accuracy, multimodality, and lane supervision and retention. Particularly noteworthy is its outstanding performance in longitudinal diversity, where our model can accurately predict the target agent's movement at different speeds without deviating from the lane. This result strongly validates the effectiveness of the proposed TI, SI, and LA modules.

V. Conclusions and Future Work

In this work, we introduce a vehicle trajectory prediction model namely IAtraj, based on contextual spatio-temporal interaction and awareness. The model takes into account the historical trajectories of the target agent and surrounding contextual information as inputs. It employs a Temporal Interaction (TI) module to comprehend the temporal dependence within historical trajectories. Simultaneously, the Spatial Interaction (SI) module adapts to the influence of neighboring agents on potential driving trajectories, while the Lane Awareness (LA) module extracts available lane information from the surrounding environment. This facilitates the generation of diverse multi-modal trajectory predictions. Additionally, the feature information undergoes filtration and screening via a gated selection mechanism. Finally, a trajectory predictor generates multi-modal trajectory sequences.

Large-scale experiments on the nuScenes dataset have validated the outstanding performance of the IAtraj model in trajectory prediction tasks, achieving superior results compared to existing studies. Moreover, extensive ablation studies have confirmed the effective representation of both the spatio-temporal interaction and awareness block, providing rich dynamic information and enhancing the understanding of spatio-temporal interactions among multiple agents.

Despite achieving excellent predictive performance in this work, there are still some directions worth exploring. For instance, especially when integrated into terminal vehicles and real-time systems, we need to consider the model's computational and memory limitations. Therefore, we are committed to developing lighter network architectures. Next, in addition to considering history trajectory and environmental factors, we should also pay more attention to predictive performance in variable scenarios, such as school zones, intersections, freeways, etc., which exhibit extensive randomness and immediate driving behaviors. Future predictive models should prioritize the modeling and application of these special scenarios.

Acknowledgment

This research was supported by the Scientific Research Fund of National Natural Science Foundation of China (Grant No. 62372168), Hunan Provincial Natural Science Foundation of China (Grant No. 2023JJ30266), Research Project on teaching reform in Hunan province(No. HNJG-2022-0791) and Hunan University of Science and Technology(No. 2022-44-8), National Social Science Funds of China (19BZX044).

References

[1] L. P. Qian, A. Feng, N. Yu, W. Xu, Y. Wu, "Vehicular networking-enabled vehicle state prediction via two- level quantized adaptive kalman filtering," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7181–7193, 2020.

[2] G. Xie, H. Gao, L. Qian, B. Huang, K. Li, J. Wang, "Vehicle trajectory prediction by integrating physics-and maneuver-based approaches using interactive multiple models," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 7, pp. 5999–6008, 2017.

[3] C. Ju, Z. Wang, C. Long, X. Zhang, D. E. Chang, "Interaction-aware kalman neural networks for trajectory prediction," in *2020 IEEE Intelligent Vehicles Symposium (IV)*, 2020, pp. 1793–1800, IEEE.

[4] J. Chen, C. Zhang, J. Luo, J. Xie, Y. Wan, "Driving maneuvers prediction based autonomous driving control by deep monte carlo tree search," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 7, pp. 7146–7158, 2020.

[5] M. Goldhammer, S. Köhler, S. Zernetsch, K. Doll, B. Sick, K. Dietmayer, "Intentions of vulnerable road users—detection and forecasting by means of machine learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 7, pp. 3035–3045, 2019.

[6] X. Shi, Y. D. Wong, C. Chai, M. Z.-F. Li, "An automated machine learning (automl) method of risk prediction for decision-making of autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 11, pp. 7145–7154, 2020.

[7] F. Altché, A. de La Fortelle, "An lstm network for highway trajectory prediction," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 2017, pp. 353–359, IEEE.

[8] K. Messaoud, I. Yahiaoui, A. Verroust-Blondet, F. Nashashibi, "Attention based vehicle trajectory prediction," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 1, pp. 175–185, 2020.

[9] B. Kim, S. H. Park, S. Lee, E. Khoshimjonov, D. Kum, J. Kim, J. S. Kim, J. W. Choi, "Lapred: Lane-aware prediction of multi-modal future trajectories of dynamic agents," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14636–14645.

[10] H. Cheng, M. Liu, L. Chen, H. Broszio, M. Sester, M. Y. Yang, "Gatraj: A graph-and attention-based multi- agent trajectory prediction model," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 205, pp. 163–175, 2023.

[11] K. Gao, X. Li, B. Chen, L. Hu, J. Liu, R. Du, Y. Li, "Dual transformer based prediction for lane change intentions and trajectories in mixed traffic environment," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 6, pp. 6203–6216, 2023.

[12] Z. Li, Y. Wang, Z. Zuo, "Interaction-aware prediction for cut-in trajectories with limited observable neighboring vehicles," *IEEE Transactions on*

Intelligent Vehicles, vol. 8, no. 3, pp. 2148–2161, 2023.

[13] N. Deo, E. Wolff, O. Beijbom, "Multimodal trajectory prediction conditioned on lane-graph traversals," in *Conference on Robot Learning*, 2022, pp. 203–212, PMLR.

[14] M. Liu, H. Cheng, L. Chen, H. Broszio, J. Li, R. Zhao, M. Sester, M. Y. Yang, "Laformer: Trajectory prediction for autonomous driving with lane-aware scene constraints," *arXiv preprint arXiv:2302.13933*, 2023.

[15] H. Cui, V. Radosavljevic, F.-C. Chou, T.-H. Lin, T. Nguyen, T.-K. Huang, J. Schneider, N. Djuric, "Multimodal trajectory predictions for autonomous driving using deep convolutional networks," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 2090–2096, IEEE.

[16] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, S. Savarese, "Sophie: An attentive gan for predicting paths compliant to social and physical constraints," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1349–1358.

[17] X. Chen, Z. Wang, Q. Hua, W.-L. Shang, Q. Luo, K. Yu, "Ai-empowered speed extraction via port-like videos for vehicular trajectory analysis," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 4, pp. 4541–4552, 2022.

[18] N. Hu, D. Zhang, K. Xie, W. Liang, M.-Y. Hsieh, "Graph learning-based spatial-temporal graph convolutional neural networks for traffic forecasting," *Connection Science*, vol. 34, no. 1, pp. 429–448, 2022.

[19] N. Hu, D. Zhang, K. Xie, W. Liang, C. Diao, K.-C. Li, "Multi-range bidirectional mask graph convolution based gru networks for traffic prediction," *Journal of Systems Architecture*, vol. 133, p. 102775, 2022.

[20] C. Diao, D. Zhang, W. Liang, K.-C. Li, Y. Hong, J.-L. Gaudiot, "A novel spatial-temporal multi-scale alignment graph neural network security model for vehicles prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 1, pp. 904–914, 2022.

[21] J. Gao, C. Sun, H. Zhao, Y. Shen, D. Anguelov, C. Li, C. Schmid, "Vectornet: Encoding hd maps and agent dynamics from vectorized representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11525–11533.

[22] Z. Sheng, Y. Xu, S. Xue, D. Li, "Graph-based spatial- temporal convolutional network for vehicle trajectory prediction in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 17654–17665, 2022.

[23] Z. Li, C. Lu, Y. Yi, J. Gong, "A hierarchical framework for interactive behaviour prediction of heterogeneous traffic participants based on graph neural network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 9102–9114, 2021.

[24] X. Zhou, W. Zhao, A. Wang, C. Wang, S. Zheng, "Spatiotemporal attention-based pedestrian trajectory prediction considering traffic-actor interaction," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 1, pp. 297–311, 2022.

[25] Z. Zhou, J. Wang, Y.-H. Li, Y.-K. Huang, "Query-centric trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17863–17873.

[26] Y. Yuan, X. Weng, Y. Ou, K. M. Kitani, "Agentformer: Agent-aware transformers for socio-temporal multi- agent forecasting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9813–9823.

[27] J. Li, H. Ma, Z. Zhang, J. Li, M. Tomizuka, "Spatio- temporal graph dual-attention network for multi-agent prediction and tracking," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 10556–10569, 2021.

[28] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11621– 11631.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[30] J. Xiang, J. Zhang, Z. Nan, "A fast and map-free model for trajectory prediction in traffics," *arXiv preprint arXiv:2307.09831*, 2023.

[31] S. Zhai, W. Talbott, N. Srivastava, C. Huang, H. Goh, R. Zhang, J. Susskind, "An attention free transformer," *arXiv preprint arXiv:2105.14103*, 2021.

[32] T. Phan-Minh, E. C. Grigore, F. A. Boulton, O. Beijbom, E. M. Wolff, "Covernet: Multimodal behavior prediction using trajectory sets," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14074–14083.

[33] Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961–971.

[34] N. Deo, M. M. Trivedi, "Convolutional social pooling for vehicle trajectory prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1468–1476.

[35] A. Psalta, V. Tsironis, K. Karantzalos, I. Spyropoulou, "Social pooling with edge convolutions on local connectivity graphs for human trajectory prediction in crowded scenes," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 2020, pp. 1–6, IEEE.

[36] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[37] X. Chen, S. Liu, R. W. Liu, H. Wu, B. Han, J. Zhao, "Quantifying arctic oil spilling event risk by integrating an analytic network process and a fuzzy comprehensive evaluation model," *Ocean & Coastal Management*, vol. 228, p. 106326, 2022.

[38] X. Zeng, M. Gao, Z. He, Y. Yang, "Trajectory prediction for surrounding traffic participants via local perception and attentive map encoding," in *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2023, pp. 1–6, IEEE.

[39] C. Wang, Y. Wang, M. Xu, D. J. Crandall, "Stepwise goal- driven networks for trajectory prediction," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2716–2723, 2022.

[40] P. Xu, J.-B. Hayet, I. Karamouzas, "Context-aware timewise vaes for real-time vehicle trajectory prediction," *IEEE Robotics and Automation Letters*, vol. 8, no. 9, pp. 5440–5447, 2023.

Xiaoliang Wang

He received a Ph.D. degree from Hunan University, China. He has worked at Xiangtan University and the Nanjing Government of China. Dr. Wang has also worked as a postdoctoral researcher at the University of Alabama. Currently, he is a professor of information technology and the department chair of Internet of Things Engineering, Hunan University of Science and Technology. His research interests include Information Security and the Internet of Things, such as VANET security, and Anonymous Authentication in Ad Hoc Networks.

Lian Zhou

He is currently pursuing a master's degree at Hunan University of Science and Technology. His research field is intelligent transportation and vehicle trajectory prediction. He has participated and won prizes in the first and second Hunan Graduate Computer Innovation Competition and fourth in the National University Computer Competition.

Kuan-Ching Li

He received a Ph.D. degree from the University of São Paulo, Brazil, in 2001. Currently, Dr. Li is appointed as a Distinguished Professor at Providence University. Besides publications in high-quality conferences and journals, he is a co-author or co-editor of more than 50 books published by well-known publishers. His research interests include parallel and distributed computing, Big Data, Blockchain, and emerging technologies. He is a fellow of the IET, a senior member of the IEEE, and a member of the AAAS.

Shiqi Zheng

He is currently pursuing a master's degree at Hunan University of Science and Technology in China. His research field is autonomous driving. He won the second prize in the China Software Cup Software Design Competition and the Hunan Province Graduate Mathematical Modeling Competition, and published a CCF C conference paper.

Huijing Fan

She is currently pursuing a graduate degree in Software Engineering at Hunan University of Science and Technology, China. She has shown a strong interest in research in the field of artificial intelligence and has focused on exploring autonomous driving in her past research, striving to delve deeper into and solve related problems.