

# Minds on the Move: Decoding Trajectory Prediction in Autonomous Driving with Cognitive Insights

Haicheng Liao\*, Chengyue Wang\*, Kaiqun Zhu, Yilong Ren, Bolin Gao, Shengbo Eben Li, *Senior Member, IEEE*, Chengzhong Xu, *Fellow, IEEE* and Zhenning Li<sup>†</sup>

**Abstract**—In mixed autonomous driving environments, accurately predicting the future trajectories of surrounding vehicles is crucial for the safe operation of autonomous vehicles (AVs). In driving scenarios, a vehicle’s trajectory is determined by the decision-making process of human drivers. However, existing models primarily focus on the inherent statistical patterns in the data, often neglecting the critical aspect of understanding the decision-making processes of human drivers. This oversight results in models that fail to capture the true intentions of human drivers, leading to suboptimal performance in long-term trajectory prediction. To address this limitation, we introduce a Cognitive-Informed Transformer (CITF) that incorporates a cognitive concept, Perceived Safety, to interpret drivers’ decision-making mechanisms. Perceived Safety encapsulates the varying risk tolerances across drivers with different driving behaviors. Specifically, we develop a Perceived Safety-aware Module that includes a Quantitative Safety Assessment for measuring the subject risk levels within scenarios, and Driver Behavior Profiling for characterizing driver behaviors. Furthermore, we present a novel module, Leanformer, designed to capture social interactions among vehicles. CITF demonstrates significant performance improvements on three well-established datasets. In terms of long-term prediction, it surpasses existing benchmarks by 12.0% on the NGSIM, 28.2% on the HighD, and 20.8% on the MoCAD dataset. Additionally, its robustness in scenarios with limited or missing data is evident, surpassing most state-of-the-art (SOTA) baselines, and paving the way for real-world applications.

**Index Terms**—Autonomous Driving, Trajectory Prediction, Perceived Safety, Mixed Autonomy Traffic, Cognitive Modeling

## I. INTRODUCTION

IN the evolving landscape of autonomous driving (AD) systems, the complex interactions between autonomous vehicles (AVs) and human-driven vehicles (HVs) present a significant challenge to achieving accurate trajectory prediction [1], [2]. The future trajectory of human-driven vehicles is essentially the result of the human driver’s decision-making process [3], [4]. Since human drivers require reaction time to adjust their behavior when facing changes in the external environment [5]–[7], the dynamics of the vehicle will not change

drastically in the short term, making **short-term** ( $\leq 2$  seconds) predictions relatively straightforward. Nevertheless, long-term prediction necessitates models that accurately estimate the impact of numerous factors on the decision-making process of human drivers, a feat that is particularly challenging to achieve [8]. Recent advancements in algorithms and the availability of driving datasets have led to significant breakthroughs in trajectory prediction [9]. However, the accuracy of long-term predictions (i.e.,  $>2$  seconds) remains a persistent challenge, primarily due to the inherent complexity of real-world driving scenarios. These challenges stem from the complex interactions between traffic agents, the impact of environmental factors like weather and road conditions, and the unpredictable nature of human driver behavior. These factors introduce significant uncertainty, making reliable long-term forecasts a persistent struggle for researchers in the field.

This backdrop prompts us to ask critical questions about the future trajectory of AD: Is the key to advancing AD not just in accumulating more data or refining algorithms, but in gaining a deeper understanding of the driving environment itself? How can we reshape our models to interpret and respond to the intricate human dynamics that underpin driving? Motivated by these questions, our research embarks on an innovative path. We propose a paradigm shift, extending beyond conventional data-driven approaches to embrace a critical yet often-neglected aspect of driving – the concept of **perceived safety**.

This concept, pivotal in shaping driving behaviors and decisions, is deeply rooted in psychological constructs, as detailed in [10]. According to the Theory of Planned Behavior, individual actions in driving are influenced by attitudes (driving behaviors towards others), subjective norms (personal evaluation of safety), and perceived behavioral control (confidence in driving ability) [11]. Further depth is added by neuroscientific research, such as studies by [12], [13] and [14], which unveil that perceived safety is an intricate blend of both conscious and instinctive responses, involving the amygdala’s emotional processing and the prefrontal cortex’s rational decision-making. Notably, this nuanced understanding of perceived safety is exemplified in diverse driving scenarios. For instance, when encountering a close car ahead, different drivers exhibit markedly varied responses. An aggressive driver, possibly influenced by sensation-seeking tendencies [15], might quickly swerve, perceiving lower risk. Conversely, a cautious driver, perhaps more risk-averse [16], might opt for a complete stop. These behaviors, far from being random, are intricately linked to each driver’s psychological profile and past experiences, revealing a significant limitation in current

<sup>†</sup> Corresponding author; \* Authors contributed equally.

Haicheng Liao, Chengyue Wang, Kaiqun Zhu, Chengzhong Xu, and Zhenning Li are with the State Key Laboratory of Internet of Things for Smart City, University of Macau, Macau. Yilong Ren is with the School of Transportation Science and Engineering, Beihang University, Beijing, China. Bolin Gao and Shengbo Eben Li are with the School of Vehicle and Mobility, Tsinghua University, Beijing, China. E-mails: zhenningli@um.edu.mo. This research is supported by the State Key Lab of Intelligent Transportation System under Project (2024-B001), Science and Technology Development Fund of Macau SAR (File no. 0021/2022/ITP, 0081/2022/A2, 001/2024/SKL), Shenzhen-Hong Kong-Macau Science and Technology Program Category C (SGDX20230821095159012), and University of Macau (SRG2023-00037-IOTSC).

AD systems: their inability to account for these complex, cognitive behavioral patterns. Overall, perceived safety and its influence on the decisions of drivers with different behaviors.

In response, our research introduces the Cognitive-Informed Transformer (CITF) that integrates the concept of perceived safety into trajectory prediction for AVs. This integration does more than add a new variable; it injects a human-centric perspective into the heart of these systems. By doing so, we aim to enhance the models' ability to interpret driving behaviors, leading to optimal long-term predictions. This approach, promises a transformative impact on the predictive capabilities of AD systems, aligning them more closely with the multifaceted nature of human driving behavior.

Overall, the key contributions of this study include:

- We introduce the Quantitative Safety Assessment (QSA) as a cornerstone component for objectively evaluating the safety of driving scenarios. In addition, we establish Driver Behavior Profiling (DBP) upon the QSA framework to differentiate between distinct driver profiles. This DBP effectively captures and interprets continuous nuances in driving behavior, while eliminating the dependence on manual labeling or predefined time windows.
- We introduce an innovative module, named Leanformer, that represents a significant advancement in understanding social interactions on the road. This lightweight transformer-based framework is adept at capturing the subtle and complex inter-vehicular interactions that occur in everyday traffic. This development reflects a paradigm shift in AD research, aligning with the latest advancements and understanding of vehicular social dynamics.
- CITF significantly outperforms the SOTA baseline models when tested on the NGSIM, MoCAD, and HighD datasets. It maintains impressive performance even when trained on only 25% of the dataset and with a much smaller number of model parameters, demonstrating its efficiency and adaptability in various traffic scenes, including highways, campuses, and busy urban locales. Importantly, in a significant stride towards practical applicability, CITF shows unparalleled resilience in scenarios with incomplete or inconsistent data.

## II. RELATED WORK

**Trajectory Prediction For Autonomous Driving.** In the field of trajectory prediction, the analysis of prediction performance is often categorized into short-term and long-term horizons [3], [17]. Early research employed physical models to represent vehicle motion dynamics, thereby estimating future trajectories. In [18], a trajectory prediction model based on the bicycle model was proposed and successfully applied to an accident warning system. While physical models achieved significant progress in short-term prediction horizons, their inherent simplicity limited their performance in long-term predictions [19]. The complexity of human driving behavior, influenced by numerous factors such as cognitive processes, interactions with surrounding vehicles, and environmental conditions, renders long-term prediction a particularly challenging task [17], [20]. In response to these challenges, researchers

have begun integrating deep learning models to account for these factors in the trajectory prediction process. Notable prior efforts [8], [21], [22] have explored the complex social dynamics among traffic participants, revealing crucial latent insights that enhance predictive accuracy. Transformer-based models [23], [24] have been increasingly employed for their ability to predict future trajectory distributions effectively. Graph Neural Networks (GNNs) are also gaining traction for capturing dynamic interactions in complex traffic scenes [25], [26]. These approaches primarily focus on understanding the temporal and spatial interplays between traffic agents from historical data to optimize accuracy. Generative models [27], including Variational Auto Encoders (VAEs), Diffusion models, and Generative Adversarial Networks (GANs), are also being explored for their potential to generate multiple future trajectory possibilities from latent distributions, offering a probabilistic perspective of future paths in this field.

**Perceived Safety Concept.** The notion of perceived safety has been a focal point in psychology and physical human-robot interaction (pHRI) studies [28]. In pHRI, it is crucial for assessing and representing individuals' perceptions of danger and comfort during interactions with autonomous systems like mobile robots [29], industrial manipulators [30], humanoid robots [31] and AVs [32]. Despite its relevance, perceived safety remains a challenging concept to quantify due to its subjective nature [33]. Our study breaks new ground in this area by proposing a novel quantitative criterion for perceived safety in self-driving trajectory prediction, drawing from Safety State Metrics (SSMs) and human decision-making processes. This innovation enables our model to more accurately interpret driving behavior and traffic conditions, thereby enhancing prediction accuracy in mixed autonomy environments.

**Driving Behavior Understanding.** Existing studies in driving behavior have formulated various criteria and metrics for detecting and representing driving patterns, using scales like the Social Value Orientation (SVO) [34], Driving Anger Scale (DAS) [35], among others [36]. While these methods have been successful, as noted by [1] and [37], they typically depend on manually annotated labels and predetermined sliding time windows for analysis. Our research diverges from these traditional approaches by proposing a dynamic, adaptive set of behavior-aware criteria. This model captures driving behavior in real-time through continuous behavioral data representation, eliminating the reliance on manual labeling in the training phase. This novel approach not only offers enhanced flexibility over fixed-category methods but also effectively addresses the challenges of label shifts and time window selection, leading to a more accurate and fluid representation of driving behavior. This advancement significantly contributes to the development of more refined and effective behavior prediction methodologies in autonomous driving systems.

## III. PROBLEM FORMULATION

In mixed autonomy traffic scenarios, trajectory prediction models within AVs are tasked with forecasting the future trajectories of all surrounding vehicles within their perception range. According to surveys by Mozaffari et al. [38] and

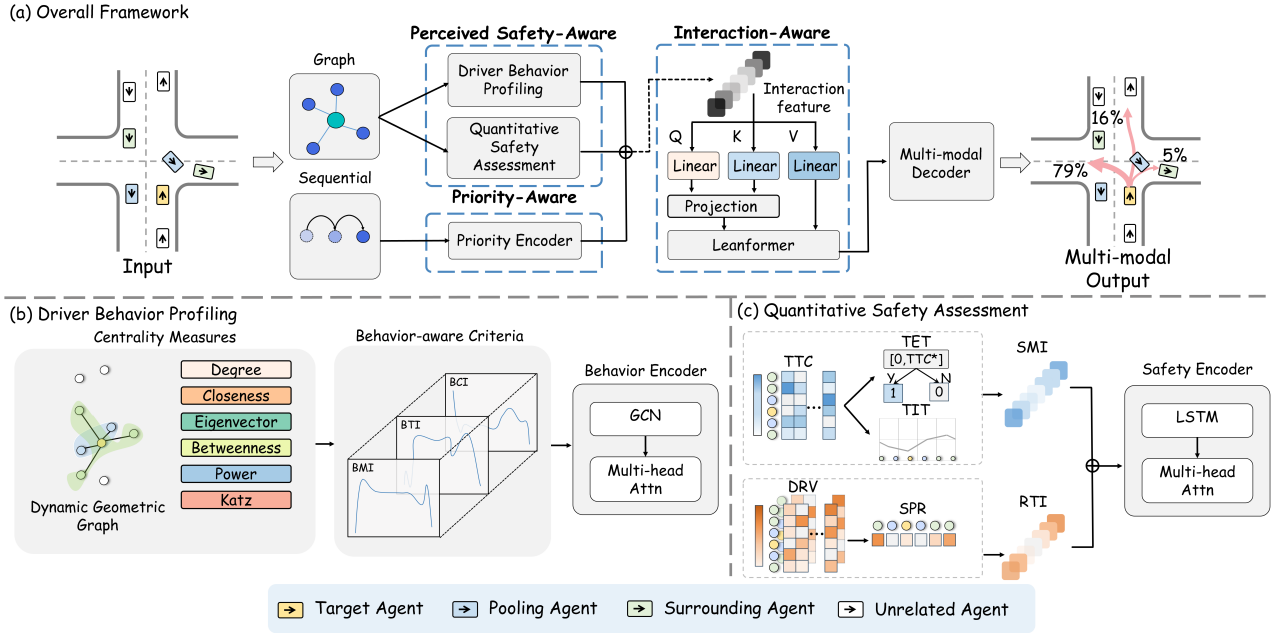


Fig. 1: Pipeline of CITEF. It is an encoder-decoder model (a) and includes four essential parts: Perceived Safety-Aware Module generates both safety and behavior features through driver behavior profiling (b) and quantitative safety assessment (c) components, respectively. These features, along with the priority feature derived from the Priority-Aware Module, are integrated into the Interaction-Aware Module, embedded by the Leanformer framework. Finally, this integration results in a high-level fusion, which is then fed into the Multimodal Decoder to produce a multimodal prediction distribution for the target vehicle.

Ding et al. [39], the single-agent prediction setting remains a prevalent approach in the field of trajectory prediction. In this setting, the model is developed by selecting one vehicle from the surrounding vehicles as the prediction target. During the evaluation phase, the model's predictive capability is assessed in a traversal manner, which treats each vehicle in the scene as the prediction target once. Adhering to this setting, we can define the terminology used in our study as follows:

- Target vehicle: The vehicle is designated as the subject of the trajectory prediction task.
- Surrounding agents: The AV and all of its perceived traffic agents, excluding the target vehicle.

In summary, our problem could be formulated as developing a trajectory prediction model that could utilize the historical states (position, velocity, etc.) of both the target vehicle  $\mathbf{X}_0^{t-t_h:t}$  and the surrounding vehicles  $\mathbf{X}_{1:n}^{t-t_h:t}$  spanning from time  $t-t_h$  to present moment  $t$ , to predict the future trajectory  $\mathbf{Y}_0^{t:t+t_f}$  of the target vehicle over the ensuing  $t_f$  time intervals.

#### A. Discretized Inputs and Outputs

Theoretically, the inputs (historical states) and outputs (future trajectories) should be represented in a continuous form. However, in practical deployment, the sensors on AVs collect data at fixed intervals. Therefore, to maintain consistency with the collected data, it is widely accepted in both academia [40], [41] and industry [42] to use discretized inputs and outputs when developing trajectory prediction models. Specifically, we define the inputs and outputs as follows:

- Inputs: The historical states  $\mathbf{X}_{0:n}^{t-t_h:t}$  of the target vehicle and its surrounding agents, consists of a sequence

of historical states  $\{\mathbf{X}_{0:n}^{t-t_h}, \mathbf{X}_{0:n}^{t-t_h+1}, \dots, \mathbf{X}_{0:n}^t\}$ . At any time  $t$ , the historical states  $\mathbf{X}_{0:n}^t$  comprise 2D position coordinates  $p_{0:n}^t$ , velocity  $v_{0:n}^t$ , and acceleration  $a_{0:n}^t$ .

- Outputs: The predicted trajectory of the target vehicle, denoted as  $\mathbf{Y}_0^{t:t+t_f}$ , consists of a sequence of predicted positions  $\{p_0^{t+1}, p_0^{t+2}, \dots, p_0^{t+t_f-1}, p_0^{t+t_f}\}$ .

#### B. Multi-modal Probabilistic Maneuver Prediction

we adopt a multimodal prediction framework to tackle the inherent uncertainty and variability in predictions. By evaluating different possible maneuvers that the target vehicle might perform, the framework computes the probability of each maneuver based on historical states  $\mathbf{X}_{0:n}^{t-t_h:t}$ , which include 2D position coordinates, velocity, and acceleration over a defined time horizon  $t_h$ . This approach generates multiple predictions while also quantifying the confidence level associated with each prediction. This allows AVs to account for and respond to the uncertainty inherent in prediction outcomes, providing a valuable advantage for decision-making processes.

In this study, we employ a hierarchical Bayesian framework to predict future trajectories. At each time step, we evaluate the probability distribution over the possible maneuvers  $\mathbf{M}$  of the target vehicle. To capture the driver's behavioral nuances, we decompose the vehicle's potential maneuvers into two distinct sub-maneuvers:  $\mathbf{M} = (\mathbf{M}_p, \mathbf{M}_v)$ . Here,  $\mathbf{M}_p$  represents the position sub-maneuver, encompassing three discrete driver decisions: left lane change  $m_l$ , right lane change  $m_r$ , and lane keeping  $m_k$ . Similarly,  $\mathbf{M}_v$  denotes the speed sub-maneuver, with three options: accelerating  $m_a$ , braking  $m_b$ , and maintaining constant speed  $m_c$ .

Following this categorization, the framework generates detailed trajectories for the vehicle conditioned on each maneuver within a predefined distributional form. At each current time  $t$ , we extend the trajectory prediction task to compute the plausible trajectory distribution  $P(Y_0^{t:t+t_f} | M, X_{0:n}^{t-t_h:t})$ . In particular, given the estimated maneuvers  $M$ , the probability distribution of the multimodal trajectory predictions  $Y_0^{t:t+t_f}$  is parameterized as a bivariate Gaussian distribution with the estimable parameters  $\Omega$ :

$$P(Y_0^{t:t+t_f} | M, X_{0:n}^{t-t_h:t}) = P_{\Omega}(Y_0^{t:t+t_f} | M, X_{0:n}^{t-t_h:t}) \quad (1)$$

$$= N(Y_0^{t:t+t_f} | \mu(X_{0:n}^{t-t_h:t}), \Sigma(X_{0:n}^{t-t_h:t}))$$

Here,  $\Omega = [\Omega^{t+1}, \dots, \Omega^{t+t_f}]$ , and  $\Omega^t = [\mu^t, \Sigma^t]$  represents the mean and variance of the distribution of predicted trajectory point at time  $t$ . Correspondingly, the multi-modal predictions are then formulated as a Gaussian Mixture Model:

$$P(Y_0^{t:t+t_f} | M, X_{0:n}^{t-t_h:t}) \quad (2)$$

$$= \sum_{\forall i} P(M_i | X_{0:n}^{t-t_h:t}) P_{\Omega}(Y_0^{t:t+t_f} | M_i, X_{0:n}^{t-t_h:t})$$

where  $M_i$  denote the  $i$ -th element in possible maneuvers  $M$ .

#### IV. TRAJECTORY PREDICTION MODEL

Figure 1 shows the hierarchical framework of CITF. Rooted in the encoder-decoder paradigm, the model seamlessly incorporates four novel modules: the Perceived Safety-Aware Module, the Priority-Aware Module, the Interaction-Aware Module, and the Multimodal Decoder. Collectively, these modules are designed to capture human-machine interactions between the target vehicle and its surrounding agents and emulate the human decision-making process during driving. Detailed overviews of these modules follow.

##### A. Perceived Safety-Aware Module

As mentioned before, perceived safety [43] plays a critical role in human decision-making during driving. The nuances in perceived safety can significantly affect human driver behavior and further impact AV's inability to account for the complex, cognitive behavioral patterns in mixed autonomy environments. Recognizing this, the Perceived Safety-Aware Module is introduced to establish precise criteria and quantify specific criteria for evaluating perceived safety. It consists of two integral components: **1) Quantitative Safety Assessment:** This component focuses on the development of physically based, measurable criteria that can accurately reflect how humans subjectively assess the level of danger; **2) Driver Behavior Profiling:** This component aims to provide in-depth, real-time analysis and profiling of the continuous driving behavior of human drivers especially those influenced by their perceived safety. Together, as shown in Table I, these components are meticulously designed to enhance AVs' understanding of perceived safety in driving contexts, allowing them to better understand and anticipate human driver responses. By incorporating this type of valuable prior knowledge, we facilitate the synthesis of human-like contextual patterns for the proposed model. This enhancement, along with the Priority-Aware Module, allows our Interaction-Aware Module to better

decipher and assimilate the intentions of traffic agents and more closely match the intricacies of human cognition and decision-making in driving scenarios, resulting in improved overall model performance.

**1) Quantitative Safety Assessment:** As shown in Figure 1 (c), this component  $H$  includes two safety indices: the Safe Magnitude Index (SMI) and the Risk Tendency Index (RTI). In a nutshell, the SMI focuses primarily on quantifying the spatio-temporal distance between different agents and the possibility of collision to evaluate the absolute safety level in real-time scenarios. Conversely, RTI tends toward a more subjective analysis, capturing dynamic shifts in safety trends and congestion conditions that reveal potential escalation or mitigation of risk over time.

**Safe Magnitude Index.** In the traffic safety domain, three metrics in SSMs have gained prominence for their comprehensive portrayal of on-road risks: Time-to-Collision, Time Exposed Time-to-Collision (TET), and Time Integrated Time-to-Collision (TIT) [44]. Originating from traffic conflict studies, these metrics are essential tools in microscopic traffic simulations to assess traffic safety. Correspondingly, we synthesize TTC, TIT, and TET into a ternary composite structure within SMI. This composite structure is introduced to evaluate the dynamics of interaction between traffic agents and to estimate the likelihood of potential collisions for each vehicle in real-time scenarios. Specifically, the SMI for an agent at a specific time  $t$  can be expressed as  $S_i^t = [TTC_i^t, TET_i^t, TIT_i^t]$ . To align these with the traffic scenarios, we made slight modifications. Using the 2D position coordinates  $p_i^t, p_j^t$  and velocity  $v_i^t, v_j^t$  for vehicles  $i$  and  $j$  at time  $t$ .

**1) Time-to-Collision:** TTC is a widely accepted measure used to evaluate the time available before two vehicles collide if they continue on their current trajectories. It offers insights into imminent collision risks and serves as an early warning indicator. The TTC for the  $i$ -th vehicle is computed as  $TTC_i^t = -\frac{d_{i,j}^t}{\dot{d}_{i,j}^t}$ , where  $d_{i,j}$  represents the distance between vehicles  $i$  and  $j$ , and  $\dot{d}_{i,j}$  is its rate of change:

$$\begin{cases} \dot{d}_{i,j}^t = \sqrt{(p_i^t - p_j^t)^\top (p_i^t - p_j^t)} \\ \dot{d}_{i,j}^t = \frac{1}{d_{i,j}^t} (p_i^t - p_j^t)^\top (v_i^t - v_j^t) \end{cases} \quad (3)$$

Accordingly, the higher the TTC value, the lower the risk of collision for the vehicle in this case.

**2) Time Exposed Time-to-Collision:** TET measures the exposure duration to critical TTC values within  $t_h$ . It is the sum product of a switching variable and a time threshold  $\tau_{sc}$  (set at 0.1s):  $TET_i^{t_k} = \sum_{t_k=t-t_h}^t \delta_i(t_k) \cdot \tau_{sc}$  with the switching variable given by:

$$\delta_i(t_k) = \begin{cases} 1 & \forall \quad 0 \leq TTC_i^{t_k} \leq TTC^* \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

In our study,  $TTC^* = 3.0s$ , delineating safety threshold.

**3) Time Integrated Time-to-Collision:** An adaptation of TET, TIT integrates the TTC profile to evaluate safety levels. It factors in the evolution of each vehicle's TET temporally:

$$TIT_i^{t_k} = \sum_{t_k=t-t_f}^{t_h} [TTC^* - TTC_i(t_k)] \cdot \tau_{sc} \quad (5)$$

TABLE I: Perceived-Safety criteria and their interpretations.

| Indies                     | Sub-indicators                          | Definition   |
|----------------------------|---|--|
| Safe Magnitude Index (SMI) | Time-to-Collision (TTC)                 | Time until a potential collision occurs between the agent and another agent moving at the current velocity               |
|                            | Time Exposed Time-to-Collision (TET)    | Cumulative sum of instances in which a driver approaches a leading vehicle with a TTC below a predefined threshold value |
|                            | Time Integrated Time-to-Collision (TIT) | Integral of the TTC profile over time if it remains below a specified threshold  |
| Risk Tendency Index (RTI)  | Subjective Risk Perception (SRP)        | Congestion level of each agent in complex traffic environment  |
|                            | Dynamic Risk Volatility (DRV)           | Rate of change of congestion level for each agent in a complex traffic environment                                       |

Elevated values of TTC, TET, and TIT imply sustained exposure to potential collision risks, underscoring a deterioration in perceived safety. Overall, the SMI provides both real-time crash risk assessment and an aggregated risk evaluation over a defined period, eliminating the need for historical crash data. It also takes into account the fluctuation and rate of change of these risks, assessing the safety benefits of AVs in mixed autonomy environments, and offering a comprehensive safety evaluation for each agent.

**Risk Tendency Index.** To further capture congestion patterns in complex traffic environments, we propose an index between the  $i$ -th and  $j$ -th vehicles at time  $t$ , denoted as subjective risk perception indicator (SPR), i.e.  $R_i^t$ , and dynamic risk volatility indicator (DRV), i.e.  $\dot{R}_{i,j}^t$ , respectively:

$$R_i^t = R_j^t = \left[ \log(R_{i,j}^t), \log(\dot{R}_{i,j}^t) \right]^T, \forall i, j \in [0, n], i \neq j \quad (6)$$

In this context, the vector  $R_{i,j}^t$  with larger values indicates an increased risk of collision, while the vector  $\dot{R}_{i,j}^t$  characterizes the dynamic congestion conditions in complex traffic scenarios. Then, the set of the safety indices  $\mathbf{H} = \{S_0^{t-t_h:t}, \mathcal{R}_0^{t-t_h:t}, \dots, S_i^{t-t_h:t}, \mathcal{R}_i^{t-t_h:t} \mid \forall i \in [1, n]\}$  serve as contextual cues and are then fed into the safety encoder for embedding into high-level safety features. The definitions of SPR and DRV are defined as follows:

$$R_{i,j}^t = R_{j,i}^t = \begin{cases} 1/e^{q_{i,j}^t}, q_{i,j}^t > 0 \\ 0, q_{i,j}^t = 0 \end{cases} \quad (7)$$

where the DRV  $\dot{R}_{i,j}^t$  represents the gradient to evaluate fluctuations in SPR  $R_{i,j}^t$  and can be expressed as follows:

$$\dot{R}_{i,j}^t = \dot{R}_{j,i}^t = \begin{cases} 1/e^{\dot{q}_{i,j}^t}, \dot{q}_{i,j}^t > 0 \\ 0, \dot{q}_{i,j}^t = 0 \end{cases} \quad (8)$$

The quantities  $q_{i,j}^t$  and  $\dot{q}_{i,j}^t$  are calculated based on several critical parameters related to the dynamics of two traffic agents. These parameters include the lateral velocity  $v_x^t$ , longitudinal velocity  $v_y^t$ , 2D position coordinate  $p_x^t$  and  $p_y^t$ , as well as the lateral  $a_x^t$  and longitudinal  $a_y^t$ . Mathematically, it can be represented as follows:

$$q_{i,j}^t = \max \left( -\frac{\Delta_{i,j} v_x^t \times \Delta_{i,j} p_x^t + \Delta_{i,j} v_y^t \times \Delta_{i,j} p_y^t}{\Delta_{i,j} v_x^2 + \Delta_{i,j} v_y^2}, 0 \right) \quad (9)$$

$$\dot{q}_{i,j}^t = -\frac{\Delta_{i,j} a_x^t \times \Delta_{i,j} p_x^t + \Delta_{i,j} a_y^t \times \Delta_{i,j} p_y^t}{\Delta_{i,j} a_x^2 + \Delta_{i,j} a_y^2} \quad (10)$$

where the  $\Delta_{i,j}(\cdot)$  denotes the difference between quantities of the  $i$ -th and  $j$ -th vehicles. A larger vector  $R_{i,j}^t$  indicates a

higher risk of collision, while the vector  $\dot{R}_{i,j}^t$  describes the dynamic congestion conditions in complex traffic scene.

**Safety Encoder.** This encoder applies the GCNs [45] to analyze the spatial layouts of traffic agents and their environmental context. Next, it enhances the scaled dot-product multi-head self-attention mechanism [46] for a nuanced analysis of temporal relationships within safety indices.

Specifically, for GCN, we employ a convolutional neural network on a fully connected interaction multigraph to capture the dynamic geometric relationships among traffic agents. This multigraph operational layer sequentially incorporates the set of safety indices  $\mathbf{H}$  as nodes. These nodes represent various security-related properties and states of the traffic agents over time. To establish the connections between these nodes, we use an adjacency matrix  $A$ , which is detailed in the following subsection. This matrix represents the edges of the graph and is critical in defining the interactions and relationships between different nodes (agents) within the graph. Formally,

$$\mathbf{Z}_i^{k+1} = \phi_{\text{ReLU}} \left( \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} \mathbf{Z}_i^k \mathbf{W}_i^k \right) \quad (11)$$

where the matrix  $\tilde{D}$  serves as the scale factor of  $\tilde{A}$ , is the degree matrix for normalizing the graph structure. It helps to balance the influence of each node based on its connectivity. The  $\mathbf{W}_i^k$  represents the trainable weight matrix of the GCN for the  $k$ -th layer, while  $\phi_{\text{ReLU}}$  is the Rectified Linear Unit (ReLU) activation function. Consequently, the matrix  $\tilde{A}$  can be defined as  $\tilde{A} = A + \lambda_A I_N$ , where  $\lambda_A$  is the weight and  $I_N$  is the identity matrix. The output of the  $k$ -th convolutional layer, denoted as  $\mathbf{Z}_i^{k+1}$ , represents the learned feature matrix of the  $i$ -th agent. Moreover, the initial feature  $\mathbf{Z}_i^0 = \phi_{\text{MLP}}(\mathbf{H})$ , where  $\phi_{\text{MLP}}$  denotes a Multi-Layer Perceptron (MLP). The MLP serves as a fully connected layer to embed the safety indices  $\mathbf{H}$  into a feature space suitable for graph convolution. In addition, we employ a tri-layer convolutional neural network that incorporates scatter and gathers operations to parallelize the learning of contextual information and spatio-temporal agent interdependencies.

Next, the feature matrix  $\mathbf{Z}_i^{k+1}$ ,  $\mathbf{Z}_i^k$  and  $\mathbf{Z}_i^{k-1}$  output from the  $(k+1)$ ,  $k$  and  $(k-1)$ -th GCNs is then converted to the query, key, value vectors, respectively, by the multi-head self-attention mechanism within the encoder to produce the high-level safety features. Formally,

$$\begin{cases} Q_i^{\text{safety}} = W^{Q^s} \phi_{\text{MLP}}(\mathbf{Z}_i^{k+1}) \\ K_i^{\text{safety}} = W^{K^s} \phi_{\text{MLP}}(\mathbf{Z}_i^k) \\ V_i^{\text{safety}} = W^{V^s} \phi_{\text{MLP}}(\mathbf{Z}_i^{k-1}) \end{cases} \quad (12)$$

where the  $W^{Q^s}, W^{K^s}, W^{V^s}$  are learnable weights that can be optimized via gradient descent. For the  $i$ -th self-attention head  $head_i$ , the formulation is as follows:

$$head_i^s = \phi_{softmax} \left( \frac{Q_i^{safety} (K_i^{safety})^\top}{\sqrt{d_s}} \right) V_i^{safety} \quad (13)$$

In the equation provided,  $\phi_{softmax}(\cdot)$  denotes the softmax activation function, while  $d_s$  represents the dimensionality of the projected key vectors. The output generated by the self-attention mechanism can be expressed as  $\alpha^{safety} = \sum_{i=1}^{h_s} head_i^s$ , where  $h_s$  is the total number of attention heads.

To increase training stability and efficiency, our model takes inspiration from ResNet [47] and incorporates Gated Linear Units (GLUs) [48] along with Layer Normalization (LN) [49] for the output of multi-head attention mechanism  $\alpha^{safety}$  to efficiently manage features. Formally,

$$\tilde{O}^{t-t_h:t} = \phi_{LN}(\phi_{MLP}(\phi_{GLUs}(\alpha))) \quad (14)$$

In particular, GLUs provide a mechanism to control the flow of information through the network, making the model more adaptable, which can be defined as:

$$\phi_{GLUs}(\alpha) = (\alpha W_1 + b_1) \odot \phi_{sigmoid}(\alpha W_2 + b_2) \quad (15)$$

where  $\alpha$  represents the safe attention coefficient from the multi-head attention mechanism,  $W_1$  and  $W_2$  are the learnable weight parameters associated with the GLUs layer,  $b_1$  and  $b_2$  are the corresponding biases,  $\odot$  denotes element-wise multiplication,  $\phi_{sigmoid}$  is the sigmoid activation function, and  $\phi_{LN}(\cdot)$  stands for Layer Normalization. Correspondingly, the output of the encoder within the Quantitative Safety Assessment is the high-level safety features, denoted as  $\tilde{O}^{t-t_h:t}_{safety}$ .

2) **Driver Behavior Profiling:** As shown in Figure 1 (b), we represent vehicles and their interactions as nodes and edges, respectively, thereby constructing a Dynamic Geometric Graph (DGG). Leveraging this graph-based framework, we employ centrality measures from graph theory to profile continuum driver behavior in an unsupervised manner.

**Dynamic Geometric Graph.** Due to the dynamic nature of traffic scenarios, the structure of the DGG evolves over time. At any given moment  $t$ , we define the DGG  $G^t = \{V^t, E^t\}$ . Specifically, the node set  $V^t = \{v_0^t, v_1^t, \dots, v_n^t\}$ , where node  $v_i^t$  represents vehicle  $i$ . The adjacency matrix  $A^t$  illustrates whether edges exist between nodes, signifying the presence of interactions between vehicles. The establishment of this matrix is based on the distances between vehicles, which can be mathematically represented as follows:

$$A^t(i, j) = \begin{cases} d(v_i^t, v_j^t) & \text{if } d(v_i^t, v_j^t) \leq r \text{ and } i \neq j \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

where  $d(v_i^t, v_j^t)$  denotes the distance between vehicle  $i$  and vehicle  $j$ , and  $r$  is a predefined threshold. The number of vehicles interacting with vehicle  $i$  is represented as  $\mathcal{N}_i^t$ .

With these configurations in place, we then apply centrality measures to assess agent behavior, identify key agents, and evaluate the overall connectivity within the traffic graph.

**Centrality Measures.** Driver behavior significantly shapes the interaction patterns between the driver and surrounding

agents, resulting in distinct spatiotemporal dynamics. Therefore, we posit that spatiotemporal dynamics can effectively differentiate between various driver behaviors. Given that centrality measures in graph theory provide a comprehensive description of the properties of nodes within a graph [50], [51], we employ centrality indicators such as degree  $J_i^t(D)$ , closeness  $J_i^t(C)$ , eigenvector  $J_i^t(E)$ , betweenness  $J_i^t(B)$ , power  $J_i^t(P)$ , and Katz  $J_i^t(K)$  centrality to characterize the spatial interaction dynamics of agent  $i$  at each moment  $t$ . To account for both the temporal and spatial dimensions of these dynamics, we further analyze the temporal evolution of these indicators and establish the Behavior-aware Criteria, enabling the continuous differentiation of diverse driving behaviors.

1) **Degree Centrality:** The number of agents a vehicle can influence reflects its significance within the traffic scene. Degree centrality  $J_i^t(D)$ , a metric that measures the number of connections a node has, is thus naturally employed to describe the importance of each vehicle  $i$ . Formally,

$$J_i^t(D) = |\mathcal{N}_i^t| + J_i^{t-1}(D) \quad (17)$$

where  $|\mathcal{N}_i^t|$  denotes the total agents in  $\mathcal{N}_i^t$ .

2) **Closeness Centrality:** The position of a vehicle within a scene also reflects its significance. It is well-recognized that vehicles located centrally exert greater influence than those at the periphery. Consequently, closeness centrality  $J_i^t(C)$ , which measures the proximity of a node to the center of the graph, is employed to characterize the importance of a vehicle as:

$$J_i^t(C) = \frac{|\mathcal{N}_i^t| - 1}{\sum_{v_j^t \in \mathcal{N}_i^t} d(v_i^t, v_j^t)} \quad (18)$$

3) **Eigenvector Centrality:** The vehicle's behavior can influence a broader set of agents through those it directly interacts with, meaning that the importance of the directly connected agents also reflects the vehicle's significance. Eigenvector centrality, which considers the importance of connected nodes, is used to assess the vehicle's importance. The eigenvector centrality of the vehicle  $i$  can be formulated as follows:

$$J_i^t(E) = \frac{\sum_{v_j^t \in \mathcal{N}_i^t} d(v_i^t, v_j^t)}{\lambda} \quad (19)$$

where  $\lambda$  is the eigenvalue [52].

4) **Betweenness Centrality:** The vehicle's influence can be transmitted to distant agents through intermediary agents, implying that vehicles frequently acting as intermediaries play a more crucial role in the network. Betweenness centrality  $J_i^t(B)$ , a metric that measures the extent to which a node serves as an intermediary within the shortest path between any two nodes, is naturally used to assess the vehicle's importance.

$$J_i^t(B) = \sum_{\forall v_s^t, v_k^t \in V^t} \frac{\sigma_{j,k}(v_i^t)}{\sigma_{j,k}} \quad (20)$$

where  $V^t$  denotes the set of all agents present in the scene,  $\sigma_{j,k}$  signifies the total number of shortest paths between agent  $v_j^t$  and agent  $v_k^t$ , and  $\sigma_{j,k}(v_i^t)$  represents the number of those paths traversing the agent  $v_i^t$ .

5) **Power Centrality:** An interaction loop is a closed loop formed by a group of agents through direct or indirect

interactions. A vehicle's participation in more interaction loops indicates greater influence within the overall traffic network. Power centrality  $\mathcal{J}_i^t(P)$ , which measures the frequency with which a node is part of closed cycles formed by edges, is used to describe the vehicle's influence.

$$\mathcal{J}_i^t(P) = \sum_k \frac{A_{ii}^k}{k!} \quad (21)$$

where  $A_{ii}^k$  denotes the  $i$ -th diagonal element of the adjacency matrix raised to the  $k$ -th power,  $k!$  signifies the factorial.

**6) Katz Centrality:** To address the limitation of degree centrality, which considers only direct interactions, we employ Katz centrality to emphasize both direct and distant interactions of the vehicle. Mathematically, the Katz centrality  $\mathcal{J}_i^t(K)$  of an agent  $v_i^t$  at time  $t$  can be formulated as:

$$\mathcal{J}_i^t(K) = \sum_k \sum_j \alpha^k A_{ij}^k + \beta^k, \forall i, j \in [0, n], \text{ where } \alpha^k < \frac{1}{\lambda_{\max}} \quad (22)$$

where  $n$  is the number of agents in the traffic scenario,  $\alpha^k$  denotes the decay factor,  $\beta^k$  represents weight for immediate neighbors, and  $A_{ij}^k$  is the  $i, j$ -th element of the  $k$ -th power of the adjacency matrix. And  $\lambda_{\max}$  denotes the largest eigenvalue of the adjacency matrix. By carefully selecting the value of the decay factor, Katz centrality can underscore the importance of closer interactions while discounting more distant connections.

**Behavior-aware Criteria.** Given the centrality metrics that capture the spatial interaction dynamics of traffic agents, we establish Behavior-aware Criteria that identify driving behavior not only based on the magnitude of these metrics but also on their temporal variation. Numerous studies have demonstrated the feasibility of this approach, showing that driving behavior can be identified using not only the instantaneous magnitude of features like speed but also their temporal derivatives, such as acceleration and jerk [53]. This approach also aligns with human intuition, as driving behaviors characterized by large and fluctuating centrality measures over short periods are more likely to be relevant to driving behavior as sudden changes in acceleration within short intervals. Inspired by the established triadic relationship between velocity, acceleration, and jerk, we introduce three continuous criteria: Behavior Magnitude Index (BMI)  $\mathcal{C}_i^t$ , which measures the influence of driving behaviors by evaluating their centrality; Behavior Tendency Index (BTI)  $\mathcal{L}_i^t$ , which quantifies behavior propensity by calculating temporal derivatives, larger derivatives suggesting higher probabilities of specific behaviors; and Behavior Curvature Index (BCI)  $\mathcal{I}_i^t$ , which uses the jerk concept to measure the intensity of driving behaviors by calculating the second-order derivatives of continuous centrality measures. At time  $t$ , the behavior  $\mathcal{J}_i^t$  for  $v_i$  can be defined as  $\mathcal{J}_i^t = [\mathcal{C}_i^t, \mathcal{L}_i^t, \mathcal{I}_i^t]^T$ . Each component meticulously evaluates the magnitude, probability, and intensity of diverse driving behaviors exhibited by the target vehicle and its surrounding agents. This assessment is conducted through the computation of threshold rates, gradients, and concavities associated with centrality measures, which capture behaviors such as lane changes, acceleration, and deceleration, as well as aggressive, neutral, or conservative driving tendencies. The underlying

rationale is that driving behaviors characterized by substantial and volatile centrality measure values over short time intervals are more likely to exert a significant influence on nearby agents, emphasizing the temporal dynamics that are integral to human drivers' decision-making processes.

**1) Behavior Magnitude Index.** The BMI is designed to quantify the scale and interconnectedness of various driving behaviors by assessing their centrality measures. The BMI encapsulates the absolute values of these measures, providing a quantitative representation of a behavior's influence on the surrounding traffic agents. Specifically, the BMI focuses on each agent's centrality measures, with a higher index indicating that a particular driving behavior exerts a more significant impact on the current traffic dynamics. Formally, we first formulate the BMI  $\mathcal{C}$  for vehicle  $i$  as follows:

$$\mathcal{C}_i^t = [|\mathcal{J}_i^t(D)|, |\mathcal{J}_i^t(C)|, |\mathcal{J}_i^t(E)|, |\mathcal{J}_i^t(B)|, |\mathcal{J}_i^t(P)|, |\mathcal{J}_i^t(K)|]^T \quad (23)$$

where  $|\cdot|$  denotes the absolute value operator.

**2) Behavior Tendency Index.** Building on the BMI, the BTI incorporates human factors in driving, particularly those behaviors that may cause significant fluctuations in centrality measures, such as aggressive driving, sudden lane changes, or abrupt braking. Specifically, the BTI aims to quantify the propensity for various driving behaviors by calculating their temporal derivatives. By capturing the temporal interaction dynamics of driving behavior, the BTI can identify instances where large gradients and local extrema suggest a higher probability of particular behaviors. This approach enables the model to estimate the likelihood of specific behaviors, even in the absence of explicit behavior classification. Mathematically,

$$\mathcal{L}_i^t = \left| \frac{\partial \mathcal{C}_i^t}{\partial t} \right| = \left[ \left| \frac{\partial \mathcal{J}_i^t(D)}{\partial t} \right|, \left| \frac{\partial \mathcal{J}_i^t(C)}{\partial t} \right|, \dots, \left| \frac{\partial \mathcal{J}_i^t(K)}{\partial t} \right| \right]^T \quad (24)$$

**3) Behavior Curvature Index.** BCI introduces the concept of jerk to quantify the potential impact of driving behavior on surrounding agents. Building upon the BTI, the BCI captures driving behavior by calculating the second derivative of sequential centrality measures. The motivation behind BCI stems from the observation that abrupt changes in BTI over short periods, such as during braking or acceleration, result in peaks in the BCI curve. Additionally, BCI considers the duration of behavior fluctuations, positing that behaviors with prolonged fluctuations have a greater impact on the traffic environment than short-term variations. For instance, a driver who frequently changes lanes or adjusts speed over an extended period may confuse and pressure other drivers, thereby significantly disrupting the dynamic traffic environment.

$$\mathcal{I}_i^t = \left| \frac{\partial \mathcal{L}_i^t}{\partial t} \right| = \left[ \left| \frac{\partial^2 \mathcal{J}_i^t(D)}{\partial^2 t} \right|, \left| \frac{\partial^2 \mathcal{J}_i^t(C)}{\partial^2 t} \right|, \dots, \left| \frac{\partial^2 \mathcal{J}_i^t(K)}{\partial^2 t} \right| \right]^T \quad (25)$$

The introduction of the BMI, BTI, and BCI provides a holistic understanding of individual driving behaviors. Additionally, our proposed behavior-aware criteria eliminate the need for manual labeling during the training phase, effectively mitigating challenges associated with dynamic behavior labels and the selection of appropriate time windows.

**Behavior Encoder.** To leverage the high-level prior knowledge embedded in the Behavior-aware Criteria, we introduce

the behavior encoder to extract behavior features. The Behavior encoder comprises two main components: the LSTM and the multi-head self-attention mechanism. The behavior  $\mathcal{J}_i^{t-t_h:t}$  are first processed by the LSTM, yielding temporal vectors:

$$\tilde{J}_i^{t-t_h:t} = \phi_{LSTM} \left( h_i^{t-t_h:t}, \phi_{MLP}(J_i^{t-t_h:t}), \phi_{MLP}(\bar{O}_{safety}^{t-t_h:t}) \right) \quad (26)$$

Here, the LSTM incrementally updates the hidden state of agent  $v_i$  on a frame-by-frame basis using shared weights. To enhance this process, we incorporate a multi-head self-attention mechanism and GLUs to calculate attention weights across diverse agent behaviors. This approach yields precise sequential behavioral features  $\bar{O}_{behavior}$ , akin to the QSA:

$$\bar{O}_{behavior}^{t-t_h:t} = \phi_{LN} \left( \phi_{MLP}(\phi_{GLUs}(\alpha^{behavior})) \right) \quad (27)$$

where  $\alpha^{behavior}$  is the output of the multi-head self-attention mechanism within the behavior encoder.

### B. Priority-Aware Module

1) *Pooling Mechanism*: In light of recent advances in cognitive studies [8], [54], it has become evident that the spatial positioning of vehicles within a scene can variably influence the behavior and decisions of a target vehicle. For instance, vehicles located directly in the anticipated trajectory path tend to exert greater influence relative to those situated behind. Furthermore, during overtaking maneuvers, vehicles positioned on the left may carry augmented significance. Recognizing these spatial intricacies, we introduce the Priority-Aware Module. This sophisticated module adeptly transforms the spatial coordinates of agents, encoding them into high-dimensional positional vectors, producing positional features.

Our pooling mechanism adeptly amalgamates dynamic positional information from the encompassing traffic scenario, effectively capturing both individual and multi-agent position vectors. This mechanism emphasizes the dynamic nuances of position data, accommodating historical agent states, denoted as  $S_i^{t_k}$ , as well as the intricate spatial interplay symbolized by  $P_{i,j}^{t_k}$ . Mathematically, these relationships are represented as:

$$S_i^{t_k} = \{p_i^{t_k} - p_i^{t_k-1}, v_i^{t_k} - v_i^{t_k-1}, a_i^{t_k} - a_i^{t_k-1}\} \quad (28)$$

Correspondingly,

$$P_{i,j}^{t_k} = \{p_i^{t_k} - p_j^{t_k}, v_i^{t_k} - v_j^{t_k}, a_i^{t_k} - a_j^{t_k}\} \quad (29)$$

By converting the position point sets into sequential vectors, this pooling mechanism effectively indicates the potential spatial relationship between the target agent and its neighboring agents at each scene and time step. This allows the module to accurately represent the necessary interactions over dynamic and spatial position characteristics of each agent.

2) *Priority Encoder*: Within this module, the encoder, which combines both LSTM and multi-head attention mechanisms, processes the dynamic position vectors. This processing involves transforming discrete position vectors into a more continuous spatio-temporal domain, thereby improving the representation of temporal and spatial dynamics. At each discrete temporal instance  $t$ , the encoder assimilates recent historical position vectors via an LSTM network:

$$O_{priority}^{t-t_h:t} = \phi_{LSTM} \left( \bar{h}_i^{t-t_h:t-1}, S_i^{t-t_h:t-1}, P_{i,j}^{t-t_h:t-1} \right) \quad (30)$$

Then, the output of the LSTM is then channeled through a multi-head attention mechanism and GLUs, similar to the quantitative safety assessment, culminating in the synthesis of refined priority features:

$$\bar{O}_{priority}^{t-t_h:t} = \phi_{LN} \left( \phi_{MLP}(\phi_{GLUs}(\alpha^{priority})) \right) \quad (31)$$

where  $\alpha^{priority}$  is the output of the multi-head attention mechanism in the priority encoder.

### C. Interaction-Aware Module

To better understand the synergistic influence of surrounding vehicles' risk levels, positions, and one's behavior on the target vehicle's future trajectory, we introduce an Interaction-Aware Module. This Module is based on a novel lightweight Transformer framework, i.e. Leanformer, which is an adaptation of the Linformer architecture [55]. Departing from traditional Transformer models that rely on fully connected weight matrices, our framework refines the Transformer design, especially in its attention mechanism. By adopting a low-rank matrix approximation, we significantly increase the computational efficiency. This approach effectively reduces the computational complexity of self-attention from  $O(n^2)$  to  $O(n \times d)$ , where  $n$  represents the sequence length and  $d$  denotes a much smaller projected dimension. Mathematically, the interaction is represented as follows:

$$O^{t-t_h:t} = \phi_{MLP}(\bar{O}_{safety}^{t-t_h:t}) \parallel \phi_{MLP}(\bar{O}_{behavior}^{t-t_h:t}) \parallel \phi_{MLP}(\bar{O}_{priority}^{t-t_h:t}) \quad (32)$$

This equation captures the integrated effect of safety perception, behavioral tendencies, and spatial positioning on the target vehicle's trajectory.

As shown in Figure 2, the sequence  $O^{t-t_h:t}$  serves as the input to the Transformer-based framework. The queries  $Q$ , keys  $K$ , and values  $V$  are obtained by performing linear transformations on the input sequence using low-rank projection matrices, as illustrated below:

$$\begin{cases} Q = W^Q(\phi_{MLP}(O^{t-t_h:t} + L_q)) \\ K = W^K(\phi_{MLP}(O^{t-t_h:t} + L_k)) \\ V = W^V(\phi_{MLP}(O^{t-t_h:t})) \end{cases} \quad (33)$$

Here,  $W^Q$ ,  $W^K$ , and  $W^V$  represent low-rank projection matrices for the queries, keys, and values, respectively. Notably, we introduce additional tokens  $L_q$  and  $L_k$  to the end of the queries  $Q$  and keys  $K$  to enhance feature representation and ensure training stability. This augmentation can be mathematically represented as follows:

$$\begin{cases} L_q = \phi_{GRU} \left( \phi_{MLP}(\bar{O}_{safety}^{t-t_h:t}) \parallel \phi_{MLP}(\bar{O}_{behavior}^{t-t_h:t}) \right) \\ L_k = \phi_{GRU} \left( \phi_{MLP}(\bar{O}_{safety}^{t-t_h:t}) \parallel \phi_{MLP}(\bar{O}_{priority}^{t-t_h:t}) \right) \end{cases} \quad (34)$$

where  $\phi_{GRU}$  denotes the GRU framework. The output matrix  $\bar{O}$  is computed as the sum of the outputs from all attention heads, denoted as  $\bar{head}_i$ , where  $i$  ranges from 1 to  $h$ , representing the total number of attention heads. Each attention head has its own set of projection matrices  $W_i^Q$ ,  $W_i^K$ , and  $W_i^V$ , while skip



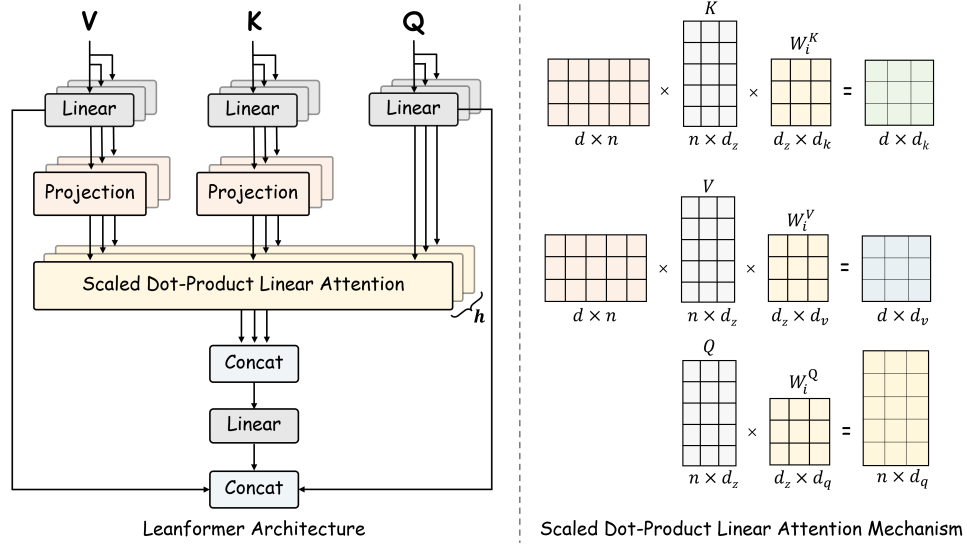


Fig. 2: Architecture of the proposed Leanformer and the scaled dot-product linear attention mechanism.

connections are also used in this framework. This is expressed mathematically as follows:

$$\bar{O} = \sum_{i=1}^h \overline{\text{head}_i} + (\phi_{MLP}(Q) \parallel \phi_{MLP}(V)) \quad (35)$$

In this equation,  $\overline{\text{head}_i}$  refers to the output of the  $i$ -th attention head. The query matrix  $Q \in \mathbb{R}^{n \times d_z}$ , key matrix  $K \in \mathbb{R}^{n \times d_z}$ , and value matrix  $V \in \mathbb{R}^{n \times d_z}$  are all of dimensionality  $n \times d_z$ .

The attention mechanism in each attention head computes a context mapping matrix  $\bar{P}$  of size  $(k \times d)$  using scaled dot-product attention. It involves the following calculations:

$$\bar{P} = \phi_{\text{softmax}} \left( \frac{QW_i^Q (U_i K W_i^K)^T}{\sqrt{d_k}} \right) \quad (36)$$

where  $U_i \in \mathbb{R}^{n \times k}$  denotes a fixed linear projection matrix, while  $d_k$  is the dimensionality of the projected key vectors.

Finally, the output matrix  $\overline{\text{head}_i}$  is obtained by multiplying  $\bar{P}$  with the projected value matrix  $FVW_i^V$  using a linear projection matrix  $F_i \in \mathbb{R}^{n \times k}$ :

$$\overline{\text{head}_i} = \bar{P} FVW_i^V = \phi_{\text{Attention}} \left( QW_i^Q, U_i K W_i^K, F_i V W_i^V \right) \quad (37)$$

This can also be expressed as follows:

$$\overline{\text{head}_i} = \underbrace{\phi_{\text{softmax}} \left( \frac{QW_i^Q (U_i K W_i^K)^T}{\sqrt{d_k}} \right)}_{\bar{P}: n \times k} \cdot \underbrace{F_i V W_i^V}_{k \times d} \quad (38)$$

where the attention weights are obtained by calculating the scaled dot product of the query and key projection matrices, followed by the softmax activation function. The resulting weights are then used to compute the weighted sum of the value projection matrix, which represents the composite interactive vectors  $\bar{O}$  fed into the Multimodal Decoder to generate the future trajectories for the target vehicle.

#### D. Multimodal Decoder

The decoder, rooted in a Gaussian Mixture Model with multimodality, employs a dedicated LSTM and a fully connected layer. It processes the composite interactive vectors  $\bar{O}$  to forecast the target vehicle's trajectory. The predicted trajectory,  $\mathbf{Y}_0^{t:t+t_f}$ , is determined by:

$$\mathbf{Y}_0^{t:t+t_f} = F_\theta (F_\theta(\bar{O})) \quad (39)$$

such that,

$$F_\theta(\cdot) = \phi_{ReLU}(\phi_{MLP}[\phi_{GN}(\phi_{LSTM}(\cdot))]) \quad (40)$$

Here,  $\phi_{GN}$  is Group Normalization, used for improved training stability. The decoder's output comprises multiple future trajectories for the vehicle.

### V. EXPERIMENTS

#### A. Experimental Setups

To validate the prediction capability of our model across different scenarios, we conduct a series of experiments on three widely used traffic datasets: NGSIM, MoCAD, and HighD. The experiments on NGSIM and HighD primarily assess the model's performance in highway scenarios with varying traffic densities, while the experiments on MoCAD focus on its ability to predict in urban and campus-like, unstructured environments. To ensure a fair comparison with existing models [21], [56], we adopt the same training and evaluation protocols as those used in prior benchmarks. Specifically, we define a reference time point, utilizing the preceding three seconds of data as input to the model and the subsequent five seconds as ground truth for supervising model training. In addition, we provide a detailed analysis of our model's performance across both **short-term** ( $\leq 2$  seconds) and **long-term** prediction horizons ( $> 2$  seconds). Moreover, we utilize the Root Mean Square Error (RMSE) as the evaluation metric.

Acknowledging a gap in existing research regarding data omissions in prediction, we develop an innovative approach to

TABLE II: Evaluation results of the proposed model and baseline methods on the **short-term** prediction horizon ( $\leq 2$  seconds). The metric is RMSE (m), with lower values indicating better performance. Cells marked with ('-') indicate data not available. **Bold** and underlined values indicate the best and the second-best performance in each category, respectively.

| Model                       | NGSIM       |             | MoCAD       |             | HighD       |             |
|-----------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                             | 1           | 2           | 1           | 2           | 1           | 2           |
| CS-LSTM [58]                | 0.61        | 1.27        | 1.45        | 1.98        | 0.22        | 0.61        |
| NLS-LSTM [59]               | 0.56        | 1.22        | 0.96        | 1.27        | 0.20        | 0.57        |
| CF-LSTM [60]                | 0.55        | 1.10        | 0.72        | 0.91        | 0.18        | 0.42        |
| iNATran [61]                | 0.39        | 0.96        | -           | -           | <b>0.04</b> | <b>0.05</b> |
| BAT [8]                     | <b>0.23</b> | <b>0.81</b> | <u>0.35</u> | <u>0.74</u> | 0.08        | 0.14        |
| MHA-LSTM [62]               | 0.41        | 1.01        | 1.25        | 1.48        | 0.19        | 0.55        |
| STDAN [3]                   | 0.39        | 0.96        | 0.62        | 0.85        | 0.19        | 0.27        |
| HLTP [12]                   | 0.41        | 0.91        | 0.55        | 0.76        | 0.09        | 0.16        |
| HLTP++ [13]                 | 0.46        | 0.98        | 0.64        | 0.86        | 0.12        | 0.18        |
| WSiP [21]                   | 0.56        | 1.23        | 0.70        | 0.87        | 0.20        | 0.60        |
| <b>CITF</b>                 | <u>0.30</u> | <b>0.81</b> | <b>0.28</b> | <b>0.63</b> | <b>0.04</b> | <u>0.09</u> |
| <b>CITF (drop 3-frames)</b> | 0.38        | 0.86        | <u>0.35</u> | 0.80        | 0.05        | 0.11        |
| <b>CITF (drop 5-frames)</b> | 0.41        | 0.90        | 0.45        | 0.94        | 0.16        | 0.30        |
| <b>CITF (drop 8-frames)</b> | 0.42        | 0.94        | 0.65        | 1.03        | 0.17        | 0.44        |
| <b>CITF (25%)</b>           | 0.42        | 0.93        | 0.55        | 0.96        | 0.08        | 0.21        |

tackle the issue of missing data. We establish the *missing* test set, which is further categorized into three subsets based on the duration of data omissions: *drop 3-frames*, *drop 5-frames*, and *drop 8-frames*. Omissions are purposefully made around the midpoint of the historical trajectory. For instance, in the *drop 5-frames* subset, data ranging from the  $(t - 8)$ -th to the  $(t - 12)$ -th frame is excluded. To manage these omissions, we employ simple linear interpolation. The evaluation results on these subsets are reported as **CITF (drop 3-frames)**, **CITF (drop 5-frames)**, and **CITF (drop 8-frames)**, respectively. Furthermore, to demonstrate the adaptability and efficiency of our model, we train it on a *limited* training set, which contains only 25% of the available training datasets. This model variant is denoted as **CITF (25%)** in this study.

### B. Training and Implementation Details

Our model was trained on an NVIDIA A40 GPU with 48GB of memory. The training process utilized a batch size of 64 and was conducted over 20 epochs. We adopted a dynamic learning rate strategy, initially set at  $10^{-3}$  and gradually reducing to  $10^{-5}$ . The Adam optimizer was employed, coupled with the CosineAnnealingWarmRestarts scheduler to manage the learning rate adjustments. Following the multi-task learning framework [57], our loss function combines the RMSE and the Negative Log-Likelihood (NLL) metrics. The RMSE metric quantifies the average Euclidean distance between predicted and ground truth trajectories, serving as a general measure of predictive accuracy. The NLL metric is particularly valuable for assessing the fidelity of trajectory predictions relative to expected maneuvers, ensuring the reliability of trajectory forecasting within action-based models.

### C. Experiment Results

1) *Performance Comparison on Short-term Horizon*: We present comparative results of our model's prediction performance on the **short-term** prediction horizon ( $\leq 2$  seconds) against existing baseline models on the NGSIM, MoCAD, and HighD datasets in Table II. Our model achieves the best or second-best performance across all three datasets. Among existing approaches, the BAT model demonstrates

the best results on the NGSIM and MoCAD datasets, while the iNATran model performs best on the HighD dataset. Therefore, a comparative analysis of these two models is both essential and representative. Compared to the BAT model, although our model slightly underperforms in the 1-second prediction horizon on the NGSIM dataset, it surpasses BAT on all other datasets. Specifically, on the MoCAD dataset, our model achieves a minimum improvement of 14.8% in short-term prediction horizons, while on the HighD dataset, the improvement is at least 35.8%. Similarly, when compared to the iNATran model, our model falls slightly behind only in the 2-second prediction horizon on the HighD dataset. However, it demonstrates significant improvements on the NGSIM dataset, achieving gains of 23.1% and 15.6% in the 1-second and 2-second prediction horizons, respectively. We further conducted a comprehensive evaluation of the robustness of our proposed model using the missing test set and compared its performance with SOTA baselines. As shown in Table II, CITF outperforms most models, including CS-LSTM and WSiP, on the *missing* test set, even surpassing their performance on the complete test set. Additionally, we assessed the ability of various baselines to handle data missingness challenges using the NGSIM *missing* dataset, with results presented in Table III. A clear trend emerges: as the proportion of missing data increases, the advantages of CITF become even more pronounced. Specifically, CITF outperforms the previous best-performing BAT model by at least 9.0% and 20.0% in the *drop 5-frames* and *drop 8-frames* scenarios, respectively. Overall, while CITF does not show a distinct advantage over previous baseline models in short-term prediction horizons, it significantly outperforms them in the presence of data omissions—a common challenge in real-world applications due to observational constraints.

2) *Performance Comparison on Long-term Horizon*: As shown in Table IV, our model achieves the best performance across all datasets for **long-term** prediction horizons ( $> 2$  seconds). On the NGSIM dataset, our model significantly surpasses all baselines from 2018 to 2024, achieving improvements of 7.8%, 15.7%, and 17.1% for the 3-second, 4-second, and 5-second prediction horizons, respectively. Similarly, results on the MoCAD dataset highlight the strong performance of our model in busy urban traffic scenarios, with improve-

TABLE III: Comparative evaluation of our model on the **short-term** prediction horizon against selected baselines on the *missing* test set of the NGSIM dataset. RMSE (m) is used as the evaluation metric. **Bold** values indicate the best performance, while underlined values indicate the second-best performance in each category.

| Model        | Drop 3-frames |             | Drop 5-frames |             | Drop 8-frames |             |
|--------------|---------------|-------------|---------------|-------------|---------------|-------------|
|              | 1s            | 2s          | 1s            | 2s          | 1s            | 2s          |
| CS-LSTM [58] | 0.67          | 1.47        | 0.75          | 1.52        | 0.84          | 1.72        |
| CF-LSTM [60] | 0.59          | 1.14        | 0.64          | 1.37        | 0.70          | 1.46        |
| WSiP [21]    | 0.60          | 1.29        | 0.69          | 1.37        | 0.76          | 1.58        |
| STDAN [3]    | 0.42          | 1.00        | <u>0.47</u>   | 1.12        | 0.57          | 1.37        |
| BAT [8]      | <b>0.28</b>   | <u>0.88</u> | 0.48          | <u>0.99</u> | <u>0.52</u>   | <u>1.15</u> |
| HLTP [12]    | 0.49          | 1.21        | 0.67          | 1.34        | 0.75          | 1.46        |
| HLTP++ [13]  | 0.48          | 1.09        | 0.63          | 1.26        | 0.70          | 1.34        |
| <b>CITF</b>  | <u>0.38</u>   | <b>0.86</b> | <b>0.41</b>   | <b>0.90</b> | <b>0.42</b>   | <b>0.94</b> |

TABLE IV: Evaluation results of the proposed model and baseline methods on the **long-term** prediction horizon ( $>2$  seconds). The accuracy is measured using RMSE (m), with lower values indicating better performance. Cells marked with ('-') indicate data not available. underlined values indicate the best and the second-best performance in each category, respectively.

| Model                       | NGSIM       |             |             | MoCAD       |             |             | HighD       |             |             |
|-----------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
|                             | 3           | 4           | 5           | 3           | 4           | 5           | 3           | 4           | 5           |
| CS-LSTM [58]                | 2.09        | 3.10        | 4.37        | 2.94        | 3.56        | 4.49        | 1.24        | 2.10        | 3.27        |
| NLS-LSTM [59]               | 2.02        | 3.03        | 4.30        | 2.08        | 2.86        | 3.93        | 1.14        | 1.90        | 2.91        |
| CF-LSTM [60]                | 1.78        | 2.73        | 3.82        | 1.73        | 2.59        | 3.44        | 1.07        | 1.72        | 2.44        |
| iNATran [61]                | 1.61        | 2.42        | 3.43        | -           | -           | -           | 0.21        | 0.54        | 1.10        |
| BAT [8]                     | 1.54        | 2.52        | 3.62        | 1.39        | 2.19        | 2.88        | <u>0.20</u> | 0.44        | 0.62        |
| MHA-LSTM [62]               | 1.74        | 2.67        | 3.83        | 2.57        | 3.22        | 4.20        | 1.10        | 1.84        | 2.78        |
| STDAN [3]                   | 1.61        | 2.56        | 3.67        | 1.62        | 2.51        | 3.32        | 0.48        | 0.91        | 1.66        |
| WSiP [21]                   | 2.05        | 3.08        | 4.34        | 1.70        | 2.56        | 3.47        | 1.21        | 2.07        | 3.14        |
| HLTP++ [13]                 | 1.52        | 2.17        | 3.02        | 1.56        | 2.40        | 3.19        | 0.30        | 0.47        | 0.75        |
| <b>CITF</b>                 | <b>1.42</b> | <b>2.04</b> | <b>2.82</b> | <b>1.09</b> | <b>1.67</b> | <b>2.28</b> | <b>0.18</b> | <b>0.30</b> | <b>0.43</b> |
| <b>CITF (drop 3-frames)</b> | <u>1.51</u> | <u>2.32</u> | <u>2.95</u> | <u>1.24</u> | <u>1.75</u> | <u>2.40</u> | 0.23        | <u>0.35</u> | <u>0.47</u> |
| <b>CITF (drop 5-frames)</b> | 1.52        | 2.40        | 3.31        | 1.30        | 1.80        | 2.72        | 0.43        | 0.64        | 0.92        |
| <b>CITF (drop 8-frames)</b> | 1.65        | 2.45        | 3.51        | 1.63        | 2.13        | 2.98        | 0.83        | 1.25        | 1.72        |
| <b>CITF (25%)</b>           | 1.55        | 2.54        | 3.30        | 1.35        | 2.22        | 3.10        | 0.41        | 0.62        | 0.92        |

TABLE V: Comparative evaluation of our model on the **long-term** prediction horizon against selected baselines on the *missing* test set of the NGSIM dataset. RMSE (m) is used as the evaluation metric. **Bold** values indicate the best performance, while underlined values indicate the second-best performance in each category.

| Model        | Drop 3-frames |             |             | Drop 5-frames |             |             | Drop 8-frames |             |             |
|--------------|---------------|-------------|-------------|---------------|-------------|-------------|---------------|-------------|-------------|
|              | 3s            | 4s          | 5s          | 3s            | 4s          | 5s          | 3s            | 4s          | 5s          |
| CS-LSTM [58] | 2.34          | 3.60        | 4.71        | 2.47          | 3.82        | 4.97        | 2.64          | 3.97        | 5.34        |
| CF-LSTM [60] | 1.82          | 2.77        | 3.91        | 1.94          | 2.83        | 3.98        | 2.21          | 3.10        | 4.47        |
| WSiP [21]    | 2.10          | 3.17        | 4.42        | 2.19          | 3.41        | 4.77        | 2.37          | 3.64        | 5.10        |
| STDAN [3]    | 1.68          | 2.64        | 3.72        | 1.81          | 2.75        | 3.88        | 2.14          | 3.04        | 4.19        |
| BAT [8]      | 1.59          | 2.59        | 3.67        | 1.74          | 2.61        | 3.84        | 1.82          | 2.69        | 3.98        |
| <b>CITF</b>  | <b>1.51</b>   | <b>2.32</b> | <b>2.95</b> | <b>1.52</b>   | <b>2.40</b> | <b>3.31</b> | <b>1.65</b>   | <b>2.45</b> | <b>3.51</b> |

ments of 21.6%, 23.7%, and 20.8% over the same horizons. Furthermore, our model demonstrates substantial performance gains on the HighD dataset, outperforming the BAT and DACR-AMTP models with remarkable improvements of up to 31.8% and 57.4% in RMSE, respectively. In the 5-second prediction horizon, our model surpasses iNATran by an impressive 60.9%. We also investigated the impact of data omissions on the model's long-term prediction performance. As shown in Table IV, even when faced with the *drop 3-frames* and *drop 5-frames* scenarios, CITF outperforms existing baselines on the majority of prediction horizons across all datasets. Specifically, at the 5-second prediction horizon, CITF (drop 3-frames) achieved improvements of 13.2%, 16.7%, and 24.2% on the NGSIM, MoCAD, and HighD datasets, respectively. Table V further illustrates CITF's performance against other models on the missing test set of the NGSIM dataset. CITF outperforms all other models across all prediction horizons, and this advantage increases as the proportion of data omissions grows. On the *drop 8-frames* test set, CITF surpasses the best existing models by 9.3%, 8.9%, and 11.8% at the 3, 4,

and 5-second prediction horizons, respectively. Overall, CITF exhibits better performance in long-term prediction horizons compared to short-term horizons. These results underscore the model's ability to capture long-term intentions and deliver accurate predictions over long-term horizons.

3) *Performance Comparison on **Limited** 25% Training Set:* To challenge our model's adaptability, we trained it using only a quarter of the available training set from the NGSIM, HighD, and MoCAD datasets, yet evaluated its performance on the *complete* test set. Impressively, as shown in Tables II and IV, even with this limited training data, our model delivered RMSE values that were notably lower than most baseline models. Such results underscore our model's efficiency and robustness in trajectory prediction. This performance indicates a promising potential: our model might substantially cut down on the data demands typically associated with training AVs, particularly in scenarios that are data-scarce. In summary, our findings attest to the model's reliability, resource efficiency, and precision in forecasting vehicle trajectories.

TABLE VI: Comparative evaluation of CITF with selected baselines. Highlighting the accuracy metric (Average RMSE (m)) and complexity measured by the number of parameters #Param. (K). **Purple** indicates the performance of our model.

| Model        | Average RMSE (m) |             |             | #Param. (K)   |
|--------------|------------------|-------------|-------------|---------------|
|              | NGSIM            | HighD       | MoCAD       |               |
| CS-LSTM [58] | 2.29             | 1.49        | 2.88        | <b>194.92</b> |
| CF-LSTM [60] | 1.99             | 1.17        | 1.88        | 387.10        |
| WSiP [21]    | 2.25             | 1.44        | 1.86        | 300.76        |
| GaVa [63]    | <u>1.65</u>      | <u>0.39</u> | <u>1.63</u> | 360.75        |
| <b>CITF</b>  | <b>1.48</b>      | <b>0.21</b> | <b>1.00</b> | <b>294.61</b> |

TABLE VII: Inference time comparison of CITF with the SOTA baselines on NGSIM. The inference time is for 10 batches with a size of 128 on two Nvidia A40 48G GPUs. **Purple** indicates the performance of our model.

| Model         | Average RMSE (m) | Inference time (s) |
|---------------|------------------|--------------------|
| CS-LSTM [58]  | 2.29             | 0.22               |
| MHA-LSTM [64] | 1.93             | <b>0.11</b>        |
| TS-GAN [65]   | 2.06             | 0.23               |
| WSiP [21]     | 2.25             | 0.25               |
| STDAN [3]     | 1.87             | 0.22               |
| <b>CITF</b>   | <b>1.48</b>      | <b>0.15</b>        |

4) *Comparative Analysis of Model Performance and Complexity*: As shown in Table VI, our model is benchmarked against several top baselines across three real-world datasets. A notable challenge in this field is the limited availability of efficiency metrics, compounded by restricted access to the source code of various models. Consequently, our comparison mainly focuses on open-source models. Although our model is not the most parameter-efficient, it surpasses all competitors by achieving the lowest average RMSE across all datasets. Remarkably, this high level of accuracy is attained with substantially reduced model complexity—using 18.3% fewer parameters than Gava. Additionally, we assess the inference speed of CITF on the NGSIM dataset. As presented in Table VII, while our model’s inference speed is slightly slower than that of the MHA-LSTM model, it ranks among the most accurate. Specifically, CITF outpaces the previous SOTA model STDAN, with inference speeds 31.8% faster, respectively. These findings highlight CITF’s ability to balance speed and accuracy, further emphasizing its lightweight, efficient, and precise performance in predicting future vehicle trajectories.

#### D. Ablation Studies

We perform a detailed ablation study to assess the specific contributions of each component within our trajectory prediction model. The summarized results are presented in Table IX. Notably, Model F, which integrates all components, consistently outperforms other variations across all evaluation metrics, underscoring the combined value of these components in optimizing performance. In contrast, Model A, which omits the Driver Behavior Profiling within the Perceived Safety-Aware Module, experiences a substantial decline in performance, particularly in short-term predictions, with reductions of at least 19.8% on the NGSIM dataset and 30.7% on the

TABLE VIII: Different components of ablation study.

| Components                    | Ablation methods |   |   |   |   |   |
|-------------------------------|------------------|---|---|---|---|---|
|                               | A                | B | C | D | E | F |
| Driver Behavior Profiling     | ✓                | ✓ | ✓ | ✓ | ✓ | ✓ |
| Perceived Safety-Aware Module | ✓                | ✓ | ✓ | ✓ | ✓ | ✓ |
| Priority-Aware Module         | ✓                | ✓ | ✗ | ✓ | ✓ | ✓ |
| Interaction-Aware Module      | ✓                | ✓ | ✓ | ✓ | ✓ | ✓ |
| Multimodal Decoder            | ✓                | ✓ | ✓ | ✓ | ✗ | ✓ |

TABLE IX: Evaluation results of the ablation analysis for different models on the NGSIM and HighD datasets, with RMSE (m) as the evaluation metric.

| Dataset | Time (s) | Model |      |      |      |      |             |
|---------|----------|-------|------|------|------|------|-------------|
|         |          | A     | B    | C    | D    | E    | F           |
| NGSIM   | 1        | 0.47  | 0.45 | 0.41 | 0.43 | 0.39 | <b>0.30</b> |
|         | 2        | 1.01  | 0.97 | 0.89 | 0.93 | 0.88 | <b>0.81</b> |
|         | 3        | 1.50  | 1.69 | 1.77 | 1.54 | 1.59 | <b>1.42</b> |
|         | 4        | 2.58  | 2.70 | 2.39 | 2.47 | 2.27 | <b>2.04</b> |
|         | 5        | 3.32  | 3.41 | 3.10 | 3.26 | 2.98 | <b>2.82</b> |
| HighD   | 1        | 0.06  | 0.05 | 0.04 | 0.05 | 0.05 | <b>0.04</b> |
|         | 2        | 0.13  | 0.14 | 0.10 | 0.12 | 0.11 | <b>0.09</b> |
|         | 3        | 0.25  | 0.25 | 0.21 | 0.24 | 0.22 | <b>0.18</b> |
|         | 4        | 0.40  | 0.42 | 0.36 | 0.38 | 0.36 | <b>0.30</b> |
|         | 5        | 0.58  | 0.67 | 0.51 | 0.55 | 0.49 | <b>0.43</b> |

HighD dataset. This underscores the critical role of Driver Behavior Profiling in improving trajectory prediction accuracy. Model B, a reduced version of Model F without the Perceived Safety-Aware Module shows a significant reduction in RMSE, especially for long-term predictions, with improvements of at least 16.0% and 28.0% on the NGSIM and HighD datasets, respectively. This highlights the importance of considering perceived safety factors in trajectory prediction, especially for long-term prediction. Model C, which uses absolute coordinates instead of relative positions in the Priority-Aware Module, displays non-negligible reductions in prediction metrics, highlighting the importance of spatial relationships in achieving accuracy. Model D, which lacks the Interaction-Aware Module, shows performance losses of at least 14.8% and 25% on the NGSIM and HighD datasets for short-term prediction, and at least 8.5% and 26.7% for long-term prediction, respectively. Finally, Model E reduces the multimodal probabilistic maneuver prediction in the Decoder, resulting in a performance degradation of at least 5.7% and 13.9% in the NGSIM and HighD datasets, respectively. This suggests their importance in improving prediction accuracy.

#### E. Qualitative Results

Figure 3 compares the performance of CITF with the top baselines, BAT and WSiP, in complex highway scenarios. All three models perform well in short-term predictions, with predicted trajectories closely aligning with the ground truth. However, as shown in Figure 3 (a), in the long-term prediction horizon, CITF (ours) successfully identifies the target vehicle’s intent to merge into an adjacent lane and predicts the overtaking maneuver, producing an accurate forecast. In contrast, both the BAT and WSiP baselines incorrectly predict that the vehicle will continue driving straight. Moreover, in Figures 3 (b-c), facing the intricate dynamics and interactions typical of long-term predictions, such as sudden lane changes, surrounding traffic congestion, and potential collisions, CITF

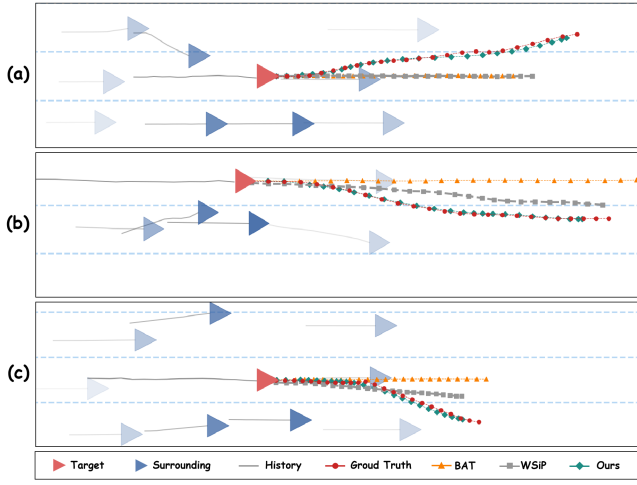


Fig. 3: Visual insights from CITF and top baselines on the NGSIM dataset, illustrating short-term and long-term predictions for three complex driving scenarios: (a) merging, and (b-c) rightward lane change. A darker blue shade indicates an increased risk to the target vehicle, and vice versa.

excels at capturing the subtle influences of driver behavior. It effectively models the traffic scene, recognizing the importance of surrounding vehicles and their impact on the target vehicle's trajectory, thus maintaining high prediction accuracy. These visual results reveal that other models struggle to accurately discern vehicle intentions in complex long-term scenarios. This also further underscores the competitive performance of CITF in short-term predictions, as well as its outstanding capabilities in long-term prediction tasks.

## VI. CONCLUSION

On the journey to fully autonomous driving, long-term trajectory prediction remains a complex challenge. This study introduces an innovative approach rooted in cognitive insights, emphasizing the critical role of perceived safety in driver decision-making. Our Perceived Safety-Aware Module harmoniously merges Quantitative Safety Assessment and Driver Behavior Profiling, offering a detailed perspective on safety perceptions in driving. Rigorous evaluations on real-world driving datasets demonstrated the adaptability of our model, even under data constraints and missing data. In long-term prediction horizons, our proposed CITF significantly outperforms existing models, highlighting the promise of combining computational advantages with human cognitive processes to enhance both the safety and efficiency of autonomous driving. Despite significant progress in prediction accuracy and efficiency, the limitations of our model in short-term low-complexity scenarios warrant further investigation. In such scenes, minimal interactions between the target and surrounding vehicles make short-term dynamics largely governed by basic kinematic principles. A promising avenue for future research is to integrate physical models with deep learning techniques to better account for varying traffic scenes and improve the robustness of the model.

## REFERENCES

- [1] W. Schwarting, A. Pierson, J. Alonso-Mora, S. Karaman, and D. Rus, "Social behavior for autonomous vehicles," *Proceedings of the National Academy of Sciences*, vol. 116, no. 50, pp. 24 972–24 978, 2019.
- [2] H. Liao, H. Shen, Z. Li, C. Wang, G. Li, Y. Bie, and C. Xu, "Gpt-4 enhanced multimodal grounding for autonomous driving: Leveraging cross-modal attention with large language models," *Communications in Transportation Research*, vol. 4, p. 100116, 2024.
- [3] X. Chen, H. Zhang, F. Zhao, Y. Hu, C. Tan, and J. Yang, "Intention-aware vehicle trajectory prediction based on spatial-temporal dynamic attention network for internet of vehicles," *IEEE TITS*, vol. 23, no. 10, pp. 19 471–19 483, 2022.
- [4] H. Liao, Y. Li, Z. Li, Z. Bian, J. Lee, Z. Cui, G. Zhang, and C. Xu, "Real-time accident anticipation for autonomous driving through monocular depth-enhanced 3d modeling," *Accident Analysis & Prevention*, vol. 207, p. 107760, 2024.
- [5] N. Arbabzadeh, M. Jafari, M. Jalayer, S. Jiang, and M. Kharbeche, "A hybrid approach for identifying factors affecting driver reaction time using naturalistic driving data," *TR Part C*, vol. 100, pp. 107–124, 2019.
- [6] Z. Li, Z. Cui, H. Liao, J. Ash, G. Zhang, C. Xu, and Y. Wang, "Steering the future: Redefining intelligent transportation systems with foundation models," *CHAIN*, vol. 1, no. 1, pp. 46–53, 2024.
- [7] H. Liao, Y. Li, C. Wang, Y. Guan, K. Tam, C. Tian, L. Li, C. Xu, and Z. Li, "When, where, and what? a benchmark for accident anticipation and localization with large language models," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 8–17.
- [8] H. Liao, Z. Li, H. Shen, W. Zeng, D. Liao, G. Li, and C. Xu, "Bat: Behavior-aware human-like trajectory prediction for autonomous driving," in *AAAI*, vol. 38, no. 9, 2024, pp. 10 332–10 340.
- [9] S. Casas, C. Gulino, S. Suo, K. Luo, R. Liao, and R. Urtasun, "Implicit latent variable model for scene-consistent motion forecasting," in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*. Springer, 2020, pp. 624–641.
- [10] M. Rubagotti, I. Tusseyeva, S. Baltabayeva, D. Summers, and A. Sandygulova, "Perceived safety in physical human–robot interaction—a survey," *Robotics and Autonomous Systems*, vol. 151, p. 104047, 2022.
- [11] I. Ajzen, "The theory of planned behavior," *Organizational behavior and human decision processes*, vol. 50, no. 2, pp. 179–211, 1991.
- [12] H. Liao, Y. Li, Z. Li, C. Wang, Z. Cui, S. E. Li, and C. Xu, "A cognitive-based trajectory prediction approach for autonomous driving," *IEEE TIV*, vol. 9, no. 4, pp. 4632–4643, 2024.
- [13] H. Liao, Y. Li, Z. Li, C. Wang, G. Li, C. Tian, Z. Bian, K. Zhu, Z. Cui, and J. Hu, "Less is more: Efficient brain-inspired learning for autonomous driving trajectory prediction," in *ECAI 2024*. IOS Press, 2024, pp. 4361–4368.
- [14] S. I. Kronemer, M. Aksent, J. Z. Ding, J. H. Ryu, Q. Xin, Z. Ding, J. S. Prince, H. Kwon, A. Khalaf, S. Forman *et al.*, "Human visual consciousness involves large scale cortical and subcortical networks independent of task report and eye movement activity," *Nature Communications*, vol. 13, no. 1, p. 7342, 2022.
- [15] M. Zuckerman, "The psychophysiology of sensation seeking," *Journal of personality*, vol. 58, no. 1, pp. 313–345, 1990.
- [16] M. Rabin, "Risk aversion and expected-utility theory: A calibration theorem," in *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 2013, pp. 241–252.
- [17] C. Wang, H. Liao, Z. Li, and C. Xu, "Wake: Towards robust and physically feasible trajectory prediction for autonomous vehicles with wavelet and kinematics synergy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [18] M. Brännström, E. Coelingh, and J. Sjöberg, "Model-based threat assessment for avoiding arbitrary vehicle collisions," *IEEE TITS*, vol. 11, no. 3, pp. 658–669, 2010.
- [19] C. Wang, H. Liao, K. Zhu, G. Zhang, and Z. Li, "A dynamics-enhanced learning model for multi-horizon trajectory prediction in autonomous vehicles," *arXiv preprint arXiv:2412.20784*, 2024.
- [20] H. Liao, S. Liu, Y. Li, Z. Li, C. Wang, Y. Li, S. E. Li, and C. Xu, "Human observation-inspired trajectory prediction for autonomous driving in mixed-autonomy traffic environments," in *ICRA*. IEEE, 2024, pp. 14 212–14 219.
- [21] R. Wang, S. Wang, H. Yan, and X. Wang, "Wsp: Wave superposition inspired pooling for dynamic interactions-aware trajectory prediction," in *AAAI*, vol. 37, 2023, pp. 4685–4692.

- [22] H. Liao, Z. Li, C. Wang, B. Wang, H. Kong, Y. Guan, G. Li, and Z. Cui, "A cognitive-driven trajectory prediction model for autonomous driving in mixed autonomy environments," in *Proceedings of the IJCAI-24*. IJCAI, 8 2024, pp. 5936–5944.
- [23] H. Liao, C. Wang, Z. Li, Y. Li, B. Wang, G. Li, and C. Xu, "Physics-informed trajectory prediction for autonomous driving under missing observation," in *Proceedings of the IJCAI-24*, 8 2024, pp. 6841–6849.
- [24] C. Wang, H. Liao, B. Wang, Y. Guan, B. Rao, Z. Pu, Z. Cui, C. Xu, and Z. Li, "Nest: A neuromodulated small-world hypergraph trajectory prediction model for autonomous driving," *arXiv preprint arXiv:2412.11682*, 2024.
- [25] L. Rowe, M. Ethier, E.-H. Dykhne, and K. Czarnecki, "Fjimp: Factorized joint multi-agent motion prediction over learned directed acyclic interaction graphs," in *CVPR*, 2023, pp. 13 745–13 755.
- [26] H. Liao, Z. Li, C. Wang, H. Shen, D. Liao, B. Wang, G. Li, and C. Xu, "Mftraj: Map-free, behavior-driven trajectory prediction for autonomous driving," in *Proceedings of IJCAI*, 2024, pp. 5945–5953.
- [27] H. Liao, X. Li, Y. Li, H. Kong, C. Wang, B. Wang, Y. Guan, K. Tam, and Z. Li, "Cdstraj: Characterized diffusion and spatial-temporal interaction network for trajectory prediction in autonomous driving," in *Proceedings of the IJCAI-24*. IJCAI, 8 2024, pp. 7331–7339, AI for Good.
- [28] J. Guiochet, M. Machin, and H. Waeselynck, "Safety-critical advanced robots: A survey," *Robotics and Autonomous Systems*, vol. 94, pp. 43–52, 2017.
- [29] X. Chen, S. He, Y. Zhang, L. C. Tong, P. Shang, and X. Zhou, "Yard crane and agv scheduling in automated container terminal: A multi-robot task allocation framework," *TR Part C*, vol. 114, pp. 241–271, 2020.
- [30] J. L. Davis, "Role-taking and robotic form: An exploratory study of social connection in human-robot interaction," *International Journal of Human-Computer Studies*, p. 103094, 2023.
- [31] B. Busch, G. Cotugno, M. Khoramshahi, G. Skaltsas, D. Turchi, L. Urbano, M. Wächter, Y. Zhou, T. Asfour, G. Deacon *et al.*, "Evaluation of an industrial robotic assistant in an ecological environment," in *IEEE RO-MAN*. IEEE, 2019, pp. 1–8.
- [32] J. Sun and J. Kim, "Joint prediction of next location and travel time from urban vehicle trajectories using long short-term memory neural networks," *TR Part C*, vol. 128, p. 103114, 2021.
- [33] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International journal of social robotics*, vol. 1, pp. 71–81, 2009.
- [34] R. O. Murphy, K. A. Ackermann, and M. J. Handgraaf, "Measuring social value orientation," *Judgment and Decision making*, vol. 6, no. 8, pp. 771–781, 2011.
- [35] J. L. Deffenbacher, E. R. Oetting, and R. S. Lynch, "Development of a driving anger scale," *Psychological reports*, vol. 74, no. 1, pp. 83–91, 1994.
- [36] O. Taubman-Ben-Ari, M. Mikulincer, and O. Gillath, "The multidimensional driving style inventory—scale construct and validation," *Accident Analysis & Prevention*, vol. 36, no. 3, pp. 323–332, 2004.
- [37] R. Chandra, U. Bhattacharya, T. Mittal, A. Bera, and D. Manocha, "Cmetric: A driving behavior measure using centrality functions," in *2020 IROS*. IEEE, 2020, pp. 2035–2042.
- [38] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakis, "Deep learning-based vehicle behavior prediction for autonomous driving applications: A review," *IEEE TITS*, vol. 23, no. 1, pp. 33–47, 2020.
- [39] Z. Ding and H. Zhao, "Incorporating driving knowledge in deep learning based vehicle trajectory prediction: A survey," *IEEE TIV*, vol. 8, no. 8, pp. 3996–4015, 2023.
- [40] X. Chen, H. Zhang, F. Deng, J. Liang, and J. Yang, "Stochastic non-autoregressive transformer-based multi-modal pedestrian trajectory prediction for intelligent vehicles," *IEEE TITS*, 2023.
- [41] H. Zhou, X. Yang, D. Ren, H. Huang, and M. Fan, "Csir: Cascaded sliding cvae with iterative socially-aware rethinking for trajectory prediction," *IEEE TITS*, 2023.
- [42] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *CVPR*, 2019, pp. 8748–8757.
- [43] G. S. Nair and C. R. Bhat, "Sharing the road with autonomous vehicles: Perceived safety and regulatory preferences," *TR Part C*, vol. 122, p. 102885, 2021.
- [44] M. M. Minderhoud and P. H. Bovy, "Extended time-to-collision measures for road traffic safety assessment," *Accident Analysis & Prevention*, vol. 33, no. 1, pp. 89–97, 2001.
- [45] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.
- [46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
- [48] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International conference on machine learning*. PMLR, 2017, pp. 933–941.
- [49] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
- [50] S. P. Borgatti, M. G. Everett, and J. C. Johnson, *Analyzing Social Networks*. SAGE Publications Limited, 2018.
- [51] J. Zhang and Y. Luo, "Degree centrality, betweenness centrality, and closeness centrality in social network," in *2017 2nd international conference on modelling, simulation and applied mathematics (MSAM2017)*. Atlantis press, 2017, pp. 300–303.
- [52] U. N. Pillai and L.-H. Lim, "Perron-frobenius theorem for nonnegative tensors," *Communications in Mathematical Sciences*, vol. 3, no. 3, pp. 531–539, 2005.
- [53] Y. L. Murphey, R. Milton, and L. Kiliaris, "Driver's style classification using jerk analysis," in *2009 IEEE workshop on computational intelligence in vehicles and vehicular systems*. IEEE, 2009, pp. 23–28.
- [54] D. P. Broadbent, G. D'Innocenzo, T. J. Ellmers, J. Parsler, A. J. Szameitat, and D. T. Bishop, "Cognitive load, working memory capacity and driving performance: A preliminary fMRI and eye tracking study," *TR Part F*, vol. 92, pp. 121–132, 2023.
- [55] S. Wang, B. Z. Li, M. Khabza, H. Fang, and H. Ma, "Linformer: Self-attention with linear complexity," *arXiv preprint arXiv:2006.04768*, 2020.
- [56] K. Gao, X. Li, B. Chen, L. Hu, J. Liu, R. Du, and Y. Li, "Dual transformer based prediction for lane change intentions and trajectories in mixed traffic environment," *IEEE TITS*, 2023.
- [57] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *CVPR*, 2018, pp. 7482–7491.
- [58] N. Deo and M. M. Trivedi, "Convolutional social pooling for vehicle trajectory prediction," in *CVPR Workshops*, 2018, pp. 1468–1476.
- [59] K. Messaoud, I. Yahiaoui, A. Verroust-Blondet, and F. Nashashibi, "Non-local social pooling for vehicle trajectory prediction," in *2019 IEEE IV*. IEEE, 2019, pp. 975–980.
- [60] X. Xie, C. Zhang, Y. Zhu, Y. N. Wu, and S.-C. Zhu, "Congestion-aware multi-agent trajectory prediction for collision avoidance," in *2021 ICRA*. IEEE, 2021, pp. 13 693–13 700.
- [61] X. Chen, H. Zhang, F. Zhao, Y. Cai, H. Wang, and Q. Ye, "Vehicle trajectory prediction based on intention-aware non-autoregressive transformer with multi-attention learning for internet of vehicles," *IEEE TIM*, vol. 71, pp. 1–12, 2022.
- [62] K. Messaoud, I. Yahiaoui, A. Verroust-Blondet, and F. Nashashibi, "Attention based vehicle trajectory prediction," *IEEE TIV*, vol. 6, no. 1, pp. 175–185, 2021.
- [63] H. Liao, S. Liu, Y. Li, Z. Li, C. Wang, B. Wang, Y. Guan, and C. Xu, "Human observation-inspired trajectory prediction for autonomous driving in mixed-autonomy traffic environments," *ICRA*, 2024.
- [64] K. Messaoud, I. Yahiaoui, A. Verroust-Blondet, and F. Nashashibi, "Attention based vehicle trajectory prediction," *IEEE TIV*, vol. 6, no. 1, pp. 175–185, 2020.
- [65] Y. Wang, S. Zhao, R. Zhang, X. Cheng, and L. Yang, "Multi-vehicle collaborative learning for trajectory prediction with spatio-temporal tensor fusion," *IEEE TITS*, vol. 23, no. 1, pp. 236–248, 2020.





**Haicheng Liao** (Student Member, IEEE) received the B.S. degree in software engineering from the University of Electronic Science and Technology of China (UESTC) in 2022. He is currently pursuing the Ph.D. degree at the State Key Laboratory of Internet of Things for Smart City and the Department of Computer and Information Science, University of Macau. Over his academic career, he has published over 20 papers. His research interests include connected autonomous vehicles and the application of deep reinforcement learning to autonomous driving.



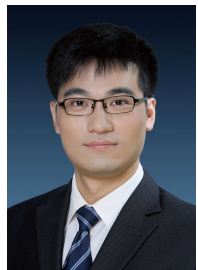
**Chengyue Wang** is currently pursuing a Ph.D. degree at the State Key Laboratory of Internet of Things for Smart City and the Department of Civil Engineering, University of Macau. He received his M.S. degree in civil engineering from the University of Illinois Urbana-Champaign (UIUC) in 2022. He received his B.E. degree in transportation engineering from Chang'an University in 2021. His research interests include connected autonomous vehicles and the application of deep reinforcement learning to autonomous driving.



**Kaiqun Zhu** received the Ph.D. degree in control science and engineering from the University of Shanghai for Science and Technology, Shanghai, China, in 2022. From 2020 to 2022, he was a visiting Ph.D. student with the Department of Computer Science, Brunel University London, Uxbridge, U.K. He is currently a Postdoctoral Fellow with the University of Macau, Macau, China. His research interests include set-membership filtering, model predictive control, neural networks, privacy preserving, and their applications in autonomous vehicles.



**Yilong Ren** (Member, IEEE) received the B.S. and Ph.D. degrees from Beihang University, Beijing, China, in 2010 and 2017, respectively. He is currently an Associate Professor with the Research Institute for Frontier Science, Beihang University. His research interests include vehicular communications, vehicular crowd sensing, and traffic Big Data.



**Bolin Gao** received the B.S. and M.S. degrees in Vehicle Engineering from Jilin University, Changchun, China, in 2007 and 2009, respectively, and the Ph.D. degree in Vehicle Engineering from Tongji University, Shanghai, China, in 2013. He is now an associate research professor at the School of Vehicle and Mobility, Tsinghua University. His research interests include the theoretical research and engineering application of the dynamic design and control of intelligent and connected vehicles, especially about the collaborative perception and tracking

method in cloud control system, intelligent predictive cruise control system on Commercial trucks with cloud control mode, as well as the test and evaluation of intelligent vehicle driving system.



**Shengbo Li** (Senior Member, IEEE) received his M.S. and Ph.D. degrees from Tsinghua University in 2006 and 2009, respectively. Before joining Tsinghua University, he had worked at Stanford University, University of Michigan, and UC Berkeley. His active research interests include intelligent vehicles and driver assistance, deep reinforcement learning, optimal control and estimation, etc. He is the author of over 130 peer-reviewed journal/conference papers and the co-inventor of over 30 patents. He is the recipient of the best (student) paper awards of IEEE ITSC, ICCAS, IEEE ICUS, CCCC, etc. His important awards include the National Award for Technological Invention of China (2013), the Excellent Young Scholar of NSF China (2016), the Young Professor of ChangJiang Scholar Program (2016), the National Award for Progress in Sci & Tech of China (2018), Distinguished Young Scholar of Beijing NSF (2018), Youth Sci & Tech Innovation Leader from MOST (2020), etc. He also serves as the Board of Governor of the IEEE ITS Society, Senior AE of IEEE OJ ITS, and AEs of IEEE ITSM, IEEE Trans ITS, Automotive Innovation, etc.



**Chengzhong Xu** (Fellow, IEEE) received the Ph.D. degree from The University of Hong Kong, in 1993. He is currently the chair professor of computer science and the dean with the Faculty of Science and Technology, University of Macau. Prior to this, he was with the faculty at Wayne State University, USA, and the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China. He has published more than 400 papers and more than 100 patents. His research interests include cloud computing and data-driven intelligent applications. He was the Best Paper awardee or the Nominee of ICPP2005, HPCA2013, HPDC2013, Cluster2015, GPC2018, UIC2018, and AIMS2019. He also won the Best Paper award of SoCC2021. He was the Chair of the IEEE Technical Committee on Distributed Processing from 2015 to 2019.



**Zhenning Li** (Member, IEEE) received his Ph.D. in Civil Engineering from the University of Hawaii at Manoa, Honolulu, Hawaii, USA, in 2019. Currently, he holds the position of Assistant Professor at the State Key Laboratory of Internet of Things for Smart City, as well as the Department of Computer and Information Science at the University of Macau, Macau. His main areas of research focus on the intersection of connected autonomous vehicles and Big Data applications in urban transportation systems. He has been honored with several awards, including the Macau Science and Technology Award, Chinese Government Award for Outstanding Self-financed Students Abroad, TRB best young researcher award and the CICTP best paper award, amongst others.