

Knowledge-Informed Multi-Agent Trajectory Prediction at Signalized Intersections for Infrastructure-to-Everything

Huilin Yin¹, Yangwenhui Xu¹, Jiaxiang Li¹, Hao Zhang¹, Gerhard Rigoll²

¹Tongji university

²Technical University of Munich

{yinhuilin, xuyangwenhui}@tongji.edu.cn

Abstract

Multi-agent trajectory prediction at signalized intersections is crucial for developing efficient intelligent transportation systems and safe autonomous driving systems. Due to the complexity of intersection scenarios and the limitations of single-vehicle perception, the performance of vehicle-centric prediction methods has reached a plateau. Furthermore, most works underutilize critical intersection information, including traffic signals, and behavior patterns induced by road structures. Therefore, we propose a multi-agent trajectory prediction framework at signalized intersections dedicated to Infrastructure-to-Everything (I2XTraj). Our framework leverages dynamic graph attention to integrate knowledge from traffic signals and driving behaviors. A *continuous signal-informed mechanism* is proposed to adaptively process real-time traffic signals from infrastructure devices. Additionally, leveraging the prior knowledge of the intersection topology, we propose a *driving strategy awareness mechanism* to model the joint distribution of goal intentions and maneuvers. To the best of our knowledge, I2XTraj represents the first multi-agent trajectory prediction framework explicitly designed for infrastructure deployment, supplying subscribable prediction services to all vehicles at intersections. I2XTraj demonstrates state-of-the-art performance on both the Vehicle-to-Infrastructure dataset V2X-Seq and the aerial-view dataset SinD for signalized intersections. Quantitative evaluations show that our approach outperforms existing methods by more than 30% in both multi-agent and single-agent scenarios.

1 Introduction

Trajectory prediction is vital in enabling autonomous vehicles to navigate safely and efficiently[Cao, 2024]. Signalized intersections, which account for a disproportionate number of traffic accidents[Ekmekci *et al.*, 2024], present particularly complex challenges for autonomous systems. Infrastructure-based trajectory prediction systems, enabled by Vehicle-to-Everything (V2X) communication, can provide

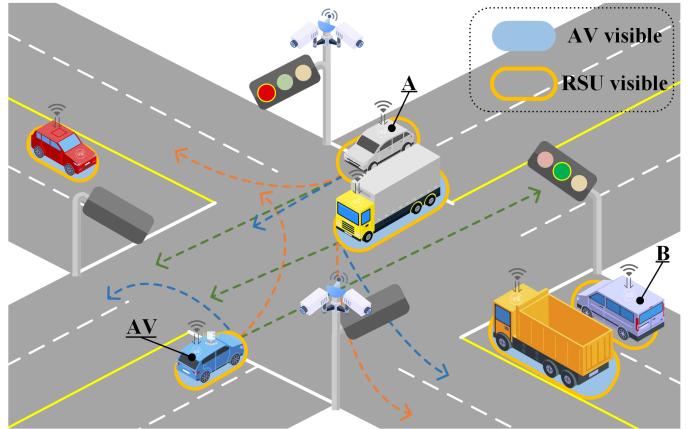


Figure 1: Schematic illustration of Infrastructure-to-Everything (I2X) trajectory prediction at a signalized intersection. Autonomous Vehicles (AVs) cannot observe all vehicles. Roadside Units (RSUs) can collect comprehensive data and knowledge, including vehicle states, real-time traffic signals, and intersection behavior patterns. Infrastructure generates accurate scene trajectory predictions shared with vehicles.

centralized trajectory prediction services to all vehicles approaching intersections. This approach not only enhances autonomous driving capabilities to reduce the occurrence of accidents but also significantly improves V2X communication dependability[Zeng *et al.*, 2024].

Many studies propose vehicle-centric trajectory prediction models for highways[Liao *et al.*, 2024a], urban roads[Gu *et al.*, 2021], and unsignalized intersections[Rowe *et al.*, 2023]. These models show promising results in various driving scenarios. However, few studies address trajectory prediction at signalized intersections, where traffic control introduces two critical challenges. On the one hand, vehicle behavior patterns at intersections differ significantly from other urban roads due to road topology and traffic control[Singh *et al.*, 2022]. This behavioral complexity requires a redundant prediction model to handle signalized intersection scenarios, placing a heavy computational burden on autonomous vehicles' limited on-board resources. Infrastructure-deployed prediction models that broadcast predictive services offer a solution by centralizing these com-

plex computations away from individual vehicles. On the other hand, due to inherent limitations in sensor coverage and field of view, individual on-board sensing systems cannot effectively capture the complete dynamic traffic environment at complex intersection scenarios. As Figure 1, while vehicles A and B are occluded from the autonomous vehicle’s (AV) field of view, they remain detectable by the roadside unit (RSU). With V2X technology, external information from other vehicles and infrastructures has been used to facilitate more accurate and reliable trajectory predictions than vehicle-based systems[Ruan *et al.*, 2023; Annunziata *et al.*, 2024]. Furthermore, the unidirectional transmission from infrastructure-to-vehicle(I2V) addresses privacy concerns associated with bidirectional vehicle-to-vehicle(V2V) and vehicle-to-infrastructure(V2I) communication. Infrastructure-to-Everything(I2X) trajectory prediction can facilitate communication device operations and enhance communication quality by enabling proactive network adjustments and optimized resource allocation[Sarkar and Taherkordi, 2023; Wang *et al.*, 2024a].

Therefore, we propose I2XTraj, an infrastructure-based trajectory prediction model for vehicles at signalized intersections. This approach also represents a novel paradigm for collaborative trajectory prediction. Our method performs trajectory prediction in a knowledge-driven manner with dynamic graph attention by leveraging real-time traffic information and prior behavior knowledge from infrastructure devices: (1) Real-time traffic signals. The *continuous signal-informed mechanism* is developed to encode signal sequences obtained directly from traffic light devices. Different types of intersections have varying rules and deployment of traffic lights. Our nonlinear continuous encoding function adaptively integrates color sequences and control directions. (2) Driving Behaviors at Intersections: The motion patterns and driving strategies of vehicles at intersections must conform to road topological structures while considering interaction events. Therefore, utilizing prior knowledge of intersection areas and vehicle driving states, we have designed a *driving strategy awareness mechanism* to estimate the distribution of such driving behaviors. The distribution represents a joint distribution of goal intentions and maneuvers for all vehicles. Based on the distribution, multimodal joint trajectory proposals are generated. (3) Multi-Agent Interaction: The *spatial-temporal-mode attention* is employed to model interaction relationships among agents, enabling fine-tuned adjustment of joint predicted trajectories. Additionally, an off-road loss is introduced to supervise the rationality of all trajectory modes. Finally, we validate our approach on the real-world dataset V2X-Seq and the signalized intersection drone dataset SinD. In single-infrastructure scenarios, our model provides reliable joint multi-agent trajectory predictions for all visible agents. In online collaborative scenarios, our method achieves more accurate single-agent prediction results for the target agent.

The principal contributions of this study are: (1) A novel trajectory prediction framework I2XTraj is proposed, which synthesizes a continuous signal-informed mechanism, driving strategy awareness mechanism, and spatial-temporal-mode attention in a knowledge-driven manner with dynamic graph attention. (2) To our knowledge, I2XTraj represents

the first collaborative trajectory prediction paradigm specifically engineered for infrastructure in intersection scenarios. (3) Our method outperforms existing state-of-the-art methods by more than 30% in V2X-Seq and 15% in SinD datasets.

2 Related Work

This section reviews trajectory prediction at signalized intersections with traffic signals and unique driving behaviors.

Trajectory Prediction for Signalized Intersections. Vehicle-based methods are constrained by their perceptual capabilities, lacking access to global trajectory information and real-time traffic control signals at intersections, leading to suboptimal performance. Consequently, cooperative trajectory prediction has been recognized as a promising solution to address these challenges[Wang *et al.*, 2024b]. In recent years, with the advancement of V2X technologies, several large-scale cooperative datasets that support prediction tasks have been released[Yu *et al.*, 2023; Yan *et al.*, 2023; Zhou *et al.*, 2024]. Researchers have established cooperative prediction paradigms at three stages: raw trajectory positions[Ruan *et al.*, 2023], cooperative features[Chen *et al.*, 2024], and layered prediction results[Annunziata *et al.*, 2024; Cao *et al.*, 2024]. However, these approaches necessitate frequent low-latency interactive communication between vehicles and infrastructure while simultaneously requiring precise trajectory prediction, imposing a substantial computational burden on vehicular platforms. Our method addresses the perceptual limitations by utilizing overhead roadside sensors, simplifying the complex task of joint prediction-communication optimization into a simple public information subscription task for vehicles.

Traffic Signals for Trajectory Prediction. Traffic signal states have been demonstrated to influence driving behavior at intersections significantly[Paul *et al.*, 2022]. In early studies, the states have been typically utilized in discrete forms. Discrete indexes have established dependencies for discontinuous vehicle motion behaviors[Zhang *et al.*, 2022]. One-hot encoding has been used to predict vehicle intentions[Cao *et al.*, 2024] and simulate vehicle behaviors[Wu *et al.*, 2024]. Dictionaries have embedded discrete states into high-dimensional spaces[Wei *et al.*, 2024]. However, these discrete methods neglect the temporal characteristics of signals. Furthermore, previous methods have been designed for standard orthogonal intersections, lacking generalizability across different intersection types. We propose a *continuous signal-informed mechanism* to encode continuous signals and adaptively process heterogeneous traffic control configurations.

Driving Behaviors for Trajectory Prediction. Effective mining of historical trajectory and traffic flow data can yield essential behavioral priors and patterns[Ding and Zhao, 2023]. Such knowledge is typically employed to design secondary tasks, encompassing goal-based[Gu *et al.*, 2021], maneuver-based[Liao *et al.*, 2024b], and interaction-based [Rowe *et al.*, 2023] trajectory prediction. Infrastructure sensor devices enable the superior collection of behavioral distributions in intersection areas. Leveraging prior knowledge of driving behaviors and intersection topologies, the driving

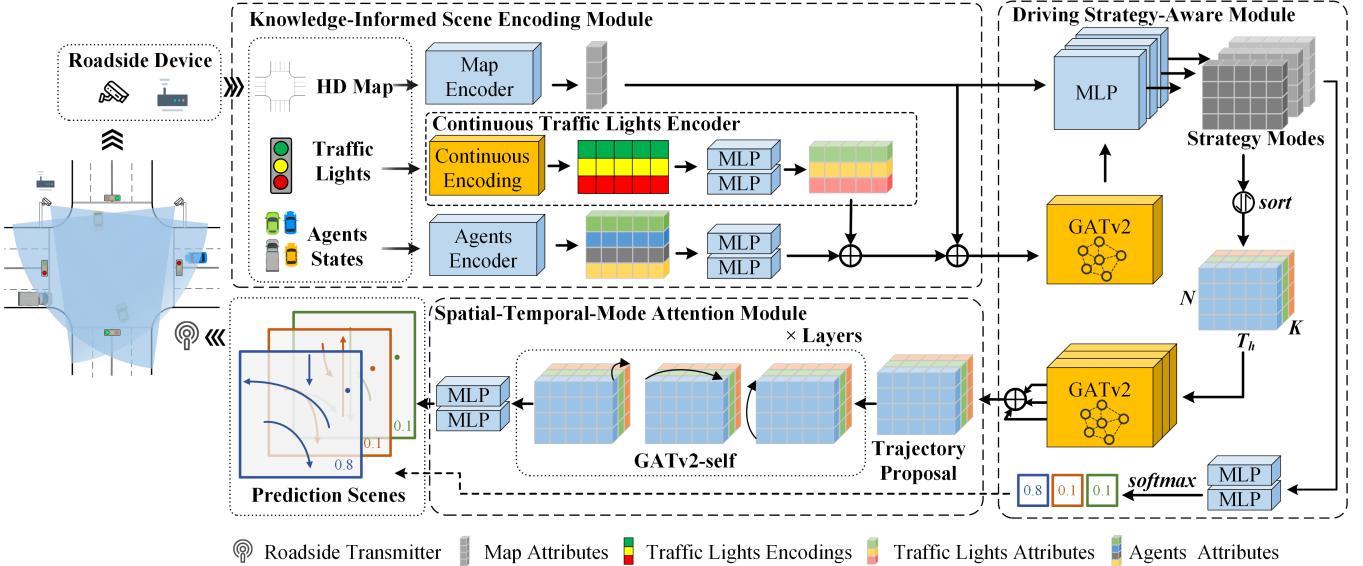


Figure 2: The overall framework of our I2XTraj. Our architecture is an infrastructure-based method, which comprises three parts: (a) Knowledge-Informed Scene Encoding Module embeds agents’ historical states with traffic signal and map knowledge. (b) Driving Strategies-Aware Module generates strategy modes based on topological features and behavioral distributions to trajectory proposals. (c) Spatial-Temporal-Mode Attention Module spans the three dimensions to generate predicted scene trajectories.

strategy awareness mechanism captures vehicles’ goal intentions and maneuvers, enabling multimodal prediction.

3 Problem Formulation

The objective of multi-agent trajectory prediction is to forecast the future trajectories of all agents with historical states. The model can directly extract the state features of agents H from the infrastructure within historical time steps T_h . Given that specific past states can lead to multiple possible futures, the goal of joint agents N prediction is to predict trajectories for any agent n under each possible scene S^k :

$$Y = \{S^k\} = \{F^{0,k}, F^{1,k}, \dots, F^{n,k}\}_{k \in [0, K-1]}, \quad (1)$$

where K denotes scene modes. Additionally, traffic light signals L and high-definition (HD) maps M are incorporated into the model as prior knowledge I . Formally, a joint future trajectory distribution among agents is represented as:

$$P(S^k | H, I) = \prod_{n=0}^{N-1} P(F^{n,k} | H, L, M), \quad (2)$$

where $F^{n,k} = \{p_x^{t_c:t_c+T_f}, p_y^{t_c:t_c+T_f}\}$ represents predicted future positions over the next T_f time steps at current time step t_c .

4 Methodology

The proposed I2XTraj framework comprises three integral components, as illustrated in Figure 4: the Knowledge-Informed Scene Encoding Module, the Driving Strategy-Aware Module, and the Spatial-Temporal-Mode Attention Module. Each component is specifically engineered to address the unique challenges of infrastructure-based trajectory prediction at signalized intersections.

4.1 Knowledge-Informed Scene Encoding Module

Infrastructure-based trajectory prediction primarily requires encoding of intersection scenarios. Recent studies have demonstrated the remarkable effectiveness of Graph Neural Networks (GNNs) and relative spatiotemporal position encoding[Tang *et al.*, 2024; Zhou *et al.*, 2023]. In the knowledge-informed scene encoding module, agent states, traffic light states, and HD maps are encoded as node attributes, while their relative spatio-temporal relationships are represented as edge attributes. This encoding method enables the framework to capture both the individual characteristics of scene elements and their intricate interconnections within the dynamic intersection environment.

Agent Encoder. The Agent Encoder module systematically encodes the agent states, including the spatial positions, motion states, geometric dimensions, and semantic attributes of agents within the scene. Specifically, a two-layer MLP is employed to embed the state of each agent at time step $t \in [t_c - T_h, t_c]$ into the agent attribute:

$$A_A^{t,n} = \text{MLP}(p_x^{t,n}, p_y^{t,n}, \theta^{t,n}, \rho^{t,n}, \varphi^{t,n}, l^{t,n}, w^{t,n}, c^{t,n}), \quad (3)$$

where $(p_x^{t,n}, p_y^{t,n})$ is the location, $\theta^{t,n}$ is the orientation, $(\rho^{t,n}, \varphi^{t,n})$ represent the velocity in polar coordinates defined with the current position as the origin and direction as the positive axis, $(l^{t,n}, w^{t,n})$ denote length and width, and $c^{t,n}$ is the type attribute.

Continuous Traffic Light Encoder. I2XTraj deployed directly on infrastructure can conveniently utilize traffic light information, including its continuous digital signals, control directions, and traffic light locations. Inspired by the attention

mechanism[Vaswani, 2017], we encode the digital signals using a nonlinear continuous function similar to positional embedding:

$$PE^{t,\ell} = \sin\left(\frac{t_{remain}^{t,d,\ell}}{T_\Omega \cdot \left(\frac{1}{3}\right)^d}\right), \quad (4)$$

where $t_{remain}^{t,d,\ell}$ is the remaining time of each color signal of a single traffic light ℓ , T_Ω is the maximum cycle time of the traffic light sequence, and $d \in \{0, 1, 2\}$ indicates the color of signals, typically following the order of red, green, and yellow.

The advantage of this approach is its adaptation to the different traffic control patterns at various intersections. Furthermore, the nonlinear function causes a sharp increase in the rate of change as the remaining time approaches zero, making the model more sensitive to the instantaneous transition of traffic light signals. Generally, the maximum cycle time T_Ω ensures the monotonicity of the values. For intersections with unknown traffic light patterns or controlled by intelligent transportation systems, a sufficiently large value has been proven to be effective.

The traffic flow direction controlled by the traffic light and the encoded signal are embedded using a two-layer MLP and then combined into traffic light attributes:

$$A_L^{t,\ell} = \text{MLP}(PE^{t,\ell}) + \text{MLP}(d^\ell), \quad (5)$$

where d^ℓ is the direction of traffic flow controlled by the traffic light. The composite feature representation encapsulates the traffic signal-governed intersection scenarios. This integrated representation captures both the instantaneous state of the intersection under traffic control and maintains the temporal context necessary for understanding the evolution of traffic patterns.

Map Encoder. Beyond the typical spatial features of lane segments such as position, length, and adjacency relationships, intersection maps contain rich semantic attributes, including interior intersection indicators, turning directions, and signal control states. A two-layer MLP is employed to encode the map attributes:

$$A_M^m = \text{MLP}(l^m, c^m), \quad (6)$$

where l^m represents the length of the centerlines, and c^m denotes the attribute features of the lane segment. The positions and coordinates of the centerlines are embedded as edge features. LaneGCN[Liang *et al.*, 2020] is employed to extract the topological relationships between lanes, utilizing graph self-attention to capture the interaction features.

Graph Edge Encoder. Similar to HPNet[Tang *et al.*, 2024], we encode the relative spatio-temporal relationships between nodes as edge features in local polar coordinate systems. A two-layer MLP is employed to encode edge attributes:

$$E_e = \text{MLP}(d^e, \phi^e, \psi^e, \delta^e), \quad (7)$$

where d^e , ϕ^e , ψ^e and δ^e represent spatial distance, the edge orientation in the reference frame, relative node orientation and temporal difference, respectively.

4.2 Driving Strategy-Aware Module

The trajectory data distributions derived from infrastructure-captured and vehicle-captured sources exhibit substantial disparities. Firstly, the field of view of roadside sensors exclusively encompasses the traffic conditions within the intersection area. Secondly, due to the complex nature of intersection environments, vehicles naturally tend to reduce their speed upon entering the intersection. Thirdly, vehicle maneuvers at intersections are constrained to a finite set of options, typically comprising stopping, through movement, left/right turns, and U-turns.

The driving awareness mechanism generates potential goal intervals for maneuvers based on the prior intersection diameter D knowledge and simple geometric operations: the stop maneuver corresponds to $[0, 1]$, the through movement corresponds to the interval exceeding the intersection diameter $[D, +\infty)$, $[1, \frac{\sqrt{2}D}{2})$ for right turns and U-turns, and $[\frac{D}{2}, D)$ for left turns. The strategies for multimodal intersection behavior are derived from latent driving intentions, thus the driving awareness mechanism maps these goal intervals into distinct modes. Corresponding to K modes, the K goal intervals $\{[0, 1], [1, \frac{D}{K-2}), [\frac{D}{K-2}, \frac{2D}{K-2}), \dots, [\frac{(k-3)D}{K-2}, D), [D, +\infty)\}$ encompasses various intersection maneuvers.

The strategy modes are predicted for all agents at the intersection in parallel. In essence, this involves mapping spatio-temporal attributes onto each goal interval by dynamic graph attention. The driving awareness decoder component comprises a self-attention module and K independent two-layer MLP decoders. Similar to the object query concept in DETR[Carion *et al.*, 2020], each independent decoder generates adaptive a trajectory anchor query $Q^{t,n}$ corresponding to each mode query $q_K^{t,n,k}$:

$$Q^{t,n} = \text{GATv2}(A_A^{t,n}, [A_A^{t,n}, E_e], [A_A^{t,n}, E_e]), \quad (8)$$

$$q_K^{t,n,k} = \text{MLP}(Q^{t,n}), \quad (9)$$

where $\text{GATv2}(Query, Key, Value)$ denotes the dynamic graph attention[Brody *et al.*, 2022]. These queries are processed through an MLP output layer to produce the probability scores P_k of strategy modes. Specifically, P_k is assumed to be the probabilities of goals falling within intervals.

$$P_k = \text{softmax}(\text{MLP}(q_K^{t,n,k})). \quad (10)$$

The mode queries are sorted based on the scores and stacked into mode anchor attributes $A_K^{t,n,k}$.

The mode anchor attributes subsequently interact with agents, traffic signals, and map features through cross-attention modules. During the dynamic process, the trajectory distribution diverges when the agent enters the intersection, and gradually converges as it exits the intersection area. To capture this trend, agent proposal query $q_A^{t,n,k}$, signal proposal query $q_L^{t,n,k}$ and map proposal query $q_M^{t,n,k}$ are generated:

$$\begin{aligned} q_A^{t,n,k} &= \text{GATv2}(A_K^{t,n,k}, [A_A^{t_c - T_h:t,n}, E_e], \\ &\quad [A_A^{t_c - T_h:t,n}, E_e]), \end{aligned} \quad (11)$$

$$q_L^{t,n,k} = \text{GATv2}(A_L^{t,n,k}, [A_L^{t_c-T_h:t,\ell}, E_e], [A_L^{t_c-T_h:t,\ell}, E_e]), \quad (12)$$

$$q_M^{t,n,k} = \text{GATv2}(A_K^{t,n,k}, [A_M^m, E_e], [A_M^m, E_e]), \quad (13)$$

These queries are summed to generate the predicted trajectory proposal attribute $Q_P^{t,n,k}$:

$$Q_P^{t,n,k} = q_A^{t,n,k} + q_L^{t,n,k} + q_M^{t,n,k}. \quad (14)$$

4.3 Spatial-Temporal-Mode Attention Module

The Spatial-Temporal-Mode Attention Module performs attention across spatiality, temporal, and mode, enabling predictions to interact with any agents, any historical timestamps, and any modes. These interactions can be either direct or multi-hop. For the spatiality, an agent A performs cross-attention with each mode at every timestamp of every neighboring agent in the scene:

$$Q_A^{t,n,k} = \text{GATv2}(Q_P^{t,n,k}, [Q_P^{t,n_{nbr},k}, E_e], [Q_P^{t,n_{nbr},k}, E_e]), \quad (15)$$

where n_{nbr} denotes all neighboring agents. For the temporal, each timestamp T performs self-attention with each mode of every agent at every historical timestamp:

$$Q_T^{t,n,k} = \text{GATv2}(Q_A^{t,n,k}, [Q_A^{t-t_c+T_h:t,n,k}, E_e], [Q_A^{t-t_c+T_h:t,n,k}, E_e]). \quad (16)$$

For the mode, each mode K performs self-attention with each mode of every agent at every historical timestamp:

$$Q_K^{t,n,k} = \text{GATv2}(Q_T^{t,n,k}, [Q_T^{t,n,0:K-1}, E_e], [Q_T^{t,n,0:K-1}, E_e]). \quad (17)$$

Multiple iterations enable the learning of broader interaction features across all dimensions while enhancing trajectory prediction accuracy[Zhou *et al.*, 2023].

Finally, a two-layer MLP is employed for trajectory decoding to generate scene prediction trajectories:

$$S^k = \text{MLP}(Q_K^{t,n,k}). \quad (18)$$

4.4 Training Objectives

While the adaptive anchor-based decoding mechanism demonstrates notable advantages in enhancing predictive flexibility, it inherently introduces rationality constraints, with boundary violations emerging as the most substantial limitation. To mitigate this issue, we implement an off-map loss function. The vector map representation transforms a rasterized drivable raster map M_r . Following this conversion, we query the spatial coordinates to generate off-map matrices for both the ground truth $O_{gt}^{t,n} = M_r(g_x^{t,n}, g_y^{t,n})$, and predicted trajectories $O_p^{t,n,k} = M_r(p_x^{t,n,k}, p_y^{t,n,k})$. To account for and exclude off-map occurrences attributable to both the vehicle's intrinsic kinematic constraints and potential map imperfections, the off-map loss is computed as follows:

$$\mathcal{L}_{off-map} = \sum_{t=t_c}^{t_c+T_f} \frac{\sum_{k=0}^{K-1} \sum_{n=0}^{N-1} [O_p^{t,n,k} \wedge O_{gt}^{t,n}]}{K \sum_{n=0}^{N-1} O_{gt}^{t,n}}. \quad (19)$$

This loss metric is designed to differentiate between legitimate trajectory deviations and those arising from prediction errors.

Following the existing works[Zhao *et al.*, 2021; Rowe *et al.*, 2023], the joint regression loss is computed utilizing the Huber loss function:

$$\mathcal{L}_{reg} = \mathcal{L}_{huber}(S^k, G), \quad (20)$$

where G donates the ground truth trajectories.

The loss for strategy probabilities P_k is computed using cross-entropy:

$$\mathcal{L}_{cls} = \mathcal{L}_{CE}(P_k, G_k), \quad (21)$$

where G donates the ground truth strategies set.

The entire model is optimized using a comprehensive loss function:

$$\mathcal{L} = (1 + \mathcal{L}_{off-map}) \mathcal{L}_{reg} + \mathcal{L}_{cls}. \quad (22)$$

In addition, in marginal prediction tasks, the joint regression loss degenerates into marginal regression loss.

5 Experiment

5.1 Experimental Setup

Dataset. To rigorously evaluate the efficacy of our model, we select two complementary datasets that explicitly incorporate traffic signal data: the real-world V2I dataset V2X-Seq[Yu *et al.*, 2023] and the drone dataset at signalized intersection SinD[Xu *et al.*, 2022]. The V2X-Seq (Single-Infrastructure, SI) set contains comprising 55,197 pure infrastructure-based scenarios. The V2X-Seq (Cooperation, C) set contains 51,146 V2I scenarios. Both sets feature trajectories of 10-second duration sampled at 10Hz, covering 672 hours of data from 28 intersections. The prediction task involves observing 5 seconds to forecast the subsequent 5 seconds. The SinD dataset contains 7 hours of continuous trajectory data sampled at 10Hz from a signalized intersection in Tianjin, China. It includes trajectories and semantic annotations for over 13,000 traffic participants, along with traffic signal states. The prediction task involves observing 1.2 seconds (12 frames) to predict the subsequent 1.2 seconds (12 frames) and 1.8 seconds (18 frames).

Metrics. The prediction performance is evaluated using a comprehensive set of trajectory metrics, including minADE, minFDE, MR, minJointADE, minJointFDE, and minJointMR. The minADE measures the average L2 distance between predicted and ground truth trajectory points, while minFDE examines the L2 distance at trajectory endpoints. The MR metric calculates the ratio of cases where minFDE exceeds 2 meters. The minJointADE evaluates the average L2 distance between predicted and ground truth trajectories across all agents, while minJointFDE focuses on the L2 distance for all agents at the final timestep. The minJointMR metric calculates the ratio of cases where minJointFDE exceeds 2 meters. The number of modes K is selected as 6.

Baselines. I2XTraj is compared to twelve state-of-the-art (SOTA) trajectory prediction models: TNT[Zhao *et al.*, 2021], DenseTNT[Gu *et al.*, 2021], HiVT[Zhou *et al.*, 2022], V2INet[Chen *et al.*, 2024], V2X-Graph[Ruan *et al.*, 2023], AIoT[Annunziata *et al.*, 2024], HPNet[Tang *et al.*, 2024] for the V2X-Seq dataset. S-GAN[Gupta *et al.*, 2018], S-LSTM[Alahi *et al.*, 2016], Trajetron+[Salzmann *et al.*, 2020], FJMP[Rowe *et al.*, 2023], KI-GAN[Wei *et al.*, 2024] for SinD dataset. As I2XTraj represents the first work to perform muti-agent prediction tasks on V2X-Seq, we re-produce and compare against HPNet, the SOTA method on the unsignalized intersection dataset INTERACTION, as our baseline. Additionally, HPNet is also adapted to the signalized intersection drone dataset SinD to establish a comparative baseline.

Implementation Details. In our implementation, each dynamic graph attention consists of four layers of multi-head attention. Our model is trained on a single A800 GPU for 32 epochs, using the AdamW[Loshchilov and Hutter, 2019] optimizer with the batch size of 4, using an initial learning rate of 5×10^{-4} , setting 30 meters radius for all local areas. The parameter size of I2XTraj is 3.18M.

5.2 Comparison with State-of-the-art

I2XTraj is compared against the strongest baselines in the V2X-Seq trajectory forecasting benchmarks. Our trajectory prediction framework is developed and evaluated using the V2X-Seq (SI). Given that only one previous study[Annunziata *et al.*, 2024] has provided a reference using single-infrastructure dataset V2X-Seq (SI) for training, we further train and validate our model on the more extensively researched collaborative dataset V2X-Seq (C).

Single-Agent Scenario. I2XTraj is compared with single-agent SOTA baselines on V2X-Seq. Our framework consistently achieves superior performance across all metrics, as shown in Table 1.

For collaborative scenarios, the PP-VIC method is employed for vehicle-infrastructure data fusion[Yu *et al.*, 2023]. PP-VIC provides trajectory data from the infrastructure to the ego vehicle in a frame-by-frame manner within the historical range. Our I2XTraj framework demonstrates significant performance improvements compared to the graph-based feature fusion method V2X-Graph, achieving a 26.25% reduction in minFDE and a 48% decrease in MR. Compared to the layer-wise fusion method AIoT, our framework demonstrates a substantial improvement of 18.94% in minADE. Although PP-VIC is not currently the best performance collaborative method, it retains complete infrastructure-captured trajectory and traffic light data, which highlights our model’s significant advantages at signalized intersections.

For single-infrastructure scenarios, our infrastructure-based I2XTraj framework underscores more pronounced advantages compared to general methods. Specifically, I2XTraj achieves a 40.16% reduction in minADE and a 47.35% reduction in minFDE compared to HiVT. When compared to AIoT, the MR decreases by 47.73%. The notably higher improvement in MR metrics emphasizes the significance of incorporating off-map loss. Although the overall performance

Method	Dataset/Fusion	minADE	minFDE	MR
TNT	C/PP-VIC	4.36	9.23	0.62
DenseTNT	C/PP-VIC	1.84	2.56	0.28
HiVT	C/PP-VIC	1.27	2.36	0.30
V2INet	C/PP-VIC	1.19	1.98	0.27
V2X-Graph	C/PP-VIC	1.12	1.98	0.30
V2X-Graph	C/FF	1.05	<u>1.79</u>	<u>0.25</u>
AIoT	C/*	<u>0.95</u>	1.87	0.27
I2XTraj	C/PP-VIC	0.77	1.32	0.13
TNT	SI/-	4.93	9.45	0.65
HiVT	SI/-	<u>1.27</u>	<u>2.83</u>	0.47
AIoT	SI/-	1.36	2.96	<u>0.44</u>
I2XTraj	SI/-	0.76	1.49	0.23

Table 1: Performance comparison of single-agent on the V2X-Seq dataset. Dataset C is the collaborative set, while SI is the single-infrastructure set. FF indicates a Feature Fusion method[Ruan *et al.*, 2023]. The asterisk (*) denotes that a layered fusion method[Annunziata *et al.*, 2024]. The dash (-) indicates without fusion. The best performance is in **bold** and the second best is underlined.

of all methods, including I2XTraj, shows some degradation compared to cooperative scenarios, this may be attributed to the strong correlation between the target agent selection and the ego vehicle in the dataset, where vehicle-captured data provides more accurate trajectories and velocities. However, I2XTraj exhibits substantially smaller average degradation compared to general models, and even a slight improvement in minADE, indicating effective utilization of infrastructure data characteristics.

Multi-Agent Scenario. I2XTraj is compared with multi-agent baselines on V2X-Seq (SI) and SinD. Our framework consistently achieves superior performance in most cases, as shown in Table 2 and Table 3.

On the V2X-Seq (SI) dataset, our I2XTraj framework demonstrates superior performance compared to HPNet, the SOTA trajectory prediction method developed for unsignalized intersections. Specifically, our framework achieves significant improvements across all metrics: a 23.53% reduction in minJointFDE, a 30.8% reduction in minJointADE, and a 19.05% reduction in minJointMR.

On the SinD dataset, our I2XTraj framework also illustrates improved prediction accuracy, with particularly notable performance in long-term prediction, achieving reductions of 9.09% and 19.23% in minADE and minFDE, respectively. While the short-term prediction shows marginal limitations, these results sufficiently support the generalizability of I2XTraj across heterogeneous data sources.

Qualitative Results The qualitative results demonstrate the effectiveness of I2XTraj in Figure 3. Compared to the (b) single-agent task, the (a) multi-agent task provides comprehensive trajectory prediction for vehicles. In (c₁), with 5 seconds remaining on the green signal, the model accurately anticipates the vehicle’s successful acceleration through the intersection. In (c₂), with 1 second remaining on the red signal, the model precisely predicts the vehicle’s starting initiation. In (c₃), where a red signal regulates the left-turn lane, non-stopping vehicles exhibit the highest probability of executing

Method	minJointFDE	minJointADE	minJointMR
HPNet	0.68	1.62	0.21
I2XTraj	0.52	1.12	0.17

Table 2: Performance comparison of multi-agent on the V2X-Seq (SI) dataset.

Method	12-12		12-18	
	minADE	minFDE	minADE	minFDE
S-GAN	1.32	2.46	1.53	2.95
S-LSTM	0.87	1.60	0.96	1.78
Trajetron++	0.37	0.93	0.70	1.91
FJMP	0.27	0.68	0.41	1.13
KI-GAN	0.05	0.12	0.11	0.26
HPNet	<u>0.09</u>	0.21	<u>0.11</u>	<u>0.25</u>
I2XTraj	<u>0.09</u>	0.12	0.10	0.21

Table 3: Performance comparison of multi-agent on the SinD dataset. The performance values of S-GAN, S-LSTM, Trajetron++, FJMP, and KI-GAN are all obtained from the original paper[Wei *et al.*, 2024].

a U-turn. The results clearly illustrate that I2XTraj effectively leverages traffic signal information to enhance predictive performance.

5.3 Ablation Study

To further demonstrate the effectiveness and outcomes of the continuous signal-informed and driving strategy awareness mechanism, we conducted ablation studies on the V2X-Seq (SI) dataset. The quantitative results of our ablation studies are presented in Table 4.

Comparing the complete framework against that without the continuous traffic light encoder, we validate the effectiveness of the continuous signal-informed (CSI) mechanism. The continuous traffic light encoding enables the model to accurately anticipate signal phase transitions, thereby precisely determining vehicles’ instantaneous dynamics during acceleration and deceleration events. Additionally, the relative positioning between traffic lights and vehicles helps the model predict specific vehicle stopping locations. The incorporation of traffic light information improved prediction accuracy by 2.61~7.44%, with particularly pronounced enhancements in final point precision.

The Driving Strategy-Aware (DSA) mechanism and map information (Map) prove crucial for the model’s scene comprehension and prediction capabilities. The improvements attributed to DSA validate the effectiveness of joint trajectory prediction based on combined the goal intents and maneuvers relationship. The knowledge-driven driving strategy prediction, followed by trajectory mode generation based on strategy confidence scores, provides causal guidance for the prediction. The enhancement brought by MAP demonstrates the significant influence of map topology on vehicle trajectories at intersections. The integration of map topology and behavioral priors enables more accurate vehicle trajectory prediction, yielding improvements of up to 8.00%.

CSI	MAP	DSA	minADE	minFDE	MR
	✓	✓	0.55/0.78	1.21/1.53	0.18/0.24
✓		✓	0.53/0.79	1.14/1.56	0.17/0.25
✓	✓		0.53/0.77	1.16/1.52	0.18/0.24
✓	✓	✓	0.52/0.76	1.12/1.49	0.17/0.23

Table 4: Ablation studies for core components in V2X-Seq (SI) dataset. Results for multi/single-agent are reported, with all metrics being joint- for the multi-agent.

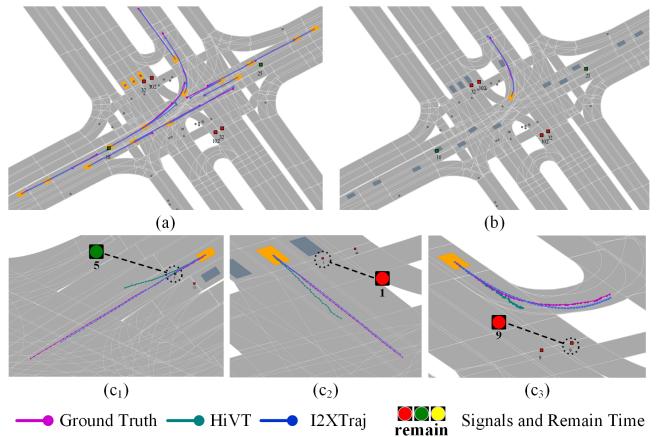


Figure 3: Qualitative results of I2XTraj on V2X-Seq. Target vehicles are depicted in orange, other vehicles are shown in gray and gray dots are non-motorized road users. (a) and (b) illustrate multi-agent and single-agent prediction, respectively. (c₁), (c₂), and (c₃) depict the impact of traffic signal phase transitions on vehicle stopping, starting, and maneuvers.

6 Conclusion

In this paper, I2XTraj has been proposed, a novel infrastructure-based multi-agent trajectory prediction framework for vehicles at signalized intersections. Our approach has introduced three key mechanisms: a continuous signal-informed mechanism that adaptively encodes traffic signal information, a driving strategy awareness mechanism that estimates behavioral distributions, and a spatial-temporal-mode attention mechanism that models multi-agent interactions. I2XTraj has demonstrated superior performance, surpassing SOTA methods by 30% and 15% on the V2X-Seq and SinD datasets, respectively. This work has not only established a new paradigm for collaborative trajectory prediction but also highlighted an Infrastructure-to-Everything approach in advancing autonomous driving systems. The success of I2XTraj has suggested promising directions for future research in knowledge-driven trajectory prediction at signalized intersections.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No.62433014. The authors would like to thank TÜV SÜD for the kind and generous support. We are also grateful for the efforts of our colleagues in the Sino-German Center of Intelligent Systems.

References

- [Alahi *et al.*, 2016] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–971, 2016.
- [Annunziata *et al.*, 2024] Daniela Annunziata, Diletta Chiaro, Pian Qi, and Francesco Piccialli. On the road to aiot: A framework for vehicle road cooperation. *IEEE Internet of Things Journal*, 2024.
- [Brody *et al.*, 2022] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *Proceedings of International Conference on Learning Representations*, 2022.
- [Cao *et al.*, 2024] Yue Cao, Wei Shangguan, Arnoud Visser, Junjie Chen, Linguo Chai, and Baigen Cai. Segment-based trajectory prediction and risk assessment for rsu-assisted cavs at signalized intersections. *IEEE Transactions on Intelligent Vehicles*, 2024.
- [Cao, 2024] Longbing Cao. Ai4tech: X-ai enabling x-tech with human-like, generative, decentralized, humanoid and metaverse ai. *International Journal of Data Science and Analytics*, 18(3):219–238, 2024.
- [Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*, pages 213–229, 2020.
- [Chen *et al.*, 2024] Xi Chen, Rahul Bhadani, and Larry Head. Conformal trajectory prediction with multi-view data integration in cooperative driving. *arXiv preprint arXiv:2408.00374*, 2024.
- [Ding and Zhao, 2023] Zhezhang Ding and Huijing Zhao. Incorporating driving knowledge in deep learning based vehicle trajectory prediction: A survey. *IEEE Transactions on Intelligent Vehicles*, 8(8):3996–4015, 2023.
- [Ekmekci *et al.*, 2024] Mustafa Ekmekci, Lee Woods, and Nima Dadashzadeh. Effects of road width, radii and speeds on collisions at three-arm priority intersections. *Accident Analysis & Prevention*, 199:107522, 2024.
- [Gu *et al.*, 2021] Junru Gu, Chen Sun, and Hang Zhao. Densentnt: End-to-end trajectory prediction from dense goal sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15303–15312, 2021.
- [Gupta *et al.*, 2018] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.
- [Liang *et al.*, 2020] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *Proceedings of the European Conference on Computer Vision*, pages 541–556, 2020.
- [Liao *et al.*, 2024a] Haicheng Liao, Zhenning Li, Huanming Shen, Wenxuan Zeng, Dongping Liao, Guofa Li, and Chengzhong Xu. Bat: Behavior-aware human-like trajectory prediction for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 10332–10340, 2024.
- [Liao *et al.*, 2024b] Haicheng Liao, Chengyue Wang, Zhenning Li, Yongkang Li, Bonan Wang, Guofa Li, and Chengzhong Xu. Physics-informed trajectory prediction for autonomous driving under missing observation. In Kate Larson, editor, *Proceedings of the Joint Conference on Artificial Intelligence*, pages 6841–6849, 2024.
- [Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of International Conference on Learning Representations*, 2019.
- [Paul *et al.*, 2022] Madhumita Paul, Indrajit Ghosh, and Md Mazharul Haque. The effects of green signal countdown timer and retiming of signal intervals on dilemma zone related crash risk at signalized intersections under heterogeneous traffic conditions. *Safety science*, 154:105862, 2022.
- [Rowe *et al.*, 2023] Luke Rowe, Martin Ethier, Eli-Henry Dykhne, and Krzysztof Czarnecki. Fjmp: Factorized joint multi-agent motion prediction over learned directed acyclic interaction graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13745–13755, 2023.
- [Ruan *et al.*, 2023] Hongzhi Ruan, Haibao Yu, Wenxian Yang, Siqi Fan, Yingjuan Tang, and Zaiqing Nie. Learning cooperative trajectory representations for motion forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2023.
- [Salzmann *et al.*, 2020] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Proceedings of the European Conference on Computer Vision*, pages 683–700, 2020.
- [Sarkar and Taherkordi, 2023] Indranil Sarkar and Amir Taherkordi. Enhancing task efficiency in vehicular fog computing: Leveraging mobility prediction and min-max optimization for reduced latency. In *Proceedings of the IEEE Vehicular Technology Conference*, pages 1–7, 2023.
- [Singh *et al.*, 2022] Mohit Kumar Singh, Bharat Kumar Pathivada, K Ramachandra Rao, and Vedagiri Perumal. Driver behaviour modelling of vehicles at signalized intersection with heterogeneous traffic. *IATSS research*, 46(2):236–246, 2022.
- [Tang *et al.*, 2024] Xiaolong Tang, Meina Kan, Shiguang Shan, Zhilong Ji, Jinfeng Bai, and Xilin Chen. Hpnet: Dynamic trajectory forecasting with historical prediction attention. In *Proceedings of the IEEE/CVF Conference on*

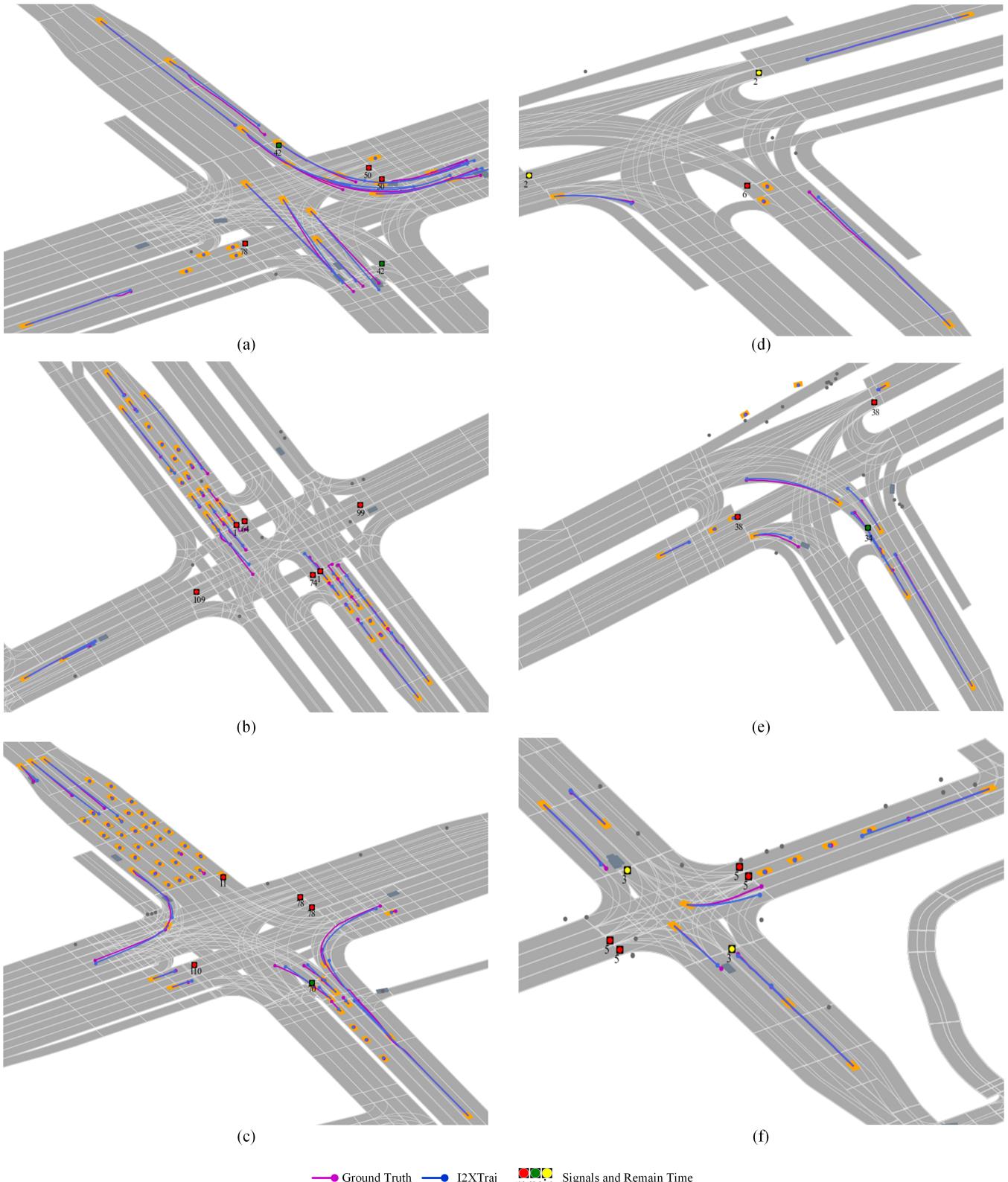
- Computer Vision and Pattern Recognition*, pages 15261–15270, 2024.
- [Vaswani, 2017] A Vaswani. Attention is all you need. In *Proceedings of the Neural Information Processing Systems*, 2017.
- [Wang *et al.*, 2024a] Chenglong Wang, Jun Peng, Lin Cai, Weirong Liu, Shuo Li, Hu He, and Zhiwu Huang. Proactive bandwidth allocation for v2x networks with multi-attentional deep graph learning. *IEEE Transactions on Wireless Communications*, 2024.
- [Wang *et al.*, 2024b] Tianqi Wang, Sukmin Kim, Ji Wenxuan, Enze Xie, Chongjian Ge, Junsong Chen, Zhenguo Li, and Ping Luo. Deepaccident: A motion and accident prediction benchmark for v2x autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5599–5606, 2024.
- [Wei *et al.*, 2024] Chuheng Wei, Guoyuan Wu, Matthew J Barth, Amr Abdelraouf, Rohit Gupta, and Kyungtae Han. Ki-gan: Knowledge-informed generative adversarial networks for enhanced multi-vehicle trajectory forecasting at signalized intersections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7115–7124, 2024.
- [Wu *et al.*, 2024] Aotian Wu, Yash Ranjan, Rahul Sengupta, Anand Rangarajan, and Sanjay Ranka. A data-driven approach for probabilistic traffic prediction and simulation at signalized intersections. In *Proceedings of IEEE Intelligent Vehicles Symposium*, pages 3092–3099, 2024.
- [Xu *et al.*, 2022] Yanchao Xu, Wenbo Shao, Jun Li, Kai Yang, Weida Wang, Hua Huang, Chen Lv, and Hong Wang. Sind: A drone dataset at signalized intersection in china. In *Proceedings of International Conference on Intelligent Transportation Systems*, pages 2471–2478, 2022.
- [Yan *et al.*, 2023] Zhijie Yan, Pengfei Li, Zheng Fu, Shaocong Xu, Yongliang Shi, Xiaoxue Chen, Yuhang Zheng, Yang Li, Tianyu Liu, Chuxuan Li, et al. Int2: Interactive trajectory prediction at intersections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8536–8547, 2023.
- [Yu *et al.*, 2023] Haibao Yu, Wenxian Yang, Hongzhi Ruan, Zhenwei Yang, Yingjuan Tang, Xu Gao, Xin Hao, Yifeng Shi, Yifeng Pan, Ning Sun, et al. V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5486–5495, 2023.
- [Zeng *et al.*, 2024] Feng Zeng, Zheng Zhang, and Jinsong Wu. Task offloading delay minimization in vehicular edge computing based on vehicle trajectory prediction. *Digital Communications and Networks*, 2024.
- [Zhang *et al.*, 2022] Yuzhen Zhang, Wentong Wang, Weizhi Guo, Pei Lv, Mingliang Xu, Wei Chen, and Dinesh Manocha. D2-tpred: Discontinuous dependency for trajectory prediction under traffic lights. In *Proceedings of the European Conference on Computer Vision*, pages 522–539, 2022.
- [Zhao *et al.*, 2021] Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Ben Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. In *Proceedings of the Conference on Robot Learning*, pages 895–904, 2021.
- [Zhou *et al.*, 2022] Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8833, 2022.
- [Zhou *et al.*, 2023] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17863–17873, 2023.
- [Zhou *et al.*, 2024] Zewei Zhou, Hao Xiang, Zhaoliang Zheng, Seth Z Zhao, Mingyue Lei, Yun Zhang, Tian-hui Cai, Xinyi Liu, Johnson Liu, Maheswari Bajji, et al. V2xpnp: Vehicle-to-everything spatio-temporal fusion for multi-agent perception and prediction. *arXiv preprint arXiv:2412.01812*, 2024.

A Inference Time

We additionally evaluate the model’s inference time and parameter scale, which are crucial factors for practical deployment. The inference latency is measured on an NVIDIA A800 GPU using the V2X-Seq validation set. I2XTraj requires a total processing time of 112.23 milliseconds per time step. Given that the model is designed for cloud-based or roadside edge computing units, this latency does not impede the real-time operation of autonomous driving systems while simultaneously reducing the computational burden on vehicle-mounted platforms.

B Quality Results

In Figure 4, (a), (b), and (c) demonstrate I2XTraj’s unoccluded capability for large-scale trajectory prediction in complex, high-density scenarios. Notably, (b) highlights the differential impacts of traffic signal transitions on vehicle behaviors at the same intersection. (d), (e), and (f) showcase the model’s predictive capabilities across diverse intersection types. (d) and (e) illustrate the model’s adaptability in three-way intersections. (f) demonstrates that I2XTraj effectively handles the interactions between vehicles and non-motorized road users at small mixed-traffic intersections. Additionally, (d) and (f) emphasize the influence of yellow signals on vehicle behavior, where vehicles preemptively decelerate and stop when encountering yellow lights. Furthermore, Figure 4 validates that the continuous signal-informed mechanism can effectively process traffic signals of varying quantities and locations.



—●— Ground Truth —●— I2XTraj ■■■■■ Signals and Remain Time

Figure 4: Qualitative results on the V2X-Seq (SI) validation set. Target vehicles are depicted in orange, other vehicles are shown in gray, and gray dots are non-motorized road users.