

Article

Occlusion-Aware Interactive End-to-End Autonomous Driving for Right-of-Way Conflicts

Jialun Yin ^{1,2,†} , Kun Zhao ^{1,†} , Xiaohan Ma ¹, Siping Yan ¹, Haoran Li ^{1,*}, Junru Yang ¹  and Yin Chen ¹

¹ Suzhou Automotive Research Institute, Tsinghua University, Suzhou 215134, China; yinjalun@tsari.tsinghua.edu.cn (J.Y.); zhaokun@tsari.tsinghua.edu.cn (K.Z.); maxiaohan@tsari.tsinghua.edu.cn (X.M.); yangjunru@tsari.tsinghua.edu.cn (J.Y.); chenying@tsari.tsinghua.edu.cn (Y.C.)

² School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China

* Correspondence: lihaoran@tsari.tsinghua.edu.cn

† These authors contributed equally to this work.

Abstract

End-to-end autonomous driving has demonstrated remarkable potential due to its strong scene-understanding capabilities. However, its performance degrades significantly in the presence of occlusions and complex multi-agent interactions, posing serious safety risks. Existing methods struggle to understand partially observed environments and accurately predict the dynamic behaviors of surrounding agents. To address these limitations, we propose OAIAD (Occlusion-Aware Interactive End-to-End Autonomous Driving), a novel end-to-end framework designed to enhance occlusion reasoning and interaction awareness. This framework specifically addresses the critical challenge of right-of-way conflicts in complex multi-agent scenarios. OAIAD employs a stereoscopic vectorized representation to explicitly model occluded areas and incorporates a module for joint optimization of trajectory prediction and planning to better capture future agent dynamics. By explicitly modeling interactive behaviors and leveraging joint trajectory optimization, OAIAD enhances the ego vehicle's ability to negotiate the right-of-way interactions in a safe and socially compliant manner, significantly reducing conflict-induced collisions. Extensive evaluations on both open- and closed-loop datasets demonstrate that OAIAD significantly improves performance in occlusion-heavy and interaction-intensive scenarios. Real-world experiments further validate the practicality and robustness of our approach, highlighting its potential for deployment in complex urban environments.

Keywords: autonomous driving; end-to-end; occlusion; vectorized planning



Academic Editor: Dan Zhang

Received: 17 September 2025

Revised: 14 October 2025

Accepted: 16 October 2025

Published: 20 October 2025

Citation: Yin, J.; Zhao, K.; Ma, X.; Yan, S.; Li, H.; Yang, J.; Chen, Y. Occlusion-Aware Interactive End-to-End Autonomous Driving for Right-of-Way Conflicts. *Machines* **2025**, *13*, 965. <https://doi.org/10.3390/machines13100965>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent years, end-to-end learning approaches have gained significant attention in the field of autonomous driving. Leveraging the rapid advancements in deep learning, end-to-end frameworks have demonstrated impressive performance by integrating perception, decision-making, and control into a unified pipeline, showing great potential in handling complex scenarios. However, despite these advancements, large-scale commercialization of autonomous vehicles continues to face critical safety challenges, particularly in scenarios with occlusions and in interactive environments such as at intersections [1].

Among these, right-of-way conflicts present a particularly demanding challenge for autonomous decision-making. These conflicts arise when multiple agents contend for the same spatial resource simultaneously, such as at unsignalized intersections, during lane merges,

or for occluded crosswalks. Resolving these conflicts requires not only a precise understanding of traffic rules but also the ability to infer other agents' intentions and generate appropriate negotiation strategies—behaviors that are often implicit and ambiguous. Traditional methods, which treat prediction and planning as separate tasks, struggle to model the tight coupling between and mutual influence of agents in such conflict scenarios, leading either to overly aggressive or overly passive behaviors that can disrupt traffic flow or compromise safety [2].

Dynamic occlusions introduce significant perception uncertainty, making it difficult for systems to construct reliable and robust environmental representations. Meanwhile, the strong coupling of multi-agent interactions imposes substantial demands on the reasoning and decision-making capabilities of autonomous systems. Although end-to-end frameworks exhibit promising generalization capabilities, ensuring safety and reliability in these challenging scenarios remains an unresolved issue [3].

Human drivers' cognitive advantages in occluded scenarios highlight critical gaps in current autonomous driving systems. As illustrated in Figure 1a,b, when visibility is limited, drivers construct three-dimensional semantic representations of occluded regions (e.g., anticipating pedestrian movements behind parked trucks) and adopt exploratory maneuvers (such as slight lateral shifts or deceleration) to minimize risk. In panel (a), the ego vehicle adjusts its trajectory to avoid an unexpected pedestrian emergence from behind a truck, demonstrating the cognitive ability to predict and mitigate risks in occluded scenarios. In panel (b), the ego vehicle decelerates to increase the distance from the truck, thereby improving forward visibility and allowing for better anticipation of hidden obstacles. This behavior highlights the importance of proactive maneuvers in dynamic environments.

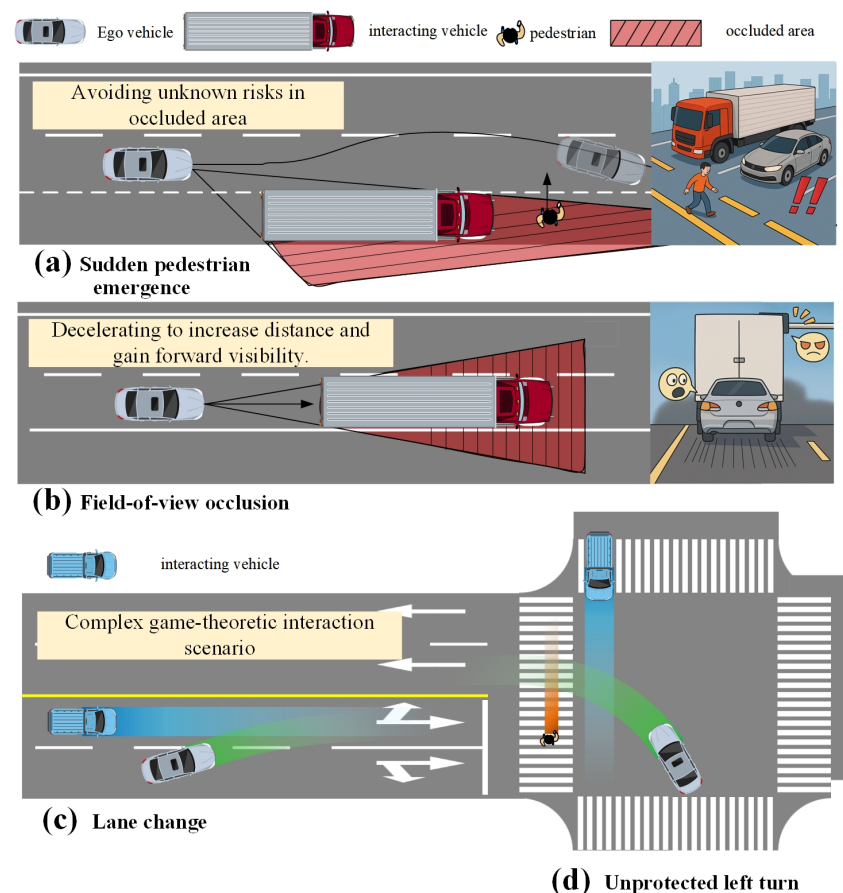


Figure 1. Illustration of visual occlusion and multi-agent interaction scenarios. Panels (a,b) demonstrate visual occlusion caused by roadside trucks; (c,d) depict potential interactions in complex multi-agent scenarios.

In contrast, current end-to-end models mainly rely on map topology and object detection boxes [4], lacking explicit modeling of the evolution of occlusion boundaries in space and time. This “geometric aphasia” limits the system’s ability to predict potential conflict zones, leading to delayed collision responses in sudden occlusion scenarios (e.g., a child running out from between parked vehicles) [5] and thus posing serious safety risks.

The bottleneck of interaction modeling in complex scenarios stems from the decoupled design between prediction and planning. Current approaches mostly treat agent trajectory prediction and ego vehicle planning as independent optimization problems, overlooking their mutual influence during dynamic interactions. In Figure 1, panel (c) illustrates a lane change scenario in which the acceleration or deceleration behavior of surrounding vehicles significantly alters the feasible trajectory space of the ego vehicle. In this case, the ego vehicle must adapt its trajectory to safely execute the lane change without causing conflicts with other vehicles. Also in Figure 1, panel (d) demonstrates an unprotected left turn scenario, where the ego vehicle faces uncertainty due to the unpredictable behaviors of surrounding traffic participants. The current MDP-based models simplify these interaction dependencies into single-step transitions, often resulting in trajectories with physical feasibility conflicts (e.g., insufficient braking distance) or violations of social norms (e.g., aggressive cutting in causing traffic disturbances).

To tackle the above challenges, this paper proposes the Occlusion-Aware Interactive Autonomous Driving (OAIAD) framework, which redefines the paradigm of environmental understanding in autonomous driving systems by integrating occlusion-aware perception and interactive planning. The contributions of the study can be summarized as follows:

- A novel approach that fuses visual features into a hierarchical dynamic graph structure is introduced, enabling 3D topological reasoning regarding occluded regions through primitive geometric decomposition;
- A hierarchical interaction framework is proposed that leverages graph attention mechanisms to explicitly model the dependencies of the multi-modal trajectories of multiple agents, generating physically and socially compatible motion plans;
- The effectiveness of the proposed method is validated through benchmark testing and a real-world field test conducted in Suzhou, China, covering both occlusion and complex interaction scenes.

OAIAD demonstrated superior performance in these challenging scenarios, particularly in cases involving emergency braking due to dynamic obstacles appearing in occluded areas and common give-way situations. Specifically, OAIAD excelled in handling high-interaction scenarios, such as merging and emergency braking under occlusion, where it showcased better adaptability and decision-making in comparison to existing methods.

2. Related Work

2.1. End-to-End Autonomous Driving

UniAD [4] represents one of the earliest end-to-end frameworks in autonomous driving, pioneering the unification of perception, prediction, and planning. By introducing a shared query interface and cross-module attention mechanisms, UniAD enabled multi-task joint optimization on the nuScenes dataset [6]. Building on this foundation, subsequent research has focused on three key directions—scene representation optimization, dynamic interaction modeling, and planning under uncertainty—each driving significant advancements.

Firstly, optimizing scene representations has been critical for enhancing efficiency and scalability. Sparse and vectorized representations have redefined information processing within end-to-end frameworks. SparseDrive [7] proposed a symmetric sparse perception module and hierarchical trajectory filtering, while VAD [8] introduced a vectorized

paradigm with instance-level constraints, achieving more efficient planning and reduced collision rates. Secondly, modeling dynamic interactions helps address the challenge of reasoning about complex multi-agent environments. GraphAD [9] proposed dynamic/static scene graphs to capture agent–map relationships, while VADv2 [10] introduced a probabilistic field function and trajectory vocabulary discretization to improve stability and interaction reasoning. Ulteriorly, Hydra-MDP [11] used more robust multitarget hydra distillation planning to solve the multi-objective optimization problem. Finally, planning under uncertainty has gained prominence for the role it plays in improving robustness. Generative approaches, such as GenAD [12], leverage variational autoencoders to model trajectory priors and generate multi-modal distributions, enabling reliable planning in complex scenarios.

In summary, the evolution of end-to-end autonomous driving can be characterized by three major trends: (1) transitioning from modular stacking architectures to goal-driven end-to-end optimization, (2) shifting from dense grid-based representations to lightweight vectorized modeling, and (3) progressing from deterministic decision-making frameworks to probabilistic multi-modal planning. These advancements collectively push autonomous driving systems toward more human-like decision-making paradigms, marking a significant step toward achieving robust, efficient, and reliable autonomous driving.

2.2. Occlusion Handling

Research on occlusion handling in autonomous driving has primarily focused on three key areas: perception enhancement, prediction and planning optimization, and end-to-end reasoning. Regarding perception, multi-modal fusion and deep learning models have emerged as dominant approaches. For example, Scene Informer and Occ3D infer occupancy and trajectory information in occluded areas using 3D voxel-based methods [13,14], while PWOC-3D [15] incorporates dynamic Bayesian networks and self-supervised learning to improve occlusion modeling. In prediction and planning, integrating deep reinforcement learning (DRL) with active perception has proven effective for managing uncertainty, while game-theoretic models and safety verification frameworks are employed to generate robust trajectories [16]. Additionally, lateral position optimization has been explored to enhance sensor visibility in occluded environments [17]. For end-to-end reasoning, advanced architectures such as ReasonNet [18] leverage transformers to integrate historical and topological context for global decision-making. Similarly, OGRIT [19] uses decision trees to achieve interpretable recognition of occluded goals, striking a balance between explainability and accuracy.

2.3. Multi-Modal Trajectory Interaction

In recent years, significant progress has been made in multi-modal interactive trajectory prediction and joint planning. From a theoretical perspective, hierarchical game-theoretic frameworks (e.g., GameFormer [20]) and structured graph-based models (e.g., FJMP [21]) have emerged as prominent approaches, modeling multi-agent interactions through Level- k reasoning and factorized representations, respectively. Regarding architectural design, the Holistic Transformer integrates sparse attention mechanisms to facilitate joint prediction and decision-making [22], while diffusion-based models, such as CDSTraj, address scene uncertainty by leveraging characterized diffusion processes [23]. End-to-end frameworks, including DIPP [24] and FusionAD [25], enable joint optimization across perception, prediction, and planning, thereby streamlining the decision-making pipeline. At the control level, DIPP [24] enhances planning-aware perception by coupling transformer-based prediction with a differentiable nonlinear optimizer, whereas ISAMP [26] employs uncertainty-aware sampling to dynamically adjust safety constraints in

real time. Despite these advancements, several challenges remain, including high inference latency, limited interaction modeling depth, and a lack of formal safety guarantees.

3. Materials and Methods

The overall architecture of the proposed Occlusion-Aware Interactive Autonomous Driving (OAIAD) framework is illustrated in Figure 2. As detailed in Section 3.1, given multi-view camera inputs, the framework first extracts visual features using a backbone network. These features are subsequently projected into a bird's-eye view (BEV) representation through a set of BEV queries, facilitating the transformation from the image domain into a top-down spatial feature space. This step enables a more structured and interpretable representation of the driving environment. To further enrich the representation, the framework employs agent, map, and occlusion queries to explicitly model their respective features. The extracted agent features are fused with ego-vehicle states and navigation commands to construct an instance-centric semantic representation. Additionally, cross-attention mechanisms are applied to integrate contextual information from the map and occlusion queries, resulting in a unified, semantically enriched scene representation that captures both spatial and contextual relationships within the environment.

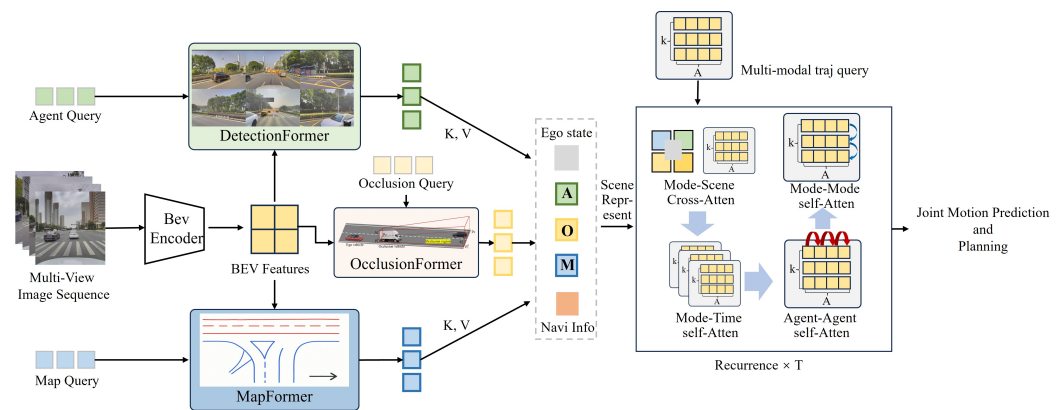


Figure 2. Overall architecture of OAIAD.

Building upon this representation, OAIAD jointly performs motion prediction and planning (as detailed in Section 3.2) through a recurrent, multi-agent, multi-modal trajectory decoding process. At each decoding step, the model facilitates interaction reasoning across agents, temporal sequences, modalities, and scene elements. This design captures the dynamic evolution of complex traffic scenarios, thereby improving both predictive accuracy and planning robustness.

The training and inference procedures of the OAIAD framework are described in Section 3.3, highlighting its end-to-end capability in terms of integrating occlusion-aware perception with multi-agent interaction modeling. This cohesive design enables the framework to address the challenges of autonomous driving in highly dynamic and occlusion-prone environments.

3.1. Scene Representation with Occlusion Awareness

3.1.1. Perception Feature Representation

Following VAD [8], we adopted a vectorized representation to model scene elements and extended it by incorporating an explicit occlusion representation module, enabling structured integration with both dynamic agents and static map elements.

Specifically, we first extracted multi-scale image features F using a CNN backbone and an FPN neck. These provide dense feature maps $F \in \mathbb{R}^{B \times C \times H \times W}$ for all views, where B is the batch size, C is the number of channels, and H and W are the height and

width of the feature maps, respectively. We then initialized a learnable BEV query tensor $B_0 \in \mathbb{R}^{H_{bev} \times W_{bev} \times C}$, where H_{bev} and W_{bev} represent the height and width of the BEV grid. This tensor is transformed into BEV features via deformable cross-attention as follows:

$$B = \text{DeformableCrossAttention}(B_0, F, F) \quad (1)$$

The goal of the deformable cross-attention mechanism is to adaptively aggregate spatial features from the image space to the BEV space while maintaining the spatial resolution and semantic information. This enables the model to flexibly learn relevant features from different views and dynamically integrate them into the BEV space, providing high-quality feature representations for subsequent tasks.

3.1.2. Vectorized Occlusion Region Modeling

To accurately characterize the occlusion regions in the ego vehicle's perception caused by preceding target vehicles, a modeling method is proposed based on geometric ray tracing. The target vehicle is abstracted as a regular cuboid model, with its position defined by the vehicle center coordinate $\mathbf{C} = (x_c, y_c, z_c)$ and the parameters length L , width W , and height H . The position and size of the target vehicle are assumed to be known or estimated from sensor data. The LiDAR sensor mounted on the ego vehicle is treated as the ray emission source $\mathbf{S} = (x_s, y_s, z_s)$, from which rays are emitted toward the target vehicle. A ray can be expressed in parametric form as

$$\mathbf{R}(t) = \mathbf{S} + t\mathbf{d}, \quad t \geq 0, \quad (2)$$

where $\mathbf{d} = \mathbf{P} - \mathbf{S}$ is the direction vector pointing to the intersection point $\mathbf{P}$ on the target vehicle's surface. The direction vector $\mathbf{d}$ is computed as the difference between the position of the ego vehicle $\mathbf{S}$ and the intersection point $\mathbf{P}$ on the target vehicle's surface. The parameter t represents the distance along the direction vector $\mathbf{d}$ from the emission source $\mathbf{S}$ to the intersection point $\mathbf{P}$ on the target vehicle's surface. By computing the intersection points for different values of t , we can determine all the intersections between the ray and the target vehicle.

The occlusion boundary is determined as follows by the pair of rays with the maximum included angle:

$$\theta_{\max} = \max_{i,j} \arccos \frac{(\mathbf{P}_i - \mathbf{S}) \cdot (\mathbf{P}_j - \mathbf{S})}{\|\mathbf{P}_i - \mathbf{S}\| \|\mathbf{P}_j - \mathbf{S}\|}, \quad (3)$$

where $\mathbf{P}_i, \mathbf{P}_j$ denote the intersection points between the cuboid surface and the rays.

Once the boundary rays have been determined, the rays are extended to the ground or the perception boundary to form a ground quadrilateral. By connecting the quadrilateral vertices with the top of the vehicle, a hexahedron is obtained, the base of which is the quadrilateral and the top surface of which is determined by height extension. This method enables the precise description of the invisible perception region caused by the target vehicle, providing quantitative input for trajectory prediction and risk assessment.

The constructed occlusion regions, as shown in Figure 3, take the geometric form of hexahedra, each represented by a set of closed vectors.

$$\mathcal{P} = \{p_0, p_1, \dots, p_n\} \quad (4)$$

where each vector segment p_i encodes both spatial and vertical occlusion attributes as follows:

$$p_i = (\mathbf{p}_s^i, h_s^i, \mathbf{p}_e^i, h_e^i), \quad i = 0, 1, \dots, n \quad (5)$$

Here

- $\mathbf{p}_s^i, \mathbf{p}_e^i \in \mathbb{R}^2$ represents the start/end BEV coordinates of the i -th edge;
- $h_s^i, h_e^i \in \mathbb{R}$ represents the occlusion height at the start and end points.

Each region is encoded as N_p vectors, resulting in the occlusion feature tensor

$$\hat{V}_c \in \mathbb{R}^{N_c \times N_p \times 6} \quad (6)$$

where N_c is the number of predicted occlusion regions, and each vector includes four spatial coordinates and two height values.

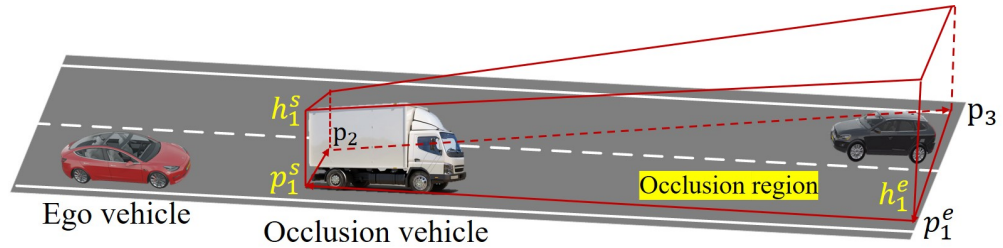


Figure 3. Illustration of occlusion region features.

3.1.3. Agent and Map Feature Representation

In parallel, we extracted features for dynamic agents and static maps as follows:

- Agent features: $\hat{V}_a \in \mathbb{R}^{N_a \times D}$, where D includes attributes like position, orientation, and category confidence;
- Map features: $\hat{V}_m \in \mathbb{R}^{N_m \times N_p \times 2}$, representing vectorized structures such as lane dividers, road boundaries, and crosswalks.

3.1.4. Query-Based Semantic Feature Extraction

To integrate the above representations into semantic embeddings, we assigned dedicated queries to each entity type as follows:

$$M = \text{CrossAttention}(M_0, B, B) \quad (7)$$

$$A = \text{CrossAttention}(A_0, B, B) \quad (8)$$

$$C = \text{CrossAttention}(C_0, B, B) \quad (9)$$

where M_0 , A_0 , and C_0 are learnable initial queries for map, agent, and occlusion, respectively. The outputs M , A , and C are context-aware embeddings that encode geometry, appearance, and spatial layout.

3.1.5. Scene Feature Fusion and Instance-Level Reasoning

Based on agent features A , we incorporate the ego-vehicle state e and navigation goal n to form an instance-centric representation

$$I_0 = \text{Concat}(e, n, A) \quad (10)$$

To model the influence between instances, we apply self-attention to I_0 as follows:

$$I_1 = \text{SelfAttention}(I_0, I_0, I_0) \quad (11)$$

We then perform cross-attention from I_1 to map and occlusion features to inject context

$$I_2 = \text{CrossAttention}(I_1, M, M) \quad (12)$$

$$I = \text{CrossAttention}(I_2, C, C) \quad (13)$$

The resulting representation $I \in \mathbb{R}^{N_a \times D}$ encodes spatial context, inter-agent interactions, topological layout, and visibility constraints, enabling robust scene understanding for downstream prediction and planning.

3.2. Joint Network for Multi-Modal Trajectory Prediction and Planning

To enable efficient trajectory prediction and joint planning for multiple agents in complex traffic scenarios, we propose a unified network that performs multi-modal trajectory prediction and planning in an end-to-end manner. This network supports the joint optimization of trajectory generation and path planning, and possesses strong capabilities regarding social interaction modeling.

As illustrated in Figure 4, during the initialization phase, we randomly generated K latent mode embeddings of dimension D , denoted as

$$\mathbf{Z}_0 \in \mathbb{R}^{K \times D} \quad (14)$$

Each row represents a candidate trajectory mode in the semantic space. To jointly model A agents per mode, each mode embedding is duplicated A times, forming a 3D tensor

$$\mathbf{Z}_{\text{init}} \in \mathbb{R}^{K \times A \times D} \quad (15)$$

Here, $\mathbf{Z}_{\text{init}}[k, :, :]$ encodes the joint future scene under mode k , while $\mathbf{Z}_{\text{init}}[k, a, :]$ represents the latent trajectory of agent a in mode k .

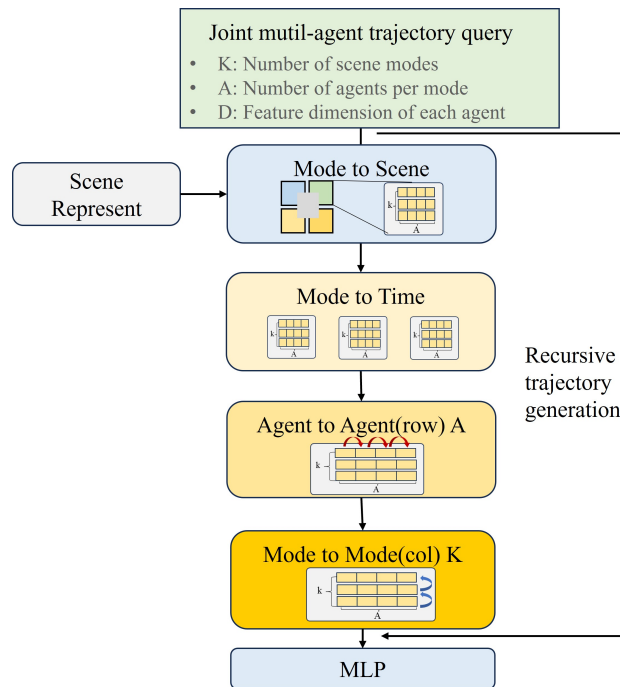


Figure 4. Joint network for multi-modal trajectory prediction and planning.

The network comprises three key modules stacked over L_{dec} decoding layers to progressively update the embedding tensor as follows:

- (1) Mode-to-Scene Cross-Attention: Incorporates scene-level features (e.g., map, occlusion, and historical trajectories) into each mode embedding to enhance environmental awareness

$$\mathbf{Z}^{(1)}[k, :, :] = \text{CrossAttn}_{\text{scene}}(\mathbf{Z}_{\text{init}}[k, :, :], \mathbf{F}_{\text{scene}}) \quad (16)$$

- (2) Mode-to-Time Cross-Attention: Captures temporal dynamics in trajectory evolution (e.g., acceleration and lane changes) with time-aware cross-attention

$$\mathbf{Z}^{(2)}[k, :, :] = \text{CrossAttn}_{\text{time}}(\mathbf{Z}^{(1)}[k, :, :], \mathbf{F}_{\text{temporal}}) \quad (17)$$

- (3) Agent Self-Attention: Models interactions between agents (e.g., yielding, following, and merging) within the same mode

$$\mathbf{Z}^{(3)}[k, :, :] = \text{SelfAttn}_{\text{agent}}(\mathbf{Z}^{(2)}[k, :, :]) \quad (18)$$

To further enhance diversity and stability across trajectory modes, we introduced a Mode Self-Attention module to enable information sharing and competition among different modes

$$\mathbf{Z}^{(4)} = \text{SelfAttn}_{\text{mode}}(\mathbf{Z}^{(3)}) \quad (19)$$

An MLP decoder then maps the final embeddings $\mathbf{Z}^{(4)}$ to generate the first 2 s of joint trajectory prediction. A recursive mechanism is used to extend predictions up to 6 s, ensuring temporal consistency and adaptive rollout.

After generating the initial trajectory proposals, we designed a trajectory refinement module to further enhance prediction accuracy and local consistency. Taking the proposed trajectories as anchors, we directly constructed an embedding tensor from trajectory points without random initialization. This tensor is then processed through a network with a similar structure, sequentially updated by temporal cross-attention, map attention, and social interaction attention.

Finally, an MLP decoder predicts 2D offsets for each anchor point, yielding fine-grained trajectory refinement and ego-vehicle decision fusion.

3.3. Training and Optimization

3.3.1. Vectorized Scene Learning

The vectorized scene learning module consists of three sub-tasks: vectorized map modeling, vectorized dynamic obstacle modeling, and vectorized occlusion region modeling. These components are designed to capture, respectively, high-definition map structures, the behavior of surrounding moving objects, and the influence of occlusion on planning.

- **Vectorized Map Modeling:** We used Manhattan distance as the regression metric to measure the geometric deviation between predicted and ground-truth map points. In addition, focal loss [27] was introduced as a classification objective to enhance the model's focus on key semantic map elements, such as lane dividers, stop lines, and crosswalks. The overall loss for this module is denoted as L_{map} .
- **Vectorized Dynamic Obstacle Modeling:** This subtask models the vectorized trajectories of dynamic targets in the traffic environment (e.g., vehicles or pedestrians) to achieve a structured representation of their behavior. The corresponding loss is denoted as L_{obj} .
- **Vectorized Occlusion Region Modeling:** This module learns the spatial distribution features of potential occlusion regions, which helps improve the model's reasoning ability for occluded or invisible targets. The loss function for this component is denoted as L_{occ} .

3.3.2. Multi-Modal Motion Prediction and Planning Constraint Modeling

Referring to QCNeXt [28], we parameterized the joint future trajectory distribution of all target agents as a mixture of Laplace distributions, formulated as follows:

$$\sum_{k=1}^K m_k \prod_{i=1}^{A'} \prod_{t=1}^{T'} f(\mathbf{p}_i^{t,x} | \boldsymbol{\mu}_{i,k}^{t,x}, \mathbf{b}_{i,k}^{t,x}) f(\mathbf{p}_i^{t,y} | \boldsymbol{\mu}_{i,k}^{t,y}, \mathbf{b}_{i,k}^{t,y}), \quad (20)$$

In this formulation, the mixture of Laplace distributions is used to model the joint future trajectory distribution of all target agents. The index k represents the k -th trajectory mode in the mixture, where K is the total number of modes. The weight of each mode is denoted by m_k , with the sum of the weights across all modes equal to 1, i.e., $\sum_{k=1}^K m_k = 1$.

The term $f(\mathbf{p}_i^{t,x} | \boldsymbol{\mu}_{i,k}^{t,x}, \mathbf{b}_{i,k}^{t,x})$ represents the Laplace probability density function (PDF) for the x -coordinate of the i -th agent's position at time step t in the k -th mode. The mean $\boldsymbol{\mu}_{i,k}^{t,x}$ and the uncertainty $\mathbf{b}_{i,k}^{t,x}$ are the predicted mean and uncertainty for the x -coordinate at time step t for the i -th agent in the k -th mode. Similarly, $f(\mathbf{p}_i^{t,y} | \boldsymbol{\mu}_{i,k}^{t,y}, \mathbf{b}_{i,k}^{t,y})$ represents the Laplace PDF for the y -coordinate, with the mean $\boldsymbol{\mu}_{i,k}^{t,y}$ and uncertainty $\mathbf{b}_{i,k}^{t,y}$ for the y -coordinate at time step t .

Each mode models a distribution with a dimensionality of $A' \times T' \times 2$, where A' is the number of predicted agents (with $A' \leq A$, where A is the total number of agents), and T' is the number of future time steps being predicted. This makes it possible to model the potential future positions of each agent in both the x and y directions across a series of time steps.

For each agent i in mode k , at each time step t , the predicted mean $\boldsymbol{\mu}_{i,k}^t = (\mu_{i,k,t}^x, \mu_{i,k,t}^y)$ and the predicted uncertainty $\mathbf{b}_{i,k}^t = (b_{i,k}^{t,x}, b_{i,k}^{t,y})$ are computed to represent the agent's expected future position and its associated uncertainty in the x and y directions, respectively.

3.3.3. Overall End-to-End Training Objective

We formulated the entire OAIAD system as an end-to-end trainable framework. The final overall loss function is defined as a weighted sum of the vectorized scene modeling loss, the multi-modal trajectory modeling loss, and the imitation learning loss, as follows:

$$L = \omega_1 L_{\text{map}} + \omega_2 L_{\text{obj}} + \omega_3 L_{\text{occ}} + \omega_4 L_{\text{propose}} + \omega_5 L_{\text{cls}}, \quad (21)$$

where ω_i are the weighting coefficients for each loss term, which can be adjusted according to specific task settings. L_{propose} is the trajectory proposal loss, and L_{cls} is a classification loss, which is defined as the negative log-likelihood of Equation (20). A scene-level winner-takes-all strategy is employed to select the mode with the lowest displacement error as the supervisory signal.

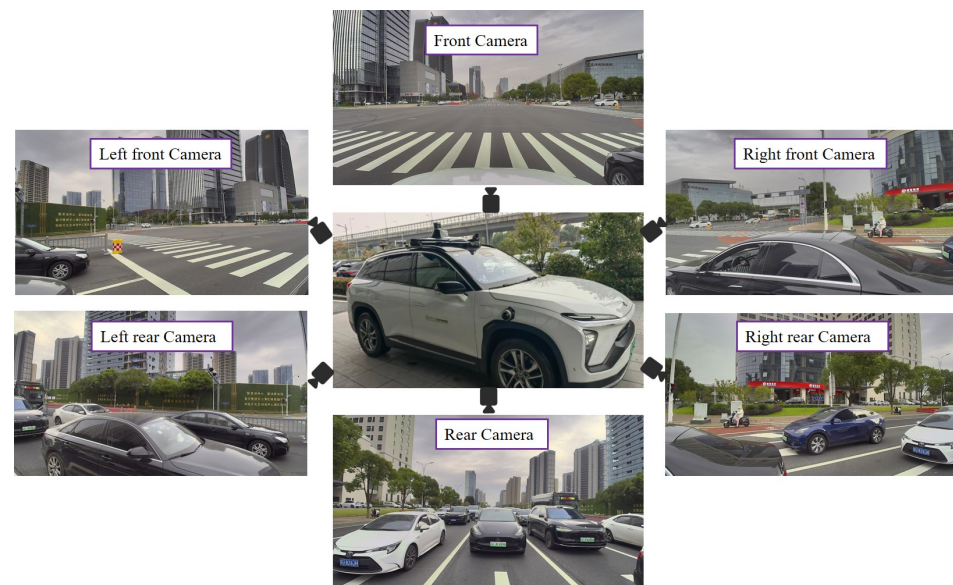
4. Experiments

4.1. Experimental Settings

We evaluated the proposed OAIAD framework on two widely used datasets. Firstly, we used the open-loop dataset, nuScenes dataset [6], which contains 1000 complex urban driving scenes, each lasting 20 s. Following the official split, 700 scenes were used for training, 150 for validation, and 150 for testing. Secondly, to quantitatively assess the model's capability in terms of handling right-of-way conflicts, we assessed closed-loop performance using the Bench2Drive dataset [29], a large-scale dataset comprising 2 million fully annotated frames collected from 13,638 short clips. These clips are uniformly distributed across 44 interactive scenarios (e.g., cutting in, overtaking, or detouring), 23 weather conditions

(e.g., sunny, foggy, or rainy), and 12 towns. For this dataset, we used 950 clips for training and reserved 50 clips for testing.

Furthermore, since the nuScenes dataset contains a limited number of dynamic occlusion scenarios and lacks systematic annotations for key aspects of the occlusions, such as occlusion type (static/dynamic), occlusion duration, and interaction patterns, we collected a dataset in the urban areas of Xiangcheng District, Suzhou, China, specifically focusing on occlusion and interaction scenarios (Figure 5a). A vehicle–road–cloud collaborative system (Figure 5b) was employed to provide high-quality ground-truth data for constructing the dataset. This dataset was specifically designed to assess the framework’s effectiveness regarding the handling of complex occlusions and interactions in real-world environments.



(a) **Experimental vehicle sensor configuration.** Visualization of multi-view images captured by the six surround-view cameras at a complex urban intersection.



(b) **Visualization of the vehicle–road–cloud collaborative system,** used to provide ground-truth labels, contains objects, maps, and trajectories.

Figure 5. The real-world test.

To ensure the robustness of the dataset, we collected data using six 120° FOV wide-angle high-definition cameras with a frame rate of 2 Hz. The data collection system was equipped with synchronized cameras to ensure precise timestamp alignment, allowing the capture of both dynamic and static occlusion scenarios. The camera configuration consisted

of six wide-angle lenses, providing comprehensive scene coverage and effectively capturing complex occlusions and interactions. The data collection process was meticulously designed to capture interaction scenarios involving traffic participants in urban environments. Ultimately, we defined scenes with an occlusion ratio greater than 20% as occlusion scenes, and scenes with more than 20 traffic participants as interaction scenes. We then selected 500 key video clips, each with a duration of 20 s, and added them to the dataset.

We used ResNet-50 as the image backbone and projected features into a BEV space of size 200×200 . Our implementation followed the VAD [8] architecture and hyperparameter setup. All BEV features, agent embeddings, and instance tokens had a hidden dimension of 256. We trained the model using the AdamW optimizer with an initial learning rate of 2×10^{-4} and a weight decay of 0.01. Experiments were conducted on six NVIDIA RTX 4090 GPUs for 60 epochs with a batch size of 8.

4.2. Performance Evaluation on Open Datasets

Firstly, as shown in the open-loop evaluation results (Table 1), OAIAD outperforms the other three baselines. The evaluation results are based on the following key metrics:

- **L2 Error:** This refers to the Euclidean distance between the predicted and actual trajectories at three different time steps (1 s, 2 s, and 3 s). The lower the L2 error, the more accurate the predicted trajectory. The “Avg.” column represents the average error across these time steps.
- **Collision Rate:** This measures the percentage of episodes in which the ego vehicle collides with obstacles. It is calculated at the 1 s, 2 s, and 3 s marks, with a lower value indicating fewer collisions. The “Avg.” column provides the average collision rate across all time steps.

Table 1. Comparison of open-loop performance on the nuScenes dataset [6].

Method	L2 (m) ↓				Collision Rate (%) ↓			
	1 s	2 s	3 s	Avg.	1 s	2 s	3 s	Avg.
UniAD [4]	0.48	0.76	1.65	1.03	0.05	0.17	1.27	0.71
VAD [8]	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22
SparseDrive-B [7]	0.29	0.55	0.91	0.58	0.01	0.02	0.13	0.06
OAIAD (Ours)	0.23	0.51	0.83	0.50	0.01	0.01	0.09	0.04

Subsequently, we conducted a comprehensive evaluation of OAIAD on the Bench2Drive dataset [29], as presented in Table 2. The results demonstrate significant improvements in both the overall driving score and success rate as follows:

- **Driving Score:** This composite score evaluates the overall performance of the vehicle across various driving tasks, including lane-keeping, speed regulation, and interaction with other vehicles. It aggregates multiple factors that contribute to the vehicle’s ability to safely and efficiently complete driving tasks.
- **Success Rate:** This measures the percentage of scenarios in which the ego vehicle successfully completes the task without any major failures, such as collisions or off-road excursions. A higher success rate reflects better overall performance in task execution.

A detailed comparison of individual capabilities reveals notable advancements in scenarios with high interaction demands, such as merging or emergency braking scenarios, predominantly caused by occlusions. These results demonstrate OAIAD’s ability to effectively fuse occlusion features and model the behavioral interactions between the ego vehicle and surrounding traffic participants, thereby generating more reasonable and adaptive driving strategies.

Table 2. Comparison of closed-loop performance on the Bench2Drive dataset [29].

Method	Multi-Ability (%) ↑					Overall ↑		
	Merging	Overtaking	Emergency Brake	Give Way	Traffic Sign	Mean	Driving Score	Success Rate
UniAD [4]	14.1	17.78	21.67	10	14.21	15.55	45.81	16.36
VAD [8]	8.11	24.44	18.64	20	19.15	18.07	42.35	15
DriveTransformer-Large [30]	17.57	35	48.36	40	52.1	38.6	60.45	30
DiffAD [31]	30	35.55	46.66	40	46.32	38.79	67.92	38.64
OAIAD (Ours)	32.68	34.46	63.47	40	42.39	42.6	68.73	48.86

In addition, to verify the correctness of OAIAD’s behavior in occlusion and interaction scenarios, we visualized and compared the results of OAIAD and VAD [8] in Figure 6. In the first two occlusion scenarios, VAD directly accelerates through the scene, whereas OAIAD adopts defensive driving strategies, such as proactive deceleration or a slight lateral movement to the right. These actions effectively increase the field of view and significantly mitigate potential risks caused by occluded areas. In the latter two interaction scenarios, due to the uncertainty of other vehicles’ intentions, multiple predicted trajectories are generated. VAD tends to stop the ego vehicle completely, exhibiting overly cautious behavior. In contrast, OAIAD enables the ego vehicle to decelerate and execute a slight rightward movement to engage with the conflicting vehicle. This adaptive interaction capability demonstrates OAIAD’s robustness in handling complex urban driving environments.

Additionally, to verify the effectiveness of incorporating occlusion region learning and joint trajectory prediction and planning, we conducted an ablation study on these two components, with the results shown in Table 3. We used VAD as the baseline for comparison. When only the occlusion modeling module was added to VAD, a performance improvement was observed. Moreover, when only the sequential prediction and planning module was replaced with the joint trajectory prediction and planning module, the performance also increased. OAIAD, which utilizes both of these components simultaneously, achieved the best performance. These results indicate that the explicit learning of occlusion regions can enhance the network’s ability to estimate risks in unseen areas, while joint trajectory prediction and planning improve the model’s capability for interactive reasoning.

Table 3. Ablation study on occlusion modeling and interaction-aware planning.

ID	Occlusion Feature	Interactive Planner	L2 (m) ↓				Collision Rate (%) ↓			
			1 s	2 s	3 s	Avg.	1 s	2 s	3 s	Avg.
1	×	×	0.41	0.70	1.05	0.72	0.07	0.17	0.41	0.22
2	✓	×	0.36	0.63	0.81	0.60	0.04	0.13	0.31	0.16
3	×	✓	0.33	0.59	0.84	0.58	0.05	0.14	0.28	0.15
4	✓	✓	0.23	0.51	0.83	0.50	0.01	0.01	0.09	0.04

4.3. Performance Evaluation on Real-World Dataset

To validate the real-world performance of OAIAD, we visualized and analyzed the model’s output in several representative scenarios. As shown in Figure 7, OAIAD demonstrates robust performance in urban environments, particularly in occlusion and complex interaction scenarios. Specifically, in an occlusion scenario where a vehicle is located in the front-left direction, OAIAD prompts the ego vehicle to decelerate and slightly shift to the right, effectively avoiding potential collision risks. This behavior is particularly important in real-world driving, especially when obstructions appear within the field of view, and obstacles may emerge unexpectedly. OAIAD is able to respond quickly and adjust its trajectory to ensure safety, thus proving its reliability in dynamic urban environments.

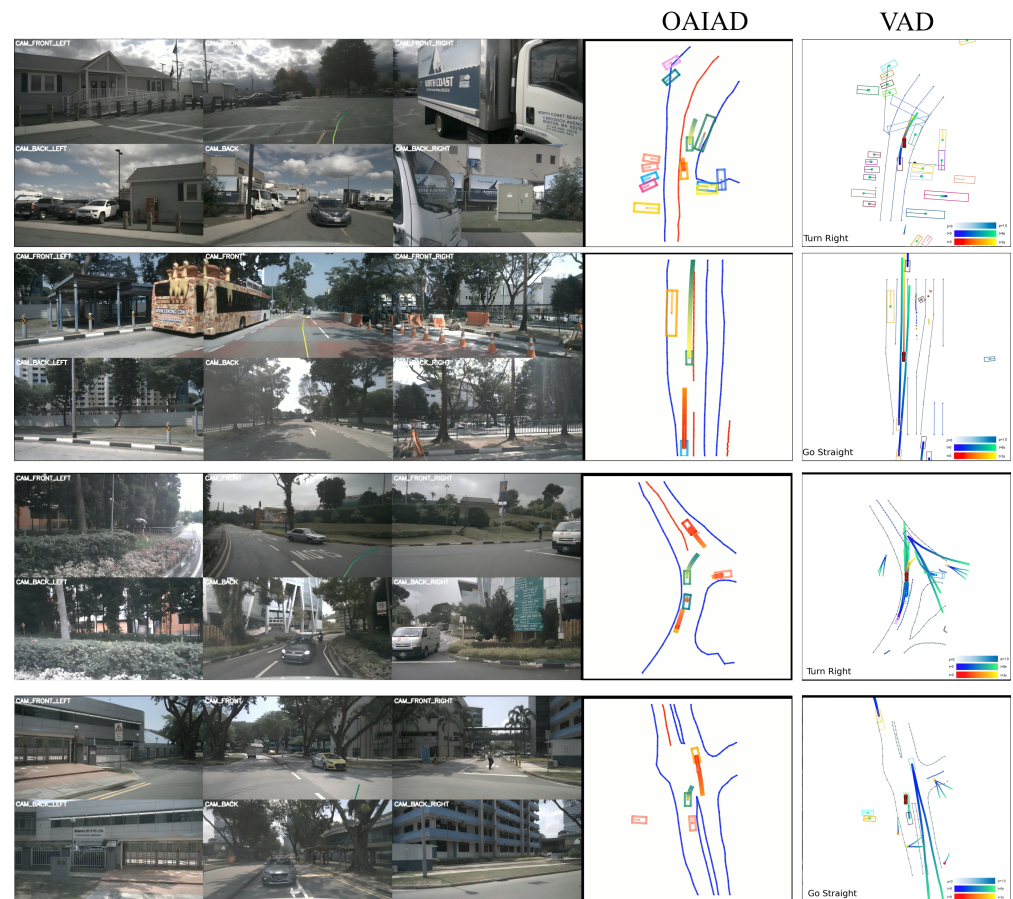


Figure 6. Visualization of planning results for OAIAD and VAD [8]. The left side displays images from six surround-view cameras, while the right side presents the results of map segmentation, object detection, motion prediction, and planning for both OAIAD and VAD. The green box and trajectory indicate the ego vehicle. Four representative scenarios are highlighted: (1) an occlusion scenario caused by a truck on the right, (2) an occlusion scenario caused by a parked bus on the left, (3) an unprotected right turn at an intersection, and (4) a conflict scenario involving a right-turning ego vehicle and a straight-moving vehicle.

Additionally, in situations where surrounding vehicles occupy both the front and rear of the adjacent left lane, OAIAD enables the ego vehicle to gradually change lanes to the left and engage in exploratory interactions with nearby vehicles. By evaluating the surrounding traffic conditions in real time and making the appropriate decisions, OAIAD successfully plans a reasonable trajectory in complex interaction environments. This capability in interaction scenarios demonstrates OAIAD's ability to model multi-agent behaviors, enabling it to handle complex urban traffic scenarios and avoid overly cautious or irrational decision-making.

The results further demonstrate OAIAD's adaptability and decision-making capability. In occlusion scenarios, OAIAD successfully predicts hidden obstacles and adjusts its trajectory, ensuring that the ego vehicle can make safe driving decisions within occluded areas. This behavior not only improves the system's robustness in complex environments but also ensures quick and accurate responses when facing unforeseen risks. Moreover, OAIAD's performance in lane-change scenarios and interactions with surrounding vehicles shows its superiority in complex traffic conditions. For example, during interactions with other vehicles, OAIAD adjusts its trajectory in real time, preventing potential risks caused by lane conflicts or violations of traffic rules.



Figure 7. Visualization of OAIAD's results on the real-world dataset. The top row shows planning results for the occlusion scenarios, while the bottom row illustrates the results for lane-change interaction scenarios. The green box and trajectory represent the ego vehicle.

To evaluate the inference efficiency of the model, we tested the average inference time of OAIAD on a single RTX 4090 GPU, with an average processing time of 230 ms (approximately 4.3 FPS) for a single frame of data, which achieves better performance than VAD with a comparable inference time.

In conclusion, OAIAD demonstrates excellent performance in real-world applications, especially in handling occlusion and multi-agent interaction scenarios. Its real-time response capability and decision-making quality enable it to cope with the complex and dynamic driving environments in the real world, further proving its potential for deployment in autonomous driving systems requiring high safety and adaptability.

5. Conclusions and Future Work

This paper presents OAIAD (Occlusion-Aware Interactive Autonomous Driving), a unified framework for complex traffic scenarios that integrates occlusion-aware perception and interactive planning into an end-to-end system. We design a multi-modal trajectory prediction and planning network with strong interaction modeling capabilities, incorporating mode-to-scene attention and recursive trajectory generation to enhance trajectory consistency and physical plausibility. Comprehensive evaluations on both open- and closed-loop datasets demonstrate that OAIAD outperforms state-of-the-art methods across multiple metrics, showcasing its superior end-to-end capability. Furthermore, real-world testing validates the effectiveness of the proposed framework at addressing challenging occlusion and complex interaction scenarios.

Crucially, our framework demonstrates superior performance in resolving right-of-way conflicts, a core challenge for urban autonomous driving. The joint optimization of prediction and planning enables OAIAD to generate interactive behaviors that are not only safe but also socially compliant, enabling the smooth and efficient negotiation of complex scenarios like unsignalized intersections. This represents a significant step toward developing autonomous systems that can seamlessly integrate into human-dominated traffic ecosystems.

Despite its strong performance across key tasks, there is still room for future improvements in OAIAD such as (1) enhancing generalization across datasets to assess transferability; (2) developing more efficient trajectory generation mechanisms to reduce the computational cost of recursion; (3) integrating VLMs (Vision Language Models) with

enhanced understanding of object occlusion and interaction dynamics improves generalization in real-world applications.

In summary, OAIAD demonstrates the potential to deeply integrate occlusion perception and interactive modeling into a unified autonomous driving framework, offering a promising direction toward safer, more efficient, and generalizable self-driving systems.

Author Contributions: Conceptualization, J.Y. (Jialun Yin); methodology, J.Y. (Jialun Yin) and K.Z.; software, K.Z.; validation, K.Z. and J.Y. (Junru Yang); formal analysis, K.Z. and X.M.; investigation, K.Z. and Y.C.; resources, J.Y. (Jialun Yin) and H.L.; data curation, X.M. and S.Y.; writing—original draft preparation, X.M. and Y.C.; writing—review and editing, K.Z., X.M. and J.Y. (Jialun Yin); visualization, X.M. and J.Y. (Junru Yang); supervision, H.L.; project administration, J.Y. (Jialun Yin); funding acquisition, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (2023YFB4302600); the Natural Science Foundation of Jiangsu Province (BK20231197, BK20220243); the Science and Technology Program of Suzhou (SYG2024057, SYC2022078); the Hubei Science and Technology Talent Service Enterprise Project (2023DJC084, 2023DJC195); and the Hubei Science and Technology Project (2024BAB087). Additionally, the study was sponsored by Tsinghua-Toyota Joint Research Institute Interdisciplinary Program and Jiangsu Science and Technology Association’s Young Science Talent Nurturing Program—Research of Collaborative End-to-End Autonomous Driving.

Data Availability Statement: Due to project confidentiality requirements involving the data used in this paper, there are currently no plans for a public release of the data, and the data cannot be provided.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fresh. Safety Considerations for Autonomous Vehicles. 2024. Available online: <https://www.sgpjbg.com.cn/baogao/169884.html> (accessed on 11 September 2025).
2. Li, M.; Li, G.; Sun, C.; Yang, J.; Li, H.; Li, J.; Li, F. A shared-road-rights driving strategy based on resolution guidance for right-of-way conflicts. *Electronics* **2024**, *13*, 3214. [[CrossRef](#)]
3. Kim, M.J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanketi, P.; et al. OpenVLA: An open-source vision-language-action model. *arXiv* **2024**, arXiv:2406.09246.
4. Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. Planning-oriented autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 17853–17862.
5. Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; Geiger, A. Occupancy networks: Learning 3d reconstruction in function space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4460–4470.
6. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuScenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11621–11631.
7. Sun, W.; Lin, X.; Shi, Y.; Zhang, C.; Wu, H.; Zheng, S. Sparsedrive: End-to-end autonomous driving via sparse scene representation. In Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA), Atlanta, GA, USA, 19–23 May 2025; pp. 8795–8801.
8. Jiang, B.; Chen, S.; Xu, Q.; Liao, B.; Chen, J.; Zhou, H.; Zhang, Q.; Liu, W.; Huang, C.; Wang, X. Vad: Vectorized scene representation for efficient autonomous driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 8340–8350.
9. Zhang, Y.; Qian, D.; Li, D.; Pan, Y.; Chen, Y.; Liang, Z.; Zhang, Z.; Zhang, S.; Li, H.; Fu, M.; et al. Graphad: Interaction scene graph for end-to-end autonomous driving. *arXiv* **2024**, arXiv:2403.19098.
10. Chen, S.; Jiang, B.; Gao, H.; Liao, B.; Xu, Q.; Zhang, Q.; Huang, C.; Liu, W.; Wang, X. Vad2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv* **2024**, arXiv:2402.13243.
11. Li, Z.; Li, K.; Wang, S.; Lan, S.; Yu, Z.; Ji, Y.; Li, Z.; Zhu, Z.; Kautz, J.; Wu, Z.; et al. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv* **2024**, arXiv:2406.06978.

12. Zheng, W.; Song, R.; Guo, X.; Zhang, C.; Chen, L. Genad: Generative end-to-end autonomous driving. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer Nature: Cham, Switzerland, 2024; pp. 87–104.
13. Lange, B.; Li, J.; Kochenderfer, M.J. Scene informer: Anchor-based occlusion inference and trajectory prediction in partially observable environments. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024; pp. 14138–14145.
14. Tian, X.; Jiang, T.; Yun, L.; Mao, Y.; Yang, H.; Wang, Y.; Wang, Y.; Zhao, H. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 64318–64330.
15. Saxena, R.; Schuster, R.; Wasenmuller, O.; Stricker, D. PWOC-3D: Deep occlusion-aware end-to-end scene flow estimation. In Proceedings of the 2019 IEEE Intelligent Vehicles Symposium (IV), Paris, France, 9–12 June 2019; pp. 324–331.
16. Zhang, Z.; Fisac, J.F. Safe occlusion-aware autonomous driving via game-theoretic active perception. *arXiv* **2021**, arXiv:2105.08169.
17. Narksri, P.; Darweesh, H.; Takeuchi, E.; Ninomiya, Y.; Takeda, K. Occlusion-aware motion planning with visibility maximization via active lateral position adjustment. *IEEE Access* **2022**, *10*, 57759–57782. [[CrossRef](#)]
18. Shao, H.; Wang, L.; Chen, R.; Waslander, S.L.; Li, H.; Liu, Y. Reasonnet: End-to-end driving with temporal and global reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 13723–13733.
19. Brewitt, C.; Tamborski, M.; Wang, C.; Albrecht, S.V. Verifiable goal recognition for autonomous driving with occlusions. In Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 1–5 October 2023; pp. 11210–11217.
20. Huang, Z.; Liu, H.; Lv, C. Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 3903–3913.
21. Wang, Y.; Tang, C.; Sun, L.; Rossi, S.; Xie, Y.; Peng, C.; Hannagan, T.; Sabatini, S.; Poerio, N.; Tomizuka, M.; et al. Optimizing diffusion models for joint trajectory prediction and controllable generation. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer Nature: Cham, Switzerland, 2024; pp. 324–341.
22. Hu, H.; Wang, Q.; Zhang, Z.; Li, Z.; Gao, Z. Holistic transformer: A joint neural network for trajectory prediction and decision-making of autonomous vehicles. *Pattern Recognit.* **2023**, *141*, 109592. [[CrossRef](#)]
23. Liao, H.; Li, X.; Li, Y.; Kong, H.; Wang, C.; Wang, B.; Guan, Y.; Tam, K.; Li, Z. Cdstraj: Characterized diffusion and spatial-temporal interaction network for trajectory prediction in autonomous driving. In Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, Jeju, Republic of Korea, 3–9 August 2024; pp. 7331–7339.
24. Huang, Z.; Liu, H.; Wu, J.; Lv, C. Differentiable integrated motion prediction and planning with learnable cost function for autonomous driving. *IEEE Trans. Neural Networks Learn. Syst.* **2023**, *35*, 15222–15236. [[CrossRef](#)] [[PubMed](#)]
25. Ye, T.; Jing, W.; Hu, C.; Huang, S.; Gao, L.; Li, F.; Wang, J.; Guo, K.; Xiao, W.; Mao, W.; et al. Fusionad: Multi-modality fusion for prediction and planning tasks of autonomous driving. *arXiv* **2023**, arXiv:2308.01006.
26. Zhou, J.; Olofsson, B.; Frisk, E. Interaction-aware motion planning for autonomous vehicles with multi-modal obstacle uncertainty predictions. *IEEE Trans. Intell. Veh.* **2023**, *9*, 1305–1319. [[CrossRef](#)]
27. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
28. Zhou, Z.; Wen, Z.; Wang, J.; Li, Y.H.; Huang, Y.K. Qcnext: A next-generation framework for joint multi-agent trajectory prediction. *arXiv* **2023**, arXiv:2306.10508.
29. Jia, X.; Yang, Z.; Li, Q.; Zhang, Z.; Yan, J. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 819–844.
30. Jia, X.; You, J.; Zhang, Z.; Yan, J. Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. *arXiv* **2025**, arXiv:2503.07656.
31. Wang, T.; Zhang, C.; Qu, X.; Li, K.; Liu, W.; Huang, C. DiffAD: A Unified Diffusion Modeling Approach for Autonomous Driving. *arXiv* **2025**, arXiv:2503.12170.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.