Department of Creative Informatics
Graduate School of Information Science and Technology
THE UNIVERSITY OF TOKYO

Master's Thesis

Cross-Attention Enhanced End-to-End Autonomous Driving in Unprotected Turns

右折時における相互注意機構を用いたエンドツーエンド自動運転

Dongyang Li
李 東洋

Supervisor: Professor Manabu Tsukada

January 2025

Abstract

Performing an unprotected turn at the intersection is a complex scenario for autonomous vehicles. It requires a comprehensive understanding of the surrounding environment, while also highly relying on the current state of the ego vehicle to make safer decisions. A conventional way to learn end-to-end autonomous driving is imitation learning, which stands for learning from expert demonstrations. While most imitation learning methods focus on imitating the expert action, they often fail to imitate a complex policy efficiently when the ego vehicle's states are crucial to the scenario because there might be arbitrary optimal actions under different states. To address this issue and investigate how vehicle states affect autonomous driving, we present a novel cross-attention enhanced imitation learning approach for end-to-end autonomous driving in unprotected turns, focusing on capturing the relationships between the ego vehicle's states and its perception of the environment. We evaluated our model in the AWSIM simulator, an open source autonomous driving simulator, and the results demonstrate that our model outperformed conventional imitation learning-based baselines in performing unprotected turn scenarios, showcasing its ability to imitate a complex policy efficiently.

概要

交差点での右折は、自動運転車にとって非常に複雑なシナリオである。これを安全に実行するには、周囲環境を包括的に理解するだけでなく、自車両（ego vehicle）の現在の状態を十分に考慮した上で安全な意思決定を行う必要がある。エンドツーエンドの自動運転を学習する従来の方法として、エキスパートデモンストレーションから学ぶ模倣学習（imitation learning）が広く用いられている。しかし、ほとんどの模倣学習手法はエキスパート行動の模倣（モデル出力）に焦点を当てているが、自車状態などの入力については研究されていない。その結果、自己車両の状態がシナリオにおいて重要である場合、異なる状態における最適な行動が任意に変化する可能性があるため、複雑な方策（policy）を効率的に模倣することが困難である。本研究では、この課題に対して、自己車両の状態が自動運転に与える影響を検討するため、右折においてエンドツーエンドの自動運転を対象とした新しい相互注意機構を用いた模倣学習手法を提案する。この手法により、自車両の状態と環境認識の関係性を効果的に捉えることができる。そして、オープンソースの自動運転シミュレーター AWSIM 上で評価した結果、提案手法が右折において従来の模倣学習ベースのベースライン手法を上回る性能を示し、複雑な方策を効率的に模倣する能力を実証した。

Contents

Contents v

List of Figures vii

List of Tables ix

List of Acronyms ix

Chapter 1 Introduction 1
1.1 Background . . . 1
1.2 Proposal and Contributions . . . 6
1.3 Organization . . . 7

Chapter 2 Related Work 8
2.1 End-to-end Autonomous Driving . . . 8
2.2 Autonomous Driving in Unprotected Turns . . . 12
2.3 Attention Learning in Autonomous Driving . . . 12

Chapter 3 Collection of Autoware Expert Dataset 15
3.1 Unprotected Turn Scenario Dataset . . . 15
3.2 Implementation of Dataset Collection Module . . . 18

Chapter 4 Proposed Cross-Attention Enhanced End-to-end Autonomous Driving 20
4.1 Problem Formation . . . 20
4.2 Vision-based Spatial Feature Extraction . . . 20
4.3 Image-State Cross-Fusion . . . 22
4.4 Imitation Learning . . . 23

Chapter 5 Experiment in AWSIM Simulator 25
5.1 Experiment Setup . . . 25
5.2 Implementation of End-to-end Autonomous Driving Agent . . . 28
5.3 Simulation Result . . . 29

Chapter 6 Conclusion 34
6.1 Summary . . . 34
6.2 Future Work . . . 34

Publications 36

References 37

List of Figures

1.1 Official Figure of SAE 6 Levels of Autonomous Driving[4] 2
1.2 A typical workflow of autonomous driving systems. Typical sensors include cameras, LiDARs, radars, IMUs, and GPS sensors. 3
1.3 Modular architecture (above) comparing to end-to-end architecture (below) in autonomous driving. 4
1.4 A typical unprotected turn scenario, featuring 2 difficulties: quick decision making required, and traffic negotiation issue. 6

2.1 The structure of Expert-Prior-RL. 9
2.2 The structure of CILRS. 10
2.3 The structure of trajectory-guided control prediction (TCP). 11
2.4 The structure of ROACH. 12
2.5 The traffic negotiation issue in unprotected turn scenarios. In the negotiation zone, the 2 vehicles must negotiate which vehicle gets the right of way, resulting in the 2 types of traffic situations. 13
2.6 Different optimal decisions under different vehicle states in unprotected turns. 13

3.1 The overview of the proposed dataset collection procedure. In phase II, a human expert intervenes and makes high-level decisions for Autoware, which generates the low-level control commands. 15
3.2 An overview of the intersections used in the dataset. Train intersections are shown in blue boxes, while validation intersection is shown in orange box. 17
3.3 The data flow chart of the dataset collection module. The relevant functions used are listed in the blocks of each of the modules. 17

4.1 An overview of our proposed cross-attention enhanced framework for imitative end-to-end autonomous driving, in unprotected turn scenarios. . . 21
4.2 Detailed structure of image patch encoder. 21
4.3 Detailed structure of state token encoder. 22
4.4 Detailed structure of cross-attention layer 23

5.1 The unprotected right turn scenario used for training and evaluation in AWSIM simulator. The white arrow depicts the desired route of the ego vehicle, while the red arrow depicts the route of NPC vehicles. 26
5.2 Visualization of the calculation process for waypoint_passed and distance_to_center. P is the position of the ego vehicle, while A and B stand for the last waypoint and current waypoint, respectively. 27
5.3 The structures and their purposes of the baselines. 28

5.4 The overview of the AD agent. The ROS topics used are listed in the middle of the figure. 29
5.5 Evaluation results of success rate (SR). 30
5.6 Evaluation result of average passing time (Time). 30
5.7 Speed, steering, and acceleration curves during a typical driving episode. The curves for both methods are smoothed by the same sliding window of size 25. 31
5.8 Attention map visualized when the ego vehicle is traveling at a speed of 3.78 m/s 32
5.9 Attention map visualized when the ego vehicle is traveling at a speed of 5.09 m/s 32

List of Tables

3.1 Specifications of the Unprotected Turn Scenario Dataset 16

4.1 Training Parameters 24

5.1 Simulation Experiment Parameters and Specifications 25
5.2 Computational Cost 33

List of Acronyms

AV Autonomous Vehicles

AD Autonomous Driving

AI Artificial Intelligence

ABS Antilock Brake System

EV Electric Vehicle

V2X Vehicle-to-everything Network

SAE Society of Autonomous Engineers

LiDAR Light Detection and Ranging

GPS Global Positioning System

RSU Road Side Unit

SLAM Simultaneous Localization and Mapping

IMU Inertial Measurement Unit

E2E-AD End-to-end Autonomous Driving

DNN Deep Neural Network

RNN Recurrent Neural Network

CNN Convolutional Neural Network

IL Imitation Learning

RL Reinforcement Learning

DDQN Double Deep-Q Network

TD3 Twin Delayed Deep Deterministic Policy Gradient

SAC Soft Actor-Critic

PID Proportional-Integral-Derivative controller

NPC Non-Player Character

MPC Model Predictive Control

RGB Red, Green, and Blue

ViT Vision Transformer

MHCA Multi-Head Cross Attention

MLP Multi-Layer Perceptron

BC Behavior Cloning

MSE Mean Squared Error

ROS2 Robot Operating System 2

Chapter 1

Introduction

1.1 Background

The evolution of vehicles has been a cornerstone of technological progress, profoundly shaping modern societies and economies. From a greater aspect, it played a significant role in shaping modern cities and in facilitating globalization[1]. From personal aspects, it granted individuals greater freedom of movement and allowed access to a wider range of opportunities[2].

From the invention of the internal combustion engine in the late 19th century to the mass production of automobiles pioneered by Ford in the early 20th century, vehicles have undergone continuous advancements to meet the growing demands for efficiency, safety, and convenience. The postwar era witnessed a surge in innovation, including the introduction of automatic transmissions, advanced braking systems, and aerodynamically optimized designs. By the late 20th century, electronic systems such as antilock braking systems (ABS), traction control, and engine management systems became integral to vehicle functionality.

In the 21st century, vehicles have embraced digital transformation, with developments in connectivity, electrification, and autonomous driving technologies redefining the transportation landscape. Emerging trends such as electric vehicles (EVs), vehicle-to-everything (V2X) communication, and advances in sensor technologies have positioned vehicles as intelligent and adaptive systems capable of addressing contemporary challenges such as environmental sustainability, urban congestion, and road safety. These developments not only mark a technological revolution but also signal a paradigm shift towards mobility solutions that are autonomous, connected, and sustainable. This thesis situates itself within this context, focusing on the cutting-edge domain of autonomous driving, where vehicles are envisioned as self-reliant entities capable of navigating complex environments with minimal human intervention.

1.1.1 Autonomous Driving Systems

Autonomous driving, often referred to as self-driving or driverless technology, represents a transformative milestone in the evolution of modern transportation. In essence, autonomous vehicles (AVs) are designed to navigate and operate without direct human intervention by leveraging advances in sensors, machine learning, robotics, and computational systems. The concept of autonomous vehicles has existed since the mid-20th century, and the past two decades have witnessed a surge in progress due to breakthroughs in artificial intelligence (AI) and hardware capabilities. Today, autonomous driving systems are at the forefront of efforts to revolutionize mobility, promising safer, more efficient, and accessible transportation solutions.

As shown in Fig. 1.1, the development of autonomous vehicles has been classified into six levels of driving automation by the Society of Automotive Engineers (SAE)[3]. At the lowest level, vehicles do not offer automation, with the human driver fully responsible for all tasks. As automation increases, systems begin to assist with specific functions such as steering and acceleration (Level 1) or even handle both simultaneously under driver supervision (Level 2). Conditional automation (Level 3) allows the vehicle to perform most tasks independently, although human intervention is required when requested. High automation (Level 4) can handle all driving tasks in specific scenarios, such as geofenced areas, while full automation (Level 5) allows vehicles to operate in any environment without human input. Although many companies aim to achieve Level 5 autonomy, most commercially available systems today operate at Levels 2 or 3.

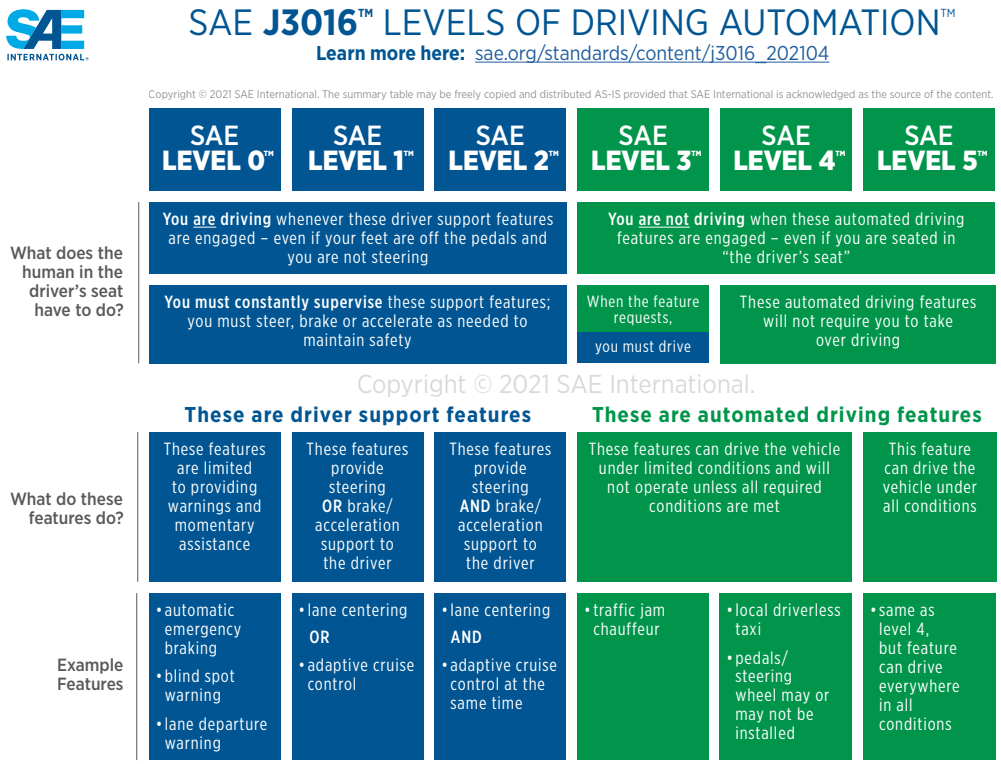


Fig. 1.1: Official Figure of SAE 6 Levels of Autonomous Driving[4]

To achieve these levels of autonomy, vehicles rely on a combination of subsystems that include perception, localization, prediction, planning, and control. Perception systems utilize sensors such as cameras, LiDAR, radar, and ultrasonic devices to identify objects, traffic signals, and road boundaries. Localization algorithms determine the vehicle's position on a map, integrating data from GPS, inertial measurement units (IMUs), and techniques like Simultaneous Localization and Mapping (SLAM). Prediction modules anticipate the behavior of surrounding agents, such as vehicles and pedestrians, while planning systems decide optimal routes and maneuvers, these modules are also commonly recognized as one planning module. Finally, control systems execute these plans by managing the vehicle's throttle, steering, and braking. Fig. 1.2 shows a typical workflow of autonomous driving, with its key components and their relevant sensors, respectively.

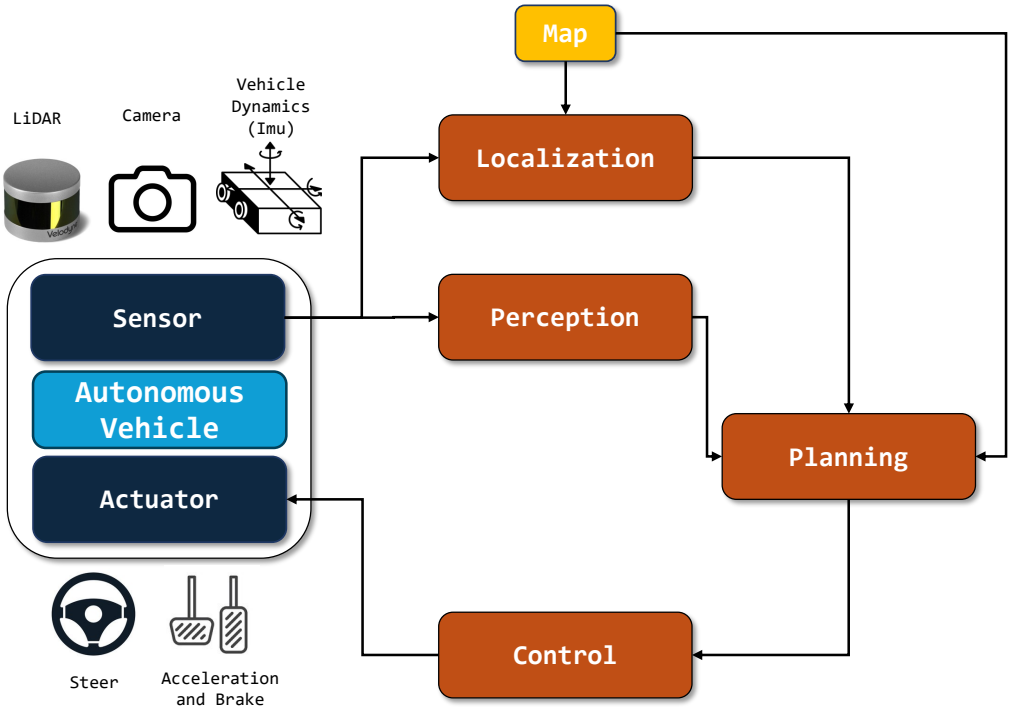


Fig. 1.2: A typical workflow of autonomous driving systems. Typical sensors include cameras, LiDARs, radars, IMUs, and GPS sensors.

1.1.2 The Evolution of Autonomous Driving Architectures

Modular Structure

Autonomous driving systems have traditionally relied on modular structures, which decompose the driving task into distinct stages, such as perception, localization, prediction & planning, and control, depicted in the upper part of Fig. 1.3. This approach mirrors the way humans drive, breaking down complex tasks into manageable sub-problems, each solved by a specialized algorithm or model.

The modular pipeline is inherently hierarchical, with outputs from one module serving as inputs to the next. For example, sensor data processed in the perception module informs localization and prediction, which then influence planning and control. While this separation of concerns provides transparency and modularity, it introduces several challenges. First, errors from earlier modules can propagate and compound through the pipeline, leading to suboptimal decisions or system failures. Second, optimizing individual modules independently does not guarantee global system optimization, as each component lacks full context about the driving task. Lastly, integrating multiple handcrafted modules often requires significant manual tuning and domain expertise, making these systems less adaptable to novel environments.

For decades, modular systems dominated the field due to their interpretability, ease of debugging, and alignment with established engineering practices. However, while this approach leverages individual algorithms to process data and generate control outputs, it suffers from certain drawbacks, notably its vulnerability to error accumulation. What's more, as the complexity of autonomous driving scenarios has grown, the limitations of modular architectures have become increasingly evident, prompting researchers to explore

alternative solutions, including end-to-end (E2E) architectures.

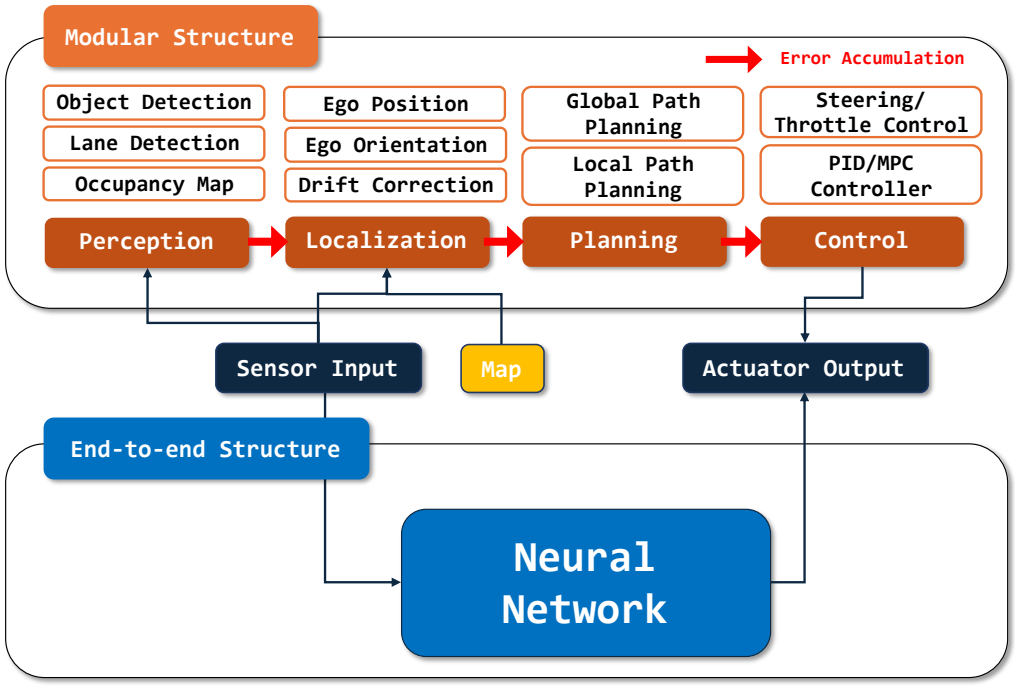


Fig. 1.3: Modular architecture (above) comparing to end-to-end architecture (below) in autonomous driving.

End-to-end Autonomous Driving

In response to the limitations of traditional modular structures, an alternative paradigm known as end-to-end driving has emerged[5]. As the lower part of Fig. 1.3 shows, this approach seeks to address the shortcomings of the modular architecture by streamlining the system, thereby enhancing efficiency and robustness. End-to-end driving entails directly mapping sensory input to control outputs, circumventing the need for intermediate processing steps. This shift is largely driven by advances in AI, computational power, and the availability of large-scale driving datasets. E2E systems bypass the need for intermediate modules by treating driving as a single learning problem, allowing the model to learn representations and decision-making strategies holistically.

The rise of convolutional neural networks (CNNs)[6] and recurrent neural networks (RNNs)[7] in the 2010s played a pivotal role in enabling E2E learning. CNNs, with their ability to process high-dimensional visual data, became the backbone of perception-driven E2E systems, while RNNs facilitated sequential decision-making by incorporating temporal dependencies. More recently, transformer-based architectures have further expanded the capabilities of E2E systems, enabling multi-modal fusion and attention-based feature prioritization. These architectures excel at capturing complex relationships between inputs, such as the spatial and temporal dynamics critical for driving in complex environments.

E2E architectures offer several advantages over modular systems. By learning directly from data, they can optimize the entire driving process rather than individual components, potentially achieving higher overall performance. The end-to-end nature also reduces the need for extensive manual engineering and integration, making these systems more scalable and adaptable to diverse driving conditions. Additionally, the use of deep learning allows

E2E models to generalize better to novel scenarios by learning latent representations that capture underlying patterns in the data.

Despite their promise, E2E systems are not without challenges. One significant drawback is the lack of interpretability. Unlike modular systems, where individual modules provide clear outputs that can be analyzed and debugged, E2E models operate as black boxes, making it difficult to understand or rectify errors. Furthermore, training E2E systems requires large amounts of high-quality data that accurately represent the complexities of real-world driving, including rare or dangerous scenarios. While at the same time, frequent situations, such as normal scenarios, contribute more to the optimization of the end-to-end autonomous driving models, making it harder to learn a good policy for rare scenarios. Thus, ensuring reliability and safety in critical situations, such as unprotected intersections or adverse weather conditions, remains a significant hurdle[8].

The transition from modular to E2E architectures reflects a broader paradigm shift in autonomous driving research, driven by the desire to overcome the limitations of traditional approaches. While modular systems remain valuable for their interpretability and domain knowledge, E2E architectures offer a compelling alternative by leveraging data-driven learning to address the challenges of complexity, error propagation, and scalability. This thesis builds upon this transition by focusing on cross-attention mechanisms within an E2E framework, aiming to enhance the model's ability to handle dynamic and uncertain scenarios, such as unprotected turns, in a more interpretable way. By addressing key limitations of E2E systems, this research contributes to bridging the gap between theory and practice in the pursuit of robust autonomous driving solutions.

1.1.3 Autonomous Driving in Unprotected Turns

Performing an unprotected turn is widely recognized as one of the most challenging maneuvers for both autonomous and human drivers[9]. As Fig. 1.4 shows, in this maneuver, the ego vehicle must execute a left (or right in certain countries like Japan) turn without the assistance of a green arrow signal, indicating that opposing traffic is permitted to traverse the intersection simultaneously. Consequently, drivers must attentively observe the flow of oncoming traffic and patiently await a sufficient gap to navigate through the intersection safely. Such decision-making demands a comprehensive grasp of the prevailing environmental conditions.

The unprotected turn scenario is inherently dangerous due to several critical factors. First, it carries a high risk of collision because vehicles making unprotected turns must cross multiple lanes of oncoming traffic, often at intersections with complex traffic patterns. This requires precise judgment of speed and distance, leaving little margin for error. Second, the scenario presents significant vehicle negotiation problems, as drivers must anticipate and respond to the actions of other vehicles, often with limited or conflicting signals. Finally, the vehicle face-off issue[10] exacerbates the danger, where two or more vehicles arrive at an intersection simultaneously, creating confusion and hesitation as drivers attempt to determine who has the right of way. These challenges collectively heighten the likelihood of accidents, making unprotected turns one of the most hazardous maneuvers in traffic systems.

Given the complexity of near-accident scenarios like unprotected turns, which require precise decision-making in dynamic environments, researchers have increasingly turned to imitation learning (IL) as a promising end-to-end learning-based approach. However, there are challenges that remain unsolved in this domain. A typical one is: While much of the existing literature has primarily focused on sensor fusion, there has been a tendency to overlook the significance of incorporating vehicle states, particularly in complex scenarios such as unprotected turns.

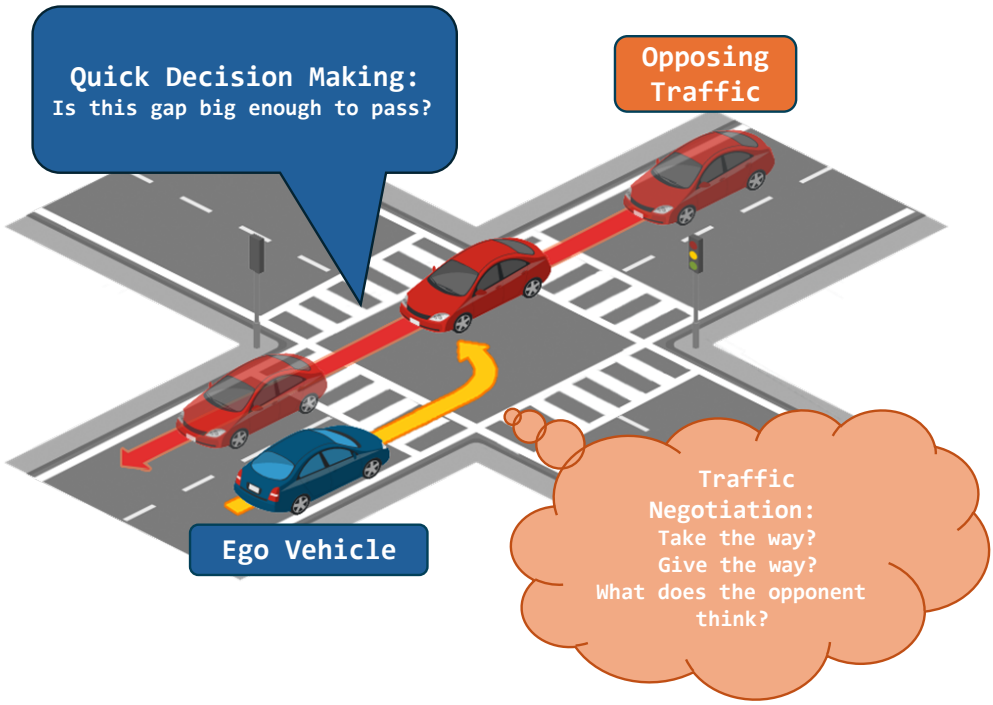


Fig. 1.4: A typical unprotected turn scenario, featuring 2 difficulties: quick decision making required, and traffic negotiation issue.

To be more specific, in contrast to human drivers, who naturally adjust their attention to the environment based on the vehicle's different states, autonomous vehicles encounter challenges stemming from inefficient or inappropriate structures for integrating state information with observations. For instance, specific studies [11] [12] [13] [14] have only attempted to integrate vehicle states into their models by directly concatenating or adding them to observation embeddings, which is proved not properly working in our experiments.

1.2 Proposal and Contributions

As discussed in the above paragraphs, performing an unprotected turn maneuver is difficult for autonomous vehicles. To achieve better performance, we need a more integrated way to understand the surrounding environment, as well as predict the intention of other road users. To solve this problem, in this thesis, we introduce a novel approach to imitation learning that integrates cross-attention mechanisms to leverage the importance of vehicle states in decision-making during unprotected turns. Our method enables the current vehicle states to guide the model's attention toward different regions of the front camera image observation through a cross-attention transformer. This architectural enhancement allows the model to dynamically focus on pertinent parts of the image input corresponding to specific vehicle states. Consequently, irrelevant portions of the observation are filtered out, resulting in a more generalized observation embedding conducive to imitation learning.

1.3 Organization

The remainder of this thesis is organized as follows: Section II presents a comprehensive review of related works pertaining to typical end-to-end autonomous driving methods, autonomous driving methods in unprotected turns, and attention learning techniques for end-to-end autonomous driving. Section III explains how we collect the dataset by featuring both accurate control and human-like decision-making, to train our model. Subsequently, Section IV elaborates on the proposed cross-attention enhanced imitation learning approach in detail. Section V presents the experiment setting and simulation results of our model, conducted using the AWSIM[15] simulator, followed by Section VI, concluding the thesis while adding discussions on its limitations and outlining potential future directions for research.

Chapter 2

Related Work

Building upon the challenges and advancements outlined in the introduction, this chapter delves into the existing body of research that has shaped end-to-end driving architectures. Specifically, we explore the evolution of end-to-end models, the incorporation of multi-modal data, and innovative techniques like attention mechanisms that enhance decision-making in dynamic and uncertain scenarios, such as unprotected intersections.

2.1 End-to-end Autonomous Driving

Traditionally, the process of training an end-to-end autonomous driving model is divided into 2 types: reinforcement learning methods and imitation learning methods. Reinforcement learning focuses on allowing the model to explore and interact with a simulated environment, learning optimal driving policies through trial and error guided by reward functions. In contrast, imitation learning relies on supervised learning principles, where the model learns to mimic the behavior of expert drivers by training on labeled datasets consisting of sensor inputs and corresponding control outputs. In this section, we discuss how these two training paradigms have been applied in the context of end-to-end autonomous driving, highlighting their respective strengths, limitations, and the advancements made to overcome challenges in each approach.

2.1.1 Reinforcement Learning Approaches

[16] proposes a hybrid approach that combines model-free deep reinforcement learning (DRL) with model-based path planning to address challenges in urban automated driving. Traditional model-based methods provide safety and interpretability but struggle with dynamic and unmodeled environments, while end-to-end DRL-based approaches excel in learning from interactions but lack hard-coded safety constraints and path consistency. This work integrates a model-based path planner, such as A*, into the state space of a DRL agent, using the distance to the nearest waypoint as a key observation. The agent is trained with a reward function designed to encourage waypoint adherence while avoiding collisions.

In contrast, [17] leverages model-free DRL to develop a driving policy without expert data or predefined models. The authors propose a novel framework that reduces the problem's complexity by using a bird-view input representation combined with visual encoding to extract low-dimensional latent states, enabling RL algorithms to handle high-dimensional observations effectively. The study implements state-of-the-art model-free RL methods, including Double Deep-Q Network (DDQN), Twin Delayed Deep Deterministic Policy Gradient (TD3), and Soft Actor-Critic (SAC), with tailored enhancements such as modified exploration strategies, reward shaping, and frame skipping.

To address the challenges of interpretability and complexity, [18] proposes a novel approach in end-to-end autonomous driving systems for urban scenarios. Unlike traditional modular frameworks, the method integrates perception, decision-making, and control in an end-to-end manner while ensuring interpretability through a latent environment model. By leveraging probabilistic graphical models and maximum entropy reinforcement learning, the proposed framework compresses high-dimensional sensor inputs, such as RGB camera images and LiDAR data, into a low-dimensional latent space. This latent space is decoded into semantic bird-view masks, providing a clear explanation of how the policy understands the driving environment. The approach significantly reduces sample complexity while enabling robust policy learning for crowded urban scenarios.

Initial driving policies in RL can also be set by a pre-trained IL-based policy. [19] applies imitative expert priors for deep reinforcement learning, which significantly improves the learning speed as well as the driving performance. Fig. 2.1 describes the proposed architecture.

To conclude, RL-based reinforcement learning offers a promising approach to autonomous driving by enabling vehicles to learn optimal driving policies directly from interactions with the environment. It can improve both the performance and transparency of end-to-end autonomous driving, making them more robust for real-world deployment.

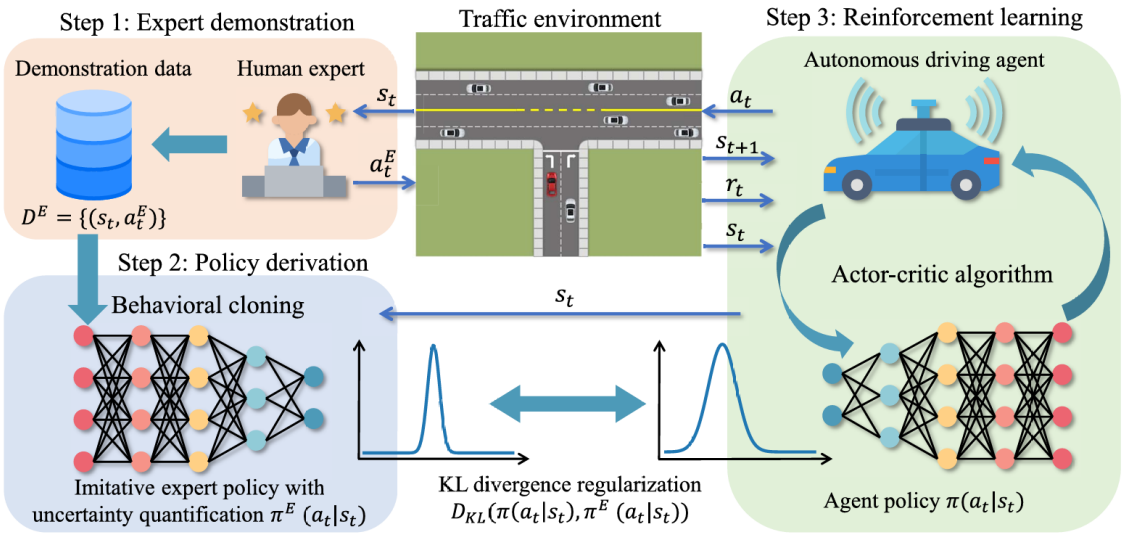


Fig. 2.1: The structure of Expert-Prior-RL.

2.1.2 Imitation Learning Approaches

Despite its promising potential, RL-based methods face several drawbacks compared to imitation-learning-based approaches. One major challenge is the high sample inefficiency of reinforcement learning, especially in complex environments like urban driving. RL algorithms often require a large number of interactions with the environment to converge to a good policy, which can be computationally expensive and time-consuming. In contrast, imitation learning (IL) involves learning from expert demonstrations to train systems to replicate expert behavior in diverse driving scenarios, allowing the model to quickly learn from high-quality, labeled data, significantly reducing the time and resources needed for training. Recent advancements in IL for end-to-end autonomous driving have shown significant improvements in both safety and efficiency.

[11] first proposed conditional imitation learning (CIL), where a driving policy is learned from demonstrations of human driving along with high-level commands (e.g., turn left, go straight), allowing the learned policy to be guided by commands at test time for autonomous urban driving. [12] and [20] followed this work and proposed CILRS and CIL++, in order to provide strong baselines on vision-based end-to-end autonomous driving with solely a camera image.

In CILRS, the model introduces resampling techniques to mitigate data sparsity in complex driving scenarios such as turns and intersections, ensuring a more balanced representation of driving behaviors during training. Additionally, CILRS incorporates a secondary task of speed prediction, enabling the network to jointly learn driving commands and vehicle speed, which enhances the model's capacity to understand vehicle dynamics and produce smoother control signals. As Fig. 2.2 shows, CILRS employs a CNN to process camera inputs and conditioned control commands (e.g., turn left, go straight). It outputs low-level control signals, including steering, throttle, and brake values. The integration of the speed prediction task acts as an auxiliary loss, guiding the network to better capture driving nuances and contextual cues in the environment. By addressing data imbalance and enhancing the understanding of vehicle states, CILRS represents a notable step forward in improving the robustness and generalizability of imitation learning-based driving systems.

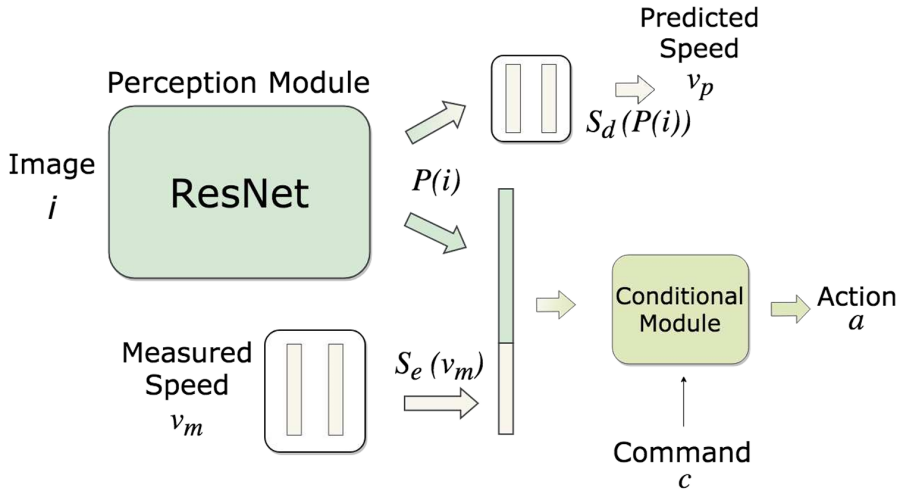


Fig. 2.2: The structure of CILRS.

For firmer design in the observation parts, [21] [22] [23] [24] [25] [26] emphasize multimodal sensor fusion to enhance the model's observation, while [13] focuses on the control parts by proposing a trajectory-guided control prediction (TCP) framework that combines trajectory planning and direct control prediction branches with a novel multi-step control branch guided by the trajectory branch. The trajectory branch serves as a guiding signal for low-level control decisions such as steering, throttle, and braking. This trajectory-conditioned approach ensures that the control outputs are closely aligned with the intended driving path, improving both the reliability and robustness of the driving policy.

The architecture of TCP is described in Fig. 2.3. First, a convolutional neural network (CNN) is used to process raw sensory inputs to extract visual features. These features are then combined with a representation of the planned trajectory, which is generated by a high-level planner. By fusing these inputs, the model can predict control commands that

respect the trajectory while adapting to real-time environmental dynamics. This integration enhances the policy's contextual awareness and allows it to make more consistent and interpretable decisions. TCP bridges the gap between traditional planning-based and end-to-end approaches, offering a balance of high performance and improved transparency in autonomous driving systems.

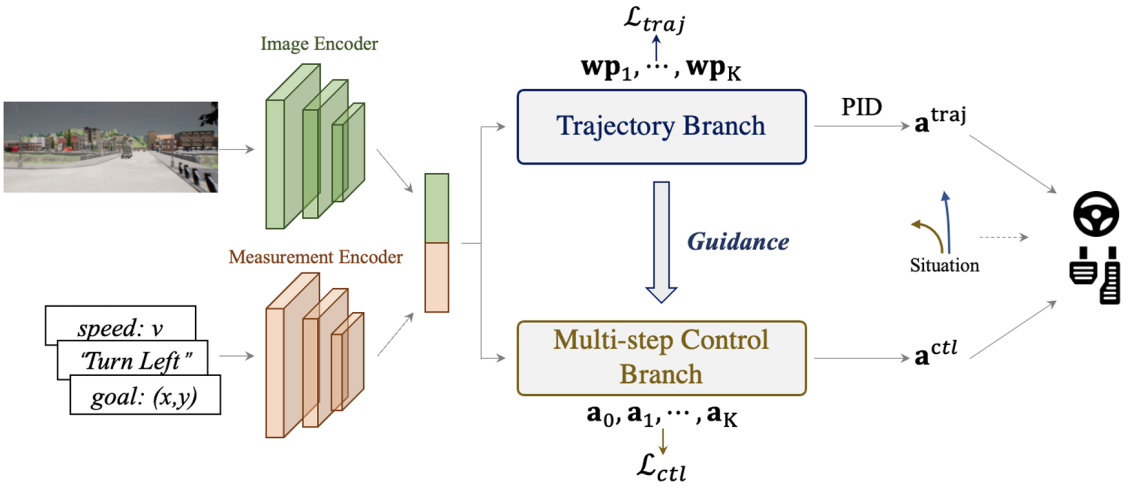


Fig. 2.3: The structure of trajectory-guided control prediction (TCP).

Another important part of imitation-learning-based approaches for end-to-end autonomous driving lies in the collection and design of the dataset. A good dataset should be filled with high-quality expert demonstrations featuring safe and efficient driving policies.

There are mainly 2 types of experts: Human experts and reinforcement-learning-based experts. For human experts, the dataset is traditionally collected by human drivers, either driving a real car or controlling the vehicle in simulated environments. For collecting expert demonstration from RL-based agents, [27], proposed by Zhang et al., presents a novel approach that combines RL and IL to achieve efficient and robust training for end-to-end autonomous driving. The core idea of this work is to use a reinforcement learning coach to generate high-quality driving data, which is then used to train an imitation learning model. This paradigm leverages the strengths of both methods: RL's ability to explore and optimize complex scenarios and IL's data efficiency and ease of training. As Fig. 2.4 shows, the framework operates in two stages. First, a reinforcement learning agent (the RL Coach) is trained in a simulated environment using RL algorithms. This coach learns an effective driving policy through interaction, generating diverse and high-quality trajectories, including challenging long-tail scenarios such as unprotected turns and dense traffic. Second, the imitation learning model is trained to replicate the behavior of the RL coach, mapping perception inputs, such as camera images, to low-level driving commands like steering, throttle, and braking.

There are still challenges in end-to-end autonomous driving using imitation learning, [12] explored the limitations of behavior cloning for end-to-end autonomous driving, demonstrating it still suffers from issues such as dataset biases, causal confusion, and high variance during training. These shortages come from insufficient structures for modeling the observation, resulting in a flawed understanding of the environment.

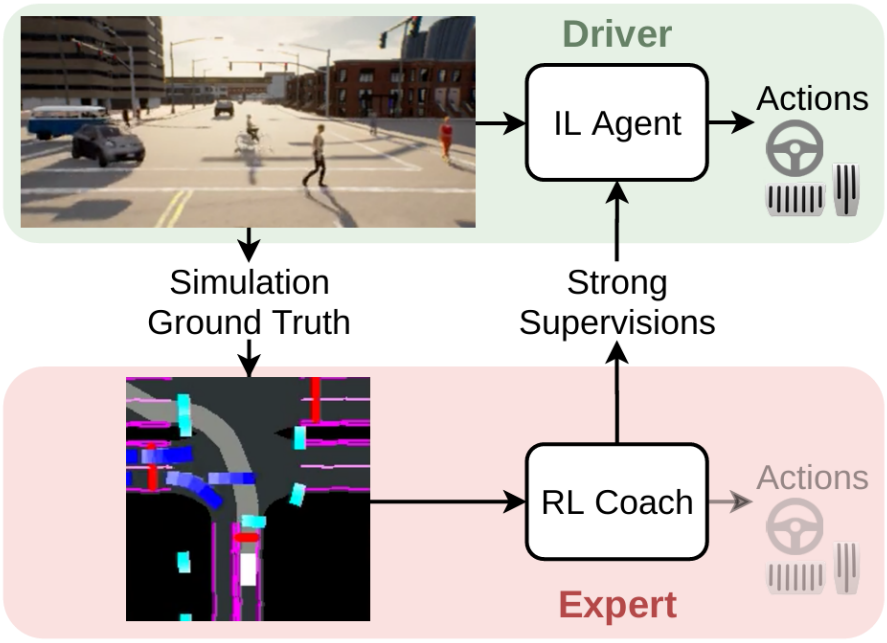


Fig. 2.4: The structure of ROACH.

2.2 Autonomous Driving in Unprotected Turns

Whether using reinforcement learning or imitation learning, it is still a challenge for end-to-end autonomous driving models to handle complex scenarios like unprotected turns, overtaking, and high-way merging. These scenarios are also called "near-accident scenarios". [28] concluded the reason why near-accident scenarios are hard for autonomous vehicles, like traffic negotiation issues, shown in Fig. 2.5. Then they stated the key point to solve the near-accident scenarios is to focus on phase transitions. [10] and [29] investigated different methods for solving the above issues.

However, an important feature in the near-accident scenarios, namely vehicle states, is still missed by the above works. After we have understood that we should focus on how to make decisions in near-accident scenarios, we should now find out what eventually affects our decision-making process. Vehicle states are exactly the key factors that influence optimal decisions because we could take different moves under different vehicle states, shown in Fig. 2.6.

2.3 Attention Learning in Autonomous Driving

The attention mechanism has also recently been adopted in autonomous driving, serving the role of explainable autonomous driving or enhancing the model's ability to understand the surrounding environment.

[30] proposed a self-attention-based model to map visual features from images collected by onboard cameras to guide potential driving actions with corresponding explanations. Furthermore, to achieve accurate control with attention, [31] proposes an end-to-end self-driving network with a sparse attention module that learns to attend to essential input regions for motion planning, improving safety and computational efficiency while enhanc-

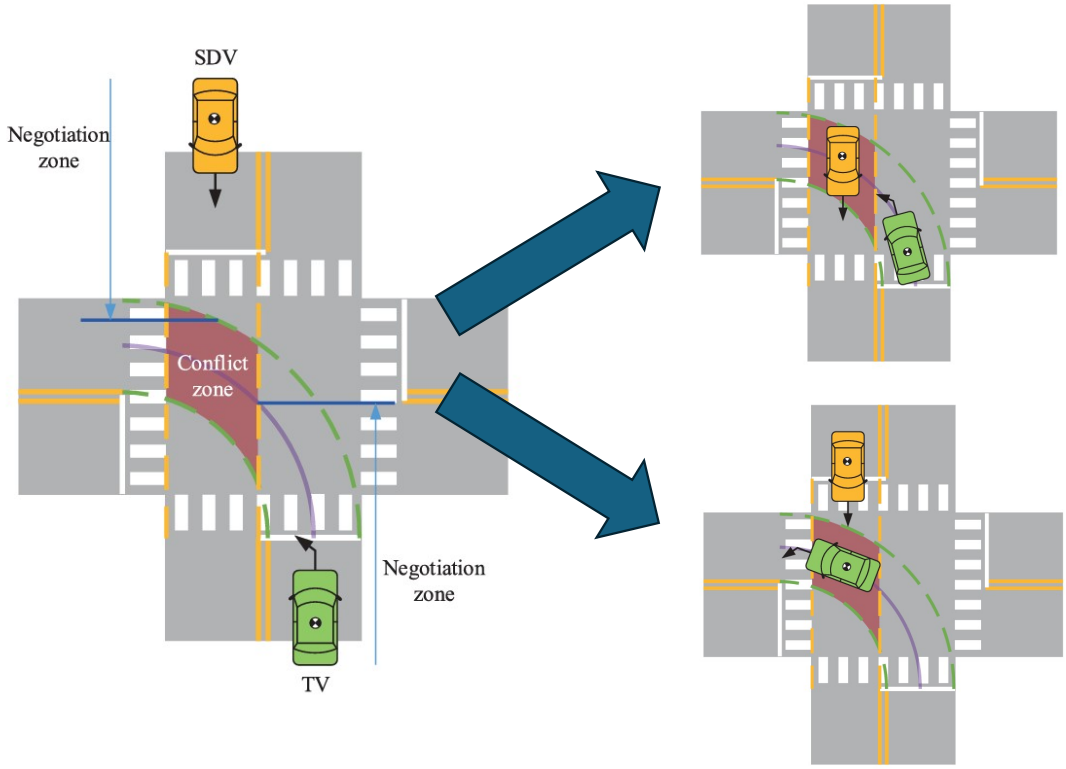


Fig. 2.5: The traffic negotiation issue in unprotected turn scenarios. In the negotiation zone, the 2 vehicles must negotiate which vehicle gets the right of way, resulting in the 2 types of traffic situations.

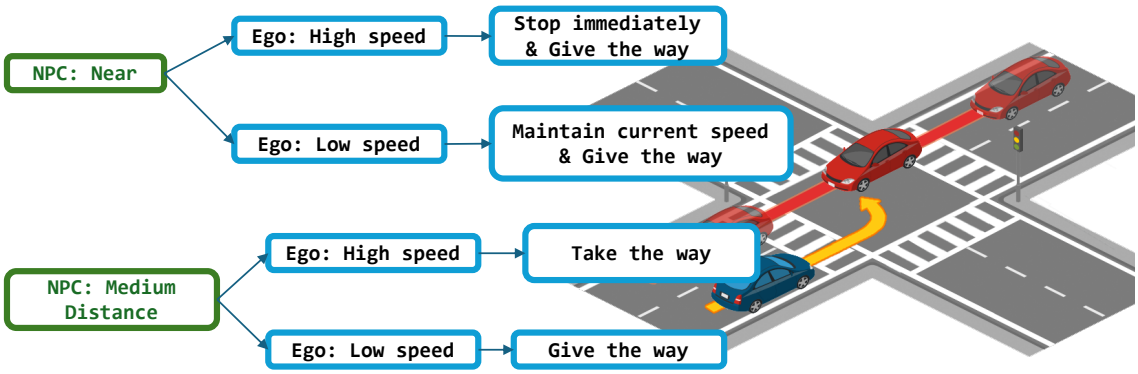


Fig. 2.6: Different optimal decisions under different vehicle states in unprotected turns.

ing interpretability.

Naturally, adapting the attention mechanism for end-to-end driving models can improve driving safety because it provides enhanced focus on critical information. [32] equipped the traditional imitation learning method with an attention model and achieved better driving performance. [33] proposed an RL-based framework with the attention mechanism for driving in unsignalized intersections, enabling the policy network to learn spatially and temporally essential features within the sensory input and to focus more attention on a few relevant surrounding vehicles at intersections. While they solely focus on leveraging

the attention, the inner motivation for guiding the attention remains unrevealed. [34] designed an attention loss for guiding the model's attention by pre-computed data, on-board sensors or a pre-trained network, which features that attention guidance has been taken into account for autonomous driving as other Computer Vision tasks did, but the importance of vehicle states in attention guiding is still not discovered.

Chapter 3

Collection of Autoware Expert Dataset

To advance the development of end-to-end autonomous driving systems, comprehensive and diverse datasets are essential. In this chapter, we introduce the construction of our unprotected turn dataset, collected in the AWSIM simulator while leveraging its realistic simulation environment and compatibility with Autoware[35], an open-source autonomous driving platform. The dataset specifically targets the unprotected turn scenarios, capturing the complexities and challenges of such maneuvers. Especially in the proposed dataset, we let a human expert intervene and supervise the decision-making process of Autoware, featuring both accurate control and human-like decision-making, which solved the nature that the rule-based method is hard to deal with complex scenarios.

3.1 Unprotected Turn Scenario Dataset

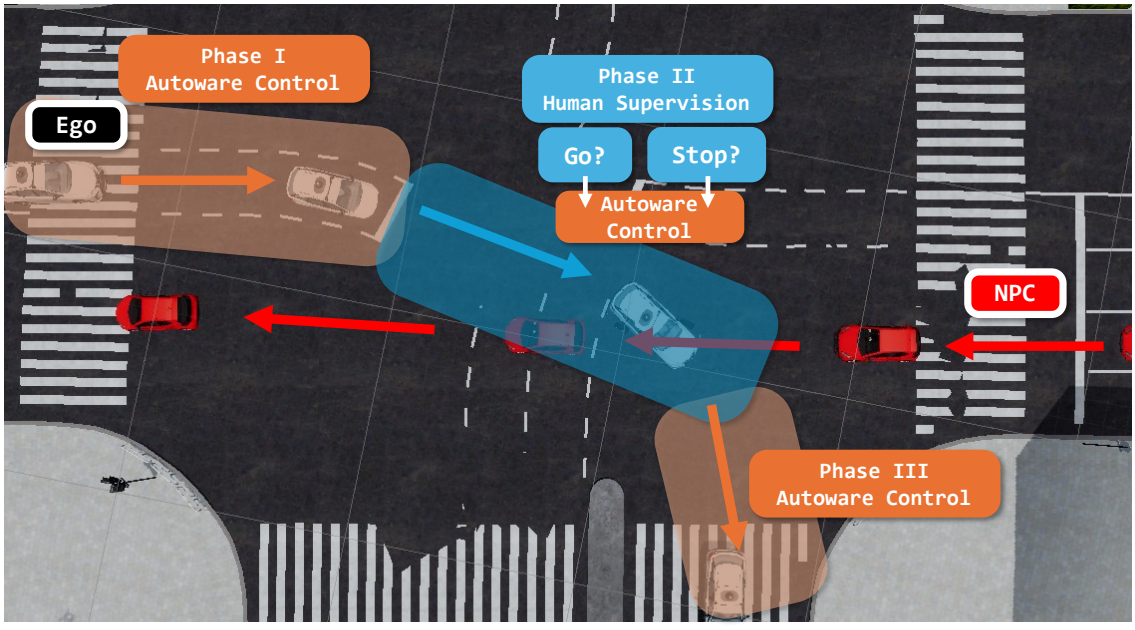


Fig. 3.1: The overview of the proposed dataset collection procedure. In phase II, a human expert intervenes and makes high-level decisions for Autoware, which generates the low-level control commands.

The dataset focuses specifically on scenarios where the ego vehicle must take right turns at intersections without dedicated traffic signals or clear priority. These scenarios are

particularly challenging due to the need for precise decision-making and interaction with oncoming traffic. To compile a thorough expert dataset for unprotected turn scenarios, it is essential to encapsulate a diverse range of expert driving behaviors.

Although Autoware cooperates with AWSIM Simulator well and provides robust planning modules for vehicle control, it is challenging to perform perfect unprotected turns by solely rule-based controllers. This is because rule-based planners are fragile to frequently changing traffic situations, making it hard to react to the changes in the environment and take the optimal next move again. The new CARLA Leaderboard 2.0 also stated this issue recently[36], and as a solution, they provide fully human expert demonstrations for the new complex scenarios. Inspired by this solution, we propose collecting the expert demonstrations from Autoware control under human expert supervision.

As Fig. 3.1 shows, during expert data collection, we utilize Autoware to drive in the designed intersection under human supervision, performing unprotected turn maneuvers. To be more specific, Autoware serves the role of driving in the desired route and velocity and recovering from any deviation of the desired route or velocity. Within this process, a human expert is required to give supervision commands of "yield to opposing traffic" or "proceed immediately" at the right time. When Autoware receives such supervision, it is forced to change its driving behavior by re-generating the desired velocity based on the given high-level command. After that, Autoware adheres to the instruction of the human expert, while at the same time, it still calculates the control command by its rule-based planners. From this method, we can collect a robust expert demonstration dataset that features both human-like decision-making and accurate and smooth control.

In practice, we chose 4 intersections in the Nishishinjuku map provided by the AWSIM simulator for collecting the dataset. Fig. 3.2 shows the location of the 4 intersections and their 6 routes of the ego vehicle. We leverage intersection 1, 2, and 4 as the training intersection, while intersection 3 is the validating and testing intersection. Throughout a total number of 165 episodes of expert demonstrations, we capture RGB images of size 900x256 from the front-facing camera of the ego vehicle, along with its current states, including longitudinal velocity, steering angle, pose, and waypoints, serving as input data. The corresponding actions, encompassing AckermannControlCommand, are recorded as labels for behavior cloning. Table 3.1 shows the specifications of the proposed dataset.

Table 3.1: Specifications of the Unprotected Turn Scenario Dataset

Intersection	**Episodes**	**Frames**	**Total Frames**
	Training Set		
Train_1_1	40	3290	
Train_1_2	25	1854	
Train_2_1	15	1165	9749
Train_2_2	15	1090	
Train_4_1	30	2350	
	Validation Set		
Val_3_1	40	3055	3055

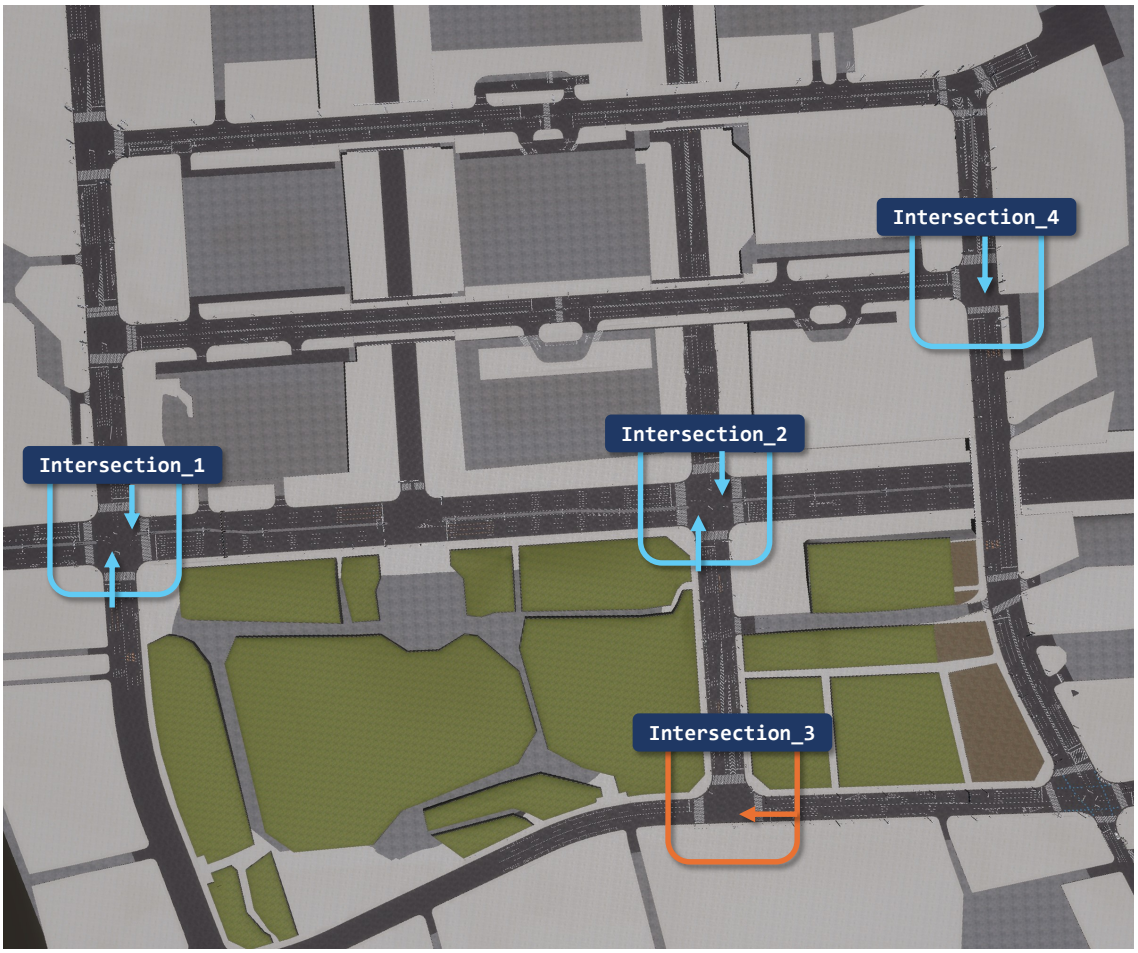


Fig. 3.2: An overview of the intersections used in the dataset. Train intersections are shown in blue boxes, while validation intersection is shown in orange box.

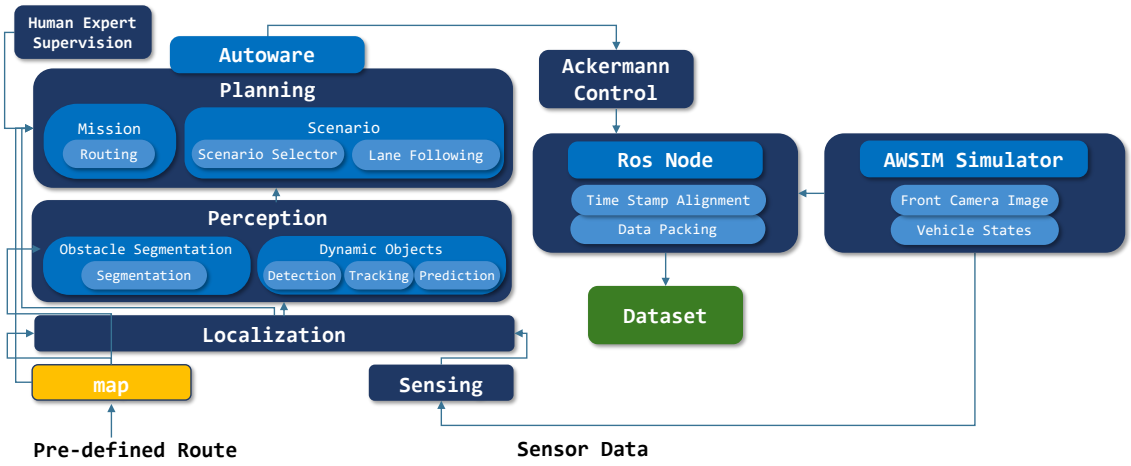


Fig. 3.3: The data flow chart of the dataset collection module. The relevant functions used are listed in the blocks of each of the modules.

3.2 Implementation of Dataset Collection Module

3.2.1 Vehicle ROS2 Interface

During the dataset collection, all sensor data are synchronized and timestamped, ensuring temporal alignment for model training and evaluation. This synchronization was achieved by leveraging the ROS2[37] communication framework used in the simulation, ensuring accurate correspondence between perception inputs and control outputs. Fig. 3.3 shows the details and data flow of the ROS2 vehicle interface implemented.

3.2.2 NPC Vehicle Control

The non-player character (NPC) vehicle control module implemented in the AWSIM simulator is actually a typical modular-structured autonomous driving system, the whole system is divided into 3 modules: Cognitive step, decision step, and control step. We modified some logic in the original NPC vehicle control module to enable more dynamic behaviors of the NPC vehicle. This modification is specially used in creating more challenging unprotected turn scenarios, as well as ensuring the reproducibility of these scenarios.

Cognitive Step

The control of NPC vehicles starts from the cognitive step. In this step, the system observes the surrounding environment of the specific NPC vehicle by using the groundtruth in the simulated environment. To be more specific, the system checks ground existence and obstacles by ray casting. It also checks the route information and traffic light states.

As a result, the system returns the next waypoint to follow, the position of the presence of obstacles in front of the NPC vehicle, and the state of the traffic light. These results serve as the main reference for the decision and control steps.

Decision Step

Based on the results of the cognitive step, the decision step outputs a short-term target point and the decision to accelerate and decelerate. Originally, the decision to accelerate or decelerate is made based on 2 constants: $minFrontVehicleDistance$ and $minStopDistance$. If the distance to the nearest front obstacle is less than $minFrontVehicleDistance$, the system will decide to decelerate, while if it is less than $minStopDistance$, the system will call an emergency stop. If neither of these 2 rules is triggered, the system will let the vehicle accelerate until it meets the required velocity or maximum velocity that the current lane allows.

The problem is that it is not realistic to make the rule be so strict as a constant, so we applied a random range for each NPC vehicle to their $minFrontVehicleDistance$, to simulate different types of driving behaviors. We remained the $minStopDistance$ a constant, to ensure that no NPC vehicles collide with the ego vehicle unconsciously. In practice, the $minFrontVehicleDistance$ is set to a random range of 2 to 5 meters, different from each NPC vehicle, and the m$inStopDistance$ is set to 1.5 meters for every NPC vehicle.

The decision will then be modified by other rules, such as red light, right of way, and turning states. The final decision will be sent to the control step for controlling the NPC vehicle.

Control Step

Given the decision of acceleration or deceleration, the control step generates the linear speed, angular speed, as well as the position and rotation of the NPC vehicles, by the predefined parameters for each type of the decision (NORMAL, SLOW, SUDDEN_STOP, ABSOLUTE_STOP, STOP).

The same problem appears here as discussed in the decision step, because the Target-Speed and Acceleration for each NPC vehicle remain constant as well. This problem is especially severe in the unprotected turn scenarios, because a constant stop distance and acceleration/deceleration will result in an equal distance between every two of the opposing NPC vehicles, and all the vehicles are driving at the same speed. In this case, the ego vehicle will face nothing dynamic regarding the invariant environment, and it will be no use to find a proper gap enough to pass safely - because the gaps are all the same.

Hence, we similarly introduce random turbulence of $(0, 1)m/s^2$ into the *Acceleration* and *Deceleration* for each NPC vehicle and each different driving mode under different decisions. Note that we also use Time-to-Collision (TTC) to the front vehicle for calculating the minimum deceleration needed to avoid collision as a restriction. When applying the deceleration to the vehicles, we use the equation below:

$$|a_{min}| = \frac{v}{TTC} \tag{3.1}$$

Chapter 4

Proposed Cross-Attention Enhanced End-to-end Autonomous Driving

In the previous chapters, we discussed the challenge that remains in end-to-end autonomous driving in terms of the lack of integrating vehicle states into perception input. We also collected an expert dataset for training the end-to-end model in the AWSIM simulator. Based on this dataset, we want to improve the performance of end-to-end autonomous driving by dynamically allocating different attention to different regions of the front camera image input. In this chapter, we explain the detailed structure of the proposed cross-attention-based end-to-end autonomous driving model. The overview of our proposed method is depicted in Fig. 4.1.

4.1 Problem Formation

Given the current state s comprised of the front camera sensor image img, the speed v, steering angle δ, and target information ($waypoints$, $target_point$), the end-to-end model is tasked with directly outputting control signals $ackermannControlCommand$, composed by acceleration command acc and steer command $steer$ to be executed in the next timestamp. In adherence to this specification, the model is expected to execute an unprotected right turn maneuver (pertinent to traffic conditions in Japan) within a randomly generated intersection traffic scenario. Notably, the model must yield to opposing traffic when necessary and navigate the intersection safely by identifying and exploiting feasible gaps between the opposing NPC vehicles.

4.2 Vision-based Spatial Feature Extraction

4.2.1 **Positional Feature Extraction**

To harness the attention guided by vehicle states across different input image regions, we commence with a feature extraction process to decompose the original camera sensor image into smaller patches. To accomplish this, we employ a pre-trained vision transformer[38], specifically the ViT-B16-224 model. This model operates on RGB images of size 224x224 and breaks the image into 14x14 patches with a patch size of 16 (pixels). We first resize the input image from (900, 256) to (224, 224) in order not to lose any information during the positional feature extraction. The specific network structure of the positional feature encoder is depicted in Fig. 4.2.

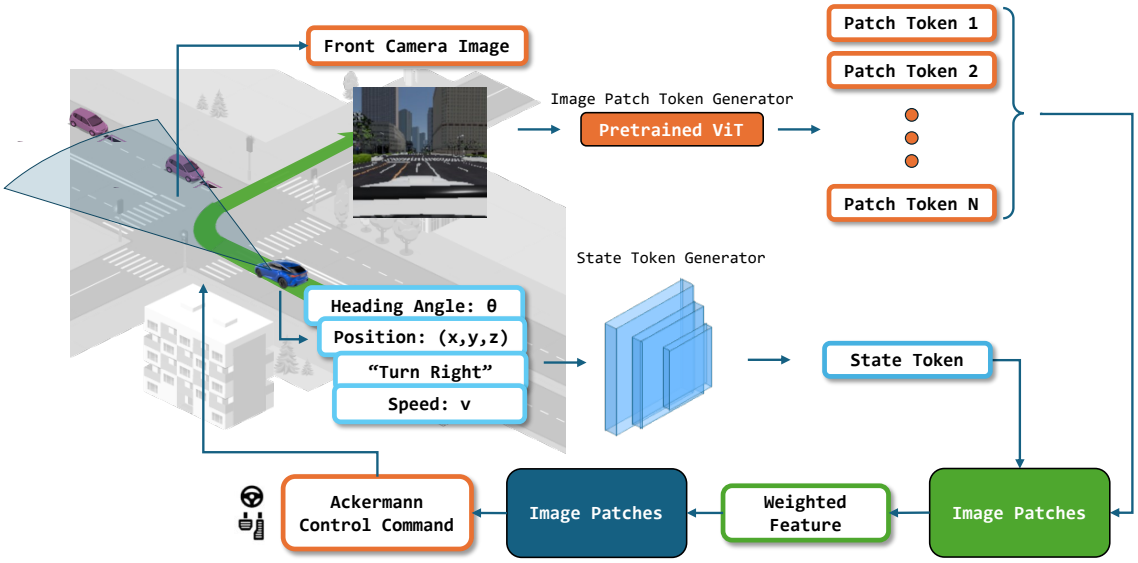


Fig. 4.1: An overview of our proposed cross-attention enhanced framework for imitative end-to-end autonomous driving, in unprotected turn scenarios.

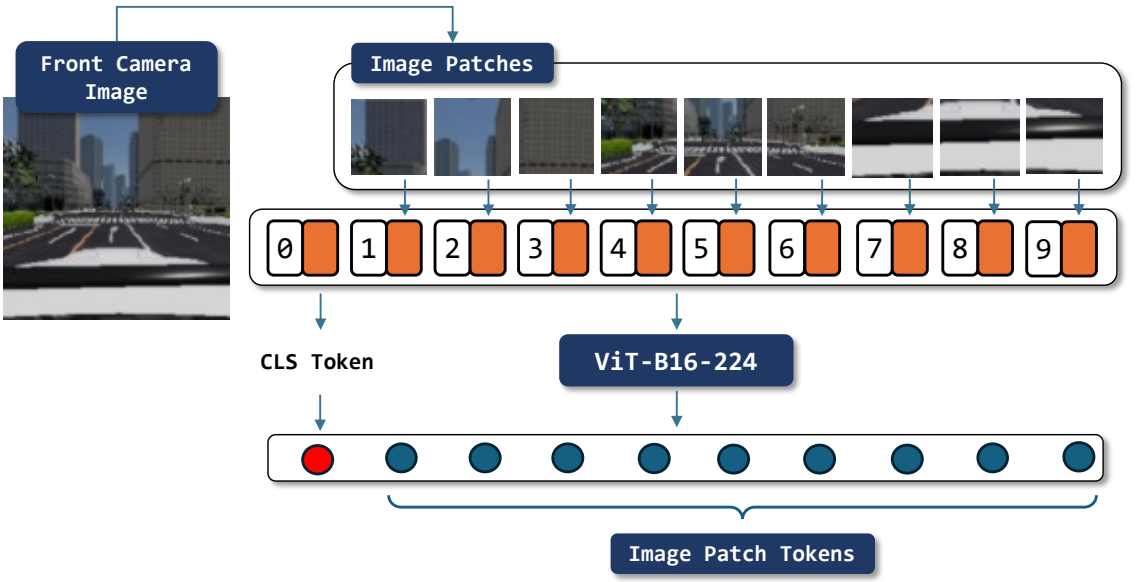


Fig. 4.2: Detailed structure of image patch encoder.

4.2.2 State Info Encoder

To ensure conformity with the constraint of maintaining consistent token embedding lengths within the cross-attention transformer, we employed a linear network featuring a sigmoid activation function. Before feeding data into the state info encoder, we first process the raw vehicle state data as Fig. 4.3 shows. Specifically, we shift the coordinate systems for the next waypoints and the target point from map_link (based on the map) to base_link (based on ego vehicle). After this shift, we injected the pose information

(position and orientation) and target information of the ego vehicle to the input of the state token encoder, in a more interpretable way.

The input vehicle states are first normalized into $[0, 1]$, then the state info encoder network projects the state information vectors of length 12 to a length of 768, thereby aligning them with the image tokens. Following this procedure, we obtain a token representing vehicle state information from each input state $s_i \in s$. The shape of the state info encoder is depicted in Fig. 4.3.

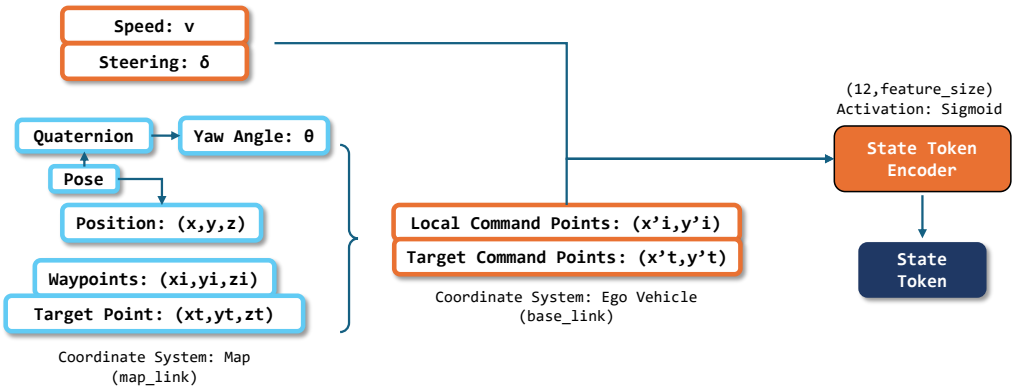


Fig. 4.3: Detailed structure of state token encoder.

4.3 Image-State Cross-Fusion

We obtain a set of image tokens and one state information token from the previous encoder networks from a given input state s_i. Then, we design a multi-head cross-attention transformer for image-state cross-fusion, using the cross-attention mechanism as introduced in [39]. By modifying the source of query, key, and value tokens, cross-attention enables the calculation of attention between tokens in different modalities. As the cross-attention transformer models the relationships between each image patch token and the state information token, we designate the state information token as the query token. In contrast, the image patch tokens are the key and value tokens. Then, an attention score matrix a can be calculated as follows:

$$a = softmax(\frac{QK^T}{\sqrt{d_k}}) \tag{4.1}$$

where QK^T represents a dot-product cross-attention operation, and d_k is the dimension of K and Q. A softmax function is applied to the calculated attention scores, yielding attention weights corresponding to each image patch in the original image. By multiplying the values from the image patches by their respective attention weights, we will form one cross-attention head as follows:

$$head = cross_attn(Q, K, V) = aV \tag{4.2}$$

To enhance the robustness and generalization capability of the model, we use 4 cross-attention heads and concatenate them. The concatenation process of multi-head cross-attention (MHCA) is as follows:

$$MHCA(X) = concat(head_1, ..., head_n) \tag{4.3}$$

Finally, a 2-layer feed-forward network is designed to process the result of MHCA to the final embedding of the given input image-state pair s_i.

The detailed structure of image-state cross-fusion is depicted in Fig. 4.4. From this structure, the model can prioritize regions with higher attention weights while attenuating the influence of irrelevant regions with lower attention weights.

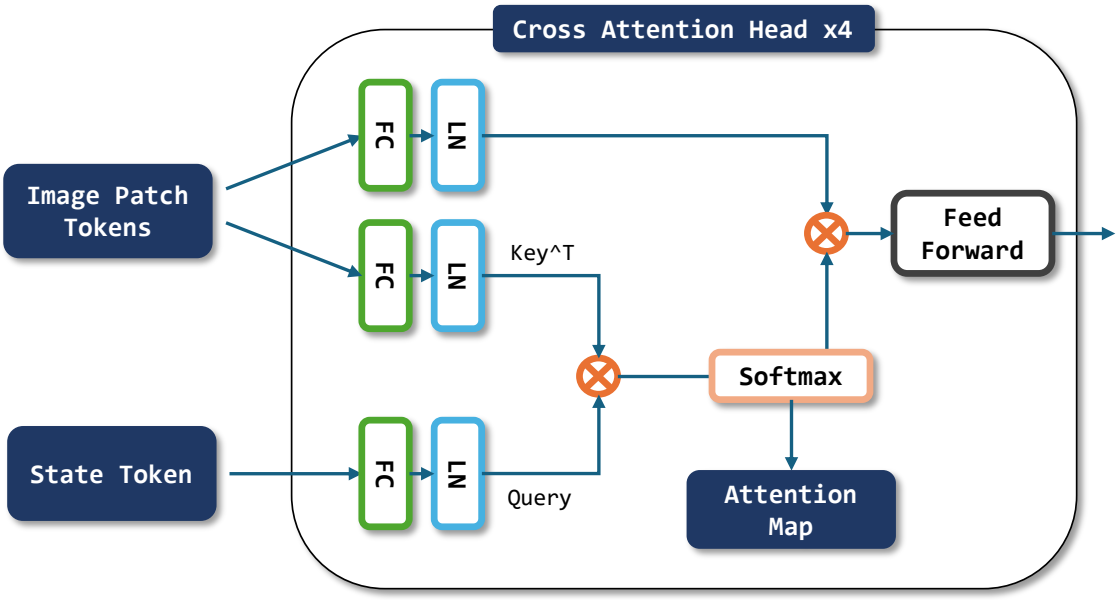


Fig. 4.4: Detailed structure of cross-attention layer

In contrast to conventional approaches that treat the entire image uniformly important and early fusion models that concatenate both image and state inputs, our cross-attention-based approach enables the model to dynamically allocate attention based on the relevance of image patches to the current vehicle state, thereby enhancing the model's discriminating power and decision-making capabilities.

4.4 Imitation Learning

The output of our proposed cross-attention layer yields a feature vector of length 768, serving as an embedding of the state input, weighted by the image-state cross-attention mechanism. To translate this feature vector into control signals, namely acceleration and steering, we design a multi-layer perceptron (MLP) followed by a Tanh active function that serves as the action head.

Subsequently, we employ a behavior cloning algorithm to facilitate the learning of an imitative driving policy from the expert demonstration dataset. The objective of BC is to solve the following optimization problem:

$$arg \ \min_{\theta} \ E_{(s,a^*)} \sim MSE(a^*(s), \pi_\theta(s)) \tag{4.4}$$

where $(s, a*)$ are state-action pairs from expert demonstration datasets, MSE represents mean squared error, calculated by:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.5}$$

where n is the total number of the data, y_i stands for the labels and $\hat{y}_i$ stands for the values predicted by the model.

We train our model separately using the proposed dataset, using the below parameters in Table 4.1. We adopt an early-stop mechanism for our method and every baseline, choosing the ones with the smallest validation losses as the final models.

Table 4.1: Training Parameters

Epoch	20
Batch Size	32
Learning Rate	0.001
Optimizer	Adam
Loss Function	MSE

Chapter 5

Experiment in AWSIM Simulator

In this chapter, we present the experimental setup and methodologies employed within the AWSIM simulator to evaluate and validate end-to-end autonomous driving systems. We first detail the design of the experimental scenarios, including both normal driving and unprotected turn situations. Additionally, we explore the various configuration settings and performance metrics used to assess the effectiveness and robustness of the system under diverse conditions. Then we evaluate the proposed cross-attention enhanced end-to-end autonomous driving system, discussing its strength compared to the baselines.

5.1 Experiment Setup

5.1.1 Scenario Design

We design the unprotected turn scenarios in Nishishinjuku-map for testing if the proposed method can drive safely and efficiently in the unprotected intersections. The software and hardware specifications for the experiment are shown in Table. 5.1.

Autoware	`awsim-stable` Branch
AWSIM-Labs	v1.3.1
ROS	ROS 2 Humble
Test Vehicle Model	Lexus RX450h 2015
Vehicle Dimensions	Width: 1895 mm, Length: 4890 mm
Operating System	Ubuntu 22.04 LTS
GPU	NVIDIA RTX 3090 24G
CPU	AMD Ryzen 5 5600X 6-Core Processor

Table 5.1: Simulation Experiment Parameters and Specifications

Unprotected Turn Scenario in Nishishinjuku-map

In our study, we designated a specific intersection located in Nishishinjuku, Tokyo, Japan, within the map provided by the AWSIM simulator. Although initially configured as a protected intersection, we manually turned off the arrow traffic lights to convert it into an unprotected intersection scenario. Within this intersection, the ego vehicle executes an unprotected right turn maneuver while ensuring safety by yielding to opposing traffic.

To emulate the complexity inherent in real-world traffic scenarios, we activated random traffic generation features within the AWSIM simulator. These traffic patterns are generated by the Route Traffic Simulator module, with vehicles spawned randomly. Each generated vehicle is controlled by the modified NPC vehicle control module within the

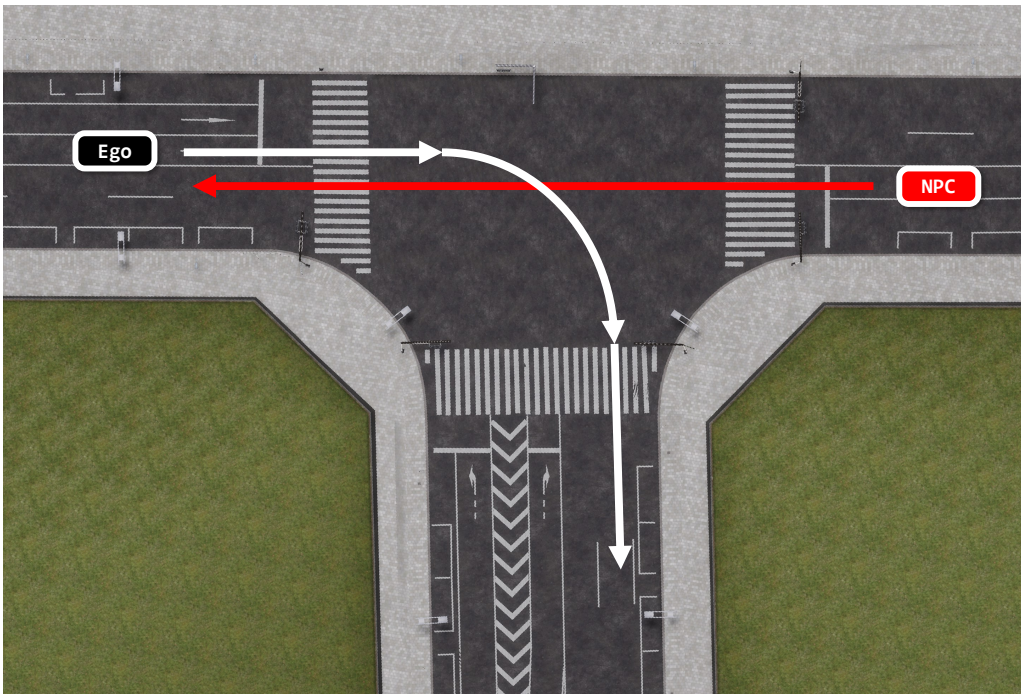


Fig. 5.1: The unprotected right turn scenario used for training and evaluation in AWSIM simulator. The white arrow depicts the desired route of the ego vehicle, while the red arrow depicts the route of NPC vehicles.

AWSIM simulator, following its predetermined route. We configured 1 type of NPC vehicle: opposing vehicles proceeding straight, as they are the core reason why unprotected turns are risky. The NPC vehicles are randomly spawned among 4 different fabrications: small car, taxi, red car, and small truck.

Furthermore, we introduced random sleep durations of 2 to 10 seconds for the ego vehicle to introduce variability in the timing of unprotected right-turn occurrences. This variation simulates different timings for initiating the unprotected right turns rather than encountering the same opposing vehicle repeatedly. The scenarios are reproducible using their respective random seeds, by setting the same seed, we can ensure that every method will face the same traffic scenarios when being evaluated. Fig. 5.1 provides an overview of the intersection scenario within the AWSIM simulator.

5.1.2 Evaluation Metrics

Safety Related Metrics

We employ success rate (SR) and average passing time (Time) as our primary evaluation metrics. We evaluate each model for 50 episodes, recording the number of successful episodes and their corresponding passing times for subsequent analysis. The success rate (SR) is calculated as follows:

$$SR = 1 - \frac{n_{collision} + n_{timeout} + n_{deviation}}{n_{episodes}} \tag{5.1}$$

where $n_{collision}$, $n_{timeout}$, and $n_{deviation}$ represent the number of failed episodes caused by collision, exceeding the 50-second time limit, and deviating more than 3.0 meters from the center of the desired route, respectively. $n_{episodes}$ is the number of total test

episodes. In practice, we implemented an episode management module to judge if one episode is successfully completed during the evaluation. The detailed logic of the episode management module follows the below algorithm. Fig. 5.2 visualizes the calculation procedure related to route information, to enable a better understanding of the episode management module.

Algorithm 1 Episode Management Algorithm

Require: Vehicle ROS node D, route information R containing waypoints
Ensure: Correctly collected evaluation result SR, $Time$

```
A ← R[CurrentWaypointNumber − 1].position
B ← R[CurrentWaypointNumber].position
P ← D.pose.position
AB ← B − A
AP ← P − A
if ProjectionLength ← (AP · AB)/(AB · AB) ≥ 1 then
    CurrentWaypointNumber ← CurrentWaypointNumber + 1
end if
if ‖AP − ProjectionLength · AB‖ > 3.0 or
        D.Collision = True or
        EpisodeTime > 50 then
    RESET()
else
    if CurrentWaypointNumber >= len(R) then
        SuccessCount ← SuccessCount + 1
        TotalTime ← TotalTime + EpisodeTime
        RESET()
    end if
end if
```

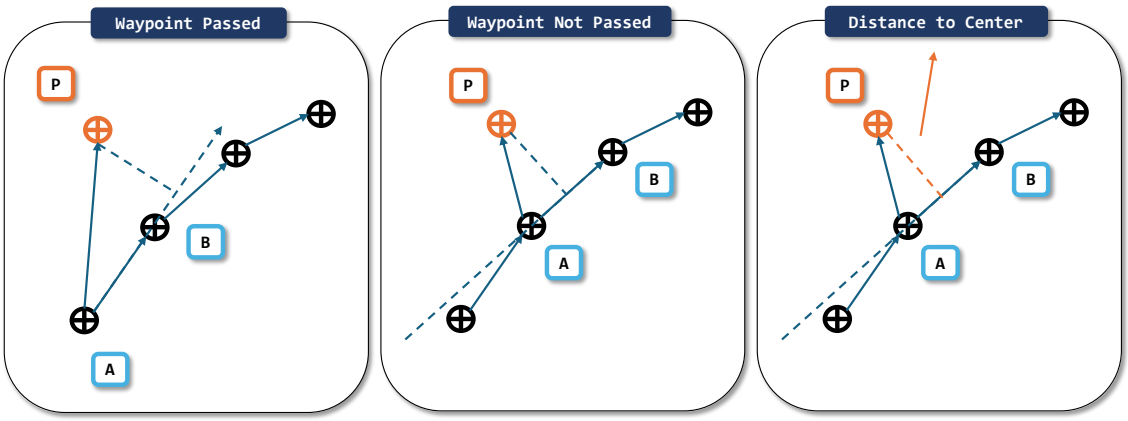


Fig. 5.2: Visualization of the calculation process for waypoint_passed and distance_to_center. P is the position of the ego vehicle, while A and B stand for the last waypoint and current waypoint, respectively.

Additionally, we visualize the ego vehicle's velocity, acceleration, steering curves, and the model's attention maps during the experiments to gain insights into its behaviors.

5.1.3 Baselines

To ensure a fair comparison, we reproduced CILRS[12] and changed its pre-trained image encoder from ResNet[40] to Vision Transformer, resulting in a model called CIL-ViT, as our baseline. To assess the significance of incorporating state embedding within the model, we introduce a baseline termed "ViT-Only". This baseline is derived from CIL-ViT by excluding the state information, resulting in only the front camera image being utilized as input. With the above 2 CIL-based baselines featuring the same vision backbone, we can clearly compare the performance of "without vehicle states", "concatenate the features", and "proposed cross-attention integration".

We also evaluated Autoware's planning and control module as the rule-based controller baseline to enable the discussion of why purely rule-based methods do not perform well in complex scenarios. Fig. 5.3 shows the structure of the 3 baselines and their corresponding purpose of comparing to the proposed method.

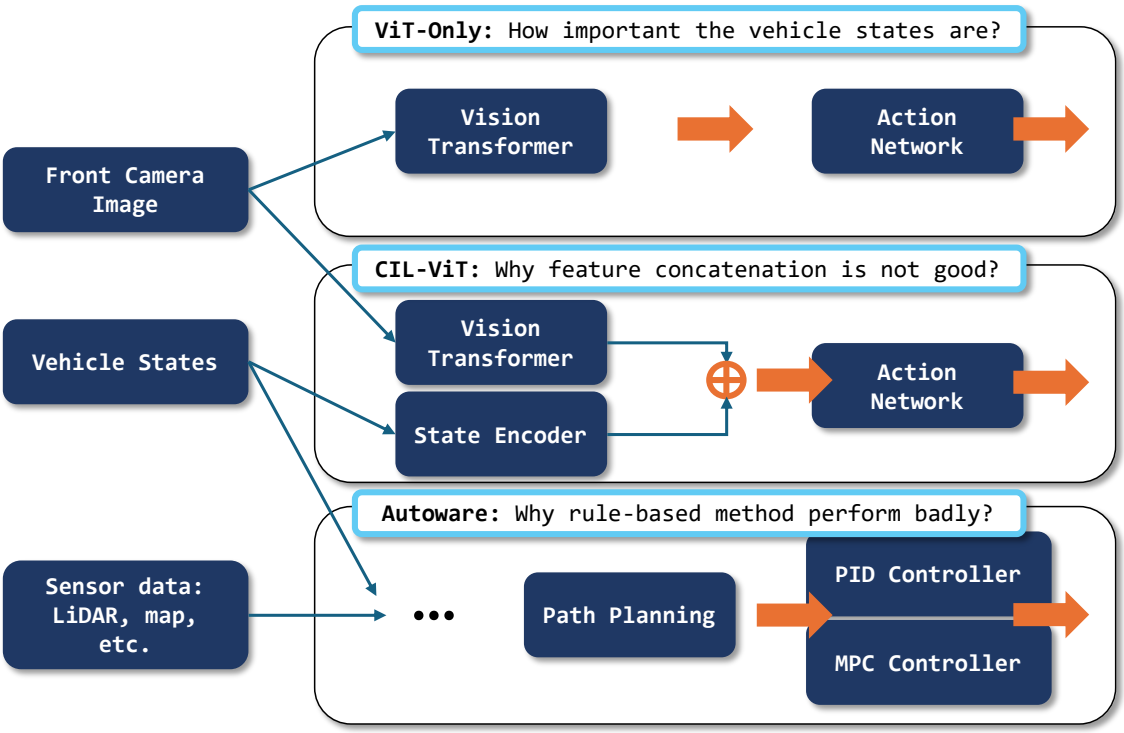


Fig. 5.3: The structures and their purposes of the baselines.

5.2 Implementation of End-to-end Autonomous Driving Agent

AWSIM simulator inherently supports native communication with ROS2 for co-simulation purposes. In our implementation, we implemented a customized ROS-based autonomous driving agent responsible for subscribing to the necessary input data from the AWSIM simulator and publishing action messages to control the ego vehicle effectively.

As Fig. 5.4 shows, the autonomous driving agent is built upon a pair of a ROS subscriber and a publisher. From the subscriber, it takes the necessary perception data as well as the collision reports from the listed topics. Then an episode management module, explained in

the above section, is used to control the simulation. The management includes asking the publisher to publish reset messages, sending the perception data to the E2E autonomous driving model, and recording the evaluation metrics during the simulation.

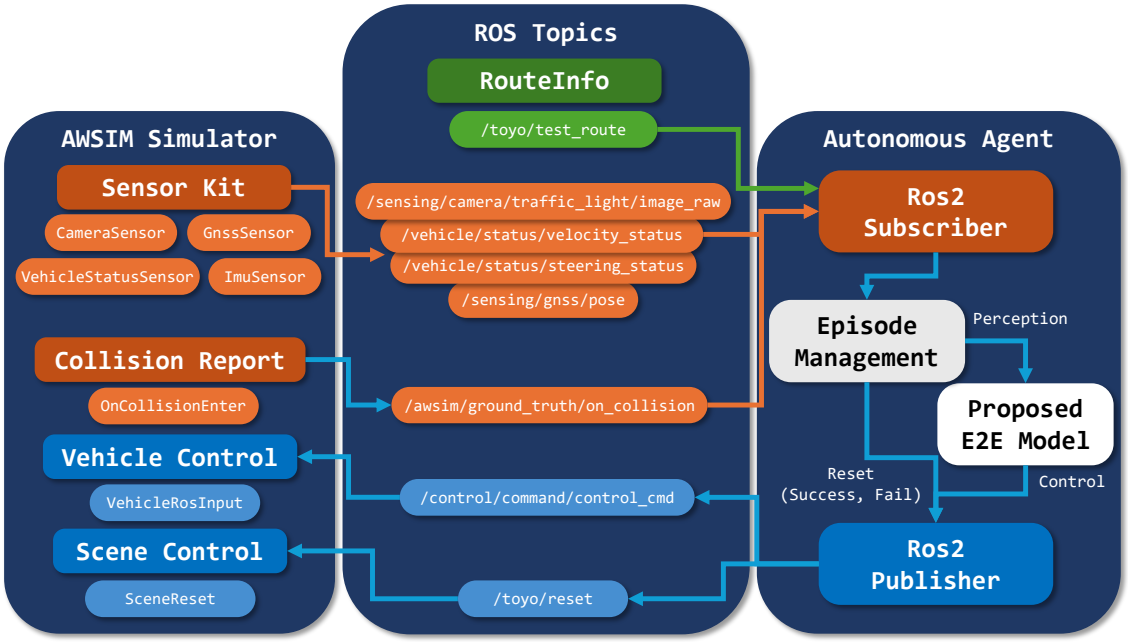


Fig. 5.4: The overview of the AD agent. The ROS topics used are listed in the middle of the figure.

5.3 Simulation Result

5.3.1 Comparing to Baselines

Success Rate and Average Time

The evaluation results for the designed scenarios of the unprotected right turn are depicted in Fig. 5.5 and Fig. 5.6. The observed 11.9% improvement over the CIL-ViT baseline in success rate proves an enhanced driving safety, highlighting the performance of our approach and the effectiveness of the proposed structure. We also noticed that CIL-ViT performed a 20% improvement over ViT-Only in success rate, indicating that vehicle states are crucial to unprotected turn scenarios that even simply concatenating the vehicle states helps much with the model's ability to drive safely. As a rule-based planner, Autoware performs the worst by a 42% success rate, we also noticed its disability of reacting fast to the opposing traffic flashing in front of the vehicle. We hypothesize that this disability comes from the fact that PID controllers tend to respond slowly to sudden appearances or disappearances of obstacles due to their reliance on feedback and integration of past errors.

The average passing time reveals that our model is slightly slower than the learning-based baselines (an average of 0.02 *seconds* and 0.74 *seconds* than CIL-ViT and ViT-Only), but 113% faster than the rule-based Autoware Controller, which tends to take an absolute stop when encountering an opposing NPC vehicle, causing the vehicle "face-off" problem. This result shows that our model is not achieving better driving safety by learning a timid driving policy.

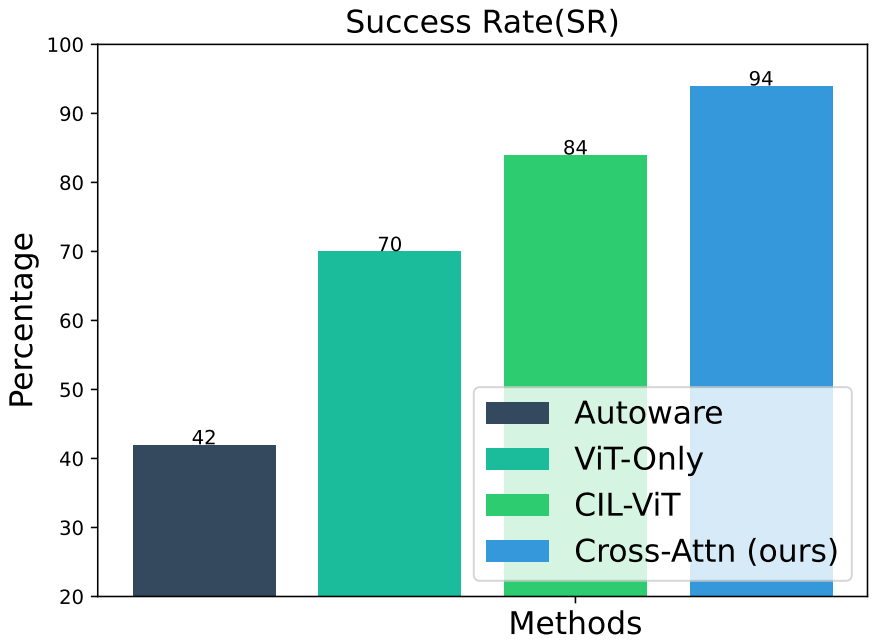


Fig. 5.5: Evaluation results of success rate (SR).

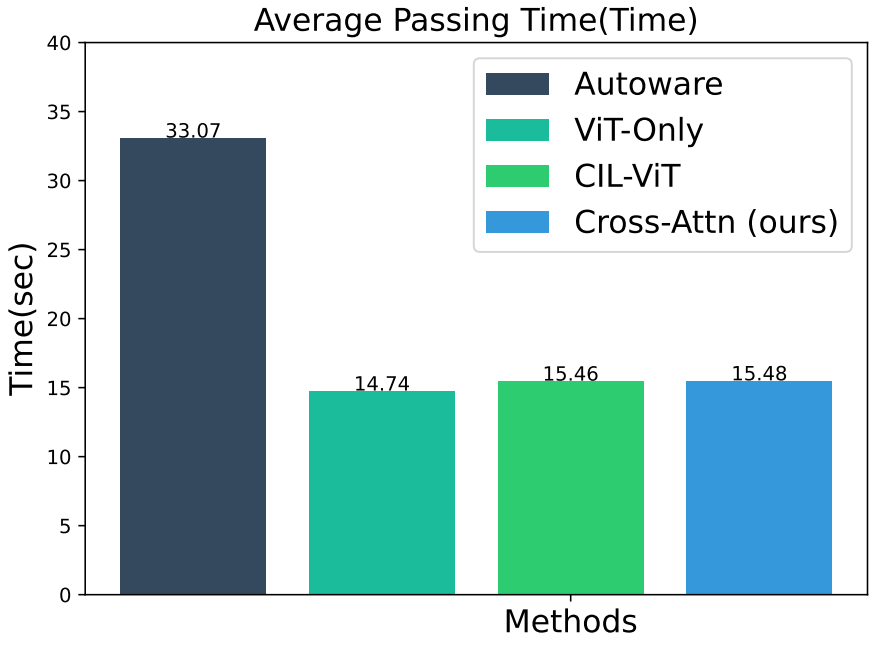


Fig. 5.6: Evaluation result of average passing time (Time).

State Curves

To further analyze the performance of our model, we captured a typical pattern of driving during the experiment by recording the vehicle state data and visualizing them through

curves. The curves in Fig. 5.7 demonstrate that our model generates smoother control commands, indicating improved stability. Especially in the steering curve, the result shows that the baseline model oscillated for 4 times to complete one turn, while the proposed method only turned 2 times. This enhanced stability can be attributed to a better understanding of the input image facilitated by the proposed cross-attention structure.

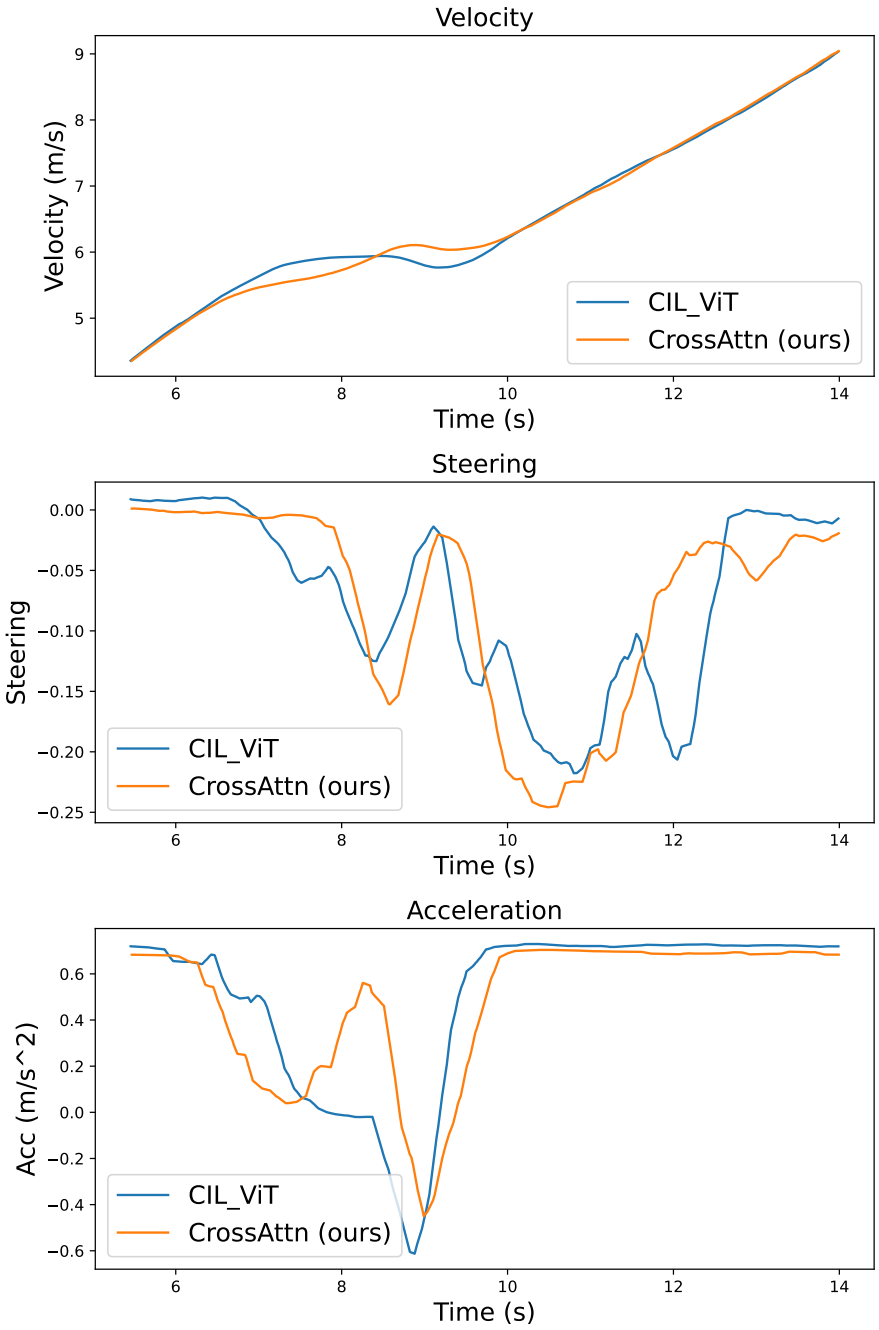


Fig. 5.7: Speed, steering, and acceleration curves during a typical driving episode. The curves for both methods are smoothed by the same sliding window of size 25.

5.3.2 Attention Visualization

We visualized attention maps in a typical driving scenario to gain insights into how the cross-attention layer operates. The attention map is depicted as a heatmap upon the original image extracted from the cross-attention layer, the color from blue to green to red stands for the attention value from lower to higher.

In Fig. 5.8, where the ego vehicle is traveling at a relatively low speed at 3.78 m/s, the model's attention is concentrated on the area directly in front of the ego vehicle and the opposing NPC vehicle, because it is the area that the ego vehicle will most likely to reach at the next timestamps. This focused attention is crucial for avoiding potential accidents. Conversely, in Fig. 5.9, with a velocity of 5.09 m/s, the model allocates more attention to the far-away parts of the road while maintaining overall awareness of the surrounding environment. This is because the presence of NPC vehicles in that area could significantly change the optimal decision - to maintain speed and pass when there's no NPC vehicle, or to decelerate earlier when there are NPC vehicles. The result shows that our cross-attention mechanism allows the model to adapt its focus based on the specific vehicle states, contributing to enhanced situational awareness and safety.

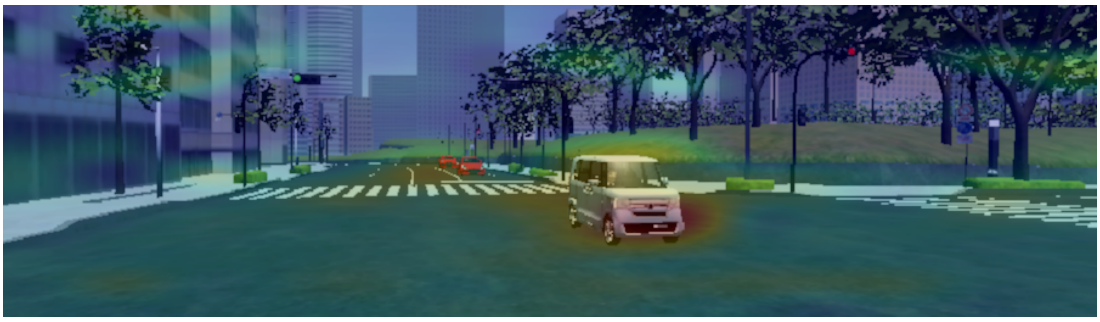


Fig. 5.8: Attention map visualized when the ego vehicle is traveling at a speed of 3.78 m/s

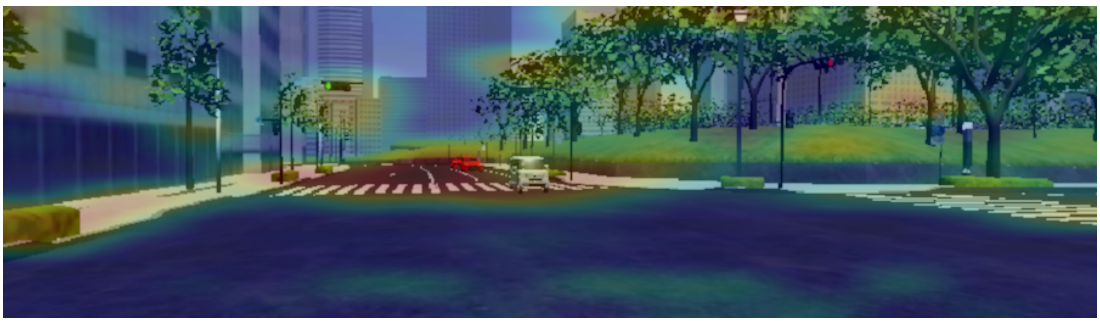


Fig. 5.9: Attention map visualized when the ego vehicle is traveling at a speed of 5.09 m/s

5.3.3 Computational Cost

Ensuring manageable computational costs is crucial for enabling real-time driving in real-world scenarios. We analyzed our model's overall latency of an average of 19.28 ms, which equates to the model being able to predict control commands approximately 51.86 times per second. The latency primarily arises from the higher computational cost of the vision transformer (ViT). However, despite this overhead, our system remains a feasible real-time end-to-end control system. Detailed latency comparison is shown in Table. 5.2.

Table 5.2: Computational Cost

Method	**Latency(ms)**	**FPS**
ViT-Only	18.42	54.29
CIL-ViT	18.47	54.14
CrossAttn(ours)	19.28	51.86

Chapter 6

Conclusion

6.1 Summary

In this thesis, we introduced a cross-attention-enhanced imitation learning framework for end-to-end autonomous driving, specifically designed to address the challenges of unprotected turns. By leveraging the relationship between vehicle states and environmental observations, our model demonstrated a significant improvement in success rates compared to baseline methods, highlighting the robustness and adaptability of our approach in complex driving scenarios.

Our proposed framework effectively integrates vehicle state information into the attention mechanism, allowing the model to dynamically allocate its focus on relevant areas of the input image. This not only enhances the model's understanding of the driving environment but also ensures smoother and safer control commands, as evidenced by our simulation results. The success of our approach showcases its potential for imitating complex driving policies with minimal reliance on extensive demonstrations.

6.2 Future Work

Despite these advances, there are several areas in which further improvements can be made. Training the framework on a more extensive and diverse dataset, encompassing various types of intersections, weather conditions, and traffic patterns, is crucial. This would help validate the model's generalizability and robustness across a wider range of scenarios. Incorporating additional sensory inputs, such as LiDAR, radar, and roadside unit (RSU) data, could further enhance the model's ability to perceive and understand complex driving environments. Exploring how these modalities can be fused with the cross-attention mechanism would be a valuable direction for future research.

While our model performed well in simulation, validating its performance in real-world scenarios is essential. This would involve deploying the model in real autonomous vehicles and testing its adaptability to dynamic and unpredictable environments. Investigating how the attention mechanism allocates focus across different driving scenarios could provide deeper insights into the model's decision-making process. Such analysis could help refine the attention mechanism to better handle edge cases, such as sudden pedestrian crossings or erratic vehicle behavior.

Conducting detailed comparisons with the latest state-of-the-art models that leverage multimodal inputs or advanced architectures could help identify specific strengths and areas for further optimization in our approach. Improving the computational efficiency of the framework, especially for deployment in real-time systems, is another critical area. This could involve designing lightweight architectures or employing model compression techniques.

By addressing these areas, future research can build upon the strong foundation laid by this study, advancing the capabilities of end-to-end autonomous driving systems in complex and dynamic environments.

Publications

(1) **Dongyang Li**, Ehsan Javanmardi, Jin Nakazato, and Manabu Tsukada. 2024. "End-to-end Vision Transformer-based Learning Approach for Autonomous Vehicle Path Planning." In Proceedings of 9^{th} Digital Signal Processing for In-Vehicle Systems and Safety.

(2) **Dongyang Li**, Ehsan Javanmardi, Naren Bao, and Manabu Tsukada. 2024. "Cross Attention Enhanced Imiatation Learning for End-to-end Autonomous Driving In Unprotected Turns." In Proceedings of 2024 International Conference on Intelligent Computing and its Emerging Applications. (**Best Paper Award**)

References

[1] Kenneth T. Jackson. *Crabgrass Frontier: The Suburbanization of the United States*. Oxford University Press, Oxford, UK, 1st edition, 1987.

[2] John Urry. The 'system' of automobility. *Theory, culture & society*, 21(4-5):25–39, 2004.

[3] SAE International. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Technical Report J3016_202104, https://www.sae.org/standards/content/j3016_202104/, April 2021.

[4] SAE International. Sae levels of driving automation™ refined for clarity and international audience, 2021. [Online; accessed 2025/01/12].

[5] Mariusz Bojarski. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[6] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021.

[7] Zachary Chase Lipton. A critical review of recurrent neural networks for sequence learning. *arXiv Preprint, CoRR, abs/1506.00019*, 2015.

[8] Ardi Tampuu, Tambet Matiisen, Maksym Semikin, Dmytro Fishman, and Naveed Muhammad. A survey of end-to-end driving: Architectures and training methods. *IEEE Trans. Neural Netw. Learn. Syst.*, 33(4):1364–1384, April 2022.

[9] Waymo. Left turns are hard for self-driving cars and people alike, 2019. [Online; accessed 2025/01/12].

[10] Jin Chen, Dihua Sun, and Min Zhao. Human-Like control for automated vehicles and avoiding "vehicle Face-Off" in unprotected left turn scenarios. *IEEE Trans. Intell. Transp. Syst.*, 24(2):1609–1618, February 2023.

[11] Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-End driving via conditional imitation learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4693–4700. IEEE, May 2018.

[12] Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. *ICCV*, pages 9328–9337, April 2019.

[13] Peng Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. *Adv. Neural Inf. Process. Syst.*, abs/2206.08129, June 2022.

[14] Kashyap Chitta, Aditya Prakash, and Andreas Geiger. NEAT: Neural ATtention fields for end-to-end autonomous driving. *ICCV*, pages 15773–15783, September 2021.

[15] TIER IV. Awsim: Open source simulator for self-driving vehicles, 2022. [Online; accessed 2025/01/12].

[16] Ekim Yurtsever, Linda Capito, Keith Redmill, and Umit Ozgune. Integrating deep reinforcement learning with model-based path planners for automated driving. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1311–1316. IEEE, 2020.

[17] Jianyu Chen, Bodi Yuan, and Masayoshi Tomizuka. Model-free deep reinforcement learning for urban autonomous driving. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 2765–2771. IEEE, October 2019.

[18] Jianyu Chen, Shengbo Eben Li, and Masayoshi Tomizuka. Interpretable end-to-end urban autonomous driving with latent deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):5068–5078, 2021.

[19] Zhiyu Huang, Jingda Wu, and Chen Lv. Efficient deep reinforcement learning with imitative expert priors for autonomous driving. *IEEE Trans Neural Netw Learn Syst*, 34(10):7391–7403, October 2023.

[20] Yi Xiao, Felipe Codevilla, Diego Porres, and Antonio M López. Scaling Vision-Based End-to-End autonomous driving with Multi-View attention learning. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1586–1593. IEEE, October 2023.

[21] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. TransFuser: Imitation with Transformer-Based sensor fusion for autonomous driving. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):12878–12895, November 2023.

[22] Hao Shao, Letian Wang, Ruobing Chen, Hongsheng Li, and Yu Liu. Safety-Enhanced autonomous driving using interpretable sensor fusion transformer. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 726–737. PMLR, 2023.

[23] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-Modal fusion transformer for end-to-end autonomous driving. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pages 7073–7083, April 2021.

[24] Weishang Wu, Xiaoheng Deng, Ping Jiang, Shaohua Wan, and Yuanxiong Guo. CrossFuser: Multi-Modal feature fusion for End-to-End autonomous driving under unseen weather conditions. *IEEE Trans. Intell. Transp. Syst.*, 24(12):14378–14392, December 2023.

[25] Qingwen Zhang, Mingkai Tang, Ruoyu Geng, Feiyi Chen, Ren Xin, and Lujia Wang. MMFN: Multi-Modal-Fusion-Net for End-to-End driving. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8638–8643. IEEE, October 2022.

[26] Chuan Jiang, Yuyan Ding, Yujun Zeng, and Haibin Xie. Multi-Modal deep CrossFusion network for autonomous driving. In *2023 China Automation Congress (CAC)*, pages 2109–2115. IEEE, November 2023.

[27] Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. End-to-end urban driving by imitating a reinforcement learning coach. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15222–15232, 2021.

[28] Zhangjie Cao, Erdem Bıyık, Woodrow Z Wang, Allan Raventos, Adrien Gaidon, Guy Rosman, and Dorsa Sadigh. Reinforcement learning based control of imitative policies for near-accident driving. *arXiv preprint arXiv:2007.00178*, 2020.

[29] Jiaqi Liu, Xiao Qi, Ying Ni, Jian Sun, and Peng Hang. Teaching autonomous vehicles to express interaction intent during unprotected left turns: A human-driving-prior-based trajectory planning approach. *arXiv preprint arXiv:2307.15950*, 2023.

[30] Jiqian Dong, Sikai Chen, Shuya Zong, Tiantian Chen, and Samuel Labi. Image transformer for explainable autonomous driving system. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2732–2737. IEEE, September 2021.

[31] Bob Wei, Mengye Ren, Wenyuan Zeng, Ming Liang, Bin Yang, and Raquel Urtasun.

Perceive, attend, and drive: Learning spatial attention for safe Self-Driving. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4875–4881. IEEE, May 2021.

[32] Luca Cultrera, Lorenzo Seidenari, Federico Becattini, P Pala, and A Bimbo. Explaining autonomous driving by learning end-to-end visual attention. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1389–1398, June 2020.

[33] Hyunki Seong, Chanyoung Jung, Seungwook Lee, and David Hyunchul Shim. Learning to drive at unsignalized intersections using attention-based deep reinforcement learning. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 559–566. IEEE, September 2021.

[34] Diego Porres, Yi Xiao, Gabriel Villalonga, Alexandre Levy, and Antonio M. López. Guiding attention in end-to-end driving models. In *2024 IEEE Intelligent Vehicles Symposium (IV)*, pages 2353–2360, 2024.

[35] Shinpei Kato, Shota Tokunaga, Yuya Maruyama, Seiya Maeda, Manato Hirabayashi, Yuki Kitsukawa, Abraham Monrroy, Tomohito Ando, Yusuke Fujii, and Takuya Azumi. Autoware on board: Enabling autonomous vehicles with embedded systems. In *Proceedings of the 9th ACM/IEEE International Conference on Cyber-Physical Systems (ICCPS)*, pages 287–296, 2018.

[36] CARLA Autonomous Driving Leaderboard. Get started with leaderboard 2.0, 2024. [Online; accessed 2025/01/12].

[37] Steven Macenski, Tully Foote, Brian Gerkey, Chris Lalancette, and William Woodall. Robot operating system 2: Design, architecture, and uses in the wild. *Sci Robot*, 7(66):eabm6074, May 2022.

[38] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[39] Ashish Vaswani, Noam M Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Adv. Neural Inf. Process. Syst.*, pages 5998–6008, June 2017.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Manabu Tsukada, for his invaluable guidance, support, and encouragement throughout my master's journey. Your expertise and constructive feedback have been instrumental in shaping this thesis and in fostering my academic and professional growth.

I am also grateful to Tsukada Lab and its members for providing a collaborative and inspiring environment. Special thanks to Professor Ehsan Javanmardi, whose insights, discussions, and suggestions made the challenges of research more manageable and the journey more enjoyable.

My heartfelt thanks go to my family for their unwavering support and understanding. To my parents, who have always believed in me and stood by me in every step of my education, whose encouragement and patience kept me motivated during the most demanding times—I could not have done this without you.

Finally, I would like to thank all my friends and mentors who have cheered me on and offered advice and support throughout this journey. To Pengfei Lin, Ye Tao, Yuze Jiang, Shangkai Zhang, Donghui Li, and Xuzong Zhang, thank you for the camaraderie, advice, and the fun times we shared during this journey, and thank you for encouraging and guiding me when I came to a foreign country for the very first time in my life.

Thank you all for being part of this transformative journey.