DME-Driver: Integrating Human Decision Logic and 3D Scene Perception in Autonomous Driving

Wencheng Han, Dongqian Guo, Cheng-Zhong Xu, Jianbing Shen*

SKL-IOTSC, CIS, University of Macau
wencheng256@gmail.com, jianbingshen@um.edu.mo

Abstract

There are two crucial aspects of reliable autonomous driving systems: the reasoning behind decision-making and the precision of environmental perception. This paper introduces Decision-Maker and Executor (DME)-Driver, a new autonomous driving system that enhances performance and robustness by fully leveraging the two crucial aspects. This system comprises two main models. The first, the Decision Maker, is responsible for providing logical driving instructions. The second, the Executor, receives these instructions and generates precise control signals for the vehicles. To ensure explainable and reliable driving decisions, we build the Decision-Maker based on a large vision language model. This model follows the logic employed by experienced human drivers and simulates making decisions in a safe and reasonable manner. On the other hand, the generation of accurate control signals relies on precise and detailed environmental perception. Therefore, a planning-oriented perception model is employed as the Executor. It translates the logical decisions made by the Decision-Maker into accurate control signals for the self-driving cars. To effectively train the proposed system, a new dataset named **H**uman-driver **B**ehavior and **D**ecision-making (HBD) dataset has been collected. This dataset encompasses a diverse range of human driver behaviors and their underlying motivations. By leveraging this dataset, our system achieves high-precision planning accuracy through a logical thinking process.

Introduction

Self-driving systems represent a significant advancement in automobile car technology, combining computer vision and artificial intelligence. These systems enable cars to perceive their surroundings (Radecki, Campbell, and Matzen 2016; Sun et al. 2020), make informed decisions (Hang et al. 2020; Isele et al. 2018), and navigate without human intervention (Kamath et al. 2023; Shao et al. 2023). The keys to these systems are sophisticated perception mechanisms and decision-making algorithms. These components allow cars to accurately understand their environment and make autonomous decisions, ensuring safe and efficient navigation. However, the complexity of these systems requires a focus on interpretability. Achieving interpretability is not just a technical challenge, but also a crucial step towards wider acceptance and integration of autonomous cars in society.

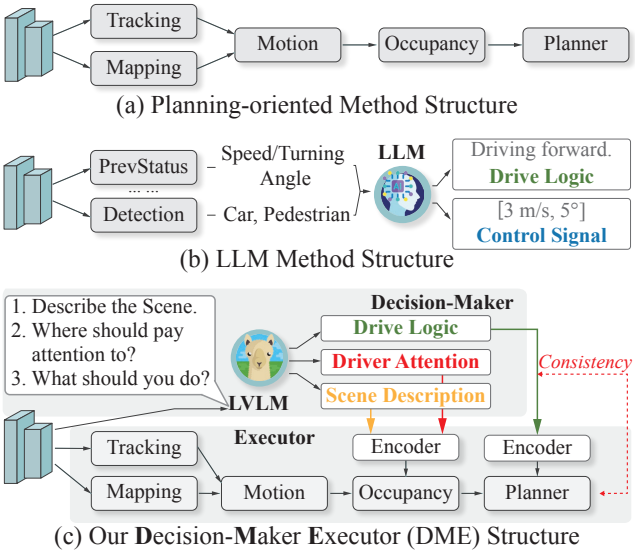

Figure 1: **Comparative Structures in Autonomous Driving Systems:** (a) Planning-oriented autonomous driving system; (b) LLM-based autonomous driving system; (c) Our DME-Driver Autonomous Driving System.

Recently, deep learning-based methods have achieved remarkable success in the realm of autonomous driving (Feng, Rosenbaum, and Dietmayer 2018; Maqueda et al. 2018). Some works (Feng et al. 2024; Gu et al. 2023; Huang et al. 2023) proposed planning-oriented autonomous driving systems that can be trained end-to-end. As illustrated in Fig. 1(a), this system (Chen et al. 2023) encompasses several critical perception modules, including tracking, mapping, motion, and occupancy detection. The output of these modules is fed into a planner that then generates the control signals for the vehicle. This approach takes full advantage of the perception models in autonomous driving cars, significantly enhancing the overall accuracy of the planning process. However, the end-to-end nature of such methods leads to a lack of reasoning in terms of human driving logic, resulting in a system that lacks interpretability. When confronted with scenarios that the model fails to resolve - which we might term “bad cases” – it becomes challenging to un-

*Corresponding author: *Jianbing Shen*.

derstand the flawed decision-making process of the system.

Taking advantage of advances in large language models (LLMs), recent approaches attempted to enhance the reasoning of autonomous driving systems by integrating LLMs into their frameworks (Li et al. 2024; Liu et al. 2023b; Ma et al. 2023; Wei et al. 2024), as shown in Fig. 1(b). Using their formidable logic reasoning capabilities and generalization, LLMs can effectively understand the behavioral logic of human drivers in various driving scenarios (Ding et al. 2024; Pan et al. 2024; Peng et al. 2023). In the face of corner cases, they can attempt to mimic human driver behavior, making rational judgments and decisions (Kim et al. 2020). Moreover, even when decision-making errors occur, the logical reasoning process of these models can be used as evidence to pinpoint the reasons of erroneous judgments and to seek solutions. However, these methods place LLMs at the core of the autonomous driving system, treating the outputs of perception modules as given conditions. These approaches did not fully utilize the various perception tasks essential for accurate environmental perception. When perception models themselves produce erroneous predictions, these errors can be accumulated in the decision-making process.

In this paper, we propose the DME (**D**ecision-**M**aker **E**xecutor)-Driver autonomous driving system, which utilizes the advantages of both LLMs and planning-oriented perception models. Our system utilizes the capability of LLMs to comprehend human driving logic, while also leveraging the precise environmental perception abilities of planning-oriented autonomous driving models for accurate vehicle control. As depicted in Fig. 1(c), our system comprises two key roles: Decision-Maker and Executor. The Decision-Maker is based on a Large Vision Language Model (LVLM), trained through imitation learning on extensive driving data and human driver behavior logic. It can fully grasp the relationship between driving scenarios and the underlying driving logic. When the vehicle encounters a new scene, the Decision-Maker can simulate human-like logical assessments of key elements, determining whether to accelerate, brake, or change lanes. Additionally, the Decision-Maker can mimic a human driver's ability to discriminate key aspects of driving scenes, providing the perception model with a reliable set of prior information. This helps the perception models to focus on elements of particular importance in the current scenario.

The Executor, on the other hand, is responsible for converting the Decision-Maker's instructions into precise vehicle control signals. It translates the high-level logic and reasoning from Decision-Maker into precise control signals. By doing so, the Executor bridges the gap between decision-making and vehicle control, enabling the system to navigate safely and efficiently in diverse driving conditions.

In summary, this paper presents four key contributions:

- **DME-Driver Autonomous Driving System:** We present the DME-Driver system, which combines the strengths of LVLMs in logical reasoning with the precise environmental sensing of planning-oriented models, improving decision-making robustness and performance in autonomous driving.
- **Human-Driver Behavior and Decision-Making (HBD) Dataset:** Leveraging both reannotated datasets and newly collected data, we developed a distinctive dataset that integrates human driver behavior logic with detailed environmental perception. It aids in understanding the relationship between driving logic and control signals.
- **Decision-Maker Model Design:** Our Decision-Maker model, based on LVLM, is capable of imitating human driver instructions and focusing on important elements in the environment, providing human-like insights for better decision-making.
- **Executor Model Formulation:** The Executor model accurately processes environmental data and translates the Decision-Maker's instructions into accurate vehicle control signals, ensuring effective and context-aware responses in various driving situations.

Empirical evaluations demonstrate that our method achieves state-of-the-art accuracy in autonomous driving planning, significantly enhancing the system's reasoning ability. Every driving decision made by the system can be traced back through logs to understand the underlying driving logic, providing a level of transparency and explainability that is unprecedented in autonomous driving systems.

Related Work

Autonomous Driving System

The development of autonomous driving technology has evolved from traditional rule-based systems to sophisticated learning-based approaches. Initially, autonomous vehicles were governed by algorithms that heavily relied on sensor-based data and predefined rules (Montremerlo et al. 2008). While these methods provided reliable results in controlled environments, they struggled with the unpredictability and complexity inherent in real-world driving conditions.

The advent of deep learning revolutionized this landscape, enabling systems to learn from vast and varied datasets of real driving scenarios (Grigorescu et al. 2020; Wu et al. 2023b,a). This shift to learning-based approaches has endowed autonomous systems with the flexibility and adaptability needed to navigate complex and dynamic environments more effectively, paving the way for more robust and versatile driving systems.

A significant trend in recent years are the development of end-to-end autonomous driving systems. These systems utilize deep neural networks to process sensory inputs directly into driving actions, seeking to streamline the autonomous driving process (Bojarski et al. 2016; Han et al. 2024). While these approaches simplify the system architecture by eliminating modular decomposition, they raise challenges in terms of interpretability and robustness. The black box nature of these systems often hinders their ability to explain decisions and adapt to novel situations, which is critical for ensuring safety and gaining user trust in real-world applications.

Data Source	Q&A Turns	Type					
		Des	Gaz	Gaz-r	Act	Act-r	3D
Look Both Ways (Kasahara, Stent, and Park 2022)	3	✓	✓	✓			
BDD-X (Yu et al. 2020)	3	✓			✓	✓	
NuScenes (Caesar et al. 2020)	2	✓			✓		✓
Newly-collected	5	✓	✓	✓	✓	✓	✓

Table 1: **HBD Dataset Comparison.** In this table, "Des" stands for *scene description*, "Gaz" represents *gaze description*, "Gaz-r" indicates *gaze understanding*, "Act" signifies *action prediction*, "Act-r" means *driving reasoning*, and "3D" represents *3D perception tasks*.

Large Language Model for Autonomous Driving

The development of Large Language Models (LLMs) has greatly enhanced natural language understanding and generation. Models such as BERT (Devlin et al. 2018) and GPT (Radford et al. 2019) laid the foundation for advanced systems like GPT-3 (Brown et al. 2020), which excel in generating coherent text and understanding language subtleties, making them valuable for various applications.

However, LLMs are not without their challenges. The computational requirements for training and running these models are substantial, raising concerns about environmental sustainability and the digital divide (Strubell, Ganesh, and McCallum 2019). Furthermore, biases in model outputs and their implications are an ongoing and active research areas (Bender et al. 2021).

For a long time, autonomous driving systems have been treated as black boxes with a lack of interpretability, making it difficult to understand how decisions are made. The development of LLMs promises to solve this problem (Ma et al. 2025; Mao et al. 2023b; Wen et al. 2023). For example, GPT-Driver (Mao et al. 2023a) transforms the GPT-3.5 model into a motion planner for autonomous driving, which demonstrates the motion planning abilities of the LLMs. DriveGPT4 (Xu et al. 2023), an interpretable end-to-end autonomous driving system based on LLMs, utilizes multimodal data such as videos, texts, and historical control signals. It generates textual responses to questions and predicts control signals for vehicle operation. Also, the reasoning abilities of LLMs improve the performance in perception and understanding tasks. LLM-AD (Elhafsi et al. 2023), a semantic anomaly detection framework utilizing LLMs' reasoning abilities, demonstrates that the LLM-based monitor aligns with human intuition in both fully end-to-end policies and classical autonomy stacks utilizing learned perception.

Human Driver Behavior and Decision-Making Dataset

Dataset Introduction

In order to thoroughly exploit the relationship between the human driving logic and robust autonomous driving, our study focuses on following four key aspects:

Human Driver Gaze. Driving on roads requires real-time responses, often based on subconscious reflexes. To comprehend these reflexive actions, we consider human gaze as an informative behavioral signal. During driving, human drivers instinctively focus on the most critical parts of the scene, which are typically directly influential in the driving logic of the current scenario. These elements could include traffic signals or other elements that might interact with the vehicle shortly. Understanding this gaze behavior is crucial for our system to recognize and prioritize important aspects of the driving environment.

Understanding of Driving Scenes. The way human drivers logically describe driving scenes provides a rich, purposeful understanding of the scenario. Unlike standard image captions that offer a global view, human drivers focus on elements and their interrelations that could impact driving decisions. Based on these logical descriptions, even another human driver can make accurate judgments about appropriate driving actions. For instance, a description: *I am in a scenario at an intersection where the vehicle is in a left-turn lane with a red light for turning.*

Decision-Making and Rationale. The decision-making process of human drivers is logical and information-rich. It encompasses the driver's synthesis of various factors to determine the appropriate action for a given scenario, along with the underlying thought process. By comprehending and emulating this aspect, an autonomous driving system can mimic human-like decision-making logic. This capability is particularly valuable in novel or challenging scenarios not encountered during training, enabling the system to make safe and reliable judgments.

Precise Control Signals. The ultimate output of an autonomous driving system should be precise control signals directly applicable to the vehicle. Regardless of robustness and interpretable the natural language-based driving logic is, it must be translated into concrete control commands. This translation is vital to ensure that the detailed understanding and decisions derived from human driving logic are effectively and safely executed by the autonomous vehicle.

Given the requirement for four distinct types of raw data, we faced the challenge that no existing open source dataset covers all these aspects. To address this gap, we reannotate three different open-source datasets, each contributing one or more of the needed data types. For example, Look-Both-Ways (Kasahara, Stent, and Park 2022) provides precise human driver gaze information collected via eye tracking devices. BDD-X (Yu et al. 2020) offers manually annotated driving decisions and their underlying reasons. Nuscenes (Caesar et al. 2020) includes detailed control signals. Using the powerful generalization capabilities of LVLM, we can extrapolate the knowledge learned from each dataset to the others, even in the absence of a single

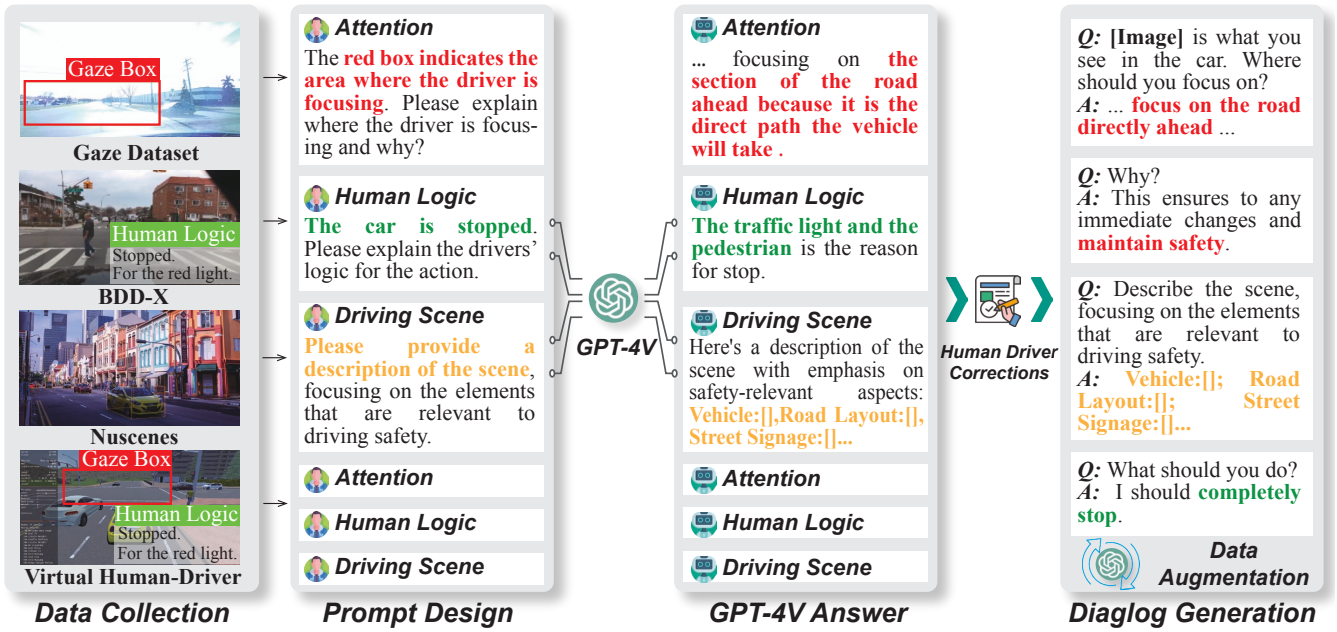

Figure 2: **The Annotation Pipeline of our HBD Dataset:** There are four key steps in our dataset creation - Data Collection, Prompt Design, Manual Corrections, and Dialogue Generation, streamlining the process from raw data gathering to structured dialogue formation.

dataset containing all types of annotation.

To further enhance the LVLM's capabilities, we collected a new sub-dataset with multi-round question-answering conversations. This sub-dataset was created using the Carla (Dosovitskiy et al. 2017) simulation engine, where human drivers operated simulated driving controls such as steering wheels and pedals. This process provided visual information and control signals for the vehicle. Additionally, drivers manually annotated their specific decisions and the reasons behind them at each time. During the driving sessions, an eye tracker was employed to capture the drivers' gaze, focusing on the objects they observed in real-time.

The re-annotated data include 591,574 images with 985,739 targets, forming 191,786 conversations. Each conversation contains 2-3 turns of question and answer exchanges. While the newly collected data comprises 2,608,038 images with more than 3 million targets in total, forming 207,150 multi-turn conversations. Each conversation includes five turns of Q&A exchanges, covering all aspects of driving logic understanding. Table 1 presents a comparison of the different data sources used in the HBD dataset. The annotation type indicates the type of annotation provided by each data source. A more detailed analysis of the HBD dataset is provided in the supplementary material.

Data Collection and Labels Generation

To efficiently process and label these data, we employ a combination of GPT-4V pre-annotation followed by manual corrections. Fig. 2 outlines the data collection pipeline for our HBD Dataset. This pipeline comprises four main steps, starting from data collection, involving Prompt Design and manual correction, to the generation of dialogues.

Step 1: Data Collection Our data collection integrates four sub-datasets, including three open-source datasets - Look-Both-Ways (Kasahara, Stent, and Park 2022), BDD-X (Yu et al. 2020), and Nuscenes (Caesar et al. 2020) - and one newly collected sub-dataset. This comprehensive combination provides a rich base of raw data encompassing various aspects of human driving behavior.

Step 2: Prompt Design To guide GPT-4V in analyzing driving behaviors from a human driver's perspective, we crafted unique prompts for each data type. For example, with human gaze data, the aim is for LVLM to understand which areas a human driver focuses on in a given scene and why. We first convert gaze points into a box indicating the attention distribution in this scene. This box is then drawn on the image. Finally, we employ GPT-4V to processes this enhanced image to infer the driver's focus and intentions.

Step 3: Manual Corrections After GPT-4V generates outputs, these data are manually reviewed and corrected. Completely incorrect responses are deleted and regenerated with altered prompts. For correct responses, any details that do not align with typical human driver thought processes are manually adjusted.

Step 4: Dialogue Generation Once all data are transformed into detailed textual information, the next step is organizing this information into dialogues, which involves three sub-steps:

First-person Conversion: All pronouns in the dialogues are converted to the first person. This is to ensure that the subsequent LVLM Decision-Maker model can process the dialogues from the perspective of the human driver.

Combining Multi-turn Dialogues: The labeling process typically involves single-question prompts. For practical use, multi-turn dialogues offer more contextual clues for the LVLM. We thus concatenate different types of Q&A information into multi-turn dialogues, as shown in Fig. 2. The number of dialogue turns varies – 2 to 3 turns for reannotated data, depending on the information available, and up to 5 turns for new collected data.

Data Augmentation: To diversify the question-answer labels and enhance the model's generalization capabilities, we utilize GPT-3.5 to rewrite the generated dialogues. This process involves changing the form of the dialogues while keeping their content consistent.

The Proposed DME-Driver Autonomous Driving System

In our DME-Driver Autonomous Driving System, as illustrated in Fig. 3, the system is divided into two main components: the Decision-Maker and the Executor. The Decision-Maker acts as the central decision-maker, synthesizing vehicle status and current visual inputs to emulate a human driver's logical judgments. Its output is expressed in natural language, providing a logical and interpretable narrative of driving decisions. However, as natural language cannot directly control a vehicle, the system incorporates the Executor network, functioning as a translator. This network converts the Decision-Maker's linguistic outputs into precise vehicle control commands. The detailed architecture and functions of both the Decision-Maker and the Executor networks are crucial to the system's effectiveness and will be elaborated in Sections 4.1 and 4.2 respectively.

Decision-Maker: Driving Logic Understanding

The Decision-Maker in our DME-Driver system is a sophisticated Large Vision Language Model (LVLM) designed to simulate the decision-making process of human drivers. In our experiments, we use LLaVA (Liu et al. 2023a) as the baseline network for the Decision-Maker. This component is engineered to process inputs from three different modalities: visual inputs from the current and previous scenes, textual inputs in the form of prompts, and current status information detailing the vehicle's operating state.

Visual Input: To process the visual input, we utilize a pretrained CLIP (Radford et al. 2021) visual encoder E_{CLIP}. This encoder converts the visual information into feature tokens. To better comprehend the context of driving scenes, we enhance the input by concatenating the previous three key frames with the current frames into an image array:

$$\mathrm{F}_v = E_{CLIP}(F_{t-3} \oplus F_{t-2} \oplus F_{t-1} \oplus F_t) \tag{1}$$

Textual and Status Input: The prompt inputs T_p and current status T_s information are handled using a methodology similar to RT-2 (Brohan et al. 2023), where a text tokenizer is employed to encode these inputs uniformly. After encoding, the tokens representing both visual and textual information are concatenated and fed into the LLaMA 2 (Touvron et al. 2023) model for processing:

$$\begin{aligned} \mathrm{F}_t &= \text{Tokenizer}(T_p) \oplus \text{Tokenizer}(T_s) \\ \mathrm{t} &= \text{LLaMA2}(\mathrm{F}_t + \mathrm{F}_v) \end{aligned} \tag{2}$$

This integrated approach allows the Decision-Maker to consider all aspects of the driving scenario, ensuring a comprehensive understanding and simulation of human-like decision-making processes. The final step involves a detokenizer, which maps the output tokens back into natural language.

Executor: The Control Signal Generator

As depicted in Fig. 3, our Executor network in the DME-Driver system is designed based on the UniAD (Hu et al. 2023) planning-oriented autonomous driving framework, featuring 4 distinct components:

Backbone Network: The initial layer of the Executor network is a backbone, which is responsible for encoding multi-view vision inputs. A image encoder is firstly employed to extract image features and then multi-view image features are transformed into Bird's Eye View (BEV) features through a process similar to EVFormer (Li et al. 2022).

Perception Modules: The next stage consists of four specialized perception modules. *TrackFormer* is designed for detecting and tracking various elements within the driving scene. *MapFormer* generates a segmented map in BEV, providing detailed spatial information about the environment. *MotionFormer* predicts the motion trajectories of each element within the scene. *OccFormer* is responsible for generating occupancy information, indicating the areas within the scene that are occupied and those that are free.

Planning Module: Following the perception modules, the Planning Module takes the output tokens from these modules as its input. This module's primary function is to generate the predicted control signals for the vehicle.

Driver Logic Encoder: Distinct from the UniAD system, our Executor network incorporates additional enhancements for the TrackFormer, the OccFormer and the Planning module. The TrackFormer and OccFormer combines textual information from scene descriptions T_{des} and gaze understanding T_{gz}, while the Planning module integrates decision making results T_{act}. Specifically, we've integrated a Bert-based text encoder E_{bert} to process corresponding textual inputs:

$$\begin{aligned} F_{t\_pcp} &= E_{bert}(T_{des}) \oplus E_{bert}(T_{gz}) \\ F_{t\_plan} &= E_{bert}(T_{act}), \end{aligned} \tag{3}$$

where F_{t_pcp} represents the text encoding for perception tasks like TrackFormer and OccFormer, while F_{t_p} represents the text encoding for the Planning module. After encoding the text, we combine the feature B generated by the backbone network with the corresponding text encoding using a transformer fusion structure named LogicalFusioner. In this structure, we consider the backbone feature as the query and the text encoding as the key and value. After the aggregating of the multi-head attention, we add a shortcut connection to the original B and produce the enhanced backbone feature B':

$$\begin{aligned} B' &= \text{LogicalFusioner}(B, T) \\ &= \text{MHA}(Q = B, K = T, V = T) + B \end{aligned} \tag{4}$$

The Driver Logic Encoder is the core module in our Executor. It gives the Executor ability for integrating decision-making and scene understanding information provided by

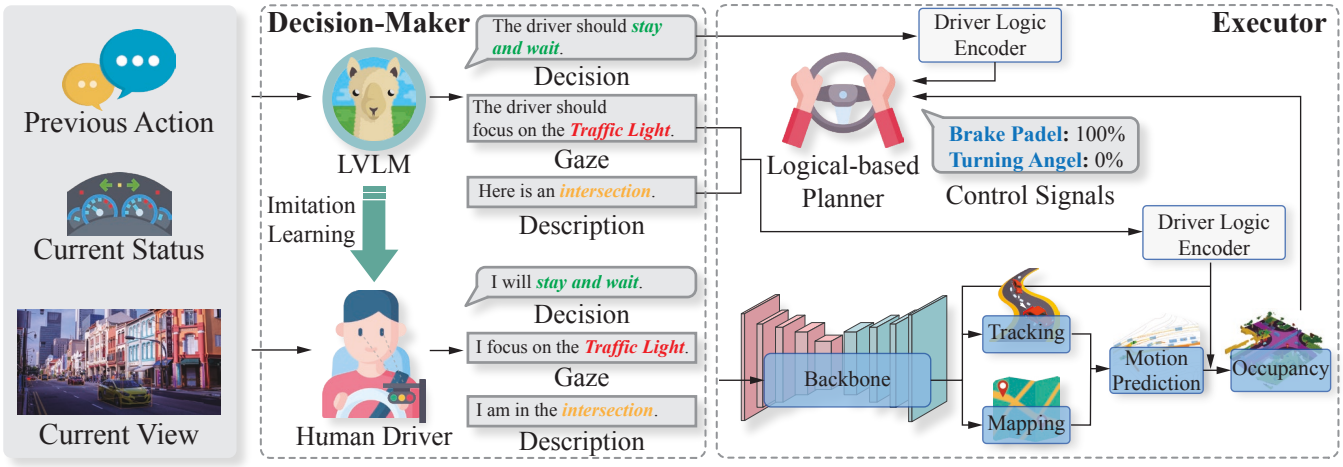

Figure 3: **The Detailed DME-Driver Autonomous Driving System:** There are two main components in our system. The **Decision-Maker** serves as the decision-maker, emulating human drivers' logic in various driving scenarios. The **Executor** then translates the Decision-Maker's instructions into precise control signals, ensuring effective execution of driving decisions.

the Decision-Maker. With this information, it can emulate human driving logic, achieving robust and accurate vehicle control. Our experiments are conducted on a workstation with 8x H800 Graphic Cards.

Training

The training of the DME-Driver system is streamlined into two essential steps. Firstly, training the Decision-Maker involves using multi-type human driver decision data to understand the driving logic. Secondly, the training focuses on the Executor, which is trained using Decision-Maker instructions and perception labels. Using these data, the executor can learn to accurately transform instructions into control signals.

Decision-Maker Training: The training of the Decision-Maker network in our DME-Driver system encompasses two critical stages: pretraining and fine-tuning. Initially, the model undergoes pretraining on diverse datasets, including 593K image-text pairs from CC3M (Changpinyo et al. 2021) and 100K video-text pairs from WebVid-10M (Bain et al. 2021), focusing on general video-text alignment. This phase involves training the video tokenizer while keeping the CLIP encoder and LLM weights fixed. The fine-tuning stage then tailors the model to the specific needs of interpretable autonomous driving. Here, the LLM is trained alongside the visual tokenizer using 39K video-text pairs, from the proposed HBD Dataset, and supplemented with 80K instruction-following image-text pairs from LLaVA (Liu et al. 2023a).

Executor Training: The training of the Executor component in our DME-Driver system primarily follows the setup utilized by UniAD (Hu et al. 2023). However, we introduce specific modifications to enhance the system's consistency. Initially, similar to UniAD, we start by jointly training the perception parts, namely the tracking and mapping modules, for six epochs. We then proceed to an end-to-end training phase, which lasts for 20 epochs and encompasses all perception, prediction, and planning modules. To ensure alignment between the output signals of the planning module and the decisions made by the Decision-Maker, we introduce an auxilary loss during the training of the planning module. This component applies a penalty whenever the control signals deviate from the Decision-Maker's decisions. Specifically, we category the Decision-Maker's decisions into distinct types, such as *moving forward*, *accelerating*, *stopped*, among others. For each of these decision types, we've established specific rules to determine whether a given control signal corresponds to one of these categories. The influence of the consistency loss is discussed in the supplementary material.

Experiment

Human Driver Logic

To assess the accuracy of the Decision-Maker in simulating human driver decision-making in driving scenarios, we conducted an evaluation using the test set of the HBD dataset. In our evaluation process, we adapted a method similar to previous works (Xu et al. 2023), utilizing an advanced version of ChatGPT to generate assessment metrics. Details about the evaluation procedure is provided in the supplementary material.

The evaluation of the Decision-Maker's accuracy was conducted across four key dimensions: **Gaze:** Assessing the accuracy of the Decision-Maker in identifying areas of focus during the driving process. **Scene Understanding:** Evaluating how precisely the Decision-Maker describes the elements present in the current driving scene. **Logic:** Determining the accuracy of the final driving decisions made by the Decision-Maker and the corresponding reasoning. To evaluate our system's performance, we compared the Decision-Maker's accuracy with that of other general-purpose large models, including LLaVA and GPT-4V, in similar scenarios. The outcomes of these comparisons are detailed in Table 3. *Examples of the decision-making results can be found in the*

Method	Input	L2(m)↓				Col. Rate(%)↓			
		1s	2s	3s	Avg.	1s	2s	3s	Avg.
NMP(Zeng et al. 2019)	Lidar	-	-	2.31	-	-	-	1.92	-
SA-NMP(Zeng et al. 2019)	Lidar	-	-	2.05	-	-	-	1.59	-
FF(Hu et al. 2021)	Lidar	0.55	1.20	2.54	1.43	0.06	0.17	1.07	0.43
EO(Khurana et al. 2022)	Lidar	0.67	1.36	2.78	1.60	0.04	0.09	0.88	0.33
ST-P3(Hu et al. 2022)	Vision	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD(Hu et al. 2023)	Vision	0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
DME-Driver	Vision	0.43	0.91	1.58	0.97	0.04	0.14	0.64	0.27

Table 2: **Planning Accuracy.** This table shows the comparison with SOTA methods on the planning task.

Method	Gaze	Scene	Logic
	ChatGPT4↑		
LLaVA-7B(Liu et al. 2023a)	45.3	59.6	43.2
LLaVA-13B(Liu et al. 2023a)	48.3	60.5	45.3
GPT-4V(Yang et al. 2023)	75.3	79.8	67.1
DME-Driver	**85.2**	**81.6**	**78.5**

Table 3: **Driving Logic Understanding.** This table contrasts our Decision-Make with other large language models using GPT-4 evaluation.

supplementary material.

Planning

The aim of this experiment is to validate the accuracy of the entire DME-Driver system in making autonomous driving decisions. Following methodologies established in previous works, we focus on evaluating the planning accuracy of the system. As shown in Table 2, the results of our experiments demonstrate that the DME-Driver system successfully harnesses the logic-driven prompts from the Decision-Maker, enhancing the decision-making precision of the planning module. This integration not only leads to higher decision accuracy in diverse driving situations but also maintains a detailed log of the decision-making process. The ability to trace back and understand the rationale behind each decision is a critical aspect of our system, adding a layer of reasoning and accountability that is crucial for autonomous driving systems. *More comprehensive experiments on other perception tasks are provided in the supplementary material.*

Module	L2(m)↓	Col.Rate(%)↓
Executor	1.03	0.31
GT+Executor	0.94	0.28
Decision-Maker + Executor	0.97	0.27

Table 4: **Ablation Study Results:** This table presents the impact of various components within our DME-Driver system, illustrating how each part contributes to the overall effectiveness and decision-making accuracy.

Ablation Study

In our ablation study of the DME-Driver system, we methodically dissected the impact of decision-making effectiveness, as shown in Table 4. We began by assessing the standalone performance of the Executor without Decision-Maker guidance, establishing a baseline. Next, we evaluated the impact of substituting the Decision-Maker's guidance with ground truth language cues, observing potential improvements. Following this, we examined the combined performance of the Decision-Maker and Executor, gauging their collaborative efficiency.

Conclusion

In addressing the challenges of interpretability and insufficient use of human driver behavior patterns in self-driving systems, this paper introduces the DME-Driver Autonomous Driving System, a novel framework comprising two integral components: the Decision-Maker and the Executor. The Decision-Maker serves as the central decision-maker, understanding and emulating human driver logic, thus ensuring each action is both logical and accountable. The Executor complements this by effectively translating the nuanced decisions of the Decision-Maker into precise vehicle control signals, harnessing the strengths of perception tasks and planning algorithms. To facilitate comprehensive training and understanding of human driver behavior, we developed the HBD dataset, rich in diverse and essential driving information such as gaze, decision logic, and operational signals. Our empirical tests showcase the system's capability to accurately simulate human driver reasoning and actions. Under the Decision-Maker's guidance, the Executor successfully converts these into operational commands, elevating the overall decision-making efficacy to a state-of-the-art level.

Acknowledgements

This work was supported partly by the FDCT grants 0102/2023/RIA2, 0154/2022/A3, 001/2024/SKL, 0121/2024/RIA2, the MYRG-CRG2022-00013-IOTSC-ICI grant, and the SRG2022-00023-IOTSC grant.

References

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-

end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1728–1738.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L. D.; Monfort, M.; Muller, U.; Zhang, J.; et al. 2016. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.; et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

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

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3558–3568.

Chen, L.; Wu, P.; Chitta, K.; Jaeger, B.; Geiger, A.; and Li, H. 2023. End-to-end autonomous driving: Challenges and frontiers. *arXiv preprint arXiv:2306.16927*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, X.; Han, J.; Xu, H.; Liang, X.; Zhang, W.; and Li, X. 2024. Holistic Autonomous Driving Understanding by Bird's-Eye-View Injected Multi-Modal Large Models. *arXiv preprint arXiv:2401.00988*.

Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; and Koltun, V. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*, 1–16. PMLR.

Elhafsi, A.; Sinha, R.; Agia, C.; Schmerling, E.; Nesnas, I. A.; and Pavone, M. 2023. Semantic anomaly detection with large language models. *Autonomous Robots*, 1–21.

Feng, D.; Rosenbaum, L.; and Dietmayer, K. 2018. Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection. In *2018 21st international conference on intelligent transportation systems (ITSC)*, 3266–3273. IEEE.

Feng, K.; Li, C.; Ren, D.; Yuan, Y.; and Wang, G. 2024. On the Road to Portability: Compressing End-to-End Motion Planner for Autonomous Driving. *arXiv preprint arXiv:2403.01238*.

Grigorescu, S.; Trasnea, B.; Cocias, T.; and Macesanu, G. 2020. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3): 362–386.

Gu, J.; Hu, C.; Zhang, T.; Chen, X.; Wang, Y.; Wang, Y.; and Zhao, H. 2023. Vip3d: End-to-end visual trajectory prediction via 3d agent queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5496–5506.

Han, W.; Tao, R.; Ling, H.; and Shen, J. 2024. Weakly Supervised Monocular 3D Object Detection by Spatial-Temporal View Consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hang, P.; Lv, C.; Xing, Y.; Huang, C.; and Hu, Z. 2020. Human-like decision making for autonomous driving: A noncooperative game theoretic approach. *IEEE Transactions on Intelligent Transportation Systems*, 22(4): 2076–2087.

Hu, P.; Huang, A.; Dolan, J.; Held, D.; and Ramanan, D. 2021. Safe local motion planning with self-supervised freespace forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12732–12741.

Hu, S.; Chen, L.; Wu, P.; Li, H.; Yan, J.; and Tao, D. 2022. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, 533–549. Springer.

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17853–17862.

Huang, Z.; Liu, H.; Wu, J.; and Lv, C. 2023. Differentiable integrated motion prediction and planning with learnable cost function for autonomous driving. *IEEE transactions on neural networks and learning systems*.

Isele, D.; Rahimi, R.; Cosgun, A.; Subramanian, K.; and Fujimura, K. 2018. Navigating occluded intersections with autonomous vehicles using deep reinforcement learning. In *2018 IEEE international conference on robotics and automation (ICRA)*, 2034–2039. IEEE.

Kamath, A.; Anderson, P.; Wang, S.; Koh, J. Y.; Ku, A.; Waters, A.; Yang, Y.; Baldridge, J.; and Parekh, Z. 2023. A New Path: Scaling Vision-and-Language Navigation with Synthetic Instructions and Imitation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10813–10823.

Kasahara, I.; Stent, S.; and Park, H. S. 2022. Look both ways: Self-supervising driver gaze estimation and road scene saliency. In *European Conference on Computer Vision*, 126–142. Springer.

Khurana, T.; Hu, P.; Dave, A.; Ziglar, J.; Held, D.; and Ramanan, D. 2022. Differentiable raycasting for self-supervised occupancy forecasting. In *European Conference on Computer Vision*, 353–369. Springer.

Kim, J.; Moon, S.; Rohrbach, A.; Darrell, T.; and Canny, J. 2020. Advisable learning for self-driving vehicles by internalizing observation-to-action rules. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9661–9670.

Li, B.; Wang, Y.; Mao, J.; Ivanovic, B.; Veer, S.; Leung, K.; and Pavone, M. 2024. Driving Everywhere with Large Language Model Policy Adaptation. *arXiv preprint arXiv:2402.05932*.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, 1–18. Springer.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744.

Liu, M.; Jiang, J.; Zhu, C.; and Yin, X.-C. 2023b. VLPD: Context-Aware Pedestrian Detection via Vision-Language Semantic Self-Supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6662–6671.

Ma, Y.; Cao, Y.; Sun, J.; Pavone, M.; and Xiao, C. 2025. Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*, 403–420. Springer.

Ma, Y.; Cui, C.; Cao, X.; Ye, W.; Liu, P.; Lu, J.; Abdelraouf, A.; Gupta, R.; Han, K.; Bera, A.; et al. 2023. Lampilot: An open benchmark dataset for autonomous driving with language model programs. *arXiv preprint arXiv:2312.04372*.

Mao, J.; Qian, Y.; Zhao, H.; and Wang, Y. 2023a. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*.

Mao, J.; Ye, J.; Qian, Y.; Pavone, M.; and Wang, Y. 2023b. A language agent for autonomous driving. *arXiv preprint arXiv:2311.10813*.

Maqueda, A. I.; Loquercio, A.; Gallego, G.; García, N.; and Scaramuzza, D. 2018. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5419–5427.

Montremerlo, M.; Beeker, J.; Bhat, S.; and Dahlkamp, H. 2008. The stanford entry in the urban challenge. *Journal of Field Robotics*, 7(9): 468–492.

Pan, C.; Yaman, B.; Nesti, T.; Mallik, A.; Allievi, A. G.; Velipasalar, S.; and Ren, L. 2024. VLP: Vision Language Planning for Autonomous Driving. *arXiv preprint arXiv:2401.05577*.

Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; Funkhouser, T.; et al. 2023. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 815–824.

Radecki, P.; Campbell, M.; and Matzen, K. 2016. All weather perception: Joint data association, tracking, and classification for autonomous ground vehicles. *arXiv preprint arXiv:1605.02196*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Shao, H.; Wang, L.; Chen, R.; Li, H.; and Liu, Y. 2023. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In *Conference on Robot Learning*, 726–737. PMLR.

Strubell, E.; Ganesh, A.; and McCallum, A. 2019. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243*.

Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2446–2454.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wei, Y.; Wang, Z.; Lu, Y.; Xu, C.; Liu, C.; Zhao, H.; Chen, S.; and Wang, Y. 2024. Editable Scene Simulation for Autonomous Driving via Collaborative LLM-Agents. *arXiv preprint arXiv:2402.05746*.

Wen, L.; Fu, D.; Li, X.; Cai, X.; Ma, T.; Cai, P.; Dou, M.; Shi, B.; He, L.; and Qiao, Y. 2023. Dilu: A knowledge-driven approach to autonomous driving with large language models. *arXiv preprint arXiv:2309.16292*.

Wu, D.; Han, W.; Wang, T.; Dong, X.; Zhang, X.; and Shen, J. 2023a. Referring multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14633–14642.

Wu, D.; Han, W.; Wang, T.; Liu, Y.; Zhang, X.; and Shen, J. 2023b. Language prompt for autonomous driving. *arXiv preprint arXiv:2309.04379*.

Xu, Z.; Zhang, Y.; Xie, E.; Zhao, Z.; Guo, Y.; Wong, K. K.; Li, Z.; and Zhao, H. 2023. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint arXiv:2310.01412*.

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1).

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zeng, W.; Luo, W.; Suo, S.; Sadat, A.; Yang, B.; Casas, S.; and Urtasun, R. 2019. End-to-end interpretable neural motion planner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8660–8669.