2024 IEEE International Conference on Mobility, Operations, Services and Technologies (MOST)

An Advanced Driving Agent with the Multimodal Large Language Model for Autonomous Vehicles

Junzhou Chen
Dept. of Computer Science
William & Mary
Williamsburg, USA
jchen57@wm.edu

Sidi Lu
Dept. of Computer Science
William & Mary
Williamsburg, USA
sidi@wm.edu

***Abstract*—As deep learning technology advances Autonomous Driving (AD), existing AD methods encounter performance limitations, especially in handling corner cases, interpretability, and verifiability, which are crucial for the safety of connected and autonomous vehicles. Multimodal Large Language Models (MLLMs) demonstrate remarkable understanding and reasoning capabilities, presenting a transformative opportunity to overcome challenges faced by traditional AD algorithms. We conduct a comprehensive study on the application of MLLMs in AD, exploring their potential to address critical challenges faced by traditional AD algorithms. We construct a Visual-Question-Answering dataset for model fine-tuning to address hallucinations and poor logic analysis issues in MLLMs. We then decompose the AD decision-making process into Scene Understanding, Prediction, and Decision, allowing MLLMs to construct Chain-of-Thought to make decisions step by step. Subsequently, we propose a new framework enabling models to perform AD tasks under conditions of limited local computing resources, few-shots, multimodality, and complex scenarios, enhancing the flexibility of future AD system deployment. Our extensive experiments and in-depth analyses demonstrate the significant advantages of MLLMs for AD. We also discuss the strengths and weaknesses of existing methods, providing a detailed outlook on MLLMs in AD.**

***Index Terms*—Autonomous driving, Multimodal Large Language Model, Chain-of-Thought**

I. Introduction

Recently, with the vast improvements in computing technologies (e.g., sensors, deep learning) and the broad deployment of communication mechanisms (e.g., 5G/6G, DSRC [1], C-V2X [2]), the advent of Tesla's Autopilot [3], Google's Waymo [4], and Baidu's Apollo [5] has propelled autonomous vehicles (AVs) to into prominence. However, it is generally accepted that the existing deep learning-based scenario understanding, prediction, and decision-making technologies for AVs heavily rely on existing data. In the context of corner cases, which may include irregular behaviors of road users, unexpected obstacles, adverse weather conditions, and complex traffic accident scenes [6], deep learning-based AV systems often fail to respond appropriately [7]–[9].

Previous work generally divided the AV's functions into three stages: scene understanding, prediction, and decision [11]. These stages are interrelated, where the results of scene understanding (e.g., identifying vehicles and pedestrians) are normally used for future behavior predictions. Then the predictions will facilitate path planning and decision-making. However, these stages both face challenges due to corner cases, and errors in any stage can lead to the failure of the overall tasks. For instance, as shown in Fig 1(a), the model may erroneously identify reflections on a waterlogged road surface as false artifacts. This issue arises in scenarios including, but not limited to, extreme weather conditions, uncommon road obstructions, rare road signs, and obscured traffic signals. Regarding prediction, since the predictive model is primarily trained on historical data, it may struggle to accurately forecast rare traffic events or emergencies. In Fig 1(b), traditional methods often fail to make correct predictions when road markings do not align with the actual traffic conditions due to construction work. In terms of planning, while plotting vehicle routes is relatively straightforward under normal conditions, it becomes more challenging during emergencies or unusual traffic situations. As depicted in Fig 1(c), when confronted

Fig. 1: Scenarios, where AVs fail to accurately analyze, predict, and make decisions, include (a) Road surfaces with standing water [10] which may incorrect positioning of the AV in relation to the image of the water surface. (b) The construction intersection where AVs may fail to predict accurately due to overly complex and unfamiliar scenes. (c) Snow-covered roads where AVs struggle to make correct decisions due to lack of information.

979-8-3503-0773-3/24/$31.00 ©2024 IEEE
DOI 10.1109/MOST60774.2024.00009

with heavy snow accumulation, the planner needs to exhibit enhanced adaptability.

In contrast, human drivers can recognize anomalies, foresee future events, and make decisions before these events occur. The core of this proactive decision-making ability lies in reasoning and extrapolating common-sense knowledge from new data. Recently, Large Language Models (LLMs) like GPT-3.5 [12], PaLM [13], and Llama [14] have gained attention for their memory and reasoning capabilities, which are similar to the human brain. These models exhibit information analysis and processing performance that is comparable or even superior to humans [15]. They also show encouraging signs in deepening the understanding of complex driving scenarios [16]. In autonomous driving, multimodal inputs such as visual, radar, and sensor data are often needed [17]. However, LLMs are limited to processing text data, which restricts their ability to handle other multimodal information like visual inputs. With the continuous introduction of Multimodal Large Language Models (MLLMs) [18]–[20], pre-trained vision-language models like CLIP [21] and CogVLM [22] demonstrate strong zero-shot generalization performance across various downstream visual tasks. This enables LLMs to effectively understand multiple data sources, including images, videos, and point clouds. The emergence of MLLMs presents a potentially transformative opportunity to overcome the challenges faced by traditional AD algorithms.

In this work, we explored the potential of MLLMs as the core of AD decision-making, particularly for corner cases. We also investigated methods for offline deployment of MLLMs in AVs. Firstly, we proposed an innovative framework to effectively integrate MLLMs into a computational platform with capabilities akin to those of an Autonomous Vehicle, thereby creating an AD Agent. We constructed a Visual-Question-Answer (VQA) dataset and fine-tuned the CogVLM model on this dataset to focus more on AD tasks. To mitigate the effects of illusions by MLLMs and enhance interpretability, we designed a step-by-step Chain-of-Thought (CoT). The CoT divides the decision-making process into scene understanding, prediction, and decision phases, each undergoing detailed experimental validation. This approach allowed us to more meticulously assess the potential value and practicality of MLLMs in various areas within the field of AD.

Specifically, our contributions are delineated as follows:

- We innovatively proposed, implemented, and evaluated a novel MLLM-based framework that is able to execute AD tasks under limited local computing resources, few-shots, multi-modality, and complex scenarios. This attempt offers new insights and possibilities for the future deployment of more flexible AV systems.
- We devised a scheme to optimize generic MLLMs as AV Agents by building a VQA dataset and designing a reasoning chain, effectively reducing model illusions and enhancing focus.
- Regarding the application of MLLMs in scene understanding, analysis, and decision-making, we conducted experimental validations on each of the three stages using a real-world dataset. We then performed comprehensive experiments in a simulated environment using highway-env [23], which clearly demonstrated the significant performance advantages of the MLLM-driven AV systems. We also discussed future directions and potential methods for further improvements.

The rest of this paper is organized as follows: Sec. II reviews related work of AV algorithms, and experimental design is detailed in Sec. IV. Extensive experimental results and our dicussions are shown in Sec. V. Sec. VI concludes the entire paper.

II. Background and Related Work

A. Traditional Autonomous Driving Technique

The levels of driving automation, as defined by the Society of Automotive Engineers [24], range from Level 0 (no automation) to Level 5 (full automation). With increasing autonomy, the need for human intervention decreases, while the requirement for the vehicle to understand its surrounding environment increases. Existing AD solutions are broadly categorized into the classic modular paradigm and end-to-end approaches [25]–[29]. The modular approach breaks down the AV task into subtasks, each executed in separate modules. This design offers advantages like modularity and functional generality, but it comes with challenges related to tuning the pipeline and managing error propagation. UniAD [8] divides the end-to-end process into three stages, including perception, prediction, and planning, and proposes a planning-oriented pipeline. These methods are usually easier to develop, but they lack interpretability, posing challenges in diagnosing errors, ensuring safety, and incorporating traffic rules. Additionally, these automated systems still fail in many driving corner cases, such as extreme weather, poor lighting conditions, or rare situations [30].

Inspired by current limitations, some research in AD is now focusing on addressing the safety of autonomous systems and enhancing their reliability [31]. Since deep neural networks are often considered black boxes, trustworthy AI aims to make systems more reliable, interpretable, and verifiable. For instance, generating adversarial safety-critical scenarios for training AD systems to better handle corner cases [32], [33]. Another approach to enhance overall safety is through vehicle-to-infrastructure and vehicle-to-vehicle communication. By leveraging information from nearby instances, the system's robustness is improved, allowing for early warnings and enhanced situational awareness [34].

B. Advancements in Multimodal Large Language Models

Compared to the limited comprehension abilities of traditional AD systems, the rapidly evolving LLMs have demonstrated a significantly enhanced capacity for understanding. LLMs refer to Transformer language models containing hundreds of billions (or more) of parameters [13], [14], [35]. They have shown capabilities in understanding natural language and solving complex tasks through text generation. LLMs can complete a series of complex tasks by leveraging prompts containing intermediate reasoning steps. There is a growing field of research utilizing LLMs to develop autonomous agents with human-like abilities [36], [37]. These agents use extensive knowledge stored in pre-trained LLMs to create coherent action plans and executable strategies. Despite LLMs' surprising zero/few-shot reasoning performance in most natural language processing (NLP) tasks, they are inherently 'blind' to visual information as they can only comprehend discrete text.

Consequently, many works have explored extending the success of instruction tuning in LLMs to multimodal development, resulting in the creation of MLLMs [38], [39]. BLIP-2 [40] links a trainable pre-trained visual encoder with an LLM through shallow alignment, mapping image information into the feature space. However, such methods exhibit poor visual understanding. Joint image-text training allows for a deep integration of image and language information. CogVLM [22] adds a trainable visual expert to the language model, not only deeply integrating image information but also effectively preserving textual information. Despite these advances, such MLLMs still face the problem of illusions [41], although this can be mitigated to some extent by increasing the number of parameters and appropriately fine-tuned them. In Sec III, we introduce a series of methods that effectively mitigate the issues above.

C. Multimodal Large Language Modals for Autonomous Driving

Current research on AD with LLMs or MLLMs is still in its exploratory phase. Existing studies primarily leverage large language models to address tasks related to AD, such as perception, reasoning, and planning [42]. LanguageMPC [43] developed an algorithm that translates LLM decisions into actionable driving commands through contextual learning using ChatGPT-3.5. DriveLikeHuman [37] delved into the process of mimicking human posture in driving, driven by knowledge. Building on this, Dilu [44] further improved the memory aspect, enabling it to continuously summarize and generalize driving information. To enable LLMs to understand multimodal information, DriveLLM [45] established a multimodal encoder that converts multimodal data into strings for LLM decision-making. However, all these methods face the challenge of information loss due to textual transmission.

GPT-4 has demonstrated a formidable capability for understanding and analyzing visual information [16], [45]. However, due to its reliance on online computational resources and lack of open-source availability, it is not suitable for integration into AV systems. DriveGPT-4 [46] proposed a multimodal model for an interpretable AV system, which takes multimodal inputs and outputs control signals. This method embeds multimodal information into the shallow features of the LLM, resulting in poor comprehension of visual information due to a scarcity of extensive visual priors. Dolphins [47] employed an Image-text jointly trained drive agent, capable of analyzing continuous image information during driving through dialogues. Nevertheless, it still experiences illusions and is unable to execute real decisions. In contrast, our approach further reduces the likelihood of illusions in the model and represents an end-to-end autonomous driving system capable of executing decisions.

III. Experiment Methodology

We have developed a closed-loop autonomous driving control system with an MLLM as the AV Agent, detailed in Sec III-A. In Sec III-B, we introduce a context-based step-by-step thought chain to reduce the occurrence of illusions and enhance the interpretability of the results. To further focus the model on the task, we established a Visual Quality Assessment (VQA) dataset, as discussed in Sec III-C. Finally, the fine-tuning and deployment of the model are presented in Sec III-D.

A. Framework Design

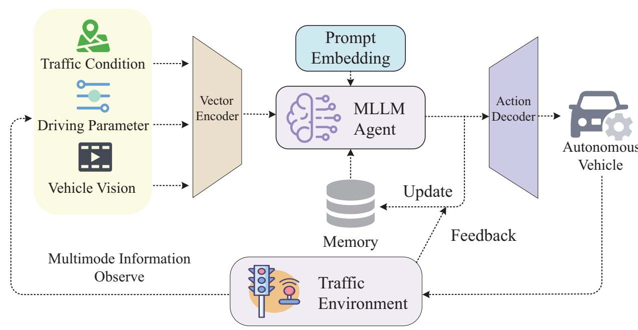

Fig. 2: An end-to-end autonomous driving framework with a Multimodal Large Language Model (MLLM) at its core. The MLLM integrates multimodal information with memory and prompts to make decisions, ultimately outputting action commands.

As discussed in Sec II-C, existing methods based on LLM encode multimodal information into character information that can be processed by the LLM, often resulting in significant information loss. To better harness the potential of MLLM as an AD Agent, we designed a more rational and easily expandable Framework for AD control decisions.

As illustrated in Fig 2 and Alg 1, we have designed and implemented an efficient and comprehensive AD decision system centered around a MLLM. Initially, we transmit task objectives, local traffic rules, and other relevant information as prompts, along with a memory database, to initialize the MLLM. This approach enables the integration of historical experience and real-time information while analyzing encoded data. In this way, the MLLM agent can make more accurate and context-adaptive decisions.

After initializing the MLLM, it begins to collect multimodal environmental information, including but not limited to traffic conditions, driving parameters, vehicle vision, sensor data, etc. We use an Encoder to transform this multimodal information into vector-style strings or image matrices that the MLLM can process. Subsequently, the MLLM receives this encoded vector and combines predefined rules and contextual data to make decisions following a thought chain approach. Once

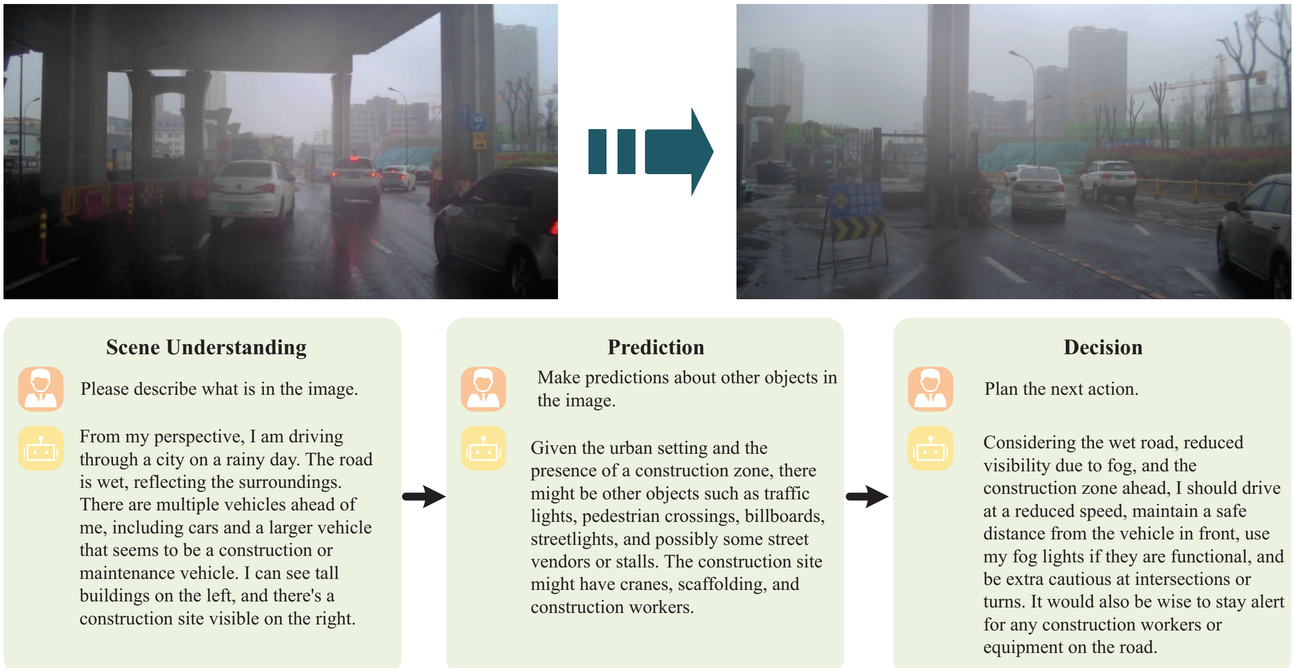

Fig. 3: Step-by-Step Thinking Chain, enabling the model to generate information progressively, resulting in more interpretable outcomes.

Algorithm 1 Automated Driving Decision-Making Process

Input: Vehicle Vision (VV), Driving Parameters (DP), Traffic Conditions (TC), Prompt, Previous Memory (Mem_old)

```
sensor_system.activate()
encoder_config.set_parameters()
MLLM_Agent.initialize(Prompt, Mem_old)
collected_data
        ← sensor_system.collect_data(VV, DP, TC, ...)
encoded_data ← vector_encoder.encode(collected_data)
decision
    ← MLLM_agent.analyze(encoded_data, Mem_old)
vehicle_command ← action_decoder.decode(decision)
vehicle_system.execute(vehicle_command)
performance ← evaluate(decision, vehicle_command)
MLLM_agent.adjust_model(performance)
environmental_impact
        ← analyze_feedback(vehicle_command)
MLLM_agent.update_model
        (environmental_impact, Mem_old)
Mem_new
        ← memory_database.store_trip_data()
MLLM_agent.iterate_model()
```

Output: Improved Decision and Action Patterns, Updated Memory (Mem_new)

a decision is made, an action decoder converts the high-level decision of the MLLM agent into executable commands. In actual execution, the system continuously optimizes the MLLM model through performance evaluation and environmental feedback analysis. This includes adjusting the model based on the results of decision-making and vehicle command execution and updating the model according to the impact of vehicle behavior on the environment. Meanwhile, new memories (Mem_new) are stored, including collected data, decisions made, and performance evaluations, providing a richer context for future decisions. This design not only enhances the accuracy of decisions and the efficiency of action patterns but also ensures that the system continuously learns and adapts over time, demonstrating a high degree of dynamic adaptability and potential for ongoing improvement.

B. Chain-of-Thought Design

Although the fine-tuned model shows more proficiency in AD tasks, it sometimes still experiences illusions. The primary reason is the large gap and weak logical connection between the data source and the task objective. Chain-of-Thought (CoT) reasoning, which generates and infers the thought process within the model's contextual window, can help the model exhibit intelligence and logical reasoning abilities. A complete CoT-inclusive prompt typically consists of three parts: Instructions, Rationale, and Exemplars. This approach significantly enhances the model's logical reasoning capabilities and the interpretability and credibility of the results through a step-by-step method.

We divided the model's reasoning process into three steps: scene understanding, prediction, and decision. First, we embedded the rationale and exemplars into the VQA dataset for the model to learn. Subsequently, during the reasoning process, we posed three questions, as shown in Fig 3, prompting the model to propose its final decision. This significantly improved the accuracy and interpretability of the decisions.

C. Visual-Question-Answer Dataset Build

Benefiting from the vast prior knowledge inherent in MLLMs, these models usually can achieve few-shot or even zero-shot learning tasks. However, traditional open-source models, in pursuit of more versatile performance, often compromise their capabilities in specific domains to enhance their generalizability. To make the model more focused on AD tasks and further address the illusion problem inherent in the model, we fine-tuned it. Due to the severe scarcity of relevant datasets, we selected numerous typical cases from BDD100k [48] to construct a VQA dataset.

Inspired by Segment Anything [49], we initially selected a few images for manual annotation. Subsequently, we employed ChatGPT-4 to annotate subsequent images automatically through context-based learning. We then manually screened and optimized the annotated content, prompting ChatGPT-4 to re-annotate any samples that did not meet standards. After four iterations, we annotated 100 cases in a VQA dataset. The dataset, exemplified by Fig 4, primarily consists of three parts: Visual, Question, and Answer. The quality of the dataset construction is crucial for enabling the model to understand the task process better and perform more professionally. Through three-part question-and-answer data cases, we established a CoT that breaks down the AD task into subtasks to be resolved step by step. To ensure that the action decoder can correctly decode the model's output, we specified the output format in the third part of the Answer, adapting it to meet the requirements of the action decoder.

D. MLLM Fine-tuning and Deployment

As mentioned in section II-B, existing MLLMs generally suffer from poor visual understanding and illusions. We chose the recently open-sourced MLLM, CogVLM, as our baseline. Specifically, CogVLM first uses a pre-trained Vision Transformer (ViT) and an MLP Adapter to extract visual and textual information into the same feature space. It then introduces a Visual expert module, as shown in Fig 5, replacing the transformer module in each layer. The visual expert module extends a new QKV matrix and FFN [50] based on the transformer to process visual feature information, freezing all other parameters and training only the newly expanded parts and the MLP Adapter. This integration allows CogVLM to merge image and text information effectively in the deep feature space of a large model, enhancing the model's visual understanding while retaining its language comprehension ability.

First, we fine-tune the CogVLM with the VQA dataset constructed in Sec III-C. To realize the framework described in Sec III-A, we then manage the MLLM using the LangChain [51] framework. To deploy CogVLM on a computing platform matching the computational power of AVs, we utilized a model with 8-bit quantization to reduce computational demands. Furthermore, to enhance the model's scalability in later stages, we provided detailed interface documentation for each module, significantly improving the model's replaceability and the framework's adaptability.

Fig. 4: Examples of the visual-question-answer (VQA) dataset.

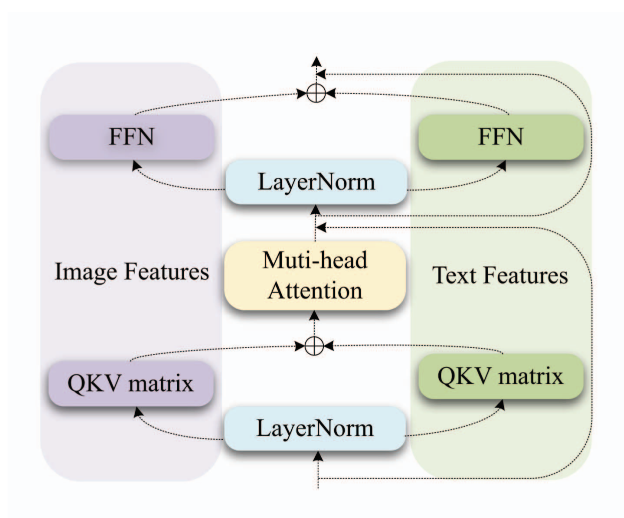

Fig. 5: Visual expert module based on transformer, the image features have a different QKV matrix and FFN, which allow CogVLM to gain a deeper visual understanding

Scene Understanding

Prompt:

This is a picture taken with a car camera. Describe what you see, especially traffic participants, road topology, and traffic lights and traffic conditions (if applicable).

CogVLM:

The image appears to be taken from the perspective of a car's rearview camera, capturing a view of a city street. On the left side of the road, there's a man riding a small truck or cart with goods. There are multiple cars parked on both sides of the road, and a few are moving. Traffic lights can be seen in the distance, and the road seems to be a two-way street with clear markings. The road is flanked by buildings, trees, and some street furniture.

Prompt:

This is a picture taken with a car camera. Describe what you see, especially traffic participants, road topology, and traffic lights and traffic conditions (if applicable).

CogVLM:

The image appears to be taken from the perspective of a car camera, capturing a traffic scene during nighttime. There are multiple vehicles on the road, including cars, trucks, and a motorcycle. There are road cones set up, indicating some kind of roadwork or obstruction. The road seems to be multi-lane, with an overpass visible in the background. Traffic lights are not clearly visible, but the overall traffic seems to be slow-moving due to the cones and the presence of multiple vehicles.

Fig. 6: Implement the fine-tuned CogVLM for traffic environments description. Purple font indicates correctly understood answers, while red fonts denote misinterpreted answers. Result demonstrates a remarkable capability for visual comprehension and description, yet it is susceptible to occasional inaccuracies in judgment due to influences from uncommon surrounding environments.

IV. Experiment Design

To gain a more comprehensive understanding of MLLM's capabilities in autonomous driving, we conducted multiple experiments focusing on the proposed method. First, we detail the specific configurations and implementation details of the experiments in Sec IV-A. In Sec IV-B- IV-D, we conduct experiments specifically targeting the three aspects of MLLM in scene understanding, prediction, and decision. Finally, in Sec IV-E, we carry out simulation experiments to compare our approach with traditional mainstream methods, providing further evaluation of our method.

A. Experimental Detail

VQA and Test Datasets. For the VQA dataset, we selected and annotated 100 images from BDD100k [48], with specific details presented in Sec III-C. Subsequently, from BDD100k, excluding the content of the VQA Dataset, we chose some complex scenarios to test the performance of MLLM. From these experiments, we selected typical cases for display in Sec IV-B- IV-D.

Model Finetuning and Experiments. We chose CogVLM [22] as our baseline and fine-tuned it using the VQA dataset on 4 $\times$ A100 GPUs. The fine-tuned CogVLM was then deployed for inference on two GeForce RTX 3090 GPUs. Initially, to further understand and test the MLLMs' capabilities in comprehending and reasoning about autonomous driving tasks, we used CogVLM to infer typical images from scene understanding, prediction, and decision, outputting interpretable text. We assessed model performance through semantic analysis of the output content. Subsequently, to test the Framework's potential for executing complete Autonomous Driving tasks, we deployed a fully autonomous driving task on Highway-Env [23], comparing and analyzing it against traditional mainstream autonomous driving methods.

B. Scene understanding based on MLLM

We conducted a series of experiments focusing on the Multi-Modal Language Models' (MLLMs) ability to scene and understand multimodal information. A classic case from these experiments is depicted in Fig 6, which comprehensively reflects the current model's capacity to comprehend visual information. In Fig 6, we demonstrate CogVLM's scene and understanding based on road images captured by onboard cameras. The model performed commendably in both instances.

The CogVLM accurately identified both dual and single carriageways, made precise assessments of the current traffic conditions, and correctly recognized the cars and pedestrians included in the images. However, it is noteworthy that in the left example of Fig 6, due to the abnormal parking of a red sedan, the model made an incorrect judgment regarding the orientation of the vehicle. In the second case, the overly complex traffic situation led to a "illusion" in the model, where it erroneously perceived the presence of a motorcycle on the road. These instances not only exhibit the model's strengths in accurately interpreting and reacting to standard

Prediction

Prompt:

Now that you are a driver, predict the movement of people and vehicles in the picture.

CogVLM:

From the picture, we can see that the traffic lights in the middle of the intersection are currently green, indicating that vehicles can pass through the intersection. In the picture, there are two vehicles waiting at the red light, while several other vehicles are passing through the intersection. On the right side of the picture, there are pedestrians walking on the pedestrian crossing, waiting to cross the road. Therefore, it can be predicted that after the vehicles in the picture pass through the intersection, there will be pedestrians crossing the road, and then vehicles waiting at the red light. After these vehicles pass through, there will be more vehicles waiting to enter the intersection.

Prompt:

Now that you are a driver, predict the movement of people and vehicles in the picture.

CogVLM:

From the picture, it can be seen that there are many vehicles driving on the road, and the traffic is relatively dense. In addition, there are many pedestrians on the roadside. Considering the rainy weather, it is expected that the pedestrians will try to avoid getting wet and may choose to stay on the sidewalks or use the pedestrian overpasses. The vehicles in the picture may also reduce their speed to ensure safety, and there may be more braking actions. Therefore, the driver should pay close attention to the surrounding vehicles and pedestrians to ensure safety.

Fig. 7: Predictions of traffic environments were made using the fine-tuned CogVLM. Purple font indicates correctly understood answers, while red fonts indicate misinterpreted answers. This demonstrates the model's predictive capabilities in scenarios of forked road intersections and rainy weather conditions.

driving environments but also highlight its current limitations in processing complex or abnormal scenarios. Such findings underline the necessity for further refinement in the model's ability to differentiate between standard and atypical traffic situations, enhancing its overall perception and decision-making accuracy in diverse driving conditions.

C. Prediction based on MLLM

In this experiment, we specifically selected scenarios involving forked roads and impaired vision due to rain for the cameras to test the predictive capabilities of our model, as illustrated in Fig 7. For these tests, the model utilized was a fine-tuned version of CogVLM. Across these scenarios, the model demonstrated robust logical reasoning abilities, effectively integrating and interpreting environmental cues for accurate prediction.

In the first scenario of Fig 7, the model initially made a correct identification of the traffic lights, deducing that the vehicle was in motion while pedestrians were waiting at the intersection. However, due to the white SUV on the right side of Case 1 slowing down to maintain distance from the vehicle ahead, indicated by its brake lights, the model inaccurately assumed that the SUV was stopping for the traffic light. In the second scenario, the model successfully recognized the severely blurred pedestrian in the image. Based on the weather conditions, it was deduced that the pedestrian would likely walk along the road to seek shelter from the rain. However, it mistakenly predicted the presence of pedestrian overpasses. Additionally, the model, taking into account factors like weather and traffic density, correctly anticipated potential braking actions from surrounding vehicles. These cases collectively showcase the potent logical reasoning capabilities of MLLMs as AD Agents. They not only exhibit the model's proficiency in dealing with complex and variable real-world driving conditions but also highlight areas where further refinement is needed, particularly in interpreting more nuanced and less predictable aspects of human and vehicular behavior.

D. Decision based on MLLM

In this experiment, we explored the role and necessity of CoT in driving decision-making by comparing the outcomes of CogVLM with and without the utilization of CoT. In the left panel of Fig 8, the model was guided using the thought process described in Sec III-B. This approach allowed the model to effectively integrate and process the information deduced from the thought chain, accurately identifying the vehicle's lane and making the correct driving decisions. This demonstrates the significant role of CoT in aiding the model to assimilate and handle complex information.

In contrast, in the right panel of Fig 8, the model was tasked to generate decisions directly without the assistance of CoT. The results indicated a marked lack of logic in the model's performance due to the broad gap between task requirements and textual information. In the absence of CoT, the model failed to correctly identify the vehicle's lane and erroneously

Fig. 8: The left case example shows CogVLM making decisions about subsequent vehicle actions using CoT, while the right case example is without using CoT. Purple font indicates correctly understood answers, and red fonts indicate misinterpreted answers.

interpreted the traffic signals, leading to incorrect driving decisions. This outcome further underscores the importance of CoT in enhancing the model's ability to comprehend complex scenarios and make rational decisions.

The juxtaposition of these two cases clearly illustrates the value of CoT in improving the quality of decisions in complex situations. CoT not only enhances the accuracy of the model's decisions but also increases the interpretability of the decision-making process, which is crucial for the safety and reliability of autonomous driving systems. Therefore, our experimental results robustly validate the necessity and effectiveness of employing CoT in the decision-making process of autonomous driving.

E. Simulation Experiment

We compared the effectiveness of our proposed method with Reinforcement Learning-based Planning (RL) and Model Predictive Control (MPC) in the Highway-env simulation environment [23], encompassing scenarios like highway, roundabout, and intersection. In these environments, we designed multiple complex real-world situations to assess the overall capability of the system. In each of the three environments, we created 50 different traffic scenarios for extensive testing. We recorded the failure probability, the probability of ineffective decisions, and the average time to complete tasks, with results presented in Table I. Subsequently, we selected some typical experimental results for display and analysis in Fig 9.

Compared to MPC and RL, which require a foundational environment for training, our method only needs context-based learning to guide the model to produce the specified output format, allowing it to complete tasks impressively. In the experiments, the MLLM-based autonomous driving system, by evaluating surrounding vehicles, distances, and traffic conditions, used a thought chain for deep logical reasoning to determine the safest and most effective overtaking path. As shown in Table 1, our method outperforms others in terms of fewer failures and ineffective actions. While RL showed the best speed in Roundabout and Highway scenarios, it also led to a high number of failures. Our method accurately predicted the surrounding driving conditions at the no-signal intersection, thereby saving driving time more effectively than other methods.

TABLE I: The data in the table are all metrics where lower values indicate better performance. "Fail" represents the percentage of task execution failures, "Inefficiency" indicates the percentage of ineffective actions, and "Average Time" denotes the time taken to complete the task under normal decision speeds, excluding decision-making time.

Scenario	Method	Fail	Inefficiency	Average Time
Intersection	RL	10.0%	6.0%	3.9s
	MPC	4.0%	6.0%	4.0s
	Ours	**0.0%**	**2.0%**	**3.8s**
Roundabout	RL	10.0%	8.0%	**4.8s**
	MPC	6.0%	4.0%	5.2s
	Ours	**0.0%**	**2.0%**	5.1s
Highway	RL	12.0%	8.0%	**18.2s**
	MPC	6.0%	4.0%	19.1s
	Ours	**0.0%**	**2.0%**	22.3s

We also selected some classic cases for display in Fig 9, providing a detailed comparison of our method against traditional methods. The MLLM-based autonomous driving method better predicts the surrounding environment. For example, in the intersection scenario, our method correctly predicted the distance of oncoming traffic, accelerating in advance to merge into the lane. In the highway scenario, our method preemp-

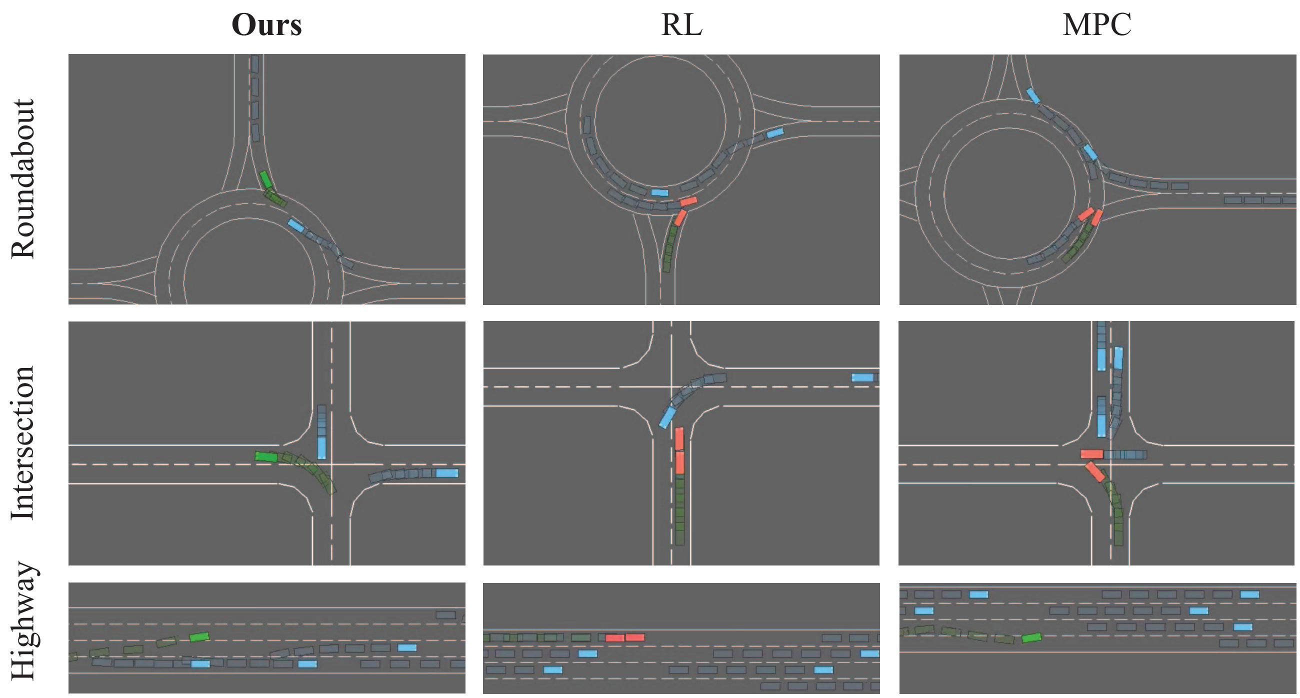

Fig. 9: Our method compared with RL and MPC in three scenarios: roundabout, intersection, and highway. In these examples, RL, motivated by time incentives, led to overly aggressive driving and resulted in rear-end collisions. MPC, on the other hand, was unable to predict many uncommon situations, leading to task failures.

tively chose a lane without vehicles for lane changing. In contrast, RL, driven by the incentive of driving time, led to collisions in its haste to reach the target. MPC unable to generalize some uncommon situations, like sudden lane changes in the roundabout scenario or slowing down of the front vehicle to yield in the intersection scenario, resulting in collisions. In the highway scenario, MPC made many unnecessary actions due to premature predictions.

V. Discussions

A. Capabilities of MLLM in Autonomous Driving

Integrating MLLMs into Autonomous Vehicles is a promising research direction. Large Language Models have demonstrated strong capabilities in information storage and logical reasoning, essential for the future development of artificial intelligence. In this paper, we conducted comprehensive experiments and explorations of CogVLM in autonomous driving scenarios. The results show that CogVLM surpasses existing autonomous driving capabilities in areas like scene understanding, prediction, and decision.

In dealing with corner cases, CogVLM leverages its advanced understanding abilities to accurately assess and predict the surrounding environment. The model also possesses robust reasoning capabilities, enabling it to infer future movement trends of objects based on their potential motives, a feature lacking in traditional methods. As demonstrated in Section IV-B, the model can judge pedestrian movement tendencies by combining surrounding traffic conditions and pedestrian actions.

The superiority of MLLMs-based AD systems in terms of interpretability and generalization capabilities stands out in comparison to other existing models. Traditional AD algorithms are predominantly black-box models, which means they suffer from a lack of transparency in their internal processes. This not only makes it challenging to understand and interpret the decision-making process of the algorithms but also often limits their adaptability in unknown or changing environments. However, MLLMs, such as CogVLM, demonstrate distinct characteristics.

In experiments, the CogVLM model has shown remarkable proficiency in completing complex tasks with just minimal fine-tuning using a small dataset. This rapid adaptation to new tasks highlights its exceptional generalization capabilities. More importantly, CogVLM provides clear logic and reasoning in every step of its decision-making process. This interpretability is crucial for AD systems, as it not only enhances the trust of developers and users in the system but also offers valuable insights for ongoing improvement of the system.

Furthermore, the transparency and adaptability of MLLMs enhance their reliability in dealing with complex real-world scenarios. Compared to traditional AD algorithms, MLLMs-based systems are better at understanding and processing linguistic information, which is vital for interaction and decision-making in AD scenarios. Therefore, this type of system is likely to play an increasingly significant role in the future development of AD technologies.

B. Limitation of MLLM in Autonomous Driving

However, in our experiments, we identified several shortcomings of MLLMs in AD and proposed some potential improvements.

Model Inference Speed: AD systems often require decision-making inferences to be made in an extremely short time. However, AD systems based on MLLMs often take several seconds or even tens of seconds to produce the correct inference, which is intolerable. For Autonomous Vehicles, text upsampling is often unnecessary. The text output process of MLLMs not only involves significant computational power loss but also information loss. Therefore, if MLLMs are to be deployed in autonomous vehicles, it might be worth considering replacing the upsampling layer of MLLMs to make them direct output vectors with control instructions.

MLLMs Illusion: Although in this paper, we have been trying to solve the illusion problem of the model through finetuning and contextual learning, as shown in Section IV-A, the model still produces erroneous outputs. The issue of model illusion is a challenging problem in the field of NLP. In the future, this could be mitigated by designing effective risk control models and self-checking structures to keep the harm caused by illusions within an acceptable range.

Spatial Continuity Reasoning: In the technology of autonomous vehicles, the capability for temporal linear reasoning of three-dimensional information is crucial, a feature currently absent in MLLMs. Autonomous driving necessitates an accurate understanding of the surrounding environment, including spatial location, object motion, and other three-dimensional information, as well as the ability to predict the changes of these factors over time. Existing MLLMs, however, are primarily adept at processing one-dimensional text and two-dimensional image data, falling short in addressing the complexities of three-dimensional spatial data.

Lack of Relevant Datasets: Although there are many large-scale autonomous driving datasets available [52], [53], they are not directly suitable for adaptation in AD with LLMs. For instance, how to generate datasets for instruction tuning and designing instruction formats based on AD datasets are still under-explored. The generation of a suitable instruction tuning dataset for MLLMs based on autonomous driving datasets is an area that remains underexplored. This involves the understanding and transformation of non-textual information in AD datasets into a format that can be processed by LLMs, while maintaining the authenticity and practicality of the data. Additionally, relevant datasets could also aid a series of downstream tasks dependent on VLMs. For instance, in autonomous driving, a Vision-Language Model (VLM) can aid in enhancing scene understanding, decision-making, and predicting the behaviors of other vehicles and pedestrians.

VI. Conclusion

This paper demonstrates that MLLMs can effectively serve as decision-making agents in AD systems. We introduced a new framework deployed on platforms with computing power comparable to Autonomous Vehicles. Focusing on the performance of MLLMs in scene understanding, prediction, and decision, we conducted extensive experiments in multimodal, few-shot, and complex scenarios, clearly displaying the significant advantages of AD systems enhanced by MLLMs. Furthermore, this paper delves into the strengths and weaknesses of existing methods and proposes detailed prospects and methods for future improvements. Our approach provides the first step towards developing safe, few-shot, locally deployable, and interpretable MLLM-based AD systems. We aspire for it to serve as an inspiration for future research in this field.

Acknowledgement

This work is supported in part by the National Science Foundation (NSF) grant CNS-2348151 and Commonwealth Cyber Initiative grant HC-3Q24-048.

References

[1] E. Moradi-Pari, D. Tian, M. Bahramgiri, S. Rajab, and S. Bai, "DSRC Versus LTE-V2X: Empirical performance analysis of direct vehicular communication technologies," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 5, pp. 4889–4903, 2023.

[2] V. Mannoni, V. Berg, S. Sesia, and E. Perraud, "A comparison of the V2X communication systems: ITS-G5 and C-V2X," in *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring)*. IEEE, 2019, pp. 1–5.

[3] A. Lyyra, K. Koskinen, C. Sørensen, and T. Marion, "Tethered architectures in cyber-physical system development: The case of Tesla's Autopilot system," *Available at SSRN 4515061*, 2023.

[4] R. Deemantha and B. Hettige, "Autonomous car: Current issues, challenges and solution: A review."

[5] S. Lu and W. Shi, "Vehicle as a mobile computing platform: Opportunities and challenges," *IEEE Network*, 2023.

[6] W. Zhou, Z. Cao, N. Deng, X. Liu, K. Jiang, and D. Yang, "Dynamically conservative self-driving planner for long-tail cases," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp. 3476–3488, 2022.

[7] Y. Huang and Y. Chen, "Autonomous driving with deep learning: A survey of state-of-art technologies," *arXiv preprint arXiv:2006.06091*, 2020.

[8] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 853–17 862.

[9] L. Chen, Y. Li, C. Huang, B. Li, Y. Xing, D. Tian, L. Li, Z. Hu, X. Na, Z. Li *et al.*, "Milestones in autonomous driving and intelligent vehicles: Survey of surveys," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1046–1056, 2022.

[10] H. Angeri. (2018) Mirror, mirror on the wall... ai (deep learning) needs to learn it all. [Online]. Available: https://harshaangeri.medium.com/mirror-mirror-on-the-wall-ai-deep-learning-needs-to-learn-it-all_-ad221bd399e1

[11] L. Chen, Y. Li, C. Huang, B. Li, Y. Xing, D. Tian, L. Li, Z. Hu, X. Na, Z. Li *et al.*, "Milestones in autonomous driving and intelligent vehicles: Survey of surveys," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1046–1056, 2022.

[12] OpenAI, "Chatgpt," *https://chat.openai.com*, 2023.

[13] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[14] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[15] J. Wang, Y. Wang, G. Xu, J. Zhang, Y. Gu, H. Jia, M. Yan, J. Zhang, and J. Sang, "An llm-free multi-dimensional benchmark for mllms hallucination evaluation," *arXiv preprint arXiv:2311.07397*, 2023.

[16] L. Wen, X. Yang, D. Fu, X. Wang, P. Cai, X. Li, T. Ma, Y. Li, L. Xu, D. Shang *et al.*, "On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving," *arXiv preprint arXiv:2311.05332*, 2023.

[17] J. Wang, J. Liu, and N. Kato, "Networking and communications in autonomous driving: A survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1243–1274, 2018.

[18] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue *et al.*, "Llama-adapter v2: Parameter-efficient visual instruction model," *arXiv preprint arXiv:2304.15010*, 2023.

[19] Z. Yin, J. Wang, J. Cao, Z. Shi, D. Liu, M. Li, L. Sheng, L. Bai, X. Huang, Z. Wang *et al.*, "Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark," *arXiv preprint arXiv:2306.06687*, 2023.

[20] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

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

[22] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song *et al.*, "Cogvlm: Visual expert for pretrained language models," *arXiv preprint arXiv:2311.03079*, 2023.

[23] E. Leurent, "An environment for autonomous driving decision-making," https://github.com/eleurent/highway-env, 2018.

[24] O.-R. A. D. O. Committee, *Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems*. SAE International, 2014.

[25] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 4693–4700.

[26] Y. Zhang, Z. Zhu, W. Zheng, J. Huang, G. Huang, J. Zhou, and J. Lu, "Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving," *arXiv preprint arXiv:2205.09743*, 2022.

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

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

[29] Y. Liu, J. Zhang, L. Fang, Q. Jiang, and B. Zhou, "Multimodal motion prediction with stacked transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7577–7586.

[30] J. Cui, L. S. Liew, G. Sabaliauskaite, and F. Zhou, "A review on safety failures, security attacks, and available countermeasures for autonomous vehicles," *Ad Hoc Networks*, vol. 90, p. 101823, 2019.

[31] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE access*, vol. 8, pp. 58 443–58 469, 2020.

[32] Y. Abeysirigoonawardena, F. Shkurti, and G. Dudek, "Generating adversarial driving scenarios in high-fidelity simulators," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8271–8277.

[33] W. Ding, B. Chen, B. Li, K. J. Eun, and D. Zhao, "Multimodal safety-critical scenarios generation for decision-making algorithms evaluation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1551–1558, 2021.

[34] W. Liu, S. Muramatsu, and Y. Okubo, "Cooperation of v2i/p2i communication and roadside radar perception for the safety of vulnerable road users," in *2018 16th International Conference on Intelligent Transportation Systems Telecommunications (ITST)*. IEEE, 2018, pp. 1–7.

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

[36] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[37] D. Fu, X. Li, L. Wen, M. Dou, P. Cai, B. Shi, and Y. Qiao, "Drive like a human: Rethinking autonomous driving with large language models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 910–919.

[38] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi, "Unified-io: A unified model for vision, language, and multi-modal tasks," *arXiv preprint arXiv:2206.08916*, 2022.

[39] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[40] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[41] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao *et al.*, "A survey on multimodal large language models for autonomous driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 958–979.

[42] C. Cui, Y. Ma, X. Cao, W. Ye, and Z. Wang, "Receive, reason, and react: Drive as you say with large language models in autonomous vehicles," *arXiv preprint arXiv:2310.08034*, 2023.

[43] H. Sha, Y. Mu, Y. Jiang, L. Chen, C. Xu, P. Luo, S. E. Li, M. Tomizuka, W. Zhan, and M. Ding, "Languagempc: Large language models as decision makers for autonomous driving," *arXiv preprint arXiv:2310.03026*, 2023.

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

[45] Y. Cui, S. Huang, J. Zhong, Z. Liu, Y. Wang, C. Sun, B. Li, X. Wang, and A. Khajepour, "Drivellm: Charting the path toward full autonomous driving with large language models," *IEEE Transactions on Intelligent Vehicles*, 2023.

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

[47] Y. Ma, Y. Cao, J. Sun, M. Pavone, and C. Xiao, "Dolphins: Multimodal language model for driving," *arXiv preprint arXiv:2312.00438*, 2023.

[48] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[49] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv:2304.02643*, 2023.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[51] H. Chase, "Langchain," https://github.com/langchain-ai/langchain, accessed: 2022-10-17.

[52] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

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