Multi-Agent Autonomous Driving Systems with Large Language Models: A Survey of Recent Advances

Anonymous ACL submission

Abstract

Autonomous Driving Systems (ADSs) are revolutionizing transportation by reducing human intervention, improving operational efficiency, and enhancing safety. Large Language Models (LLMs), known for their exceptional planning and reasoning capabilities, have been integrated into ADSs to assist with driving decision-making. However, LLM-based single-agent ADSs face three major challenges: limited perception, insufficient collaboration, and high computational demands. To address these issues, recent advancements in *LLM-based multi-agent ADSs* have focused on improving inter-agent communication and cooperation. This paper provides a frontier survey of LLM-based multi-agent ADSs. We begin with a background introduction to related concepts, followed by a categorization of existing LLM-based approaches based on different agent interaction modes. We then discuss agent-human interactions in scenarios where LLM-based agents engage with humans. Finally, we summarize key applications, datasets, and challenges in this field to support future research[1].

1 Introduction

Autonomous driving systems (ADSs) are redefining driving behaviors, reshaping global transportation networks, and driving a technological revolution (Yurtsever et al., 2020). Traditional ADSs primarily rely on data-driven approaches (as detailed in Appendix A.1), often focusing on system development while overlooking dynamic interactions with the environment. To enhance engagement with diverse and complex driving scenarios, agentic roles have been incorporated into ADSs (Durante et al., 2024) using methods such as reinforcement learning (Zhang et al., 2024b) and active learning (Lu et al., 2024). Despite notable progress, these methods struggle with "long-tail" scenarios, where rare but critical driving situations—such as sudden obstacles—pose significant challenges to model performance. Furthermore, their "black-box" nature limits interpretability, making their decisions difficult to trust.

LLM-based single-agent ADSs help overcome the limitations of data-driven methods (Wang et al., 2024a). Pre-trained on vast, multi-domain datasets, LLMs excel in knowledge transfer and generalization (Achiam et al., 2023), enabling strong performance in traffic scenarios under zero-shot settings, thus addressing the long-tail issue (Yang et al., 2023). Moreover, techniques such as Reinforcement Learning from Human Feedback (RLHF) and Chain-of-Thought (CoT) (Zhao et al., 2023), enhance language-based interaction and logical reasoning, allowing LLMs to make human-like, real-time decisions while providing interpretable and trustworthy feedback across various driving conditions. For instance, Drive-Like-a-Human (Fu et al., 2024) builds a closed-loop system comprising environment, agent, memory, and expert modules. The agent interacts with the environment, reflects on expert feedback, and ultimately accumulates experience. DiLu (Wen et al., 2024) replaces human experts with a reflection module and integrates an LLM-based reasoning engine to enable continuous decision-making. Agent-Driver (Mao et al., 2024) designs a tool library to collect environmental data and uses LLMs' cognitive memory and reasoning to improve planning.

However, as shown in Figure 1, researchers have identified three critical limitations of LLM-based single-agent ADSs in complex traffic environments: ❶ **Limited Perception:** LLMs can only respond to sensor inputs and lack predictive and generalization capabilities. As a result, LLM-based single-agent ADSs cannot complement incomplete sensor information and thus miss critical information in driving scenarios, such as pedestrians or vehicles

[1] We provide an open-source library for future study via the following link: `https://anonymous.4open.science/r/LLM-based_Multi-agent_ADS-3A5C/README.md`

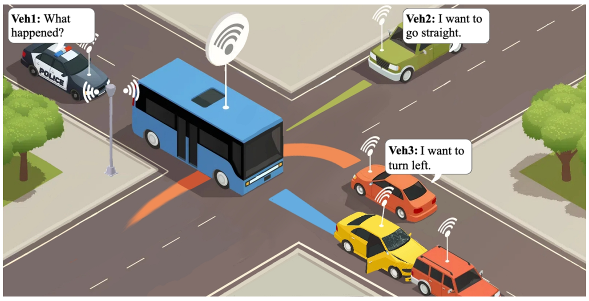

Figure 1: Limitations of LLM-based single-agent ADSs. At an intersection without traffic lights, an accident has occurred ahead, causing Veh1 to be stuck. Due to **limited perception**, Veh1 is unable to assess the situation and cannot proceed. Veh2 intends to go straight, and Veh3 wants to turn left. However, due to **insufficient collaboration**, they are also unable to navigate the intersection efficiently. Furthermore, due to **high computing demands**, the lightweight agent on Veh1 struggles to handle the complex driving scenario and has to rely on a more powerful cloud-based agent for assistance.

hidden in complex intersection environments (Hu et al., 2024c). ❷ **Insufficient Collaboration:** A single LLM-based agent cannot coordinate with other vehicles or infrastructure, leading to suboptimal performance in scenarios requiring multi-agent interactions, such as lane merging or navigating roundabouts (Hu et al., 2021). ❸ **High Computational Demands:** With billions of parameters in LLMs, these methods demand substantial independent computational resources, making real-time deployment challenging, particularly in resource-limited in-vehicle systems (Cui et al., 2023).

To address these limitations, LLM-based multi-agent ADSs enable distinct agents to communicate and collaborate, enhancing safety and performance. First, LLMs enhance contextual awareness by allowing agents to share data, extend their perceptual range, and enhance the detection of occluded objects in complex environments (Hu et al., 2024c). Second, real-time coordination between LLM-based agents mitigates insufficient collaboration, enabling joint decision-making in scenarios such as lane merging and roundabout navigation, ultimately leading to safer and more efficient driving operations (Hu et al., 2021). Third, LLMs optimize computational efficiency by distributing tasks among agents, reducing individual workloads, and enabling real-time processing in resource-limited systems (Cui et al., 2023).

As LLM capabilities continue to advance, they are playing an increasingly significant role in ADS as intelligent driving assistants. Several reviews have focused on two primary aspects: ***i)*** the integration of LLMs into data-driven methods (Yang et al., 2023; Li et al., 2023) and ***ii)*** the applications of specific LLM types, such as vision-based (Zhou et al., 2024b) and multimodal-based (Fourati et al., 2024; Cui et al., 2024c) models in ADSs. However, no comprehensive survey has systematically examined the emerging field of LLM-based multi-agent ADSs. This gap motivates us to provide a thorough review that consolidates existing knowledge and offers insights to guide future research and the development of advanced ADSs.

In this study, we present a comprehensive survey of LLM-based multi-agent systems. Specifically, Section 2 introduces the core concepts of LLM-based multi-agent ADSs, including *agent environments and profiles*, *inter-agent interaction mechanisms*, and *agent-human interactions*. Section 3 provides a structured review of the state-of-the-art in multi-agent ADS, categorizing existing studies into three key interaction types: *multi-vehicle interaction*, *vehicle-infrastructure interaction*, and *vehicle-assistant interaction*. As agent capabilities continue to grow, human-vehicle co-driving is becoming the dominant autonomous driving paradigm, with human involvement playing an increasingly vital role. Humans collaborate with agents by providing guidance or supervising their behavior. Therefore, we consider humans as special virtual agents and examine human-agent interactions in Section 4. Section 5 explores various applications, while Section 6 compiles a comprehensive collection of public datasets and open-source resources. Section 7 discusses existing challenges and future research directions and Section 8 concludes the study.

2 LLM-based Agents for ADS

2.1 LLM-based Single-agent ADS

Achieving human-level driving is an ultimate goal of ADS. As shown in Figure 2(a), the LLM-based single agent retrieves past driving experiences from the memory, integrates them with real-time environmental information for reasoning, and makes driving decisions. Additionally, the driving agent reflects on its decision and updates its memory accordingly, ensuring safe and efficient driving actions. However, the complex and dynamic nature of real-world driving scenarios, where interactions with other vehicles significantly impact decision-making, suggests that neglecting these interactions

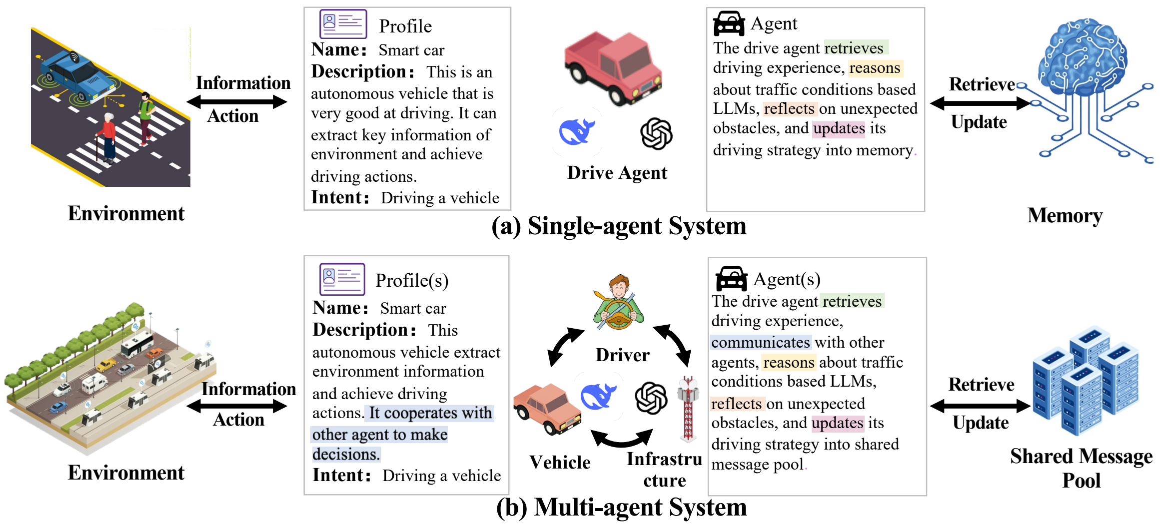

Figure 2: Overview of LLM-based (a) single- and (b) multi-agent ADSs, with key terms and differences highlighted.

can lead to suboptimal or unsafe driving outcomes.

2.2 LLM-based Multi-agent ADS

With interactions among multiple agents, LLM-based multi-agent ADS leverages collective intelligence and specialized skills, with each agent playing a distinct role, communicating and collaborating within the system. This enhances the efficiency and safety of autonomous driving. Below, we introduce the LLM-based multi-agent ADS, as shown in Figure 2(b), and provide a detailed analysis of its three key modules: Agent Environment and Profile, LLM-based Multi-Agent Interaction, and LLM-based Agent-Human Interaction.

2.2.1 Agent Environment and Profile

Similar to the single-agent architecture in Figure 2(a), multi-agent systems first obtain relevant information from their environments, enabling them to make informed decisions and take appropriate actions. The environmental conditions define the settings and necessary context for agents in LLM-based multi-agent ADS to operate effectively. Generally, there are two environment types, *i.e.*, physical environment and simulation environment.

- ***Physical environment***. It represents the real-world setting where driver agents gather information using various sensors, such as cameras and LiDAR, and interact with other traffic participants.However, due to the high cost of vehicles and strict regulations on public roads, collecting large amounts of data in real world is impractical.

- ***Simulation environment***. As a viable alternative, the simulation environment provides a simulated setting constructed by humans. It can accurately model specific conditions without incurring the high costs and complexities associated with real-world data collection, allowing agents to freely test actions and strategies across a variety of scenarios (Dosovitskiy et al., 2017).

In LLM-based multi-agent systems, each agent is assigned distinct roles with specific functions through profiles, enabling them to collaborate on complex driving tasks or simulate intricate traffic scenarios. These profiles are crucial in defining the functionality of the agent, its interaction with the environment, and its collaboration with other agents. Existing work (Li et al., 2024) generates agent profiles using three types of methods: Pre-defined, Model-generated, and Data-derived.

- ***Pre-defined methods***. In these cases, system designers explicitly define agent profiles based on prior knowledge and the analysis of complex scenarios (Chen et al., 2024a). Each agent has unique attributes and behavior patterns that can be adjusted based on the scenario. In driving environments, the objectives of ADS require the collaboration of vehicle agents, infrastructure agents, and drivers. In particular, ❶ Vehicle agents denote various types of autonomous vehicles, traveling according to preset routes and traffic rules, while communicating and collaborating with other vehicles and driver agents. ❷ Infrastructure agents, *e.g.*, traffic lights, road condition monitors, and parking facilities, provide real-time traffic information and instructions, influencing the behavior of driver and vehicle agents.

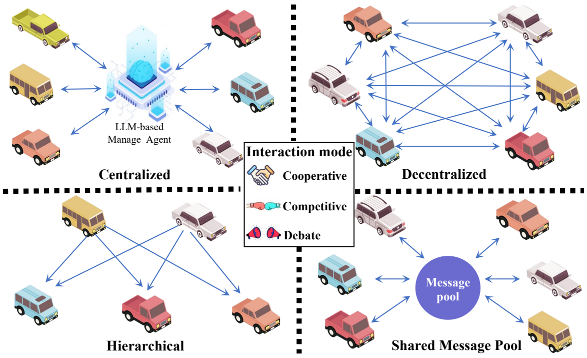

Figure 3: Different interaction modes and interaction structures.

- ***Model-generated methods***. These approaches create agent profiles using advanced LLMs based on the interaction context and the goals that need to be accomplished (Zhou et al., 2024c).

- ***Data-derived Profile***. They design agent profiles based on pre-existing datasets (Guo et al., 2024).

2.2.2 LLM-based Multi-Agent Interaction

In LLM-based multi-agent ADS, effective information exchange and action coordination between agents are essential to improve collective intelligence and solve complex traffic scenarios. Agent interactions are influenced by both the interaction mode and the underlying interaction structure.

- ***The interaction mode*** of LLM-based multi-agent ADS can be classified as: *cooperative*, *competitive*, and *debate* mode. ❶ In cooperative mode, agents work together to achieve shared objectives by exchanging information (Chen et al., 2024d; Jin et al., 2024). ❷ In competitive mode, agents strive to accomplish their individual goals and compete with others (Yao et al., 2024). ❸ The Debate mode enables agents to debate with each other, propose their own solutions, criticize the solutions of other agents, and collaboratively identify optimal strategies (Liang et al., 2024).

- ***The interaction structure*** delineates the architecture of communication networks within LLM-based multi-agent ADS, including *centralized*, *decentralized*, *hierarchical*, and *shared message pool* structures, as shown in Figure 3. Specifically, ❶ the centralized interaction structures defines a central agent or a group of central agents to manage interactions among all agents (Zhou et al., 2024c). ❷ The decentralized interaction structure allows for direct communication between agents, with all agents being equal to each other (Hu et al., 2024b). ❸ Hierarchical structures focus on interactions within a layer or with adjacent layers (Ohmer et al., 2022). ❹ The shared memory interaction structure maintains a shared message pool, allowing agents to send and extract the necessary information (Jiang et al., 2024). We provide a more detailed introduction to LLM-based multi-agent ADSs based on their interaction structures and modes in Section 3.

2.2.3 LLM-based Agent-Human Interaction

Recent studies have shown that human-machine co-driving systems leverage LLMs to improve agent-human interactions, enabling autonomous vehicles to communicate and collaborate seamlessly with human drivers through natural language (Feng et al., 2024). This capability allows vehicles to better understand and respond to human intent, provide context-aware responses, enhance driving safety and comfort, and offer personalized recommendations based on driver preferences. Furthermore, humans play a crucial role in guiding and supervising agent behavior, enhancing the agents' capabilities while ensuring safety and compliance with legal standards. We explore the role of humans as special virtual agents in LLM-based multi-agent ADS and examine the intricate dynamics of agent-human interactions in Section 4.

3 LLM-based Multi-Agent Interaction

Mutual interaction is central to multi-agent ADSs, enabling systems to solve complex problems beyond the capabilities of a single agent. Through information exchange and coordinated decision-making, multiple agents effectively complete shared tasks and achieve overarching objectives (Li et al., 2024). This section reviews recent studies on multi-agent ADSs, emphasizing interactions among vehicles, infrastructures, and assisted agents in driving scenarios. As shown in Figure 4, we categorize existing methods into three interaction types: *multi-vehicle interaction*, *vehicle-infrastructure interaction*, and *vehicle-assistant interaction*.

3.1 Multi-Vehicle Interaction

Multi-vehicle interactions involve multiple autonomous vehicles powered by LLMs exchanging real-time information, such as locations, speeds, sensor data, and intended trajectories. By sharing partial observations of the environment or negotiating maneuvers, multiple vehicles overcome the inherent limitations of single-agent ADS, such

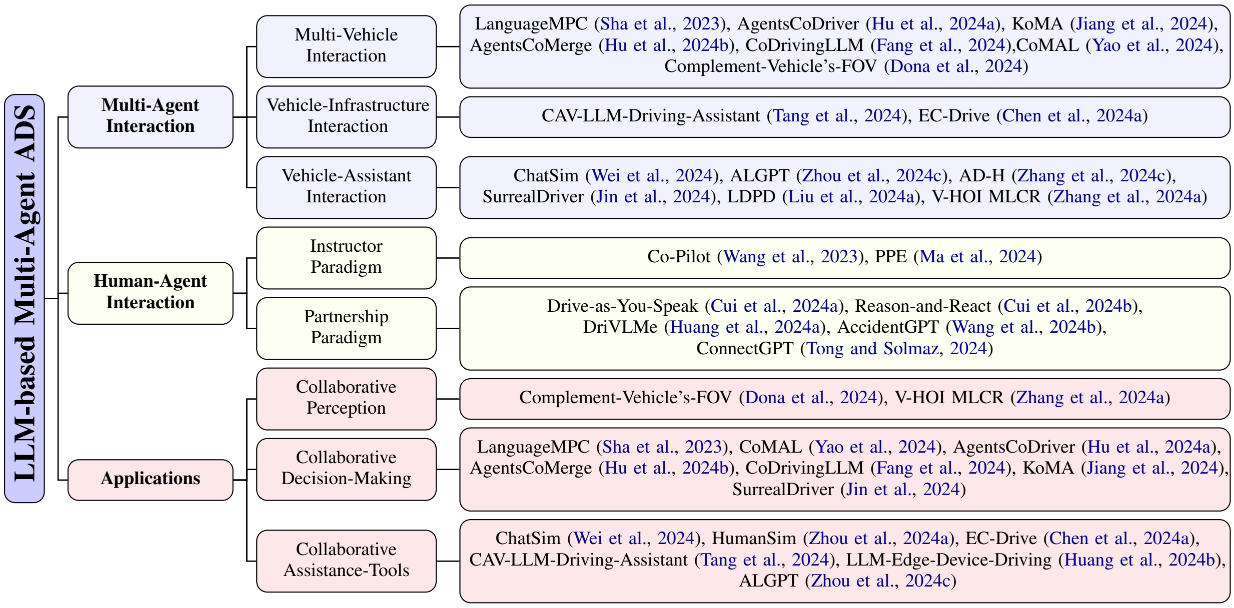

Figure 4: A taxonomy of LLM-based Multi-Agent Autonomous Driving Systems.

as restricted perception and lack of collaboration. Typically, these interactions operate in a cooperative mode. LanguageMPC (Sha et al., 2023) employs a centralized structure, where a central agent acts as the "brain" of the fleet, providing coordination and control commands to each vehicle agent. In contrast, other decentralized approaches (Fang et al., 2024; Dona et al., 2024) treat all agents equally, allowing direct communication between multiple agents. For instance, AgentsCoDriver (Hu et al., 2024a) designs a communication module that generates messages for inter-agent communication when the agent deems it necessary. AgentsCoMerge (Hu et al., 2024b) and CoDrivingLLM (Fang et al., 2024) incorporate agent communication into the reasoning process, facilitating intention sharing and negotiation before decision-making. Additionally, KoMA (Jiang et al., 2024) and CoMAL (Yao et al., 2024) build a shared memory pool, allowing agents to send and retrieve the necessary information to facilitate interaction between agents.

3.2 Vehicle-Infrastructure Interaction

The interaction between vehicles and external agents, such as traffic lights, roadside sensors, and LLM-powered control centers, not only helps autonomous vehicles make more intelligent decisions but also alleviates on-board computing requirements. This enables LLM-based multi-agent ADSs to operate effectively in real-world environments. EC-Drive (Chen et al., 2024a) proposes an Edge-Cloud collaboration framework with a hierarchical interaction structure. The edge agent processes real-time sensor data and makes preliminary decisions under normal conditions. When anomalies are detected or the edge agent generates a low-confidence prediction, the system flags these instances and uploads them to the cloud agent equipped with LLMs. The cloud agent then performs detailed reasoning to generate optimized decisions and combines them with the output of the edge agent to update the driving plan. Following a similar architecture, Tang et al. (2024) uses agents deployed on remote clouds or network edges to assist connected driving agents in handling complex driving decisions.

3.3 Vehicle-Assistant Interaction

Beyond the interactions between the primary agents in driving scenarios, additional interactions among assisted agents play a crucial role in LLM-based multiagent ADSs. Both ChatSim (Wei et al., 2024) and ALGPT (Zhou et al., 2024c) employ a manager (PM) agent to interpret user instructions and coordinate tasks among other agents. ChatSim (Wei et al., 2024) adopts a centralized structure in which the PM agent decouples an overall demand into specific subtasks and dispatches instructions to other team agents. Similarly, the PM agent in ALGPT (Zhou et al., 2024c) formulates a work plan upon receiving user commands and assembles an agent team with the plan. Specifically, agents no longer communicate point-to-point with each other but instead communicate through a shared message pool, greatly improving efficiency.

Additionally, hierarchical agent architectures further enhance the performance and effectiveness of LLM-based multi-agent ADSs. AD-H (Zhang et al., 2024c) assigns high-level reasoning tasks to the multimodal LLM-based planner agent while delegating low-level control signal generation to a lightweight controller agent. These agents interact through mid-level commands generated by the multimodal LLMs. In LDPD (Liu et al., 2024a), the teacher agent leverages the LLM for complex cooperative decision reasoning and trains smaller student agents via its own decision demonstrations to achieve cooperative decision-making. SurrealDriver (Jin et al., 2024) introduces a CoachAgent to evaluate DriverAgent's driving behavior and provide guidelines for continuous improvement.

Different from the conventional collaborative interaction mode, V-HOI (Zhang et al., 2024a) proposes a hybrid interaction mode that blends collaboration with debate. It establishes various agents across different LLMs to evaluate reasoning logic from different aspects, enabling cross-agent reasoning. This process culminates in a debate-style integration of responses from various LLMs, improving predictions for enhanced decision-making.

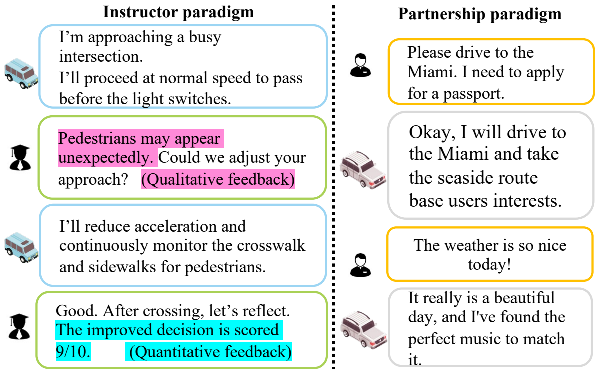

Figure 5: Two modes of agent-human interaction.

4 LLM-based Agent-Human Interaction

Depending on the roles of human assume when interacting with agents, we classify current methods as: *instructor paradigm* and *partnership paradigm*.

4.1 Instructor Paradigm

In Figure 5, the instructor paradigm involves agents interacting with humans in a conversational manner, where humans act as "tutors" to offer quantitative and qualitative feedback to improve the agent's decision-making (Li et al., 2017). Quantitative feedback typically includes binary evaluations or ratings, while qualitative feedback consists of language suggestions for refinement. The agent incorporates these suggestions to adapt and enhance its performance in complex driving scenarios. For instance, Wang et al. (2023) propose "Expert-Oriented Black-box Tuning", a method where domain experts provide feedback to optimize model performance. Similarly, Ma et al. (2024) present a human-guided learning pipeline that integrates driver feedback to refine agent decision-making.

4.2 Partnership Paradigm

In Figure 5, the partnership paradigm emphasizes collaboration, where agents and humans interact as equals to accomplish complex driving tasks. In this paradigm, agents assist in decision-making by adapting to individual driver preferences and real-time traffic conditions. For instance, Talk2Drive (Cui et al., 2023), DaYS (Cui et al., 2024a) and Receive (Cui et al., 2024b) utilize memory modules to store human-vehicle interactions, enabling a more personalized driving experience based on individual driver preferences, such as overtaking speed and following distance. Additionally, infrastructure agents in AccidentGPT (Wang et al., 2024b) and ConnectGPT (Tong and Solmaz, 2024) connect vehicles to monitor traffic conditions, identify potential hazards, and provide proactive safety warnings, blind spot alerts, and driving suggestions through agent-human interaction.

5 Applications

5.1 Collaborative Perception

Despite significant advancements in the perception modules of ADS, LLM-based single-agent ADS continues to face substantial challenges, including constrained sensing ranges and persistent occlusion issues (Han et al., 2023). These two key limitations hinder their comprehensive understanding of the driving environment and can lead to suboptimal decision-making, especially in complex and dynamic traffic scenarios (Hu et al., 2024c).

(Dona et al., 2024) propose a multi-agent cooperative framework that enhances the ego vehicle's field-of-view (FOV) by integrating complementary visual perspectives through inter-vehicle dialogues mediated by onboard LLMs, significantly expanding the ego vehicle's environmental comprehension. However, in complex road scenarios, reliance on a single LLM can lead to erroneous interpretations and hallucinatory predictions when processing complex traffic situations. To address this limitation, V-HOI MLCR (Zhang et al., 2024a) introduces a collaborative debate framework among different LLMs for video-based Human-Object Interaction (HOI) detection tasks. This framework first implements a Cross-Agent Reasoning scheme, assigning distinct roles to various agents within an LLM to conduct reasoning from multiple perspectives. Subsequently, a cyclic debate mechanism is employed to evaluate and aggregate responses from multiple agents, culminating in the final outcome.

5.2 Collaborative Decision-Making

After obtaining environmental information, the ADS performs three core functions: route planning, trajectory optimization, and real-time decision-making. In complex traffic scenarios such as roundabout navigation and lane merging, LLM-based multi-agent systems enable coordinated motion planning through three key mechanisms: ❶ real-time intention sharing between agents, ❷ adaptive communication protocols, and ❸ dynamic negotiation frameworks. This collaborative architecture allows ADS to precisely coordinate their trajectories, maneuver strategies, and environmental interactions while maintaining operational safety.

LanguageMPC (Sha et al., 2023) uses LLMs to perform scenario analysis and decision-making. Additionally, it introduces a multi-vehicle control method where distributed LLMs govern individual vehicle operations, while a central LLM facilitates multi-vehicle communication and coordination. AgentsCoDriver (Hu et al., 2024a) presents a comprehensive LLM-based multi-vehicle collaborative decision-making framework with life-long learning capabilities, moving the field towards practical applications. This framework consists of five parts, as follows: the observation module, cognitive memory module, and reasoning engine support the high-level decision-making process for AD; the communication module enables negotiation and collaboration among vehicles; and the reinforcement reflection module reflects the output and decision-making process. Similarly, AgentsCoMerge (Hu et al., 2024b) combines vision-based and text-based scene understanding to gather essential environmental information and incorporates a hierarchical planning module to allow agents to make informed decisions and effectively plan trajectories. Instead of directly interacting with each other, agents in KoMA (Jiang et al., 2024) analyze and infer the intentions of surrounding vehicles via an interaction module to enhance decision-making. It also introduces a shared memory module to store successful driving experiences and a ranking-based reflection module to review them.

5.3 Other Collaborative Assistance-Tools

The long-term data accumulation in both industry and academia has enabled great success in highway driving and automatic parking (Liu et al., 2024b). However, collecting real-world data remains costly, especially for multi-agents or customized scenarios. Additionally, the uncontrollable nature of real scenarios makes it challenging to capture certain corner cases. To address these issues, many LLM-based studies focus on simulating multi-agent ADS, offering a cost-effective alternative to real-world data collection. For example, ChatSim (Wei et al., 2024) provides editable photo-realistic 3D driving scenario simulations via natural language commands and external digital assets. The system leverages multiple LLM agents with specialized roles to decompose complex commands into specific editing tasks, introducing novel McNeRF and Mclight methods that generate customized high-quality output. HumanSim (Zhou et al., 2024a) integrates LLMs to simulate human-like driving behaviors in multi-agent systems via pre-defined driver characters. By employing navigation strategies, HumanSim facilitates behavior-level control of vehicle movements, making it easier to generate corner cases in multi-agent environments.

Although many innovative studies have explored the application of LLM-based multi-agent ADS, significant technical challenges remain in deploying LLMs locally on autonomous vehicles due to their huge computational resource requirements (Sun et al., 2024). To address these issues, Tang et al. (2024) apply remote LLMs to provide assistance for connected autonomous vehicles, which communicate between themselves and with LLMs via vehicle-to-everything technologies. Moreover, this study evaluates LLMs' comprehension of driving theory and skills in a manner akin to human driver tests. However, remote LLM deployment can introduce inference latency, posing risks in emergency scenarios. To further improve system efficiency, Chen et al. (2024a) introduce a novel edge-cloud collaborative ADS with drift detection capabilities, using small LLMs on edge devices and GPT-4 on cloud to process motion planning data and complex inference tasks, respectively.

In addition, ALGPT (Zhou et al., 2024c) uses a muti-agent cooperative framework to enable open-

Table 1: Single-agent and multi-agent autonomous driving datasets.

Datasets	Dataset Type	Sensor Type	Tasks
KITTI (Geiger et al., 2012)	Single-agent	Camera, LiDAR	2D/3D detection, tracking, depth estimation
nuScenes (Geiger et al., 2020)	Single-agent	Cameras, LiDAR, Radars	3D detection, tracking, trajectory forecasting
BDD100K (Yu et al., 2020)	Single-agent	Camera	Object detection, lane detection, segmentation
Waymo (Sun et al., 2020)	Single-agent	Camera, LiDAR, Radars	2D/3D detection, tracking, domain adaptation
BDD-X (Kim et al., 2018)	Single-agent	BDD	Object detection, driving scenario captioning
nuScenes-QA (Qian et al., 2024)	Single-agent	nuScenes	3D detection, tracking, visual QA
DriveLM (Sima et al., 2025)	Single-agent	nuScenes, Waymo	Multi-modal planning, question answering
DAIR-V2X (Yu et al., 2022)	Multi-agent	Camera, LiDAR (multi-vehicle)	Cooperative perception, tracking
TUMTraf-V2X (Zimmer et al., 2024)	Multi-agent	Multi-vehicle camera, LiDAR	Cooperative perception, multi-agent tracking
V2V4Real (Xu et al., 2023)	Multi-agent	Multi-vehicle camera, LiDAR	Cooperative detection, tracking
V2XSet (Xu et al., 2022)	Multi-agent	Multi-vehicle camera, LiDAR	Multi-agent detection, tracking

vocabulary and multimodal auto-annotation for autonomous driving. ALGPT introduces a Standard Operating Procedure that clarifies the role of each agent and shares project documentation, thereby enhancing the effectiveness of multi-agent interactions. Moreover, ALGPT establishes a specialized knowledge base for each type of agent, using CoT and In-Context Learning (Brown et al., 2020).

6 Datasets

We organize the latest state-of-the-art open-source work to foster research of more advanced ADSs. And we summarize mainstream ADS datasets in Table 1. More details are listed in Appendix A.3.

7 Challenges and Future Directions

This section explores key open challenges and potential opportunities for future research.

- ***Hallucination Problem.*** It refers to LLMs generating outputs that are factually incorrect or nonsensical (Huang et al., 2023). In complex driving scenarios, a single driving agent's hallucinations in an LLM-based multi-agent ADS can be accepted and further propagated by other agents in the network via the inter-agent communication, potentially leading to serious accidents. Consequently, detecting and mitigating hallucinations at the individual agent level and managing the flow of information between agents are crucial issues for future research (Fan et al., 2024).

- ***Multi-Modality Ability.*** Agents in current multi-agent systems primarily use LLMs for scene understanding and decision making. These methods convert the outputs of perception algorithms into textual representations through manual prompts or interpreters, which are then fed into an LLM to produce decisions. This approach heavily depends on the performance of the perception algorithm and can lead to loss of environmental information (Gao et al., 2023). Therefore, integrating language understanding with the ability to process and fuse multiple data modalities to develop a multimodal multi-agent ADS represents a promising direction for future research.

- ***Scalability Problem.*** LLM-based multi-agent ADS can scale up by adding more agents to handle increasingly complex driving scenarios. However, more LLM agents increase the demand for computing resources, while their interactions impose strict requirements on communication efficiency, which is critical for real-time decision-making (Huang et al., 2024b). Therefore, under limited computing resources, it is crucial to develop a system architecture that supports distributed computing and efficient communication, as well as agents capable of adapting to various environments and tasks, to optimize multi-agent ADS within resource constraints.

8 Conclusion

This paper systematically outlines LLM-based multi-agent ADS and comprehensively reviews the latest research in this field. Our study first traces the development trajectory of LLM-based multi-agent ADS from single-agent ADS to multi-agent ADS. Subsequently, we provide a detailed description of the LLM-based multi-agent ADS from the perspectives of agent-environments and profiles, inter-agent interaction mechanisms, and agent-human interactions. We also systematically classify and introduce existing studies from the perspectives of multi-agent interaction, agent-human interaction, and different applications. Finally, this paper provides comprehensive public datasets and open source codes, and deeply explores the current challenges and future research directions of LLM-based multi-agent ADS. We hope that this review can bring new inspiration and ideas to future research on LLM-based multi-agent ADS.

Limitations

Emerging Research and Limited Data. As the field of LLM-based multi-agent ADS is still relatively new, existing research is limited, which may restrict the scope of our classification and analysis. **Some Unverified Work.** Since LLM-based multi-agent ADS is a novel topic, some of the papers summarized in this review are from unreviewed arXiv preprints. As these works have not been formally published, their conclusions may require further investigation to confirm their validity. **Limited Discussion on Real-world Applications.** As LLM-based multi-agent ADS is still in the theoretical stage, although many companies have begun deploying practical applications in this area, this review does not cover discussions on real-world deployments due to a lack of up-to-date internal information from these companies.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv:2303.08774*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *In Proc. of NeurIPS*, 33:1877–1901.

Jiao Chen, Suyan Dai, Fangfang Chen, Zuohong Lv, and Jianhua Tang. 2024a. Edge-cloud collaborative motion planning for autonomous driving with large language models. *arXiv:2408.09972*.

Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. 2024b. End-to-end autonomous driving: Challenges and frontiers. *IEEE TPAMI*, 46(12):10164–10183.

Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. 2024c. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *Proc. of ICRA*, pages 14093–14100. IEEE.

Pei Chen, Shuai Zhang, and Boran Han. 2024d. Comm: Collaborative multi-agent, multi-reasoning-path prompting for complex problem solving. In *Proc. of NAACL-HLT (Findings)*, pages 1720–1738. Association for Computational Linguistics.

C Cui, Z Yang, Y Zhou, Y Ma, J Lu, L Li, Y Chen, J Panchal, and Z Wang. 2023. Personalized autonomous driving with large language models: Field experiments. *arXiv:2312.09397*.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. 2024a. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In *Proc. of WACV*, pages 902–909.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. 2024b. Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles. *IEEE ITS Mag*, 16(4):81–94.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024c. A survey on multimodal large language models for autonomous driving. In *Proc. of WACV*, pages 958–979.

Malsha Ashani Mahawatta Dona, Beatriz Cabrero-Daniel, Yinan Yu, and Christian Berger. 2024. Tapping in a remote vehicle's onboard llm to complement the ego vehicle's field-of-view. *arXiv:2408.10794*.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. Carla: An open urban driving simulator. In *Proc. of CoRL*, pages 1–16. PMLR.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. 2024. Agent ai: Surveying the horizons of multimodal interaction. *arXiv:2401.03568*.

Jiaqi Fan, Jianhua Wu, Hongqing Chu, Quanbo Ge, and Bingzhao Gao. 2024. Hallucination elimination and semantic enhancement framework for vision-language models in traffic scenarios. *arXiv:2412.07518*.

Shiyu Fang, Jiaqi Liu, Mingyu Ding, Yiming Cui, Chen Lv, Peng Hang, and Jian Sun. 2024. Towards interactive and learnable cooperative driving automation: a large language model-driven decision-making framework. *arXiv:2409.12812*.

Xueyang Feng, Zhiyuan Chen, Yujia Qin, Yankai Lin, Xu Chen, Zhiyuan Liu, and Ji-Rong Wen. 2024. Large language model-based human-agent collaboration for complex task solving. In *Proc. of EMNLP (Findings)*, pages 1336–1357. Association for Computational Linguistics.

Sonda Fourati, Wael Jaafar, Noura Baccar, and Safwan Alfattani. 2024. Xlm for autonomous driving systems: A comprehensive review. *arXiv:2409.10484*.

Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. 2024. Drive like a human: Rethinking autonomous driving with large language models. In *Proc. of WACVW*, pages 910–919.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv:2304.15010*.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proc. of CVPR*, pages 11621–11631.

Andreas Geiger, Philip Lenz, et al. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. of CVPR*, pages 3354–3361.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. In *Proc. of IJCAI*, pages 8048–8057. ijcai.org.

Yushan Han, Hui Zhang, Huifang Li, Yi Jin, Congyan Lang, and Yidong Li. 2023. Collaborative perception in autonomous driving: Methods, datasets, and challenges. *IEEE ITS Mag*, 15(6):131–151.

Senkang Hu, Zhengru Fang, Yiqin Deng, Xianhao Chen, and Yuguang Fang. 2021. Collaborative autonomous driving—a survey of solution approaches and future challenges. *Sensors*, 21(11):3783.

Senkang Hu, Zhengru Fang, Zihan Fang, Yiqin Deng, Xianhao Chen, and Yuguang Fang. 2024a. Agentscodriver: Large language model empowered collaborative driving with lifelong learning. *arXiv:2404.06345*.

Senkang Hu, Zhengru Fang, Zihan Fang, Yiqin Deng, Xianhao Chen, Yuguang Fang, and Sam Kwong. 2024b. Agentscomerge: Large language model empowered collaborative decision making for ramp merging. *arXiv:2408.03624*.

Senkang Hu, Zhengru Fang, et al. 2024c. Collaborative perception for connected and autonomous driving: Challenges, possible solutions and opportunities. *arXiv:2401.01544*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv:2311.05232*.

Yidong Huang, Jacob Sansom, Ziqiao Ma, Felix Gervits, and Joyce Chai. 2024a. Drivlme: Enhancing llm-based autonomous driving agents with embodied and social experiences. In *Proc. of IROS*, pages 3153–3160. IEEE.

Yizhou Huang, Yihua Cheng, and Kezhi Wang. 2024b. Efficient driving behavior narration and reasoning on edge device using large language models. *arXiv:2409.20364*.

Kemou Jiang, Xuan Cai, Zhiyong Cui, Aoyong Li, Yilong Ren, Haiyang Yu, Hao Yang, Daocheng Fu, Licheng Wen, and Pinlong Cai. 2024. Koma: Knowledge-driven multi-agent framework for autonomous driving with large language models. *IEEE TIV*, pages 1–15.

Ye Jin, Ruoxuan Yang, Zhijie Yi, Xiaoxi Shen, Huiling Peng, Xiaoan Liu, Jingli Qin, Jiayang Li, Jintao Xie, Peizhong Gao, et al. 2024. Surrealdriver: Designing llm-powered generative driver agent framework based on human drivers' driving-thinking data. In *Proc. of IROS*, pages 966–971. IEEE.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. 2018. Textual explanations for self-driving vehicles. In *Proc. of ECCV*, pages 563–578.

Jiwei Li, Alexander H Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2017. Dialogue learning with human-in-the-loop. In *Proc. of ICLR*.

Xin Li, Yeqi Bai, Pinlong Cai, Licheng Wen, Daocheng Fu, Bo Zhang, Xuemeng Yang, Xinyu Cai, Tao Ma, Jianfei Guo, et al. 2023. Towards knowledge-driven autonomous driving. *arXiv:2312.04316*.

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proc. of EMNLP*, pages 17889–17904. Association for Computational Linguistics.

Jiaqi Liu, Chengkai Xu, Peng Hang, Jian Sun, Mingyu Ding, Wei Zhan, and Masayoshi Tomizuka. 2024a. Language-driven policy distillation for cooperative driving in multi-agent reinforcement learning. *arXiv:2410.24152*.

Mingyu Liu, Ekim Yurtsever, Jonathan Fossaert, Xingcheng Zhou, Walter Zimmer, Yuning Cui, Bare Luka Zagar, and Alois C Knoll. 2024b. A survey on autonomous driving datasets: Statistics, annotation quality, and a future outlook. *IEEE TIV*, pages 1–29.

Han Lu, Xiaosong Jia, Yichen Xie, Wenlong Liao, Xiaokang Yang, and Junchi Yan. 2024. Activead: Planning-oriented active learning for end-to-end autonomous driving. *arXiv:2403.02877*.

Yunsheng Ma, Xu Cao, Wenqian Ye, Can Cui, Kai Mei, and Ziran Wang. 2024. Learning autonomous driving tasks via human feedbacks with large language models. In *Proc. of EMNLP (Findings)*, pages 4985–4995.

Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. 2023. Gpt-driver: Learning to drive with gpt. *arXiv:2310.01415*.

Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. 2024. A language agent for autonomous driving. In *Proc. of COLM*.

Xenia Ohmer, Marko Duda, and Elia Bruni. 2022. Emergence of hierarchical reference systems in multi-agent communication. In *Proc. of COLING*, pages 5689–5706. International Committee on Computational Linguistics.

Mingxing Peng, Xusen Guo, Xianda Chen, Meixin Zhu, Kehua Chen, Xuesong Wang, Yinhai Wang, et al. 2024. Lc-llm: Explainable lane-change intention and trajectory predictions with large language models. *arXiv:2403.18344*.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. 2024. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proc. of AAAI*, pages 4542–4550.

Axel Sauer, Nikolay Savinov, and Andreas Geiger. 2018. Conditional affordance learning for driving in urban environments. In *Proc. of CoRL*, pages 237–252. PMLR.

Hao Sha, Yao Mu, Yuxuan Jiang, Li Chen, Chenfeng Xu, Ping Luo, Shengbo Eben Li, Masayoshi Tomizuka, Wei Zhan, and Mingyu Ding. 2023. Languagempc: Large language models as decision makers for autonomous driving. *arXiv:2310.03026*.

Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. 2022. Motion transformer with global intention localization and local movement refinement. *In Proc. of NeurIPS*, 35:6531–6543.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. 2025. Drivelm: Driving with graph visual question answering. In *Proc. of ECCV*, pages 256–274.

Hao Sun, Jiayi Wu, Hengyi Cai, Xiaochi Wei, Yue Feng, Bo Wang, Shuaiqiang Wang, Yan Zhang, and Dawei Yin. 2024. Adaswitch: Adaptive switching between small and large agents for effective cloud-local collaborative learning. In *Proc. of EMNLP*, pages 8052–8062. Association for Computational Linguistics.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. 2020. Scalability in perception for autonomous driving: Waymo open dataset. In *Proc. of CVPR*, pages 2446–2454.

Zuoyin Tang, Jianhua He, Dashuai Pe, Kezhong Liu, Tao Gao, and Jiawei Zheng. 2024. Test large language models on driving theory knowledge and skills for connected autonomous vehicles. In *Proc. of MobiArch*, pages 1–6.

Kailin Tong and Selim Solmaz. 2024. Connectgpt: Connect large language models with connected and automated vehicles. In *Proc. of IEEE IV*, pages 581–588.

Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. 2023. Scene as occupancy. In *Proc. of ICCV*, pages 8406–8415.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024a. A survey on large language model based autonomous agents. *FCS*, 18(6):186345.

Lening Wang, Yilong Ren, Han Jiang, Pinlong Cai, Daocheng Fu, Tianqi Wang, Zhiyong Cui, Haiyang Yu, Xuesong Wang, Hanchu Zhou, et al. 2024b. Accidentgpt: A v2x environmental perception multi-modal large model for accident analysis and prevention. In *Proc. of IEEE IV*, pages 472–477. IEEE.

Shiyi Wang, Yuxuan Zhu, Zhiheng Li, Yutong Wang, Li Li, and Zhengbing He. 2023. Chatgpt as your vehicle co-pilot: An initial attempt. *IEEE TIV*, 8(12):4706–4721.

Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. 2021. DETR3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Proc. of CoRL*.

Yuxi Wei, Zi Wang, Yifan Lu, Chenxin Xu, Changxing Liu, Hao Zhao, Siheng Chen, and Yanfeng Wang. 2024. Editable scene simulation for autonomous driving via collaborative llm-agents. In *Proc. of CVPR*, pages 15077–15087.

Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao MA, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. 2024. Dilu: A knowledge-driven approach to autonomous driving with large language models. In *Proc. of ICLR*.

Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, et al. 2023. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proc. of CVPR*, pages 13712–13722.

Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. 2022. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *Proc. of ECCV*, pages 107–124.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. 2024. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE RA-L*.

Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. 2023. Llm4drive: A survey of large language models for autonomous driving. In *NeurIPS 2024 Workshop on Open-World Agents*.

Huaiyuan Yao, Longchao Da, Vishnu Nandam, Justin Turnau, Zhiwei Liu, Linsey Pang, and Hua Wei. 2024. Comal: Collaborative multi-agent large language models for mixed-autonomy traffic. *arXiv:2410.14368*.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arXiv:2303.10420*.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proc. of CVPR*, pages 2636–2645.

Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. 2022. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proc. of CVPR*, pages 21361–21370.

Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. 2020. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, pages 58443–58469.

Hang Zhang, Wenxiao Zhang, Haoxuan Qu, and Jun Liu. 2024a. Enhancing human-centered dynamic scene understanding via multiple llms collaborated reasoning. *arXiv:2403.10107*.

Ruiqi Zhang, Jing Hou, Florian Walter, Shangding Gu, Jiayi Guan, Florian Röhrbein, Yali Du, Panpan Cai, Guang Chen, and Alois Knoll. 2024b. Multi-agent reinforcement learning for autonomous driving: A survey. *arXiv:2408.09675*.

Zaibin Zhang, Shiyu Tang, Yuanhang Zhang, Talas Fu, Yifan Wang, Yang Liu, Dong Wang, Jing Shao, Lijun Wang, and Huchuan Lu. 2024c. Ad-h: Autonomous driving with hierarchical agents. *arXiv:2406.03474*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv:2303.18223*.

Lingfeng Zhou, Mohan Jiang, and Dequan Wang. 2024a. Humansim: Human-like multi-agent novel driving simulation for corner case generation. In *ECCV 2024 Workshop on MPCC-AD*.

Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C Knoll. 2024b. Vision language models in autonomous driving: A survey and outlook. *IEEE TIV*, pages 1–20.

Yijie Zhou, Xianhui Cheng, Qiming Zhang, Lei Wang, Wenchao Ding, Xiangyang Xue, Chunbo Luo, and Jian Pu. 2024c. Algpt: Multi-agent cooperative framework for open-vocabulary multi-modal auto-annotating in autonomous driving. *IEEE TIV*, pages 1–15.

Walter Zimmer, Gerhard Arya Wardana, Suren Sritharan, Xingcheng Zhou, Rui Song, and Alois C Knoll. 2024. Tumtraf v2x cooperative perception dataset. In *Proc. of CVPR*, pages 22668–22677.

A Appendix

A.1 Data-driven Autonomous Driving System

Traditional ADS rely on data-driven approaches, which are categorized into modular and end-to-end frameworks (Chen et al., 2024b). **Modular-based systems** break the entire autonomous driving process into separate components, such as *perception module*, *prediction module*, and *planning module*. Perception modules are responsible for obtaining information about the vehicle's surrounding environment, aiming to identify and locate important traffic elements such as obstacles, pedestrians, and vehicles near the autonomous vehicle, usually including tasks such as object detection (Wang et al., 2021) and object occupancy prediction (Tong et al., 2023). Prediction modules estimate the future motions of surrounding traffic participants based on the information provided by the perception module, usually including tasks such as trajectory prediction and motion prediction (Shi et al., 2022). Planning module aims to derive safe and comfortable driving routes and decisions through the results of perception and prediction (Sauer et al., 2018). Each module is individually developed and integrated into onboard vehicles to achieve safe and efficient autonomous driving functions. Although modular methods have achieved remarkable results in many driving scenarios, the stacking design of multiple modules can lead to the loss of key information during transmission and introduce redundant calculations. Furthermore, due to the inconsistency in the optimization objectives of each module, the modular-based system may accumulate errors, which can negatively impact the vehicle's overall decision-making performance. **End-to-end-based systems** integrate the entire driving process into a single neural network, and then directly optimize the entire driving pipeline from sensor inputs to produce driving actions (Chen et al., 2024b). However, this approach introduces the "black box" problem, meaning a lack of transparency in the decision-making process, complicating interpretation and validation.

A.2 LLMs in Autonomous Driving System

As shown in Figure 6, 7, LLMs, with their powerful open-world cognitive and reasoning capabilities, have shown significant potential in ADSs (Yang et al., 2023; Li et al., 2023). LC-LLM (Peng et al., 2024) is an explainable lane change prediction model that leverages LLMs to process driving scenario information as natural language prompts. By incorporating CoT reasoning and supervised finetuning, it not only predicts lane change intentions and trajectories but also provides transparent and reliable explanations for its predictions. GPT-Driver (Mao et al., 2023) regards the motion planning task as a language modeling problem, using a fine-tuned GPT-3.5 model (Ye et al., 2023) to generate driving trajectories. DriveGPT4 (Xu et al., 2024) introduces an interpretable end-to-end autonomous driving system that uses multimodal LLMs to process multi-frame video inputs and textual queries, enabling vehicle action interpretation and low-level control prediction. By employing a visual instruction tuning dataset and mixfinetuning strategy, it provides a novel approach to directly map sensory inputs to actions, achieving superior performance in autonomous driving tasks. Driving with LLM (Chen et al., 2024c) integrates vectorized numeric data with pre-trained LLMs to improve context understanding in driving scenarios and enhances the interpretability of driving decisions.

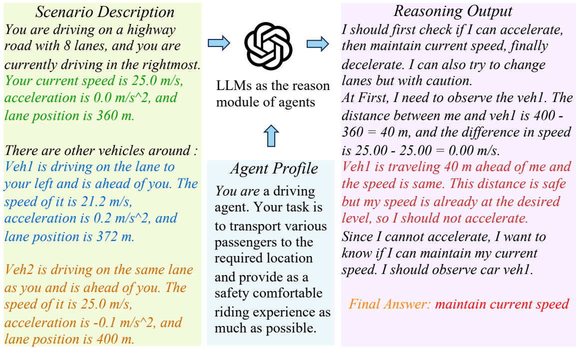

Figure 6: An example of an LLM-based single-agent ADS (Wen et al., 2024).

A.3 Datasets

Single-agent Autonomous Driving Dataset. Single-agent datasets are obtained from a single reference agent, which can be the ego vehicle or roadside infrastructure, using various sensors. Mainstream singel-agent autonomous driving datasets like KITTI (Geiger et al., 2012), nuScenes (Geiger et al., 2020), and Waymo (Sun et al., 2020) provide comprehensive multimodal sensor data, enabling researchers to develop and benchmark algorithms for multiple tasks such as object detection, tracking, and segmentation.

In addition to these foundational datasets, newer ones like BDD-X (Kim et al., 2018), Driv-

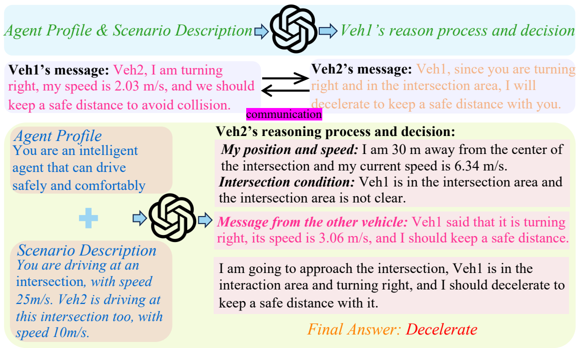

Figure 7: The communication among multiple agents in an LLM-based multi-agent system.

eLM (Sima et al., 2025), and nuScenes-QA (Qian et al., 2024) introduce action descriptions, detailed captions, and question-answer pairs that can be used to interact with LLMs. Combining language information with visual data can enrich semantic and contextual understanding, promote a deeper understanding of driving scenarios, and enhance the safety and interaction capabilities of autonomous vehicles.

Multi-agent Autonomous Driving Dataset. Beyond single-vehicle view datasets, integrating more viewpoints of traffic elements, such as drivers, vehicles and infrastructures into the data also brings advantages to AD systems. Multi-agent autonomous driving datasets, such as DAIR-V2X (Yu et al., 2022), V2XSet (Xu et al., 2022), V2V4Real (Xu et al., 2023), and TUMTraf-V2X (Zimmer et al., 2024) typically include data from multiple vehicles or infrastructure sensors, capturing the interactions and dependencies between different agents and additional knowledge regarding the environments. These datasets are essential for researching and developing cooperative perception, prediction, and planning strategies that enable vehicles to overcome the limitations of single agent datasets such as limited field of view (FOV) and occlusion.