drones

MDPI

Article

Applications of Large Language Models and Multimodal Large Models in Autonomous Driving: A Comprehensive Review

Jing Li [1], Jingyuan Li [2], Guo Yang [3], Lie Yang [4,*], Haozhuang Chi [4] and Lichao Yang [5]

[1] Independent Researcher, Nanchang 330000, China; lijingformal@outlook.com
[2] School of Vehicle and Mobility, Tsinghua University, Beijing 100084, China; lijingyu20@mails.tsinghua.edu.cn
[3] Department of Industrial and Manufacturing Systems Engineering, The University of Hong Kong, Hong Kong 999077, China; yangguo@hku.hk
[4] School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore 639798, Singapore; chih0001@e.ntu.edu.sg
[5] Faculty of Engineering and Applied Science, Cranfield University, Cranfield MK43 0AL, UK; lichao.yang@cranfield.ac.uk
* Correspondence: lie.yang@ntu.edu.sg

Abstract: The rapid development of large language models (LLMs) and multimodal large models (MLMs) has introduced transformative opportunities for autonomous driving systems. These advanced models provide robust support for the realization of more intelligent, safer, and efficient autonomous driving. In this paper, we present a systematic review on the integration of LLMs and MLMs in autonomous driving systems. First, we provide an overview of the evolution of LLMs and MLMs, along with a detailed analysis of the architecture of autonomous driving systems. Next, we explore the applications of LLMs and MLMs in key components such as perception, prediction, decision making, planning, multitask processing, and human–machine interaction. Additionally, this paper reviews the core technologies involved in integrating LLMs and MLMs with autonomous driving systems, including multimodal fusion, knowledge distillation, prompt engineering, and supervised fine tuning. Finally, we provide an in-depth analysis of the major challenges faced by autonomous driving systems powered by large models, offering new perspectives for future research. Compared to existing review articles, this paper not only systematically examines the specific applications of LLMs and MLMs in autonomous driving systems but also delves into the key technologies and potential challenges involved in their integration. By comprehensively organizing and analyzing the current literature, this review highlights the application potential of large models in autonomous driving and offers insights and recommendations for improving system safety and efficiency.

Keywords: large language models; multimodal large models; autonomous driving; comprehensive review

Academic Editor: Pablo Rodríguez-Gonzálvez

Received: 27 January 2025
Revised: 2 March 2025
Accepted: 21 March 2025
Published: 24 March 2025

Citation: Li, J.; Li, J.; Yang, G.; Yang, L.; Chi, H.; Yang, L. Applications of Large Language Models and Multimodal Large Models in Autonomous Driving: A Comprehensive Review. *Drones* **2025**, *9*, 238. https://doi.org/10.3390/drones9040238

1. Introduction

The core objective of autonomous vehicles is to achieve intelligent driving without human intervention [1], a goal that is revolutionizing the traditional human-centric transportation system. As an integrated innovation of artificial intelligence, computer vision, and sensor fusion, autonomous driving technology not only holds the potential to significantly enhance traffic safety but also lays the foundation for the comprehensive upgrade of intelligent transportation systems by improving traffic efficiency and fostering societal convenience [2].

Autonomous driving systems predominantly follow two architectural paradigms: modularized architectures and end-to-end architectures [3]. Modularized architectures divide the system into distinct components like perception, trajectory prediction, planning, and control, with each module addressing a specific task [4]. For example, the perception module processes sensor data to interpret the environment, while the planning module determines the optimal driving path. This design ensures clear functional separation and facilitates collaborative development but can suffer from cumulative errors due to inter-module information exchange. Conversely, end-to-end architectures directly derive driving decisions from sensor inputs, minimizing information loss and enabling superior global optimization [3,5]. Despite these advantages, end-to-end systems are highly data-dependent, requiring extensive high-quality datasets, and their "black-box" nature challenges safety and reliability assurance.

With advancements in deep learning and sensor technology, autonomous driving has shifted from rule-based methods to data-driven approaches. Deep learning enables systems to autonomously learn complex driving behaviors from large-scale datasets, significantly enhancing generalization capabilities [6]. However, these methods remain limited in handling long-tail scenarios, impeding broader adoption. Recently, large language models (LLMs) have introduced new opportunities for autonomous driving. Through self-supervised training on extensive datasets, LLMs exhibit remarkable language generation, reasoning, and contextual learning capabilities, adapting to diverse downstream tasks without fine tuning [7]. Yet, LLMs' reliance on textual input restricts their applicability in vision-centric domains like autonomous driving. Multimodal large models (MLMs) address this limitation by integrating visual data with language models to enhance visual–linguistic interaction and understanding [8]. This integration facilitates complex tasks such as scene perception, path planning, and decision-making, advancing the capabilities of autonomous systems.

LLMs and MLMs, with their notable strengths, are expected to address critical challenges in autonomous driving systems, such as long-tail scenarios and the "black-box" nature of decision making. In perception, LLMs and MLMs, when integrated with sensor data (e.g., bird's eye view (BEV) and LiDAR), enhance the understanding of traffic scenes [9–11]. Moreover, the strong image comprehension and reasoning capabilities of MLMs have been directly applied to scene understanding in autonomous driving, significantly improving system safety [12–14]. In trajectory prediction and planning tasks, LLMs excel due to their contextual learning abilities. Using techniques such as prompt engineering and chain-of-thought reasoning, LLMs can efficiently perform interpretable trajectory prediction [15,16] and path planning [17]. Additionally, to meet user preferences and enhance interaction, autonomous driving systems require transparent decision making and the ability to adjust driving decisions based on user instructions. For instance, Co-Pilot [18] employs GPT3.5 as a human–machine interaction tool, showcasing its effectiveness in intent understanding and task execution, and highlighting the potential of LLMs in improving collaborative driving experiences. By leveraging zero-shot and few-shot learning, LLMs can efficiently parse user commands and determine the requirements of autonomous driving systems, further enhancing interaction efficiency and system safety [19]. Furthermore, employing GPT-4 for sentiment analysis enables more natural and personalized human–machine interactions [20]. Autonomous driving systems based on LLMs and MLMs have also achieved innovative breakthroughs in task processing capabilities, no longer being limited to traditional models that handle only a single driving task. For example, DOLPHINS [21] utilizes OpenFlamingo to unify path planning, control signal generation, and language generation.

The integration of LLMs and MLMs into autonomous driving systems leverages core technologies such as multimodal fusion, knowledge distillation, prompt engineering, and supervised fine tuning. These advancements enhance model adaptability to driving scenarios and improve integration efficiency. LLMs and MLMs showcase their potential by enabling reasoning capabilities across driving modules, assisting task execution, and managing multitasking in end-to-end systems, advancing safety, efficiency, and user experience in autonomous driving. Despite their promise, LLM-based systems face significant challenges. High-quality data are crucial for training, yet balancing data quality with user privacy remains a challenge. Efficient alignment between visual and textual modalities is critical for accurate comprehension and decision making, as misalignment can lead to perception errors and safety risks. The hallucination issue in LLMs, where models generate incorrect outputs, poses severe threats to safety-critical systems, making its mitigation a key research focus. Additionally, adversarial attacks exploiting subtle input perturbations further jeopardize system stability, necessitating robust defense mechanisms to ensure reliability and safety in autonomous driving systems.

Despite numerous challenges, the introduction of LLMs and MLMs has revolutionized autonomous driving systems, driving significant technological advancements in the field. The application of LLMs and MLMs in autonomous driving has recently emerged as a prominent research focus, gaining widespread attention and achieving remarkable results, with related scientific publications growing exponentially [22]. To address this rapidly evolving field, several researchers have conducted reviews, which can be broadly categorized into three types: application-oriented [23,24], background-oriented [25–28], and model-oriented surveys [2,29]. Application-oriented reviews systematically introduce the use of LLMs and MLMs in tasks such as perception, prediction, planning, and control while analyzing techniques like prompt engineering and reinforcement learning. However, they often lack detailed discussions on the developmental trajectory of LLMs and MLMs in autonomous driving, making it difficult to fully grasp their advantages and limitations. Background-oriented reviews focus on foundational knowledge, elaborating on modularized and end-to-end architectures and the core properties of LLMs/MLMs, yet they provide limited insights into practical applications. Model-oriented reviews highlight LLM- and MLM-based models, offering example-driven insights but insufficiently addressing challenges these models face, thereby limiting inspiration for future research directions. While the existing reviews provide valuable perspectives, their varying emphases and limitations leave room for further refinement and comprehensive exploration.

The main contributions of this paper are as follows:

1. Compared to application-oriented reviews, this paper not only systematically categorizes and summarizes the key applications of LLMs and MLMs in the field of autonomous driving but also provides a comprehensive overview of the development history of LLMs, MLMs, and autonomous driving technologies. This study not only helps researchers gain a deeper understanding of the technological evolution of LLMs and autonomous driving from independent development to deep integration but also provides clear guidance on identifying key technological entry points and determining priority development directions.
2. Compared to background-oriented reviews, this paper provides a comprehensive analysis of the core technologies of LLMs and MLMs in autonomous driving systems, including prompt engineering, instruction fine tuning, knowledge distillation, and multimodal fusion, supported by detailed case studies. The results indicate that these technologies enhance model adaptability, system real-time performance, and perception efficiency. This integration of theory and practice facilitates a clearer understanding of technological applications, aids researchers in developing targeted

technical solutions, and promotes the optimization and innovation of autonomous driving technology.

3. Compared to model-oriented reviews, this paper conducts an in-depth exploration of the key challenges faced by autonomous driving systems based on LLMs and MLMs, including hallucination issues, multimodal alignment challenges, training data limitations, and adversarial attacks. These analyses not only highlight the limitations of current technologies but also provide researchers with insights into potential solutions, offering essential theoretical support and practical references to advance the field.

2. The Current Development Status of Large Models

Large models refer to artificial intelligence models trained on extensive datasets with an exceptionally large number of parameters. These models exhibit remarkable generalization capabilities and the potential to adapt to a wide range of downstream tasks [30,31]. The category of large models typically includes LLMs and MLMs. With their outstanding emergent abilities, large models have been widely applied across various domains, including autonomous driving, where they play a critical role in modules such as perception, prediction, planning, and multitask processing [14,16,32,33]. Therefore, gaining a thorough understanding of the developmental trajectory of large models is essential for grasping their technical advancements and application value. In the following sections, we will outline the development histories of LLMs and MLMs.

2.1. The Development of LLMs

Language is a human communication system based on grammatical rules, serving not only as a means of self-expression but also playing a crucial role in facilitating interpersonal communication and human–machine interaction [34,35]. Developing models capable of understanding, generating, and predicting natural language is of great significance, as these models assist humans in performing complex and time-consuming tasks, such as machine translation, text generation, and information retrieval. In recent years, the field of LLMs has developed rapidly, with various types of LLMs emerging, as shown in Figure 1. Early language models were based on rule-based and statistical approaches [36,37], but these methods suffered from poor scalability and high data requirements and were eventually supplanted by neural-network-based methods. Among these, recurrent neural networks (RNNs) [38] and long short-term memory (LSTM) networks [39] emerged as dominant methods for handling sequential data. However, these models faced limitations in capturing long-range dependencies. The introduction of the Transformer [40] revolutionized the field of natural language processing (NLP). By incorporating attention mechanisms, the Transformer efficiently captured global contextual information, significantly enhancing the handling of long-range dependencies. Building on this foundation, the bidirectional encoder representations from Transformers (BERT) [41] established the "pretraining-fine-tuning" paradigm by leveraging large-scale corpus pretraining followed by task-specific fine tuning. This approach greatly improved model performance on new tasks.

The research has shown that increasing the scale of training data and expanding model parameters further enhance model performance. For instance, GPT-2 with 1.5 billion parameters [42] demonstrated exceptional text generation capabilities, while GPT-3 with 175 billion parameters [43] achieved near-human levels of contextual understanding, text fluency, and coherence. These large-scale language models, often referred to as LLMs, exhibit emergent abilities, such as in-context learning and zero-shot learning, enabling them to perform new tasks with specific prompts without task-specific fine tuning, even in the absence of prior exposure to certain task data. The commercialization of the GPT

series has driven the widespread application of these powerful language models across various domains, spurring the development of open-source LLMs. Models such as LLaMA (7B/13B/30B/65B parameters) [44], Vicuna (7B/13B parameters), and LIMA (65B parameters) [45] have gained popularity for their technical characteristics and transparency, making them widely applicable to multilingual, multimodal, and multitask scenarios. These open-source models not only reduce research and development costs but also facilitate a deeper understanding of model architectures and underlying technologies among researchers.

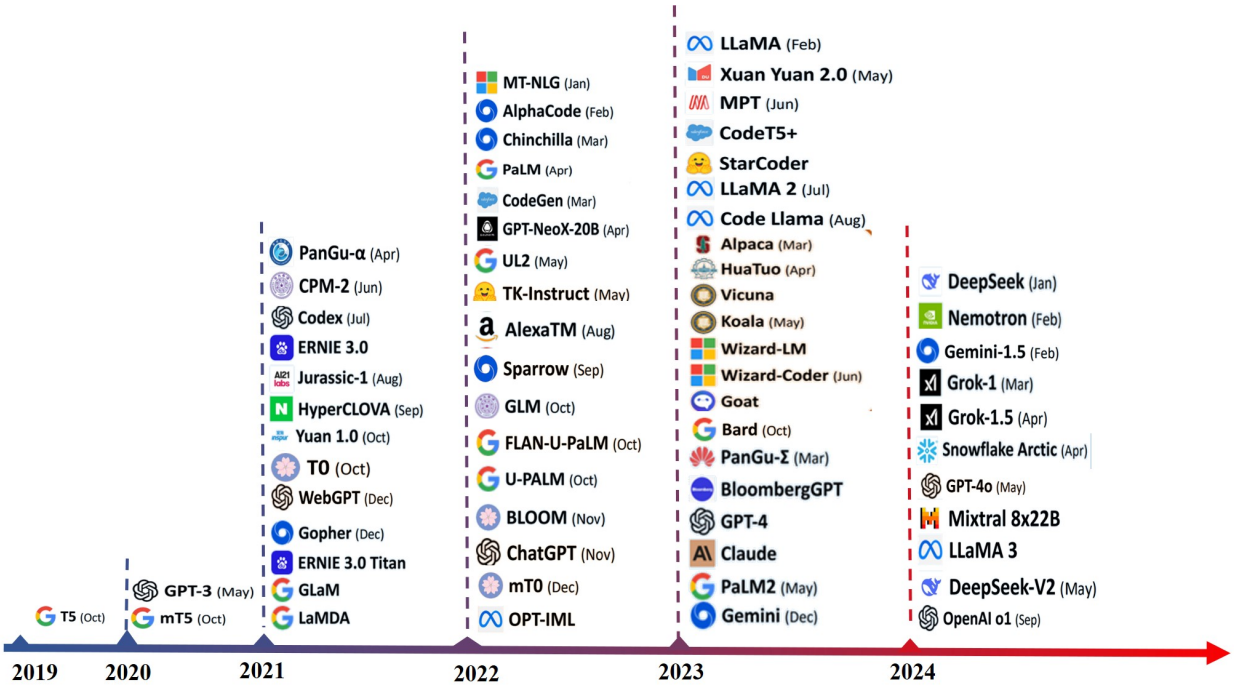


Figure 1. The development timeline of some very representative LLMs. We have categorized and summarized the well-known large language models (LLMs) released globally from 2019 to 2024 based on their publication year, with the icons preceding each model name representing their respective companies or institutions. The analysis reveals a significant surge in the number of LLMs starting from 2021 as numerous enterprises and research institutions worldwide have actively engaged in LLM development and application.

LLMs, with their robust capabilities in language understanding, reasoning, and question-answering, have become core tools for addressing challenges across various domains, including autonomous driving. In autonomous driving systems, LLMs not only enhance perception to effectively handle complex scenarios [10,13,46] but also play a significant role in improving the interpretability of trajectory prediction [15,47,48] and planning [17,49]. By generating natural language explanations, LLMs increase the transparency of system decision making while strengthening human interaction capabilities [18–20,50]. This makes autonomous driving systems more intelligent and aligned with user needs, thereby enhancing user trust and overall experience.

2.2. The Development of MLM

Vision and language are two key modalities for human interaction and problem solving. Advances in vision–language tasks have paralleled progress in their respective domains. Language models have evolved from rule-based and statistical methods to neural networks, culminating in Transformer-based architectures and LLMs. These models transitioned from single-task solutions to mastering reasoning capabilities via the "pretraining-finetuning" paradigm, marking a significant leap in language processing.

Similarly, computer vision has advanced rapidly with deep learning. Convolutional neural networks (CNNs) were once the dominant approach for visual feature processing. However, early vision–language models (VLMs) were primarily task-oriented, relying on simple feature concatenation for cross-modal interaction. Transformers revolutionized natural language processing and extended their success to vision (e.g., ViT [51]) and multimodal models (e.g., ViLT [52]), enhancing cross-modal interaction and representation learning. However, due to the limitations of supervised data, these models still require fine tuning when applied to downstream tasks. Inspired by BERT, CLIP [53] used contrastive learning on 400 million image–text pairs to generate generalized multimodal representations. These models excel in understanding tasks without fine tuning but are limited in generative capabilities. The rise of LLMs has led to integrating vision models with LLMs to achieve multimodal understanding and generation [26]. These models are referred to as MLMs and have experienced rapid development over the past three years. Some representative MLMs are shown in Figure 2.

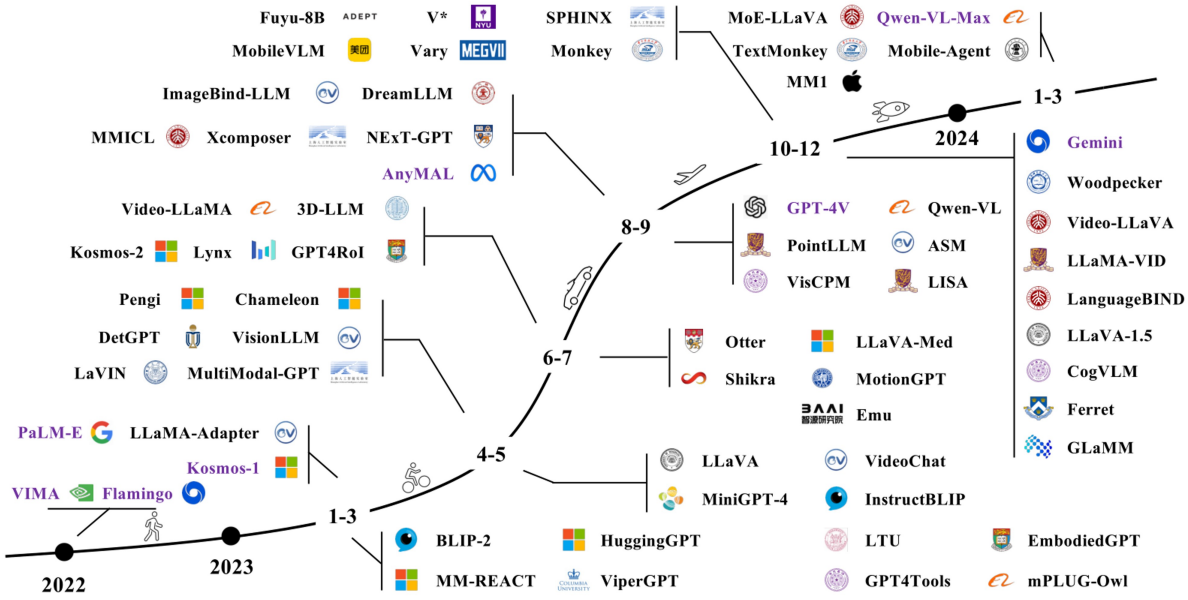


Figure 2. The development timeline of some very representative MLMs. This figure is sourced from reference [54].

A practical approach involves connecting frozen visual encoders to LLMs through intermediate modules, leveraging LLM reasoning and knowledge for enhanced perception. Flamingo [55] connects vision models to LLMs via intermediate modules to avoid costly end-to-end training. BLIP-2 [56] introduces a Query Transformer (Q-Former) to improve cross-modal interaction. Mini-GPT4 [57] utilizes Vicuna to enhance generative capabilities, while LLaVA [58] incorporates visual instructions for better conversational abilities. InternVL [59] addresses the vision–language data gap with its large-scale InternViT-6B encoder, and Video-LLaMA [60] integrates video and audio for broader multimodal understanding. Additionally, commercial models like OpenAI's GPT-4 [61] and Google's Gemini [62] further drive vision–language advancements across industries.

In autonomous driving, systems typically include perception, prediction, and planning modules [23,26,27]. MLMs enhance perception, improving performance in complex scenarios while increasing interpretability and task diversity [8]. MLMs also serve as unified frameworks for multiple driving tasks [63] or as conversational assistants bridging the cognitive gap between systems and human drivers [21].

2.3. Challenges Faced by LLMs and MLMs

Through extensive training on large-scale datasets, LLMs possess a rich repository of world knowledge, mastery of complex linguistic syntax and semantics, and robust contextual reasoning capabilities, enabling them to interact naturally with humans. MLMs extend the applicability of these models from unimodal to multimodal domains, allowing them to integrate information from diverse modalities (e.g., vision, speech, and 3D data) and combine visual information with language modalities to perform cross-modal understanding and generation tasks. Despite the rapid advancements driven by their inherent strengths, these models face several significant challenges:

1. Hallucination Issues: Hallucination refers to the phenomenon where a LLM generates information that appears plausible but is inaccurate or fabricated [64]. This issue poses a substantial threat to high-stakes domains such as autonomous driving, where precision is critical. Addressing hallucination requires developing diverse methods, such as constructing hallucination detection benchmarks and introducing paradigms like reinforcement learning from human feedback (RLHF) or knowledge distillation.
2. Modal Alignment and Fusion Challenges: A substantial gap exists between the pixel-level features of vision and the semantic features of language. This disparity becomes even more pronounced with multimodal data (e.g., video, audio, and language), making alignment across modalities particularly challenging. The degree of modal alignment directly influences model performance, marking it as a critical area of ongoing research.
3. Security and Privacy Concerns: Input data, such as videos and audio, often contain sensitive information, raising significant privacy concerns. Ensuring data security while simultaneously improving model performance is a pressing issue that demands immediate and effective solutions.
4. Explainability issue: LLMs and VLMs, due to their high-dimensional complexity, exhibit black-box characteristics, affecting trust and transparency. Explainable AI (XAI) seeks to unveil their reasoning processes, enhancing explainability and reliability through methods like model-based design, post hoc explanations (e.g., SHAP, LIME), and causal inference. However, XAI faces challenges such as difficulty in explaining complex architectures, trade-offs between explainability and performance, and lack of standardized evaluation criteria. Future research should focus on cross-modal collaboration, improved explainability methods, and standardized evaluation frameworks to enhance AI transparency and usability.

3. Current Development Status and Challenges of Autonomous Driving

3.1. The Development of Autonomous Driving

An autonomous driving vehicle, also known as a self-driving vehicle, represents a groundbreaking innovation in the field of transportation [5,65]. Its origins can be traced back to the 1950s, when basic vehicle control was achieved using radio control and simple mechanical devices, laying the foundation for modern autonomous driving technology.

According to the classification by the Society of Automotive Engineers (SAE), autonomous driving technology is divided into six levels (L0–L5) [66,67], representing the progression from fully manual driving to full autonomy, illustrating the step-by-step evolution of autonomous driving systems. The definitions of each level are as follows:

L0 (No Automation): The vehicle is entirely controlled by the driver with no automation functionality.

L1 (Driver Assistance): The system provides assistance under specific conditions, aiding the driver in performing certain tasks but lacking independent driving capabilities.

L2 (Partial Automation): The system can simultaneously control both lateral (steering) and longitudinal (acceleration and braking) movements; however, the driver must remain alert and be ready to take over at any time.

L3 (Conditional Automation): The system can fully assume driving tasks in specific scenarios, with the driver required to intervene only when prompted.

L4 (High Automation): The vehicle can achieve full autonomy within limited environments (such as specific geographic areas or weather conditions) without requiring driver supervision for safety.

L5 (Full Automation): The vehicle operates independently under all conditions and in any environment, requiring no human intervention.

Early autonomous driving technologies relied on rule-based algorithms, guided by a set of predefined rules constructed from human expertise to inform vehicle decision making. The inaugural autonomous vehicle competition in 2004 marked a significant milestone, catalyzing rapid advancements in this field [68]. In 2009, Google launched its autonomous driving project, signaling the beginning of the technology's journey toward commercialization. Around 2010, breakthroughs in sensor technologies and the rise of deep learning facilitated a paradigm shift from rule-based algorithms to data-driven approaches. These advancements enabled vehicles to learn from large-scale datasets, enhancing their ability to understand and predict environmental changes. During this phase, autonomous driving technology expanded from closed testing environments to open urban roads. Google conducted extensive highway tests, Tesla introduced autonomous driving features, and Uber's autonomous vehicle accident sparked widespread discussions about safety.

As the technology matured, autonomous driving systems evolved from single-scene applications to multi-scene adaptability and from single-modality approaches to multi-modal integration. This evolution, coupled with the development of regulatory frameworks, has propelled the industry's growth [69,70]. In 2018, Waymo launched the world's first commercial autonomous ride-hailing service, marking the onset of the commercialization of autonomous driving technologies. Subsequently, companies such as Baidu Apollo and AutoX introduced autonomous taxi services around 2020. Simultaneously, Vehicle-to-Everything (V2X) technologies were progressively implemented, enabling real-time information sharing between vehicles, infrastructure, pedestrians, and networks, thereby enhancing road safety and traffic efficiency [71].

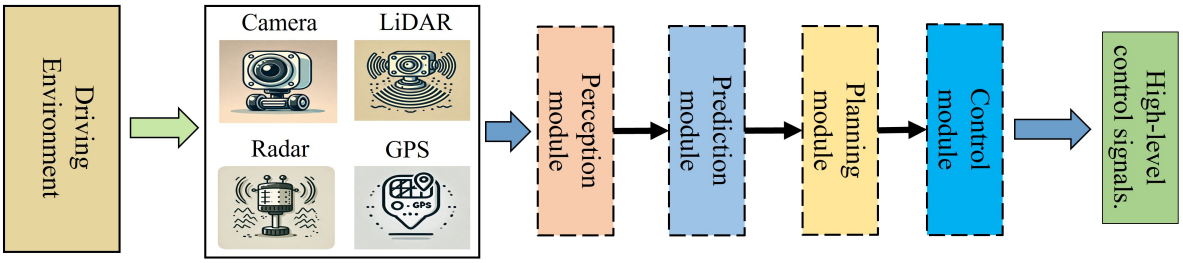
Despite the adaptability and generalization capabilities of data-driven autonomous driving systems, their heavy reliance on data presents significant challenges. These include high data annotation costs, inadequate coverage of long-tail scenarios by limited datasets [3], and the "black-box" nature of deep learning models, which compromises interpretability. Such issues undermine public trust in autonomous driving systems. Future advancements must focus on improving system interpretability, reducing data costs, and addressing performance gaps in long-tail scenarios to enable the widespread adoption and reliable development of autonomous driving technologies.

3.2. System Architectures for Autonomous Driving

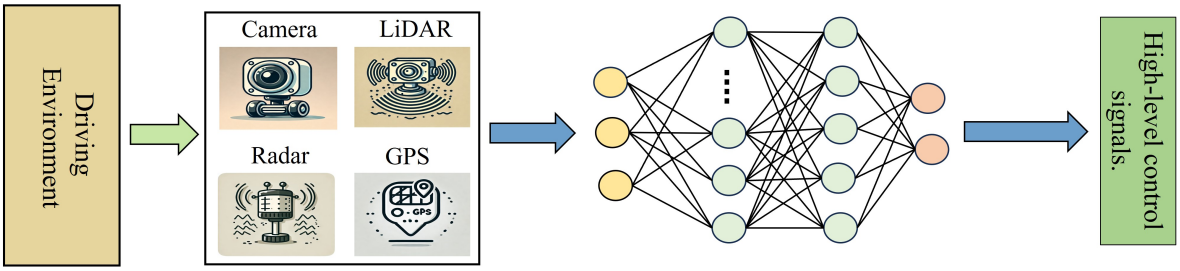
Autonomous driving systems can be categorized into modularized architectures and end-to-end architectures based on their design frameworks. These two approaches differ significantly in terms of system design, functional decomposition, and development philosophy, each offering distinct advantages and disadvantages suitable for varying scenarios and requirements. The choice of architecture plays a critical role in ensuring the safety and reliability of vehicle systems [5]. The following sections provide a systematic overview of modularized and end-to-end architectures in autonomous driving systems.

3.2.1. Modularized Autonomous Driving Methods

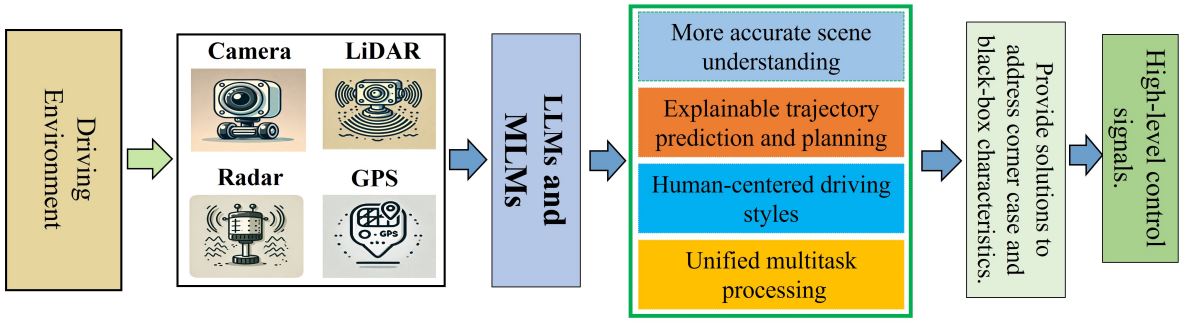
The modularized architecture is a classical design approach for autonomous driving systems, characterized by decomposing the overall driving task into multiple independent functional modules, including perception, prediction, planning, and control (as shown in Figure 3a). Each module is responsible for a specific task and operates sequentially to complete the entire process from perception to control [32]. This architecture offers high interpretability in system design, facilitating development and debugging. However, there are certain limitations. Due to serial computation between modules, the system may suffer from error accumulation and communication delays. Additionally, as each module is optimized independently, achieving global optimality is challenging, potentially leading to suboptimal overall performance.



(a) Modular autonomous driving system



(b) Traditional end-to-end autonomous driving system



(c) Autonomous Driving System with LLMs and MLMs. LLMs and MLMs possess robust comprehension capabilities, enabling them to deeply understand data collected by various sensors. Additionally, they exhibit strong reasoning and decision-making abilities, allowing them to address corner cases based on their understanding of the environment and the powerful reasoning and decision-making skills.

Figure 3. Three common autonomous driving system architectures. (**a**) The architecture of modularized autonomous driving system; (**b**) The architecture of end-to-end autonomous driving system; (**c**) The architecture of autonomous driving system based on LLMs and MLMs.

1. Perception Module

The perception module serves as the foundation of the entire autonomous driving system [26], tasked with collecting and processing environmental information from various sensors such as cameras, LiDAR, and radar. It provides semantic descriptions and geo-

metric information about the static and dynamic surroundings of the vehicle, serving as a critical input source for the decision-making, planning, and control modules. The primary tasks of the perception module include 2D and 3D object detection, semantic segmentation, drivable area segmentation, and multi-object tracking, all aimed at achieving real-time environmental understanding and interaction to ensure safety, comfort, and driving efficiency.

With advancements in sensor technology and deep learning, the perception module of autonomous driving systems has evolved from rule-based single-sensor methods to deep-learning-driven multi-sensor fusion approaches [72]. A single sensor offers limited environmental understanding and cannot capture the intricate relationships between traffic scenes and objects [73]. To address this, researchers integrate data from multiple sensors—such as cameras, LiDAR, and radar—leveraging their complementary advantages to enhance system robustness in complex environments. Among multi-sensor fusion methods, bird's eye view (BEV) representation has emerged as a pivotal approach. BEV transforms data from various sensors into a unified top-down environmental representation, providing a comprehensive and accurate depiction of object spatial relationships. The adoption of deep learning techniques has further accelerated the development of BEV-based perception. Transformer-based BEV models, for example, have demonstrated remarkable performance in perception tasks [27]. From the early stages of detecting obstacles and traffic signs to the current capability of comprehending the entire surrounding environment, the perception module has made significant progress, laying a reliable foundation for autonomous systems to "understand" their surroundings effectively. High-quality perception results are crucial for the safety and reliability of autonomous driving systems, as they directly influence vehicle decision making and behavior.

2. Prediction Module

The prediction module is responsible for forecasting the future behaviors and trajectories of road users, such as pedestrians and vehicles, serving as a critical bridge between the perception and planning modules [26]. Based on the outputs of the perception module, the prediction module extracts features such as the historical trajectories of target objects, the relative position and velocity of the ego vehicle, and intention cues. Using these features, it predicts potential future behaviors (e.g., lane changes or turns) and trajectories of target objects over a given time horizon, assigning probabilities to different trajectories.

Early prediction methods primarily relied on rule-based and physics-based models, which performed well in simple scenarios but exhibited significant limitations in complex environments. With the rise of deep learning, data-driven trajectory prediction methods have become predominant [74,75]. These methods learn complex behavioral patterns from large-scale datasets, offering superior adaptability and robustness. However, the "black-box" nature of deep learning models has reduced the interpretability of trajectory predictions, potentially undermining user trust in the system. To address this issue, researchers have recently introduced LLMs into the prediction module. By generating natural language explanations, LLMs enhance the transparency of prediction results, thereby increasing system reliability and user trust [15,47,76]. Moreover, research focusing on the interactive behaviors among multiple agents has been rapidly advancing, enabling systems to more accurately predict dynamic behaviors in complex traffic scenarios.

3. Planning Module

The planning module is tasked with generating the vehicle's trajectory over a specified time horizon based on the surrounding environment and its own state. It is one of the core components of an autonomous driving system, typically divided into two sub-tasks: path planning and behavioral planning [5]. Path planning involves determining the optimal route from the current position to a target destination, accounting for factors such as safety,

comfort, efficiency, and feasibility. For instance, the trajectory must avoid collisions (safety), produce smooth paths to prevent sudden braking or sharp turns (comfort), ensure timely arrival at the destination while adhering to traffic regulations (efficiency), and be practically executable by the vehicle (feasibility). Behavioral planning, on the other hand, determines specific driving actions, such as lane changes, deceleration, or stopping, to adapt to the current driving environment.

Current planning methods are generally categorized into rule-based approaches [77–79] and learning-based approaches [17,80,81]. Rule-based methods rely on predefined rule sets and human driving experience to guide vehicle behavior. These rules typically incorporate traffic regulations, common driving habits, and safety constraints. The main advantages of rule-based methods include their simplicity, clarity, and high interpretability, making them well-suited for straightforward and routine driving scenarios [77,78]. However, they lack robustness in complex traffic environments and struggle to handle diverse driving conditions and rare edge cases. Additionally, rule-based methods require redesigning rules when encountering new scenarios, resulting in limited adaptability. The advent of deep learning and reinforcement learning has facilitated the development of learning-based planning methods, which can learn complex driving strategies and planning techniques from large-scale data, bypassing the need for manually designed rules [17]. Among these methods, imitation learning and reinforcement learning are the two primary techniques [80]. Imitation learning derives planning strategies by emulating human driving behaviors, leveraging approaches like behavior cloning or inverse reinforcement learning to capture human decision-making logic. While this method is straightforward and computationally efficient, it depends heavily on high-quality demonstration data and is prone to propagating errors from the training data into the model. Reinforcement learning, on the other hand, enables the vehicle to learn optimal strategies through interaction with the environment, utilizing trial and error to explore sophisticated planning methods without requiring manually labeled data [6]. However, this approach imposes high demands on environmental modeling and reward function design, and its training process is computationally intensive.

4. Control Module

The control module acts as the "executor" of the entire autonomous driving system, converting high-level planning instructions into low-level control signals (e.g., acceleration, deceleration, and steering) to ensure the vehicle follows the planned trajectory [32]. The performance of the control module directly impacts the safety, stability, and comfort of the vehicle's operation. Its inputs include the predicted trajectory from the planning module, vehicle state information, and environmental data provided by the perception module. Advances in perception and localization technologies have jointly driven significant progress in the development of control modules.

In modularized autonomous driving systems, the perception, prediction, planning, and control modules are designed independently, offering strong interpretability and facilitating system development and debugging [82]. However, this architecture requires independent designs for each module, increasing the system's overall complexity. Additionally, delays and errors in information transmission between modules can potentially affect real-time performance and decision-making accuracy.

3.2.2. End-to-End Autonomous Driving Methods

End-to-end autonomous driving refers to an approach that directly maps raw sensor data to vehicle control commands (e.g., acceleration, steering, braking). This method integrates perception, prediction, planning, and control into a unified system, optimized in a differentiable manner [3,83] (as shown in Figure 3b). Compared to traditional modularized architectures, end-to-end approaches address autonomous driving tasks in a more

direct manner, avoiding the error accumulation and inefficiencies caused by inter-module information transfer.

Early attempts at end-to-end autonomous driving employed simple neural networks to learn driving behaviors. In 1989, the ALVINN system demonstrated the capability to map camera data to steering controls, achieving end-to-end autonomous driving. However, due to the limited training data, the ALVINN system exhibited poor generalization ability and could only operate in specific scenarios. The rise of deep learning and advancements in computing hardware have brought new possibilities to end-to-end autonomous driving. In 2016, NVIDIA introduced DAVE-2, which used convolutional neural networks (CNNs) to map camera images to steering angles. DAVE-2 achieved impressive results in simulated environments, demonstrating the potential of deep learning in end-to-end autonomous driving [32]. This phase of learning primarily relied on imitation learning, where models were trained by mimicking human drivers' behaviors. However, this process required extensive driving data, making the models susceptible to data biases and limiting their generalization capabilities. Moreover, relying on a single camera restricted the richness of scene information, leading to less accurate driving decisions. To address these limitations, the integration of multi-sensor data, including cameras, LiDAR, radar, and GPS, has enhanced the models' environmental perception capabilities [84]. Additionally, the creation of multimodal public datasets has further advanced research in end-to-end autonomous driving. To overcome the limitations of imitation learning, reinforcement learning and self-supervised learning have been introduced into end-to-end autonomous driving systems. These approaches reduce the dependence on large amounts of labeled data and improve adaptability to complex environments. Recently, the rapid development of LLMs has enabled their powerful language understanding and reasoning capabilities to be widely applied in end-to-end autonomous driving systems. This integration not only enhances the system's interpretability but also improves the vehicle's driving style, making autonomous driving behavior more aligned with that of human drivers.

Compared to traditional modularized autonomous driving systems, the end-to-end approach operates as a unified system that is fully differentiable and optimized exclusively for the final task. This clear and direct optimization avoids the issues of error accumulation and inefficiencies caused by information transfer between modules [85]. However, end-to-end systems heavily rely on large amounts of training data, posing significant challenges in constructing high-quality datasets [86]. Furthermore, due to the "black-box" nature of end-to-end models, their interpretability is relatively weak, making fault diagnosis and system optimization more difficult.

3.2.3. Current Limitations and Challenges

Autonomous driving systems have undergone a complex and winding journey, evolving from early conceptual proposals by research institutions to ongoing commercial exploration. Advances in computing hardware, sensor technology, deep learning, and large-scale annotated data have brought transformative progress to the field of autonomous vehicles. Currently, autonomous driving technology has achieved significant developments at Level 2 and certain Level 3 capabilities [1]. However, the comprehensive realization of Level 4 and Level 5 autonomy continues to face multiple challenges related to technology, cost, and regulatory frameworks [65]. The primary challenges include the following:

1. Limitations of Perception Systems

Extreme weather conditions (e.g., heavy rain, snowstorms, dense fog) and rare long-tail scenarios can significantly degrade or even completely impair the performance of perception systems. This impacts the depth and accuracy of the system's understanding

of driving scenes. Enhancing the robustness of perception systems to handle diverse and extreme environments remains a critical technical challenge.

2. Insufficient Real-Time Decision Making

Autonomous driving systems must make rapid decisions in dynamic driving scenarios. However, the complex architectures of existing models often struggle to meet real-time requirements. Developing lightweight, efficient model architectures to improve system responsiveness and computational efficiency is essential for enabling real-time decision making.

3. Limited Scale and Diversity of Publicly Available Annotated Datasets

Data-driven autonomous driving systems rely heavily on large, high-quality annotated datasets. However, the scale and diversity of publicly available datasets are significantly inferior to those of mainstream vision–language datasets. Additionally, the high cost of data annotation and the heavy reliance on human resources impose substantial burdens on system development. Future research should focus on developing low-cost annotation technologies or leveraging automated annotation methods to facilitate the creation of high-quality datasets, supporting broader application scenarios.

4. Ethical and Legal Challenges

Autonomous driving systems collect vast amounts of user data during operation, raising sensitive issues concerning privacy protection. Clear mechanisms for safeguarding privacy need to be established. Furthermore, in cases of accidents involving autonomous vehicles, the determination of liability is complex and sensitive, as existing legal frameworks are not yet fully adapted to the rapid development of autonomous driving technology. Establishing comprehensive certification systems and legal regulations to standardize the application and promotion of autonomous driving technologies is an urgent societal necessity.

The future development of autonomous driving technology is both filled with opportunities and fraught with challenges. Achieving breakthroughs in technology and advancing societal standardization must proceed in parallel to drive the comprehensive realization of Level 4 and Level 5 autonomy. To address these challenges, some researchers have attempted to develop autonomous driving systems based on LLMs and MLMs to advance the comprehensive realization of higher-level autonomous driving (as shown in Figure 3c).

4. Applications of Large Models in the Field of Autonomous Driving

With significant breakthroughs achieved by LLMs and MLMs across various tasks, some researchers have explored their application in the field of autonomous driving to advance the technology to higher levels. This year has seen a surge in studies leveraging LLMs and MLMs for autonomous driving. In Table 1, we provide a systematic review and summary of some representative studies in this area. In this section, we will systematically introduce the applications of LLMs and MLMs in perception, prediction, decision making and planning, multitasking, and scenario generation.

Table 1. Some representative studies on the application of LLMs and MLMs in the field of autonomous driving.

Models	Year	Backbones	Tasks	Descriptions
LiDAR-LLM [10]	2023	LLaMA2-7B	Perception	The integration of LLM enables the model to comprehend and reason about 3D scenes, while also generating rational plans and explanations.
Talk2BEV [9]	2023	Flan5XXL and Vicuna-13b	Perception	The model utilizes MLM to generate textual descriptions for each object in the BEV map, creating a semantically enhanced BEV representation, which is then input into the LLM. Through prompt engineering, this enables understanding of complex scenes.
BEV-TSR [11]	2024	LIama and GPT-3	Perception	The semantically enhanced BEV map, along with prompts constructed from a knowledge graph, is fed into the LLM to enhance the model's understanding of complex scenes.
OmniDrive [46]	2024	LLaVA v1.5	Perception	Compress multi-view high-resolution video features into a 3D representation, and then input it into an LLM for 3D perception, reasoning, and planning.
DRIVEVLM [13]	2024	Qwen-VL	Perception	DriveVLM directly leverages large multimodal models to analyze images in driving scenarios, providing outputs such as scene descriptions, scene analyses, and planning results. DriveVLM-Dual combines traditional 3D perception with MLM to compensate for the spatial reasoning and real-time inference limitations of MLM.
GPT4V-AD [12]	2023	GPT-4V	Perception	The powerful image understanding capabilities of GPT-4V are applied to autonomous driving perception systems, aiming to evaluate its comprehension and reasoning abilities in driving scenarios, as well as its capacity to simulate driver behavior.
CarLLaVA [14]	2024	LLaVA and LLaMA	Perception	The model optimizes longitudinal and lateral control performance in autonomous driving systems by integrating high-resolution visual encoding with a semi-disentangled output representation, thereby enhancing both scene understanding and control capabilities.
LC-LLM [15]	2024	Llama-2-13b-chat	Prediction	Driving scenes are transformed into natural language prompts and input into LLMs for intention prediction and trajectory forecasting. By incorporating chain-of-thought reasoning, the interpretability of the predictions is further enhanced.
LLM-PCMP [76]	2024	GPT-4V	Prediction	Structured traffic scene information is transformed into visual prompts and combined with textual prompts, which are then input into GPT-4V. The model outputs understanding of the driving scene, which is utilized to enhance traditional motion prediction.
LG-Traj [47]	2024	Not clearly	Prediction	LG-Traj leverages LLMs to extract motion cues from historical trajectories, facilitating the analysis of pedestrian movement patterns (e.g. linear motion, curvilinear motion, or stationary behavior).

Table 1. *Cont.*

Models	Year	Backbones	Tasks	Descriptions
Traj-LLM [16]	2024	GPT-2	Perception	Traj-LLM first transforms agent and scene features into a format understandable by LLMs through sparse context joint encoding and leverages parameter-efficient fine-tuning (PEFT) techniques to train the model for trajectory prediction tasks.
GPT-Driver [17]	2023	GPT-3.5	Decision Making and Planning	The vehicle's self-information, perception results, and prediction information are converted into linguistic prompts to guide GPT-3.5 in generating trajectory predictions based on natural language descriptions.
LLM-ASSIST [87]	2023	GPT-3 and GPT-4	Decision Making and Planning	The model leverages LLMs to address the limitations of rule-based planners by generating safe trajectories or providing optimized parameters in scenarios where the proposals from the planners fail to meet safety thresholds.
VELMA [48]	2023	GPT-3 and GPT-4	Decision Making and Planning	VELMA leverages the contextual learning capabilities of LLMs to interpret navigation instructions and associate them with identified landmarks (e.g., cafes or parks), thereby enabling the system to make informed and rational decisions.
DaYS [33]	2023	GPT-4	Decision Making and Planning	The vehicle's state and perception data, along with the driver's natural language commands, are input into a large language model (LLM) to generate real-time planning decisions.
DiLu [88]	2024	GPT-4 and GPT-3.5	Decision Making and Planning	In addition to utilizing the LLM as the core decision-making tool, the model incorporates a reflection module and a memory module, which are, respectively, used for evaluating and improving decisions and storing experiences.
MTD-GPT [89]	2023	GPT-2	Decision Making and Planning	MTD-GPT leverages the sequence modeling capabilities of GPT to represent each driving task (e.g., left turn, straight, right turn) as sequential data. By predicting future actions based on historical state sequences, it aims to address the multitask decision-making challenges faced by autonomous vehicles at unsignalized intersections.
RRaR [90]	2023	GPT-4	Decision Making and Planning	An LLM is employed as the decision-making core, taking perception data and user instructions as inputs and incorporating chain-of-thought reasoning to generate interpretable dynamic driving plans.
EoLLM [91]	2023	GPT-4 and LLaMA	Decision Making and Planning	Simulation experiments are conducted to evaluate the spatial awareness and decision-making capabilities of the LLM, as well as its ability to comply with traffic regulations.
DOLPHINS [21]	2023	OpenFlamingo	Multitasking	The visual encoder in a vision–language model is utilized to process video information from driving scenarios, while the text encoder handles textual instructions and manages the interaction between visual and textual features. Through training, the model enables multiple tasks, such as interpretable path planning and control signal generation.

Table 1. *Cont.*

Models	Year	Backbones	Tasks	Descriptions
EMMA [92]	2024	Gemini	Multitasking	Non-sensor data are converted into natural language and combined with data from cameras as input to the Gemini model. Different branches and loss functions are designed for various tasks, enabling the model to handle 3D object detection, road map estimation, and motion planning tasks through training.
TrafficGPT [19]	2024	GPT-3.5 Turbo	Multitasking	The model extracts information from multimodal traffic data and leverages an LLM to decompose complex user instructions into multiple subtasks, mapping them to specific traffic foundation models (TFMs). The system dynamically allocates the required TFMs as needed, thereby avoiding redundant calls and resource conflicts.
ESR [93]	2024	GPT-4	Multitasking	The model decomposes the hazard analysis and risk assessment (HARA) process into multiple subtasks, including scene description, hazard event definition, and safety goal generation. Each subtask leverages specific prompt designs to optimize the output of the LLM.
LMDrive [94]	2024	LLaMA	Multitasking	LMDrive is a closed-loop, end-to-end autonomous driving framework. It continuously processes vehicle states, sensor data, and environmental information, leveraging a large language model (LLM) for real-time interpretation and prediction of control signals to execute corresponding actions.
DriveGPT4 [8]	2024	LLaMA2	Multitasking	Video and textual data, after being processed, are input into the LLM, which interprets dynamic scenes based on the input and performs behavior reasoning and control signal prediction.
DKD [95]	2023	GPT-3.5	Scenarios generation	The model is the first to utilize an LLM for automating the construction of driving scene ontologies. Through interactions with ChatGPT, it facilitates the definition of concepts, attributes, and relationships in the autonomous driving domain, thereby enabling ontology construction.
TARGET [96]	2023	GPT-4	Scenarios generation	The model leverages an LLM to interpret complex traffic rules, with the parsed information used to generate templated scenario scripts. These scripts are executed in a simulator to test the system for issues such as rule violations, collisions, or timeouts.
ADEPT [97]	2023	GPT-3	Scenarios generation	ADEPT extracts information from real-world traffic accident reports, utilizing GPT-3's question-answering capabilities to translate the reports into structured data, which are then combined with scenario templates to generate test scenarios.

Table 1. *Cont.*

Models	Year	Backbones	Tasks	Descriptions
OmniTester [98]	2024	GPT-3.5 and GPT-4	Scenario generation	OmniTester is a driving scenario generation model. After the LLM parses the user's natural language description into a structured scene representation, the retrieval module matches traffic network regions based on preprocessed map data and converts the data into an XML format compatible with simulation platforms.
LCTGen [99]	2023	GPT-4	Scenario generation	LCTGen is the first scenario generation model based on LLMs, capable of receiving natural language inputs and generating dynamic traffic scenarios.
TransGPT [100]	2024	Vicuna	Scenarios generation	TransGPT is a traffic-specific large model, where the single-modal TransGPT-SM is designed to answer traffic-related questions, and the multi-modal TransGPT-MM can process image–text inputs to generate traffic-related textual outputs.

4.1. Large Models for Perception

Perception is the core of autonomous driving systems, responsible for processing sensor data from cameras, LiDAR, and other sources to provide critical information for planning and decision-making. It directly impacts the system's safety and reliability. However, traditional perception methods often fall short in scene understanding and object relationship modeling, while lacking interpretability, which undermines user trust. With the development of LLMs and MLMs, their extensive world knowledge, robust reasoning capabilities, and contextual learning abilities have been integrated into perception modules, significantly enhancing the understanding of complex scenes [24,25,101].

Compared to 2D information, 3D scenes provide richer spatial information, which is crucial for perception and decision-making in autonomous driving systems. However, current MLMs primarily rely on 2D image inputs, which significantly limits their ability to understand 3D scenes. To address this limitation, LiDAR-LLM [10] proposed a framework that integrates LLMs with 3D LiDAR data, the network structure of this method is shown in Figure 4a. This framework extracts 3D voxel features using a LiDAR encoder, generates bird's-eye view (BEV) features, and maps these BEV features into an LLM-compatible semantic space using a Visual Alignment Transformer (VAT), achieving effective multimodal feature alignment. The framework employs a three-stage training process to progressively enhance the model's ability to describe 3D scenes, understand object positions and semantic relationships, and perform complex reasoning tasks. This approach enables efficient LLM-based comprehension of 3D point cloud data, pioneering a novel pathway for 3D scene understanding in autonomous driving.The experimental results validate the effectiveness of this method. In the 3D captioning task, LiDAR-LLM achieved a BLEU-4 score of 19.26%, significantly outperforming Mini-GPT4 (2.63%) and LLaMA-AdapterV2 (7.45%). In the 3D grounding task, LiDAR-LLM attained an ACC-5 accuracy of 63.1%, which is seven times that of Mini-GPT4 and five times that of LLaMA-AdapterV2. These results demonstrate the superior performance of LiDAR-LLM in 3D semantic understanding tasks, highlighting its potential for applications in autonomous driving perception systems.

Additionally, converting multi-sensor data into BEV representations is a common perception method. However, traditional BEV maps lack semantic information, limiting their capability to deeply understand complex scenes. To address this, Talk2BEV [9] leverages MLMs to generate semantically enhanced BEV maps. By creating textual descriptions for each object, it enriches both geometric and semantic information, enabling the system to interpret natural language queries and respond to visual and spatial questions. Talk2BEV does not require retraining or fine tuning of models, is compatible with various MLMs, and offers excellent adaptability and task scalability. Similarly, BEV-TSR [11] focuses on enhancing the semantic capabilities of BEV representations but employs a different approach, the network structure of this method is shown in Figure 4c. It achieves deep fusion of BEV features and language embeddings through two methods: shared feature space alignment and text-generation-based alignment enhancement. Additionally, it incorporates knowledge graphs as structured prompts to further improve LLMs' understanding of scene-specific semantics. The experimental results demonstrate that on the nuScenes-Retrieval Easy dataset, BEV-TSR achieved a Top-1 accuracy of 85.78% in the scene-to-text retrieval task, significantly outperforming SigLIP-Base in front view (36.84%) and surrounding view (43.33%). This method effectively enhances the semantic representation of BEV, leading to significant improvements in text-to-scene retrieval, semantic reasoning, and scene understanding tasks, thereby providing an innovative solution for semantic-driven perception and decision-making in autonomous driving.

Several innovative approaches have been proposed to enhance perception and reasoning capabilities in autonomous driving systems. OmniDrive [46] utilizes a sparse query

mechanism to compress multi-view high-resolution video features into 3D representations, which are then processed by an LLM for 3D perception, reasoning, and planning. This method effectively extends the 2D reasoning capability to 3D scenarios. The model is first pre-trained on 2D tasks to initialize the visual-language alignment module (Q-Former) and subsequently fine tuned on 3D driving tasks, significantly improving semantic understanding and interpretable reasoning in dynamic environments. Addressing the limitations of traditional 3D perception methods in handling long-tail scenarios and fine-grained semantic relationships, DRIVEVLM [13] integrates the visual understanding and reasoning capabilities of MLMs, employing a chain-of-thought (CoT) mechanism to semantically encode complex scenarios; the framework of this method is shown in Figure 4b. Through critical object analysis and hierarchical planning, it improves planning accuracy. Additionally, the proposed DriveVLM-Dual system combines traditional 3D perception with MLMs, mitigating MLMs' shortcomings in spatial and real-time reasoning to achieve efficient driving in complex environments.

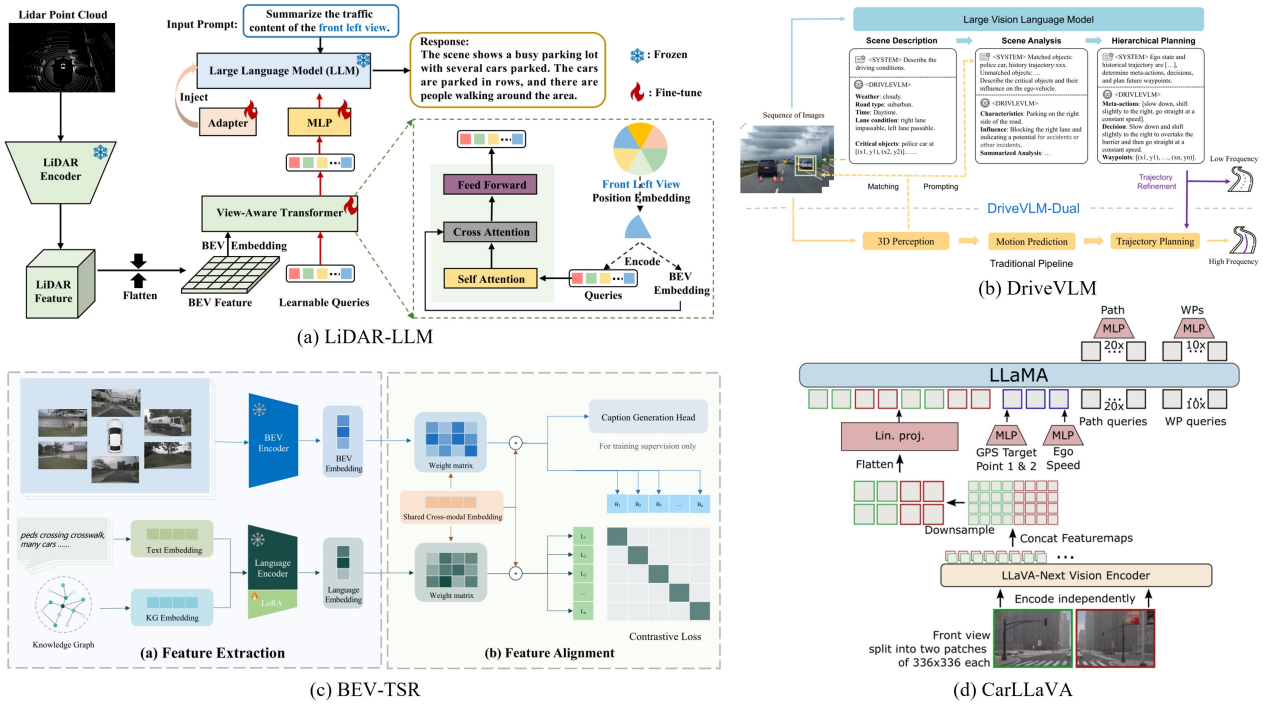


Figure 4. Some representative studies related to the application of LLMs and MLMs in perception.

In terms of perception capabilities, GPT-4V [12] demonstrates outstanding image understanding performance, being applied to scenario comprehension, reasoning, and simulated driver behavior testing. It excels in identifying time, weather, and traffic participant behaviors and surpasses traditional methods in causal reasoning and basic scene processing, though further improvement is needed for dynamic scenarios. Future research should focus on enhancing multi-modal data fusion and temporal reasoning to optimize performance in complex driving scenarios.

Furthermore, while multi-sensor data fusion can achieve high-precision perception, it incurs substantial resource costs. Single-camera data, although economical, have limited capture capability. To address this, CarLLaVA [14] introduces an end-to-end closed-loop driving system relying solely on camera input; the network structure of this method is shown in Figure 4d. The system employs the LLaVA-NeXT framework to integrate high-resolution image features with the generative capabilities of language models for lateral and longitudinal control, significantly improving driving performance in complex scenarios

while eliminating reliance on expensive sensors and labeled data. These studies provide new technological pathways for multi-modal fusion, semantic understanding, and the efficient implementation of autonomous driving systems.

In summary, the integration of LLMs and MLMs with autonomous driving perception in recent years has provided multiple innovative approaches for understanding complex scenarios. From the 3D point cloud comprehension in LiDAR-LLM to the BEV representation enhancements in Talk2BEV and BEV-TSR [9–11], and further to the advancements in multimodal and semantic reasoning explored by methods such as OmniDrive, DRIVEVLM, and GPT-4V [12–14], these studies demonstrate that the incorporation of LLMs and MLMs is significantly enhancing the semantic understanding and decision-making capabilities of autonomous driving systems.

4.2. Large Models for Prediction

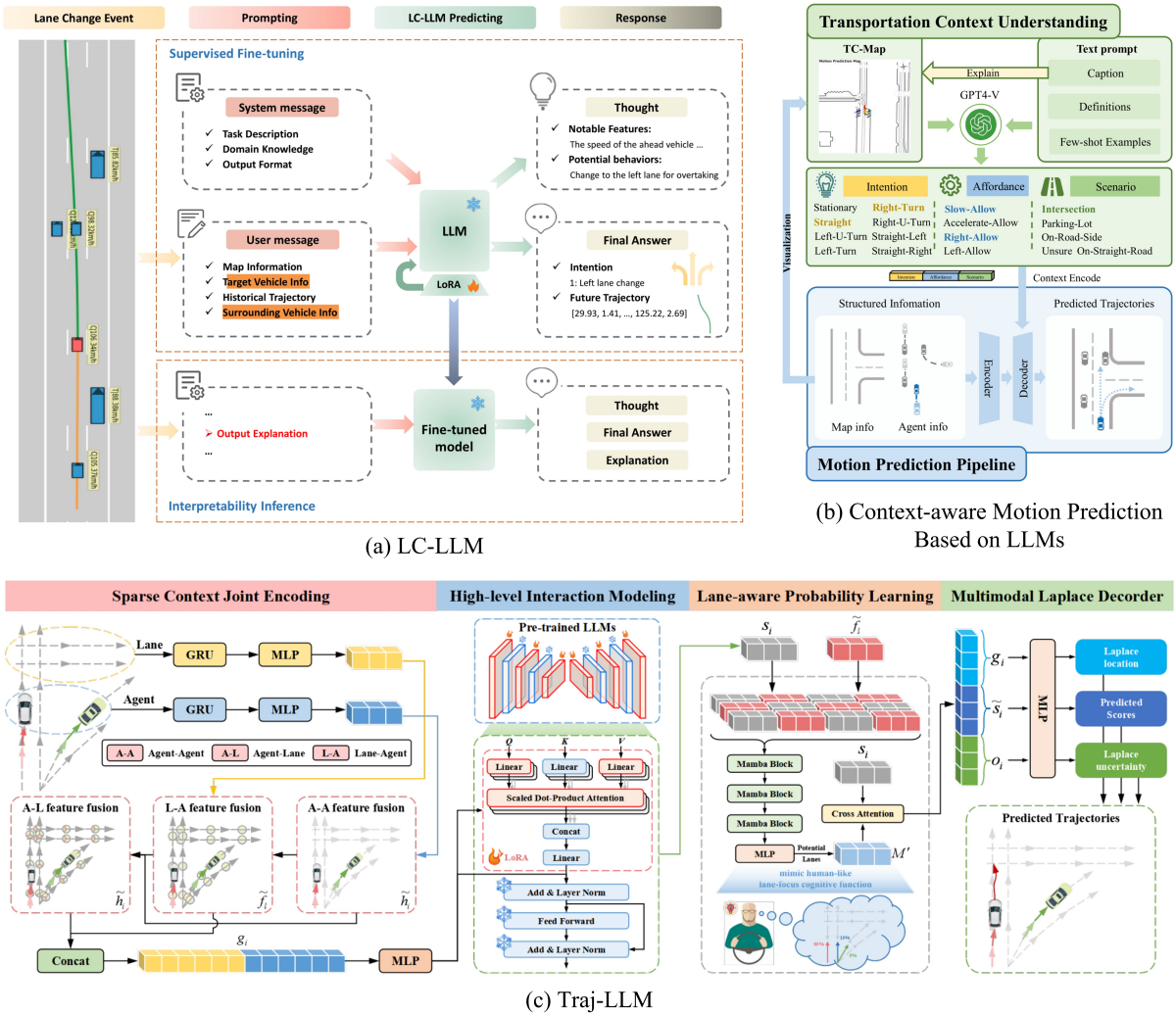
Prediction is a critical module in autonomous driving systems that forecasts the future trajectories of surrounding vehicles and pedestrians over a given time horizon. By taking perception information and the vehicle's state as inputs, it outputs the behavioral intentions (e.g., lane-changing, driving straight, stopping) and potential trajectories of dynamic participants. This provides essential references for decision making and planning, enhancing the vehicle's capability to handle complex scenarios [15,16,47,76].

Trajectory prediction based on deep learning has achieved significant progress; however, these methods often lack interpretability. Given their robust understanding and reasoning abilities, LLMs are promising candidates for enhancing trajectory prediction modules in autonomous driving. Peng et al. [15] were the first to propose using LLMs for intention prediction and trajectory prediction in autonomous driving, redefining the lane-change prediction task as a language modeling problem; the framework of this method is shown in Figure 5a. Driving scene information was transformed into natural language prompts and input into an LLM (Llama-2-13b-chat). Subsequently, the LLM was fine tuned using the low-rank adaptation (LoRA) strategy and supervised fine tuning to better adapt and understand driving scene information for lane-change prediction tasks. To enhance interpretability, a chain-of-thought (CoT) mechanism was introduced during the fine-tuning stage, enabling the generation of intermediate reasoning steps for prediction results. The experimental results demonstrated that the LC-LLM model outperformed baseline models in lane-change intention prediction, lateral trajectory prediction, and longitudinal trajectory prediction. This work provides novel ideas and methodologies for developing prediction modules in autonomous driving systems.

Unlike LC-LLM, which uses LLMs as the core prediction model, ref. [76] proposes an assistive approach that employs MLMs, such as GPT-4V, to understand traffic environments and provide contextual information, thereby enhancing the performance of traditional motion prediction models. The framework of this method is shown in Figure 5b. The authors designed a method to visualize structured traffic scene information, such as vector map data and historical trajectories, generating image prompts in the form of traffic context maps (TC-Maps) and corresponding text prompts. These prompts are input into GPT-4V for scene understanding, producing enriched environmental information, including behavioral intentions, which are then integrated with a classical motion prediction model, Motion Transformer (MTR). By organically combining contextual information with motion prediction, the proposed model significantly improves trajectory prediction accuracy.

Pedestrian trajectory prediction is a critical task in autonomous driving planning, but the diversity of pedestrian behaviors and the complexity of their interactions with the surrounding environment make it highly challenging. LG-Traj [47] introduces a framework that leverages LLMs to extract motion cues from historical trajectories to aid in analyzing

pedestrian movement patterns, such as linear motion, curvilinear motion, or stationary behavior. A Gaussian mixture model (GMM) is used to cluster potential future trajectories and identify probable motion patterns. The LLM then generates detailed motion cues based on these clusters, which are used to predict multiple future paths. These cues provide deeper motion context to the model, significantly enhancing the accuracy of pedestrian trajectory prediction and improving adaptability to diverse movement patterns.



(a) LC-LLM

(b) Context-aware Motion Prediction Based on LLMs

(c) Traj-LLM

Figure 5. Some representative studies related to the application of LLMs and MLMs in prediction.

Traditional LLM-based trajectory prediction models often rely on prompt engineering to adapt to trajectory-related tasks. Traj-LLM [16] introduces an innovative approach that directly leverages the reasoning capabilities of LLMs for trajectory prediction, eliminating the need for traditional prompt engineering and providing a more versatile and easily adaptable solution; the network structure of this method is shown in Figure 5c. Traj-LLM employs a sparse context joint encoding mechanism to transform agent and scene features into formats understandable by LLMs and uses parameter-efficient fine-tuning (PEFT) techniques to train the model for trajectory prediction tasks. Another significant innovation of Traj-LLM is the introduction of the Mamba module, inspired by human driving experience. This module learns the probabilistic distribution of lane positions, aiding the model in generating lane-consistent motion states, thereby significantly enhancing its understanding and prediction capabilities in complex scenarios. The experimental results demonstrate that Traj-LLM excels in few-shot learning, outperforming baseline methods relying on

complete datasets even when trained with only 50% of the data, highlighting its strong generalization ability.

The aforementioned models demonstrate the extensive applications of LLMs in trajectory prediction for autonomous driving, encompassing the development of core predictive models, enhancement of auxiliary modules, modeling the diversity of pedestrian trajectories, and optimization of few-shot learning. These studies provide novel insights and technical pathways for the prediction modules in autonomous driving, significantly advancing the field of trajectory prediction.

4.3. Large Models for Decision Making and Planning

Decision making and planning are core modules of autonomous driving systems, responsible for formulating driving strategies and generating specific operational paths based on information provided by perception and prediction modules. These modules ensure the vehicle can complete driving tasks safely and efficiently [102,103].

Traditional rule-based planning methods exhibit poor adaptability, making them less effective in handling extreme driving conditions. Data-driven methods, while capable of addressing more complex scenarios, often lack interpretability and transparency. Leveraging the exceptional understanding and reasoning capabilities of LLMs, GPT-Driver [17] introduces an innovative approach that redefines motion planning as a language modeling problem (as shown in Figure 6a). This approach converts vehicle-specific information, perception results, and prediction data into language prompts to guide GPT-3.5 in generating trajectory predictions described in natural language. During training, prompt engineering is employed to transform raw data into text formats that LLMs can comprehend. A chain-of-thought reasoning mechanism is used to analyze key obstacles, enabling the derivation of reasonable driving decisions. Experimental results demonstrate that this method not only significantly improves planning performance but also enhances decision interpretability. In the trajectory prediction task, it achieves a 79.1% reduction in error compared to ST-P3, while substantially lowering the collision rate, showcasing more precise trajectory forecasting and enhanced safety robustness. These findings further validate the superior advantages of this approach in autonomous driving motion planning and provide a promising direction for developing more interpretable intelligent driving decision-making systems.

Unlike GPT-Driver, which relies entirely on LLMs for planning, Sharan et al. [87] proposed a hybrid planning approach that combines traditional rule-based planners with the reasoning capabilities of LLMs. Rule-based methods perform reliably in conventional scenarios but exhibit limitations in complex environments. The authors utilize LLMs to address the shortcomings of rule-based planners by generating safe trajectories or optimizing parameters when the proposals from the rule-based planner fail to meet safety thresholds. This method combines the stability of rule-based approaches with the flexibility of LLMs, significantly outperforming single-method approaches in complex scenarios, providing a more reliable solution for autonomous driving planning.

Navigation, often regarded as high-level or global planning, is a critical component of autonomous driving systems. Schumann et al. [48] introduced VELMA, the first city-scale vision-and-language navigation (VLN) system based on LLMs. VELMA leverages the contextual learning capabilities of LLMs to interpret navigation instructions and associate them with identified landmarks (e.g., cafes or parks), enabling the system to make informed decisions. By integrating the reasoning abilities of LLMs with visual inputs, VELMA offers a novel solution to complex urban navigation tasks, demonstrating the potential of embodied agents in real-world applications.

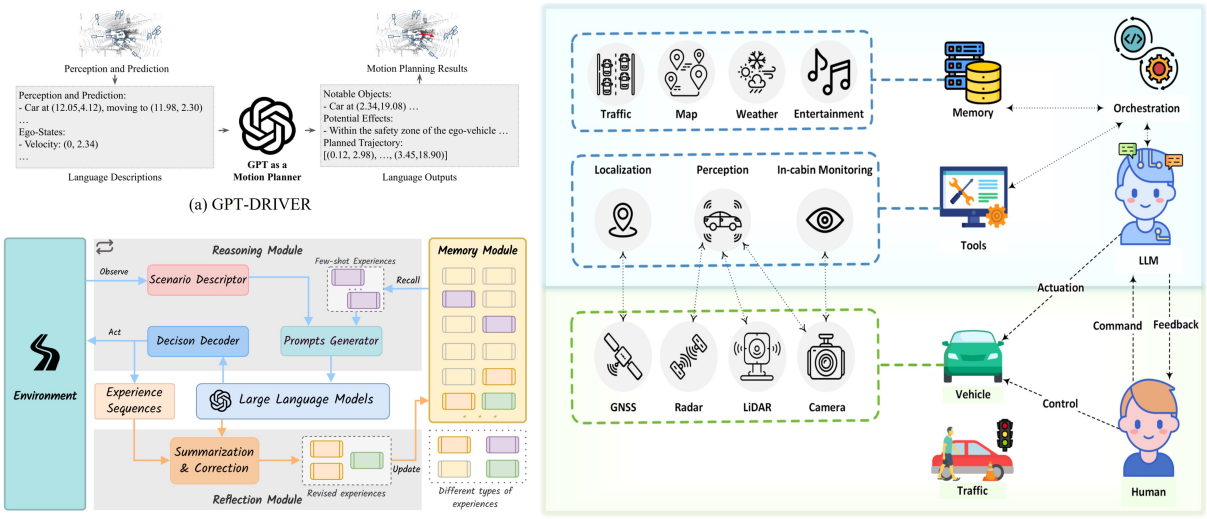


Figure 6. Some representative studies related to the application of LLMs and MLMs in decision making and planning.

Additionally, Cui et al. [33] proposed a framework that integrates LLMs into autonomous driving systems, positioning them as the "decision-making brain" (as shown in Figure 6c). Combined with the perception module ("eyes") and control module ("hands"), the LLM not only processes natural language commands from the driver but also generates real-time planning decisions by incorporating vehicle state and perception data. Furthermore, the system enhances transparency and user trust by explaining its decision logic in natural language, offering a novel solution for human–machine collaborative autonomous driving systems.

Similarly, leveraging LLMs as the "decision-making brain", ref. [90] introduced a chain-of-thought reasoning approach to interpret user instructions and integrate them with perception data to generate dynamic driving plans. This method enables LLMs to intelligently assess the surrounding environment and provide detailed explanations for unsafe or infeasible commands, thereby improving transparency. Moreover, the model emphasizes real-time interaction and personalized driving styles, delivering a superior and more customized driving experience for users.

The studies by [33,90] thoroughly validated the extensive potential of LLMs in decision making for autonomous driving systems. Building on this, ref. [91] specifically evaluated the capabilities of LLMs in spatial-aware decision making (SADM) and following traffic rules (FTR). Through simulation tests and real-world vehicle experiments, the research demonstrated that LLMs, particularly GPT-4, excel in understanding environmental information, generating safe driving decisions, and adhering to traffic rules while providing transparent decision rationales. However, LLaMA-2 and GPT-3.5 were found to be less effective than GPT-4 in both scene comprehension and decision making.

Unlike traditional models that rely on LLMs' commonsense knowledge bases and training data for driving decisions, DiLu [88] introduced a knowledge-driven framework that integrates a reasoning module, a reflection module, and a memory module, with LLMs as the core reasoning component (as shown in Figure 6b). Mimicking human learning and reflective processes, DiLu significantly enhances adaptability and decision quality. The framework employs the memory module to store experiences, while the reflection module evaluates and refines decisions. The reasoning workflow is as follows: the LLM first generates driving decisions based on perception data. The reflection module then assesses these decisions, identifies unsafe behaviors, and uses the LLM to refine them into

safer alternatives. The improved decisions are stored in the memory module as experience, enabling rapid adaptation to similar scenarios in the future. By leveraging this mechanism, DiLu demonstrated outstanding safety and generalization capabilities in complex driving scenarios, achieving performance comparable to reinforcement learning methods that require extensive training but with significantly fewer experiential data. This work presents a novel direction for knowledge-driven research in autonomous driving.

In addition to their knowledge reasoning capabilities, the sequence modeling abilities of LLMs hold significant value in multitask decision making. MTD-GPT [89] leverages GPT's sequence modeling capabilities to represent each driving task (e.g., left turn, straight, right turn) as sequential data. By predicting future actions based on historical state sequences, MTD-GPT aims to address the multitask decision-making challenges faced by autonomous vehicles at unsignalized intersections. Experimental results demonstrated that MTD-GPT significantly outperformed traditional single-task methods in both task success rates and generalization capabilities, offering an efficient solution for multitask decision making in autonomous driving.

The extensive knowledge base and robust reasoning capabilities of LLMs have shown immense potential in decision making and planning for autonomous driving, providing innovative solutions to this field [33,88,90]. Furthermore, by integrating LLM-based decision-making frameworks with human-like reasoning and learning abilities, research has demonstrated how LLMs can continuously improve their decision-making capabilities in closed-loop driving tasks by learning from past experiences. These innovative approaches inject fresh momentum into decision making and planning modules for autonomous driving, offering researchers new perspectives and ideas for future development.

4.4. Large Models for Multitasking

The architectural design of autonomous driving systems is undergoing a paradigm shift from modularized to end-to-end approaches. While modularized designs are prone to information loss and error accumulation during data transfer, end-to-end approaches address this issue by directly transforming sensor data into control signals [3,85,86]. However, the primary drawback of end-to-end methods lies in their lack of interpretability, which undermines human trust in the system. The emergent capabilities of LLMs and MLMs offer a novel solution. LLMs not only enable unified multitask processing but also enhance the transparency and interpretability of decision-making processes, thereby increasing user trust. For example, DOLPHINS [21] is a multitask autonomous driving framework based on the vision–language model OpenFlamingo, the framework of this method is shown in Figure 7a. The model utilizes a pretrained vision encoder to process driving video data, converting it into visual features, while a text encoder (e.g., LLaMA or MPT) processes textual instructions. A gated cross-attention layer facilitates enhanced interaction between the visual and textual modalities. During training, the model is trained on video–text interleaved datasets, enabling it to handle various autonomous driving tasks. Specific parts of the model are fine tuned using the LoRA strategy to improve learning efficiency and reduce computational demands. Additionally, the model incorporates the GCoT dataset for training, endowing it with fine-grained reasoning capabilities. DOLPHINS is capable of handling path planning, control signal generation, and language generation tasks, achieving unified multitask processing in autonomous driving. The experimental results indicate that DOLPHINS achieves a 30.9% reduction in collision rate and a 19.9% decrease in L2 error for 3-s trajectory prediction compared to DriveGPT-4. Additionally, in behavior understanding and reasoning tasks, DOLPHINS outperforms DriveGPT-4 with significantly higher BLEU-4 and BERT Score metrics. These results highlight DOLPHINS' ability to seamlessly integrate path planning, language generation, and behavioral

reasoning, demonstrating its potential to redefine multitask autonomous driving frameworks while significantly advancing the intelligence and decision-making transparency of autonomous systems.

Unlike DOLPHINS, which focuses on language reasoning and control signal generation, EMMA [92] emphasizes multitask processing for perception and behavior decision making tasks. The model utilizes Gemini to transform non-sensor inputs into natural language, which is then combined with visual data captured by cameras and input into the Gemini model. The training data encompass tasks such as object detection, motion planning, and road map estimation. The model extracts general features through a vision encoder, providing critical contextual information for each task. Each task is equipped with a dedicated branch and loss function, while a unified loss function is employed for overall optimization. The collaborative training across tasks enables EMMA to outperform independently trained models in 3D object detection, road map estimation, and motion planning.

Existing traffic foundation models (TFMs), while powerful, are typically specialized for single tasks and lack multi-turn interaction capabilities, making them inadequate for handling complex tasks. To address this issue, TrafficGPT [19] introduces a modularized framework based on LLMs. This system extracts information from multimodal traffic data and uses LLMs to decompose complex user instructions into multiple subtasks, which are then mapped to specific TFMs. By invoking TFMs as needed, the system avoids redundant calls and resource conflicts. TrafficGPT demonstrates exceptional performance in handling complex tasks, providing an efficient solution for traffic management and planning.

To address complex environmental interactions and potential high-risk events, autonomous driving systems must decompose tasks into multiple subtasks to enhance analysis and processing capabilities. Nouri et al. [93] proposed breaking down the hazard analysis and risk assessment (HARA) process into subtasks, including scene description, hazard event definition, and safety goal generation. The framework of this method is shown in Figure 7d. Each subtask is optimized through specific prompt designs to refine the outputs of LLMs. Additionally, the model incorporates validation from safety experts in the automotive industry to ensure the quality and practical applicability of the LLM-generated results. This human–machine collaborative approach highlights the potential of LLMs to accelerate safety engineering processes, offering innovative directions for autonomous driving safety design.

One key advantage of human drivers is their ability to quickly adjust driving behavior based on real-time feedback. To narrow this gap, researchers are continually working to improve the real-time adaptability of autonomous driving systems. DOLPHINS integrates a feedback mechanism that receives environmental feedback and error detection, enabling self-correction and adjustment. This allows the model to react swiftly and adjust decisions in dynamic environments, enhancing driving safety. In comparison, LMDrive [94] introduces the first closed-loop, end-to-end autonomous driving framework; the framework of this method is shown in Figure 7b. This framework continuously collects vehicle state, sensor data, and environmental information, enabling real-time interpretation and control signal prediction within an LLM, which then executes the corresponding actions. The closed-loop design, combined with natural language integration, demonstrates exceptional performance in real-time responsiveness and interactivity, providing a promising direction for the future development of autonomous driving technology.

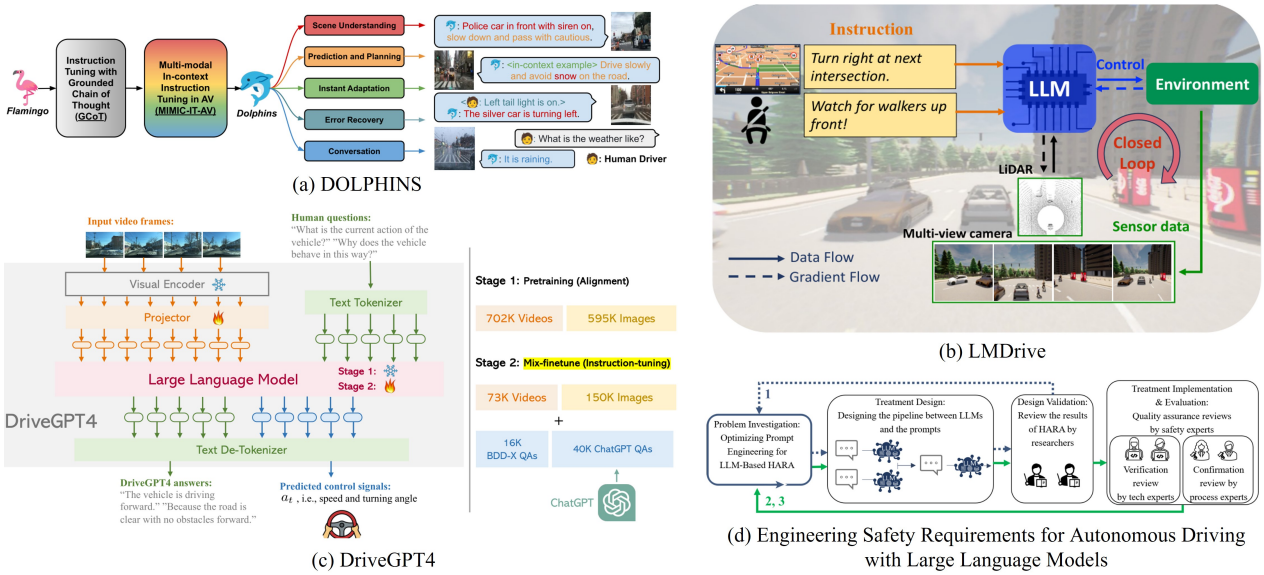


Figure 7. Some representative studies related to the application of LLMs and MLMs in multitasking.

Interpretability and transparency are critical for building human trust in autonomous driving systems. As a result, recent LLM-based autonomous driving frameworks have focused extensively on enhancing system interpretability. DriveGPT4 [8] is the first framework to apply MLM to interpretable end-to-end autonomous driving; the framework of this method is shown in Figure 7c. Video and text data are processed and input into the LLM, which analyzes dynamic scenes, performs behavior reasoning, and predicts control signals. Through a two-stage training process (pretraining and fine tuning), DriveGPT4 demonstrates exceptional performance in multimodal data fusion and alignment. Its decision-making logic is clear and interpretable, providing users with transparent explanations of model behavior. This transparency significantly enhances the system's credibility and usability.

Whether through unified multitask processing or improved human interaction, LLM-based autonomous driving frameworks have overcome the limitations of traditional models, successfully addressing many complex tasks that were previously challenging. These frameworks not only enhance the efficiency and intelligence of autonomous driving systems but also strengthen user trust by improving interpretability and transparency. As these technologies mature, LLM-based approaches are driving autonomous driving systems toward greater safety, reliability, and human-centric design, laying a solid foundation for achieving fully autonomous driving.

4.5. Large Models for Scenario Generation

Safety verification and validation (V&V) refers to the process of verifying and validating the perception, decision-making, and control modules of autonomous driving systems. Its purpose is to ensure that autonomous systems can operate safely across various scenarios and comply with predefined safety standards and regulatory requirements [104,105]. This process typically includes distance-based testing, coverage-based testing, and scenario-based testing. Among these, scenario-based testing, which evaluates the behavior and reactions of autonomous systems through simulations or recreations of different driving scenarios, is one of the primary methods for V&V in autonomous systems [106,107]. Traditional scenario-based methods initially relied on manually designed test scenarios, which lacked diversity. As autonomous systems have become increasingly complex, manually

created scenarios are no longer sufficient for validation, leading to the widespread adoption of automated scenario generation.

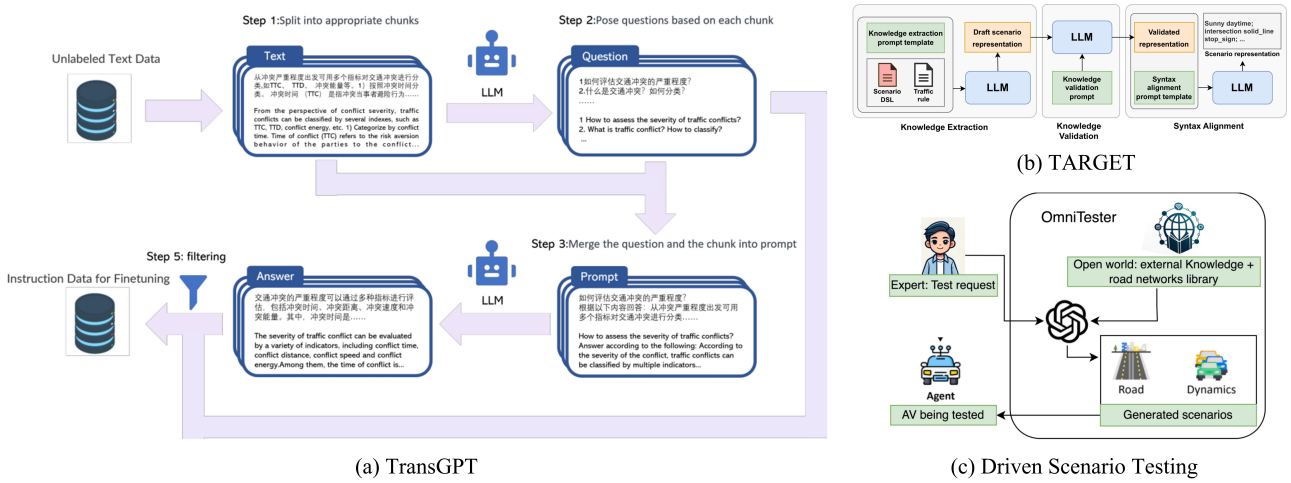
Leveraging the extensive world knowledge of LLMs, concepts and relationships relevant to autonomous driving can be extracted to construct driving scenario ontologies, which serve as the foundation for scenario generation. A driving scenario ontology is a structured representation of domain knowledge in autonomous driving, used to describe concepts and their interrelations in scenarios, thereby enhancing vehicles' perception and decision-making capabilities in diverse environments. Ref. [95] is the first model to utilize LLMs for automated construction of driving scenario ontologies. Through interaction with ChatGPT, the model helps define concepts, attributes, and relationships within the domain of autonomous driving, facilitating ontology construction. Specific prompts are designed to guide the LLM to output information in task-specific formats, and the ontology is iteratively refined. To ensure the effectiveness of the construction process, the model incorporates human intervention and a web-based assistant for real-time optimization. This approach provides a significant reference point for employing LLMs in driving scenario construction, advancing the capabilities of automated safety verification and validation.

In addition to extracting autonomous driving-related knowledge from LLMs to construct test scenario ontologies, leveraging LLMs' strong understanding and reasoning capabilities for test scenario generation is also worth exploring. In [95], LLMs are utilized to interpret complex traffic regulations and convert them into machine-readable formats that comply with DSL syntax. In contrast, ref. [96] extends the functionality of LLMs to interpreting complex traffic rules and converting them into machine-readable formats that comply with DSL syntax; the framework of this method is shown in Figure 8b. The parsed information is then used to generate templated scenario scripts, which are executed in simulators to test for rule violations, collisions, or timeouts. The experimental results indicate that TARGET's automated scenario generation process is significantly more efficient than traditional manual methods, enabling the creation of a large number of scenarios for comprehensive testing. This framework represents a major advancement in autonomous driving testing by being the first to utilize LLMs (e.g., GPT-4) to parse traffic rules and generate formal test scenario representations.

Unlike the aforementioned approaches to generating test scenarios, ADEPT [97] extracts information from real-world traffic accident reports, utilizing GPT-3's question-answering capabilities to translate these reports into structured data. These data are then combined with scenario templates to generate test scenarios. These scenarios are input into the CARLA simulator for testing, and ADEPT also supports adversarial testing by generating rule-violating behaviors or modifying scenario parameters to challenge system robustness. ADEPT not only provides a novel method for scenario generation but also enhances system safety and robustness through adversarial testing, advancing autonomous driving systems toward higher safety standards.

Although scenario-template-based methods can generate a large number of test scenarios, they often lack diversity and fail to cover extreme cases. To address this limitation, Lu et al. [98] proposed OmniTester, which retrieved relevant road network data from real-world map databases to ensure that generated scenarios align with actual traffic environments. The framework of OmniTester is shown in Figure 8c. After parsing user-provided natural language descriptions into structured scenario representations using an LLM, the retrieval module matches traffic network regions based on preprocessed map data and converts the data into XML format compatible with simulation platforms. The generator integrates the structured data and map information into the SUMO simulation platform, simulating traffic flow, participant behavior, and signal control. OmniTester is a flexible framework that constructs complete autonomous driving test scenarios through natural

language descriptions, map retrieval, and simulation generation. It effectively produces rare scenarios and improves the quality of generated scenarios through iterative refinement.



(a) TransGPT

(b) TARGET

(c) Driven Scenario Testing

Figure 8. Some representative studies related to the application of LLMs and MLMs in scenario generation.

The primary purpose of test scenarios is to evaluate the safety of autonomous driving systems in specific situations, such as adherence to traffic rules or emergency pedestrian avoidance. The evaluation criteria are well-defined, distinguishing these scenarios from the general driving environments encountered by the system. Driving environments include traffic participants, road infrastructure, and environmental conditions, which serve as inputs to the perception module and provide critical support for the prediction and planning modules. LCTGen [99] is the first LLM-based scenario generation model that accepts natural language input to produce dynamic traffic scenarios. The model uses an LLM to parse natural language into structured representations, a retrieval module to select suitable maps, and a generator to create realistic traffic scenarios. By offering diverse scenario generation solutions, LCTGen reduces the complexity of manual operations and enhances scalability in autonomous driving scenario generation.

In addition, LLMs have a wide range of applications in intelligent transportation, such as traffic problem reasoning, route planning, and scenario generation. However, general-purpose LLMs, trained on broad corpora, may lack specialized knowledge in the transportation domain. To address this, TransGPT [100] introduces transportation-specific large models, including the single-modal TransGPT-SM and the multimodal TransGPT-MM. The framework of this method is shown in Figure 8a. TransGPT-SM, optimized based on ChatGLM2-6B, is designed to answer traffic-related questions, while TransGPT-MM, built on VisualGLM-6B, processes image-text inputs to generate traffic-related text. Both variants are fine tuned with domain-specific data and demonstrate exceptional performance in tasks such as scenario generation and question answering, offering new tools for intelligent transportation applications.

LLM-based approaches have brought significant advancements to the safety verification and scenario generation of autonomous driving systems. By constructing driving scenario ontologies, interpreting traffic rules, and extracting information from accident reports, these methods enable automated and diverse test scenario generation, while improving system robustness and efficiency [10,96]. From OmniTester's real-world map retrieval to LCTGen's dynamic scenario generation and the transportation-focused TransGPT models, these innovative methods not only expand the scope of testing coverage

but also enhance the safety and reliability of autonomous driving technologies [98–100]. Together, they lay a solid foundation for the future development of the field.

5. Key Technologies for Integrating LLMs and MLMs with Autonomous Driving Systems

With their rich world knowledge, powerful reasoning capabilities, and contextual learning ability, LLMs and MLMs offer new opportunities for the intelligent development of autonomous driving systems. Integrating these large models into autonomous driving involves a series of key technologies, including prompt engineering, instruction tuning, knowledge distillation, and multimodal fusion, all of which play a crucial role in enhancing system performance and adaptability [108–111].

(1) Prompt Engineering: Optimizes LLMs for autonomous driving tasks while constraining output formats to ensure accuracy, controllability, and alignment with human expectations [109].

(2) Instruction Tuning: A supervised fine-tuning method that optimizes LLMs and MLMs for domain-specific tasks in autonomous driving. This enhances the model's expertise and improves human–machine interaction, making it more applicable to real-world scenarios [110].

(3) Knowledge Distillation: Transfers knowledge from large, complex teacher models to smaller, more efficient student models, reducing model complexity and computational costs. This improves decision-making efficiency and real-time responsiveness, addressing the demands of autonomous driving systems for fast and reliable inference [108].

(4) Multimodal Fusion: Integrates visual, linguistic, and sensor data, significantly enhancing system perception in complex driving environments. This provides strong support for environment understanding, behavior prediction, and decision planning in autonomous driving [111].

This section reviews these key technologies from four aspects: prompt engineering, instruction fine tuning, knowledge distillation, and multimodal fusion, and analyzes their core value for autonomous driving systems through specific cases, providing references for future research and applications.

5.1. Prompt Engineering

Prompt engineering [109], a technique emerging alongside the development of foundational models [112], involves adjusting and optimizing inputs to guide models in generating desired outputs. Through prompt engineering, models can improve the quality of generated results, control the style of outputs, and quickly adapt to new tasks. As a key method within the framework of efficient fine tuning (EFT), prompt engineering requires minimal parameter adjustments to efficiently complete tasks, making it widely adopted and applied. In LLM- and MLM-based autonomous driving systems, prompt engineering plays a crucial role in leveraging the world knowledge and reasoning capabilities of LLMs, thereby enhancing perception, prediction, planning, and interpretability.

With its exceptional image understanding capabilities, GPT-4V has been applied to image analysis in autonomous driving scenarios. Wen et al. [12] designed refined textual prompts to guide GPT-4V in understanding, reasoning, and decision-making based on scene images. Experimental results demonstrated that GPT-4V performed well in basic scene understanding and reasoning tasks but faced challenges in more complex tasks such as spatial reasoning and direction recognition.

To improve predictive capabilities, tasks such as lane-changing and trajectory prediction—key components of autonomous driving systems—have been optimized using prompt engineering. Ref. [15] redefined lane-change intention and future trajectory predic-

tion tasks as language modeling problems. Scene information was converted into natural language descriptions, which served as prompts for LLMs. Additional system messages were incorporated as part of the prompts to define the task and output format. This structured prompting enabled the LLM to comprehend input content and generate high-quality predictions and explanations.

Prompt engineering has also played a vital role in enabling more natural human–machine interaction. Current autonomous driving systems face challenges in adjusting driving styles based on verbal commands. Ref. [19] proposed a human-centered autonomous driving approach by parsing user natural language commands into multiple key questions and transforming them into binary classification tasks for system operation modules (e.g., perception, navigation, and in-cabin monitoring). These questions, structured as prompts, guide the LLM in step-by-step reasoning and classification, offering a novel perspective for understanding and executing user intentions in complex or emergency scenarios.

Prompt engineering has further advanced the unification of multitask execution in autonomous driving systems. For instance, ref. [63] combined risk-object detection and motion planning tasks by posing questions like "Which object poses the highest risk?" alongside visual features as inputs to the LLM, guiding the model to simultaneously perform perception and planning. Similarly, ref. [92] used task-specific prompts to transform autonomous driving tasks into visual question answering (VQA) problems, such as detecting 3D objects, estimating drivable lanes, or analyzing road conditions ahead. By designing staged prompts for chain-of-thought reasoning, this method generated interpretable trajectory predictions and decision-making processes. Likewise, ref. [13] utilized carefully designed prompts to guide VLMs in completing scene description, scene analysis, and hierarchical planning tasks. The model first generated language descriptions of weather, time, and road types through prompts, then further analyzed object attributes and behaviors, and finally, integrated path planning information to produce detailed decision-making descriptions.

Prompt engineering, with its flexibility and efficiency, enables LLMs to rapidly adapt to various autonomous driving tasks without requiring model retraining. However, designing effective prompts tailored to specific tasks and model characteristics remains a significant challenge. Complex reasoning tasks may struggle to fully unlock the model's potential through prompts alone, leading to less accurate results. Additionally, LLMs may generate unrealistic outputs based on poorly constructed prompts. Future research should focus on enhancing the reliability and adaptability of prompt engineering to provide more robust solutions for complex driving tasks. This includes developing strategies to mitigate inaccuracies, improving prompt design methodologies, and ensuring the consistency of outputs across diverse scenarios.

5.2. Instruction Fine-Tuning

Instruction fine tuning refers to a supervised learning technique that fine tunes LLMs using specific natural language instructions to enable the models to better understand and execute human task directives, thereby improving their performance across a variety of downstream tasks [110]. In LLM-based autonomous driving systems, instruction fine tuning is frequently employed to adapt LLMs for better comprehension of natural language commands and to flexibly handle complex driving tasks, providing critical support for the development of more intelligent autonomous driving systems.

Traditional end-to-end autonomous driving systems often suffer from a lack of interpretability, leading to reduced human trust and limiting commercial adoption. Enhancing model interpretability is thus an urgent priority. Xu et al. [8] addressed this issue through instruction fine tuning in their proposed DriveGPT4, which provides explanatory descriptions of vehicle behavior. Specifically, after pretraining, the model freezes the visual encoder

and performs a two-stage fine tuning process on the projection module linking the visual encoder and the LLM, as well as on the LLM itself. In the first stage, the model is fine tuned using 223 K general instruction data to retain general visual understanding capabilities and reduce hallucinations. In the second stage, it is fine tuned with 56 K video–text autonomous driving instruction data to enhance its understanding of domain-specific tasks. By training on both text generation loss and control signal prediction loss, the model simultaneously performs vehicle behavior description, reasoning explanation, and prediction of speed and steering angles, significantly improving domain adaptability and multitask capabilities. Similarly, ref. [21] employed a staged instruction fine tuning approach. In the first stage, the model is optimized with grounded chain-of-thought (GCoT) instructions using a dataset of 32 K video-instruction-answer triplets constructed from image descriptions, spatial position explanations, and reasoning processes generated by ChatGPT. This stage enriches the model's fine-grained reasoning abilities. In the second stage, a contextual learning mechanism transfers the model's general reasoning capabilities to autonomous driving tasks, enabling efficient adaptation to new tasks. This fine-tuning method not only enhances the model's task adaptability but also significantly improves its interpretability, providing a solid foundation for multitask collaboration.

In complex driving scenarios, human–machine interaction can effectively address decision-making challenges arising from the complexity of the environment, such as processing navigation and passenger language instructions. Ref. [94] proposed a language-aware autonomous driving system that generates future trajectory points based on navigation instructions. The researchers designed a dataset containing approximately 64 K data fragments, combining navigation commands with sensor data to generate matching trajectories and prediction signals. The fine-tuned model demonstrated the ability to accurately interpret diverse navigation instructions, enabling it to handle complex tasks in real-world driving scenarios.

LiDAR, with its rich spatial information, serves as a critical data source for the perception module in autonomous driving systems, making the enhancement of 3D scene understanding essential for improving system safety and efficiency. Ref. [10] transformed the complex task of 3D scene understanding into a language modeling problem to achieve multimodal alignment and task execution. The model was initially trained on 420 K 3D captioning data to align LiDAR data with language embeddings. Subsequently, it was trained on 280 K 3D grounding data to improve object classification and position prediction capabilities. Finally, instruction fine tuning was conducted using the nuScenes-QA dataset, enhancing the model's 3D spatial reasoning ability and enabling it to generate rich and reasonable responses.

Instruction fine tuning is widely applied in LLM-based autonomous driving systems, as it unifies task descriptions through natural language instructions, simplifies multitask learning processes, and provides both behavioral descriptions and reasoning explanations. This significantly enhances the transparency and interpretability of the system. Furthermore, by handling interactions such as navigation commands, models demonstrate improved capabilities in addressing complex scenario challenges. However, constructing diverse and highly relevant autonomous driving instruction datasets remains a significant challenge. Further research and exploration are required to fully realize the potential of instruction fine tuning and to provide reliable solutions for more complex autonomous driving tasks.

5.3. Knowledge Distillation

Knowledge distillation refers to the technique of transferring knowledge from a superior but complex teacher model to a simpler student model, thereby reducing resource consumption and enhancing knowledge transfer [108,113]. In the field of autonomous driving, real-time performance is a critical requirement. However, the temporal features of dynamic driving scenarios make direct computation resource-intensive, reducing the efficiency of perception and decision making. Through knowledge distillation, the multimodal fusion capabilities, long-term sequence modeling and memory capabilities, and contextual learning abilities of teacher models can be transferred to student models. This approach not only enhances the adaptability of the model but also reduces dependency on labeled data while improving the system's real-time performance and multitask processing capabilities.

To address the issue of existing models overemphasizing spatial information while neglecting temporal features, Zheng et al. [114] proposed a time-focused knowledge distillation method named TempDistiller. This method utilizes masked feature reconstruction to extract long-term memory from the teacher model and employs KL divergence to constrain the student model to learn the relationships between different frames captured by the teacher model, thereby capturing the motion relationships of dynamic objects. Experimental results demonstrated that even with a reduced number of input frames, the model's ability to detect dynamic targets improved significantly.

Although LLMs are widely applied in autonomous driving systems due to their strong reasoning capabilities, their high computational demands and limited real-time performance constrain their use in time-sensitive scenarios. LDPD [115] leverages knowledge distillation to transfer the complex collaborative decision-making knowledge embedded in LLM teacher models to student agents composed of multiple smaller networks. The teacher model provides high-quality learning guidance, enabling student models to progressively learn autonomous exploration and decision making. This approach significantly enhances the decision-making capabilities of networked autonomous vehicles in complex scenarios while reducing computational overhead and improving system efficiency.

Moreover, diverse and complex driving scenarios are critical for the verification and testing of autonomous driving systems. However, the ontologies required to generate these scenarios are typically manually constructed by domain experts, a time-consuming and inflexible process. Given their broad cross-domain knowledge, Tang et al. [95] utilized LLMs to construct scenario ontologies for autonomous driving. By employing prompt engineering and iterative tasks for knowledge distillation, they extracted domain-specific concepts, definitions, attributes, and relationships to build ontologies for verification and testing purposes. This automated scenario generation method eliminates the reliance on manual construction, enhances diversity, and provides effective support for autonomous driving systems.

Traditional knowledge models typically rely on teacher models to guide student models in a static manner, which is not conducive to the continuous optimization of student model performance. To address this, Liu et al. [116] proposed a dynamic knowledge distillation approach. This method leverages LLMs to generate representative samples (e.g., rare or hard-to-obtain samples), which are annotated and provided to the student model as new training data. Additionally, the system dynamically adjusts the sample generation strategy by interacting with the student model and analyzing its weaknesses. This mechanism significantly enhances the quality of knowledge transfer, enabling the student model to efficiently complete few-shot learning tasks while maintaining continuous performance improvement.

In summary, knowledge distillation has become a prominent research direction in LLM-based autonomous driving systems, particularly in reducing model complexity and

improving real-time performance. However, challenges such as knowledge loss during distillation and imbalances between performance and scalability remain critical issues to address. Future research should focus on further optimizing knowledge transfer mechanisms to overcome these limitations.

5.4. Multimodal Data Fusion

Multimodal data fusion is a foundational capability of LLMs and MLMs in autonomous driving, focusing on unifying different modalities (e.g., vision, language, and sensor data) into a shared semantic space to enable efficient information understanding and utilization. As a necessary approach for handling complex environments, multimodal data fusion significantly enhances the comprehensiveness, robustness, and accuracy of perception while providing critical support for efficient decision making and planning [111,117]. This technology forms the foundation for building safe, efficient, and intelligent autonomous driving systems.

Autonomous driving systems relying solely on independent perception may result in biased message descriptions, whereas integrating inputs from multiple sensors can substantially improve the accuracy of perception. Duan et al. [118] proposed a joint representation method that fuses camera and LiDAR data into a unified feature representation. This method employs a Swin Transformer for feature fusion, preserving both spatial and semantic information. The hierarchical structure of the Swin Transformer effectively handles spatial relationships, retaining the geometric information from LiDAR data while integrating the semantic features from images. Based on sensor inputs and the current state of the vehicle, the model constructs corresponding natural language prompts. Subsequently, the LLM generates vehicle behavior decisions based on these prompts. Accurate perception information greatly enhances the model's ability to produce precise driving decisions. Furthermore, ref. [119] addressed the geometric and semantic losses in traditional sensor fusion methods by mapping features from cameras and LiDAR into a shared bird's-eye view (BEV) space. This approach enables tasks such as 3D object detection and BEV map segmentation, demonstrating improved performance in integrating geometric and semantic features for autonomous driving applications.

Unlike the previously mentioned fusion methods, Ma et al. [120] proposed a novel approach to fusing visual content through multiple forms. Camera images are converted into two types of features: one representing bird's-eye view (BEV) features containing road structures and obstacles and the other capturing dynamic interaction information, contextual details, and temporal backgrounds as video features extracted via a visual-language model (VLM). The Planning Transformer module combines BEV and video features, enabling the understanding of objects' spatial positions while capturing semantic information, thereby achieving accurate predictions of driving trajectories. Similarly focusing on BEV fusion, ref. [9] proposed a method that integrates multi-view images with LiDAR point cloud data to generate BEV representations. This language-enhanced map aligns image descriptions with positional data using a visual-language model, encompassing objects' geometric information and semantic descriptions. Without requiring additional training or fine tuning, this method supports tasks such as visual reasoning, spatial understanding, and decision making, demonstrating versatility and efficiency.

However, forcing the fusion of LiDAR data and image data into a bird's-eye view (BEV) space can result in spatial distribution inconsistencies between modalities, leading to false positive or false negative detection results. Fu et al. [121] proposed a semantic flow alignment (SFA) module that addresses the issues caused by depth and perspective differences by spatially aligning the BEV features of LiDAR and camera data. The SFA module performs spatial consistency alignment on features prior to fusion, effectively

improving the accuracy of 3D object detection and offering a novel approach to achieving multimodal fusion.

Overall, multimodal data fusion technologies significantly enhance the perception and decision-making capabilities of autonomous driving systems in complex environments by integrating visual, linguistic, and sensor data. From joint feature representation to semantic-guided alignment, these methods provide diverse pathways for implementing multimodal fusion, driving intelligent driving technologies toward greater efficiency and reliability.

6. Current Challenges

Although the introduction of LLMs and MLMs has enhanced the ability of autonomous driving systems to handle long-tail events and increased the transparency of decision-making processes, autonomous driving systems based on LLMs and MLMs still face numerous challenges. These challenges not only impact the safety and intelligence of such systems but may also hinder further technological advancements. In-depth research into these challenges can help researchers more accurately identify current limitations and potential avenues for improvement, accelerating the maturity of autonomous driving systems. This chapter provides a detailed analysis of four key issues: the demand for annotated datasets, optimization of visual-text alignment, detection and mitigation of hallucinations, and adversarial attacks and defenses. It explores strategies to effectively address these challenges and pave the way for future advancements.

6.1. The Urgent Need for Carefully Annotated Datasets

Carefully annotated datasets are essential for model training, but linguistic annotations in dynamic scenes are particularly costly and complex. Unlike general image–text datasets, the autonomous driving domain requires handling intricate dynamic scenarios and multimodal data, making manual annotation more challenging. Commonly used datasets in the field of autonomous driving are summarized in Table 2 and mainly include the following:

(1) KITTI [122], a foundational dataset in early autonomous driving research, contains 7481 training images and 80,256 3D bounding boxes, suitable for tasks like object detection and semantic segmentation.
(2) nuScenes [123], focusing on urban driving scenarios, comprises data from multiple sensors and includes 1000 scene segments, supporting object detection and tracking tasks.
(3) Waymo Open Dataset [124], consisting of 1150 training scenes and 750 validation scenes, facilitates object detection, tracking, and semantic segmentation.
(4) Specialized datasets include ApolloScape [125], covering large-scale complex driving scenarios; CADC, focused on adverse weather conditions; nuScenes-QA [126], designed for video question-answering tasks; and NuPrompt [127], built on nuScenes for 3D object detection involving multiple objects.

Despite increasing scale and diversity, datasets specifically designed for instruction fine tuning remain relatively scarce. Instruction fine tuning enables the transfer of LLMs' extensive reasoning capabilities to the autonomous driving domain, enhancing the model's adaptability to new tasks. Refs. [8,10,21,94] have constructed instruction fine-tuning datasets for autonomous driving, significantly improving models' adaptability.

Table 2. Some representative datasets in the field of autonomous driving.

Dataset	Year	Size	Sensor Modalities				Task Types
			Camera	LiDAR	Radar	Others	
KITTI [122]	2012	15,000 images and point clouds.	Front-view	✓	✗	GPS/IMU	2D/3D Object Detection, Semantic Segmentation, Object Tracking
nuScenes [123]	2020	1000 scenes with 40,000 annotated frames.	360°	✓	✓	GPS/IMU	3D Object Detection, Object Tracking, Scene Understanding
Waymo Open Dataset [124]	2020	1150 scenes with 200 frames per scene scene, 3D and 2D bounding boxes.	360°	✓	✗	GPS/IMU	2D/3D Object Detection, Semantic Segmentation, Object Tracking, Motion Planning
ApolloScape [125]	2019	140 K images with per-pixel semantic mask, 89,430 annotated objects in total.	Front-view	✓	✗	GPS/IMU	3D Object Detection, Semantic Segmentation, Motion Prediction, Lane Detection
CADC [128]	2020	75 driving sequences, with 56,000 images, 7000 point clouds.	360°	✓	✗	GPS/IMU	3D Object Detection, Semantic Segmentation, Object Tracking
ONCE [129]	2021	1 million LiDAR scenes, 7 million images with 2D/3D bounding boxes.	360°	✓	✗	GPS/IMU	2D/3D Object Detection, Semantic Segmentation, Object Tracking
Drama [130]	2023	17,785 driving scenarios, 17,066 captions, 77,639 questions, 102,830 answers.	Front-view	✗	✗	IMU/CAN	visual question answering about video and object, Image Captioning
DriveLM [131]	2023	DriveLM-nuScenes: 4871 video frames, 450 K QA pairs DriveLM-CARLA: 64,285 frames, 1.5M QA pairs.	Based on nuScenes and CARLA.				Visual Question Answering, Perception Tasks, Behavior Prediction, Planning Tasks, Trajectory Prediction, Behavior Classification
NuPrompt [127]	2023	35,367 textual description for 3D objects.	Based on nuScenes.				Multi-Object Tracking, 3D Object Localization, Trajectory Prediction, 3D Object Detection
nuScenes-QA [126]	2023	34 K visual scenes and 460 K QA pairs.	Based on nuScenes.				Visual Question Answering, Scene Understanding, Spatio-temporal Reasoning, HD Map Assistance
NuInstruct [132]	2023	91 K instruction-response pairs in total.	Based on nuScenes.				Object Detection, Object Tracking, Scene Understanding, Driving Planning
Reason2Drive [133]	2023	600 K video-text pairs.	Based on nuScenes, Waymo and ONCE.				Visual Question Answering, Multi-Object Tracking, Commonsense Reasoning, Object Detection

However, many studies remain focused on single tasks or single-modal information, limiting comprehensive understanding of full driving scenarios. To address this limitation, NuInstruct [132], built on the nuScenes dataset, includes 91,355 instruction–response pairs. By converting multimodal information (video, LiDAR, IMU data) into a structured database, relevant scene data are extracted using SQL queries. Diverse instructions are then generated using GPT-4 or templates. NuInstruct covers tasks including perception, prediction, risk assessment, and planning, providing a comprehensive benchmark for testing and evaluation.

To address the "black-box" problem of traditional end-to-end autonomous driving systems, Nie et al. [133] created the benchmark dataset Reason2Drive, comprising over 600 K video–text pairs. By integrating and standardizing data from publicly available datasets

such as nuScenes, Waymo, and ONCE into an object-centric database, Reason2Drive unifies annotation formats. Manually curated question modules are used to address object-level and scene-level tasks, while GPT-4 is employed for automatic annotation of perception, prediction, and reasoning tasks. This significantly reduces manual workload and improves consistency and diversity. The dataset spans various driving environments (e.g., urban, highways, rural roads) and dynamic scenarios, providing crucial support for enhancing the interpretability and chain-of-thought reasoning capabilities of autonomous driving systems.

Despite the progress in scale and diversity achieved by these datasets, they still struggle to fully cover real-world long-tail scenarios, and annotating multimodal and dynamic scenes remains challenging. Issues such as inconsistent annotation quality and insufficient diversity remain critical challenges. In the future, constructing larger-scale datasets to encompass more dynamic scenarios and long-tail events will be essential. Additionally, employing GPT-assisted annotation or developing more efficient annotation methods will be crucial to improve data quality and annotation efficiency, thereby providing a stronger foundation for the advancement of autonomous driving technologies.

6.2. The Alignment of Visual Information and Text in Autonomous Driving Scenarios

In recent years, the application of LLMs in autonomous driving has significantly improved system interpretability and generalization capabilities [134]. However, enabling LLMs to better understand semantic information from visual data and make reasonable decisions remains a critical challenge. Aligning visual information from sensors with textual information such as user instructions, in-vehicle navigation, and map data not only enhances the system's capabilities in scene understanding, reasoning, decision making, and user interaction but also reduces the likelihood of hallucinations.

To enhance the interpretability of LLM-based autonomous driving systems and address planning and decision-making failures caused by long-tail events, Tian et al. [135] transformed driving scenes into object-level tokens, which were then aligned with the language model to achieve more efficient semantic understanding and reasoning. The study proposed a method for training adapters based on question-answering tasks to align the latent token space with the textual embedding space, a crucial step in enabling LLMs to comprehend visual information effectively.

In the field of autonomous driving, semantic information-based bird's-eye view (BEV) perception methods have played a significant role in improving system interpretability. By describing 3D scenes in natural language, these methods not only enhance the explanatory capabilities of models but also provide crucial support for the decision-making process. Perception tasks and description tasks are inherently complementary, and their joint alignment can significantly improve system performance in scene understanding and decision-making. Ma et al. [136] proposed a multimodal task alignment framework that offers an effective solution for multimodal collaboration in autonomous driving scenarios. This framework aligns semantic information generated by BEV perception with natural language descriptions and implements cross-modal alignment between perception outputs and description outputs to achieve deeper task synergy. This alignment mechanism substantially improves perception accuracy and reduces hallucinations caused by information misinterpretation or bias, providing a solid foundation for the reliability and safety of decision-making. By employing this method, autonomous driving systems can interpret complex scenes more accurately, improving their adaptability to dynamic driving environments and their ability to handle unexpected situations.

In addition to the alignment of object-level tokens derived from driving scenes with textual embeddings [135] and the alignment of BEV perception with natural language descriptions [136], video–text alignment is also a crucial method for enhancing the inter-

pretability of autonomous driving systems. To this end, ref. [133] introduced a new benchmark dataset, Reason2Drive, along with related methodologies to advance interpretable reasoning and chain-of-decision processes. This dataset comprises 600 K video–text pairs covering perception, prediction, and reasoning tasks, providing diverse and complex task support for safe and reliable autonomous driving. Furthermore, PiTe [137] proposed a pixel-temporal alignment strategy based on trajectory data to achieve fine-grained spatial and temporal alignment between vision and language, significantly improving video understanding and multimodal task performance.

Overall, vision–text alignment is a foundational technique for promoting multimodal fusion and enhancing LLMs' ability to extract and understand visual information. However, the inherent differences between dense, pixel-based visual data and discrete, semantic textual data present significant challenges. Moreover, the dynamic and temporal nature of autonomous driving scenarios demands fine-grained alignment between vision and language in the temporal dimension. Future research should focus on developing more efficient alignment methods while covering a broader range of scenarios to further enhance the safety and interpretability of autonomous driving systems.

6.3. Detection and Mitigation of Hallucinations

Hallucination is one of the inherent limitations of LLMs, referring to cases where the model generates outputs that are inconsistent with the real world or entirely incorrect after receiving input [64,138,139]. When MLMs incorporate visual information, the hallucination issue may become even more pronounced. For instance, the model might over-rely on the prior knowledge of LLMs, disregarding visual input and directly producing an answer, or it may generate erroneous reasoning due to misinterpretation of visual content. In safety-critical applications such as autonomous driving, hallucinations pose severe risks. Therefore, mitigating hallucination issues in LLMs and MLMs within autonomous driving systems and enhancing model reliability has become a critical focus of the current research.

In autonomous driving systems, pedestrian detection is a fundamental task [140]. Yet, even in this relatively mature domain, hallucination remains an issue for LLM-based systems. To address this, Dona et al. [141] categorized hallucination types specific to pedestrian detection and proposed strategies for mitigating them. The authors identified three primary hallucination types in advanced driver assistance systems/autonomous driving (ADAS/AD) scenarios: false negatives (failing to detect an existing pedestrian), false positives (incorrectly detecting a pedestrian), and refusal to process (misjudgments caused by content policy restrictions). To tackle these challenges, the authors proposed methods such as consistency checking (BO3), temporal historical voting (THV), and physical plausibility verification. The experimental results demonstrated that these strategies effectively reduce hallucinations and improve the reliability of LLMs, particularly in scenarios leveraging temporal sequence data, where contextual information significantly enhances system performance.

To systematically investigate the hallucination problem in LLM-based autonomous driving systems, researchers have proposed automated methods for generating hallucination detection benchmarks. Wu et al. [142] developed the first automated benchmark for hallucination in VLMs, aiming to reduce reliance on manually designed scenarios. This method employs a large-scale automated pipeline, including scene generation, image manipulation (e.g., anomaly insertion and context removal), question construction, and hallucination detection, to generate pairs of (image, question). This approach not only reduces costs but also significantly increases the diversity of test cases, enabling researchers to delve deeper into the mechanisms that trigger hallucinations in models. The benchmark provides

a valuable tool for assessing the sensitivity and robustness of LLM-based autonomous driving systems to hallucination issues.

Through scene generation, image manipulation (e.g., anomaly insertion, paired insertion, and context removal), question construction, and hallucination detection, the automated pipeline systematically creates diverse (image, question) pairs. This method is designed to provoke hallucinations in LVLMs caused by language priors, aiding researchers in identifying common failure patterns and optimizing models. By emphasizing the induction of hallucinations in specific scenarios, the benchmark serves to evaluate the robustness of models against hallucination issues, providing a structured framework for improvement.

In addition to the development of evaluation benchmarks, constructing datasets for addressing hallucination issues has become a critical research direction. Gunjal et al. [138] proposed a hallucination detection dataset named M-HalDetect, comprising 16,000 image–description pairs with fine-grained annotations (down to the clause level). Using this dataset, researchers trained a reward model for hallucination detection and rejection sampling. This model demonstrated transferability to other MLMs, effectively reducing hallucination occurrences.

Unlike most models that rely on reinforcement learning with external knowledge, Liang et al. [143] focus on reducing hallucinations within the internal mechanisms of LLMs. They propose a hallucination mitigation approach based on enhancing the model's self-awareness, first defining four distinct knowledge states of LLMs and developing an automated hallucination detection tool, DreamCatcher, which assesses content authenticity through knowledge probing and consistency checking. Furthermore, they introduce a reinforcement learning framework based on knowledge feedback (RLKF), which optimizes the model through reinforcement learning, enabling it to better utilize internal knowledge while minimizing unnecessary hallucinations. Experimental results demonstrate that RLKF significantly improves model performance in knowledge reasoning and truthfulness evaluation tasks, providing an efficient solution for hallucination mitigation.

Although considerable progress [142–145] has been made in hallucination detection and mitigation for general-purpose LLMs and MLMs, these studies are largely limited to generic models. Models specifically designed for hallucination detection and mitigation in the complex scenarios of autonomous driving remain relatively scarce, especially for rare driving events requiring specialized benchmarks and training datasets. Future research should focus on developing hallucination detection methods tailored to autonomous driving and expanding their application to more diverse scenarios. Furthermore, challenges inherent to LLMs themselves must be addressed. For instance, the strong prior knowledge of LLMs can sometimes overshadow visual inputs, causing perception results to deviate from reality. Current reinforcement learning with human feedback (RLHF) or self-supervised learning methods can only partially alleviate this issue. To address these challenges, more efficient multimodal alignment mechanisms need to be designed to maximize consistency between visual and textual content, thereby reducing hallucination occurrences. Additionally, incorporating knowledge-enhancement modules and self-supervised mechanisms can ensure that model-generated responses align more closely with real-world rules, providing stronger assurances for the safety and reliability of autonomous driving systems.

6.4. Adversarial Attacks and Defense Strategies

Adversarial attacks refer to the introduction of imperceptible perturbations into input data, causing deep models to make incorrect predictions or decisions [146]. Such attacks pose significant risks in autonomous driving systems and large-scale language models, as they may lead to erroneous judgments in critical scenarios, potentially resulting in severe traffic accidents [28,147,148]. MLMs, which combine pretrained visual encoders with LLMs,

are particularly vulnerable, further increasing the risks of adversarial exploitation [149]. Consequently, defending against adversarial attacks in LLM- and MLM-based autonomous driving systems has become a critical area of research.

In autonomous driving scenarios, adversarial attacks often target the perception, prediction, and planning modules simultaneously. For instance, attacking input images in the perception module can prevent detectors from recognizing pedestrians; injecting false information into LiDAR point cloud data can disrupt sensor perception; and coordinated interference with multimodal data can compromise the entire decision-making chain. More advanced attacks target end-to-end autonomous driving models, introducing global perturbations that affect outputs across all modules. These attack methods pose significant threats to system safety, making the development of robust defense mechanisms an urgent challenge [149–153].

To address adversarial attacks on VLMs in autonomous driving, Zhang et al. [146] proposed the ADvLM framework. This framework aims to handle the diversity of textual instructions and the temporal characteristics of visual scenes. In the textual modality, ADvLM employs semantic invariance induction to ensure the effectiveness of attacks across diverse textual instructions. In the visual modality, attention mechanisms are used to identify key frames most influential to driving decisions, maintaining consistency and generalizability of the perturbations. Experimental results demonstrated that ADvLM achieves significant attack efficacy in white-box, black-box, and real-world physical environments, highlighting the security vulnerabilities of autonomous driving VLMs and providing valuable insights for the design of future defense mechanisms.

In terms of perception attacks, attackers can disrupt the target detection and tracking modules, leading the system to make incorrect decisions. For instance, fabricating objects, removing real objects, or misclassifying objects can significantly impair the normal operation of the perception module. To address this, the HUDSON [151] framework was proposed. HUDSON transforms perception data into domain-specific language (DSL) and uses a chain-of-thought prompting engine to detect inconsistencies in context, temporal sequences, and spatial relationships. It further employs causal reasoning to analyze discrepancies between decisions and perception results, thereby generating safe driving decisions. In various attack scenarios, HUDSON demonstrated high attack detection rates and safe decision-making rates, significantly enhancing the robustness of autonomous driving systems.

Autonomous driving systems are evolving from a modularized architecture to an end-to-end framework, where perception, prediction, and planning tasks are integrated into a unified model. This integration reduces information loss and error accumulation between modules [154–156], enabling state-of-the-art performance. However, such models are more susceptible to adversarial attacks, making adversarial training an effective approach to improving their robustness. Traditional adversarial training methods typically involve the inclusion of adversarial samples [157,158] or focus on adversarial training for individual modules, such as 3D object detection [159,160] and trajectory prediction [161–164]. However, adversarial training specifically for end-to-end autonomous driving systems remains relatively scarce, which contrasts with the growing trend of end-to-end systems as a research hotspot. Thus, exploring adversarial training for end-to-end systems holds significant importance. MA2T [153] is the first model to conduct adversarial training specifically for end-to-end autonomous driving systems. This model introduces noise into each module and employs a unified loss function during training to address inconsistencies in the training objectives of different modules. Additionally, a dynamic weight adjustment mechanism adaptively modifies the loss weights based on each module's contribution to the final output. The experimental results demonstrate that MA2T significantly enhances

the model's robustness under adversarial attacks, reducing collisions and trajectory deviations. This provides a novel approach to advancing the safety research of end-to-end autonomous driving systems.

Integrating multimodal sensor data has become a key trend in improving the performance of autonomous driving systems. The combination of LiDAR and camera data can significantly enhance 3D object detection capabilities. However, this integration also introduces new vulnerabilities to adversarial attacks. Yang et al. [152] demonstrated that adding a small number of adversarial points to point cloud data can render vehicles undetectable, thereby posing a serious threat to system safety. Their experiments revealed that the effectiveness of such attacks is closely related to the number of adversarial points, the target's distance, and the angle of observation, offering new insights into studying system robustness.

Although autonomous driving systems integrating LLMs and MLMs have made some progress in adversarial attack research, this field is still in its early stages, with significant untapped potential. Current attack methods primarily focus on single modalities, making them inadequate for addressing the complexities of multimodal data. Furthermore, as attack techniques evolve, existing defense mechanisms may become less effective, necessitating dynamic updates to defense strategies. The lack of standardized evaluation frameworks for adversarial attacks and defenses also limits the ability to conduct cross-comparative studies. Future research should focus on multimodal adversarial training, optimizing defense mechanisms, and establishing standardized evaluation frameworks to comprehensively enhance the safety and robustness of autonomous driving systems.

7. Conclusions

This paper provides a systematic review of the applications and key technologies of LLMs and MLMs in autonomous driving systems. It focuses on analyzing the practical applications of LLM-based autonomous driving frameworks in core tasks such as perception, prediction, planning, multitask processing, and human–machine interaction. Through illustrative case studies, the paper highlights the value and significance of LLMs in various tasks. Furthermore, this paper presents a comprehensive analysis of key technologies in LLM-based and MLM-based autonomous driving systems, including prompt engineering, instruction tuning, knowledge distillation, and multimodal data fusion. Through specific case studies, it explores how these technologies contribute to enhancing system safety, improving real-time decision making, optimizing operational efficiency, and increasing adaptability to diverse driving tasks. While LLM-based autonomous driving frameworks have addressed challenges such as limited performance in long-tail scenarios and poor interpretability due to the "black-box" nature of traditional models, their integration frameworks still face numerous challenges. The paper systematically identifies major challenges, including hallucination issues, difficulties in multimodal alignment, vulnerabilities to adversarial attacks, and insufficient dataset scale and diversity. It also presents current solutions to these challenges through detailed case studies and proposes potential future research directions, offering insights into resolving these critical issues. By analyzing the applications, key technologies, and challenges of LLM- and MLM-based autonomous driving frameworks, this paper provides valuable insights and inspiration for researchers, facilitating advancements toward safer, more efficient, and more intelligent L4 and L5 autonomous driving systems.

Author Contributions: Conceptualization, J.L. (Jing Li) and L.Y. (Lie Yang); methodology, J.L. (Jing Li) and L.Y. (Lie Yang); formal analysis, J.L. (Jing Li) and J.L. (Jingyuan Li); investigation, J.L. (Jing Li) and L.Y. (Lie Yang); resources, L.Y. (Lichao Yang), G.Y., and H.C.; data curation, J.L. (Jing Li) and G.Y.; writing—original draft preparation, J.L. (Jing Li); writing—review and editing, J.L. (Jing Li) and L.Y. (Lie Yang); visualization, L.Y. (Lie Yang) and J.L. (Jingyuan Li); supervision, L.Y. (Lie Yang); project administration, L.Y. (Lie Yang); funding acquisition, L.Y. (Lie Yang). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Without any developmental data availability.

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

References

1. Jo, K.; Kim, J.; Kim, D.; Jang, C.; Sunwoo, M. Development of autonomous car—Part I: Distributed system architecture and development process. *IEEE Trans. Ind. Electron.* **2014**, *61*, 7131–7140. [CrossRef]
2. Li, Y.; Katsumata, K.; Javanmardi, E.; Tsukada, M. Large Language Models for Human-like Autonomous Driving: A Survey. *arXiv* **2024**, arXiv:2407.19280.
3. Chen, L.; Wu, P.; Chitta, K.; Jaeger, B.; Geiger, A.; Li, H. End-to-end autonomous driving: Challenges and frontiers. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 10164–10183
4. Parekh, D.; Poddar, N.; Rajpurkar, A.; Chahal, M.; Kumar, N.; Joshi, G.P.; Cho, W. A review on autonomous vehicles: Progress, methods and challenges. *Electronics* **2022**, *11*, 2162. [CrossRef]
5. Zhao, J.; Zhao, W.; Deng, B.; Wang, Z.; Zhang, F.; Zheng, W.; Cao, W.; Nan, J.; Lian, Y.; Burke, A.F. Autonomous driving system: A comprehensive survey. *Expert Syst. Appl.* **2024**, *242*, 122836. [CrossRef]
6. Elallid, B.B.; Benamar, N.; Hafid, A.S.; Rachidi, T.; Mrani, N. A comprehensive survey on the application of deep and reinforcement learning approaches in autonomous driving. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 7366–7390. [CrossRef]
7. Chen, Z.; Xu, L.; Zheng, H.; Chen, L.; Tolba, A.; Zhao, L.; Yu, K.; Feng, H. Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models. *Comput. Mater. Contin.* **2024**, *80*, 1753–1808. [CrossRef]
8. Xu, Z.; Zhang, Y.; Xie, E.; Zhao, Z.; Guo, Y.; Wong, K.Y.K.; Li, Z.; Zhao, H. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robot. Autom. Lett.* **2024**, *9*, 8186–8193. [CrossRef]
9. Choudhary, T.; Dewangan, V.; Chandhok, S.; Priyadarshan, S.; Jain, A.; Singh, A.K.; Srivastava, S.; Jatavallabhula, K.M.; Krishna, K.M. Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving. *arXiv* **2023**, arXiv:2310.02251.
10. Yang, S.; Liu, J.; Zhang, R.; Pan, M.; Guo, Z.; Li, X.; Chen, Z.; Gao, P.; Guo, Y.; Zhang, S. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. *arXiv* **2023**, arXiv:2312.14074.
11. Tang, T.; Wei, D.; Jia, Z.; Gao, T.; Cai, C.; Hou, C.; Jia, P.; Zhan, K.; Sun, H.; Fan, J.; et al. BEV-TSR: Text-Scene Retrieval in BEV Space for Autonomous Driving. *arXiv* **2024**, arXiv:2401.01065.
12. Wen, L.; Yang, X.; Fu, D.; Wang, X.; Cai, P.; Li, X.; Ma, T.; Li, Y.; Xu, L.; Shang, D.; et al. On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving. *arXiv* **2023**, arXiv:2311.05332.
13. Tian, X.; Gu, J.; Li, B.; Liu, Y.; Wang, Y.; Zhao, Z.; Zhan, K.; Jia, P.; Lang, X.; Zhao, H. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv* **2024**, arXiv:2402.12289.
14. Renz, K.; Chen, L.; Marcu, A.M.; Hünermann, J.; Hanotte, B.; Karnsund, A.; Shotton, J.; Arani, E.; Sinavski, O. CarLLaVA: Vision language models for camera-only closed-loop driving. *arXiv* **2024**, arXiv:2406.10165.
15. Peng, M.; Guo, X.; Chen, X.; Zhu, M.; Chen, K.; Wang, X.; Wang, Y. LC-LLM: Explainable Lane-Change Intention and Trajectory Predictions with Large Language Models. *arXiv* **2024**, arXiv:2403.18344.
16. Lan, Z.; Liu, L.; Fan, B.; Lv, Y.; Ren, Y.; Cui, Z. Traj-llm: A new exploration for empowering trajectory prediction with pre-trained large language models. *IEEE Trans. Intell. Veh.* 2024, *early access*.
17. Mao, J.; Qian, Y.; Ye, J.; Zhao, H.; Wang, Y. Gpt-driver: Learning to drive with gpt. *arXiv* **2023**, arXiv:2310.01415.
18. Wang, S.; Zhu, Y.; Li, Z.; Wang, Y.; Li, L.; He, Z. ChatGPT as your vehicle co-pilot: An initial attempt. *IEEE Trans. Intell. Veh.* **2023**, *8*, 4706–4721.
19. Yang, Y.; Zhang, Q.; Li, C.; Marta, D.S.; Batool, N.; Folkesson, J. Human-centric autonomous systems with llms for user command reasoning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 988–994.
20. Liao, H.; Shen, H.; Li, Z.; Wang, C.; Li, G.; Bie, Y.; Xu, C. Gpt-4 enhanced multimodal grounding for autonomous driving: Leveraging cross-modal attention with large language models. *Commun. Transp. Res.* **2024**, *4*, 100116.

21. Ma, Y.; Cao, Y.; Sun, J.; Pavone, M.; Xiao, C. Dolphins: Multimodal language model for driving. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer: Berlin/Heidelberg, Germany, 2025; pp. 403–420.
22. Fourati, S.; Jaafar, W.; Baccar, N.; Alfattani, S. XLM for Autonomous Driving Systems: A Comprehensive Review. *arXiv* **2024**, arXiv:2409.10484.
23. Gao, H.; Wang, Z.; Li, Y.; Long, K.; Yang, M.; Shen, Y. A survey for foundation models in autonomous driving. *arXiv* **2024**, arXiv:2402.01105.
24. Yang, Z.; Jia, X.; Li, H.; Yan, J. Llm4drive: A survey of large language models for autonomous driving. *arXiv* **2023**. arXiv:2312.00438.
25. Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Zhou, Y.; Liang, K.; Chen, J.; Lu, J.; Yang, Z.; Liao, K.D.; 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, Waikoloa, HI, USA, 3–8 January 2024; pp. 958–979.
26. Zhou, X.; Liu, M.; Yurtsever, E.; Zagar, B.L.; Zimmer, W.; Cao, H.; Knoll, A.C. Vision language models in autonomous driving: A survey and outlook. *IEEE Trans. Intell. Veh.* 2024, *early access*.
27. Zhu, Y.; Wang, S.; Zhong, W.; Shen, N.; Li, Y.; Wang, S.; Li, Z.; Wu, C.; He, Z.; Li, L. Will Large Language Models be a Panacea to Autonomous Driving? *arXiv* **2024**, arXiv:2409.14165.
28. Zhou, X.; Liu, M.; Zagar, B.L.; Yurtsever, E.; Knoll, A.C. Vision language models in autonomous driving and intelligent transportation systems. *arXiv* **2023**, arXiv:2310.14414.
29. Huang, Y.; Chen, Y.; Li, Z. Applications of large scale foundation models for autonomous driving. *arXiv* **2023**, arXiv:2311.12144.
30. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–45.
31. Liang, Z.; Xu, Y.; Hong, Y.; Shang, P.; Wang, Q.; Fu, Q.; Liu, K. A Survey of Multimodel Large Language Models. In Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering, Xi' an, China, 26–28 January 2024; pp. 405–409.
32. Huang, Y.; Chen, Y. Autonomous driving with deep learning: A survey of state-of-art technologies. *arXiv* **2020**, arXiv:2006.06091.
33. Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Wang, Z. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 902–909.
34. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A survey of large language models. *arXiv* **2023**, arXiv:2303.18223.
35. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A comprehensive overview of large language models. *arXiv* **2023**, arXiv:2307.06435.
36. Bar-Hillel, Y. The present status of automatic translation of languages. *Adv. Comput.* **1960**, *1*, 91–163.
37. Holt, A.W.; Turanski, W. Man-to-machine communication and automatic code translation. In Proceedings of the Western Joint IRE-AIEE-ACM Computer Conference, San Francisco, CA, USA, 3–5 May 1960; pp. 329–339.
38. Arevalo, J.; Solorio, T.; Montes-y Gómez, M.; González, F.A. Gated multimodal units for information fusion. *arXiv* **2017**, arXiv:1702.01992.
39. Graves, A. Long short-term memory. In *Supervised Sequence Labelling with Recurrent Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 37–45.
40. Vaswani, A. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
41. Kenton, J.D.M.W.C.; Toutanova, L.K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the naacL-HLT, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, p. 2.
42. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog.* **2019**, *1*, 9.
43. Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; et al. Language models are few-shot learners. *arXiv* **2020**, arXiv:2005.14165.
44. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
45. Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. Lima: Less is more for alignment. *Adv. Neural Inf. Process. Syst.* **2024**, *36*.
46. Wang, S.; Yu, Z.; Jiang, X.; Lan, S.; Shi, M.; Chang, N.; Kautz, J.; Li, Y.; Alvarez, J.M. OmniDrive: A Holistic LLM-Agent Framework for Autonomous Driving with 3D Perception, Reasoning and Planning. *arXiv* **2024**, arXiv:2405.01533.
47. Chib, P.S.; Singh, P. LG-Traj: LLM Guided Pedestrian Trajectory Prediction. *arXiv* **2024**, arXiv:2403.08032.

48. Schumann, R.; Zhu, W.; Feng, W.; Fu, T.J.; Riezler, S.; Wang, W.Y. Velma: Verbalization embodiment of llm agents for vision and language navigation in street view. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 26–27 February 2024; Volume 38, pp. 18924–18933.
49. Sha, H.; Mu, Y.; Jiang, Y.; Chen, L.; Xu, C.; Luo, P.; Li, S.E.; Tomizuka, M.; Zhan, W.; Ding, M. Languagempc: Large language models as decision makers for autonomous driving. *arXiv* **2023**, arXiv:2310.03026.
50. Atakishiyev, S.; Salameh, M.; Goebel, R. Incorporating Explanations into Human-Machine Interfaces for Trust and Situation Awareness in Autonomous Vehicles. *arXiv* **2024**, arXiv:2404.07383.
51. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
52. Kim, W.; Son, B.; Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. In Proceedings of the International Conference on Machine Learning. PMLR, Virtual, 18–24 July 2021; pp. 5583–5594.
53. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning. PMLR, Virtual, 18–24 July 2021; pp. 8748–8763.
54. Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A survey on multimodal large language models. *arXiv* **2023**, arXiv:2306.13549.
55. Alayrac, J.B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 23716–23736.
56. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International Conference on Machine Learning. PMLR, Honolulu, HI, USA, 23–29 July 2023; pp. 19730–19742.
57. Zhu, D.; Chen, J.; Shen, X.; Li, X.; Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv* **2023**, arXiv:2304.10592.
58. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **2024**, *36*.
59. Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024; pp. 24185–24198.
60. Zhang, H.; Li, X.; Bing, L. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv* **2023**, arXiv:2306.02858.
61. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
62. Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A.M.; Hauth, A.; Millican, K.; et al. Gemini: A family of highly capable multimodal models. *arXiv* **2023**, arXiv:2312.11805.
63. Ding, X.; Han, J.; Xu, H.; Zhang, W.; Li, X. Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving. *arXiv* **2023**, arXiv:2309.05186.
64. McDonald, D.; Papadopoulos, R.; Benningfield, L. Reducing llm hallucination using knowledge distillation: A case study with mistral large and mmlu benchmark. *Authorea Prepr.* **2024**. [CrossRef]
65. Taeihagh, A.; Lim, H.S.M. Governing autonomous vehicles: Emerging responses for safety, liability, privacy, cybersecurity, and industry risks. *Transp. Rev.* **2019**, *39*, 103–128. [CrossRef]
66. Khan, M.A.; Sayed, H.E.; Malik, S.; Zia, T.; Khan, J.; Alkaabi, N.; Ignatious, H. Level-5 autonomous driving—Are we there yet? a review of research literature. *ACM Comput. Surv. (CSUR)* **2022**, *55*, 1–38. [CrossRef]
67. Barabas, I.; Todoruţ, A.; Cordoş, N.; Molea, A. Current challenges in autonomous driving. *Iop Conf. Ser. Mater. Sci. Eng.* **2017**, *252*, 012096.
68. Thrun, S. Toward robotic cars. *Commun. ACM* **2010**, *53*, 99–106.
69. Zablocki, É.; Ben-Younes, H.; Pérez, P.; Cord, M. Explainability of deep vision-based autonomous driving systems: Review and challenges. *Int. J. Comput. Vis.* **2022**, *130*, 2425–2452.
70. Bachute, M.R.; Subhedar, J.M. Autonomous driving architectures: Insights of machine learning and deep learning algorithms. *Mach. Learn. Appl.* **2021**, *6*, 100164.
71. Wang, W.; Wang, L.; Zhang, C.; Liu, C.; Sun, L. Social interactions for autonomous driving: A review and perspectives. *Found. Trends® Robot.* **2022**, *10*, 198–376.
72. Han, Y.; Zhang, H.; Li, H.; Jin, Y.; Lang, C.; Li, Y. Collaborative perception in autonomous driving: Methods, datasets, and challenges. *IEEE Intell. Transp. Syst. Mag.* **2023**, *15*, 131–151.
73. Zhou, Y.; Liu, L.; Zhao, H.; López-Benítez, M.; Yu, L.; Yue, Y. Towards deep radar perception for autonomous driving: Datasets, methods, and challenges. *Sensors* **2022**, *22*, 4208. [CrossRef]
74. Mozaffari, S.; Al-Jarrah, O.Y.; Dianati, M.; Jennings, P.; Mouzakitis, A. Deep learning-based vehicle behavior prediction for autonomous driving applications: A review. *IEEE Trans. Intell. Transp. Syst.* **2020**, *23*, 33–47.

75. Liu, J.; Mao, X.; Fang, Y.; Zhu, D.; Meng, M.Q.H. A survey on deep-learning approaches for vehicle trajectory prediction in autonomous driving. In Proceedings of the 2021 IEEE International Conference on Robotics and Biomimetics (ROBIO), Sanya, China, 27–31 December 2021; pp. 978–985.
76. Zheng, X.; Wu, L.; Yan, Z.; Tang, Y.; Zhao, H.; Zhong, C.; Chen, B.; Gong, J. Large Language Models Powered Context-aware Motion Prediction in Autonomous Driving. In Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Abu Dhabi, United Arab Emirates, 14–18 October 2024; pp. 980–985.
77. Treiber, M.; Hennecke, A.; Helbing, D. Congested traffic states in empirical observations and microscopic simulations. *Phys. Rev. E* **2000**, *62*, 1805.
78. Thrun, S.; Montemerlo, M.; Dahlkamp, H.; Stavens, D.; Aron, A.; Diebel, J.; Fong, P.; Gale, J.; Halpenny, M.; Hoffmann, G.; et al. Stanley: The robot that won the DARPA Grand Challenge. *J. Field Robot.* **2006**, *23*, 661–692.
79. Bacha, A.; Bauman, C.; Faruque, R.; Fleming, M.; Terwelp, C.; Reinholtz, C.; Hong, D.; Wicks, A.; Alberi, T.; Anderson, D.; et al. Odin: Team victortango's entry in the darpa urban challenge. *J. Field Robot.* **2008**, *25*, 467–492.
80. Codevilla, F.; Müller, M.; López, A.; Koltun, V.; Dosovitskiy, A. End-to-end driving via conditional imitation learning. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 4693–4700.
81. Rhinehart, N.; McAllister, R.; Kitani, K.; Levine, S. Precog: Prediction conditioned on goals in visual multi-agent settings. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 27 October–2 November 2019; pp. 2821–2830.
82. Wang, Y.; Mao, Q.; Zhu, H.; Deng, J.; Zhang, Y.; Ji, J.; Li, H.; Zhang, Y. Multi-modal 3d object detection in autonomous driving: A survey. *Int. J. Comput. Vis.* **2023**, *131*, 2122–2152.
83. Chib, P.S.; Singh, P. Recent advancements in end-to-end autonomous driving using deep learning: A survey. *IEEE Trans. Intell. Veh.* **2023**, *9*, 103–118. [CrossRef]
84. Coelho, D.; Oliveira, M. A review of end-to-end autonomous driving in urban environments. *IEEE Access* **2022**, *10*, 75296–75311.
85. Wang, T.H.; Maalouf, A.; Xiao, W.; Ban, Y.; Amini, A.; Rosman, G.; Karaman, S.; Rus, D. Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024; pp. 6687–6694.
86. Jia, X.; Wu, P.; Chen, L.; Xie, J.; He, C.; Yan, J.; Li, H. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 21983–21994.
87. Sharan, S.; Pittaluga, F.; Chandraker, M. Llm-assist: Enhancing closed-loop planning with language-based reasoning. *arXiv* **2023**, arXiv:2401.00125.
88. Wen, L.; Fu, D.; Li, X.; Cai, X.; Ma, T.; Cai, P.; Dou, M.; Shi, B.; He, L.; Qiao, Y. Dilu: A knowledge-driven approach to autonomous driving with large language models. *arXiv* **2023**, arXiv:2309.16292.
89. Liu, J.; Hang, P.; Qi, X.; Wang, J.; Sun, J. Mtd-gpt: A multi-task decision-making gpt model for autonomous driving at unsignalized intersections. In Proceedings of the 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 24–28 September 2023; pp. 5154–5161.
90. Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Wang, Z. Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles. *IEEE Intell. Transp. Syst. Mag.* **2024**, *16*, 81–94.
91. Tanahashi, K.; Inoue, Y.; Yamaguchi, Y.; Yaginuma, H.; Shiotsuka, D.; Shimatani, H.; Iwamasa, K.; Inoue, Y.; Yamaguchi, T.; Igari, K.; et al. Evaluation of large language models for decision making in autonomous driving. *arXiv* **2023**, arXiv:2312.06351.
92. Hwang, J.J.; Xu, R.; Lin, H.; Hung, W.C.; Ji, J.; Choi, K.; Huang, D.; He, T.; Covington, P.; Sapp, B.; et al. Emma: End-to-end multimodal model for autonomous driving. *arXiv* **2024**, arXiv:2410.23262.
93. Nouri, A.; Cabrero-Daniel, B.; Törner, F.; Sivencrona, H.; Berger, C. Engineering safety requirements for autonomous driving with large language models. In Proceedings of the 2024 IEEE 32nd International Requirements Engineering Conference (RE), Reykjavik, Iceland, 24–28 June 2024; pp. 218–228.
94. Shao, H.; Hu, Y.; Wang, L.; Song, G.; Waslander, S.L.; Liu, Y.; Li, H. Lmdrive: Closed-loop end-to-end driving with large language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 May 2024; pp. 15120–15130.
95. Tang, Y.; Da Costa, A.A.B.; Zhang, X.; Patrick, I.; Khastgir, S.; Jennings, P. Domain knowledge distillation from large language model: An empirical study in the autonomous driving domain. In Proceedings of the 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 24–28 September 2023; pp. 3893–3900.
96. Deng, Y.; Yao, J.; Tu, Z.; Zheng, X.; Zhang, M.; Zhang, T. Target: Automated scenario generation from traffic rules for testing autonomous vehicles. *arXiv* **2023**, arXiv:2305.06018.

97. Wang, S.; Sheng, Z.; Xu, J.; Chen, T.; Zhu, J.; Zhang, S.; Yao, Y.; Ma, X. ADEPT: A testing platform for simulated autonomous driving. In Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering, Rochester, MI, USA, 10–14 October 2022; pp. 1–4.
98. Lu, Q.; Wang, X.; Jiang, Y.; Zhao, G.; Ma, M.; Feng, S. Multimodal large language model driven scenario testing for autonomous vehicles. *arXiv* **2024**, arXiv:2409.06450.
99. Tan, S.; Ivanovic, B.; Weng, X.; Pavone, M.; Kraehenbuehl, P. Language conditioned traffic generation. *arXiv* **2023**, arXiv:2307.07947.
100. Wang, P.; Wei, X.; Hu, F.; Han, W. Transgpt: Multi-modal generative pre-trained transformer for transportation. *arXiv* **2024**, arXiv:2402.07233.
101. Fu, D.; Li, X.; Wen, L.; Dou, M.; Cai, P.; Shi, B.; Qiao, Y. Drive like a human: Rethinking autonomous driving with large language models. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2024; pp. 910–919.
102. Wang, P.; Zhu, M.; Zheng, X.; Lu, H.; Zhong, H.; Chen, X.; Shen, S.; Wang, X.; Wang, Y.; Wang, F.Y. Bevgpt: Generative pre-trained foundation model for autonomous driving prediction, decision-making, and planning. *IEEE Trans. Intell. Veh.* 2024, *early access*.
103. Guan, Y.; Liao, H.; Li, Z.; Hu, J.; Yuan, R.; Li, Y.; Zhang, G.; Xu, C. World models for autonomous driving: An initial survey. *IEEE Trans. Intell. Veh.* 2024, *early access*.
104. Zhang, J.; Li, J. Testing and verification of neural-network-based safety-critical control software: A systematic literature review. *Inf. Softw. Technol.* **2020**, *123*, 106296.
105. Tahir, Z.; Alexander, R. Coverage based testing for V&V and safety assurance of self-driving autonomous vehicles: A systematic literature review. In Proceedings of the 2020 IEEE International Conference On Artificial Intelligence Testing (AITest), Oxford, UK, 3–6 August 2020; pp. 23–30.
106. Feng, D.; Harakeh, A.; Waslander, S.L.; Dietmayer, K. A review and comparative study on probabilistic object detection in autonomous driving. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 9961–9980.
107. Singh, S.; Saini, B.S. Autonomous cars: Recent developments, challenges, and possible solutions. *Iop Conf. Ser. Mater. Sci. Eng.* **2021**, *1022*, 012028.
108. Cantini, R.; Orsino, A.; Talia, D. Xai-driven knowledge distillation of large language models for efficient deployment on low-resource devices. *J. Big Data* **2024**, *11*, 63.
109. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Conditional prompt learning for vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16816–16825.
110. Ren, M.; Cao, B.; Lin, H.; Liu, C.; Han, X.; Zeng, K.; Wan, G.; Cai, X.; Sun, L. Learning or self-aligning? rethinking instruction fine-tuning. *arXiv* **2024**, arXiv:2402.18243.
111. Zhang, X.; Li, Z.; Gong, Y.; Jin, D.; Li, J.; Wang, L.; Zhu, Y.; Liu, H. Openmpd: An open multimodal perception dataset for autonomous driving. *IEEE Trans. Veh. Technol.* **2022**, *71*, 2437–2447.
112. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the opportunities and risks of foundation models. *arXiv* **2021**, arXiv:2108.07258.
113. Xu, X.; Li, M.; Tao, C.; Shen, T.; Cheng, R.; Li, J.; Xu, C.; Tao, D.; Zhou, T. A survey on knowledge distillation of large language models. *arXiv* **2024**, arXiv:2402.13116.
114. Zheng, H.; Cao, D.; Xu, J.; Ai, R.; Gu, W.; Yang, Y.; Liang, Y. Distilling Temporal Knowledge with Masked Feature Reconstruction for 3D Object Detection. *arXiv* **2024**, arXiv:2401.01918.
115. Liu, J.; Xu, C.; Hang, P.; Sun, J.; Ding, M.; Zhan, W.; Tomizuka, M. Language-Driven Policy Distillation for Cooperative Driving in Multi-Agent Reinforcement Learning. *arXiv* **2024**, arXiv:2410.24152.
116. Liu, C.; Kang, Y.; Zhao, F.; Kuang, K.; Jiang, Z.; Sun, C.; Wu, F. Evolving Knowledge Distillation with Large Language Models and Active Learning. *arXiv* **2024**, arXiv:2403.06414.
117. Alaba, S.Y.; Gurbuz, A.C.; Ball, J.E. Emerging Trends in Autonomous Vehicle Perception: Multimodal Fusion for 3D Object Detection. *World Electr. Veh. J.* **2024**, *15*, 20. [CrossRef]
118. Duan, Y.; Zhang, Q.; Xu, R. Prompting Multi-Modal Tokens to Enhance End-to-End Autonomous Driving Imitation Learning with LLMs. *arXiv* **2024**, arXiv:2404.04869.
119. Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D.L.; Han, S. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 2774–2781.
120. Ma, Z.; Sun, Q.; Matsumaru, T. Bidirectional Planning for Autonomous Driving Framework with Large Language Model. *Sensors* **2024**, *24*, 6723. [CrossRef] [PubMed]
121. Fu, J.; Gao, C.; Wang, Z.; Yang, L.; Wang, X.; Mu, B.; Liu, S. Eliminating Cross-modal Conflicts in BEV Space for LiDAR-Camera 3D Object Detection. *arXiv* **2024**, arXiv:2403.07372.

122. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
123. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11621–11631.
124. Mei, J.; Zhu, A.Z.; Yan, X.; Yan, H.; Qiao, S.; Chen, L.C.; Kretzschmar, H. Waymo open dataset: Panoramic video panoptic segmentation. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 53–72.
125. Huang, X.; Cheng, X.; Geng, Q.; Cao, B.; Zhou, D.; Wang, P.; Lin, Y.; Yang, R. The apolloscape dataset for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 954–960.
126. Qian, T.; Chen, J.; Zhuo, L.; Jiao, Y.; Jiang, Y.G. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 4542–4550.
127. Wu, D.; Han, W.; Wang, T.; Liu, Y.; Zhang, X.; Shen, J. Language prompt for autonomous driving. *arXiv* **2023**, arXiv:2309.04379.
128. Pitropov, M.; Garcia, D.E.; Rebello, J.; Smart, M.; Wang, C.; Czarnecki, K.; Waslander, S. Canadian adverse driving conditions dataset. *Int. J. Robot. Res.* **2021**, *40*, 681–690. [CrossRef]
129. Mao, J.; Niu, M.; Jiang, C.; Liang, H.; Chen, J.; Liang, X.; Li, Y.; Ye, C.; Zhang, W.; Li, Z.; et al. One million scenes for autonomous driving: Once dataset. *arXiv* **2021**, arXiv:2106.11037.
130. Malla, S.; Choi, C.; Dwivedi, I.; Choi, J.H.; Li, J. Drama: Joint risk localization and captioning in driving. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 1043–1052.
131. Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Beißwenger, J.; Luo, P.; Geiger, A.; Li, H. Drivelm: Driving with graph visual question answering. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer: Berlin/Heidelberg, Germany, 2025; pp. 256–274.
132. Ding, X.; Han, J.; Xu, H.; Liang, X.; Zhang, W.; Li, X. Holistic Autonomous Driving Understanding by Bird's-Eye-View Injected Multi-Modal Large Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 13668–13677.
133. Nie, M.; Peng, R.; Wang, C.; Cai, X.; Han, J.; Xu, H.; Zhang, L. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer: Berlin/Heidelberg, Germany, 2025; pp. 292–308.
134. Chen, L.; Sinavski, O.; Hünermann, J.; Karnsund, A.; Willmott, A.J.; Birch, D.; Maund, D.; Shotton, J. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024; pp. 14093–14100.
135. Tian, R.; Li, B.; Weng, X.; Chen, Y.; Schmerling, E.; Wang, Y.; Ivanovic, B.; Pavone, M. Tokenize the world into object-level knowledge to address long-tail events in autonomous driving. *arXiv* **2024**, arXiv:2407.00959.
136. Ma, Y.; Yaman, B.; Ye, X.; Tao, F.; Mallik, A.; Wang, Z.; Ren, L. MTA: Multimodal Task Alignment for BEV Perception and Captioning. *arXiv* **2024**, arXiv:2411.10639.
137. Liu, Y.; Ding, P.; Huang, S.; Zhang, M.; Zhao, H.; Wang, D. PiTe: Pixel-Temporal Alignment for Large Video-Language Model. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer: Berlin/Heidelberg, Germany, 2025; pp. 160–176.
138. Gunjal, A.; Yin, J.; Bas, E. Detecting and preventing hallucinations in large vision language models. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 26–27 February 2024; Volume 38, pp. 18135–18143.
139. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
140. Liu, M.; Zhu, C.; Ren, S.; Yin, X.C. Unsupervised multi-view pedestrian detection. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 1034–1042.
141. Dona, M.A.M.; Cabrero-Daniel, B.; Yu, Y.; Berger, C. Evaluating and Enhancing Trustworthiness of LLMs in Perception Tasks. *arXiv* **2024**, arXiv:2408.01433.
142. Wu, X.; Guan, T.; Li, D.; Huang, S.; Liu, X.; Wang, X.; Xian, R.; Shrivastava, A.; Huang, F.; Boyd-Graber, J.L.; et al. AUTOHALLUSION: Automatic Generation of Hallucination Benchmarks for Vision-Language Models. *arXiv* **2024**, arXiv:2406.10900.
143. Liang, Y.; Song, Z.; Wang, H.; Zhang, J. Learning to trust your feelings: Leveraging self-awareness in llms for hallucination mitigation. *arXiv* **2024**, arXiv:2401.15449.
144. Jiang, C.; Jia, H.; Dong, M.; Ye, W.; Xu, H.; Yan, M.; Zhang, J.; Zhang, S. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 525–534.

145. Manakul, P.; Liusie, A.; Gales, M.J. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv* **2023**, arXiv:2303.08896.
146. Zhang, T.; Wang, L.; Zhang, X.; Zhang, Y.; Jia, B.; Liang, S.; Hu, S.; Fu, Q.; Liu, A.; Liu, X. Visual Adversarial Attack on Vision-Language Models for Autonomous Driving. *arXiv* **2024**, arXiv:2411.18275.
147. Shalev-Shwartz, S.; Shammah, S.; Shashua, A. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv* **2016**, arXiv:1610.03295.
148. Yurtsever, E.; Lambert, J.; Carballo, A.; Takeda, K. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access* **2020**, *8*, 58443–58469. [CrossRef]
149. Liu, D.; Yang, M.; Qu, X.; Zhou, P.; Cheng, Y.; Hu, W. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv* **2024**, arXiv:2407.07403.
150. Bhagwatkar, R.; Nayak, S.; Bashivan, P.; Rish, I. Improving Adversarial Robustness in Vision-Language Models with Architecture and Prompt Design. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 12–16 November 2024; pp. 17003–17020.
151. Song, R.; Ozmen, M.O.; Kim, H.; Bianchi, A.; Celik, Z.B. Enhancing llm-based autonomous driving agents to mitigate perception attacks. *arXiv* **2024**, arXiv:2409.14488.
152. Yang, B.; Ji, X.; Jin, Z.; Cheng, Y.; Xu, W. Exploring Adversarial Robustness of LiDAR-Camera Fusion Model in Autonomous Driving. In Proceedings of the 2023 IEEE 7th Conference on Energy Internet and Energy System Integration (EI2), Hangzhou, China, 15–18 December 2023; pp. 3634–3639.
153. Zhang, T.; Wang, L.; Kang, J.; Zhang, X.; Liang, S.; Chen, Y.; Liu, A.; Liu, X. Module-wise Adaptive Adversarial Training for End-to-end Autonomous Driving. *arXiv* **2024**, arXiv:2409.07321.
154. Sadat, A.; Casas, S.; Ren, M.; Wu, X.; Dhawan, P.; Urtasun, R. Perceive, predict, and plan: Safe motion planning through interpretable semantic representations. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXIII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 414–430.
155. Luo, W.; Yang, B.; Urtasun, R. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3569–3577.
156. Liang, M.; Yang, B.; Zeng, W.; Chen, Y.; Hu, R.; Casas, S.; Urtasun, R. Pnpnet: End-to-end perception and prediction with tracking in the loop. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11553–11562.
157. Mądry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. *Statistics* **2017**, *arXiv:1706.06083* .
158. Liu, A.; Liu, X.; Yu, H.; Zhang, C.; Liu, Q.; Tao, D. Training robust deep neural networks via adversarial noise propagation. *IEEE Trans. Image Process.* **2021**, *30*, 5769–5781.
159. Li, X.; Liu, J.; Ma, L.; Fan, X.; Liu, R. Advmono3d: Advanced monocular 3d object detection with depth-aware robust adversarial training. *arXiv* **2023**, arXiv:2309.01106.
160. Zhang, Y.; Hou, J.; Yuan, Y. A comprehensive study of the robustness for lidar-based 3d object detectors against adversarial attacks. *Int. J. Comput. Vis.* **2024**, *132*, 1592–1624.
161. Cao, Y.; Xiao, C.; Anandkumar, A.; Xu, D.; Pavone, M. Advdo: Realistic adversarial attacks for trajectory prediction. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 36–52.
162. Zhang, Q.; Hu, S.; Sun, J.; Chen, Q.A.; Mao, Z.M. On adversarial robustness of trajectory prediction for autonomous vehicles. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 15159–15168.
163. Zhang, T.; Wang, L.; Li, H.; Xiao, Y.; Liang, S.; Liu, A.; Liu, X.; Tao, D. Lanevil: Benchmarking the robustness of lane detection to environmental illusions. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 5403–5412.
164. Zhang, X.; Liu, A.; Zhang, T.; Liang, S.; Liu, X. Towards robust physical-world backdoor attacks on lane detection. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, VIC, Australia, 28 October–1 November 2024; pp. 5131–5140.

Applications of large language models and multimodal large models in autonomous driving: a comprehensive review

Li, Jing

2025-04-01

Li J, Li J, Yang G, et al., (2025) Applications of large language models and multimodal large models in autonomous driving: a comprehensive review. Drones, Volume 9, Issue 4, April 2025, Article number 238

https://doi.org/10.3390/drones9040238