A Survey on Multimodal Large Language Models for Autonomous Driving

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tong Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, Ziran Wang, Chao Zheng, Purdue University, West Lafayette, IN, USA 47907
Tencent T Lab, Beijing, China
University of Illinois Urbana-Champaign, Champaign, IL, USA 61801
University of Virginia, Charlottesville, VA, USA 22903
New York University, New York, NY, USA 11201
PediaMed AI, Shenzhen, China
SambaNova Systems, Inc, Palo Alto, CA, USA 94303
Objective, Inc, San Francisco, CA, USA 94110

Abstract

With the emergence of Large Language Models (LLMs) and Vision Foundation Models (VFMs), multimodal AI systems benefiting from large models have the potential to equally perceive the real world, make decisions, and control tools as humans. In recent months, LLMs have shown widespread attention in autonomous driving and map systems. Despite its immense potential, there is still a lack of a comprehensive understanding of key challenges, opportunities, and future endeavors to apply in LLM driving systems. In this paper, we present a systematic investigation in this field. We first introduce the background of Multimodal Large Language Models (MLLMs), the multimodal models development using LLMs, and the history of autonomous driving. Then, we overview existing MLLM tools for driving, transportation, and map systems together with existing datasets and benchmarks. Moreover, we summarized the works in The 1st WACV Workshop on Large Language and Vision Models for Autonomous Driving (LLVM-AD), which is the first workshop of its kind regarding LLMs in autonomous driving. To further promote the development of this field, we also discuss several important problems regarding using MLLMs in autonomous driving systems that need to be solved by both academia and industry.

1. Introduction

Large Language Models (LLMs) have gained significant attention recently, showing remarkable potential in emulating human-like intelligence. The advancements have sparked a surge of enthusiasm in Multimodal Large Language Models (MLLMs) [199], which merge the sophisticated reasoning prowess of LLMs with image, video, and audio data. The modality alignment enables them to execute a diverse array of tasks with higher proficiency, including categorizing images, matching text to corresponding videos, and speech detection. Besides, [174] proved that LMs can deal with easy tasks within the robotics domain including basic logical, geometrical, and mathematical rea-
You are a professional self-driving agent and good at observing LiDAR point clouds and making driving decisions. I will show you some BEV point cloud images. Please use simple, clear, structured language to describe the road condition, surroundings, abnormal situation and driving strategy.

**GPT-4V Answer:** The BEV LiDAR image suggests a multi-lane road, but the exact number of lanes and presence of traffic are unclear. There are no visible obstacles, pedestrians, or significant environmental features in the immediate vicinity of the vehicle. The image lacks detail on road surface quality and surrounding signage or traffic lights. In this scenario, maintaining a steady speed while staying within the detected lanes, and being alert for unseen hazards is advisable.

As the AI system on an autonomous vehicle, your role is to analyze the driving context and make informed decisions. You will receive an image from the forward-facing camera of the ego vehicle. Please use concise and organized language to describe the current environment, any anomalies, and plan the next steps.

**GPT-4V Answer:** Driving in an urban setting, the forward-facing camera captures a clear day with multi-story buildings lining both sides of a road, with some patches appearing wet or muddy. The road, populated with cars, motorcycles, and a three-wheeled vehicle. An anomaly to note is the three-wheeled vehicle on the left, which is carrying a heavy load that might affect its stability. Given the situation, the next steps involve maintaining a safe distance from this vehicle, monitoring motorcycles for sudden movements, scanning for traffic signals or signages, and adjusting speed to match the traffic flow.

Make a right lane change.

**GPT-4 Code Generation:**

```python
def right_lane_change():
    target_lane = get_right_lane(vehicle=ego)
    while True:
        if is_safe_enter_lane(target_lane):
            break
        yield autopilot()
    set_target_lane(target_lane)
```

Figure 2. Exploring GPT-4V [127] to understand driving scenes and make driving actions. Our findings reveal that while GPT-4V adeptly identifies scene components such as objects, it falls short in recognizing critical traffic elements like lane information. This underscores the significant challenges yet to be overcome in advancing multimodal language models for reliable autonomous vehicle navigation.

...
real-time text-based information sources, such as HD maps, traffic reports, and weather updates, enabling the vehicle to attain a more comprehensive understanding of its surroundings [30]. A good example is to improve the navigation in the vehicle-mounted maps. LLMs can process real-time traffic data to identify congested routes and suggest alternative paths, ultimately optimizing navigation for efficiency and safety [159]. For motion planning, LLMs play a role by utilizing their natural language understanding and reasoning [110]. They facilitate user-centric communication and enable passengers to express their intentions and preferences using everyday language. Additionally, LLMs also process textual data sources such as maps, traffic reports, and real-time information, and then make high-level decisions for optimized route planning [124]. In the context of motion control, LLMs, firstly, enable the customization of controller parameters to align with driver preferences, achieving personalization in the driving experience [150]. Additionally, LLMs can provide transparency by explaining each step of the motion control process.

MLLMs represent the next level of LLMs, bringing together the power of language understanding with the capability to process and integrate diverse data modalities [39, 199]. Within the landscape of autonomous driving, the significance of MLLMs is huge and transformative. Vehicles equipped with MLLMs can deal with information from textual input with other features captured by onboard cameras and other sensors, offering easier learning of complex traffic scenes and driving behaviors. Beyond autonomous driving, MLLMs can also significantly enhance personalized human-vehicle interaction through voice communication and user preference analysis. In future SAE L4-L5 autonomous vehicles, passengers could communicate their requests while driving using language, gestures, or even gazes, with the MLLMs offering real-time in-cabin feedback by integrating visual displays or voice responses.

In our pursuit to bridge the domains of autonomous driving and advanced modeling, we co-organized the inaugural Workshop on Large Language and Vision Models for Autonomous Driving (LLVM-AD) at the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). This event is designed to enhance collaboration between academic researchers and industry professionals, exploring the possibility and challenges of implementing multimodal large language models in the field of autonomous driving. LLVM-AD also launched a follow-up open-source real-world traffic language understanding dataset, catalyzing practical advancements.

The main contributions of this paper are summarized as follows:

- A brief overview of the background of current MLLMs and autonomous driving technologies is provided.
- The benefits of using LLMs and MLLMs in autonomous driving are outlined, highlighting their roles and current works in perception, motion planning, motion control, and recently declared industry applications.
- Datasets relevant to autonomous driving are summarized, with an emphasis on driving language datasets for traffic scenes.
- The accepted papers from the WACV LLVM-AD Workshop are reviewed, providing insights into future directions of LLMs and MLLMs in autonomous driving.

As Figure 1 shows, our survey paper aims to provide a comprehensive overview of MLLMs for autonomous driving and discuss growing trends, and future directions. The following two sections provide a brief description of the developmental history of autonomous driving and MLLMs separately. Section 4 presents current published works about MLLMs for autonomous driving in perception, motion planning, and motion control. Section 5 introduces related autonomous driving industry applications utilizing MLLMs. In the last three sections, we summarize the papers in the 1st WACV LLVM-AD workshop and discuss potential research directions for LLMs and MLLMs for autonomous driving.

2. Development of Autonomous Driving

The quest for autonomous driving has been a progressive journey, marked by a continuous interplay between visionary aspirations and technological capabilities. The first wave of comprehensive research on autonomous driving started in the late 20th century. For example, the Autonomous Land Vehicle (ALV) project launched by Carnegie Mellon University utilized sensor readings from stereo cameras, sonars, and the ERIM laser scanner to perform tasks like lane keeping and obstacle avoidance [70, 134]. However, these researches were constrained by limited sensor accuracy and computation capabilities.

The last two decades have seen rapid improvements in autonomous driving systems. A classification system published by the Society of Automotive Engineers (SAE) in 2014 defined six levels of autonomous driving systems [28]. The classification method has now been widely acknowledged and illustrated important milestones for the research and development progress. The introduction of Deep Neural Networks (DNNs) has also played a significant role [48, 85]. Backed by deep learning, computer vision has been crucial for interpreting complex driving environments, offering state-of-the-art solutions for problems such as object detection, scene understanding, and vehicle localization [65, 90, 136]. Deep Reinforcement Learning (DRL) has
The autonomous driving technology has evolved significantly over several decades. It began with early explorations and advancements like the ALV Project by Carnegie Mellon University [70,172], Mitsubishi Debonair, the first to offer LiDAR-based ADAS system [120], and the winner of 2005 DARPA Grand Challenge Stanley by Stanford University [166]. This was followed by recent achievements after the introduction of a standardized level of automation [28] and rapid progress in Deep Neural Networks. Autonomous driving platform-wise, various open-source and commercialized software solutions are introduced, such as Tesla Autopilot [118], NVIDIA DRIVE, Autoware.AI [73,74], Baidu Apollo [8], and PonyAlpha [135]. Regulatory and service-wise, autonomous driving technology is receiving increasing government acceptance and public acknowledgment, with numerous companies receiving permits to operate autonomous driving vehicles on public roads in designated regions while more vehicles with autonomous driving capabilities are being mass-produced [49]. Overall, it demonstrates the evolution and increasing sophistication of AD systems over several decades.

Additionally, improved sensor accuracy and computation power improvements allow larger models with more accurate results to be run on the vehicle. With such improvements, more L1 to L2 level Advanced Driver Assistance Systems (ADAS) like lane centering and adaptive cruise control are now available on everyday vehicles [11, 21]. Companies like Waymo, Zoox, Cruise, and Baidu are also rolling out Robotaxis with Level 3 or higher autonomy. Nevertheless, such autonomous systems still fail in many driving edge cases such as extreme weather, bad lighting conditions, or rare situations [32].

3. Development of Multimodal Language Models

3.1. Development of Language Models

The development of language models has been a journey marked by significant breakthroughs. Since the early 1960s, many linguists, most renowned Noam Chomsky, attempted to model natural languages [24]. Early efforts focused mainly on rule-based approaches [9, 56, 123]. However, in the late 1980s and early 1990s, the spotlight shifted onto statistic models, such as N-gram [13], hidden Markov models [40], which relied on counting the frequency of words and sequences in text data. The 2000s witnessed the introduction of neural networks into natural language modeling. Recurrent Neural Networks (RNNs) [148] and Long Short-Term Memory (LSTM) networks [55] were used for various NLP tasks.

Despite their potential, early neural models had limitations in capturing long-range dependencies and struggled with complex language tasks. In 2013, Tomas Mikolov and his team at Google introduced Word2Vec [113], a groundbreaking technique for representing words as dense vectors, providing a better understanding of semantic relationships between words. This laid down the foundation for the rise of deep learning [27,162], which eventually led to the pivotal work, Attention is all you need [173], which kick-started the new era of large language models. [14, 25, 34, 141, 142].
3.2. Advancements in Large Language Models

LLMs are a category of Transformer-based language models known for their extensive number of parameters, often numbering in the hundreds of billions. These models are trained on vast amounts of internet data, which enables them to perform a wide range of language tasks, primarily through text generation. Some well-known examples of LLMs include GPT-3 [14], PaLM [25], LLaMA [169], and GPT-4 [126]. One of the most notable characteristics of LLMs is their emergent abilities, such as in-context learning (ICL) [14], instruction following [129], and reasoning with chain-of-thought (CoT) [184].

There is a growing area of research that utilizes LLMs to develop autonomous agents with human-like capabilities. These agents leverage the extensive knowledge stored in pre-trained LLMs to create coherent action plans and executable policies [2, 39, 60, 61, 96, 176]. Embodied language models [39] directly integrate real-world sensor data with language models, establishing a direct connection between words and perceptual information. Voyager [176] introduces lifelong learning by incorporating three main components: an automatic curriculum that promotes exploration, a skill library to store and retrieve complex behaviors, and an iterative prompting mechanism to generate executable code for embodied control. Voxposer [61] utilizes LLMs to generate robot trajectories for a wide range of manipulation tasks, guided by open-ended instructions and objects.

In parallel with these advancements, the use of LLMs in the field of autonomous driving is gaining momentum. Recent research [41, 68] has investigated the application of LLMs to comprehend driving environments. These studies have demonstrated the impressive ability of LLMs to handle complex scenarios by converting visual information into text representation, enabling LLMs to interpret the surrounding world. Similarly, in RRR [30], authors propose a human-centric autonomous driving framework that breaks down user commands into a series of intermediate reasoning steps, accompanied by a detailed list of action descriptions to accomplish the objective.

3.3. Early Efforts in Modality Fusion

Over the past few decades, the fusion of various modalities such as vision, language, video, and audio has been a key objective in artificial intelligence (AI). Initial efforts in this domain focused on simple tasks, such as image or video captioning and text-based image retrieval, which were mostly rule-based and relied on hand-crafted features. A classic example of early AI problems in the 1970s and 1980s was the "Blocks World" [158], where the goal was to rearrange colored blocks on a table based on textual instructions. This early attempt bridged vision (understanding block configurations) with language (interpreting and executing instructions), even though it was not based on deep learning.

3.4. Advancements in Vision-Language Models

In the following years, the field of multimodal models saw significant advancements. Over the last decade, the advent of deep learning has revolutionized approaches to visual-language tasks. Convolutional Neural Networks (CNNs) [83] became the de facto standard for image and video processing, while Recurrent Neural Networks (RNNs) [55, 148] emerged as the go-to models for processing sequential data, such as natural languages. During this period, popular tasks included image and video captioning, which involves generating descriptive sentences for images and videos, and visual question answering (VQA), where models answer questions related to visual data. Typical vision-language models employed joint embeddings, with image features (processed by CNNs) and text features (processed by RNNs or Transformers [173]) mapped to a shared semantic space to facilitate multimodal learn-
ing [6, 72, 111, 175]. Beyond vision and language, researchers also proposed models for other modalities, such as audio, speech, and 3D data. For instance, Mroueh et al. (2015) developed a deep multimodal learning model for audio-visual speech recognition that utilizes CNNs for visual data and RNNs for audio data [119]. Arandjelović and Zisserman (2017) explored the relationship between visual and auditory data by developing a model that learns shared representations from unlabeled videos, using CNNs for both image and audio processing [7]. Furthermore, Qi et al. (2016) introduced models that process 3D data, including point clouds, for object classification tasks, employing CNNs to learn representations from volumetric data and multiple 2D views of 3D objects [137]. These works highlight the potential of multimodal learning in capturing complex relationships between different types of data, leading to richer and more accurate representations.

3.5. Pre-Training and Multimodal Transformers

Building on this momentum, the field of multimodal models has continued to evolve, with researchers exploring the potential of pre-training multimodal models on extensive datasets before fine-tuning them on specific tasks. This approach has resulted in significant performance improvements across a range of applications. Inspired by the success of pre-trained NLP models like BERT [34], T5 [142], and GPTs [14, 140], researchers developed multimodal Transformers that can process cross-modality inputs such as text, image, audio, pointcloud [45, 51, 59]. Notable examples of visual-language models include CLIP [139], ViLBERT [100], VisualBERT [95], SimVLM [181], BLIP-2 [94] and Flamingo [3], which were pre-trained on large-scale cross-modal datasets comprising images and languages. Other works have explored the use of multimodal models for tasks such as video understanding [210], audio-visual scene understanding [4], and even 3D data processing [53]. Pre-training allows the models to align different modalities and enhance the representation learning ability of the model encoder. By doing so, these models aim to create systems that can generalize across tasks without the need for task-specific training data. Furthermore, the evolution of multimodal models has also given rise to new and exciting possibilities. For instance, DALL-E [144] extends the GPT-3 architecture to generate images from textual descriptions, Stable Diffusion [145] and ControlNet [204] utilized CLIP and UNet-based diffusion model to generate images controlled by text prompt. They showcase the potential for using multimodal models in many application scenarios such as healthcare [97], civil engineering [133], robotics [71] and, art [80].

3.6. Emergence of Multimodal Large Language Models

Recently, MLLMs have emerged as a significant area of research. These models leverage the power of LLMs, such as ChatGPT [125], InstructGPT [129], FLAN [26, 183], and OPT-IML [64] to perform tasks across multiple modalities such as text and images. They exhibit surprising emergent capabilities, such as writing stories based on images and performing OCR-free math reasoning, which are rare in traditional methods. This suggests a potential path to artificial general intelligence. Key techniques and applications in MLLMs include Multimodal Instruction Tuning, which tunes the model to follow instructions across different modalities [98, 197, 209]; Multimodal In-Context Learning, which allows the model to learn from the context of multimodal data [38, 52, 101, 102, 196]; Multimodal Chain of Thought, which enables the model to maintain a chain of thought across different modalities [43, 54, 146, 206]; and LLM-Aided Visual Reasoning (LAVR), which uses LLMs to aid in visual reasoning tasks [52, 101, 154, 178, 188, 196, 205, 211]. MLLMs are more in line with the way humans perceive the world, offering a more user-friendly interface and supporting a larger spectrum of tasks compared to LLMs. The recent progress of MLLMs has been ignited by the development of GPT-4V [127], which, despite not having an open multimodal interface, has shown amazing capabilities. The research community has made significant efforts to develop capable and open-sourced MLLMs, exhibiting surprising practical capabilities.

4. Multimodal Language Models for Autonomous Driving

In the autonomous driving industry, MLLMs have the potential to understand traffic scenes, improve the decision-making process for driving, and revolutionize the interaction between humans and vehicles. These models are trained on vast amounts of traffic scene data, allowing them to extract valuable information from different sources like maps, videos, and traffic regulations. As a result, they can enhance a vehicle’s navigation and planning, ensuring both safety and efficiency. Additionally, they can adapt to changing road conditions with a level of understanding that closely resembles human intuition.

4.1. Multimodal Language Models for Perception

Traditional perception systems are often limited in their ability to recognize only a specific set of predefined object categories. This restricts their adaptability and requires the cumbersome process of collecting and annotating new data to recognize different visual concepts. As a result, their generality and usefulness are undermined. In contrast, a new paradigm is emerging that involves learning from raw text-


<table>
<thead>
<tr>
<th>Model</th>
<th>Year</th>
<th>Backbone</th>
<th>Task</th>
<th>Modality</th>
<th>Learning</th>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Driving with LLMs [22]</td>
<td>2023</td>
<td>LLaMA</td>
<td>Perception Control</td>
<td>Vector Language</td>
<td>FT Vector</td>
<td>Query Actions</td>
<td>Response</td>
</tr>
<tr>
<td>GAIA-1 [57]</td>
<td>2023</td>
<td>-</td>
<td>Planning</td>
<td>Vision Language</td>
<td>PT Video</td>
<td>Prompt Video</td>
<td>Video</td>
</tr>
<tr>
<td>LMaZP [60]</td>
<td>2022</td>
<td>GPT-3</td>
<td>Planning</td>
<td>Language ICL</td>
<td>Text</td>
<td>Plan</td>
<td></td>
</tr>
<tr>
<td>Dilu [185]</td>
<td>2023</td>
<td>GPT-3.5</td>
<td>Planning</td>
<td>Language ICL</td>
<td>Text</td>
<td>Action</td>
<td></td>
</tr>
<tr>
<td>DaYS [31]</td>
<td>2023</td>
<td>GPT-4</td>
<td>Planning</td>
<td>Language ICL</td>
<td>Text</td>
<td>Code</td>
<td></td>
</tr>
<tr>
<td>RRR [30]</td>
<td>2023</td>
<td>GPT-4</td>
<td>Planning</td>
<td>Language ICL</td>
<td>Text</td>
<td>Action</td>
<td></td>
</tr>
<tr>
<td>DlaH [42]</td>
<td>2023</td>
<td>GPT-3.5</td>
<td>Planning</td>
<td>Language ICL</td>
<td>Text</td>
<td>Action</td>
<td></td>
</tr>
<tr>
<td>GPT-Driver [110]</td>
<td>2023</td>
<td>GPT-3.5</td>
<td>Planning</td>
<td>Vision Language ICL</td>
<td>Text</td>
<td>Trajectory</td>
<td></td>
</tr>
<tr>
<td>SurrealDriver [68]</td>
<td>2023</td>
<td>GPT-4</td>
<td>Planning</td>
<td>Language ICL</td>
<td>Text</td>
<td>Action</td>
<td></td>
</tr>
<tr>
<td>LanguageMPC [150]</td>
<td>2023</td>
<td>GPT-3.5</td>
<td>Planning</td>
<td>Language ICL</td>
<td>Text</td>
<td>Action</td>
<td></td>
</tr>
<tr>
<td>DriveGPT4 [193]</td>
<td>2023</td>
<td>Llama 2</td>
<td>Planning</td>
<td>Vision Language ICL</td>
<td>Text</td>
<td>Action</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Summary of recent research on MLLMs for autonomous driving. The main backbone for current models are LLaMA [168], Llama 2 [169], GPT-3.5 [125], GPT-4 [126], Flan5XXL [26], Vicuna-13b [165]. FT, ICL and PT refer to fine-tuning, in-context learning and pretrained respectively.

Multimodal Large Language Models (MLLMs) have gained significant interest due to their proficiency in analyzing non-textual data like images and point clouds through text analysis [3, 139, 170, 203]. These advancements have greatly improved zero-shot and few-shot image classification [130, 139], segmentation [79, 103], and object detection [115].

Pioneering models like CLIP [139] have shown that training to match images with captions can effectively create image representations from scratch. Building on this, Liu et al. introduced LLaMa [98], which combines a vision encoder with an LLM to enhance the understanding of both visual and linguistic concepts. Zhang et al. further extended this work with Video-LLaMa [203], enabling MLLMs to process visual and auditory information from videos. This represents a significant advancement in machine perception by integrating linguistic and visual modalities.

Furthermore, researchers have explored the use of vectorized visual embeddings to equip MLLMs with environmental perception capabilities, particularly in autonomous driving scenarios. DriveGPT4 [193] interprets video inputs to generate driving-related textual responses. HiLMD [37] focuses on incorporating high-resolution details into MLLMs, improving hazard identification and intention prediction. Similarly, Talk2BEV [35] leverages pre-trained image-language models to combine Bird’s Eye View (BEV) maps with linguistic context, enabling visuo-linguistic reasoning in autonomous vehicles.

At the same time, progress in autonomous driving is not limited to discriminative perception models; generative models are also gaining popularity. One example is the Generative AI for Autonomy model (GAIA-1), which generates realistic driving scenarios by integrating video, text, and action inputs. This generative world model can anticipate various potential outcomes based on the vehicle’s maneuvers, showcasing the sophistication of generative models in adapting to the changing dynamics of the real world [57]. Similarly, UniSim [194] aims to replicate real-world interactions by combining diverse datasets, including objects, scenes, actions, motions, language, and motor controls, into a unified video generation framework. Moreover, the Waymo Open Sim Agents Challenge (WOSAC) [50, 117] is the first public challenge to develop simulations with realistic and interactive agents.
4.2. Multimodal Language Models for Planning and Control

The use of language in planning and control tasks has a longstanding history in robotics, dating back to the use of lexical parsing in natural language for early demonstrations of human-robot interaction [187], and it has been widely studied being used in the robotics area. There exists comprehensive review works on this topic [104, 164]. It has been well-established that language acts as a valuable interface for non-experts to communicate with robots [82]. Moreover, the ability of robotic systems to generalize to new tasks through language-based control has been demonstrated in various works [2, 66]. Achieving specific planning or control tasks or policies, including model-based [5, 121, 153], imitation learning [105, 155], and reinforcement learning [47, 67, 116], has been extensively explored.

Due to the significant ability in zero-shot learning [167], in-context learning [114] and reasoning [184], many works showed that LLMs could enable reasoning of planning [152, 176] and perceiving the environment with textual description [157] to develop user in the loop robotics [174]. [81] broke down natural language commands into sequences of executable actions through a combination of text completion and semantic translation to control the robot. SayCan [2] utilized weighted LLMs to produce reasonable actions and control robots while [62] uses environmental feedback, LLMs can develop an inner monologue, enhancing their capacity to engage in more comprehensive processing within robotic control scenarios. Socratic Models [202] employs visual language models to replace perceptual information within the language prompts used for robot action generation. [96] introduces an approach that uses LLMs to directly generate policy code for robots to do control tasks, specify feedback loops, and write low-level control primitives.

In autonomous driving, LLMs could serve as the bridge to support human-machine interactions. For general purposes, LLMs can be task-agnostic planners. In [60], the authors discovered that pre-trained LLMs contain actionable knowledge for coherent and executable action plans without additional training. Huang et al. [61] proposed the use of LLMs for converting arbitrary natural language commands or task descriptions into specific and detail-listed objectives and constraints. [185] proposed integrating LLMs as decision decoders to generate action sequences following chain-of-thoughts prompting in autonomous vehicles. In [31], authors showcased that LLMs can decompose arbitrary commands from drivers to a set of intermediate phases with a detailed list of descriptions of actions to achieve the objective.

Meanwhile, it is essential to enhance the safety and explainable of autonomous driving. The multimodal language model provides the potential to comprehend its surroundings and the transparency of the decision process. [77] showed that video-to-text models can help generate textual explanations of the environment aligned with downstream controllers. Deruyttere et al. [33] compared baseline models and showed that LLMs can identify specific objects in the surroundings that are related to the commands or descriptions in natural language. For the explainability of the model, Xu et al. [193] proposed to integrate LLMs to generate explanations along with the planned actions. In [31], the authors proposed a framework where LLMs can provide descriptions of how they perceive and react to environmental factors, such as weather and traffic conditions.

Furthermore, the LLMs in autonomous driving can also facilitate the fine-tuning of controller parameters, aligning them with the driver’s preferences and thus resulting in a better driving experience. [150] integrates LLMs into low-level controllers through guided parameter matrix adaptation.

Besides the development of LLMs, great progress has also been witnessed in MLLMs. The MLLMs have the potential to serve as a general and safe planner model for autonomous driving. The ability to process and fuse visual signals such as images enhanced navigation tasks by combining visual cues and linguistic instructions [69, 84]. Interoperability challenges have historically been an issue for autonomous planning processes [23, 46]. However, recent advancements in addressing interoperability challenges in autonomous planning have leveraged the impressive reasoning capabilities of MLLMs during the planning phases of autonomous driving [22, 41]. In one notable approach, Chen et al. [22] integrated vectorized object-level 2D scene representations into a pre-trained LLM with adapters, enabling direct interpretation and comprehensive reasoning about various driving scenarios. Additionally, Fu et al. [41] employed LLMs for reasoning and translated this reasoning into actionable driving behaviors, showing the versatility of LLMs in enhancing autonomous driving planning. Additionally, GPT-Driver [110] reformulated motion planning as a language modeling problem and utilized LLM to describe highly precise trajectory coordinates and its internal decision-making process in natural language in motion planning. SurrealDriver [68] simulated MLLM-based generative driver agents that can perceive complex traffic scenarios and generate corresponding driving maneuvers. [76] investigated the utilization of textual descriptions along with pre-trained language encoders for motion prediction in autonomous driving.

4.3. Industrial Applications

The integration of MLLMs in the autonomous driving industry has been developed by several significant initiatives. Wayve introduces LINGO-1, which enhances the learning and explainability of foundational driving models by inte-
grating vision, language, and action [182]. They also developed GAIA-1, a generative world model for realistic driving scenario generation, offering fine-grained control over vehicle behavior and scene features [57].

Tencent T Lab generated traffic, map, and driving-related context from their HD map AI system [163], creating MAPLM, a large map and traffic scene dataset for scene understanding.

Waymo’s contribution, MotionLM, improved motion prediction in multi-agent environments. By conceptualizing continuous trajectories as discrete motion tokens, it transfers multi-agent motion prediction to a language modeling task [149]. This approach transforms the dynamic interaction of road agents into a manageable sequence-to-sequence prediction problem.

Research from the Bosch Center focuses on using natural language for enhanced scene understanding and predicting future behaviors of surrounding traffic [76]. Meanwhile, researchers from the Hong Kong University of Science and Technology and Huawei Noah’s Ark Lab have leveraged MLLMs to integrate various autonomous driving tasks, including risk object localization and intention and suggestion prediction from videos [57].

These developments in industry illustrate the expanding role of MLLMs in enhancing the capabilities and functionalities of autonomous driving systems, marking a significant improvement in vehicle intelligence and situational awareness.

5. Datasets and Benchmarks

5.1. Vision Datasets for Autonomous Driving

Publicly available datasets have played a crucial role in advancing autonomous driving technologies. Tab. 3 provides a comprehensive overview of the latest representative datasets for autonomous driving. In the past, datasets mainly focused on 2D annotations, like bounding boxes and masks, primarily for RGB camera images [131, 171]. However, achieving autonomous driving capabilities that can match human performance requires precise perception and localization in the 3D environment. Unfortunately, extracting depth information from purely 2D images poses significant challenges.

To enable robust 3D perception or mapping, researchers have created many multimodal datasets. These datasets include not only camera images but also data from 3D sensors like radar and LiDAR. An influential example in this field is the KITTI dataset [44], which provides multimodal sensor data, including front-facing stereo cameras and LiDAR. KITTI also includes annotations of 3D boxes and covers tasks such as 3D object detection, tracking, stereo, and optical flow. Subsequently, NuScenes [15] and the Waymo Open dataset [161] have emerged as representative multi-modal datasets. These datasets set new standards by offering a large number of scenes. These datasets represent a significant advancement in the availability of large data for advancing research in autonomous driving.

5.2. Multimodal-Language Datasets for Traffic Scene

Several pioneering studies have explored language-guided visual understanding in driving scenarios. These studies either enhance existing datasets with additional textual information or create new datasets independently. The former category includes works such as Talk2Car [33], nuScenes-QA [138], DriveLM [29], and NuPrompt [189]. Among these, Talk2Car [33] stands out as the first object referral dataset, which contains natural language commands for autonomous vehicles. On the other hand, datasets like BDD-X [77] and DRAMA [109] were independently created. DRAMA [109] specifically focuses on video and object-level inquiries regarding driving hazards and associated objects. This dataset aims to enable visual captioning through free-form language descriptions and uses both closed and open-ended responses to multi-tiered questions. It allows for the evaluation of various visual captioning abilities in driving contexts.

Despite the advancements in language comprehension in traffic scenes with MLLMs, their performance is still far below the human level. This is because traffic data-text pairs contain diverse modalities, such as 3D point clouds, panoramic 2D imagery, high-definition map data, and traffic regulations. These elements significantly differ from conventional domain contexts and question-answer pairs, highlighting the unique challenges of deploying MLLMs in that autonomous driving context. The datasets mentioned above are limited in terms of scale and quality, which hinders efforts to fully address these emerging challenges.

6. LLVM-AD Workshop Summary

The 1st LLVM-AD is held together with WACV 2024 on Jan 8th, 2024 in Waikoloa, Hawaii. We seek to bring together academia and industry professionals in a collaborative exploration of applying MLLMs to autonomous driving. Through a half-day in-person event, the workshop
will showcase regular and demo paper presentations and invited talks from famous researchers in academia and industry. Additionally, LLVM-AD will launch two open-source real-world traffic language understanding datasets, catalyzing practical advancements. The workshop will host two challenges based on this dataset to assess the capabilities of language and computer vision models in addressing autonomous driving challenges.

6.1. Multimodal Large Language Models for Autonomous Driving Challenges

MAPLM Dataset. Tencent’s THMA HD Map AI labeling system is utilized to create descriptive paragraphs from HD map labels, offering nuanced portrayals of traffic scenes [163]. Participants worked with various data modalities, including 2D camera images, 3D point clouds, and Bird’s Eye View (BEV) images, enhancing our understanding of the environment. This innovative initiative explores the intersection of computer vision, AI-driven mapping, and natural language processing, highlighting the transformative potential of Tencent’s THMA technology in reshaping our understanding and navigation of our surroundings.

UCU Dataset. The primary objective of this challenge is the development of algorithms that are proficient in understanding drivers’ commands and instructions represented as natural language input. These commands and instructions could encompass a diverse array of command types, ranging from safety-oriented instructions such as “engage the emergency brakes” or “adjust headlight brightness”, to driving operational instructions such as “shift to park mode” or “set the cruise control to 70 mph”, and comfort-related requests such as “turn up the AC” or “turn off seat heating”. The scope of commands can even be extended to vehicle-specific instructions like “open sunroof” or “enable ego mode”.

6.2. Workshop Summary

Nine papers were accepted in the inaugural Workshop on Large Language and Vision Models for Autonomous Driving (LLVM-AD) at the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). They cover topics on MLLMs for autonomous driving focusing on integrating LLMs into user-vehicle interaction, motion planning, and vehicle control. Several papers explored the novel use of LLMs to enhance human-like interaction and decision-making in autonomous vehicles. For example, “Drive as You Speak” [31] and “Drive Like a Human” [41] presented frameworks where LLMs interpret and reason in complex driving scenarios, mimicking human behavior. “Human-Centric Autonomous Systems With LLMs” [195] emphasized the importance of user-centric design, utilizing LLMs to interpret user commands. This approach represents a significant shift towards more intuitive and human-centric autonomous systems.

In addition to LLM integration, the workshop featured methodologies in vision-based systems and data processing. “A Safer Vision-based Autonomous Planning System for Quadrotor UAVs” [208] and “VLAAD” [132] demonstrated advanced approaches to object detection and trajectory planning, enhancing the safety and efficiency of UAVs and autonomous vehicles.

Optimizing technical processes was also a significant focus. For instance, “A Game of Bundle Adjustment” [10] introduced a novel approach to improving 3D reconstruction efficiency, while “Latency Driven Spatially Sparse Optimization” [201] and “LIP-Loc” [156] explored advancements in CNN optimization and cross-modal localization, respectively. These contributions represent notable progress towards more efficient and accurate computational models in autonomous systems.

Furthermore, the workshop presented innovative approaches to data handling and evaluation. For example, NuScenes-MQA [53] introduced a dataset annotation technique for autonomous driving. Collectively, these papers...
illustrate a significant stride in integrating language models and advanced technologies into autonomous systems, paving the way for more intuitive, efficient, and human-centric autonomous vehicles.

7. Discussion

New Datasets for Multimodal Large Language Models in Autonomous Driving. Despite the success of LLMs in language understanding, applying them to autonomous driving presents a unique challenge. This is due to the necessity for these models to integrate and interpret inputs from diverse modalities, such as panoramic images, 3D point clouds, and HD map annotations. The current limitations in data scale and quality mean that existing datasets struggle to address all these challenges comprehensively. Furthermore, almost all multimodal LLMs like GPT-4V [127] have been pre-trained on a wealth of open-source datasets including traffic and driving scenes, the visual-language datasets annotated from nuScenes may not provide a robust benchmark for visual-language understanding in driving scene. Consequently, there is an urgent need for new, large-scale datasets that encompass a wide range of traffic and driving scenarios, including numerous corner cases, to effectively test and enhance these models in autonomous driving applications.

Hardware Support for Large Language Models in Autonomous Driving. In the use case of LLMs as the planner for autonomous driving, the perception reasoning for the LLMs and the subsequent control decision should be generated in real-time with low latency in order to meet safety requirements for autonomous driving. The number of (Floating-point operations per second)FLOPs of the LLMs has a positive correlation with the latency as well as the power consumption, which should be of consideration if LLMs are hosted in the vehicle. For LLMs deployed remotely, the bandwidth of perception information and control decision transfer will be a great challenge.

Another use case for LLMs in autonomous driving is a navigation planner [143, 151]. Unlike driving planners, the tolerance of response time for the LLMs is much higher, and the number of queries for navigation planners is far less in general. Consequently, the hardware performance demand is easier to meet, and even moving the host to remote servers is a reasonable proposal.

The user-vehicle interaction could also be a use case of LLMs in autonomous driving [31]. LLMs could interpret drivers’ intentions into control commands given to the vehicle. For intentions unrelated to driving, e.g. entertainment control, the high latency of the response from LLMs could be accepted. However, if the intentions involve taking over autonomous driving, then the hardware requirements would be similar to the counterpart of using LLMs as an autonomous driving planner where LLMs are expected to respond with low latency.

LLMs in the applications of autonomous driving could potentially be compressed, which reduces the computation power requirements and the latency and lowers the HW limitation. However, the current effort in this field is still undeveloped.

Using Large Language Models for Understanding HD Maps. HD maps play a crucial role in autonomous vehicle technology, as they provide essential information about the physical environment in which the vehicle operates. The semantic map layer from the HD map is of utmost importance as it captures the meaning and context of the physical surroundings. To effectively encode this valuable information into the LLMs-powered next-generation autonomous driving, it is important to find a way to represent and comprehend the details of the environment in the language space.

Inspired by transformer-based language models, Tesla proposes a special language that they developed for encoding lanes and their connectivities. In this language of lanes, the words and tokens represent the lane positions in 3D space. The ordering of the tokens and predicted modifiers in the tokens encode the connectivity relationships between these lanes. Producing a lane graph from the model output sentence requires less post-processing than parsing a segmentation mask or a heatmap [20]. Pre-trained models (PTMs) have become a fundamental backbone for downstream tasks in natural language processing and computer vision. Baidu Maps has developed a system called ERNIE-GeoL, which has already been deployed in production. This system applies generic PTMs to geo-related tasks at Baidu Maps since April 2021, resulting in significant performance improvements for various downstream tasks [58].

Tencent has developed an HD Map AI system called THMA which is an innovative end-to-end, AI-based, active learning HD map labeling system capable of producing and labeling HD maps with a scale of hundreds of thousands of kilometers [163] [207]. To promote the development of this field, they proposed the MAPLM [86] dataset containing over 2 million frames of panoramic 2D images, 3D LiDAR point cloud, and context-based HD map annotations, and a new question-answer benchmark MAPLM-QA.

User-Vehicle Interaction with Large Language Models. Non-verbal language interpretation is also an important aspect to consider for user-autonomy teaming. Driver distraction poses a critical road safety challenge, including all activities such as smartphone use, eating, and interacting with passengers that divert attention from driving. According to the National Highway Traffic Safety Administration (NHTSA), distractions were a factor in 8.1% of the 38,824 vehicle-related fatalities in the U.S. in 2020 [160].
issue becomes more pressing as semi-autonomous driving systems, particularly SAE Level 3 systems, gain prominence, requiring drivers to be ready to take control when prompted [147].

To detect and mitigate driver distraction, driver action recognition strategies are commonly employed. These strategies involve continuous monitoring using sensors like RGB and infrared cameras, coupled with deep learning algorithms to identify and classify driver actions. Significant advancements have been made in this field [12, 107, 108, 128, 190].

Assessing the driver’s cognitive state is also crucial, as it greatly indicates distraction levels. Physiological monitoring, such as through EEG signals, can provide insights into a driver’s cognitive state [177, 179], but the intrusiveness of such sensors and their impact on regular driving patterns must be taken into account. Besides, behavior monitoring works such as through facial analysis, gaze, human pose, and motion [17, 18, 87–89, 91] can also be used to analyze driver’s driving status. Furthermore, current datasets on driver action recognition often lack mental state annotations required to train models in recognizing these states from sensory data, highlighting the need for semi-supervised learning methods to address this relatively unexplored challenge [106].

**Personalized Autonomous Driving.** The integration of LLMs into autonomous vehicles marks a paradigm shift characterized by continuous learning and personalized engagement. LLMs can continuously learn from new data and interactions, adapting to changing driving patterns, user preferences, and evolving road conditions. This adaptability results in a refined and increasingly adept performance over time. Moreover, LLMs have the capability to be precisely fine-tuned or in-context learned to match individual driver preferences, furnishing personalized assistance that significantly improves the driving experience. This personalized approach enriches the driving experience, providing assistance that not only offers information but also aligns closely with the distinct requirements and subtleties of each driver.

Recent studies [30, 31] have indicated the potential for LLMs to enhance real-time personalization in driving simulations, demonstrating their capacity to adapt driving behaviors in response to spoken commands. As the LLM-based personalized assistance in autonomous driving is not well-developed, there are numerous opportunities for further research. Most recent studies focus on utilizing LLMs in the simulation environment instead of real vehicles. Integrating LLMs into actual vehicles is an exciting area of potential, moving beyond simulations to affect real-world driving experiences. Additionally, future investigations could also explore the development of LLM-driven virtual assistants that align with drivers’ individual preferences, the employment of LLMs for the enhancement of safety features like fatigue detection, the application of these models in predictive vehicle maintenance, and the personalization of routing to align with drivers’ unique inclinations. Furthermore, LLMs have the potential for personalizing in-vehicle entertainment, learning from drivers’ behaviors to improve the driving experience.

**Trustworthy and Safety for Autonomous Driving.** Another crucial takeaway is enhancing transparency and trust. When the vehicle makes a complex decision, such as overtaking another vehicle on a high-speed, two-lane highway, passengers and drivers might naturally have questions or concerns. In these instances, the LLM doesn’t just execute the task but also articulates the reasoning behind each step of the decision-making process. By providing real-time, detailed explanations in understandable language, the LLM demystifies the vehicle’s actions and underlying logic. This not only satisfies the innate human curiosity about how autonomous systems work but also builds a higher level of trust between the vehicle and its occupants.

Moreover, the advantage of “zero-shotting” was particularly evident during the complex overtaking maneuver on a high-speed Indiana highway. Despite the LLM not having encountered this specific set of circumstances before—varying speeds, distances, and even driver alertness—it was able to use its generalized training to safely and efficiently generate a trajectory for the overtaking action. With some uncertainty estimation techniques [112, 191, 198], this can ensure that even in dynamic or edge case scenarios, the system can make sound judgments while keeping the user informed, therefore building confidence in autonomous technology.

To sum up, LLMs demonstrate their potential to revolutionize autonomous driving by enhancing safety, transparency, and user experience. Tasked with complex commands like overtaking, the LLM considered real-time data from multiple vehicle modules to make informed decisions, clearly articulating these to the driver. The model also leveraged its zero-shot learning capabilities to adapt to new scenarios, providing personalized, real-time feedback. Overall, the LLM proved effective in building user trust and improving decision-making in autonomous vehicles, emphasizing its utility in future automotive technologies.

**8. Conclusion**

In this survey, we explored the pattern of integrating multimodal large language models (MLLMs) into the next generation of autonomous driving systems. Our study began with an overview of the development of both MLLMs and autonomous driving, which have traditionally been considered distinct fields but are now increasingly intercon-
nected. Then, we conducted an extensive literature review on the specific algorithms and applications of multimodal language models for autonomous driving and then focused on the current state of research and benchmarking datasets that apply MLLMs to autonomous driving. A significant highlight of our study was the synthesis of key insights and findings from the first LLVM-AD workshop such as proposing new datasets and improving current MLLMs algorithms on autonomous driving. Finally, we engaged in a forward-looking discussion on vital research themes and the promising potential for enhancing MLLMs in autonomous driving. We discussed both challenges and opportunities that lie ahead, aiming to show the pathway for further exploration. In general, this paper serves as a valuable resource for researchers in the autonomous driving area. It offers a comprehensive understanding of the significant role and vast potential that MLLMs hold in revolutionizing the landscape of autonomous transportation. We hope this paper could facilitate research in integrating MLLMs with autonomous driving in the future.

Acknowledgments

We would like to express our gratitude for the support received from the Purdue University Digital Twin Lab (https://purduedigitaltwin.github.io/), Tencent T Lab, and PediaMed AI (http://pediamedai.github.io/) for their contribution to this survey paper.

References


[15] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan,
Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In CVPR, pages 11621–11631, 2020. 9, 10


[21] Long Chen, Yuchen Li, Chao Huang, Bai Li, Yang Xing, Daxin Tian, Li Li, Zhongxu Hu, Xiaoliang Na, Zixuan Li, Siyu Teng, Chen Lv, Jinhun Wang, Dongpu Cao, Nanning Zheng, and Fei-Yue Wang. Milestones of surveys. IEEE Transactions on Intelligent Vehicles, 8(2):1046–1056, 2023. 4


[35] Vikrant Dewangan, Tushar Choudhary, Shivam Chandhok, Shubham Priyadarshan, Anushka Jain, Arun K. Singh,
[71] Xuhui Kang, Wenqian Ye, and Yen-Ling Kuo. Imagined subgoals for hierarchical goal-conditioned policies. In CoRL 2023 Workshop on Learning Effective Abstractions for Planning (LEAP), 2023. 6


[178] Teng Wang, Jinrui Zhang, Junjie Fei, Hao Zheng, Yunlong Tang, Zhe Li, Mingqi Gao, and Shanshan Zhao. Caption anything: Interactive image description with diverse multimodal controls, 2023. 6


[183] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai,
