A Survey on Multimodal Large Language Models for Autonomous Driving

Can Cui[1]*, Yunsheng Ma[1]*, Xu Cao[3,6]*, Wenqian Ye[4,6]*, Yang Zhou[5], Kaizhao Liang[7], Jintai Chen[3], Juanwu Lu[1], Zichong Yang[1], Kuei-Da Liao[8], Tianren Gao[7], Erlong Li[2], Kun Tang[2], Zhipeng Cao[2], Tong Zhou[2], Ao Liu[2], Xinrui Yan[2], Shuqi Mei[2], Jianguo Cao[6]†, Ziran Wang[1]†, Chao Zheng[2]†

[1] Purdue University, West Lafayette, IN, USA 47907
[2] Tencent T Lab, Beijing, China
[3] University of Illinois Urbana-Champaign, Champaign, IL, USA 61801
[4] University of Virginia, Charlottesville, VA, USA 22903
[5] New York University, New York, NY, USA 11201
[6] PediaMed AI, Shenzhen, China
[7] SambaNova Systems, Inc, Palo Alto, CA, USA 94303
[8] 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.

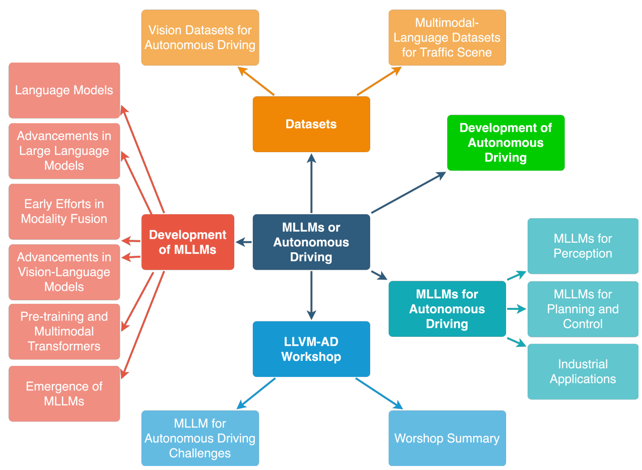


Figure 1. This survey paper focuses on the use of Multimodal Large Language Models (MLLMs) in the advancement of autonomous driving. The organization of the paper will delve into various aspects related to this topic.

1. Introduction

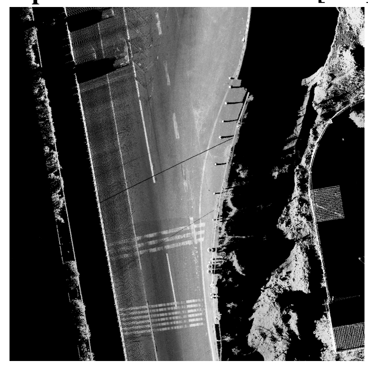
Large Language Models (LLMs) have gained significant attention recently, showing remarkable potential in emulating human-like intelligence. The advancements have sparked a surge of enthusiasm in Multimodal Large Language Models (MLLMs) [199], which merge the sophisticated reasoning prowess of LLMs with 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 LLMs can deal with easy tasks within the robotics domain including basic logical, geometrical, and mathematical rea-

*Equal contribution.
†Correspondence authors.

Question / Prompt: 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.

Input LiDAR Point Cloud: [163]



Question / Prompt: 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.

Input Driving Front View:

Question / Prompt: Make a right lane change.

GPT-4 Code Genration:

```
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)
```

Simulation [92]:

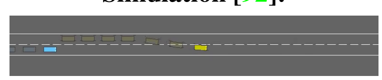


Figure 2. Exploring GPT-4V [127] to understand driving scenes and make driving actions. Our findings reveal that while GPT-4V adeptly identifies scene components such as objects, it falls short in recognizing critical traffic elements like lane information. This underscores the significant challenges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

soning, to complex tasks such as aerial navigation, manipulation, and embodied agents. However, the integration of LLMs into the realm of transportation and autonomous vehicles is at a pioneering stage. Merging linguistic communication with multimodal sensory inputs like panoramic images, LiDAR point clouds, and driving actions could revolutionize the foundation models that govern current autonomous driving systems.

Recently, the emergence of more capable foundation models has made SAE L3 driving automation practicable [28]. However, the integration of multimodal LLMs in autonomous driving has not followed these advancements, and one natural question is, do LLM-based models like GPT-4, PaLM-2, and LLaMA-2 have the potential to enhance autonomous driving? Figure 2 shows us a very good example. It is undeniable that integrating LLMs into the autonomous vehicle industry can bring a significant paradigm shift in vehicle intelligence, decision-making, and passenger interaction [30, 31], offering a more user-centric, adaptable, and trustworthy future of transportation.

In the context of autonomous driving, LLMs will offer a transformative impact across crucial modules: perception, motion planning, and motion control [180]. In terms of perception, LLMs can harness external APIs to access 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

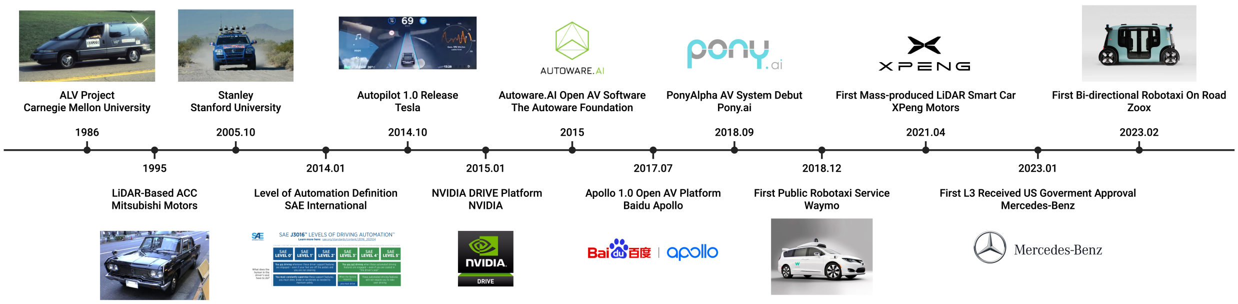


Figure 3. The figure outlines the chronological development of autonomous driving technology. It begins with representative early exploration and advancements like the ALV Project by Carnegie Mellon University [70, 172], Mitsubishi Debonair the first to offer LiDAR-based ADAS system [120], and winner of 2005 DARPA Grand Challenge Stanley by Stanford University [166]. It then showcases 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 are 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 played a pivotal role in enhancing the control strategies of autonomous vehicles, refining motion planning, and decision-making processes to adapt to dynamic and uncertain driving conditions [16, 75, 78, 93],. Moreover, 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].

Inspired by current limitations, part of the research on autonomous driving is now focusing on addressing the safety of autonomous systems and enhancing the safety of autonomous systems [200]. As Deep Neural Networks are often considered black boxes, trustworthy AI aims at making the system more reliable, explainable, and verifiable. For example, generating adversarial safety-critical scenarios for training autonomous driving systems such that the system is more capable of handling cases with low probability [1, 36]. Another way to improve the overall safety is through vehicle-to-infrastructure and vehicle-to-vehicle communication. With information from nearby instances, the system will have improved robustness and can receive early warnings [99, 122]. Meanwhile, as Large Language Models show their powerful reasoning and scene-understanding capability, research is being conducted to utilize them to improve the safety and overall performance of the autonomous driving system.

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].

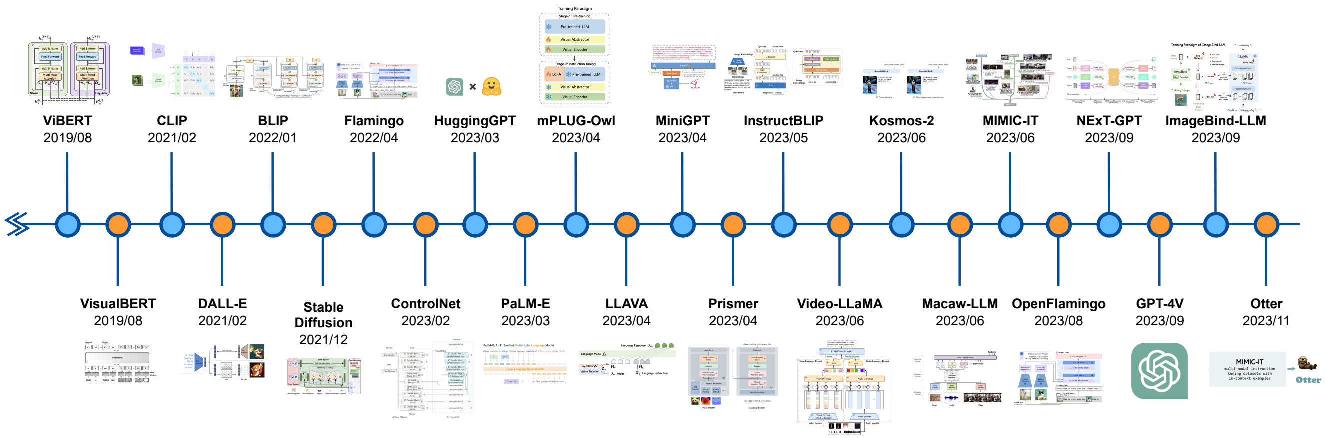


Figure 4. A timeline of recent advancements in Multimodal Large Language Models (MLLMs).

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 learning [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 tex-

Model	Year	Backbone	Task	Modality	Learning	Input	Output
Driving with LLMs [22]	2023	LLaMA	Perception Control	Vector Language	FT	Vector Query	Response Actions
Talk2BEV [35]	2023	Flan5XXL Vicuna-13b	Perception Planning	Vision Language	ICL	Image Query	Response
GAIA-1 [57]	2023	-	Planning	Vision Language	PT	Video Prompt	Video
LMaZP [60]	2022	GPT-3 Codex	Planning	Language	ICL	Text	Plan
Dilu [185]	2023	GPT-3.5 GPT-4	Planning Control	Language	ICL	Text	Action
DaYS [31]	2023	GPT-4	Planning	Language	ICL	Text	Code
RRR [30]	2023	GPT-4	Planning Control	Language	ICL	Text	Action
DlaH [42]	2023	GPT-3.5	Planning Control	Language	ICL	Text	Action
GPT-Driver [110]	2023	GPT-3.5	Planning	Vision Language	ICL	Text	Trajectory
SurrealDriver [68]	2023	GPT-4	Planning Control	Language	ICL	Text	Text Action
LanguageMPC [150]	2023	GPT-3.5	Planning	Language	ICL	Text	Action
DriveGPT4 [193]	2023	Llama 2	Planning Control	Vision Language	ICL	Image Text Action	Text Action

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.

tual descriptions and various modalities, providing a richer source of supervision.

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. HiLM-D [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 integrating 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 [37].

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 multimodal 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.

Dataset	Year	RGB	LiDAR	Text	Map
KITTI [44]	2012	15K	15K	✗	✗
nuScenes [15]	2019	1.4M	400K	✓	✓
Argo1 [19]	2019	107K	22K	✗	✓
Waymo Open [161]	2019	1M	200K	✗	✓
Argo2 [186]	2021	5.4M	6M	✗	✓
V2V4Real [192]	2023	40K	20K	✗	✓

Table 2. Comparison of representative autonomous driving datasets.

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

Dataset	Year	QA	Caption	Scenario	Text	Modality		
						Image.	Point Cloud.	Map Info.
BDD-X [77]	2018	✗	✓	7K	26K	✓	✗	✗
Talk2Car [33]	2019	✗	✓	34K	12K	✓	✗	✗
DRAMA [109]	2023	✗	✓	18K	102K	✓	✗	✗
nuScenes-QA [138]	2023	✓	✗	340K	460K	✓	✓	✗
NuPrompt [189]	2023	✗	✓	34K	35K	✓	✓	✗
DriveLM [29]	2023	✓	✓	34K	375K	✓	✗	✗
MAPLM [86, 163]	2023	✓	✓	2M	16M	✓	✓	✓

Table 3. Multimodal-Language datasets for self-driving can be split to two types: (1) Added additional texts for existing nuScenes [15] dataset such as Talk2Car [33], nuScenes-QA [138], DriveLM [29], and NuPrompt [189]; (2) independent collected datasets such as BDD-X [77], and DRAMA [109].

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 [63] 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]. This 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].

Personlized 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 personalization 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 interconnected. 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 contributions to this survey paper.

References

[1] Yasasa Abeysirigoonawardena, Florian Shkurti, and Gregory Dudek. Generating adversarial driving scenarios in high-fidelity simulators. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8271–8277, 2019. 4

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, 2022. arXiv:2204.01691. 5, 8

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. In *NeurIPS*. arXiv, 2022. arXiv:2204.14198. 6, 7

[4] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Zeeuw, Hervé Jégou, and Andrew Zisserman. Self-supervised multimodal versatile networks. In *Advances in Neural Information Processing Systems*, pages 18718–18730, 2020. 6

[5] Jacob Andreas, Dan Klein, and Sergey Levine. Learning with Latent Language, 2017. arXiv:1711.00482. 8

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 6

[7] Relja Arandjelović and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 609–617, 2017. 6

[8] Baidu. Baidu Apollo Project Repository. `https://github.com/ApolloAuto/apollo`. Accessed: 2023-11-11. 4

[9] Yehoshua Bar-Hillel. The present status of automatic translation of languages. *Advances in computers*, 1:91–163, 1960. 4

[10] Amir Belder, Refael Vivanti, and Ayellet Tal. A game of bundle adjustment-learning efficient convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8428–8437, 2023. 10

[11] Klaus Bengler, Klaus Dietmayer, Berthold Farber, Markus Maurer, Christoph Stiller, and Hermann Winner. Three decades of driver assistance systems: Review and future perspectives. *IEEE Intelligent Transportation Systems Magazine*, 6(4):6–22, 2014. 4

[12] Mahdi Biparva, David Fernández-Llorca, Rubén Izquierdo Gonzalo, and John K. Tsotsos. Video Action Recognition for Lane-Change Classification and Prediction of Surrounding Vehicles. *IEEE Transactions on Intelligent Vehicles*, 7(3):569–578, Sept. 2022. 12

[13] Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–480, 1992. 4

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *NeurIPS*, volume 33, pages 1877–1901, 2020. 4, 5, 6

[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

[16] Peide Cai, Hengli Wang, Yuxiang Sun, and Ming Liu. DQ-GAT: Towards Safe and Efficient Autonomous Driving With Deep Q-Learning and Graph Attention Networks. *IEEE Transactions on Intelligent Transportation Systems*, 23(11):21102–21112, 2022. 4

[17] Xu Cao, Xiaoye Li, Liya Ma, Yi Huang, Xuan Feng, Zening Chen, Hongwu Zeng, and Jianguo Cao. Aggpose: Deep aggregation vision transformer for infant pose estimation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5045–5051, 7 2022. 12

[18] Xu Cao, Wenqian Ye, Elena Sizikova, Xue Bai, Megan Coffee, Hongwu Zeng, and Jianguo Cao. Vitasd: Robust vision transformer baselines for autism spectrum disorder facial diagnosis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 12

[19] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3D Tracking and Forecasting With Rich Maps. In *CVPR*, pages 8748–8757, 2019. 9

[20] Kevin Chen. Analyzing tesla ai day 2022. [EB/OL]. https://kevinchen.co/blog/tesla-ai-day-2022/. 11

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

[22] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving, 2023. arXiv:2310.01957. 7, 8

[23] Pranav Singh Chib and Pravendra Singh. Recent Advancements in End-to-End Autonomous Driving using Deep Learning: A Survey, 2023. arXiv:2307.04370. 8

[24] Noam Chomsky. *Aspects of the Theory of Syntax*. MIT press, 2014. 4

[25] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling Language Modeling with Pathways, Oct. 2022. arXiv:2204.02311. 4, 5

[26] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling Instruction-Finetuned Language Models, 2022. arXiv:2210.11416. 6, 7

[27] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 4

[28] On-Road Automated Driving (ORAD) Committee. *Taxonomy and Definitions for Terms Related to On-Road Motor Vehicle Automated Driving Systems*, 2014. 2, 3, 4

[29] DriveLM Contributors. Drivelm: Drive on language. https://github.com/OpenDriveLab/DriveLM, 2023. 9, 10

[30] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Receive, Reason, and React: Drive as You Say with Large Language Models in Autonomous Vehicles, 2023. arXiv:2310.08034. 2, 3, 5, 7, 12

[31] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. 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 (WACV) Workshops*, 2024. 2, 7, 8, 10, 11, 12

[32] Jin Cui, Lin Shen Liew, Giedre Sabaliauskaite, and Fengjun Zhou. A review on safety failures, security attacks, and available countermeasures for autonomous vehicles. *Ad Hoc Networks*, 90:101823, 2019. 4

[33] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. Talk2Car: Taking Control of Your Self-Driving Car. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2088–2098, 2019. 8, 9, 10

[34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4, 6

[35] Vikrant Dewangan, Tushar Choudhary, Shivam Chandhok, Shubham Priyadarshan, Anushka Jain, Arun K. Singh,

Siddharth Srivastava, Krishna Murthy Jatavallabhula, and K. Madhava Krishna. Talk2BEV: Language-enhanced Bird's-eye View Maps for Autonomous Driving, 2023. arXiv:2310.02251. 7

[36] Wenhao Ding, Baiming Chen, Bo Li, Kim Ji Eun, and Ding Zhao. Multimodal safety-critical scenarios generation for decision-making algorithms evaluation. *IEEE Robotics and Automation Letters*, 6(2):1551–1558, April 2021. 4

[37] Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. HiLM-D: Towards High-Resolution Understanding in Multimodal Large Language Models for Autonomous Driving, 2023. arXiv:2309.05186. 7, 9

[38] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 6

[39] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An Embodied Multimodal Language Model, Mar. 2023. arXiv:2303.03378. 3, 5

[40] Shai Fine, Yoram Singer, and Naftali Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine learning*, 32:41–62, 1998. 4

[41] Daocheng Fu, Xin Li, Licheng Wen, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2024. 5, 8, 10

[42] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive Like a Human: Rethinking Autonomous Driving with Large Language Models. *arXiv preprint arXiv:2307.07162*, 2023. 7

[43] Jiaxin Ge, Hongyin Luo, Siyuan Qian, Yulu Gan, Jie Fu, and Shanghang Zhang. Chain of thought prompt tuning in vision language models, 2023. 6

[44] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, June 2012. 9

[45] Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, and Anurag Arnab. Audiovisual masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16144–16154, 2023. 6

[46] Prashant Gohel, Priyanka Singh, and Manoranjan Mohanty. Explainable AI: current status and future directions, July 2021. arXiv:2107.07045 [cs]. 8

[47] Prasoon Goyal, Scott Niekum, and Raymond J. Mooney. PixL2R: Guiding Reinforcement Learning Using Natural Language by Mapping Pixels to Rewards, Nov. 2020. arXiv:2007.15543 [cs, stat]. 8

[48] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020. 3

[49] Mercedes-Benz Group. Certification for SAE Level 3 system for U.S. market. `https://group.mercedes-benz.com/innovation/product-innovation/autonomous-driving/drive-pilot-nevada.html`, Jan. 2023. Accessed: 2023-11-11. 4

[50] Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, John D. Co-Reyes, Rishabh Agarwal, Rebecca Roelofs, Yao Lu, Nico Montali, Paul Mougin, Zoey Yang, Brandyn White, Aleksandra Faust, Rowan McAllister, Dragomir Anguelov, and Benjamin Sapp. Waymax: An Accelerated, Data-Driven Simulator for Large-Scale Autonomous Driving Research. In *NeurIPS*. arXiv, 2023. arXiv:2310.08710 [cs]. 7

[51] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023. 6

[52] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962, 2023. 6

[53] Xinyu Han, Jianhui Lai, Kuiyuan Yang, Xiaojuan Li, Yujun Zhang, Dahua Lin, and Hao Zeng. Occuseg: Occupancy-aware 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2918–2927, 2020. 6

[54] Vaishnavi Himakunthala, Andy Ouyang, Daniel Rose, Ryan He, Alex Mei, Yujie Lu, Chinmay Sonar, Michael Saxon, and William Yang Wang. Let's think frame by frame: Evaluating video chain of thought with video infilling and prediction, 2023. 6

[55] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4, 5

[56] Anatol W Holt and WJ Turanski. Man-to-machine communication and automatic code translation. In *Papers presented at the May 3-5, 1960, western joint IRE-AIEE-ACM computer conference*, pages 329–339, 1960. 4

[57] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. GAIA-1: A Generative World Model for Autonomous Driving, Sept. 2023. arXiv:2309.17080 [cs]. 7, 9

[58] Jizhou Huang, Haifeng Wang, Yibo Sun, Yunsheng Shi, Zhengjie Huang, An Zhuo, and Shikun Feng. Ernie-geol: A geography-and-language pre-trained model and its applications in baidu maps. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3029–3039, 2022. 11

[59] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that lis-

ten. *Advances in Neural Information Processing Systems*, 35:28708–28720, 2022. 6

[60] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents, Mar. 2022. arXiv:2201.07207 [cs]. 5, 7, 8

[61] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. VoxPoser: Composable 3D Value Maps for Robotic Manipulation with Language Models, 2023. arXiv:2307.05973 [cs]. 5, 8

[62] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. Inner Monologue: Embodied Reasoning through Planning with Language Models, 2022. arXiv:2207.05608. 8

[63] Yuichi Inoue, Yuki Yada, Kotaro Tanahashi, and Yu Yamaguchi. Nuscenes-mqa: Integrated evaluation of captions and qa for autonomous driving datasets using markup annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2024. 10

[64] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. Opt-iml: Scaling language model instruction meta learning through the lens of generalization, 2023. 6

[65] Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer vision for autonomous vehicles: Problems, datasets and state of the art. *Foundations and Trends® in Computer Graphics and Vision*, 12(1–3):1–308, 2020. 3

[66] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. BC-Z: Zero-Shot Task Generalization with Robotic Imitation Learning, Feb. 2022. arXiv:2202.02005 [cs]. 8

[67] Yiding Jiang, Shixiang Gu, Kevin Murphy, and Chelsea Finn. Language as an Abstraction for Hierarchical Deep Reinforcement Learning, Nov. 2019. arXiv:1906.07343 [cs, stat]. 8

[68] Ye Jin, Xiaoxi Shen, Huiling Peng, Xiaoan Liu, Jingli Qin, Jiayang Li, Jintao Xie, Peizhong Gao, Guyue Zhou, and Jiangtao Gong. SurrealDriver: Designing Generative Driver Agent Simulation Framework in Urban Contexts based on Large Language Model, Sept. 2023. arXiv:2309.13193 [cs]. 5, 7, 8

[69] Aishwarya Kamath, Peter Anderson, Su Wang, Jing Yu Koh, Alexander Ku, Austin Waters, Yinfei Yang, Jason Baldridge, and Zarana Parekh. A New Path: Scaling Vision-and-Language Navigation with Synthetic Instructions and Imitation Learning, Apr. 2023. arXiv:2210.03112 [cs]. 8

[70] Takeo Kanade, Chuck Thorpe, and William Whittaker. Autonomous land vehicle project at CMU. In *Proceedings of the 1986 ACM fourteenth annual conference on Computer science - CSC '86*, pages 71–80, Cincinnati, Ohio, United States, 1986. ACM Press. 3, 4

[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

[72] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 6

[73] Shinpei Kato, Eijiro Takeuchi, Yoshio Ishiguro, Yoshiki Ninomiya, Kazuya Takeda, and Tsuyoshi Hamada. An open approach to autonomous vehicles. *IEEE Micro*, 35(6):60–68, Nov 2015. 4

[74] Shinpei Kato, Shota Tokunaga, Yuya Maruyama, Seiya Maeda, Manato Hirabayashi, Yuki Kitsukawa, Abraham Monrroy, Tomohito Ando, Yusuke Fujii, and Takuya Azumi. Autoware on board: Enabling autonomous vehicles with embedded systems. In *2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems (ICCPS)*, pages 287–296, April 2018. 4

[75] Alex Kendall, Jeffrey Hawke, David Janz, Przemyslaw Mazur, Daniele Reda, John-Mark Allen, Vinh-Dieu Lam, Alex Bewley, and Amar Shah. Learning to drive in a day. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8248–8254, 2019. 4

[76] Ali Keysan, Andreas Look, Eitan Kosman, Gonca Gürsun, Jörg Wagner, Yu Yao, and Barbara Rakitsch. Can you text what is happening? integrating pre-trained language encoders into trajectory prediction models for autonomous driving, 2023. 8, 9

[77] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *Proceedings of the European conference on computer vision (ECCV)*, pages 563–578, 2018. 8, 9, 10

[78] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2022. 4

[79] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 7

[80] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. Large-scale text-to-image generation models for visual artists' creative works. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 919–933, 2023. 6

[81] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large Language Models are Zero-Shot Reasoners. In *NeurIPS*, volume 35, pages 22199–22213, 2022. 8

[82] Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. Toward understanding natural language directions. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 259–266, Mar. 2010. ISSN: 2167-2148. 8

[83] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 5

[84] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding, Oct. 2020. arXiv:2010.07954 [cs]. 8

[85] Sampo Kuutti, Richard Bowden, Yaochu Jin, Phil Barber, and Saber Fallah. A survey of deep learning applications to autonomous vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, 22(2):712–733, 2021. 3

[86] Tencent T Lab. Maplm: A real-world large-scale vision-language dataset for map and traffic scene understanding. [EB/OL]. `https://github.com/LLVM-AD/MAPLM/`. 10, 11

[87] Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global-local correlation for egocentric gaze estimation. *arXiv preprint arXiv:2208.04464*, 2022. 12

[88] Bolin Lai, Miao Liu, Fiona Ryan, and James M Rehg. In the eye of transformer: Global–local correlation for egocentric gaze estimation and beyond. *International Journal of Computer Vision*, pages 1–18, 2023. 12

[89] Bolin Lai, Fiona Ryan, Wenqi Jia, Miao Liu, and James M Rehg. Listen to look into the future: Audio-visual egocentric gaze anticipation. *arXiv preprint arXiv:2305.03907*, 2023. 12

[90] Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds, 2019.

[91] Sangmin Lee, Hak Gu Kim, Dae Hwi Choi, Hyung-Il Kim, and Yong Man Ro. Video prediction recalling long-term motion context via memory alignment learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3054–3063, 2021. 12

[92] Edouard Leurent. An Environment for Autonomous Driving Decision-Making, 2018. 2

[93] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J. Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, Michael Sokolsky, Ganymed Stanek, David Stavens, Alex Teichman, Moritz Werling, and Sebastian Thrun. Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 163–168, 2011. 4

[94] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 6

[95] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 6

[96] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as Policies: Language Model Programs for Embodied Control. In *ICRA*, 2023. 5, 8

[97] Kaizhao Liang, Xu Cao, Kuei-Da Liao, Tianren Gao, Wenqian Ye, Zhengyu Chen, Jianguo Cao, Tejas Nama, and Jimeng Sun. Pie: Simulating disease progression via progressive image editing, 2023. 6

[98] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 6, 7

[99] Weijie Liu, Shintaro Muramatsu, and Yoshiyuki Okubo. Cooperation of v2i/p2i communication and roadside radar perception for the safety of vulnerable road users. In *2018 16th International Conference on Intelligent Transportation Systems Telecommunications (ITST)*, pages 1–7, 2018. 4

[100] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[101] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*, 2023. 6

[102] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity, 2022. 6

[103] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022. 7

[104] Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel. A Survey of Reinforcement Learning Informed by Natural Language, June 2019. arXiv:1906.03926 [cs, stat]. 8

[105] Corey Lynch and Pierre Sermanet. Language Conditioned Imitation Learning over Unstructured Data, July 2021. arXiv:2005.07648 [cs]. 8

[106] Yunsheng Ma and Ziran Wang. ViT-DD: Multi-Task Vision Transformer for Semi-Supervised Driver Distraction Detection. *IEEE Intelligent Vehicles Symposium*, 2023. 12

[107] Yunsheng Ma, Wenqian Ye, Xu Cao, Amr Abdelraouf, Kyungtae Han, Rohit Gupta, and Ziran Wang. CEM-Former: Learning to Predict Driver Intentions from In-Cabin and External Cameras via Spatial-Temporal Transformers. *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, May 2023. 12

[108] Yunsheng Ma, Liangqi Yuan, Amr Abdelraouf, Kyungtae Han, Rohit Gupta, Zihao Li, and Ziran Wang. M2DAR: Multi-View Multi-Scale Driver Action Recognition with Vision Transformer. *Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2023. 12

[109] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1043–1052, 2023. 9, 10

[110] Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. GPT-Driver: Learning to Drive with GPT, Oct. 2023. arXiv:2310.01415 [cs]. 3, 7, 8

[111] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014. 6

[112] Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, and Jie Zhou. Prevent the language model from being overconfident in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3456–3468, 2021. 12

[113] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 4

[114] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?, Oct. 2022. arXiv:2202.12837 [cs]. 8

[115] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European Conference on Computer Vision*, pages 728–755. Springer, 2022. 7

[116] Dipendra Misra, John Langford, and Yoav Artzi. Mapping Instructions and Visual Observations to Actions with Reinforcement Learning, July 2017. arXiv:1704.08795 [cs]. 8

[117] Nico Montali, John Lambert, Paul Mougin, Alex Kuefler, Nick Rhinehart, Michelle Li, Cole Gulino, Tristan Emrich, Zoey Yang, Shimon Whiteson, Brandyn White, and Dragomir Anguelov. The Waymo Open Sim Agents Challenge, July 2023. arXiv:2305.12032 [cs]. 7

[118] Tesla Motors. Model S Owner's Manual [Online]. https://www.tesla.com/sites/default/files/model_s_owners_manual_touchscreen_7.1_das_ap_north_america_r20160112_en_us.pdf. Accessed: 2023-11-11. 4

[119] Youssef Mroueh, Tom Sercu, and Vaibhava Goel. Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2130–2134. IEEE, 2015. 6

[120] Marc Nabhan. *Models and algorithms for the exploration of the space of scenarios: toward the validation of the autonomous vehicle*. PhD thesis, Université Paris-Saclay, 2020. 4

[121] Suraj Nair, Eric Mitchell, Kevin Chen, Brian Ichter, Silvio Savarese, and Chelsea Finn. Learning Language-Conditioned Robot Behavior from Offline Data and Crowd-Sourced Annotation, Oct. 2021. arXiv:2109.01115 [cs]. 8

[122] Ying Ni, Shihan Wang, Liuyan Xin, Yiwei Meng, Juyuan Yin, and Jian Sun. A v2x-based approach for avoiding potential blind-zone collisions between right-turning vehicles and pedestrians at intersections. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–6, 2020. 4

[123] Anthony G Oettinger. *Automatic language translation: Lexical and technical aspects, with particular reference to Russian*. Harvard University Press, 1960. 4

[124] Mohammad Omama, Pranav Inani, Pranjal Paul, Sarat Chandra Yellapragada, Krishna Murthy Jatavallabhula, Sandeep Chinchali, and Madhava Krishna. Alt-pilot: Autonomous navigation with language augmented topometric maps, 2023. 3

[125] OpenAI. ChatGPT, 2023. https://openai.com/blog/chatgpt. 6, 7

[126] OpenAI. GPT-4 Technical Report, Mar. 2023. 5, 7

[127] OpenAI. Gpt-4v(ision) system card. https://openai.com/research/gpt-4v-system-card, 2023. 2, 6, 11

[128] Chaojie Ou and Fakhri Karray. Enhancing Driver Distraction Recognition Using Generative Adversarial Networks. *IEEE Transactions on Intelligent Vehicles*, 5(3):385–396, Sept. 2020. 12

[129] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, volume 35, pages 27730–27744, 2022. 5, 6

[130] Jishnu Jaykumar P, Kamalesh Palanisamy, Yu-Wei Chao, Xinya Du, and Yu Xiang. Proto-CLIP: Vision-Language Prototypical Network for Few-Shot Learning, July 2023. arXiv:2307.03073 [cs]. 7

[131] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Spatial as deep: Spatial cnn for traffic scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 9

[132] SungYeon Park, MinJae Lee, JiHyuk Kang, Hahyeon Choi, Yoonah Park, Juhwan Cho, Adam Lee, and Dong-Kyu Kim. Vlaad: Vision and language assistant for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2024. 10

[133] Joern Ploennigs and Markus Berger. Ai art in architecture. *AI in Civil Engineering*, 2(1):8, 2023. 6

[134] Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. *Advances in neural information processing systems*, 1, 1988. 3

[135] Pony.ai. Pony.ai. https://pony.ai/story?lang=en. Accessed: 2023-11-11. 4

[136] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation, 2017. 3

[137] Charles R Qi, Hao Su, Matthias Niessner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 5648–5656, 2016. 6

[138] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. *arXiv preprint arXiv:2305.14836*, 2023. 9, 10

[139] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 6, 7

[140] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training, 2018. 6

[141] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 4

[142] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020. 4, 6

[143] Abhinav Rajvanshi, Karan Sikka, Xiao Lin, Bhoram Lee, Han-Pang Chiu, and Alvaro Velasquez. Saynav: Grounding large language models for dynamic planning to navigation in new environments. *arXiv preprint arXiv:2309.04077*, 2023. 11

[144] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 6

[145] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 6

[146] Daniel Rose, Vaishnavi Himakunthala, Andy Ouyang, Ryan He, Alex Mei, Yujie Lu, Michael Saxon, Chinmay Sonar, Diba Mirza, and William Yang Wang. Visual chain of thought: Bridging logical gaps with multimodal infillings, 2023. 6

[147] SAE On-Road Automated Vehicle Standards Committee and others. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles. *SAE International: Warrendale, PA, USA*, 2018. 12

[148] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997. 4, 5

[149] Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S Refaat, Rami Al-Rfou, and Benjamin Sapp. Motionlm: Multi-agent motion forecasting as language modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8579–8590, 2023. 9

[150] Hao Sha, Yao Mu, Yuxuan Jiang, Li Chen, Chenfeng Xu, Ping Luo, Shengbo Eben Li, Masayoshi Tomizuka, Wei Zhan, and Mingyu Ding. Languagempc: Large language models as decision makers for autonomous driving. *arXiv preprint arXiv:2310.03026*, 2023. 3, 7, 8

[151] Dhruv Shah, Michael Equi, Blazej Osinski, Fei Xia, Brian Ichter, and Sergey Levine. Navigation with large language models: Semantic guesswork as a heuristic for planning. *arXiv preprint arXiv:2310.10103*, 2023. 11

[152] Dhruv Shah, Błażej Osiński, brian ichter, and Sergey Levine. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 492–504. PMLR, 14–18 Dec 2023. 8

[153] Pratyusha Sharma, Balakumar Sundaralingam, Valts Blukis, Chris Paxton, Tucker Hermans, Antonio Torralba, Jacob Andreas, and Dieter Fox. Correcting Robot Plans with Natural Language Feedback. In *Robotics: Science and Systems XVIII*. Robotics: Science and Systems Foundation, June 2022. 8

[154] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580*, 2023. 6

[155] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. CLIPort: What and Where Pathways for Robotic Manipulation, 2021. arXiv:2109.12098. 8

[156] Sai Shubodh, Mohammad Omama, Husain Zaidi, Udit Singh Parihar, and Madhava Krishna. Lip-loc: Lidar image pretraining for cross-modal localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2024. 10

[157] Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11523–11530, 2023. 8

[158] John Slaney and Sylvie Thiébaux. Blocks world revisited. *Artificial Intelligence*, 125(1-2):119–153, 2001. 5

[159] N. N. Sriram, Tirth Maniar, Jayaganesh Kalyanasundaram, Vineet Gandhi, Brojeshwar Bhowmick, and K Madhava Krishna. Talk to the vehicle: Language conditioned autonomous navigation of self driving cars. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5284–5290, 2019. 3

[160] Timothy Stewart. Overview of Motor Vehicle Crashes in 2020. Technical report, National Highway Traffic Safety Administration, 2022. 11

[161] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Sheng Zhao, Shuyang Cheng, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in Per-

ception for Autonomous Driving: Waymo Open Dataset. In *CVPR*, 2020. arXiv:1912.04838 [cs, stat]. 9

[162] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. 4

[163] Kun Tang, Xu Cao, Zhipeng Cao, Tong Zhou, Erlong Li, Ao Liu, Shengtao Zou, Chang Liu, Shuqi Mei, Elena Sizikova, et al. Thma: Tencent hd map ai system for creating hd map annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15585–15593, 2023. 2, 9, 10, 11

[164] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. Robots That Use Language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):25–55, May 2020. _eprint: https://doi.org/10.1146/annurev-control-101119-071628. 8

[165] The Vicuna Team. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, 2023. 7

[166] Sebastian Thrun, Mike Montemerlo, Hendrik Dahlkamp, David Stavens, Andrei Aron, James Diebel, Philip Fong, John Gale, Morgan Halpenny, Gabriel Hoffmann, Kenny Lau, Celia Oakley, Mark Palatucci, Vaughan Pratt, Pascal Stang, Sven Strohband, Cedric Dupont, Lars-Erik Jendrossek, Christian Koelen, Charles Markey, Carlo Rummel, Joe van Niekerk, Eric Jensen, Philippe Alessandrini, Gary Bradski, Bob Davies, Scott Ettinger, Adrian Kaehler, Ara Nefian, and Pamela Mahoney. Stanley: The robot that won the darpa grand challenge. *Journal of Field Robotics*, 23(9):661–692, 2006. 4

[167] Catherine Tong, Jinchen Ge, and Nicholas D. Lane. Zero-Shot Learning for IMU-Based Activity Recognition Using Video Embeddings. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(4):180:1–180:23, Dec. 2022. 8

[168] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, Feb. 2023. arXiv:2302.13971 [cs]. 7

[169] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. arXiv:2307.09288 [cs]. 5, 7

[170] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal Few-Shot Learning with Frozen Language Models. In *NeurIPS*, volume 34, pages 200–212, 2021. 7

[171] TuSimple. Tusimple benchmark. `https://github.com/TuSimple/tusimple-benchmark`. Accessed: 2022-03-04. 9

[172] Carnegie Mellon University. The Robot Hall of Fame. `http://www.robothalloffame.org/inductees/08inductees/navlab.html`. Accessed: 2023-11-11. 4

[173] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4, 5

[174] Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. ChatGPT for Robotics: Design Principles and Model Abilities, July 2023. arXiv:2306.17582 [cs]. 1, 8

[175] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 6

[176] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An Open-Ended Embodied Agent with Large Language Models, May 2023. arXiv:2305.16291 [cs]. 5, 8

[177] Shouyi Wang, Yiqi Zhang, Changxu Wu, Felix Darvas, and Wanpracha Art Chaovalitwongse. Online Prediction of Driver Distraction Based on Brain Activity Patterns. *IEEE Transactions on Intelligent Transportation Systems*, 16(1):136–150, Feb. 2015. 12

[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

[179] Yu-Kai Wang, Tzyy-Ping Jung, and Chin-Teng Lin. EEG-Based Attention Tracking During Distracted Driving. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 23(6):1085–1094, Nov. 2015. 12

[180] Ziran Wang, Yougang Bian, Steven E. Shladover, Guoyuan Wu, Shengbo Eben Li, and Matthew J. Barth. A survey on cooperative longitudinal motion control of multiple connected and automated vehicles. *IEEE Intelligent Transportation Systems Magazine*, 12(1):4–24, 2020. 2

[181] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 6

[182] Wayve. LINGO-1: Exploring Natural Language for Autonomous Driving, Sept. 2023. 9

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

and Quoc V. Le. Finetuned language models are zero-shot learners, 2022. 6

[184] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*, 2022. 5, 8

[185] Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. Dilu: A knowledge-driven approach to autonomous driving with large language models. *arXiv preprint arXiv:2309.16292*, 2023. 7, 8

[186] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next Generation Datasets for Self-Driving Perception and Forecasting. In *NeurIPS*, 2021. arXiv:2301.00493 [cs]. 9

[187] Terry Winograd. Procedures as a Representation for Data in a Computer Program for Understanding Natural Language. *AI Technical Reports*, 1971. 8

[188] Chenfei Wu, Shengming Yin, Weizhen Qi, Xiaodong Wang, Zecheng Tang, and Nan Duan. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*, 2023. 6

[189] Dongming Wu, Wencheng Han, Tiancai Wang, Yingfei Liu, Xiangyu Zhang, and Jianbing Shen. Language prompt for autonomous driving. *arXiv preprint arXiv:2309.04379*, 2023. 9, 10

[190] Yang Xing, Chen Lv, Huaji Wang, Dongpu Cao, Efstathios Velenis, and Fei-Yue Wang. Driver Activity Recognition for Intelligent Vehicles: A Deep Learning Approach. *IEEE Transactions on Vehicular Technology*, 68(6):5379–5390, June 2019. 12

[191] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023. 12

[192] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, Hongkai Yu, Bolei Zhou, and Jiaqi Ma. V2V4Real: A Real-world Large-scale Dataset for Vehicle-to-Vehicle Cooperative Perception. In *CVPR*, 2023. 9

[193] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee. K. Wong, Zhenguo Li, and Hengshuang Zhao. DriveGPT4: Interpretable End-to-end Autonomous Driving via Large Language Model, Oct. 2023. arXiv:2310.01412. 7, 8

[194] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning Interactive Real-World Simulators, Oct. 2023. arXiv:2310.06114 [cs]. 7

[195] Yi Yang, Qingwen Zheng, Ci Li, Daniel L.S. Marta, Nazre Batool, and John Folkesson. Human-centric autonomous systems with llms for user command reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2024. 10

[196] Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action, 2023. 6

[197] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023. 6

[198] Wenqian Ye, Yunsheng Ma, Xu Cao, and Kun Tang. Mitigating Transformer Overconfidence via Lipschitz Regularization. *Conference on Uncertainty in Artificial Intelligence*, 2023. 12

[199] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 1, 3

[200] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE Access*, 8:58443–58469, 2020. 4

[201] Giorgos Zampokas, Christos-Savvas Bouganis, and Dimitrios Tzovaras. Latency driven spatially sparse optimization for multi-branch cnns for semantic segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2024. 10

[202] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language, May 2022. arXiv:2204.00598 [cs]. 8

[203] Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. In *EMNLP*. arXiv, 2023. arXiv:2306.02858 [cs, eess]. 7

[204] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 6

[205] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Hongsheng Li, Yu Qiao, and Peng Gao. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners, 2023. 6

[206] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models, 2023. 6

[207] Chao Zheng, Xu Cao, Kun Tang, Zhipeng Cao, Elena Sizikova, Tong Zhou, Erlong Li, Ao Liu, Shengtao Zou, Xinrui Yan, and Shuqi Mei. High-definition map automatic annotation system based on active learning. *AI Magazine*, 2023. 11

[208] Jiageng Zhong, Ming Li, Yinliang Chen, Zihang Wei, Fan Yang, and Haoran Shen. Safer vision-based autonomous planning system for quadrotor uavs with dynamic obstacle trajectory prediction. In *Proceedings of the IEEE/CVF*

Winter Conference on Applications of Computer Vision (WACV) Workshops, 2024. 10

[209] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. 6

[210] Linjie Zhu, Jieyu Xu, Yi Yang, and Alexander G Hauptmann. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8743–8752, 2020. 6

[211] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyao Zeng, Shanghang Zhang, and Peng Gao. Pointclip v2: Adapting clip for powerful 3d open-world learning. *arXiv preprint arXiv:2211.11682*, 2022. 6