LLM4Drive: A Survey of Large Language Models for Autonomous Driving

Zhenjie Yang[1*] **Xiaosong Jia**[1*] **Hongyang Li**[1,2]

Junchi Yan[1†]

[*] Equal contributions, order decided by a coin toss. [†] Correspondence author

[1] Dept. of CSE & School of AI & MoE Key Lab of AI, Shanghai Jiao Tong University

[2] OpenDriveLab

Abstract

Autonomous driving technology, a catalyst for revolutionizing transportation and urban mobility, has the tend to transition from rule-based systems to data-driven strategies. Traditional module-based systems are constrained by cumulative errors among cascaded modules and inflexible pre-set rules. In contrast, end-to-end autonomous driving systems have the potential to avoid error accumulation due to their fully data-driven training process, although they often lack transparency due to their "black box" nature, complicating the validation and traceability of decisions. Recently, large language models (LLMs) have demonstrated abilities including understanding context, logical reasoning, and generating answers. A natural thought is to utilize these abilities to empower autonomous driving. By combining LLM with foundation vision models, it could open the door to open-world understanding, reasoning, and few-shot learning, which current autonomous driving systems are lacking. In this paper, we systematically review the research line about *(Vision) Large Language Models for Autonomous Driving ((V)LLM4Drive)*. This study evaluates the current state of technological advancements, distinctly outlining the principal challenges and prospective directions for the field. For the convenience of researchers in academia and industry, we provide real-time updates on the latest advances in the field as well as relevant open-source resources via the designated link: `https://github.com/Thinklab-SJTU/Awesome-LLM4AD`.

1 Introduction

Autonomous driving is rapidly reshaping our understanding of transportation, heralding a new era of technological revolution. This transformation means not only the future of transportation but also a fundamental shift across various industries. In conventional autonomous driving systems, algorithms typically adopt the modular design [1, 2, 3], with separate components responsible for critical tasks such as perception [4, 5], prediction [6, 7, 8, 9], and planning [10, 11, 12, 13]. Specifically, the perception component handles object detection [4, 5], tracking [14], and sophisticated semantic segmentation tasks [15]. The prediction component analyzes the external environment [16] and estimates the future states of the surrounding agents [17, 18]. The planning component, often reliant on rule-based decision algorithms [10], determines the optimal and safest route to a predetermined destination. While the module-based approach provides reliability and enhanced security in a variety of scenarios, it also presents challenges. The decoupled design between system components may lead

This work was supported by NSFC (92370201, 62222607) and Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102.

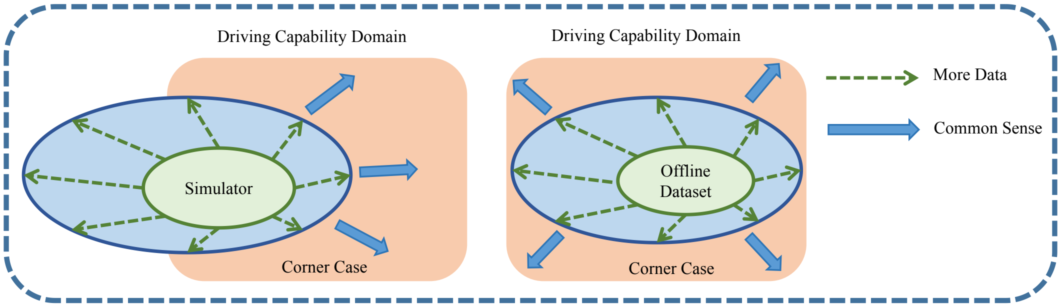

Figure 1: The limitation of current autonomous driving paradigm (green arrow) and where LLMs can potentially enhance autonomous driving ability (blue arrow).

to key information loss during transitions and potentially redundant computation as well. Additionally, errors may accumulate within the system due to inconsistencies in optimization objectives among the modules, affecting the vehicle's overall decision-making performance [19].

Rule-based decision systems, with their inherent limitations and scalability issues, are gradually giving way to data-driven methods. End-to-end autonomous driving solutions are increasingly becoming a consensus in the field [20, 21, 22, 23, 24, 25]. By eliminating integration errors between multiple modules and reducing redundant computations, the end-to-end system enhances the expression of visual [26] and sensory information while ensuring greater efficiency. However, this approach also introduces the “black box" problem, meaning a lack of transparency in the decision-making process, complicating interpretation and validation.

Simultaneously, the explainability of autonomous driving has become an important research focus [27]. Although smaller language models (like early versions of BERT [28] and GPT [29]) employed in massive data collection from driving scenarios help address this issue, they often lack sufficient generalization capabilities to perform optimally. Recently, large language models [30, 31] have demonstrated remarkable abilities in understanding context, generating answers, and handling complex tasks. They are also now integrated with multimodal models [32, 33, 34, 35, 36]. This integration achieves a unified feature space mapping for images, text, videos, point clouds, etc. Such consolidation significantly enhances the system's generalization capabilities and equips them with the capacity to quickly adapt to new scenarios in a zero-shot or few-shot manner.

In this context, developing an interpretable and efficient end-to-end autonomous driving system has become a research hotspot [19]. Large language models, with their extensive knowledge base and exceptional generalization, could facilitate easier learning of complex driving behaviors. By leveraging the visual-language model (VLM)'s robust and comprehensive capabilities of open-world understanding and in-context learning [37, 38, 33, 34], it becomes possible to address the long-tail problem for perception networks, assist in decision-making, and provide intuitive explanations for these decisions.

This paper aims to provide a comprehensive overview of this rapidly emerging research field, analyze its basic principles, methods, and implementation processes, and introduce in detail regarding the application of LLMs for autonomous driving. Finally, we discuss related challenges and future research directions.

2 Motivation of LLM4AD

In today's technological landscape, large language models such as GPT-4 and GPT-4V [30, 39] are drawing attention with their superior contextual understanding and in-context learning capabilities. Their enriched common sense knowledge has facilitated significant advancements in many downstream tasks. We ask the question: *how do these large models assist in the domain of autonomous driving, especially in playing a critical role in the decision-making process?*

In Fig. 1, we give an intuitive demonstration of the limitation of current autonomous driving paradigm and where LLMs can potentially enhance autonomous driving ability. We summarize two primary aspects of driving skills. The orange circle represents the ideal level of driving competence, akin to that possessed by an experienced human driver. There are two main methods to acquire such proficiency:

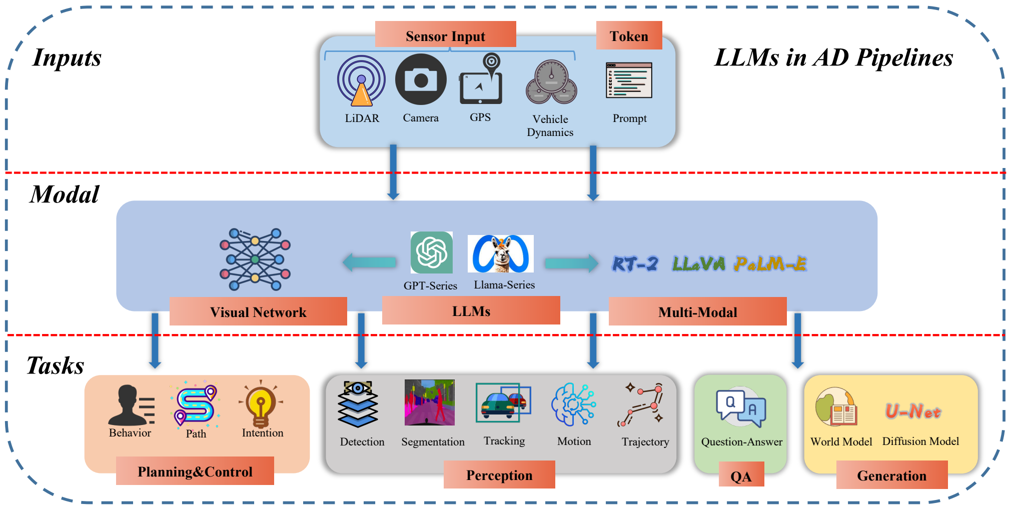

Figure 2: LLMs in Autonomous Driving Pipelines.

one, through learning-based techniques within simulated environments; and two, by learning from offline data through similar methodologies. It's important to note that due to discrepancies between simulations and the real world, these two domains are not fully the same, i.e. sim2real gap [40]. Concurrently, offline data serves as a subset of real-world data since it's collected directly from actual surroundings. However, it is difficult to fully cover the distribution as well due to the notorious long-tailed nature [41] of autonomous driving tasks.

The final goal of autonomous driving is to elevate driving abilities from a basic green stage to a more advanced blue level through extensive data collection and deep learning. However, the high cost associated with data gathering and annotation, along with the inherent differences between simulated and real-world environments, mean there's still a gap before reaching the expert level of driving skills. In this scenario, if we can effectively utilize the innate common sense embedded within large language models, we might gradually narrow this gap. Intuitively, by adopting this approach, we could progressively enhance the capabilities of autonomous driving systems, bringing them closer to, or potentially reaching, the ideal expert level of driving proficiency. Through such technological integration and innovation, we anticipate significant improvements in the overall performance and safety of autonomous driving.

The application of large language models in the field of autonomous driving indeed covers a wide range of task types, combining depth and breadth with revolutionary potential. LLMs in autonomous driving pipelines is shown in the Fig. 2.

3 Application of LLM4AD

In the following sections, we divide existing works based on the perspective of applying LLMs: planning, perception, question answering, and generation. The corresponding taxonomy tree is shown in Fig. 3.

3.1 Planning & Control

Large language models (LLMs) have achieved great success with their open-world cognitive and reasoning capabilities [42, 43, 29, 44, 30]. These capabilities could provide a transparent explanation of the autonomous driving decision-making process, significantly enhancing system reliability and user trust in the technology [45, 46, 47, 27, 48, 49, 50, 51]. Within this domain, based on whether tuning the LLM, related research can be categorized into two main types: fine-tuning pre-trained models and prompt engineering.

3.1.1 Fine-tuning pre-trained models

In the application of fine-tuning pre-trained models, MTD-GPT [52] translates multi-task decision-making problems into sequence modeling problems. Through training on a mixed multi-task dataset, it addresses various decision-making tasks at unsignaled intersections. Although this approach outperforms the performance of single-task decision-making RL models, the used scenes are limited to unsignaled intersections, which might be enough to demonstrate the complexity of the real world application. Driving with LLMs [36] designs an architecture that fuses vectorized inputs into LLMs with a two-stage pretraining and fine-tuning method. Due to the limitation of vectorized representations, their method are only tested in the simulation. DriveGPT4 [35] presents a multimodal LLM based on Valley [53] and develops a visual instruction tuning dataset for interpretable autonomous driving. Besides predicting a vehicle's basic control signals, it also responds in real-time, explaining why the action is taken. It outperforms baseline models in a variety of QA tasks while the experiments about planning is simple. GPT-Driver [54] transforms the motion planning task into a language modeling problem. It exceeds the UniAD[25] in the L2 metric. Nevertheless, since it uses past speed and acceleration information, there is concern about unfair comparison with UniAD. Additionally, L2 only reflects the fitting degree of the driving route and might not reflect the driving performance [11]. Agent-Driver [55] leverages LLMs common sense and robust reasoning capabilities to improve the capabilities of planning by designing a tool library, a cognitive memory, and a reasoning engine. This paradigm achieves better results on the nuScenes dataset. Meanwhile, shortening the inference time is also an urgent problem. DriveLM [56] uses a trajectory tokenizer to process ego-trajectory signals to texts, making them belong to the same domain space. Such a tokenizer can be applied to any general vision language models. Moreover, they utilize a graph-structure inference with multiple QA pairs in logical order, thus improving the final planning performance. [57] adapts LLMs as a vehicle "Co-Pilot" of driving, which can accomplish specific driving tasks with human intention satisfied based on the information provided. It lacks verification in complex interaction scenarios. LMDrive [58] designs a multi-modal framework to predict the control signal and whether the given instruction is completed. It adopts Resnet [59] as the vision encoder which has not been through an image-text alignment pretraining. In addition, it introduces a benchmark LangAuto which includes approximately 64K instruction-following data clips in CARLA. The LangAuto benchmark tests the system's ability to handle complex instructions and challenging driving scenario. DriveMLM [60] adopts a multi-modal LLM(Multi-view image, Point cloud, and prompt) to generate high-level decision commands and uses Apollo as a planner to get the control signal. Moreover, the training data generated by experts and uses GPT-3.5 to increase data diversity. It achieves 76.1 driving score on the CARLA Town05 Long, which reaches the level of classic end-to-end autonomous driving. KoMA [61] is a knowledge-driven multi-agent framework in which each agent is powered by large language models. These agents analyze and infer the intentions of surrounding vehicles to enhance decision-making. AsyncDriver [62] is an asynchronous LLM-enhanced framework where the inference frequency of LLM is controllable and can be decoupled from the real-time planner. It has good closed-loop evaluation performance in challenging scenarios of nuPlan. PlanAgent [63] extracts bird's-eye view (BEV) representation and generates a text description input based on the lane map through an environment transformation module. It uses a reasoning engine module to perform a hierarchical thinking chain to guide driving scene understanding, motion command generation, and planning code writing. AGENTSCODRIVER [64] is an LLM-powered framework for multi-vehicle collaborative driving with lifelong learning, enabling communication and collaboration among driving agents in complex traffic scenarios, featuring a reasoning engine, cognitive memory, reinforcement reflection, and a communication module. DriveVLM [65] leverages Vision-Language Models to enhance scene understanding and planning capabilities for autonomous driving, while DriveVLM-Dual synergizes these advancements with traditional 3D perception and planning approaches to effectively address spatial reasoning and computational challenges, demonstrating superior performance in complex and dynamic driving scenarios. RAG-Driver [66], a Multi-Modal Large Language Model with Retrieval-augmented In-context Learning, provides explainable and generalizable end-to-end driving by producing numerical control signals, along with explanations and justifications for driving actions, and demonstrates impressive zero-shot generalization to unseen environments without additional training. LLaDA [67] designs a training-free mechanism to assist human drivers and adapt autonomous driving policies to new environments. VLP [68] is a Vision-Language-Planning model intended to enhance autonomous driving systems (ADS) by incorporating two novel components: ALP and SLP. ALP (Agent-wise Learning Paradigm) aligns the generated bird's-eye-view (BEV) with the true BEV map, improving self-driving BEV reasoning. SLP (Self-Driving-Car-Centric

Learning Paradigm) aligns the ego vehicle's query features with its textual planning features, enhancing self-driving decision-making. DME-Driver [69] enhances decision logic explainability and environmental perception accuracy by using a vision language model for decision-making and a planning-oriented perception model for generating precise control signals, effectively translating human-like driving logic into actionable commands, and achieving high-precision planning accuracy through the comprehensive HBD dataset.

3.1.2 prompt engineering

In the prompt engineering perspective, some methods tried to tap into the deep reasoning potential of the LLMs through clever prompt design. DiLu [70] designs a framework of LLMs as agents to solve closed-loop driving tasks. This method introduces a memory module to record experience, to leverage LLMs to facilitate reasoning and reflection processes. DiLu exhibits strong generalization capabilities compared with SOTA RL-based methods. However, the reasoning and reflection processes require multiple rounds of question-answering, and its inference time cannot be ignored. Similarly, Receive Reason and React [71] and Drive as You Speak [72] integrate the language and reasoning capabilities of LLMs into autonomous vehicles. In addition to memory and reflection processes, these methods introduce additional raw sensor information such as camera, GNSS, lidar, and radar. However, the inference speed is unsolved as well. Furthermore, SurrealDriver [73] divides the memory module into short-term memory, long-term guidelines, and safety criteria. Meanwhile, it interviews 24 drivers and uses their detailed descriptions of driving behaviors as chain-of-thought prompts to develop a 'coach agent' module. However, there is a lack of comparison with traditional algorithms to prove that large language models indeed bring performance improvements. LanguageMPC [74] also designs a chain-of-thought framework for LLMs in driving scenarios and it integrates with low-level controllers by guided parameter matrix adaptation. Although its performance exceeds MPC and RL-based methods in the simplified simulator environments, it lacks validation in complex environments. TrafficGPT [75] is a fusion of ChatGPT and traffic foundation models which can tackle complex traffic-related problems and provide insightful suggestions. It leverages multimodal data as a data source, offering comprehensive support for various traffic-related tasks. Talk2BEV [76] introduces a large vision-language model (LVLM) interface for bird's-eye view (BEV) maps in autonomous driving contexts. It does not require any training or fine-tuning, only relying on pre-trained image-language models. In addition, it presents a benchmark for evaluating subsequent work in LVLMs for AD applications. Talk2Drive [72] utilizes human verbal commands and makes autonomous driving decisions based on contextual information to meet humanly personalized preferences for safety, efficiency, and comfort. AccidentGPT [77] integrates multi-vehicle collaborative perception to improve environmental understanding and collision avoidance, offering advanced safety features like proactive remote safety warnings and blind spot alerts. It also supports traffic police and management agencies by providing real-time intelligent analysis of traffic safety factors.

3.2 Perception

Large language models have demonstrated their unique value and strong capabilities in "perception" tasks [78, 79, 80, 81, 82]. Especially in environments where data is relatively scarce, these models can rely on their few-shot learning characteristics to achieve fast and accurate learning and reasoning [83, 84]. This learning ability is of significance in the perception stage of the autonomous driving system, which greatly improves the system's adaptability and generalization capabilities in changing and complex driving environments. PromptTrack [85] fuses cross-modal features in a prompt reasoning branch to predict 3D objects. It uses language prompts as semantic cues and combines LLMs with 3D detection tasks and tracking tasks. Although it achieves better performance compared to other methods, the advantages of LLMs do not directly affect the tracking task. Rather, the tracking task serves as a query to assist LLMs in performing 3D detection tasks. HiLM-D [86] incorporates high-resolution information into multimodal large language models for the Risk Object Localization and Intention and Suggestion Prediction (ROLISP) task. It combines LLMs with 2D detection tasks and obtains better performance in detection tasks and QA tasks compared to other multi-modal large models such as Video-LLaMa [87], eP-ALM [88]. It is worth noting to point out one potential limitation of the dataset: each video contains only one risk object, which might not capture the complexity of real-world scenarios. [89] integrates pre-trained language models as text-based input encoders for the autonomous driving trajectory prediction task. Joint encoders(image and text) over both modalities perform better than using a single encoder in isolation. While the joint model significantly improves the baseline, its performance has not reached the state-of-the-art level yet [90, 91]. LC-LLM [92] is designed for lane change prediction, leveraging LLM capabilities to understand complex scenarios, enhancing prediction performance, and providing explainable predictions by generating explanations for lane change intentions and trajectories. AIDE [93] introduces a paradigm for an automatic data engine, incorporating automatic data querying and labeling using VLM, and continual learning with pseudo labels. It introduces a new benchmark to evaluate such automated data engines for AV perception that allows combined insights across multiple paradigms of open vocabulary detection, semi-supervised, and continual learning. Context-aware Motion Prediction [94] designs and conducts prompt engineering to enable GPT4-V to comprehend complex traffic scenarios. It combines the context information outputted by GPT4-V with MTR [95] to enhance motion prediction.

Dataset	Task	Size	Annotator	Description
BDD-X [96]	Planning VQA	77 hours, 6970 videos, 8.4M frames, 26228 captions	Human	Ego-vehicle actions description and explanation.
HAD [97]	Planning Perception	30 hours, 5744 videos 22366 captions	Human	Joint action description for goal-oriented advice and attention description for stimulus-driven advice.
Talk2Car [45]	Planning Perception	15 hours, 850 videos of 20s each 30k frames, 11959 captions	Human	Object referral dataset that contains commands written in natural language for self-driving cars.
DriveLM [56]	Perception Prediction Planning VQA	In Carla, 18k frames and 3.7M QA pairs;In nuScenes, 4.8k frames and 450k QA pairs	Human Rule-Based	P3 with reasoning logic; Connect the QA pairs in a graph-style structure; Use “What if"-style questions.
DRAMA [48]	VQA	91 hours, 17785 videos, 77639 question, 102830 answering, 17066 captions	Human	Joint risk localization with visual reasoning of driving risks in a free-form language description.
Rank2Tell [98]	Perception VQA	several hours, 118 videos of 20s each	Human	Joint important object identification, important object localization ranking, and reasoning.
NuPrompt [85]	Perception	15 hours, 35367 prompts for 3D objects	Huamn GPT3.5	Object-centric language prompt set for perception tasks.
NuScenes-QA [99]	VQA	15 hours, Train(24149 scences, 459941 QA pairs), Test(6019 scences, 83337 QA pairs)	Rule-Based	Leverage 3D annotations(object category, position, orientation, relationships information) and designed question templates to construct QA pairs.
Reason2Drive [100]	Perception Prediction VQA	600K video-text pairs	Human GPT-4	Composed of nuScenes, Waymo and ONCE, with driving instructions.
LingoQA [100]	VQA	419.9k QA pairs, 28k scenarios	Rule-Based GPT-3.5/4 Software Human	Contains reasoning pairs in addition to object presence, description, and localisation.
NuInstruct [101]	Perception Prediction VQA	91k QA pairs, 17 subtasks	GPT-4 Human	Integrates multi-view information, requiring responses from multiple perspectives, with balanced view distribution for perception tasks.
OpenDV-2K [102]	Perception Prediction VQA	2059 hours of videos paired with texts(1747 hours from YouTube and 312 hours from public datasets).	BLIP-2	A large-scale multimodal dataset for autonomous driving, to support the training of a generalized video prediction model.

Table 1: Description of different datasets regarding LLM4AD.

3.3 Question Answering

Question-Answering is an important task that has a wide range of applications in intelligent transportation, assisted driving, and autonomous vehicles [103, 104]. It mainly reflects through different question and answer paradigms, including traditional QA mechanism [105] and more detailed visual QA methods [35]. [105] constructs the domain knowledge ontology by "chatting" with ChatGPT. It develops a web-based assistant to enable manual supervision and early intervention at runtime and it guarantees the quality of fully automated distillation results. This question-and-answer system enhances the interactivity of the vehicle, transforms the traditional one-way human-machine interface into an interactive communication experience, and might be able to cultivate the user's sense of participation and control. These sophisticated models [105, 35], equipped with the ability to parse, understand, and generate human-like responses, are pivotal in real-time information processing and provision. They design comprehensive questions related to the scene, including but not limited to vehicle states, navigation assistance, and understanding of traffic situations. [106] provides a human-centered perspective and gives several key insights through different prompt designs to enable LLMs to achieve AD system requirements within the cabin. Dolphins [107] enhances reasoning capabilities through the innovative Grounded Chain of Thought (GCoT) process and specifically adapts to the driving domain by building driving-specific command data and command adjustments. LingoQA [108] develops a QA benchmark and datasets, details are in 3.5 and 4. EM-VLM4AD [109] is an efficient, lightweight, multi-frame vision language model for Visual Question Answering in autonomous driving, and it only requires much less memory and floating point operations than DriveLM [56]. [110] proposes a prototype of a pipeline of prompts and LLMs that receives an item definition and outputs solutions in the form of safety requirements. Hybrid Reasoning [111] uses Large Language Models (LLMs) with inputs from image-detected objects and sensor data, including parameters like object distance, car speed, direction, and location, to generate precise brake and speed control values based on weather conditions. TransGPT [112] is a novel large language model for the transportation domain that comes in two variants—TransGPT-SM for single-modal data and TransGPT-MM for multi-modal data—designed to enhance traffic analysis and modeling by generating synthetic traffic scenarios, explaining traffic phenomena, answering traffic-related questions, offering recommendations, and creating comprehensive traffic reports.

3.4 Generation

In the realm of "generation" task, large language models leverage their advanced knowledge-base and generative capabilities to create realistic driving videos or intricate driving scenarios under specific environmental factors [113, 114]. This approach offers revolutionary solutions to the challenges of data collection and labeling for autonomous driving, also constructing a safe and easily controllable setting for testing and validating the decision boundaries of autonomous driving systems. Moreover, by simulating a variety of driving situations and emergency conditions, the generated content becomes a crucial resource for refining and enriching the emergency response strategies of autonomous driving systems.

The common generative models include the Variational Auto-Encoder(VAE) [115], Generative Adversarial Network(GAN) [116], Normalizing Flow(Flow)[117], and Denoising Diffusion Probabilistic Model(Diffusion)[118]. With diffusion models have recently achieved great success in text-to-image [119, 120, 121], some research has begun to study using diffusion models to generate autonomous driving images or videos. DriveDreamer [122] is a world model derived from real-world driving scenarios. It uses text, initial image, HDmap, and 3Dbox as input, then generates high-quality driving videos and reasonable driving policies. Similarly, Driving Diffusion [123] adopts a 3D layout as a control signal to generate realistic multi-view videos. GAIA-1 [124] leverages video, text, and action inputs to generate traffic scenarios, environmental elements, and potential risks. In these methods, text encoder both adopt CLIP [78] which has a better alignment between image and text. In addition to generating autonomous driving videos, traffic scenes can also be generated. CTG++ [125] is a scene-level diffusion model that can generate realistic and controllable traffic. It leverages LLMs for translating a user query into a differentiable loss function and use a diffusion model to transform the loss function into realistic, query compliant trajectories. MagicDrive [126] generates highly realistic images, exploiting geometric information from 3D annotations by independently encoding road maps, object boxes, and camera parameters for precise, geometry-guided synthesis. This approach effectively solves the challenge of multi-camera view consistency. Although it achieves better performance in terms of generation fidelity compared to BEVGen [127] and BEVControl [128], it also faces huge challenges in some complex scenes, such as night views and unseen weather conditions. ADriver-I [129] combines Multimodal Large Language Models(MLLM) and Video Diffusion Model(VDM) to predict the control signal of current frame and the future frames. It shows impressive performance on nuScenes and their private datasets. However, MLLM and VDM are trained separately, which fails to optimize jointly. Driving into the Future [130] develops a multiview world model, named Drive-WM, which is capable of generating high-quality, controllable, and consistent multi-view videos in autonomous driving scenes. It explores the potential application of the world model in end-to-end planning for autonomous driving. ChatScene [131] designs an LLM-based agent that generates and simulates challenging safety-critical scenarios in CARLA, improving the collision avoidance capabilities and robustness of autonomous vehicles. REvolve [132] is an evolutionary framework utilizing GPT-4 to generate and refine reward functions for autonomous driving through human feedback. The reward function is used for RL, and the score is achieved closely by human driving standards. GenAD [102] is a large-scale video prediction model for autonomous driving that uses extensive web-sourced data and novel temporal reasoning blocks to handle diverse driving scenarios, generalize to unseen datasets in a zero-shot manner, and adapt for action-conditioned prediction or motion planning. DriveDreamer-2 [133] builds on DriveDreamer with a Large Language Model (LLM), generates customized and high-quality multi-view driving videos by converting user queries into agent trajectories and HDMaps, enhancing training for driving perception methods. ChatSim [134] enable editable photo-realistic 3D driving scene simulations via natural language commands with external digital assets, leverages a large language model agent collaboration framework and novel multi-camera neural radiance field and lighting estimation methods to produce scene-consistent, high-quality outputs. LLM-Assisted Light [135] integrates the human-mimetic reasoning capabilities of LLMs, enabling the signal control algorithm to interpret and respond to complex traffic scenarios with the nuanced judgment typical of human cognition. It developed a closed-loop traffic signal control system, integrating LLMs with a comprehensive suite of interoperable tools. LangProp [136] is a framework that iteratively optimizes code generated by large language models (LLMs) using both supervised and reinforcement learning, automatically evaluating code performance, catching exceptions, and feeding results back to the LLM to improve code generation for autonomous driving in CARLA. These methods explore the customized authentic generations of autonomous driving data. Although these diffusion-based models achieved good results on video and image-generated metrics, it is still unclear whether they could really be used in closed-loop to really boost the performance of the autonomous driving system.

3.5 Evaluation & Benchmark

In terms of evaluation, On the Road with GPT-4V [137] conducts a comprehensive and multi-faceted evaluation of GPT-4V in various autonomous driving scenarios, including Scenario Understanding, Reasoning, and Acting as a Driver. GPT-4V performs well in scene understanding, intent recognition and driving decision-making. It is good at handling out-of-distribution situations, can accurately assess the intentions of traffic participants, use multi-view images to comprehensively perceive the environment, and accurately identify dynamic interactions between traffic participants. However, GPT-4V still has certain limitations in direction recognition, interpretation of traffic lights, and non-English traffic signs. GPT-4V Takes the Wheel [138] evaluates the potential of GPT-4V for autonomous pedestrian behavior prediction using publicly available datasets. Although GPT-4V has made significant advances in AI capabilities for pedestrian behavior prediction, it still has shortcomings compared with leading traditional domain-specific models.

In terms of benchmark, LMDrive [58] introduces LangAuto(Language-guided Autonomous Driving) CARLA benchmark. It covers various driving scenarios in 8 towns and takes into account 16 different environmental conditions. It contains three tracks: LangAuto track (updates navigation instructions based on location and is divided into sub-tracks of different route lengths), LangAuto-Notice track (adds notification instructions based on navigation instructions), and LangAuto-Sequential track (Combining consecutive instructions into a single long instruction). In addition, LangAuto also uses three main evaluation indicators: route completion, violation score, and driving score to comprehensively evaluate the autonomous driving system's ability to follow instructions and driving safety. LingoQA [108] developed LingoQA which is used for evaluating video question-answering models for autonomous driving. The evaluation system consists of three main parts: a GPT-4-based evaluation, which determines whether the model's answers are consistent with human answers; and the Lingo-Judge metric, which evaluates the model using a trained text classifier called Lingo-Judge Accuracy of answers; and correlation analysis with human ratings. This analysis involves multiple human annotators rating responses to 17 different models on a scale of 0 to 1, which are interpreted as the likelihood that the response accurately solves the problem. Reason2Drive [100] introduces the protocol to measure the correctness of the reasoning chains to resolve semantic ambiguities. The evaluation process includes four key metrics: Reasoning Alignment, which measures the extent of overlap in logical reasoning; Redundancy, aimed at identifying any repetitive steps; Missing Step, which focuses on pinpointing any crucial steps that are absent but necessary for problem-solving; and Strict Reason, which evaluates scenarios involving visual elements. LaMPilot [139] is an benchmark test used to evaluate the instruction execution capabilities of autonomous vehicles, including three parts: a simulator, a data set, and an evaluator. It employs Python Language Model Programs (LMPs) to interpret human-annotated driving instructions and execute and evaluate them within its framework. [140] evaluated the two core capabilities of large language models (LLMs) in the field of autonomous driving: first, the spatial awareness decision-making ability, that is, LLMs can accurately identify the spatial layout based on coordinate information; second, the ability to follow traffic rules to ensure that LLMs Ability to strictly abide by traffic laws while driving. [141] tests three OpenAI LLMs and several other LLMs on UK Driving Theory Test Practice Questions and Answers, and only GPT-4o passed the test, indicating that the performance of LLMs still needs to be further improved. [142] has developed an LLM-based safe autonomous driving framework, which evaluates and enhances the performance of existing LLM-AD methods in driving safety, sensitive data usage, token consumption and alignment scenarios by integrating security assessment agents. DriveSim [143] is a specialized simulator that creates diverse driving scenarios to test and benchmark MLLMs' understanding and reasoning of real-world driving scenes from a fixed in-car camera perspective. OmniDrive [144] introduces a comprehensive benchmark for visual question-answering tasks related to 3D driving, ensuring strong alignment between agent models and driving tasks through scene description, traffic regulation, 3D grounding, counterfactual reasoning, decision-making, and planning. AIDE [93] proposes an automatic data engine design paradigm, which features automatic data query and labeling using VLM and continuous learning with pseudo-labels. It also introduces a new benchmark to evaluate such automatic data engines for self-driving car perception, providing comprehensive insights across multiple paradigms including open vocabulary detection, semi-supervised learning, and continuous learning. ELM [145] is proposed to understand driving scenes over long-scope space and time, showing promising generalization performance in handling complex driving scenarios. LimSim++ [146] introduces an open-source evaluation platform for (M)LLM in autonomous driving, supporting scenario understanding, decision-making, and evaluation systems.

4 Datasets in LLM4AD

Traditional datasets such as nuScenes dataset [147, 148] lack action description [149], detailed caption, and question-answering pairs which are used to interact with LLMs. The BDD-x [96], Rank2Tell [98], DriveLM [56], DRAMA [48], NuPrompt [85] and NuScenes-QA [99] datasets represent key developments in LLM4AD research, each bringing unique contributions to understanding agent behaviors and urban traffic dynamics through extensive, diverse, and situation-rich annotations. We give a summary of each dataset in Table 1. We give detailed descriptions below.

BDD-X Dataset [96]: With over 77 hours of diverse driving conditions captured in 6,970 videos, this dataset is a collection of real-world driving behaviors, each annotated with descriptions and explanations. It includes 26K activities across 8.4M frames and thus provides a resource for understanding and predicting driver behaviors across different conditions.

Honda Research Institute-Advice Dataset (HAD) [97]: HAD offers 30 hours of driving video data paired with natural language advice and videos integrated with can-bus signal data. The advice includes Goal-oriented advice(top-down signal) which is designed to guide the vehicle in a navigation task and Stimulus-driven advice(bottom-up signal) which highlights specific visual cues that the user expects the vehicle controller to actively focus on.

Talk2Car [45]: The Talk2Car dataset contains 11959 commands for the 850 videos of the nuScenes [147, 148] training set as 3D bounding box annotations. Of the commands, 55.94% were from videos recorded in Boston, while 44.06% were from Singapore. On average, each command contains 11.01 words, which includes 2.32 nouns, 2.29 verbs, and 0.62 adjectives. Typically, there are about 14.07 commands in every video. It is a object referral dataset that contains commands written in natural language for self-driving cars.

DriveLM Dataset [56]: This dataset integrates human-like reasoning into autonomous driving systems, enhancing Perception, Prediction, and Planning (P3). It employs a "Graph-of-Thought" structure, encouraging a futuristic approach through “What if" scenarios, thereby promoting advanced, logic-based reasoning and decision-making mechanisms in driving systems.

DRAMA Dataset [48]: Collected from Tokyo's streets, it includes 17,785 scenario clips captured using the video camera, each clipped to 2 seconds in duration. It contains different annotations: Video-level Q/A, Object-level Q/A, Risk object bounding box, Free-form caption, separate labels for ego-car intention, scene classifier, and suggestions to the driver.

Rank2Tell Dataset [98]: It is captured from a moving vehicle on highly interactive traffic scenes in the San Francisco Bay Area. It includes 116 clips (20s each) of 10FPS captured using an instrumented vehicle equipped with three Point Grey Grasshopper video cameras with a resolution of 1920×1200 pixels, a Velodyne HDL-64E S2 LiDAR sensor, and high precision GPS. The dataset includes Video-level Q/A, Object-level Q/A, LiDAR and 3D bounding boxes (with tracking), Field of view from 3 cameras (stitched), important object bounding boxes (multiple important objects per frame with multiple levels of importance- High, Medium, Low), free-form captions (multiple captions per object for multiple objects), ego-car intention.

NuPrompt Dataset [85]: It represents an expansion of the nuScenes dataset, enriched with annotated language prompts specifically designed for driving scenes. This dataset includes 35,367 language prompts for 3D objects, averaging 5.3 instances per object. This annotation enhances the dataset's practicality in autonomous driving testing and training, particularly in complex scenarios requiring linguistic processing and comprehension.

NuScenes-QA dataset [99]: It is a dataset in autonomous driving, containing 459,941 question-answer pairs from 34,149 distinct visual scenes. They are partitioned into 376,604 questions from 28,130 scenes for training, and 83,337 from 6,019 scenes for testing. NuScenes-QA showcases a wide array of question lengths, reflecting different complexity levels, making it challenging for AI models. Beyond sheer numbers, the dataset ensures a balanced range of question types and categories, from identifying objects to assessing their behavior, such as whether they are moving or parked. This design inhibits the model's tendency to be biased or rely on linguistic shortcuts.

Reason2Drive [100]: It consists of nuScenes, Waymo, and ONCE datasets with 600,000 video-text pairs labeled by humans and GPT-4. It provides a detailed representation of the driving scene through a unique automatic annotation mode, capturing various elements such as object types, visual and kinematic attributes, and their relationship to the ego vehicle. It has been enhanced with GPT-4 to include complex question-answer pairs and detailed reasoning narratives.

LingoQA [108]: This dataset is a large-scale, a diverse collection for autonomous driving, containing approximately 419,000 question-answer pairs, covering both action and scenery subsets. It provides rich information about driving behavior, environmental perception, and road conditions through high-quality videos and detailed annotations. It features complex questions and free-form answers, leveraging GPT-3.5/4 to enhance the diversity and depth of content. The driving capabilities covered include actions, reasons, attention, recognition, positioning, etc., which are particularly suitable for improving the understanding and decision-making capabilities of autonomous driving systems.

NuInstruct [101]: It is a dataset featuring 91K multi-view video-QA pairs spanning 17 subtasks, each requiring comprehensive information such as temporal, multi-view, and spatial data, thereby significantly raising the complexity of the challenges. It developed a SQL-based method that automatically generates instruction-response pairs, inspired by the logical progression of human decision-making.

OpenDV-2K [102]: This dataset is a large-scale multimodal dataset for autonomous driving, comprising 2059 hours of curated driving videos, including 1747 hours from YouTube and 312 hours from public datasets, with automatically generated language annotations to support generalized video prediction model training.

5 Conclusion

In this paper, we have provided a comprehensive survey on LLM4AD. We classify and introduce different applications employing LLMs for autonomous driving and summarize the representative approaches in each category. At the same time, we summarize the latest datasets related to LLM4AD. We will continue to monitor developments in the field and highlight future research directions.

References

[1] Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. Pnpnet: End-to-end perception and prediction with tracking in the loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11553–11562, 2020.

[2] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3569–3577, 2018.

[3] Abbas Sadat, Sergio Casas, Mengye Ren, Xinyu Wu, Pranaab Dhawan, and Raquel Urtasun. Perceive, predict, and plan: Safe motion planning through interpretable semantic representations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 414–430. Springer, 2020.

[4] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, pages 1–18. Springer, 2022.

[5] Zhijian Liu, Haotian Tang, Alexander Amini, Xingyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *ICRA*, 2023.

[6] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35:6531–6543, 2022.

[7] Xiaosong Jia, Li Chen, Penghao Wu, Jia Zeng, Junchi Yan, Hongyang Li, and Yu Qiao. Towards capturing the temporal dynamics for trajectory prediction: a coarse-to-fine approach. In *CoRL*, 2022.

[8] Xiaosong Jia, Penghao Wu, Li Chen, Yu Liu, Hongyang Li, and Junchi Yan. Hdgt: Heterogeneous driving graph transformer for multi-agent trajectory prediction via scene encoding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.

[9] Xiaosong Jia, Shaoshuai Shi, Zijun Chen, Li Jiang, Wenlong Liao, Tao He, and Junchi Yan. Amp: Autoregressive motion prediction revisited with next token prediction for autonomous driving. *arXiv preprint arXiv:2403.13331*, 2024.

[10] Martin Treiber, Ansgar Hennecke, and Dirk Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical review E*, 62(2):1805, 2000.

[11] Daniel Dauner, Marcel Hallgarten, Andreas Geiger, and Kashyap Chitta. Parting with misconceptions about learning-based vehicle motion planning. In *CoRL*, 2023.

[12] Qifeng Li, Xiaosong Jia, Shaobo Wang, and Junchi Yan. Think2drive: Efficient reinforcement learning by thinking in latent world model for quasi-realistic autonomous driving (in carla-v2). In *ECCV*, 2024.

[13] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *NeurIPS 2024 Datasets and Benchmarks Track*, 2024.

[14] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision (ECCV)*, 2022.

[15] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022.

[16] Xiaosong Jia, Liting Sun, Masayoshi Tomizuka, and Wei Zhan. Ide-net: Interactive driving event and pattern extraction from human data. *IEEE Robotics and Automation Letters*, 6(2):3065–3072, 2021.

[17] Xiaosong Jia, Liting Sun, Hang Zhao, Masayoshi Tomizuka, and Wei Zhan. Multi-agent trajectory prediction by combining egocentric and allocentric views. In *Conference on Robot Learning*, pages 1434–1443. PMLR, 2022.

[18] Ming Zhang, Shenghan Zhang, Zhenjie Yang, Lekai Chen, Jinliang Zheng, Chao Yang, Chuming Li, Hang Zhou, Yazhe Niu, and Yu Liu. Gobigger: A scalable platform for cooperative-competitive multi-agent interactive simulation. In *International Conference on Learning Representations*, 2023.

[19] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *arXiv preprint arXiv:2306.16927*, 2023.

[20] Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline, 2022.

[21] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Pattern Analysis and Machine Intelligence (PAMI)*, 2023.

[22] Dian Chen and Philipp Krähenbühl. Learning from all vehicles. In *CVPR*, 2022.

[23] Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, and Hongyang Li. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving, 2023.

[24] Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, and Hongyang Li. Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving, 2023.

[25] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023.

[26] Penghao Wu, Li Chen, Hongyang Li, Xiaosong Jia, Junchi Yan, and Yu Qiao. Policy pre-training for autonomous driving via self-supervised geometric modeling. In *The Eleventh International Conference on Learning Representations*, 2022.

[27] Bu Jin, Xinyu Liu, Yupeng Zheng, Pengfei Li, Hao Zhao, Tong Zhang, Yuhang Zheng, Guyue Zhou, and Jingjing Liu. Adapt: Action-aware driving caption transformer, 2023.

[28] 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.

[29] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[30] OpenAI. Gpt-4 technical report, 2023.

[31] 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, 2023.

[32] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

[34] 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, 2023.

[35] 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, 2023.

[36] 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.

[37] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[38] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023.

[39] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision), 2023.

[40] Sebastian Höfer, Kostas Bekris, Ankur Handa, Juan Camilo Gamboa, Melissa Mozifian, Florian Golemo, Chris Atkeson, Dieter Fox, Ken Goldberg, John Leonard, et al. Sim2real in robotics and automation: Applications and challenges. *IEEE transactions on automation science and engineering*, 18(2):398–400, 2021.

[41] Ashesh Jain, Luca Del Pero, Hugo Grimmett, and Peter Ondruska. Autonomy 2.0: Why is self-driving always 5 years away? *arXiv preprint arXiv:2107.08142*, 2021.

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

[43] 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.

[44] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[45] 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)*. Association for Computational Linguistics, 2019.

[46] Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny. Grounding human-to-vehicle advice for self-driving vehicles, 2019.

[47] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions, 2023.

[48] 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.

[49] Qifeng Li, Xiaosong Jia, Shaobo Wang, and Junchi Yan. Think2drive: Efficient reinforcement learning by thinking in latent world model for quasi-realistic autonomous driving (in carla-v2). *arXiv preprint arXiv:2402.16720*, 2024.

[50] Yazhe Niu, Yuan Pu, Zhenjie Yang, Xueyan Li, Tong Zhou, Jiyuan Ren, Shuai Hu, Hongsheng Li, and Yu Liu. Lightzero: A unified benchmark for monte carlo tree search in general sequential decision scenarios. *Advances in Neural Information Processing Systems*, 36, 2024.

[51] Yuan Pu, Yazhe Niu, Jiyuan Ren, Zhenjie Yang, Hongsheng Li, and Yu Liu. Unizero: Generalized and efficient planning with scalable latent world models. *arXiv preprint arXiv:2406.10667*, 2024.

[52] Jiaqi Liu, Peng Hang, Xiao qi, Jianqiang Wang, and Jian Sun. Mtd-gpt: A multi-task decision-making gpt model for autonomous driving at unsignalized intersections, 2023.

[53] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability, 2023.

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

[55] Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving, 2023.

[56] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023.

[57] Shiyi Wang, Yuxuan Zhu, Zhiheng Li, Yutong Wang, Li Li, and Zhengbing He. Chatgpt as your vehicle co-pilot: An initial attempt. *IEEE Transactions on Intelligent Vehicles*, pages 1–17, 2023.

[58] Hao Shao, Yuxuan Hu, Letian Wang, Steven L. Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models, 2023.

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[60] Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, et al. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023.

[61] Kemou Jiang, Xuan Cai, Zhiyong Cui, Aoyong Li, Yilong Ren, Haiyang Yu, Hao Yang, Daocheng Fu, Licheng Wen, and Pinlong Cai. Koma: Knowledge-driven multi-agent framework for autonomous driving with large language models. *arXiv preprint arXiv:2407.14239*, 2024.

[62] Yuan Chen, Zi-han Ding, Ziqin Wang, Yan Wang, Lijun Zhang, and Si Liu. Asynchronous large language model enhanced planner for autonomous driving. *arXiv preprint arXiv:2406.14556*, 2024.

[63] Yupeng Zheng, Zebin Xing, Qichao Zhang, Bu Jin, Pengfei Li, Yuhang Zheng, Zhongpu Xia, Kun Zhan, Xianpeng Lang, Yaran Chen, et al. Planagent: A multi-modal large language agent for closed-loop vehicle motion planning. *arXiv preprint arXiv:2406.01587*, 2024.

[64] Senkang Hu, Zhengru Fang, Zihan Fang, Xianhao Chen, and Yuguang Fang. Agentscodriver: Large language model empowered collaborative driving with lifelong learning. *arXiv preprint arXiv:2404.06345*, 2024.

[65] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

[66] Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828*, 2024.

[67] Boyi Li, Yue Wang, Jiageng Mao, Boris Ivanovic, Sushant Veer, Karen Leung, and Marco Pavone. Driving everywhere with large language model policy adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14948–14957, June 2024.

[68] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14760–14769, 2024.

[69] Wencheng Han, Dongqian Guo, Cheng-Zhong Xu, and Jianbing Shen. Dme-driver: Integrating human decision logic and 3d scene perception in autonomous driving. *arXiv preprint arXiv:2401.03641*, 2024.

[70] 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.

[71] 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. *arXiv preprint arXiv:2310.08034*, 2023.

[72] 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. *arXiv preprint arXiv:2309.10228*, 2023.

[73] 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, 2023.

[74] 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.

[75] Siyao Zhang, Daocheng Fu, Zhao Zhang, Bin Yu, and Pinlong Cai. Trafficgpt: Viewing, processing and interacting with traffic foundation models. *arXiv preprint arXiv:2309.06719*, 2023.

[76] 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.

[77] Lening Wang, Han Jiang, Pinlong Cai, Daocheng Fu, Tianqi Wang, Zhiyong Cui, Yilong Ren, Haiyang Yu, Xuesong Wang, and Yinhai Wang. Accidentgpt: Accident analysis and prevention from v2x environmental perception with multi-modal large model. *arXiv preprint arXiv:2312.13156*, 2023.

[78] 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, 2021.

[79] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

[80] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[81] Tianyu Li, Li Chen, Huijie Wang, Yang Li, Jiazhi Yang, Xiangwei Geng, Shengyin Jiang, Yuting Wang, Hang Xu, Chunjing Xu, Junchi Yan, Ping Luo, and Hongyang Li. Graph-based topology reasoning for driving scenes. *arXiv preprint arXiv:2304.05277*, 2023.

[82] Hongyang Li, Chonghao Sima, Jifeng Dai, Wenhai Wang, Lewei Lu, Huijie Wang, Jia Zeng, Zhiqi Li, Jiazhi Yang, Hanming Deng, Hao Tian, Enze Xie, Jiangwei Xie, Li Chen, Tianyu Li, Yang Li, Yulu Gao, Xiaosong Jia, Si Liu, Jianping Shi, Dahua Lin, and Yu Qiao. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. *arXiv preprint arXiv:2209.05324*, 2022.

[83] Jishnu Jaykumar P, Kamalesh Palanisamy, Yu-Wei Chao, Xinya Du, and Yu Xiang. Proto-clip: Vision-language prototypical network for few-shot learning, 2023.

[84] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models, 2023.

[85] Dongming Wu, Wencheng Han, Tiancai Wang, Yingfei Liu, Xiangyu Zhang, and Jianbing Shen. Language prompt for autonomous driving, 2023.

[86] 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.

[87] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[88] Mustafa Shukor, Corentin Dancette, and Matthieu Cord. ep-alm: Efficient perceptual augmentation of language models, 2023.

[89] 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.

[90] Nachiket Deo, Eric M. Wolff, and Oscar Beijbom. Multimodal trajectory prediction conditioned on lane-graph traversals, 2021.

[91] Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Gohome: Graph-oriented heatmap output for future motion estimation, 2021.

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

[93] Mingfu Liang, Jong-Chyi Su, Samuel Schulter, Sparsh Garg, Shiyu Zhao, Ying Wu, and Manmohan Chandraker. Aide: An automatic data engine for object detection in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14695–14706, 2024.

[94] Xiaoji Zheng, Lixiu Wu, Zhijie Yan, Yuanrong Tang, Hao Zhao, Chen Zhong, Bokui Chen, and Jiangtao Gong. Large language models powered context-aware motion prediction. *arXiv preprint arXiv:2403.11057*, 2024.

[95] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement, 2023.

[96] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[97] Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny. Grounding human-to-vehicle advice for self-driving vehicles. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[98] Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Behzad Dariush, Chiho Choi, and Mykel Kochenderfer. Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. *arXiv preprint arXiv:2309.06597*, 2023.

[99] 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.

[100] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving, 2023.

[101] Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13668–13677, 2024.

[102] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14662–14672, 2024.

[103] Li Xu, He Huang, and Jun Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9878–9888, June 2021.

[104] Li Xu, He Huang, and Jun Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events, 2021.

[105] Yun Tang, Antonio A. Bruto da Costa, Jason Zhang, Irvine Patrick, Siddartha Khastgir, and Paul Jennings. Domain knowledge distillation from large language model: An empirical study in the autonomous driving domain, 2023.

[106] Yi Yang, Qingwen Zhang, Ci Li, Daniel Simões Marta, Nazre Batool, and John Folkesson. Human-centric autonomous systems with llms for user command reasoning, 2023.

[107] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving, 2023.

[108] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, and Oleg Sinavski. Lingoqa: Video question answering for autonomous driving, 2023.

[109] Akshay Gopalkrishnan, Ross Greer, and Mohan Trivedi. Multi-frame, lightweight & efficient vision-language models for question answering in autonomous driving. *arXiv preprint arXiv:2403.19838*, 2024.

[110] Ali Nouri, Beatriz Cabrero-Daniel, Fredrik Torner, Hakan Sivencrona, and Christian Berger. Engineering safety requirements for autonomous driving with large language models, 2024.

[111] Mehdi Azarafza, Mojtaba Nayyeri, Charles Steinmetz, Steffen Staab, and Achim Rettberg. Hybrid reasoning based on large language models for autonomous car driving. *arXiv preprint arXiv:2402.13602*, 2024.

[112] Peng Wang, Xiang Wei, Fangxu Hu, and Wenjuan Han. Transgpt: Multi-modal generative pre-trained transformer for transportation, 2024.

[113] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators, 2023.

[114] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation, 2023.

[115] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[116] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[117] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows, 2016.

[118] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[119] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[120] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[121] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[122] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023.

[123] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. *arXiv preprint arXiv:2310.07771*, 2023.

[124] 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. *arXiv preprint arXiv:2309.17080*, 2023.

[125] Ziyuan Zhong, Davis Rempe, Yuxiao Chen, Boris Ivanovic, Yulong Cao, Danfei Xu, Marco Pavone, and Baishakhi Ray. Language-guided traffic simulation via scene-level diffusion, 2023.

[126] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control, 2023.

[127] Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. *arXiv preprint arXiv:2301.04634*, 2023.

[128] Kairui Yang, Enhui Ma, Jibin Peng, Qing Guo, Di Lin, and Kaicheng Yu. Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout. *arXiv preprint arXiv:2308.01661*, 2023.

[129] Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. Adriver-i: A general world model for autonomous driving, 2023.

[130] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving, 2023.

[131] Jiawei Zhang, Chejian Xu, and Bo Li. Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15459–15469, 2024.

[132] Rishi Hazra, Alkis Sygkounas, Andreas Persson, Amy Loutfi, and Pedro Zuidberg Dos Martires. Revolve: Reward evolution with large language models for autonomous driving. *arXiv preprint arXiv:2406.01309*, 2024.

[133] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv preprint arXiv:2403.06845*, 2024.

[134] Yuxi Wei, Zi Wang, Yifan Lu, Chenxin Xu, Changxing Liu, Hao Zhao, Siheng Chen, and Yanfeng Wang. Editable scene simulation for autonomous driving via collaborative llm-agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15077–15087, 2024.

[135] Maonan Wang, Aoyu Pang, Yuheng Kan, Man-On Pun, Chung Shue Chen, and Bo Huang. Llm-assisted light: Leveraging large language model capabilities for human-mimetic traffic signal control in complex urban environments, 2024.

[136] Shu Ishida, Gianluca Corrado, George Fedoseev, Hudson Yeo, Lloyd Russell, Jamie Shotton, Joao F. Henriques, and Anthony Hu. Langprop: A code optimization framework using large language models applied to driving. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

[137] Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, Zheng Zhu, Shaoyan Sun, Yeqi Bai, Xinyu Cai, Min Dou, Shuanglu Hu, Botian Shi, and Yu Qiao. On the road with gpt-4v(ision): Early explorations of visual-language model on autonomous driving, 2023.

[138] Jia Huang, Peng Jiang, Alvika Gautam, and Srikanth Saripalli. Gpt-4v takes the wheel: Evaluating promise and challenges for pedestrian behavior prediction, 2023.

[139] Yunsheng Ma, Can Cui, Xu Cao, Wenqian Ye, Peiran Liu, Juanwu Lu, Amr Abdelraouf, Rohit Gupta, Kyungtae Han, Aniket Bera, James M. Rehg, and Ziran Wang. Lampilot: An open benchmark dataset for autonomous driving with language model programs, 2023.

[140] Kotaro Tanahashi, Yuichi Inoue, Yu Yamaguchi, Hidetatsu Yaginuma, Daiki Shiotsuka, Hiroyuki Shimatani, Kohei Iwamasa, Yoshiaki Inoue, Takafumi Yamaguchi, Koki Igari, Tsukasa Horinouchi, Kento Tokuhiro, Yugo Tokuchi, and Shunsuke Aoki. Evaluation of large language models for decision making in autonomous driving, 2023.

[141] Zuoyin Tang, Jianhua He, Dashuai Pei, Kezhong Liu, and Tao Gao. Testing large language models on driving theory knowledge and skills for connected autonomous vehicles. *arXiv preprint arXiv:2407.17211*, 2024.

[142] Xiangrui Kong, Thomas Braunl, Marco Fahmi, and Yue Wang. A superalignment framework in autonomous driving with large language models. *arXiv preprint arXiv:2406.05651*, 2024.

[143] Shiva Sreeram, Tsun-Hsuan Wang, Alaa Maalouf, Guy Rosman, Sertac Karaman, and Daniela Rus. Probing multimodal llms as world models for driving. *arXiv preprint arXiv:2405.05956*, 2024.

[144] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *arXiv preprint arXiv:2405.01533*, 2024.

[145] Yunsong Zhou, Linyan Huang, Qingwen Bu, Jia Zeng, Tianyu Li, Hang Qiu, Hongzi Zhu, Minyi Guo, Yu Qiao, and Hongyang Li. Embodied understanding of driving scenarios. *arXiv preprint arXiv:2403.04593*, 2024.

[146] Daocheng Fu, Wenjie Lei, Licheng Wen, Pinlong Cai, Song Mao, Min Dou, Botian Shi, and Yu Qiao. Limsim++: A closed-loop platform for deploying multimodal llms in autonomous driving. 2024.

[147] 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. *arXiv preprint arXiv:1903.11027*, 2019.

[148] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *arXiv preprint arXiv:2109.03805*, 2021.

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

[150] Yixuan Wang, Ruochen Jiao, Chengtian Lang, Sinong Simon Zhan, Chao Huang, Zhaoran Wang, Zhuoran Yang, and Qi Zhu. Empowering autonomous driving with large language models: A safety perspective, 2023.

[151] Can Cui, Zichong Yang, Yupeng Zhou, Yunsheng Ma, Juanwu Lu, and Ziran Wang. Large language models for autonomous driving: Real-world experiments, 2023.

[152] 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, 2023.

[153] Maonan Wang, Aoyu Pang, Yuheng Kan, Man-On Pun, Chung Shue Chen, and Bo Huang. Llm-assisted light: Leveraging large language model capabilities for human-mimetic traffic signal control in complex urban environments. *arXiv preprint arXiv:2403.08337*, 2024.

[154] SP Sharan, Francesco Pittaluga, Manmohan Chandraker, et al. Llm-assist: Enhancing closed-loop planning with language-based reasoning. *arXiv preprint arXiv:2401.00125*, 2023.

[155] Pranjal Paul, Anant Garg, Tushar Choudhary, Arun Kumar Singh, and K Madhava Krishna. Lego-drive: Language-enhanced goal-oriented closed-loop end-to-end autonomous driving. *arXiv preprint arXiv:2403.20116*, 2024.

A Large Language Models for Autonomous Driving Research Tree

In Figure 3, we gives the VLLM4AD Research Tree.

B Ethical Statement

When applying LLMs to the field of autonomous driving, we must deeply consider their potential ethical implications. First, the illusion of the model may cause the vehicle to misunderstand the external environment or traffic conditions, thus causing safety hazards. Second, model discrimination and bias may lead to vehicles making unfair or biased decisions in different environments or when facing different groups. Additionally, false information and errors in reasoning can cause a vehicle to adopt inappropriate or dangerous driving behaviors. Inductive advice may leave the vehicle vulnerable to external interference or malicious behavior. Finally, privacy leakage is also a serious issue, as vehicles may inadvertently reveal sensitive information about the user or the surrounding environment. To sum up, we strongly recommend that before deploying a large language model to an autonomous driving system, an in-depth and detailed ethical review should be conducted to ensure that its decision-making logic is not only technically accurate but also ethically appropriate. At the same time, we call for following the principles of transparency, responsibility, and fairness to ensure the ethics and safety of technology applications. We call on the entire community to work together to ensure reliable and responsible deployment of autonomous driving technology based on large language models.

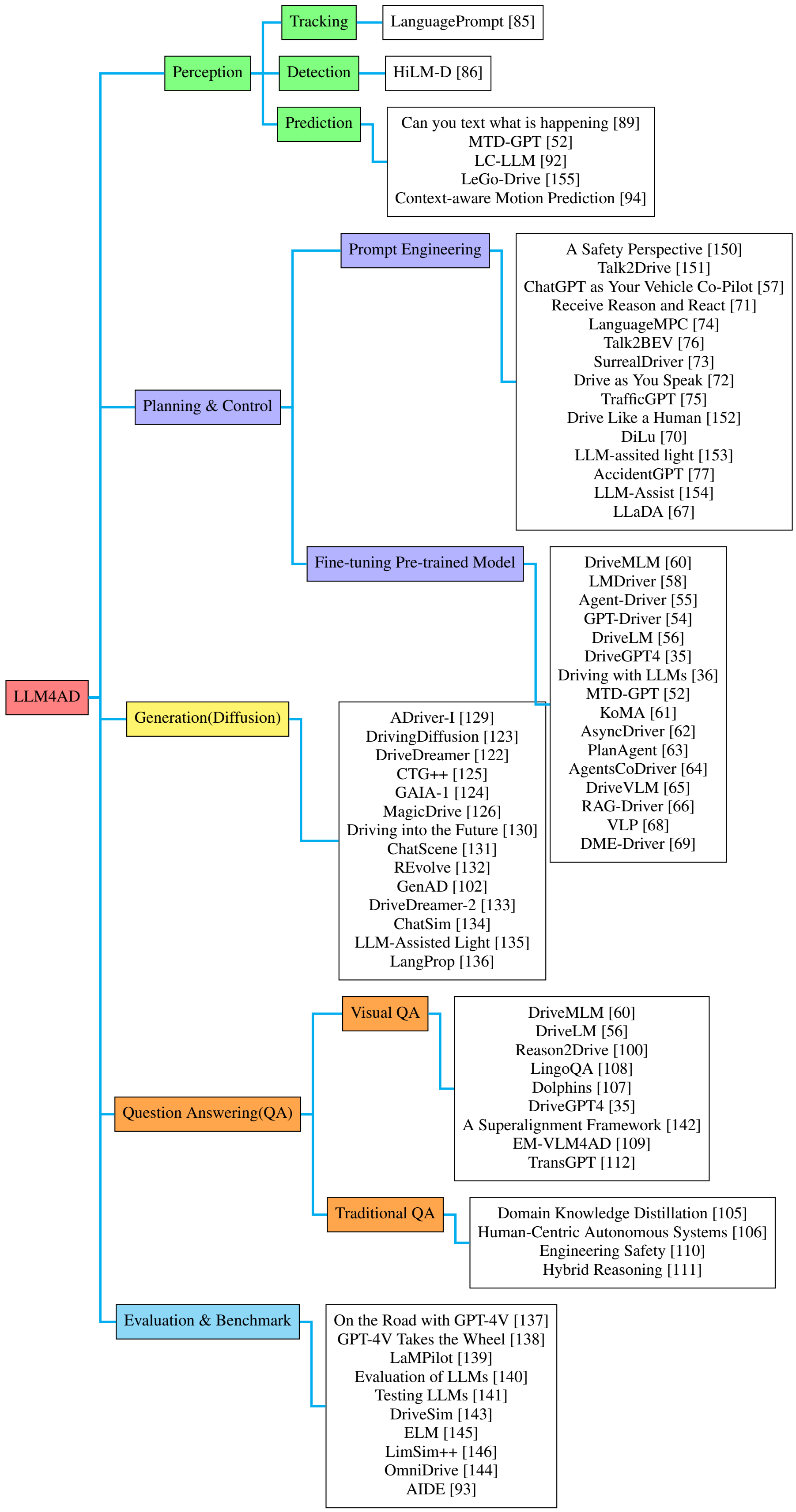

Figure 3: Vision Large Language Models for Autonomous Driving Research Tree