Large (Vision) Language Models for Autonomous Vehicles: Current Trends and Future Directions

Hanlin Tian[1*], *Student Member, IEEE,* Kethan Reddy[1*], Yuxiang Feng[1†], *Member, IEEE,*
Mohammed Quddus[1], Yiannis Demiris[2], *Senior Member, IEEE,* and Panagiotis Angeloudis[1]

***Abstract*—As autonomous vehicles (AVs) continue to advance, the integration of Large (Vision) Language Models (L(V)LMs) into AV systems has become increasingly significant. These models excel in natural language processing and visual reasoning, making them invaluable for enhancing the capabilities of AV systems across various domains. This survey provides a comprehensive and up-to-date overview of current research and developments in the application of L(V)LMs in autonomous driving, focusing on four key areas: modular integration, end-to-end integration, data generation, and platforms and datasets. We critically assess various methodologies, outcomes, and the strengths and limitations of these approaches, highlighting the gaps in existing surveys and how our work addresses them by providing detailed analyses of practical integration strategies and real-world implementations. Additionally, we explore future research directions, including the practical integration of L(V)LMs with existing AV systems, addressing regulatory and ethical challenges, and enhancing Vehicle-to-Everything (V2X) communication. This survey aims to inform and guide future innovations in the field by offering detailed insights into L(V)LM model choices, computational trade-offs, and task-specific requirements for autonomous driving.**

***Index Terms*—Autonomous vehicles, Natural language processing, Computer vision.**

I. Introduction

Large Language Models (LLMs) like GPT-4 [1] and Claude 3.5 [2] are advanced AI models designed to understand and generate human-like text. They excel in tasks such as text generation, translation, and summarisation. Large Vision-Language Models (LVLMs) on the other hand, such as CLIP [3], BLIP-2 [4], and LLaVA [5], integrate visual understanding with language processing, enabling tasks like image captioning and visual reasoning. These capabilities make LLMs and LVLMs especially valuable in AVs, where visual and textual comprehension, human-machine interaction, and explainability are crucial. Henceforth throughout this paper, we reference LLMs **or** LVLMs as **L(V)LMs** as shorthand to avoid repetition. Some explicit examples of how these models can bolster the AV task are as follows: improved natural language interfaces that interpret and respond to commands in a conversational manner, context-aware decision-making to predict traffic patterns, real-time semantic understanding of driving environments to explain AV decisions, etc.

To further expound on the utility of L(V)LMs in the context of real-world applications concerning the AV task, concrete examples can be observed through industry adoption. The recent success of Wayve can, in part, be owed to the front-facing product showcase of *LINGO-1* that demonstrates how an open-loop driving model can perform visual question answering (VQA) on tasks such as perception, counterfactuals, planning, reasoning and attention, lending credence to the possibility of AI-explainability. Iterating on the framework, Wayve's *LINGO-2* is the first closed-loop vision-language-action driving model (VLAM) tested on public roads [6]. Beyond Wayve, NVIDIA utilizes LLMs to comprehend complex and long-tailed open-world driving scenarios, addressing a broad spectrum of AV tasks [7].

Furthermore, numerous research institutes and communities are actively exploring L(V)LM integration within the AV context, reflecting the growing interest in this field. For instance, *DriveLM* [8] studies how LVLMs trained on web-scale data can be integrated into end-to-end driving systems to boost generalisation. Another example is *GenAD* [9], which is a generalised video prediction model for AVs that supports action-conditioned prediction and planning. This underscores the versatility of L(V)LMs, *Drive as you speak* uses LLMs for voice commands [10], but because of this breadth of potential use cases, it is becoming increasingly difficult to keep track of cutting-edge research in this niche field.

Existing surveys on Large (Vision) Language Models in autonomous vehicles offer valuable insights but have notable gaps. *LLM4Drive* [11] overviews applications in perception, prediction, and planning but lacks deep exploration of practical integration challenges and broader implications like regulatory and ethical considerations. *A Survey for Foundation Models in Autonomous Driving* [12] categorizes foundation models but does not extensively analyze integration challenges or computational trade-offs involved in deployment. *Towards Knowledge-Driven Autonomous Driving* [13] proposes a conceptual framework emphasizing knowledge integration to enhance cognition and learning but does not specifically focus on the role of L(V)LMs or provide detailed analyses of their applications within AV systems. *Large Language Models for Mobility in Transportation Systems* [14] focuses on mobility forecasting, neglecting other critical AV aspects like perception and control. Collectively, while these surveys highlight advancements, they underexplore the practical integration of L(V)LMs into existing AV systems, especially regarding regulatory, ethical, and Vehicle-to-Everything (V2X) communication considerations.

[1] H. Tian, K. Reddy, Y. Feng, M. Quddus, and P. Angeloudis are with the Centre for Transport Engineering and Modelling, Department of Civil and Environmental Engineering, Imperial College London, UK

[2] Y. Demiris is with the Personal Robotics Laboratory, Department of Electrical and Electronic Engineering, Imperial College London, UK

[*] Equal contribution

[†] Corresponding author: Yuxiang Feng (e-mail: y.feng19@imperial.ac.uk)

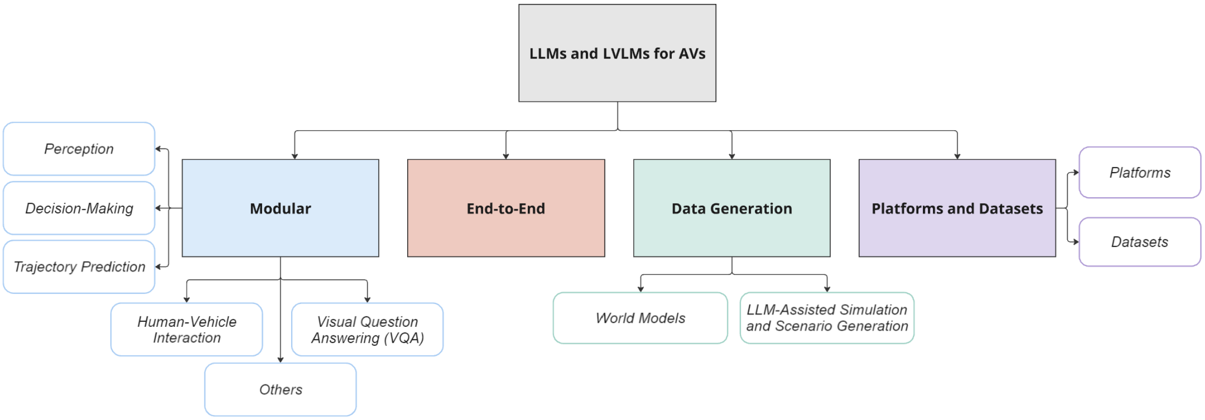


Fig. 1. Diagrammatic representation of the key areas for integrating L(V)LMs into AV architectures, categorized into modular integration, end-to-end integration, data generation, and platforms and datasets.

To address these gaps, this survey offers a comprehensive overview of current research developments in L(V)LM applications for AVs. We critically assess various approaches, highlighting methodologies, outcomes, strengths, and limitations, with a focus on practical integration into existing AV systems in real-world contexts. By defining tasks suited for L(V)LMs in AVs and accounting for the unique requirements and constraints of AV systems, we aim to provide clearer guidance for future research and development in this field. The main contributions of this work can be summarised as follows:

- Provide an up-to-date review, as of June 2024, of 62 papers on the integration and impact of L(V)LMs in autonomous vehicles.
- Delineated and detailed the L(V)LM models, platforms, datasets and benchmarks used in each paper.
- Propose further research directions and innovative approaches to address the current challenges in the integration and application of L(V)LMs in AVs.

The outline of this paper is as follows: In Section II, a review of existing literature related to L(V)LM adoption in the AV context is provided. Section III elucidates the selection, curation, and assimilation process for choosing appropriate papers. Section IV explores recent advancements in modular integration of L(V)LMs within existing AV architectures. Section V highlights L(V)LM applications in end-to-end integration within AV systems and frameworks. Section VI examines how L(V)LM models can be leveraged for data generation across various AV domains. Section VII overviews typically utilised AV platforms and datasets, in addition to new L(V)LM-specific benchmarks to stress test the performance and efficiency of AV L(V)LM pipelines. Section VIII expands on particular potential shortcomings and research directions pertaining to L(V)LMs in AV use cases. And the key conclusions are presented in Section IX.

II. Literature Review

The scope of the literature review primarily encompasses three sub-areas, which reflect how L(V)LMs are typically incorporated in the AV domain. **L(V)LMs** systems can be tacked onto AV architectures, dividing tasks like perception and decision-making into specialised components to create intelligent agents capable of advanced comprehension and interaction. **End-to-End AV** architectures represent state-of-the-art solutions for autonomous vehicles, providing a seamless integration of perception, prediction, and planning into a unified framework. And finally, the integration of L(V)LMs into data generation frameworks is particularly promising. Due to their extensive object-relation ontologies and ability to translate abstract scenarios into structured, machine-readable formats. And when these models work in tandem with **diffusion models**, they enable realistic images and videos to sharpen the effectiveness of simulation-based testing.

A. Related Work on L(V)LMs

LLMs have evolved significantly in recent years, exhibiting advanced capabilities in text generation, comprehension, and other areas [15]. The recent release of Llama 3 [16] showcases the developmental and adoption speed typically exhibited in the AI field, now considered one of the most widely used open-source LLM models. Its robustness and versatility have made it a popular choice in academic research, supporting a wide range of studies across various disciplines.

LVLMs advance the integration of natural language processing and computer vision, enabling holistic interpretation of multimodal data. LLaVA (Large Language and Vision Assistant) [5] excels in tasks like image captioning, visual question answering, and multimodal translation by aligning visual and linguistic representations through extensive pre-training on image-text pairs. Similarly, Qwen-VL [17] enhances large language models with vision capabilities, achieving superior performance in complex tasks requiring detailed scene understanding and interaction, essential for applications like AVs.

B. Related Work on End-to-End Autonomous Vehicles

End-to-end autonomous vehicles integrate perception, prediction, and planning into a single framework. Early work,

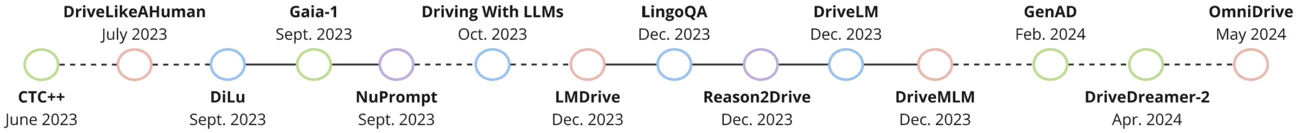


Fig. 2. Timeline of some select papers incorporating L(V)LMs for AVs. **Blue** indicates L(V)LM integration within a modular framework, **red** signals L(V)LM end-to-end architectures, **green** refers to L(V)LM data generation pipelines, and **purple** references L(V)LM platforms and datasets.

such as ALVINN, used neural networks for direct steering control from sensor inputs [18]. This concept was revitalised by NVIDIA's end-to-end learning system in 2016, which employed CNNs to process raw camera data and output steering commands [19].

Modern systems predominantly use imitation learning (IL) and reinforcement learning (RL). IL methods, like Conditional Imitation Learning (CIL), learn from human driving data and condition the policy on high-level commands [20]. RL approaches, using algorithms like Deep Q-Networks (DQN) and Asynchronous Advantage Actor-Critic (A3C), optimise policies through simulated interactions [21].

Recent advancements in end-to-end AV systems have demonstrated their potential benefits. For instance, UniAD [22] highlights how such systems can enhance transportation efficiency and safety through comprehensive case studies and empirical evidence. These systems show performance improvements and the ability to effectively manage complex driving scenarios, making them a promising direction for future research and development in AV technology.

C. *Related Work on Diffusion Models*

In addition to the prolific success of the transformer architecture for language generation, diffusion models responsible for tabula rasa image and video generation have also skyrocketed in popularity. This is evidenced by the rate of adoption of DALL-E [23], and Stable Diffusion [24] across various industry vertical use cases. Diffusion models operate by progressively transforming random noise into coherent images or videos through iterative refinement [25]. The flexibility of diffusion models in accepting diverse input modalities, such as text prompts and sketches, enables rapid prototyping and customised outputs [26].

In the context of AV research, latent diffusion models are trained on high-level scene components and dynamics by mapping textual driving descriptions, video frames, road infrastructure data, and actions to a shared latent space, then are decoded into high-quality, temporally consistent video frames. These models can generate realistic scenarios based on specific instructions, demonstrating coherent scene generation and prolonged temporal consistency [27], [28]. These are referred to as World Models. A full technical overview of World Models is outlined in section VI-B.

III. Methodological Approach

The methodological approach opted for this survey involved a comprehensive and systematic review of literature related to L(V)LMs in the field of AVs. We conducted extensive searches, predominantly between Sept. 2023 and June 2024 (Fig. 2), across various academic databases including IEEE Xplore, Google Scholar, and arXiv. The selection criteria focused on identifying studies that explicitly discussed LLMs or LVLMs and their applications, challenges, and advancements in AVs.

A. *Search Strategy and Selection Criteria*

A systematic review was employed to ensure a replicable, transparent procedure. To locate relevant studies on L(V)LMs for AVs, a set of keywords was first identified as the research pillars. These keywords included "LLM," "large language models", "vision language models," "autonomous vehicles," "self-driving," and various combinations of these terms. Although overlaps might occur during the search, utilising multiple databases ensures comprehensive coverage of as many pertinent articles as reasonably possible. This process resulted in the selection of 62 papers published before June 1, 2024, for this review. We excluded studies that **solely utilised vision transformers** to better focus on the unique contributions of L(V)LMs, which offer LLM-enhanced commonsense reasoning and knowledge utilisation capabilities for AVs.

B. *Quality Assessment and Methodological Limitations*

The quality of the studies was evaluated based on peer-review status, and the strength of research methodologies and findings. Peer-reviewed papers published in conferences and journals were prioritised as they typically ensure high quality and rigour. Additionally, we focused more on peer-reviewed and published papers to ensure a higher level of reliability and scholarly integrity.

Despite these measures, the fast-paced nature of L(V)LM research in AVs imposes certain methodological limitations on our survey. The emergence of new findings post-review potentially narrows the survey's scope.

C. *Classification of L(V)LM Applications in Autonomous Vehicles*

The application of L(V)LMs in autonomous vehicles is diverse. This survey classifies the research by their fields to provide a structured overview, as depicted in Fig. 1:

- **Modular Integration** approaches bolster specific components like perception, planning, and control.
- **End-to-end Integration** systems process sensory data directly and make driving decisions, integrating all tasks into a single cohesive model for holistic responses.

- **Data Generation** frameworks create diverse and realistic training scenarios, which are crucial for developing robust autonomous vehicle models.
- **Platforms and Datasets** establish standards for evaluating the performance of autonomous vehicle systems, essential for measuring progress and identifying areas for improvement.

IV. Modular Integration

Autonomous vehicle systems often adopt modular architecture, dividing the overall task into specific components such as perception, decision-making, trajectory prediction, and human-vehicle interaction. This structured approach allows for the specialised development and optimisation of each component. Modular systems contrast with end-to-end approaches, which aim to streamline all tasks into a single unified framework.

In this section, we explore the various modules that comprise an autonomous vehicle system, detailing recent advancements and methodologies in each area. This section is divided into several subsections: **Perception**, **Decision-making**, **Trajectory Prediction**, **Human-Vehicle Interaction**, **Visual Question Answering (VQA)**, and **others**.

A. Perception

Perception systems aim to improve the vehicle's environmental understanding and navigation through advanced scene interpretation, object recognition, and contextual reasoning. Key tasks include improving visual and spatial reasoning to deduce spatial relationships and visual attributes of objects within the scene. Visual anomaly detection which involves the identification and understanding of contextually irregular situations or anomalies to enhance safety is also important, as illustrated in Fig. 3. Recent advancements incorporating L(V)LMs into perception systems have improved visual scene understanding and reasoning.

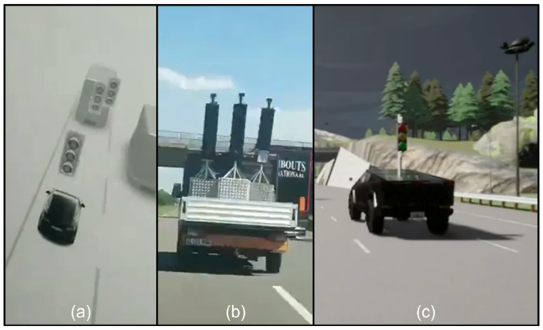


Fig. 3. An example of perception tasks in autonomous vehicles: a semantic anomaly where traffic lights appear to pass through a car, later revealed as inactive lights in transport. Image from [29].

a) Semantic Anomaly Detection with Large Language Models [29]: This paper tackles identifying semantic anomalies in vision-based policies. By converting visual observations into textual descriptions, the framework detects contextually irregular situations, such as inactive traffic lights being transported (Fig. 3). Evaluated in the CARLA simulator, this approach effectively identifies and reasons about semantic anomalies, aligning with human judgment.

b) Zelda [30]: It is an LVLM for video analytics, enabling natural language queries and improving result relevance and diversity. Zelda reduces redundant results and enhances accuracy by employing a sophisticated prompting strategy and semantic-rich embeddings. Evaluated with the VIVA engine on the BDD dataset, Zelda shows improvements in retrieval precision and efficiency compared to traditional systems.

c) Talk2BEV [31]: It is an LVLM interface for bird's-eye view (BEV) images in autonomous vehicle contexts to ameliorate visual reasoning and spatial understanding. The framework augments BEV maps with image-language features, utilising pre-trained LVLMs without additional training.

Summary on Perception: Leveraging pre-trained VLMs, the reviewed papers demonstrate improvements in perception tasks for autonomous driving. These models effectively tackle challenges in object detection, classification, semantic anomaly detection, and visual reasoning. The use of common evaluation metrics such as accuracy, IoU, and mAP across various simulation platforms like CARLA and datasets like BDD-X validates their performance. Experiments on platforms like CARLA and datasets such as BDD-X highlight the practical benefits of these models, validating improvements in perception tasks and emphasising the need for continued research to enhance efficiency and real-time processing across diverse driving scenarios. Notably, Talk2BEV and Zelda leverage local GPUs (Nvidia T4 and A100) for high-performance computation.

B. Decision-making

Decision-making in autonomous driving involves the process by which a vehicle determines the safest and most efficient actions to take in response to its environment. This process includes navigating through complex, dynamic scenarios, predicting the behaviour of other road users, and planning trajectories that ensure both safety and compliance with traffic regulations. Integrating Large (Vision) Language Models (L(V)LMs) into these systems enhances the vehicle's ability to interpret nuanced situations, apply advanced reasoning, and make more informed decisions. Fig. 4 illustrates an example framework for integrating LLMs into the decision-making process of an AV system.

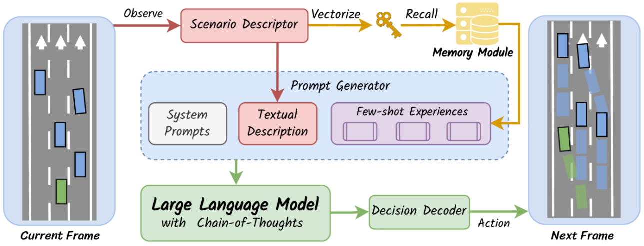


Fig. 4. Example framework for integrating LLMs into decision-making in AV systems. Image from [38].

Recent progress suggests that models that incorporate L(V)LMs into the decision-making module boost the safety, efficiency, and adaptability of autonomous vehicles by aligning complex scenario outcomes (or proposed trajectories) closer to

TABLE I
TABLE FOR PERCEPTION PAPERS

Paper	Submission Time	Specific Tasks	LLM Model	Dataset or Simulator
Semantic Anomaly Detection with Large Language Models [31]	May 2023	Semantic anomaly detection in vision-based policies	GPT-3.5 [32]	CARLA [33]
Zelda [30]	May 2023	Video analytics	VIVA [34]	BDD-X [35]
Talk2BEV [31]	Oct 2023	Visual reasoning, spatial understanding, decision-making	BLIP-2 [4] , MiniGPT-4 [36] , InstructBLIP [37]	Talk2BEV-Bench [31]

expected human outcomes, and does so through their implicitly learned object relation ontologies and semantic parsing of human experience. Additionally, L(V)LM architectures can model scenarios as sequence problems, which easily permits the unification and coupling of natural language with trajectory data and information to process in latent space. As further detailed below, these capabilities lead to safer, more efficient, and context-aware AV systems.

a) MTD-GPT [39]: Models the multi-task decision-making problem of autonomous vehicles crossing unsignalised intersections as a sequence modelling problem. The proposed pipeline trains single-task decision-making expert algorithms through reinforcement learning, which then provides an expert dataset that is transformed into sequential data for offline training of the MTD-GPT model.

b) Empowering Autonomous Driving with LLMs [41]: This paper integrates LLMs with a model predictive control trajectory planner to enhance safety. Utilising the LLM as a decision-maker for lane changes simplifies the highway driving task when modelled as an MPC problem by removing the discrete decision-making steps, thus reducing the computational complexity required at the controller level.

c) Receive, Reason, and React [43]: Enables LLMs to serve as a decision-making module in autonomous vehicles when requested by the human driver. A structured language generator is utilised to formulate observations received from the perception and localisation modules into format-specific contexts that can be ingested by an LLM. This then enables the LLM to make definite categorical decisions for highway overtaking and on-ramp merging tasks, such as “change lanes”, “accelerate”, “decelerate”, etc.

d) LanguageMPC [44]: Is a chain-of-thought framework for LLMs to handle driving scenarios, dividing the complex decision-making process into numerous sub-problems for the LLM to make informed decisions based on the state of the environment, traffic rules, and the logical reasoning ability of the LLM. The textual outputs of the LLM are converted to mathematical representations, namely a weight matrix, an observation matrix and an action bias, which are used as parameters for an MPC that directly controls the vehicle.

e) DiLu [38]: A framework for AV that integrates LLMs to enable decision-making based on common-sense knowledge. DiLu consists of a driver agent, interactive environment, and memory component. The agent employs a reasoning module to query experiences from memory, and a reflection module to refine decisions.

f) A Language Agent for Autonomous Driving [46]: Transforms the traditional perception-prediction-planning pipeline by integrating a versatile tool library for dynamic function calls, a cognitive memory storing common sense and experiential knowledge, and a reasoning engine capable of chain-of-thought reasoning, task planning, motion planning, and self-reflection.

g) LLM Multimodal Traffic Forcasting [48]: Studies the application of L(V)LMs for traffic accident forecasting. The proposed framework leverages deep learning methods, including transformers, alongside traditional models like ARIMA and Prophet, to predict traffic accidents using rich datasets. The model integrates LLMs coupled with LLaVA for real-time multimodal data processing.

h) Driving with LLMs [51]: Proposes a multimodal LLM architecture that merges vectorised numeric data with pre-trained LLMs to better contextually understand driving scenarios. The approach aligns numeric vector modalities with LLMs through vector captioning, demonstrating improved decision-making compared to traditional methods.

i) CAVG [52]: Addresses visual grounding in AVs using LLMs. The Context-Aware Visual Grounding (CAVG) model integrates multiple encoders—text, emotion, image, context, and cross-modal—with a multimodal decoder to capture contextual semantics and human emotional features. The CAVG model employs multi-head cross-modal attention mechanisms and a Region-Specific Dynamic (RSD) layer for enhanced attention modulation.

j) LLM-Assist [55]: Presents a hybrid planning framework for AV that integrates rule-based planners and LLM-based planners. The system leverages the common-sense reasoning capabilities of LLMs to generate robust and well-reasoned plans. LLM-Assist uses a conventional rule-based planner, PDM-Closed, for common scenarios and invokes the LLM-based planner for complex situations where rule-based methods struggle. The LLM provides trajectory or parameter adjustments to the base planner.

k) Evaluation of LLMs for Decision Making in Autonomous Driving [57]: Evaluates LLMs for decision-making in AV, focusing on spatial-aware decision-making and adherence to traffic rules. The study assesses different LLMs' performance in both simulated real-world traffic conditions and actual vehicle deployment. The results indicate that GPT-4 outperforms the other models in both accuracy and reasoning, particularly in complex scenarios requiring ethical judgments and traffic rule compliance.

TABLE II
TABLE FOR DECISION MAKING PAPERS

Paper	Submission Time	Specific Tasks	LLM Model	Dataset or Simulator
MTD-GPT [39]	Sep 2023	Multi-Task Decision-Making, Uncontrolled Intersection	GPT-2	OpenAI Gym [40]
Empowering Autonomous Driving with LLMs [41]	Nov 2023	Decision Making, Lane Change	GPT-4 [1]	HighwayEnv [42]
Receive, Reason and React [43]	Apr 2024	Decision Making, Highway Overtaking, On-ramp Merging	GPT-4 [1]	HighwayEnv [42]
LanguageMPC [44]	Oct 2023	Decision Making, Intersection, Obstacle Avoidance	GPT-3.5 [32]	IdSim [45]
DiLu [38]	Sep 2023	Knowledge-Driven Decision Making	Not Specified	HighwayEnv [42]
A Language Agent for Autonomous Driving [46]	Nov 2023	Decision Making, Motion Planning	GPT-3.5 [32]	nuScenes [47]
LLM Multimodal Traffic Forecasting [48]	Oct 2023	Traffic Accident Forecasting	GPT-4 [1] LLaMa2 [16] Zephyr-7b-α [49]	Fatality Analysis Reporting System (FARS) [50]
Driving with LLMs [51]	Oct 2023	Decision Making, QA Task	GPT-3.5 [32]	Custom 2D Simulator
CAVG [52]	Oct 2023	Visual Grounding	GPT-4 [1], BERT [53]	Talk2Car [54]
LLM-Assist [55]	Dec 2023	Decision Making	GPT-3 [32], GPT-4 [1]	nuPlan [56]
Evaluation of LLMs for Decision Making in Autonomous Driving [57]	Dec 2023	Decision Making, Traffic Rules	LLaMa2 [16] GPT-3.5 [32], GPT-4 [1]	Field test
Hybrid Reasoning Based on LLMs for Autonomous Car Driving [58]	Feb 2024	Decision Making, Adverse Weather	GPT-4 [1]	CARLA [33]
Multi-Modal GPT-4 Aided Action Planning and Reasoning for Self-driving Vehicles [59]	Mar 2024	Decision Making, Action Planning, Reasoning	GPT-4 [1]	CARLA [33]

l) Hybrid Reasoning Based on LLMs for Autonomous Car Driving [58]: Leverages LLMs for enhancing AV decision-making under various weather conditions. The study integrates both common-sense and arithmetic reasoning to process multimodal data from object detection and sensors. By evaluating nine distinct scenarios, the framework demonstrates the capability of LLMs to improve decision accuracy and response times, particularly in adverse conditions such as heavy rain.

m) Multi-Modal GPT-4 Aided Action Planning and Reasoning for Self-driving Vehicles [59]: Leverages GPT-4 for action planning and reasoning in autonomous vehicles using multi-modal data. The system processes time-series data from a monocular camera through a graph-of-thought (GoT) structure, enabling robust policy learning and generating natural language rationales.

Summary on Decision Making: The reviewed frameworks leverage L(V)LMs' (mostly LLMs) advanced reasoning capabilities to simplify and manage decision-making. Models like MTD-GPT and LanguageMPC handle multi-task decision-making and logical reasoning for trajectory planning effectively. Frameworks such as DiLu and LLM-Assist demonstrate the use of common-sense knowledge and hybrid planning for accurate and context-aware decision-making. Key evaluation metrics include accuracy, response time, and adherence to traffic rules. However, limitations such as decision-making latency and the presence of hallucinations in models like DiLu highlight the need for further research into LLM compression and optimisation. Most of the aforementioned papers use ChatGPT API, suggesting a shift towards using local L(V)LMs to address latency issues.

C. Trajectory Prediction

Trajectory prediction aims to forecast the future positions of traffic participants over a given time horizon. This is distinct from AV decision-making as trajectory prediction entails forecasting *the sequence* of kinematic measures (i.e. position, velocity, acceleration, pose, etc.) associated with traffic participants, whereas decision-making refers to taking appropriate action at a specific timestep. Accurate trajectory prediction is essential for safe and efficient motion planning, collision avoidance, and overall autonomous vehicle operation. The goal of trajectory prediction is to minimise errors and miss rates, thereby enhancing the autonomous vehicle's ability to navigate complex and dynamic traffic scenarios safely and efficiently.

a) GPT-driver: [60] Transforms motion planning for autonomous vehicles into a language modelling problem.

TABLE III
TABLE FOR TRAJECTORY PREDICTION PAPERS

Paper	Submission Time	Specific Tasks	LLM Model	Dataset or Simulator
GPT-driver [60]	Oct 2023	Trajectory Prediction, Motion planning	GPT-3.5 [32]	nuScenes [47]
LLaDA [61]	Feb 2024	Traffic Rule Assistance for Tourists, AV Motion Plan Adaptation	GPT-4 [1]	nuScenes [47], nuPlan [56]
Diffusion-ES [62]	Feb 2024	Trajectory Optimisation, Zero-Shot Instruction Following	GPT-3.5 [32]	nuPlan [56]
LC-LLM [63]	Mar 2024	Lane Change Intention, Trajectory Predictions	LLaMA2-7B [16]	highD [64]

Using GPT-3.5, it reformulates planner inputs and outputs as language tokens, generating driving trajectories through natural language descriptions. The method employs a prompting-reasoning-finetuning strategy, allowing the LLM to describe precise trajectory coordinates in its decision-making process.

b) LLaDA: [61] Uses LLMs to adapt driving behaviour to new environments, customs, and laws. It processes inputs like execution plans, local traffic codes, and scene descriptions to generate adaptive driving instructions. Demonstrated on the nuScenes dataset, LLaDA improves motion planning by integrating a Traffic Rule Extractor (TRE) with LLMs like GPT-4 to apply relevant traffic regulations effectively.

c) LC-LLM [63]: Leverages the reasoning and self-explanation capabilities of LLMs to predict lane change intentions and future trajectories of vehicles. By reformulating the lane change prediction task as a language modelling problem, the model processes driving scenario information as natural language prompts and fine-tunes the LLMs for this specific task. The model employs chain-of-thought reasoning to enhance prediction transparency and reliability.

d) Diffusion-ES: [62] Optimises trajectory planning for AV using a gradient-free method combined with trajectory denoising. This approach uses a diffusion model to sample and mutate trajectories guided by a reward function, optimising non-differentiable language-shaped reward functions generated by LLM-prompting to follow complex driving instructions.

Summary on Trajectory Prediction: These models leverage both language and visual data to improve the accuracy and interpretability of trajectory forecasts. Evaluation metrics such as Average Displacement Error (ADE), Final Displacement Error (FDE), and Miss Rate (MR) are crucial for assessing model performance. Frameworks like GPT-driver and LC-LLM highlight the use of language modelling for generating precise trajectory coordinates and decision explanations. The reviewed models demonstrate enhanced motion planning and collision avoidance capabilities.

D. Human-vehicle Interaction

The primary task in human-vehicle interaction using LLMs is to enable seamless communication between the driver and the autonomous system through natural language. This includes interpreting verbal commands, providing context-aware responses, and adapting to individual driver preferences to ensure safe, comfortable, and efficient driving experiences.

a) DriveAsYouSpeak: [10] Utilises LLMs to interpret verbal commands from drivers and translate them into executable actions, supported by sensory tools like perception modules and localisation systems to provide real-time environmental awareness.

b) ChatGPT as Your Vehicle Co-Pilot: [65] Explores using ChatGPT to enhance human-machine interaction in AV. This framework integrates ChatGPT as a ”Co-Pilot” to interpret and fulfil human driving intentions, handling tasks such as path tracking control and trajectory planning based on natural language commands.

c) Dolphins: [7] Introduces a vision-language model for autonomous vehicles, built on OpenFlamingo. It processes video data, text instructions, and control signals to generate informed driving decisions. Enhanced by the Grounded Chain of Thought (GCoT) process and specialised instruction tuning using the BDD-X dataset, Dolphins excels in behaviour comprehension, control signal forecasting, and conversational understanding.

d) Talk2Drive: [69] Leverages LLMs to enable autonomous vehicles to understand and execute natural verbal commands, enhancing personalisation in driving experiences. Using a memory module, it translates verbal commands into executable controls and adapts to individual driver preferences, reducing driver takeover rates and improving trust in the autonomous system through field experiments.

Summary on Human-vehicle Interaction: Integrating L(V)LMs into human-vehicle interaction systems significantly enhances communication between drivers and autonomous vehicles. These advancements facilitate natural language interactions, leading to more intuitive, personalised, and safer driving experiences. The highlighted studies demonstrate the capabilities of LLMs in interpreting verbal commands, providing real-time responses, and adapting to driver preferences. Evaluation metrics such as safety, comfort, efficiency, user satisfaction, and system robustness are crucial for assessing the effectiveness of these systems. Continued research in this area will further refine these interactions, making AVs more reliable and user-friendly.

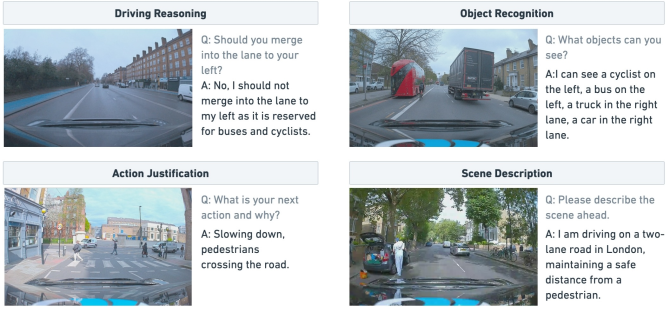
E. VQA

Visual Question Answering (VQA) combines computer vision and natural language processing to answer questions based on visual input (images or videos). In autonomous vehicles, VQA systems analyse sensory data to provide insights

TABLE IV
TABLE FOR HUMAN-VEHICLE INTERACTION PAPERS

Paper	Submission Time	Specific Tasks	LLM Model	Dataset or Simulator
DriveAsYouSpeak [10]	Sep 2023	Natural Language Interaction, Adaptive Decision-Making	ChatGPT-4 [1]	Not Mentioned
ChatGPT as Your Vehicle Co-Pilot [65]	Oct 2023	Path Tracking, Trajectory Planning, Human-Machine Interaction	ChatGPT-3.5 Turbo [32]	Simulink [66], CarSim [67]
Dolphins [7]	Dec 2023	Decision Making, Holistic Driving Understanding	CLIP [3] LLaMA [16] MPT [68]	BDD-X [35] DriveLM [8]
Talk2Drive [69]	Dec 2023	Verbal Commands Controls, Personalized Driving Preferences	GPT-4 [1]	Autoware [70]

into the environment, objects, and scenarios. By explaining the vehicle's actions and decisions, VQA enhances transparency and trust, improving user confidence in safety. It also aids in debugging and refining autonomous systems.



Q: Should you merge into the lane to your left?
A: No, I should not merge into the lane to my left as it is reserved for buses and cyclists.

Q: What objects can you see?
A:I can see a cyclist on the left, a bus on the left, a truck in the right lane, a car in the right lane.

Q: What is your next action and why?
A: Slowing down, pedestrians crossing the road.

Q: Please describe the scene ahead.
A: I am driving on a two-lane road in London, maintaining a safe distance from a pedestrian.

Fig. 5. Examples of Visual Question Answering (VQA) tasks in AV, showcasing driving reasoning, object recognition, action justification, and scene description. Adapted from [71].

The task of VQA involves interpreting sensory data from the vehicle's environment to answer natural language questions. These questions can pertain to driving scenarios, object identification, action justification, and scene description. The input to a VQA system typically includes images or videos from the vehicle's cameras and a natural language question. The output is a textual answer that accurately reflects the information derived from the visual input. Evaluating open-ended textual dialogues in VQA is challenging due to the ambiguity and subjectivity of correct answers. Common metrics for evaluating question-answering models include BLEU, METEOR, and CIDEr (Fig. 5).

a) LingoQA: [71] Introduces a VQA benchmark for AVs with an evaluation dataset and a classifier-based metric called Lingo-Judge, which correlates highly with human evaluations. This benchmark facilitates efficient exploration and improvement of VQA systems by providing rapid feedback and reliable evaluation.

b) DriveLM: [8] Proposes Graph Visual Question Answering (GVQA) for AVs, using vision-language models to enhance reasoning through interconnected question-answer pairs. DriveLM-Data, built on nuScenes and CARLA, trains DriveLM-Agent for GVQA and end-to-end driving.

c) LiDAR-LLM: [72] Adopts Large Language Models for 3D LiDAR data in AVs, performing tasks like 3D captioning and question answering. A three-stage training strategy aligns LiDAR data with language embeddings, enhancing spatial comprehension. Evaluations are conducted using the NuScenes-QA dataset.

d) EM-VLM4AD: [73] Is an efficient vision-language model for VQA in AVs, integrating traffic scene images with the T5 language model. It generates accurate answers to safety-related questions while minimising memory and computational needs, demonstrating superior efficiency.

Summary on VQA: To conclude, the integration of L(V)LMs into AV VQA systems has improved the ability to understand and respond to natural language queries about driving environments. This enhancement facilitates better transparency, debugging, and refinement of autonomous systems. The reviewed frameworks demonstrate various applications of VQA, such as the LingoQA benchmark for evaluating vision-language models, DriveLM's graph-based VQA, and LiDAR-LLM's 3D data captioning and question answering. Evaluation metrics like BLEU, METEOR, and CIDEr are essential for assessing the quality of VQA systems, ensuring they provide accurate and contextually relevant responses. Continued advancements in VQA will further enhance the interpretability and reliability of autonomous driving systems.

F. Others

This section highlights innovative applications of AV L(V)LMs that do not fit into traditional categories like Perception or VQA.

a) TrafficGPT: [77] Integrates ChatGPT with the Traffic Fundamentals Model to enhance traffic data analysis and task decomposition. This framework allows interactive feedback and progressive task completion, demonstrating LLMs' potential in complex traffic management.

b) R2T-LLM: [78] Leverages LLMs for traffic flow prediction by transforming multimodal traffic data into natural language descriptions. It captures complex spatiotemporal patterns and external factors, providing accurate predictions and intuitive interpretability.

c) AgentsCoDriver: [79] Introduces a modular architecture for vehicle collaboration, featuring components like Observation, Reasoning Engine, Memory, Iterative Reinforcement Reflection, and Communication Protocol. It outperforms existing systems like DiLu [38] in collaborative vehicle operations.

TABLE V
TABLE FOR VISUAL QUESTION ANSWERING PAPERS

Paper	Submission Time	Specific Tasks	LLM Model	Dataset or Simulator
LingoQA [71]	Dec 2023	Video Question Answering, Evaluating Vision-Language Models	Vicuna v1.5 7B [74]	LingoQA [71]
DriveLM [8]	Dec 2023	Graph-based Visual Question Answering, Planning	BLIP-2 [4]	DriveLM-Data [8]
LiDAR-LLM [72]	Dec 2023	3D Captioning & Grounding, 3D Question Answering, Zero-Shot Planning	LLaMA2-7B [16]	nuScenes [47] , NuScenes-QA [75]
EM-VLM4AD [73]	Mar 2024	Visual Question Answering for ADS Safety	T5 LM [76]	DriveLM-Data [8]

d) DriveCmd: [80] Explores using LLMs to interpret user commands in AV. By leveraging LLMs' reasoning capabilities, it improves the understanding and response to in-cabin commands, enhancing human-vehicle interaction.

e) BALD: BALD [81] A framework for backdoor attacks against LLM-enabled decision-making systems, exploring vulnerabilities during fine-tuning. It proposes word injection, scenario manipulation, and knowledge injection as attack mechanisms. Experiments demonstrate the effectiveness of these attacks, highlighting security risks in L(V)LM-based decision-making systems.

V. END-TO-END INTEGRATIONS

AV End-to-End systems represent an integrated approach where a single model processes sensory data and directly outputs driving decisions [82]. Unlike traditional modular architectures that decompose the driving task into separate components such as perception, planning, and control, end-to-end systems streamline these processes into a unified framework. This holistic approach is designed to enhance the execution of driving tasks, providing a more adaptive and cohesive response to the dynamic environments encountered on the road [22] (depicted in Fig. 6).

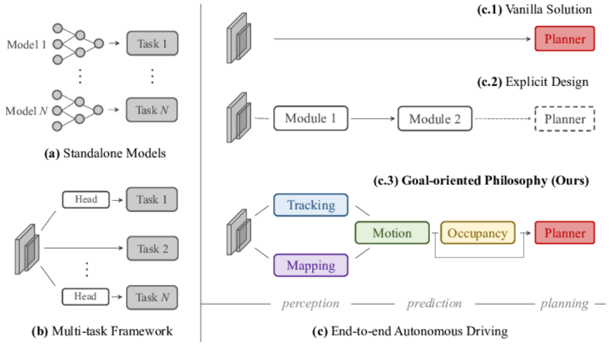


Fig. 6. Comparison of various algorithm framework designs: Standalone Models, Multi-task Frameworks, and End-to-End AVs. Adapted from [22].

The rationale behind adopting end-to-end systems lies in their potential to reduce the complexity and error propagation inherent in modular systems. By learning directly from data, end-to-end models can optimise the entire driving process holistically, rather than optimising individual components in isolation. This can lead to more robust performance, as the system learns to handle a wide range of scenarios through continuous learning and adaptation.

Moreover, end-to-end systems facilitate a more straightforward training pipeline. Instead of requiring extensive engineering to integrate and fine-tune different modules, a single model can be trained to learn the driving task comprehensively. This approach not only simplifies the development process but also enables the system to better generalise from training data to real-world driving situations.

a) DriveLikeAHuman: [83] Explores the potential of LLMs to emulate human driving behaviour, addressing the limitations of traditional optimisation-based and modular AV systems in handling long-tail corner cases. The study leverages GPT-3.5 to build a closed-loop system capable of interpreting complex driving environments, reasoning about potential actions, and remembering past experiences to improve decision-making.

b) HiLM-D: [84] Presents an innovative approach that, for the first time, leverages singular multimodal large language models (MLLMs) to consolidate multiple AV tasks from videos. The HiLM-D system focuses on Risk Object Localisation and Intention and Suggestion Prediction (ROLISP), using a dual-branch architecture. The low-resolution reasoning branch processes videos to identify and describe risk objects, while the high-resolution perception branch enhances detection accuracy by generating detailed feature maps that highlight potential risks.

c) DriveGPT4: [86] Similar to HiLM-D, this model introduces an interpretable End-to-End AV system that uses MLLMs. This system integrates video processing and textual queries to facilitate the interpretation of vehicle actions, provide reasoning, and address user questions effectively. DriveGPT4 predicts low-level vehicle control signals in an end-to-end manner, utilising a bespoke visual instruction tuning dataset specifically tailored for AV applications.

d) On the Road with GPT-4V(ision): [83] Explores the use of the GPT-4V(ision) model in AV, presenting an innovative approach to building a system that drives like a human. The study addresses the limitations of previous optimisation-based AV systems in dealing with long-tail corner cases, which often result from catastrophic forgetting of global optimisation. The authors identify three necessary abilities for an AV (AD) system: reasoning, interpretation, and memorisation.

TABLE VI
TABLE FOR END-TO-END AV PAPERS

Paper	Submission Time	Specific Tasks	LLM Model	Dataset or Simulator
DriveLikeAHuman [83]	Jul 2023	Reasoning, Interpretation, Memorization	GPT-3.5 [32]	HighwayEnv [42]
HiLM-D [84]	Sep 2023	Risk Object Localization, Intention and Suggestion Prediction (ROLISP)	Custom MLLM	DRAMA [85], DRAMA-ROLISP [84]
DriveGPT4 [86]	Oct 2023	Interpretation, Open-loop Control Signal Prediction	GPT-4 [1]	BDD-X [35]
On the Road with GPT-4V(ision) [87]	Nov 2023	Reasoning, Interpretation, Act-as-a-Driver	GPT-4V [88]	nuScenes [47], Waymo Open dataset [89], BDD-X [35], CODA [90], D2-city [91], CCD [92], TSDD [93], ADD [94], DAIR-V2X [95], CitySim [96], CARLA [33]
LMDrive [97]	Dec 2023	Closed-loop Driving Performance under Language Instructions	LLaVA-v1.5, Vicuna-v1.5	CARLA [33]
DriveMLM [98]	Dec 2023	Behavioral Planning, Decision-Making	LLaMA2-7B [16] EVA-CLIP [99]	CARLA [33]
DriveVLM [100]	Mar 2024	Interpretation, Planning, Chain-of-Thought Process Integration	Qwen-VL [17]	SUP-AD [100], nuScenes [47]
OmniDrive [101]	May 2024	Scene Description, Scene Analysis, Meta Actions, Decision Description, Trajectory Waypoints	Q-Former3D [4]	OmniDrive-nuScenes [101], NuScenes-QA [75]
Co-driver [102]	May 2024	Adjustable Driving Behaviours, Interpretation	Qwen-VL [17]	CARLA [33]

e) LMDrive: [97] Introduces a novel language-guided, end-to-end, closed-loop AV framework. The study aims to overcome the limitations of previous AV systems by integrating LLMs with multi-modal sensor data to enable human-like interaction and advanced reasoning in complex driving scenarios. LMDrive processes and integrates multi-modal sensor data with natural language instructions, facilitating effective communication with humans and navigation software.

f) DriveMLM: [98] An LLM-based framework designed for AV that bridges the gap between linguistic decision outputs and actionable vehicle control signals. This framework aligns the decision states with the behavioural planning module of the Apollo system, allowing for effective closed-loop driving in realistic simulators like CARLA. The DriveMLM model leverages a multi-modal tokenizer and a multi-modal LLM (MLLM) decoder to process inputs such as multi-view images, LiDAR point clouds, traffic rules, and user instructions.

g) DriveVLM: [100] An AV system that leverages Vision-Language Models (VLMs) for enhanced scene understanding and planning capabilities. The system integrates a Chain-of-Thought (CoT) process with modules for scene description, scene analysis, and hierarchical planning. This approach allows DriveVLM to linguistically depict driving environments, analyse critical objects, and formulate step-by-step plans, addressing challenges in object perception, intention-level prediction, and task-level planning.

h) OmniDrive: [101] Presents a holistic framework for end-to-end AV by utilising LLM agents capable of 3D perception, reasoning, and planning. The framework introduces OmniDrive-Agent, a novel 3D vision-language model that employs sparse queries to convert visual data into 3D representations before processing with an LLM. This setup allows for the joint encoding of dynamic objects and static map elements, creating a comprehensive world model essential for effective decision-making in complex 3D environments. Additionally, OmniDrive-nuScenes, a new benchmark, includes tasks such as scene description, traffic regulation adherence, 3D grounding, counterfactual reasoning, and planning, all designed to evaluate the model's 3D spatial reasoning and planning capabilities.

i) Co-driver: [102] An innovative AV assistant system designed to provide human-like behaviour and understanding in complex road scenes using Visual Language Models (VLMs). The system utilises the CARLA simulator and ROS2 to validate its effectiveness, operating on a single Nvidia 4090 24G GPU. It combines the capabilities of VLMs to offer adjustable driving behaviours based on visual perception inputs. The authors contribute a new dataset containing images and corresponding prompts for fine-tuning the VLM module.

Summary on End-to-End Systems: AV End-to-End systems represent a paradigm shift from traditional modular architectures by integrating perception, planning, and control into a single, cohesive model. These systems, leveraging L(V)LMs, aim to enhance decision-making, reasoning, and adaptability in complex driving scenarios. Papers like DriveLikeAHuman and HiLM-D explore the use of LLMs to emulate human-like behaviour, addressing long-tail corner cases and improving risk object localization and intention prediction. DriveGPT4 and LMDrive further push the boundaries by incorporating interpretable models that predict vehicle control signals and follow natural language instructions in real time.

The trend toward end-to-end systems highlights their potential to reduce error propagation inherent in modular approaches, offering more robust and adaptive performance. These systems simplify the training pipeline, allowing a single model to learn from data holistically, which is crucial for handling diverse and unpredictable driving environments. However, the challenge remains in balancing the **interpretability** and **scalability** of these models with their computational demands.

VI. Data Generation

Data generation for AV simulation-based testing encompasses methodologies that create realistic, diverse, and safety-critical scenarios to test vehicles' operational responses under various road layouts, traffic configurations, and environmental conditions. The success and versatility of L(V)LMs naturally dovetail into their incorporation with existing data generation frameworks. This is due, for the most part, to their implicit object-relation ontologies garnered through an internet-scale corpus of natural text and their aptitude in parsing abstract scenario relations into readable formats.

For this reason, this section delineates two broad framework types: *LLM-Assisted Simulation and Scenario Generation*, and *World Models*. The former refers to any framework that incorporates an L(V)LM that does not align with the World Model pipeline.

A. LLM-Assisted Simulation and Scenario Generation

LLM-assisted simulation and scenario generation frameworks leverage the capabilities of LLMs to create realistic and diverse driving scenarios for testing and validating autonomous vehicles. These frameworks use LLMs to interpret natural language commands, generate complex driving scenarios, and provide interactive feedback for dynamic adjustments. The success and versatility of LLMs in parsing abstract scenario relations into JSON-readable formats (or other specifications) make them highly suitable for this task. This approach enhances the realism and variety of simulated environments, making them more effective for evaluating AV systems' responses under various road layouts, traffic configurations, and environmental conditions.

a) CTG++: CTG++ [103] Addresses multi-agent consistency in traffic scene generation using a scene-level conditional diffusion model and a spatial-temporal transformer to model all traffic participants' trajectories. It operates on past and future trajectories, using GPT-4 to translate user queries into loss functions for traffic-compliant simulations, outperforming baselines in generating realistic scenarios.

b) Domain Knowledge Distillation from Large Language Model: This framework [104] uses ChatGPT to automate the construction of domain knowledge ontologies for AVs. It extracts and organises knowledge through prompt engineering and iterative LLM interactions, improving the robustness and scalability of ontology construction for testing and validating AVs, surpassing traditional manual methods.

c) SurrealDriver: SurrealDriver [106] Utilises GPT-4 to simulate generative driver agents in urban contexts, following design guidelines for scene understanding, safety, short-term memory, and long-term driving rules. The framework's CoachAgent guides DriverAgent to mimic human decision-making, reducing collision rates and increasing human-likeness in practical simulations.

d) LimSim++: LimSim++ [107] Integrates scenario information from SUMO and visual content from CARLA, using multimodal prompts for MLLMs to perform driving tasks. It features continuous learning with evaluation, reflection, and memory modules, supporting various driving scenarios with high completion rates and driving scores.

e) ChatSim: ChatSim [109] Enables editable, photo-realistic 3D driving scene simulations via natural language commands. It uses a collaborative LLM-agent framework and employs McNeRF for scene rendering and McLight for realistic lighting. Experiments on the Waymo dataset show its ability to generate realistic, customised driving scenarios.

f) TrafficGPT: TrafficGPT [77] Leverages ChatGPT and specialised Traffic Foundation Models (TFMs) for urban traffic management. It analyses live traffic feeds to identify congestion and suggests alternative routes, using historical data to predict peak times. The framework supports task decomposition and interactive feedback for dynamic adjustments.

g) SeGPT: SeGPT [110] Integrates foundation models, vehicle operating systems, and advanced infrastructure to generate diverse scenarios with human-like driving guidelines. It significantly improves trajectory prediction models' performance under challenging conditions through realistic scenario generation.

h) CRITICAL: CRITICAL [112] A closed-loop framework that enhances AV safety resilience by targeting RL agent learning gaps with real-world traffic dynamics, scenario generation, and LLM analysis. Using Proximal Policy Optimisation (PPO) and HighwayEnv, it demonstrates noticeable performance improvements in AV safety and training.

i) ChatScene: ChatScene [114] Uses LLMs to generate safety-critical scenarios for AVs. It transforms unstructured language instructions into detailed traffic scenarios, breaking them down into sub-scenarios for simulators like CARLA. This approach enhances AV safety and reliability by creating diverse and complex scenarios.

Summary on LLM-Assisted Simulation and Scenario Generation: The integration of L(V)LMs into AV simulation frameworks has elevated the sophistication and realism of testing environments. These models, particularly advanced versions like GPT-4 and LLaMA, enable the creation of diverse, human-like driving scenarios and real-time adjustments that were previously difficult to achieve. This shift allows for more nuanced testing, improving AV systems' ability to respond to

TABLE VII
TABLE FOR LLM-ASSISTED SIMULATION AND SCENARIO GENERATION

Paper	Submission Time	Specific Tasks	LLM Model	Dataset or Simulator
CTG++ [103]	Jun, 2023	Queryable Traffic Compliant Simulations, Fine-Grained Control, Trajectory Modelling	GPT-4 [1]	nuScenes [47]
Domain Knowledge Distillation from Large Language Model [104]	Jul, 2023	Domain Knowledge Distillation	GPT-3.5 [1]	OpenXOntology [105]
SurrealDriver [106]	Sep, 2023	Simulating Human-like Behaviours, Learning from Experts, Adhering to Long-Term Safety Guidelines	GPT-4 [1]	CARLA [33]
LimSim++ [107]	Feb, 2024	Interpretation, Decision-Making, Framework Enhancement	GPT-3.5 [1], GPT-4 [1], GPT-4V [88]	SUMO [108], CARLA [33]
ChatSim [109]	Feb, 2024	Editable Photo-Realistic 3D Driving Scene Simulations via Natural Language Commands	GPT-4 [1]	Waymo Open Dataset [89]
TrafficGPT [77]	Mar, 2024	Closed-loop Reasoning, Decision-Making Support, Task Decomposition, Feedback Adaption	GPT-3.5 Turbo [1]	SUMO [108]
SeGPT [110]	Mar, 2024	Interpretation, Framework Enhancement	GPT-4 [1]	Interaction Dataset [111]
CRITICAL [112]	Apr, 2024	Critical Scenario Generation, Closed-loop Training and Performance Augmentation, Safety Analysis	Mistral-7B-Instruct [113]	HighwayEnv [42]
ChatScene [114]	May, 2024	Safety-Critical Scenario Generation	GPT-3.5 [1]	CARLA [33]

complex, real-world conditions. Moreover, L(V)LMs enhance the adaptability and safety of AVs by enabling dynamic scenario customisation and targeted safety testing, reflecting a broader trend toward making AV simulations more aligned with real-world demands and fidelities.

B. World Models

World Models represent an advanced approach to data generation for autonomous vehicle simulation, primarily operating on the latent diffusion model (LDM) pipeline. The LDM pipeline starts with an autoregressive transformer that predicts high-level scene components and dynamics by mapping inputs (video frames, text, and actions) to a shared latent space and temporally sequencing this representation. A video diffusion decoder then translates these latent representations into high-quality, realistic video frames with temporal consistency. The autoregressive transformer facilitates this multi-modal integration, allowing for text and action conditioning. World Models can generate coherent scenes with realistic object interactions and placements, showcasing contextual awareness of underlying road and world rules. They can also produce novel and diverse images and/or videos with prolonged temporal consistency, extending beyond specific training instances. These capabilities make World Models ideal for testing AV systems in a wide range of scenarios.

a) Gaia-1: Researchers at Wayve developed GAIA-1 [27], which operates on a LDM pipeline. The LDM starts with an autoregressive transformer that maps inputs to a shared latent space and sequences them temporally. A video diffusion decoder decodes these representations into consistent video frames. This transformer also provides fine-grained control over the outputs (such as vehicle dynamics and scene features) while the diffusion decoder addresses the common issue of temporal inconsistency in video generation. GAIA-1 supports text and action conditioning, enabling it to generate realistic scenarios based on specific instructions.

b) Magicdrive: Magicdrive [116] Generates high-fidelity street-view images and videos with 3D geometry control, creating training datasets to enhance perception tasks like BEV segmentation and 3D object detection. It provides multi-level scene control, adjusting attributes like foreground object orientations and background layouts while maintaining multi-camera consistency. The pipeline encodes and integrates geometric conditions (3D bounding boxes, road maps) and shares information across multiple camera views using cross-attention and additive encoder branches.

c) Drive-WM: Drive-WM [118] Predicts future events in AV scenarios, integrating with existing end-to-end AV planning models. It uses temporal and multi-view encoding layers to process image sequences across multiple views, incorporating images, text, 3D layouts, and actions for flexible video generation. Drive-WM supports safe planning by evaluating future trajectories with image-based reward functions, improving robustness in out-of-distribution situations for zero-shot path planning.

d) GenAD: GenAD [9] A generalised video prediction model for AV. Utilising the OpenDV-2K dataset, it employs temporal reasoning blocks with causal temporal attention and decoupled spatial attention to model dynamic interactions. GenAD supports action-conditioned prediction and planning, outperforming non-LDM models in video prediction quality and zero-shot generalisation, reducing prediction errors and

TABLE VIII
TABLE FOR WORLD MODEL PAPERS

Paper	Submission Time	Specific Tasks	LLM Model	Dataset or Simulator
Gaia-1 [27]	Sep, 2023	Text & Action Conditioning, Contextual Awareness (Road & World Rules), Scene & Dynamics Control	Encoded Text via T5-large Model [115]	Proprietary Driving Data (London, UK between 2019 and 2023) ≈ 4,700 hours at 25Hz.
Magicdrive [116]	Oct, 2023	BEV Segmentation, 3D Object Detection, Multi-Camera Consistency	Pre-trained CLIP Text Encoder [117]	nuScenes [47]
Drive-WM [118]	Nov, 2023	Flexible Integration with Existing End-to-End AV Planning Models	Pre-trained CLIP Text Encoder [117]	nuScenes [47]
GenAD [9]	Feb, 2024	Action-Conditioned Prediction & Planning, Zero-Shot Generalisation	Pre-trained CLIP Text Encoder [117]	nuScenes [47], OpenDV-2K Dataset [9]
ADriver-I [119]	Nov, 2023	Vision-Action Pairs unifying Visual Features and Control Signals, Adaptive Generation	Vicuna-7B-1.5 [74]	nuScenes [47]
DriveDreamer-2 [28]	Apr, 2024	Action-Conditioned Prediction & Planning, Zero-Shot Generalisation	Finetuned GPT-3.5 [1]	nuScenes [47]

enhancing simulation consistency.

e) ADriver-I: ADriver-I [119] is a Multi-modal Large Language Model (MLLM) that unifies visual features and control signals to construct a general world- and diffusion model. It integrates perception, prediction, planning, and control into a cohesive system processing vision-action pairs to predict and generate future driving scenarios. Using Stable Diffusion 2.1, it generates future video frames based on predicted control signals and historical frames. Trained on the nuScenes dataset and a large private dataset, ADriver-I adapts to various driving conditions without extensive prior information.

f) DriveDreamer-2: DriveDreamer-2 [28] Generates high-quality driving videos and realistic driving policies through a two-stage training process. The first stage processes road structural features and traffic metadata to contextualise real-world scenes. The second stage, ActionFormer, couples sequential HDMaps and 3D bounding boxes to encoded driving actions, which are then input to a diffusion model. DriveDreamer-2 adds a finetuned GPT-3.5 to convert text prompts to agent trajectories for diverse, multi-view driving simulations. Both models are trained on the nuScenes dataset.

Summary on World Models World Models utilise L(V)LMs (frequently LLMs) in conjunction with LDM to enhance autonomous vehicle simulations. Key models like Gaia-1 and Magicdrive focus on generating high-quality, temporally consistent video frames and enhancing perception tasks with 3D geometry control, respectively. Drive-WM supports safe planning through multi-view inputs, while CTG++ excels in generating multi-agent consistent traffic scenarios. *World Models* have the potential to upturn current data generation schemas by offering continual realistic data generation throughout the typical AV stack, and are expected to be intimately linked (in a closed-loop fashion) with AV evaluation frameworks. However, it is also worth noting that these models require substantial computational resources.

VII. PLATFORMS, BENCHMARKS, AND DATASETS

Autonomous vehicle research relies heavily on both simulation platforms and comprehensive datasets to develop, test, and validate various components of driving systems. Simulation platforms provide controlled environments to test algorithms under diverse conditions, while datasets offer real-world data essential for training and benchmarking these systems. For a more detailed exploration, this section is divided into two subsections: *Platforms*, *Benchmarks and Datasets*. The former discusses key simulation environments, and the latter reviews significant datasets critical to advancing autonomous vehicle technology.

A. Platforms

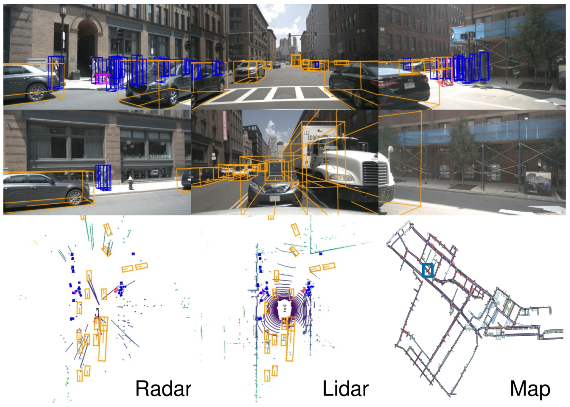
Simulation platforms play a crucial role in autonomous vehicle research. Two notable platforms are CARLA [33], and HighwayEnv [42]. They are integral to advancing the field of AV by providing robust environments for comprehensive testing and development.

a) CARLA: CARLA [33] An open-source simulator designed for developing, training, and validating autonomous urban driving systems. It offers a highly configurable and realistic environment with diverse conditions, traffic scenarios, and sensor setups like cameras, LiDAR, and radar. CARLA's realistic urban layouts and dynamic elements make it ideal for testing algorithms in perception, planning, and control.

b) HighwayEnv: HighwayEnv [42] A lightweight simulation environment tailored for reinforcement learning algorithms in highway driving scenarios. It simulates realistic traffic conditions and vehicle dynamics, supporting tasks such as lane keeping, lane changing, and car-following. HighwayEnv's simplicity and efficiency facilitate rapid experimentation and iterative development.

B. Benchmarks and Datasets

Datasets are crucial for the training and benchmarking of AV systems. Prominent datasets like the Waymo Open Dataset and NuScenes [47] (shown in Fig. 7) provide extensive multi-sensor data, including high-resolution images, lidar point clouds, and detailed annotations. These datasets have been instrumental in advancing algorithms for object detection, tracking, and prediction in real-world scenarios. The NuScenes dataset, in particular, supports multi-modal sensor fusion research and includes a robust benchmark and leaderboard, fostering a competitive environment for researchers.



"Ped with pet, bicycle, car makes a u-turn, lane change, peds crossing crosswalk"

Fig. 7. Examples of the NuScenes Dataset. Adapted from [47].

In addition to these foundational datasets, newer ones like NuScenes-QA [75] and OmniDrive-nuScenes [101] expand the scope to cognitive tasks, such as visual question answering and counterfactual reasoning. Benchmarks like BDD-X [35], which focuses on interpretability and reasoning, and DRAMA [85], which targets risk perception and explainability, highlight the diverse challenges in AV. Future dataset development should integrate diverse data sources, capture dynamic and real-time scenarios, and include human factors and ethical considerations to support the advancement of robust and intelligent AV systems.

a) BDD-X: BDD-X [35] A benchmark dataset for evaluating the interpretability and reasoning capabilities of AV systems, featuring diverse scenarios with textual descriptions, questions, and answers to enhance systems' understanding and explanations.

b) NuScenes: NuScenes [47] Is a comprehensive dataset for AV, providing real-world driving data with multi-sensor information, aiding in the development and testing of perception and decision-making systems in various scenarios.

c) DRAMA: DRAMA [85] Focuses on risk perception and explainability in driving scenes, featuring 17,785 interactive scenarios with video-level and object-level questions about driving risks and important objects, enhancing situational awareness.

d) NuScenes-QA: NuScenes-QA [75] Extends the NuScenes dataset with a question-answering framework to evaluate an agent's understanding and reasoning in driving contexts, offering visual question-answering tasks for detailed environmental queries.

e) NuPrompt: NuPrompt [120] Integrates language prompts into driving scenarios within 3D environments. Built on NuScenes, it includes 35,367 descriptions matched with objects, supporting cross-modal understanding and object-tracking tasks.

f) Driving QA: Driving QA [121] Consisting of 160k QA pairs derived from 10k driving scenarios. These scenarios are collected using a 2D simulator and an RL agent, providing a robust dataset for training the model in interpreting scenarios, answering questions, and making decisions.

g) Talk2BEV-Bench: Talk2BEV-Bench evaluates large vision-language models for AV using 1000 annotated bird's-eye view (BEV) scenes and over 20,000 question-answer pairs. It assesses competencies like instance attributes, counting, and spatial reasoning, using metrics like accuracy and regression.

h) Reason2Drive: Reason2Drive [122] Offers over 600,000 video-text pairs for interpretable reasoning in AV, emphasising sequential perception, prediction, and reasoning with data from nuScenes, Waymo, and ONCE.

i) LaMPilot: LaMPilot [123] Integrates LLMs into AV for interpreting user commands via code generation. The LaMPilot-Bench evaluates LLM-based agents across various driving scenarios for adaptability and accuracy.

j) LangAuto: LangAuto [97] Benchmarks AV systems' ability to follow natural language instructions with dynamically updated navigation commands, testing understanding and execution in dynamic environments.

k) LingoQA: LingoQA [71] Provides a benchmark for video question answering in AV with 28K scenarios and 419K QA pairs, highlighting performance gaps between humans and models. It includes Lingo-Judge for accurate evaluation.

l) SUP-AD: SUP-AD [100] A comprehensive dataset designed to evaluate AV' capabilities in scene understanding and meta-action planning within complex driving scenarios. It includes 1,000 video clips across more than 40 different driving scenario categories, with detailed annotations covering scene descriptions.

m) CODA-LM: CODA-LM [124] The first benchmark designed for the automated and systematic evaluation of Large Vision-Language Models (LVLMs) on self-driving corner cases. Based on the CODA dataset, CODA-LM comprises 9,768 real-world driving scenarios with extensive annotations, including 41,722 textual annotations for critical road entities and 21,537 annotations specifically for corner cases.

n) OmniDrive-nuScenes: OmniDrive-nuScenes [101] Enhances NuScenes with 3D spatial understanding and counterfactual reasoning tasks, assessing autonomous systems' decision-making and planning abilities through simulated trajectories and outcomes.

Summery on Platforms, Benchmarks and Datasets Simulation platforms like CARLA and datasets such as NuScenes and Waymo Open Dataset are foundational for advancing autonomous vehicle (AV) technologies, offering realistic scenarios and multi-sensor data for tasks like 3D detection and tracking. Specialised benchmarks like NuScenes-QA, DRAMA, and BDD-X expand into cognitive tasks, risk perception, and interpretability. However, there's a growing need for more comprehensive resources, such as the emerging CODA-LM, that address a broader range of tasks, integrate complex scenarios, and consider human factors.

TABLE IX
TABLE FOR BENCHMARKS AND DATASETS

Dataset Name	Submission Time	Specific Tasks	Data Format	Datasize
BDD-X [35]	Jul, 2018	Introspective and rationalization explanations for vehicle behavior	Video, text	6,984 videos 77 hours 26,228 annotations
NuScenes [47]	Mar, 2019	3D detection and tracking, multi-modal sensor fusion	Images, LiDAR, radar	1,000 scenes 1.4M images 1.3M LiDAR sweeps 1.4M annotations
DRAMA [85]	Sep, 2022	Joint risk localization and captioning in driving scenes	Video, text	17,785 scenarios 91 hours 35,038 visual attributes
NuScenes-QA [75]	May, 2023	Visual question answering for AV	Images, LiDAR, Q&A pairs	34K scenes 460K Q&A pairs
NuPrompt [120]	Sep, 2023	Object-centric language prompts for 3D driving scenes	Images, text	35,367 prompts 188,445 instances
Driving QA Dataset [121]	Oct, 2023	Question answering for driving scenarios, control commands	Text, control commands	160K Q&A pairs 10K driving scenarios
Talk2BEV-Bench [31]	Oct, 2023	Evaluates LVLMs for instance attributes, counting, spatial reasoning	BEV images, Q&A pairs	1,000 scenes 20,000 Q&A pairs
Reason2Drive [122]	Dec, 2023	Chain-based reasoning for AV	Video, text	600K video-text pairs
LaMPilot [123]	Dec, 2023	User instruction following, code generation	Semi-human annotated traffic scenes	4,900 scenarios
LangAuto [97]	Dec, 2023	Closed-loop driving, language instruction following	Multi-modal sensor data, navigation and notice instructions	64K data clips
LingoQA [71]	Dec, 2023	Video question answering for AV	Video, Q&A pairs	28K scenarios 419.9K Q&A pairs
SUP-AD [100]	Feb, 2024	Scene understanding, meta-action planning, hierarchical planning	Multi-view videos, 3D perception, scene descriptions, meta-actions, decision descriptions, waypoints	1,000 video clips
CODA-LM [124]	Apr, 2024	Automated evaluation of LVLMs on self-driving corner cases	Real-world driving scenarios, textual annotations	9,768 scenarios 63,259 textual annotations
OmniDrive-nuScenes [101]	May, 2024	Perception, reasoning, and planning in 3D domain with QA pairs	Q&A pairs, text	341,490 conversations 34,149 descriptions 34,149 keywords 135,948 planning tasks

VIII. CHALLENGES AND FUTURE RESEARCH DIRECTIONS

The integration of Large Language Models (LLMs) and Large Vision-Language Models (LVLMs) into autonomous vehicles (AVs) has significantly advanced adaptive decision-making, trajectory prediction, traffic accident forecasting, human-vehicle interaction, and targeted high-fidelity data generation. L(V)LMs have fundamentally shifted how research institutions and industry approach the dynamic driving task (DDT) problem, regardless of the integration approach—be it modular integration, end-to-end integration, data generation methods, or the choice of specific platforms and datasets (as outlined and explored throughout the paper). However, several challenges remain to be addressed before regulatory and authorisation agencies accept these technologies for commercial or civilian applications.

A. Hallucinations

Hallucinations refer to the generation of outputs by L(V)LMs that are factually incorrect or nonsensical. This issue is particularly critical in AVs, where accurate and reliable information is paramount for safety. Current L(V)LMs can occasionally produce hallucinations when interpreting complex driving scenarios or making decisions based on incomplete or ambiguous data. For example, an LVLM might misinterpret a visual input, leading to incorrect obstacle detection. Future research should focus on developing methods to detect and mitigate hallucinations, such as incorporating stronger verification mechanisms, enhancing training data quality, and leveraging multi-modal inputs to cross-validate information.

B. Latency

Real-time decision-making is crucial for the safety and efficiency of autonomous driving. However, the computational complexity of L(V)LMs can introduce delays that are unacceptable in dynamic driving environments. High latency can result in delayed responses to critical situations, potentially leading to accidents. Future research should explore optimising model architectures and leveraging hardware accelerations, such as GPUs and TPUs on-premises (within the vehicle), to reduce inference times. Approaches like model compression, quantisation, and pruning can also help in reducing computational overhead. Additionally, hybrid strategies that balance the use of on-board processing and edge/cloud computing could help mitigate latency issues while maintaining performance.

C. Ethical and Regulatory Considerations

Ensuring that these systems make unbiased decisions, especially in life-critical scenarios, is required for regulatory compliance. There are concerns about how L(V)LMs may inadvertently perpetuate biases present in training data, affecting decision-making in scenarios involving pedestrians or other vehicles. Additionally, data privacy and the security of AI systems are paramount, as vulnerabilities could be exploited maliciously. Future research should address these challenges by developing frameworks for ethical AI, implementing bias mitigation strategies, ensuring compliance with regulations, and implementing robust security measures to protect against attacks. Collaboration with regulatory bodies to establish standards and guidelines for the deployment of L(V)LMs in AVs is also essential.

D. Vehicle-to-Everything (V2X) Communication

V2X communication involves the exchange of information between vehicles and various entities like infrastructure, pedestrians, and other vehicles. Integrating L(V)LMs into V2X communication frameworks presents an opportunity to enhance the context-awareness and decision-making capabilities of AVs. For instance, L(V)LMs can interpret complex messages and predict traffic patterns based on shared data. Future research should investigate the integration of L(V)LMs into V2X communication, focusing on improving the reliability, security, and efficiency of these interactions. This includes developing protocols that ensure timely and accurate information exchange while safeguarding against misinformation and malicious interference.

IX. Conclusion

This survey has provided a comprehensive overview of the integration and impact of Large (Vision) Language Models (L(V)LMs) in autonomous vehicles (AVs), focusing on four key areas: modular integration, end-to-end integration, data generation, and platforms and datasets. We have examined the current state of research, highlighting significant advancements and innovations that L(V)LMs bring to various aspects of AV systems, including perception, decision-making, trajectory prediction, and human-vehicle interaction. These models offer substantial benefits in terms of enhanced contextual understanding, improved decision-making processes, and more intuitive human-vehicle interactions.

Despite these advancements, several challenges remain, such as addressing issues of hallucinations, latency, and ethical considerations. The scalability and adaptability of L(V)LMs to different driving environments, as well as their integration into Vehicle-to-Everything (V2X) communication frameworks, are areas that require further research and development. Our survey fills gaps left by previous reviews by providing detailed analyses of practical integration strategies and real-world implementations, offering insights into computational trade-offs and task-specific requirements.

As we move forward, it is crucial to continue exploring these challenges and developing robust solutions that ensure the safe, efficient, and reliable deployment of L(V)LMs in autonomous driving systems. By addressing these issues, we can unlock the full potential of L(V)LMs, leading to a future where autonomous vehicles are not only smarter and more capable but also safer and more trustworthy.

Acknowledgment

The authors would like to thank Zhengyu Wu, Leandro Parada, and Kevin Qu for their dedicated efforts and assistance.

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.

[2] Anthropic, “Claude 3.5,” https://www.anthropic.com/claude.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[4] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[5] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *Advances in neural information processing systems*, vol. 36, 2024.

[6] W. AI, “Lingo 2: driving with language,” https://wayve.ai/thinking/lingo-2-driving-with-language/, 2023, accessed: 2024-08-07.

[7] Y. Ma, Y. Cao, J. Sun, M. Pavone, and C. Xiao, “Dolphins: Multimodal language model for driving,” *arXiv preprint arXiv:2312.00438*, 2023.

[8] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, P. Luo, A. Geiger, and H. Li, “Drivelm: Driving with graph visual question answering,” *arXiv preprint arXiv:2312.14150*, 2023.

[9] J. Yang, S. Gao, Y. Qiu, L. Chen, T. Li, B. Dai, K. Chitta, P. Wu, J. Zeng, P. Luo *et al.*, “Generalized predictive model for autonomous driving,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 662–14 672.

[10] C. Cui, Y. Ma, X. Cao, W. Ye, and Z. 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*, 2024, pp. 902–909.

[11] Z. Yang, X. Jia, H. Li, and J. Yan, “Llm4drive: A survey of large language models for autonomous driving,” *arXiv e-prints*, pp. arXiv–2311, 2023.

[12] H. Gao, Y. Li, K. Long, M. Yang, and Y. Shen, “A survey for foundation models in autonomous driving,” *arXiv preprint arXiv:2402.01105*, 2024.

[13] X. Li, Y. Bai, P. Cai, L. Wen, D. Fu, B. Zhang, X. Yang, X. Cai, T. Ma, J. Guo *et al.*, “Towards knowledge-driven autonomous driving,” *arXiv preprint arXiv:2312.04316*, 2023.

[14] Z. Zhang, Y. Sun, Z. Wang, Y. Nie, X. Ma, P. Sun, and R. Li, “Large language models for mobility in transportation systems: A survey on forecasting tasks,” *arXiv preprint arXiv:2405.02357*, 2024.

[15] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, “A survey of large language models,” *arXiv preprint arXiv:2303.18223*, 2023.

[16] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[17] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," 2023.

[18] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," *Advances in neural information processing systems*, vol. 1, 1988.

[19] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang *et al.*, "End to end learning for self-driving cars," *arXiv preprint arXiv:1604.07316*, 2016.

[20] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 4693–4700.

[21] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning*. PMLR, 2016, pp. 1928–1937.

[22] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 853–17 862.

[23] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Dall·e: Creating images from text," https://github.com/openai/DALL-E, 2021, accessed: 2024-07-09.

[24] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," https://github.com/CompVis/stable-diffusion, 2022, accessed: 2024-07-09.

[25] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *arXiv preprint arXiv:2006.11239*, 2020.

[26] P. Dhariwal and A. Q. Nichol, "Diffusion models beat gans on image synthesis," *arXiv preprint arXiv:2105.05233*, 2021.

[27] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado, "Gaia-1: A generative world model for autonomous driving," *arXiv preprint arXiv:2309.17080*, 2023.

[28] G. Zhao, X. Wang, Z. Zhu, X. Chen, G. Huang, X. Bao, and X. Wang, "Drivedreamer-2: Llm-enhanced world models for diverse driving video generation," *arXiv preprint arXiv:2403.06845*, 2024.

[29] A. Elhafsi, R. Sinha, C. Agia, E. Schmerling, I. A. Nesnas, and M. Pavone, "Semantic anomaly detection with large language models," *Autonomous Robots*, vol. 47, no. 8, pp. 1035–1055, 2023.

[30] F. Romero, C. Winston, J. Hauswald, M. Zaharia, and C. Kozyrakis, "Zelda: Video analytics using vision-language models," *arXiv preprint arXiv:2305.03785*, 2023.

[31] V. Dewangan, T. Choudhary, S. Chandhok, S. Priyadarshan, A. Jain, A. K. Singh, S. Srivastava, K. M. Jatavallabhula, and K. M. Krishna, "Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving," *arXiv preprint arXiv:2310.02251*, 2023.

[32] OpenAI, "ChatGPT 3.5," 2023, https://platform.openai.com/docs/models/gpt-3-5.

[33] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[34] F. Romero, J. Hauswald, A. Partap, D. Kang, M. Zaharia, and C. Kozyrakis, "Optimizing video analytics with declarative model relationships," *Proceedings of the VLDB Endowment*, vol. 16, no. 3, pp. 447–460, 2022.

[35] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 563–578.

[36] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[37] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=vvoWPYqZJA

[38] L. Wen, D. Fu, X. Li, X. Cai, T. Ma, P. Cai, M. Dou, B. Shi, L. He, and Y. Qiao, "Dilu: A knowledge-driven approach to autonomous driving with large language models," *arXiv preprint arXiv:2309.16292*, 2023.

[39] J. Liu, P. Hang, X. Qi, J. Wang, and J. Sun, "Mtd-gpt: A multi-task decision-making gpt model for autonomous driving at unsignalized intersections," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023, pp. 5154–5161.

[40] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

[41] Y. Wang, R. Jiao, C. Lang, S. S. Zhan, C. Huang, Z. Wang, Z. Yang, and Q. Zhu, "Empowering autonomous driving with large language models: A safety perspective," *arXiv preprint arXiv:2312.00812*, 2023.

[42] E. Leurent, "An environment for autonomous driving decision-making," https://github.com/eleurent/highway-env, 2018.

[43] C. Cui, Y. Ma, X. Cao, W. Ye, and Z. Wang, "Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles," *IEEE Intelligent Transportation Systems Magazine*, 2024.

[44] H. Sha, Y. Mu, Y. Jiang, L. Chen, C. Xu, P. Luo, S. E. Li, M. Tomizuka, W. Zhan, and M. Ding, "Languagempc: Large language models as decision makers for autonomous driving," *arXiv preprint arXiv:2310.03026*, 2023.

[45] Y. Liu, Q. Zhang, and D. Zhao, "A reinforcement learning benchmark for autonomous driving in intersection scenarios," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2021, pp. 1–8.

[46] J. Mao, J. Ye, Y. Qian, M. Pavone, and Y. Wang, "A language agent for autonomous driving," *arXiv preprint arXiv:2311.10813*, 2023.

[47] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[48] I. de Zarzà, J. de Curtò, G. Roig, and C. T. Calafate, "Llm multimodal traffic accident forecasting," *Sensors*, vol. 23, no. 22, p. 9225, 2023.

[49] H. H. Team, "Zephyr 7b alpha," https://huggingface.co/HuggingFaceH4/zephyr-7b-alpha, 2024.

[50] National Highway Traffic Safety Administration, "Fatality Analysis Reporting System (FARS) 2020," 2020, u.S. Department of Transportation. Accessed: [Month Day, Year]. [Online]. Available: https://static.nhtsa.gov/nhtsa/downloads/FARS/2020/National/FARS2020NationalCSV.zip

[51] L. Chen, O. Sinavski, J. Hünermann, A. Karnsund, A. J. Willmott, D. Birch, D. Maund, and J. Shotton, "Driving with llms: Fusing object-level vector modality for explainable autonomous driving," *arXiv preprint arXiv:2310.01957*, 2023.

[52] H. Liao, H. Shen, Z. Li, C. Wang, G. Li, Y. Bie, and C. Xu, "Gpt-4 enhanced multimodal grounding for autonomous driving: Leveraging cross-modal attention with large language models," *Communications in Transportation Research*, vol. 4, p. 100116, 2024.

[53] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[54] T. Deruyttere, S. Vandenhende, D. Grujicic, L. Van Gool, and M.-F. Moens, "Talk2car: Taking control of your self-driving car," *arXiv preprint arXiv:1909.10838*, 2019.

[55] S. Sharan, F. Pittaluga, M. Chandraker *et al.*, "Llm-assist: Enhancing closed-loop planning with language-based reasoning," *arXiv preprint arXiv:2401.00125*, 2023.

[56] K. T. e. a. H. Caesar, J. Kabzan, "Nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles," in *CVPR ADP3 workshop*, 2021.

[57] K. Tanahashi, Y. Inoue, Y. Yamaguchi, H. Yaginuma, D. Shiotsuka, H. Shimatani, K. Iwamasa, Y. Inoue, T. Yamaguchi, K. Igari *et al.*, "Evaluation of large language models for decision making in autonomous driving," *arXiv preprint arXiv:2312.06351*, 2023.

[58] M. Azarafza, M. Nayyeri, C. Steinmetz, S. Staab, and A. Rettberg, "Hybrid reasoning based on large language models for autonomous car driving," *arXiv preprint arXiv:2402.13602*, 2024.

[59] F. Chi, Y. Wang, P. Nasiopoulos, and V. C. Leung, "Multi-modal gpt-4 aided action planning and reasoning for self-driving vehicles," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 7325–7329.

[60] J. Mao, Y. Qian, H. Zhao, and Y. Wang, "Gpt-driver: Learning to drive with gpt," *arXiv preprint arXiv:2310.01415*, 2023.

[61] B. Li, Y. Wang, J. Mao, B. Ivanovic, S. Veer, K. Leung, and M. Pavone, "Driving everywhere with large language model policy adaptation," *arXiv preprint arXiv:2402.05932*, 2024.

[62] B. Yang, H. Su, N. Gkanatsios, T.-W. Ke, A. Jain, J. Schneider, and K. Fragkiadaki, "Diffusion-es: Gradient-free planning with diffusion

for autonomous driving and zero-shot instruction following," *arXiv preprint arXiv:2402.06559*, 2024.

[63] M. Peng, X. Guo, X. Chen, M. Zhu, K. Chen, X. Wang, Y. Wang *et al.*, "Lc-llm: Explainable lane-change intention and trajectory predictions with large language models," *arXiv preprint arXiv:2403.18344*, 2024.

[64] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems," in *2018 21st international conference on intelligent transportation systems (ITSC)*. IEEE, 2018, pp. 2118–2125.

[65] S. Wang, Y. Zhu, Z. Li, Y. Wang, L. Li, and Z. He, "Chatgpt as your vehicle co-pilot: An initial attempt," *IEEE Transactions on Intelligent Vehicles*, 2023.

[66] S. Documentation, "Simulation and model-based design," 2020. [Online]. Available: https://www.mathworks.com/products/simulink.html

[67] R. Johansson, D. Williams, A. Berglund, and P. Nugues, "Carsim: a system to visualize written road accident reports as animated 3d scenes," in *Proceedings of the 2nd Workshop on Text Meaning and Interpretation*, 2004, pp. 57–64.

[68] M. N. Team. (2023) Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-05-05. [Online]. Available: www.mosaicml.com/blog/mpt-7b

[69] C. Cui, Z. Yang, Y. Zhou, Y. Ma, J. Lu, and Z. Wang, "Large language models for autonomous driving: Real-world experiments," *arXiv preprint arXiv:2312.09397*, 2023.

[70] S. Kato, E. Takeuchi, Y. Ishiguro, Y. Ninomiya, K. Takeda, and T. Hamada, "An open approach to autonomous vehicles," *IEEE Micro*, vol. 35, no. 6, pp. 60–68, 2015.

[71] A.-M. Marcu, L. Chen, J. Hünermann, A. Karnsund, B. Hanotte, P. Chidananda, S. Nair, V. Badrinarayanan, A. Kendall, J. Shotton *et al.*, "Lingoqa: Video question answering for autonomous driving," *arXiv preprint arXiv:2312.14115*, 2023.

[72] S. Yang, J. Liu, R. Zhang, M. Pan, Z. Guo, X. Li, Z. Chen, P. Gao, Y. Guo, and S. Zhang, "Lidar-llm: Exploring the potential of large language models for 3d lidar understanding," *arXiv preprint arXiv:2312.14074*, 2023.

[73] A. Gopalkrishnan, R. Greer, and M. Trivedi, "Multi-frame, lightweight & efficient vision-language models for question answering in autonomous driving," *arXiv preprint arXiv:2403.19838*, 2024.

[74] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *See https://vicuna. lmsys. org (accessed 14 April 2023)*, vol. 2, no. 3, p. 6, 2023.

[75] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4542–4550.

[76] J. Cho, J. Lei, H. Tan, and M. Bansal, "Unifying vision-and-language tasks via text generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1931–1942.

[77] S. Zhang, D. Fu, W. Liang, Z. Zhang, B. Yu, P. Cai, and B. Yao, "Trafficgpt: Viewing, processing and interacting with traffic foundation models," *Transport Policy*, vol. 150, pp. 95–105, 2024.

[78] X. Guo, Q. Zhang, J. Jiang, M. Peng, H. F. Yang, and M. Zhu, "Towards responsible and reliable traffic flow prediction with large language models," *Available at SSRN 4805901*, 2024.

[79] S. Hu, Z. Fang, Z. Fang, X. Chen, and Y. Fang, "Agentscodriver: Large language model empowered collaborative driving with lifelong learning," *arXiv preprint arXiv:2404.06345*, 2024.

[80] Y. Yang, Q. Zhang, C. Li, D. S. Marta, N. Batool, and J. Folkesson, "Human-centric autonomous systems with llms for user command reasoning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 988–994.

[81] R. Jiao, S. Xie, J. Yue, T. Sato, L. Wang, Y. Wang, Q. A. Chen, and Q. Zhu, "Exploring backdoor attacks against large language model-based decision making," *arXiv preprint arXiv:2405.20774*, 2024.

[82] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," *arXiv preprint arXiv:2306.16927*, 2023.

[83] D. Fu, X. Li, L. Wen, M. Dou, P. Cai, B. Shi, and Y. 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*, 2024, pp. 910–919.

[84] X. Ding, J. Han, H. Xu, W. Zhang, and X. Li, "Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving," *arXiv preprint arXiv:2309.05186*, 2023.

[85] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, and J. Li, "Drama: Joint risk localization and captioning in driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1043–1052.

[86] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K. K. Wong, Z. Li, and H. Zhao, "Drivegpt4: Interpretable end-to-end autonomous driving via large language model," *arXiv preprint arXiv:2310.01412*, 2023.

[87] L. Wen, X. Yang, D. Fu, X. Wang, P. Cai, X. Li, M. Tao, Y. Li, X. Linran, D. Shang *et al.*, "On the road with gpt-4v (ision): Explorations of utilizing visual-language model as autonomous driving agent," in *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

[88] "Gpt-4v(ision) system card," 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:263218031

[89] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[90] K. Li, K. Chen, H. Wang, L. Hong, C. Ye, J. Han, Y. Chen, W. Zhang, C. Xu, D.-Y. Yeung *et al.*, "Coda: A real-world road corner case dataset for object detection in autonomous driving," in *European Conference on Computer Vision*. Springer, 2022, pp. 406–423.

[91] Z. Che, G. Li, T. Li, B. Jiang, X. Shi, X. Zhang, Y. Lu, G. Wu, Y. Liu, and J. Ye, "D^2-city: a large-scale dashcam video dataset of diverse traffic scenarios," *arXiv preprint arXiv:1904.01975*, 2019.

[92] W. Bao, Q. Yu, and Y. Kong, "Uncertainty-based traffic accident anticipation with spatio-temporal relational learning," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2682–2690.

[93] B. J. University, "Chinese traffic sign database," http://www.nlpr.ia.ac.cn/pal/trafficdata/detection.html.

[94] Z. Wu, X. Chen, H. Wei, F. Song, and T. Xu, "Add: An automatic desensitization fisheye dataset for autonomous driving," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106766, 2023.

[95] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan *et al.*, "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 361–21 370.

[96] O. Zheng, M. Abdel-Aty, L. Yue, A. Abdelraouf, Z. Wang, and N. Mahmoud, "Citysim: a drone-based vehicle trajectory dataset for safety-oriented research and digital twins," *Transportation research record*, vol. 2678, no. 4, pp. 606–621, 2024.

[97] H. Shao, Y. Hu, L. Wang, S. L. Waslander, Y. Liu, and H. Li, "Lmdrive: Closed-loop end-to-end driving with large language models," *arXiv preprint arXiv:2312.07488*, 2023.

[98] W. Wang, J. Xie, C. Hu, H. Zou, J. Fan, W. Tong, Y. Wen, S. Wu, H. Deng, Z. Li *et al.*, "Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving," *arXiv preprint arXiv:2312.09245*, 2023.

[99] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva: Exploring the limits of masked visual representation learning at scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 358–19 369.

[100] X. Tian, J. Gu, B. Li, Y. Liu, C. Hu, Y. Wang, K. Zhan, P. Jia, X. Lang, and H. Zhao, "Drivevlm: The convergence of autonomous driving and large vision-language models," *arXiv preprint arXiv:2402.12289*, 2024.

[101] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi, N. Chang, J. Kautz, Y. Li, and J. M. Alvarez, "Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning," *arXiv preprint arXiv:2405.01533*, 2024.

[102] Z. Guo, A. Lykov, Z. Yagudin, M. Konenkov, and D. Tsetserukou, "Co-driver: Vlm-based autonomous driving assistant with human-like behavior and understanding for complex road scenes," *arXiv preprint arXiv:2405.05885*, 2024.

[103] Z. Zhong, D. Rempe, Y. Chen, B. Ivanovic, Y. Cao, D. Xu, M. Pavone, and B. Ray, "Language-guided traffic simulation via scene-level diffusion," in *Conference on Robot Learning*. PMLR, 2023, pp. 144–177.

[104] Y. Tang, A. A. B. Da Costa, X. Zhang, I. Patrick, S. Khastgir, and P. Jennings, "Domain knowledge distillation from large language model: An empirical study in the autonomous driving domain," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023, pp. 3893–3900.

[105] ASAM, "Asam openxontology," 2023. [Online]. Available: https://www.asam.net/standards/asam-openxontology/

[106] Y. Jin, X. Shen, H. Peng, X. Liu, J. Qin, J. Li, J. Xie, P. Gao, G. Zhou, and J. Gong, "Surrealdriver: Designing generative driver agent simulation framework in urban contexts based on large language model," *arXiv preprint arXiv:2309.13193*, 2023.

[107] D. Fu, W. Lei, L. Wen, P. Cai, S. Mao, M. Dou, B. Shi, and Y. Qiao, "Limsim++: A closed-loop platform for deploying multimodal llms in autonomous driving," *arXiv preprint arXiv:2402.01246*, 2024.

[108] D. Krajzewicz, G. Hertkorn, C. Rössel, and P. Wagner, "Sumo (simulation of urban mobility)-an open-source traffic simulation," in *Proceedings of the 4th middle East Symposium on Simulation and Modelling (MESM20002)*, 2002, pp. 183–187.

[109] Y. Wei, Z. Wang, Y. Lu, C. Xu, C. Liu, H. Zhao, S. Chen, and Y. Wang, "Editable scene simulation for autonomous driving via collaborative llm-agents," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 077–15 087.

[110] X. Li, E. Liu, T. Shen, J. Huang, and F.-Y. Wang, "Chatgpt-based scenario engineer: A new framework on scenario generation for trajectory prediction," *IEEE Transactions on Intelligent Vehicles*, 2024.

[111] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kümmerle, H. Königshof, C. Stiller, A. de La Fortelle, and M. Tomizuka, "INTERACTION Dataset: An INTERnational, Adversarial and Cooperative moTION Dataset in Interactive Driving Scenarios with Semantic Maps," *arXiv:1910.03088 [cs, eess]*, Sep. 2019.

[112] H. Tian, K. Reddy, Y. Feng, M. Quddus, Y. Demiris, and P. Angeloudis, "Enhancing autonomous vehicle training with language model integration and critical scenario generation," *arXiv preprint arXiv:2404.08570*, 2024.

[113] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[114] J. Zhang, C. Xu, and B. Li, "Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 459–15 469.

[115] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[116] R. Gao, K. Chen, E. Xie, L. Hong, Z. Li, D.-Y. Yeung, and Q. Xu, "Magicdrive: Street view generation with diverse 3d geometry control," *arXiv preprint arXiv:2310.02601*, 2023.

[117] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *arXiv preprint arXiv:2112.10752*, 2022.

[118] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. Zhang, "Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 749–14 759.

[119] F. Jia, W. Mao, Y. Liu, Y. Zhao, Y. Wen, C. Zhang, X. Zhang, and T. Wang, "Adriver-i: A general world model for autonomous driving," *arXiv preprint arXiv:2311.13549*, 2023.

[120] D. Wu, W. Han, T. Wang, Y. Liu, X. Zhang, and J. Shen, "Language prompt for autonomous driving," *arXiv preprint arXiv:2309.04379*, 2023.

[121] L. Chen, O. Sinavski, J. Hünermann, A. Karnsund, A. J. Willmott, D. Birch, D. Maund, and J. Shotton, "Driving with llms: Fusing object-level vector modality for explainable autonomous driving," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024.

[122] M. Nie, R. Peng, C. Wang, X. Cai, J. Han, H. Xu, and L. Zhang, "Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving," *arXiv preprint arXiv:2312.03661*, 2023.

[123] Y. Ma, C. Cui, X. Cao, W. Ye, P. Liu, J. Lu, A. Abdelraouf, R. Gupta, K. Han, A. Bera *et al.*, "Lampilot: An open benchmark dataset for autonomous driving with language model programs," *arXiv preprint arXiv:2312.04372*, 2023.

[124] Y. Li, W. Zhang, K. Chen, Y. Liu, P. Li, R. Gao, L. Hong, M. Tian, X. Zhao, Z. Li *et al.*, "Automated evaluation of large vision-language models on self-driving corner cases," *arXiv preprint arXiv:2404.10595*, 2024.

BIOGRAPHIES

Hanlin Tian is a postgraduate researcher at the Centre for Transport Engineering and Modelling, Imperial College London. He received a BEng in Computer Science from Shandong University and an MSc in Computer Engineering from New york University. His main research interests include computer vision and autonomous vehicles.

Kethan Reddy is a Research Associate at the Centre for Transport Engineering and Modelling, Imperial College London. He received an MPhys. in Physics from the University of Kent and an MSc. in Machine Learning from University College London. His main research interests include artificial intelligence for robotics and autonomous vehicles.

Yuxiang Feng is a Research Associate and Lab Manager at the Centre for Transport Engineering and Modelling, Imperial College London. He received a BEng in Mechanical Engineering from Tongji University and an MSc in Mechatronics and PhD in Automotive Engineering from the University of Bath. His main research interests include environment perception, sensor fusion and artificial intelligence for robotics and autonomous vehicles.

Mohammed Quddus received the B.Sc. degree in civil engineering from Bangladesh University of Engineering and Technology in 1998, the master's degree in transportation engineering from the National University of Singapore in 2001, and the Ph.D. degree from Imperial College London in 2006. He joined the School of Architecture, Building and Civil Engineering, Loughborough University, U.K., in 2006, as a Lecturer, where he was a Professor of intelligent transport systems (ITS) in 2013. In 2021, he moved to Imperial College London as the Chair Professor of ITS. He has authored over 200 technical papers in international refereed journals and conference proceedings. His research interests include connected and autonomous vehicles, AI, and statistical modeling. He is an Associate Editor of Transportation Research—C: Emerging Technologies.

Yiannis Demiris (SM'03) received the B.Sc. (Hons.) degree in artificial intelligence and computer science and the Ph.D. degree in intelligent robotics from the Department of Artificial Intelligence, University of Edinburgh, Edinburgh, U.K., in 1994 and 1999, respectively. He is a Professor with the Department of Electrical and Electronic Engineering, Imperial College London, London, U.K., where he is the Royal Academy of Engineering Chair in Emerging Technologies, and the Head of the Personal Robotics Laboratory. His current research interests include human-robot interaction, machine learning, user modeling, and assistive robotics. Prof. Demiris is a Fellow of the Institution of Engineering and Technology (IET), and the British Computer Society (BCS).

Panagiotis Angeloudis is Reader and Head of the Transport Systems and Logistics Laboratory (TSL), based in the Centre for Transport Studies (CTS) at Imperial College London. Before establishing TSL, Panagiotis held a JSPS Research Fellowship at Kyoto University. He previously obtained a PhD in Transportation at Imperial College London and spent periods as a research analyst at DP World and the United Nations in Geneva. His research focuses on the study of networks, optimisation methods and multi-agent systems, as well as their applications in autonomous transport systems, urban infrastructure and logistics.