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

Hanlin Tian, *Graduate Student Member, IEEE*, Kethan Reddy, Yuxiang Feng, *Member, IEEE*, Mohammed Quddus, Yiannis Demiris, *Senior Member, IEEE*, and Panagiotis Angeloudis, *Member, IEEE*

***Abstract*—As autonomous vehicles (AVs) advance, the integration of Large (Vision) Language Models (LLMs and VLMs) has emerged as a promising approach to enhance AV capabilities in perception, planning, decision-making, and data generation. However, the practical challenges of incorporating LLMs and VLMs into AV systems, including computational efficiency, real-time processing, and ethical considerations, remain underexplored. This survey aims to provide a comprehensive review of the current research on LLM and VLM applications in AVs, focusing on the following key areas: modular integration, end-to-end integration, data generation, evaluation platforms, datasets, and benchmarks. We systematically analyse 77 recent papers published before Sep 2025, detailing their methodologies and models. Our findings highlight the potential of LLMs and VLMs to improve AV system performance while acknowledging limitations. This survey offers researchers and practitioners a panoptic view of the classification and progression of LLMs and VLMs in the AV sphere, while systematically distilling models to their core components. We envision this survey as a central reference for AV researchers navigating this rapidly evolving landscape to accelerate future research.**

***Index Terms*—Autonomous vehicles, natural language processing, computer vision.**

I. Introduction

LARGE Language Models (LLMs) like GPT-4 [1] 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 (VLMs) on the other hand, such as LLaVA [2], integrate visual understanding with language processing, enabling tasks like image captioning and visual reasoning. These capabilities make LLMs and VLMs especially valuable in autonomous vehicles (AVs), where visual and textual comprehension, human-machine interaction, and explainability are crucial. Henceforth throughout this paper, we reference LLMs **or** VLMs as **VLMs** as shorthand to avoid repetition **unless** the model **only** takes texts as an input. Explicit examples of how these models can enhance AV tasks include natural language interfaces that can interpret and respond to conversation commands, as well as context-aware decision-making mechanisms that predict and adapt to evolving traffic patterns.

Concrete examples can be observed through industry adoption to further expound on the utility of VLMs in the context of real-world applications concerning the AV task. 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. Beyond Wayve, NVIDIA utilises LLMs to comprehend complex and long-tailed open-world driving scenarios, addressing a broad spectrum of AV tasks [3].

Furthermore, numerous research institutes and communities are actively exploring VLM integration within the AV context, reflecting the growing interest in this field. For instance, *DriveLM* [4] studies how VLMs trained on web-scale data can be integrated into End-to-End driving systems to boost generalisation. Another example is *GenAD* [5], which is a generalised video prediction model for AVs that supports action-conditioned prediction and planning. This underscores the versatility of VLMs, *Drive as you speak* uses LLMs for voice commands [6], 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 Language Models and Vision Language Models in AVs offer valuable insights, but have notable gaps. *LLM4Drive* [7] overviews applications in perception, prediction, and planning but lacks deep exploration of practical integration challenges and broader implications such as regulatory and ethical considerations. *A Survey for Foundation Models in Autonomous Driving* [8] categorises foundation models but does not extensively analyse integration challenges or computational trade-offs involved in deployment. *Towards Knowledge-Driven Autonomous Driving* [9] proposes a conceptual framework emphasising knowledge integration to enhance cognition and learning, but does not specifically focus on the role of VLMs or provide detailed analyses of their applications within AVs. *Large Language Models for Mobility in Transportation Systems* [10] focuses on mobility forecasting, neglecting other critical AV aspects such as perception and control. Collectively, while these surveys highlight advancements, they

Received 8 October 2024; revised 26 March 2025 and 2 September 2025; accepted 1 November 2025. This work was supported by the Innovate U.K. under Grant 10063539. The Associate Editor for this article was C. Lv. *(Hanlin Tian and Kethan Reddy contributed equally to this work.) (Corresponding author: Yuxiang Feng.)*

Hanlin Tian, Kethan Reddy, Yuxiang Feng, Mohammed Quddus, and Panagiotis Angeloudis are with the Centre for Transport Engineering and Modelling, Department of Civil and Environmental Engineering, Imperial College London, SW7 2AZ London, U.K. (e-mail: y.feng19@imperial.ac.uk).

Yiannis Demiris is with the Personal Robotics Laboratory, Department of Electrical and Electronic Engineering, Imperial College London, SW7 2AZ London, U.K.

Digital Object Identifier 10.1109/TITS.2025.3628969

underexplored the practical integration of VLMs into existing AV systems, especially regarding regulatory, ethical, and Vehicle-to-Everything (V2X) communication considerations.

To address these gaps, this survey offers a comprehensive overview of current research developments in VLMs applications for AVs. We critically assess various approaches and methodologies with a focus on practical integration into existing AV systems in real-world contexts. By defining tasks suited for VLMS in AVs and accounting for the unique requirements and constraints of autonomous driving 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:

- Conducted a comprehensive and up-to-date review (as of Sep 2025) of 77 papers on the integration and impact of VLMs in AVs, providing a structured analysis across the following key categories: modular integration, End-to-End integration, data generation, evaluation platforms, datasets, and benchmarks.
- Systematically discussed the VLMs models used in each paper, including the specific LLM models employed (e.g. GPT-4, LLaMA).
- Propose further research directions and innovative approaches to address the current challenges in the integration and application of VLMs in AVs.

The outline of this paper is as follows: In Section II, a review of existing literature related to VLM 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 VLMs within existing AV architectures. Section V highlights VLM applications in End-to-End integration within AV systems and frameworks. Section VI examines how VLM models can be leveraged for data generation across various AV domains. Section VII overviews typically utilised AV platforms and datasets, in addition to new VLM-specific benchmarks to stress test the performance and efficiency of AV VLM pipelines. Section VIII expands on particular potential shortcomings and research directions pertaining to VLMs 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 VLMs are typically incorporated in the AV domain.

First, VLM systems can be tacked onto existing AV architectures. This VLM integration method applies to modular AV architectures that divide tasks such as perception and decision-making into specialised components to create intelligent agents capable of advanced comprehension and interaction. Second, and recently becoming more prominent in the literature, VLMs are being integrated into End-to-End AV architectures. End-to-End AV architectures represent state-of-the-art solutions for AVs, providing a seamless integration of perception, prediction, and planning into a unified framework. And finally, the integration of VLMs into data generation frameworks is particularly promising. This is due to their extensive object-relations naturally embedded in human language, as well as their ability to translate abstract scenarios into structured, machine-readable formats [11].

A. Related Work on Integrating VLMs With Modular AVs

LLMs have evolved significantly in recent years, exhibiting advanced capabilities in text generation, comprehension, and other areas [12]. The rapid iteration of the Llama 3 series, with successive releases like Llama 3.1, 3.2, and beyond, showcases the accelerated development pace now typical in the AI field [13].

VLMs advance the integration of natural language processing and computer vision, enabling holistic interpretation of multimodal data. LLaVA (Large Language and Vision Assistant) [2] 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 [14] 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 Combining VLMs With End-to-End AVs

End-to-End AVs integrate perception, prediction, and planning into a single framework. Early work, such as ALVINN, used neural networks for direct steering control from sensor inputs [15]. 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 [16].

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 [17]. RL approaches, using algorithms like Deep Q-Networks (DQN) and Asynchronous Advantage Actor-Critic (A3C), optimise policies through simulated interactions [18].

Recent advancements in End-to-End AV systems have demonstrated their potential benefits. For instance, UniAD [19] 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 Merging VLMs With Generation 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 [20], and Stable Diffusion [21] across various industry vertical use cases. In AV applications, the main focus for generative capabilities lies in visual generation, hence the terms "diffusion models" and "generation models" are used interchangeably in this paper. Diffusion models operate by

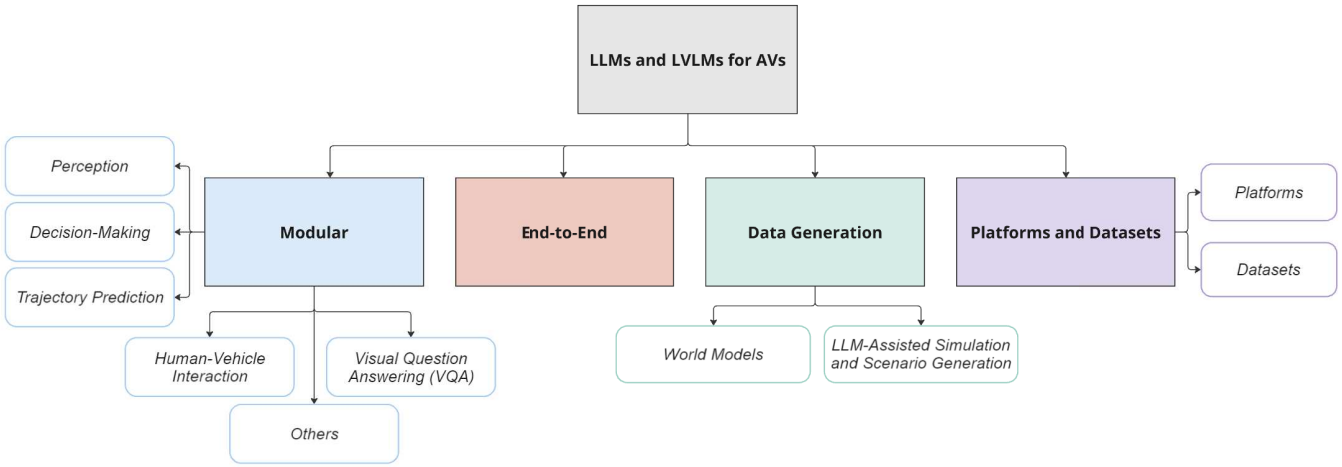

Fig. 1. Diagrammatic representation of the key areas for integrating VLMs into AV architectures, categorised into modular integration, End-to-End integration, data generation, and platforms and datasets.

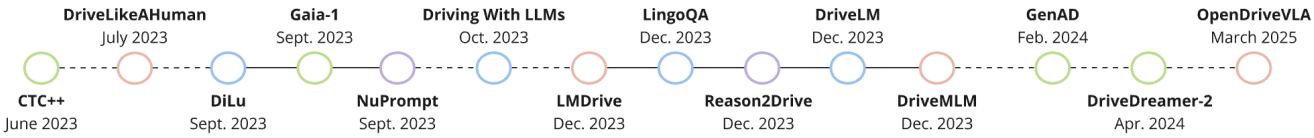

Fig. 2. Timeline of some select papers incorporating VLMs for AVs. Blue indicates VLM integration within a modular framework, red signals VLM End-to-End architectures, green refers to VLM data generation pipelines, and purple references VLM platforms and datasets. A dashed line indicates different publication months, while a solid line indicates the same publication month.

progressively transforming random noise into coherent images or videos through iterative refinement [22]. The flexibility of diffusion models in accepting diverse input modalities, such as text prompts and sketches, enables rapid prototyping and customised outputs [23].

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 [24], [25]. 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 VLMs in the field of AVs. We conducted extensive searches, predominantly between Sep 2023 and Sep 2025 (Fig. 2), across various academic databases including IEEE Xplore, Google Scholar, and arXiv. The selection criteria focused on identifying studies that explicitly discussed VLMs 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 VLMs 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 77 papers published before Sep 1, 2025, for this review. We excluded studies that **solely utilised vision transformers** to better focus on the unique contributions of VLMs, 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 VLM 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 VLM Applications in AVs

The application of VLMs in AVs is diverse. This survey classifies the research by its fields to provide a structured overview, as depicted in Fig. 6:

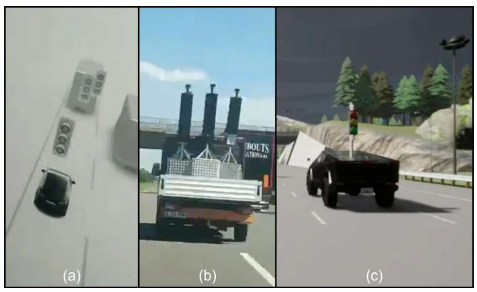

Fig. 3. An example of perception tasks in AVs: a semantic anomaly where traffic lights appear to pass through a car, later revealed as inactive lights in transport. Image from [26].

- **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 AV models.
- **Platforms and Datasets** establish standards for evaluating the performance of AV systems.

IV. Modular Integration

AVs often adopt modular architecture, dividing the overall task into specific components such as perception, decision-making, trajectory prediction, human-vehicle interaction, etc. This structured approach allows for the specialised development and optimisation of each component.

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

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 VLMs into perception systems have improved visual scene understanding and reasoning.

1) Semantic Anomaly Detection with Large Language Models [26]: 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). This approach effectively identifies and reasons about semantic anomalies, aligning with human judgment.

2) Zelda [27]: A vision-language model for video analytics that processes natural language queries like "cars during daytime at intersections." It improves retrieval relevance and diversity using synonym-based prompt generation and filters low-quality frames using descriptive cues (e.g., "blurry"). Evaluated on BDD with VIVA, Zelda shows higher precision and efficiency than traditional systems.

3) Talk2BEV [28]: Uses BLIP-2 to augment bird's-eye view (BEV) maps with detailed object and scene descriptions, capturing semantics like vehicle types, colours, and spatial relations. This allows for flexible, language-based querying. For instance, VLMs interpret user queries and utilise geometric data from BEVs to perform tasks such as calculating distances or identifying objects in specific spatial contexts. This capability eliminates the need for task-specific models and allows for zero-shot generalisation across diverse driving scenarios to help with decision-making.

4) ChatBEV [29]: Introduces a modular multi-stage pipeline combining BEV spatial reasoning with language-conditioned action planning. The system leverages a frozen pre-trained BEV encoder to extract semantic map features, which are then encoded into discrete tokens using a vector quantisation layer. ChatBEV achieves zero-shot performance on the nuScenes dataset and enables spatial querying capabilities such as "Where is the red car?" or "What is the next action?".

5) X-Driver [30]: Builds a cross-modal driving agent that interprets scenes and generates trajectory-level plans using an X-VLM framework. The system extracts layered visual features from BEV and semantic maps, processes them through a multi-scale encoder, and fuses these with structured traffic rule prompts embedded in a language-aware planner. High-level intentions and trajectories are decoded via a chain-of-thought (CoT) enhanced LLM.

Summary on Perception: A key trend in perception is the translation of visual data into a language-compatible format to leverage the reasoning capabilities of VLMs. Semantic Anomaly Detection converts scenes to text to identify outliers via linguistic reasoning. Talk2BEV enriches geometric Bird's-Eye View maps with semantic descriptions, enabling natural language queries about spatial relationships. Zelda optimises this process for video analytics, using language prompts to filter and retrieve relevant frames. While demonstrating powerful scene understanding and anomaly detection, these methods often face challenges in computational efficiency for real-time deployment, pointing to a need for future work on optimisation and lightweight models.

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 VLMs into these systems enhances the vehicle's ability to interpret nuanced situations, apply advanced reasoning, and make more informed decisions.

TABLE I
TABLE FOR PERCEPTION PAPERS

Paper	Submission Time	Specific Tasks	LLM Model	Dataset or Simulator
Semantic Anomaly Detection with Large Language Models [26]	May 2023	Semantic anomaly detection in vision-based policies	GPT-3.5 [31]	CARLA [32]
Zelda [27]	May 2023	Video analytics	VIVA [33]	BDD-X [34]
Talk2BEV [28]	Oct 2023	Visual reasoning, spatial understanding, decision-making	BLIP-2 [35] , MiniGPT-4 [36] , InstructBLIP [37]	Talk2BEV-Bench [28]
ChatBEV [29]	Mar 2025	BEV Scene Reasoning, Spatial Planning	Vicuna-7B, GPT-3.5	nuScenes
X-Driver [30]	Jun 2025	Scene Interpretation, Planning Trajectory Control	X-VLM	CARLA-XDrive

Recent advances indicate that integrating VLMs into the decision-making modules of AVs can enhance safety, efficiency, and adaptability by aligning complex scenario outcomes or proposed trajectories more closely with human expectations, leveraging their implicitly learned object-relational ontologies and semantic parsing of human experience. Moreover, VLM 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.

1) MTD-GPT [38]: Models the multi-task decision-making problem of AVs crossing unsignalised intersections as a sequence modelling problem. RL trains single-task expert models using domain-specific reward functions to generate high-quality state-action-reward data. GPT is then trained offline on this expert dataset, abstracting the problem as a sequence modelling task. This combination allows MTD-GPT to handle multi-task decision-making effectively, where RL alone would lack generalisation and GPT alone would lack domain-specific expertise. While not strictly employing LLMs (only GPT-based), we thought this publication was a notable precursor to subsequent publications.

2) DiLu [40]: Consists of a driver agent, an interactive environment, and a memory component, with the LLM playing a central role in decision-making. The Reasoning Module leverages the LLM to query past experiences from the memory module, combining scenario descriptions with few-shot examples to generate step-by-step reasoning and decisions, which are decoded into actions for the ego vehicle. The Reflection Module refines these decisions by analysing errors, identifying reasoning flaws, and suggesting corrections, which are stored back in memory for continuous learning. DiLu uses an LLM as its core reasoner, allowing it to reflect on its knowledge and environment. This leads to more dynamic and flexible decision-making than what is typically possible with traditional reinforcement learning.

3) Empowering Autonomous Driving with LLMs [42]: Integrates an LLM with a Model Predictive Controller to simplify lane-change decisions. The LLM acts as a high-level decision-maker, interpreting textual descriptions of driving scenarios to recommend lane-change actions (e.g., "left lane"). This approach offloads discrete decision-making from the MPC, reducing computational complexity while maintaining safety.

4) LanguageMPC [43]: A CoT 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.

5) LLM Multimodal Traffic Forecasting [45]: Integrates time-series models (ARIMA, Prophet) with LLMs (GPT-4, Zephyr-7b-α) and a VLM (LLaVA) for traffic accident forecasting. The framework processes structured data, textual reports, and real-time images to provide context-aware risk assessments and driving recommendations.

6) Driving with LLMs [48]: 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.

7) A Language Agent for Autonomous Driving [49]: Transforms the traditional perception-prediction-planning pipeline by integrating a library for dynamic function calls, cognitive memory, and a reasoning engine. By dynamically invoking functions from a tool library, the LLM extracts and synthesises environmental information such as detected objects, predicted trajectories, and maps. This data is further augmented by a cognitive memory system storing common sense and past driving experiences, enabling the LLM to perform CoT reasoning, hierarchical task planning, and motion planning.

8) Evaluation of LLMs for Decision Making in Autonomous Driving [51]: 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 [38]	Sep 2023	Multi-Task Decision-Making, Uncontrolled Intersection	GPT-2	OpenAI Gym [39]
DiLu [40]	Sep 2023	Knowledge-Driven Decision Making	Not Specified	HighwayEnv [41]
Empowering Autonomous Driving with LLMs [42]	Nov 2023	Decision Making, Lane Change	GPT-4 [1]	HighwayEnv [41]
LanguageMPC [43]	Oct 2023	Decision Making, Intersection, Obstacle Avoidance	GPT-3.5 [31]	IdSim [44]
LLM Multimodal Traffic Forecasting [45]	Oct 2023	Decision Making, Traffic Accident Forecasting	GPT-4 [1] LLaMa2 [13] Zephyr-7b-α [46]	Fatality Analysis Reporting System (FARS) [47]
Driving with LLMs [48]	Oct 2023	Decision Making, QA Task	GPT-3.5 [31]	Custom 2D Simulator
A Language Agent for Autonomous Driving [49]	Nov 2023	Decision Making, Motion Planning	GPT-3.5 [31]	nuScenes [50]
Evaluation of LLMs for Decision Making in Autonomous Driving [51]	Dec 2023	Decision Making, Traffic Rules	LLaMa2 [13] GPT-3.5 [31], GPT-4 [1]	Field test
Multi-Modal GPT-4 Aided Action Planning and Reasoning for Self-driving Vehicles [52]	Mar 2024	Decision Making, Action Planning, Reasoning	GPT-4 [1]	CARLA [32]
Receive, Reason and React [53]	Apr 2024	Decision Making, Highway Overtaking, On-ramp Merging	GPT-4 [1]	HighwayEnv [41]
Sce2DriveX [54]	Feb 2025	Scene Encoding, Meta Action Planning, CoT Reasoning	Vicuna-v1.5-7b	nuScenes, CARLA Bench2Drive
AgentThink [55]	Jun 2025	Tool-Augmented Reasoning, Dynamic Tool Use, Autonomous VQA	Qwen2.5-VL-7B, GPT-4o [1]	DriveLMM-o1, DriveMLLM, BDD-X, DriveBench, Navsim

9) Multi-Modal GPT-4 Aided Action Planning and Reasoning for Self-driving Vehicles [52]: Leverages GPT-4 for action planning and reasoning in AVs 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.

10) Receive, Reason, and React [53]: Enables LLMs to serve as a decision-making module in AVs 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.

11) Sce2DriveX [54]: Proposes a human-like CoT reasoning framework that achieves progressive reasoning from multi-view scene understanding to driving actions. The method employs separate encoders for multi-view videos and BEV maps, aligning them into a unified visual feature space using LanguageBind's contrastive learning approach. These features are projected into a Vicuna-v1.5-7B backbone, which generates comprehensive outputs including scene understanding, meta-action reasoning, behaviour interpretation, motion planning, and control signals.

12) AgentThink [55]: Introduces a tool-augmented reasoning framework for autonomous driving that enhances VLMs with dynamic decision-making capabilities. Like AlphaDrive, AgentThink employs a two-stage training process of supervised fine-tuning and Group Relative Policy Optimisation (GRPO). The architecture utilises a structured dataset generation pipeline that integrates a modular driving tool library (visual info, detection, prediction, mapping) into reasoning chains, enabling tool-triggered verifications within reasoning steps. During inference, the model dynamically decides when and how to invoke tools to improve decision validity and completeness.

Summary on Decision Making: LLMs are being integrated into decision-making through two primary paradigms: reasoning-enhanced planning and tool-augmented agency. The first paradigm uses the LLM as a reasoning engine to decompose problems, as seen in DiLu's memory-reflection loop and LanguageMPC's CoT-to-MPC parameter conversion. The second paradigm treats the LLM as an agent that can call specialised tools (perception, prediction, control), exemplified by A Language Agent for Autonomous Driving and AgentThink. A consistent challenge across all frameworks is balancing the LLM's reasoning strength with the need for safety, low latency, and reliability, often addressed through hybrid designs that combine LLMs with traditional, verifiable controllers (e.g., MPC in Empowering Autonomous Driving).

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 AV operation. The goal of trajectory prediction is to minimise errors and miss rates, thereby enhancing the AV's ability to navigate complex and dynamic traffic scenarios safely and efficiently.

1) GPT-driver [56]: Transforms motion planning for AVs into a language modelling problem. This approach reformulates planner inputs, such as perception data, ego-vehicle states, and high-level driving goals, into structured language tokens. The LLM then generates trajectory outputs as natural language descriptions of waypoints. The method employs a prompting-reasoning-finetuning strategy, allowing the LLM to describe precise trajectory coordinates in its decision-making process. GPT-Driver can therefore generate interpretable driving trajectories.

2) LLaDA [57]: The framework employs a the Traffic Rule Extractor (TRE), which uses LLMs like GPT-4, to filter and extract context-specific traffic rules from extensive local traffic codes using inputs such as execution plans and descriptions of unexpected situations. These refined rules, combined with the LLM's reasoning capabilities, enable the generation of motion plans that comply with regional regulations in a zero-shot manner. Additionally, LLaDA integrates VLMs like GPT-4V to process visual scene data, converting it into textual descriptions to further refine context understanding. Demonstrated on datasets like nuScenes, LLaDA significantly improves motion planning accuracy, reduces trajectory prediction errors, and ensures compliance with localised driving norms.

3) Diffusion-ES [59]: Optimises trajectory planning for AV using a gradient-free method combined with trajectory denoising. The framework generates initial trajectory samples using a diffusion model, then refines them through an LLM-assisted denoising process. LLMs transform natural language instructions into reward functions, produces trajectories that align with nuanced driving scenarios. For example, LLMs map instructions like "yield to car 8 and change lanes to the left" into executable Python-based reward shaping functions, enabling Diffusion-ES to dynamically adapt its trajectory planning. This approach achieves high performance results on the nuPlan benchmark.

4) LC-LLM [60]: Uses an LLM to predict lane-change intentions and trajectories. It converts driving scenarios (vehicle states, traffic, map data) into text prompts. Using CoT reasoning, the model infers behaviors—like an overtake due to a slow truck—and outputs both intention and trajectory tokens. On the highD dataset, it outperforms traditional models in accuracy.

5) Occ-LLM [62]: Introduces an LLM for autonomous driving that processes 3D occupancy grids—a universal representation of scene geometry and semantics. To handle the sparsity and imbalance of occupancy data, it uses a novel Motion Separation VAE (MS-VAE) to encode dynamic and static elements separately. The resulting tokens are fed into an LLM (Llama2) trained for multiple tasks: 4D occupancy forecasting, ego-vehicle planning, and scene question-answering. It sets a new state-of-the-art on nuScenes, significantly outperforming prior work like OccWorld in forecasting IoU (32.52% vs. 26.63%) and reducing planning error (L2 of 0.28m vs. 1.17m).

Summary on Trajectory Prediction: The integration of LLMs marks a shift towards more interpretable and context-aware autonomous driving systems. The reviewed approaches demonstrate two primary patterns:

- Using LLMs to **directly output** plans or predictions in a structured format (e.g., GPT-Driver, LC-LLM, Occ-LLM).
- Using LLMs as **reasoning engines** to guide other processes, such as rule filtering for LLaDA or reward-shaping for Diffusion-ES.

While demonstrating promise, a key challenge remains validating their reliability in real-world conditions. Furthermore, their natural language capabilities provide a crucial foundation for the next critical challenge: fostering effective **human–vehicle interaction**.

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.

1) DriveAsYouSpeak [6]: LLMs interpret verbal commands, such as "overtake the car in front," by integrating data from various sensory modules, including perception (speed and traffic conditions), localisation, and in-cabin monitoring (driver attentiveness). The system also employs memory modules to store historical data and driver preferences, enabling personalised and adaptive responses. For instance, it can recall a driver's comfort level with overtaking speeds to tailor trajectory plans accordingly. This combination of continuous learning and transparent decision-making boosts trust and improves user experience in AVs.

2) ChatGPT as Your Vehicle Co-Pilot [63]: Effectively identical to *DriveAsYouSpeak*, but differs with the inclusion of fail-safe mechanisms that monitor and filter erroneous outputs from ChatGPT. These mechanisms include rule-based validators and expert-designed heuristics to cross-check LLM-generated plans against predefined safety parameters, such as collision avoidance and adherence to traffic laws. For instance, if the LLM suggests a trajectory that violates a safe distance from other vehicles, the system flags and corrects the output before execution. This ensures that the vehicle operates within safety margins at all times, even when encountering unexpected or edge-case scenarios.

3) Dolphins [3]: Building on OpenFlamingo, it employs the Grounded Chain of Thought (GCoT) process to enable step-by-step reasoning, which allows it to interpret complex

TABLE III
TABLE FOR TRAJECTORY PREDICTION PAPERS

Paper	Submission Time	Specific Tasks	LLM Model	Dataset or Simulator
GPT-driver [56]	Oct 2023	Trajectory Prediction, Motion planning	GPT-3.5 [31]	nuScenes [50]
LLaDA [57]	Feb 2024	Traffic Rule Assistance for Tourists, AV Motion Plan Adaptation	GPT-4 [1]	nuScenes [50], nuPlan [58]
Diffusion-ES [59]	Feb 2024	Trajectory Optimisation, Zero-Shot Instruction Following	GPT-3.5 [31]	nuPlan [58]
LC-LLM [60]	Mar 2024	Lane Change Intention, Trajectory Predictions	LLaMA2-7B [13]	highD [61]
Occ-LLM [62]	Feb 2025	4D Occupancy Forecasting, Scene QA, Self-Ego Planning	Llama2 [13]	nuScenes [50]

TABLE IV
TABLE FOR HUMAN-VEHICLE INTERACTION PAPERS

Paper	Submission Time	Specific Tasks	LLM Model	Dataset or Simulator
DriveAsYouSpeak [6]	Sep 2023	Natural Language Interaction, Adaptive Decision-Making	ChatGPT-4 [1]	Not Mentioned
ChatGPT as Your Vehicle Co-Pilot [63]	Oct 2023	Path Tracking, Trajectory Planning, Human-Machine Interaction	ChatGPT-3.5 Turbo [31]	Simulink [64], CarSim [65]
Dolphins [3]	Dec 2023	Decision Making, Holistic Driving Understanding	CLIP [66] LLaMA [13] MPT [67]	BDD-X [34] DriveLM [4]
Towards Human-Centric Autonomous Driving [68]	May 2025	Instruction Following, Motion Planning, Safety-Critical Decision Making	GPT-4o-mini [1]	Highway-Env [41]

driving scenarios, such as recognising dynamic behaviours of other vehicles or understanding weather conditions and traffic signals. Using instruction tuning based on the BDD-X dataset, Dolphins is optimised for tasks such as behaviour comprehension, control signal forecasting, and conversational engagement. For example, it can process a driver's command like "drive cautiously in the rain" by analysing visual inputs to generate context-aware responses and plans.

4) Towards Human-Centric Autonomous Driving [68]: Proposes a "fast-slow" architecture where GPT-4o-mini interprets natural language commands into structured directives for a reinforcement learning (RL) agent handling real-time control. The framework enables adaptive, user-preferred driving while maintaining safety via an override mechanism. Evaluated in a custom Highway-Env simulator, it outperforms RL and LLM baselines in success rate, safety, and instruction adherence.

Summary on Human-vehicle Interaction: LLM-based HVI systems enhance driving experiences through natural language interfaces, personalization, and contextual adaptation. Key challenges include ensuring safety via validation mechanisms (e.g., rule-based checkers) and balancing user preferences with operational constraints. Metrics such as safety compliance, user satisfaction, and instruction accuracy are critical for evaluation. Future work may focus on multi-modal integration and real-world deployment.

E. VQA

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

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. 4).

1) LingoQA [69]: The framework processes short video segments and incorporates multi-frame analysis via CLIP and Q-Former to extract and align visual features with language representations. The use of a learned text classifier, Lingo-Judge, provides a robust, automated evaluation metric that outperforms traditional methods such as BLEU and CIDEr. This metric evaluates models' answers for correctness, focusing on factual accuracy and nuanced reasoning.

2) DriveLM [4]: Proposes Graph Visual Question Answering (GVQA) for AVs. GVQA models driving scenarios as interconnected question-answer pairs arranged in a directed graph that spans perception, prediction, and planning stages.

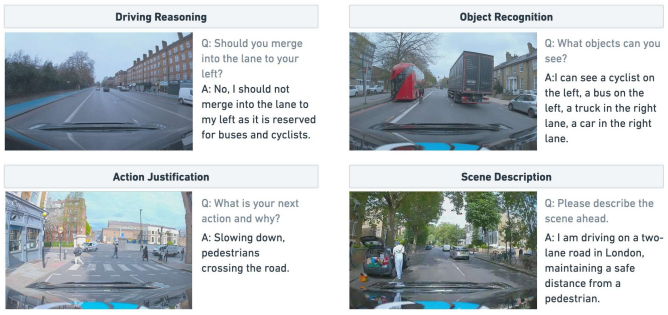

Fig. 4. Examples of Visual Question Answering (VQA) tasks in AV, showcasing driving reasoning, object recognition, action justification, and scene description. Adapted from [69].

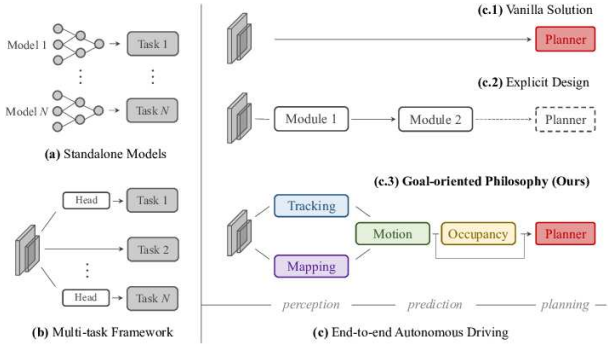

Fig. 5. Comparison of various algorithm framework designs: Standalone Models, Multi-task Frameworks, and End-to-End AVs. Adapted from [19].

DriveLM-Data, built on datasets like nuScenes and CARLA, provides extensive graph-structured annotations for training, linking visual inputs with natural language queries. DriveLM-Agent, the baseline model, utilises graph prompting and trajectory tokenisation to predict trajectories. This approach improves interpretability and zero-shot generalisation.

3) LiDAR-LLM [70]: Adopts Large Language Models for 3D LiDAR data in AVs, performing tasks like 3D captioning and question answering. A three-stage training strategy progressively aligns sparse 3D point cloud data with the language embedding space, enabling effective cross-modal reasoning. Tasks like 3D captioning, grounding, and question-answering are performed by generating 700k high-quality LiDAR-text pairs, aligning descriptions of complex outdoor scenarios to textual queries. Evaluations are conducted using the NuScenes-QA dataset.

4) EM-VLM4AD [71]: This framework aggregates visual data from six camera views into a unified embedding, fused with text embeddings, and processed by a fine-tuned T5-Base model. This architecture enables accurate responses to safety-critical questions, such as identifying key objects or predicting traffic agent behaviour, while reducing memory and computation compared to larger models. Evaluated on the DriveLM dataset, EM-VLM4AD outperforms larger systems like DriveLM-Agent in metrics like ROUGE-L and CIDEr.

5) DriveLMM-o1 [72]: Introduces a large multimodal model and benchmark for step-by-step reasoning in autonomous driving scenarios. The model is built by fine-tuning InternVL2.5-8B using LoRA on a novel dataset. It processes stitched multi-view images using dynamic image patching to handle high-resolution inputs efficiently. The model is designed to answer questions requiring complex reasoning about perception, prediction, and planning. It generates detailed, CoT rationales before providing a final answer, improving transparency and accuracy

6) MPDrive [73]: Proposes a marker-based prompt learning framework to enhance spatial understanding in autonomous driving VQA. It utilises a detection expert to overlay visual markers (numerical labels on object masks) on driving scenes, converting coordinate prediction into simpler marker index prediction. The architecture includes Marker ControlNet for feature fusion and Perception-Enhanced Spatial Prompt Learning for dual-granularity spatial prompts. Evaluated on DriveLM and CODA-LM datasets, it achieves state-of-the-art performance in spatial perception metrics and language consistency.

Summary on VQA: The reviewed VQA systems for autonomous driving demonstrate a clear evolution from basic question-answering towards complex, structured reasoning. LingoQA and EM-VLM4AD both aggregate multi-view visual data but differ in focus: the former introduces a novel automated evaluation metric (Lingo-Judge), while the latter prioritizes computational efficiency. LiDAR-LLM expands modality integration by aligning 3D point clouds with language. DriveLM pioneers a graph-based structure to explicitly model the relationships between perception, prediction, and planning questions. Finally, DriveLMM-o1 advances this further by focusing explicitly on generating and evaluating detailed, step-by-step rationales, significantly improving answer accuracy and model transparency. Collectively, these works highlight a field moving beyond simple captioning towards building explainable and trustworthy reasoning systems for AVs.

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 [76]. Unlike traditional modular architectures that decompose the driving task into separate components,[1] End-to-End systems streamline these processes into a unified framework. This holistic approach is designed to enhance the execution of driving tasks, providing more adaptive and cohesive responses in the dynamic environments encountered on the road [19] (depicted in Fig. 5).

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

[1]Such as perception, planning, and control.

TABLE V
TABLE FOR VISUAL QUESTION ANSWERING PAPERS

Paper	Submission Time	Specific Tasks	LLM Model	Dataset or Simulator
LingoQA [69]	Dec 2023	Video Question Answering, Evaluating Vision-Language Models	Vicuna v1.5 7B [74]	LingoQA [69]
DriveLM [4]	Dec 2023	Graph-based Visual Question Answering, Planning	BLIP-2 [35]	DriveLM [4]
LiDAR-LLM [70]	Dec 2023	3D Captioning & Grounding, 3D Question Answering, Zero-Shot Planning	LLaMA2-7B [13]	nuScenes [50] , NuScenes-QA [11]
EM-VLM4AD [71]	Mar 2024	Visual Question Answering for ADS Safety	T5 LM [75]	DriveLM-Data [4]
DriveLMM-o1 [72]	Mar 2025	Step-by-Step Reasoning, VQA	InternVL2.5-8B	DriveLMM-o1
[73]	Apr 2025	Spatial Understanding, Visual Question Answering (VQA), Marker-Based Coordinate Prediction	Detection-Expert (StreamPETR)	DriveLM [4] , CODA-LM

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 [77]*

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 [78]*

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 [80]*

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) [81]*

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.

E. *LMDrive [91]*

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 [92]*

An LLM-based framework designed for AVs 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 tokeniser and a multi-modal LLM decoder to process inputs such as multi-view images, LiDAR point clouds, traffic rules, and user instructions.

G. *DriveVLM [94]*

The system integrates a 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 [95]*

Presents a holistic framework for End-to-End AV by utilising LLM agents capable of 3D perception, reasoning, and

TABLE VI
TABLE FOR END-TO-END AV PAPERS

Paper	Submission Time	Specific Tasks	LLM Model	Dataset or Simulator
DriveLikeAHuman [77]	Jul 2023	Reasoning, Interpretation, Memorization	GPT-3.5 [31]	HighwayEnv [41]
HiLM-D [78]	Sep 2023	Risk Object Localization, Intention and Suggestion Prediction (ROLISP)	Custom MLLM	DRAMA [79], DRAMA-ROLISP [78]
DriveGPT4 [80]	Oct 2023	Interpretation, Open-loop Control Signal Prediction	GPT-4 [1]	BDD-X [34]
On the Road with GPT-4V(ision) [81]	Nov 2023	Reasoning, Interpretation, Act-as-a-Driver	GPT-4V [82]	nuScenes [50], Waymo Open dataset [83], BDD-X [34], CODA [84], D2-city [85], CCD [86], TSDD [87], ADD [88], DAIR-V2X [89], CitySim [90], CARLA [32]
LMDrive [91]	Dec 2023	Closed-loop Driving Performance under Language Instructions	LLaVA-v1.5, Vicuna-v1.5 [74]	CARLA [32]
DriveMLM [92]	Dec 2023	Behavioral Planning, Decision-Making	LLaMA2-7B [13] EVA-CLIP [93]	CARLA [32]
DriveVLM [94]	Mar 2024	Interpretation, Planning, Chain-of-Thought Process Integration	Qwen-VL [14]	SUP-AD [94], nuScenes [50]
OmniDrive [95]	May 2024	Scene Description, Scene Analysis, Meta Actions, Decision Description, Trajectory Waypoints	Q-Former3D from BLIP-2 [35]	OmniDrive-nuScenes [95], NuScenes-QA [11]
Co-driver [96]	May 2024	Adjustable Driving Behaviours, Interpretation	Qwen-VL [14]	CARLA [32]
Senna [97]	Oct 2024	Multi-view Scene Understanding, Meta-Action Planning, Trajectory Generation	Vicuna v1.5 [74]	nuScenes [50], DriveX [97]
DriVLMe [98]	Oct 2024	Social Dialogue Navigation, Long-Horizon Planning	Vicuna-7B [74]	SDN [99], CARLA [32]
WiseAD [100]	Dec 2024	Scene Description, Object Recognition, Action Justification, Risk Analysis, Driving Suggestions, Trajectory Planning	MobileLLaMA	CARLA [32], LingoQA [69], DRAMA [79], BDDX [34], DriveLM [94],
VLM-E2E [101]	Feb 2025	Attention-guided Planning, Multimodal (BEV-Text) Fusion, Semantic Occupancy Prediction	BLIP-2 [35]	nuScenes [50]
Orion [102]	Mar 2025	Closed-loop Planning, VQA, Multi-modal Trajectory Generation	Vicuna-v1.5 [74]	Bench2Drive [103]
AlphaDrive [104]	Mar 2025	High-Level Planning, Reasoning	Qwen2VL-2B [14] GPT-4o [1]	MetaAD [104]
OpenDriveVLA [105]	Mar 2025	Trajectory Planning, Driving Question Answering (QA), Agent-Environment-Ego Interaction	Qwen2.5-VL [106]	nuScenes [11]
SOLVE [107]	May 2025	Trajectory prediction, Feature-level VLM/E2E synergy, Trajectory Chain-of-Thought (T-CoT)	LLaVA v1.5	nuScenes [11], CARLA [32]

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.

I. Co-driver [96]

An innovative AV assistant system designed to provide human-like behaviour and understanding in complex road scenes using 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.

J. Senna

[97] Proposes a structured autonomous driving system that decouples high-level planning from low-level trajectory prediction. The framework consists of Senna-VLM, which processes multi-view images to generate high-level planning decisions in natural language (e.g., "accelerate and turn left"), and Senna-E2E, which uses these language-based decisions to predict precise trajectories. This approach avoids the limitations of using VLMs for numerical trajectory prediction while leveraging their strengths in reasoning and scene understanding

K. DriVLMe [98]

Addresses long-horizon driving and social navigation via a dialogue-augmented VLM planner trained on the SDN benchmark. It augments image inputs with explicit linguistic interaction history, using a Vicuna-7B model to track and resolve dialogue-based navigation goals. Its planner is trained to integrate these multimodal cues with driving intents, producing trajectory and language rationales concurrently. DriVLMe demonstrates strong generalisation in both route planning and natural language generation across the CARLA simulator, excelling in interpreting driver-to-passenger exchanges and executing context-aware goals with minimal intervention.

L. WiseAD [100]

Processes short video clips and target waypoints, answering a range of driving-related questions (scene description, object recognition, action justification, risk analysis, driving suggestions) alongside generating future trajectory waypoints in textual form. A joint learning protocol aligns planning behaviour with accumulated driving expertise, significantly reducing critical accidents. Evaluated in CARLA, WiseAD achieves state-of-the-art closed-loop performance, outperforming prior works such as VAD and Roach. WiseAD also demonstrates superior knowledge transfer on in-domain and out-of-domain QA benchmarks (LingoQA, BDDX, DriveLM, HAD).

M. VLM-E2E [101]

Enhances End-to-End autonomous driving by leveraging VLMs to infuse driver attentional semantics into BEV representations. The framework uses BLIP-2 to generate textual descriptions of critical driving cues from a front-view image, which are refined with ground truth data to mitigate VLM hallucination. A novel BEV-Text learnable weighted fusion strategy dynamically balances geometric and semantic features. This enriched representation significantly improves perception, prediction, and planning performance.

N. Orion [102]

Employs QT-Former, a query-based temporal module that compresses multi-frame visual inputs and retrieves relevant long-term scene information. This information is fused with scene reasoning in a Vicuna v1.5 LLM, fine-tuned with LoRA, which produces a planning token that guides a generative planner to generate multi-modal driving trajectories. Evaluated on the Bench2Drive benchmark, Orion achieves a Driving Score (DS) of 77.74 and a Success Rate (SR) of 54.62%, surpassing the previous state-of-the-art DriveTransformer. Orion also demonstrates superior Multi-Ability performance (mean 54.72%), excelling in complex skills such as Overtaking and Emergency Braking.

O. AlphaDrive [104]

Integrates VLM architecture with reinforcement learning for high-level autonomous driving planning. AlphaDrive employs a two-stage training pipeline: supervised fine-tuning distils reasoning traces from a large model (e.g., GPT-4o), followed by RL using Group Relative Policy Optimization (GRPO) to refine the planning policy. The architecture combines a vision encoder and cross-modal adapter with an LLM-based planner, producing structured, multi-modal planning decisions formatted with thinking tokens. Four custom rewards (accuracy, action-weighted, diversity, and format) enable AlphaDrive to generate diverse and safe trajectory plans.

P. OpenDriveVLA [105]

Presents a VLA model that unifies vision-language understanding and trajectory generation for End-to-End autonomous driving. The model processes multi-view images through a visual-centric encoder to extract structured tokens for the scene, dynamic agents, and map elements. A hierarchical alignment module projects these 2D and 3D visual tokens into a unified semantic space for a Qwen2.5 LLM. OpenDriveVLA is trained in multiple stages, including feature alignment, driving instruction tuning, and an autoregressive agent-env-ego interaction process to model dynamic relationships. The final model generates future trajectories by autoregressively predicting tokenised waypoints conditioned on visual perceptions, ego state, and driving commands.

Q. SOLVE [107]

Introduces a synergistic framework that integrates a VLM with an End-to-End driving model through a shared Sequential Q-Former for feature-level knowledge exchange and a Trajectory Chain-of-Thought process for refined planning. The system employs a LLaVA-1.5 model to interpret complex scenes and generate trajectories via a coarse-to-fine selection and refinement mechanism from a trajectory bank. A temporal decoupling strategy allows the VLM's high-quality asynchronous predictions to guide the real-time E2E planner.

Summary on End-to-End Systems: AV End-to-End systems represent a paradigm shift by leveraging VLMs to streamline autonomous driving pipelines. These systems directly map raw sensory inputs to driving decisions, bypassing traditional modular decompositions and offering improved

adaptability, context awareness, and real-time reasoning in complex driving environments.

Several models focus on human-like driving behaviour and reasoning to augment scenario fidelity. **DriveLikeAHuman** and **On the Road with GPT-4V** use LLMs to mimic human decision processes, incorporating memory and contextual reasoning for rare events. Likewise, **DriveVLM** applies a CoT framework for linguistic scene description and interpretable planning. Other frameworks prioritise multi-modal integration and low-level control. **HiLM-D** and **DriveGPT4** use MLLMs to extract high-risk features and generate control signals. **DriveMLM** and **LMDrive** combine language instructions with multi-sensor inputs for improved planning, while **OmniDrive** and **Senna** develop 3D spatial world models for dynamic scene understanding.

A growing number of these systems adopt Vision-Language Action (VLA) designs. Unlike conventional VLMs that align through shared embeddings, VLAs integrate vision and language as co-dependent reasoning streams throughout the planning pipeline. Systems like **AlphaDrive** and **Orion** use planning tokens, cross-modal adapters, and temporal memory to fuse scene context with language reasoning. These designs enable human-readable rationales for decisions, improving interpretability. Advancements like **OpenDriveVLA** demonstrate scaling across subtasks through modular attention and dialogue-conditioned planning. VLAs thus represent a distinct architectural evolution beyond traditional VLMs, offering stronger generalisation and causal reasoning for ambiguous conditions.

The trend towards unified vision-language-action architectures reduces error propagation across modules and simplifies training by enabling the system to learn holistically from large-scale multi-modal data. However, a challenge remains in balancing the interpretability and scalability of these models with their substantial computational demands and the critical need for real-time inference. Inference times can span several seconds. To mitigate this, aggressive visual token compression (Senna-VLM's Driving Vision Adapter), strategic truncation of costly Chain-of-Thought reasoning, hardware-aware strategies such as deploying severely optimised models on embedded SoCs (NVIDIA OrinX) via PEFT, and dual-processor setups aim to compress latency to under 500 milliseconds. Continued work in model compression is not only beneficial but essential for deploying VLMs in practical autonomous vehicle settings.

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 VLMs 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 VLM 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 AVs. 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.

1) CTG++ [108]: 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.

2) Domain Knowledge Distillation from Large Language Model [109]: This framework 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.

3) SurrealDriver [111]: 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.

4) LimSim++ [112]: 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.

5) ChatSim [114]: 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.

6) TrafficGPT [115]: 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.

TABLE VII
TABLE FOR LLM-ASSISTED SIMULATION AND SCENARIO GENERATION

Paper	Submission Time	Specific Tasks	LLM Model	Dataset or Simulator
CTG++ [108]	Jun, 2023	Queryable Traffic Compliant Simulations, Fine-Grained Control, Trajectory Modelling	GPT-4 [1]	nuScenes [50]
Domain Knowledge Distillation from Large Language Model [109]	Jul, 2023	Domain Knowledge Distillation	GPT-3.5 [1]	OpenXOntology [110]
SurrealDriver [111]	Sep, 2023	Simulating Human-like Behaviours, Learning from Experts, Adhering to Long-Term Safety Guidelines	GPT-4 [1]	CARLA [32]
LimSim++ [112]	Feb, 2024	Interpretation, Decision-Making, Framework Enhancement	GPT-3.5 [1], GPT-4 [1], GPT-4V [82]	SUMO [113], CARLA [32]
ChatSim [114]	Feb, 2024	Editable Photo-Realistic 3D Driving Scene Simulations via Natural Language Commands	GPT-4 [1]	Waymo Open Dataset [83]
TrafficGPT [115]	Mar, 2024	Closed-loop Reasoning, Decision-Making Support, Task Decomposition, Feedback Adaption	GPT-3.5 Turbo [1]	SUMO [113]
SeGPT [116]	Mar, 2024	Interpretation, Framework Enhancement	GPT-4 [1]	Interaction Dataset [117]
CRITICAL [118]	Apr, 2024	Critical Scenario Generation, Closed-loop Training and Performance Augmentation, Safety Analysis	Mistral-7B-Instruct [119]	HighwayEnv [41]
ChatScene [120]	May, 2024	Safety-Critical Scenario Generation	GPT-3.5 [1]	CARLA [32]
DiffVLA [121]	Jun 2025	VLM-Guided Driving Command Generation, Multi-modal Diffusion Planning	Senna-VLM [97]	NAVSIM [122]
ReasonPlan [123]	May 2025	Self-Supervised Next Scene Prediction, Supervised Decision Chain-of-Thought Reasoning	Qwen-0.5B [14]	Bench2Drive [103]

7) SeGPT [116]: 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.

8) CRITICAL [118]: 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.

9) ChatScene [120]: 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.

10) DiffVLA [121]: The framework uses Senna-VLM [97] to interpret multi-view images and output high-level driving commands, which are combined with a diffusion planner to produce future trajectories. This enhances realism and behavioural diversity of the scenarios generated, making it suitable for simulating complex interactions in AV testing. Evaluated on the NAVSIMv2 benchmark, the model achieves an extended PDM score of 45.0, demonstrating its effectiveness in generating reactive driving behaviours.

11) ReasonPlan [123]: It integrates a self-supervised Next Scene Prediction (NSP) task to enhance visual-language alignment and a supervised Decision Chain-of-Thought (DeCoT) process for interpretable planning. The DeCoT process also includes reasoning steps for not only scene understanding and traffic sign recognition, but also meta action prediction. The model is trained on the PDR dataset and achieves a Driving Score of 64.01 on the Bench2Drive benchmark and demonstrates strong zero-shot generalisation on the DriveOcclusionSim (DOS) benchmark.

Summary on LLM-Assisted Simulation and Scenario Generation: These models, particularly advanced versions like GPT-4 and LLaMA, enable the creation of diverse, human-like driving scenarios that were previously difficult to achieve. Moreover, VLMs 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 fidelities.

Several models focus on scenario generation and multi-agent consistency to create realistic traffic interactions. **CTG++** leverages conditional diffusion and spatial-temporal transformers to model traffic participant behaviours, and uses GPT-4 to generate traffic-compliant

scenarios. Similarly, **ChatScene** enhances safety-critical AV testing by breaking down unstructured language instructions into detailed, simulator-ready scenarios, ensuring comprehensive safety evaluations. **SurrealDriver** introduces a CoachAgent–DriverAgent framework, mimicking human driving behaviours to reduce collision rates and improve realism in urban simulations.

Other frameworks integrate LLMs with multimodal inputs to refine scenario realism. **LimSim++** ingests scenario information from SUMO and visual content from CARLA to support continuous learning and evaluation to improve scenario diversity. **ChatSim** enables editable, photo-realistic 3D scene simulations, by employing McNeRF for scene rendering and McLight for realistic lighting. **TrafficGPT** integrates Traffic Foundation Models (TFMs) with ChatGPT to analyse live traffic feeds, predict congestion, and suggest optimal routes, demonstrating dynamic traffic scenario management.

DiffVLA and **ReasonPlan** demonstrate enhanced planning. DiffVLA uses a VLM to derive driving commands from multi-view images, guiding a diffusion planner for trajectory generation. ReasonPlan employs a multimodal LLM with next-scene prediction and structured reasoning (DeCoT) to inform its trajectory output. Both showcase how visual perception combined with language reasoning improves interpretable driving behaviour.

These advancements underscore the growing role of VLMs in bridging the gap between simulated and real-world driving environments. By enabling richer scenario diversity, adaptive traffic modelling, and interactive simulation refinement, these models contribute to more robust AV validation via scenario generation.

B. World Models

In the context of AV data generation, a World Model refers to a cognitive representation of the environment that enables an AV to perceive, predict, and interact with its surroundings effectively. This model integrates sensor data, prior knowledge, and learnt patterns to simulate real-world dynamics, allowing for scenario-based safety validation and enhanced decision-making. World Models facilitate the creation of synthetic driving scenarios, reducing dependency on extensive real-world data collection while enabling scalable testing of edge cases. World Models generate synthetic driving scenarios, road networks, and 3D infrastructure by leveraging latent diffusion models (LDMs). 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.

It is crucial to note that we **only** include world models where natural language input is an option to the end-user. In this vein, we could not include a handful of World Models (**GEM** [130], for example).

1) Gaia-1 [24]: Researchers at Wayve developed GAIA-1, 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.

2) Magicdrive [125]: 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.

3) Drive-WM [127]: 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.

4) GenAD [5]: 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 enhancing simulation consistency.

5) ADriver-I [128]: 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.

6) DriveDreamer-2 [25]: 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

TABLE VIII
TABLE FOR WORLD MODEL PAPERS

Paper	Submission Time	Specific Tasks	LLM Model	Dataset or Simulator
Gaia-1 [24]	Sep, 2023	Text & Action Conditioning, Contextual Awareness (Road & World Rules), Scene & Dynamics Control	Encoded Text via T5-large Model [124]	Proprietary Driving Data (London, UK between 2019 and 2023) $\approx$ 4,700 hours at 25Hz.
Magicdrive [125]	Oct, 2023	BEV Segmentation, 3D Object Detection, Multi-Camera Consistency	Pre-trained CLIP Text Encoder [126]	nuScenes [50]
Drive-WM [127]	Nov, 2023	Flexible Integration with Existing End-to-End AV Planning Models	Pre-trained CLIP Text Encoder [126]	nuScenes [50]
GenAD [5]	Feb, 2024	Action-Conditioned Prediction & Planning, Zero-Shot Generalisation	Pre-trained CLIP Text Encoder [126]	nuScenes [50], OpenDV-2K Dataset [5]
ADriver-I [128]	Nov, 2023	Vision-Action Pairs unifying Visual Features and Control Signals, Adaptive Generation	Vicuna-7B-1.5 [74]	nuScenes [50]
DriveDreamer-2 [25]	Apr, 2024	Action-Conditioned Prediction & Planning, Zero-Shot Generalisation	Finetuned GPT-3.5 [1]	nuScenes [50]
FutureSightDrive [129]	May 2025	Spatio-Temporal Visual CoT Reasoning, Future Frame & Perception Prediction, Inverse Dynamics Planning	Qwen2-VL-2B [14]	nuScenes [50]

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.

7) FutureSightDrive [129]: A world model that uses a spatio-temporal CoT. This world model predicts future image frames with integrated lane dividers and 3D detections to enforce physical constraints, but can also act as an inverse dynamics model that uses these predictions for trajectory planning. Key innovations include a unified pre-training paradigm that activates visual generation in existing LLMs by expanding their vocabulary with VQ-VAE tokens, requiring minimal data. This visual reasoning approach avoids the information loss of text-based CoT. Evaluated on nuScenes, FutureSightDrive achieves strong performance in planning, frame generation, and scene understanding.

Summary on World Models: World Models leverage latent diffusion models (LDMs), multi-modal integration, and autoregressive transformers to generate realistic driving scenarios for data augmentation, safety validation, and synthetic testing in autonomous driving. These works can be broadly categorised based on their emphasis on scenario generation, prediction and planning, and multimodal integration.

Several models focus on high-fidelity scenario generation for training and validation purposes. For instance, **Gaia-1**'s LDM pipeline allows for text-conditioned scenario generation with precise control over vehicle dynamics and scene features. Similarly, **MagicDrive** specialises in 3D geometry-aware scene generation, improving BEV segmentation and 3D object detection by maintaining multi-camera consistency and fine-tuned object placement.

Other frameworks emphasise predictive capabilities and trajectory simulation. **Drive-WM** integrates temporal and multi-view encoding to predict future events in AV scenarios, supporting zero-shot planning by generating image-based reward functions for future trajectory validation. **GenAD**, on the other hand, applies causal temporal and spatial attention mechanisms to model dynamic scene interactions, excelling in action-conditioned video prediction for more realistic future state forecasting. **FutureSightDrive** extends this direction by introducing a spatio-temporal visual CoT, where the VLM itself acts as a world model to generate future image frames, which then serve as intermediate reasoning steps for inverse dynamics trajectory planning.

The integration of VLM World Models with multimodal data to enhance contextual understanding and planning can be observed. **ADriver-I** unifies visual perception and control signals into a world model that not only processes vision-action pairs but also predicts future driving scenarios using Stable Diffusion 2.1. Meanwhile, **DriveDreamer-2** extends world models by incorporating HDMaps, 3D bounding boxes, and finetuned GPT-3.5 to generate both synthetic driving videos and agent trajectories based on textual prompts.

However, generating long-term temporally consistent video sequences continues to be computationally expensive, and diffusion models require significant training data to maintain realism across diverse scenarios. Additionally, the reliance on large-scale private datasets (e.g., nuScenes) limits accessibility and generalisation across varied driving environments.

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

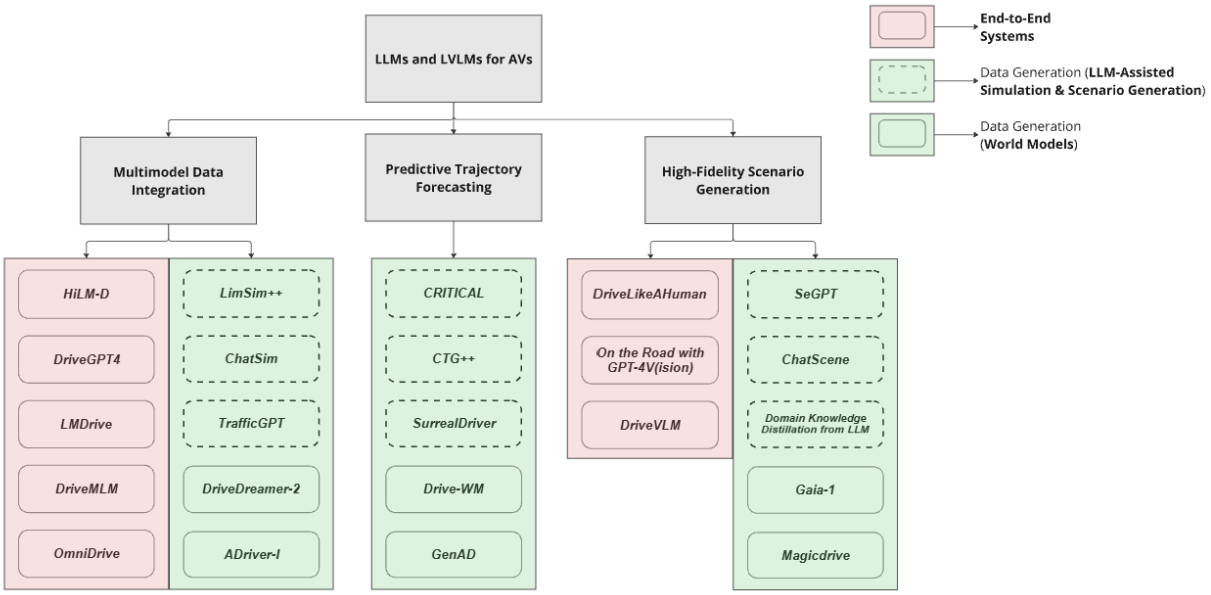

Fig. 6. Academic literature related to End-to-End VLM models and Data Generation VLM models relevant to AV applications can be predominantly bucketed into three emergent subclasses, those being: Multimodel Data Integration, Predictive Trajectory Forecasting, and High-Fidelity Scenario Generation. This fact can be gleaned from the summaries of the respective sections. We encourage AV research hubs to explore more subclasses, for example: Adversarial and Robustness Testing, Multi-Agent Coordination and Interaction Synthesis, Uncertainty-Aware Risk Assessment, etc.

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 [32], and HighwayEnv [41]. They are integral to advancing the field of AV by providing robust environments for comprehensive testing and development.

1) CARLA [32]: 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.

2) HighwayEnv [41]: 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 [50] (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,

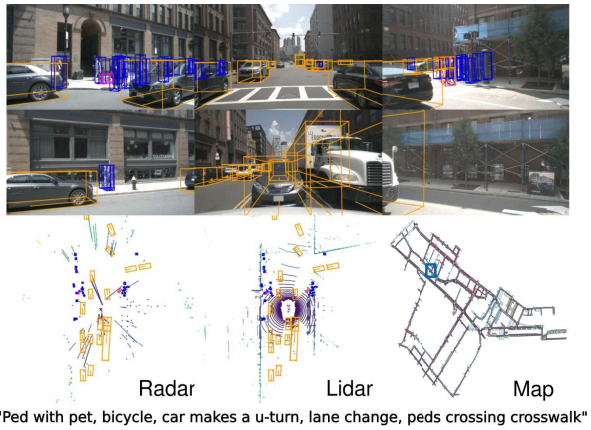

Fig. 7. Examples of the NuScenes Dataset. Adapted from [50].

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.

In addition to these foundational datasets, newer ones like NuScenes-QA [11] and OmniDrive-nuScenes [95] expand the scope to cognitive tasks, such as visual question answering and counterfactual reasoning. Benchmarks like BDD-X [34], which focuses on interpretability and reasoning, and DRAMA [79], 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.

1) BDD-X [34]: A benchmark dataset for evaluating the interpretability and reasoning capabilities of AV systems,

TABLE IX
TABLE FOR BENCHMARKS AND DATASETS

Dataset Name	Submission Time	Specific Tasks	Data Format	Datasize
BDD-X [34]	Jul, 2018	Introspective and rationalization explanations for vehicle behavior	Video, text	6,984 videos 77 hours 26,228 annotations
NuScenes [50]	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 [79]	Sep, 2022	Joint risk localization and captioning in driving scenes	Video, text	17,785 scenarios 91 hours 35,038 visual attributes
NuScenes-QA [11]	May, 2023	Visual question answering for AV	Images, LiDAR, Q&A pairs	34K scenes 460K Q&A pairs
NuPrompt [131]	Sep, 2023	Object-centric language prompts for 3D driving scenes	Images, text	35,367 prompts 188,445 instances
Driving QA Dataset [132]	Oct, 2023	Question answering for driving scenarios, control commands	Text, control commands	160K Q&A pairs 10K driving scenarios
Talk2BEV-Bench [28]	Oct, 2023	Evaluates LVLMs for instance attributes, counting, spatial reasoning	BEV images, Q&A pairs	1,000 scenes 20,000 Q&A pairs
Reason2Drive [133]	Dec, 2023	Chain-based reasoning for AV	Video, text	600K video-text pairs
LaMPilot [134]	Dec, 2023	User instruction following, code generation	Semi-human annotated traffic scenes	4,900 scenarios
LangAuto [91]	Dec, 2023	Closed-loop driving, language instruction following	Multi-modal sensor data, navigation and notice instructions	64K data clips
LingoQA [69]	Dec, 2023	Video question answering for AV	Video, Q&A pairs	28K scenarios 419.9K Q&A pairs
SUP-AD [94]	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 [135]	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 [95]	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
VLADBench [136]	Mar 2025	Benchmarking, Planning, QA Evaluation	GPT-4, Qwen-VL	VLADBench (Carla, nuScenes, BDD-X)

featuring diverse scenarios with textual descriptions, questions, and answers to enhance systems' understanding and explanations.

2) NuScenes [50]: Is a comprehensive dataset for AV driving, providing real-world driving data with multi-sensor information, aiding in the development and testing of perception and decision-making systems in various scenarios.

3) DRAMA [79]: 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.

4) NuScenes-QA [11]: 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.

5) NuPrompt [131]: 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.

6) Driving QA [132]: Consists 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.

7) Talk2BEV-Bench [28]: 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.

8) Reason2Drive [133]: 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.

9) LaMPilot [134]: LaMPilot integrates LLMs into AVs for interpreting user commands via code generation. The LaMPilot-Bench evaluates LLM-based agents across various driving scenarios for adaptability and accuracy.

10) LangAuto [91]: Benchmarks AV systems' ability to follow natural language instructions with dynamically updated navigation commands, testing understanding and execution in dynamic environments.

11) LingoQA [69]: 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.

12) SUP-AD [94]: 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.

13) CODA-LM [135]: The first benchmark designed for the automated and systematic evaluation of VLMs 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.

14) OmniDrive-nuScenes [95]: Enhances NuScenes with 3D spatial understanding and counterfactual reasoning tasks, assessing autonomous systems' decision-making and planning abilities through simulated trajectories and outcomes.

Summary on Platforms, Benchmarks and Datasets Simulation platforms like CARLA and datasets such as NuScenes and Waymo Open Dataset are foundational for advancing 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.

Although a direct comparison of representative studies on a common benchmark would be ideal, the diversity and specificity of datasets used across the surveyed literature currently limit this possibility. To mitigate this limitation, we summarise key methodologies, datasets, evaluation metrics, and reported results in structured tables throughout the manuscript. This approach enables readers to intuitively compare and understand the relative strengths, weaknesses, and applicability of each methodology discussed.

VIII. Challenges and Future Research Directions

The integration of LLMs and VLMs into AVs has advanced various aspects of autonomous driving, including adaptive decision-making, trajectory prediction, traffic accident forecasting, human-vehicle interaction, and targeted data generation. These advances have redefined how industry and research institutions approach the dynamic driving task (DDT), regardless of integration strategies such as modular, end-to-end, or data-centric approaches. However, several critical challenges remain before regulatory agencies can broadly authorise these technologies for civilian or commercial deployment. Drawing on insights from prior sections, this section synthesises major challenges and identifies future research avenues to ensure the safe, efficient, and ethically sound deployment of VLM-powered AV systems.

A. Computational Trade-Offs

Computational trade-offs for deploying VLMs in AVs primarily revolve around balancing inference efficiency, accuracy, and real-time responsiveness.

A critical bottleneck when deploying VLMs is their prohibitively long inference time, which, as evidenced by models like DriVLMe [98], can extend to several seconds and severely disrupt the closed-loop control required for safe operation at high environment sampling frequencies. A primary focus is on efficient visual token compression, particularly for the high-dimensional input from multi-camera surround-view systems. Initiatives such as the Senna-VLM's [97] Driving Vision Adapter directly address this by avoiding simple feature projections; they perform aggressive encoding and compression of image tokens to curb sequence length while preserving critical spatial and contextual information for driving scenarios. This naturally reduces inference latency in transformer-based models. Chain-of-Thought (CoT) is computationally expensive by design since recursive prompt calls are inherent in CoT, highlighting a known trade-off between its reasoning and temporal cost, and encouraging investigations into optimally truncating CoT recursion. The selection of a computationally small model, such as WiseA [100], provides a foundational advantage for on-vehicle deployment. However, a common interim solution involves temporal decoupling, where the VLM operates at a deliberately reduced frequency (2 Hz), offloading high-rate control to a low-level motion planner and leveraging available historical context to compensate.

From a hardware perspective, the foundational training of most of these VLMs is conducted on NVIDIA A100s, NVIDIA A800s, or Nvidia RTX 4090 GPUs, which provide the necessary computational throughput and memory. However, for deployment on energy-constrained vehicle hardware, the models are severely optimised to run on embedded systems-on-a-chip (SoCs) like the NVIDIA OrinX. Parameter-Efficient Fine-Tuning (PEFT) methods like Low-Rank Adaptation (LoRA) and Quantised (Q)LoRA, which freeze most of the original model and only train small adaptable components, further ease computational and memory load. Additionally, AVs are often designed to operate asynchronously within a dual-processor setup, where a high-frequency controller handles immediate vehicle dynamics on one chip, while the VLM provides higher-level, slower reasoning on a secondary chip.

The culmination of these hardware- and architectural-focused optimisations is the compression of inference latency to under 500 milliseconds, making it feasible for integrated,

real-time automotive applications. Future efforts should specifically explore adaptive model architectures and optimised hardware-software co-design strategies, thus achieving an optimal equilibrium between computational efficiency, accuracy, and cost-effectiveness for real-world AV deployment.

B. Hallucinations

Hallucinations refer to the generation of outputs by VLMs that are factually incorrect or nonsensical. This issue is particularly critical in AVs, where accurate and reliable information is paramount for safety. Current VLMs can occasionally produce hallucinations when interpreting complex driving scenarios or making decisions based on incomplete or ambiguous data. For example, an VLM might misinterpret a visual input, leading to incorrect obstacle detection. Although reinforcement learning from human feedback (RLHF) and game-theoretic strategy fusion [137] have been applied to reduce hallucinations in AVs, studies explicitly addressing hallucination in VLMs for AVs remain scarce. One possible reason is that current research in this domain is still largely exploratory, focused on showcasing the potential of VLMs for driving scene understanding, captioning, and interactive explanation, rather than ensuring their robustness for deployment. In many cases, VLMs are used as auxiliary tools rather than as components of the safety-critical decision-making pipeline, which may deprioritise concerns such as hallucination in early-stage work. Moreover, some researchers may assume that existing multi-sensor fusion frameworks, already widely adopted in AV perception stacks, could inherently mitigate hallucination risks by cross-validating information from LiDAR, radar, and cameras. Recent work has begun to address this challenge more directly. For instance, Nullu [138], a training-free method that mitigates object hallucinations by identifying a "hallucination subspace" in the model's feature space and projecting model weights to its null space. This approach effectively suppresses the model's internal prior biases that lead to hallucinations while adding zero inference overhead, making it particularly suitable for real-time AV applications.

Nonetheless, 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. Specialised hallucination mitigation frameworks for AV-specific VLMs, combined with multi-modal sensor fusion and AV-focused benchmarks, will be crucial for ensuring robustness and safety in real-world deployments.

C. Latency

Real-time decision-making is crucial for the safety and efficiency of autonomous driving. However, the computational complexity of VLMs can introduce delays that are unacceptable in dynamic driving environments. High latency can result in delayed responses to critical situations, potentially leading to accidents. Recent efforts have explored architectural and algorithmic innovations to reduce inference latency. Techniques such as persistent batching, blocked key-value caches, tensor parallelism, and optimised CUDA kernels have shown promise in accelerating vision-language inference pipelines [139]. Additional methods, including patch selection, token compression, and speculative decoding, have been proposed to reduce the computational footprint without compromising output quality [140]. Nonetheless, achieving consistently low-latency inference in real-world, resource-constrained AV settings remains an open challenge. Future research must therefore 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.

D. Ethical and Regulatory Considerations

Current regulatory frameworks for AVs, such as ISO 26262, primarily address aspects of functional safety and cybersecurity within conventional rule-based systems. However, these regulations do not fully capture the complexities introduced by dynamic, adaptive, and learning-based decision-making approaches that characterise VLMs [141]. To enhance global regulatory compliance and ensure AV adaptability across varying local traffic norms and laws, future guidelines should incorporate criteria that recognise and accommodate the flexible reasoning capacities inherent to VLM-driven systems.

Transparency and interpretability in AV systems have been increasingly emphasised due to their critical role in regulatory approval, accountability, and fostering public trust. Recent advancements, such as the integration of LLM-based reasoning and traffic rule extraction into AV decision-making frameworks [57] and graph-based topology reasoning methods [142], represent significant steps towards enhancing transparency. However, further improvements are necessary to systematically embed explainability into VLM architectures. Techniques from explainable artificial intelligence (XAI), including SHAP (Shapley Additive Explanations), could be effectively integrated to provide clear, human-readable rationales for AV decisions, particularly in complex or ambiguous scenarios.

Finally, ethical considerations in AV decision-making extend beyond extreme moral dilemmas, such as the widely discussed trolley problem, to everyday traffic scenarios involving nuanced risk distribution among road users. Recent empirical findings illustrate that human drivers often accept personal risks to protect others, reflecting altruistic tendencies currently absent in many AV algorithms [143]. Thus, AV decision-making frameworks utilising VLMs should be developed to mirror these human-like risk redistribution behaviours.

E. Vehicle-to-Everything (V2X) Communication

V2X communication involves the exchange of information between vehicles and various entities like infrastructure, pedestrians, and other vehicles. Integrating VLMs into V2X communication frameworks presents an opportunity to

enhance the context-awareness and decision-making capabilities of AVs. For instance, VLMs can interpret complex messages and predict traffic patterns based on shared data. Future research should investigate the integration of VLMs 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 Language Models Large Vision-Language Models (LLMs and VLMs) in autonomous vehicles (AVs), focusing on following key areas: modular integration, end-to-end integration, data generation, evaluation platforms, datasets, and benchmarks. We have examined the current state of research, highlighting significant advancements and innovations that VLMs 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 VLMs 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, as well as offering insights into computational trade-offs.

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 VLMs in autonomous driving systems. By addressing these issues, we can unlock the full potential of VLMs, 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 et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[2] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 1–11.

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

[4] C. Sima et al., "DriveLM: Driving with graph visual question answering," 2023, *arXiv:2312.14150*.

[5] J. Yang et al., "Generalized predictive model for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 14662–14672.

[6] 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 *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2024, pp. 902–909.

[7] Z. Yang, X. Jia, H. Li, and J. Yan, "LLM4Drive: A survey of large language models for autonomous driving," 2023, *arXiv:2311.01043*.

[8] H. Gao, Z. Wang, Y. Li, K. Long, M. Yang, and Y. Shen, "A survey for foundation models in autonomous driving," 2024, *arXiv:2402.01105*.

[9] X. Li et al., "Towards knowledge-driven autonomous driving," 2023, *arXiv:2312.04316*.

[10] Z. Zhang et al., "Large language models for mobility analysis in transportation systems: A survey on forecasting tasks," 2024, *arXiv:2405.02357*.

[11] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y. Jiang, "NuScenes-QA: A multi-modal visual question answering benchmark for autonomous driving scenario," in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 4542–4550.

[12] W. Xin Zhao et al., "A survey of large language models," 2023, *arXiv:2303.18223*.

[13] H. Touvron et al., "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[14] J. Bai et al., "Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond," 2023, *arXiv:2308.12966*.

[15] D. A. Pomerleau, "ALVINN: An autonomous land vehicle in a neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 1, 1988, pp. 1–9.

[16] M. Bojarski et al., "End to end learning for self-driving cars," 2016, *arXiv:1604.07316*.

[17] F. Codevilla, M. Müller, A. López, V. Koltun, and A. Dosovitskiy, "End-to-end driving via conditional imitation learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 4693–4700.

[18] V. Mnih et al., "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1928–1937.

[19] Y. Hu et al., "Planning-oriented autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 17853–17862.

[20] A. Ramesh et al. (2021). *DALL-E: Creating Images From Text*. Accessed: Jul. 9, 2024. [Online]. Available: https://github.com/openai/DALL-E

[21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. (2022). *High-Resolution Image Synthesis With Latent Diffusion Models*. Accessed: Jul. 9, 2024. [Online]. Available: https://github.com/CompVis/stable-diffusion

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

[23] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," 2021, *arXiv:2105.05233*.

[24] A. Hu et al., "GAIA-1: A generative world model for autonomous driving," 2023, *arXiv:2309.17080*.

[25] G. Zhao et al., "DriveDreamer-2: LLM-enhanced world models for diverse driving video generation," 2024, *arXiv:2403.06845*.

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

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

[28] T. Choudhary et al., "Talk2BEV: Language-enhanced bird's-eye view maps for autonomous driving," 2023, *arXiv:2310.02251*.

[29] Q. Xu, S. Chen, G. Chen, Y. Wang, and Y. Zhang, "ChatBEV: A visual language model that understands BEV maps," 2025, *arXiv:2503.13938*.

[30] W. Liu, J. Zhang, B. Zheng, Y. Hu, Y. Lin, and Z. Zeng, "X-driver: Explainable autonomous driving with vision-language models," 2025, *arXiv:2505.05098*.

[31] OpenAI. (2023). *ChatGPT 3.5*. Accessed: Jul. 9, 2024. [Online]. Available: https://platform.openai.com/docs/models/gpt-3-5

[32] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. 1st Annu. Conf. Robot Learn.*, vol. 78, 2017, pp. 1–16.

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

[34] J. Kim et al., "Textual explanations for self-driving vehicles," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 563–578.

[35] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023, pp. 19730–19742.

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

[37] W. Dai et al., "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," in *Proc. 37th Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2023, pp. 1–11.

[38] 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 *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2023, pp. 5154–5161.

[39] G. Brockman et al., "OpenAI gym," 2016, *arXiv:1606.01540*.

[40] L. Wen et al., "DiLu: A knowledge-driven approach to autonomous driving with large language models," 2023, *arXiv:2309.16292*.

[41] E. Leurent. (2018). *An Environment for Autonomous Driving Decision-Making*. Accessed: Aug. 9, 2024. [Online]. Available: https://github.com/eleurent/highway-env

[42] Y. Wang et al., "Empowering autonomous driving with large language models: A safety perspective," 2023, *arXiv:2312.00812*.

[43] H. Sha et al., "LanguageMPC: Large language models as decision makers for autonomous driving," 2023, *arXiv:2310.03026*.

[44] Y. Liu, Q. Zhang, and D. Zhao, "A reinforcement learning benchmark for autonomous driving in intersection scenarios," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2021, pp. 1–8.

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

[46] HuggingFace H4 Team. (2024). *Zephyr 7b Alpha*. Accessed: Jul. 9, 2024. [Online]. Available: https://huggingface.co/HuggingFaceH4/zephyr-7b-alpha

[47] National Highway Traffic Safety Administration. (2020). *Fatality Analysis Reporting System (FARS) 2020*. Accessed: Jul. 9, 2024. [Online]. Available: https://static.nhtsa.gov/nhtsa/downloads/FARS/2020/National/FARS2020NationalCSV.zip

[48] L. Chen et al., "Driving with LLMs: Fusing object-level vector modality for explainable autonomous driving," 2023, *arXiv:2310.01957*.

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

[50] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.

[51] K. Tanahashi et al., "Evaluation of large language models for decision making in autonomous driving," 2023, *arXiv:2312.06351*.

[52] F. Chi, Y. Wang, P. Nasiopoulos, and V. C. M. Leung, "Multi-modal GPT-4 aided action planning and reasoning for self-driving vehicles," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2024, pp. 7325–7329.

[53] 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 Intell. Transp. Syst. Mag.*, vol. 16, no. 4, pp. 81–94, Jul. 2024.

[54] R. Zhao et al., "Sce2DriveX: A generalized MLLM framework for scene-to-drive learning," 2025, *arXiv:2502.14917*.

[55] K. Qian et al., "AgentThink: A unified framework for tool-augmented chain-of-thought reasoning in vision-language models for autonomous driving," 2025, *arXiv:2505.15298*.

[56] J. Mao, Y. Qian, J. Ye, H. Zhao, and Y. Wang, "GPT-driver: Learning to drive with GPT," 2023, *arXiv:2310.01415*.

[57] B. Li et al., "Driving everywhere with large language model policy adaptation," 2024, *arXiv:2402.05932*.

[58] H. Caesar et al., "NuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles," 2021, *arXiv:2106.11810*.

[59] B. Yang et al., "Diffusion-ES: Gradient-free planning with diffusion for autonomous driving and zero-shot instruction following," 2024, *arXiv:2402.06559*.

[60] M. Peng, X. Guo, X. Chen, M. Zhu, and K. Chen, "LC-LLM: Explainable lane-change intention and trajectory predictions with large language models," 2024, *arXiv:2403.18344*.

[61] 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 *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 2118–2125.

[62] T. Xu, H. Lu, X. Yan, Y. Cai, B. Liu, and Y. Chen, "Occ-LLM: Enhancing autonomous driving with occupancy-based large language models," 2025, *arXiv:2502.06419*.

[63] S. Wang, Y. Zhu, Z. Li, Y. Wang, L. Li, and Z. He, "ChatGPT as your vehicle co-pilot: An initial attempt," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 12, pp. 4706–4721, Dec. 2023.

[64] S. Documentation. (2020). *Simulation and Model-Based Design*. Accessed: Sep. 12, 2024. [Online]. Available: https://www.mathworks.com/products/simulink.html

[65] R. Johansson, D. Williams, A. Berglund, and P. Nugues, "Carsim: A system to visualize written road accident reports as animated 3D scenes," in *Proc. 2nd Workshop Text Meaning Interpretation-TextMean*, 2004, pp. 57–64.

[66] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[67] M. N. Team. (2023). *Introducing MPT-7b: A New Standard for Open-Source, Commercially Usable LLMs*. Accessed: May 5, 2023. [Online]. Available: https://www.mosaicml.com/blog/mpt-7b

[68] C. Xu, J. Liu, Y. Guo, Y. Zhang, P. Hang, and J. Sun, "Towards human-centric autonomous driving: A fast-slow architecture integrating large language model guidance with reinforcement learning," 2025, *arXiv:2505.06875*.

[69] A.-M. Marcu et al., "LingoQA: Visual question answering for autonomous driving," 2023, *arXiv:2312.14115*.

[70] S. Yang et al., "LiDAR-LLM: Exploring the potential of large language models for 3D LiDAR understanding," 2023, *arXiv:2312.14074*.

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

[72] A. Ishaq et al., "DriveLMM-o1: A Step-by-step reasoning dataset and large multimodal model for driving scenario understanding," 2025, *arXiv:2503.10621*.

[73] Z. Zhang et al., "MPDrive: Improving spatial understanding with marker-based prompt learning for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 12089–12099.

[74] W.-L. Chiang et al. (2023). *Vicuna: An Open-Source Chatbot Impressing GPT-4 With 90%* ChatGPT Quality*. Accessed: Apr. 14, 2023. [Online]. Available: https://vicuna.lmsys.org

[75] J. Cho, J. Lei, H. Tan, and M. Bansal, "Unifying vision-and-language tasks via text generation," in *Proc. 38th Int. Conf. Mach. Learn.*, vol. 139, Jul. 2021, pp. 1931–1942.

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

[77] D. Fu et al., "Drive like a human: Rethinking autonomous driving with large language models," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2024, pp. 910–919.

[78] 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," 2023, *arXiv:2309.05186*.

[79] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, and J. Li, "DRAMA: Joint risk localization and captioning in driving," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 1043–1052.

[80] Z. Xu et al., "DriveGPT4: Interpretable end-to-end autonomous driving via large language model," 2023, *arXiv:2310.01412*.

[81] L. Wen et al., "On the road with GPT-4V(ision): Early explorations of visual-language model on autonomous driving," 2023, *arXiv:2311.05332*.

[82] (2023). *GPT-4v(Ision) System Card*. Accessed: Nov. 16, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:263218031

[83] P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 2446–2454.

[84] K. Li et al., "CODA: A real-world road corner case dataset for object detection in autonomous driving," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2022, pp. 406–423.

[85] Z. Che et al., "D^2-city: A large-scale dashcam video dataset of diverse traffic scenarios," 2019, *arXiv:1904.01975*.

[86] W. Bao, Q. Yu, and Y. Kong, "Uncertainty-based traffic accident anticipation with spatio-temporal relational learning," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2682–2690.

[87] B. J. University. *Chinese Traffic Sign Database*. Accessed: Nov. 29, 2024. [Online]. Available: http://www.nlpr.ia.ac.cn/pal/trafficdata/detection.html

[88] Z. Wu, X. Chen, H. Wei, F. Song, and T. Xu, "ADD: An automatic desensitization fisheye dataset for autonomous driving," *Eng. Appl. Artif. Intell.*, vol. 126, Nov. 2023, Art. no. 106766.

[89] H. Yu et al., "DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 21361–21370.

[90] 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," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2678, no. 4, pp. 606–621, Apr. 2024.

[91] 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," 2023, *arXiv:2312.07488*.

[92] W. Wang et al., "DriveMLM: Aligning multi-modal large language models with behavioral planning states for autonomous driving," 2023, *arXiv:2312.09245*.

[93] Y. Fang et al., "EVA: Exploring the limits of masked visual representation learning at scale," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19358–19369.

[94] X. Tian et al., "DriveVLM: The convergence of autonomous driving and large vision-language models," 2024, *arXiv:2402.12289*.

[95] S. Wang et al., "Omnidrive: A holistic LLM-agent framework for autonomous driving with 3D perception, reasoning and planning," 2024, *arXiv:2405.01533*.

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

[97] B. Jiang et al., "Senna: Bridging large vision-language models and end-to-end autonomous driving," 2024, *arXiv:2410.22313*.

[98] Y. Huang, J. Sansom, Z. Ma, F. Gervits, and J. Chai, "DriVLMe: Enhancing LLM-based autonomous driving agents with embodied and social experiences," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2024, pp. 3153–3160.

[99] Z. Ma et al., "DOROTHIE: Spoken dialogue for handling unexpected situations in interactive autonomous driving agents," 2022, *arXiv:2210.12511*.

[100] S. Zhang, W. Huang, Z. Gao, H. Chen, and C. Lv, "WiseAD: Knowledge augmented end-to-end autonomous driving with vision-language model," 2024, *arXiv:2412.09951*.

[101] P. Liu, H. Liu, H. Liu, X. Liu, J. Ni, and J. Ma, "VLM-E2E: Enhancing end-to-end autonomous driving with multimodal driver attention fusion," 2025, *arXiv:2502.18042*.

[102] H. Fu et al., "ORION: A holistic end-to-end autonomous driving framework by vision-language instructed action generation," 2025, *arXiv:2503.19755*.

[103] X. Jia, Z. Yang, Q. Li, Z. Zhang, and J. Yan, "Bench2Drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving," 2024, *arXiv:2406.03877*.

[104] B. Jiang, S. Chen, Q. Zhang, W. Liu, and X. Wang, "AlphaDrive: Unleashing the power of VLMs in autonomous driving via reinforcement learning and reasoning," 2025, *arXiv:2503.07608*.

[105] X. Zhou, X. Han, F. Yang, Y. Ma, and A. C. Knoll, "OpenDriveVLA: Towards end-to-end autonomous driving with large vision language action model," 2025, *arXiv:2503.23463*.

[106] S. Bai et al., "Qwen2.5-VL technical report," 2025, *arXiv:2502.13923*.

[107] X. Chen, L. Huang, T. Ma, R. Fang, S. Shi, and H. Li, "SOLVE: Synergy of language-vision and end-to-end networks for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 12068–12077.

[108] Z. Zhong et al., "Language-guided traffic simulation via scene-level diffusion," in *Proc. Conf. Robot Learn.*, 2023, pp. 144–177.

[109] 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 *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2023, pp. 3893–3900.

[110] ASAM. (2023). *Asam Openxontology*. Accessed: Sep. 22, 2024. [Online]. Available: https://www.asam.net/standards/asam-openxontology

[111] Y. Jin et al., "Surrealdriver: Designing generative driver agent simulation framework in urban contexts based on large language model," 2023, *arXiv:2309.13193*.

[112] D. Fu et al., "LimSim++: A closed-loop platform for deploying multimodal LLMs in autonomous driving," 2024, *arXiv:2402.01246*.

[113] D. Krajzewicz, G. Hertkorn, C. Rössel, and P. Wagner, "SUMO (simulation of urban mobility)—An open-source traffic simulation," in *Proc. 4th Middle East Symp. Simul. Model.*, 2002, pp. 183–187.

[114] Y. Wei et al., "Editable scene simulation for autonomous driving via collaborative LLM-agents," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 15077–15087.

[115] S. Zhang et al., "TrafficGPT: Viewing, processing and interacting with traffic foundation models," *Transp. Policy*, vol. 150, pp. 95–105, May 2024.

[116] 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 Trans. Intell. Vehicles*, vol. 9, no. 3, pp. 4422–4431, Mar. 2024.

[117] W. Zhan et al., "INTERACTION dataset: An INTERnational, adversarial and cooperative moTION dataset in interactive driving scenarios with semantic maps," 2019, *arXiv:1910.03088*.

[118] 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," 2024, *arXiv:2404.08570*.

[119] A. Q. Jiang et al., "Mistral 7B," 2023, *arXiv:2310.06825*.

[120] J. Zhang, C. Xu, and B. Li, "ChatScene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 15459–15469.

[121] A. Jiang et al., "DiffVLA: Vision-language guided diffusion planning for autonomous driving," 2025, *arXiv:2505.19381*.

[122] D. Dauner et al., "NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, 2024, pp. 28706–28719.

[123] X. Liu et al., "ReasonPlan: Unified scene prediction and decision reasoning for closed-loop autonomous driving," 2025, *arXiv:2505.20024*.

[124] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2022.

[125] R. Gao et al., "MagicDrive: Street view generation with diverse 3D geometry control," 2023, *arXiv:2310.02601*.

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

[127] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 14749–14759.

[128] F. Jia et al., "ADriver-I: A general world model for autonomous driving," 2023, *arXiv:2311.13549*.

[129] S. Zeng et al., "FutureSightDrive: Thinking visually with spatio-temporal CoT for autonomous driving," 2025, *arXiv:2505.17685*.

[130] M. Hassan et al., "GEM: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control," 2024, *arXiv:2412.11198*.

[131] D. Wu et al., "Language prompt for autonomous driving," 2023, *arXiv:2309.04379*.

[132] L. Chen et al., "Driving with LLMs: Fusing object-level vector modality for explainable autonomous driving," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2024, pp. 14093–14100.

[133] M. Nie et al., "Reason2Drive: Towards interpretable and chain-based reasoning for autonomous driving," 2023, *arXiv:2312.03661*.

[134] Y. Ma et al., "LaMPilot: An open benchmark dataset for autonomous driving with language model programs," 2023, *arXiv:2312.04372*.

[135] K. Chen et al., "Automated evaluation of large vision-language models on self-driving corner cases," 2024, *arXiv:2404.10595*.

[136] Y. Li et al., "Fine-grained evaluation of large vision-language models in autonomous driving," 2025, *arXiv:2503.21505*.

[137] Y. Feng, K. Zhang, H. Ouchouid, A. Kaniamparambil, I. Souflas, and P. Angeloudis, "Strategic fusion of vision language models: Shapley-credited context-aware dawid-skene for multi-label tasks in autonomous driving," 2025, *arXiv:2510.01126*.

[138] L. Yang, Z. Zheng, B. Chen, Z. Zhao, C. Lin, and C. Shen, "Nullu: Mitigating object hallucinations in large vision-language models via HalluSpace projection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 14635–14645.

[139] C. Cui et al., "On-board vision-language models for personalized autonomous vehicle motion control: System design and real-world validation," 2024, *arXiv:2411.11913*.

[140] J. Huang, Y. Jin, L. An, and J. Park, "LiteVLM: A low-latency vision-language model inference pipeline for resource-constrained environments," 2025, *arXiv:2506.07416*.

[141] E. Nassif, H. Tian, E. Candela, Y. Feng, P. Angeloudis, and W. Y. Ochieng, "Safety standards for autonomous vehicles: Challenges and way forward," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2023, pp. 3004–3009.

[142] T. Li et al., "Graph-based topology reasoning for driving scenes," 2023, *arXiv:2304.05277*.

[143] S. Krügel and M. Uhl, "The risk ethics of autonomous vehicles: An empirical approach," *Sci. Rep.*, vol. 14, no. 1, p. 960, Jan. 2024.

Hanlin Tian (Graduate Student Member, IEEE) received the B.Eng. degree in computer science from Shandong University and the M.Sc. degree in computer engineering from New York University. He is currently a Postgraduate Researcher with the Centre for Transport Engineering and Modelling, Imperial College London. His research interests include computer vision and autonomous vehicles.

Kethan Reddy received the M.Phys. degree in physics from the University of Kent and the M.Sc. degree in machine learning from University College London. He is currently a Research Associate with the Centre for Transport Engineering and Modelling, Imperial College London. His research interests include artificial intelligence for robotics and autonomous vehicles.

Yuxiang Feng (Member, IEEE) received the B.Eng. degree in mechanical engineering from Tongji University and the M.Sc. degree in mechatronics and Ph.D. degree in automotive engineering from the University of Bath. He is currently a Research Fellow and the Laboratory Manager with the Centre for Transport Engineering and Modelling, Imperial College London. His 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 Loughborough University in 2006 as a Lecturer, where he has been a Professor of intelligent transport systems (ITS) since 2013. In 2021, he moved to Imperial College London as the Chair Professor of ITS. He is currently with the Chair in Intelligent Transport Systems, Imperial College London. 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.

Yiannis Demiris (Senior Member, IEEE) received the B.Sc. degree (Hons.) 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 currently a Professor of human-centred robotics with Imperial College London, where he holds the Royal Academy of Engineering Chair of Emerging Technologies (Personal Assistive Robotics) and is the Head of the Personal Robotics Laboratory. His current research interests include human–robot interaction, machine learning, user modeling, and assistive robotics. He is a fellow of the Institution of Engineering and Technology (IET) and the British Computer Society (BCS).

Panagiotis Angeloudis (Member, IEEE) received the Ph.D. degree in transportation from Imperial College London. He is currently a Professor of transport systems and logistics (TSL) with Imperial College London. Before establishing TSL, he held a JSPS Research Fellowship at Kyoto University. He spent periods as a Research Analyst at DP World and the United Nations in Geneva. His research interests include the study of networks, optimization methods, and multi-agent systems, as well as their applications in autonomous transport systems, urban infrastructure, and logistics.