Engineering Applications of Artificial Intelligence 176 (2026) 114805

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Survey paper

Foundation models for autonomous driving: A comprehensive survey

Sonda Fourati [a], Wael Jaafar [a,*], Noura Baccar [b], Safwan Alfattani [c], Rami Langar [a]

[a] École de technologie supérieure (ÉTS), University of Quebec, 1100, Notre-Dame West, Montreal, H3C 1K3, Quebec, Canada
[b] Mediterranean Institute of Technology (MedTech), Campus MSB-MedTech, Les Jardins du Lac II, 1053 Tunis, Tunisia
[c] King AbdulAziz University (KAU), Rabigh, 25732, Saudi Arabia

ARTICLE INFO

Keywords:
Large Language Models
Vision-Language Models
Multimodal large language models
Vision Foundation Models
Foundation Models
Cross-modal Language Models
Autonomous Driving Systems
Reinforcement Learning from Human Feedback
Prompt engineering
Safety alignment and verification
Edge deployment for real-time inference
Perception, prediction, planning, and control
Decision making and planning
Datasets and simulators implemented artificial intelligence
Application of artificial intelligence

ABSTRACT

Large Language Models (LLMs) have showcased remarkable proficiency in various information-processing tasks. They excel at data extraction, literature summarization, content generation, predictive modeling, decision-making, and system control. Moreover, Vision-Language Models (VLMs) and Multimodal LLMs (MLLMs), collectively referred to in this work as Cross-modal Language Models (XLMs), integrate multiple data modalities with language understanding, thereby advancing Autonomous Driving Systems (ADS). On the implemented Artificial Intelligence (AI) side, we analyze core techniques such as prompt engineering, supervised fine-tuning, reinforcement learning from human feedback, knowledge distillation, quantization and pruning, and safety alignment/verification, together with edge-aware deployment strategies. On the application of AI side, we map XLMs capabilities to the driving stack, including perception, prediction, planning, control, and human–machine interaction/vehicle-to-everything, and summarize how XLMs improve scene understanding, intent forecasting, decision-making, and closed-loop control by coupling natural-language reasoning with multimodal sensory inputs, such as panoramic images, Light Detection and Ranging (LiDAR), and radar. In this survey, we synthesize the state of XLMs for ADS: we review the relevant literature on ADS and XLMs, including their architectures, tools, and frameworks. We then compare deployment approaches across the driving stack and summarize datasets, simulators, and benchmarks for both open- and closed-loop evaluation. Finally, we analyze key challenges, such as grounding and hallucination, long-tail robustness, real-time and resource constraints, safety alignment and verification, and data governance and privacy, and outline research directions toward safe, efficient, and trustworthy XLM-enabled ADS.

1. Introduction

1.1. General context

In today's dynamic and constantly evolving world, the public sector faces numerous challenges, particularly in ensuring safety and conserving resources. In an effort to make roads safer, reduce human driving errors, increase mobility for those unable to drive, and enhance transportation efficiency, Autonomous Driving Systems(ADS), also known as self-driving cars, is transforming our ways of traveling (Kiss and Garai-Fodor, 2024). However, achieving fully autonomous driving faces several challenges, such as unreliable perception and decision-making in complex and unpredictable traffic scenarios. To navigate these challenges, governments and public organizations are leveraging novel technologies, such as Generative Artificial Intelligence (GAI) (Cao et al., 2023). The latter stands out as a potential solution across multiple disciplines, such as public safety (Ooi et al., 2023; Mahor et al., 2023; Anderljung et al., 2023). GAI techniques, such as Large Language Models (LLMs) utilize advanced algorithms to generate human-like text or content based on input prompts or patterns in the data they were trained on Chang et al. (2024) and Hadi et al. (2023). LLMs are initially trained on a large dataset of text data to learn semantics and general language patterns. This pre-training process involves tasks like predicting the next word in a sentence (language modeling) or filling in missing words based on the surrounding context. Embedding of customized data into LLMs allows for utilizing the strengths of pre-trained language models while adapting them to specific applications or domains (Ge et al., 2023; M. Chen et al., 2023; Wu et al., 2024).

Despite the LLMs' strong reasoning performance on several Natural Language Processing (NLP) tasks (Huang, 2024), they are "blind" to visuals, since they are limited to understanding discrete text. In contrast, Vision-Language Models (VLMs) can see well but are typically weaker in reasoning compared to LLMs (J. Wang et al., 2023; Maaz

* Corresponding author.
E-mail addresses: sonda.fourati.1@ens.etsmtl.ca (S. Fourati), wael.jaafar@etsmtl.ca (W. Jaafar), noura.baccar@medtech.tn (N. Baccar), smalfattani@kau.edu.sa (S. Alfattani), eami.langar@etsmtl.ca (R. Langar).

https://doi.org/10.1016/j.engappai.2026.114805
Received 22 August 2025; Received in revised form 29 December 2025; Accepted 8 April 2026
Available online 11 April 2026

et al., 2023; Y. Bai et al., 2024; P. Xu et al., 2023; Y. Zhao et al., 2024; Li et al., 2024a). Nevertheless, the complementarity of VLMs and LLMs paves the way toward a new field known as Multimodal LLMs (MLLMs) (D. Zhang et al., 2024; Yin et al., 2023). Indeed, MLLMs can combine image, video, and audio data with the advanced reasoning capabilities of LLMs; thus, they are well-equipped to execute a wide range of tasks, such as speech recognition, image classification, and text-to-video matching. We abbreviate the aforementioned language models as XLMs. Throughout this survey, the term XLM (Cross/Extended Language Model) is used as a unifying shorthand to denote any foundation model integrating language with additional modalities (vision, text, or sensor data).

Research on applying XLMs to autonomous driving has only recently begun and remains in its early stages. For instance, recent efforts such as J. Li et al. (2025) have surveyed the integration of LLMs and multimodal large models in autonomous driving systems. Authors in Wu et al. (2025) provide a multi-agent perspective, and Cui et al. in Cui et al. (2024d) propose a benchmark for LLM-augmented ADS. Besides that, authors of Cui et al. (2024c) highlighted the advantages of MLLMs for ADS and overviewed recently published related work. Also, authors in S. Zhang et al. (2024) and Zheng et al. (2023) introduced traffic safety decision-making through multi-modality representation learning. They suggested that MLLMs can potentially contribute to various aspects of traffic safety research. In addition, recent work, including (Jiang et al., 2025; Zhao et al., 2025; Cui et al., 2025; Wang et al., 2025; Azdam et al., 2025) investigated the use of intelligent agents powered by XLMs toward autonomous driving.

1.2. Motivations, objectives, and contributions

The rapid progress of XLMs has opened new frontiers for ADS. LLMs have shown strong abilities in understanding context, handling complex tasks, and generating answers. Moreover, their integration with multimodal data to build MLLM systems can enhance generalization and adaptation to novel scenarios with complex driving behaviors. With VLMs, the perception issue can be tackled, and the decision-making process becomes more transparent.

In contrast to prior works (Yang et al., 2023; Huang et al., 2023; Gao et al., 2024; X. Zhou et al., 2024; Cui et al., 2024c; Luo et al., 2024) that typically focus on a single family (e.g., VLMs) or provide high-level overviews without deployment guidance, our survey *unifies* **LLM**, **VFM**, **VLM**, and **MLLM** perspectives under the **XLM** umbrella and maps them *explicitly* to the ADS stack (perception → prediction → planning → HMI/V2X). We further synthesize practical components, datasets, simulators, and open-source toolchains, together with *deployment constraints* (latency, memory/KV-cache, edge scheduling) and *safety alignment* (hallucination mitigation, verification, and uncertainty-aware abstention). As summarized in Table 2, earlier surveys do not simultaneously cover (i) a cross-family XLM taxonomy aligned to ADS tasks, (ii) closed- *and* open-loop evaluation guidance, and (iii) actionable recommendations for reliability and security.

Why this matters now: XLMs are rapidly transitioning from proof-of-concept to *agentic*, tool-using components within ADS. A consolidated treatment that links model design to *how* they are evaluated and *where* they integrate in the stack is essential for reproducible progress and safe deployment.

Since 2023, several researchers have been investigating ADS with LLM, VLM, or MLLM. However, to the best of our knowledge, there is no comprehensive work that gathered state-of-the-art works focusing on using LLM, VLM, and MLLM in ADS together, nor did they evaluate the practical deployment of XLMs for ADS. Hence, the main objective of this survey is to draw the full picture of the state-of-the-art regarding the use of XLMs for ADS and to bridge the gap between their theory and practical use in ADS.

Contributions: Specifically, we provide:

- **A unified, task-aligned taxonomy** of LLM/VFM/VLM/MLLM (XLM) capabilities mapped to ADS modules perception, prediction, planning, Human–Machine Interface (HMI)/Vehicle to Everything (V2X), clarifying *where* each class adds value.
- **A comparative synthesis** of recent (2022–2025) methods with *cross-family* coverage, unlike prior surveys that center on a single modality or model class.
- Extensive study on recent datasets, tools, frameworks, and benchmarks that enable the practical implementation of XLMs in ADS, including guidance for *open- vs. closed-loop* evaluation and simulator-in-the-loop testing.
- Identification of **related challenges and future research directions** to promote ADS deployment assisted with XLMs, with a dedicated subsection on *hallucination mitigation, verification, and uncertainty-aware abstention* for safety.
- **Standardization recommendations** (checklists for data coverage, metrics, and robustness reporting) are intended to reduce fragmentation across XLM-for-ADS studies.

Together, these elements position our survey as a *single, practice-oriented reference* that consolidates XLM modeling, evaluation, and safety considerations for ADS, complementing and extending the scope of prior surveys (Cui et al., 2024c; Yang et al., 2023; Huang et al., 2023; Gao et al., 2024; X. Zhou et al., 2024; Luo et al., 2024).

1.3. Organization of the paper

The remainder of this survey is structured as follows. Section 2 summarizes related surveys that studied XLMs, also known as, Foundation Models (FM), used to achieve ADS. Section 3 introduces the background knowledge related to ADS and XLMs. Section 4 highlights the role of XLM in mitigating ADS challenges. Section 5 presents the proposed taxonomy. Sections 6, 7, and 8 explore the technological advancements and recent works on the use of LLMs, VLMs, and MLLMs, respectively, toward AD. Section 9 assesses datasets and benchmarks that could be utilized to enable the practical deployment and testing of XLMs in ADS. Section 10 presents practical frameworks for integrating XLMs within ADS, detailing system architecture, deployment trends, and technical constraints, and providing a comparative performance analysis of recent models on ADS-relevant tasks. Section 11 identifies open challenges and outlines directions for future research. Finally, Section 12 concludes the survey. Fig. 1 provides a visual guide to this organization: it *maps* the taxonomy (Section 5) onto the ADS stack (perception → prediction → planning → control/HMI/V2X), and cross-references the sections where each capability class (LLM/VFM/VLM/MLLM) is developed (Sections 6–8) and evaluated (Section 9), thereby helping readers navigate from concepts to datasets and deployment (Section 10).

1.4. List of acronyms

In Table 1, we present the list of acronyms used in the survey.

2. Related works and research methodology

2.1. Overview of existing surveys

Recent surveys and studies have explored the application of XLMs within ADS. These investigations assessed the potential of XLMs to enhance the ADS's predictive and decision-making capabilities. For instance, a categorization and an analysis of recent works applying foundation models to ADS was provided by Gao et al. in Gao et al. (2024). They established a taxonomy based on modality and functions in ADS, then discussed techniques to adapt FMs to ADS, including in-context learning, fine-tuning, Reinforcement Learning (RL), and visual instruction tuning. They also highlighted the limitations of FMs, such

Table 1
List of acronyms.

Acronym	Description	Acronym	Description
ACC	Adaptive Cruise Control	AD	Autonomous Driving
ADS	Autonomous Driving Systems	AIDE	Automatic Data Engine
AMF	Access and Mobility Management Function	AV	Autonomous Vehicle
Beit	Bidirectional Encoder Representation From Image Transformers	BLIS	Blind Spot Information System
CAN	Controller Area Network	CLIP	Contrastive Language-Image Pretraining
CNN	Convolutional Neural Network	CoT	Chain-Of-Thought
CrossViT	Cross-Attention Multi-Scale Vision Transformer	DRL	Deep Reinforcement Learning
E2E	End-To-End	EM-VLM4AD	Efficient, Lightweight Multi-Frame VLM for VQA in AD
FFNN	Feed-Forward Neural Network	FM	Foundation Model
GAI	Generative Artificial Intelligence	GPU	Graphical Processing Unit
GPS	Global Positioning System	HDT	Human Digital Twin
HiLM-D	High-Resolution Understanding In MLLMs for AD	HMI	Human–Machine Interaction
IDS	Instruction Detection System	IMU	Inertial Measurement Unit
LC-LLM	Lane Change Large Language Model	LiDAR	Light Detection and Ranging
LIN	Local Interconnect Protocol	LKA	Lane Keeping Assist
LLM	Large Language Model	LLM4AD	Large-scale Language Models for AD
MAE	Mean Absolute Error	MLLM	Multimodal Large Language Model
MLM	Masked Language Modeling	MLP	Lightweight Multilayer Perceptron
MOST	Media Oriented System Transport	MPC	Model Predictive Control
MuRAG	Multimodal Retrieval-Augmented Generation	NDD	Natural Denoising Diffusion
NLP	Natural Language Processing	ONCE	One Million Scenes
pFedLVM	Personalized Federated Learning Large Vision Model	PPO	Proximal Policy Optimization
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses	PVT	Pyramid Vision Transformer
RAG	Retrieval-Augmented Generation	ReAc	Reasoning and Acting
ResNet	Residual Network	RL	Reinforcement Learning
RLHF	Reinforcement Learning with Human Feedback	RTK	Real-Time-Kinetic
SADM	Spatial-Aware Decision Making	SAE	Society of Automotive Engineers
SAM	Segment Anything Model	SMF	Session Management Function
SUMO	Simulation of Urban Mobility	T2T-ViT	Tokens-To-Token Vision Transformer
TPU	Tensor Processing Unit	TRS	Traffic Rules Satisfaction
UDM	Unified Data Management	uniMVM	Unified Multi-View Model
UNITER	Universal Image-Text Representation	UPF	User Plane Function
V2D	Vehicle-To-Device	V2G	Vehicle-To-Grid
V2I	Vehicle-To-Infrastructure	V2N	Vehicle-To-Network
V2P	Vehicle-To-Pedestrian	V2V	Vehicle-To-Vehicle
VLM	Vision-Language Model	VLM4AD	Simulation of Urban Mobility
ViT	Vision Transformer	VQA	Visual Question Answering
XLM	Any Language Model	ZOD	Zenseact Open Dataset

as hallucination, latency, and efficiency, and finally proposed relevant research directions. Similarly, Yang et al. (2023) surveyed the use of LLMs for Autonomous Driving (LLM4AD). Specifically, the authors explored various applications of LLMs within ADS and detailed prominent methodologies in each category. They also presented the most recent datasets pertinent to LLM4AD.

Alternatively, authors of X. Zhou et al. (2024) provided a detailed survey on integrating VLMs within ADS. They classified existing studies based on VLM types and application domains while considering five major aspects: Perception and understanding, navigation and planning, decision-making and control, End-to-End (E2E) autonomous driving, and data generation. Moreover, the survey consolidated emerging vision-language tasks and metrics, analyzed classic and language-enhanced AD datasets, explored potential applications and technological advances, and discussed benefits, challenges, and research gaps for ADS. In Cui et al. (2024c), Cui et al. provided a comprehensive study of integrating MLLMs into ADS. They focused on MLLM's capability to process multimodal data for perception, motion planning, and motion control, and identified future research directions. In the same context, authors in Luo et al. (2024) provided a taxonomy of the use of FMs in autonomous vehicles. Moreover, they examined data augmentation and model optimization of LLMs and VLMs to enhance autonomous driving decisions. Finally, authors of Huang et al. (2023) discussed the architectures of several solutions that apply FMs in ADS.

Although existing surveys provide valuable insights into XLMs and their role in advancing ADS research, many overlook the practical integration of XLMs within real-world ADS frameworks. Furthermore, a rigorous comparative evaluation of the proposed XLM-based approaches for ADS remains largely absent. Most surveys tend to focus narrowly on a single class of XLMs, for example, addressing either LLMs or VLMs, but rarely both.

To address these gaps, we present here an in-depth study of XLM-ADS integration, offering a comparative analysis across diverse XLM-based methods for autonomous driving. Unlike prior surveys focusing on a single modality, our work unifies LLM, VLM, and MLLM paradigms for autonomous driving under a coherent and holistic taxonomy. Similar to recent hybrid approaches in transportation modeling, such as the Bi-LSTM fusion framework for passenger flow prediction (Balasubramani and Natarajan, 2024), this integrative view enables cross-modal reasoning and a more comprehensive understanding of intelligent mobility systems. Additionally, we place particular emphasis on the most recent state-of-the-art contributions published over the last two years (2023–2024). In this context, our survey aims to explore and provide answers to the following research questions:

- **RQ0:** What are the fundamental principles of ADS and XLM?
- **RQ1:** What are the main AD challenges and issues that can be tackled with XLM-based methods?
- **RQ2:** How can LLMs be integrated into ADS to improve decision-making, situational awareness, and advanced driver-assistance systems?
- **RQ3:** How can VFMs/VLMs be optimized for real-time object detection and identification, and obstacle recognition within AD environments?
- **RQ4:** How can MLLMs be employed to improve visual and linguistic information integration in ADS, and what are the potential applications of MLLMs to enhance human–vehicle interaction systems?
- **RQ5:** What are the best practices for datasets and tools to evaluate XLMs in the context of ADS?

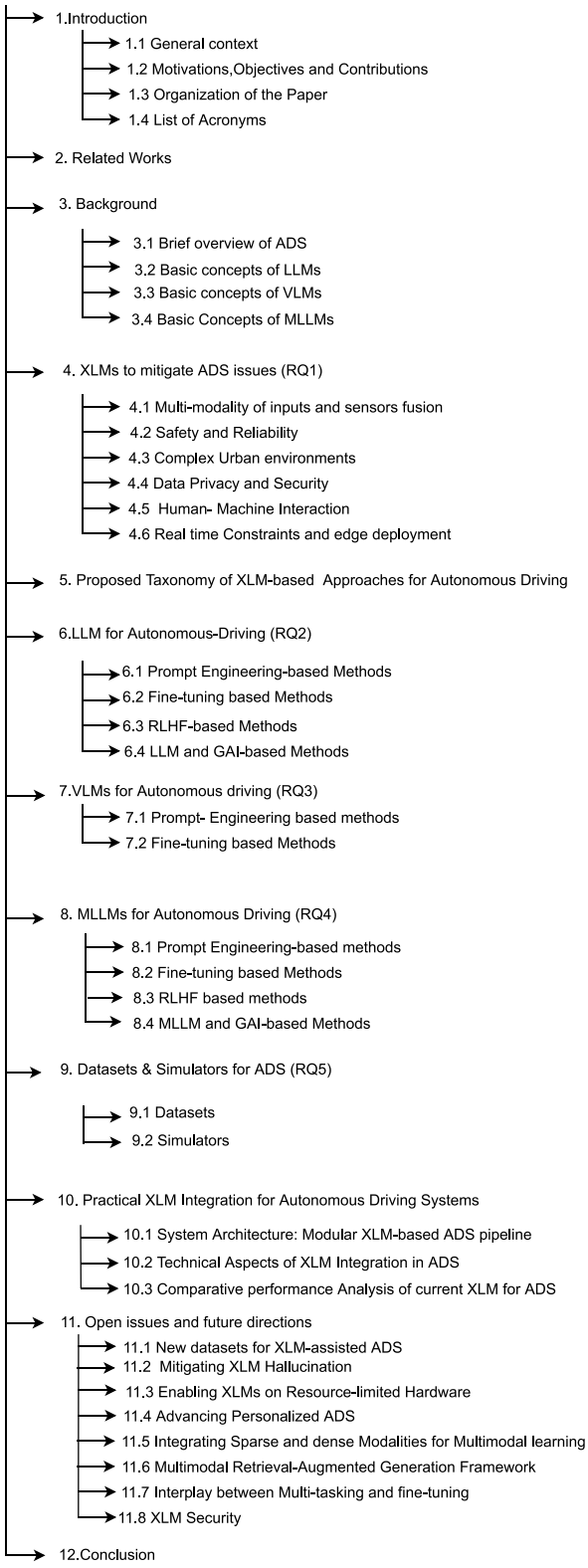

Fig. 1. Paper organization.

- **RQ6:** What are the open issues and future research directions for integrating XLMs into ADS?

In Table 2, we highlight the contributions of the previous surveys, compared to ours. It contrasts prior surveys across two axes: (i) model families (LLM, VLM, MLLM) and (ii) *scope dimensions* including taxonomy design, architectural coverage, datasets, simulators, and future directions. A ✓ denotes explicit coverage; ✗ indicates that the aspect is not covered (or only tangentially). The table shows that existing reviews typically emphasize a *single* family (e.g., VLMs) or provide high-level taxonomies without linking them to datasets/simulators or deployment guidance. In contrast, the row "Ours" evidences our *cross-family* treatment (LLM+VFM/VLM+MLLM) and end-to-end scope (taxonomy → architecture → data/sim → evaluation), which addresses the integration gaps in Yang et al. (2023), Huang et al. (2023), Gao et al. (2024), X. Zhou et al. (2024) and Luo et al. (2024).

2.2. Review methodology and search strategy

This section outlines the research methodology adopted to identify, select, and analyze relevant studies addressing the scope and research questions introduced in the Related Works subsection. It details the systematic process used to collect, filter, and synthesize publications pertinent to foundation models and their applications in autonomous driving systems.

For our research methodology, we employed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework (Kim et al., 2023) to ensure a rigorous and transparent literature review process. PRISMA provides a comprehensive and systematic approach to identify, screen, and include relevant studies to our research questions.

2.2.1. Identification of relevant studies

We designed a reproducible search to identify studies aligned with the scope of this survey (foundation models for ADS) and with the research questions introduced in the *Related Works* subsection. Searches were executed across *IEEE Xplore*, *ACM Digital Library*, *Scopus*, *arXiv*, and *Google Scholar*, covering the period *January 2019–October 2025* and restricted to English. We complemented database results with backward/forward snowballing from key surveys and benchmark papers.

To ensure precision and coverage, we used standardized Boolean templates with quoted phrases and synonyms. Representative query patterns include:

- `("autonomous driving" OR ADS) AND ("large language model" OR LLM)`
- `("autonomous driving" OR ADS) AND ("vision- language model" OR VLM)`
- `("autonomous driving" OR ADS) AND ("multimodal large language model" OR MLLM OR "multimodal LLM")`
- `("foundation model" OR "vision foundation model" OR VFM) AND ("perception" OR "planning" OR "prediction") AND ("autonomous driving" OR ADS)`
- `("datasets" OR "benchmarks" OR "simulators") AND (LLM OR VLM OR MLLM) AND ("autonomous driving" OR ADS)`
- `("frameworks" OR toolchain OR deployment) AND (LLM OR VLM OR MLLM) AND ("autonomous driving" OR ADS)`
- `(hallucination OR reliability OR "safety alignment" OR verification) AND (LLM OR VLM OR MLLM) AND ("autonomous driving" OR ADS)`

2.2.2. Inclusion criteria

(i) the study addresses LLM/VLM/MLLM/VFM methods or deployments relevant to ADS modules (perception, prediction, planning, HMI/V2X); (ii) peer-reviewed venues or well-cited preprints with technical details (architecture, data, metrics); (iii) release of datasets or simulators or frameworks or reproducible evaluation.

Table 2
Summary of relevant surveys. Abbrev.: Taxo. = Taxonomy, Arch. = Architecture, Data. = Datasets, Sim. = Simulators, Fut. = Future Directions. Symbols: ✓ covered, ✗ not covered.

Ref.	Objective	LLM	VLM	MLLM	Taxo.	Arch.	Data.	Sim.	Fut.
Yang et al. (2023)	Review of technical achievements of LLMs for AD	✓	✗	✗	✓	✗	✓	✓	✗
Huang et al. (2023)	Architectures of FMs for AD	✓	✓	✓	✗	✓	✗	✗	✗
Gao et al. (2024)	Analysis of works on applying FMs to ADS	✓	✓	✓	✗	✗	✗	✗	✓
X. Zhou et al. (2024)	Analysis of the potential of VLMs for ADS	✗	✓	✗	✓	✗	✓	✗	✓
Cui et al. (2024c)	Comprehensive study on integrating MLLMs with ADS	✗	✗	✓	✗	✗	✓	✓	✗
Luo et al. (2024)	Taxonomy of FMs in AD and optimization methods for enhanced decisions	✗	✗	✓	✓	✗	✗	✗	✓
Ours	XLMs review for ADS (LLM+VFM/VLM+MLLM) with task-aligned taxonomy, deployment, and evaluation	✓	✓	✓	✓	✓	✓	✓	✓

2.2.3. Exclusion criteria

(i) non-technical commentaries without methods or results; (ii) narrowly domain-specific works with no ADS linkage; (iii) duplicate versions. When multiple versions existed, the most complete and recent version was retained.

2.2.4. PRISMA process

We followed a PRISMA-style selection. First, we conducted broad multi-database searches and removed duplicates. Next, two authors screened titles/abstracts against the inclusion/exclusion criteria. Eligible records underwent full-text assessment, focusing on methodological transparency (model/design details), ADS relevance, and evaluative evidence (datasets, metrics, or deployment reports). The PRISMA flow (Fig. 2) documents counts at each stage.

2.2.5. Mapping selected papers to research questions

The correspondence between representative studies and the addressed research questions is summarized in Table 3, providing a structured mapping that illustrates how selected works contribute to each identified research theme.

3. Background (RQ0)

3.1. Brief overview of autonomous driving systems

ADS is considered a significant advancement in automotive technology that revolutionizes the transportation experience (Parekh et al., 2022). The main factors motivating ADS's development are safety and mobility enhancements, traffic congestion reduction, and fuel efficiency improvement. AD systems are categorized by the Society Of Automotive Engineers (SAE) into six levels of automation, ranging from driver assistance to full self-driving.

3.1.1. Architecture of autonomous driving systems

A typical architecture of ADS should include a combination of sensors, cameras, communication modules, and sophisticated algorithms to navigate and control the vehicle without human intervention. As illustrated in Fig. 3, it involves five layers, namely the perception layer, the processing and decision layer, the control and actuation layer, the cybersecurity layer, and the communication layer, described as follows:

- **Perception:** Multi-sensor sensing and ego-localization using cameras, LiDAR/radar, GNSS/RTK, IMU, and ultrasonics, producing scene and self-state estimates for downstream planning.

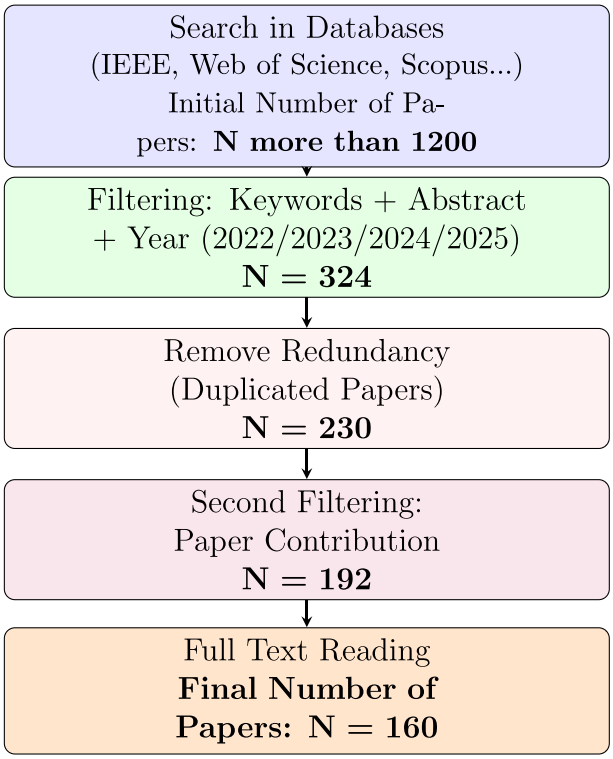

Fig. 2. Followed PRISMA steps.

- **Processing & Decision:** Core autonomy stack combining perception (detection/tracking/lanes), prediction/behavior modeling, localization/mapping, and motion planning to output driving intents and trajectories.
- **Control & Actuation:** Low-level execution of steering/throttle/braking, including common ADAS controllers (e.g., ACC, LKA) and trajectory tracking under vehicle dynamics constraints.
- **Communication:** Intra-vehicle networking (e.g., CAN/LIN/FlexRay/MOST) and inter-vehicle V2X (V2V/V2I/V2P/V2N/V2D/V2G) enabling cooperation, infrastructure awareness, and fleet connectivity.
- **Cybersecurity:** Cross-layer protection via encryption, authentication, access control, and IDS/monitoring to ensure operational safety and integrity of sensing, networking, and control.

Table 3
Mapping of examples of selected papers to the research questions.

RQ	Related papers	Motivation
RQ0	Janai et al. (2020), Parekh et al. (2022), Ahangar et al. (2021), M. Liu et al. (2024), Song et al. (2023), Yazgan et al. (2024), Chang et al. (2024), W.X. Zhao et al. (2023), Xu et al. (2024b) and Y. Wang et al. (2024)	Understand the basic concepts, architectures, and techniques related to ADS and XLMs.
RQ1	J. Wu et al. (2023), Su et al. (2023), Nouri et al. (2024), Y. Wang et al. (2023), Cui et al. (2023a), Luo et al. (2024), Y. Yang et al. (2024) and Aldeen et al. (2024)	Learn ADS challenges and requirements that could be mitigated using XLMs.
RQ2	Mao et al. (2023b), Chen et al. (2024a), Huang et al. (2023), Gao et al. (2024), Sha et al. (2023), G. Zhao et al. (2024), Peng et al. (2024), R. Yang et al. (2024), Wen et al. (2023), Ananthajothi et al. (2023), Murtaza et al. (2024), Z. Zhou et al. (2024), Tanahashi et al. (2023), Azarafza et al. (2024), H. Tian et al. (2024), T. Wang et al. (2024), Miceli-Barone et al. (2023), Jin et al. (2023) and X. Wang et al. (2023)	Analyze in depth how LLMs are integrated into ADS to improve decision-making, situational awareness, and ADAS.
RQ3	Yan et al. (2024), Shukor et al. (2023), D. Wu et al. (2023), Xie et al. (2024), Cho et al. (2024), Kou et al. (2024), X. Zhou et al. (2024), Gopalkrishnan et al. (2024), Li et al. (2024b), W. Han et al. (2024), Guo et al. (2024), X. Tian et al. (2024), Y. Huang et al. (2024), Liang et al. (2024) and Nie et al. (2023)	Study in depth how VLMs are integrated into ADS to enhance real-time object detection and identification and obstacle recognition.
RQ4	Cui et al. (2024c), S. Wang et al. (2024), Ding et al. (2024), Ray and Ohn-Bar (2024), Jia et al. (2023), Sreeram et al. (2024), Luo et al. (2024), Guan et al. (2024), Yuan et al. (2024), W. Wang et al. (2023), Chen et al. (2024b), Liao et al. (2024), Wei et al. (2024a), Park et al. (2024), Ding et al. (2023), Xu et al. (2024) and Guo et al. (2024)	Analyze in depth how MLLMs are integrated into ADS.
RQ5	Ding et al. (2024), Fu et al. (2024), Kong et al. (2024), Cui et al. (2024b,a), Wei et al. (2024b), Ge et al. (2024), Cui et al. (2023b), Ma et al. (2024a), Song et al. (2023), Khan (2024), Nie et al. (2023), Mao et al. (2023a), Malla et al. (2023), Inoue et al. (2024) and Mao et al. (2023b)	Assess key datasets, tools, frameworks, and benchmarks for deploying XLMs within ADS.
RQ6	Pi et al. (2024), Duan et al. (2024), Mao et al. (2023b), Z. Bai et al. (2024), G. Bai et al. (2024) and Xu et al. (2024b)	Identify open issues and future research directions to stimulate the integration of XLMs within ADS.

3.1.2. Autonomous driving datasets

Over the past decade, numerous datasets have enabled progress across ADS perception, prediction, and planning. For practical selection, we prioritize (i) *modalities* (Cam/LiDAR/Radar/Map), (ii) *scale & diversity* (cities, weather, day/night), and (iii) *task fit* (2D/3D detection, tracking, segmentation, forecasting, end-to-end control) (M. Liu et al., 2024). To improve readability, we group representative datasets by their *primary role* in ADS research, while detailed dataset-level characteristics and extended comparisons are provided in Appendix.

- **Classical perception baselines:** CamVid (Brostow et al., 2008) and KITTI (Geiger et al., 2012) remain widely used for segmentation and multi-task benchmarking.
- **Large-scale multimodal 3D driving:** nuScenes (Caesar et al., 2020), Waymo Open (Mei et al., 2022), and ZOD (Alibeigi et al., 2023) provide rich multi-sensor annotations supporting 3D detection/tracking and temporal reasoning.
- **Forecasting and map-centric benchmarks:** Argoverse (Chang et al., 2019) emphasizes HD maps and motion forecasting, enabling prediction- and planning-oriented evaluation.
- **Robustness and diversity at scale:** BDD100K (Yu et al., 2020) offers broad geographic and environmental diversity (weather/time-of-day), supporting robustness-oriented benchmarking.
- **Long-tail and corner cases:** CODA (K. Li et al., 2022) and Boreas (Burnett et al., 2023) target rare events and distribution shift (seasonal/weather repeats), useful for safety–critical validation.
- **Massive-scale pretraining sources:** ONCE (Mao et al., 2021) supports large-scale representation learning and long-tail coverage for 3D perception.

3.2. Basic concepts of large language models

LLMs are advanced language models with massive parameter sizes and exceptional learning capabilities (Chang et al., 2024; W.X. Zhao et al., 2023). They can understand contexts and nuances to support tasks across domains such as NLP, which covers text generation, translation, personalized chatbots, text classification, sentiment analysis, and question answering. In particular, question-answering is a crucial technology for HMI, which has proven its usefulness in search engines, intelligent customer service, and question-answering (Q&A) systems.

The LLM process starts with pre-training, in which it is exposed to a large dataset of text from a variety of sources, including books, papers, and websites. Unsupervised learning allows the LLM model to anticipate the next word in a phrase based on the context of previous words while comprehensively respecting grammar, syntax, and semantic rules. Indeed, LLMs tokenize text, map tokens to embeddings with

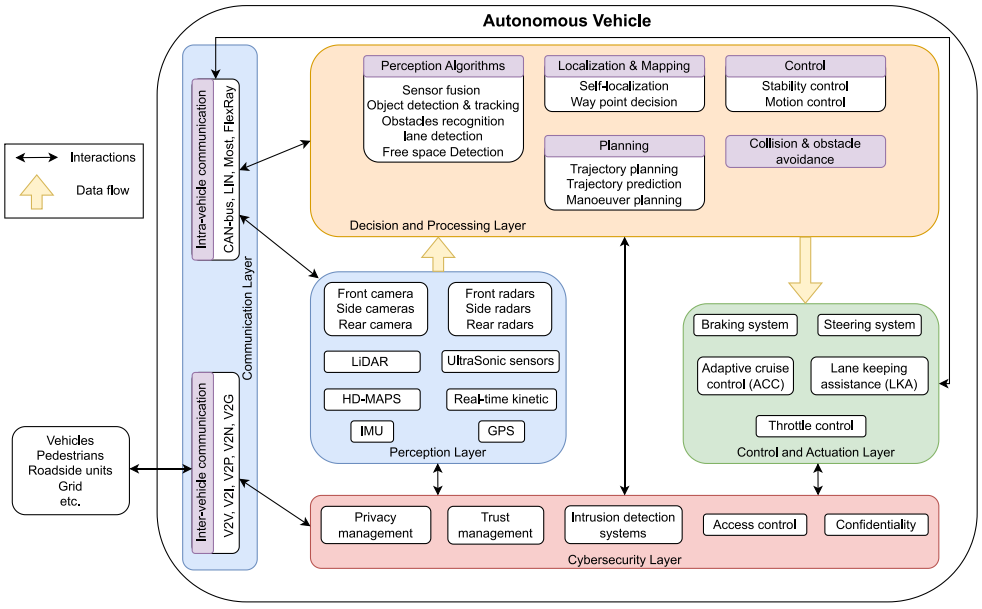

Fig. 3. A typical ADS architecture.

positional cues, and pass them through stacked Transformer blocks (attention + feed-forward) with normalization and dropout; a projection head converts hidden states to next-token probabilities, trained via cross-entropy. Instruction/fine-tuning or adapters customize behavior without changing this core pipeline.

3.2.1. Common large language models architectures

The LLM architecture is built around the Transformer framework (Vaswani et al., 2017). The two primary parts of a Transformer are the encoder and/or decoder. To find complex correlations between tokens, it breaks down at first the input data into tokens. Then, the latter are subjected to simultaneous mathematical operations. Through this approach, the system is enabled to identify and extract patterns in a way that is comparable to human cognition when faced with a similar question.

Besides, LLMs are characterized by the tasks they are designed for, including text generation, text classification, and text summarizing, and their training strategies, e.g., self-supervised, unsupervised, distillation, fine-tuning, etc. LLMs deploy large-scale text corpora during the training process that necessitates significant computational resources of advanced hardware like Tensor Processing Units (TPUs) and Graphical Processing Units (GPUs). Therefore, the required hardware is another criterion that needs to be considered when selecting LLMs for specific tasks.

3.2.2. Interacting with large language models

It can be achieved using techniques like prompt engineering, fine-tuning, zero-shot, and Reinforcement Learning From Human Feedback (RLHF). These methods enhance the model's performance and adapt it to specific tasks or domains. They are described as follows:

- **Prompt Engineering:** Prompt engineering for LLMs involves crafting inputs (prompts) to guide the model in generating the desired responses (Marvin et al., 2023). Prompting strategies span *chain-of-thought (CoT)* and *self-consistency* for reasoning quality, *ReAct* for reasoning-with-tools, and *contextual/dynamic prompts* plus *knowledge-generation/transfer-style prompts* to adapt pretrained models to new tasks and scenarios.
- **Fine-tuning:** Fine-tuning LLMs involves training them on a specific dataset to customize them to respond to requests in a specific context (Z. Han et al., 2024). It stimulates the model to generate consistent outputs and reduces hallucinations. Fine-tuning can be realized in an unsupervised, supervised, or instruction-based manner.
- **Zero-shot, One-shot, and Few-shot Learning:** Recent studies suggested that LLMs exhibit high levels of generalization, enabling them to apply their acquired knowledge to new tasks not included in their original training process. This capability is known as zero-shot learning (Sun et al., 2023). When the model is provided with a single example to illustrate the task, it corresponds to one-shot learning, while few-shot learning is when the model is provided with a few examples to better understand the task requirements and format.
- **Reinforcement Learning from Human Feedback:** RLHF is an advanced technique of fine-tuning, where feedback collected from users regarding the model's responses is used to enable the LLM model to learn from it and improve its future responses (Chan et al., 2024).
- **Multi-modal Integration:** LLMs could be enhanced by integrating them with other data modalities such as images, audio, or structured data. Indeed, combining image data with textual prompts leads to a comprehensive model capable of understanding and generating responses based on both text and visual inputs. Also, integrating structured data, e.g., tables and databases, with textual inputs, leads to more informed and accurate responses.

In Table 4, we summarize the main characteristics of common LLM architectures.

3.3. Basic concepts of vision-language models

Vision-Language Models (VLMs) are designed to jointly process and reason over visual and textual information. To establish a clear understanding of their foundation and evolution, we first present an overview of Vision Foundation Models (VFMs), followed by a detailed examination of VLMs and their integration mechanisms.

3.3.1. Vision-foundation models

VFMs are advanced neural network architectures developed to process and understand visual data such as images and videos. They have demonstrated high performance in various computer vision tasks,

Table 4
Representative LLM architectures. Abbrev.: Arch. = Architecture, Param. = Parameters, Pretrain. = Pre-training, Durat. = Training duration, Enc. = Encoder, Dec. = Decoder.

Year	Name	Arch.	Param.	Pretrain.	Datasets	Hardware	Durat.
2019	DistilBERT	Enc.	66M	Self-supervised; distillation	BookCorpus; Wikipedia	NVIDIA V100	90 h
2019	RoBERTa	Enc.	125–355M	Self-supervised	BookCorpus; OpenWebText; CC-News; Stories	TPU v4 cluster	≈2 wks
2019	Sentence-BERT	Enc.	110M	Fine-tuning	SNLI; Multi-Genre NLI	NVIDIA V100	–
2019	T5	Enc.-Dec.	60M–11B	Self-supervised	C4	1024 TPU v3	–
2020	GPT-3	Dec.	175B	Unsupervised	CommonCrawl; WebText2; Books; Wikipedia	NVIDIA A100	–
2021	GLM	Enc.	110M–130B	Self-supervised	BookCorpus; Wikipedia	1024 TPU v4	60 d
2021	HuBERT	Enc.-Dec.	281M–2.8B	Self-supervised	Libri-Light; LibriSpeech	–	–
2022	InstructGPT	Dec.	175B	RLHF	CommonCrawl; WebText2; Books; Wikipedia	992 A100 80G	–
2022	PaLM	Dec.	54B	Unsupervised	Web, code, books, news	6144 TPU v4	120 d
2023	Whisper	Enc.-Dec.	39M–1.15B	Self-supervised	(Speech corpora)	–	–
2023	LLaMA	Dec.	7–70B	Self-supervised	CommonCrawl; C4; GitHub; Wikipedia; arXiv	2000 A100 80G	21–25 d

including object detection, segmentation, and image classification. Although they are not VLMs, as they do not include any language modality, they often serve as the visual backbone in VLMs and MLLMs (J. Li et al., 2023; Yu et al., 2022).

The fundamental building blocks of VFMs are Convolutional Neural Networks (CNNs).

A typically adopted CNN structure for VLMs is the Residual Network (ResNet). The latter enables the training of deeper neural networks than CNNs. Moreover, Transformer models have been used for vision-based tasks. Specifically, Vision Transformers (ViTs) divide an image into fixed-size patches, linearly embed them into vectors, and then process them by Transformer encoders and self-attention mechanisms to capture relationships between patches, as shown in Fig. 4 (Dosovitskiy et al., 2020). Positional encodings are added to the patch embeddings to retain spatial information, allowing the model to differentiate between spatial positions.

Following ViT, several vision Transformer variants have been proposed to enhance training and inference efficiency in image recognition and generation tasks. Among the ViT architecture variants, we highlight the following. First, DeiT incorporated distillation strategies in ViT to outperform standard CNNs without requiring pre-training on large-scale datasets. Also, Tokens-To-Token Vision Transformer (T2T-ViT) (N. Li et al., 2023), Swin Transformer (Liu et al., 2022), ViT Masked AutoEncoder (VitMAE) (He et al., 2022), VitDet that combines ViT and Deit (Y. Li et al., 2022), Pyramid Vision Transformer-v2 (PVT) (W. Wang et al., 2022), and Cross-Attention Multi-Scale Vision Transformer (CrossViT) (Chandrasiri and Talagala, 2023) focused on improving the network architecture of ViT, mainly to make it lighter and fast-processing. In contrast, AdaViT (Meng et al., 2022) reduced the computation cost of ViT through E2E adaptive computation, while DETR (Liu et al., 2025) achieved it with a hybrid CNN-Transformer architecture. In Xie et al. (2021), SegFormer has been proposed, which is a semantic segmentation model that unifies Transformers with Lightweight Multilayer Perception (MLP) decoders. In addition, DINOv2 exploited training on a large ViT model, before distilling it into smaller models that can achieve higher performances (Oquab et al., 2023). Also, the bidirectional encoder representation from image Transformers, also known as, Beit, introduced a self-supervised vision representation model (Rajesh et al., 2024). Specifically, it executes masked image modeling to pre-train vision Transformers and fine-tune the model parameters. Moreover, Transformer variants like Mobile-ViT (MehtaS, 2022), TinyViT (Wu et al., 2022), and ED-ViT (X. Liu et al., 2024) have been developed specifically to enable real-time performance on edge devices. Finally, the VFMs can learn exclusively from visual data (Y. Bai et al., 2024). In Table 5, we summarize the characteristics of VFM architectures.

3.3.2. Vision-language models

VLMs are ML architectures designed to jointly process and integrate visual data (e.g., images or videos) and natural language text. These models are trained on paired image–text datasets to align visual and textual representations in a shared embedding space or to enable cross-modal tasks, such as image captioning, visual question answering, and image–text retrieval. Typically, VLMs consist of:

- A **visual encoder**, such as a CNN or a vision Transformer (e.g., ViT), that extracts semantic features from images.
- A **text encoder**, often a Transformer-based architecture, that processes textual inputs.
- A **fusion mechanism** or **alignment objective** that learns joint representations across the two modalities.

The key distinction between VFMs and VLMs is that VFMs operate on visual data exclusively and are optimized for vision-centric tasks, while VLMs explicitly integrate visual and textual modalities, enabling cross-modal understanding and alignment.

The Universal Image-Text Representation (UNITER) is among the first VLMs to combine images and text embedding (Y.-C. Chen et al., 2023). It has been trained for masked language modeling, image–text matching, and masked region modeling. Similarly, the Simple Visual Language Model (SimVLM) has been proposed for the joint representation of images and text, while reducing the training complexity using large-scale weak supervision (Wang et al., 2021). OpenAI has developed two VLMs, called Contrastive Language-Image Pretraining (CLIP) and DALL-E (Vayadande et al., 2024). CLIP integrates visuals

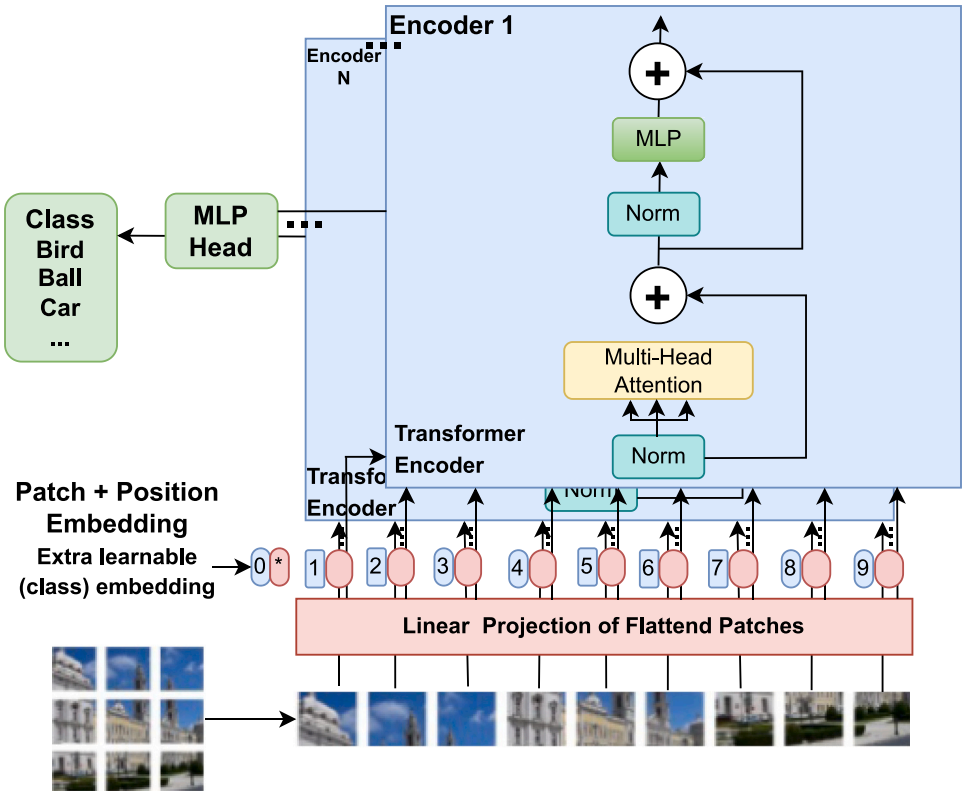

Fig. 4. Structure of ViT.

and their textual information in training and contrastive learning to create a shared embedding space for texts and images. Hence, CLIP can perform zero-shot learning to recognize and classify new images. As shown in Fig. 5, CLIP architecture involves three main modules: (1) Contrastive pre-training to learn matching images with their corresponding text description using text and image encoders, (2) classifier creation for a new task without additional training on task-specific data. It requires two steps. First, label text encoding where desired labels for the new task are converted into text descriptions. Then, text encoder utilization where text descriptions are fed through the pre-trained text encoder to generate their respective embeddings. Finally, (3) zero-shot prediction, in which a new image is encoded to generate its embedding, then the latter is compared to the embeddings of the label texts generated in the previous stage, hence predicting the right label for the embedded image. On the other hand, DALL-E generates images from textual descriptions, emphasizing creative visual content synthesis (Ramesh et al., 2022). It is built with three modules, namely (1) a Transformer model that captures the long-range dependencies in sequences, (2) a text encoder to convert text into a series of embeddings that capture the semantic meaning of the text, and (3) an image encoder that generates images from embeddings. DALL-E2 and DALL-E3 are new versions of DALL-E (Ramesh et al., 2022). As an example, we illustrate in Fig. 6 the DALL-E2 architecture. The latter uses an encoder–decoder pipeline that encodes the text description into a CLIP embedding representing both the text and image content. Then, the model decodes the embedding back to an image using a diffusion model. Also, authors of Jia et al. (2021) proposed a VLM named "A Large-scale ImaGe and Noisy-text embedding (ALIGN)" to learn joint visual and textual representations from noisy web data. It uses a dual-encoder architecture, where an efficient Net-based visual encoder and a BERT-based text encoder are trained from scratch using a contrastive objective. ALIGN leverages 1.8 billion noisy image–text pairs, collected with minimal filtering, to scale up vision-language representation learning. Despite its simplicity, ALIGN achieves state-of-the-art performance on cross-modal retrieval benchmarks and demonstrates strong zero-shot classification capabilities on ImageNet and its variants. Unlike MLLMs, ALIGN does not include a generative language model or support open-ended reasoning. Instead, it focuses on efficient cross-modal alignment, similar to CLIP. It scales CLIP-like training to massive image–text datasets. Authors of J. Li et al. (2022) designed a unified vision-language pre-training (VLP) framework, called Bootstrapping Language-Image Pre-training (BLIP), which addresses the challenge of learning from noisy web image–text data. Unlike earlier VLMs that excelled at either understanding-based tasks (e.g., retrieval, classification) or generation-based tasks (e.g., captioning), BLIP is designed for both, using a novel *Multimodal Mixture of Encoder–Decoder (MED)* architecture. It supports three operational modes: Unimodal encoding, image-grounded encoding, and image-grounded decoding. To improve the quality of training data, BLIP introduces a bootstrapping mechanism, *CapFilt*, which combines a caption generator and a filter to produce diverse synthetic captions and remove low-quality ones. This results in a cleaner, more informative dataset for pre-training. The model jointly optimizes contrastive learning, image–text matching, and language modeling objectives. Empirically, BLIP achieves state-of-the-art results on a wide range of benchmarks, including image–text retrieval, captioning, VQA, and zero-shot transfer to video-language tasks. Given its encoder–decoder design and training approach, BLIP is best categorized as an advanced VLM.

Finally, Li et al. (2021) proposed the VLM ALign the image and text representations BEfore Fusing (ALBEF). It presents a novel framework for vision-language representation learning, introducing a two-stage training strategy. First, it aligns the unimodal image and text representations using a contrastive learning objective. Then, the aligned features are fused using a multimodal encoder for downstream vision-language tasks. To enhance training stability and effectiveness, ALBEF uses a *momentum distillation* mechanism, where a momentum encoder provides soft targets for contrastive learning, resulting in improved representation alignment. Compared to prior approaches, ALBEF demonstrates both higher accuracy and faster inference speed across multiple

Table 5
Summary of VFM architectures.

Year	Model name	Model Arch.	Oriented tasks	Model parameters	Pre-training method and dataset	Testing datasets
2020	DETR	Encoder-Decoder	Object detection; Instance segmentation; Panoptic segmentation	40 Million	Supervised; COCO-2017	COCO-2017
2020	ViT	Encoder	Image classification	86–632 Million	Supervised and self-supervised; ImageNet-21K	ImageNet-1K
2021	DeiT	Encoder	Image classification	5–88 Million	Distilled; ImageNet-1K	ImageNet-1K, CIFAR-10/100, Flowers, Cars, iNat-18/19
2021	SegFormer	Encoder-Decoder	Segmentation	3.7–82 Million	Supervised; ImageNet-1K	Cityscapes, ADE20K, COCO-Stuff
2021	Swin Transformer	Encoder	Image classification; Object detection; Semantic segmentation; Video classification	26 Million	Self-supervised; ImageNet-22K	ImageNet-1K, ADE20K, COCO, Object365 v2
2021	BEiT	Encoder	Image classification; Semantic segmentation	86–632 Million	Self-supervised; ImageNet-1K	ImageNet-1K, ADE20K
2021	T2T-ViT	Transformer-based	Image classification	6.9–65 Million	From scratch; ImageNet-1K; T2T mechanism	ImageNet-1K
2021	PVT	Encoder	Classification; Object detection; Semantic segmentation	PVT-S: 25M, PVT-M: 44M	Supervised	ImageNet-1K, COCO, ADE20K
2021	CrossViT	Cross-attention Vision Transformer	Image classification	Variable (multi-scale)	Supervised; ImageNet-1K	ImageNet-1K
2022	AdaViT	ViT	Image classification	Variable (adaptive)	Supervised; ImageNet-1K	ImageNet-1K
2022	ViTMAE	Encoder	Image classification	86–632 Million	Self-supervised; ImageNet-1K	COCO, ADE20K, iNat Places
2022	ViTDet	Encoder	Object detection	86–632 Million	MAE minimization; ImageNet-1K	COCO
2023	DINOv2	Encoder	Image classification	1.1 Billion	Self-supervised; LVD-142M	ImageNet-1K, ImA, Oxford-II
2023	LVM	Encoder-Decoder	Semantic segmentation; Depth; Normals; Edge detection	300M–3B	Self-supervised; UVD-v1	Kinetics-700, ImageNet
2022	MobileViT	CNN + Transformer blocks	Image classification on edge devices	5.6 Million (for MobileViT-S)	Supervised; ImageNet-1K	ImageNet, CIFAR-10/100
2022	TinyViT	Lightweight ViT + Efficient token mixing	Image classification; Semantic segmentation; Object detection	5–21 Million	Supervised; ImageNet-1K + distillation	ImageNet-1K, ADE20K, COCO
2024	ED-ViT	Early-exit Distilled ViT	Efficient inference for vision tasks	22.8 Million	Supervised + Knowledge distillation; ImageNet	ImageNet-1K

vision-language tasks, including image–text retrieval, visual question answering, and captioning. Table 6 summarizes the studied VLMs above.

3.4. Basic concepts of multimodal large language models

MLLMs extend the capabilities of traditional LLMs by enabling them to process and reason over multiple modalities such as text, images, audio, and video. Unlike early VLMs, which are typically trained for alignment or retrieval tasks, MLLMs aim to support a wide range of generation-based and reasoning-based tasks in a unified framework.

MLLMs are characterized by the following core architectural components:

- **Modality Encoder:** A pre-trained model (e.g., ViT or CNN) that transforms raw visual inputs (images, frames) into embeddings.
- **Language Backbone:** A powerful pre-trained LLM (e.g., GPT, PaLM) that interprets and generates textual data.
- **Multimodal Interface:** A connector that aligns visual embeddings with the language model's input space. This can be implemented via projection layers, cross-attention modules, or learned adapters.

Recent advances in MLLMs have predominantly adopted Transformer-based architectures to facilitate cross-modal interaction. These architectures are often trained using combinations of contrastive objectives (e.g., ITC), language modeling (LM), and matching tasks (e.g., ITM). Compared to VLMs, which mostly focus on representation learning and discriminative understanding, MLLMs are designed for generative tasks such as visual question answering, image-grounded dialogue, visual storytelling, and code generation from visual input.

Examples of recent MLLMs include *MiniGPT-4*, and *PaLM-E* that demonstrated strong performance on both image-to-text and text-to-image generation tasks. These models integrate large-scale visual pre-training with autoregressive text generation, enabling rich multimodal

Table 6
Summary of VLM architectures.

Year	Model name	Arch.	Knowledge representation	Learn. Obj.	Param.	Pre-training datasets
2020	UNITER	Encoder	Joint visual-language embedding	Masked Language Modeling (MLM), Image–Text Matching (ITM), Image–Text Contrastive learning (ITC)	110M	COCO, VG, Conceptual Captions, SBU
2021	ALIGN	Dual Encoder	Contrastive image–text alignment	ITC	–	1.8B noisy image–text pairs
2021	CLIP	Dual Encoder	Joint embedding via contrastive learning	ITC	400M to 1.6B	400M+ image–text pairs (web data)
2021	DALL· E 2	Text-to-Image Decoder	Diffusion decoder conditioned on CLIP embeddings	Text-to-image generation	3.5B (approx.)	YFCC100M, internal curated datasets
2021	ALBEF	Encoder	Align before fuse architecture; Momentum distillation	ITC, ITM, LM	210M	COCO, VG, Conceptual Captions, SBU, CC12M
2022	BLIP	Encoder–decoder	Bootstrapped caption filtering (CapFilt)	ITC, ITM, LM	ViT-B:224M, ViT-L:361M	COCO, VG, CC, SBU, LAION
2022	BEiT-3	Unified Transformer	Multimodal encoder using a shared Transformer backbone for vision and text	MLM, ITM	2B	ImageNet-1K, JFT-3B, COCO, Visual Genome, VQA v2.0

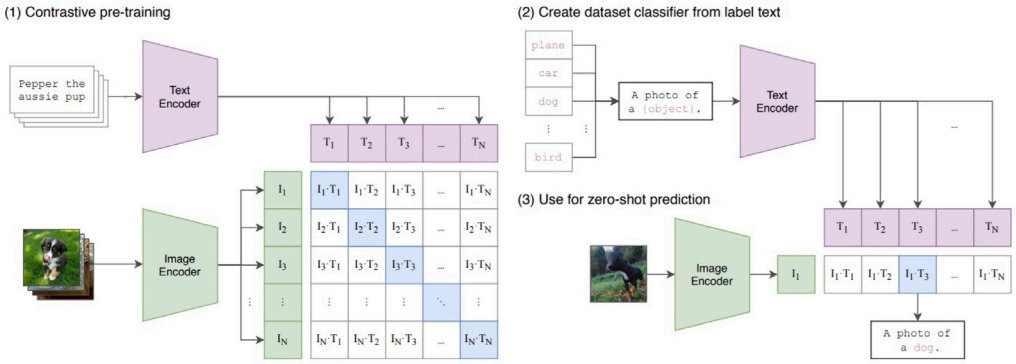

Fig. 5. Architecture of CLIP framework.

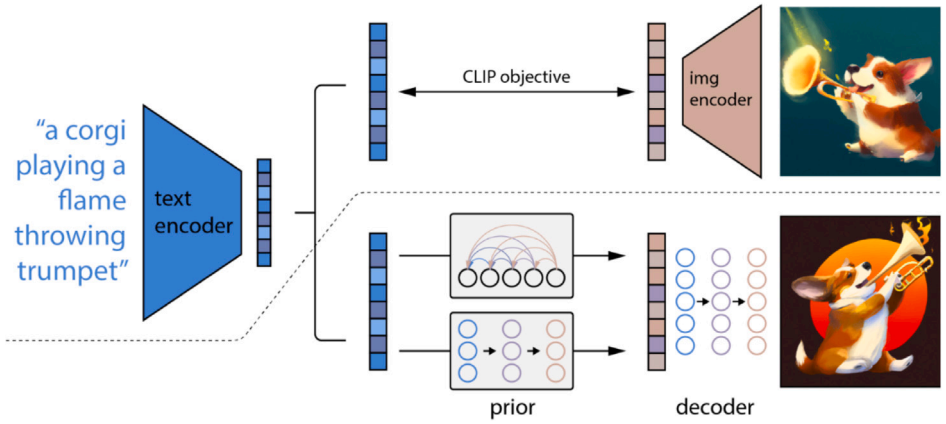

Fig. 6. Architecture of DALL-E2 framework.

reasoning capabilities beyond the scope of VLMs. For instance, PaLM-E (W. Wang et al., 2023) is a large embodied multimodal model that addresses a variety of embodied reasoning tasks, such as robotics tasks (Driess et al., 2023b). The latter is among the large MLLMs with 562 billion parameters. Finally, authors of S. Huang et al. (2024) proposed KOSMOS-1, an MLLM that can perceive general modalities and follow instructions, e.g., zero-shot.

Autonomous driving requires a broader perception stack that includes, in addition to visual/textual data, LiDAR, radar, and inertial measurements. These additional modalities provide complementary spatial and depth information that are critical for robust driving in adverse conditions, such as fog, night, rain, etc., where camera performance may degrade. To address this issue, recent multimodal models started integrating non-visual data into ADS Transformer-based architectures to enhance scene understanding and decision-making. For instance, TransFuser (Chitta et al., 2022), is one of the earliest Transformer-based models to fuse LiDAR and Red-Green-Blue (RGB) images for AD. The model uses separate CNN backbones to extract features from both modalities and merge them using a cross-attention

Table 7
Summary of MLLM architectures.

Year	Model name	Arch.	Knowledge representation	Learn. Obj.	Param.	Pre-training datasets
2021	SimVLM	Encoder-Decoder	Unified vision-text prefix modeling	PrefixLM, generation	{86, 307, 1200}M	ALIGN
2023	PaLM-E	Encoder-Decoder	PaLM + ViT; real-world robot tasks	Language + visual scene understanding	562B	LAION, WebVid-10M, HD-VILA, ACAI
2023	Flamingo	Perceiver + Frozen LLM	Visual prefix + causal language modeling	Few-shot multimodal generation	80B	Internally curated web image–text pairs
2023	MiniGPT-4	Encoder–Decoder	ViT + Q-Former + LLaMA-13B	Image-to-text generation, instruction following	13B	CC3M, SBU, LAION, and instruction tuning data
2024	KOSMOS-1	Encoder-Decoder	Unified multimodal inputs; prompts + dialogue	In-context learning, generation	1.6B	Web image–text pairs (scale not public)
2021	TransFuser	CNN + Transformer Fusion	LiDAR + RGB fusion via cross-attention	Waypoint prediction for autonomous driving	–	Carla simulator, real-world driving data
2021	Voxel Transformer	Transformer-based 3D encoder	Voxelized LiDAR for 3D perception	3D object detection in BEV	–	KITTI, nuScenes
2023	VoxelFormer	Dual-attention Transformer	BEV from voxelized LiDAR; semantic + spatial attention	Accurate BEV-level 3D detection	–	nuScenes
2023	BEVFusion	Multi-sensor fusion with Transformer	Unified BEV across LiDAR, camera, radar	Multi-task (3D detection, semantics)	–	nuScenes, Argoverse

Transformer block. The fused representation is used to generate way-points for autonomous vehicle control. Similarly, VoxelFormer, presented in Z. Li et al. (2023), is a Transformer framework designed for 3D object detection in Bird's Eye View (BEV). It takes voxelized LiDAR data as input and applies a dual spatial and semantic attention mechanism to aggregate features across the 3D space. Compared to CNN-based pipelines, VoxelFormer improves object detection accuracy and spatial consistency, which is essential for AD tasks like motion planning and tracking. Finally, authors of Z. Liu et al. (2023) introduced BEVFusion, which is a unified framework that combines LiDAR, camera, and radar inputs into a shared BEV representation. The model includes modality-specific encoders followed by a Transformer block that aggregates information spatially and temporally, allowing for high-performance 3D detection and semantic map prediction. In Table 7, we summarize the characteristics of selected MLLM architectures, while Table 8 presents a comparison between LLM, VFM, VLM, and MLLM models.

4. Foundation models to mitigate challenges in autonomous driving systems (RQ1)

To guarantee the development of safe, reliable, and efficient ADS, and to ensure seamless integration of AVs into transportation networks, several challenges should be addressed. We describe below the most relevant challenges related to ADS.

4.1. Multi-modality of inputs and sensors fusion

To perceive their environment, AVs rely on a variety of sensors, including visual sensors (using cameras), proximity sensors (using LiDAR and Radar), ultrasonic sensors, navigation signals (with GPS), language instructions, and HD-Maps. Moreover, for robust and efficient perception, the large volume of gathered multi-modal data should be combined. This task is known as sensor fusion. Indeed, accurate scene understanding requires that multi-modal data is synchronized to the same spatial and temporal coordinates. However, given their heterogeneity, achieving efficient sensor fusion is challenging. Recent studies started investigating mechanisms of multi-sensor fusion and cooperative perception. They were surveyed in Xiang et al. (2023), Singh (2023), X. Zhao et al. (2023) and Hasanujjaman et al. (2023). However, with the advent of MLLMs, further investigation is needed to enable more efficient sensor fusion. For instance, authors of Choi et al. (2023) proposed semantics-guided Transformer-based sensor fusion to improve way-point predictions. Concretely, *VLMs* learn joint vision-language embeddings that align camera cues with textual/map priors, while *MLLMs* fuse image, LiDAR, and radar tokens via cross-attention to produce time-aligned scene graphs consumed by planners. Temporal consistency can be improved by using sequence models with temporal positional encodings and KV-cache alignment, allowing multi-sensor streams to be synchronized to common timestamps for downstream perception and prediction.

4.2. Safety and reliability

Designing systems that can manage in real-time sensor failures, unexpected situations, and weather conditions, and software bugs, without jeopardizing AD safety, is a critical issue. To overcome such challenges, ADS algorithms should be trained in a diversity of situations and conditions to improve their environment perception. Accordingly, XLMs are a key enabler for enhanced perception and decision-making, particularly in critical situations. In this context, authors of Nouri et al. (2024) proposed a solution that generates safety requirements via the use of a pipeline of prompts and LLMs that receive item definitions. The pipeline also reviews the requirements' dataset to identify redundant or contradictory requirements. In Y. Wang et al. (2023), the authors examined how LLMs can be integrated into ADS. They proposed techniques that utilize LLMs to make intelligent decisions in behavioral planning, and included a safety verifier for contextual learning to enhance driving performance and safety. Finally, authors in L. Wang et al. (2023) presented AccidentGPT, an E2E accident analysis and prevention framework based on the perception of the vehicle-to-everything environment and the use of MLLMs. Their objective was to improve traffic safety during the transition from manual driving to AD. In practice, *LLMs* encode traffic rules and generate natural-language constraints/rationales that act as a guard layer for planners; *VLMs* provide grounded evidence (e.g., sign and hazard cues) with confidence scores; and *MLLMs* support counterfactual reasoning and uncertainty-aware *abstention* (requesting human takeover) to reduce unsafe actions.

4.3. Complex urban environments

Navigating through complex urban environments with dynamic components such as pedestrians, vehicles, and cyclists, presents numerous challenges for ADS. Furthermore, the latter should understand

Table 8
Comparison between LLM, VFM, VLM, and MLLM. *Terminology: we use XLM as an umbrella term that includes LLM, VFM, VLM, and MLLM.*

	LLM	VFM	VLM	MLLM
Definition	Models that process and generate text using deep learning	Models that process visual inputs to extract hierarchical features or representations	Models that align and integrate vision and language features for understanding and retrieval	Unified models that integrate multiple modalities (text, images, audio) into a single reasoning framework
Main functions	Text understanding, generation, translation, tool use	Visual recognition, detection, segmentation, and representation learning	Multimodal alignment, image–text retrieval, visual grounding	Multimodal generation, reasoning, dialogue, question answering
Input types	Text	Images or videos	Images + Text	Images, Text, Video, Audio; can include maps/LiDAR/radar via adapters
Output types	Text	Class labels, segmentation maps, detection outputs	Image–text pairs, captions, matching scores	Text, image captions, dialogue, actions, code
Relevant applications	Chatbots, translation, summarization, code generation	Object detection, segmentation, image classification, video analysis	Image captioning, visual question answering, text-to-image retrieval	Multimodal dialogue, robotics, visual instruction following, video Q&A
Model architecture	Transformer-based decoders (e.g., GPT) or encoder–decoders (e.g., T5)	CNNs, ViTs, Swin Transformers	Vision encoder + projection/fusion → LLM (shared or dual towers)	LLM backbone + multi-modal encoders/adapters (cross-attn, projection, tokenizers)
Training data	Massive text corpora (e.g., books, web data)	Large-scale image/video datasets (e.g., ImageNet, COCO)	Image–text pairs (e.g., LAION, CC12M)	Multimodal instruction-tuned sets; video/audio; *driving logs* when available
Strengths	Fluent and context-aware text generation and understanding	High performance on pure visual tasks; reusable in downstream applications	Effective for cross-modal alignment and understanding (image–text retrieval, captioning)	Rich cross-modal reasoning; temporal context; tool use and planning
Weaknesses	Limited to text modality; cannot interpret visual/audio data	Cannot handle text or multi-modal tasks directly	Limited generative reasoning ability; mostly alignment-focused	High training cost and complexity; safety/latency constraints in deployment
Cross-domain transferability	Transfer within textual domains and tasks	Transfer across visual domains and tasks (e.g., detection, segmentation)	Transfer between vision and language tasks	Transfer across vision, language, audio, and instruction-based tasks
ADS impact	HMI/driver interaction; NL policy/spec extraction; log triage; data labeling; simulator scripting	Perception backbone for detection/segmentation/BEV; rare-object recall; adverse-weather robustness	Scene captioning; hazard/affordance description; grounding and visual QA for explainability	Multi-sensor grounding (RGB/LiDAR/ /maps); intention prediction; planner-in-the-loop suggestions; NL tasking of tools; closed-/open-loop evaluation

and adhere to local traffic laws and conventions, which can vary between regions. Accordingly, authors of Luo et al. (2024) investigated multi-modal multi-task visual understanding FMs, specifically designed for road scenes. These models leverage multi-modal and multi-task learning capabilities to process and fuse data from diverse sources, enabling them to handle various driving-related tasks with adaptability. Here, *VLMs* enhance fine-grained perception under occlusion and adverse weather, while *LLMs* lift perceptual outputs into rule-consistent intents (e.g., yielding, lane selection). *MLLMs* jointly reason over agent interactions and scene context (e.g., crosswalk occupancy + pedestrian intent) to stabilize trajectory prediction and negotiation in dense traffic.

4.4. Data privacy and security

Given the large amount of data generated and processed by AVs, there is a serious risk of data/sensor breaches, alteration, and/or eavesdropping. XLMs can be exploited to protect the AVs' data. For instance, authors of Aldeen et al. (2024) proposed the use of MLLMs to mitigate natural denoising diffusion (NDD) attacks on traffic signs and integrate them into ADS. Privacy can be preserved by *on-vehicle PEFT/LoRA* fine-tuning and *federated* or *DP*-aware adaptation, avoiding raw data sharing. From a security angle, *VLMs* help detect spoofed/perturbed signage and sensor-level attacks, while *LLMs* support anomaly triage, audit trail generation, and red-teaming of instruction channels (HMI/V2X).

4.5. Human–Machine interaction

The development of AVs brings new HMI opportunities and challenges. Indeed, as AVs evolve, understanding and responding to human intent becomes a significant requirement. Therefore, a smooth and intuitive interaction between AVs and drivers, passengers, and other road users, is required to realize large-scale adoption of ADS. Designing interfaces that allow passengers to understand the AV's actions and intentions is necessary. The integration of chat-bots, voice-to-text, text-to-voice, text-to-image, and image-to-text functionalities into AVs would enhance HMI and make it more intuitive and natural. In this context, authors of Cui et al. (2023a) studied how integrating LLMs with Human Digital Twin (HDT) can change HMI for AD. Similarly, Yang et al. highlighted in Y. Yang et al. (2024) the benefits of integrating LLMs into ADS. Specifically, they conducted experiments using various LLM models and prompt designs to evaluate their effectiveness in few-shot multivariate binary classification tasks. Results have shown that GPT-4 is the most accurate one in task understanding and responding, compared to other LLMs such as CodeLlama. Finally, the authors of Xu et al. (2024) introduced in DriveGPT4, an interpretable E2E ADS based on LLM. They showed that DriveGPT4 can process textual queries and multi-frame video inputs, thus facilitating the interpretation of vehicle actions with reasoning. Operationally, *LLMs* enable dialogue for intent clarification and post-hoc explanations of actions; *VLMs* ground these explanations in visual evidence (frames, saliency); and *MLLMs* support multimodal queries from occupants or first responders, improving trust and accountability.

4.6. Real-time constraints and edge deployment

Although XLMs exhibit remarkable performance in perception, captioning, and reasoning tasks, their large-scale architectures pose significant latency and memory challenges in real-time systems. Autonomous vehicles require rapid and deterministic inference, often within 50–100 ms, to ensure safe perception and control. In contrast, many existing XLMs operate with high latency and resource demands, making them impractical for direct deployment on embedded platforms without adaptation. To mitigate these issues, recent research has explored multiple optimization strategies. These include: (i) model compression (e.g., pruning and quantization) to reduce size and inference cost, (ii) knowledge distillation to train lightweight student models from larger teacher models, (iii) hardware-aware architecture design for better performance on GPUs or edge accelerators, and (iv) edge deployment frameworks such as NVIDIA Jetson, Qualcomm Cloud AI 100, and EdgeTPU, which support XLM inference under constrained power budgets. For instance, Transformer variants like MobileViT (MehtaS, 2022), TinyViT (Wu et al., 2022), and ED-ViT (X. Liu et al., 2024) have been developed specifically to enable real-time operation on edge devices. Some works (Tang et al., 2023; Zang et al., 2022; Y. Xu et al., 2023) have also proposed hybrid pipelines, where lightweight vision encoders handle low-latency tasks and heavier XLMs are called selectively or asynchronously for non-real-time tasks. In ADS terms, this maps to keeping the *control loop* on fast vision encoders (and distilled *planner* heads), while invoking richer *LLM/MLLM* reasoning via speculative decoding, mixed precision, and KV-cache offloading only when slack is available, thus meeting 50–100 ms deadlines without sacrificing capability.

5. Proposed taxonomy of foundation models based approaches for autonomous driving systems

This section introduces our proposed taxonomy for the application of XLMs in the context of AD. The taxonomy is designed to address three research questions, namely RQ2, RQ3, and RQ4. The detailed approaches related to the proposed taxonomy will be discussed in the remaining sections of the survey. Specifically, Section 6 divides the use of LLMs in AD into four main categories as follows:

1. **Prompt engineering-based methods:** Within this category, studied contributions are classified according to the provided AD tasks, which are planning and control, perception, multi-tasking, and question-answering.
2. **Fine-tuning-based methods of pre-trained models:** Under this category, we consider approaches that fine-tuned pre-trained LLMs for precise planning and control tasks, lane change maneuvers, and path planning.
3. **RLHF-based methods:** We classified the studied solutions within this category into approaches that utilize RLHF to improve decision-making in planning and control, and to generate scenarios to test and improve the AV's response to critical situations.
4. **LLM and GAI-based methods:** This category explores the integration of LLMs with GAI to design advanced AD solutions.

Subsequently, Section 7 categorizes the use of VLMs in AD into two categories as follows:

1. **Prompt engineering-based methods:** Here, we discuss solutions to improve the AV's perception of its environment using VLMs including VLMs' implementations to answer visual and scene interpretation queries.
2. **Fine-tuning-based methods:** This category includes frameworks based on VLMs and makes use of fine-tuning techniques to improve accuracy in perceiving the driving environment, as well as frameworks that adjust VLMs for improved question-answering tasks related to visual data.

Finally, Section 8 focuses on the integration of MLLMs in ADS. The studied methods in this section are classified into four categories as follows:

1. **Prompt engineering-based methods:** Under this category, studied approaches include solutions that use MLLMs to enhance the environment perception and those that improve the performances of Q&A tasks.
2. **Fine-tuning-based methods:** Studied frameworks in this category aim to enhance planning and control tasks, trajectory planning, perception tasks, and question-answering.
3. **RLHF-based methods:** Studies addressed here use RLHF to train agents for better perception, planning, and control, and to enhance waypoint prediction.
4. **MLLM and GAI-based methods:** This category investigates the integration of MLLMs with GAI to develop more efficient AD solutions.

The proposed taxonomy is synthesized in Fig. 7, which conceptually maps the roles of LLMs, VLMs, and MLLMs within the autonomous driving stack. Each color, coded block indicates the primary contribution area, ranging from perception to HMI/V2X, thereby illustrating how different classes of XLMs support complementary ADS functionalities. In the sequel, we detail the aforementioned three sections.

6. Large language models for autonomous driving (RQ2)

Various approaches have been proposed for integrating LLMs with ADS.

For instance, the authors of Tanahashi et al. (2023) quantitatively evaluated the Spatial-Aware Decision-Making (SADM) and Traffic Rules Satisfaction (TRS) abilities of LLMs. Moreover, to implement LLMs within ADS, several strategies have been developed, including methods based on prompt engineering, fine-tuning, RLHF, and GAI.

6.1. Prompt-engineering-based methods

Prompt engineering uses queries (prompts) to guide the output of LLMs. In this section, we discuss the related work that proposed LLM-based prompt engineering toward AD task provisioning. Planning and control are critical components that determine the AV's ability to navigate and make decisions in real time. Recent advancements in LLMs and prompt engineering enabled novel approaches to enhance planning and control in ADS. For instance, Zhou et al. integrated in Z. Zhou et al. (2024) LLMs with RL to enhance AD agents, making them more efficient. The proposed framework integrated GPT-3.5-turbo with an RL agent based on deep Q-Learning to take driving control actions. The LLM is used as a proxy for reward calculations, i.e., it interprets textual prompts (e.g., task description, objective, and last outcome) to generate reward signals that influence the RL agent's behavior. Results showed that RL agents guided by LLMs achieved more balanced and human-like behaviors compared to traditional RL agents. Also, authors of Wen et al. (2023) designed the DiLu framework, illustrated in Fig. 8, that integrated LLM in ADS to develop four modules, namely an AD simulation environment module, an AD memory module to acquire and save experience from past driving scenarios, a reasoning module that generates reasoning chains and provides AD decisions, and a reflection module to correct the reasoning process for future driving scenarios. Authors of Sha et al. (2023) designed "LanguageMPC", which is an LLM prompt engineering-based framework, that uses LLMs in reasoning and understanding high-level information with Model Predictive Control (MPC) to execute specific driving actions.

In Azarafza et al. (2024), the authors proposed a generative driver agent simulation framework to perceive complex traffic scenarios and provide realistic driving maneuvers. Their framework uses LLMs to generate responses based on input prompts, enabling the driver agent to comprehend complex driving scenarios.

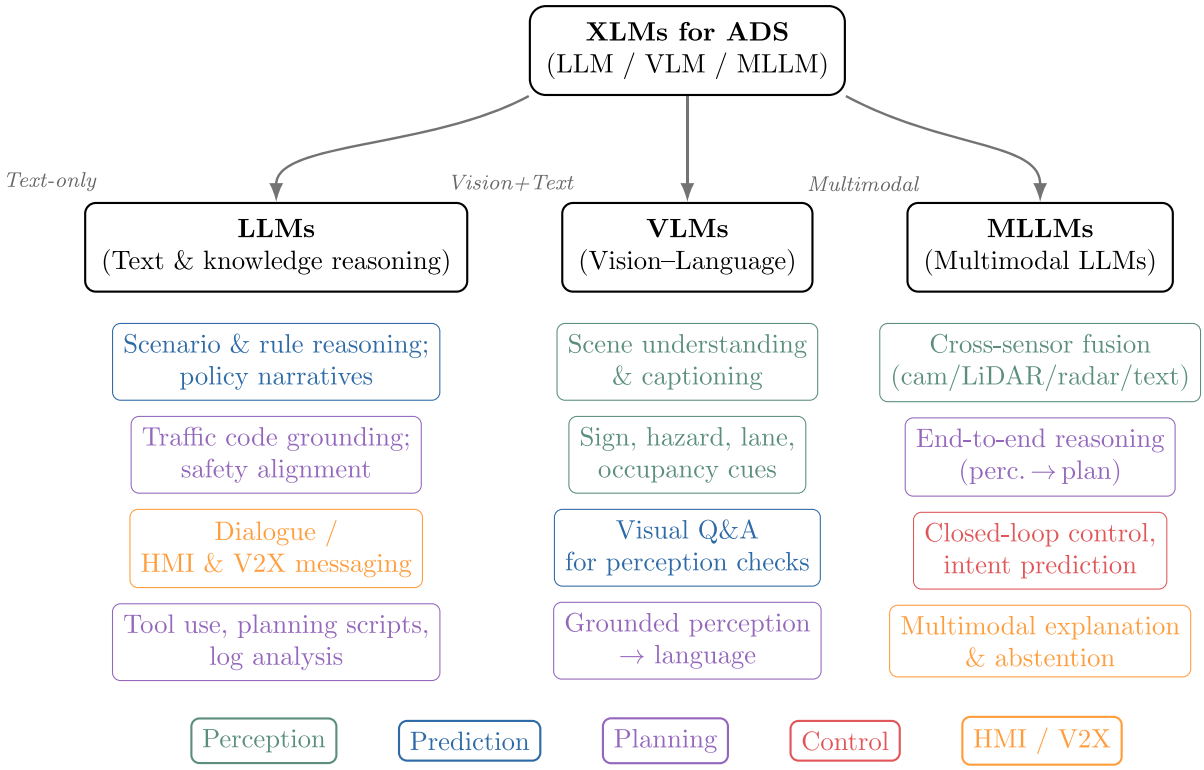

Fig. 7. Conceptual taxonomy of **XLMs** (LLMs, VLMs, MLLMs) aligned with the ADS stack. Each capability block is color-tagged by its primary contribution area: Perception, Prediction, Planning, Control, and HMI/V2X.

The authors compared the accuracy of LLMs to human-generated ground truth using the CARLA simulator and found that combining detected objects and sensor data in the LLM provides precise brake information.

The SurrealDriver framework proposed in Jin et al. (2023) and illustrated in Fig. 9, is a generative driver agent simulator that uses LLMs. It is composed of three main modules, namely DriverAgent, CoachAgent, and short-term memory, it performs multiple tasks, including perception and control. Within the SurrealDriver framework, the LLM processes the collected environment data parameters to comprehend the driving situation and make decisions guided by predefined requirements and safety rules. The DriverAgent module generates JSON-formatted commands to control the vehicle such as maintaining speed, lane changing, stopping, speeding up, and slowing down. The CoachAgent module is based on the use of CoT prompts of feedback from 24 drivers to refine the DriverAgent behavior and make it similar to human driving. Finally, the short-term memory stores recent driving behaviors to ensure continuity and consistency in decision-making. Similarly, authors of Miceli-Barone et al. (2023) proposed a system that generates self-driving simulation scenarios using natural language dialogue with GPT-4. It allows iterative refinement of scenarios through user-induced language-based instructions and corrections, which are translated by the LLM into executable simulation code.

LLMs for question-answering focus on designing and optimizing prompts to respond to various questions. Accordingly, authors of Chen et al. (2024a) proposed "LLM-driver", shown in Fig. 10, a framework that integrates numeric vector modalities into pre-trained LLMs. This method uses object-level 2D scene representations to fuse vector data into a pre-trained LLM with adapters. The model can interpret driving situations and generate adequate actions. A dataset with 160,000 Q&A pairs generated by GPT-3.5 from 10,000 driving scenarios was introduced and associated with high-quality control commands, collected from an RL driving agent, to fine-tune the model.

The model's performance was evaluated in terms of action prediction Mean Absolute Error (MAE), traffic light detection accuracy, and normalized errors of acceleration, brake pressure, and steering.

6.2. Fine-tuning-based methods

Wang et al. developed in T. Wang et al. (2024) DriveCoT, an E2E driving dataset using the CARLA simulator. It included complex driving scenarios (e.g., lane-changing and high-speed driving), and incorporated a CoT labeling scheme to provide reasoning processes for driving decisions. Also, they designed a DriveCoT-agent model trained on the DriveCoT dataset. The latter showcased strong performances in open-loop and closed-loop evaluations.

In Peng et al. (2024), the authors proposed Lane Change Large Language Model (LC-LLM), an explainable lane change prediction model reconceptualized as a language modeling problem and solved using LLMs. The core of the LC-LLM framework is a fine-tuned LLM for the lane change and trajectory prediction tasks for highway AD. Through experiments, it is shown that LC-LLM can accurately predict lane change intentions and trajectories compared to benchmarks based on long Short-Term Memory (LSTM) or Transformers.

6.3. Reinforcement learning with human feedback-based methods

LLMs with RLHF represent a cutting-edge approach to enhance vehicle decision-making and safety. The integration of human feedback into the training process increases the model's performance via a better understanding of the complex driving environments and scenes.

Authors in R. Yang et al. (2024) developed LLM-based driver agents endowed with reasoning and decision-making capabilities aligned with human driving behaviors. Their multi-alignment framework integrates demonstrations, human feedback, and reinforcement signals to progressively adapt the model's driving policy toward human-like actions. In addition, the multi-alignment framework introduces a Coach Agent that evaluates past driving behaviors and formulates driving guidelines. The framework's effectiveness was validated in the CARLA simulator, showing significant improvements in behavioral consistency and scenario adaptability.

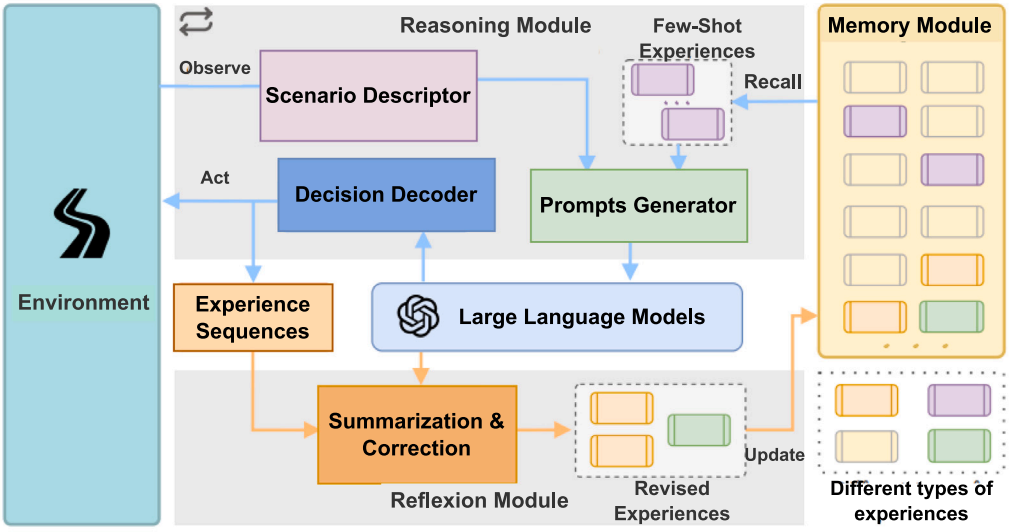

Fig. 8. Architecture of DiLu framework.

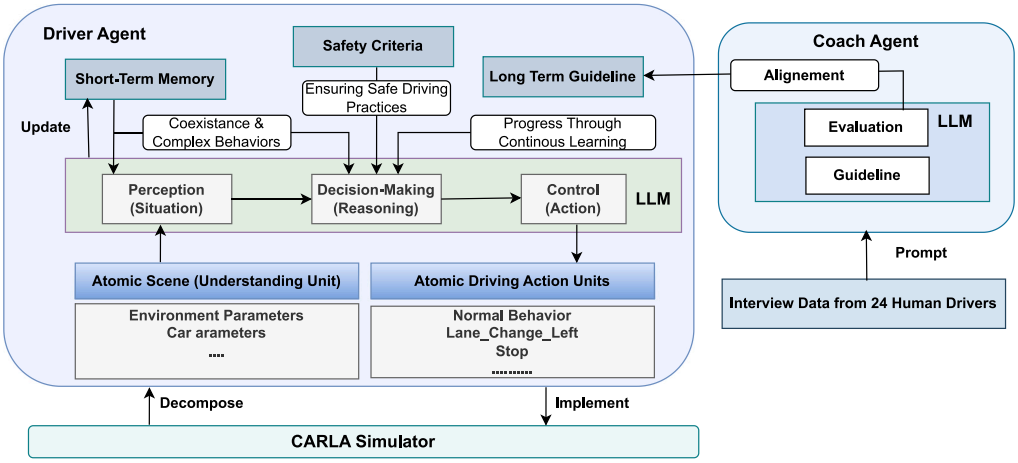

Fig. 9. Architecture of the SurrealDriver framework.

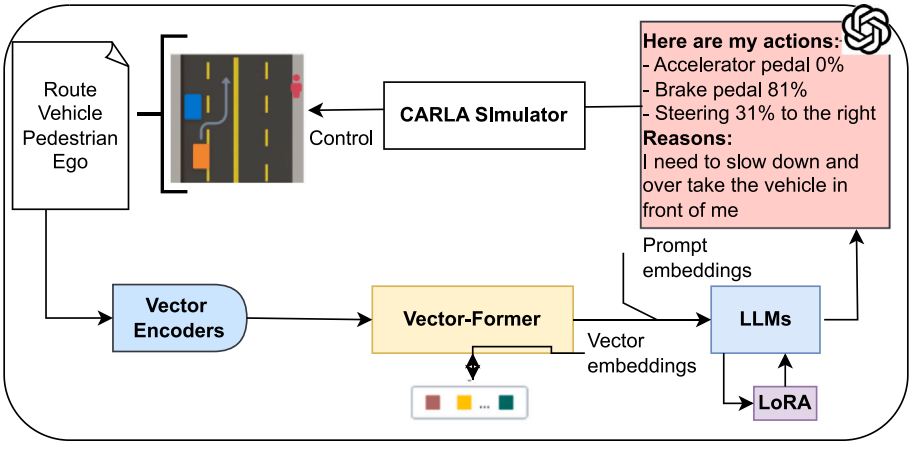

Fig. 10. Architecture of LLM-driver framework.

Similarly, H. Tian et al. (2024) introduced a closed-loop reinforcement learning framework named CRITICAL, designed to enhance both training and evaluation of autonomous driving agents. CRITICAL automatically generates challenging and high-risk driving scenarios. An integrated LLM module refines and diversifies the generated scenarios by leveraging historical training data and domain knowledge from real-world traffic. It also uses real-world traffic dynamics data from the highD dataset that classifies real-world traffic into normal, aggressive,

Table 9
Comparative study of LLM approaches for AD.

Year	Ref.	Model	Datasets and tools	Trainable modules	AD services
2023	Tanahashi et al. (2023)	GPT-3.5-turbo-1106	HighwayEnv	RL agent integrating LLM	Human-like driving behavior; Dynamic adaptation to driving conditions; Customized driving styles.
2023	Sha et al. (2023)	ChatGPT-3.5	IdSim	MPC; RL agent	Single-vehicle decision-making; Multi-vehicle coordination; Driving behavior modulation
2023	Jin et al. (2023)	GPT-3; GPT-4	Driving Behavior Data, simulation data generated with CARLA; CARLA simulator, NLP library	DriverAgent (Perception, control, and decision-making); CoachAgent (Driving scenario generation and evaluation)	ADS simulation and testing; Safety assurance; Real-time monitoring
2023	Chen et al. (2024a)	GPT-3.5	160k QA driving pairs dataset and Control Commands Dataset (15 virtual environments); 2D driving simulator	PPO-based RL agent; LoRA modules	Perception, action prediction, and driving Q&A
2024	Z. Zhou et al. (2024)	GPT-3; GPT-4; BERT; T5	CitySim dataset; Highway-env, graph representation and box plots	Few-shot learning model	Closed-loop driving task.
2024	Azarafza et al. (2024)	GPT-4	CARLA simulator and YOLOv8 object detector	Common sense knowledge; Common sense reasoning	Object detection and localization in diverse weather conditions.
2024	T. Wang et al. (2024)	Unspecified	Private dataset generated with CARLA; CARLA simulator	Unspecified	Interpretability of E2E ADS: Perception, planning, prediction, and reasoning
2024	Peng et al. (2024)	LLaMA-7B; LLaMA-13B; LLaMA-70B	highD dataset; LoRA,	Supervised fine-tuning	Lane change intention prediction; Explainable predictions
2024	R. Yang et al. (2024)	GPT-4	Private dataset from human drivers; CARLA simulator, NLP library, RL library	Language understanding; Behavioral cloning; RL agent; Decision-making	Behavioral alignment; Human-in-the-loop system
2024	H. Tian et al. (2024)	Mistral-7B-Instruct	HighD Dataset; HighwayEnv Simulation, and LongChain	Scenario generation; PPO-based RL agent; closed-loop feedback	AD scenario generation; Dynamic scenario modification; Safety measures analysis
2024	G. Zhao et al. (2024)	GPT-3.5	Trajectory-to-HDMap Dataset and Multi-View Video Dataset; Python libraries	Function Library; Embedding HDMaps and 3D Boxes; UniMVM module	Perception; Video generation of driving scenarios; AD scenario simulation

and defensive driving behaviors. Besides that, to quantify scenario severity, surrogate safety indicators such as time-to-collision and a unified risk index were employed. In addition, the framework utilizes PPO to train the RL agent within the HighwayEnv simulation environment. CRITICAL establishes a feedback loop where data from training episodes, including failure reports and risk metrics, is continuously collected and analyzed. This feedback is used to modify the environment configuration, generating new critical scenarios that challenge the RL agent.

6.4. Large language model and generative artificial intelligence-based methods

The DriveDreamer-2 framework proposed in G. Zhao et al. (2024), which is an extension of DriveDreamer (X. Wang et al., 2023), was designed to generate user-customized synthetic and realistic multi-view driving videos used for training, testing and validation of ADS efficiency. Specifically, it generates driving videos by combining structured conditions (e.g., HD-Maps and 3D boxes) with image features. The system uses encoders to embed HD-Maps, 3D boxes, and image frames into latent space features, which are then processed to produce the final videos. Then, using the Unified Multi-View Model (UniMVM), the spatial and temporal coherence of generated videos is enhanced. The architecture of DriverDream-2 is presented in Fig. 11. In Table 9, we present a comparative study of the above works.

6.5. Lessons learned from large language model integration in autonomous driving

Across the surveyed approaches, several practical lessons emerge. Prompt-based pipelines offer transparent reasoning traces and fast iteration in simulators, yet their behavior is brittle to prompt phrasing and context drift; without explicit grounding and guardrails, generalization degrades under distribution shift. Fine-tuning of pretrained language models improves task specificity and explainability (for example, chain-of-thought driving rationales or lane-change intent), but it demands costly supervision and can overfit to the operational design domain, reducing robustness in novel scenes. Reinforcement learning with human feedback aligns policies with human preferences and supports closed-loop decision-making, though it remains challenging to scale safely due to sample efficiency, exploration risk, and strict real-time control deadlines. Generative data pipelines broaden scenario coverage and enable rare-event validation, while raising questions about realism, controllability, and measurable gaps between synthetic and real distributions. Overall, deployment-oriented systems increasingly favor hybrid designs that couple lightweight edge perception and control with deferred high-level reasoning, explicit safety constraints, and continual feedback; striking a workable balance among interpretability, latency, and reliability, and motivating evaluation protocols that report both task performance and system-level metrics.

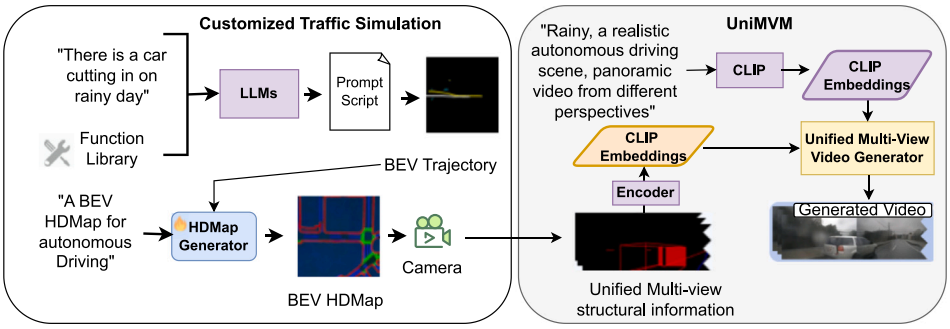

Fig. 11. Architecture of DriverDream-2 framework.

7. Vision-language models for autonomous driving (RQ3)

7.1. Prompt-engineering-based method

In Guo et al. (2024), Guo et al. proposed the Co-driver framework, based on ViT, illustrating the potential of prompt engineering in planning and control, and trajectory prediction tasks. Co-driver employs prompt engineering to understand visual inputs and generate driving instructions, which are then fed to a Deep RL (DRL) agent for driving actions.

Also, it predicts trajectories through the integration of map contexts and past vehicle positions using scene encoding and path classification.

The Co-driver architecture is presented in Fig. 12.

7.2. Fine-tuning-based methods

Perception in ADS is crucial for interpreting and understanding the vehicle's surroundings, thus accurately detecting and recognizing objects, pedestrians, and road conditions. This yields informed decisions in real-time by the ADS, thus improving the overall safety and performance. In X. Tian et al. (2024), Tian et al. proposed DriveVLM and DriveVLM-Dual frameworks. DriveVLM uses a ViT encoder to process images and extract visual features. Then, a text tokenizer and an LLM, supported by CoT, process textual data and generate scene descriptions and driving decisions. Finally, CoT decisions are converted into vehicle control commands for execution. DriveVLM focused on the scene understanding and planning tasks, while DriveVLM-Dual extended the work to integrate VLMs with conventional AD methods, thus strengthening the spatial understanding of the driving environment and speeding real-time inference. Moreover, the corner Case Object Detection and Analysis for Large Models (CODA-LM) framework was proposed in Li et al. (2024b) to evaluate VLMs' performances in complex driving scenarios. It provides a detailed analysis of the VLMs' strengths and weaknesses when handling corner cases, thus providing insights into the areas that need further investigation to enable AD. Finally, Kou et al. proposed in Kou et al. (2024) the Personalized Federated Learning Large Vision Model (PFedLVM) framework aiming to improve perception by leveraging Federated Learning (FL) and personalization. Specifically, LVM is deployed only within the aggregation server, the latent feature-based FL is used to exchange compressed feature maps, while personalized learning ensures that each AV model learns from the others but preserves its unique characteristics. Its architecture is presented in Fig. 13.

Authors of Gopalkrishnan et al. (2024) proposed "Efficient, Lightweight Multi-Frame Vision-Language Model for Visual Question Answering in Autonomous Driving", also known as, EM-VLM4AD, a VLM model focusing on performing Q&A tasks for AD. To produce interpretable responses, EM-VLM4AD integrates multiple camera views into a unified visual representation and combines it with text embeddings generated with ViT. Specifically, it uses a fine-tuned T5-Medium model and an 8-bit quantized T5-Large model fine-tuned via LoRA to align with the concatenated multi-view image and text embeddings. The EM-VLM4AD architecture is shown in Fig. 14.

In Table 10, we present a comparative study of the presented works.

7.3. Lessons learned from vision-language models in autonomous driving

Across the surveyed vision-language approaches, several deployment-relevant lessons emerge. Prompt-driven systems can translate rich scene context into human-readable rationales and high-level commands, but their behavior is sensitive to prompt design and visual grounding quality; without explicit spatial anchoring and temporal memory, they drift under occlusion, adverse weather, or viewpoint change. Fine-tuned models improve robustness on target tasks (e.g., multi-view scene understanding, trajectory hints), yet they demand carefully curated supervision, strict camera/LiDAR time-sync, and calibration to avoid brittle cross-sensor fusion. Multi-frame designs help stabilize perception but introduce latency-throughput trade-offs that must be balanced against real-time control budgets; practical stacks therefore partition work, keeping low-latency detectors/planners on the edge while reserving vision-language reasoning for asynchronous guidance, explanation, and operator interaction. Federated or personalized training mitigates fleet heterogeneity and privacy risk, though feature-only aggregation can cap ultimate accuracy without periodic central refreshing. Evaluation needs to go beyond open-loop VQA or captioning: closed-loop metrics (route completion, intervention rate), robustness probes (lighting, weather, corner cases), and safety monitors (rule checks, abstention) are essential to reveal failure modes.

8. Multimodal large language models for autonomous driving (RQ4)

8.1. Prompt-engineering-based methods

In D. Wu et al. (2023), Wu et al. proposed a new large-scale language prompt dataset for driving scenes, called "NuPrompt", specializing in 3D objects and is built on NuScenes dataset (Caesar et al., 2020) for multi-view 3D object detection. Also, they introduced an approach to 3D object detection and tracking by integrating cross-modal features within prompt reasoning, called "PromptTrack". This approach outperformed traditional object detection methods due to its novel fusion of multi-modal inputs. The architecture of PromptTrack is illustrated in Fig. 15, where the Transformer decoder processes each frame's visual attributes and inquiries to generate decoded questions. Also, the past reasoning module improves and refines tracking based on historical queries, whereas the future reasoning module facilitates cross-frame query propagation. Finally, the prompt reasoning branch predicts prompt-related tracks. Authors of Ding et al. (2023) proposed "High-Resolution Understanding in MLLMs for Autonomous Driving", also known as, HiLM-D, as an efficient technique to incorporate a high-resolution information into MLLMs for risk object localization

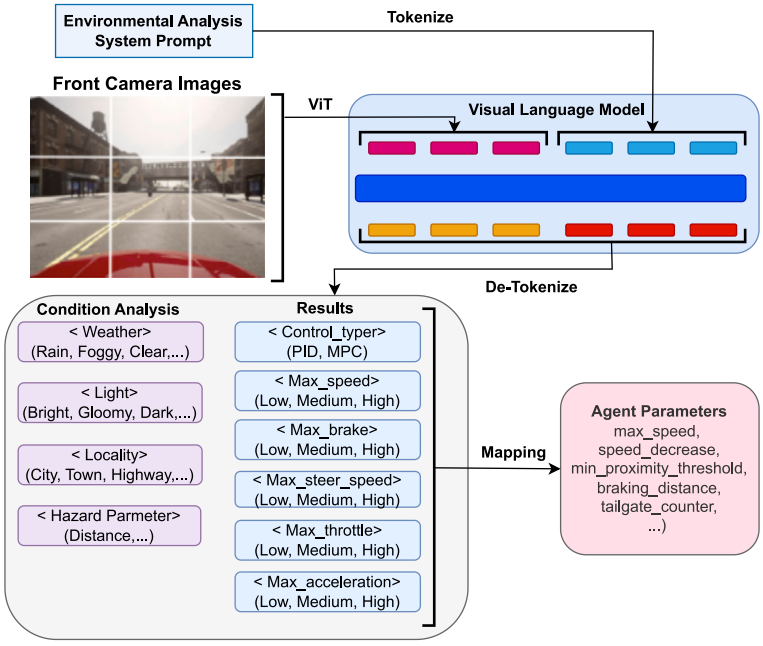

Fig. 12. Architecture of Co-driver framework.

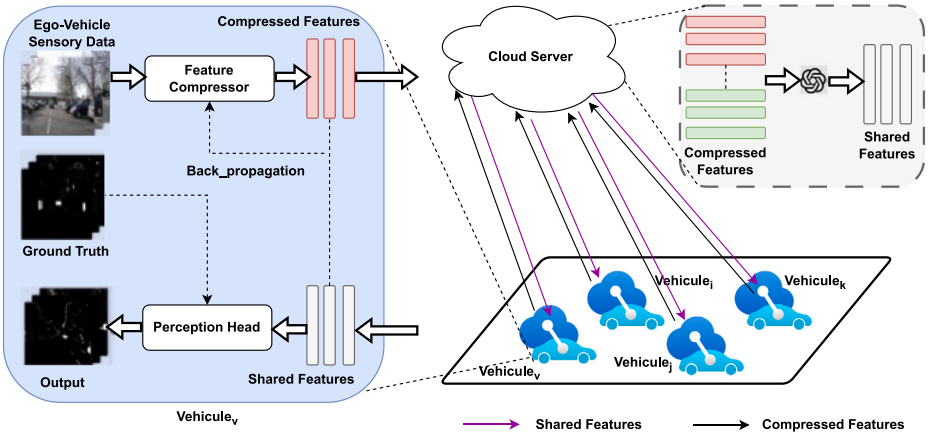

Fig. 13. Architecture of pFedLVM framework.

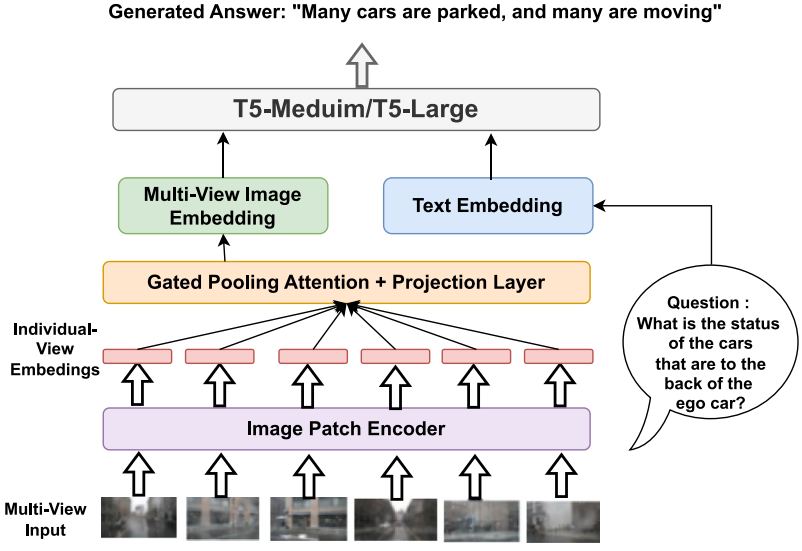

Fig. 14. Architecture of EM-VLM4AD framework.

Table 10
Comparative study of VLM approaches for AD.

Year	Ref.	Model	Datasets and tools	Trainable modules	AD services
2024	Guo et al. (2024)	Qwen-VL-9.6B	Customized dataset with image sets and corresponding prompts; CARLA simulator and ROS2	Qwen-VL fine-tuning using Quantized Low-Rank Adaptation (QLoRA)	Adjustable driving behaviors; Trajectory and lane prediction; Planning and control
2024	X. Tian et al. (2024)	Qwen-VL	SUP-AD dataset and NuScenes	ViT encoder; LLM	Scene understanding; Motion planning; Decision-making; Trajectory planning
2024	Li et al. (2024b)	Flamingo	Ego4D, Waymo Open Dataset, NuScenes, and BDD100K; Cityscapes, OpenCV, PyTorch, COCO API, and Detectron2	ViT encoder; CLIP	Object detection; Scene understanding; Path planning; Obstacle avoidance.
2024	Kou et al. (2024)	ViT	Cityscapes and CamVid	Feature extraction; Compression; Backpropagation	Object detection; Semantic segmentation; Vehicle-specific behavior modeling
2024	Gopalkrishnan et al. (2024)	T5-Medium; T5-Large; ViT	DriveLM dataset; LoRA,	Attention and Projection; LLM fine-tuning	Question-answering

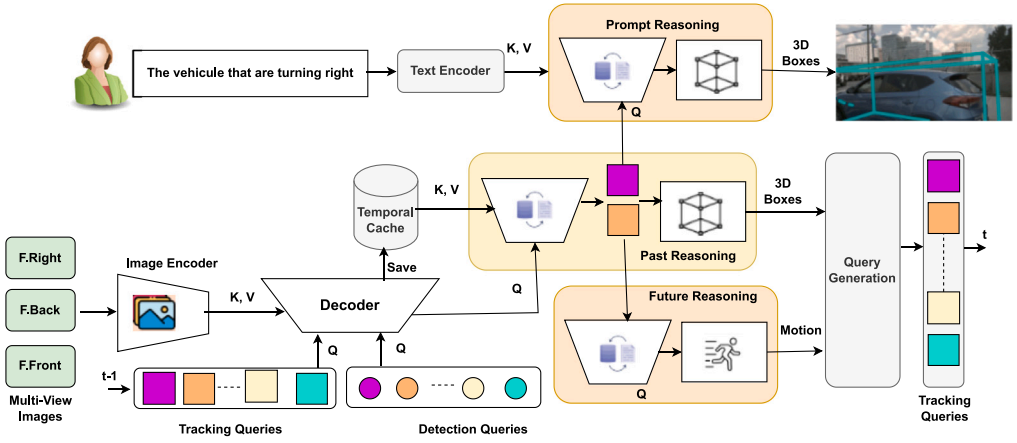

Fig. 15. Architecture of PromptTrack framework.

and intention prediction. Through experiments, the authors proved the superiority of HiLM-D in localizing obstacles and predicting vehicles' intentions compared to two benchmarks, namely eP-ALM (Shukor et al., 2023) and Video-LLaMA (Zhang et al., 2023). Finally, RAG-driver is a recent framework based on prompt engineering that generates driving explanations where the MLLM is responding to queries about driving scenarios (Yuan et al., 2024). It leverages a retrieval engine that looks for similar driving experiences to the one under analysis, then the MLLM processes both the current query and the in-context learning samples that have been retrieved in its memory to provide action explanation, justification, and next control signal prediction.

8.2. Fine-tuning-based methods

In W. Wang et al. (2023), the authors proposed DriveMLM, an LLM-based framework performing close-loop AD in realistic simulators. Indeed, it integrates three modules: (1) the "Behavioral Planning" module that utilizes MLLM to incorporate user commands, driving rules, and sensor inputs for driving decisions and explanations, (2) the "Behavioral Planning States Alignment", which aligns LLM's linguistic decision outputs with the behavioral planning module, and (3) the "Data Engine" which collects data with decision states and explanations for model training and evaluation. According to the authors, DriveMLM improves driving scores compared to Apollo baseline, on the "CARLA Town05 Long" benchmark. The DriveMLM framework is illustrated in Fig. 16. Moreover, Han et al. designed in W. Han et al. (2024) an ADS named Decision-Making and Execution-driver (DME-Driver) that generates NLP-based driving decisions based on correlation vehicle status and visual inputs, which are then converted into control commands. It is based on a pre-trained CLIP visual encoder to convert visual information into feature tokens, and a text tokenizer to encode prompt inputs and current status information. The latter are processed using LLaMA 2 model. It has been shown that DME-Driver outperforms GPT-4V in terms of accuracy across several tasks, including gaze (85.2% vs. 75.3% for GPT-4V), scene understanding (86.5% vs. 79.2% for GPT-4V), and logic (80.3% vs. 65.4% for GPT-4V).

DriVLMe, presented in Y. Huang et al. (2024), is a fine-tuning MLLM for trajectory planning. It is a video-language-model-based AD agent facilitating communication between humans and AVs that perceive the environment and navigate. Specifically, it learns from embodied experiences in a simulated environment and social experiences from real human dialogue to find the shortest paths from the agent's current location to the destination specified by the MLLM.

In addition, Liao et al. proposed in Liao et al. (2024) VLM2Scene, a fine-tuning method for AD perception. It shifts from traditional point-level contrastive learning to region-level learning to address the inherent sparsity and noise in LiDAR point clouds. To do so, it leverages region masks derived from the Segment Anything Model (SAM). VLM2Scene fine-tunes the learning process to enhance the model's perception accuracy and robustness by introducing a semantic-filtered region-learning and a region-semantic assignment strategy. Similarly, Liang et al. (2024) designs the Automatic Data Engine (AIDE) that automates data labeling, model training, and driving scenario generation for model evaluation, aiming to improve AV perception models continuously.

From a multi-tasking perspective, Ding et al. presented their BEV-InMLLM framework in Ding et al. (2024), which they introduced as

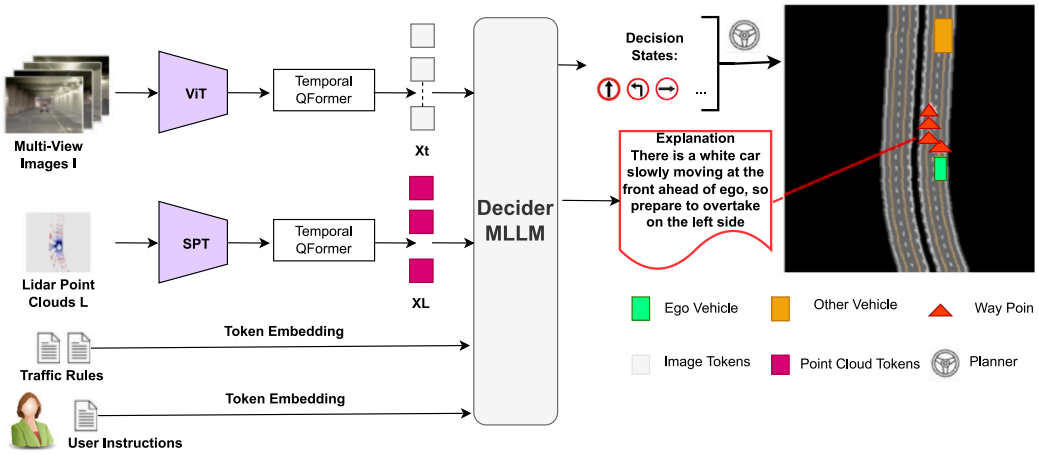

Fig. 16. Architecture of DriveMLM framework.

NuInstruct, a novel dataset comprising 91,000 multi-view video-QA pairs across 17 subtasks. One of the main contributions of this work consists of injecting Bird's-Eye-View (BEV) data into MLLM to fine-tune the models. The proposed BEV-InMLLM approach uses a specialized multi-view Q-Former to handle multi-view video inputs and capture temporal appearance information across different views, thus accurately localizing objects and estimating distances. Furthermore, the framework OmniDrive features a novel 3D vision-language model that utilizes sparse queries to extract and compress visual representations into 3D before inputting them into a language model and a comprehensive Visual Question Answering (VQA) module that includes traffic regulation, scene description, 3D grounding, decision-making, planning, and counterfactual reasoning (S. Wang et al., 2024). Based on the experiments, OmniDrive demonstrated high reasoning and planning capabilities in complex 3D scenes. Finally, DriveGPT4 proposed fine-tuned MLLMs for question-answering in ADS (Xu et al., 2024). DriveGPT4 is an interpretable E2E ADS that processes both multi-frame video inputs and textual queries to explain taken AD actions to the human user.

8.3. Reinforcement learning with human feedback-based methods

Mao et al. presented in Mao et al. (2023b) Agent-driver, an MLLM-based intelligent agent for AD that integrates: (1) a versatile tool library for dynamic perception and prediction, (2) a cognitive memory for human knowledge, and (3) a reasoning engine that emulates human decision-making using CoT technique. Through experimentation, Agent-driver is proven to outperform state-of-the-art AD methods in terms of interpretability and few-shot learning. Also, authors in Ray and Ohn-Bar (2024) proposed a feedback-guided E2E sensorimotor driving agent with MLLM to provide a language interface for user control and refinement. The objective is to train the sensorimotor agent to map front camera images and AV's state information encoded as language tokens and predict a set of future waypoints. The training phase involves two stages: (1) The privileged agent takes ground truth environmental information and provides rich supervision for training the sensorimotor agent, and (2) the sensorimotor agent is fine-tuned with prompt-based feedback to enable efficient failure reasoning. Simulation results demonstrated the efficiency of this method. Further architectural details of this framework are illustrated in Fig. 17.

Authors in Z. Li et al. (2025) propose a multimodal deep reinforcement learning framework for collision-free vehicle navigation that decouples perception and control while leveraging camera, LiDAR, and IMU streams to build a rich state for policy learning. A cross-domain self-attention module strengthens visual-LiDAR fusion; a recurrent deduction mechanism aggregates local observations into globally consistent decisions; and a Self-Assessment Gradient Model (SAGM) stabilizes and accelerates policy optimization. Policies are trained in CARLA and transferred to a physical platform via model compression and parameter freezing, narrowing the sim-to-real gap. Across diverse weather and lighting conditions, the method outperforms strong RL baselines in success rate, stability, and safety (collisions/off-lane), with ablations confirming the benefits of multimodal fusion and stacked cross-domain attention.

Authors in C. Xu et al. (2025) propose a practical DRL pipeline for AD that replaces raw sensor inputs with pre-defined multimodal representations, a denoised BEV map, a cleaned semantic front view, and compact ego/neighbor state vectors, so the agent learns state–action couplings faster and with fewer spurious cues. shared feature-extraction module compresses these high-dimensional inputs once per step and feeds SAC actor–critic heads, cutting redundant computation (≈3× speedup over embedded feature blocks) while preserving end-to-end training. A safety/efficiency/comfort reward design guides behavior. In CARLA, the multimodal fusion setting (BEV+front view+states) consistently outperforms raw-image baselines and strong fusion alternatives (e.g., BEV-centric models), remaining robust under injected noise and achieving higher returns with tighter lane keeping and higher average speeds. The results show that structuring perception into task-aligned images and sharing features across networks materially improve convergence and closed-loop driving quality.

In Huang et al. (2025), VLM-RL is proposed as a unified framework that couples pre-trained VLMs with reinforcement learning to eliminate manual reward engineering for autonomous driving. Its key idea is a Contrasting Language Goal (CLG)-as-Reward paradigm, where positive and negative natural-language goals are contrasted to produce semantic reward signals from image observations, complemented by a hierarchical reward synthesis that fuses CLG rewards with vehicle-state information for stability. The method further improves training efficiency via batch reward computation from replay buffers, and remains plug-and-play with standard RL algorithms. Extensive CARLA experiments show strong gains over prior baselines, including lower collision rates, higher route completion, and robust generalization to unseen scenarios, demonstrating a practical path toward scalable, semantics-guided RL driving policies.

8.4. Multimodal large language model and generative artificial intelligence-based methods

Authors of Jia et al. (2023) introduced the notion of the interleaved vision-action pair, which unifies the format of visual features and control signals, leading to the general world model, called ADriver-I. The latter can predict the control signal of the current frame based on the vision-action pairs as inputs. Then, it predicts and generates future frames based on the generated control signals with the vision-action

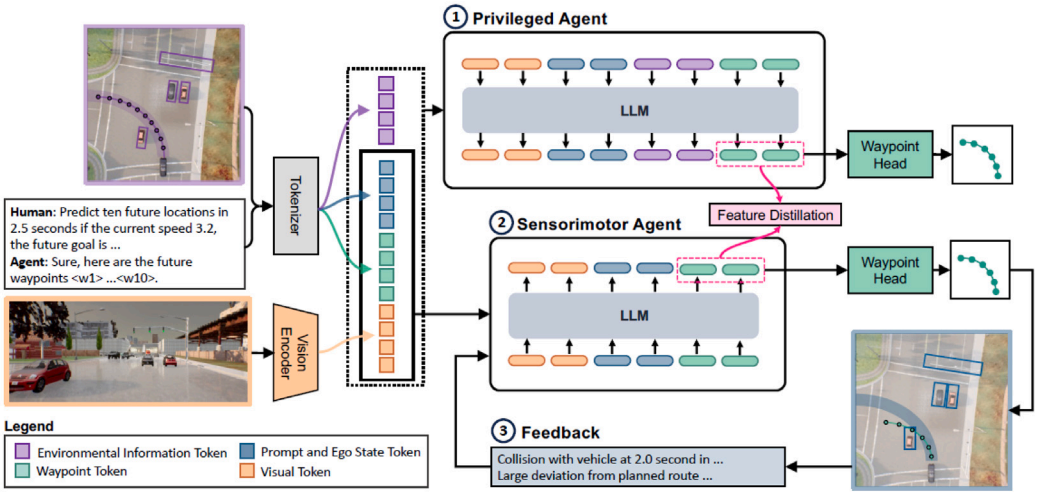

Fig. 17. Architecture of the Feedback-Guided-AD.

pairs. Finally, Table 11 presents a comparative study of the discussed works in Section 9.

8.5. Lessons learned from multimodal large language models in autonomous driving

Across prompt-engineering, fine-tuning, RLHF, and generative pipelines, several patterns emerge. Prompt-driven multimodal models are fast to prototype and excel at producing explanations and high-level rationales, but they are brittle without explicit spatial grounding, temporal memory, and retrieval—sync errors between cameras, LiDAR, radar, and CAN/IMU quickly degrade tracking and intent inference. Fine-tuned systems close this gap: injecting high-resolution crops or region-level supervision, structuring perception into bird's-eye-view or multi-view query formers, and aligning language outputs with control interfaces significantly improves risk localization, trajectory hints, and scene understanding; however, these gains depend on carefully curated supervision, tight cross-sensor calibration, and task-specific adapters. Closed-loop controllers based on small, instruction-tuned planners paired with frozen visual encoders achieve practical latency on embedded hardware, provided the stack partitions workload (low-latency perception/control on the edge; asynchronous language reasoning for guidance, explanation, and operator interaction). RLHF, scenario generation, and coach-style reflection reduce unsafe behaviors and improve human-like policy shaping, yet raise efficiency and cost concerns. Generative world models and language-conditioned video synthesis are promising for rare-event training and counterfactual testing, but they require safety constraints and verification monitors before influencing real vehicles.

9. Datasets & simulators for autonomous driving systems (RQ5)

9.1. Datasets

ADS research relies on datasets that capture diverse scenes, provide robust annotations, and ensure reliable cross-sensor synchronization. While recent corpora target language-vision tasks, *practical deployment requires* not only coverage of long-tail events but also tight multi-modal sync (e.g., camera-LiDAR-radar-CAN/IMU), stable licensing for model/benchmark reuse, and evaluation splits that reflect closed- *and* open-loop settings. Below we briefly profile four representative, *recent* FM-oriented datasets and critically assess their gaps with respect to deployment constraints (latency, safety validation, regulatory acceptance).

NuInstruct (Ding et al., 2024): A language-driven, multi-view video QA corpus (91k pairs across 17 subtasks) structured along perception →prediction→risk→planning. *Limitations:* Potential *label-source bias* from human validation, uneven geographic and diversity, and limited grounding to raw low-level signals (e.g., radar) may weaken safety validation and controller transfer.

DRAMA (Malla et al., 2023): 17,785 short (2 s) interactive scenarios from Tokyo with synchronized camera, CAN, and IMU signals; focused on braking responses. *Limitations:* short horizons constrain long-term reasoning; city/weather bias; camera-only perception (no LiDAR/radar) limits multimodal transfer and robustness analysis for adverse conditions.

NuScenes-MQA (Inoue et al., 2024): QA derived from nuScenes annotations (1.46M items), enabling sentence generation + VQA over a richly annotated multimodal log. *Limitations:* inherits nuScenes class/scene imbalance and may over-emphasize *presence/location* queries versus causal/temporal reasoning; language items are tied to annotation schema, which can cap free-form reasoning coverage.

MaLaMPilot (Ma et al., 2024b): The MaLaMPilot dataset is an open benchmark designed to advance language-conditioned autonomous driving. It integrates synchronized multimodal inputs (RGB, LiDAR, and textual instructions) to support perception, reasoning, and planning tasks. Unlike previous datasets that focus on perception or QA, MaLaMPilot explicitly connects natural language with low-level driving actions, promoting interpretable and instruction-based control evaluation.

Across these datasets, the main gaps are (i) long-tail hazards and rare interventions; (ii) end-to-end *closed-loop* evaluation hooks (simulator-in-the-loop, route-level metrics); (iii) full-stack sync (vision-LiDAR-radar-CAN/IMU-GNSS) for robust sensor fusion; and (iv) governance signals (license, privacy, reusability) (see Table 12).

9.2. Simulators

The complexity and criticality of AD necessitate rigorous evaluation and benchmarking. We review here relevant tools and platforms created for this purpose. First, LimSim++ has been recently proposed as an advanced closed-loop simulation platform for AD (Fu et al., 2024). It includes detailed simulations of traffic flow, traffic control, road infrastructure, and environmental conditions, thus enabling robust evaluation of AD performances. Users can use LimSim++ in different ways: (1) Prompt engineering, e.g., a user can create appropriate scenario descriptions and prompt cues for custom scenarios to facilitate the use of MLLMs for vehicle control, (2) model evaluation, e.g., MLLM for AD performance evaluation, and (3) ADS framework improvement,

Table 11
Comparative study of MLLM approaches for AD.

Year	Ref.	Model	Datasets and tools	Trainable modules	AD services
2023	Xu et al. (2024)	LlaMA-2; LlaVA; Valley	CC3M and WebVid-2M datasets	Video tokenizer (based on Valley); CLIP encoder; Mix-fine-tune module; Text tokenizer/de-tokenizer	Question-answering task; Control
2023	Ding et al. (2023)	BLiP-2; Q-Former; MiniGPT-4	DRAMA dataset; Pytorch libraries	Visual encoder; Query detection; Incorporation module ST-adapter	Risky object detection
2023	W. Wang et al. (2023)	LlaMA-7B; Fine-tuned GD-MAE on ONCE	CARLA-generated dataset; CARLA simulator	Visual encoder: ViT-g/14 from EVA-CLIP; Q-Former; Token embedding	Perception; Reasoning; Question-answering; Planning
2023	Mao et al. (2023b)	Llama-2-7B; GPT-3.5-turbo-1106; GPT-3.5-turbo-0613	NuScenes dataset	Detection; Prediction; Occupancy; Mapping	Task planning; Motion planning; Reasoning; Driving actions
2023	Jia et al. (2023)	Diffusion Models; Vicuna-7B-1.5; LlaVA-7B-1.5	NuScenes and private datasets	ViT; CLIP-ViT-Large; Multi-layer perceptrons	Future scene prediction; Action prediction
2024	Yuan et al. (2024)	Small MLLM-7B	BDD-X and Spoken-SAX datasets	Clip4clip; VIT-B/32	Q&A for planning; Perception Control
2024	W. Han et al. (2024)	LlaMA-2; CLIP	HBD dataset	Perception modules (TrackFormer, MapFormer, MotionFormer, and OccFormer); Planning module	Decision-making; Scene understanding; Motion prediction; Vehicle control
2024	Y. Huang et al. (2024)	Vicuna-7B v1.1	For closed-loop: CARLA simulator; For open-loop: Situated dialogue navigation and BDD-X datasets	Video tokenizer; Text tokenizer; LLM backbone; CLIP encoder; Route planning	Dialogue tasks; Route planning
2024	Liao et al. (2024)	CLIP; BLIP-2; SAM	NuScenes, KITTI, and Waymo datasets	3D network (E3D); Region semantic concordance regularization; Region caption prompts	Perception; Contextual awareness
2024	Liang et al. (2024)	Dense captioning models; OWL-ViT; CLIP; Otter; ChatGPT	LVIS dataset; OWL-v2	Pseudo-labeling models; Fine-tuning models	Perception; Object detection
2024	Ding et al. (2024)	BLIP-2; MiniGPT4; Video Llama; BEV Extractor	NuInstruct dataset	Q-Former; Injection module;	Perception; Risk assessment; Prediction; Planning.
2024	S. Wang et al. (2024)	BLIP-2; LlaVa 1.5; GPT-4V; Lora	NuScenes dataset	Depth-first algorithm; Q-Former 3D MLLM	VQA generation; Perception-action alignment; Decision-making; Planning
2024	Ray and Ohn-Bar (2024)	LLaVA-7B; LlaMA	NuScenes dataset; CARLA simulator	CLIP; Vit; Token prediction; Vision encoder; Language encoder; Waypoint prediction	Future location prediction
2024	Wei et al. (2024a)	7 LLM agents, each for a specific task	Waymo dataset	McNeRF and McLight for background and foreground rendering	Editing 3D driving scenes

through modification of the ADS sub-modules in closed-loop mode. LimSim++ is composed of three main modules as follows: (1) an information integration module that feeds in scenarios provided by Simulation of Urban Mobility (SUMO) and visual contents from the CARLA simulator, (2) an MLLM prompt engine to understand scenarios and tasks, and (3) a continuous learning module, which allows the driver agent make behavioral decisions. The architecture of LimSim++ is presented in Fig. 18. Moreover, ChatSim is an AD scene simulation that has drawn attention due to its significant potential to produce accurate data for driving scenarios (Wei et al., 2024a). It allows the generation and editing of realistic and customized 3D driving scenes using collaborative LLM.

10. Practical foundation models integrations for autonomous driving systems (RQ5)

Despite the theoretical promise of XLMs in enhancing perception, reasoning, and control within ADS, practical integration remains an open challenge due to latency constraints, resource limitations, and a lack of real-time guarantees. This section identifies current implementation patterns and design strategies for deploying XLMs within real-world and simulated ADS stacks, highlighting system architecture, data flow, optimization methods, and open bottlenecks.

10.1. System architecture: Modular cross-language model-based autonomous driving system pipeline

As illustrated in Fig. 19, practical integration of XLMs in ADS adopts a modular design pattern composed of the following layers:

1. **Sensor Acquisition and Synchronization Layer**: Multimodal raw data is collected from heterogeneous sensor suites typically mounted on the vehicle, including RGB cameras (for vision), LiDAR (for depth and structure), radar (for velocity estimation and occlusion handling), GPS/RTK (for localization), IMU (for motion tracking), and microphone arrays (for auditory cues such as sirens). To ensure temporal coherence across modalities, a hardware-synchronized time-stamping mechanism is required.

Table 12
Comparison of recent ADS/XLM datasets. Abbreviations: Perc. = Perception; Pred. = Prediction; Risk = Risk Assessment; Plann. = Planning; Multi-v. = Multi-view; Tempo. = Temporal; Multi-obj. = Multi-objective; Dist. = Distance; Loc. = Location-based; Road = Road semantics; Sync. = cross-sensor synchronization.

Aspect	NuInstruct	DRAMA	NuScenes-MQA	MaLaMPilot
Supported tasks				
Perc.	✓	✓	✓	✓
Pred.	✓	✓	✓	×
Risk	✓	✓	✓	×
Plann.	✓	×	✓	✓
Included information				
Multi-v.	✓	✓	✓	✓
Tempo.	✓	×	✓	✓
Multi-obj.	✓	✓	✓	✓
Dist.	✓	×	✓	×
Loc.	✓	✓	✓	✓
Road	✓	×	✓	×
Size	91k	100k	1.4M	–
Sync.	~camera views	camera+CAN/IMU	Full (nuScenes)	RGB+LiDAR sync
Key limitations (deploym.-relevant)	Label-source bias; limited radar/low-level signals; uneven geo diversity; fewer causality/temporal stress tests.	Short horizon (2s) hinders planning validation; city/weather bias (Tokyo); no LiDAR/radar; limited adverse-condition.	Over-reliance on presence/location QA; inherits nuScenes imbalance; limited free-form reasoning stress tests.	Benchmarks instruction→action grounding for planning; fewer adverse-weather/long-tail stress tests.

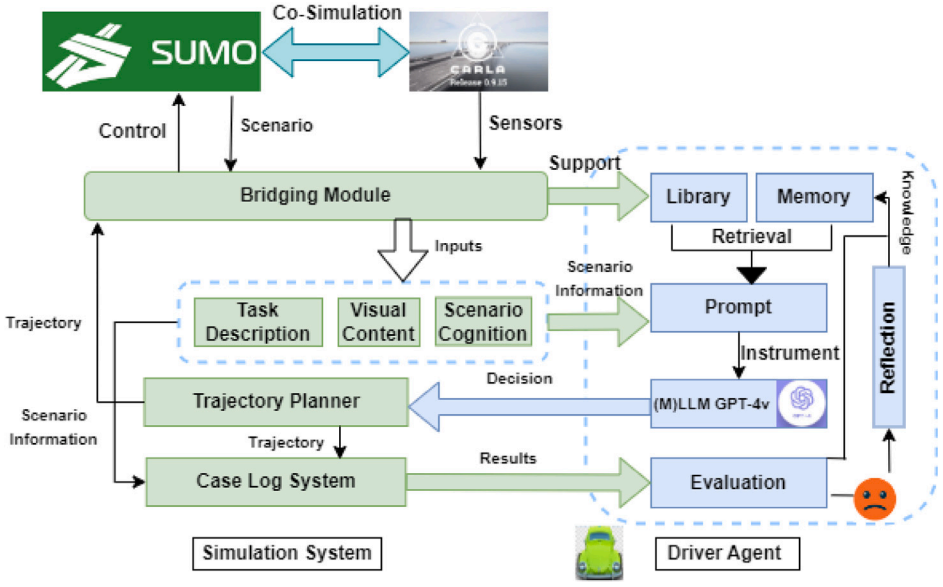

Fig. 18. Architecture of LimSim++.

Sensor fusion is achieved through extended Kalman filters (EKF) or Unscented Kalman Filters (UKF), and for spatial alignment, Iterative Closest Point (ICP) or Normal Distributions Transform (NDT) methods could be used to register LiDAR frames. These steps yield a temporally and spatially unified sensor snapshot at a given time.

2. **Preprocessing and Feature Encoding Layer**: Image frames are resized and normalized before being passed through transformer-based encoders like ViT. These generate spatially-aware feature maps or global embeddings. Point clouds are either voxelized using VoxelNet or projected into bird's-eye-view (BEV) grids using techniques like PointPillars, then encoded via sparse 3D CNNs (e.g., MinkowskiNet). Feature fusion across modalities may occur through early, middle, or late fusion strategies depending on model design. Regarding natural language commands, scenario descriptions, or route instructions, are tokenized using subword encoders such as Byte-Pair Encoding (BPE) or SentencePiece. Positional encodings are added before feeding the token sequences into transformer-based LLMs (e.g., BERT, T5, or GPT-3), where self-attention layers model inter-token dependencies. If speech is the input, it is first converted to text via an ASR model (e.g., Whisper or wav2vec2) before tokenization.
3. **Core XLM Inference Engine Layer**: The core inference engine leverages a range of XLM architectures tailored to different input modalities and operational roles within the ADS stack. LLMs, such as GPT, T5, and LLaMA, are used to process tokenized natural language instructions, contextual traffic rules, or high-level mission plans. Depending on the architecture, decoder-only models (e.g., GPT) are optimized for generative tasks like instruction synthesis, while encoder–decoder models (e.g., T5) are more suited for sequence-to-sequence applications such as route summarization and plan translation. VLMs operate on visual embeddings paired with linguistic prompts, enabling cross-modal attention mechanisms to align spatial perception data with textual queries. These models support scene understanding, image

captioning, and interactive visual reasoning, which are critical for perception-driven planning and interoperability. Alternatively, MLLMs extend this capability by fusing both textual and visual tokens within a shared transformer backbone. These architectures include modality-specific encoders and shared attention layers, facilitating tightly coupled reasoning over multimodal inputs.

4. **Planning and Decision Layer**: The output from the XLM (e.g., intent labels, action probabilities, or natural language plans) is transformed into intermediate decision representations. For example, a decoded output "Turn left in 30 m" may be converted into a cost-based waypoint selection using Frenet coordinates. Trajectory planners then optimize motion trajectories considering dynamic constraints, lane geometry, and traffic predictions, while safety constraints are imposed via rule-checkers or Control Barrier Functions (CBF).
5. **Control Interface Layer** It executes commands by translating planned trajectories into low-level actuation. Controllers include: (1) Proportional–Integral–Derivative (PID)/Stanley controllers for lane-following; (2) Pure Pursuit or Model Predictive Control (MPC) for trajectory tracking; and (3) Learned control policies from deep RL or imitation learning. Control signals (e.g., steering angle θ, throttle τ, and brake b) are transmitted via intra-vehicular communication protocol, such as CAN bus, to the vehicular actuation system, ensuring responsiveness to XLM-guided high-level decisions.

Role of XLMs in the pipeline: As shown in Fig. 19, foundation models can be deployed in complementary roles across the ADS stack: (i) *VLMs* transform sensory observations into semantically rich visual embeddings for perception and situation understanding; (ii) *LLMs* convert user/system prompts into tokenized instructions, infer driving intents, and provide high-level reasoning over contextual cues; and (iii) *MLLMs* integrate both modalities to perform grounded multimodal inference, producing structured outputs (e.g., intentions, waypoints) that feed the planning and decision layer. This modular view reflects common industrial practice where high-level reasoning is supported by foundation models while safety-critical actuation remains within verified control modules.

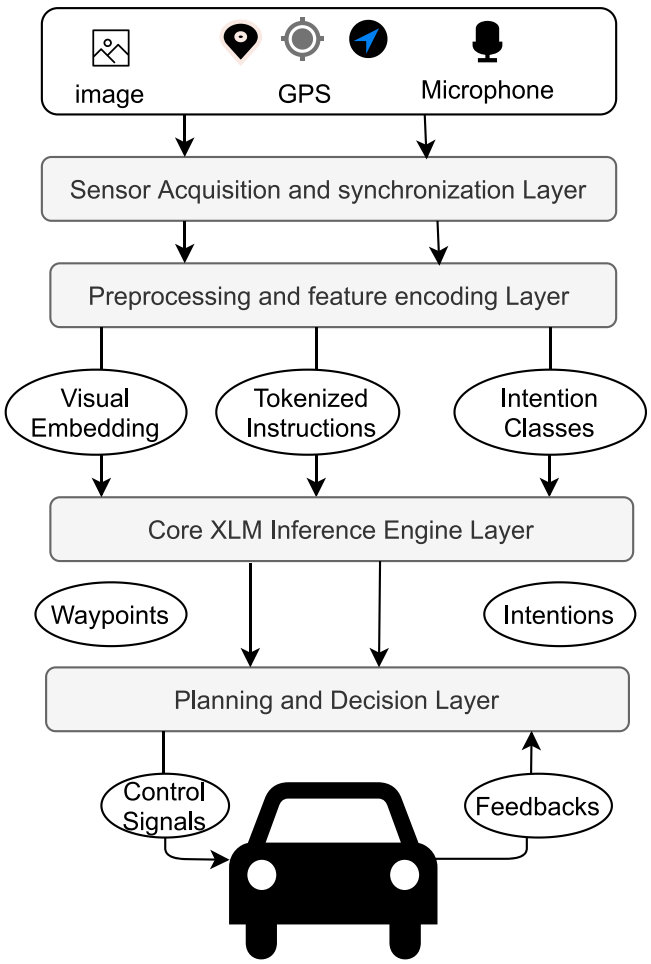

Fig. 19. End-to-end interaction of foundation models in an ADS pipeline. Sensor streams (e.g., camera, GPS, microphone/V2X voice commands) are synchronized and encoded into multimodal representations. VLMs primarily support perception and scene understanding by producing visual embeddings; LLMs handle instruction understanding, reasoning, and intent generation from tokenized prompts; and MLLMs unify vision-language inference to jointly ground instructions in the observed scene and output structured driving decisions (e.g., intentions, waypoints) for the planning and decision layer. The pipeline closes the loop through feedback signals from the vehicle/environment.

10.2. Technical aspects of cross-language model integration in autonomous driving systems

The deployment of XLMs within ADS introduces a unique set of technical design considerations spanning hardware, software, scheduling, memory, and security.

10.2.1. Edge inference hardware design

Deploying large-scale XLMs such as GPT, BLIP-2, or PaLM-E on embedded platforms remains constrained by memory and latency limitations. Onboard inference is often infeasible for large models exceeding billions of parameters. For instance, PaLM-E (with 562 billion parameters) requires over 100 GB of VRAM for full-precision deployment, which exceeds the available memory of most edge computing platforms in autonomous vehicles. These memory demands are further exacerbated by the simultaneous encoding of high-dimensional sensory inputs, such as multi-camera video streams and LiDAR point clouds, which can collectively produce tens of gigabytes of data per second.

To tackle this issue, several strategies are adopted. For instance, quantization, e.g., int4, int8 (Frantar et al., 2022), can reduce memory usage by 4× to 8× with minimal impact on accuracy when applied selectively to non-critical layers. Also, fine-tuning techniques such as Low-Rank Adaptation (LoRA) and quantized LoRA (QLoRA) can tune large models through lightweight adapter layers while keeping the main weights frozen, thus reducing the training/inference footprint (Dettmers et al., 2023). Peak memory usage can be lowered with activation checkpointing and memory paging at intermediate tensors during backpropagation. In addition, FlashAttention and fused-kernel libraries help reduce memory consumption during attention computation (H. Xu et al., 2023). These processes can be supported by high-throughput edge accelerators like NVIDIA Jetson AGX Orin and AMD Versal AI (NVIDIA, 2022; Xilinx, 2023).

Finally, hybrid cloud–edge deployments can host core reasoning modules in powerful remote data centers with asynchronous control loops, while mission-critical inference, e.g., emergency braking or lane switching, remains within the vehicle on dedicated computing units.

10.2.2. Edge inference software design

Real-time XLM integration within ADS stacks demands tight coupling between perception, reasoning, and control tasks. ROS 2 and Apollo Cyber RT frameworks offer suitable middleware for distributed scheduling, but are not natively optimized for XLM large-model inference. Hence, custom microservices must manage token streaming, context window shifting, and asynchronous prompt completion using REpresentational State Transfer (RESTful) application architecture.

In parallel, Tokenizer and embedding pipelines should be cached and executed using vectorized tensor engines (e.g., ONNX Runtime with TensorRT) to minimize decoding overhead. In addition, techniques such as attention pruning, early exit strategies, and input token reduction via vision-language summarization can significantly reduce the number of inference cycles (Zhou et al., 2022; X. Wang et al., 2022). Finally, safety-critical services must include deterministic fallbacks for prompt failures or output uncertainty.

10.2.3. Timing guarantee and scheduling design

ADS systems operate under stringent latency requirements, e.g., control loops run at 20–100 ms per cycle. Such a target cannot be achieved by current XLMs, as they require 300–600 ms for each operation. Hence, techniques like lazy decoding, selective prompting, and frozen backbone caching can be employed to reduce delays. Also, schedulers must support task prioritization between perception updates and language reasoning modules. To reduce the computational overhead, event-triggered inference is recommended, for instance, reasoning is triggered on red lights or pedestrian crossings only.

10.2.4. Security and robustness design

Given that XLMs are susceptible to adversarial prompting and hallucination, additional safety measures are needed. For instance, input filtering using formal grammars or scene validation ensures that only well-formed and contextually relevant inputs are processed, while constraint decoding techniques, such as rule-based output post-processing, enforce adherence to expected output formats or boundaries (P. Liu et al., 2023). Safety monitors can also be employed to implement reject options when outputs fall below a confidence threshold or exhibit uncertainty. To support human-in-the-loop (HiTL) supervision, critical for debugging, trust, and liability, explainability mechanisms such as attention heatmaps, textual rationales, and symbolic traces can be integrated

To provide deeper insight into the practical applicability of XLMs within ADS, we conduct a comparative analysis focused on existing models with visual and multimodal capabilities, as these are critical for perception and semantic reasoning. Table 13 presents a comparative analysis of representative VLMs evaluated across consistent performance and deployment criteria, including supported modalities, task relevance, inference latency, model size, and VQA-v2[1] Accuracy as a unified performance benchmark.

According to Table 13, PaLM-E, with its extensive multimodal capabilities and 562 billion parameters, demonstrates the highest relevance for ADS tasks involving long-horizon planning and complex goal abstraction, and the highest accuracy of 75%, although with an inference latency exceeding 1500 ms, making it unattractive for real-time deployment (Driess et al., 2023a). In contrast, GIT offers the lowest latency (350 ms) and a compact architecture (345M parameters only), which is suitable for time-sensitive tasks such as traffic sign detection, even though its inference accuracy on the VQA-v2 is about 13% lower than that of PaLM-E (W. Wang et al., 2022). Models like BLIP-2 (J. Li et al., 2023), MiniGPT-4 (Zhu et al., 2023), and BEiT-3 (Bao et al., 2021) achieve a trade-off between semantic grounding and reasoning capabilities, with moderate latency, inference accuracy, and parameter size. These results reinforce the need to align model selection with deployment constraints and target tasks within the ADS stack.

[1] VQA v2.0 is a dataset with open-ended questions about images. Answering the questions requires an understanding of vision, language, and common-sense knowledge.

10.3. Task-model fit and deployment guidance

Across recent works, three systematic gaps hinder comparisons: (i) missing or inconsistent information regarding latency/jitter on representative hardware, especially edge and automotive System on a Chip (SoC), (ii) limited closed-loop evaluations beyond CARLA, and (iii) sparse reporting of memory footprints under realistic image or video resolutions.

10.3.1. Perception task

For real-time tracking, segmentation, or detection that feeds a traditional planner, we recommend compact vision backbones such as distilled Transformers with INT8/FP16 support on edge hardware. Aim for median (p_{50}) per-frame latency of $\leq$30–50 ms with minimal jitter to preserve stable controller update rates. In this case, we can invoke VLM selectively (event-triggered or at low sampling rates) for explanations or cross-checks, rather than on every frame, so that the total time remains tightly bounded.

10.3.2. Low-level control and near-field avoidance

For throttle/steer/brake actuation and near-field collision avoidance, employ tiny planners or distilled MLLMs paired with fixed visual encoders; quantize weights/activations and constrain the context window to upper-bound the KV cache footprint. Keep the control loop entirely on-device, offloading large-language reasoning to a side channel or higher-level thread. Report task-level outcomes such as Driving Score (DS), Route Completion (RC), and Infraction Score (IS), together with p_{50}/p_{95} end-to-end latency and jitter, as unstable timing often correlates with degraded DS in practice.

10.3.3. Longer-horizon reasoning and prediction (hundreds of ms acceptable)

For route-level decision-making, intent prediction, situational explanation, and operator interaction, VLMs/MLLMs are appropriate but should execute asynchronously alongside the hard real-time loop under budget-aware scheduling. Employ speculative decoding, vocabulary pruning, and cache reuse, and batch non-urgent queries to smooth computational load. Assess both task accuracy and operator utility, including explanation quality, calibrated uncertainty, and abstention rates, to ensure that slower yet richer reasoning delivers measurable value without destabilizing the control stack.

10.3.4. Human–machine interaction and cooperative driving messages

Favor instruction-tuned, safety-aligned LLMs/VLMs, and implement explicit abstention mechanisms with verification hooks for any safety–critical utterance. Maintain a strict logical separation of HMI and V2X messaging from the hard real-time control loop, and gate any action-triggering communication through verified contracts or policy guards to ensure correctness and traceability.

10.3.5. Deployment guidance

Quantize early, distill when feasible, and profile on the target hardware using realistic image resolutions and token budgets; always report p_{50}/p_{95} latency and jitter. Favor event or uncertainty-triggered MLLM/VLM invocations over per-frame calls to keep compute bounded. Keep perception and control pipelines deterministic and lean, delegating complex reasoning to non-blocking background threads. Finally, pair task accuracy with resource footprint (GPU utilization, memory, power) and timing characteristics (latency/jitter) to provide a truthful picture of deployability.

Table 13
Performance comparison of representative cross-language models (LLMs, VLMs, MLLMs) in autonomous-driving-relevant tasks.

Model	Family/Modality	Primary task(s)	Benchmark/Setting	Device/Precision	Latency (ms)	Params	Key metric(s)/Notes
Closed-loop (control-in-the-loop)							
DriveGPT4-V2	**MLLM**	End-to-end control; short-horizon planning	CARLA Longest6	A800/FP16	*as reported*	0.5–1.5B	DS: 70, RC: 91, IS: 0.77
DriveVLM-Dual	**VLM**	Perception→ ‘ reasoning + assistive control	Real-vehicle tests	Dual Orin X/mixed	~410	–	On-road qualitative + internal eval
LMDrive	**MLLM**	Multimodal fusion → text policy → low-level actuation	CARLA	–/–	–	7B	Reports DS/RC/IS in paper
Open-loop (offline perception/reasoning)							
BLIP-2	**VLM** (Image+Text)	Scene captioning; visual reasoning	VQA-v2, COCO	A100/V100/FP16	~400	~3B + LM	VQA-v2: **65.2%**
MiniGPT-4	**MLLM** (Image+Text)	Dialogue + reasoning over images	VQA-v2, custom	A100/V100/ FP16	~500	13B	VQA-v2: **59.6%**
GIT	**VLM**	Image→text (sign/scene cues)	VQA-v2, COCO	A100/V100/ FP16	~350	345M	VQA-v2: **61.2%**
PaLM-E	**MLLM**	Long-horizon abstraction/task planning	VQA-v2, robotics tasks	TPU/GPU/BF16	>1500	562B	VQA-v2: **75.0%**
KOSMOS-2	**MLLM**	Grounded language-vision tasks	VQA/grounding	A100/V100/FP16	~800	1.6B	*as reported*
BEiT-3	**MLLM**	Perception + captioning	VQA, COCO, ImageNet	A100/V100/FP16/ INT8	~600	~2B	*as reported*

Notes. *Closed-loop* rows involve control-in-the-loop evaluations; *open-loop* rows are offline perception/reasoning. **DS** = Driving Score, **RC** = Route Completion (%), **IS** = Infraction Score (0–1, higher is better). Latency is end-to-end model inference (p50) on the stated *Device/Precision* unless noted. “*as reported*” indicates values copied verbatim from the cited source to avoid inconsistencies across setups.

10.4. Comparative performance analysis of current cross-language models for autonomous driving system tasks

To provide deeper insight into the practical applicability of XLMs within ADS, we conduct a comparative analysis focused on existing models with visual and multimodal capabilities, as these are critical for perception and semantic reasoning. Beyond offline accuracy alone, we report deployment-aware fields, *loop type* (open vs. closed), *device/precision*, *latency p50*, *jitter p95*, *throughput*, *parameter size*, and *memory/KV-cache*, to reflect real-time viability. Table 13 therefore presents a comparative analysis of representative models under consistent reporting categories, including supported modalities, task relevance, timing characteristics, and ADS-relevant metrics (e.g., DS/RC/IS for closed-loop, mAP/mIoU/VQA for open-loop).

10.5. Industry and applied case studies

Recent prototypes illustrate how foundation models are being evaluated beyond benchmarks and into pilot or pre-production contexts.

Wayve (LINGO-2, GAIA-1): Wayve’s LINGO-2 (AI, 2023) demonstrates a closed-loop vision-language-action driving model *tested on public roads*, linking natural-language rationales with vehicle control for explainability and interactive supervision. Complementarily, GAIA-1 (Hu et al., 2023) explores generative world-modeling for driving, indicating how language-conditioned reasoning can sit atop video-centric priors for planning and evaluation.

DriveVLM-Dual: VLM-centric stacks such as DriveVLM-Dual (X. Tian et al., 2024) represent among the first VLM-based autonomous driving frameworks *tested on a real vehicle platform*. The system integrates two NVIDIA Orin X processors operating asynchronously: a high-frequency end-to-end driving controller on Orin X-1 and the DriveVLM perception-reasoning module on Orin X-2. This dual configuration enables parallel visual reasoning and control with low-latency feedback; on-road tests report an average inference time of ~410 ms on Orin X.

DriveGPT4-V2: Concurrently, MLLM-driven controllers such as *DriveGPT4-V2* (Z. Xu et al., 2025) show closed-loop, *end-to-end* control in complex urban settings using tiny-scale planners (e.g., Qwen-0.5B, TinyLLaMA) instruction-tuned for multimodal understanding, while keeping a SigLIP 384 × 384 visual encoder fixed. Training is end-to-end (except for the encoder), and inference employs PID controllers to convert predicted targets into throttle/steer/brake commands, enabling real-time execution.

Tesla FSD v12 (vision-only, end-to-end): Publicly described as a large end-to-end video-model replacing hand-engineered modules, Tesla’s FSD v12 (Tesla, 2024) situates learned planning and control within a production fleet. While details remain proprietary, its vision-only approach underscores both the promise and verification burden of end-to-end policies at scale.

NVIDIA DRIVE Thor (centralized automotive compute): DRIVE Thor (NVIDIA Corporation, 2022) exemplifies the hardware trend toward centralized, mixed-criticality SoCs (cockpit+ADAS+AD) with safety and high AI throughput, practical enablers for co-locating low-latency perception with larger XLM reasoning components under automotive safety constraints.

OpenPilot (open-source ADAS with learned components): Comma.ai’s OpenPilot (comma.ai, 2025) provides an open, widely deployed L2 ADAS stack with end-to-end learning components and community fleet telemetry. Although not a full foundation-model stack, it offers a pragmatic testbed for lightweight policies, rapid iteration, and on-road validation.

Ecosystem toolchains and regulation: Industrial pipelines increasingly leverage NVIDIA DRIVE infrastructure and Omniverse-based simulation (e.g., world foundation models such as Cosmos) to stress-test long-tail scenarios before road trials. While L3 production systems (e.g., Mercedes-Benz DRIVE PILOT) evidence regulatory progress, *most foundation-model use in ADS remains pilot-scale*, with emerging best practices emphasizing hybrid architectures (classical perception +

Table 14
Industrial and applied deployments of XLM-based autonomous driving systems.

System	Model type	Execution platform	Deployment level	Key contribution
Wayve (LINGO-2, GAIA-1)	VLM/World Model	GPU-based onboard compute	Public-road testing	Language-conditioned driving and explainability
DriveVLM-Dual	VLM (dual Orin)	NVIDIA Orin X (edge)	Real-vehicle prototype	Low-latency multimodal reasoning
DriveGPT4-V2	MLLM + PID control	Embedded GPU	Urban driving trials	End-to-end reasoning and control
Tesla FSD v12	Vision-only FM	Custom ASIC (HW4)	Large-scale fleet	End-to-end learned driving policy
OpenPilot	Lightweight DNNs	Commodity hardware	Large user fleet	Open ADAS experimentation

XLM reasoning), rigorous simulator-in-the-loop evaluation, and staged on-road deployment.

Industrial Readiness and Deployment Maturity: Beyond algorithmic performance, the transition of foundation models from research prototypes to deployable autonomous driving systems critically depends on system-level constraints such as latency, compute budgets, safety certification, and integration with legacy automotive stacks. Recent industrial efforts increasingly emphasize *hybrid architectures*, where foundation models augment, but do not fully replace, classical perception and control pipelines.

To better contextualize industrial maturity, Table 14 summarizes representative real-world deployments and pilot systems, highlighting their execution environments (simulation vs. on-road), hardware constraints, model scale, and deployment readiness. Notably, systems such as DriveVLM-Dual and DriveGPT4-V2 demonstrate that multimodal reasoning can be executed under real-time constraints using edge-grade hardware (e.g., NVIDIA Orin), while Tesla FSD v12 exemplifies large-scale fleet learning under tightly controlled proprietary settings.

These observations indicate that current industrial practice favors hybrid and modular architectures, where FMs provide high-level semantic reasoning, explanation, and planning. At the same time, safety-critical control loops remain tightly optimized and verifiable. This trend suggests that near-term deployment of FMs in autonomous driving will continue to emphasize co-design between learning-based reasoning and classical control, supported by extensive simulation and staged real-world validation.

11. Open issues and future directions (RQ6)

Despite the great progress and advancements in XLMs for ADS, their deployment in real-world systems is still facing critical challenges.

11.1. New datasets for cross-language model-assisted autonomous driving systems

The XLMs that will be used to enable ADS must synthesize and interpret inputs from multiple modalities, including 3D point clouds, panoramic images, and HD map annotations. Current datasets are limited in scale, quality, and diversity to achieve precise ADS functionalities. Most multimodal LLMs, such as GPT-4V, have been pre-trained on open-source datasets that include driving and traffic scenes. However, the vision-language datasets derived from NuScenes, for instance, do not provide a sufficiently robust benchmark for visual-language understanding in AD contexts. Hence, there is an urgent need to create extensive, diverse, and scalable datasets covering any traffic or driving situation, particularly critical and rare events. In addition, such new datasets should be accompanied by high-quality annotations, e.g., object labels and semantic information, that help in understanding complex scenes and increase the training and evaluation precision. These enhanced datasets would be critical for thoroughly testing and improving the performance of MLLMs in AD applications. Indeed, it is essential to consider several factors for the development of these new datasets:

- **Diversity of AD Scenarios**: Datasets should include a wide variety of balanced driving conditions, including different weather, lighting, and traffic patterns, to ensure models can generalize across all possible real-world situations. This includes various environments such as urban, suburban, and rural settings, as well as different weather conditions (rain, snow, fog) and times of day (day, night, dawn, dusk). The inclusion of a wide range of traffic situations, such as heavy congestion, light traffic, and unusual events like accidents or road closures, is crucial for evaluating the model's reliability across diverse contexts
- **Annotation Quality:** High-quality, detailed annotations are crucial for training and evaluating models effectively. This includes not only object labels but also semantic information that can help in complex scenes understanding, accurate map data with annotations for road features, traffic signals, and lane markings.
- **Integration of Multimodal Data:** Effective integration of data from different modalities (e.g., images, point clouds, maps) is necessary for creating comprehensive datasets that reflect the true complexity of driving environments. In addition, developing adaptive fusion mechanisms that dynamically adjust the weighting of different modalities based on the context and current driving conditions, and prioritize the most relevant data, to improve the overall performance and safety still an open issue requiring more investigation. In addition, future research could aim to enhance the integration of various modalities, facilitating seamless information exchange and collaborative reasoning across different sensors. Additionally, utilizing knowledge distillation from robust foundation models trained on extensive datasets presents a promising avenue.
- **Edge Cases and Rare Events:** New datasets should be given special attention to capture critical events, such as sudden pedestrian crossings or unexpected obstacles, or sudden braking of a vehicle in front of the ego vehicle, which are crucial for the safe operation of AVs. Therefore, research in data generation and in artificial intelligence generative computing algorithms could modify and augment existing datasets. This enables efficient data expansion and customization, tailoring training sets to specific driving scenarios, especially critical scenarios.
- **Scalability and Accessibility:** The datasets should be large enough and balanced to train advanced models and accessible to researchers and developers to foster collaboration and innovation in the field.

Addressing these considerations will help in building robust datasets that can advance the development and deployment of XLLMs in ADS, contributing to safer and reliable AV.

11.2. Mitigating hallucination in foundation models

The hallucination of XLMs refers to the phenomenon where the outputs (e.g., generated text response) are inconsistent with the corresponding visual content, which might be critical in the AD context.

Recently, new approaches have been proposed to address hallucination from data, model, training, and inference (Z. Bai et al., 2024). Nevertheless, more in-depth work is needed to mitigate the hallucination effect. Improvement directions would encompass diversifying the training datasets and enhancing their quality, developing systems that explicitly enforce consistency between modalities during training and inference, proposing novel XLM models that natively reduce the hallucination events, e.g., using RLHF, and finally designing simulators and/or testbeds for hallucination assessment. In the following, some directions that could contribute to mitigating hallucination:

- **Enhancing Dataset Quality and Diversity:** Introducing counterfactual data, reducing noise and errors in the existing database, and guaranteeing data diversity are good directions to overcome hallucination from data.
- **Cross-Modal Consistency Networks:** Develop networks that explicitly enforce consistency between modalities during both training and inference phases. This reduces the likelihood of hallucinations and ensures a more coherent understanding of the driving environment. Ensuring that generated content remains consistent and contextually relevant to the input modality requires sophisticated techniques for capturing and modeling cross-modal relationships.
- **Design of New Models and Frameworks:** To deal with hallucination from the model, new advanced architectures should be designed, and additional learning objectives should be considered. RLHF could improve safety and reduce instances of hallucinations in model predictions. However, MLLMs and LVM-based frameworks that go under the RLHF category are still limited. Furthermore, designing mechanisms for the model to learn from its mistakes and improve over time should be integrated into the frameworks.
- **Design of Tesbeds for Hallucination Assessments:** The evaluation of hallucinations in MLLMs within the context of AD necessitates the creation of specialized testbeds. These testbeds must be meticulously designed to identify and analyze instances where the models generate inaccurate or misleading outputs. To accurately assess hallucinations, testbeds should encompass a broad spectrum of realistic driving scenarios, integrating data from multiple sources to mimic the complex sensory inputs received by AVs. Besides that, Ground Truth Annotations are very important. Therefore, testbeds should include thoroughly annotated datasets that serve as ground truth references. These annotations must be precise and cover all relevant aspects of the driving environment, including object locations, movements, and interactions. This ground truth data is vital for comparing the model's predictions and identifying hallucinations. In addition, a significant portion of the testbed should be dedicated to edge cases and anomalies. Critical scenarios, such as sudden pedestrian crossings, unexpected obstacles, or erratic behavior from other vehicles, are essential for testing the model's ability to handle unpredictable events without generating hallucinations. Furthermore, to assess hallucinations, a combination of quantitative and qualitative metrics should be employed, and dedicated metrics should be defined such as (i) accuracy metrics: Measuring the percentage of correct predictions versus hallucinations, (ii) False Positive Rates: Identifying instances where the model falsely identifies objects or events, and (iii) Human-in-the-Loop Evaluations: Involving human evaluators to qualitatively assess the plausibility of the model's outputs.

11.3. Enabling foundation models on resource-limited hardware

Given the complexity of XLM methods, it is difficult to deploy them on capacity-limited hardware.

To bypass this issue, several strategies can be developed. For instance, FL techniques can be leveraged to train models across distributed data sources, e.g., AVs, without requiring centralized data, processing, and storage. Until today, limited efforts have been conducted to bring LlaMa-7B to edge devices such as computer systems and smartphones (Xu et al., 2024a). Hence, further research is required to support XLM in edge computing systems. Moreover, the design of scalable and energy-efficient XLM architectures that run on low-capacity hardware without significantly sacrificing performance would be a key enabler of cost-effective XLM-assisted ADS. Such architectures would potentially consider latency and memory optimization, model compression, and knowledge distillation techniques.

11.4. Advancing personalized autonomous driving systems

Integrating XLMs into ADS marks a paradigm characterized by continuous learning and personalized engagement. Indeed, XLMs can continuously learn from new data and interactions, thus adapting to changing driving patterns, user preferences, and evolving road conditions. This adaptability results in a refined and enhanced performance over time.

However, real-time personalization in ADS is lacking, which opens numerous opportunities to deploy and validate XLM-assisted personalized AD frameworks. Moreover, the development of XLM-driven virtual assistants that align with drivers' individual preferences, with safety features, like fatigue detection, and maintenance specifications, can be explored.

11.5. Integrating sparse and dense modalities for multimodal learning

Despite recent progress, fusing heterogeneous modalities such as LiDAR and images remains a challenge due to differences in resolution, noise characteristics, and data structure (e.g., point clouds vs. dense grids). Moreover, integrating these modalities with LLMs or MLLMs for reasoning remains an open research direction. Some early-stage work explored incorporating BEV or point cloud features into XLM pipelines via prompt engineering (Fourati et al., 2025) or adapter modules. However, such work is still in its infancy.

11.6. Multimodal retrieval-augmented generation frameworks

Although retrieval-augmented generation (RAG) techniques, which incorporate relevant external knowledge during text generation resulting in more accurate and contextually relevant outputs, are well addressed in LLMs, Multimodal RAG (MuRAG) is under-explored, especially in ADS (Dai et al., 2024; Yuan et al., 2024). Indeed, the latter is expected to combine information retrieval with multimodal data processing and generative capabilities to enhance the AV's understanding and response to complex driving scenarios.

11.7. Interplay between multi-tasking and fine-tuning

When an XLM in ADS multi-tasks, e.g., used for scene understanding and trajectory prediction, it is hard to fine-tune it efficiently. Novel approaches are needed to coordinate task-specific adjustments while maintaining shared representations among tasks. Also, researchers should explore collaborative sensory modalities in XLM development to obtain a holistic understanding of the driving environment, regarding multi-tasking.

11.8. Foundation models security, and ethical considerations

11.8.1. Security issues

As XLMs continue to evolve, several critical risks have been identified, including prompt injection, data poisoning, model inversion, and catastrophic forgetting. These vulnerabilities pose substantial barriers to the safe and large-scale deployment of multimodal models in real-world ADS. To mitigate such threats, it is essential to integrate **robust security mechanisms** throughout the development pipeline.

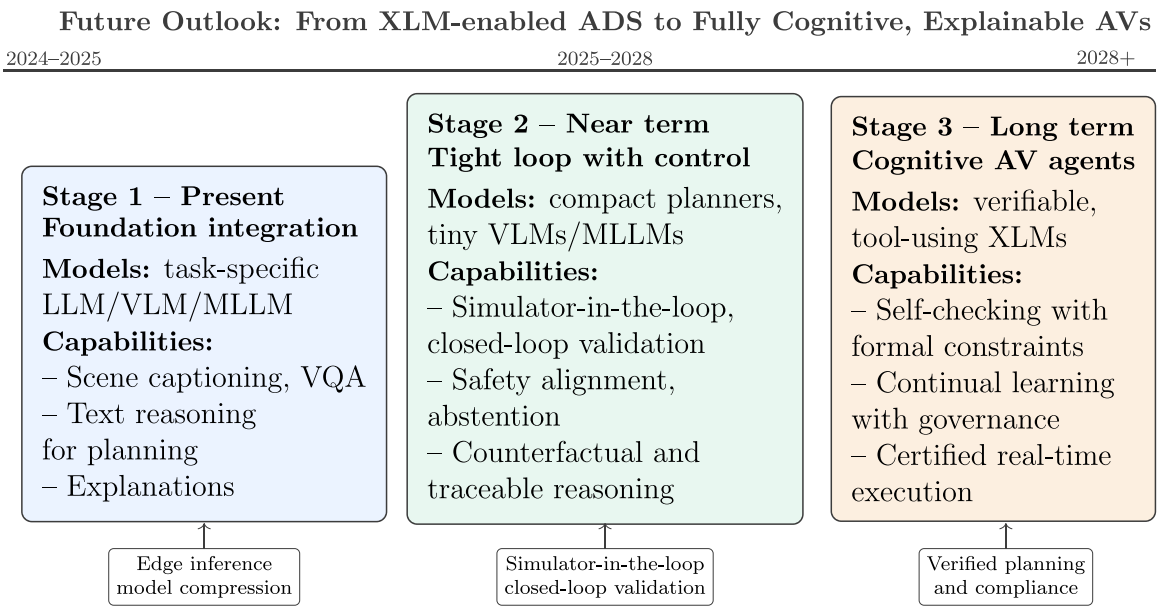

Fig. 20. Future outlook for XLM-assisted ADS. The timeline illustrates the evolution from current foundation-model integration (Stage 1) to near-term closed-loop control (Stage 2), and ultimately to long-term cognitive and explainable autonomous vehicles (Stage 3).

Promising research directions include the adoption of frameworks such as *MLLM-Protector*, adversarially robust training, fine-grained access control for sensor-language interfaces, and data encryption techniques to safeguard both training and inference stages. Furthermore, the incorporation of anomaly detection and red-teaming evaluation can help ensure continuous monitoring and reliability of deployed systems.

11.8.2. Ethical and societal considerations

The integration of large foundation models into safety-critical systems such as AD introduces complex ethical and societal challenges that extend beyond technical performance. A central concern lies in the *opacity and interpretability* of LLMs, VLMs, and MLLMs. These models operate as high-dimensional black boxes whose internal reasoning processes remain largely inaccessible, making it difficult to trace or justify specific driving actions. In critical events such as collisions or near-misses, this lack of transparency complicates *accountability*, accident attribution, and legal liability. Explainability frameworks and post-hoc reasoning tools are therefore essential to ensure that human operators, regulators, and insurers can understand and audit model decisions.

Another major issue is algorithmic bias and fairness. Indeed, foundation models are trained on large-scale, web-scraped, or sensor-collected datasets that may reflect imbalanced demographic, geographic, or environmental representations. When transferred to ADS, such biases can manifest as unequal recognition accuracy across weather conditions, regions, or vulnerable road users (e.g., pedestrians, cyclists, wheelchair users). Ethically aligned dataset curation, domain balancing, and bias auditing pipelines are necessary to prevent discrimination and ensure equitable safety for all road participants.

Value alignment and moral reasoning also remain open research problems. While MLLMs can reason contextually using natural-language prompts, they lack embedded moral or normative grounding. For example, a model might prioritize collision avoidance in one scenario but fail to reason about ethical trade-offs involving multiple agents or unexpected human behaviors. Embedding normative reasoning through constrained optimization, human feedback (RLHF), or hybrid rule-based safety governors represents a key step toward ethically aware ADS behavior.

Finally, privacy and data governance are critical ethical dimensions. Training multimodal foundation models for ADS requires collecting vast amounts of real-world data from cameras, microphones, LiDAR, and GPS sensors, often capturing personally identifiable information or sensitive contexts (faces, license plates, geolocation traces). Robust anonymization, federated learning, and on-device training can mitigate such risks, ensuring that ethical compliance and data protection are integral to the development pipeline. Collectively, addressing these ethical considerations is essential to build *trustworthy, transparent, and socially acceptable* autonomous driving systems powered by foundation models.

11.8.3. Reliability and safety assurance challenges

Beyond security and ethics, deploying XLMs in safety–critical ADS raises reliability challenges that can directly impact operational safety. Key issues include (i) *distribution shift* across locations, weather, sensor degradation, and rare events; (ii) *hallucinated or ungrounded outputs* in language-based reasoning that may conflict with physical constraints; and (iii) limited *formal verifiability* of end-to-end learned policies. To improve trustworthiness, recent practice increasingly combines FM reasoning with *safety monitors* and *rule-based/verified governors* (e.g., runtime constraint checking, fallback planners, and emergency braking), alongside systematic *scenario-based validation* in simulation and staged on-road trials. These measures help ensure that high-level model outputs remain consistent with safety envelopes, regulatory expectations, and the strict latency constraints of real-time vehicle control.

11.9. Short- and long-term research priorities

To guide future research and development, it is essential to distinguish between short-term actionable objectives and long-term transformative directions for XLM-assisted Autonomous Driving Systems (ADS). While several technical advances have been achieved in perception, reasoning, and control, the sustainable deployment of XLMs in safety-critical domains requires both incremental optimization and systemic evolution across regulation, verification, and ethical governance.

Short-term priorities (1–3 years): Immediate efforts should focus on bridging the gap between current research prototypes and deployable systems. Key priorities include: (i) establishing unified datasets and benchmarking protocols that accurately reflect multimodal synchronization, real-time constraints, and safety–critical edge cases; (ii) reducing inference latency and computational overhead via model compression, quantization, and distillation for embedded execution; (iii)

Table A.15
Summary of existing autonomous-driving datasets. Modalities: Cam. = camera, Li. = LiDAR, Ra. = radar, GPS/IMU.

Ref.	Name	Modalities	Scale	Highlights	Annotations
Brostow et al. (2008)	CamVid	Cam. (video)	701 images	960 × 720; daytime; urban traffic	Pixel-wise segmentation; 32 classes
Geiger et al. (2012)	KITTI	Cam., Li., GPS/IMU/RTK	15k images	Diverse scenes; daytime driving	80k 3D boxes; 15k 2D boxes
XinyuHuang et al. (2020) and ApolloScape (2018)	ApolloScape	Cam., Li., GPS/IMU	100k frames	Dense urban traffic; varied weather; cm-level accuracy	Complete 3D annotation
Caesar et al. (2020)	nuScenes	360°C am., Li., Ra., GPS/IMU	1k scenes; 23 obj. classes	Rich sensor suite; diverse conditions	Dense multimodal annotations
Mei et al. (2022)	Waymo Open	Cam., Li.	1k seg.; 2860 seq.; 100k images	High-res data; diverse scenarios; 8 tracking, 25 semantic classes	Detailed 2D/3D annotation
Chang et al. (2019)	Argoverse	Cam., Li., GPS/IMU	324 sequences	Motion forecasting; 3D tracking; map priors	3D boxes for multiple objects
Yu et al. (2020)	BDD100K	Cam., GPS	100k video seq.	Diverse weather/time/geography	2D boxes; pixel-level labels (10 classes)
K. Li et al. (2022)	CODA	Cam., Li.	1500 scenes	Corner-case focus; merged from KITTI/nuScenes/ONCE; 34 classes	Manually verified object-level boxes
Alibeigi et al. (2023)	ZOD	Cam., Li., GPS/IMU	100k frames; 1473 seq.; 29 scenes	High-res sensors; broad coverage	2D/3D boxes; instance segmentation; 156 traffic-sign classes
Burnett et al. (2023)	Boreas	Cam., Li., Ra., GPS/IMU	~350 km driving	Multi-modal; cm-level accuracy; clear weather	Ground-truth poses; 3D labels
Mao et al. (2021)	ONCE	Cam., Li.	1M scenes; 144 h; 7M images; 417k 3D boxes	Large-scale; diverse conditions; high temporal resolution	16k scenes with 3D GT; 769k 2D boxes

enhancing interpretability and explainability, especially in perception-decision coupling; (iv) integrating uncertainty estimation, hallucination mitigation, and verification layers to increase reliability; and (v) reinforcing cybersecurity, privacy-preserving learning, and adversarial robustness in connected environments. These directions represent the foundational steps necessary to ensure trustworthy, verifiable, and resource-aware XLM integration within ADS pipelines.

Long-term priorities (3–10 years): Over the next decade, research should converge toward scalable, self-adaptive, and regulation-compliant ADS architectures. This includes (i) defining certification frameworks and safety validation protocols for foundation-model-driven control loops; (ii) achieving transparent human–AI co-driving paradigms with interpretable and auditable decision chains; (iii) developing dynamic retrieval-augmented and continual-learning mechanisms for evolving road contexts; (iv) advancing personalized and context-aware driving agents capable of cross-domain reasoning; and (v) establishing ethical and societal governance models that ensure accountability, fairness, and inclusivity in global deployment. Long-term progress also depends on collaboration between academia, industry, and regulators to co-design validation suites and open simulation environments for testing foundation models under realistic operational conditions.

Overall, these complementary short and long-term priorities delineate a clear roadmap toward safe, efficient, and ethically aligned XLM-based autonomous driving, where foundation models evolve from experimental research tools to certified components of real-world intelligent mobility systems. A forward-looking perspective summarizing this evolution toward fully cognitive and explainable autonomous vehicles is illustrated in Fig. 20.

12. Conclusion

In this survey, we reviewed XLMs and their integration into ADS, analyzing recent works from architectural, dataset, and conceptual perspectives. We examined how XLMs address key ADS challenges such as multi-modal data fusion, safety, reliability, and complex environment understanding, and discussed core techniques including prompt engineering, fine-tuning, RLHF, and GAI. Our synthesis highlights the growing utility of XLMs across ADS tasks like perception, planning & control, multi-tasking, and question answering, while emphasizing the importance of curated datasets and simulation tools. Beyond synthesis, we identified areas where XLMs already deliver practical value. Indeed, VLM/MLLM stacks are maturing for perception-reasoning handoffs, lightweight LLM planners paired with PID/MPC controllers enable closed-loop control, and simulator-in-the-loop evaluation provides a viable safety gate before road testing. We also relate model complexity to deployment constraints, offering guidance on when compact planners (e.g., sub-1B XLM models) or hybrid pipelines are preferable to larger XLMs for real-time control. From an industrial standpoint, our taxonomy and comparison tables serve as a blueprint for mapping XLM capabilities to ADS modules (perception → prediction → planning → control → HMI/V2X), helping practitioners align model selection, datasets, and simulators with certification, verification, and resource budgets. Looking ahead, we argue that progress will hinge on three axes, namely (i) verifiable and uncertainty-aware XLMs (e.g., abstention, safety alignment, formalized monitors), (ii) resource-aware deployment (e.g., compression, distillation, scheduling on automotive SoCs), and (iii) evolution of datasets and benchmarks toward synchronized multimodality and long-tail scenario coverage. Together, these insights offer a practical roadmap for transitioning XLM-enabled ADS from promising prototypes to trustworthy, standards-ready systems.

CRediT authorship contribution statement

Sonda Fourati: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Wael Jaafar:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Methodology, Funding acquisition, Formal analysis, Conceptualization. **Noura Baccar:** Writing – review & editing, Validation, Supervision. **Safwan Alfattani:** Writing – review &

editing, Writing – original draft, Validation, Methodology, Funding acquisition. **Rami Langar:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is funded in part by the Mitacs Globalink Research Award (GRA) program and in part by the Deanship of Scientific Research at King Abdulaziz University.

Appendix. Summary of existing autonomous-driving datasets

See Table A.15.

Data availability

No data was used for the research described in the article.

References

Ahangar, M.N., Ahmed, Q.Z., Khan, F.A., Hafeez, M., 2021. A survey of autonomous vehicles: Enabling communication technologies and challenges. Sensors 21 (3), 706.

AI, W., 2023. Lingo-2: Driving with language. https://wayve.ai/thinking/lingo-2-driving-with-language/. (Accessed 31 October 2025).

Aldeen, M., MohajerAnsari, P., Ma, J., Chowdhury, M., Cheng, L., Pesé, M.D., 2024. WIP: A first look at employing large multimodal models against autonomous vehicle attacks. In: Proc. Symp. Veh. Secu. Priv. (VehicleSec).

Alibeigi, M., Ljungbergh, W., Tonderski, A., Hess, G., Lilja, A., Lindström, C., Motorniuk, D., Fu, J., Widahl, J., Petersson, C., 2023. Zenseact Open Dataset: A large-scale and diverse multimodal dataset for autonomous driving. In: Proc. IEEE/CVF Int. Conf. Comput. Vis.. pp. 20178–20188.

Ananthajothi, K., GS, S.S., Saran, J., 2023. LLM's for autonomous driving: A new way to teach machines to drive. In: Proc. Int. Conf. Mob. Netw. Wireless Commun.. ICMNWC, IEEE, pp. 1–6.

Anderljung, M., Barnhart, J., Leung, J., Korinek, A., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., et al., 2023. Frontier AI regulation: Managing emerging risks to public safety. arXiv preprint arXiv:2307.03718.

ApolloScape, 2018. ApolloScape dataset. URL http://apolloscape.auto/. (Accessed 28 August 2024).

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

Azdam, S., Doma, P., Arab, A.M., 2025. ManeuverGPT agentic control for safe autonomous stunt maneuvers. arXiv preprint arXiv:2503.09035.

Bai, G., Chai, Z., Ling, C., Wang, S., Lu, J., Zhang, N., Shi, T., Yu, Z., Zhu, M., Zhang, Y., et al., 2024. Beyond efficiency: A systematic survey of resource-efficient large language models. arXiv preprint arXiv:2401.00625.

Bai, Y., Geng, X., Mangalam, K., Bar, A., Yuille, A., Darrell, T., Malik, J., Efros, A.A., 2024. Sequential modeling enables scalable learning for large vision models. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.. CVPR, pp. 22861–22872.

Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z., 2024. Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930.

Balasubramani, K., Natarajan, U.M., 2024. Improving bus passenger flow prediction using Bi-LSTM fusion model and SMO algorithm. Babylon. J. Artif. Intell. 2024, 73–82.

Bao, H., Dong, L., Piao, S., Wei, F., 2021. Beit: BERT pre-training of image transformers. arXiv preprint arXiv:2106.08254.

Brostow, G.J., Fauqueur, J., Cipolla, R., 2008. Segmentation and recognition using structure from motion point clouds. In: Proc. Europ. Conf. Comput. Vis.. Springer, pp. 44–57.

Burnett, K., Yoon, D.J., Wu, Y., Li, A.Z., Zhang, H., Lu, S., Qian, J., Tseng, W.-K., Lambert, A., Leung, K.Y., et al., 2023. Boreas: A multi-season autonomous driving dataset. Int. J. Robot. Res. 42 (1–2), 33–42.

Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., 2020. nuScenes: A multimodal dataset for autonomous driving. In: Proc. IEEE/CVF Conf. Comput. Vis. Patt. Recogn.. pp. 11621–11631.

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P.S., Sun, L., 2023. A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT. arXiv preprint arXiv:2303.04226.

Chan, A.J., Sun, H., Holt, S., van der Schaar, M., 2024. Dense reward for free in reinforcement learning from human feedback. arXiv preprint arXiv:2402.00782.

Chandrasiri, M., Talagala, P.D., 2023. Cross-vit: Cross-attention vision transformer for image duplicate detection. In: 2023 8th International Conference on Information Technology Research. ICITR, IEEE, pp. 1–6.

Chang, M.-F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., et al., 2019. Argoverse: 3D tracking and forecasting with rich maps. In: Proc. IEEE/CVF Conf. Comput. Vis. Patt. Recogn.. pp. 8748–8757.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al., 2024. A survey on evaluation of large language models. ACM Trans. Intell. Syst. Technol. 15 (3), 1–45.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J., 2023. Supplementary material uniter: universal image-text representation learning. ReCALL 1 (5), 10.

Chen, L., Sinavski, O., Hünermann, J., Karnsund, A., Willmott, A.J., Birch, D., Maund, D., Shotton, J., 2024a. Driving with LLMs: Fusing Object-Level vector modality for explainable autonomous driving. In: Proc. IEEE Int. Conf. Robot. Automa.. ICRA, pp. 14093–14100. http://dx.doi.org/10.1109/ICRA57147.2024.10611018.

Chen, M., Tao, Z., Tang, W., Qin, T., Yang, R., Zhu, C., 2023. Enhancing emergency Decision-making with knowledge graphs and large language models. arXiv preprint arXiv:2311.08732.

Chen, L., Zhang, Y., Ren, S., Zhao, H., Cai, Z., Wang, Y., Wang, P., Meng, X., Liu, T., Chang, B., 2024b. PCA-Bench: Evaluating multimodal large language models in Perception-Cognition-Action chain. arXiv preprint arXiv:2402.15527.

Chitta, K., Prakash, A., Jaeger, B., Yu, Z., Renz, K., Geiger, A., 2022. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. IEEE Trans. Pattern Anal. Mach. Intell. 45 (11), 12878–12895.

Cho, J.H., Ivanovic, B., Cao, Y., Schmerling, E., Wang, Y., Weng, X., Li, B., You, Y., Krähenbühl, P., Wang, Y., et al., 2024. Language-image models with 3D understanding. arXiv preprint arXiv:2405.03685.

Choi, H.-S., Jeong, J., Cho, Y.H., Yoon, K.-J., Kim, J.-H., 2023. Semantics-guided Transformer-based sensor fusion for improved waypoint prediction. arXiv preprint arXiv:2308.02126.

comma.ai, I., 2025. OpenPilot – An open-source ADAS driving software system. https://comma.ai/openpilot. (Accessed 05 November 2025).

Cui, Y., Lin, H., Yang, S., Wang, Y., Huang, Y., Chen, H., 2025. Chain-of-Thought for autonomous driving: A comprehensive survey and future prospects. arXiv preprint arXiv:2505.20223.

Cui, C., Ma, Y., Cao, X., Ye, W., Wang, Z., 2023a. Human-autonomy teaming on autonomous vehicles with large language Model-Enabled human digital twins. In: Proc. IEEE/ACM Symp. Edge Comput.. SEC, IEEE, pp. 319–324.

Cui, C., Ma, Y., Cao, X., Ye, W., Wang, Z., 2024a. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.. pp. 902–909.

Cui, C., Ma, Y., Cao, X., Ye, W., Wang, Z., 2024b. Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles. IEEE Intell. Transp. Syst. Mag..

Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.-D., et al., 2024c. A survey on multimodal large language models for autonomous driving. In: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.. pp. 958–979.

Cui, C., Ma, Y., Yang, Z., Zhou, Y., Liu, P., Lu, J., Li, L., Chen, Y., Panchal, J.H., Abdelraouf, A., et al., 2024d. Large language models for autonomous driving (llm4ad): Concept, benchmark, experiments, and challenges. arXiv preprint arXiv:2410.15281.

Cui, C., Yang, Z., Zhou, Y., Ma, Y., Lu, J., Wang, Z., 2023b. Large language models for autonomous driving: Real-world experiments. arXiv preprint arXiv:2312.09397.

Dai, X., Guo, C., Tang, Y., Li, H., Wang, Y., Huang, J., Tian, Y., Xia, X., Lv, Y., Wang, F.-Y., 2024. VistaRAG: Toward safe and trustworthy autonomous driving through retrieval-augmented generation. IEEE Trans. Intell. Veh..

Dettmers, T., et al., 2023. QLoRA: Efficient finetuning of quantized LLMs. arXiv preprint arXiv:2305.14314.

Ding, X., Han, J., Xu, H., Liang, X., Zhang, W., Li, X., 2024. Holistic autonomous driving understanding by Bird's-Eye-View injected Multi-Modal large models. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.. CVPR, pp. 13668–13677.

Ding, X., Han, J., Xu, H., Zhang, W., Li, X., 2023. HiLM-D: Towards High-Resolution understanding in multimodal large language models for autonomous driving. arXiv preprint arXiv:2309.05186.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Driess, D., Nguyen, A., Xia, F., et al., 2023a. PaLM-E: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 URL https://arxiv.org/abs/2303.03378.

Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al., 2023b. Palm-E: An embodied multimodal language model. In: Proc. Int. Conf. ML. ICML, pp. 8469–8488.

Duan, Y., Zhang, Q., Xu, R., 2024. Prompting Multi-Modal tokens to enhance End-to-End autonomous driving imitation learning with LLMs. arXiv preprint arXiv:2404.04869.

Fourati, S., Jaafar, W., Baccar, N., 2025. A novel MLLM-based approach for autonomous driving in different weather conditions. Comput. AI Connect. 2, 1–11.

Frantar, E., et al., 2022. GPTQ: Accurate post-training quantization for generative pre-trained transformers. Adv. Neural Inf. Process. Syst..

Fu, D., Lei, W., Wen, L., Cai, P., Mao, S., Dou, M., Shi, B., Qiao, Y., 2024. LimSim++: A Closed-Loop platform for deploying multimodal LLMs in autonomous driving. arXiv preprint arXiv:2402.01246.

Gao, H., Li, Y., Long, K., Yang, M., Shen, Y., 2024. A survey for foundation models in autonomous driving. arXiv preprint arXiv:2402.01105.

Ge, J., Chang, C., Zhang, J., Li, L., Na, X., Lin, Y., Li, L., Wang, F.-Y., 2024. LLM-based operating systems for automated vehicles: A new perspective. IEEE Trans. Intell. Veh..

Ge, J., Sun, S., Owens, J., Galvez, V., Gologorskaya, O., Lai, J.C., Pletcher, M.J., Lai, K., 2023. Development of a liver Disease-Specific large language model chat interface using retrieval augmented generation. MedRxiv.

Geiger, A., Lenz, P., Stiller, C., Urtasun, R., 2012. KITTI vision benchmark suite. URL http://www.cvlibs.net/datasets/kitti/. (Accessed 28 August 2024).

Gopalkrishnan, A., Greer, R., Trivedi, M., 2024. Multi-Frame, lightweight & efficient Vision-Language models for question answering in autonomous driving. arXiv preprint arXiv:2403.19838.

Guan, Y., Liao, H., Li, Z., Hu, J., Yuan, R., Li, Y., Zhang, G., Xu, C., 2024. World models for autonomous driving: An initial survey. IEEE Trans. Intell. Veh..

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

Hadi, M.U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M.B., Akhtar, N., Wu, J., Mirjalili, S., et al., 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. Authorea Preprint.

Han, Z., Gao, C., Liu, J., Zhang, S.Q., et al., 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. arXiv preprint arXiv:2403.14608.

Han, W., Guo, D., Xu, C.-Z., Shen, J., 2024. DME-Driver: Integrating human decision logic and 3D scene perception in autonomous driving. arXiv preprint arXiv:2401.03641.

Hasanujjaman, M., Chowdhury, M.Z., Jang, Y.M., 2023. Sensor fusion in autonomous vehicle with traffic surveillance camera system: detection, localization, and AI networking. Sensors 23 (6), 3335.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: Proc. IEEE/CVF Conf. Comput. Vis. Patt. Recogn.. pp. 16000–16009.

Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G., 2023. GAIA-1: a generative world model for autonomous driving. pp. 1–25, arXiv preprint arXiv:2309.17080.

Huang, Y., 2024. Leveraging large language models for enhanced NLP task performance through knowledge distillation and optimized training strategies. arXiv preprint arXiv:2402.09282.

Huang, Y., Chen, Y., Li, Z., 2023. Applications of large scale foundation models for autonomous driving. arXiv preprint arXiv:2311.12144.

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Patra, B., et al., 2024. Language is not all you need: Aligning perception with language models. Adv. Neural Inf. Process. Syst. 36.

Huang, Y., Sansom, J., Ma, Z., Gervits, F., Chai, J., 2024. DriVLMe: Exploring foundation models as autonomous driving agents that perceive, communicate, and navigate. In: Proc. Vis. & Lang. for Autonom. Driv. Robot. Wrkshp.. pp. 1–12.

Huang, Z., Sheng, Z., Qu, Y., You, J., Chen, S., 2025. Vlm-rl: A unified vision language models and reinforcement learning framework for safe autonomous driving. Transp. Res. C 180, 105321.

Inoue, Y., Yada, Y., Tanahashi, K., Yamaguchi, Y., 2024. NuScenes-MQA: Integrated evaluation of captions and QA for autonomous driving datasets using markup annotations. In: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.. pp. 930–938.

Janai, J., Güney, F., Behl, A., Geiger, A., et al., 2020. Computer vision for autonomous vehicles: Problems, datasets and state of the art. Found. Trends Comput. Graph. Vis. 12 (1–3), 1–308.

Jia, F., Mao, W., Liu, Y., Zhao, Y., Wen, Y., Zhang, C., Zhang, X., Wang, T., 2023. Adriver-I: A general world model for autonomous driving. arXiv preprint arXiv:2311.13549.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In: Proc. Int. Conf. Mach. Learn.. ICML, PMLR, pp. 4904–4916.

Jiang, S., Huang, Z., Qian, K., Luo, Z., Zhu, T., Zhong, Y., Tang, Y., Kong, M., Wang, Y., Jiao, S., et al., 2025. A survey on Vision-Language-Action models for autonomous driving. arXiv preprint arXiv:2506.24044.

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

Khan, M.H., 2024. Multi-agent LLM Framework for Conversational Data Retrieval from Structured Data. Intern. Report, TECHNISCHE UNIVERSITäT MüNCHEN.

Kim, W., Kim, J.H., Cha, Y.K., Chong, S., Kim, T.J., 2023. Completeness of reporting of systematic reviews and meta-analysis of diagnostic test accuracy (DTA) of radiological articles based on the PRISMA-DTA reporting guideline. Acad. Radiol. 30 (2), 258–275.

Kiss, G., Garai-Fodor, M., 2024. The role of Self-Driving vehicles in sustainability and road safety from a generation-specific perspective. Decis. Mak.: Appl. Manag. Eng. 7 (1), 568–584.

Kong, X., Braunl, T., Fahmi, M., Wang, Y., 2024. A superalignment framework in autonomous driving with large language models. arXiv preprint arXiv:2406.05651.

Kou, W.-B., Lin, Q., Tang, M., Xu, S., Ye, R., Leng, Y., Wang, S., Chen, Z., Zhu, G., Wu, Y.-C., 2024. pFedLVM: A large vision model (LVM)-Driven and latent feature-based personalized federated learning framework in autonomous driving. arXiv preprint arXiv:2405.04146.

Li, N., Chen, Y., Li, W., Ding, Z., Zhao, D., Nie, S., 2023. Bvit: Broad attention-based vision transformer. IEEE Trans. Neural Netw. Learn. Syst. 35 (9), 12772–12783.

Li, K., Chen, K., Wang, H., Hong, L., Ye, C., Han, J., Chen, Y., Zhang, W., Xu, C., Yeung, D.-Y., et al., 2022. CODA: A real-world road corner case dataset for object detection in autonomous driving. In: Europ. Conf. Comput. Vis.. Springer, pp. 406–423.

Li, J., Li, D., Savarese, S., Hoi, S.C.H., 2023. BLIP-2: Bootstrapped Language-Image pretraining with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 URL https://arxiv.org/abs/2301.12597.

Li, J., Li, D., Xiong, C., Hoi, S., 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Proc. Int. Conf. Mach. Learn.. ICML, PMLR, pp. 12888–12900.

Li, J., Li, J., Yang, G., Yang, L., Chi, H., Yang, L., 2025. Applications of large language models and multimodal large models in autonomous driving: A comprehensive review. Drones.

Li, Y., Mao, H., Girshick, R., He, K., 2022. Exploring plain vision transformer backbones for object detection. In: Europ. Conf. Comput. Vis.. Springer, pp. 280–296.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H., 2021. Align before fuse: Vision and language representation learning with momentum distillation. Adv. Neural Inf. Process. Syst. 34, 9694–9705.

Li, Z., Shang, T., Xu, P., 2025. Multi-Modal attention perception for intelligent vehicle navigation using deep reinforcement learning. IEEE Trans. Intell. Transp. Syst..

Li, Y., Tian, W., Jiao, Y., Chen, J., Jiang, Y.-G., 2024a. Eyes can deceive: Benchmarking counterfactual reasoning abilities of Multi-modal large language models. arXiv preprint arXiv:2404.12966.

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

Li, Z., Zhang, C., Ma, W.-C., Zhou, Y., Huang, L., Wang, H., Lim, S., Zhao, H., 2023. VoxelFormer: Bird's-eye-view feature generation based on dual-view attention for multi-view 3D object detection. arXiv preprint arXiv:2304.01054.

Liang, M., Su, J.-C., Schulter, S., Garg, S., Zhao, S., Wu, Y., Chandraker, M., 2024. AIDE: An automatic data engine for object detection in autonomous driving. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.. CVPR, pp. 14695–14706.

Liao, G., Li, J., Ye, X., 2024. VLM2Scene: Self-supervised Image-Text-LiDAR learning with foundation models for autonomous driving scene understanding. In: Proc. AAAI Conf. Artifi. Intelli.. pp. 3351–3359.

Liu, P., Bai, Y., et al., 2023. Training language models to follow rules with reinforcement learning from human feedback. arXiv preprint arXiv:2305.18290.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H., 2022. Video swin transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3202–3211.

Liu, X., Song, Y., Li, X., Sun, Y., Lan, H., Liu, Z., Jiang, L., Li, J., 2024. ED-ViT: Splitting vision transformer for distributed inference on edge devices. arXiv preprint arXiv:2410.11650.

Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L., Han, S., 2023. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In: IEEE Int. Conf. Robot. Automa.. ICRA, IEEE, pp. 2774–2781.

Liu, M., Yurtsever, E., Fossaert, J., Zhou, X., Zimmer, W., Cui, Y., Zagar, B.L., Knoll, A.C., 2024. A survey on autonomous driving datasets: Statistics, annotation quality, and a future outlook. IEEE Trans. Intell. Veh..

Liu, F., Zheng, Q., Tian, X., Shu, F., Jiang, W., Wang, M., Elhanashi, A., Saponara, S., 2025. Rethinking the multi-scale feature hierarchy in object detection transformer (DETR). Appl. Soft Comput. 175, 113081.

Luo, S., Chen, W., Tian, W., Liu, R., Hou, L., Zhang, X., Shen, H., Wu, R., Geng, S., Zhou, Y., Shao, L., Yang, Y., Gao, B., Li, Q., Wu, G., 2024. Delving into Multi-Modal Multi-Task foundation models for road scene understanding: From learning paradigm perspectives. IEEE Trans. Intell. Veh. 1–25. http://dx.doi.org/10.1109/TIV.2024.3406372.

Ma, Y., Cui, C., Cao, X., Ye, W., Liu, P., Lu, J., Abdelraouf, A., Gupta, R., Han, K., Bera, A., Rehg, J.M., Wang, Z., 2024a. LaMPilot: An open benchmark dataset for autonomous driving with language model programs. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.. CVPR, pp. 15141–15151. http://dx.doi.org/10.1109/CVPR52733.2024.01434.

Ma, Y., Cui, C., Cao, X., Ye, W., Liu, P., Lu, J., Abdelraouf, A., Gupta, R., Han, K., Bera, A., et al., 2024b. Lampilot: An open benchmark dataset for autonomous driving with language model programs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15141–15151.

Maaz, M., Rasheed, H., Khan, S., Khan, F.S., 2023. Video-ChatGPT: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424.

Mahor, V., Rawat, R., Kumar, A., Garg, B., Pachlasiya, K., et al., 2023. IoT and artificial intelligence techniques for public safety and security. In: Smart Urban Comput. Appl.. River Publishers, pp. 111–126.

Malla, S., Choi, C., Dwivedi, I., Choi, J.H., Li, J., 2023. Drama: Joint risk localization and captioning in driving. In: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.. pp. 1043–1052.

Mao, J., Niu, M., Jiang, C., Liang, H., Chen, J., Liang, X., Li, Y., Ye, C., Zhang, W., Li, Z., et al., 2021. One million scenes for autonomous driving: Once dataset. arXiv preprint arXiv:2106.11037.

Mao, J., Qian, Y., Zhao, H., Wang, Y., 2023a. GPT-driver: Learning to drive with GPT. arXiv preprint arXiv:2310.01415.

Mao, J., Ye, J., Qian, Y., Pavone, M., Wang, Y., 2023b. A language agent for autonomous driving. arXiv preprint arXiv:2311.10813.

Marvin, G., Hellen, N., Jjingo, D., Nakatumba-Nabende, J., 2023. Prompt engineering in large language models. In: Proc. Int. Conf. Data Intelli. Cogn. Informat.. Springer, pp. 387–402.

MehtaS, R.M.M., 2022. MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer.

Mei, J., Zhu, A.Z., Yan, X., Yan, H., Qiao, S., Chen, L.-C., Kretzschmar, H., 2022. Waymo open dataset: Panoramic video panoptic segmentation. In: Proc. Europ. Conf. Comput. Vis.. Springer, pp. 53–72.

Meng, L., Li, H., Chen, B.-C., Lan, S., Wu, Z., Jiang, Y.-G., Lim, S.-N., 2022. AdaViT: Adaptive vision transformers for efficient image recognition. In: Proc. IEEE/CVF Conf. Comput. Vis. Patt. Recogn.. pp. 12309–12318.

Miceli-Barone, A.V., Lascarides, A., Innes, C., 2023. Dialogue-based generation of self-driving simulation scenarios using Large Language Models. arXiv preprint arXiv:2310.17372.

Murtaza, M., Cheng, C.-T., Fard, M., Zeleznikow, J., 2024. Transforming driver education: A comparative analysis of LLM-Augmented training and conventional instruction for autonomous vehicle technologies. Int. J. Artif. Intell. Educ. 1–38.

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

Nouri, A., Cabrero-Daniel, B., Törner, F., Sivencrona, H.a., Berger, C., 2024. Engineering safety requirements for autonomous driving with large language models. In: Proc. IEEE Int. Requir. Engineer. Conf.. RE, pp. 218–228. http://dx.doi.org/10.1109/RE59067.2024.00029.

NVIDIA, 2022. Jetson AGX orin series modules. URL https://developer.nvidia.com/embedded/jetson-agx-orin.

NVIDIA Corporation, 2022. NVIDIA DRIVE Thor^{TM} — Centralized car computer unifying cluster, infotainment, automated driving and parking in a single cost-saving system. https://nvidianews.nvidia.com/news/nvidia-unveils-drive-thor-centralized-car-computer-unifying-cluster-infotainment-automated-driving-and-parking-in-a-single-cost-saving-system, (Accessed 05 November 2025). (2022 (published) / 2024 (press re-release)).

Ooi, K.-B., Tan, G.W.-H., Al-Emran, M., Al-Sharafi, M.A., Capatina, A., Chakraborty, A., Dwivedi, Y.K., Huang, T.-L., Kar, A.K., Lee, V.-H., et al., 2023. The potential of generative artificial intelligence across disciplines: Perspectives and future directions. J. Comput. Inf. Syst. 1–32.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.

Parekh, D., Poddar, N., Rajpurkar, A., Chahal, M., Kumar, N., Joshi, G.P., Cho, W., 2022. A review on autonomous vehicles: Progress, methods and challenges. Electron. 11 (14), 2162.

Park, S., Lee, M., Kang, J., Choi, H., Park, Y., Cho, J., Lee, A., Kim, D., 2024. VLAAD: Vision and language assistant for autonomous driving. In: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.. pp. 980–987.

Peng, M., Guo, X., Chen, X., Zhu, M., Chen, K., Wang, X., Wang, Y., et al., 2024. LC-LLM: Explainable Lane-Change intention and trajectory predictions with large language models. arXiv preprint arXiv:2403.18344.

Pi, R., Han, T., Xie, Y., Pan, R., Lian, Q., Dong, H., Zhang, J., Zhang, T., 2024. MLLM-Protector: Ensuring MLLM's safety without hurting performance. arXiv preprint arXiv:2401.02906.

Rajesh, N., Yalavarthy, A.B., Vamsi, V.K., Saranya, G., 2024. BEiT transformer models to aid in the early detection of parkinson illness. In: 2024 International Conference on Advances in Computing, Communication and Applied Informatics. ACCAI, IEEE, pp. 1–7.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M., 2022. Hierarchical text-conditional image generation with clip latents. p. 3, arXiv preprint arXiv:2204.06125, 1 (2).

Ray, J.Z.Z.H.A., Ohn-Bar, E., 2024. Feedback-Guided autonomous driving. In: Proc. IEEE/CVF Conf. Comput. Vis. Patt. Recogn. (CVPR). pp. 15000–15011.

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

Shukor, M., Dancette, C., Cord, M., 2023. EP-ALM: Efficient perceptual augmentation of language models. In: Proc. IEEE/CVF Int. Conf. Comput. Vis.. pp. 22056–22069.

Singh, A., 2023. Transformer-based sensor fusion for autonomous driving: A survey. In: Proc. IEEE/CVF Int. Conf. Comput. Vis.. pp. 3312–3317.

Song, Z., He, Z., Li, X., Ma, Q., Ming, R., Mao, Z., Pei, H., Peng, L., Hu, J., Yao, D., et al., 2023. Synthetic datasets for autonomous driving: A survey. IEEE Trans. Intell. Veh..

Sreeram, S., Wang, T.-H., Maalouf, A., Rosman, G., Karaman, S., Rus, D., 2024. Probing multimodal LLMs as world models for driving. arXiv preprint arXiv:2405.05956.

Su, H., Brooks, J., Jia, Y., 2023. Development and evaluation of comfort assessment approaches for passengers in autonomous vehicles. SAE Int. J. Adv. Curr. Pr. Mob. 5 (2023-01-0788), 2068–2077.

Sun, J., Shaib, C., Wallace, B.C., 2023. Evaluating the zero-shot robustness of instruction-tuned language models. arXiv preprint arXiv:2306.11270.

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

Tang, B., Wu, W., Wang, Y., Wu, Y., Zheng, S., Liu, Z., 2023. MobileSAM: Lightweight segment anything model with MobileViT backbone. arXiv preprint arXiv:2306.00989.

Tesla, I., 2024. Tesla finally releases FSD v12: its last hope for self-driving. Electrek URL https://electrek.co/2024/01/22/tesla-releases-fsd-v12-last-hope-self-driving/.

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

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Vayadande, K., Pednekar, C.B., Khune, P.A., Prabhavalkar, V.S., Dange, V.R., 2024. GPT-3-and DALL-E-Powered applications: A complete survey. In: How Machine Learning is Innovatin Today's World: A Concise Technical Guide. pp. 329–341.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al., 2023. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In: Proc. IEEE/CVF Conf. Comput. Vis. Patt. Recogn.. pp. 19175–19186.

Wang, Y., Chen, W., Han, X., Lin, X., Zhao, H., Liu, Y., Zhai, B., Yuan, J., You, Q., Yang, H., 2024. Exploring the reasoning abilities of multimodal large language models (MLLMs): A comprehensive survey on emerging trends in multimodal reasoning. arXiv preprint arXiv:2401.06805.

Wang, L., Jiang, H., Cai, P., Fu, D., Wang, T., Cui, Z., Ren, Y., Yu, H., Wang, X., Wang, Y., 2023. AccidentGPT: Accident analysis and prevention from V2X environmental perception with multi-modal large model. arXiv preprint arXiv:2312.13156.

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

Wang, W., Li, H., Liu, J., Wang, H., Hou, L., et al., 2022. GIT: A generative image-to-text transformer for vision and language. arXiv preprint arXiv:2205.14100 URL https://arxiv.org/abs/2205.14100.

Wang, Y., Liu, Q., Jiang, Z., Wang, T., Jiao, J., Chu, H., Gao, B., Chen, H., 2025. RAD: Retrieval-augmented decision-making of Meta-Actions with vision-language models in autonomous driving. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3838–3848.

Wang, X., Liu, S., Yu, L., Shen, F., Wang, Z., Yu, G., Shen, D., Yuille, A.L., 2022. VLC-BERT: Vision-language Pre-training of BERT using large-scale weakly supervised data. arXiv preprint arXiv:2209.07865.

Wang, J., Liu, Z., Zhao, L., Wu, Z., Ma, C., Yu, S., Dai, H., Yang, Q., Liu, Y., Zhang, S., et al., 2023. Review of large vision models and visual prompt engineering. Meta-Radiology 100047.

Wang, T., Xie, E., Chu, R., Li, Z., Luo, P., 2024. DriveCoT: Integrating chain-of-thought reasoning with End-to-End driving. arXiv preprint arXiv:2403.16996.

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

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L., 2022. Pvt v2: Improved baselines with pyramid vision transformer. Comput. Vis. Media 8 (3), 415–424.

Wang, S., Yu, Z., Jiang, X., Lan, S., Shi, M., Chang, N., Kautz, J., Li, Y., Alvarez, J.M., 2024. OmniDrive: A holistic LLM-Agent framework for autonomous driving with 3D perception, reasoning and planning. arXiv preprint arXiv:2405.01533.

Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y., 2021. SimVLM: Simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904.

Wang, X., Zhu, Z., Huang, G., Chen, X., Lu, J., 2023. DriveDreamer: Towards real-world-driven world models for autonomous driving. arXiv preprint arXiv:2309.09777.

Wei, Y., Wang, Z., Lu, Y., Xu, C., Liu, C., Zhao, H., Chen, S., Wang, Y., 2024a. Editable scene simulation for autonomous driving via collaborative LLM-Agents. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.. CVPR, pp. 15077–15087.

Wei, Y., Wang, Z., Lu, Y., Xu, C., Liu, C., Zhao, H., Chen, S., Wang, Y., 2024b. Editable scene simulation for autonomous driving via collaborative LLM-Agents. In: Proc. IEEE/CVF Conf. Comput. Vis. Patt. Recogn.. CVPR, pp. 15077–15087. http://dx.doi.org/10.1109/CVPR52733.2024.01428.

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

Wu, J., Gao, B., Gao, J., Yu, J., Chu, H., Yu, Q., Gong, X., Chang, Y., Tseng, H.E., Chen, H., et al., 2023. Prospective role of foundation models in advancing autonomous vehicles. Research.

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

Wu, Y., Li, D., Chen, Y., Jiang, R., Zou, H.P., Fang, L., Wang, Z., Yu, P.S., 2025. Multi-agent autonomous driving systems with large language models: A survey of recent advances. arXiv preprint arXiv:2502.16804.

Wu, K., Li, W., Xiao, X., 2024. AccidentGPT: Large Multi-Modal foundation model for traffic accident analysis. arXiv preprint arXiv:2401.03040.

Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., Yuan, L., 2022. TinyViT: Fast pretraining distillation for small vision transformers. In: Europ. Conf. Comp. Vis.. Springer, pp. 68–85.

Xiang, C., Feng, C., Xie, X., Shi, B., Lu, H., Lv, Y., Yang, M., Niu, Z., 2023. Multi-sensor fusion and cooperative perception for autonomous driving: A review. IEEE Intell. Transp. Syst. Mag..

Xie, Y., Chen, H., Meyer, G.P., Lee, Y.J., Wolff, E.M., Tomizuka, M., Zhan, W., Chai, Y., Huang, X., 2024. Cohere3D: Exploiting temporal coherence for unsupervised representation learning of Vision-based autonomous driving. arXiv preprint arXiv:2402.15583.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 34, 12077–12090.

Xilinx, A., 2023. Versal AI edge series. URL https://www.amd.com/en/products/soc/versal-ai-edge.

XinyuHuang, P., XinjingCheng, D., Geng, Q., Yang, R., 2020. The ApolloScape open dataset for autonomous driving and its application. IEEE Trans. Pattern Anal. Mach. Intell. 42 (10).

Xu, Z., Bai, Y., Zhang, Y., Li, Z., Xia, F., Wong, K.-Y.K., Wang, J., Zhao, H., 2025. DriveGPT4-V2: Harnessing large language model capabilities for enhanced closed-loop autonomous driving. In: Proc. IEEE/CVF Conf. Comput. Vis. Patt. Recogn.. CVPR, pp. 17261–17270. http://dx.doi.org/10.1109/CVPR52734.2025.01609.

Xu, M., Cai, D., Wu, Y., Li, X., Wang, S., 2024a. FwdLLM: Efficient FedLLM using forward gradient. ArXiv arXiv:2308.13894.

Xu, Y., Dai, W., Wang, F., Wu, L., Liang, X., 2023. LiteVLM: Vision-language foundation models at Light-Speed. In: Proc. IEEE/CVF Conf. Comp. Vis. Patt. Recogn.. CVPR, pp. 1–6.

Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., Luo, P., 2023. LVLM-eHUB: A comprehensive evaluation benchmark for large vision-language models. arXiv preprint arXiv:2306.09265.

Xu, M., Yin, W., Cai, D., Yi, R., Xu, D., Wang, Q., Wu, B., Zhao, Y., Yang, C., Wang, S., et al., 2024b. A survey of resource-efficient LLM and multimodal foundation models. arXiv preprint arXiv:2401.08092.

Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.-Y.K., Li, Z., Zhao, H., 2024. DriveGPT4: Interpretable End-to-End autonomous driving via large language model. IEEE Robot. Autom. Lett. 9 (10), 8186–8193. http://dx.doi.org/10.1109/LRA.2024.3440097.

Xu, C., Zhao, H., Lu, X., Sun, K., Gao, B., Chen, H., 2025. A deep reinforcement learning method for autonomous driving integrating Multi-Modal fusion. IEEE Trans. Intell. Transp. Syst..

Xu, H., et al., 2023. FusedMM: A fused matrix multiplication kernel for Memory-Efficient attention. arXiv preprint arXiv:2302.05442.

Yan, X., Zhang, H., Cai, Y., Guo, J., Qiu, W., Gao, B., Zhou, K., Zhao, Y., Jin, H., Gao, J., et al., 2024. Forging vision foundation models for autonomous driving: Challenges, methodologies, and opportunities. arXiv preprint arXiv:2401.08045.

Yang, Z., Jia, X., Li, H., Yan, J., 2023. LLM4Drive: A survey of large language models for autonomous driving. ArXiv E-Prints, arXiv–2311.

Yang, R., Zhang, X., Fernandez-Laaksonen, A., Ding, X., Gong, J., 2024. Driving style alignment for LLM-powered driver agent. arXiv preprint arXiv:2403.11368.

Yang, Y., Zhang, Q., Li, C., Marta, D.S., Batool, N., Folkesson, J., 2024. Human-centric autonomous systems with LLMs for user command reasoning. In: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.. pp. 988–994.

Yazgan, M., Akkanapragada, M.V., Marius Zöllner, J., 2024. Collaborative perception datasets in autonomous driving: A survey. In: Proc. IEEE Intelli. Veh. Symp.. IV, pp. 2269–2276. http://dx.doi.org/10.1109/IV55156.2024.10588870.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E., 2023. A survey on multimodal large language models. arXiv preprint arXiv:2306.13549.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T., 2020. BDD100k: A diverse driving dataset for heterogeneous multitask learning. In: Proc. IEEE/CVF Conf. Comput. Vis. Patt. Recogn.. pp. 2636–2645.

Yu, J., Zhai, X., Xiao, T., Barret, D., Li, C., Kolesnikov, A., Beyer, L., Funke, F., Zhou, Y., Zhai, S., et al., 2022. CoCa: Contrastive captioners are Image-Text foundation models. arXiv preprint arXiv:2205.01917.

Yuan, J., Sun, S., Omeiza, D., Zhao, B., Newman, P., Kunze, L., Gadd, M., 2024. RAG-Driver: Generalisable driving explanations with Retrieval-Augmented In-Context learning in Multi-Modal large language model. arXiv preprint arXiv:2402.10828.

Zang, Z., Li, Z., Zhao, Q., Zhou, P., Liu, Y., Liang, D., Zhang, Z., Liu, Z., 2022. Multi-modal prompt learning for visual grounding. IEEE Trans. Multimed..

Zhang, S., Fu, D., Liang, W., Zhang, Z., Yu, B., Cai, P., Yao, B., 2024. TrafficGPT: Viewing, processing and interacting with traffic foundation models. Transp. Policy 150, 95–105.

Zhang, H., Li, X., Bing, L., 2023. Video-Llama: An instruction-tuned audio-visual language model for video understanding. In: Proc. Conf. Empir. Methods Nat. Lang. Process.: Syst. Demo.. pp. 1–11.

Zhang, D., Yu, Y., Li, C., Dong, J., Su, D., Chu, C., Yu, D., 2024. MM-LLMs: Recent advances in multimodal large language models. arXiv preprint arXiv:2401.13601.

Zhao, X., Fang, Y., Min, H., Wu, X., Wang, W., Teixeira, R., 2023. Potential sources of sensor data anomalies for autonomous vehicles: An overview from road vehicle safety perspective. Expert Syst. Appl. 121358.

Zhao, Y., Pang, T., Du, C., Yang, X., Li, C., Cheung, N.-M.M., Lin, M., 2024. On evaluating adversarial robustness of large vision-language models. Adv. Neural Inf. Process. Syst. 36.

Zhao, G., Wang, X., Zhu, Z., Chen, X., Huang, G., Bao, X., Wang, X., 2024. DriveDreamer-2: LLM-Enhanced world models for diverse driving video generation. arXiv preprint arXiv:2403.06845.

Zhao, Y., Zhou, J., Bi, D., Mihalj, T., Hu, J., Eichberger, A., 2025. A survey on the application of large language models in scenario-based testing of automated driving systems. arXiv preprint arXiv:2505.16587.

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

Zheng, O., Abdel-Aty, M., Wang, D., Wang, C., Ding, S., 2023. TrafficSafetyGPT: Tuning a pre-trained large language model to a domain-specific expert in transportation safety. arXiv preprint arXiv:2307.15311.

Zhou, J., He, J., Zhao, W., Li, W., Wen, S., Liu, Y., Liu, Z., 2022. Dynamic early exit for efficient inference of Transformer-Based models. IEEE Trans. Neural Netw. Learn. Syst. 33 (8), 3675–3687.

Zhou, X., Liu, M., Yurtsever, E., Zagar, B.L., Zimmer, W., Cao, H., Knoll, A.C., 2024. Vision language models in autonomous driving: A survey and outlook. IEEE Trans. Intell. Veh..

Zhou, Z., Zhang, J., Zhang, J., Wang, B., Shi, T., Khamis, A., 2024. In-context learning for automated driving scenarios. arXiv preprint arXiv:2405.04135.

Zhu, D., et al., 2023. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 URL https://arxiv.org/abs/2304.10592.