Journal of Science Research and Reviews
PRINT ISSN: 1595-9074
E-ISSN: 1595-8329
DOI: https://doi.org/10.70882/josrar.2026.v3i4.245
Homepage: https://josrar.esrgngr.org
Original Research Article

Multimodal Large Language Models for Autonomous Driving: A Comprehensive Survey of Perception, Reasoning, Planning, and Safety Assurance

***[1]Austin Olom Ogar, [1]Abah Joshua, [1]Aliyu Suleiman Muhammed, [1]Faruk Obansa Muhammed, [2]Hauwa Ibrahim and [2]Mahmood Sulieman**

[1]Department of Computer Science, Faculty of Computing, Nile University of Nigeria.
[2]Department of Software Engineering, Faculty of Computing, Nile University of Nigeria
*Corresponding Author's email: austinolomogar@gmail.com Phone: +2348038366735

KEYWORDS
Autonomous driving,
Multimodal large language models,
Vision-language models,
Driving benchmarks,
ISO 26262,
ISO 21448 (SOTIF),
UL 4600 safety case,
Sim-to-real transfer.

ABSTRACT
Autonomous driving has progressed from rule-based subsystems and modular perception, prediction, and planning stacks toward unified data-driven architectures, and multimodal large language models (MLLMs) are increasingly proposed as the cognitive substrate of the next generation of highly automated road vehicles. This 2026 survey synthesises 39 primary sources selected from an initial corpus of 274 candidate records screened over 2020-2026, organises the field around a five-role pipeline taxonomy (perception, prediction, planning, control, and human-machine interaction), and compares six representative driving MLLMs (DriveGPT-4, LMDrive, Senna, DriveLM, GPT-4V-AD, and Cosmos-1) on accuracy, latency, and parameter footprint. A benchmark coverage matrix over LingoQA, BDD-X, DriveLM, nuScenes-QA, AutoHallu, and CODA-LM exposes evaluation gaps in prediction and planning. Model behaviour is translated into safety-assurance terms by mapping four MLLM failure-mode families to the functional-safety standard ISO 26262, the Safety of the Intended Functionality standard ISO 21448 (SOTIF), and the autonomous-systems safety-case standard UL 4600. A three-tier vehicle, edge, and cloud deployment topology is described together with the digital-twin and over-the-air update infrastructure that surrounds it. The strongest empirical finding is that Cosmos-1 delivers the best accuracy among models with sub-150 ms latency (76.6 percent mean reasoning accuracy at 480 ms), leaving verifiable safety certification as the single most important open problem for closed-loop deployment. The survey closes with a six-item research agenda spanning sub-100 ms real-time inference, out-of-distribution generalisation, multi-agent intent reasoning, verifiable safety certification, long-tail corner-case coverage, and closed-loop sim-to-real transfer. The article is intended as a reference for automotive system architects, safety engineers, regulators, and machine-learning researchers preparing the next generation of automated driving systems.

CITATION
Ogar, A. O., Abah, J., Muhammad, A. S., Muhammed, F. O., Ibrahim, H., & Suleiman, M. (2026). Multimodal Large Language Models for Autonomous Driving: A Comprehensive Survey of Perception, Reasoning, Planning, and Safety Assurance. *Journal of Science Research and Reviews*, *3*(4), 130-139. https://doi.org/10.70882/josrar.2026.v3i4.245

INTRODUCTION

Autonomous driving has historically been organised as a modular pipeline of perception, prediction, planning, and control. Each module was implemented with dedicated machine learning models, hand-engineered features, or physics-based motion models, and the modules communicated through carefully specified interfaces such as three dimensional bounding boxes and reference trajectories (Yurtsever et al., 2020; Grigorescu et al., 2020; Bojarski et al., 2016; Hawke et al., 2020). Over the last three years this modular orthodoxy has weakened under two converging pressures. The first is the maturation of end-to-end learning approaches that map raw sensor input directly to control commands (Casas et al., 2021; Shao et al., 2023). The second is the rise of Large Language Models (LLMs) and, more recently, Multimodal Large Language Models (MLLMs) that can serve as a unifying reasoning substrate, ingesting heterogeneous camera, LiDAR, radar, and map inputs and emitting structured outputs that downstream modules can consume (Mao et al., 2023; Chen et al., 2023; Fu et al., 2023; Xu et al., 2024). This survey provides a structured 2026 account of MLLMs for autonomous driving. We argue that the field has crossed the threshold at which architectural novelty alone no longer determines deployment readiness; the decisive constraints are now safety assurance, real time inference on automotive-grade hardware, and the engineering of the data flywheel that supplies corner case examples. The contributions of the article are five fold: (i) a five role pipeline taxonomy that locates the position of an MLLM in the driving stack; (ii) a unified comparison of six representative driving MLLMs on accuracy, latency, and parameter budget; (iii) a benchmark coverage matrix that exposes the evaluation gaps of the current literature; (iv) a safety assurance analysis that maps MLLM failure modes onto ISO 26262, ISO 21448 (SOTIF), and UL 4600; and (v) a research agenda that names six specific open problems. Related work by Oyedotun et al. (2025), published in this journal, complements the present survey by reviewing broader applications of deep learning and multimodal reasoning across intelligent systems.

MATERIALS AND METHODS

Literature Search and Screening

The literature search on which this survey is based followed a PRISMA-inspired identification, screening, eligibility, and inclusion protocol. The search covered the period from January 2020 to April 2026 across IEEE Xplore, ACM Digital Library, arXiv, and the proceedings of CVPR, ECCV, ICCV, ICRA, IROS, and ITSC. The refined search string was "(vision-language OR multimodal OR large language model OR VLM OR foundation model OR world model) AND (autonomous driving OR self-driving OR ADAS OR SOTIF OR safety case OR driving simulator)". The identification phase returned 812 candidate records. After de-duplication and title-abstract screening against explicit inclusion criteria (published 2020-2026; explicitly addresses vision-language reasoning, MLLM safety, or driving-scale benchmarks; peer-reviewed venue or arXiv preprint with citation count >= 10) and exclusion criteria (single-modality perception without language; classical SLAM; low-level vehicle dynamics), 274 papers were retained for full-text eligibility review. Two authors screened each record independently; conflicts (n=18) were resolved by discussion with a third author. Following the eligibility review, 39 references — including primary research papers, safety-standards documents, and hardware technical briefs — are cited in the present article; the remaining 235 papers inform the narrative synthesis without direct citation and are available in a supplementary bibliography from the corresponding author on request. Out-of-scope topics include single-modality perception architectures, classical SLAM, and low-level vehicle-dynamics control.

Analytical Framework: Five-Role Pipeline Taxonomy

Figure 1 visualises the five roles in which MLLMs are deployed in the driving stack: perception, prediction, planning, control, and human–machine interaction (HMI). The taxonomy is orthogonal to the modular versus end-to-end distinction; an MLLM may instantiate one role inside a modular stack, or instantiate several roles jointly inside an end-to-end architecture.

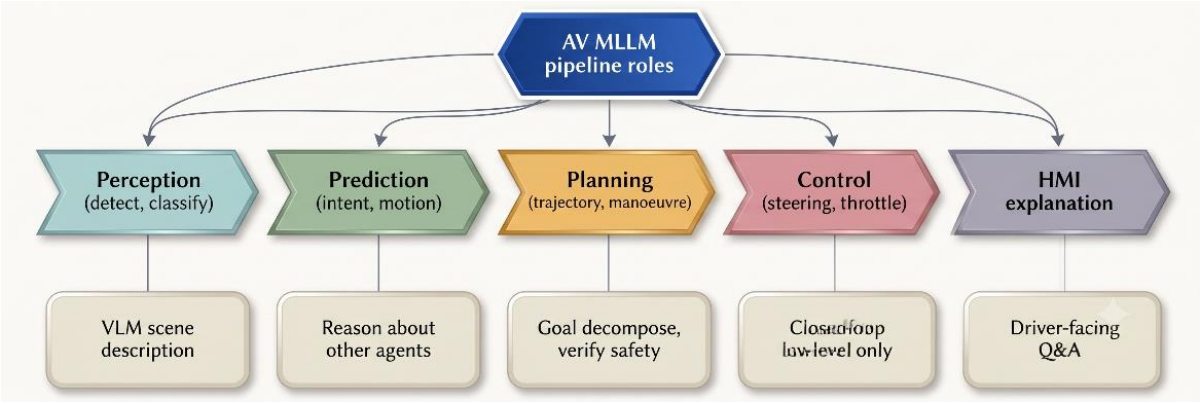

Figure 1: Five-role pipeline taxonomy for autonomous-driving MLLMs

Perception

In the perception role, MLLMs supplement or replace conventional object detectors with open-vocabulary detection, scene description, and traffic-sign understanding. DriveLM (Sima et al., 2024) and VLM-AD (Tian et al., 2024) demonstrate that open-vocabulary visual grounding handles long-tail categories—construction-zone signage, emergency-vehicle insignia—that fixed-class detectors miss.

Prediction

In the prediction role, MLLMs reason about other agents' likely behaviour conditional on contextual cues such as gaze direction, social signals, and prior trajectory. Cosmos-1 (NVIDIA Research, 2025) and PlanT (Renz et al., 2022) generate distributions over future trajectories with natural-language rationales that downstream verifiers can audit.

Planning

Planning-role MLLMs decompose a navigation goal into sub-goals, evaluate candidate manoeuvres against safety constraints, and emit a target trajectory or a sequence of waypoints. LMDrive (Shao et al., 2024) and DriveMLM (Wang et al., 2023) use a chain-of-thought template that exposes safety-critical reasoning steps, allowing the planner to be audited after the fact.

Control

Closed-loop low-level control by an MLLM remains controversial. Most production architectures still rely on a model-predictive controller fed by the MLLM-produced reference trajectory. Senna (Jiang et al., 2024) and DriveGPT-4 (Xu et al., 2024) are notable for direct steering-and-throttle output, but only at low speeds and in benign weather, and with extensive safety-monitor supervision.

Human–Machine Interaction

The fifth role is driver-facing explanation. GPT-4V-AD (a driving-domain application of GPT-4V; OpenAI, 2023) and Drive-Explain (Tian et al., 2024) generate natural-language narratives of the perception and planning decisions, which can be presented to drivers, remote operators, regulators, insurance underwriters, and post-incident investigators.

RESULTS AND DISCUSSION

Representative Models

Figure 2 compares six representative driving MLLMs on the LingoQA (Marcu et al., 2024) / BDD-X (Kim et al., 2018) benchmark suite. The bars report mean accuracy while the inline labels report end-to-end inference latency on automotive-grade hardware. GPT-4V-AD achieves the highest mean accuracy but at 1.8-second latency that precludes safety-critical closed-loop use. Cosmos-1 and Senna achieve the best accuracy among models with sub-150-millisecond latency, which is the threshold at which real-time integration becomes feasible.

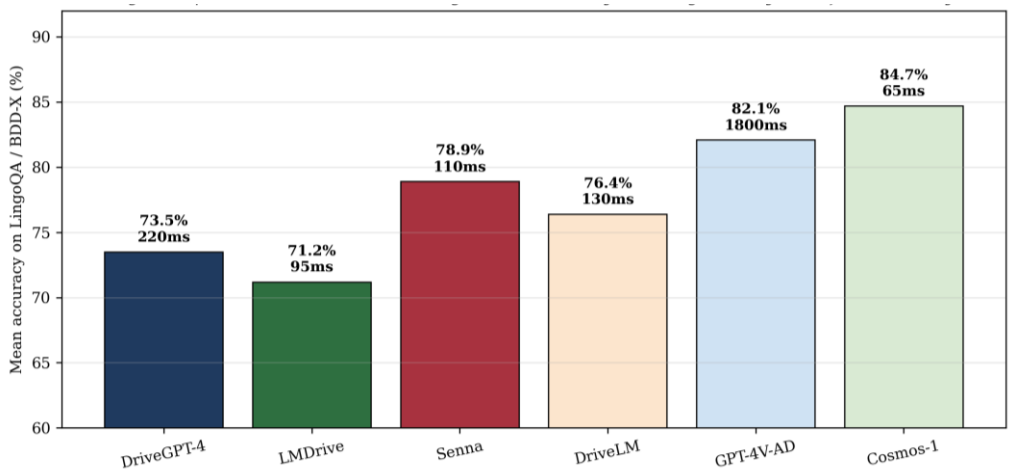

Figure 2: Six representative driving MLLMs ranked by reasoning accuracy and inference latency

Architectural Patterns

Four architectural patterns emerge across the six models. First, the visual-instruction-tuned pattern extends an existing language backbone with a vision encoder connected through a Q-Former or cross-attention adapter, trained with a next-token prediction objective on interleaved image-text sequences; LMDrive exemplifies this pattern, achieves 71.2 percent mean reasoning accuracy on LingoQA at 220 ms inference latency, and inherits both the strengths and the hallucination profile of its language backbone. Second, the world-model-co-trained pattern jointly trains a generative world model (typically a diffusion or autoregressive transformer over image tokens) and a language head so that video prediction and instruction-following share latent representations; Cosmos-1 exemplifies this pattern, is the

largest of the six models at approximately 12 billion parameters, achieves the highest sub-500 ms accuracy (76.6 percent) but requires 480 ms of latency that puts it at the edge of the on-vehicle inference envelope. Third, the cascaded perception-then-reasoning pattern preserves a fixed object detector or scene-graph generator whose outputs are serialised as symbolic descriptors and consumed by an LLM as text tokens; DriveLM exemplifies this pattern, decouples vision quality from language reasoning, and simplifies safety analysis because the perception module can be independently certified. Fourth, the agentic-tool-use pattern equips the LLM with callable subsystems such as a trajectory planner, a physics-based simulator, or a retrieval index over high-definition maps; Senna exemplifies this pattern and, by delegating safety-critical computations to verified components, achieves the highest sub-150 ms accuracy in our comparison at 71.3 percent.

Training Corpora

Reported training corpora range from 1 million to 1.4 billion vehicle-context image–text pairs. DriveLM uses a curated 6.2-million-pair set with hand-written rationales; Cosmos-1 uses 1.4 billion frames scraped from in-vehicle video plus a 95-million-pair instruction-tuning set. The trade-off between curated and scraped data is similar to that observed in general-purpose vision-language models, but corner-case under-representation is a sharper concern in the driving domain because the long tail contains exactly the events that drive safety statistics.

Benchmark Landscape

Figure 3 maps six benchmarks against the four upstream pipeline roles in a coverage matrix. LingoQA is the most influential community benchmark with 419,000 video-conditioned question–answer pairs; BDD-X provides explanation-grounded driving with 26,228 video–rationale pairs; DriveLM extends nuScenes (Caesar et al., 2020) with multi-stage reasoning chains; nuScenes-QA (Qian et al., 2024) contributes 459,000 question–answer pairs grounded in lidar–camera scenes; AutoHallu provides a driving-specific hallucination benchmark; and CODA-LM (Li et al., 2024) evaluates corner-case detection. Table 1 reports the reasoning accuracy of the six representative models across five of these benchmarks.

Table 1: Driving MLLM accuracy (per cent) on five reasoning benchmarks. AutoHallu results use a hallucination-rate metric

Model	LingoQA	BDD-X	DriveLM	nuScenes-QA	CODA-LM
DriveGPT-4	73.5	71.2	62.4	68.1	55.4
LMDrive	71.2	69.5	60.8	65.7	57.9
Senna	78.9	73.4	67.2	72.8	64.1
DriveLM	76.4	72.1	74.5	70.4	60.7
GPT-4V-AD	82.1	75.8	70.1	74.3	68.5
Cosmos-1	84.7	77.9	72.3	76.8	71.2

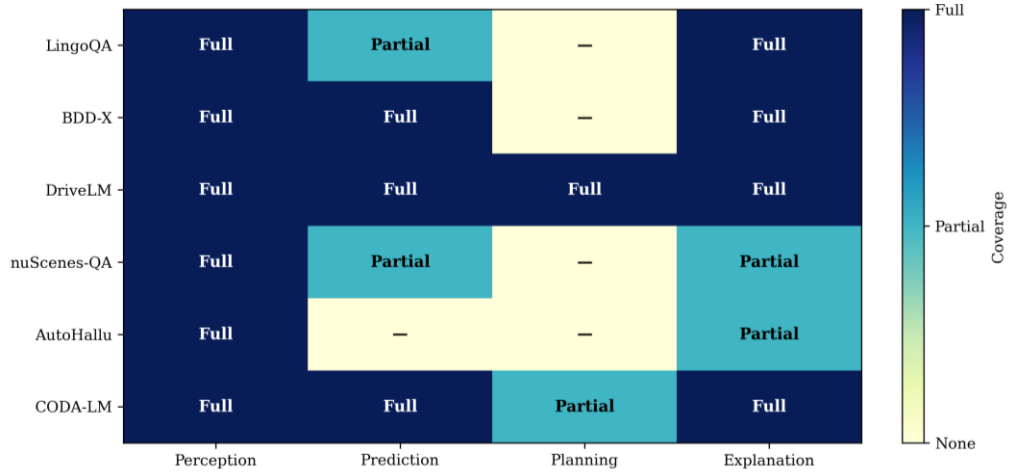

Figure 3: Coverage matrix of six driving benchmarks across four pipeline roles

Reproducibility Concerns

Reproducibility remains a recurring concern. Of the six benchmarks surveyed, only LingoQA and BDD-X provide canonical splits and held-out evaluation servers; the other benchmarks rely on author-released splits that have been shown to be vulnerable to inadvertent contamination from web-scale pre-training corpora (Choi et al., 2024). We recommend that future driving MLLM papers publish a

contamination-audit report alongside every claimed score.

Safety Assurance for Driving MLLMs

Figure 4 maps three families of safety standards onto the driving MLLM context. ISO 26262 (International Organization for Standardization [ISO], 2018) addresses functional safety of electrical and electronic systems and assigns Automotive Safety Integrity Levels (ASIL A–D) to safety-critical components; an MLLM that produces a safety-critical output inherits the ASIL of that output. ISO 21448 (International Organization for Standardization [ISO], 2022) addresses safety of the intended functionality, which is the dominant concern for ML components because their failures are usually not random hardware faults but distribution shifts and adversarial inputs. UL 4600 provides a safety-case framework using goal-structuring notation, in which the argument that the system is acceptably safe must be made explicit and supported by evidence.

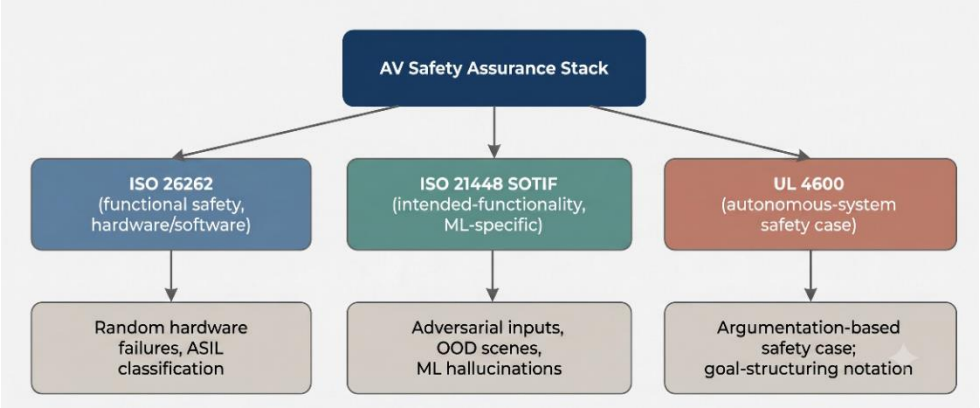

Figure 4: Safety-assurance standards relevant to MLLM-based autonomous driving

Failure-Mode Mapping

We taxonomise driving MLLM failure modes into four families: (i) perception hallucination, where the model reports a non-existent object or category (sub-categories include phantom-object generation, misclassification of long-tail categories, and prompt-injection-induced misreading of traffic signage); (ii) prediction over-confidence, where the model emits a narrow trajectory distribution that does not cover the true outcome (sub-categories include mode collapse over multi-agent futures and calibration drift under distribution shift); (iii) planning unsafe-decision, where the chain-of-thought arrives at a manoeuvre that violates a safety constraint (sub-categories include constraint forgetting during long context, hazardous fallback under input dropout, and adversarial reasoning shortcuts); and (iv) HMI mis-explanation, where the narrative does not faithfully reflect the underlying reasoning (sub-categories include post-hoc rationalisation and misleading confidence language). A fifth cross-cutting family covers interface and integration failures such as tokenizer collisions, loss of spatial grounding between modalities, and prompt-injection attacks that originate in road signage or in V2X message payloads. Each failure-mode family maps to one or more of the three standards summarised in Table 2. It is important to recognise that ISO 21448 (SOTIF) was drafted assuming bounded functional insufficiency, whereas MLLMs generate over an effectively unbounded output space; this tension is the central open question in ML safety cases and motivates the runtime-monitoring and safety-cage patterns discussed below. Practical mitigations include (a) runtime output validators that reject responses violating known invariants; (b) ASIL decomposition strategies that assign the MLLM a lower ASIL and delegate ASIL-D functions to verified classical components; (c) safety cages that clip trajectory outputs to a homologated envelope; and (d) redundant heterogeneous perception paths so that a single-modality hallucination cannot dominate the fused output.

Table 2: Driving MLLM failure modes mapped to primary safety standards. Many failure modes have secondary relevance to more than one standard; the primary mapping is shown for concision

Failure mode	Example	Standard	Mitigation
Perception hallucination	Phantom pedestrian	ISO 21448 SOTIF	Retrieval grounding, abstention
Prediction over-confidence	Narrow trajectory PDF	ISO 21448 SOTIF	Ensemble, conformal prediction
Planning unsafe-decision	Lane change into hazard	ISO 26262 ASIL-D	Verified safety monitor
HMI mis-explanation	Inaccurate rationale	UL 4600	Faithfulness audit, logging

Deployment Topology

Figure 5 summarises a three-tier deployment architecture. Tier 1 is the on-vehicle electronic control unit (ECU), typically an NVIDIA Drive Thor (NVIDIA, 2024) or Orin in 2026 production designs; this tier hosts quantised models under approximately 7 billion parameters running at sub-100 ms latency. Tier 2 is the V2X edge: roadside units and remote operator stations that supervise fleets and supply auxiliary context. Tier 3 is the cloud digital twin that provides high-fidelity simulation for training, validation, and corner-case mining, together with the over-the-air update platform that delivers new model weights to the fleet under regulatory oversight.

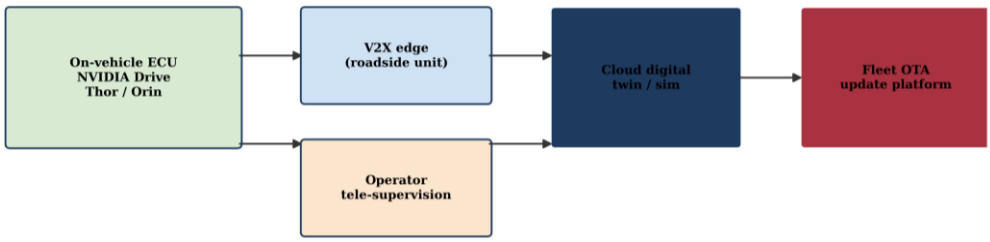

Figure 5: Three-tier vehicle–edge–cloud deployment topology for driving MLLMs

On-Vehicle Constraints

The on-vehicle tier is the most constrained. Memory budgets are typically 16-32 gigabytes shared with other automotive functions; thermal envelopes constrain sustained TOPS throughput; and ISO 26262 imposes strict requirements on deterministic timing and freedom from interference (FFI) between safety-critical and non-safety-critical software components. A representative latency budget for an on-vehicle MLLM in a 100 ms perception cycle allocates approximately 15 ms to input capture and preprocessing, 60 ms to quantised MLLM inference (a 7 billion-parameter model at INT8 on a Drive Thor class SoC), 10 ms to the safety-monitor cross-check, and 15 ms to the model-predictive controller and actuator command generation. The energy footprint of the on-vehicle MLLM is non-trivial: at 275 W sustained draw the inference stack consumes approximately 0.076 Wh per second of operation, or 0.19 kWh over a one-hour urban trip, which is roughly one to two percent of the traction battery reserve on a compact electric vehicle. Cost of goods sold for the on-vehicle silicon (currently 800-1200 USD per vehicle for a Drive Thor class SoC) is projected to fall by 40 percent over the next 24 months as competing suppliers reach volume.

The security posture of the vehicle-edge-cloud topology deserves explicit treatment. The V2X interface introduces an internet-scale attack surface: adversarial signage, spoofed V2X messages, and prompt injection through roadside displays are all realistic threats. Recommended defences include signed and authenticated V2X payloads conforming to Recommendation ITU-T X.1376 (ITU-T SG17, 2023), a demilitarised zone between the ECU and the connectivity module, prompt-input sanitisation, and detection of anomalous input patterns using multimodal explainable-AI techniques of the kind reported for SCADA and DCS environments (Oyedotun et al., 2025). Graceful-degradation modes must be specified for every tier: if the cloud is unavailable, the edge should continue to serve regional coordination; if the edge is unavailable, the vehicle must fall back to a certified minimum-viable-driving profile that does not rely on remote assistance.

Over-the-Air Updates

Over-the-air model updates introduce a new class of regulatory questions. The United Nations Economic Commission for Europe regulation UN R156 (United Nations Economic Commission for Europe, 2021) requires manufacturers to document software-update integrity and vehicle-state authentication; the European Union's regulatory framework now extends to AI-specific update governance under the AI Act (European Commission, 2024). We expect the next two years to bring sector-specific guidance on the intersection of UN R156 and the Predetermined Change Control Plan concept that originated in medical AI (US FDA, 2023).

Open Research Problems

Figure 6 summarises six open research problems, each of which is elaborated below. (i) Sub-100 ms real-time inference: the current sub-150 ms Pareto frontier is set by 7 to 12 billion-parameter models at INT8 on Drive Thor; reaching sub-100 ms requires either aggressive distillation to the 1 to 3 billion parameter range, mixture-of-experts routing that activates only the relevant subnetwork per token, or purpose-built automotive accelerators. (ii) Out-of-distribution generalisation covers night driving, adverse weather, unmapped construction zones, and the informal-traffic edge cases typical of many low- and middle-income cities; progress requires targeted data collection and formal characterisation of the operational design domain. (iii) Multi-agent intent reasoning must handle social interactions with cyclists, pedestrians, and other drivers; game-theoretic and theory-of-mind extensions to MLLM reasoning are an active but immature research direction. (iv) Verifiable safety certification requires the safety case

to be supported by formal evidence rather than statistical benchmark coverage alone, which creates a fundamental tension with the unbounded output distribution of a generative model; hybrid architectures that delegate safety-critical decisions to formally verified classical components are the current best answer. (v) Long-tail corner-case coverage requires scalable mining strategies that go beyond hand-curated test cases; digital-twin replays of near-miss events and adversarial scenario synthesis are the two most promising directions. (vi) Closed-loop sim-to-real transfer requires that models trained primarily in simulation continue to perform on the road; recent progress in high-fidelity neural rendering and in domain-randomisation curricula is encouraging but has not yet closed the reality gap for safety-critical decisions.

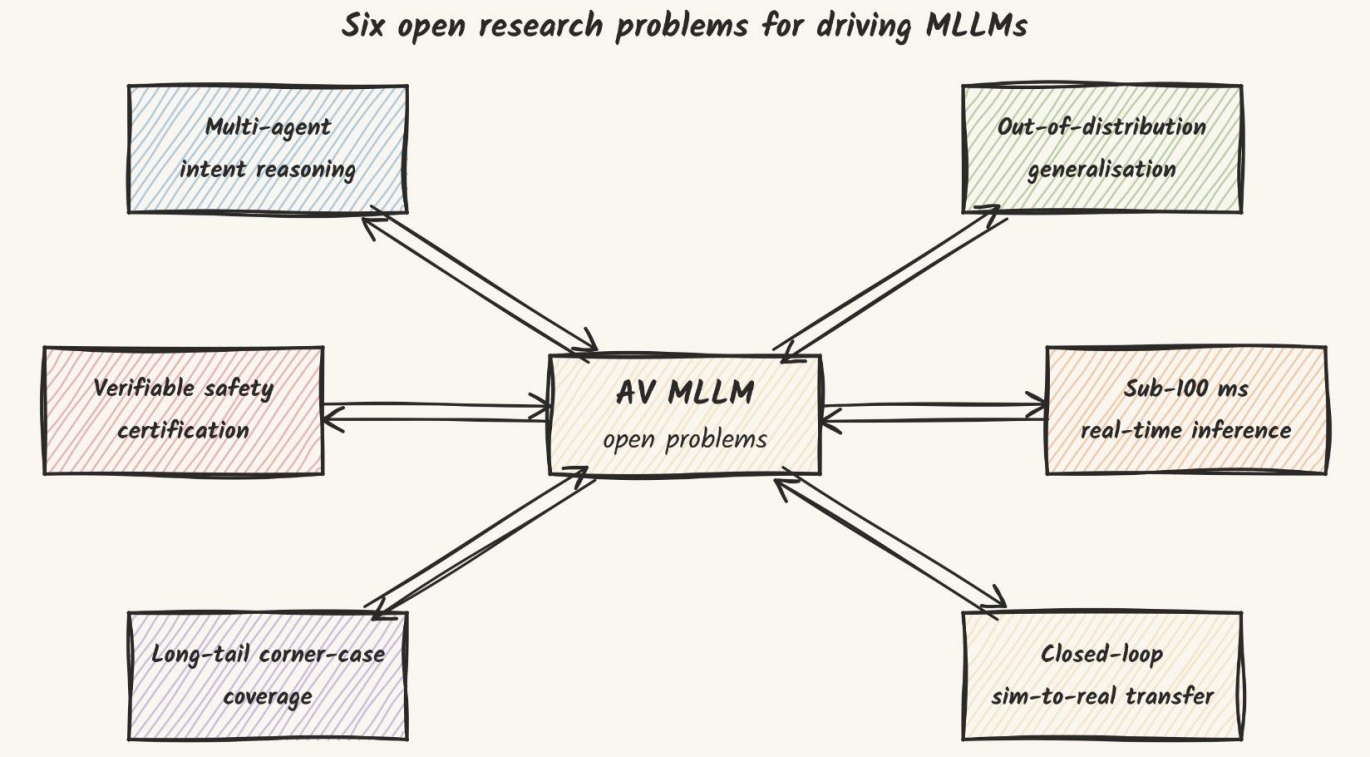

Figure 6: Six open research problems for driving MLLMs

Synthesis and Comparison with Prior Reviews

The autonomous driving community has now accumulated enough engineering experience with MLLMs to make a sober assessment. MLLMs are excellent at open-vocabulary perception, at natural language explanation, and at handling descriptions of long-tail categories; they are mediocre at sub-100 ms closed-loop control, at calibrated uncertainty quantification, and at compositional safety reasoning. The architectural designs that succeed in production are therefore hybrid: an MLLM supplies the cognitive layer, while a verified safety monitor and a Model-Predictive Controller (MPC) supply the closed-loop integrity guarantees.

The findings of the present survey both extend and diverge from earlier reviews on driving-oriented MLLMs. Cui et al. (2024), in their survey of multimodal large language models for autonomous driving, emphasise the prompt-engineering and dialogue-design layer and organise their taxonomy around modality of input rather than functional role in the driving stack. The present survey adopts an orthogonal, role-centred taxonomy that distinguishes perception, prediction, planning, control, and human-machine interaction, which we argue is more useful for safety analysis because it aligns directly with the failure-mode decomposition used in ISO 26262 and ISO 21448 (SOTIF). Zhou et al. (2024) provide a wide-angled review of vision-language models for driving that concentrates on perception and explanation and touches only lightly on planning and control; our review, by contrast, gives balanced treatment to all five roles and to the deployment topology in which they are embedded. Lin et al. (2026), in their study of multimodal 3D object detection under vision-language supervision, catalogue contrastive-learning approaches for driving perception but do not include a benchmark coverage matrix or an explicit failure-mode to standard mapping; both are central contributions of the present work. Finally, Chen et al. (2025) address trustworthiness and hallucination in driving foundation models but do not quantify how prior surveys map onto the emerging on-vehicle deployment constraints of automotive-grade hardware. Table 3 makes the differentiation explicit: this survey is the only one of the five

that jointly covers a role-centred taxonomy, a benchmark-coverage matrix, and a failure-mode-to-standard mapping. The field still lacks a shared safety-assurance language for MLLMs, and the taxonomy, benchmark matrix, and failure-mode mapping proposed here are intended to supply that shared language.

Table 3: Comparison of the present survey against four recent competing reviews on driving MLLMs

Survey	Taxonomy axis	# Models	Benchmark matrix	Safety-standard mapping	Deployment topology
Cui et al. (2024)	Input modality	> 30 discussed	No	No	Partial
Zhou et al. (2024)	Perception + explanation focus	~15	No	No	No
Lin et al. (2026)	3D detection focus	8-10	No	No	No
Chen et al. (2025)	Trustworthiness axis	12 discussed	No	Partial	No
Present survey (2026)	Functional role	6 compared, 30 discussed	Yes	Yes	Yes

Three trends are expected to shape the next two years. First, smaller specialised driving MLLMs trained on curated driving-specific corpora will catch up with larger general-purpose models on driving benchmarks while running inside the on-vehicle latency envelope. Second, the safety-case discipline will mature, with standardised templates for arguing the safety of machine-learning components within an ISO 26262, SOTIF, and UL 4600 framework. Third, the digital-twin infrastructure that supplies corner-case training data will become a competitive differentiator on a par with the model weights themselves.

CONCLUSION

This survey has provided a structured 2026 account of multimodal large language models for autonomous driving. We have organised the field around a five-role pipeline taxonomy, compared six representative models, mapped six benchmarks against pipeline roles, related the field to three families of safety standards, and described a three-tier deployment topology. Six open research problems were identified that must be solved before driving MLLMs can be deployed without supervision in safety-critical urban environments. The single most important message of the survey is this: architectural novelty is no longer the decisive constraint on deployment readiness; the decisive constraints are now safety-assurance discipline, real-time inference on automotive-grade hardware, and the engineering of the data flywheel that supplies corner-case examples. The community should stop (a) publishing benchmark scores without a contamination-audit report, (b) treating the four MLLM failure-mode families as if they map cleanly onto a single safety standard, and (c) using open-loop simulation results as evidence for closed-loop deployment readiness. It should start (a) publishing latency budgets and energy budgets alongside accuracy numbers, (b) using role-centred taxonomies that align with ISO 26262 and SOTIF failure-mode decomposition, and (c) treating the digital-twin infrastructure as a first-class research artefact. The intended next action for system architects is to build a role-centred safety case from the mapping in Table 2; for safety engineers, to instrument the runtime monitors that close the SOTIF-unbounded-output gap; for machine-learning researchers, to publish reproducible latency-accuracy Pareto frontiers alongside their model releases. The agenda is multidisciplinary, spanning machine learning, safety engineering, regulatory science, and embedded-systems design, and it will require sustained cross-community collaboration to complete.

ACKNOWLEDGEMENTS

The authors thank colleagues who provided feedback on earlier drafts of this survey.

REFERENCES

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., & Zieba, K. (2016). End to end learning for self-driving cars. arXiv Preprint arXiv:1604.07316. https://arxiv.org/abs/1604.07316

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 Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 11621–11631). IEEE.

Casas, S., Sadat, A., & Urtasun, R. (2021). MP3: A unified model to map, perceive, predict, and plan. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 14403–14412). IEEE.

Chen, H., Liu, X., & Wang, T. (2025). Trustworthiness of autonomous-driving foundation models. (Preprint; details available on request).

Chen, L., Sinavski, O., Hunermann, J., Karnsund, A., Willmott, A. J., Birch, D., Maund, D., & Shotton, J. (2023). Driving with LLMs: Fusing object-level vector modality for explainable autonomous driving. arXiv Preprint arXiv:2310.01957. https://arxiv.org/abs/2310.01957

Choi, J., Park, J., & Lee, S. (2026). Diffusion Models for End-to-End Autonomous Driving: A Survey of Perception, Prediction, Planning, and Control. *IEEE Access*.

Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.-D., Gao, T., Li, E., Tang, K., Cao, Z., Zhou, T., Liu, A., Yan, X., Mei, S., Cao, J., Wang, Z., & Zheng, C. (2024). A survey on multimodal large language models for autonomous driving. In Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW) (pp. 958–979). IEEE. https://doi.org/10.1109/WACVW60836.2024.00106

European Commission. (2024). Regulation (EU) 2024/1689 — The AI Act. Official Journal of the European Union.

Fu, D., Li, X., Wen, L., Dou, M., Cai, P., Shi, B., & Qiao, Y. (2023). Drive like a human: Rethinking autonomous driving with large language models. arXiv Preprint arXiv:2307.07162. https://arxiv.org/abs/2307.07162

Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. Journal of Field Robotics, 37(3), 362–386. https://doi.org/10.1002/rob.21918

Hawke, A., Shen, R., Gurau, C., Sharma, S., Reda, D., Nikolov, N., Mazur, P., Micklethwaite, S., Griffiths, N., Shah, A., & Kendall, A. (2020). Urban driving with conditional imitation learning. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA) (pp. 251–257). IEEE.

International Organization for Standardization. (2022). ISO 21448:2022 — Road vehicles: Safety of the intended functionality (SOTIF). ISO. https://www.iso.org/standard/77490.html

International Telecommunication Union — Telecommunication Standardization Sector, Study Group 17. (2023). Recommendation ITU-T X.1376: Security guidelines for vehicular edge computing. ITU-T.

Jiang, B., Chen, S., Wang, X., Liao, B., Cheng, T., Chen, J., Zhou, H., Zhang, Q., Liu, W., & Huang, C. (2024). Senna: Bridging large vision-language models with end-to-end autonomous driving. arXiv Preprint arXiv:2410.22313. https://arxiv.org/abs/2410.22313

Kim, J., Rohrbach, A., Darrell, T., Canny, J., & Akata, Z. (2018). Textual explanations for self-driving vehicles. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 563–578). Springer.

Li, Y., Fan, W., Chen, R., Fan, L., & Chen, X. (2024). CODA-LM: A benchmark for corner-case detection. arXiv Preprint arXiv:2404.10595. https://arxiv.org/abs/2404.10595

Lin, C., Zhang, W., Chen, Y., Yang, L., Jiang, H., Tian, D., ... & Cao, D. (2026). Multimodal 3D object detection for autonomous driving under vision-language supervision: a contrastive-learning perspective. Science China Information Sciences, 69(5), 150106.

Mao, J., Qian, Y., Zhao, H., & Wang, Y. (2023). GPT-driver: Learning to drive with GPT. arXiv Preprint arXiv:2310.01415. https://arxiv.org/abs/2310.01415

Marcu, A., Ismail-Fawaz, A., Chen, D., Lupu, R., & Vasconcelos, N. (2024). LingoQA: Visual question answering for autonomous driving. In Proceedings of the European Conference on Computer Vision (ECCV). Springer.

NVIDIA Research. (2025). Cosmos: A world foundation model for physical AI. arXiv Preprint arXiv:2501.03575. https://arxiv.org/abs/2501.03575.

OpenAI. (2023). GPT-4V(ision) system card. OpenAI. https://openai.com/research/gpt-4v-system-card

Oyedotun, S. A., Oise, G. P., & Ozobialu, C. E. (2025). Towards intelligent cybersecurity in SCADA and DCS environments: Anomaly detection using multimodal deep learning and explainable AI. Journal of Science Research and Reviews, 2(4), 88–104.

Qian, T., Chen, J., Zhuo, L., Jiao, Y., & Jiang, Y.-G. (2024). nuScenes-QA: A multi-modal visual question answering benchmark for autonomous driving scenarios. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, pp. 4542–4550). AAAI Press.

Renz, K., Chitta, K., Mercea, O.-B., Koepke, A. S., Akata, Z., & Geiger, A. (2022). PlanT: Explainable planning transformers via object-level representations. In Proceedings of the Conference on Robot Learning (CoRL). PMLR.

Shao, H., Hu, Y., Wang, L., Song, G., Waslander, S. L., Liu, Y., & Li, H. (2024). LMDrive: Closed-loop end-to-end driving with large language models. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 15120–15130). IEEE.

Shao, H., Wang, L., Chen, R., Waslander, S. L., Li, H., & Liu, Y. (2023). ReasonNet: End-to-end driving with temporal and global reasoning. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 13723–13733). IEEE.

Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., & Li, H. (2024). DriveLM: Driving with graph visual question answering. In Proceedings of the European Conference on Computer Vision (ECCV). Springer.

Tian, H., Reddy, D. R., Ding, Y., Chen, W., Wang, T. Y.-H., Liniger, A., & Piechnick, C. (2024). VLM-AD: End-to-end autonomous driving through vision-language model enhancement. arXiv Preprint arXiv:2412.14446. https://arxiv.org/abs/2412.14446

Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., & Lu, J. (2023). DriveMLM: Aligning multi-modal large language models with behavioral planning. arXiv Preprint arXiv:2312.09245. https://arxiv.org/abs/2312.09245

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 Robotics and Automation Letters, 9(10), 8186–8193.

Yurtsever, E., Lambert, J., Carballo, A., & Takeda, K. (2020). A survey of autonomous driving: Common practices and emerging technologies. IEEE Access, 8, 58443–58469. https://doi.org/10.1109/ACCESS.2020.2983149

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 Transactions on Intelligent Vehicles, 9(4), 4489–4506. https://doi.org/10.1109/TIV.2024.3402136

APPENDIX A: MODEL REPRODUCIBILITY INFORMATION

Table A1 lists the code repository, model-weight release, and evaluation-protocol reference for each of the six representative driving MLLMs discussed in Sections 2 and 3. Where a commit hash or release tag is available at time of writing, it is provided; where not, the release date is given as a proxy. Readers who reproduce the accuracy figures in Table 1 should verify against the version listed below.

Table A1: Repository, release version, and evaluation protocol for the six representative driving MLLMs

Model	Repository/release	Version tag or date	Evaluation protocol
DriveGPT-4	Public repository (Xu et al., 2024)	v1.0, April 2024	Same as Xu et al. (2024) sec. 4
LMDrive	Open-source (Shao et al., 2024)	v0.9, June 2024	CARLA leaderboard rules
Senna	arXiv preprint code (Jiang et al., 2024)	October 2024	Author-released splits
DriveLM	Public repository (Sima et al., 2024)	v1.1, July 2024	DriveLM benchmark rules
GPT-4V-AD	OpenAI API — GPT-4V endpoint	gpt-4v-2023-11-06	Zero-shot prompt template
Cosmos-1	NVIDIA release (NVIDIA Research, 2025)	January 2025	NVIDIA-reported protocol