Scilight

Transactions on Artificial Intelligence
https://www.sciltp.com/journals/tai

Survey

Recent Advances in Autonomous Driving Safety

Shuguang Wang [1,*], Hongzong Li [2] and Guanyi Zhao [1]

[1] Department of Computer Science, City University of Hong Kong, Hong Kong
[2] Hong Kong Generative AI R&D Center, Hong Kong University of Science and Technology, Hong Kong
* Correspondence: sgwang6-c@my.cityu.edu.hk

How To Cite: Wang, S.; Li, H.; Zhao, G. Recent Advances in Autonomous Driving Safety. *Transactions on Artificial Intelligence* **2026**, *2*(1), 161–177. https://doi.org/10.53941/tai.2026.100010

Received: 12 April 2026
Revised: 24 April 2026
Accepted: 9 May 2026
Published: 21 May 2026

Abstract: This paper reviews recent advancements in autonomous driving safety, focusing on the evolution of autonomous driving systems from modular pipelines to end-to-end (E2E) frameworks and emerging vision-language-action (VLA) models. For modular systems, this paper analyzes how to mitigate error propagation between decoupled modules using multi-sensor redundancy and formal verification. For end-to-end systems, this paper delves into learning-based motion planning. It emphasizes safety innovations to address the lack of transparency in deep learning, such as interpretable cost maps and world-model-based simulations. For VLA models, this paper investigates integrating vision language models (VLMs) to enhance high-level semantic reasoning and understanding of long-tail driving scenarios. It discusses safety guardrail technologies, such as chain of thought (CoT) reasoning, to ensure that logic aligns with driving regulations. Finally, this paper summarizes current challenges and outlines future research directions, providing a systematic reference for building safe and reliable autonomous driving systems.

Keywords: autonomous driving; safety validation; end-to-end; vision-language-action (VLA)

1. Introduction

Autonomous driving (AD) technology provides new mobility solutions through intelligent perception and decision-making, helping to improve traffic efficiency [1]. However, the deployment of AD systems is fundamentally constrained by high safety requirements. Unlike general AI applications, a single failure in autonomous driving systems can lead to catastrophic consequences, such as loss of life and significant property damage [2]. Ensuring safety is no longer just a functional requirement, but the core prerequisite for public trust and regulatory approval [3,4].

The technological landscape of AD has experienced a significant transformation. Traditional *modular systems* decomposed the driving task into several components: perception, prediction, and planning [5]. While this architecture offers interpretability and formal safety guarantees through rule-based frameworks such as Responsibility-Sensitive Safety (RSS) [2], it suffers from cascading errors where small upstream miscalculations lead to downstream failures [6,7]. To tackle these limitations, *End-to-end (E2E)* autonomous driving systems have gained significant focus [8]. By optimizing the entire pipeline from raw pixels to control commands, E2E systems such as UniAD [9] significantly improved driving generalization. However, this integration introduced the "black-box" [10,11], making it nearly impossible to audit the internal logic of a vehicle's decision in safety-critical driving scenarios. The phenomenon of *causal confusion*, wherein learned policies associate control actions with spurious features, further undermines the reliability of E2E approaches [12–14].

The latest frontier is the integration of *Vision-Language-Action (VLA)* models, which leverage the reasoning power of Language Models [15–17]. VLA models promise a breakthrough in handling complex scenarios. However, they introduce safety risks such as hallucinations where the model generates logically plausible but physically impossible or dangerous trajectories [18]. Current research is thus pivoting toward hybrid safety architectures, focusing on safety-alignment through reinforcement learning from human feedback, interpretability through

Chain-of-Thought (CoT) reasoning [19], and the development of high-fidelity world models to simulate risky environments [20,21].

The main contributions of this work can be summarized as follows:

(1) We provide a comprehensive taxonomy of autonomous driving safety across three architectures: modular, end-to-end, and VLA-based.
(2) We analyze the unique safety vulnerabilities inherent in each architecture, tracing the evolution from sensor-level uncertainty to high-level cognitive hallucinations.
(3) We categorize state-of-the-art safety enhancement methodologies, including formal verification and LLM-based safety guardrails.
(4) We identify key open challenges and propose future directions for developing safe foundation models for autonomous driving.

The rest of the paper is organized as follows: Section 2 reviews the architectural evolution and establishes a failure taxonomy. Section 3 analyzes the safety mechanisms and limitations of classical modular pipelines. Section 4 explores the transition to end-to-end learning and its challenges. Section 5 delves into the cutting-edge VLA architectures, focusing on semantic safety and reasoning alignment. Section 6 concludes with future research directions.

2. Autonomous Driving Architectures

There are currently three main framework designs for autonomous driving (AD) systems. Each paradigm redefines the boundary between human-designed prior knowledge and data-driven learning, thereby introducing distinct safety features. Figure 1 illustrates the high-level dataflow of each paradigm.

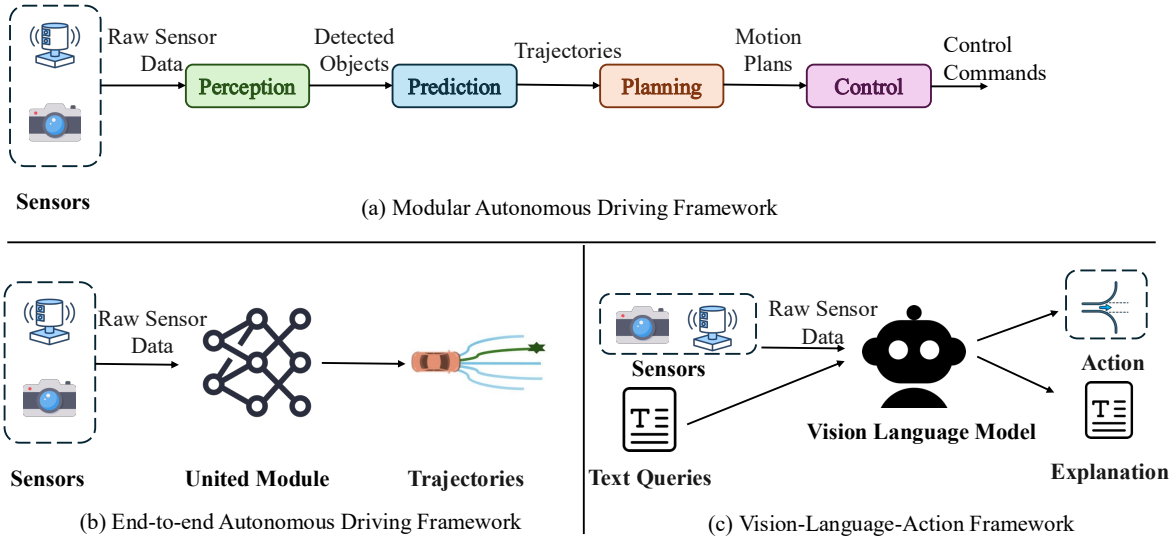

Figure 1. Comparison of the designs of the autonomous driving frameworks. (**a**) The modular framework deploys separate modules for different tasks; (**b**) The end-to-end framework unites modules mapping raw sensor data directly to planned trajectories; (**c**) The VLA leverages foundation models to produce driving actions and language explanations.

2.1. Modular Paradigm

The classical modular framework decomposes the driving task into a serial pipeline of functional modules: *perception* (object detection, semantic segmentation, and lane estimation), *prediction* (trajectory forecasting of surrounding agents), *planning* (route and motion planning), and *control* (lateral and longitudinal actuation). Each module exposes a well-defined interface, usually in the form of structured representations such as 3-D bounding boxes or occupancy grids, to its downstream consumer.

This architecture has been the backbone of major open-source stacks and industrial deployments. Apollo [5] and Autoware [22] adopt this paradigm, offering modular libraries where each component can be independently upgraded or validated. On the perception front, the evolution from PointPillars [23] to BEVFusion [24] demonstrates a clear trend toward multi-modal bird's-eye-view (BEV) representations that unify camera and LiDAR features in a shared spatial coordinate system. For prediction and planning, recent work such as GameFormer [25] formulates multi-agent interaction as a game-theoretic process. Planscope [26] introduces a decision scope framework that leverages wavelet transformations and a Transformer-based decoder to filter unpredictable noise in motion planning.

The principal advantage of the modular paradigm is *interpretability*: each module produces intermediate outputs that can be independently audited against formal safety specifications. The Responsibility-Sensitive Safety (RSS) framework [2], for example, defines mathematically verifiable safe distances that can be directly computed from the structured outputs of the prediction module. However, this architecture suffers from two well-documented limitations. First, the hand-designed interfaces between modules act as *information bottlenecks*, discarding fine-grained sensory details (e.g., subtle texture cues indicating road surface conditions) that may be safety-relevant but are not captured by the predefined representation schema [6]. Second, errors introduced by an upstream module propagate and *compound* through the pipeline: a missed detection in perception leads to an absent trajectory in prediction, which in turn causes the planner to produce a collision trajectory [6].

2.2. End-to-end Paradigm

End-to-end (E2E) architectures address the information bottleneck by replacing the manually designed inter-module interfaces with learnable, differentiable connections, thereby enabling gradient-based optimization of the entire pipeline toward a unified driving objective. The work of Bojarski et al. [8] demonstrated that a single convolutional network could learn a direct mapping from front-facing camera images to steering commands via behavioral cloning. Subsequent work has significantly scaled both the input modality and the output representation.

A recent development in this paradigm is UniAD [9], which retains interpretable intermediate tasks (perception, prediction and motion planning) within a single Transformer-based architecture, but optimizes all tasks jointly through a shared query-based representation. VAD [27] further simplifies the pipeline by representing the driving scene as vectorized agents and map features, achieving state-of-the-art planning performance with reduced computational overhead. More recently, GenAD [28] frames planning as a generative modeling problem using a variational autoencoder on latent trajectory distributions, enabling the model to capture multi-modal future trajectories. SparseDrive [29] pushes efficiency further by employing a fully sparse architecture that eliminates dense BEV feature maps while maintaining competitive safety metrics.

Despite these advances, E2E still face numerous challenges in security verification. The lack of explicit intermediate representations makes it difficult to pinpoint the specific cause of planning failures [10]. Moreover, Codevilla et al. [13] and De et al. [12] identified the phenomenon of *causal confusion*, wherein E2E models learn spurious correlations (e.g., associating the brake light of a leading vehicle with deceleration rather than the actual traffic signal state), leading to catastrophic failures in novel environments. The out-of-distribution (OOD) generalization of E2E models remains an open concern, as demonstrated by the significant performance degradation observed when models trained on nuScenes are evaluated on geographically distinct datasets [30].

2.3. Vision-Language-Action Paradigm

The most recent paradigm integrates Vision-Language-Action (VLA) models into the driving stack. These architectures leverage the pre-trained world knowledge and reasoning capabilities of foundation models to address scenarios that require semantic understanding beyond geometric pattern matching.

Early explorations such as DriveGPT4 [31] and GPT-Driver [15] employed LLMs as high-level reasoning engines that convert visual observations into textual scene descriptions and then generate planning decisions through language-based inference. DriveLM [17] introduced a graph-structured visual question answering paradigm that decomposes the driving decision into a chain of perception, prediction, and planning sub-questions, each answered by a VLM and linked by logical dependencies. DriveVLM [16] proposed a dual-system architecture where a VLM handles long-horizon scene understanding and a conventional spatial planner refines the trajectory under kinematic constraints.

More tightly integrated VLA architectures have emerged subsequently. Senna [19] decouples high-level planning reasoning and low-level trajectory prediction into an VLM-based planner and a conventional E2E action planner. This work demonstrates that VLM-generated reasoning can serve as an effective planning prior. LeapAD [32] draws on cognitive science to implement a dual-process model: a fast, experience-driven "System 1" for routine driving and a slow, deliberate "System 2" powered by an VLM for novel situations. Emma [33] from Waymo unifies multiple driving tasks within a single multimodal language model by formulating all outputs as language tokens in a shared vocabulary.

While VLA models demonstrate remarkable improvements on long-tail scenarios, they introduce safety risks that are different from those of prior paradigms. The most prominent risk is *hallucination*: the model may generate linguistically coherent but physically infeasible or environmentally inconsistent outputs [18]. For instance, a VLA model might reason that "the pedestrian has crossed the street" while the pedestrian remains occluded and stationary. Furthermore, the autoregressive token generation process in VLMs introduces non-negligible latency, posing challenges for real-time tasks [16]. The alignment between the model's learned representation of "safety" and the physical constraints of vehicle dynamics is an active area of research.

2.4. Taxonomy of Safety-Critical Failures in Autonomous Driving

To enable a systematic analysis of challenges across the three architectural paradigms, we propose a hierarchical taxonomy of safety-critical failures organized into three layers: perception and representation failures, decision and planning failures, and cognitive and logic failures. Each layer corresponds to an increasingly abstract stage of the information processing pipeline, and while all three paradigms are susceptible to failures at each layer, the *dominant failure mode* shifts upward as the architecture evolves from modular to VLA. A summary of the taxonomy is presented in Table 1.

Table 1. Taxonomy of safety-critical failures in autonomous driving. For each failure layer we list the sub-categories, a mechanistic description, the architectural paradigms that are affected (**M** = Modular, **E** = End-to-End, **V** = VLA), and representative references.

Failure Layer	Sub-Category	Description	M	E	V	Key References
Perception & Representation	Detection Error	Missed/false detections due to occlusions or adverse weather.	✓	✓	✓	[34,35]
	Temporal Instability	Intermittent detections across frames causing track fragmentation and jerky control.	✓	✓	✓	[36]
	Domain Shift	Performance drop when deployment conditions (weather/lighting) differ from training.	✓	✓	✓	[37]
	Adversarial Attack	Misclassification caused by physical perturbations like patches or modified road signs.	✓	✓	✓	[38,39]
Decision & Planning	Causal Confusion	Misattributing control to spurious features (e.g., UI elements) instead of road logic.		✓	✓	[12,13]
	Frozen Robot	Excessive caution in dense traffic leading to deadlocks or secondary risks.	✓	✓		[40,41]
	Out-of-Distribution	Failure to handle rare, aggressive, or culturally specific agent behaviors.	✓	✓	✓	[11,42]
Cognitive & Logic	Semantic Gap	High-level reasoning fails to map to feasible physical actions.			✓	[17]
	Hallucination	Generative models creating plausible but non-existent driving states (e.g., phantom clear paths).		✓	✓	[43]
	Reasoning Drift	Chain-of-thought logic diverging from sensory input or action output under latency.			✓	[19,44]

2.4.1. Perception and Representation Failures

Perception failures occur when the system constructs an inaccurate or incomplete representation of the surrounding environment. We identify four principal sub-categories.

Detection error. Small, rare, or partially occluded objects remain a persistent challenge. Peri et al. [34] showed that state-of-the-art 3-D detectors miss up to 40% of pedestrians at distances beyond 50 m under adverse weather conditions. Conversely, sensor noise and reflections can generate phantom objects that trigger unnecessary emergency braking. BEVFusion [24] mitigates some of these issues through camera-LiDAR feature alignment in BEV space, but degradation under sensor dropout (e.g., LiDAR failure in heavy rain) remains significant [35].

Temporal inconsistency. Reliable tracking requires temporally consistent detections across frames. Flicker detection occurs when an object is detected in frame t, missed in frame $t + 1$, and then re-detected in frame $t + 2$. This leads to unstable trajectory prediction and oscillating planning behavior. The Perception benchmark [45] has pointed out the gap between per-frame detection accuracy and tracking-level consistency as a critical safety metric.

Domain shift and environmental degradation. Models trained on data from clear-weather, daytime conditions in specific geographic regions exhibit significant performance drops when deployed in novel domains. Sakaridis et al. [37] quantified this degradation systematically across fog, rain, night, and snow conditions. Despite advances in domain-adaptive perception, the long tail of environmental conditions remains unresolved.

Adversarial vulnerability. Adversarial attacks in the real world pose a real threat to the perception module. Cao et al. [38] demonstrated that strategically placed adversarial patches can cause LiDAR-based detectors to completely ignore stop signs. The transferability of such attacks across different detector architectures further amplifies the risk [39].

2.4.2. Decision and Planning Failures

Even when a system maintains an accurate perception of the environment, its planning module may produce unsafe trajectories due to algorithmic limitations or distributional gaps. We identify three principal sub-categories.

Causal confusion. In imitation learning and end-to-end architectures, models may learn to associate control actions with spurious features. These spurious features are only related to correct driving behavior, but do not have a causal relationship. De Haan et al. [12] provided a formal treatment of this phenomenon and showed that standard behavioral cloning is particularly susceptible. For example, a model trained on data where the ego vehicle consistently decelerates when a specific dashboard indicator is visible may learn to associate braking with the indicator rather than with the preceding vehicle's deceleration. Recent counterfactual data augmentation techniques [12] have shown promise in mitigating this issue, but a principled, general solution remains elusive.

Frozen robot problem. Safety-oriented planners impose high costs on any trajectory that approaches other vehicles, which can be overly conservative in congested traffic. If a trajectory that simultaneously satisfies all safety constraints cannot be found, vehicles may be "frozen" indefinitely. This overly conservative strategy goes against the expectations of surrounding drivers and could potentially cause rear-end collisions [40]. Recent work by Cheng et al. [41] addresses this through a contrastive learning framework that jointly optimizes for safety and progress.

Distributional mismatch in interactive scenarios. The behavior of surrounding agents in the real world often deviates significantly from the distributions captured in training data. Aggressive lane changes, jaywalking pedestrians, and non-standard vehicle maneuvers (e.g., U-turns at undesignated locations) constitute distributional tail events that are underrepresented in standard datasets such as nuScenes [46] and Waymo Open Dataset [45]. Suo et al. [42] and Zhong et al. [47] proposed simulation-based approaches to synthesize such adversarial interactions, but the sim-to-real gap in agent behavior modeling persists.

2.4.3. Cognitive and Logic Failures

Unique to architectures incorporating large language or multimodal models, cognitive failures arise from a disconnect between the model's high-level semantic reasoning and the physical or spatial reality of the driving scene.

Semantic misalignment and grounding failure. A VLA model may correctly interpret the semantic content of a traffic sign or verbal instruction but fail to translate that understanding into a spatially grounded control action. For example, DriveLM [17] demonstrated that while VLMs can accurately describe a "no left turn" sign, the downstream planning module may still generate a left-turn trajectory when the sign's spatial location is not correctly associated with the ego vehicle's current lane. This grounding gap is a manifestation of the broader symbol grounding problem in AI, now manifesting in safety-critical physical systems.

World model hallucination. Generative world models, such as GAIA-1 [20] and DriveDreamer [21], can synthesize future driving scenarios for training and evaluation. However, when these models are used for online prediction or planning, they may generate seemingly reasonable but actually non-existent environmental states. The presence of hallucinations further increases the complexity of detection, as each hallucinated token conditions subsequent generation [18].

Instruction drift and reasoning inconsistency. In VLA systems that employ chain-of-thought (CoT) reasoning, the intermediate reasoning steps may gradually diverge from the initial perception input, a phenomenon we term *instruction drift*. This can manifest as a conflict between the model's verbalized reasoning (e.g., "I should yield to the pedestrian") and its final action output (e.g., an acceleration command). Jiang et al. [19] observed that even when the reasoning is factually correct, the action decoder may fail to faithfully condition on it, particularly under time pressure when the reasoning chain is truncated for latency reduction.

2.5. Scope and Relation to Existing Surveys

The literature on autonomous driving includes extensive review work, so it is necessary to clarify how this paper supplements rather than repeats existing work. We position this survey along three dimensions: architectural scope, analytical perspective, and temporal coverage.

2.6. Literature Selection Strategy

To ensure the comprehensiveness of this survey, we adopted a structured literature selection strategy. The databases consulted include IEEE Xplore, ACM Digital Library, Springer Link, Web of Science, Google Scholar, and the arXiv preprint repository. Searches were conducted using Boolean combinations of the following keyword groups: ("autonomous driving" OR "self-driving" OR "automated vehicle") AND ("safety" OR "safety-critical" OR "collision avoidance" OR "formal verification" OR "robustness") AND ("end-to-end" OR "modular pipeline" OR

"vision-language model" OR "VLA" OR "foundation model" OR "world model"). The scope covers publications from 2017 through 2026.

A three-stage filtering procedure was then applied. In the first stage, titles and abstracts were screened for direct relevance to safety mechanisms or safety-aware design of autonomous driving systems; works focusing exclusively on perception accuracy or planning efficiency without an explicit safety dimension were excluded. In the second stage, full-text review assessed methodological depth and contribution significance, with priority given to publications in top-tier venues. High-impact arXiv preprints that have demonstrably influenced subsequent peer-reviewed work, as evidenced by citation count and community adoption, were also retained. In the third stage, forward and backward citation tracing was performed on selected core papers to capture additional relevant works missed by the keyword search. To maintain currency, we conducted a supplementary search in early 2026 specifically targeting VLA-based safety and world-model-based safety verification, areas that have seen rapid growth in the most recent literature cycle.

2.6.1. Comparison with Perception-Centric and Module-Specific Surveys

A significant strand of survey literature focuses on individual modules within the modular pipeline [48]. For example, Teng et al. [49] survey motion prediction techniques with an emphasis on interaction-aware forecasting. While these surveys offer invaluable depth within their respective domains, they do not address *cross-module* failure propagation or the safety implications of replacing modular interfaces with learned representations. The present work explicitly traces how safety risks *transform* as architectural boundaries dissolve.

2.6.2. Comparison with End-to-end and Foundation Model Surveys

Recent surveys have emerged to track the rapid development of E2E and foundation model-based driving systems. Chen et al. [6] provide a comprehensive overview of E2E autonomous driving methods, categorizing approaches by their learning paradigm (imitation vs. reinforcement) and output representation (waypoint vs. control). Li et al. [50] survey the application of vision-language models in driving, covering scene understanding, decision-making, and data generation. Yang et al. [51] focus specifically on the role of LLMs in autonomous driving across the full stack. However, these surveys primarily focus on functionality such as accuracy, rather than systematically analyzing the failure modes introduced by these architectures. Our work fills this gap by treating safety as the primary lens through which each architecture is analyzed.

2.6.3. Distinguishing Contributions of This Survey

In contrast to the above, this survey makes three specific contributions to the literature. First, it provides a *cross-paradigm* analysis that spans modular, E2E, and VLA architectures, enabling direct comparison of safety mechanisms and vulnerabilities across paradigms. Second, it introduces a *failure-mode-centric taxonomy* (Section 2.4) that categorizes safety risks not by sensor modality or network architecture, but by the *layer of abstraction* at which the failure ranges from low-level sensor noise to high-level cognitive hallucination. Third, it covers literature up to early 2026, capturing the latest advancements in VLA safety alignment, generative world models for security verification, and emerging benchmarks for security-critical assessment.

3. Safety Mechanisms in Modular Paradigm

The modular framework facilitates the insertion of security mechanisms at well-defined interfaces due to its clear decomposition into modules with different functions. This section examines three complementary lines of defence: enhancing the perception through multimodal fusion and principled uncertainty quantification (Section 3.1), imposing mathematically verifiable safety constraints on the prediction and planning stages (Section 3.2), mitigating the cascading errors that arise at inter-module boundaries (Section 3.3).

3.1. Perception Robustness: Sensor Fusion and Uncertainty Quantification

The perception module is the first and most important component in a modular system. No downstream component can correct for undetected obstacles or seriously misjudged trajectories. Hence, perception robustness is a necessary condition for system-level safety. Two interrelated research thrusts address this requirement: *multi-modal sensor fusion*, which exploits the complementary failure modes of heterogeneous sensors, and *uncertainty quantification*, which endows each perception output with a calibrated confidence measure that downstream modules can consume as a risk signal.

3.1.1. Multi-Modal Sensor Fusion Strategies

Modern autonomous vehicles typically combine cameras, LiDAR, and millimetre-wave radar, each offering a distinct trade-off between spatial resolution, range, and robustness to environmental degradation. Fusion architectures are commonly categorised by the stage at which sensor streams are combined. *Early fusion* concatenates raw or minimally processed sensor data into a shared representation before any task-specific processing. PointPainting [23] exemplifies this strategy by projecting semantic segmentation scores from a camera backbone onto LiDAR point clouds, enriching geometric data with appearance cues prior to 3-D detection. While straightforward, early fusion is sensitive to spatial and temporal calibration errors between sensors, and a malfunction in one modality can corrupt the fused representation entirely. *Late fusion* independently processes each sensor stream through modality-specific detectors and merges the resulting object-level hypotheses via association and consensus algorithms. This approach is inherently more resilient to single-sensor failure and permits independent validation of each modality. *Deep fusion* strikes a middle ground by learning to combine intermediate feature representations across modalities within a shared latent space. BEVFusion [24] unifies camera and LiDAR features in a bird's-eye-view (BEV) coordinate system through a differentiable view transformation, enabling joint spatial reasoning while retaining modality-specific encoders. TransFusion [52] employs a transformer decoder that attends to both LiDAR BEV features and camera image features, using cross-attention to resolve ambiguities that neither modality can address alone.

A critical yet often overlooked aspect of fusion is *graceful degradation* under sensor failure. In practice, a LiDAR unit may become occluded by mud or snow, or a camera may be temporarily blinded by direct sunlight. Robust fusion architectures must detect such degradation and reconfigure the fusion weights accordingly. MetaBEV [53] addresses this by learning a meta-fusion strategy that dynamically re-weights modality contributions based on an estimated quality score for each sensor stream, achieving significantly smaller performance drops under simulated sensor outages compared to static fusion baselines.

3.1.2. Uncertainty Quantification in Perception

Even deterministic perceptual outputs struggle to convey effective information about model confidence. For safety-critical downstream modules, the *absence of uncertainty information is itself a safety vulnerability*: a planner that treats a low-confidence detection identically to a high-confidence one may allocate insufficient safety margin.

Bayesian deep learning provides a principled framework for uncertainty estimation. Monte Carlo Dropout (MC-Dropout) [54] approximates Bayesian inference by performing multiple stochastic forward passes at test time with dropout enabled, interpreting the variance of predictions as epistemic (model) uncertainty.

Deep ensemble methods [55] train multiple instances of the same architecture with different random initializations and aggregate their predictions, capturing both aleatoric (data) and epistemic uncertainty. While computationally expensive at inference, Miller et al. [56] demonstrated that a small ensemble of three to five members suffices to yield well-calibrated uncertainty estimates for object detection tasks relevant to autonomous driving.

More recently, *evidential deep learning* [57] has emerged as a single-forward-pass alternative to ensembles. Rather than predicting class probabilities directly, the network outputs the parameters of a Dirichlet distribution over class probabilities, enabling simultaneous estimation of both the predicted class and the associated evidential uncertainty. Amini et al. [58] extended this framework to regression tasks, demonstrating its applicability to depth estimation and velocity prediction in driving contexts.

A practical challenge is *calibration*: the requirement that a model's predicted confidence accurately reflects its empirical accuracy. Overconfidence in uncalibrated deep networks [59] is particularly dangerous, as they provide false assurance to the planner. Post-hoc calibration techniques such as temperature scaling [59] and isotonic regression can significantly improve calibration without retraining, and have been integrated into several production perception stacks.

3.1.3. Robustness Enhancement Under Adverse Conditions

Beyond fusion and uncertainty quantification, targeted robustness enhancement for adverse conditions constitutes the third pillar of perception safety. Domain-adaptive training strategies, such as those evaluated on the ACDC benchmark [37], employ style transfer or feature alignment to bridge the distribution gap between clear-weather training data and degraded deployment conditions. RoboBEV [60] provides a benchmark for evaluating BEV-based perception under corruptions including camera noise, LiDAR beam drop, and temporal misalignment, revealing that even top-performing models lose up to 30% of their nominal mAP under moderate corruption levels. Adversarial training incorporating worst-case perturbations during the optimization process has been shown to improve robustness against both natural corruptions and deliberate adversarial attacks [39], though at a modest cost to clean-condition performance.

3.2. Formal Safety Verification in Prediction and Planning

While perception robustness minimises the likelihood of an incorrect world model reaching the planner, it cannot eliminate this possibility entirely. Formal safety verification provides a complementary layer of defence by mathematically guaranteeing that the planning output satisfies specified safety constraints, irrespective of the prediction module's imperfections, within an explicitly bounded uncertainty set.

The RSS framework [2] formalises a set of common-sense driving rules as mathematical constraints over longitudinal and lateral safe distances. Given worst-case assumptions about the response times and braking capabilities of both the ego vehicle and surrounding agents, RSS derives closed-form *proper response* conditions that, if satisfied, guarantee the ego vehicle will not be at fault in any collision.

The principal attraction of RSS lies in its *deterministic safety guarantee*: any planner whose output is projected onto the RSS-admissible set is provably safe under the stated assumptions. Intel's Mobileye has integrated RSS as a post-planning safety filter in its commercial stack, demonstrating real-world viability. However, the framework's assumptions, particularly the requirement that all surrounding vehicles comply with minimum deceleration braking requirements, can be violated by drivers with abnormal or aggressive driving behavior, creating a gap between formal assurances and real-world safety. Furthermore, overly conservative RSS parameters can exacerbate the frozen robot problem (Section 2.4.2), as the admissible trajectory set shrinks to the empty set in dense traffic.

3.3. Mitigating Error Propagation in Modular Pipelines

The advantage of modular pipelines is their ability to clearly separate concerns, but this is also their weakness. Because each module communicates with subsequent modules through a fixed-format interface, errors introduced at any stage can propagate undetected to all downstream computations. Error propagation in a serial pipeline is governed by two mechanisms. The first is *hard error compounding*: a categorical mistake by an upstream module deterministically produces a corresponding failure in all subsequent stages, since the prediction module has no input to forecast and the planner has no obstacle to avoid. The second mechanism is *soft error amplification*: even when a detection is not entirely missed, small localisation or velocity estimation errors can be magnified through nonlinear downstream processing [61].

When the perception module outputs a deterministic point estimate rather than a full probability distribution at the module interface, it may lead to harmful soft-error amplification. The prediction module, receiving a single bounding box centre as input, has no mechanism to account for the upstream localisation uncertainty and may produce a confidently incorrect trajectory forecast. Ivanovic and Pavone [62] demonstrated that incorporating perception uncertainty as an explicit input to the trajectory predictor significantly improves forecast calibration and reduces the rate of overconfident predictions that lead to unsafe plans.

Several recent studies have attempted to quantify the information loss incurred at module interfaces. Caesar et al. [46] reported that the nuScenes detection evaluation protocol, which considers only the top-K detections per frame, systematically underestimates the impact of missed detections on downstream planning safety. Philion and Fidler [61] introduced the *Lift-Splat-Shoot* framework, which explicitly lifts 2-D image features into 3-D space by predicting a depth distribution per pixel rather than a point estimate. This richer intermediate representation preserves depth uncertainty across the perception–planning interface, providing the planner with a more faithful picture of the perceptual uncertainty landscape.

Two strategies have been proposed to mitigate inter-module error propagation while retaining the interpretability benefits of the modular paradigm.

Probabilistic interfaces. Rather than transmitting point estimates, modules can communicate full posterior distributions or, at minimum, confidence-annotated outputs. Ivanovic and Pavone [62] proposed the Trajectron framework, in which the perception module emits Gaussian mixture distributions over object states, and the prediction module conditions its recurrent network on these distributions, naturally propagating upstream uncertainty into the trajectory forecast. ProbabilisticBEV [63] extends this idea to the BEV representation itself, encoding each cell not as a binary occupancy flag but as a categorical distribution over semantic classes with associated confidence.

Differentiable interfaces with joint fine-tuning. A pragmatic compromise between full end-to-end learning and strict modularity is to retain the modular structure but replace discrete interfaces with differentiable connections and perform joint fine-tuning across adjacent modules. UniAD [9] exemplifies this strategy, preserving interpretable intermediate outputs while allowing gradients to flow across module boundaries during training. This approach significantly reduces error accumulation. It enables the prediction module to implicitly compensate for systematic biases in the perception module without sacrificing the ability to validate each stage.

3.4. Discussion

The modular safety approaches described above share an architectural assumption: driving tasks can be decomposed into independent phases. This decomposition enables formal verification about individual components but also introduces vulnerabilities at module interfaces. Upstream perception errors can potentially violate the preconditions upon which downstream safety guarantees depend, and such cascading failures are rarely explicitly modeled in current models. Comparing along the evaluation dimension within this paradigm, rule-based methods such as RSS offer worst-case deterministic guarantees but rely on strong assumptions, often leading to overly conservative trajectories and reduced traffic efficiency. Learning-enhanced safety filters relax these assumptions, but their safety properties degrade to limits that are only valid within the training distribution.

Almost all modular safety approaches have a common drawback: a lack of systematic evaluation under distributional shifts. Most experiments are conducted based on nuScenes or CARLA in general scenarios; very few methods have been stress-tested against adversarial perturbations. Therefore, it remains uncertain whether the safety margins reported in controlled environments can be generalized to real-world deployment environments.

4. Safety Mechanisms in End-to-end Architectures

End-to-end (E2E) learning architectures for autonomous driving aim to establish a direct mapping from raw sensory inputs to control commands, thereby bypassing the traditional modular pipeline of perception, prediction, and planning. While this paradigm promises to achieve globally optimized driving strategies, it also presents significant safety challenges. The main challenges stem from the lack of explicit causal relationships in the underlying representations and the inherent difficulty in validating black-box models. This section systematically reviews the safety mechanisms proposed to bridge the gap between E2E performance and safety requirements.

4.1. Safety-Aware Representation Learning and Feature Encoding

A fundamental strategy for embedding safety within E2E systems involves the design of learning objectives that explicitly capture safety-critical environmental features. Early E2E models, notably the imitation learning framework by Bojarski et al. [8], were criticized for optimizing solely for action similarity, which often leads to the neglect of long-tail safety constraints such as road boundaries or vulnerable road users.

To mitigate this, recent literature has shifted toward auxiliary task learning, where the primary control task is jointly trained with safety-oriented perception objectives, including semantic segmentation, depth estimation, and object detection [64,65]. These auxiliary losses serve as an inductive bias, forcing the network to maintain high-fidelity internal representations of the spatial constraints necessary for collision avoidance. Furthermore, the adoption of Bird's-Eye-View (BEV) representations has emerged as a dominant trend, as it provides a structured spatial manifold that facilitates explicit reasoning about occupancy and agent interactions [66]. Advanced techniques, such as safety-aware motion prediction [67], further refine this space by maximizing the distance between embeddings of nominal and hazardous scenarios.

4.2. Interpretability via Attention Mechanisms and Cost Map Synthesis

The black-box nature of deep neural networks remains a main bottleneck for the certification and public trust of E2E driving systems. Consequently, interpretability is no longer treated as an auxiliary feature but as a core safety requirement.

Attention visualization has become a standard diagnostic tool. By mapping spatial attention weights in convolutional or transformer-based backbones, researchers can verify if the model's saliency aligns with objects like traffic signals or pedestrians [68,69]. Beyond mere visualization, some approaches incorporate natural language rationalization to provide human-readable justifications for specific maneuvers [70]. A more structurally transparent paradigm involves the intermediate prediction of cost maps or indicators [71,72]. Rather than direct command regression, these models output a spatial risk field, allowing for a trajectory selection process that is both inspectable and subject to classical optimization constraints [73].

4.3. Constrained Reinforcement Learning and Safety Shielding

In the context of Reinforcement Learning (RL) based policies, safety is increasingly formalized through Constrained Markov Decision Processes (CMDPs). Here, the optimization objective is reformulated to maximize rewards while strictly adhering to safety-related cost thresholds [74]. Algorithmic frameworks such as Constrained Policy Optimization (CPO) [75] and Lagrangian-based refinements [76] have demonstrated significant efficacy in reducing collision rates in complex scenarios like intersection navigation [77].

To provide hard safety guarantees that RL alone cannot ensure, the community has explored Shielding and Safety Filters. These mechanisms, often grounded in Control Barrier Functions (CBFs) or temporal logic specifications, act as an execution-time monitor [78]. By projecting the learned policy's output onto a verified safe control set, these hybrid architectures ensure forward invariance of safe states, effectively combining the high-level agility of neural policies with the rigorous guarantees of model-based control [79].

4.4. Robustness to Adversarial Attacks and Distributional Shift

E2E models exhibit a known vulnerability to adversarial perturbations—subtle, often imperceptible input modifications that can induce catastrophic control failures [80,81]. Current defense research focuses on adversarial training and randomized smoothing, aiming to provide provable robustness bounds against both digital and physical-world attacks [82].

Equally critical is the challenge of Distributional Shift, where performance degrades under novel weather or geographic conditions. To enhance out-of-distribution (OOD) resilience, practitioners employ domain randomization [83,84] and test-time adaptation [85]. Moreover, uncertainty-aware architectures, such as ensemble-based models, have shown promise in detecting OOD inputs, allowing the system to trigger conservative fallback behaviors when confidence levels drop below safety thresholds [55,86].

4.5. Discussion

End-to-end safety approaches exhibit different risks compared to modular approaches. End-to-end architectural integration comes at the cost of interpretability. When safety failures occur, it is often impossible to attribute them to specific perception misjudgments or planning flaws, making root cause analysis more difficult. Among end-to-end approaches, imitation learning methods are limited by the distributional constraints of their datasets. They struggle to handle corner cases that are poorly represented in the dataset. Reinforcement learning methods can explore rare hazardous scenarios through simulation, but designing rewards for safety scenarios remains an open challenge, and the transfer from simulation to reality introduces its own distributional gaps. From a benchmarking perspective, the industry's over-reliance on simulator's closed-loop evaluation has raised concerns about robustness. While simulator's traffic scenarios are diverse, they do not fully capture the complexities of real-world driving environments. Recently introduced more challenging benchmarks, such as NavSim and Bench2Drive, have partially addressed this deficiency, but standardized safety-specific evaluation protocols still lack.

5. Safety Mechanisms in Vision-Language-Action (VLA) Models

The emergence of Vision-Language-Action (VLA) models marks a paradigm shift toward leveraging large-scale pretrained multimodal models for autonomous driving. While VLA models introduce semantic reasoning and zero-shot reasoning capabilities, they also introduce new failure modes. The most representative failure mode is hallucination, thus requiring specialized security measures.

5.1. Mitigation of Hallucination Failures

A critical vulnerability in VLA-based driving is the propensity for hallucination [87]. In the driving domain, hallucinations can be taxonomized along two axes. *Object-level hallucinations* occur when the model fabricates entities that do not exist in the scene. For instance, AD reports a phantom pedestrian on an empty sidewalk or invents a vehicle in an adjacent lane, thereby triggering unnecessary emergency braking or evasive maneuvers. *Logical hallucinations*, by contrast, involve incorrect causal or spatiotemporal reasoning over correctly perceived entities: the model may observe a green traffic light yet erroneously conclude that the intersection is unsafe due to a nonexistent conflicting traffic flow [16].

To improve reliability, current research employs Retrieval-Augmented Generation (RAG) to anchor model outputs to verified environmental data [88]. In the autonomous driving context, RAG pipelines typically index high-definition (HD) map databases and historical driving logs; at inference time, the VLA model's scene description is used to query a vector store of map tiles and prior traversal records for the current road segment, and the retrieved context—lane topology, speed limits, intersection geometry, and previously observed traffic patterns—is injected into the language prompt to constrain generation toward factually consistent outputs [89]. This mechanism is effective for correcting logical hallucinations, as verified map semantics provide hard constraints on permissible maneuver reasoning.

Beyond retrieval-based mitigation, many state-of-the-art VLAs incorporate object-level verification modules that cross-check the model's language output against high-precision perception pipelines to ensure factual alignment [17]. Quantitative evaluation of hallucination severity is typically conducted using Visual Question Answering

(VQA) accuracy on driving-scene benchmarks [17]. As illustrated in Figure 2, the Driving VLM Benchmark provides a comprehensive assessment across different tasks: perception, prediction, planning, and robustness, specifically focusing on the model's visual grounding capabilities. Together, these detection and evaluation mechanisms provide a layered defense against the propagation of hallucinated information into safety-critical planning stages.

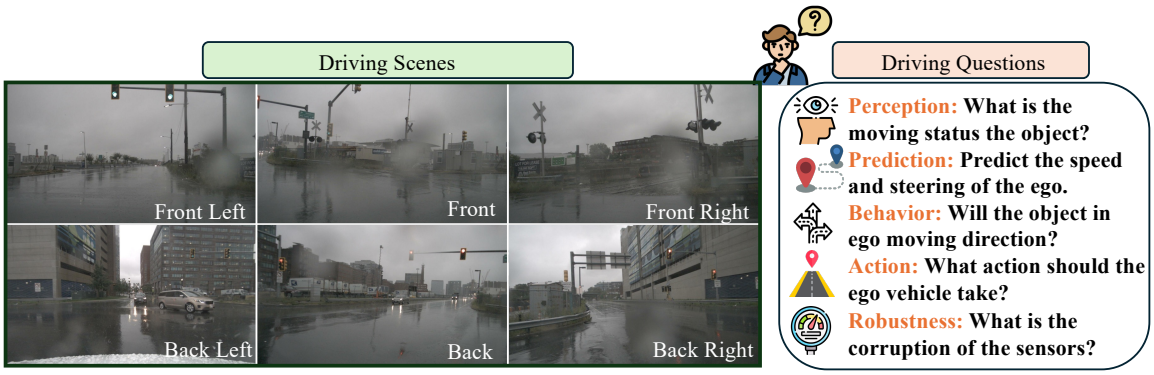

Figure 2. Driving VLM Benchmark Examples. This benchmark assesses the reliability and visual grounding of VLMs in autonomous driving across tasks: perception, prediction, planning, and robustness.

5.2. Transparency through Chain-of-Thought (CoT) Reasoning

By requiring the model to verbalize its internal reasoning, CoT provides a decision trace from scene perception to risk assessment and finally to action [90]. In practice, VLA driving systems structure the CoT as a three-stage pipeline mirroring autonomous driving stacks. In the *Perception* stage, the model generates a structured scene description that enumerates detected agents, their estimated velocities, and relevant elements such as traffic signals and lane markings. The *Prediction* stage then articulates anticipated future behaviors of surrounding agents, for example: "the cyclist on the right is decelerating and likely intending to turn." Finally, the *Planning* stage synthesizes these observations and forecasts into explicit actions, such as: "therefore, maintain current lane and reduce speed to 25 km/h to yield." This decomposition enables engineers to localize failures to a specific reasoning stage rather than treating the model as a monolithic black box [91]. For example, ColaVLA [92] integrates cognitive reasoning with hierarchical planning to map multi-view visual tokens directly to discrete driving actions. The framework optimizes decision-making by pruning visual context and employing meta-queries for refined latent deliberation.

Empirical evidence suggests that CoT-augmented models excel in high-reasoning scenarios, such as navigating unprotected left turns [44,91]. However, the *faithfulness* of these explanations remains an active area of critical inquiry [93]. Language models can generate fluent explanations, but these explanations are post-hoc rationalizations of decisions made based on various justifications. This phenomenon is particularly concerning in safety-critical applications, as inaccurate explanations can lead human to believe that dangerous actions are entirely justified.

5.3. Human-Centric Safety Alignment

While supervised fine-tuning on expert demonstrations provides a strong behavioral prior for VLA-based driving systems, it cannot fully capture the context-dependent nature of human safety preferences. Reinforcement Learning from Human Feedback (RLHF) addresses this gap by explicitly learning a reward model from human evaluative judgments, which is then used to fine-tune the driving policy via policy gradient methods [94]. This paradigm has been increasingly adopted in the autonomous driving domain to encode preferences that are difficult to specify through hand-crafted cost functions [95].

The foundation of RLHF in the driving domain lies in constructing a reward model that faithfully captures multifaceted human preferences. In practice, human annotators are presented with pairs of driving trajectories and asked to indicate which is safer or more comfortable. A key challenge is the inherently multi-dimensional nature of driving quality: unlike dialogue tasks where a single helpfulness score may suffice, driving safety alignment must simultaneously account for collision avoidance, traffic rule compliance, and passenger comfort [16,17]. Direct Preference Optimization (DPO) has emerged as a lightweight alternative that bypasses explicit reward modeling entirely, instead optimizing the policy directly on preference pairs through a classification-style objective. Results in VLA-based driving indicate that DPO achieves alignment quality comparable to full RLHF at substantially reduced training cost, making it attractive for iterative safety refinement across diverse deployment regions [96,97].

5.4. Discussion

VLA-based driving systems introduce a completely new dimension to the field of safety. VLA models can generate language explanations for their decisions, providing interpretability not available in purely numerical end-to-end models. However, whether these explanations faithfully reflect the model's internal decision-making process or merely constitute post hoc reasonable explanations remains an open question. Furthermore, there are currently no mature mitigation strategies with formal guarantees. Existing VLA models cannot provide formal safety guarantees, and validation is limited to small-scale benchmarks. The computational cost of large language model inference further limits their real-time applicability, creating a latency-safety tradeoff. These observations suggest that while VLA-based systems have considerable potential, they are currently considered as auxiliary inference modules rather than independent controllers.

6. Conclusions and Future Research Directions

This section identifies the most pressing open problems and outlines promising directions for future research.

6.1. Conclusions

This survey has provided a comprehensive overview of the safety evolution in autonomous driving architectures, tracing the paradigm shift from traditional modular pipelines to E2E frameworks and the recent emergence of VLA models. We found that while modular architectures offer high interpretability and benefit from formal verification methods like RSS, they are limited by the risk of cascading errors where upstream perception failures propagate through the pipeline. The transition to E2E systems has significantly improved generalization by jointly optimizing the entire stack. However, these systems introduce a black-box dilemma and are susceptible to causal confusion, making their decision-making logic difficult to audit in safety-critical scenarios. The integration of VLM marks the latest frontier, enhancing high-level semantic reasoning and long-tail scene understanding. Nevertheless, they introduce hallucinations and non-negligible inference latency that challenge real-time control loops. In summary, achieving safe autonomous driving requires bridging the gap between high-level cognitive reasoning and low-level physical constraints.

6.2. Future Research Directions

To develop more robust and transparent autonomous agents, we propose the following four key research directions:

- Safety-Alignment for Embodied AI: Future research should focus on safety alignment by embedding physical constraints (e.g., vehicle dynamics and collision boundaries) directly into the reward functions of foundation models. Techniques like Reinforcement Learning from Human Feedback (RLHF) can be further explored to align AI reasoning with human driving ethics.
- High-Fidelity World Models for Validation: To address the scarcity of safety-critical data in the real world, the development of generative world models (e.g., GAIA-1 [20], DriveDreamer) is essential. These models can synthesize complex adversarial scenarios to perform rigorous stress-testing of autonomous systems in high-fidelity simulations.
- Standardized Safety Benchmarks for Foundation Models: There is a critical lack of standardized benchmarks tailored for VLA and E2E architectures. Future efforts must establish multi-dimensional evaluation metrics that encompass perception robustness, logical consistency, and out-of-distribution (OOD) generalization across diverse geographic and environmental domains.
- Efficiency and Dual-System Architectures: To mitigate the latency of large models, research into lightweight architectures and dual-process models (e.g., a fast, rule-based system for routine maneuvers and a slow, VLM-powered system for complex reasoning) is vital. This balance will ensure that safety-critical decisions are made within milliseconds without sacrificing cognitive depth.

By moving in these directions, the field can advance towards a framework that combines safety verification with the intelligence of modern foundational models.

Author Contributions

S.W.: Conceptualization, investigation, writing, and revision; H.L.: Investigation, writing, and revision; G.Z: Writing, and revision. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used Gemini for language polishing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Yurtsever, E.; Lambert, J.; Carballo, A.; et al. A Survey of Autonomous Driving: Common Practices and Emerging Technologies. *IEEE Access* **2020**, *8*, 58443–58469.
2. Shalev-Shwartz, S.; Shammah, S.; Shashua, A. On a Formal Model of Safe and Scalable Self-Driving Cars. *arXiv* **2017**, arXiv:1708.06374.
3. Chen, N.; Zhang, S.; Zhou, X.; et al. Fast Robustness Enhancement for Dynamic IIoT Topology with Adaptive Bayesian Learning. *IEEE Trans. Mob. Comput.* **2025**, *24*, 10886–10899.
4. Chen, N.; Qiu, T.; Zhou, X.; et al. LEGO-Motif: Enhancing IoT Topology Robustness with Evolutionary Motif-Based Generation. *IEEE Trans. Syst. Man Cybern. Syst.* **2026**, 1–14.
5. Raju, V.M.; Gupta, V.; Lomate, S. Performance of Open Autonomous Vehicle Platforms: Autoware and Apollo. In Proceedings of the 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), Bombay, India, 29–31 March 2019; pp. 1–5.
6. Chen, L.; Wu, P.; Chitta, K.; et al. End-to-End Autonomous Driving: Challenges and Frontiers. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 10164–10183.
7. Wang, S.; Zhou, Q.; Wu, K.; et al. Interventional Root Cause Analysis of Failures in Multi-Sensor Fusion Perception Systems. In Proceedings of the Network and Distributed System Security (NDSS) Symposium 2025, San Diego, CA, USA, 24–28 February 2025.
8. Bojarski, M.; Del Testa, D.; Dworakowski, D.; et al. End to End Learning for Self-Driving Cars. *arXiv* **2016**, arXiv:1604.07316.
9. Hu, Y.; Yang, J.; Chen, L.; et al. Planning-Oriented Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 17853–17862.
10. Atakishiyev, S.; Salameh, M.; Yao, H.; et al. Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions. *IEEE Access* **2024**, *12*, 101603–101625.
11. Wang, S.; Zhou, Q.; Wu, K.; et al. REDOUBT: Duo Safety Validation for Autonomous Vehicle Motion Planning. In Proceedings of the The Thirty-Ninth Annual Conference on Neural Information Processing Systems, San Diego, CA, USA, 2–7 December 2025.
12. De Haan, P.; Jayaraman, D.; Levine, S. Causal Confusion in Imitation Learning. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019.
13. Codevilla, F.; Santana, E.; López, A.M.; et al. Exploring the Limitations of Behavior Cloning for Autonomous Driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9329–9338.
14. Luo, Y.; Zheng, J.; Chen, N.; et al. Taming Language Models for Predicting Stock Price with Causality Guidance. *IEEE Trans. Emerg. Top. Comput. Intell.* **2026**, 1–13. https://doi.org/10.1109/TETCI.2026.3663488.
15. Mao, J.; Qian, Y.; Ye, J.; et al. GPT-Driver: Learning to Drive with GPT. *arXiv* **2023**, arXiv:2310.01415.
16. Tian, X.; Gu, J.; Li, B.; et al. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. *arXiv* **2024**, arXiv:2402.12289.
17. Sima, C.; Renz, K.; Chitta, K.; et al. DriveLM: Driving with Graph Visual Question Answering. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 256–274.
18. Liu, H.; Xue, W.; Chen, Y.; et al. A Survey on Hallucination in Large Vision-Language Models. *arXiv* **2024**, arXiv:2402.00253.
19. Jiang, B.; Chen, S.; Liao, B.; et al. Senna: Bridging Large Vision-Language Models and End-to-End Autonomous Driving. *arXiv* **2024**, arXiv:2410.22313.

20. Hu, A.; Russell, L.; Yeo, H.; et al. Gaia-1: A Generative World Model for Autonomous Driving. *arXiv* **2023**, arXiv:2309.17080.
21. Wang, X.; Zhu, Z.; Huang, G.; et al. DriveDreamer: Towards Real-World-Drive World Models for Autonomous Driving. In *European Conference on Computer Vision, Proceedings of the 18th European Conference, Milan, Italy, 29 September–4 October 2024*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 55–72.
22. Kato, S.; Tokunaga, S.; Maruyama, Y.; et al. Autoware on Board: Enabling Autonomous Vehicles with Embedded Systems. In Proceedings of the 2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems (ICCPS), Porto, Portugal, 11–13 April 2018; pp. 287–296.
23. Lang, A.H.; Vora, S.; Caesar, H.; et al. PointPillars: Fast Encoders for Object Detection from Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12697–12705.
24. Liu, Z.; Tang, H.; Amini, A.; et al. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; pp. 2774–2781.
25. Huang, Z.; Liu, H.; Lv, C. GameFormer: Game-Theoretic Modeling and Learning of Transformer-Based Interactive Prediction and Planning for Autonomous Driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 3903–3913.
26. Xin, R.; Cheng, J.; Liu, H.; et al. PlanScope: Learning to Plan Within Decision Scope for Urban Autonomous Driving. *IEEE Robot. Autom. Lett.* **2026**, *11*, 3246–3253.
27. Jiang, B.; Chen, S.; Xu, Q.; et al. VAD: Vectorized Scene Representation for Efficient Autonomous Driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 8340–8350.
28. Yang, J.; Gao, S.; Qiu, Y.; et al. GenAD: Generalized Predictive Model for Autonomous Driving. *arXiv* **2024**, arXiv:2403.09630.
29. Sun, W.; Lin, X.; Shi, Y.; et al. SparseDrive: End-to-End Autonomous Driving via Sparse Scene Representation. In Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA), Atlanta, GA, USA, 19–23 May 2025; pp. 8795–8801.
30. Chen, L.; Li, Y.; Huang, C.; et al. Milestones in Autonomous Driving and Intelligent Vehicles: Survey of Surveys. *IEEE Trans. Intell. Veh.* **2022**, *8*, 1046–1056.
31. Xu, Z.; Zhang, Y.; Xie, E.; et al. DriveGPT4: Interpretable End-to-End Autonomous Driving via Large Language Model. *IEEE Robot. Autom. Lett.* **2024**, *9*, 8186–8193.
32. Ma, Y.; Wei, T.; Zhong, N.; et al. LeapVAD: A Leap in Autonomous Driving via Cognitive Perception and Dual-Process Thinking. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *37*, 1963–1977.
33. Hwang, J.J.; Xu, R.; Lin, H.; et al. EMMA: End-to-End Multimodal Model for Autonomous Driving. *arXiv* **2024**, arXiv:2410.23262.
34. Peri, N.; Dave, A.; Ramanan, D.; et al. Towards Long-Tailed 3D Detection. In Proceedings of The 6th Conference on Robot Learning; Auckland, New Zealand, 14–18 December 2022; pp. 1904–1915.
35. Dong, Y.; Kang, C.; Zhang, J.; et al. Benchmarking Robustness of 3D Object Detection to Common Corruptions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 1022–1032.
36. Gu, Y.; Wang, Q.; Qin, X. Real-Time Streaming Perception System for Autonomous Driving. In Proceedings of the 2021 China Automation Congress (CAC), Beijing, China, 22–24 October 2021; pp. 5239–5244.
37. Sakaridis, C.; Dai, D.; Van Gool, L. ACDC: The Adverse Conditions Dataset with Correspondences for Semantic Driving Scene Understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10765–10775.
38. Cao, Y.; Wang, N.; Xiao, C.; et al. Invisible for Both Camera and LiDAR: Security of Multi-Sensor Fusion Based Perception in Autonomous Driving Under Physical-World Attacks. In Proceedings of the 2021 IEEE Symposium on Security and Privacy (SP), online, 24–27 May 2021; pp. 176–194.
39. Tu, J.; Ren, M.; Manivasagam, S.; et al. Physically Realizable Adversarial Examples for LiDAR Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 13716–13725.
40. Trautman, P.; Krause, A. Unfreezing the Robot: Navigation in Dense, Interacting Crowds. In Proceedings of them 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 18–22 October 2010; pp. 797–803.
41. Cheng, J.; Chen, Y.; Chen, Q. Pluto: Pushing the Limit of Imitation Learning-Based Planning for Autonomous Driving. *arXiv* **2024**, arXiv:2404.14327.
42. Suo, S.; Regalado, S.; Casas, S.; et al. TrafficSim: Learning to Simulate Realistic Multi-Agent Behaviors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10400–10409.
43. Xie, S.; Kong, L.; Dong, Y.; et al. Are VLMs Ready for Autonomous Driving? An Empirical Study from the Reliability, Data and Metric Perspectives. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, HI, USA, 19–20 October 2025; pp. 6585–6597.

44. NVIDIA; Wang, Y.; Luo, W.; et al. Alpamayo-R1: Bridging Reasoning and Action Prediction for Generalizable Autonomous Driving in the Long Tail. *arXiv* **2025**, arXiv:2511.00088.
45. Sun, P.; Kretzschmar, H.; Dotiwalla, X.; et al. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2446–2454.
46. Caesar, H.; Bankiti, V.; Lang, A.H.; et al. nuScenes: A Multimodal Dataset for Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–June 19 2020; pp. 11621–11631.
47. Zhong, Z.; Rempe, D.; Chen, Y.; et al. Language-Guided Traffic Simulation via Scene-Level Diffusion. In Proceedings of The 7th Conference on Robot Learning, Atlanta, GA, USA, 6–9 November 2023; pp. 144–177.
48. Sun, C.; Zhang, R.; Lu, Y.; et al. Toward Ensuring Safety for Autonomous Driving Perception: Standardization Progress, Research Advances, and Perspectives. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 3286–3304.
49. Teng, S.; Hu, X.; Deng, P.; et al. Motion Planning for Autonomous Driving: The State of the Art and Future Perspectives. *IEEE Trans. Intell. Veh.* **2023**, *8*, 3692–3711.
50. Jiang, S.; Huang, Z.; Qian, K.; et al. A Survey on Vision-Language-Action Models for Autonomous Driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–20 October 2025; pp. 4524–4536.
51. Yang, Z.; Jia, X.; Li, H.; et al. LLM4Drive: A Survey of Large Language Models for Autonomous Driving. *arXiv* **2023**, arXiv:2311.01043.
52. Bai, X.; Hu, Z.; Zhu, X.; et al. TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1090–1099.
53. Ge, C.; Chen, J.; Xie, E.; et al. MetaBEV: Solving Sensor Failures for BEV Detection and Map Segmentation. *arXiv* **2023**, arXiv:2304.09801.
54. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of The 33rd International Conference on Machine Learning, New York, NY, USA, 20–22 June 2016; pp. 1050–1059.
55. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, USA, 4–9 December 2017; pp. 6405–6416.
56. Miller, D.; Nicholson, L.; Dayoub, F.; et al. Dropout Sampling for Robust Object Detection in Open-Set Conditions. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018; pp. 3243–3249.
57. Sensoy, M.; Kaplan, L.; Kandemir, M. Evidential Deep Learning to Quantify Classification Uncertainty. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018; pp. 3183–3193.
58. Amini, A.; Schwarting, W.; Soleimany, A.; et al. Deep Evidential Regression. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 14927–14937.
59. Guo, C.; Pleiss, G.; Sun, Y.; et al. On Calibration of Modern Neural Networks. In Proceedings of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; pp. 1321–1330.
60. Xie, S.; Kong, L.; Zhang, W.; et al. RoboBEV: Towards Robust Bird's Eye View Perception Under Corruptions. *arXiv* **2023**, arXiv:2304.06719.
61. Philion, J.; Fidler, S. Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D. In Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 194–210.
62. Ivanovic, B.; Pavone, M. The Trajectron: Probabilistic Multi-Agent Trajectory Modeling with Dynamic Spatiotemporal Graphs. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 2375–2384.
63. Reading, C.; Harakeh, A.; Chae, J.; et al. Categorical Depth Distribution Network for Monocular 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8555–8564.
64. Hecker, S.; Dai, D.; Van Gool, L. End-to-End Learning of Driving Models with Surround-View Cameras and Route Planners. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 435–453.
65. Ishihara, K.; Kanervisto, A.; Miura, J.; et al. Multi-Task Learning with Attention for End-to-End Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 2902–2911.
66. Hu, S.; Chen, L.; Wu, P.; et al. ST-P3: End-to-End Vision-Based Autonomous Driving via Spatial-Temporal Feature Learning. In Proceedings of the 17th European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 533–549.
67. Ren, X.; Yang, T.; Li, L.E.; et al. Safety-Aware Motion Prediction with Unseen Vehicles for Autonomous Driving. In Proceedings

of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 15731–15740.
68. Zablocki, É.; Ben-Younes, H.; Pérez, P.; et al. Explainability of Deep Vision-Based Autonomous Driving Systems: Review and Challenges. *Int. J. Comput. Vis.* **2022**, *130*, 2425–2452.
69. Cultrera, L.; Seidenari, L.; Becattini, F.; et al. Explaining Autonomous Driving by Learning End-to-End Visual Attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 340–341.
70. Kim, J.; Rohrbach, A.; Darrell, T.; et al. Textual Explanations for Self-Driving Vehicles. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 563–578.
71. Sadat, A.; Casas, S.; Ren, M.; et al. Perceive, Predict, and Plan: Safe Motion Planning Through Interpretable Semantic Representations. In Proceedings of the European Conference on Computer Vision; Glasgow, UK, 23-28 August 2020; pp. 414–430.
72. Chen, D.; Zhou, B.; Koltun, V.; et al. Learning by Cheating. In Proceedings of the Conference on Robot Learning, online, 16–18 November 2020; pp. 66–75.
73. Zeng, W.; Luo, W.; Suo, S.; et al. End-to-End Interpretable Neural Motion Planner. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8660–8669.
74. Altman, E. *Constrained Markov Decision Processes*; Routledge: Oxfordshire, UK, 2021.
75. Achiam, J.; Held, D.; Tamar, A.; et al. Constrained Policy Optimization. In Proceedings of the 34 th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017; pp. 22–31.
76. Tessler, C.; Mankowitz, D.J.; Mannor, S. Reward Constrained Policy Optimization. *arXiv* **2018**, arXiv:1805.11074.
77. Isele, D.; Nakhaei, A.; Fujimura, K. Safe Reinforcement Learning on Autonomous Vehicles. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 1–6.
78. Alshiekh, M.; Bloem, R.; Ehlers, R.; et al. Safe Reinforcement Learning via Shielding. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 2669–2678.
79. Cheng, R.; Orosz, G.; Murray, R.M.; et al. End-to-End Safe Reinforcement Learning Through Barrier Functions for Safety-Critical Continuous Control Tasks. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 3387–3395.
80. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. *arXiv* **2014**, arXiv:1412.6572.
81. Eykholt, K.; Evtimov, I.; Fernandes, E.; et al. Robust Physical-World Attacks on Deep Learning Visual Classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1625–1634.
82. Madry, A.; Makelov, A.; Schmidt, L.; et al. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv* **2017**, arXiv:1706.06083.
83. Tobin, J.; Fong, R.; Ray, A.; et al. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 23–30.
84. Zhou, Z.; Cai, T.; Zhao, S.Z.; et al. AutoVLA: A Vision-Language-Action Model for End-to-End Autonomous Driving with Adaptive Reasoning and Reinforcement Fine-Tuning. *Adv. Neural Inf. Process. Syst.* **2026**, *38*, 27920–27956.
85. Wang, D.; Shelhamer, E.; Liu, S.; et al. Tent: Fully Test-Time Adaptation by Entropy Minimization. *arXiv* **2020**, arXiv:2006.10726.
86. Hendrycks, D.; Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *arXiv* **2016**, arXiv:1610.02136.
87. Li, Y.; Du, Y.; Zhou, K.; et al. Evaluating Object Hallucination in Large Vision-Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 292–305.
88. Lewis, P.; Perez, E.; Piktus, A.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
89. Liu, F.; Lin, K.; Li, L.; et al. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. *arXiv* **2023**, arXiv:2306.14565.
90. Wei, J.; Wang, X.; Schuurmans, D.; et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
91. Mao, J.; Ye, J.; Qian, Y.; et al. A Language Agent for Autonomous Driving. *arXiv* **2023**, arXiv:2311.10813.
92. Peng, Q.; Chen, X.; Yang, C.; et al. ColaVLA: Leveraging Cognitive Latent Reasoning for Hierarchical Parallel Trajectory Planning in Autonomous Driving. *arXiv* **2025**, arXiv:2512.22939.
93. Turpin, M.; Michael, J.; Perez, E.; et al. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 74952–74965.
94. Ouyang, L.; Wu, J.; Jiang, X.; et al. Training Language Models to Follow Instructions with Human Feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744.
95. Lu, Y.; Tu, J.; Ma, Y.; et al. ReAL-AD: Towards Human-Like Reasoning in End-to-End Autonomous Driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–20 October 2025; pp. 27783–27793.

96. Li, Y.; Javanmardi, E.; Thompson, S.; et al. PrefDrive: Enhancing Autonomous Driving Through Preference-Guided Large Language Models. In Proceedings of the 2025 IEEE Intelligent Vehicles Symposium (IV), Cluj-Napoca, Romania, 22–25 June 2025; pp. 1689–1696.
97. Fu, H.; Zhang, D.; Zhao, Z.; et al. ORION: A Holistic End-to-End Autonomous Driving Framework by Vision-Language Instructed Action Generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025.