A Survey of Reasoning in Autonomous Driving Systems: Open Challenges and Emerging Paradigms

Anonymous authors
Paper under double-blind review

Abstract

The development of high-level autonomous driving (AD) is shifting from perception-centric limitations to a more fundamental bottleneck, i.e., a deficit in robust and generalizable reasoning. Although current systems manage structured environments, they consistently falter in long-tail scenarios and complex social interactions that require human-like judgment. Meanwhile, the advent of large language and multimodal models (LLMs and MLLMs) presents a transformative opportunity to integrate a powerful cognitive engine into AD systems, moving beyond pattern matching toward genuine comprehension. However, a systematic framework to guide this integration is critically lacking. To bridge this gap, we provide a comprehensive review of this emerging field, arguing that reasoning must be elevated from a modular component to the central cognitive core. Specifically, we first propose a novel Cognitive Hierarchy to deconstruct the monolithic driving task based on its cognitive and interactive complexity. Based on that, we further derive and systematize seven core reasoning challenges, such as the responsiveness-reasoning trade-off and social game. Furthermore, we conduct a dual-perspective review of the state-of-the-art, analyzing both system-centric approaches to architecting intelligent agents and evaluation-centric practices for their validation. Our analysis reveals a clear trend toward holistic and interpretable "glass-box" agents. In conclusion, we identify a fundamental and unresolved tension between the high-latency, deliberative nature of LLM-based reasoning and the millisecond-scale, safety-critical demands of vehicle control. For future work, the primary objective is to bridge the symbolic-to-physical gap, including verifiable neuro-symbolic architectures, robust reasoning under uncertainty, and scalable models for implicit social negotiation.

1 Introduction

> *"The eye sees only what the mind is prepared to comprehend."*
>
> Robertson Davies (1952)

Autonomous driving (AD) aims to build a transportation system that is safer, more efficient, and more accessible (Oviedo-Trespalacios et al., 2019; Muhammad et al., 2020). The primary bottleneck is shifting from the physical limitations of perception and control systems to a deficit in robust and generalizable reasoning (He & Lv, 2023). This challenge manifests in scenarios requiring intricate situational understanding and commonsense, such as navigating temporary traffic control (Ghosh et al., 2024) or compensating for perception system degradation (Matos et al., 2025). However, integrating reasoning ability into AD systems remains underexplored (Plebe et al., 2024). To bridge the gap, we provide a comprehensive view on the integration of **large language and multimodal models (LLMs and MLLMs)** as a cognitive engine to address these **reasoning deficits in AD systems**.

For AD systems, the central and remaining challenge is shifting from perception to reasoning, specifically for the large-scale real-world deployments (Chen et al., 2021). As reported in the individual reports from Waymo and Cruise, a more fundamental challenge for AD is the lack of advanced reasoning, such as "planning discrepancy" and "prediction discrepancy" (Boggs et al., 2020). With the development of hardware, the

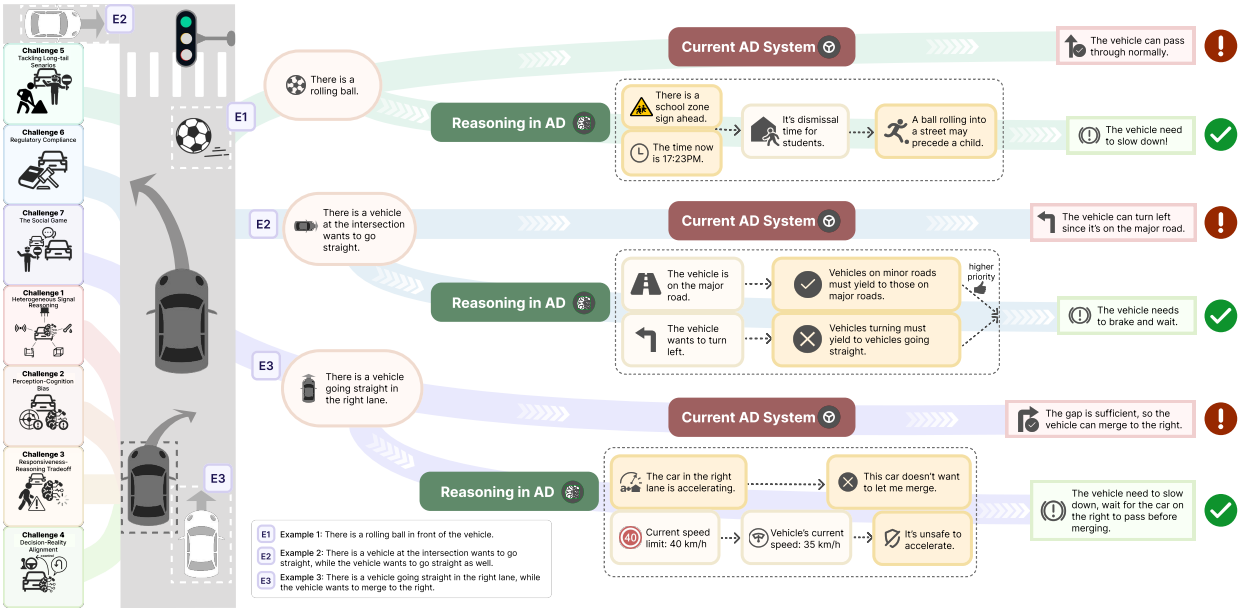

Figure 1: Challenges and examples for AD systems. Reasoning capability is necessary for suitable reactions.

bottlenecks are no longer the perception ability but the *absence of an integrated reasoning framework* (Chen et al., 2021; Mahmood & Szabolcsi, 2025; Xu & Sankar, 2024). Therefore, the current primary object of AD is to develop a cohesive cognitive architecture for advanced reasoning ability based on the individual components.

Fortunately, LLMs and MLLMs exhibit remarkable reasoning capabilities, providing a promising solution to the AD reasoning bottleneck. Through pre-training on huge data, these LLMs and MLLMs exhibit powerful emergent reasoning abilities (Webb, 2023; Yin et al., 2024) and can further leverage a rich repository of commonsense knowledge for complex situational assessment (Zhang et al., 2024a; Wei et al., 2024b). Although their internal mechanisms differ fundamentally from human cognition, they can effectively simulate deliberative thought processes and thus address the reasoning deficit in AD (Xu et al., 2025b). As shown in Fig. 1, the reasoning capacity allows AD system for propoer reaction to slove challenging scenarios. For instance, when a ball rolls onto a street, a system with such reasoning can infer that an unseen child may follow and thus prompt the vehicle to decelerate preemptively. Such capacity of probabilistic and context-aware inference is crucial for navigating the inherent unpredictability of real-world environments, and its profound potential as a new cognitive architecture motivates the systematic review presented in this survey.

To systematically analyze the central role of reasoning in autonomous driving and to establish a clear foundation for the subsequent discussion, this survey makes the following key contributions:

- **Articulation of the Core Reasoning Deficit in AD.** We systematically position the integration of reasoning as the central and unsolved challenge for next-generation autonomous systems. We move beyond broad AI-in-driving surveys to specifically articulate why LLM-based reasoning is essential for overcoming documented real-world failures (e.g., mishandling novel construction zones or misinterpreting human social cues), establishing a focused foundation for the review.
- **A Novel Cognitive Hierarchy for Driving.** We propose a new conceptual framework to deconstruct the monolithic driving task based on cognitive and interactive complexity. This hierarchy comprises three distinct levels: (1) the **Sensorimotor Level** (vehicle-to-environment), (2) the **Egocentric Reasoning Level** (vehicle-to-agents), and (3) the **Social-Cognitive Level** (vehicle-in-society). This framework provides a principled methodology for analyzing the required reasoning capabilities at each layer.
- **A New Taxonomy of Core Reasoning Challenges.** Building directly upon our cognitive hierarchy, we derive and systematize the seven most urgent challenges that impede the deployment

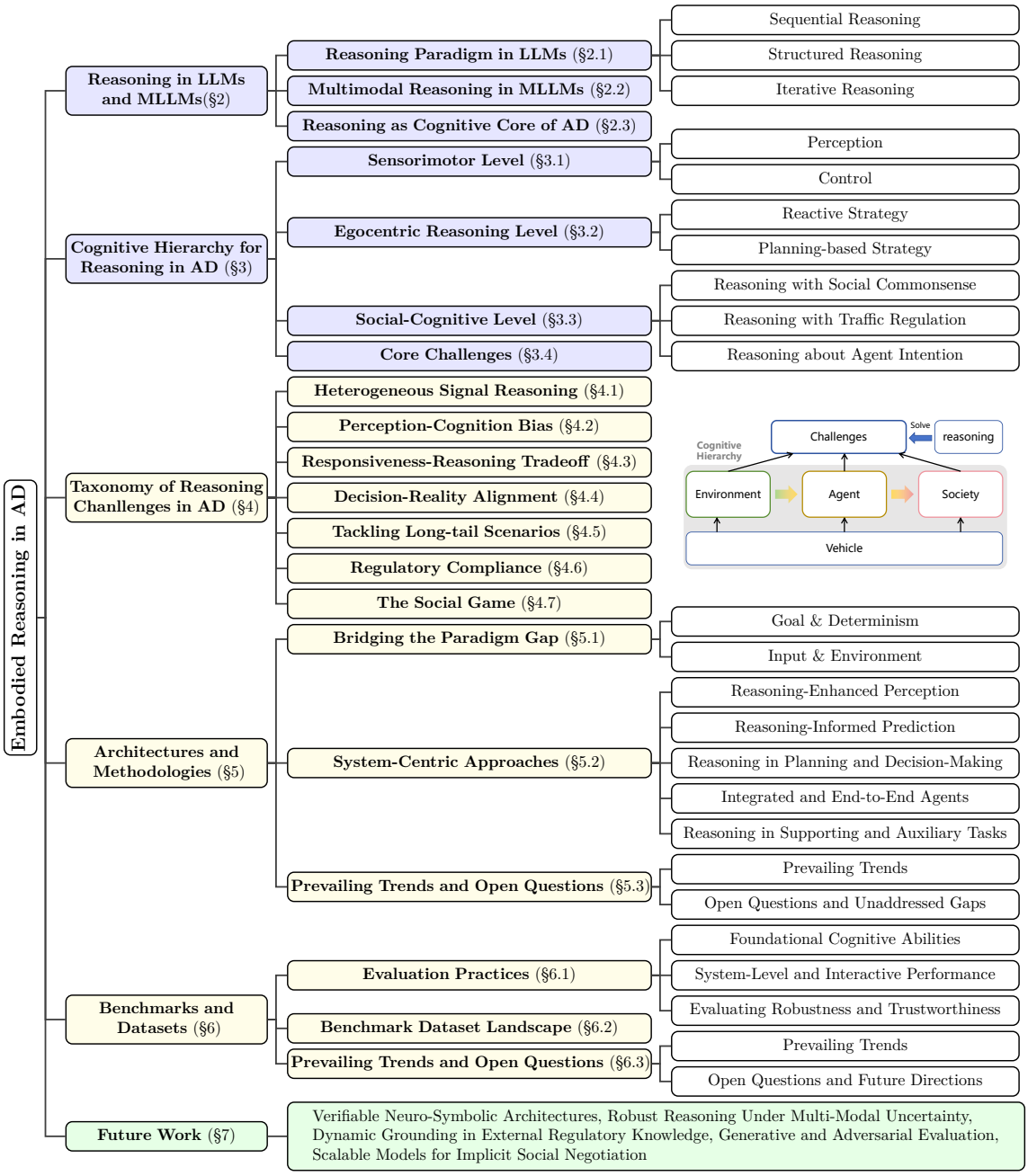

Figure 2: The outline of the survey on the reasoning in autonomous driving systems.

of LLM-based reasoning in AD. Our framework allows us to analyze how these challenges (e.g., Responsiveness-Reasoning Tradeoff, The Social Game) manifest with different priorities at each cognitive level, providing a structured problem space for the research community.

- **A Dual-Perspective Taxonomy and Analysis of the State-of-the-Art.** We provide a comprehensive review of current research through a structured analysis that distinguishes between the intelligent agent and its validation. We explore the field from two complementary viewpoints: (1) **system-centric approaches** and (2) **evaluation-centric practices**. Our analysis reveals a clear trend toward holistic, interpretable "glass-box" agents but identifies a critical gap in methods for verifying their real-time safety and social compliance.

The detailed structures of this survey are presented as follows and summarized in Fig. 2. Sec. 2 provides the preliminary concepts of reasoning paradigms within LLMs and MLLMs. Sec. 3 introduces our new cognitive hierarchy to deconstruct the autonomous driving task. Building on this framework, Sec. 4 details our taxonomy of the seven core reasoning challenges that must be addressed. We then conduct a comprehensive review of the state-of-the-art from two perspectives: Sec. 5 surveys system-centric approaches, analyzing current architectures and methodologies, while Sec. 6 reviews evaluation-centric practices, covering the critical benchmarks and datasets. Finally, Sec. 7 concludes the survey by summarizing key findings and proposing promising directions for future research.

2 Preliminary: Reasoning in LLMs and MLLMs

Reasoning, a cornerstone of human intelligence, refers to the systematic process of forming conclusions or decisions from evidence and prior knowledge. It is integral to cognitive functions such as problem-solving, decision-making, and critical analysis. In the domain of artificial intelligence, enhancing the reasoning capabilities of large language models (LLMs) and multimodal large language models (MLLMs) represents a critical step in the pursuit of artificial general intelligence (AGI) (Huang & Chang, 2023; Huang & Zhang, 2024; Han et al., 2025; Yan et al., 2025; Qu et al., 2025; Liu et al., 2025a; Zhang et al., 2024c). This advancement allows models to transition from pattern recognition and instruction execution to more complex problem-solving. Specifically for MLLMs, reasoning facilitates the coherent integration of information from diverse modalities, such as vision and text, to form a unified understanding (Wang et al., 2024d; 2025c; Li et al., 2025b; Shou et al., 2025).

2.1 Reasoning Paradigms in LLMs

To address complicated challenges in natural language processing, such as answering long-tail questions and performing task planning (Ichter et al., 2022; Yao et al., 2023b), reasoning mechanisms in LLMs decompose problems into a series of manageable, intermediate steps. This is fundamentally achieved by prompting the model to generate an explicit thought process before arriving at a conclusion (Kojima et al., 2022), a technique pioneered by methods like Chain-of-Thought (CoT) (Wei et al., 2022). Building upon this principle, existing reasoning paradigms can be categorized based on the architecture of the generated thought process (Zhang et al., 2025b). This classification reflects an evolution in methodology designed to handle increasingly sophisticated tasks. These paradigms are broadly grouped into three categories: sequential reasoning, which constructs a single linear path (Zhou et al., 2023; Yao et al., 2023b); structured reasoning, which explores multiple paths in parallel (Wang et al., 2023a; Yao et al., 2023a); and iterative reasoning, which refines a solution through feedback loops (Madaan et al., 2023; Shinn et al., 2023b).

Sequential Reasoning. Sequential reasoning represents the foundational paradigm, employing a linear, unidirectional process to generate a single, continuous chain of thought (Yao et al., 2023b; Creswell et al., 2023). This approach emulates human-like step-by-step problem-solving and serves as the basis for more advanced reasoning structures (Nye et al., 2021). The most direct implementation is the Chain-of-Thought (CoT) method (Wei et al., 2022), which instructs the model to produce a linear thought process, often activated through simple prompts (e.g., "Let's think step by step") (Kojima et al., 2022) or few-shot exemplars. Subsequent research has focused on enhancing the robustness and reliability of this linear path. Strategies include ensemble approaches like Self-Consistency (Wang et al., 2023b) and CoRe (Zhu et al., 2023), which select the most consistent answer from multiple reasoning chains, and self-correction methods like Chain of Verification (CoV) (Dhuliawala et al., 2024), which introduces an explicit step to verify and refine the generated reasoning.

Structured Reasoning. To address complicated problems that require exploration beyond a single line of thought, structured reasoning transcends the limitations of linear approaches (Wei et al., 2022; Kojima et al., 2022) by concurrently exploring multiple reasoning paths (Wang et al., 2023a; Creswell et al., 2023). This paradigm organizes these paths within non-linear data structures, such as trees or graphs, enabling a model to deliberately consider, evaluate, and integrate diverse lines of thought (Shinn et al., 2023a; Hao et al., 2023). An extension of the linear chain is the tree structure, implemented in the Tree of Thoughts (ToT)

framework (Yao et al., 2023a), which expands a single thought chain into a tree of intermediate states (Ding et al., 2025b; Wu et al., 2025b). A foresight module then evaluates each branch to systematically navigate the solution space. To facilitate more sophisticated reasoning, graph-based structures like Graph of Thoughts (GoT) (Besta et al., 2024) advance beyond trees by allowing paths to not only branch but also merge (Jin et al., 2024). The capacity to synthesize information from different branches provides graph-based reasoning with superior flexibility for problems that demand a holistic analysis.

Iterative Reasoning. In contrast to the generative and exploratory nature of the preceding paradigms, iterative reasoning introduces a feedback loop for refinement and self-correction (Shinn et al., 2023a; He et al., 2025; Yu et al., 2025). This approach targets higher solution accuracy and robustness (Yao et al., 2023a), operating on a generate-and-refine cycle where an initial solution undergoes progressive improvement. This paradigm can be categorized by the source of feedback. The fundamental form uses an internal loop, where a model critiques and revises its own output, as seen in Self-Refine (Madaan et al., 2023). A more advanced form incorporates external feedback from tools like search engines, as exemplified by the ReAct framework (Yao et al., 2023b), which interleaves thought, action, and observation. A third form leverages experiential feedback for long-term improvement. Methods like Chain of Hindsight (CoH) (Liu et al., 2024a) analyze past performance on a task to generate an improved reasoning path, which is then used to guide the model on similar future tasks, establishing a meta-level learning loop (Shinn et al., 2023a).

2.2 Reasoning Paradigms in MLLMs

The paradigm of reasoning enables LLMs to perform complicated logical inference (Wei et al., 2022). For this intelligence to become applicable in the physical world, however, reasoning must transcend purely linguistic symbols and integrate with sensory modalities such as vision (Zitkovich et al., 2023). MLLMs are pivotal in achieving this integration. Applying reasoning within MLLMs extends the capabilities of these models beyond abstract textual processing, allowing them to perceive and understand the physical world directly (Yang et al., 2023). By transforming raw visual pixels into semantic concepts, MLLMs facilitate complex reasoning grounded in visual information (Zhang et al., 2024d; Pham & Ngo, 2025).

In the course of the development of multimodal reasoning, numerous models emerge, establishing new paradigms for applications. For example, in reasoning-based segmentation, models no longer depend on predefined semantic labels. Instead, these models can interpret and execute complex instructions, such as "segment the fruit in the image with the highest vitamin C content," a task that requires both functional understanding and world knowledge (Lai et al., 2024; Xia et al., 2024; Liu et al., 2025d; 2024b; Wei et al., 2025; 2024a). Furthermore, reasoning capabilities expand from two-dimensional visual understanding to three-dimensional scene comprehension (Hong et al., 2023). In 3D multimodal reasoning, models integrate data from multi-view images, depth maps, and point clouds (Guo et al., 2023). These inputs enable the models to infer spatial structures, manage occlusions, and reason about physical feasibility (Azuma et al., 2022). Through the incorporation of explicit spatial representations and 3D encoders into language models, these systems can handle complex physical reasoning tasks, such as "determining whether a robotic arm can grasp an object without colliding with others" (Zhao et al., 2025; Huang et al., 2025; Driess et al., 2023; Zitkovich et al., 2023).

Currently, cutting-edge multimodal models, including the Qwen-VL series (Bai et al., 2025), Google Gemini (Anil et al., 2023), and GPT-5 (OpenAI, 2023), significantly advance these capabilities. The reasoning mechanisms and cross-modal spatial modeling inherent to these models are instrumental in bridging the gap between 2D perception and 3D understanding (Hong et al., 2023). This progress lays the foundation for intelligent agents capable of interacting with the physical world and comprehending complex scenes (Zitkovich et al., 2023). Such spatial reasoning abilities also provide crucial technical support for domains like AD, where the real-time understanding of dynamic 3D environments is essential (Xu et al., 2024b; Mao et al., 2023).

2.3 A New Paradigm: Reasoning as the Cognitive Core of Autonomous Driving

Traditional AD systems offen rely on a modular pipeline that consists of perception, prediction, planning, and control (Huang & Chen, 2020). While this architecture demonstrates success in structured environments, it suffers from fundamental limitations. These include significant information loss between discrete modules (Bojarski et al., 2016; Zhou et al., 2024), an over-reliance on predefined rules, and inherent brittleness when encountering uncertain or ambiguous scenarios.

This survey explores a new paradigm that moves beyond the traditional pipeline. We propose that reasoning should not be treated as another sequential module but instead be elevated to the role of a cognitive core for the entire system (Mao et al., 2023; Xu et al., 2024b). The function of this core is not to replace existing modules but rather to understand, coordinate, and empower them. This approach transforms the system from a rigid, linear process into an integrated, intelligent agent capable of holistic scene comprehension and adaptive decision-making. To fully appreciate the necessity of this paradigm shift, we contrast it with preceding frameworks. Traditional symbolic AI, while strong in formal logic, proved too brittle and unscalable for the dynamic, unstructured nature of real-world driving (Bhuyan et al., 2024), struggling with the combinatorial explosion of rules required for open environments (Bouchard et al., 2022). Causal inference models, though theoretically robust in distinguishing correlation from causation, face insurmountable challenges in practice due to the immense complexity and the presence of unobserved confounding variables inherent in traffic scenarios, making the construction of a complete causal graph infeasible (Zhang et al., 2020). Similarly, pre-LLM neuro-symbolic systems, which attempted to merge neural perception with symbolic logic, were hampered by significant integration challenges, often requiring laborious manual design of interfaces and domain-specific languages (Sun et al., 2020), while failing to fundamentally solve the symbol grounding problem (Harnad, 1990). These persistent limitations across prior paradigms underscore the need for a new approach, positioning LLM-based reasoning not merely as an alternative, but as a necessary evolutionary step to overcome these deep-rooted obstacles.

3 A Cognitive Hierarchy for Reasoning in Autonomous Driving

A monolithic view of "driving" is insufficient for targeted intervention. To map the advanced reasoning capabilities of a central Cognitive Core to concrete operational challenges, a granular deconstruction of the driving task itself is required. Drawing inspiration from the evolving complexity of the interaction between the vehicle, its environment, and human agents (Wilde, 1976; Wang et al., 2022), we propose a new conceptual framework to hierarchically structure driving tasks. As shown in Fig. 3, this framework comprises three distinct levels: (1) the Sensorimotor Level (vehicle-to-environment), (2) the Egocentric Reasoning Level (vehicle-to-agents), and (3) the Social-Cognitive Level (vehicle-in-society). This hierarchy provides a principled methodology to deconstruct the monolithic concept of "driving" into layers of increasing cognitive and interactive complexity.

3.1 Sensorimotor Level (Vehicle-to-Environment)

This level corresponds to the most fundamental operations of driving, representing atomic actions. These tasks typically involve a direct mapping from a perceptual input to an executive output and require minimal complex decision-making. The core capabilities at this level are illustrated by the following representative examples:

- **Perception (Cognition & Sensing).** Perception refers to the identification of elements in the environment and the determination of their states. This category includes foundational modules for object detection and vehicle localization, analogous to a human driver visually recognizing another vehicle or a traffic signal.
- **Control (Cognition & Actuation).** Control pertains to the execution of direct physical commands. In an autonomous system, this involves carrying out basic vehicle commands for steering, acceleration, and braking, similar to a human driver responding to a simple directive to press a pedal or turn the steering wheel.

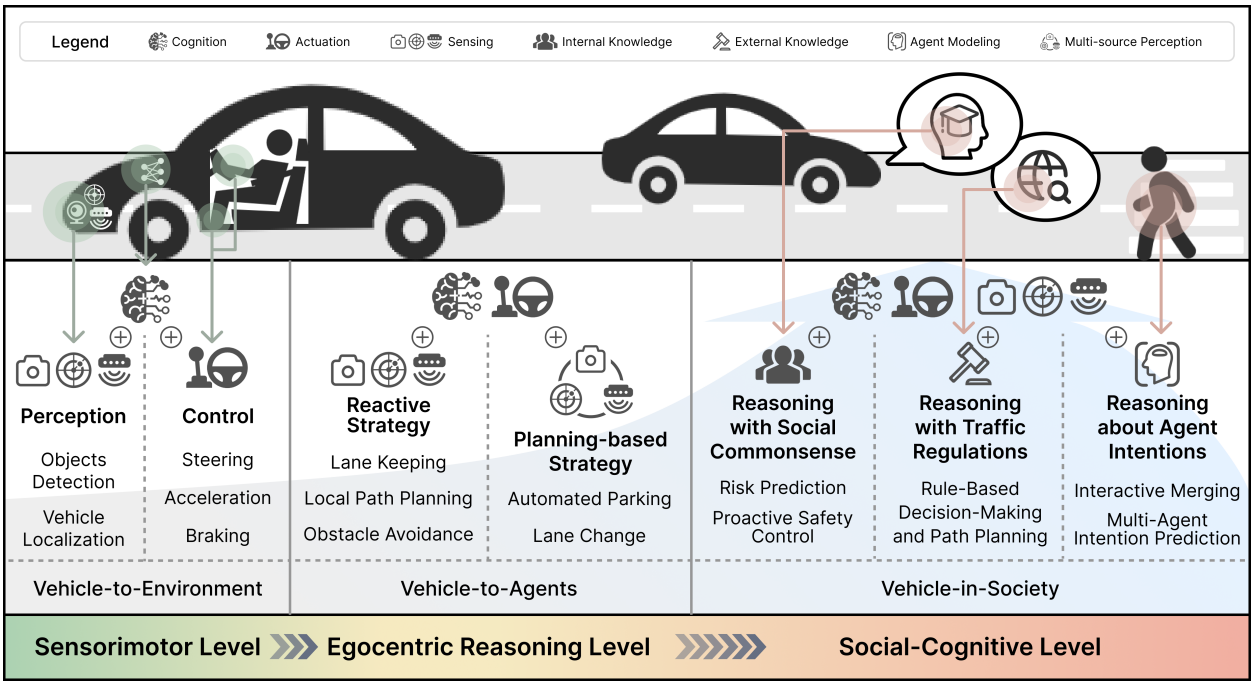

Figure 3: The proposed Cognitive Hierarchy for reasoning in autonomous driving. This framework deconstructs the monolithic "driving" task into three distinct levels of increasing cognitive and interactive complexity: (1) the Sensorimotor Level (vehicle-to-environment), (2) the Egocentric Reasoning Level (vehicle-to-agents), and (3) the Social-Cognitive Level (vehicle-in-society).

3.2 Egocentric Reasoning Level (Vehicle-to-Agents)

At this level, the system reasons from an egocentric perspective, focusing on its immediate interactions with other agents. This requires the integration of multiple internal functions to perform a closed-loop operation. Task completion relies primarily on the coordination between perception and control, allowing the system to react to other agents based on observed data, without a deep understanding of complex social rules or underlying intent. Reasoning at this level manifests in several key strategies, including but not limited to the following:

- **Reactive Strategy (Cognition & Perception & Actuation).** A reactive strategy involves the system adjusting the state of the vehicle in direct response to real-time dynamic data from sensors. For instance, when the perception system detects that a preceding vehicle is decelerating, it autonomously engages the brakes to maintain a safe following distance and lane centering. Another example occurs when the system identifies a static obstacle on the road and rapidly plans and executes a local avoidance maneuver, which requires precise coordination of steering and speed.
- **Planning-based Strategy (Cognition & Multi-source Perception & Actuation).** A planning-based strategy involves the system executing a complete sequence of complex actions to achieve a clear objective within a specific scenario. This process depends on a comprehensive model of the environment, constructed from fused sensor data. For example, during automated parking, the system integrates information from cameras and radar to understand the position and boundaries of a parking space. It then plans a coherent series of steering, forward, and reverse movements to guide the vehicle into the location, governed primarily by geometric and kinematic principles.

3.3 Social-Cognitive Level (Vehicle-in-Society)

This level represents the ultimate tasks in achieving full autonomy. It requires the system to emulate the advanced cognitive abilities of a human driver by incorporating an understanding of social commonsense,

traffic regulations, and the predicted intentions of other agents. This constitutes a profound level of social reasoning, demanding that the system operate as a socially aware participant within the dynamic traffic environment. Key challenges at this level are exemplified by the following forms of reasoning:

- **Reasoning with Social Commonsense (Cognition & Perception & Actuation & Internal Knowledge).** This capability involves leveraging an internal world model to infer unstated context and anticipate likely outcomes. For instance, upon perceiving a ball rolling into the roadway, the system infers a high probability of a child following. Consequently, the system preemptively reduces vehicle speed. This response is not a simple perception-control reflex but an anticipatory judgment derived from a sophisticated world model.
- **Reasoning with Traffic Regulations (Cognition & Perception & Actuation & External Knowledge).** This form of reasoning requires the system to query and integrate an external knowledge base of traffic laws to guide planning and decision-making. For example, when navigating an intersection that lacks traffic signals, the system must retrieve and apply the relevant right-of-way regulation to determine the appropriate action. This involves applying a formal rule from an external source to a dynamic, real-world context.
- **Reasoning about Agent Intentions (Cognition & Perception & Actuation & Agent Modeling).** This capability involves modeling the behavior of other intelligent agents to navigate complex, interactive scenarios. To merge into dense traffic, for example, the system must predict whether an adjacent vehicle will yield by observing the dynamics of that vehicle, such as changes in speed and trajectory. Based on this prediction, the system determines whether to accelerate for the merge or to slow down and await a new opportunity. Such interactions often require game-theoretic reasoning to model the reciprocal actions of intelligent agents.

3.4 From Cognitive Hierarchy to Core Challenges

Current AD systems, aligned with society of automotive engineers (SAE) Levels 2 and 3 (Standard, 2021), have established foundational capabilities at both the Sensorimotor and Egocentric Reasoning levels. While proficient in basic perception-to-action loops and rule-based planning, their reasoning at the egocentric level often lacks the flexibility to manage novel or highly interactive scenarios (Badue et al., 2021). The critical barrier to achieving higher automation (L4 and beyond), however, remains the largely unaddressed Social-Cognitive Level, which requires navigating the complex, unwritten protocols of human social interaction in traffic.

Traditional methodologies, from intricate rule-based systems to purely data-driven models, are often too rigid to address the full spectrum of these higher-level reasoning challenges. Recent breakthroughs in LLMs and MLLMs, however, offer a more powerful and unified cognitive engine capable of enhancing Egocentric Reasoning while enabling Social-Cognitive capabilities (Hu et al., 2023b). Their extensive world knowledge provides the context needed to anticipate agent intentions and understand implicit social norms. Furthermore, their sophisticated instruction-following abilities facilitate more robust planning in complex interactive scenarios and promote interpretable vehicle behavior. Critically, the architecture of these models is highly amenable to reinforcement learning, allowing for continuous adaptation in both direct agent interactions and broader social contexts (Ouyang et al., 2022). Collectively, these attributes position advanced reasoning models as a transformative technology for mastering the higher levels of the cognitive hierarchy.

However, adapting these general-purpose reasoning engines to the autonomous driving systems remains a formidable challenge. A naive integration of such models is insufficient, necessitating a structured approach to identify and categorize the key difficulties. To establish this foundation, the following section introduces a comprehensive taxonomy of the reasoning challenges that must be addressed. This taxonomy serves as a framework to systematically analyze the limitations of current approaches and to guide future research toward the development of autonomous vehicles.

Figure 4: The taxonomy of seven core reasoning challenges in autonomous driving. These challenges are categorized by their corresponding cognitive level: C1–C4 in Egocentric Reasoning level, while C5–C7 in Social-Cognitive level. Each numbered scenario illustrates a specific challenge analyzed in the text.

4 A Taxonomy of Reasoning Challenges in AD Systems

Achieving higher levels of vehicle autonomy requires mastering the complex reasoning demands of both the Egocentric Reasoning (C1–C4) and Social-Cognitive (C5–C7) levels. To provide a structured approach for tackling this multifaceted problem, this section introduces a taxonomy that deconstructs it into seven fundamental challenges. These challenges encompass the spectrum of higher-level cognition, from navigating direct agent interactions to understanding implicit social contexts. Each of these challenges is visually represented by the scenarios in Fig. 4 and is subsequently analyzed in detail.

4.1 C1 Challenge #1: Heterogeneous Signal Reasoning

An AD system must process and synthesize a multitude of heterogeneous data streams, including cameras, LiDAR, radar, and HD maps, to build a coherent representation of its environment. For example, a central reasoning engine must integrate these disparate signals, from raw sensor data to 3D point clouds. The core challenge lies in fusing this diverse information to support complex high-level reasoning.

This challenge manifests in several key areas of current model limitations. For instance, language models natively struggle with non-textual, numerical data like LiDAR point clouds, which necessitates robust mechanisms for cross-modal alignment. Furthermore, many contemporary vision models are fundamentally 2D-centric, creating a critical need for architectures that perform true 3D spatial reasoning across multiple views. The computational burden of processing high-resolution video streams also requires efficient information compression techniques that preserve temporal coherence. Finally, to effectively leverage the reasoning capabilities inherent to large language models, a shift is required from dense, unstructured feature representations to object-level tokens that align the perceptual representation with the symbolic reasoning process (Hu et al., 2023a).

These advanced reasoning capabilities are essential for critical driving tasks. For example, to understand complex spatial relationships, such as localizing an object "behind and to the left of a blue van" or determining if an approaching vehicle is facing forward, the system must move beyond simple 2D object detection. It must fuse images from multiple viewpoints with 3D geometric data to reason about the pose, orientation, and relative positioning of objects within a unified 3D space. Moreover, reasoning over high-resolution video streams is critical for tracking multiple objects and interpreting behavior over time, such as identifying a vehicle running a red light. This requires the system to dynamically associate visual cues with specific actors while maintaining temporal consistency.

4.2 C2 Challenge #2: Perception-Cognition Bias

The reliability of an AD system is fundamentally challenged by uncertainty originating from both imperfect sensory input and fallible cognitive models. Extrinsic factors, such as adverse weather and sensor malfunctions, degrade perceptual data, while intrinsic factors, like model "hallucinations," introduce cognitive errors. Overcoming these challenges requires a robust reasoning mechanism capable of cross-modal validation, compensatory inference, and reality-checking to maintain a stable and accurate understanding of the environment.

Reasoning is essential for mitigating uncertainty from extrinsic sources. For example, adverse weather like rain and fog severely compromises optical sensors such as cameras and LiDAR. Simultaneously, the signals from radar, while capable of penetrating such conditions, are difficult to interpret due to complex reflective properties (Zhang et al., 2023b; Matos et al., 2024). A reasoning module must therefore dynamically weigh and fuse evidence from these disparate sources to form a coherent environmental representation. Similarly, when a primary sensor is occluded or fails, the system must use compensatory reasoning. With sensor performance degraded, the system may infer the presence of a concealed hazard from a lead vehicles sudden braking and decelerate preemptively, even without direct visual confirmation.

The system must also employ reasoning to counteract intrinsic cognitive failures, most notably model hallucinations. An MLLM might generate non-existent objects, such as a phantom traffic light on an open road, or omit critical, perceived objects like traffic cones from a descriptive inference. To address this, a reasoning layer must continuously perform reality-checking by cross-validating the outputs from the model against raw sensor data and map information. This process allows the system to filter fabricated entities, correct omissions, and align the final decision with the ground truth of the physical world.

4.3 C3 Challenge #3: Responsiveness-Reasoning Tradeoff

AD systems face a fundamental tradeoff between real-time responsiveness and the latency inherent to deep reasoning. In critical situations, a system must react within milliseconds. However, achieving this speed is complicated by three primary factors. First, the complicated reasoning processes of large models (LLMs and MLLMs) are computationally intensive and time-consuming (Tian et al., 2025a). Second, reliance on remote foundational models for inference introduces significant network latency, which is often incompatible with the stringent real-time demands of driving (Krentsel et al., 2024; Tahir & Parasuraman, 2025). Third, the massive data streams from high-resolution, multi-modal sensors impose a heavy computational load on onboard hardware, further hindering the efficient execution of sophisticated models (Wang et al., 2024c; Brown et al., 2023).

This tension between deliberative "slow thinking" and reactive "fast thinking" necessitates the development of dual-process architectures. Such a system must dynamically arbitrate between a fast, reactive controller for immediate responses and a deep, deliberative engine for complex planning and prediction (Serban et al., 2018). This challenge is particularly evident in scenarios that require the system to balance immediate action with strategic foresight. For example, a vehicle may need to execute an emergency maneuver to avoid a sudden obstacle, an action requiring the immediate activation of a reactive module. If this occurs during a complex, game-theoretic negotiation like a lane merge, the deliberative module must simultaneously evaluate the long-term consequences associated with its actions, demanding seamless coordination between the two processes. This tradeoff also manifests acutely during high-speed merging maneuvers. When entering dense highway traffic, the system must identify a safe gap in a fleeting time window, a task that requires both rapid perception of vehicle dynamics and deep reasoning about the intentions of other drivers. The final decision must synthesize these inputs with an assessment of the status of the ego-vehicle, where a slight delay risks a missed opportunity and a misjudgment in reasoning could lead to a collision.

4.4 C4 Challenge #4: Decision-Reality Alignment

A core challenge for AD systems is grounding high-level, semantic decisions in the physical reality of the vehicle and its environment. A significant gap often exists between the abstract reasoning of large models and the concrete requirements of motion planning, leading to inconsistencies between a conceptual decision and the final executable trajectory (Yao et al., 2024). For instance, a chain-of-thought output like "change to the right lane to avoid the obstacle" may be conceptually sound but physically infeasible if it disregards the kinodynamic constraints of the vehicle (e.g., turning radius, tire adhesion) or the geometric constraints of the environment. This misalignment can produce plans that are unsafe or impossible to execute. Therefore, a reasoning mechanism is imperative to ensure that high-level strategic decisions are continuously validated against real-world physical laws and constraints.

The necessity of this reasoning is evident in common yet complicated scenarios. For example, when maneuvering on a narrow road, a high-level decision to "reverse to yield" is insufficient. The reasoning module must

further assess the physical viability of this plan by evaluating the clearance behind the vehicle and ensuring the intended trajectory does not conflict with roadside obstacles. This challenge is amplified in dynamic and uncertain conditions, such as navigating a sharp turn on a slippery surface. A high-level plan for a sharp turn may be invalidated by the physical reality of reduced tire adhesion. Reasoning must connect the semantic goal of "navigate the turn" to the physical state of the environment, inferring a drastically reduced coefficient of friction for the road surface. This inference is then used to compute a safe speed for the vehicle, preventing a loss of vehicle control that could occur if the decision were based solely on the posted speed limit.

4.5 C5 Challenge #5: Tackling Long-tail Scenarios

The long-tail distribution of driving scenarios poses a fundamental challenge to prevailing data-driven methodologies, which falter in novel or data-scarce situations. While current models perform well in common scenarios, their efficacy sharply degrades when encountering rare events like temporary construction zones or extreme weather (Bogdoll et al., 2022; Tian et al., 2024b). This problem stems from two interconnected limitations: the infeasibility of comprehensive data collection and the inherent brittleness of pattern-matching mechanisms. Consequently, the key to handling long-tail events lies not in accumulating more data, but in equipping the system with robust reasoning capabilities to make sound decisions without direct experiential precedent.

The first limitation is data scarcity. Long-tail events are, by definition, rare, making the acquisition of sufficient real-world training data logistically impossible (Zhang et al., 2023a). While data synthesis offers a partial remedy, generating high-fidelity, physically realistic simulation data remains a formidable challenge, and it is impossible to enumerate all potential corner cases beforehand (Dosovitskiy et al., 2017; Wang et al., 2021). The existence of these "unknown unknowns" places a ceiling on any strategy that relies on exhaustive data coverage. Therefore, an autonomous system must be able to generalize from underlying commonsense principles rather than merely interpolating from seen data. The second limitation is the operational mechanism of current models, which relies on pattern matching rather than logical inference. This approach is effective at generalizing within the training distribution but becomes brittle when confronted with out-of-distribution long-tail events (Grigorescu et al., 2020). Such scenarios often involve dynamic rule changes and conflicting information, which demand causal and logical deliberation that pattern matching cannot provide.

The critical role of reasoning is evident in how a system handles such scenarios. For instance, when a pedestrian suddenly emerges from an occluded area, a system limited to pattern matching may fail without a direct precedent. In contrast, a reasoning-enabled system can proactively apply defensive driving principles, using commonsense knowledge that occlusions create blind spots or inferring that a wrong-way vehicle signals a downstream anomaly to reduce speed. This inferential capability is also essential for resolving rule conflicts. As an example, in the temporary construction zone, hand signals from a traffic officer instruct the vehicle to stop, contradicting a green traffic light. The challenge here is not perceptual but decision-theoretic. The system must apply a known hierarchy of authority (e.g., human officer $\succ$ traffic signal $\succ$ map data) to deduce that standing regulations are temporarily superseded and execute the action corresponding to the highest-priority command.

4.6 C6 Challenge #6: Regulatory Compliance

An AD system must navigate a complicated and dynamic web of traffic regulations to ensure safe and legal operation. This requirement presents a significant reasoning challenge, which can be deconstructed into two primary dimensions: the complexity of interpretation and the context-dependency of application. The difficulty in interpretation stems from a regulatory landscape that includes not only thousands of legal statutes but also numerous local ordinances and unwritten driving conventions (Koopman & Wagner, 2017; Lin, 2015). A scenario in which a vehicle must reason about legal statutes, local customs, and regional standards to determine proper bus-lane use highlights this complexity. Compounding this issue, traffic laws vary significantly across different jurisdictions, necessitating robust system capabilities for retrieving and interpreting applicable rules in any given context (Dixit et al., 2016). Furthermore, the application of these

rules is highly dependent on the immediate situation. The system must continuously analyze the behavior of other road users, current road conditions, and environmental factors to select and prioritize relevant regulations.

Addressing this challenge demands a multi-step reasoning process: from scene perception, the system must retrieve relevant regulations, judge their applicability, and resolve any conflicts among them to yield a safe and lawful decision. Numerous driving scenarios illustrate this requirement. For instance, consider unprotected turns at intersections, which research indicates account for a substantial portion of traffic accidents (Li et al., 2019). In these settings, the system must accurately infer the assignment of right-of-way and integrate this inference with predictions regarding the intent of other vehicles and a comprehensive risk assessment. Another critical scenario involves the system response to emergency vehicles or temporary traffic controllers. The detection of such an event must trigger a specialized reasoning protocol to retrieve and execute overriding rules, such as yielding to an ambulance, even if this action requires temporarily violating standard traffic laws. This highlights the system need for a robust capability to dynamically prioritize among conflicting regulations.

4.7 C7 Challenge #7: The Social Game

In mixed-traffic environments, an autonomous vehicle must operate not merely as a law-abiding agent but as a socially intelligent participant. This requires the system to reason about the implicit, non-verbal communication of other road users to infer their intentions and anticipate their actions. Existing models, however, often treat human drivers as passive agents, failing to interpret the subtle modulations in speed and headway used to signal intent. This can lead to interactions that are overly conservative, inefficient, or dangerously aggressive (Li et al., 2024a; Poots, 2024).

Beyond inferring the intent of others, the system's own actions must be legible and its decisions transparent. The absence of human-like cues can make the behavior of the vehicle unpredictable to pedestrians, creating unsafe and uncomfortable situations. Furthermore, the opaque nature of end-to-end models undermines the trust from passengers and regulators, as the rationale behind decisions is not accessible (Rezwana & Lownes, 2024; Zhanguzhinova et al., 2023). Therefore, a reasoning-enabled system must not only interpret social cues but also generate behavior that is socially compliant and readily understandable.

The application of social reasoning is critical across three primary interaction domains. First, in vehicle-to-vehicle interactions like merging, the system must decode and respond to subtle cues. Human drivers use changes in speed and headway to implicitly signal the intent to yield or proceed, and the autonomous system must participate in this dynamic negotiation (Naiseh et al., 2025; Nozari et al., 2024). Second, in vehicle-to-pedestrian scenarios, such as the unsignalized crosswalk, the system must infer the intent to cross from posture and movement. It must then project its own yielding intent through clear and considerate actions, such as early and smooth deceleration, to establish a shared understanding. Finally, for vehicle-to-passenger interaction, the system must provide transparent explanations for its actions. The ability to articulate the rationale for a sudden maneuver in natural languagelinking sensor inputs to a high-level decisionis crucial for building trust with the occupants of the vehicle.

Summary & Discussion. The preceding analysis of seven core challenges collectively underscores a necessary paradigm shift in the development of autonomous systems: a move from pure perception and control to sophisticated, human-like reasoning. To operate safely and effectively in a complex world, an autonomous system must master a spectrum of reasoning capabilities. The essential challenges that we have detailed are summarized as follows:

Egocentric Reasoning Level (several challenges remain unresolved):

- **Heterogeneous Signal Reasoning:** Systems must fuse disparate data types to build a coherent and unified world model as the prerequisite for all subsequent reasoning.
- **Perception-Cognition Bias:** Reasoning is required to validate information and compensate for failures arising from environmental factors or intrinsic model limitations, ensuring the world model is reliable.

- **Responsiveness-Reasoning Tradeoff:** The high latency inherent to large models must be reconciled with the millisecond-level reaction times required for driving, pointing toward hybrid architectures that balance deliberation and reaction.
- **Decision-Reality Alignment:** High-level semantic decisions must be continuously aligned with the kinodynamic constraints of the vehicle and the physical laws of the environment to ensure all plans are executable.

Social-Cognitive Level (largely unsolved):

- **Tackling Long-tail Scenarios:** In novel edge cases where direct experience is absent, systems must rely on commonsense and social context to navigate situations that defy standard patterns.
- **Regulatory Compliance:** Systems must interpret and adhere to the complex and dynamic set of formal societal rules embodied in traffic laws to ensure all actions are legally compliant.
- **The Social Game:** The system must infer the implicit, informal rules of human interaction, interpreting the intent of other agents and engaging in legible, socially acceptable negotiation.

This comprehensive analysis reveals a critical insight. While the initial set of challenges represents profound engineering hurdles in building and grounding a reliable agent, the ultimate bottleneck to achieving human-like autonomy lies in navigating the complexities of the social world. Addressing these latter challenges requires a fundamental move beyond pattern matching. It demands the deep, structured reasoning capabilities promised by large models. The pivotal open question for future research is therefore how to integrate these powerful reasoning engines into AD systems, guaranteeing the absolute reliability and real-time performance required for deployment in the physical world.

5 System-Centric Approaches: Architectures and Methodologies

5.1 Bridging the Paradigm Gap: From General Reasoning to Situated Autonomy

A fundamental paradigm gap separates the reasoning required for AD from the general reasoning capabilities of LLMs and MLLMs. This disparity manifests across several critical dimensions, most notably in their core objectives, operational domains, and input modalities.

Goal & Determinism. The primary mandate of an AD system is to ensure safety and control within the physical world. This mandate requires a reasoning framework that is highly deterministic and predictable. Given identical perceptual input, the decisions of the system must remain stable and consistent, as its logic is grounded in rigorous physical laws, traffic regulations, and control theory. Conversely, the core objective of an LLM or MLLM is to operate within the informational domain. The reasoning process is inherently probabilistic and creative, prioritizing semantic coherence and plausibility over physical precision. As a result, an LLM or MLLM can generate multiple distinct yet valid responses to a single prompt.

Input & Environment. The input to autonomous driving consists of real-time, high-dimensional, and continuous physical world data from sensors (e.g. cameras, LiDAR, and millimeter wave radar). Real physical environments are required to be processed that are full of noise and uncertainty. In contrast, the input to LLMs mainly comprises discrete symbolic textual data. LLMs operate within a relatively closed virtual world constructed by linguistic symbols. Even when external information is acquired through tool calling, their interaction mode remains structured.

In summary, reasoning in autonomous driving is a deterministic, safety-critical process engineered for precise action in the physical world. The reasoning of LLMs and MLLMs, conversely, is a probabilistic, generative process designed for semantic flexibility in the informational world. This fundamental paradigm gap necessitates innovative approaches to adapt and ground the powerful general reasoning within the specialized context of autonomous systems. The following taxonomy provides a systematic review of these emerging system-centric innovations.

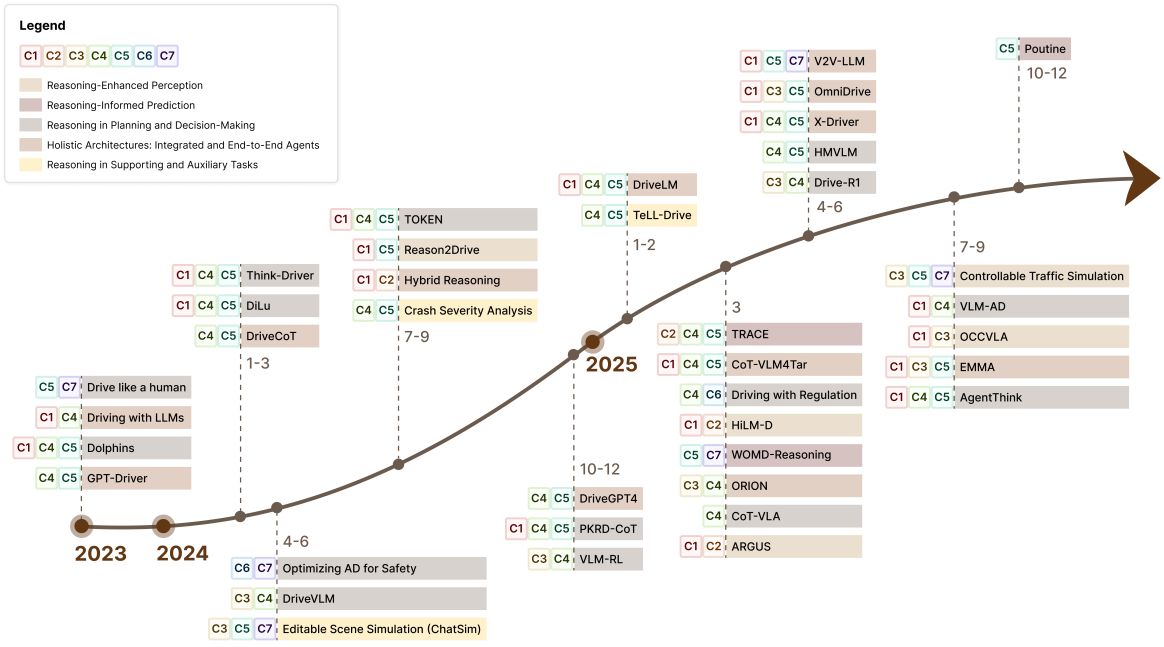

Figure 5: The chronological evolution of major methods in autonomous driving. This timeline highlights the rapid progression of the field and provides historical context for the thematic complexity taxonomy introduced in Sec. 5.2.

5.2 A Task-Based Taxonomy within System-Centric Approaches

A significant body of research now focuses on bridging the paradigm gap by adapting foundation models for the specific demands of autonomous driving. These efforts universally leverage techniques such as Chain-of-Thought (CoT) reasoning and structured representations to endow systems with more robust capabilities. To provide a clear narrative, we structure our taxonomy to reflect a logical progression from enhancing individual components to building holistic intelligent agents. As depicted in Fig. 5, the timeline presents the chronological release of each major method, illustrating the rapid pace at which this domain is evolving. Our taxonomy groups the research landscape into thematically ordered stages of rising complexity.

Our taxonomy commences with a review of methods that enhance the foundational modules for environmental understanding: perception (Sec. 5.2.1) and prediction (Sec. 5.2.2). It then transitions to the core action-generation module responsible for creating safe and coherent behaviors: planning and decision-making (Sec. 5.2.3). Subsequently, we explore holistic architectures that transcend the traditional modular pipeline by creating more integrated and end-to-end agents (Sec. 5.2.4). Our review culminates in an examination of the supporting and auxiliary tasks that enable robust system development and validation (Sec. 5.2.5). Throughout this taxonomy, we systematically connect the discussed approaches to the core challenges (C1–C7) identified in Sec. 4.

5.2.1 Reasoning-Enhanced Perception

Recent research in perception primarily focuses on moving beyond simple object detection towards a deeper, contextual understanding of the driving scene. The goal is to enhance the visual comprehension, robustness, and structured representation capabilities of MLLMs.

Key works in this area exemplify several strategic directions. One direction involves architectural innovations to improve fine-grained perception. For instance, HiLM-D (Ding et al., 2025a) employs a two-stream framework to better interpret and perceive high-risk objects, addressing how standard models often fail to

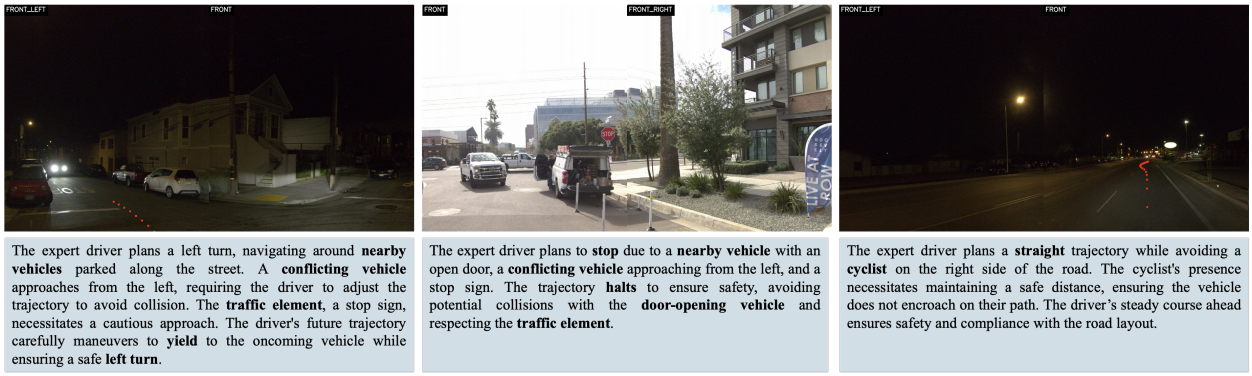

Figure 6: Demonstration of Model-Predicted Trajectories. Figure referenced from Rowe et al. (2025)

capture critical multi-scale details (C1, C2). OccVLA (Liu et al., 2025b) proposes a novel occupancy-visual-language framework, which achieves more critical fine-grained understanding of 3D spatial semantics and addresses the inference latency issue caused by the massive number of parameters in autonomous driving models based on 3D vision-language models (C1, C3). A second strategy focuses on data-centric contributions to enable robust evaluation. Reason2Drive (Nie et al., 2024) contributes a large-scale video-text dataset designed specifically for the structured assessment of perception and reasoning. This work directly addresses data scarcity and provides a new benchmark for verifying genuine scene comprehension (C1, C5). A third direction centers on creating an explicit link between visual evidence and language-based reasoning. ARGUS (Man et al., 2025) utilizes bounding boxes as explicit signals to guide the visual attention of the model, improving performance in tasks that require precise visual grounding (C1, C2).

5.2.2 Reasoning-Informed Prediction

In the context of AD, prediction involves forecasting the future behavior of other traffic participants. Current research in this area increasingly focuses on enhancing robustness in long-tail scenarios, augmenting data with domain knowledge, and ensuring that predictions align with safe decision-making.

Several distinct research strategies highlight this trend. One strategy focuses on improving predictive robustness in situations with sparse or uncertain observations. For example, TRACE (Puthumanaillam et al., 2025) improves behavior prediction by using Tree-of-Thought (ToT) and counterfactual criticism, which allows the model to generate and evaluate multiple hypothetical reasoning paths (C2, C4, C5). Another strategy involves injecting explicit domain knowledge into the models. WOMD-Reasoning (Li et al., 2024c) accomplishes this by generating millions of question-answer pairs based on traffic rules, thereby enriching the data resources for both prediction and decision-making (C5, C7). A third direction seeks to ensure consistency between prediction and subsequent planning. As shown in Fig. 6, Poutine (Rowe et al., 2025) utilizes vision-language-trajectory pre-training and preference-based reinforcement learning to align the model's predictions with established driving preferences and safety regulations (C2, C5).

5.2.3 Reasoning in Planning and Decision-Making

This category reviews works that focus specifically on the core action-generation module of the autonomous driving stack, which is responsible for planning the ego-vehicle's trajectory and making tactical decisions. Research in this domain has evolved significantly beyond traditional trajectory optimization. The central goal is to create a reasoning process that is not just optimal, but also compliant with external rules, coherent in its internal logic, and adaptive to an open-world environment. We organize our comprehensive review around these three central themes.

Embedding External Constraints for Compliant Behavior. A primary research thrust is the explicit integration of external rules and safety norms into the reasoning framework. This marks a paradigm shift from optimizing for pure performance to optimizing for compliance and social acceptability. These approaches

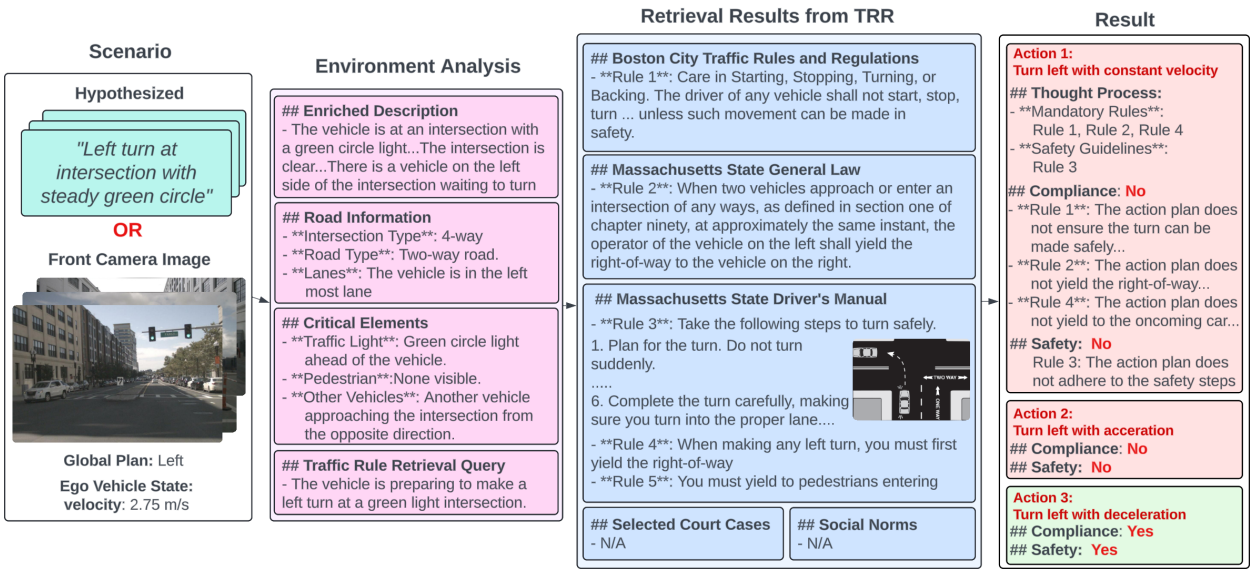

Figure 7: Demonstration of Consistency between Perception and Reasoning. Figure referenced from Cai et al. (2024)

ensure that an agent's behavior is grounded in the complex realities of the road. For instance, some methods directly tackle legal and safety verification. As shown in Fig. 7, Driving with Regulation (Cai et al., 2024) integrates a regulation retrieval mechanism with CoT reasoning, enabling the system to generate decisions that are certifiably compliant with local traffic laws (C4, C6). Optimizing AD for Safety (Sun et al., 2024) combines RLHF with LLMs, using physical and physiological feedback to train an agent that prioritizes human safety (C6, C7). Extending beyond hard constraints, other works focus on social compliance. Drive Like a Human (Fu et al., 2024) emphasizes that systems should exhibit human-like driving styles and be able to interact naturally with their environment, addressing the crucial challenges of social interaction and long-tail events (C5, C7).

Ensuring Internal Coherence through Structured Reasoning. Another significant body of research focuses on ensuring consistency between the model's high-level semantic reasoning and its final, low-level executable plan. This addresses the critical challenge of reasoning-action alignment. Chain-of-Thought has become a cornerstone technique for achieving this coherence by making the "thought process" explicit. Works like CoT-VLA (Zhao et al., 2025), Think-Driver (Zhang et al., 2024b), and PKRD-CoT (Luo et al., 2024) incorporate explicit reasoning chains to generate transparent and justifiable driving instructions, directly improving interpretability and the alignment between reasoning and action (C1, C4, C5). To further ground this reasoning process, many methods utilize structured representations. For example, TOKEN (Tian et al., 2024a) decomposes scenes into object-level tokens, providing a scaffold that helps the LLM apply its common-sense knowledge to long-tail planning scenarios (C1, C4, C5). HMVLM (Wang et al., 2025a) employs a multi-stage CoT process that moves from scene understanding to decision-making to trajectory inference, ensuring a logical flow (C4, C5). A variety of VLM-based architectures also contribute to this theme. For instance, VLM-AD (Xu et al., 2024a) uses a VLM as a teacher to provide rich reasoning labels, while DriveVLM (Tian et al., 2024c) and Dolphins (Ma et al., 2024a) use hybrid systems and multimodal CoT to achieve human-like adaptability. Finally, reinforcement learning (RL) is increasingly used to fine-tune this alignment (C1, C3, C4, C5). VLM-RL (Huang et al., 2024) uses a pre-trained VLM to generate dense reward signals, while Drive-R1 (Li et al., 2025a) connects CoT reasoning to the RL policy to improve decision quality (C3, C4).

Achieving Open-World Adaptability through Learning and Interaction. Furthermore, the recent works are advancing planning and decision-making models from closed systems to open agents that can learn from experience and interact with external knowledge. This pushes the boundaries of adaptability, especially for handling novel situations. DiLu (Wen et al., 2024) pioneers this direction by enabling a model

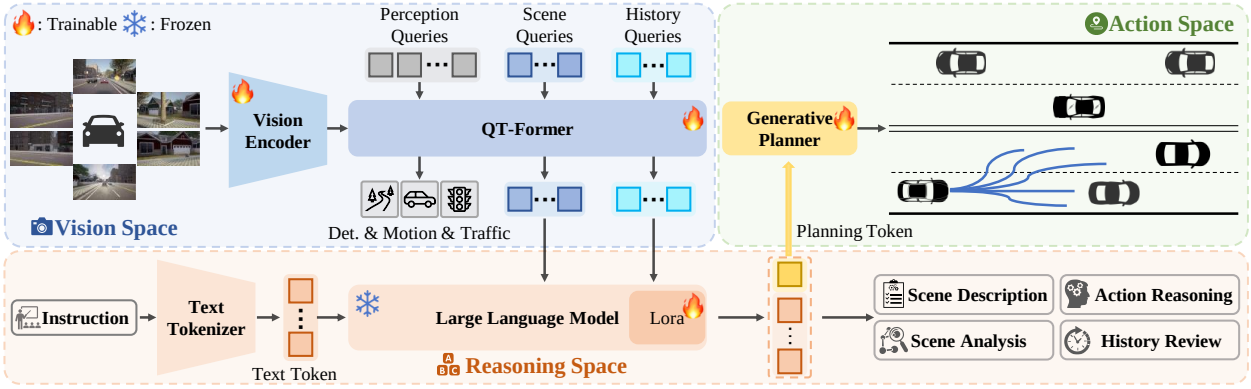

Figure 8: Demonstration of End-to-End Seamless and Coherent Information Flow. Figure referenced from Fu et al. (2025)

to accumulate experience and correct past mistakes through internal memory and reflection mechanisms (C1, C4, C5). This transforms decision-making from a series of independent, one-time inferences into a dynamic process of continuous learning. Pushing this concept further, AgentThink (Qian et al., 2025) trains an agent to dynamically call external tools (e.g., calculators, search engines). This significantly enhances the depth and flexibility of its reasoning process, allowing it to solve complex problems that require external knowledge or precise computation, thereby evolving the agent from a "closed-world planner" to an "open-world problem solver" (C1, C4, C5).

5.2.4 Holistic Architectures: Integrated and End-to-End Agents

This category reviews works that move beyond enhancing individual modules to architecting more holistic intelligent agents. The primary motivation is to overcome the information bottlenecks and error propagation inherent in traditional, sequential pipelines. The research in this area follows a trajectory, beginning with the tight coupling of key modules, progressing towards fully unified end-to-end reasoning systems, and finally expanding the very scope of what a holistic agent can do.

Integrating Perception and Decision-Making. The first step toward holistic design involves creating architectures that directly integrate the perception and decision-making modules. These works reframe visual reasoning not as a passive representation learning task, but as an active, decision-oriented explanatory process. For instance, some research focuses on breaking through visual reasoning bottlenecks in complex scenarios. DriveLM (Sima et al., 2024b) uses a graph-structured VQA dataset to handle complex scene interactions, while CoT-VLM4Tar (Ren et al., 2025) employs CoT reasoning to resolve abnormal traffic situations like phantom jams (C1, C4, C5). Other works improve the grounding of reasoning through structured multimodal inputs. Driving with LLMs (Chen et al., 2024) achieves this by fusing object geometric vectors with language, while Hybrid Reasoning (Azarafza et al., 2024) integrates mathematical and commonsense reasoning to improve robustness in adverse weather (C1, C2, C4). Collectively, these approaches aim to improve closed-loop performance and interpretability, as exemplified by X-Driver (Liu et al., 2025c), which uses an autoregressive model to generate decision commands from vision-language inputs (C1, C4, C5).

The Evolution to End-to-End Reasoning Agents. The ultimate goal of this research direction is the development of fully end-to-end agents that map sensor inputs directly to control outputs. A critical evolution in this domain is the shift from opaque "black-box" models to interpretable "glass-box" systems. The foundation for this shift is often laid by new datasets, such as DriveCoT (Wang et al., 2024b), which provides end-to-end data with explicit CoT labels (C4, C5). Building on this, architectures like GPT-Driver (Mao et al., 2023) and DriveGPT4 (Xu et al., 2024c) pioneer the "glass-box" concept by generating explicit natural language justifications alongside vehicle actions (C4, C5). These models must also handle diverse inputs; EMMA (Hwang et al., 2025), for instance, fuses non-perceptual inputs like navigation instructions with visual data, though this highlights the trade-off between reasoning depth and real-time performance (C1, C3, C5).

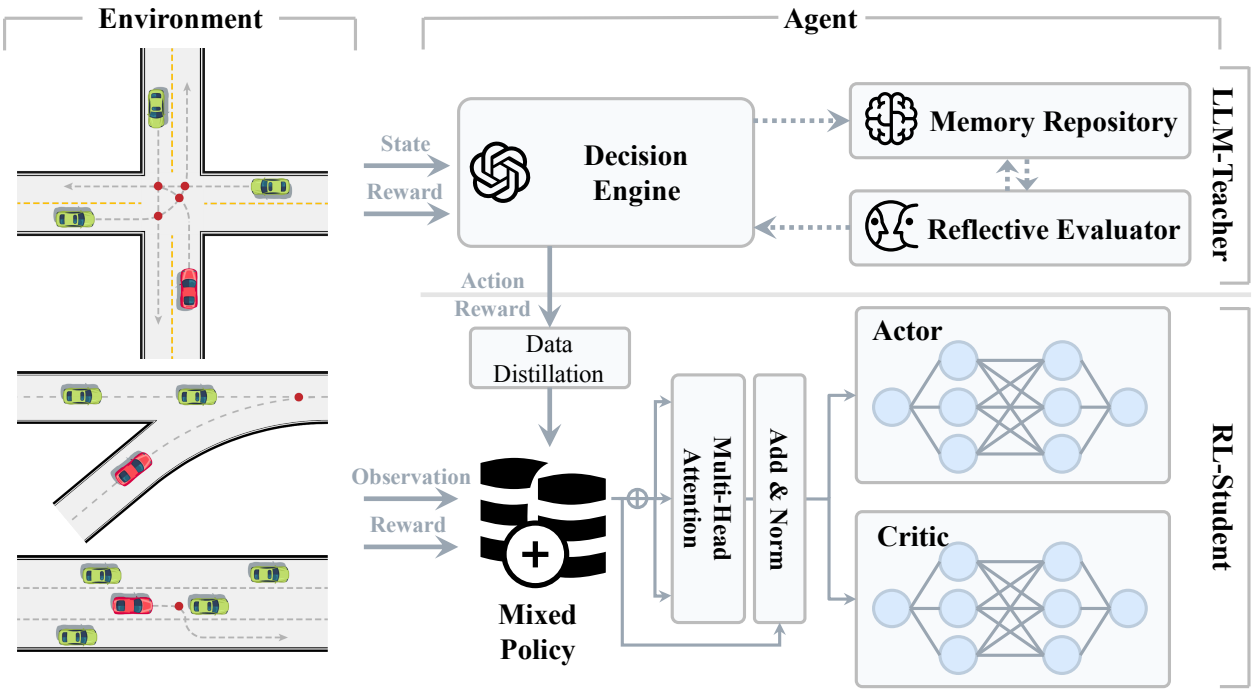

Figure 9: Demonstration of Optimization Decision-Making in Teacher-Student Models. Figure referenced from Xu et al. (2025a)

Perhaps the most critical technical challenge is ensuring consistency between high-level semantic reasoning and low-level numerical trajectories. As shown in Fig. 8, ORION (Fu et al., 2025) directly tackles this semantic-numerical alignment gap with a generative planner, improving closed-loop scores and real-world applicability (C3, C4).

Expanding the Scope of Holistic Reasoning. Beyond single-agent architectures, cutting-edge research is expanding the scope of reasoning to handle more complex interactions and scenarios. One major frontier is shifting the focus from performance in average scenarios to robust reasoning in extreme conditions. OmniDrive (Wang et al., 2024a) exemplifies this by using a dataset built on counterfactual reasoning to evaluate decision-making in rare, long-tail events, pushing the field to address the most difficult cases (C1, C3, C5). Another frontier is the expansion from single-vehicle intelligence to collective, multi-agent intelligence. V2V-LLM (Chiu et al., 2025) introduces a vehicle-to-vehicle question-answering framework that enables inter-vehicle perceptual fusion and cooperative planning. This represents a critical step towards addressing complex social interaction challenges and realizing the potential of swarm intelligence on the road (C1, C5, C7).

5.2.5 Reasoning in Supporting and Auxiliary Tasks

Beyond the core driving pipeline, reasoning is also being applied to crucial supporting tasks that form the development and validation ecosystem. These efforts are critical for creating robust and scalable solutions. The research in this area can be broadly categorized into innovations in learning paradigms and the development of reasoning-driven tools for data generation and analysis.

Learning Paradigms. Some research focuses on improving the training process itself, aiming to make learning more efficient and generalizable. As shown in Fig. 9, TeLL-Drive (Xu et al., 2025a) introduces a teacher-student framework where a powerful teacher LLM guides a more compact student deep reinforcement learning (DRL) policy. This approach improves the robustness of the learned agent and enhances its ability to transfer knowledge across different driving conditions. Such innovations address the core challenges of

reasoning-decision alignment (C4) and long-tail generalization (C5) by focusing on the scalability of the training paradigm itself, rather than just the model architecture.

Reasoning-Driven Simulation and Analysis. A major focus in auxiliary tasks is the generation and analysis of high-fidelity, diverse, and controllable scenarios, which is essential for safely testing agents in rare or dangerous situations. Several works leverage the generative capabilities of LLMs to create rich simulation environments. For instance, Editable Scene Simulation (ChatSim) (Wei et al., 2024c) allows developers to use natural language to edit and compose realistic 3D scenes for simulation training (C3, C5, C7). Similarly, Controllable Traffic Simulation (Liu et al., 2024c) uses hierarchical CoT reasoning to generate complex and controllable traffic scenarios (C3, C5, C7). In a related direction, other tools use reasoning for offline data analysis. Crash Severity Analysis (Zhen et al., 2024) applies CoT reasoning to assess traffic accident reports, providing deeper, interpretable insights into event causation and severity (C4, C5).

5.3 Discussion and Insights: Prevailing Trends and Open Questions

Prevailing Trends. This review of system-centric approaches reveals several dominant trends in the effort to integrate reasoning into autonomous systems:

- **Holistic Architectures:** A clear trajectory exists away from optimizing isolated modules and toward designing holistic, reasoning-centric architectures. Techniques like CoT are applied across the entire stack to serve as a cognitive backbone.
- **Internal Coherence:** A strong focus is placed on ensuring internal coherence. CoT and structured representations are widely adopted to bridge the gap between high-level semantic decisions and low-level executable actions, a direct response to the Decision-Reality Alignment (C4) challenge.
- **Interpretability:** An architectural shift is underway from opaque "black-box" models to interpretable "glass-box" systems. These systems provide explicit justifications for their actions, addressing demands for transparency in The Social Game (C7).
- **Open-World Adaptability:** An emerging trend targets open-world adaptability. Methods that incorporate memory (e.g., DiLu) or external tools (e.g., AgentThink) signify a move from "closed-world planners" to "open-world problem solvers" capable of handling novel Long-tail Scenarios (C5).
- **Compliance over Optimization:** A significant paradigm shift involves prioritizing compliance over pure optimization. A growing body of work explicitly embeds external constraints, such as traffic laws (C6) and social norms (C7), directly into the decision-making framework.

Open Questions and Unaddressed Gaps. Despite this rapid progress, our analysis identifies critical gaps that remain significant barriers to deployment:

- **Reconciling Deliberation and Reaction Time:** The high computational latency inherent to deliberative reasoning processes is fundamentally incompatible with the millisecond-scale reaction times required for safety-critical operations. A core open question is how to design safe and verifiable arbitration mechanisms that can reliably reconcile deep, "slow" reasoning with "fast" reactive control.
- **The Symbolic-to-Physical Grounding Gap:** High-level symbolic reasoning, such as a "safe merge" command, must be reliably grounded in vehicle control trajectories that are both physically feasible and numerically stable. However, a robust and generalizable methodology to guarantee this grounding has yet to be established. This gap between symbolic planning and sub-symbolic execution remains a primary obstacle.
- **The Formal Safety Assurance Gap:** A significant methodological gap persists in the formal verification and validation of reasoning-based agents. The field currently lacks mature frameworks to provide formal safety assurances for these systems, particularly concerning their behavior in the unbounded problem space of Long-tail Scenarios (C5), often referred to as the "unknown unknowns" challenge.
- **Modeling Implicit Social Dynamics:** Modeling complex social interactions remains in a nascent stage. Current approaches are often limited to explicit rule-following and fail to capture the nuanced, implicit communication of human agents. The development of scalable and generalizable models

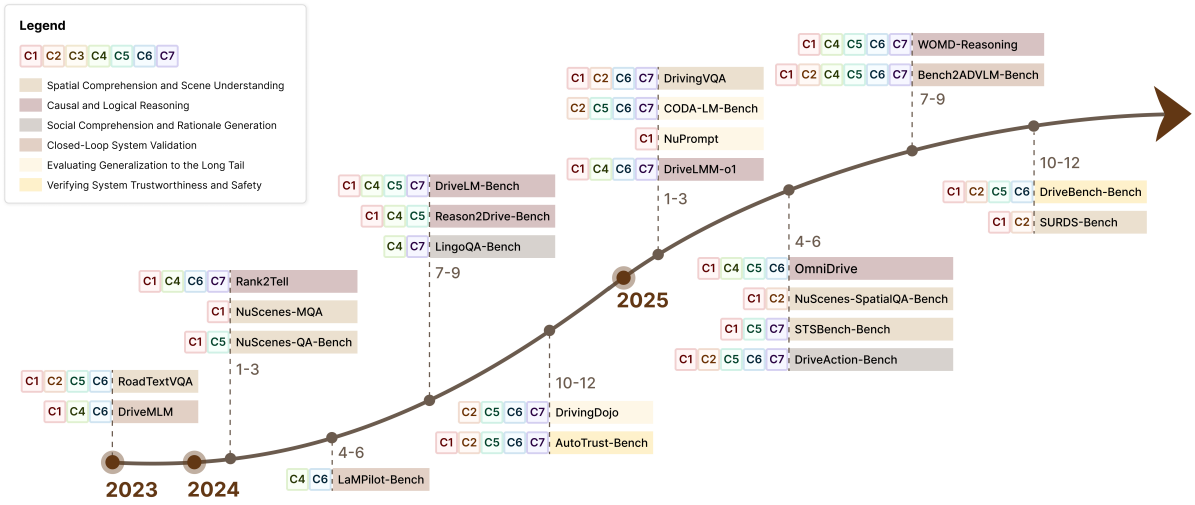

Figure 10: The chronological evolution of benchmarks and datasets for autonomous driving reasoning. This timeline illustrates the rapid acceleration of evaluation-centric research and provides the historical context for the thematic taxonomy discussed in Sec. 6.1.

capable of implicit, fluid social negotiation, as distinct from mere legal compliance, constitutes a primary research frontier.

In conclusion, the integration of reasoning has injected a powerful new paradigm into autonomous driving research, opening promising avenues for solving long-standing challenges in long-tail scenarios and social interaction. However, the true viability of these system-centric innovations hinges on the ability to rigorously and reliably measure their performance. The development of such evaluation methods is therefore a critical and parallel research frontier, which we explore in the following section.

6 Evaluation-Centric Practices: Benchmarks and Datasets

Complementing the system-centric innovations discussed previously, this section examines the parallel and equally critical frontier of evaluation. Progress in developing robust reasoning agents is intrinsically tied to the methodologies used to validate their performance. In the domain of autonomous driving reasoning, the construction of benchmarks and datasets is not merely a supporting activity; it constitutes a critical research area that delineates the boundaries for model learning and ensures rigorous scientific validation. This section provides a systematic review of these evaluation-centric practices, organized by the increasing complexity of the capabilities they are designed to measure.

6.1 A Thematic Taxonomy of Evaluation Practices

Traditional evaluation metrics for AD, such as collision rates or trajectory error, are adept at quantifying physical performance but are insufficient to assess the cognitive capabilities of a reasoning-based agent. These metrics can confirm what a system does, but often fail to reveal why it makes a certain decision or whether its underlying reasoning is sound. To address this evaluation gap, a new generation of benchmarks and datasets is emerging, which prioritize the measurement of cognitive skills over purely physical outcomes.

This section organizes a review of these evaluation-centric practices. To provide historical context for this review, Fig. 10 chronologically marks the release of each major benchmark, offering a visualization of the rapid evolution of this research domain. Our taxonomy organizes the landscape into a thematic progression of increasing complexity. We begin with benchmarks designed to evaluate foundational cognitive abilities (Sec. 6.1.1). We then examine platforms that assess system-level and interactive performance in dynamic

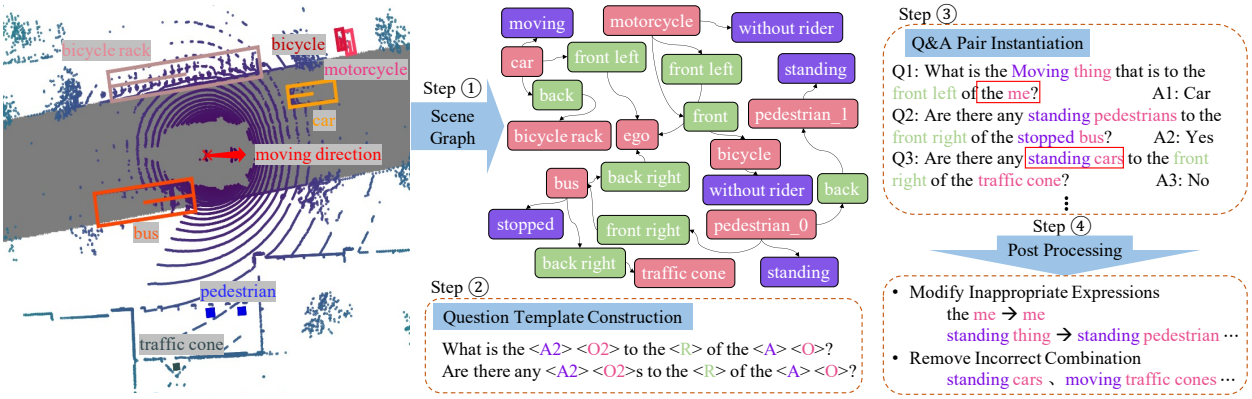

Figure 11: Data construction flow of NuScenes-QA. Figure referenced from Qian et al. (2024).

closed-loop environments (Sec. 6.1.2). Finally, we review benchmarks that focus on robustness and trustworthiness, specifically by stress-testing agents in long-tail and safety-critical scenarios (Sec. 6.1.3). Throughout this taxonomy, we systematically connect the discussed benchmarks and datasets to the core challenges (C1–C7) identified in Sec. 4.

6.1.1 Evaluating Foundational Cognitive Abilities

The ability to perform complex actions in the real world is predicated on a set of foundational cognitive skills. We review benchmarks designed to measure these core competencies, the essential precursors to intelligent driving behavior. These evaluations typically operate in an open-loop, question-answering format, focusing on what the agent understands about a static or pre-recorded scene, rather than how it acts within it.

Spatial Comprehension and Scene Understanding. This category of benchmarks evaluates an agent's ability to accurately perceive and model three-dimensional spatial relationships, a fundamental requirement for any physical agent. As a pioneering work, NuScenes-QA-Bench (Qian et al., 2024) established the first large-scale VQA benchmark for autonomous driving as shown in Fig. 11 (C1, C5). This foundation has inspired a family of more advanced evaluations. For instance, NuScenes-MQA (Inoue et al., 2024) scales the volume of questions (C1), while NuScenes-SpatialQA-Bench (Tian et al., 2025b) focuses specifically on complex spatial reasoning by generating answers directly from ground-truth LiDAR data, thereby mitigating the impact of perception model errors (C1, C2). To further refine this evaluation, SURDS-Bench (Guo et al., 2025) systematically assesses fine-grained spatial capabilities across distinct categories such as orientation and depth (C1, C2). This area is also supplemented by other benchmarks with specialized focuses, including RoadTextVQA (Tom et al., 2023) for scene text comprehension (C1, C2, C5, C6) and STSBench-Bench (Fruhwirth-Reisinger et al., 2025) for evaluating spatio-temporal reasoning in complex multi-agent interactions (C1, C5, C7).

Causal and Logical Reasoning. Moving beyond spatial awareness ("what" and "where"), this class of benchmarks assesses an agent's ability to perform multi-step and logical inference to understand "how" and "why" events unfold. The primary methodology here is the evaluation of explicit reasoning chains. For example, DriveLM-Bench (Sima et al., 2024a) adopts a graph-structured VQA framework that spans the entire pipeline from object detection to planning decisions (C1, C4, C6). Similarly, Reason2Drive-Bench (Nie et al., 2024) provides over 600,000 video-text pairs that explicitly decompose the driving process into perception, prediction, and reasoning steps (C1, C4, C5). Other datasets, like DriveLMM-o1 (Ishaq et al., 2025) (C1, C4, C6, C7) and WOMD-Reasoning (Li et al., 2024b) (C1, C4, C5, C6, C7), provide millions of annotated question-answer pairs focused on reasoning about interactions governed by traffic rules and human intentions. DrivingVQA (Corbière et al., 2025) specifically tests rule-based logical reasoning by deriving questions from official driving theory exams (C1, C2, C6, C7). More advanced forms of reasoning are also targeted; as shown in Fig. 12, OmniDrive (Wang et al., 2025b) employs counterfactual reasoning

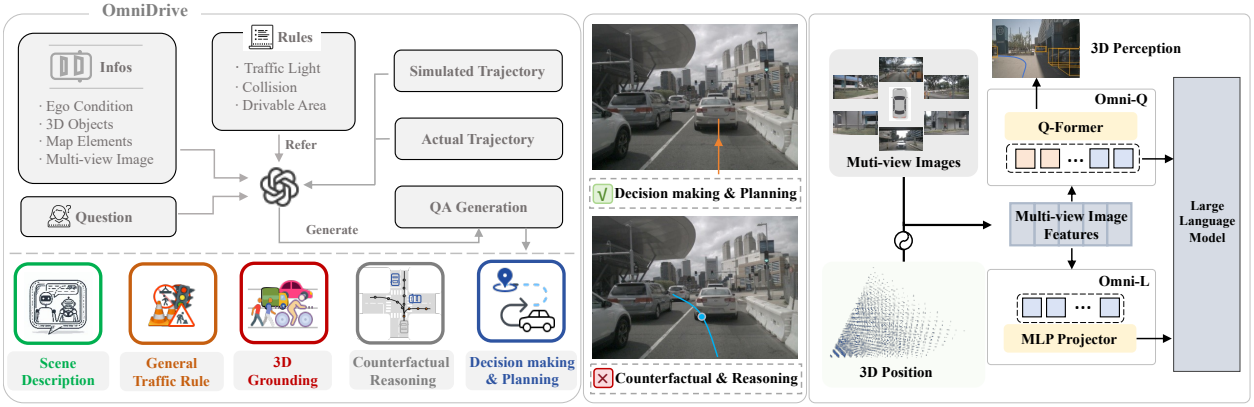

Figure 12: OmniDrive is a holistic vision-language dataset for autonomous driving. Figure referenced from Wang et al. (2025b).

to link plans with explanations (C1, C4, C5, C6), while Rank2Tell (Sachdeva et al., 2024) is designed for multimodal importance ranking and rationale generation (C1, C4, C6, C7). Collectively, these benchmarks test an agent's logical consistency and its capacity for causal explanation.

Social Comprehension and Rationale Generation. This final category of foundational benchmarks evaluates the highest level of cognitive skill: understanding the intentions of other agents and justifying the ego-vehicle's own actions. These capabilities are crucial for ensuring social compliance and building human trust (C7). LingoQA-Bench (Marcu et al., 2024) provides nearly 420,000 action-justification VQA pairs, requiring a model not only to predict an action but also to explain the reason behind it (C4, C7). Similarly, DriveAction-Bench (Hao et al., 2025) employs a tree-structured evaluation framework to assess whether a model's decisions are both human-like and logically sound (C1, C2, C5, C6, C7). These benchmarks collectively push models beyond being passive observers of the environment towards becoming active, explainable, and socially aware participants in traffic.

6.1.2 Evaluating System-Level and Interactive Performance

While foundational benchmarks are essential for assessing cognitive skills in isolation, the true viability of a driving agent can only be determined by evaluating its performance "in the loop," where its actions have direct and immediate consequences. This section reviews the platforms and datasets designed for this stage of validation.

Closed-Loop System Validation. A system-level evaluation is to ensure that an agent's high-level reasoning translates into physically executable and effective low-level control. This directly tests Decision-Reality Alignment (C4). Benchmarks in this category provide comprehensive, long-duration data that allows for the validation of the entire system stack, from perception to action. An example is DriveMLM (Guo et al., 2024), which offers a large-scale dataset of 280 hours of driving data with multi-modal inputs (C1, C4, C6). LaMPilot-Bench (Ma et al., 2024b) integrates Large Language Models into autonomous driving systems by generating executable code as driving policies, evaluating language-driven autonomous driving agents (C4, C6). The explicit goal of such datasets is to provide the necessary resources to bridge the critical gap between abstract, language-based decisions and the continuous control signals required for safe driving.

Interactive and Critical Scenarios. Beyond ensuring internal coherence, a robust evaluation must also assess how the agent performs when interacting with other dynamic agents, especially in safety-critical situations. The focus here shifts from general driving to the agent's ability to adapt and react under pressure. Bench2ADVLM-Bench (Zhang et al., 2025a) exemplifies this approach by offering a dedicated benchmark specifically designed for closed-loop testing in 220 curated "threat-critical" scenarios (C1, C2, C4, C5, C6,

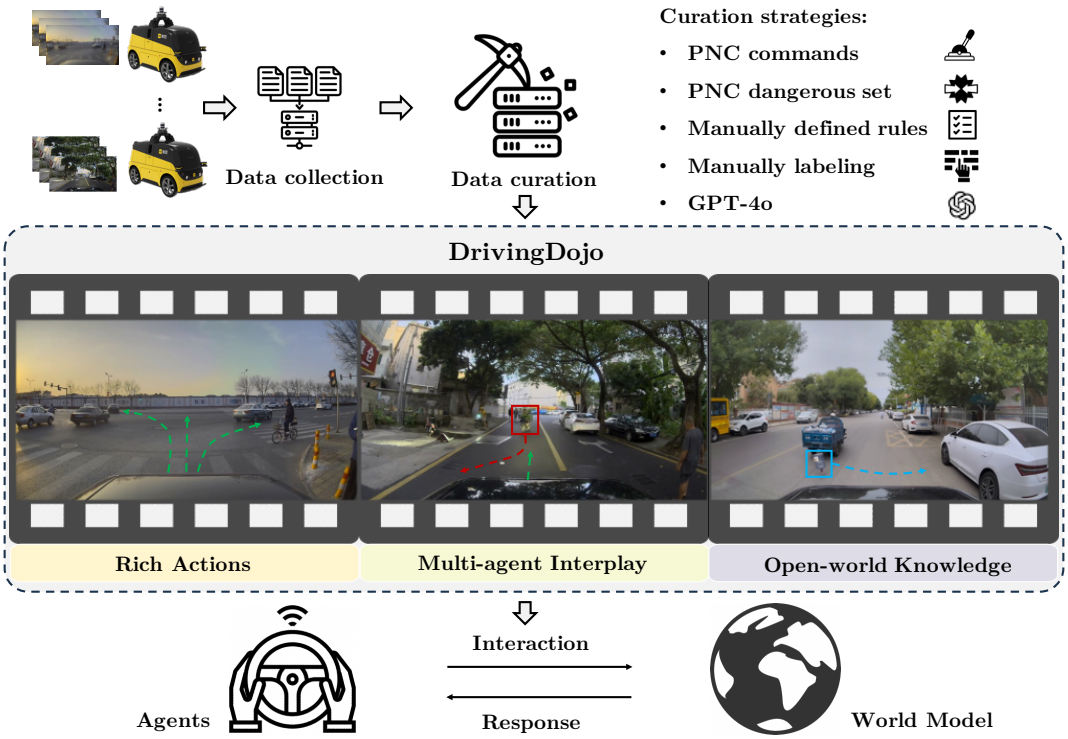

Figure 13: Enhancing interactive and knowledge-enriched learning of world models. Figure referenced from Wang et al. (2024e).

C7). This scenario-based methodology, often operationalized within interactive simulation platforms, is essential for rigorously assessing a model's performance and adaptability when safety is on the line.

6.1.3 Evaluating Robustness and Trustworthiness

Building upon the assessment of system performance in standard interactive settings, this final stage of evaluation moves to the most demanding frontier: assessing an agent's behavior under stress, in the face of novelty, and against its own internal biases. These benchmarks are designed to move beyond average-case performance and probe the limits of a system's reliability, which is a prerequisite for establishing public trust and ensuring safe deployment.

Evaluating Generalization to the Long Tail. A fundamental limitation of any data-driven model is its performance on events that are rare in its training distribution. We review benchmarks designed specifically to measure robustness against such long-tail scenarios, directly addressing a core challenge for autonomous driving (C5). These datasets provide a controlled environment for testing a model's ability to generalize its reasoning to novel situations. For example, as shown in Fig. 13, DrivingDojo (Wang et al., 2024e) is an interactive dataset curated with thousands of video clips containing rare events like crossing animals and falling debris, as well as complex multi-agent interactions such as cut-ins (C2, C5, C6, C7). Similarly, CODA-LM-Bench (Chen et al., 2025) provides extensive annotations for corner cases involving abnormal pedestrian behavior and unique traffic signs (C2, C5, C6, C7). Other works, such as NuPrompt (Wu et al., 2025a), address this challenge by introducing language-prompted 3D object tracking, which tests and improves a model's ability to generalize to unseen or rare attribute combinations (C1,C5).

Verifying System Trustworthiness and Safety. Beyond handling external novelty, a reliable agent must also possess intrinsic trustworthiness. This category features benchmarks that move beyond conventional accuracy metrics to systematically evaluate system safety, reliability, and robustness against internal

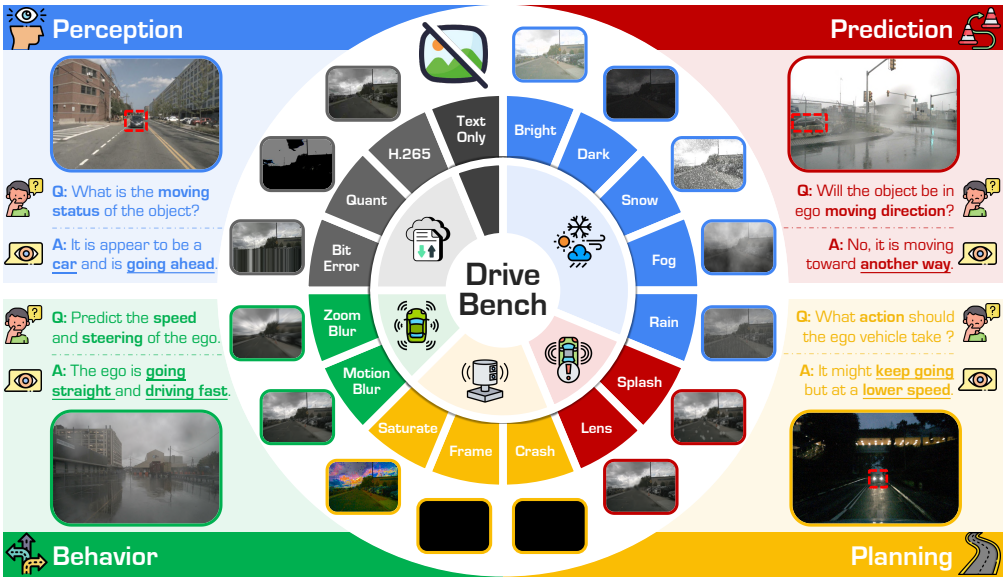

Figure 14: Overview of Drive Bench. Figure referenced from Xie et al. (2025).

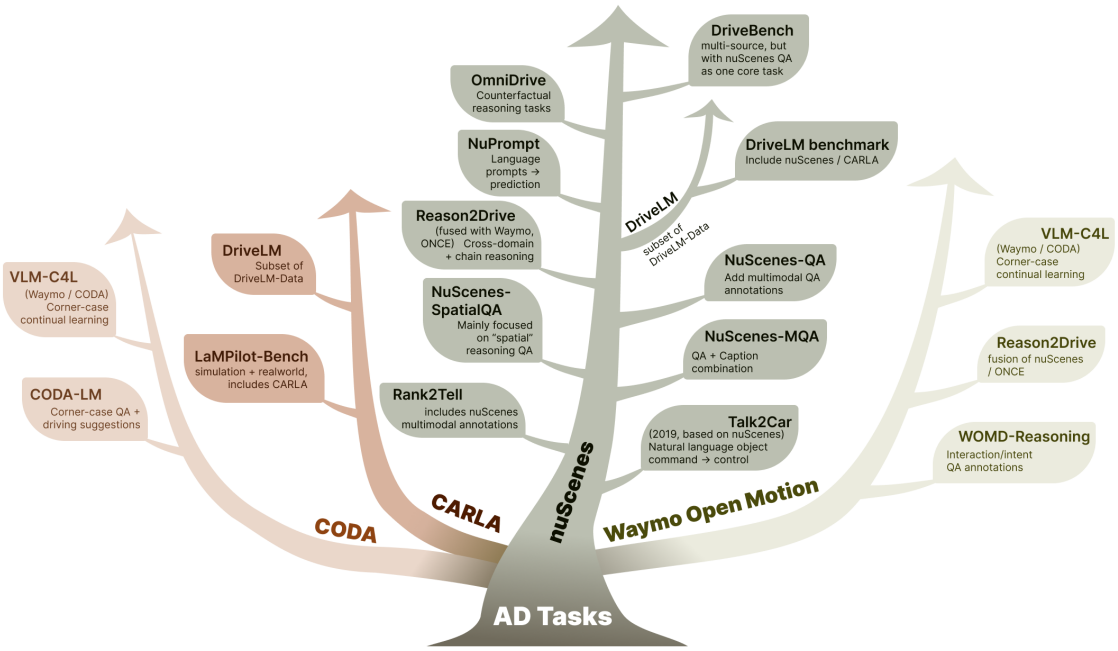

Figure 15: A genealogical chart of the autonomous driving benchmark ecosystem. We illustrate the derivative and inheritance relationships among key datasets, highlighting how foundational platforms (e.g., NuScenes) have inspired a "family" of specialized benchmarks designed to evaluate specific reasoning capabilities.

failures. This directly corresponds to the core challenges of Perception-Cognition Bias (C2) and Regulatory Compliance (C6). AutoTrust-Bench (Xing et al., 2024) introduces the first comprehensive benchmark for VLM trustworthiness in driving, establishing a systematic protocol that covers robustness to data corruption, fairness, privacy, and safety evaluation (C1, C2, C5, C6, C7). Complementing this, as shown in Fig. 14, DriveAction-Bench (Xie et al., 2025) focuses specifically on model reliability and corruption robustness by testing the full pipeline under a variety of challenging conditions (C1, C2, C5, C6, C7). Furthermore, other datasets contribute to this theme by design. For example, DrivingVQA (Corbière et al., 2025) (C1, C2, C6,

Table 1: A comprehensive overview and comparison of key benchmarks for autonomous driving reasoning. This table summarizes the primary attributes of each benchmark, including its publication **Date**, core reasoning **Challenges** (C1–C7) it addresses, associated **Keywords**, **Data Scale**, and community engagement metrics (GitHub **Stars**). Repository **Links** are also provided for reference.

Date	Dataset	Challenges	Keywords	Data Scale	Stars	Link
2023.08	RoadTextVQA (Tom et al., 2023)	C1,C2,C5,C6	Driving VideoQA; Scene Text; Road Signs	3,222 videos / 10,500 QA	5	
2023.12	DriveMLM (Guo et al., 2024)	C1,C4,C6	Closed-loop Driving; Multi-modal Input; Decision Alignment	280h driving / 50k routes / 30 scenarios / 4 cams + LiDAR	179	
2024.01	Rank2Tell (Sachdeva et al., 2024)	C1,C4,C6,C7	Importance Ranking; Language Explanations; Ego-centric	116 clips / 3 cams + LiDAR + CAN	-	-
2024.01	NuScenes-MQA (Inoue et al., 2024)	C1	3D VQA; Markup-QA; Spatial Reasoning	1.46M QA / 34k scenarios / 6 cams	31	
2024.02	NuScenes-QA-Bench (Qian et al., 2024)	C1,C5	Multi-modal VQA; 3D Scene Graph	34K scenes / 460K QA	210	
2024.06	LaMPilot-Bench (Ma et al., 2024b)	C4,C6	Language Model Programs ; Driving Instructions; Code-as-Policy	4,900 scenes (500 test)	31	
2024.09	DriveLM-Bench (Sima et al., 2024a)	C1,C4,C5,C7	Graph VQA; End-to-end Autonomous Driving	19,200 frames / 20,498 QA pairs	1.2k	🤗
2024.09	LingoQA-Bench (Marcu et al., 2024)	C4,C7	ActionJustification VQA; End-to-End Benchmark	28K scenarios / 419.9K QA pairs	183	
2024.09	Reason2Drive-Bench (Nie et al., 2024)	C1,C4,C5	Chain-based Reasoning; Cross Datasets	600K+ video-text pairs	92	
2024.12	DrivingDojo (Wang et al., 2024e)	C2,C5,C6,C7	Driving World Model; Interactive Dataset	18k videos	73	🤗
2024.12	AutoTrust-Bench (Xing et al., 2024)	C1,C2,C5,C6,C7	Trust/safety/robustness/privacy/fairness	10K+ driving scenes; 18K+ QA pairs	48	
2025.01	DrivingVQA (Corbière et al., 2025)	C1,C2,C6,C7	VQA; Chain-of-Thought	3,931 multiple-choice problems	0	🤗
2025.02	CODA-LM-Bench (Chen et al., 2025)	C2,C5,C6,C7	Corner cases; LVLM evaluation	9,768 scenarios; over 63K annotations	92	🤗
2025.02	NuPrompt (Wu et al., 2025a)	C1,C5	Language Prompt; 3D Perception; Multi-Object Tracking (MOT)	40.1K language prompts / avg. 7.4 tracklets per prompt	149	
2025.03	DriveLMM-o1 (Ishaq et al., 2025)	C1,C4,C6,C7	Step-by-Step Reasoning	18k train / 4k test QA	66	🤗
2025.04	NuScenes-SpatialQA-Bench (Tian et al., 2025b)	C1,C2	Spatial Reasoning	3.5M QA	18	🤗
2025.06	STSBench-Bench (Fruhwirth-Reisinger et al., 2025)	C1,C5,C7	Spatio-temporal Reasoning	971 MCQs / 43 scenes	10	🤗
2025.06	DriveAction-Bench (Hao et al., 2025)	C1,C2,C5,C6,C7	Tree-structured Evaluation; Human-like Decisions	16,185 QA / 2,610 scenes	-	🤗
2025.06	OmniDrive (Wang et al., 2025b)	C1,C4,C5,C6	Counterfactual reasoning	-	486	
2025.07	WOMD-Reasoning (Li et al., 2024b)	C1,C4,C5,C6,C7	Interaction Reasoning; Traffic Rules	3M QA	37	
2025.08	Bench2ADVLM-Bench (Zhang et al., 2025a)	C1,C2,C4,C5,C6,C7	Closed-loop evaluation; Dual-system adaptation	220 threat-critical scenarios	2	
2025.10	DriveBench-Bench (Xie et al., 2025)	C1,C2,C5,C6	Reliability; Corruption Robustness	19,200 frames / 20,498 QA	112	🤗
2025.12	SURDS-Bench (Guo et al., 2025)	C1,C2	Spatial Reasoning; Spatial Understanding; Driving VQA	41,080 train QA/ 9,250 test QA	54	

C7) and DriveLMM-o1 (Ishaq et al., 2025) (C1, C4, C6, C7) enhance safety by grounding decisions in explicit traffic rules, while NuScenes-SpatialQA-Bench (Tian et al., 2025b) improves robustness by forcing models to reason from ground-truth sensor data rather than relying on potentially fallible intermediate perception modules (C1, C2).

6.2 Comprehensive Benchmark Dataset Landscape

To synthesize the preceding categorical discussion and provide a high-level, comparative view, this subsection presents resources that map the evaluation ecosystem. We have constructed a dataset genealogy in Fig. 15. This chart reveals the derivative and inheritance relationships between different benchmarks, illustrating how foundational datasets like NuScenes have spawned an entire "family" of specialized evaluations. This

genealogical view, complemented by a comprehensive comparison table (see Table 1), serves as a quick reference for researchers to navigate the relationships and core attributes of existing benchmarks.

6.3 Discussion and Insights: Prevailing Trends and Open Questions

Prevailing Trends. Our review of the evaluation-centric landscape highlights several key trends that shape how the field measures progress:

- **From Physical Outcomes to Cognitive Processes:** A clear shift is underway from metrics based on physical outcomes (e.g., collision rates, trajectory error) to evaluations targeting the cognitive process itself. Benchmarks increasingly focus on assessing the quality of rationale generation, the logical consistency of reasoning chains, and the understanding of causality, directly addressing the need for interpretability (C7).
- **From Static VQA to Dynamic Simulation:** The evaluation paradigm is evolving from static, open-loop, question-answering formats (e.g., NuScenes-QA) to dynamic, closed-loop, interactive simulations. This progression is essential for assessing system-level behavior and the critical Responsiveness-Reasoning Tradeoff (C3).
- **Deliberate Curation of the Long Tail:** There is a growing recognition that performance on "average" scenarios is insufficient. A dominant trend is the deliberate and resource-intensive curation of benchmarks (e.g., DrivingDojo, CODA-LM) specifically designed to test generalization and robustness in rare, safety-critical, and Long-tail Scenarios (C5).
- **Emergence of Holistic Trustworthiness:** Evaluation is moving beyond simple task accuracy. New benchmarks (e.g., AutoTrust-Bench) establish comprehensive protocols to measure holistic trustworthiness, including robustness to data corruption (C2), fairness, and adherence to safety and Regulatory Compliance (C6).

Open Questions and Future Directions. Despite this progress, the evaluation of reasoning agents still faces significant unsolved challenges:

- **The Methodological Scalability of Evaluation:** The reliance on finite, curated benchmarks, even for long-tail events, represents a methodological limitation. These benchmarks primarily validate performance against a pre-defined set of scenarios. A fundamental gap persists in developing generative evaluation processes that can automatically and scalably probe for novel system failure modes. Such methodologies are essential to progress from testing "known unknowns" to addressing the more complex challenge of "unknown unknowns" inherent to Long-tail Scenarios (C5).
- **Measuring Complex Social Dynamics:** The assessment of The Social Game (C7) remains in a nascent stage. Current benchmarks are restricted to evaluating explicit rationale generation or simplified intent inference within highly constrained contexts. A significant open question is how to formulate reproducible and scalable metrics for the implicit, fluid, and multi-agent negotiation that defines true social compliance, as distinct from mere adherence to explicit legal rules.
- **The Predictive Validity of Open-Loop Metrics:** A significant disconnect persists between model performance on open-loop, foundational benchmarks and demonstrated competence in closed-loop, interactive environments. The predictive validity of high scores on isolated cognitive tasks, such as static VQA, remains an unverified hypothesis. It is unclear if this performance effectively translates to a system's ability to manage real-time Decision-Reality Alignment (C4).
- **Simulation Fidelity and Resource Overhead:** The practical utility of high-fidelity, interactive evaluation is constrained by its significant computational and financial resource overhead. A formidable engineering challenge persists in developing evaluation platforms that are simultaneously scalable, photorealistic, and physically accurate. This is particularly true for platforms that must model the complex sensor interactions (C1) and environmental degradations, such as adverse weather, associated with Perception-Cognition Bias (C2).

In summary, the true viability of these system-centric innovations hinges on the ability to rigorously and reliably measure their performance. The development of such evaluation methods is therefore a critical and parallel research frontier, which will be essential for guiding the field toward truly intelligent and deployable autonomous systems.

7 Conclusion and Future Work

This survey argues that the progression toward higher levels of vehicle autonomy is fundamentally a problem of reasoning. We introduce a taxonomy that deconstructs this problem into seven core challenges (C1–C7), spanning from Egocentric Reasoning (C1–C4) to the more complicated Social-Cognitive (C5–C7) domain. Our analysis posits that while egocentric challenges represent profound engineering hurdles, the ultimate bottleneck to achieving human-like operational capabilities lies in mastering the social-cognitive challenges.

Our review reveals a clear, accelerating response to this paradigm shift. In system-centric approaches (Sec. 5), research moves decisively from optimizing isolated modules toward holistic, reasoning-centric architectures. The shift to interpretable "glass-box" agents underscores this trend. In parallel, evaluation-centric practices (Sec. 6) are evolving. Focus is shifting from physical outcomes to assessing cognitive processes, rationale generation, and robustness in curated long-tail scenarios.

Despite progress in methodology and evaluation, our analysis concludes a fundamental tension remains unresolved. This tension exists between the powerful, deliberative, but high-latency symbolic reasoning offered by large models, and the millisecond-scale, physically-grounded, and safety-critical demands of real-world vehicle control. Reconciling these two worlds (the abstract and the physical, the deliberative and the reactive) is the central, most pressing objective for future AD research. Based on the critical gaps identified in our analysis, we outline the following primary directions for future research that are essential for bridging the gap from reasoning-capable models to road-ready AD agents.

Verifiable Neuro-Symbolic Architectures. The Responsiveness-Reasoning Tradeoff (C3) and the Decision-Reality Alignment (C4) gap represent the immediate architectural barriers. Future work should move beyond simple dual-process concepts (i.e., "fast" and "slow" thinking). A critical frontier is developing verifiable neuro-symbolic architectures. These systems must not only provide an arbitration mechanism to manage latency but also offer formal assurances that abstract symbolic plans (e.g., "yield to the merging vehicle") are reliably, safely grounded in the sub-symbolic, continuous control space of the vehicle.

Robust Reasoning Under Multi-Modal Uncertainty. A foundational challenge remains in how reasoning engines handle the noisy, asynchronous, and often contradictory data from heterogeneous signals (C1). While many models assume clean inputs, real-world operation demands robustness to the Perception-Cognition Bias (C2). Future research should develop architectures that can explicitly reason about data uncertainty, fuse conflicting evidence, and perform compensatory inference to maintain a stable world model, especially during sensor degradation or failure in adverse conditions.

Dynamic Grounding in External Regulatory Knowledge. True Regulatory Compliance (C6) on a global scale requires more than pre-trained knowledge. Current models lack a mechanism to adapt to the varied legal frameworks of different jurisdictions. Future work should explore "open-world" systems that can dynamically query, retrieve, and interpret knowledge bases of formal traffic laws. A key challenge will be grounding this symbolic, legal reasoning into the agent's real-time decision-making process.

Generative and Adversarial Evaluation. Tackling Long-tail Scenarios (C5) and the methodological limitations of evaluation highlight a critical need for new verification and validation paradigms. Current benchmarks, while valuable, are finite and curated, primarily testing against "known unknowns." Future research should invest in generative and adversarial evaluation frameworks. Such systems would leverage simulation and world models not just to test agents, but to automatically discover novel, systemic failure modes, thereby moving the field closer to addressing the challenge of "unknown unknowns."

Scalable Models for Implicit Social Negotiation. Mastering The Social Game (C7) remains the ultimate research frontier. Current approaches, in both methods and evaluation, are largely nascent and often limited to explicit Regulatory Compliance (C6). The next generation of research must move beyond legal compliance to model the fluid, implicit, and non-verbal social game of human interaction. This requires developing scalable models that can infer latent human intent, negotiate multi-agent interactions, and generate behavior that is not only safe but also socially legible and predictable to other road users.

References

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023.

Mehdi Azarafza, Mojtaba Nayyeri, Charles Steinmetz, Steffen Staab, and Achim Rettberg. Hybrid reasoning based on large language models for autonomous car driving. In *ICCMA*, pp. 14–22. IEEE, 2024.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*, pp. 19107–19117. IEEE, 2022.

Claudine Badue, Rânik Guidolini, Raphael Vivacqua Carneiro, Pedro Azevedo, Vinicius B. Cardoso, Avelino Forechi, Luan F. R. Jesus, Rodrigo Ferreira Berriel, Thiago M. Paixão, Filipe Wall Mutz, Lucas de Paula Veronese, Thiago Oliveira-Santos, and Alberto F. De Souza. Self-driving cars: A survey. *Expert Syst. Appl.*, 165:113816, 2021.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. In *AAAI*, pp. 17682–17690. AAAI Press, 2024.

Bikram Pratim Bhuyan, Amar Ramdane-Cherif, Ravi Tomar, and T. P. Singh. Neuro-symbolic artificial intelligence: a survey. *Neural Comput. Appl.*, 36(21):12809–12844, 2024.

Daniel Bogdoll, Maximilian Nitsche, and J. Marius Zöllner. Anomaly detection in autonomous driving: A survey. In *CVPR Workshops*, pp. 4487–4498. IEEE, 2022.

Alexandra M Boggs, Ramin Arvin, and Asad J Khattak. Exploring the who, what, when, where, and why of automated vehicle disengagements. *Accident Analysis & Prevention*, 136:105406, 2020.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars. *CoRR*, abs/1604.07316, 2016.

Frédéric Bouchard, Sean Sedwards, and Krzysztof Czarnecki. A rule-based behaviour planner for autonomous driving. In *RuleML+RR*, volume 13752 of *Lecture Notes in Computer Science*, pp. 263–279. Springer, 2022.

Nicholas E. Brown, Farhang Motallebi Araghi, Johan Fanas Rojas, Sherif Ayantayo, Richard T. Meyer, Zachary D. Asher, Ali Riza Ekti, Chieh Ross Wang, Nicholas A. Goberville, and Ben Feinberg. Evaluation of autonomous vehicle sensing and compute load on a chassis dynamometer. In *ITSC*, pp. 1989–1995. IEEE, 2023.

Tianhui Cai, Yifan Liu, Zewei Zhou, Haoxuan Ma, Seth Z. Zhao, Zhiwen Wu, and Jiaqi Ma. Driving with regulation: Interpretable decision-making for autonomous vehicles with retrieval-augmented reasoning via LLM. *CoRR*, abs/2410.04759, 2024.

Kai Chen, Yanze Li, Wenhua Zhang, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. Automated evaluation of large vision-language models on self-driving corner cases. In *WACV*, pp. 7817–7826. IEEE, 2025.

Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *ICRA*, pp. 14093–14100. IEEE, 2024.

Qiping Chen, Yinfei Xie, Shifeng Guo, Jie Bai, and Qiang Shu. Sensing system of environmental perception technologies for driverless vehicle: A review of state of the art and challenges. *Sensors and Actuators A: Physical*, 319:112566, 2021.

Hsu-Kuang Chiu, Ryo Hachiuma, Chien-Yi Wang, Stephen F. Smith, Yu-Chiang Frank Wang, and Min-Hung Chen. V2V-LLM: vehicle-to-vehicle cooperative autonomous driving with multi-modal large language models. *CoRR*, abs/2502.09980, 2025.

Charles Corbière, Simon Roburin, Syrielle Montariol, Antoine Bosselut, and Alexandre Alahi. DRIVINGVQA: analyzing visual chain-of-thought reasoning of vision language models in real-world scenarios with driving theory tests. *CoRR*, abs/2501.04671, 2025.

Antonia Creswell, Murray Shanahan, and Irina Higgins. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *ICLR*. OpenReview.net, 2023.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *ACL (Findings)*, pp. 3563–3578. Association for Computational Linguistics, 2024.

Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. Hilm-d: Enhancing mllms with multi-scale high-resolution details for autonomous driving. *Int. J. Comput. Vis.*, 133(8):5379–5395, 2025a.

Yifu Ding, Wentao Jiang, Shunyu Liu, Yongcheng Jing, Jinyang Guo, Yingjie Wang, Jing Zhang, Zengmao Wang, Ziwei Liu, Bo Du, Xianglong Liu, and Dacheng Tao. Dynamic parallel tree search for efficient LLM reasoning. In *ACL (1)*, pp. 11233–11252. Association for Computational Linguistics, 2025b.

Vinayak Dixit, Sai Chand, and Divya Nair. Autonomous vehicles: Disengagements, accidents and reaction times. *PLOS ONE*, 11:e0168054, 12 2016. doi: 10.1371/journal.pone.0168054.

Alexey Dosovitskiy, Germán Ros, Felipe Codevilla, Antonio M. López, and Vladlen Koltun. CARLA: an open urban driving simulator. In *CoRL*, volume 78 of *Proceedings of Machine Learning Research*, pp. 1–16. PMLR, 2017.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8469–8488. PMLR, 2023.

Christian Fruhwirth-Reisinger, Dusan Malic, Wei Lin, David Schinagl, Samuel Schulter, and Horst Possegger. Stsbench: A spatio-temporal scenario benchmark for multi-modal large language models in autonomous driving. *CoRR*, abs/2506.06218, 2025.

Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. In *WACV (Workshops)*, pp. 910–919. IEEE, 2024.

Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. ORION: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. *CoRR*, abs/2503.19755, 2025.

Anurag Ghosh, Robert Tamburo, Shen Zheng, Juan R. Alvarez-Padilla, Hailiang Zhu, Michael Cardei, Nicholas Dunn, Christoph Mertz, and Srinivasa G. Narasimhan. Roadwork dataset: Learning to recognize, observe, analyze and drive through work zones. *CoRR*, abs/2406.07661, 2024.

Sorin Mihai Grigorescu, Bogdan Trasnea, Tiberiu T. Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *J. Field Robotics*, 37(3):362–386, 2020.

Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Chenming Zhang, Shuai Liu, and Long Chen. Drivemllm: A benchmark for spatial understanding with multimodal large language models in autonomous driving. *CoRR*, abs/2411.13112, 2024.

Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Dujun Nie, Wenke Huang, Chenming Zhang, Shuai Liu, Hao Zhao, and Long Chen. Surds: Benchmarking spatial understanding and reasoning in driving scenarios with vision language models. In *NeurIPS*, 2025.

Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, and Pheng-Ann Heng. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *CoRR*, abs/2309.00615, 2023.

Longzhen Han, Awes Mubarak, Almas Baimagambetov, Nikolaos Polatidis, and Thar Baker. A survey of generative categories and techniques in multimodal large language models. *CoRR*, abs/2506.10016, 2025.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *EMNLP*, pp. 8154–8173. Association for Computational Linguistics, 2023.

Yuhan Hao, Zhengning Li, Lei Sun, Weilong Wang, Naixin Yi, Sheng Song, Caihong Qin, Mofan Zhou, Yifei Zhan, Peng Jia, and Xianpeng Lang. Driveaction: A benchmark for exploring human-like driving decisions in VLA models. *CoRR*, abs/2506.05667, 2025.

Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

Jiayi He, Hehai Lin, Qingyun Wang, Yi R. Fung, and Heng Ji. Self-correction is more than refinement: A learning framework for visual and language reasoning tasks. In *ACL (Findings)*, pp. 6405–6421. Association for Computational Linguistics, 2025.

Xiangkun He and Chen Lv. Towards safe autonomous driving: Decision making with observation-robust reinforcement learning. *Automotive Innovation*, 6(4):509–520, 2023.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In *NeurIPS*, 2023.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *CVPR*, pp. 17853–17862. IEEE, 2023a.

Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *CVPR*, pp. 17853–17862. IEEE, 2023b.

Jiaxing Huang and Jingyi Zhang. A survey on evaluation of multimodal large language models. *CoRR*, abs/2408.15769, 2024.

Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. In *ACL (Findings)*, pp. 1049–1065. Association for Computational Linguistics, 2023.

Ting Huang, Zeyu Zhang, and Hao Tang. 3d-r1: Enhancing reasoning in 3d vlms for unified scene understanding. *CoRR*, abs/2507.23478, 2025.

Yu Huang and Yue Chen. Autonomous driving with deep learning: A survey of state-of-art technologies. *CoRR*, abs/2006.06091, 2020.

Zilin Huang, Zihao Sheng, Yansong Qu, Junwei You, and Sikai Chen. VLM-RL: A unified vision language models and reinforcement learning framework for safe autonomous driving. *CoRR*, abs/2412.15544, 2024.

Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. EMMA: end-to-end multimodal model for autonomous driving. *Trans. Mach. Learn. Res.*, 2025, 2025.

Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do as I can, not as I say: Grounding language in robotic affordances. In *CoRL*, volume 205 of *Proceedings of Machine Learning Research*, pp. 287–318. PMLR, 2022.

Yuichi Inoue, Yuki Yada, Kotaro Tanahashi, and Yu Yamaguchi. Nuscenes-mqa: Integrated evaluation of captions and QA for autonomous driving datasets using markup annotations. In *WACV (Workshops)*, pp. 930–938. IEEE, 2024.

Ayesha Ishaq, Jean Lahoud, Ketan More, Omkar Thawakar, Ritesh Thawkar, Dinura Dissanayake, Noor Ahsan, Yuhao Li, Fahad Shahbaz Khan, Hisham Cholakkal, Ivan Laptev, Rao Muhammad Anwer, and Salman H. Khan. Drivelmm-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding. *CoRR*, abs/2503.10621, 2025.

Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Zheng Li, Ruirui Li, Xianfeng Tang, Suhang Wang, Yu Meng, and Jiawei Han. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. In *ACL (Findings)*, pp. 163–184. Association for Computational Linguistics, 2024.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022.

Philip Koopman and Michael Wagner. Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intell. Transp. Syst. Mag.*, 9(1):90–96, 2017.

Alexander Krentsel, Peter Schafhalter, Joseph E. Gonzalez, Sylvia Ratnasamy, Scott Shenker, and Ion Stoica. Managing bandwidth: The key to cloud-assisted autonomous driving. *CoRR*, abs/2410.16227, 2024.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: reasoning segmentation via large language model. In *CVPR*, pp. 9579–9589. IEEE, 2024.

Xiao Li, Kaiwen Liu, H. Eric Tseng, Anouck Girard, and Ilya V. Kolmanovsky. Interaction-aware decision-making for autonomous vehicles in forced merging scenario leveraging social psychology factors. In *ACC*, pp. 285–291. IEEE, 2024a.

Xiaofeng Li, Alexander Weber, Adrian Cottam, and Yao-Jan Wu. Impacts of changing from permissive/protected left-turn to protected-only phasing: Case study in the city of tucson, arizona. *Transportation Research Record: Journal of the Transportation Research Board*, 2673(4):616626, 2019.

Yiheng Li, Cunxin Fan, Chongjian Ge, Zhihao Zhao, Chenran Li, Chenfeng Xu, Huaxiu Yao, Masayoshi Tomizuka, Bolei Zhou, Chen Tang, et al. Womd-reasoning: A large-scale dataset for interaction reasoning in driving. *arXiv preprint arXiv:2407.04281*, 2024b.

Yiheng Li, Chongjian Ge, Chenran Li, Chenfeng Xu, Masayoshi Tomizuka, Chen Tang, Mingyu Ding, and Wei Zhan. Womd-reasoning: A large-scale language dataset for interaction and driving intentions reasoning. *CoRR*, abs/2407.04281, 2024c.

Yue Li, Meng Tian, Dechang Zhu, Jiangtong Zhu, Zhenyu Lin, Zhiwei Xiong, and Xinhai Zhao. Drive-r1: Bridging reasoning and planning in vlms for autonomous driving with reinforcement learning. *CoRR*, abs/2506.18234, 2025a.

Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, Shouzheng Huang, Xinping Zhao, Borui Jiang, Lanqing Hong, Longyue Wang, Zhuotao Tian, Baoxing Huai, Wenhan Luo, Weihua Luo, Zheng Zhang, Baotian Hu, and Min Zhang. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *CoRR*, abs/2505.04921, 2025b.

Patrick Lin. *Why Ethics Matters for Autonomous Cars*, pp. 69–85. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.

Hanmeng Liu, Zhizhang Fu, Mengru Ding, Ruoxi Ning, Chaoli Zhang, Xiaozhang Liu, and Yue Zhang. Logical reasoning in large language models: A survey. *CoRR*, abs/2502.09100, 2025a.

Hao Liu, Carmelo Sferrazza, and Pieter Abbeel. Chain of hindsight aligns language models with feedback. In *ICLR*. OpenReview.net, 2024a.

Ruixun Liu, Lingyu Kong, Derun Li, and Hang Zhao. Occvla: Vision-language-action model with implicit 3d occupancy supervision. *CoRR*, abs/2509.05578, 2025b.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: marrying DINO with grounded pre-training for open-set object detection. In *ECCV (47)*, volume 15105 of *Lecture Notes in Computer Science*, pp. 38–55. Springer, 2024b.

Wei Liu, Jiyuan Zhang, Binxiong Zheng, Yufeng Hu, Yingzhan Lin, and Zengfeng Zeng. X-driver: Explainable autonomous driving with vision-language models. *CoRR*, abs/2505.05098, 2025c.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *CoRR*, abs/2503.06520, 2025d.

Zhiyuan Liu, Leheng Li, Yuning Wang, Haotian Lin, Zhizhe Liu, Lei He, and Jianqiang Wang. Controllable traffic simulation through llm-guided hierarchical chain-of-thought reasoning. *CoRR*, abs/2409.15135, 2024c.

Xuewen Luo, Fan Ding, Yinsheng Song, Xiaofeng Zhang, and Junn Yong Loo. Pkrd-cot: A unified chain-of-thought prompting for multi-modal large language models in autonomous driving. *CoRR*, abs/2412.02025, 2024.

Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In *ECCV (45)*, volume 15103 of *Lecture Notes in Computer Science*, pp. 403–420. Springer, 2024a.

Yunsheng Ma, Can Cui, Xu Cao, Wenqian Ye, Peiran Liu, Juanwu Lu, Amr Abdelraouf, Rohit Gupta, Kyungtae Han, Aniket Bera, et al. Lampilot: An open benchmark dataset for autonomous driving with language model programs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15141–15151, 2024b.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*, 2023.

Ali Mahmood and Róbert Szabolcsi. A systematic review on risk management and enhancing reliability in autonomous vehicles. *Machines*, 13(8):646, 2025.

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Argus: Vision-centric reasoning with grounded chain-of-thought. In *CVPR*, pp. 14268–14280. Computer Vision Foundation / IEEE, 2025.

Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with GPT. *CoRR*, abs/2310.01415, 2023.

Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision*, pp. 252–269. Springer, 2024.

Francisco Matos, Jorge Bernardino, João Durães, and João Carlos Cunha. A survey on sensor failures in autonomous vehicles: Challenges and solutions. *Sensors*, 24(16):5108, 2024.

Francisco Matos, João Durães, and João Cunha. Simulating the effects of sensor failures on autonomous vehicles for safety evaluation. In *Informatics*, volume 12, pp. 94. MDPI, 2025.

Khan Muhammad, Amin Ullah, Jaime Lloret, Javier Del Ser, and Victor Hugo C De Albuquerque. Deep learning for safe autonomous driving: Current challenges and future directions. *IEEE Transactions on Intelligent Transportation Systems*, 22(7):4316–4336, 2020.

Mohammad Naiseh, Jediah R. Clark, Tugra Akarsu, Yaniv Hanoch, Mario Brito, Mike Wald, Thomas Webster, and Paurav Shukla. Trust, risk perception, and intention to use autonomous vehicles: an interdisciplinary bibliometric review. *AI Soc.*, 40(2):1091–1111, 2025.

Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. In *European Conference on Computer Vision*, pp. 292–308. Springer, 2024.

Sheida Nozari, Ali Krayani, Pablo Marin, Lucio Marcenaro, David Martín Gómez, and Carlo S. Regazzoni. Exploring action-oriented models via active inference for autonomous vehicles. *EURASIP J. Adv. Signal Process.*, 2024(1):92, 2024.

Maxwell I. Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models. *CoRR*, abs/2112.00114, 2021.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Oscar Oviedo-Trespalacios, Verity Truelove, Barry Watson, and Jane A Hinton. The impact of road advertising signs on driver behaviour and implications for road safety: A critical systematic review. *Transportation research part A: policy and practice*, 122:85–98, 2019.

Tan-Hanh Pham and Chris Ngo. Multimodal chain of continuous thought for latent-space reasoning in vision-language models. *CoRR*, abs/2508.12587, 2025.

Alice Plebe, Henrik Svensson, Sara Mahmoud, and Mauro Da Lio. Human-inspired autonomous driving: A survey. *Cogn. Syst. Res.*, 83:101169, 2024.

Jordan Poots. RACE-SM: reinforcement learning based autonomous control for social on-ramp merging. *CoRR*, abs/2403.03359, 2024.

Gokul Puthumanaillam, Paulo Padrao, Jose Fuentes, Pranay Thangeda, William E. Schafer, Jae Hyuk Song, Karan Jagdale, Leonardo Bobadilla, and Melkior Ornik. TRACE: A self-improving framework for robot behavior forecasting with vision-language models. *CoRR*, abs/2503.00761, 2025.

Kangan Qian, Sicong Jiang, Yang Zhong, Ziang Luo, Zilin Huang, Tianze Zhu, Kun Jiang, Mengmeng Yang, Zheng Fu, Jinyu Miao, Yining Shi, He Zhe Lim, Li Liu, Tianbao Zhou, Hongyi Wang, Huang Yu, Yifei Hu, Guang Li, Guang Chen, Hao Ye, Lijun Sun, and Diange Yang. Agentthink: A unified framework for tool-augmented chain-of-thought reasoning in vision-language models for autonomous driving. *CoRR*, abs/2505.15298, 2025.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4542–4550, 2024.

Xiaoye Qu, Yafu Li, Zhaochen Su, Weigao Sun, Jianhao Yan, Dongrui Liu, Ganqu Cui, Daizong Liu, Shuxian Liang, Junxian He, Peng Li, Wei Wei, Jing Shao, Chaochao Lu, Yue Zhang, Xian-Sheng Hua, Bowen Zhou, and Yu Cheng. A survey of efficient reasoning for large reasoning models: Language, multimodality, and beyond. *CoRR*, abs/2503.21614, 2025.

Tianchi Ren, Haibo Hu, Jiacheng Zuo, Xinhong Chen, Jianping Wang, Chun Jason Xue, Jen-Ming Wu, and Nan Guan. Cot-vlm4tar: Chain-of-thought guided vision-language models for traffic anomaly resolution. *CoRR*, abs/2503.01632, 2025.

Saki Rezwana and Nicholas Lownes. Interactions and behaviors of pedestrians with autonomous vehicles: A synthesis. *Future Transportation*, 4(3):722–745, 2024.

Luke Rowe, Rodrigue de Schaetzen, Roger Girgis, Christopher Pal, and Liam Paull. Poutine: Vision-language-trajectory pre-training and reinforcement learning post-training enable robust end-to-end autonomous driving. *CoRR*, abs/2506.11234, 2025.

Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel J. Kochenderfer, Chiho Choi, and Behzad Dariush. Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning. In *WACV*, pp. 7498–7507. IEEE, 2024.

Alexandru Constantin Serban, Erik Poll, and Joost Visser. A standard driven software architecture for fully autonomous vehicles. In *ICSA Companion*, pp. 120–127. IEEE Computer Society, 2018.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: language agents with verbal reinforcement learning. In *NeurIPS*, 2023a.

Noah Shinn, Beck Labash, and Ashwin Gopinath. Reflexion: an autonomous agent with dynamic memory and self-reflection. *CoRR*, abs/2303.11366, 2023b.

Yuntao Shou, Tao Meng, Wei Ai, and Keqin Li. Multimodal large language models meet multimodal emotion recognition and reasoning: A survey. *CoRR*, abs/2509.24322, 2025.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pp. 256–274. Springer, 2024a.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pp. 256–274. Springer, 2024b.

SAE Standard. J3016_202104; taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. *SAE: Warrendale, PA, USA*, 2021.

Jiankai Sun, Hao Sun, Tian Han, and Bolei Zhou. Neuro-symbolic program search for autonomous driving decision module design. In *CoRL*, volume 155 of *Proceedings of Machine Learning Research*, pp. 21–30. PMLR, 2020.

Yuan Sun, Navid Salami Pargoo, Peter J. Jin, and Jorge Ortiz. Optimizing autonomous driving for safety: A human-centric approach with llm-enhanced RLHF. *CoRR*, abs/2406.04481, 2024.

Nazish Tahir and Ramviyas Parasuraman. Edge computing and its application in robotics: A survey. *CoRR*, abs/2507.00523, 2025.

Chunlin Tian, Xinpeng Qin, Kahou Tam, Li Li, Zijian Wang, Yuanzhe Zhao, Minglei Zhang, and Chengzhong Xu. CLONE: customizing llms for efficient latency-aware inference at the edge. In *USENIX ATC*, pp. 563–585. USENIX Association, 2025a.

Kexin Tian, Jingrui Mao, Yunlong Zhang, Jiwan Jiang, Yang Zhou, and Zhengzhong Tu. Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. *CoRR*, abs/2504.03164, 2025b.

Ran Tian, Boyi Li, Xinshuo Weng, Yuxiao Chen, Edward Schmerling, Yue Wang, Boris Ivanovic, and Marco Pavone. Tokenize the world into object-level knowledge to address long-tail events in autonomous driving. In *CoRL*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3656–3673. PMLR, 2024a.

Ran Tian, Boyi Li, Xinshuo Weng, Yuxiao Chen, Edward Schmerling, Yue Wang, Boris Ivanovic, and Marco Pavone. Tokenize the world into object-level knowledge to address long-tail events in autonomous driving. In *CoRL*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3656–3673. PMLR, 2024b.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. In *CoRL*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4698–4726. PMLR, 2024c.

George Tom, Minesh Mathew, Sergi Garcia-Bordils, Dimosthenis Karatzas, and C. V. Jawahar. Reading between the lanes: Text videoqa on the road. In *ICDAR (6)*, volume 14192 of *Lecture Notes in Computer Science*, pp. 137–154. Springer, 2023.

Daming Wang, Yuhao Song, Zijian He, Kangliang Chen, Xing Pan, Lu Deng, and Weihao Gu. HMVLM: multistage reasoning-enhanced vision-language model for long-tailed driving scenarios. *CoRR*, abs/2506.05883, 2025a.

Jingkang Wang, Ava Pun, James Tu, Sivabalan Manivasagam, Abbas Sadat, Sergio Casas, Mengye Ren, and Raquel Urtasun. Advsim: Generating safety-critical scenarios for self-driving vehicles. In *CVPR*, pp. 9909–9918. Computer Vision Foundation / IEEE, 2021.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and José M. Álvarez. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *CoRR*, abs/2405.01533, 2024a.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and José M. Álvarez. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *CVPR*, pp. 22442–22452. Computer Vision Foundation / IEEE, 2025b.

Tianqi Wang, Enze Xie, Ruihang Chu, Zhenguo Li, and Ping Luo. Drivecot: Integrating chain-of-thought reasoning with end-to-end driving. *CoRR*, abs/2403.16996, 2024b.

Wenshuo Wang, Letian Wang, Chengyuan Zhang, Changliu Liu, Lijun Sun, et al. Social interactions for autonomous driving: A review and perspectives. *Foundations and Trends® in Robotics*, 10(3-4):198–376, 2022.

Xu Wang, Mohammad Ali Maleki, Muhammad Waqar Azhar, and Pedro Trancoso. Moving forward: A review of autonomous driving software and hardware systems. *CoRR*, abs/2411.10291, 2024c.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*. OpenReview.net, 2023a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*. OpenReview.net, 2023b.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *CoRR*, abs/2503.12605, 2025c.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *CoRR*, abs/2401.06805, 2024d.

Yuqi Wang, Ke Cheng, Jiawei He, Qitai Wang, Hengchen Dai, Yuntao Chen, Fei Xia, and Zhao-Xiang Zhang. Drivingdojo dataset: Advancing interactive and knowledge-enriched driving world model. In *NeurIPS*, 2024e.

Taylor Webb. Large-scale ai language systems display an emergent ability to reason by analogy, 2023.

Cong Wei, Yujie Zhong, Haoxian Tan, Yingsen Zeng, Yong Liu, Zheng Zhao, and Yujiu Yang. Instructseg: Unifying instructed visual segmentation with multi-modal large language models. *CoRR*, abs/2412.14006, 2024a.

Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Jie Hu, Dengjie Li, Zheng Zhao, and Yujiu Yang. Hyperseg: Hybrid segmentation assistant with fine-grained visual perceiver. In *CVPR*, pp. 8931–8941. Computer Vision Foundation / IEEE, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

Xiao Wei, Haoran Chen, Hang Yu, Hao Fei, and Qian Liu. Guided knowledge generation with language models for commonsense reasoning. In *EMNLP (Findings)*, pp. 1103–1136. Association for Computational Linguistics, 2024b.

Yuxi Wei, Zi Wang, Yifan Lu, Chenxin Xu, Changxing Liu, Hao Zhao, Siheng Chen, and Yanfeng Wang. Editable scene simulation for autonomous driving via collaborative llm-agents. In *CVPR*, pp. 15077–15087. IEEE, 2024c.

Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. Dilu: A knowledge-driven approach to autonomous driving with large language models. In *ICLR*. OpenReview.net, 2024.

Gerald JS Wilde. Social interaction patterns in driver behavior: An introductory review. *Human factors*, 18(5):477–492, 1976.

Dongming Wu, Wencheng Han, Yingfei Liu, Tiancai Wang, Cheng-Zhong Xu, Xiangyu Zhang, and Jianbing Shen. Language prompt for autonomous driving. In *AAAI*, pp. 8359–8367. AAAI Press, 2025a.

Haoyuan Wu, Xueyi Chen, Rui Ming, Jilong Gao, Shoubo Hu, Zhuolun He, and Bei Yu. Totrl: Unlock LLM tree-of-thoughts reasoning potential through puzzles solving. *CoRR*, abs/2505.12717, 2025b.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. GSVA: generalized segmentation via multimodal large language models. In *CVPR*, pp. 3858–3869. IEEE, 2024.

Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. *CoRR*, abs/2501.04003, 2025.

Shuo Xing, Hongyuan Hua, Xiangbo Gao, Shenzhe Zhu, Renjie Li, Kexin Tian, Xiaopeng Li, Heng Huang, Tianbao Yang, Zhangyang Wang, Yang Zhou, Huaxiu Yao, and Zhengzhong Tu. Autotrust: Benchmarking trustworthiness in large vision language models for autonomous driving. *CoRR*, abs/2412.15206, 2024.

Chengkai Xu, Jiaqi Liu, Shiyu Fang, Yiming Cui, Dong Chen, Peng Hang, and Jian Sun. Tell-drive: Enhancing autonomous driving with teacher llm-guided deep reinforcement learning. *CoRR*, abs/2502.01387, 2025a.

Cong Xu and Ravi Sankar. A comprehensive review of autonomous driving algorithms: Tackling adverse weather conditions, unpredictable traffic violations, blind spot monitoring, and emergency maneuvers. *Algorithms*, 17(11):526, 2024.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, Chenyang Shao, Yuwei Yan, Qinglong Yang, Yiwen Song, Sijian Ren, Xinyuan Hu, Yu Li, Jie Feng, Chen Gao, and Yong Li. Towards large reasoning models: A survey of reinforced reasoning with large language models. *CoRR*, abs/2501.09686, 2025b.

Yi Xu, Yuxin Hu, Zaiwei Zhang, Gregory P. Meyer, Siva Karthik Mustikovela, Siddhartha Srinivasa, Eric M. Wolff, and Xin Huang. VLM-AD: end-to-end autonomous driving through vision-language model supervision. *CoRR*, abs/2412.14446, 2024a.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics Autom. Lett.*, 9(10):8186–8193, 2024b.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics Autom. Lett.*, 9(10):8186–8193, 2024c.

Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. In *ACL (Findings)*, pp. 11798–11827. Association for Computational Linguistics, 2025.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v(ision). *CoRR*, abs/2309.17421, 2023.

Ruoyu Yao, Yubin Wang, Haichao Liu, Rui Yang, Zengqi Peng, Lei Zhu, and Jun Ma. Calmm-drive: Confidence-aware autonomous driving with large multimodal model. *CoRR*, abs/2412.04209, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*, 2023a.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *ICLR*. OpenReview.net, 2023b.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Yiyao Yu, Yuxiang Zhang, Dongdong Zhang, Xiao Liang, Hengyuan Zhang, Xingxing Zhang, Mahmoud Khademi, Hany Hassan Awadalla, Junjie Wang, Yujiu Yang, and Furu Wei. Chain-of-reasoning: Towards unified mathematical reasoning in large language models via a multi-paradigm perspective. In *ACL (1)*, pp. 24914–24937. Association for Computational Linguistics, 2025.

Junzhe Zhang, Daniel Kumor, and Elias Bareinboim. Causal imitation learning with unobserved confounders. In *NeurIPS*, 2020.

Kunpeng Zhang, Shipu Wang, Ning Jia, Liang Zhao, Chunyang Han, and Li Li. Integrating visual large language model and reasoning chain for driver behavior analysis and risk assessment. *Accident Analysis & Prevention*, 198:107497, 2024a.

Qiming Zhang, Meixin Zhu, and Hao Frank Yang. Think-driver: From driving-scene understanding to decision-making with vision language models. In *European Conference on Computer Vision Workshop*, 2024b.

Tianyuan Zhang, Ting Jin, Lu Wang, Jiangfan Liu, Siyuan Liang, Mingchuan Zhang, Aishan Liu, and Xianglong Liu. Bench2advlm: A closed-loop benchmark for vision-language models in autonomous driving. *CoRR*, abs/2508.02028, 2025a.

Xuanliang Zhang, Dingzirui Wang, Longxu Dou, Qingfu Zhu, and Wanxiang Che. A survey of table reasoning with large language models. *Frontiers Comput. Sci.*, 19(9):199348, 2025b.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. LLM as a mastermind: A survey of strategic reasoning with large language models. *CoRR*, abs/2404.01230, 2024c.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):10795–10816, 2023a.

Yuxiao Zhang, Alexander Carballo, Hanting Yang, and Kazuya Takeda. Perception and sensing for autonomous vehicles under adverse weather conditions: A survey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 196:146177, February 2023b. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2022.12.021. URL `http://dx.doi.org/10.1016/j.isprsjprs.2022.12.021`.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Trans. Mach. Learn. Res.*, 2024, 2024d.

Symbat Zhanguzhinova, Emese Makó, Attila Borsos, Ágoston Pál Sándor, and Csaba Koren. Communication between autonomous vehicles and pedestrians: An experimental study using virtual reality. *Sensors*, 23 (3):1049, 2023.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Tsung-Yi Lin, Gordon Wetzstein, Ming-Yu Liu, and Donglai Xiang. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *CVPR*, pp. 1702–1713. Computer Vision Foundation / IEEE, 2025.

Hao Zhen, Yucheng Shi, Yongcan Huang, Jidong J. Yang, and Ninghao Liu. Leveraging large language models with chain-of-thought and prompt engineering for traffic crash severity analysis and inference. *Comput.*, 13(9):232, 2024.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. Least-to-most prompting enables complex reasoning in large language models. In *ICLR*. OpenReview.net, 2023.

Jian Zhou, Yulong Gao, Björn Olofsson, and Erik Frisk. Uncertainty-aware decision-making and planning for autonomous forced merging. In *CDC*, pp. 1095–1102. IEEE, 2024.

Xinyu Zhu, Junjie Wang, Lin Zhang, Yuxiang Zhang, Yongfeng Huang, Ruyi Gan, Jiaxing Zhang, and Yujiu Yang. Solving math word problems via cooperative reasoning induced language models. In *ACL (1)*, pp. 4471–4485. Association for Computational Linguistics, 2023.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong T. Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: vision-language-action models transfer web knowledge to robotic control. In *CoRL*, volume 229 of *Proceedings of Machine Learning Research*, pp. 2165–2183. PMLR, 2023.