Beyond Textual Chain-of-Thought: A Survey on Action-Grounded Reasoning in Autonomous Driving

Zhengxu Tang*, Xiaozhou Zhang*, Guofeng Cui, Ziyu Gong, Zi Wang, Yunfei Shi, Ruifeng Deng, Chengzhi Qi, Ke Chen, Sachin Patil, Tianjun Xiao, Langechuan Liu, Pichao Wang†
NVIDIA

Abstract

Chain-of-thought (CoT) reasoning powers generative models by eliciting intermediate steps before producing an answer. In autonomous driving, the answer is a continuous action. Thus its reasoning must share the same spatiotemporal structure as the physical world. This survey studies the resulting shift from textual CoT to action-grounded reasoning. Surveying 171 papers, including 130 method papers and 41 benchmarks, datasets, surveys, and analysis papers, we propose a representation-centered taxonomy that treats the form of the intermediate state as the organizing axis. We systematize the 130 methods into four categories: language-based, visual-spatial, latent-dynamic, and externalized reasoning, further divided into 13 subtypes tied to distinct regions of interests. Our synthesis shows that the open frontier of reasoning in driving agents lies in intermediate representations that can be grounded in the real world, coupled to real-time action, and verified under safety-critical systems. Project page: https://github.com/tangzhengxu/awesome-av-cot.

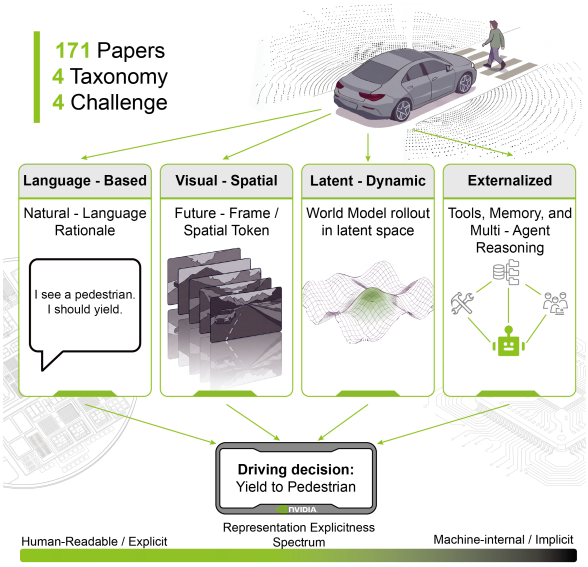


Figure 1: Overview of action-grounded reasoning representations in autonomous driving. Different representations can connect the same scene to a driving decision. They differ in interpretability, grounding, action coupling, latency, and closed-loop reliability.

1 Introduction

Autonomous driving has cycled through several architectural generations, all driven by the same pressure: better generalization. As autonomous driving systems increasingly incorporate large language models (LLMs), vision-language models (VLMs), and vision-language-action (VLA) models for scene understanding, decision explanation, and action generation (Hwang et al., 2024; Tian et al., 2025b), they aim to escape the local minimum of pure end-to-end driving through explicit thinking traces, made possible by stronger backbones pretrained on large-scale multimodal data. But what should a driving model think before it acts? We need a reasoning layer flexible enough to bridge scene interpretation and trajectory generation, yet explicit enough to be supervised and audited.

Chain-of-thought reasoning offers a natural starting point for such a layer. In language models, CoT encourages step-by-step reasoning before an answer (Wei et al., 2022; Wang et al., 2023b; Yao et al., 2023), and early driving methods similarly use textual chains to describe scenes, identify critical objects, infer interactions, and derive plans (Mao et al., 2023; Wang et al., 2024a). However, driving reshapes the role of CoT in three ways. First, the output is a trajectory, not a textual answer, and the two domains have different similarity structures. Second, reasoning must be grounded in geometry, motion, and traffic rules. Third, reasoning must run under real-time and closed-loop

* Equal Contribution.
† Corresponding Author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 11519–11561
October 24-29, 2026 ©2026 Association for Computational Linguistics

constraints, where verbose or unstable traces compound rather than reduce error.

Driving CoT is moving beyond textual chains in response to these challenges. Recent methods express intermediate reasoning as stage-wise language plans, reflective revisions, compressed textual traces, future visual states, occupancy structures, spatial crops, latent world-model rollouts, dynamics tokens, retrieved memories, traffic-rule knowledge, tool outputs, or messages exchanged among agents. Although these methods differ architecturally, they answer the same question: in what form should a driving model represent intermediate reasoning so that it can be grounded, executed, and evaluated? We call such forms action-grounded reasoning representations: identifiable intermediate structures that organize driving-relevant information before a final decision is produced.

This representation-centered view differs from existing surveys; Table 1 gives a side-by-side comparison. Our claim is not that earlier surveys omit visual, latent, externalized, or safety-related work. They provide complementary model-, task-, pipeline-, and capability-centered views (Cui et al., 2026, 2025; Jiang et al., 2025c; Hu et al., 2025; Yu et al., 2026). Our survey instead asks what identifiable artifact appears between observation and a driving-relevant output, how it connects to the downstream task, and what evidence is needed to evaluate it. This common unit allows textual chains, imagined futures, latent rollouts, retrieval traces, tool calls, and multi-agent messages to be compared within a unified framework.

We include a method only when its intermediate structure plays an explicit reasoning role between observation and action: it generates or supervises an intermediate reasoning trace; it uses an intermediate state as a decision-relevant reasoning representation rather than a generic feature; or its intermediate representation can be inspected, intervened on, retrieved, compressed, distilled, refined, or communicated as part of decision-making. This boundary admits broader forms of driving CoT while excluding ordinary perception, prediction, or planning modules that do not model reasoning as an identifiable intermediate process.

Our contributions are threefold. First, we propose a representation-centered framework for driving CoT, shifting the focus from textual to action-grounded reasoning representations. Second, we organize 130 method papers into four representation families and 13 subtypes, identifying how language-based, visual-spatial, latent-dynamic, and externalized reasoning occupy different points in the trade-off space. Third, we synthesize benchmark evidence and open challenges, showing that the central obstacles for driving CoT are faithful action coupling, perceptual grounding, adaptive reasoning cost, closed-loop reliability, and safety verification.

2 From Textual CoT to Action-Grounded Representations

The notion of CoT is often associated with visible natural-language reasoning (Wei et al., 2022; Wang et al., 2023b; Yao et al., 2023). This association is useful for language tasks, where an intermediate text trace can expose how a model decomposes a question before generating an answer. In autonomous driving, however, the same definition is too narrow. A driving system may reason through language, but it may also reason through spatial evidence, future states, latent dynamics, retrieved memories, or tool outputs (Zhang et al., 2023; Ahn et al., 2022; Huang et al., 2023; Zitkovich et al., 2023). What matters is not whether the reasoning is written in natural language, but whether it forms a decision-relevant intermediate structure bridging observation and action.

Driving changes the output of reasoning. In standard language tasks, a reasoning chain connects a question to a textual answer (Wei et al., 2022). In driving, the output may be a trajectory, control command, maneuver decision, or planner-ready intervention (Mao et al., 2023; Tian et al., 2025b; Wen et al., 2024). This shifts not just the output format but the underlying domain: text and action carry different similarity structures. A trash bin and a stalled car are linguistically distant, yet operationally identical to avoid. A textual explanation that does not influence the planner may be interpretable but not action-grounded; conversely, a latent rollout that is not human-readable may still be valuable if it improves collision avoidance, comfort, and route completion (Yang et al., 2026; Zheng et al., 2025b). Driving CoT should be evaluated by its connection to action, not by textual coherence alone.

Driving changes the grounding of reasoning. Language CoT mainly operates over symbolic statements and commonsense knowledge (Wei et al., 2022). Driving reasoning must additionally be grounded in metric geometry, lane topology, map

Survey	Scope	Primary organizing axis	Treatment of intermediate representations
LLM4AD (Cui et al., 2026)	LLMs for autonomous driving	LLM roles, tasks, benchmarks, and pipeline integration	Reasoning is one capability among several
Chain-of-Thought for Autonomous Driving (Cui et al., 2025)	Driving CoT methods	CoT paradigms, methods, and tasks	Closest in scope, but mainly organized around CoT methodology
Jiang et al. (2025c)	VLA models for autonomous driving	Architectures and system components	Representations are discussed through VLA architecture
Hu et al. (2025)	Past and future VLA systems	System paradigms and action generation	Focuses on model evolution and end-to-end integration
Yu et al. (2026)	Reasoning in autonomous driving	Cognitive capabilities and reasoning challenges	Organizes reasoning by capability and system challenge
Ours	Decision-relevant intermediate reasoning	Language, visual-spatial, latent-dynamic, and externalized artifacts	Uses intermediate-artifact form as the paper-level axis across 13 subtypes

Table 1: Side-by-side comparison with prior surveys.

priors, object locations, occlusion, motion, interaction intent, and traffic rules (Tian et al., 2025a; Godbole et al., 2025). A plausible chain can be unsafe if it mislocates a pedestrian, ignores a blind spot, or misunderstands right-of-way; this grounding pressure explains why many methods move beyond text. Some methods attach language reasoning to visual regions, object crops, masks, or bounding boxes (Corbiere et al., 2025; Li et al., 2025b). Others imagine future visual states or occupancy structures (Wang et al., 2024b; Zeng et al., 2026; Yang et al., 2025). Still others compress scene evolution into latent world models or dynamics tokens (Zheng et al., 2025b; Shang et al., 2026; Lu et al., 2026). These designs are not alternative explanation formats but attempts to move reasoning closer to the variables that determine safe driving.

Driving changes the deployment constraints of reasoning. Language reasoning can be long and reflective in offline tasks, but driving operates under real-time and closed-loop constraints (Zhang et al., 2025d; Jia et al., 2024): a slow, detailed trace may help debugging yet harm execution, and the vehicle's action changes the next observation, so reasoning errors compound through interaction. This motivates compressed, distilled, latent, and adaptive forms of CoT. Some systems use explicit textual reasoning as supervision during training and internalize it into faster representations at inference time (Feng et al., 2025; Liu et al., 2025e). Others trigger expensive reasoning only in uncertain or high-risk scenarios (Qian et al., 2024; Luo et al., 2025c). The central question is not whether a model can reason more, but whether it reasons at the right level of detail under the right conditions.

3 Representation-Centered Taxonomy

Figure 6 (Appendix B) gives a corpus-level view: growth accelerates after 2025, and while language-based methods dominate, externalized, latent-dynamic, and visual-spatial representations now form substantial branches. This pattern motivates our four-way organization into language-based (Mao et al., 2023; Tian et al., 2025b; Yuan et al., 2025; Feng et al., 2025), visual-spatial (Wang et al., 2024b; Yang et al., 2025; Corbiere et al., 2025; Li et al., 2025b), latent-dynamic (Zheng et al., 2025b; Shang et al., 2026; Yang et al., 2026), and externalized (Wen et al., 2024; Luo et al., 2025a; Qian et al., 2025a; Chiu et al., 2025) representations. These categories are not a strict temporal progression; they reflect different design priorities: interpretability, grounding, action coupling, efficiency, knowledge access, and collaboration.

We assign the primary label according to the form in which the decision-relevant intermediate artifact is consumed at the reasoning-to-output interface, rather than its surface serialization or provenance. Tool provenance is noted separately where relevant, but does not by itself determine the primary family. For example, a textual coordinate description that subsequent reasoning consumes as tokens is language-based despite its spatial content; the same values parsed into a bounding box, object-state tuple, occupancy element, or waypoint consumed by prediction or planning are visual-spatial, even if serialized as text; and a tool-returned bounding box consumed as structured spatial state likewise remains visual-spatial, with its tool provenance noted separately. This section discusses each subtype at the level of common design patterns and representative trade-offs; Table 2 gives one representative method per subtype, and full paper-level comparisons, including backbone, benchmark, and assigned subtype, are provided in Appendix B.

Subtype	Representative	Intermediate artifact	Backbone	Benchmark/task	Source-reported within-study evidence
Language-based					
Descriptive	ORION (Fu et al., 2025a)	Scene description and causal rationale aligned with trajectory head	Vicuna	Bench2Drive closed-loop	Driving Score 77.74; Success Rate 54.62%
Procedural	DriveVLM (Tian et al., 2025b)	Stage-wise perception–prediction–planning language plan	Qwen	nuScenes open-loop planning	Average L2 0.40 m; 0.31 m for the dual-system variant in the same study
Reflective	AutoDrive-R^2 (Yuan et al., 2025)	Critique followed by a revised reasoning trace	Qwen	nuScenes open-loop planning	Average L2 0.19 m; matched base model 1.45 m
Compressed	AdaThinkDrive (Luo et al., 2025c)	Adaptive think/no-think textual trace	InternVL	NAVSIM v1 navtest	PDMS 90.3 with adaptive CoT switching
Visual-spatial					
Predictive	FSDrive (Zeng et al., 2026)	Predicted future-scene representation	Qwen	nuScenes open-loop planning	Average L2 0.28 m; collision rate 0.10%
Action-grounded	RIV-CoT (Corbiere et al., 2025)	Retrieved and cropped visual evidence interleaved with the textual chain	Qwen+LLaVA	DrivingVQA	F1 68.8 with interleaved visual CoT
Latent-dynamic					
Rollout-based	World4Drive (Zheng et al., 2025b)	Intention-conditioned latent world-model rollouts	E2E CNN/Trans.	NAVSIM v1 navtest	PDMS 85.1 without perception annotation
Tokenized	OneVL (Lu et al., 2026)	Four visual and two language latent positions	Qwen	NAVSIM planning	PDM score 88.84 at 4.46 s; explicit-CoT baseline 88.29 at 6.58 s
Optimization-based	ReCogDrive (Li et al., 2025e)	Diffusion-refined trajectory states over cognitive features	Qwen+InternVL	NAVSIM v1 navtest	PDMS 90.8
Externalized					
Retrieval-augmented	DiLu (Wen et al., 2024)	Retrieved past driving experiences injected into the prompt	GPT	HighwayEnv closed-loop	Success rate improves as the experience memory accumulates
Knowledge-grounded	LeAD (Zhang et al., 2025e)	Traffic-law and map priors guiding hierarchical planning	GPT	CARLA Leaderboard V1	Driving Score 71.96 (closed-loop)
Tool-mediated	AgentThink (Qian et al., 2025a)	Typed tool calls and returned observations	Qwen	DriveLMM-o1 reasoning	Reasoning score 79.68; untuned backbone 51.77
Cooperative	CoLMDriver (Liu et al., 2025a)	Inter-vehicle natural-language negotiation messages	InternVL	CARLA InterDrive closed-loop	Driving Score 88.53 across interactive scenarios

Table 2: One representative method per subtype. Representatives are selected for the clarity of their intermediate artifact and the availability of a documented within-study comparison, not for the highest reported score. Each result retains the source paper's model configuration, split, horizon, inputs, and metric definition; rows are not ranked or compared across incompatible protocols, and baselines are included only when reported within the same study or under a matched setting. Appendix B remains the complete corpus inventory.

3.1 Language-Based Representations

Language-based representations are the most direct extension of standard CoT to driving (Wei et al., 2022; Mao et al., 2023; Xu et al., 2024), expressing intermediate reasoning as scene descriptions, object analyses, risk explanations, decision justifications, driving plans, self-reflections, or compressed textual traces (Tian et al., 2025b; Wang et al., 2024a; Yuan et al., 2025; Jiang et al., 2025b). Their main advantage is inspectability: language traces can be read, aligned with traffic rules, supervised by annotations, and used for passenger-facing explanations (Hwang et al., 2024; Wen et al., 2024; Cui et al., 2024). Their main weakness is action coupling: language describes geometry and motion indirectly (Tian et al., 2025a; Song et al., 2025).

Descriptive language reasoning. Descriptive language reasoning verbalizes scene semantics, critical objects, risk factors, and driving decisions before producing an action or explanation (Mao et al., 2023; Xu et al., 2024; Sha et al., 2025; Ma et al., 2024; Hwang et al., 2024). This subtype makes driving models more transparent by exposing what the model appears to notice and why it chooses a maneuver (Zheng et al., 2024; Yao et al., 2024; Zhang et al., 2024; Xing et al., 2025; Fu et al., 2025a; Chahe and Zhou, 2025; Liu et al., 2025d; Lu et al., 2025). It is especially useful for driving question answering, explanation, and instruction-following (Chi et al., 2026; Zarghani et al., 2025; Liu et al., 2025c; Peng et al., 2025a; Yu et al., 2025; Wu and Luo, 2025; Zhang et al., 2026a; Hu et al., 2025; Guo et al., 2026; Gao et al., 2026; Xie et al., 2025). However, description alone does not guarantee action coupling: a model may describe a scene correctly yet output a poor trajectory. The core risk is post-hoc rationalization: a text trace may look like reasoning even when the action was produced by a separate planner or action head.

Procedural language reasoning. Procedural language reasoning imposes an explicit reasoning structure, often following perception, prediction, planning, and control stages (Wang et al., 2023a; Tian et al., 2025b; Wang et al., 2024a; Luo et al.,

2024; Mandalika et al., 2025). Compared with descriptive reasoning, procedural reasoning is more action-aware because it organizes information in a decision-relevant order and can be connected to planners or behavior states (Sharan et al., 2023; Peng et al., 2025b; Wang et al., 2025b; Zhao et al., 2025; Liao et al., 2025b; Diao et al., 2025). Its limitation is rigidity: fixed templates may over-reason in simple scenarios and under-specify complex negotiation, while perception or grounding errors can still propagate through a well-structured but incorrect chain. A promising direction is adaptive procedure selection across lane following, merging, occluded hazards, and emergencies (Wang et al., 2025a; Liu et al., 2025f; Wang et al., 2025f; Liu et al., 2025g; Han et al., 2026; Tao et al., 2026; Ghosh et al., 2026; Gu et al., 2026). Hybrid procedural reasoning can combine symbolic calculation with commonsense guidance. Azarafza et al. (2024) provide an LLM with detected objects and vehicle-state signals, then use an explicit sequence of arithmetic and commonsense reasoning steps to produce brake and speed commands in CARLA. We assign procedural-language reasoning as the primary subtype because this staged procedure is the intermediate artifact that structures control generation, and record knowledge-grounded reasoning as a secondary tag because the procedure also invokes commonsense driving knowledge.

Reflective language reasoning. Reflective language reasoning adds critique, revision, rollback, or failure learning to the textual reasoning process (Yuan et al., 2025; Li et al., 2025c; Zhang et al., 2025c). This turns CoT from a one-pass explanation into an error-correction mechanism, valuable in interactive or high-risk scenarios where an initial plan may be inconsistent, unsafe, or incomplete (Peng et al., 2025d; Luo et al., 2026a). The central trade-off is latency: reflection requires additional generation, verification, or optimization steps, making it difficult to deploy at every time step. A practical reflective system must decide when deeper reasoning is necessary and when fast execution is safer; reflection is thus best viewed as a risk-triggered mechanism for difficult scenes rather than a universal inference mode.

Compressed language reasoning. Compressed language reasoning shortens, distills, internalizes, or selectively activates textual reasoning to reduce inference cost (Huang et al., 2024; Jiang et al., 2025b; Qiao et al., 2025; Liu et al., 2025e; Feng et al., 2025; Wasif et al., 2026). These methods reveal that textual CoT may be most useful not as a verbose inference-time output, but as a training-time scaffold (Zhou et al., 2026b; Li et al., 2026b; Zheng et al., 2025a; Luo et al., 2025c; Wang et al., 2025d). A model can learn from explicit reasoning traces and later execute behavior through shorter language, latent states, or direct action heads (Zhang et al., 2025a; Li et al., 2025a; Fu et al., 2025b; Zhang et al., 2025f; Zhao et al., 2026; Chen et al., 2026; Ye et al., 2026). This blurs the boundary with latent-dynamic reasoning and reframes the goal: not longer chains, but better transfer from linguistic reasoning to action-effective representations that preserves the benefits of explicit supervision after reasoning is internalized.

3.2 Visual-Spatial Representations

Visual-spatial representations move intermediate reasoning closer to the perceptual structure of driving scenes, treating future frames, occupancy maps, object crops, masks, bounding boxes, or trajectory sketches as reasoning carriers. This addresses a key weakness of text: language describes spatial relations, but does not preserve metric geometry, occlusion, lane topology, or temporal evolution.

Predictive visual-spatial reasoning. Predictive visual-spatial reasoning represents intermediate reasoning as future visual or occupancy states (Wang et al., 2024b; Gao et al., 2024; Yang et al., 2025; Chen et al., 2025b). These methods allow the model to see before acting by simulating how the scene may evolve under possible ego actions (Zeng et al., 2026; Xiong et al., 2026; Zhou et al., 2026a; Wang et al., 2026a), which is attractive for interaction-heavy scenarios such as merging, unprotected turns, and pedestrian crossings. The limitation is that visual plausibility does not imply planning usefulness: a realistic future may fail to preserve the geometry, uncertainty, or agent behavior needed for safe control, so the key question is whether predicted futures causally improve driving decisions rather than video or occupancy metrics.

Action-grounded visual-spatial reasoning. Action-grounded visual-spatial reasoning localizes the evidence most relevant to the current action, such as critical-object crops, boxes, masks, depth-aware spatial features, or trajectory sketches (Corbiere et al., 2025; Chen et al., 2025a; Li et al., 2025b). Compared with full future generation, this subtype is more compact and more

directly tied to the final maneuver, focusing the model on the pedestrian to avoid, the vehicle to yield to, or the trajectory region to refine (Zhang et al., 2026b,c). Its risk is evidence selection: if the crop, box, mask, or sketch misses the true risk factor, later reasoning may become confidently wrong. This shifts the burden from complete simulation to faithful selection of decision-relevant evidence; for safety-critical deployment, the system must also know when its selected evidence is incomplete.

3.3 Latent-Dynamic Representations

Latent-dynamic representations compress reasoning into latent states, dynamics tokens, rollouts, or iterative optimization processes. They are less interpretable than text or images, but often more compatible with planning and control: reasoning need not be human-readable to be decision-relevant.

Rollout-based latent-dynamic reasoning. Rollout-based latent-dynamic reasoning uses latent world models to simulate future scene evolution before action selection (ByteDance Seed, 2025; Zheng et al., 2025b; Lin et al., 2025; Tan et al., 2025). This preserves thinking through possible futures without pixel-level generation cost: latent rollouts represent temporal dynamics more compactly than videos and more action-relevantly than language (Liao et al., 2025c; Li et al., 2026a; Liu et al., 2026; Luo et al., 2026b). The main challenge is verification: a human can inspect a textual chain or predicted frame, but not a latent rollout. Future systems need probing, decoding, consistency checks, or intervention tests to determine whether latent reasoning is faithful, grounded, and safe.

Tokenized latent-dynamic reasoning. Tokenized latent-dynamic reasoning uses compact discrete codes or continuous latent positions explicitly trained to carry information about future dynamics, actions, or intermediate reasoning (Ganai et al., 2026; Wang et al., 2026b; Shang et al., 2026; Lu et al., 2026). DynVLA quantizes future ego and environment dynamics into codes supervised through future-scene reconstruction and action prediction (Shang et al., 2026), while OneVL uses a small fixed set of visual and language latent positions trained through reconstruction objectives and consumed before action generation (Lu et al., 2026). Unlike generic hidden features, these carriers are identifiable in the architecture and are explicitly supervised, decoded, or passed to a downstream module. We also clarify that HiST-VLA is a hybrid boundary case: it produces an explicit granular command, a coarse trajectory, and one aggregate trajectory-confidence score, rather than attaching a confidence value to each primitive or representing the explicit command labels as learned dynamics tokens (Wang et al., 2026b). This subtype inherits the sequential structure of textual CoT while removing the requirement that each token be a word, but its open problem is semantic alignment: without probing or intervention, it is unclear whether latent carriers encode risk, intention, affordance, motion, or statistical shortcuts.

Optimization-based latent-dynamic reasoning. Optimization-based latent-dynamic reasoning treats reasoning as iterative refinement of latent action states, often through diffusion, reinforcement fine-tuning, parallel decoding, or progressive trajectory optimization (Jiang et al., 2025a; Gao et al., 2025b; Peng et al., 2025c). In this view, the reasoning trace is not a sentence or image, but a trajectory of internal optimization states. This is attractive because planning is naturally an optimization problem: a candidate trajectory can be refined to better satisfy safety, comfort, and goal constraints (Li et al., 2025e; Ma et al., 2025; Yang et al., 2026). Its limitation is observability: if refinement improves scores but cannot be inspected or causally linked to risk reduction, it is hard to certify as reasoning rather than opaque optimization—an important boundary for future work.

3.4 Externalized Representations

Externalized representations move part of reasoning outside the model's internal activations, using retrieval, memory, traffic rules, tools, or communication among agents instead of parametric knowledge alone. Their strength is long-tail coverage and modularity; their weakness is reliability, because external sources may be stale, irrelevant, delayed, inconsistent, or unsafe.

Retrieval-augmented reasoning. Retrieval-augmented reasoning uses retrieved regulations, demonstrations, cases, experiences, or vehicle-to-everything information as intermediate evidence for driving decisions (Cai et al., 2026; Luo et al., 2025a; Han et al., 2025; Wang et al., 2025e). Retrieval helps with long-tail scenarios where parametric memory is insufficient, such as rare traffic rules, unusual road layouts, or previously seen risky interactions (Chang et al., 2026; Luo et al., 2025b; Gan et al., 2025; Patrikar et al.,

2025), and makes part of reasoning auditable because retrieved evidence can be inspected. The new failure mode is retrieving irrelevant, outdated, or misleading evidence and reasoning confidently from it, so retrieval quality, source reliability, and fallback behavior are central to this subtype.

Knowledge-grounded reasoning. Knowledge-grounded reasoning uses explicit rules, maps, scene graphs, risk constraints, or other structured priors that cannot be inferred reliably from the current observation alone (Cui et al., 2024; Jiang et al., 2024; Xu et al., 2025a; Fang et al., 2025; Xu et al., 2025b). SafeDrive turns risk knowledge into decision constraints (Zhou et al., 2026c); PlanAgent and LeAD expose route or map priors to planning (Zheng et al., 2026; Zhang et al., 2025e); and KLDrive constructs a scene knowledge graph for constrained reasoning (Tian et al., 2026). We distinguish this subtype from retrieval by the role of the intermediate artifact: retrieval is primary when memory access and example selection constitute the reasoning process, whereas knowledge-grounded reasoning is primary when a rule, graph, or structured prior itself guides the decision. Memory-centric methods such as DiLu and Agent-Driver are therefore retrieval-primary, with a secondary knowledge tag where appropriate (Wen et al., 2024; Mao et al., 2024). Mechanisms reported by individual papers, such as confidence-weighted guidance or lower-level overrides, are described as system-specific safeguards rather than evidence of a general solution to knowledge–scene conflict (Qian et al., 2024, 2025b; Liu et al., 2025b; Luo et al., 2025d; Wang et al., 2026c; Zhang et al., 2025b).

Tool-mediated reasoning. Tool-mediated reasoning delegates part of the reasoning process to callable modules such as planners, simulators, map queries, rule checkers, or structured APIs (Qian et al., 2025a; Goba et al., 2025), making driving CoT more executable: instead of only stating a plan, the model can call a tool that computes, checks, or simulates part of the decision. Tool use also offers modularity, since specialized components handle geometry, rules, or optimization more reliably than a general-purpose model. The cost is interface reliability: a call may fail, return delayed output, or be invoked in the wrong context, so tool-mediated CoT requires validation, uncertainty handling, and safe fallback policies.

Cooperative reasoning. Cooperative reasoning represents intermediate reasoning as messages exchanged among vehicles, agents, or infrastructure components (Hu et al., 2024; Chiu et al., 2025). Many driving risks are distributed: no single vehicle observes the whole scene, and cooperative negotiation can improve merging, intersection handling, or occluded-risk awareness. Communication exposes intent and shares local observations, making reasoning more socially and spatially informed (Liu et al., 2025a; Gao et al., 2025a; Hou et al., 2025). However, cooperation introduces synchronization, authentication, bandwidth, and conflict-resolution challenges: the system must decide which messages to trust and how to remain safe when communication is missing or adversarial.

4 Cross-Category Analysis

The four categories are not competing replacements; they answer the same pressure: intermediate reasoning must be both meaningful to humans and useful for action. Language-based reasoning is easiest to inspect but weakest in metric grounding (Mao et al., 2023; Tian et al., 2025b; Hwang et al., 2024); visual-spatial reasoning is more grounded but can be costly or incomplete (Corbiere et al., 2025; Li et al., 2025b; Zeng et al., 2026); latent-dynamic reasoning is closer to control but harder to verify (Zheng et al., 2025b; Yang et al., 2026; Luo et al., 2026b); and externalized reasoning expands access to knowledge and cooperation but depends on external sources and interfaces (Wen et al., 2024; Luo et al., 2025a; Qian et al., 2025a).

Interpretability versus action coupling. A central tension is that the most interpretable representations are not always the most action-effective. Textual chains expose the model's apparent logic but may remain loosely connected to the final trajectory (Song et al., 2025; Tian et al., 2025b; Mao et al., 2023); latent rollouts and optimization states can directly influence planning but are difficult to inspect (Yang et al., 2026; Li et al., 2025e; Gao et al., 2025b). Visual and externalized representations occupy a middle ground: crops, masks, retrieved cases, and tool outputs are easier to inspect than latent states, but still require mechanisms to verify that the final action actually depends on them (Corbiere et al., 2025; Zhang et al., 2026b; Chang et al., 2026; Patrikar et al., 2025; Qian et al., 2025a). Future systems may therefore need dual representations: a compact internal representation

for control and a faithful external trace for monitoring and explanation.

Grounding versus coverage. Another tension is between grounding and coverage. Visual-spatial reasoning grounds decisions in the current scene but may miss long-tail knowledge such as rare rules or unusual interactions (Corbiere et al., 2025; Li et al., 2025b; Tian et al., 2025a; Godbole et al., 2025); retrieval and knowledge-grounded reasoning provide this coverage but may introduce stale or irrelevant information (Cai et al., 2026; Luo et al., 2025a; Chang et al., 2026; Wen et al., 2024); latent world models encode temporal dynamics but may fail silently under distribution shift (Wang et al., 2024b; Gao et al., 2024). A robust driving CoT system should integrate perception-grounded evidence, learned dynamics, and external knowledge.

Reasoning depth versus latency. Reasoning is not free: reflection, retrieval, tool use, world-model rollout, and multi-agent communication all increase computation or communication cost (Yuan et al., 2025; Li et al., 2025c; Luo et al., 2025a; Qian et al., 2025a; Gao et al., 2025a). A system that reasons too slowly may fail in urgent situations, while one that reasons too little may miss rare hazards. This points to adaptive reasoning as a unifying direction (Luo et al., 2025c; Qian et al., 2024; Xie et al., 2025; Ghosh et al., 2026): estimate uncertainty, risk, and conflict, then decide whether to use fast reactive control, short language reasoning, retrieved knowledge, latent simulation, tool-mediated verification, or cooperative communication.

From representation choice to system design. Representation choice is a system-design decision rather than a purely modeling one; four questions make it concrete. **Trigger**: When is additional reasoning activated—for example, under uncertainty, ambiguous right of way, or conflict between scene evidence and a retrieved rule (Luo et al., 2025c; Qian et al., 2024)? **Interface**: Where does the intermediate artifact affect the task—for example, maneuver selection, trajectory scoring, constraint construction, or candidate rollout—rather than appearing only as a post-hoc explanation? **Validation**: What is checked before the artifact is used, such as entity grounding for language traces, geometric and temporal consistency for spatial states, rollout consistency for latent states, or source and scene compatibility for retrieved information? **Fallback**: What alternative system behavior is used when the artifact is missing, inconsistent, uncertain, or late, and under what assumptions is that alternative evaluated? Inconsistencies among perception, the intermediate artifact, and the planned action should be checked separately; the planned action is not treated as ground truth. This checklist is design and evaluation guidance, not evidence that the surveyed systems already satisfy deployment-level safety requirements.

5 Evaluation Landscape

Existing benchmarks evaluate different fragments of driving reasoning. QA and explicit-reasoning datasets test whether models can describe scenes, answer questions, and produce structured reasoning traces (Sima et al., 2024; Wang et al., 2025c; Wei et al., 2025a; Ishaq et al., 2025; Zhang et al., 2025d; Zeng et al., 2025). Spatial grounding and robustness benchmarks test whether models localize objects, understand relations, and remain stable under perturbations (Tian et al., 2025a; Cannons et al., 2025; Li et al., 2025f). Planning benchmarks evaluate whether reasoning improves motion decisions, collision avoidance, route, and progress (Dauner et al., 2024; Jia et al., 2024), and closed-loop benchmarks expose whether reasoning remains reliable when actions affect future observations (Tanahashi et al., 2023; Wei et al., 2025b).

Despite this progress, evaluation remains fragmented: QA benchmarks rarely test action consequences, planning benchmarks rarely inspect intermediate reasoning, and closed-loop benchmarks rarely evaluate reasoning faithfulness or evidence reliability (details in Appendix C). Evaluation should jointly measure grounding, action faithfulness, reasoning cost, and closed-loop safety.

QA and planning metrics measure different parts of a reasoning system. Exact-match or multiple-choice accuracy measures final-answer correctness, but does not show whether the answer is visually grounded or used by the planner. Trajectory displacement metrics compare a prediction with a logged future and may penalize other safe behaviors. Open-loop collision estimates depend on the footprint, horizon, object extrapolation, and collision protocol, and do not model how other agents respond to the ego vehicle. Closed-loop and composite scores cover more aspects of driving, but remain specific to the simulator, benchmark version, scenario set, and aggregation rule. We therefore report these metrics as source-specific evidence

rather than treating them as directly comparable measures of reasoning quality. A compact metric glossary is provided in Appendix C.2.

6 Open Challenges

Faithful action coupling. The most important question for driving CoT is not whether the reasoning trace is plausible, but whether it actually affects action: a model may generate a convincing explanation the planner ignores, or act safely for reasons that differ from the stated chain. Future work should evaluate causal coupling through intervention: changing the reasoning representation should predictably change the action, and changing irrelevant text should not. This is especially important for language-based explanations, where post-hoc rationalization can be mistaken for faithful reasoning. Two safeguards make such intervention tests meaningful. First, the intervention must preserve artifact validity rather than create an implausible or out-of-distribution state. Second, the expected behavioral change is assessed at the artifact's abstraction level, not necessarily by exact waypoint correspondence: high-level reasoning may affect maneuver class, yielding behavior, risk profile, trajectory family, control decision, or closed-loop outcome through a downstream planner, vehicle dynamics, map constraints, or a safety override.

Adaptive reasoning budget. Routine lane following may require fast reactive control, while ambiguous intersections or rare hazards may require deeper deliberation, retrieval, or simulation. A major challenge is to allocate reasoning budget adaptively: reason more when uncertainty, risk, or conflict is high, and less when the action is obvious, which requires uncertainty estimation, risk-aware triggering, early exiting, and safe fallback.

Closed-loop safety evaluation. Open-loop benchmarks are insufficient because errors compound through interaction: a trace that improves offline question-answering may still destabilize closed-loop driving if it delays action, overreacts to spurious risks, or produces inconsistent plans across time. Closed-loop evaluation should measure not only collision and route completion, but also temporal stability of reasoning, recovery from incorrect intermediate states, and robustness under distribution shift, with monitors for abnormal reasoning such as missing risk factors, impossible future states, irrelevant retrievals, failed tool calls, or conflicting agent messages.

6.1 Representation-Specific Verification

Three properties of a reasoning artifact should be kept distinct: *grounding*, whether the artifact correctly reflects the scene, its geometry, rules, and dynamics; *action-faithfulness*, whether the downstream action actually relies on it; and *closed-loop safety*, whether the system remains safe under interaction, latency, and error accumulation. A grounded artifact may be ignored by the planner, while an erroneous artifact may be relied upon and produce an unsafe action, so verification mechanisms must be representation-specific rather than limited to final-action metrics. For language-based artifacts, check whether referenced objects, relations, maps, and rules agree with the scene, and whether valid changes to action-relevant content are accompanied by compatible downstream changes at the corresponding abstraction level. For visual-spatial artifacts, check geometric validity, lane topology, temporal consistency, and whether decision-critical spatial evidence is covered. For latent-dynamic artifacts, decode or probe motion and risk variables, test transition consistency, and report uncertainty or rollout disagreement. For externalized artifacts, check source provenance, freshness, context match, tool-output validity, latency, and conflicts between external information and local observations. Inconsistencies among perception, the intermediate artifact, and the planned action are flagged for separate checking rather than assuming that any one component is ground truth. These are evaluation targets, not evidence that current systems provide formal safety guarantees.

7 Conclusion

Driving is action-oriented, perception-based, and latency-constrained, so intermediate reasoning has expanded beyond text into language-based, visual-spatial, latent-dynamic, and externalized representations: language supports inspection, visual evidence improves grounding, latent dynamics strengthen action coupling, and externalized reasoning unlocks knowledge and cooperation. No single representation suffices; future systems should combine them to be grounded, causally linked to behavior, efficient, and monitored in closed-loop driving. The future of driving CoT is not longer textual chains, but reasoning representations that can be executed and trusted in safety-critical systems.

Limitations

This survey has several limitations. First, the field of CoT and foundation models for autonomous driving is evolving rapidly, and new arXiv papers and technical reports continue to appear at a high frequency. Our corpus therefore reflects the publicly available literature at the time of writing and may not cover every very recent work. Second, our taxonomy assigns each method to its primary reasoning representation. This improves clarity, but inevitably simplifies hybrid systems that combine language reasoning, visual grounding, latent planning, retrieval, tool use, or agent communication. Alternative categorizations may be reasonable for methods whose reasoning-action pipeline relies on multiple representations. Finally, our scope is centered on intermediate reasoning representations rather than all foundation-model-based autonomous driving research. We include perception, prediction, planning, simulation, VLA training, and world-modeling works only when their intermediate states explicitly serve a decision-relevant reasoning role. This boundary enables a focused analysis of the transition from textual CoT to action-grounded reasoning, but it also excludes some important autonomous driving methods outside this focus. This survey conducts no author-run ablation, intervention study, or matched-condition comparison and therefore makes no causal claim about the superiority or action-faithfulness of any representation family. Its original contribution is a corpus-level analysis of representation-action interfaces and of the evidence used to evaluate them.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, and 1 others. 2022. Do as I can, not as I say: Grounding language in robotic affordances. In *Conference on Robot Learning (CoRL)*.

Hidehisa Arai, Keita Miwa, Kento Sasaki, Kohei Watanabe, Yu Yamaguchi, Shunsuke Aoki, and Issei Yamamoto. 2025. Covla: Comprehensive vision-language-action dataset for autonomous driving. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1933–1943. IEEE.

Mehdi Azarafza, Mojtaba Nayyeri, Charles Steinmetz, Steffen Staab, and Achim Rettberg. 2024. Hybrid reasoning based on large language models for autonomous car driving. In *2024 12th International Conference on Control, Mechatronics and Automation (ICCMA)*, pages 14–22. IEEE.

ByteDance Seed. 2025. Uniugp: Unifying understanding, generation, and planning for end-to-end autonomous driving (bytedance seed). *arXiv preprint arXiv:2512.09864*.

Caesar, Holger, Bankiti, Varun, Lang, Alex H., Vora, Sourabh, Liong, Venice Erin, Xu, Qiang, Krishnan, Anush, Pan, Yu, Baldan, Giancarlo, Beijbom, and Oscar. 2020. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Caesar, Holger, Kabzan, Juraj, Tan, Kok Seang, Fong, Whye Kit, Wolff, Eric, Lang, Alex, Fletcher, Luke, Beijbom, Oscar, Omari, and Sammy. 2021. nuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles. In *CVPR Workshop on Autonomous Driving*.

Tianhui Cai, Yifan Liu, Zewei Zhou, Haoxuan Ma, Seth Z Zhao, Zhiwen Wu, Xu Han, Zhiyu Huang, and Jiaqi Ma. 2026. Driving with regulation: Trustworthy and interpretable decision-making for autonomous driving with retrieval-augmented reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38287–38295.

Kevin Cannons, Saeed Ranjbar Alvar, Mohammad Asiful Hossain, Ahmad Rezaei, Mohsen Gholami, Alireza Heidarikhazaei, Zhou Weimin, Yong Zhang, and Mohammad Akbari. 2025. From segments to scenes: Temporal understanding in autonomous driving via vision-language models (huawei). *arXiv preprint arXiv:2512.05277*.

Amirhosein Chahe and Lifeng Zhou. 2025. Reasondrive: Efficient visual question answering for autonomous vehicles with reasoning-enhanced small vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3870–3879.

Cheng Chang, Jingwei Ge, Jiazhe Guo, Zelin Guo, Binghong Jiang, and Li Li. 2026. Driving-rag: Driving scenarios embedding, search, and rag applications. *Automotive Innovation*, pages 1–13.

Jiayi Chen, Shuai Wang, Guangxu Zhu, and Chengzhong Xu. 2026. Bridging large-model reasoning and real-time control via agentic fast-slow planning. *arXiv preprint arXiv:2604.01681*.

Xuesong Chen, Linjiang Huang, Tao Ma, Rongyao Fang, Shaoshuai Shi, and Hongsheng Li. 2025a. Solve: Synergy of language-vision and end-to-end networks for autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12068–12077.

Yuntao Chen, Yuqi Wang, and Zhaoxiang Zhang. 2025b. Drivinggpt: Unifying driving world modeling and

planning with multi-modal autoregressive transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 26890–26900.

Haohan Chi, Huan-ang Gao, Ziming Liu, Jianing Liu, Chenyu Liu, Jinwei Li, Kaisen Yang, Yangcheng Yu, Zeda Wang, Wenyi Li, and 1 others. 2026. Impromptu vla: Open weights and open data for driving vision-language-action models. *Advances in Neural Information Processing Systems*, 38.

Hsu-kuang Chiu, Ryo Hachiuma, Chien-Yi Wang, Stephen F Smith, Yu-Chiang Frank Wang, and Min-Hung Chen. 2025. V2v-llm: Vehicle-to-vehicle cooperative autonomous driving with multi-modal large language models. *arXiv preprint arXiv:2502.09980*.

Charles Corbiere, Simon Roburin, Syrielle Montariol, Antoine Bosselut, and Alexandre Alahi. 2025. Retrieval-based interleaved visual chain-of-thought in real-world driving scenarios. *arXiv preprint arXiv:2501.04671*.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. 2024. Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles. *IEEE Intelligent Transportation Systems Magazine*, 16(4):81–94.

Can Cui, Yunsheng Ma, Sung-Yeon Park, Zichong Yang, Yupeng Zhou, Peiran Liu, Juanwu Lu, Juntong Peng, Jiaru Zhang, Ruqi Zhang, and 1 others. 2026. Llm4ad: Large language models for autonomous driving—concept, review, benchmark, experiments, and future trends. *Proceedings of the IEEE*.

Yixin Cui, Haotian Lin, Shuo Yang, Yixiao Wang, Yanjun Huang, and Hong Chen. 2025. Chain-of-thought for autonomous driving: A comprehensive survey and future prospects. *arXiv preprint arXiv:2505.20223*.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, and 1 others. 2024. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37:28706–28719.

Muxi Diao, Lele Yang, Hongbo Yin, Zhexu Wang, Yejie Wang, Daxin Tian, Kongming Liang, and Zhanyu Ma. 2025. Driverx: A vision-language reasoning model for cross-task autonomous driving. *arXiv preprint arXiv:2505.20665*.

Shiyu Fang, Jiaqi Liu, Chengkai Xu, Chen Lv, Peng Hang, and Jian Sun. 2025. Interact, instruct to improve: A llm-driven parallel actor-reasoner framework for enhancing autonomous vehicle interactions. *IEEE Transactions on Intelligent Transportation Systems*.

Bowen Feng, Zhiting Mei, Julian Ost, Filippo Ghilotti, Baiang Li, Roger Girgis, Anirudha Majumdar, and Felix Heide. 2025. Verdi: Vlm-embedded reasoning for autonomous driving. *arXiv preprint arXiv:2505.15925*.

Mohamed Amine Ferrag, Abderrahmane Lakas, and Merouane Debbah. 2026. Agentdrive: An open benchmark dataset for agentic ai reasoning with llm-generated scenarios in autonomous systems. *arXiv preprint arXiv:2601.16964*.

Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. 2025a. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24823–24834.

Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Hongwei Xie, Bing Wang, Guang Chen, Dingkang Liang, and Xiang Bai. 2025b. Minddrive: A vision-language-action model for autonomous driving via online reinforcement learning. *arXiv preprint arXiv:2512.13636*.

Wenbin Gan, Minh-Son Dao, and Koji Zettsu. 2025. Case-based reasoning augmented large language model framework for decision making in realistic safety-critical driving scenarios. *Safety Science*, 201:107234.

Milan Ganai, Katie Luo, Jonas Frey, Clark Barrett, and Marco Pavone. 2026. Self-supervised bootstrapping of action-predictive embodied reasoning. *arXiv preprint arXiv:2602.08167*.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. 2024. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596.

Tian Gao, Celine Tan, Catherine Glossop, Timothy Gao, Jiankai Sun, Kyle Stachowicz, Shirley Wu, Oier Mees, Dorsa Sadigh, Sergey Levine, and 1 others. 2026. Steervla: Steering vision-language-action models in long-tail driving scenarios. *arXiv preprint arXiv:2602.08440*.

Xiangbo Gao, Yuheng Wu, Rujia Wang, Chenxi Liu, Yang Zhou, and Zhengzhong Tu. 2025a. Langcoop: Collaborative driving with language. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4226–4237.

Yu Gao, Anqing Jiang, Yiru Wang, Wang Jijun, Hao Jiang, Zhigang Sun, Heng Yuwen, Wang Shuo, Hao Zhao, and Sun Hao. 2025b. Diffvla++: Bridging cognitive reasoning and end-to-end driving through metric-guided alignment. *arXiv preprint arXiv:2510.17148*.

Anurag Ghosh, Srinivasa Narasimhan, Manmohan Chandraker, and Francesco Pittaluga. 2026. Rad-lad:

Rule and language grounded autonomous driving in real-time. *arXiv preprint arXiv:2603.28522*.

Omar Y Goba, Ahmed Y Gado, Catherine M Elias, and Ahmed Hussein. 2025. From prompts to pavement: Lmms-based agentic behavior-tree generation framework for autonomous vehicles. In *2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1637–1643. IEEE.

Mihir Godbole, Xiangbo Gao, and Zhengzhong Tu. 2025. Drama-x: A fine-grained intent prediction and risk reasoning benchmark for driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 815–820.

Yi Gu, Yan Wang, Yuxiao Chen, Yurong You, Wenjie Luo, Yue Wang, Wenhao Ding, Boyi Li, Heng Yang, Boris Ivanovic, and 1 others. 2026. Accelerating structured chain-of-thought in autonomous vehicles. *arXiv preprint arXiv:2602.02864*.

Ziang Guo, Feng Yang, Xuefeng Zhang, Jiaqi Guo, Kun Zhao, Yixiao Zhou, Peng Lu, Sifa Zheng, and Zufeng Zhang. 2026. Listen, look, drive: Coupling audio instructions for user-aware vla-based autonomous driving. *arXiv preprint arXiv:2601.12142*.

Xu Han, Zhiwen Wu, Xin Xia, and Jiaqi Ma. 2025. Traffic regulation-aware path planning with regulation databases and vision-language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11024–11030. IEEE.

Yuxuan Han, Kunyuan Wu, Qianyi Shao, Renxiang Xiao, Zilu Wang, Cansen Jiang, Yi Xiao, Liang Hu, and Yunjiang Lou. 2026. Applevlm: End-to-end autonomous driving with advanced perception and planning-enhanced vision-language models. *arXiv preprint arXiv:2602.04256*.

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

Xinmeng Hou, Wuqi Wang, Long Yang, Hao Lin, Jinglun Feng, Haigen Min, and Xiangmo Zhao. 2025. Driveagent: Multi-agent structured reasoning with llm and multimodal sensor fusion for autonomous driving. *IEEE Robotics and Automation Letters*.

Senkang Hu, Zhengru Fang, Zihan Fang, Yiqin Deng, Xianhao Chen, and Yuguang Fang. 2024. Agentscodriver: Large language model empowered collaborative driving with lifelong learning. *arXiv preprint arXiv:2404.06345*.

Tianshuai Hu, Xiaolu Liu, Song Wang, Yiyao Zhu, Ao Liang, Lingdong Kong, Guoyang Zhao, Zeying Gong, Jun Cen, Zhiyu Huang, and 1 others. 2025. Vision-language-action models for autonomous driving: Past, present, and future. *arXiv preprint arXiv:2512.16760*.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, and 1 others. 2023. Inner monologue: Embodied reasoning through planning with language models. In *Conference on Robot Learning*, pages 1769–1782. PMLR.

Zhijian Huang, Tao Tang, Shaoxiang Chen, Sihao Lin, Zequn Jie, Lin Ma, Guangrun Wang, and Xiaodan Liang. 2024. Making large language models better planners with reasoning-decision alignment. In *European Conference on Computer Vision*, pages 73–90. Springer.

Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, and 1 others. 2024. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262, Accepted by TMLR*.

Ayesha Ishaq, Jean Lahoud, Ketan More, Omkar Thawakar, Ritesh Thawkar, Dinura Dissanayake, Noor Ahsan, Yuhao Li, Fahad Shahbaz Khan, Hisham Cholakkal, and 1 others. 2025. Drivelmm-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 20501–20508. IEEE.

Xiaosong Jia, Yuqian Shao, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. 2026. Bench2drive-vl: Benchmarks for closed-loop autonomous driving with vision-language models. *arXiv preprint arXiv:2604.01259*.

Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. 2024. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *Advances in Neural Information Processing Systems*, 37:819–844.

Anqing Jiang, Yu Gao, Zhigang Sun, Yiru Wang, Jijun Wang, Jinghao Chai, Qian Cao, Yuweng Heng, Hao Jiang, Yunda Dong, and 1 others. 2025a. Diffvla: Vision-language guided diffusion planning for autonomous driving. *arXiv preprint arXiv:2505.19381*.

Bo Jiang, Shaoyu Chen, Qian Zhang, Wenyu Liu, and Xinggang Wang. 2025b. Alphadrive: Unleashing the power of vlms in autonomous driving via reinforcement learning and reasoning. *arXiv preprint arXiv:2503.07608*.

Kemou Jiang, Xuan Cai, Zhiyong Cui, Aoyong Li, Yilong Ren, Haiyang Yu, Hao Yang, Daocheng Fu, Licheng Wen, and Pinlong Cai. 2024. Koma: Knowledge-driven multi-agent framework for autonomous driving with large language models. *IEEE Transactions on Intelligent Vehicles*.

Sicong Jiang, Zilin Huang, Kangan Qian, Ziang Luo, Tianze Zhu, Yang Zhong, Yihong Tang, Menglin Kong, Yunlong Wang, Siwen Jiao, and 1 others.

2025c. A survey on vision-language-action models for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4524–4536.

Jingyu Li, Junjie Wu, Dongnan Hu, Xiangkai Huang, Bin Sun, Zhihui Hao, Xianpeng Lang, Xiatian Zhu, and Li Zhang. 2026a. Sgdrive: Scene-to-goal hierarchical world cognition for autonomous driving. *arXiv preprint arXiv:2601.05640*.

Lin Li, Yuxin Cai, Jianwu Fang, Jianru Xue, and Chen Lv. 2025a. Covlm-rl: Critical object-oriented reasoning for autonomous driving using vlm-guided reinforcement learning. *arXiv preprint arXiv:2512.09349*.

Peizheng Li, Zhenghao Zhang, David Holtz, Hang Yu, Yutong Yang, Yuzhi Lai, Rui Song, Andreas Geiger, and Andreas Zell. 2025b. Spacedrive: Infusing spatial awareness into vlm-based autonomous driving. *arXiv preprint arXiv:2512.10719*, 2.

Pengxiang Li, Yinan Zheng, Yue Wang, Huimin Wang, Hang Zhao, Jingjing Liu, Xianyuan Zhan, Kun Zhan, and Xianpeng Lang. 2025c. Discrete diffusion for reflective vision-language-action models in autonomous driving. *arXiv preprint arXiv:2509.20109*.

Yiheng Li, Cunxin Fan, Chongjian Ge, Seth Z Zhao, Chenran Li, Chenfeng Xu, Huaxiu Yao, Masayoshi Tomizuka, Bolei Zhou, Chen Tang, and 1 others. 2025d. Womd-reasoning: A large-scale dataset for interaction reasoning in driving. In *International Conference on Machine Learning*, pages 34288–34311. PMLR.

Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, and 1 others. 2025e. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. *arXiv preprint arXiv:2506.08052*.

Yue Li, Meng Tian, Zhenyu Lin, Jiangtong Zhu, Dechang Zhu, Haiqiang Liu, Yueyi Zhang, Zhiwei Xiong, and Xinhai Zhao. 2025f. Fine-grained evaluation of large vision-language models in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9431–9442.

Yue Li, Meng Tian, Dechang Zhu, Jiangtong Zhu, Zhenyu Lin, Zhiwei Xiong, and Xinhai Zhao. 2026b. Drive-r1: Bridging reasoning and planning in vlms for autonomous driving with reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 6708–6716.

Dacheng Liao, Mengshi Qi, Peng Shu, Zhining Zhang, Yuxin Lin, Liang Liu, and Huadong Ma. 2025a. Robodrivevlm: A novel benchmark and baseline towards robust vision-language models for autonomous driving. *arXiv preprint arXiv:2512.01300*.

Haicheng Liao, Hanlin Kong, Bonan Wang, Chengyue Wang, Kanye Ye Wang, Zhengbing He, Chengzhong Xu, and Zhenning Li. 2025b. Cot-drive: Efficient motion forecasting for autonomous driving with llms and chain-of-thought prompting. *IEEE Transactions on Artificial Intelligence*, 7(2):625–641.

Haicheng Liao, Huanming Shen, Bonan Wang, Yongkang Li, Yihong Tang, Chengyue Wang, Dingyi Zhuang, Kehua Chen, Hai Yang, Chengzhong Xu, and 1 others. 2025c. Think before you drive: World model-inspired multimodal grounding for autonomous vehicles. *arXiv preprint arXiv:2512.03454*.

Hongbin Lin, Yiming Yang, Yifan Zhang, Chaoda Zheng, Jie Feng, Sheng Wang, Zhennan Wang, Shijia Chen, Boyang Wang, Yu Zhang, and 1 others. 2025. Futurex: Enhance end-to-end autonomous driving via latent chain-of-thought world model. *arXiv preprint arXiv:2512.11226*.

Changxing Liu, Genjia Liu, Zijun Wang, Jinchang Yang, and Siheng Chen. 2025a. Colmdriver: Llm-based negotiation benefits cooperative autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 25951–25960.

Haichao Liu, Haoren Guo, Pei Liu, Benshan Ma, Yuxiang Zhang, Jun Ma, and Tong Heng Lee. 2025b. Vlm-udmc: Vlm-enhanced unified decision-making and motion control for urban autonomous driving. *arXiv preprint arXiv:2507.15266*.

Lin Liu, Ziying Song, Caiyan Jia, Hangjun Ye, Xiaoshuai Hao, Long Chen, and 1 others. 2026. Driveworld-vla: Unified latent-space world modeling with vision-language-action for autonomous driving. *arXiv preprint arXiv:2602.06521*.

Pei Liu, Qingtian Ning, Xinyan Lu, Haipeng Liu, Weiliang Ma, Dangen She, Peng Jia, Xianpeng Lang, and Jun Ma. 2025c. Omnireason: A temporal-guided vision-language-action framework for autonomous driving. *arXiv preprint arXiv:2509.00789*.

Wei Liu, Jiyuan Zhang, Binxiong Zheng, Yufeng Hu, Yingzhan Lin, and Zengfeng Zeng. 2025d. Xdriver: Explainable autonomous driving with vision-language models. *arXiv preprint arXiv:2505.05098*.

Wenru Liu, Pei Liu, and Jun Ma. 2025e. Dsdrive: Distilling large language model for lightweight end-to-end autonomous driving with unified reasoning and planning. *arXiv preprint arXiv:2505.05360*.

Xueyi Liu, Zuodong Zhong, Qichao Zhang, Yuxin Guo, Yupeng Zheng, Junli Wang, Dongbin Zhao, Yun-Fu Liu, Zhiguo Su, Yinfeng Gao, and 1 others. 2025f. Reasonplan: Unified scene prediction and decision reasoning for closed-loop autonomous driving. In *Conference on Robot Learning*, pages 3051–3068. PMLR.

Yudong Liu, Spencer Hallyburton, Jiwoo Kim, Yueqian Lin, Yiming Li, Qinsi Wang, Hui Ye, Jingwei Sun,

Miroslav Pajic, Yiran Chen, and 1 others. 2025g. Llavida: A large language vision driving assistant for explicit reasoning and enhanced trajectory planning. *arXiv preprint arXiv:2512.18211*.

Jinghui Lu, Jiayi Guan, Zhijian Huang, Jinlong Li, Guang Li, Lingdong Kong, Yingyan Li, Han Wang, Shaoqing Xu, Yuechen Luo, and 1 others. 2026. Onevl: One-step latent reasoning and planning with vision-language explanation. *arXiv preprint arXiv:2604.18486*.

Yuhang Lu, Jiadong Tu, Yuexin Ma, and Xinge Zhu. 2025. Real-ad: Towards human-like reasoning in end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27783–27793.

Xuewen Luo, Fan Ding, Yinsheng Song, Xiaofeng Zhang, and Junnyong Loo. 2024. Pkrd-cot: A unified chain-of-thought prompting for multi-modal large language models in autonomous driving. In *International Conference on Neural Information Processing*, pages 62–76. Springer.

Xuewen Luo, Chenxi Liu, Fan Ding, Fengze Yang, Yang Zhou, Junnyong Loo, and Hwa Hui Tew. 2025a. Senserag: Constructing environmental knowledge bases with proactive querying for llm-based autonomous driving. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 989–996.

Xuewen Luo, Fengze Yang, Fan Ding, Xiangbo Gao, Shuo Xing, Yang Zhou, Zhengzhong Tu, and Chenxi Liu. 2025b. V2x-unipool: Unifying multimodal perception and knowledge reasoning for autonomous driving. *arXiv preprint arXiv:2506.02580*.

Yuechen Luo, Qimao Chen, Fang Li, Shaoqing Xu, Jaxin Liu, Ziying Song, Zhi-xin Yang, and Fuxi Wen. 2026a. Unleashing vla potentials in autonomous driving via explicit learning from failures. *arXiv preprint arXiv:2603.01063*.

Yuechen Luo, Fang Li, Shaoqing Xu, Yang Ji, Zehan Zhang, Bing Wang, Yuannan Shen, Jianwei Cui, Long Chen, Guang Chen, and 1 others. 2026b. Last-vla: Thinking in latent spatio-temporal space for vision-language-action in autonomous driving. *arXiv preprint arXiv:2603.01928*.

Yuechen Luo, Fang Li, Shaoqing Xu, Zhiyi Lai, Lei Yang, Qimao Chen, Ziang Luo, Zixun Xie, Shengyin Jiang, Jiaxin Liu, and 1 others. 2025c. Adathinkdrive: Adaptive thinking via reinforcement learning for autonomous driving. *arXiv preprint arXiv:2509.13769*.

Ziang Luo, Kangan Qian, Jiahua Wang, Yuechen Luo, Jinyu Miao, Zheng Fu, Yunlong Wang, Sicong Jiang, Zilin Huang, Yifei Hu, and 1 others. 2025d. Mtr-drive: Memory-tool synergistic reasoning for robust autonomous driving in corner cases. *arXiv preprint arXiv:2509.20843*.

Enhui Ma, Jiahuan Zhang, Guantian Zheng, Tao Tang, Shengbo Eben Li, Yuhang Lu, Xia Zhou, Xueyang Zhang, Yifei Zhan, Kun Zhan, and 1 others. 2026. Drivecombo: Benchmarking compositional traffic rule reasoning in autonomous driving. *arXiv preprint arXiv:2603.01637*.

Yingzi Ma, Yulong Cao, Wenhao Ding, Shuibai Zhang, Yan Wang, Boris Ivanovic, Ming Jiang, Marco Pavone, and Chaowei Xiao. 2025. dvlm-ad: Enhance diffusion vision-language-model for driving via controllable reasoning. *arXiv preprint arXiv:2512.04459*.

Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. 2024. Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*, pages 403–420. Springer.

Sriram Mandalika, Athira Nambiar, and 1 others. 2025. Primedrive-cot: A precognitive chain-of-thought framework for uncertainty-aware object interaction in driving scene scenario. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5293–5301.

Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. 2023. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*.

Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. 2024. A language agent for autonomous driving. In *First Conference on Language Modeling*.

Angel Martinez-Sanchez, Parthib Roy, and Ross Greer. 2026. Natural language instructions for scene-responsive human-in-the-loop motion planning in autonomous driving using vision-language-action models. *arXiv preprint arXiv:2602.04184*.

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

Jay Patrikar, Apoorva Sharma, Sushant Veer, Boyi Li, Sebastian Scherer, and Marco Pavone. 2025. The case for negative data: From crash reports to counterfactuals for reasonable driving. *arXiv preprint arXiv:2509.18626*.

Jinghan Peng, Jingwen Wang, Xing Yu, and Dehui Du. 2025a. The system description of cps team for track on driving with language of cvpr 2024 autonomous grand challenge. *arXiv preprint arXiv:2509.11071*.

Mingxing Peng, Xusen Guo, Xianda Chen, Kehua Chen, Meixin Zhu, Long Chen, and Fei-Yue Wang. 2025b. Lc-llm: Explainable lane-change intention and trajectory predictions with large language models. *Communications in Transportation Research*, 5:100170.

Qihang Peng, Xuesong Chen, Chenye Yang, Shaoshuai Shi, and Hongsheng Li. 2025c. Colavla: Leveraging cognitive latent reasoning for hierarchical parallel trajectory planning in autonomous driving. *arXiv preprint arXiv:2512.22939*.

Zhenghao Peng, Wenhao Ding, Yurong You, Yuxiao Chen, Wenjie Luo, Thomas Tian, Yulong Cao, Apoorva Sharma, Danfei Xu, Boris Ivanovic, and 1 others. 2025d. Counterfactual vla: Self-reflective vision-language-action model with adaptive reasoning. *arXiv preprint arXiv:2512.24426*.

Kangan Qian, Sicong Jiang, Yang Zhong, Ziang Luo, Zilin Huang, Tianze Zhu, Kun Jiang, Mengmeng Yang, Zheng Fu, Jinyu Miao, and 1 others. 2025a. Agentthink: A unified framework for tool-augmented chain-of-thought reasoning in vision-language models for autonomous driving. *arXiv preprint arXiv:2505.15298*, 1(2):3.

Kangan Qian, Ziang Luo, Sicong Jiang, Zilin Huang, Jinyu Miao, Zhikun Ma, Tianze Zhu, Jiayin Li, Yangfan He, Zheng Fu, and 1 others. 2025b. Fasionad++: Integrating high-level instruction and information bottleneck in fat-slow fusion systems for enhanced safety in autonomous driving with adaptive feedback. *arXiv preprint arXiv:2503.08162*.

Kangan Qian, Zhikun Ma, Yangfan He, Ziang Luo, Tianyu Shi, Tianze Zhu, Jiayin Li, Jianhui Wang, Ziyu Chen, Xiao He, and 1 others. 2024. Fasionad: Fast and slow fusion thinking systems for human-like autonomous driving with adaptive feedback. *arXiv preprint arXiv:2411.18013*.

Zhijie Qiao, Haowei Li, Zhong Cao, and Henry X Liu. 2025. Lightemma: Lightweight end-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2505.00284*.

Hao Sha, Yao Mu, Yuxuan Jiang, Li Chen, Chenfeng Xu, Ping Luo, Shengbo Eben Li, Masayoshi Tomizuka, Wei Zhan, and Mingyu Ding. 2025. Languagempc: Large language models as decision makers for autonomous driving. *arXiv preprint arXiv:2310.03026*.

Shuyao Shang, Bing Zhan, Yunfei Yan, Yuqi Wang, Yingyan Li, Yasong An, Xiaoman Wang, Jierui Liu, Lu Hou, Lue Fan, and 1 others. 2026. Dynvla: Learning world dynamics for action reasoning in autonomous driving. *arXiv preprint arXiv:2603.11041*.

SP Sharan, Francesco Pittaluga, Manmohan Chandraker, and 1 others. 2023. Llm-assist: Enhancing closed-loop planning with language-based reasoning. *arXiv preprint arXiv:2401.00125*.

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

Xurui Song, Shuo Huai, JingJing Jiang, Jiayi Kong, and Jun Luo. 2025. More than meets the eye? uncovering the reasoning-planning disconnect in training vision-language driving models. *arXiv preprint arXiv:2510.04532*.

Shuhan Tan, Kashyap Chitta, Yuxiao Chen, Ran Tian, Yurong You, Yan Wang, Wenjie Luo, Yulong Cao, Philipp Krahenbuhl, Marco Pavone, and 1 others. 2025. Latent chain-of-thought world modeling for end-to-end driving. *arXiv preprint arXiv:2512.10226*.

Kotaro Tanahashi, Yuichi Inoue, Yu Yamaguchi, Hidetatsu Yaginuma, Daiki Shiotsuka, Hiroyuki Shimatani, Kohei Iwamasa, Yoshiaki Inoue, Takafumi Yamaguchi, Koki Igari, Tsukasa Horinouchi, Kento Tokuhiro, Yugo Tokuchi, and Shunsuke Aoki. 2023. Evaluation of large language models for decision making in autonomous driving. *arXiv preprint arXiv:2312.06351*.

Zecong Tang, Zixu Wang, Yifei Wang, Weitong Lian, Tianjian Gao, Haoran Li, Tengju Ru, Lingyi Meng, Zhejun Cui, Yichen Zhu, and 1 others. 2026. Autodridm: An explainable benchmark for decision-making of vision-language models in autonomous driving. *arXiv preprint arXiv:2601.14702*.

Ximeng Tao, Pardis Taghavi, Dimitar Filev, Reza Langari, and Gaurav Pandey. 2026. Navidrivevlm: Decoupling high-level reasoning and motion planning for autonomous driving. *arXiv preprint arXiv:2603.07901*.

Kexin Tian, Jingrui Mao, Yunlong Zhang, Jiwan Jiang, Yang Zhou, and Zhengzhong Tu. 2025a. Nuscenes-spatialqa: A spatial understanding and reasoning benchmark for vision-language models in autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4567–4576.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. 2025b. Drivevlm: The convergence of autonomous driving and large vision-language models. In *Conference on Robot Learning*, pages 4698–4726. PMLR.

Ye Tian, Jingyi Zhang, Zihao Wang, Xiaoyuan Ren, Xiaofan Yu, Onat Gungor, and Tajana Rosing. 2026. Kldrive: Fine-grained 3d scene reasoning for autonomous driving based on knowledge graph. *arXiv preprint arXiv:2603.21029*.

Daming Wang, Yuhao Song, Zijian He, Kangliang Chen, Xing Pan, Lu Deng, and Weihao Gu. 2025a. Hmvlm: Multistage reasoning-enhanced vision-language model for long-tailed driving scenarios. *arXiv preprint arXiv:2506.05883*.

Dingrui Wang, Marc Kaufeld, and Johannes Betz. 2025b. Dualad: Dual-layer planning for reasoning in autonomous driving. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12057–12063. IEEE.

Guoqing Wang, Pin Tang, Xiangxuan Ren, Guodongfang Zhao, Bailan Feng, and Chao Ma. 2026a. Learning vision-language-action world models for autonomous driving. *arXiv preprint arXiv:2604.09059*.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. 2025c. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *Proceedings of the computer vision and pattern recognition conference*, pages 22442–22452.

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

Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, and 1 others. 2023a. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*.

Yan Wang, Wenjie Luo, Junjie Bai, Yulong Cao, Tong Che, Ke Chen, Yuxiao Chen, Jenna Diamond, Yifan Ding, Wenhao Ding, and 1 others. 2025d. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*.

Yiru Wang, Zichong Gu, Yu Gao, Anqing Jiang, Zhigang Sun, Shuo Wang, Yuwen Heng, and Hao Sun. 2026b. Hist-vla: A hierarchical spatio-temporal vision-language-action model for end-to-end autonomous driving. *arXiv preprint arXiv:2602.13329*.

Yujin Wang, Quanfeng Liu, Zhengxin Jiang, Tianyi Wang, Junfeng Jiao, Hongqing Chu, Bingzhao Gao, and Hong Chen. 2025e. Rad: Retrieval-augmented decision-making of meta-actions with vision-language models in autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3838–3848.

Yujin Wang, Tianyi Wang, Quanfeng Liu, Wenxian Fan, Junfeng Jiao, Christian Claudel, Yunbing Yan, Bingzhao Gao, Jianqiang Wang, and Hong Chen. 2026c. Kept: Knowledge-enhanced prediction of trajectories from consecutive driving frames with vision-language models. *Communications in Transportation Research*, 6(1):9640012.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. 2024b. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759.

Zhaohui Wang, Tengbo Yu, and Hao Tang. 2025f. Cot4ad: A vision-language-action model with explicit chain-of-thought reasoning for autonomous driving. *arXiv preprint arXiv:2511.22532*.

Dawood Wasif, Terrence J. Moore, Chandan K. Reddy, Frederica Free-Nelson, Seunghyun Yoon, Hyuk Lim, Dan Dongseong Kim, and Jin-Hee Cho. 2026. Drivemind: A dual visual language model-based reinforcement learning framework for autonomous driving. *arXiv preprint arXiv:2506.00819*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Maolin Wei, Wanzhou Liu, and Eshed Ohn-Bar. 2025a. Driveqa: Passing the driving knowledge test. *arXiv preprint arXiv:2508.21824*.

Zhaoyang Wei, Chenhui Qiang, Bowen Jiang, Xumeng Han, Xuehui Yu, and Zhenjun Han. 2025b. Adˆ 2-bench: A hierarchical cot benchmark for mllm in autonomous driving under adverse conditions. *arXiv preprint arXiv:2506.09557*.

Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. 2024. Dilu: A knowledge-driven approach to autonomous driving with large language models. In *International Conference on Learning Representations*, volume 2024, pages 34503–34522.

Aodi Wu and Xubo Luo. 2025. Enhancing vision-language models for autonomous driving through task-specific prompting and spatial reasoning. *arXiv preprint arXiv:2510.24152*.

Jialong Xie, Yi Yang, Xing Jiabin, Jianyun Xu, Sheng Yang, and Fengyu Zhou. 2025. Deliberation meets reaction: A dual-expert vla framework for autonomous driving. Unpublished manuscript. Available at https://openreview.net/forum?id=rHcWxVrDFV.

Shuo Xing, Chengyuan Qian, Yuping Wang, Hongyuan Hua, Kexin Tian, Yang Zhou, and Zhengzhong Tu. 2025. Openemma: Open-source multimodal model for end-to-end autonomous driving. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1001–1009.

Zhexiao Xiong, Xin Ye, Burhan Yaman, Sheng Cheng, Yiren Lu, Jingru Luo, Nathan Jacobs, and Liu Ren. 2026. Unidrive-wm: Unified understanding, planning and generation world model for autonomous driving. *arXiv preprint arXiv:2601.04453*.

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

Chengkai Xu, Jiaqi Liu, Yicheng Guo, Yuhang Zhang, Peng Hang, and Jian Sun. 2025b. Towards human-centric autonomous driving: A fast-slow architecture integrating large language model guidance with reinforcement learning. *arXiv preprint arXiv:2505.06875*.

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

Pengxuan Yang, Ben Lu, Zhongpu Xia, Chao Han, Yinfeng Gao, Teng Zhang, Kun Zhan, XianPeng Lang, Yupeng Zheng, and Qichao Zhang. 2026. Worldrft: Latent world model planning with reinforcement fine-tuning for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 11649–11657.

Yu Yang, Jianbiao Mei, Yukai Ma, Siliang Du, Wenqing Chen, Yijie Qian, Yuxiang Feng, and Yong Liu. 2025. Driving in the occupancy world: Vision-centric 4d occupancy forecasting and planning via world models for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9327–9335.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Yuqi Ye, Zijian Zhang, Junhong Lin, Shangkun Sun, Changhao Peng, and Wei Gao. 2026. Autodrive-p3: Unified chain of perception–prediction–planning thought via reinforcement fine-tuning. In *The Fourteenth International Conference on Learning Representations*.

Kejin Yu, Yuhan Sun, Taiqiang Wu, Ruixu Zhang, Zhiqiang Lin, Yuxin Meng, Junjie Wang, and Yujiu Yang. 2026. A survey of reasoning in autonomous driving systems: Open challenges and emerging paradigms. *arXiv preprint arXiv:2603.11093*.

Seungjun Yu, Junsung Park, Youngsun Lim, and Hyunjung Shim. 2025. Robust driving qa through metadata-grounded context and task-specific prompts. *arXiv preprint arXiv:2510.19001*.

Zhenlong Yuan, Chengxuan Qian, Jing Tang, Rui Chen, Zijian Song, Lei Sun, Xiangxiang Chu, Yujun Cai, Dapeng Zhang, and Shuo Li. 2025. Autodrive-r2: Incentivizing reasoning and self-reflection capacity for vla model in autonomous driving. *arXiv preprint arXiv:2509.01944*.

Abolfazl Zarghani, Amirhossein Ebrahimi, and Amir Malekesfandiari. 2025. Multimodal framework for explainable autonomous driving: Integrating video, sensor, and textual data for enhanced decision-making and transparency. *arXiv preprint arXiv:2507.07938*.

Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. 2026. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. *Advances in Neural Information Processing Systems*, 38:67299–67318.

Tong Zeng, Longfeng Wu, Liang Shi, Dawei Zhou, and Feng Guo. 2025. Are vision llms road-ready? a comprehensive benchmark for safety-critical driving video understanding. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5972–5983.

Dapeng Zhang, Fei Shen, Rui Zhao, Yinda Chen, Peng Zhi, Chenyang Li, Rui Zhou, and Qingguo Zhou. 2026a. Coc-vla: Delving into adversarial domain transfer for explainable autonomous driving via chain-of-causality visual-language-action model. *Advances in Neural Information Processing Systems*, 38:70912–70939.

Dapeng Zhang, Zhenlong Yuan, Zhangquan Chen, Chih-Ting Liao, Yinda Chen, Fei Shen, Qingguo Zhou, and Tat-Seng Chua. 2025a. Reasoning-vla: A fast and general vision-language-action reasoning model for autonomous driving. *arXiv preprint arXiv:2511.19912*.

Liangdong Zhang, Yiming Nie, Haoyang Li, Fanjie Kong, Baobao Zhang, Shunxin Huang, Kai Fu, Chen Min, and Liang Xiao. 2026b. A vision-language-action model with visual prompt for off-road autonomous driving. *arXiv preprint arXiv:2601.03519*.

Lingjun Zhang, Yujian Yuan, Changjie Wu, Xinyuan Chang, Xin Cai, Shuang Zeng, Linzhe Shi, Sijin Wang, Hang Zhang, and Mu Xu. 2026c. Minddriver: Introducing progressive multimodal reasoning for autonomous driving. *arXiv preprint arXiv:2602.21952*.

Ruifei Zhang, Junlin Xie, Wei Zhang, Weikai Chen, Xiao Tan, Xiang Wan, and Guanbin Li. 2025b. Adadrive: Self-adaptive slow-fast system for language-grounded autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5112–5121.

Songyan Zhang, Wenhui Huang, Zhan Chen, Chua Jiahao Collister, Qihang Huang, and Chen Lv. 2025c. Openread: Reinforced open-ended reasoning for end-to-end autonomous driving with llm-as-critic. *arXiv preprint arXiv:2512.01830*.

Songyan Zhang, Wenhui Huang, Zihui Gao, Hao Chen, and Chen Lv. 2024. Wisead: Knowledge augmented end-to-end autonomous driving with vision-language model. *arXiv preprint arXiv:2412.09951*.

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

Yuhang Zhang, Jiaqi Liu, Chengkai Xu, Peng Hang, and Jian Sun. 2025e. Lead: The llm enhanced planning system converged with end-to-end autonomous driving. *arXiv preprint arXiv:2507.05754*.

Zhenguo Zhang, Haohan Zheng, Yishen Wang, Le Xu, Tianchen Deng, Xuefeng Chen, Qu Chen, Bo Zhang, and Wuxiong Huang. 2025f. Omnidrive-r1: Reinforcement-driven interleaved multi-modal chain-of-thought for trustworthy vision-language autonomous driving. *arXiv preprint arXiv:2512.14044*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Chang Zhao, Zheming Yang, Yunqing Hu, Qi Guo, Zijian Wang, Pengcheng Li, and Wen Ji. 2026. Thinkdrive: Chain-of-thought guided progressive reinforcement learning fine-tuning for autonomous driving. *arXiv preprint arXiv:2601.04714*.

Rui Zhao, Qirui Yuan, Jinyu Li, Haofeng Hu, Yun Li, Zhenhai Gao, and Fei Gao. 2025. Sce2drivex: A generalized mllm framework for scene-to-drive learning. *IEEE Robotics and Automation Letters*.

Peiru Zheng, Yun Zhao, Zhan Gong, Hong Zhu, and Shaohua Wu. 2024. Simplellm4ad: An end-to-end vision-language model with graph visual question answering for autonomous driving. *arXiv preprint arXiv:2407.21293*.

Weicheng Zheng, Xiaofei Mao, Nanfei Ye, Pengxiang Li, Kun Zhan, Xianpeng Lang, and Hang Zhao. 2025a. Driveagent-r1: Advancing vlm-based autonomous driving with active perception and hybrid thinking. *arXiv preprint arXiv:2507.20879*.

Yupeng Zheng, Zebin Xing, Qichao Zhang, Bu Jin, Pengfei Li, Yuhang Zheng, Zhongpu Xia, Yaran Chen, and Dongbin Zhao. 2026. Planagent: A multi-modal large language agent for closed-loop vehicle motion planning. *IEEE Transactions on Cognitive and Developmental Systems*.

Yupeng Zheng, Pengxuan Yang, Zebin Xing, Qichao Zhang, Yuhang Zheng, Yinfeng Gao, Pengfei Li, Teng Zhang, Zhongpu Xia, Peng Jia, and 1 others. 2025b. World4drive: End-to-end autonomous driving via intention-aware physical latent world model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 28632–28642.

Yang Zhou, Xiaofeng Wang, Hao Shao, Letian Wang, Guosheng Zhao, Jiangnan Shao, Jiagang Zhu, Tingdong Yu, Zheng Zhu, Guan Huang, and 1 others. 2026a. Drivedreamer-policy: A geometry-grounded world-action model for unified generation and planning. *arXiv preprint arXiv:2604.01765*.

Zewei Zhou, Tianhui Cai, Seth Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. 2026b. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *Advances in Neural Information Processing Systems*, 38:27920–27956.

Zhiyuan Zhou, Heye Huang, Boqi Li, Shiyue Zhao, Yao Mu, and Jianqiang Wang. 2026c. Safedrive: Knowledge-and data-driven risk-sensitive decision-making for autonomous vehicles with large language models. *Accident Analysis & Prevention*, 224:108299.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, and 1 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR.

A Four Shifts in Driving CoT

The main body of this survey organizes driving CoT methods by the form of their intermediate reasoning representation. In this appendix, we step back from individual categories and examine the corpus from a broader historical perspective. Based on the 130 method papers surveyed in this work, we identify four shifts that jointly characterize the evolution of action-grounded reasoning in autonomous driving: an infrastructure shift from closed APIs to open backbones, a cognitive shift in what models reason about, an action shift from verbal explanation to trajectory generation, and a deployment shift from always-on reasoning to cost-aware adaptive reasoning. These shifts are not independent. Open backbones enable fine-tuning and reinforcement learning; stronger training pipelines make action-oriented reasoning possible; action-oriented reasoning exposes the faithfulness gap between explanation and control; and deployment constraints force the field to reconsider when, how, and how much a driving model should reason.

A.1 The Infrastructure Shift: From Closed APIs to Open Backbones

The earliest wave of driving CoT methods was largely built on general-purpose large language models. Systems such as GPT-Driver and LanguageMPC serialized driving scenes into textual prompts and queried GPT-style models to obtain decisions, waypoints, or control-related outputs (Mao et al., 2023; Sha et al., 2025). This design was natural at the time: closed APIs provided strong language reasoning ability, required little training infrastructure, and allowed researchers to quickly test whether language-style reasoning could be useful for driving. However, this approach also imposed a clear ceiling. Closed APIs restrict access to model weights, limit gradient-based adaptation, make latency difficult to control, and prevent tight integration with perception and planning modules.

The corpus shows a clear transition away from this paradigm. In 2023, a substantial fraction of methods still relied on GPT-3.5, GPT-4, or other closed models. By 2025 and 2026, however, the dominant backbones are open or semi-open vision-language models such as LLaVA, Qwen-VL, InternVL, and their driving-specialized variants (Tian et al., 2025b; Zheng et al., 2024; Xing et al., 2025; Liu et al., 2025g). This change is not merely a preference for open-source implementation. It changes the kind of research that becomes possible. Once model weights are available, researchers can perform supervised fine-tuning on driving-specific reasoning traces, distill reasoning into smaller models, align models with trajectory-level objectives, or optimize reasoning behavior with reinforcement learning (Feng et al., 2025; Liu et al., 2025e; Li et al., 2026b; Zheng et al., 2025a; Zhang et al., 2025f).

The rise of reinforcement learning further strengthens this trend. Methods such as Drive-R1, DriveAgent-R1, OmniDrive-R1, MindDrive, and AutoDrive-P3 use GRPO, reward-driven optimization, or related alignment strategies to improve reasoning and planning performance (Li et al., 2026b; Zheng et al., 2025a; Fu et al., 2025b; Zhang et al., 2025f; Ye et al., 2026). Such methods would be difficult or impossible to implement with closed APIs because the optimization requires direct access to model parameters or at least to trainable policy components. In this sense, the move toward open backbones is not only an engineering shift; it is a methodological precondition for reasoning alignment.

A more recent development is the emergence of world-model and VLA-style backbones. Instead of treating language models as the central reasoning engine, some methods use video diffusion models, latent world models, or vision-language-action models to encode physical dynamics more directly (Wang et al., 2025d; ByteDance Seed, 2025; Zheng et al., 2025b; Luo et al., 2026b). This suggests a deeper infrastructure transition: the backbone of driving CoT is moving from language-centric intelligence toward physically grounded, action-conditioned foundation models. The likely future is not a pure LLM planner wrapped around perception modules, but an integrated stack in which language, vision, dynamics, and action are jointly represented.

The infrastructure shift therefore has a clear implication. Early driving CoT asked whether a general LLM could reason about driving when given a textual description of the scene. Current work increasingly asks how to train, align, and deploy specialized open models whose internal representations are grounded in perception, dynamics, and action. This shift explains why the field has rapidly moved from prompting to fine-tuning, from text-only models to multimodal models, and from language reasoning to action-grounded reasoning.

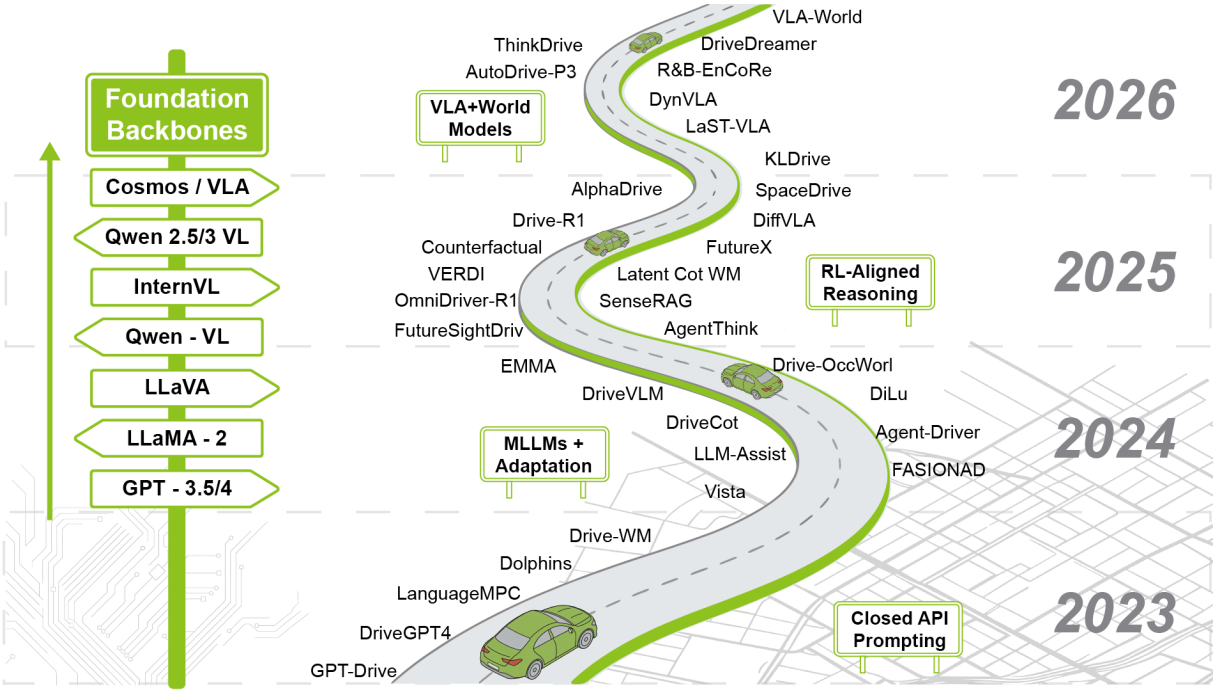


Figure 2: The infrastructure shift in driving CoT. Representative methods are placed along a timeline spanning three phases: closed-API LLMs (2023), open-weights multimodal models (2024–2025), and vision-language-action models and world models (2025–2026). Foundation models are listed on the left; the transition from prompting to SFT to RL alignment tracks directly with the shift from closed to open backbones.

A.2 The Cognitive Shift: What Do Models Actually Reason About?

Although many papers claim to perform reasoning, the content of that reasoning varies widely. Some models reason about what is visible in the scene; some reason about future motion; some reason about traffic rules; some reason about risk, uncertainty, or social interaction. To understand this variation, we classify the intermediate reasoning content in the surveyed methods into several cognitive targets, including perception and description, spatial and geometric understanding, motion and intent prediction, decision and planning, risk and safety, traffic rules, reflection and causality, world-model prediction, and social negotiation.

The most common target is decision and planning. Many methods organize their reasoning chain around the sequence from perception to prediction to planning and finally to behavior selection (Wang et al., 2023a; Tian et al., 2025b; Wang et al., 2024a; Luo et al., 2024; Liao et al., 2025b). This is understandable because autonomous driving ultimately requires an action. A reasoning trace that does not affect maneuver choice, trajectory generation, or control remains incomplete from a driving perspective. Procedural methods such as DriveVLM, DriveCoT, PKRD-CoT, and CoT4AD explicitly structure reasoning around driving-relevant stages rather than free-form explanation (Tian et al., 2025b; Wang et al., 2024a; Luo et al., 2024; Wang et al., 2025f).

The second major target is perception and scene description. Methods such as DriveGPT4, Dolphins, EMMA, OpenEMMA, and ReasonDrive use language to describe objects, road context, ego behavior, and scene-level risk factors (Xu et al., 2024; Ma et al., 2024; Hwang et al., 2024; Xing et al., 2025; Chahe and Zhou, 2025). This form of reasoning is useful for interpretability and human interaction, but it also exposes a limitation: describing a scene is not the same as reasoning about how to act. A model may correctly mention a pedestrian, vehicle, or lane boundary while still failing to produce a safe trajectory. Therefore, perception-description reasoning is necessary but not sufficient for action-grounded CoT.

Spatial and geometric reasoning forms another important target. Driving decisions depend on distances, lanes, drivable areas, occlusions, right-of-way geometry, and relative motion. Language-only reasoning is weak at preserving such metric structure, which motivates methods that incorpo-

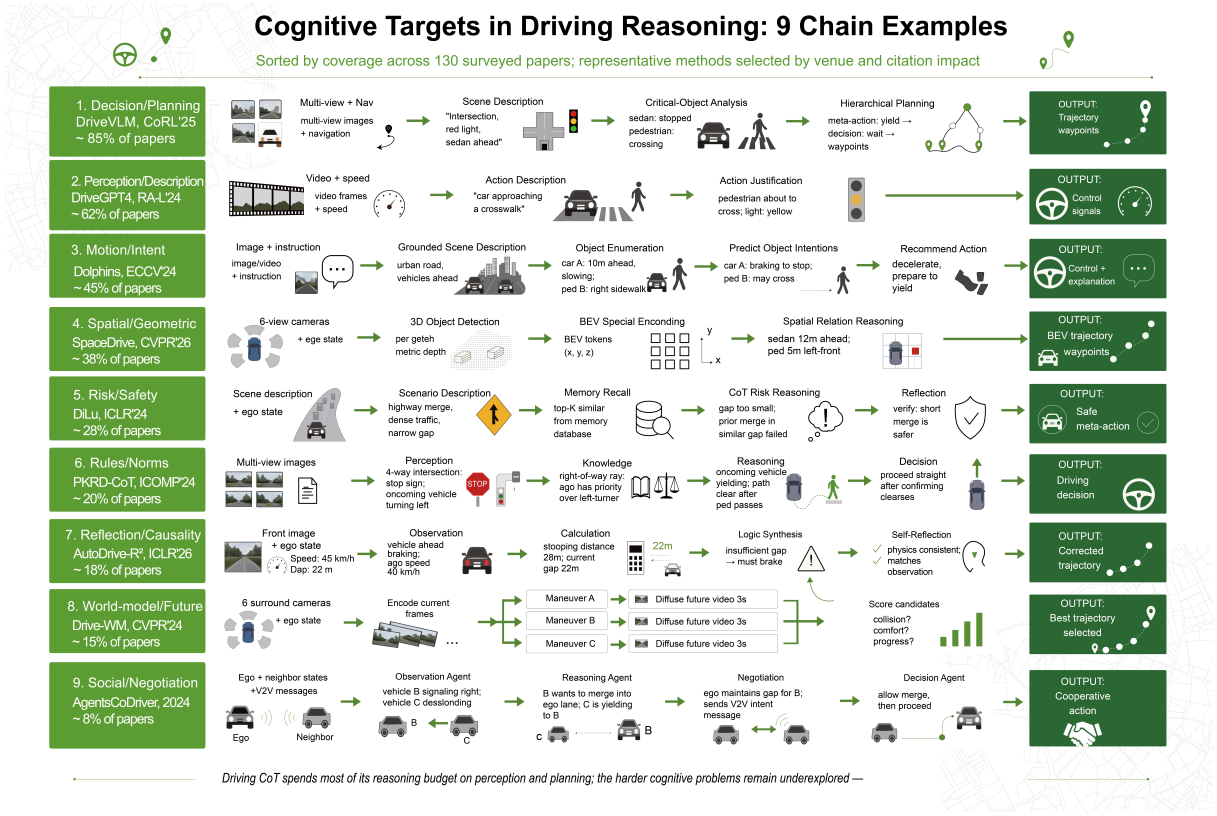


Figure 3: Nine cognitive targets of intermediate reasoning in driving CoT, ordered by frequency. Each strip shows a representative reasoning chain from one method. Strip height is proportional to the number of methods addressing each target, illustrating the concentration on decision/planning and perception tasks and the scarcity of rule reasoning and social negotiation.

rate BEV features, object-level crops, occupancy maps, or spatially grounded visual prompts (Yang et al., 2025; Corbiere et al., 2025; Li et al., 2025b; Chen et al., 2025a; Zhang et al., 2026b). Benchmarks such as NuScenes-SpatialQA also show that spatial reasoning remains a major weakness of current VLMs in driving scenarios (Tian et al., 2025a). This suggests that spatial reasoning should not be treated as a minor subtask of language explanation; it is one of the central bottlenecks for reliable driving CoT.

A smaller but increasingly important group of methods focuses on future-oriented reasoning. World-model-based approaches generate future frames, occupancy states, or latent rollouts to support planning under uncertainty (Wang et al., 2024b; Gao et al., 2024; Zeng et al., 2026; Zheng et al., 2025b; Xiong et al., 2026). These methods shift reasoning from "what is happening now" to "what may happen next." This is a crucial cognitive transition because many driving decisions cannot be made from the current frame alone. Merging, yielding, unprotected turns, occluded pedestrians, and aggressive neighboring vehicles all require forecasting. However, future prediction also introduces the risk of compounding error: a visually plausible future may not be physically or behaviorally useful for planning.

Reflection and causality represent another emerging target. Instead of producing a single reasoning chain, reflective methods revise, critique, or counterfactually evaluate possible decisions (Yuan et al., 2025; Li et al., 2025c; Zhang et al., 2025c; Peng et al., 2025d; Luo et al., 2026a). This changes CoT from a static explanation into an error-correction process. In safety-critical driving, this is especially valuable because the first plan may be unsafe, incomplete, or inconsistent with traffic context. The limitation is cost: reflection usually requires extra inference steps, additional verification, or multiple candidate plans. As a result, reflective reasoning is more suitable for difficult or uncertain scenarios than for every frame.

The least developed cognitive targets are rules, norms, risk, and social negotiation. Yet these are precisely the areas where foundation models

could provide the most value beyond traditional perception and planning pipelines. Rule-grounded methods such as PKRD-CoT, DiLu, KoMA, and related agent-based approaches attempt to inject traffic laws, right-of-way reasoning, human preference, or multi-agent interaction into the decision process (Luo et al., 2024; Wen et al., 2024; Jiang et al., 2024; Zheng et al., 2026; Mao et al., 2024). Cooperative methods further extend reasoning across vehicles or infrastructure through message exchange (Hu et al., 2024; Chiu et al., 2025; Liu et al., 2025a; Gao et al., 2025a; Hou et al., 2025). However, these topics remain underrepresented compared with perception and planning. This imbalance suggests that current driving CoT still spends much of its reasoning budget on tasks that overlap with conventional modules, while the harder cognitive problems of norm interpretation, negotiation, and risk-aware decision-making remain relatively open.

The cognitive shift therefore reveals a deeper issue. The field is moving from descriptive reasoning toward action-oriented reasoning, but not all cognitively important targets have advanced equally. Future driving CoT should place more emphasis on the problems that classical pipelines handle poorly: ambiguous right-of-way, social interaction, rare traffic rules, causal failure analysis, and uncertainty-aware risk reasoning.

A.3 The Action Shift: From Talking to Doing

The most visible evolution in driving CoT is the shift from textual explanation to trajectory-oriented output. Early methods often produced scene descriptions, driving advice, control commands, or high-level maneuvers (Xu et al., 2024; Sha et al., 2025; Ma et al., 2024). These outputs were useful for demonstrating that language models could reason about driving scenes, but they did not always close the loop to executable behavior. More recent methods increasingly output waypoints, trajectories, or planner-compatible action representations (Tian et al., 2025b; Wang et al., 2025f; Zhou et al., 2026b; Zhang et al., 2025a; Ye et al., 2026). This marks an important change: CoT is no longer treated merely as a text-generation mechanism for explanation, but as an intermediate process that should support driving behavior.

However, outputting a trajectory does not automatically mean that the reasoning trace faithfully caused the action. The corpus reveals several forms of weak coupling between reasoning and behavior. The first is parallel or decoupled generation. In some architectures, the model produces a reasoning text and a trajectory from shared visual features, but the trajectory does not explicitly depend on the generated reasoning. In such cases, the text may explain the action after the fact rather than determine it. This creates the risk of post-hoc rationalization, where the reasoning appears coherent but is not causally linked to the final driving decision.

The second weak-coupling pattern is training-only reasoning. Methods such as VERDI, DSDrive, and other distillation-based approaches use explicit textual reasoning during training but compress or remove it at inference time for efficiency (Feng et al., 2025; Liu et al., 2025e; Qiao et al., 2025). This design is practically motivated: verbose language reasoning is too slow for real-time driving. Yet it raises a critical question. If reasoning is internalized into hidden states, how can we verify that the deployed model preserves the causal and semantic benefits of the original reasoning traces? A faster model may inherit the performance gains of CoT supervision without preserving the faithfulness of the reasoning process.

The third pattern is QA-only reasoning. Driving QA and explanation benchmarks are valuable because they test whether models can answer questions, describe scenes, and explain decisions (Sima et al., 2024; Wang et al., 2025c; Wei et al., 2025a; Zhang et al., 2025d; Zeng et al., 2025). However, QA performance does not guarantee planning performance. A model can correctly answer a question about a scene and still fail to generate a safe trajectory. Conversely, a planner may produce a safe trajectory while giving an incomplete or misleading explanation. This creates a reasoning-planning disconnect that has been observed in recent analysis work (Song et al., 2025).

The fourth pattern is reward-driven decoupling. With the rise of reinforcement fine-tuning (RFT), models can learn to exploit the reward function rather than achieving the true objective — a well-known reward-hacking failure mode amplified by the strong optimization capability of VLMs and the misalignment between the true goal and the reward proxy. In driving, this typically manifests as a decoupling between the CoT trace and the predicted vehicle behavior. Because directly mapping visual inputs to trajectories is more straightforward than predicting a CoT and then conditioning the trajectory on it, the model tends to take the shortcut and ignore its own reasoning when only the trajectory is

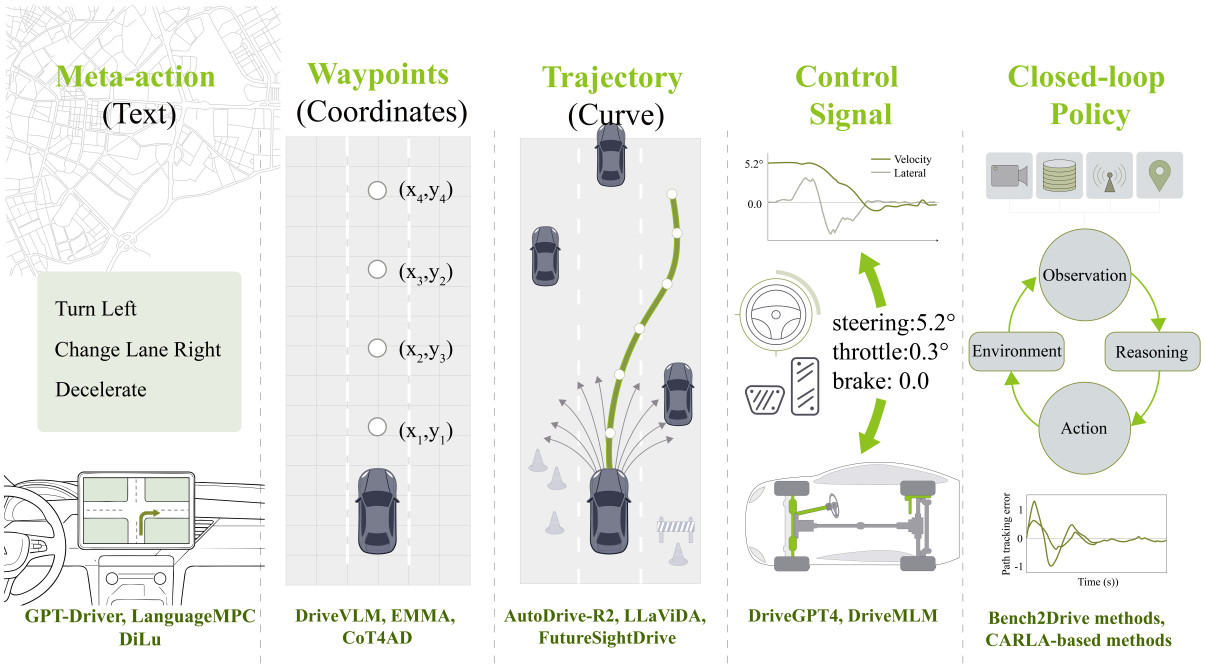


Figure 4: Five forms of action output in driving CoT methods. From left to right: meta-action as high-level text commands, discrete waypoint coordinates in bird's-eye view, continuous trajectory curves, low-level control signals, and closed-loop policies with observation-action feedback. Representative methods are listed below each type.

rewarded. Symmetrically, when the textual CoT is rewarded independently, the model can satisfy the linguistic reward through format manipulation or hallucinated content. Without a consistency penalty linking the two heads, the model's articulated intent, claiming to yield to a vulnerable road user for instance, can directly contradict the trajectory it actually executes.

This action shift therefore exposes the central challenge of driving CoT: faithfulness. A useful reasoning representation should not merely accompany an action; it should constrain, guide, or verify the action. Future evaluation should include causal intervention tests. For example, if a critical object is removed from the reasoning trace, the trajectory should change in a predictable way. If irrelevant text is perturbed, the action should remain stable. If a retrieved rule contradicts the scene, the system should either reject it or explain why it is inapplicable. Such tests are necessary because conventional open-loop metrics cannot tell whether the intermediate reasoning actually matters.

The field has made real progress from talking about driving to producing driving actions. But the harder transition is from action output to action-faithful reasoning. A method that outputs waypoints on nuScenes or NAVSIM is still not fully action-grounded if the reasoning trace can be removed, corrupted, or replaced without changing the trajectory. The next stage of driving CoT should therefore move beyond trajectory generation alone and evaluate whether reasoning is causally coupled to closed-loop behavior.

A.4 The Deployment Shift: The Cost of Thinking

Reasoning is useful only if it can be deployed under realistic constraints. Autonomous driving requires low latency, temporal consistency, robustness to distribution shift, and safe fallback behavior. These requirements change how CoT should be designed. In general language tasks, longer reasoning may improve accuracy. In driving, however, longer reasoning can be dangerous if it delays action, accumulates errors, or produces unstable decisions across time.

Different reasoning representations face different deployment bottlenecks. Linguistic CoT is highly interpretable but slow because autoregressive text generation introduces substantial latency (Tian et al., 2025b; Wang et al., 2024a; Yuan et al., 2025). It also struggles with fine-grained spatial grounding, since language is a lossy medium for metric geometry and temporal motion (Tian et al., 2025a; Godbole et al., 2025). Externalized CoT, including retrieval, tool use, and multi-agent

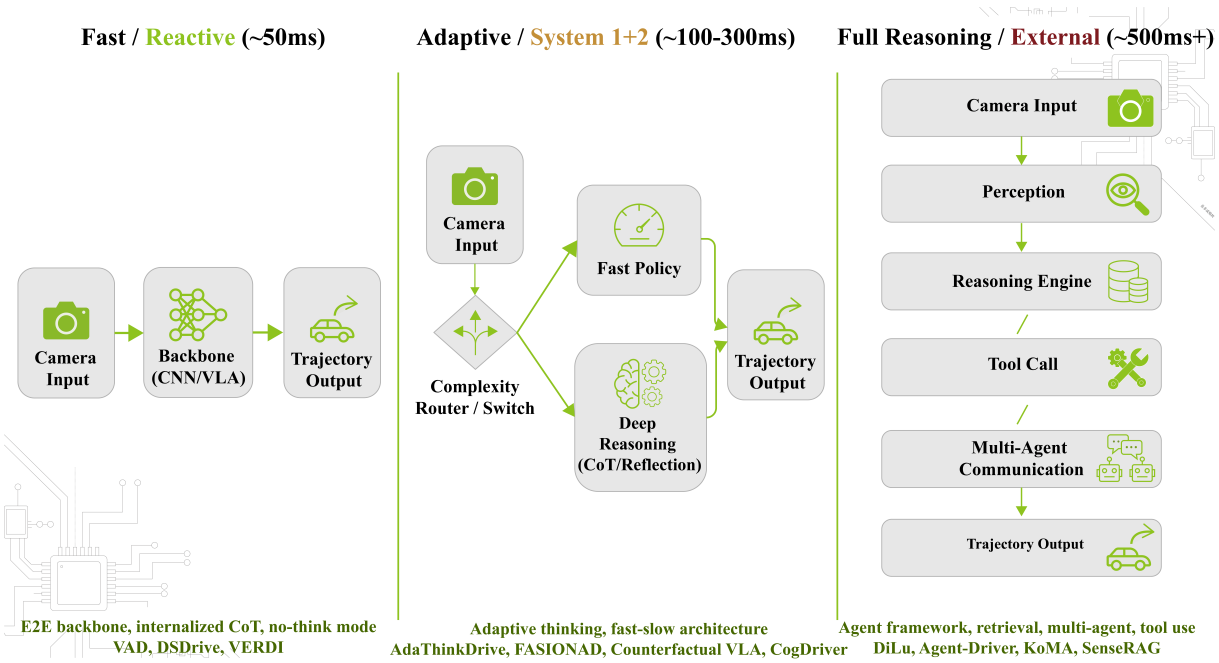


Figure 5: Three deployment strategies for driving CoT, illustrating the trade-off between reasoning depth and inference latency. Left: fast reactive policies that internalize reasoning into a single forward pass (∼50 ms). Middle: adaptive systems that route simple scenes to a fast path and complex scenes to a deliberative reasoning path (∼100–300 ms). Right: full external reasoning pipelines involving memory, tool calls, and multi-agent communication (∼500 ms+).

communication, can improve long-tail knowledge and modularity, but it introduces additional delays and interface failures (Luo et al., 2025a; Chang et al., 2026; Qian et al., 2025a; Chiu et al., 2025). Retrieved evidence may be irrelevant, a tool may be called in the wrong context, or a communication message may be delayed or inconsistent.

Visual-spatial reasoning has a different bottleneck. Future frames, occupancy maps, masks, crops, and trajectory sketches can ground reasoning in scene evidence, but they may suffer from error accumulation and incomplete evidence selection (Wang et al., 2024b; Yang et al., 2025; Zeng et al., 2026; Corbiere et al., 2025; Li et al., 2025b). A generated future may look realistic while failing to preserve the interaction variables needed for safe planning. A selected crop may focus on a visible vehicle while missing an occluded pedestrian. Thus, visual reasoning improves grounding but does not automatically guarantee completeness or safety.

Latent-dynamic reasoning is often more compatible with planning and control, but its deployment challenge is verification. Latent rollouts, dynamics tokens, and optimization states are difficult to inspect directly (Zheng et al., 2025b; Shang et al., 2026; Yang et al., 2026; Gao et al., 2025b). This creates a blind spot: latent reasoning may be action-effective but hard to certify. If a latent rollout leads to an unsafe decision, it is difficult to determine whether the failure came from perception, dynamics modeling, reward design, or trajectory optimization. For production systems, this lack of interpretability cannot be ignored simply because the representation is internal.

The rise of reinforcement learning introduces another deployment challenge: reward design. Recent methods use GRPO, DPO, or reward-driven optimization to align reasoning and planning (Li et al., 2026b; Zheng et al., 2025a; Zhang et al., 2025f; Fu et al., 2025b; Ye et al., 2026). Trajectory-level objectives can reward collision avoidance, route progress, comfort, and displacement error. But reasoning quality is harder to reward. A reasoning trace should be grounded, concise, faithful, risk-aware, and causally connected to action. No widely accepted reward captures all of these properties. As a result, RL may improve planning metrics while leaving the reasoning process under-constrained.

These bottlenecks motivate adaptive reasoning. Instead of running full CoT at every frame, several recent methods activate deeper reasoning only when the scene is difficult, uncertain, or risky (Luo

et al., 2025c; Qian et al., 2024; Peng et al., 2025d; Xie et al., 2025). Fast-slow architectures use lightweight reactive policies for routine driving and slower deliberative modules for complex scenarios. Counterfactual or reflective systems trigger additional reasoning when the model is uncertain or when candidate plans conflict. Hierarchical systems use high-level language reasoning at low frequency and low-level control at high frequency (Zhang et al., 2025e). This design better matches the structure of driving: most frames do not require deep reasoning, but rare frames may require substantial deliberation.

The deployment shift therefore changes the central design question. The goal is not to build a vehicle that always reasons more. The goal is to build a vehicle that reasons at the right time, with the right representation, at the right computational cost. In routine scenarios, compact latent or reactive policies may be sufficient. In ambiguous intersections, occluded hazards, rule conflicts, or multi-agent negotiation, deeper language, retrieval, simulation, or tool-mediated reasoning may be necessary. Future driving CoT should therefore be evaluated not only by reasoning accuracy, but also by reasoning efficiency, trigger reliability, fallback safety, and closed-loop stability.

Summary. The four shifts described above show that driving CoT is undergoing a transition from language-prompted explanation to action-grounded, trainable, and deployment-aware reasoning. The infrastructure shift enables model adaptation; the cognitive shift expands what counts as reasoning; the action shift demands causal coupling between intermediate representations and behavior; and the deployment shift forces reasoning to become adaptive and cost-aware. Together, these trends support the central claim of this survey: the future of CoT in autonomous driving is not longer textual chains, but intermediate representations that can be grounded, executed, verified, and trusted under real driving constraints.

B Full Comparison Table

Table 4 lists the papers in our corpus and groups the 130 method papers according to their primary reasoning representation. The remaining benchmark, dataset, survey, and analysis papers are listed separately when applicable. Backbone is reported at the model-family level; "–" indicates that the specific model was not identifiable from publicly available sources. Because many systems combine multiple representations, the category assigned in the table reflects the representation that plays the most central role in the reasoning-action pipeline. For hybrid systems, the Subtype column lists the primary label with a secondary tag in parentheses, where (K) denotes a knowledge-grounded and (R) a retrieval-augmented secondary role.

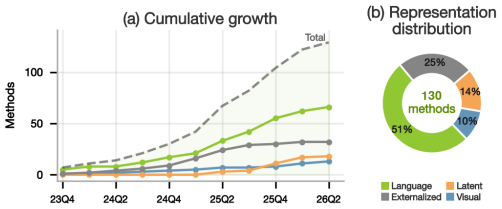


Figure 6: Temporal growth and representation distribution of the 130 method papers. Left: cumulative number of methods by primary reasoning representation from 2023Q4 to 2026Q2. Right: distribution over the four categories.

C Benchmark Landscape

This appendix expands the compact discussion in Section 5. The benchmark literature does not evaluate a single unified notion of "reasoning." Instead, different datasets test different parts of the reasoning-action pipeline: language explanations, spatial grounding, robustness, trajectory planning, and closed-loop behavior.

C.1 Evaluation Focus

Reasoning and QA benchmarks. DriveLM (Sima et al., 2024), Reason2Drive (Nie et al., 2024), DriveLMM-o1 (Ishaq et al., 2025), AD^2-Bench (Wei et al., 2025b), DriveQA (Wei et al., 2025a), WOMD-Reasoning (Li et al., 2025d), DriveCombo (Ma et al., 2026), and AgentDrive (Ferrag et al., 2026) evaluate whether models can describe driving scenes, answer structured questions, follow traffic rules, or produce step-wise explanations. These benchmarks are useful for language-based representations, but most remain open-loop and do not directly test whether a reasoning trace changes the final trajectory.

Spatial grounding and robustness benchmarks. NuScenes-SpatialQA (Tian et al., 2025a), VLAD-Bench (Li et al., 2025f), DVBench (Zeng et al., 2025), DRAMA-X (Godbole et al., 2025), and RoboDriveVLM (Liao et al., 2025a) focus on object grounding, spatial relations, intent prediction,

safety-critical perception, and robustness under corruption. They are especially important for visual-spatial reasoning because they test whether intermediate representations are grounded in the correct objects, locations, and relations before action is produced.

Planning and action benchmarks. nuScenes (Caesar et al., 2020), nuPlan (Caesar et al., 2021), NAVSIM (Dauner et al., 2024), CoVLA (Arai et al., 2025), DriveAction (Hao et al., 2025), doScenes (Martinez-Sanchez et al., 2026), and AutoDriDM (Tang et al., 2026) evaluate trajectory prediction, planning quality, action prediction, or instruction-conditioned driving outputs. These benchmarks connect reasoning representations to motion outputs, but they often report final planning metrics without evaluating whether the intermediate reasoning is faithful to the action.

Closed-loop benchmarks. Bench2Drive (Jia et al., 2024), Bench2ADVLM (Zhang et al., 2025d), and Bench2Drive-VL (Jia et al., 2026) place driving models in interactive simulation or real/sim closed-loop settings. They measure route completion, driving score, infractions, and safety outcomes, making them the closest benchmarks to deployment-oriented evaluation. However, even these benchmarks usually evaluate behavior rather than the correctness, stability, or faithfulness of the intermediate reasoning representation.

Coverage gaps. The current landscape is fragmented. QA benchmarks evaluate reasoning quality without action consequences; planning benchmarks evaluate actions without inspecting intermediate reasoning; grounding benchmarks expose perceptual errors without measuring downstream control impact; and closed-loop benchmarks rarely measure latency, retrieval reliability, tool failure, or multi-agent message consistency. The taxonomy in Section 3 defines four reasoning *artifacts*: language, visual-spatial, latent-dynamic, and externalized. While the evaluation literature defines several orthogonal *dimensions* along which any artifact can be tested: (i) grounding correctness, whether the artifact encodes the right scene content; (ii) action coupling, whether the artifact actually drives the produced action; (iii) reasoning cost, whether the artifact is affordable in real-time deployment; (iv) closed-loop reliability, whether the artifact remains stable when actions affect future observations; and (v) faithfulness under intervention, whether perturbing the artifact predictably perturbs the action. A useful way to read the current literature is as a matrix of *artifact* $\times$ *dimension* cells. Existing benchmarks cluster in a narrow band of this matrix: language artifacts probed for grounding correctness, or arbitrary artifacts scored by open-loop trajectory error. Latent-dynamic representations are barely tested for grounding or faithfulness, because their internal states are not directly readable; externalized representations are rarely tested for source reliability, delay, or conflict handling. Most cells in the matrix are empty, and this emptiness, rather than the absence of any single benchmark, is what makes cross-representation comparison difficult.

Implications. This fragmentation suggests that future benchmarks should move from *task-level scoring* to *representation-conditional evaluation*. A driving model should be judged not only by whether it answers correctly, plans safely, or completes a route, but also by whether its intermediate reasoning artifact is scenario-appropriate and causally useful for the final action. For language artifacts, this means evaluating object- and rule-level faithfulness; for visual-spatial artifacts, whether the selected region is causally relevant to the maneuver; for latent-dynamic artifacts, whether internal rollouts or state transitions can be probed or perturbed; and for externalized artifacts, whether source reliability, delay, and conflict handling are explicitly measured.

From coverage to causality. In our opinion, filling the matrix above requires two evaluation primitives that current protocols lack. The first is *intervention-based faithfulness*: perturbing or ablating a part of the reasoning artifact and checking whether the action changes in a scene-consistent way. The probe is naturally representation-specific (swap the object for language, replace the case for retrieval, intervene on the rolled-out state for latent rollouts, inject a known-wrong response for tools), and without it post-hoc rationalization is statistically indistinguishable from causal reasoning. The second is *stratified evaluation by scenario difficulty*: aggregate L2 or collision rates dilute the value of CoT, because reasoning's marginal benefit lies in long-tail cases, not in routine lane following. Reporting reasoning-conditional metrics by difficulty stratum, paired with cost measurements (latency, retrieval calls, tool invocations, communication rounds), turns the matrix from a coverage chart

into a causal one and points toward *representation-conditional protocols*, rather than yet another end-to-end driving suite. Two safeguards apply to such interventions: the perturbation must preserve artifact validity rather than create an implausible or out-of-distribution state, and the expected behavioral change is assessed at the artifact's abstraction level rather than by exact waypoint correspondence. These representation-specific probes are proposed operational diagnostics for evaluating the surveyed systems, not results claimed or run by this survey.

C.2 Metric Glossary

Table 3 summarizes the most common metric families in the evaluation landscape and their main limitations.

Metric	What it measures	Main limitation
Exact-match / multiple-choice accuracy	Agreement with a predefined answer	Does not establish grounding or planner use
L2 / ADE / FDE	Distance from the logged trajectory at selected or averaged future steps	Penalizes alternative safe futures; definitions and horizons vary
Open-loop collision rate	Predicted geometric overlap with obstacles	Protocol-dependent and does not capture interactive reactions
Closed-loop or composite score, including PDMS/EPDMS	Progress, compliance, comfort, and infractions under a benchmark protocol	Simulator- and version-specific; aggregate values may hide component failures

Table 3: Metric glossary for the evaluation landscape.

C.3 Benchmark Inventory

Table 6 lists the benchmark, dataset, and analysis papers used in this survey. Compared with Table 5, which groups benchmarks by evaluation focus, this inventory serves as a fine-grained appendix-level reference. It records the individual benchmark purpose, task setting, main output format, and reported metrics whenever they are identifiable from the paper. Since benchmarks with similar high-level goals may still evaluate different parts of the reasoning-action pipeline, the inventory helps readers locate whether each benchmark supports language rationales, spatial grounding, trajectory quality, robustness, or closed-loop safety.

Table 4: Full comparison of method papers, grouped by primary reasoning representation.

Year	Paper	Subtype	Backbone	Training	Evaluation
I. Language-based CoT					
2024	Making Large Language Models Better Planners with Reasoning-Decision Alignment (Huang et al., 2024)	Compressed	LLaMA	SFT	nuScenes, DriveLM
2025	AlphaDrive (Jiang et al., 2025b)	Compressed	Qwen	RL+SFT	MetaAD
2025	LightEMMA (Qiao et al., 2025)	Compressed	Qwen+LLaMA	Prompt	nuScenes
2025	DSDrive (Liu et al., 2025e)	Compressed	Qwen+LLaMA	Distill	CARLA
2025	VERDI (Feng et al., 2025)	Compressed	Qwen	Distill	nuScenes,HugSim
2025	DriveMind (Wasif et al., 2026)	Compressed	CLIP-based	RL+Distill	CARLA
2025	AutoVLA (Zhou et al., 2026b)	Compressed	Qwen	RL+SFT	nuScenes, NAVSIM, Waymo E2E, Bench2Drive
2025	Drive-R1 (Li et al., 2026b)	Compressed	InternVL	RL+SFT	nuScenes, DriveLM
2025	DriveAgent-R1 (Zheng et al., 2025a)	Compressed	Qwen	RL+SFT	nuScenes
2025	AdaThinkDrive (Luo et al., 2025c)	Compressed	InternVL	RL+SFT	NAVSIM
2025	Alpamayo-R1 (AR1) (Wang et al., 2025d)	Compressed	Cosmos-Reason	RL+SFT	open-loop + closed-loop
2025	Reasoning-VLA (Zhang et al., 2025a)	Compressed	Qwen	RL+SFT	nuScenes, NAVSIM
2025	COVLM-RL (Li et al., 2025a)	Compressed	InternVL	RL+Prompt	CARLA
2025	MindDrive (Fu et al., 2025b)	Compressed	Qwen	RL+SFT	Bench2Drive
2025	OmniDrive-R1 (Zhang et al., 2025f)	Compressed	Qwen	RL	DriveLMM-o1
2026	ThinkDrive (Zhao et al., 2026)	Compressed	Qwen	RL+SFT	DrivingVQA
2026	Bridging Large-Model Reasoning (Chen et al., 2026)	Compressed	InternVL+DeepSeek	SFT+Prompt	CARLA
2026	AutoDrive-P3 (Ye et al., 2026)	Compressed	Qwen	RL+SFT	nuScenes, NAVSIM
2023	GPT-Driver (Mao et al., 2023)	Descriptive	GPT	SFT	nuScenes
2023	DriveGPT4 (Xu et al., 2024)	Descriptive	LLaMA	SFT	BDD-X
2023	LanguageMPC (Sha et al., 2025)	Descriptive	GPT	Prompt	IDSim
2023	Dolphins (Ma et al., 2024)	Descriptive	OpenFlamingo	instruction tuning	BDD-X
2024	SimpleLLM4AD (Zheng et al., 2024)	Descriptive	InternVL	SFT	DriveLM-nuScenes
2024	EMMA (Hwang et al., 2024)	Descriptive	Gemini	SFT	nuScenes, Waymo
2024	CALMM-Drive (Yao et al., 2024)	Descriptive	GPT	Prompt	nuPlan
2024	WiseAD (Zhang et al., 2024)	Descriptive	LLaMA	SFT	CARLA
2024	OpenEMMA (Xing et al., 2025)	Descriptive	Qwen+LLaMA	Prompt	nuScenes
2025	ORION (Fu et al., 2025a)	Descriptive	Vicuna	SFT	Bench2Drive
2025	ReasonDrive (Chahe and Zhou, 2025)	Descriptive	Qwen+LLaMA	SFT+Distill	DriveLM
2025	X-Driver (Liu et al., 2025d)	Descriptive	LLaVA	SFT	CARLA,Bench2Drive
2025	Impromptu VLA (Chi et al., 2026)	Descriptive	Qwen	SFT+Distill	nuScenes,NeuroNCAP
2025	Multimodal Framework for Explainable Autonomous Driving (Zarghani et al., 2025)	Descriptive	VideoMAE	SFT	nuScenes, BDD-X
2025	ReAL-AD (Lu et al., 2025)	Descriptive	LLaMA+Qwen	SFT	nuScenes, Bench2Drive
2025	OmniReason (Liu et al., 2025c)	Descriptive	LLaVA	SFT+Distill	nuScenes,Bench2Drive
2025	The System Description of CPS (Peng et al., 2025a)	Descriptive	LLaVA	SFT	DriveLM-nuScenes
2025	Robust Driving QA through Meta (Yu et al., 2025)	Descriptive	Qwen	Prompt	driving QA benchmark
2025	Enhancing Vision-Language Mode (Wu and Luo, 2025)	Descriptive	Qwen	Prompt	RoboSense Challenge at IROS 2025
2025	CoC-VLA (Zhang et al., 2026a)	Descriptive	LLaVA	SFT	nuScenes + CARLA
2025	E3AD (Hu et al., 2025)	Descriptive	Qwen	SFT	Talk2Car, DrivePilot, MoCAD,Talk2Car-Trajectory
2026	Listen, Look, Drive (Guo et al., 2026)	Descriptive	Qwen	SFT	nuScenes
2026	SteerVLA (Gao et al., 2026)	Descriptive	InternVL	SFT	Bench2Drive
2026	Deliberation Meets Reaction (Xie et al., 2025)	Descriptive	InternVL	SFT	Bench2Drive
2023	DriveMLM (Wang et al., 2023a)	Procedural	LLaMA	SFT	CARLA
2024	LLM-Assist (Sharan et al., 2023)	Procedural	LLaMA+GPT	Prompt	nuPlan

Continued on next page

Table 4: Full comparison of method papers, grouped by primary reasoning representation.

Year	Paper	Subtype	Backbone	Training	Evaluation
2024	DriveVLM (Tian et al., 2025b)	Procedural	Qwen	SFT	nuScenes
2024	DriveCoT (Wang et al., 2024a)	Procedural	VideoSwin-Transformer	SFT + Distill	CARLA
2024	LC-LLM (Peng et al., 2025b)	Procedural	LLaMA	SFT	highD
2024	DualAD (Wang et al., 2025b)	Procedural	GLM+GPT	Prompt	nuPlan
2024	PKRD-CoT (Luo et al., 2024)	Procedural	Qwen+GPT	Prompt	nuScenes
2025	Sce2DriveX (Zhao et al., 2025)	Procedural	Vicuna	SFT	nuScenes, Bench2Drive
2025	CoT-Drive (Liao et al., 2025b)	Procedural	GPT	Distill	NGSIM, HighD, MoCAD, ApolloScape,nuScenes
2025	PRIMEDrive-CoT (Mandalika et al., 2025)	Procedural	Bayesian Graph Neural Network	SFT	DriveCoT
2025	DriveRX (Diao et al., 2025)	Procedural	DeepSeek	RL	DriveBench, DriveLM-Hard
2025	HMVLM (Wang et al., 2025a)	Procedural	Qwen-VL	SFT	Waymo
2025	ReasonPlan (Liu et al., 2025f)	Procedural	Qwen	SFT+Self-sup	Bench2Drive
2025	CoT4AD (Wang et al., 2025f)	Procedural	LLaMA	SFT	nuScenes, Bench2Drive
2025	LLaViDA (Liu et al., 2025g)	Procedural	LLaVA	SFT	nuScenes
2026	AppleVLM (Han et al., 2026)	Procedural	Janus Pro	SFT	CARLA
2026	NaviDriveVLM (Tao et al., 2026)	Procedural	Qwen-VL	SFT	nuScenes
2026	RAD-LAD (Ghosh et al., 2026)	Procedural	Qwen	3-stage curriculum	nuPlan
2026	Accelerating Structured Chain-of-Thought in Autonomous Vehicles (Gu et al., 2026)	Procedural	Qwen	SFT	Internal
2024	Hybrid Reasoning Based on Large Language Models for Autonomous Car Driving (Azarafza et al., 2024)	Procedural (K)	GPT	Prompt	CARLA
2025	AutoDrive-R^2 (Yuan et al., 2025)	Reflective	Qwen	RL+SFT	nuScenes, NAVSIM, Waymo
2025	Discrete Diffusion for Reflective Vision-Language-Action Models in Autonomous Driving (Li et al., 2025c)	Reflective	LLaDA-V	SFT	NAVSIM
2025	OpenREAD (Zhang et al., 2025c)	Reflective	Qwen	RL+SFT	nuScenes
2025	Counterfactual VLA (Peng et al., 2025d)	Reflective	Qwen	SFT	Internal
2026	Unleashing VLA Potentials in Autonomous Driving via Explicit Learning from Failures (Luo et al., 2026a)	Reflective	Qwen+InternVl	RL+SFT	NAVSIM
II. Visual CoT					
2025	Retrieval-Based Interleaved Visual Chain-of-Thought in Real-World Driving Scenarios (Corbiere et al., 2025)	Action-grounded	Qwen+LLaVA	SFT	DrivingVQA
2025	SOLVE (Chen et al., 2025a)	Action-grounded	LLaMA	SFT	nuScenes
2025	SpaceDrive (Li et al., 2025b)	Action-grounded	Qwen	SFT	nuScenes, Bench2Drive
2026	A Vision-Language-Action Model with Visual Prompt for OFF-Road Autonomous Driving (Zhang et al., 2026b)	Action-grounded	Qwen	Prompt	RELLIS-3D (off-road).
2026	MindDriver (Zhang et al., 2026c)	Action-grounded	Qwen	RL+SFT	nuScenes, CARLA, Bench2Drive
2023	Driving into the Future (Wang et al., 2024b)	Predictive	World/Diff	Diffusion	nuScenes
2024	Vista (Gao et al., 2024)	Predictive	World/Diff	Diffusion	nuScenes, OpenDV

Continued on next page

Table 4: Full comparison of method papers, grouped by primary reasoning representation.

Year	Paper	Subtype	Backbone	Training	Evaluation
2024	Driving in the Occupancy World (Yang et al., 2025)	Predictive	BEVFormer-style	SFT	nuScenes
2024	DrivingGPT (Chen et al., 2025b)	Predictive	LLaMA	SFT	NAVSIM, nuPlan
2025	FutureSightDrive (Zeng et al., 2026)	Predictive	Qwen	unified pretraining	nuScenes, NAVSIM, DriveLM
2026	UniDrive-WM (Xiong et al., 2026)	Predictive	Vicuna	SFT	nuScenes, Bench2Drive
2026	DriveDreamer-Policy (Zhou et al., 2026a)	Predictive	Qwen	SFT	NAVSIM
2026	Learning Vision-Language-Action World Models for Autonomous Driving (Wang et al., 2026a)	Predictive	Qwen	RL+SFT	nuScenes
III. Latent / Simulative CoT					
2025	DiffVLA (Jiang et al., 2025a)	Optimization-based	LLaMA	SFT	NAVSIM
2025	ReCogDrive (Li et al., 2025e)	Optimization-based	Qwen + InternVL	GRPO + SFT	NAVSIM, CARLA, Bench2Drive
2025	DiffVLA++ (Gao et al., 2025b)	Optimization-based	LLaMA	SFT	NAVSIM
2025	dVLM-AD (Ma et al., 2025)	Optimization-based	SigLIP-based	SFT	nuScenes, Waymo
2025	WorldRFT (Yang et al., 2026)	Optimization-based	E2E CNN/Trans	RL+SFT	nuScenes, NAVSIM
2025	ColaVLA (Peng et al., 2025c)	Optimization-based	LLaMA	SFT	nuScenes
2025	UniUGP (ByteDance Seed, 2025)	Rollout-based	Qwen	SFT	nuScenes, Waymo, DriveLM
2025	World4Drive (Zheng et al., 2025b)	Rollout-based	E2E CNN/Trans	SFT	nuScenes, NAVSIM
2025	FutureX (Lin et al., 2025)	Rollout-based	model-agnostic	SFT	NAVSIM, CARLA
2025	Latent Chain-of-Thought World Modeling for End-to-End Driving (Tan et al., 2025)	Rollout-based	Qwen	RL+SFT	PhysicalAI-AV
2025	Think Before You Drive (Liao et al., 2025c)	Rollout-based	BERT	SFT	Talk2Car, MoCAD, DrivePilot, RefCOCO/+/g
2026	SGDrive (Li et al., 2026a)	Rollout-based	InternVL	RL+SFT	NAVSIM
2026	DriveWorld-VLA (Liu et al., 2026)	Rollout-based	InternVL	SFT	nuScenes, NAVSIM
2026	LaST-VLA (Luo et al., 2026b)	Rollout-based	InternVL	RL+SFT	NAVSIM
2026	R&B-EnCoRe (Ganai et al., 2026)	Tokenized	Qwen+LLaMA	2-stage variational	LIBERO-90, nuScenes
2026	HiST-VLA (Wang et al., 2026b)	Tokenized	LLaMA	SFT	NAVSIM
2026	DynVLA (Shang et al., 2026)	Tokenized	EMU3	RL+SFT	NAVSIM, CARLA, Bench2Drive
2026	OneVL (Lu et al., 2026)	Tokenized	Qwen	SFT	NAVSIM
IV. Externalized / Distributed CoT					
2024	AgentsCoDriver (Hu et al., 2024)	Cooperative	GPT	Prompt	highway-env multi-vehicle
2025	V2V-LLM (Chiu et al., 2025)	Cooperative	LLaVA	SFT	V2V-QA
2025	CoLMDriver (Liu et al., 2025a)	Cooperative	InternVL	SFT+Prompt	CARLA
2025	LangCoop (Gao et al., 2025a)	Cooperative	Qwen+LLaMA	Prompt	CARLA
2025	DriveAgent (Hou et al., 2025)	Cooperative	LLaMA	SFT	Internal
2023	Receive, Reason, and React (Cui et al., 2024)	Knowledge-grounded	GPT	Prompt	HighwayEnv
2024	PlanAgent (Zheng et al., 2026)	Knowledge-grounded	GPT	Prompt	nuPlan
2024	KoMA (Jiang et al., 2024)	Knowledge-grounded	GPT	Prompt	HighwayEnv
2024	FASIONAD (Qian et al., 2024)	Knowledge-grounded	Qwen	SFT	nuScenes, CARLA

Continued on next page

Table 4: Full comparison of method papers, grouped by primary reasoning representation.

Year	Paper	Subtype	Backbone	Training	Evaluation
2024	SafeDrive (Zhou et al., 2026c)	Knowledge-grounded	GPT	Prompt	HighD, InD
2025	TeLL-Drive (Xu et al., 2025a)	Knowledge-grounded	GPT	RL+SFT	HighwayEnv
2025	FASIONAD++ (Qian et al., 2025b)	Knowledge-grounded	Qwen	SFT	nuScenes, CARLA, Bench2Drive
2025	Interact, Instruct to Improve (Fang et al., 2025)	Knowledge-grounded	LLaMA	Prompt	simulated heterogeneous HVs + real world
2025	Towards Human-Centric Autonomous Driving (Xu et al., 2025b)	Knowledge-grounded	GPT	RL+Prompt	HighwayEnv
2025	LeAD (Zhang et al., 2025e)	Knowledge-grounded	Qwen	Prompt	CARLA
2025	VLM-UDMC (Liu et al., 2025b)	Knowledge-grounded	GPT	SFT+Prompt	simulation + real-world
2025	MTRDrive (Luo et al., 2025d)	Knowledge-grounded	Qwen	RL+SFT	NAVSIM
2025	KEPT (Wang et al., 2026c)	Knowledge-grounded	Qwen	SFT	nuScenes
2025	AdaDrive (Zhang et al., 2025b)	Knowledge-grounded	LLaMA	SFT	LangAuto
2026	KLDrive (Tian et al., 2026)	Knowledge-grounded	Qwen	Prompt	nuScenes
2024	Driving with Regulation (Cai et al., 2026)	Retrieval-augmented	GPT	Prompt	DriveReg
2024	DiLu (Wen et al., 2024)	Retrieval-augmented (K)	GPT	Prompt	HighwayEnv
2024	A Language Agent for Autonomous Driving (Mao et al., 2024)	Retrieval-augmented (K)	GPT	SFT	nuScenes
2025	SenseRAG (Luo et al., 2025a)	Retrieval-augmented	LLaMA	Prompt	real-world V2X
2025	Traffic Regulation-aware Path (Han et al., 2025)	Retrieval-augmented	LLaVA	Prompt	simulation + real-world
2025	RAD (Wang et al., 2025e)	Retrieval-augmented	Qwen	SFT	nuScenes
2025	Driving-RAG (Chang et al., 2026)	Retrieval-augmented	GPT	Prompt	CitySim
2025	V2X-UniPool (Luo et al., 2025b)	Retrieval-augmented	GPT	Prompt	DAIR-V2X (real-world V2X)
2025	Case-based Reasoning Augmented (Gan et al., 2025)	Retrieval-augmented	LLaMA	Prompt	Internal
2025	The Case for Negative Data (Patrikar et al., 2025)	Retrieval-augmented	GPT+Mistral	Prompt	nuScenes
2025	AgentThink (Qian et al., 2025a)	Tool-mediated	Qwen	RL+SFT	DriveLMM-o1
2026	From Prompts to Pavement (Goba et al., 2025)	Tool-mediated	GPT	Prompt	CARLA

Table 5: Benchmark landscape for action-grounded reasoning in autonomous driving.

Focus	Representative benchmarks	Setting	Main output	Primary metrics	Best used for
Reasoning / QA	DriveLM (Sima et al., 2024); Reason2Drive (Nie et al., 2024); DriveLMM-o1 (Ishaq et al., 2025); AD^2-Bench (Wei et al., 2025b); DriveQA (Wei et al., 2025a); Drive-Combo (Ma et al., 2026); AgentDrive (Ferrag et al., 2026)	Mostly open-loop QA	Answers, rationales, rule decisions, staged reasoning	Accuracy, reasoning score, per-step correctness, language metrics	Language-based reasoning and rule reasoning

Focus	Representative benchmarks	Setting	Main output	Primary metrics	Best used for
Spatial grounding	NuScenes-SpatialQA (Tian et al., 2025a); VLAD-Bench (Li et al., 2025f); DVBench (Zeng et al., 2025); DRAMA-X (Godbole et al., 2025)	Open-loop perception / QA	Object grounding, spatial relations, intent, risk labels	Spatial accuracy, detection, intent accuracy, MCQ accuracy	Visual-spatial grounding and perception-faithfulness checks
Robustness	RoboDriveVLM (Liao et al., 2025a); DVBench (Zeng et al., 2025); AD^2-Bench (Wei et al., 2025b)	Open-loop perturbation tests	Predictions or answers under corruption, weather, or safety-critical cases	Trajectory error, collision, robustness drop, adverse-condition accuracy	Failure modes under distribution shift
Planning / trajectory	nuScenes (Caesar et al., 2020); nuPlan (Caesar et al., 2021); NAVSIM (Dauner et al., 2024); CoVLA (Arai et al., 2025); DriveAction (Hao et al., 2025); doScenes (Martinez-Sanchez et al., 2026)	Open-loop planning and action prediction	Waypoints, trajectories, actions, instruction-conditioned plans	L2/ADE, collision, PDMS/EPDMS, action accuracy	Action coupling and trajectory quality
Closed-loop driving	Bench2Drive (Jia et al., 2024); Bench2ADVLM (Zhang et al., 2025d); Bench2Drive-VL (Jia et al., 2026); nuPlan (Caesar et al., 2021)	Closed-loop simulation or real/sim evaluation	Interactive driving behavior over routes or scenarios	Driving score, success rate, route completion, infractions	Closed-loop reliability and safety consequences

Table 6: Individual benchmark and dataset inventory.

Benchmark	Year	Focus	What it evaluates	Setting	Main metrics
Reason2Drive (Nie et al., 2024)	2023	Reasoning / QA	Chain-based driving reasoning across perception, prediction, and planning questions	Open-loop QA	Aggregated reasoning score; answer accuracy
Evaluation of Large Language Models (Tanahashi et al., 2023)	2023	Rule reasoning	Spatially aware decision-making and traffic-rule compliance with text-prompted LLMs	Open-loop decision tasks	Collision avoidance; rule-compliance accuracy
DriveLM (Sima et al., 2024)	2023	Reasoning / QA	Graph-structured visual QA over perception, prediction, planning, and behavior stages	Open-loop Graph-VQA	QA accuracy; BLEU/ROUGE/CIDEr; planning-oriented QA
OmniDrive (Wang et al., 2025c)	2024	Multimodal QA / planning	3D-aware vision-language alignment with DriveLM-style QA and planning outputs	Open-loop QA + nuScenes planning	QA accuracy; L2; collision
Bench2Drive (Jia et al., 2024)	2024	Closed-loop driving	Interactive closed-loop driving across disentangled CARLA scenarios	Closed-loop simulation	Driving score; route completion; infractions; success rate
CoVLA (Arai et al., 2025)	2024	VLA dataset	Paired vision-language-action data for trajectory and language-conditioned driving	Open-loop trajectory + language	Trajectory error; caption / language quality
DriveLMM-o1 (Ishaq et al., 2025)	2025	Step-wise reasoning	Step-wise visual reasoning faithfulness across perception, prediction, and planning	Open-loop VQA	Final-answer accuracy; reasoning-quality score
VLADBench (Li et al., 2025f)	2025	Fine-grained QA	Closed-form QA from traffic knowledge to scene and decision reasoning	Open-loop QA	Per-domain and per-aspect accuracy
NuScenes-SpatialQA (Tian et al., 2025a)	2025	Spatial grounding	Qualitative and quantitative spatial reasoning in driving scenes	Open-loop QA	Spatial accuracy; distance / relation correctness
DVBench (Zeng et al., 2025)	2025	Safety reasoning	VLLM perception and reasoning over safety-critical driving situations	Open-loop MCQ	Accuracy over hierarchical ability categories

Benchmark	Year	Focus	What it evaluates	Setting	Main metrics
WOMD-Reasoning (Li et al., 2025d)	2025	Interaction reasoning	Interaction, traffic-rule, and trajectory-intent reasoning over driving scenes	Open-loop QA + prediction	QA quality; interaction-prediction accuracy
DriveAction (Hao et al., 2025)	2025	Action coupling	Action-rooted vision-language-action coupling and modality necessity	Open-loop action prediction	Action accuracy; modality ablations
AD^2-Bench (Wei et al., 2025b)	2025	CoT evaluation	CoT reasoning quality of MLLMs under adverse weather and complex scenes	Open-loop CoT evaluation	Per-step CoT accuracy; end-to-end CoT correctness
Bench2ADVLM (Zhang et al., 2025d)	2025	Closed-loop ADVLM	Closed-loop behavior of autonomous-driving VLM agents in simulation and physical tests	Closed-loop sim + real	Driving score; safety score; stage-wise closed-loop metrics
DRAMA-X (Godbole et al., 2025)	2025	Grounding / intent	Object detection and directional intent prediction for risk reasoning	Open-loop multi-task evaluation	Detection mAP; intent accuracy; risk reasoning score
DriveQA (Wei et al., 2025a)	2025	Rule QA	Rule, sign, and right-of-way knowledge for LLMs and MLLMs	Open-loop QA	Per-category accuracy
Reasoning-Planning Disconnect (Song et al., 2025)	2025	Faithfulness analysis	Whether language reasoning causally improves downstream planning quality	Open-loop with nuPlan metrics	Planning scores under CoT removal / intervention
RoboDriveVLM (Liao et al., 2025a)	2025	Robustness	Robustness of VLM-based trajectory prediction under sensor and visual corruptions	Open-loop corruption tests	Trajectory error; collision under corruption
From Segments to Scenes (Cannons et al., 2025)	2025	Scenario dataset	Natural-language scene construction from driving segments	Dataset / benchmark	Scene-level coverage; segment-to-scene quality
AutoDriDM (Tang et al., 2026)	2026	Decision-centric QA	Perception-to-decision ability boundary across object, scene, and decision levels	Open-loop QA	Per-dimension accuracy; perception-decision consistency
AgentDrive (Ferrag et al., 2026)	2026	Agentic reasoning	Agentic LLM reasoning across physics, policy, hybrid, and safety dimensions	Open-loop MCQ + simulation	Per-dimension MCQ accuracy; simulation safety
doScenes (Martinez-Sanchez et al., 2026)	2026	Instruction planning	Natural-language instruction-conditioned trajectory planning quality	Open-loop trajectory prediction	ADE; instruction-ablation metrics
DriveCombo (Ma et al., 2026)	2026	Compositional rules	Compositional multi-rule traffic reasoning under conflicting rules	Open-loop QA	Per-level accuracy; rule-conflict resolution
Bench2Drive-VL (Jia et al., 2026)	2026	Closed-loop VLM4AD	Closed-loop VLM4AD behavior with auto-generated behavior questions	Closed-loop driving + QA	Driving score; planning / QA metrics

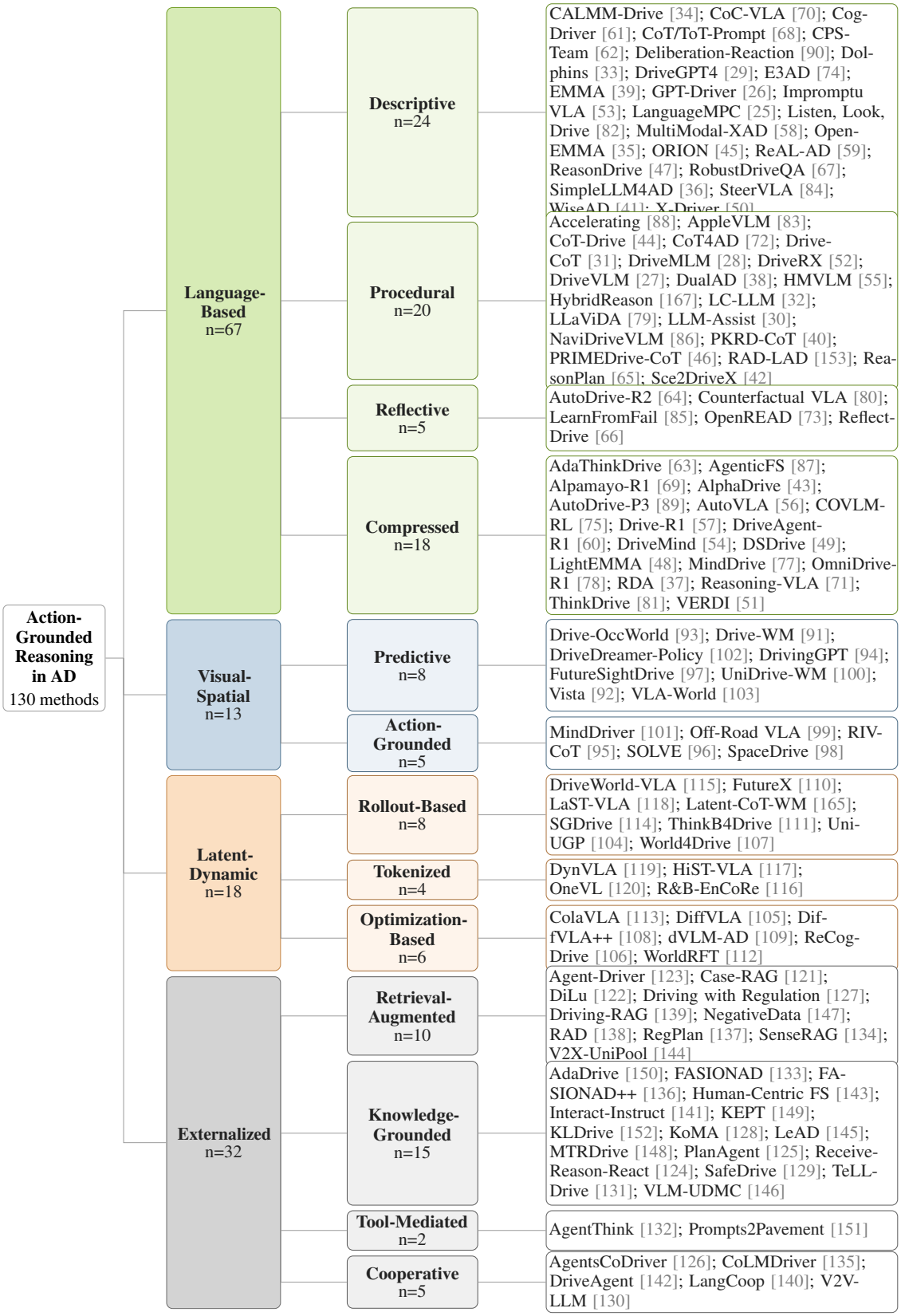


Figure 7: Full taxonomy of 130 method papers organized by primary intermediate reasoning representation. Numbers in brackets are reference indices.

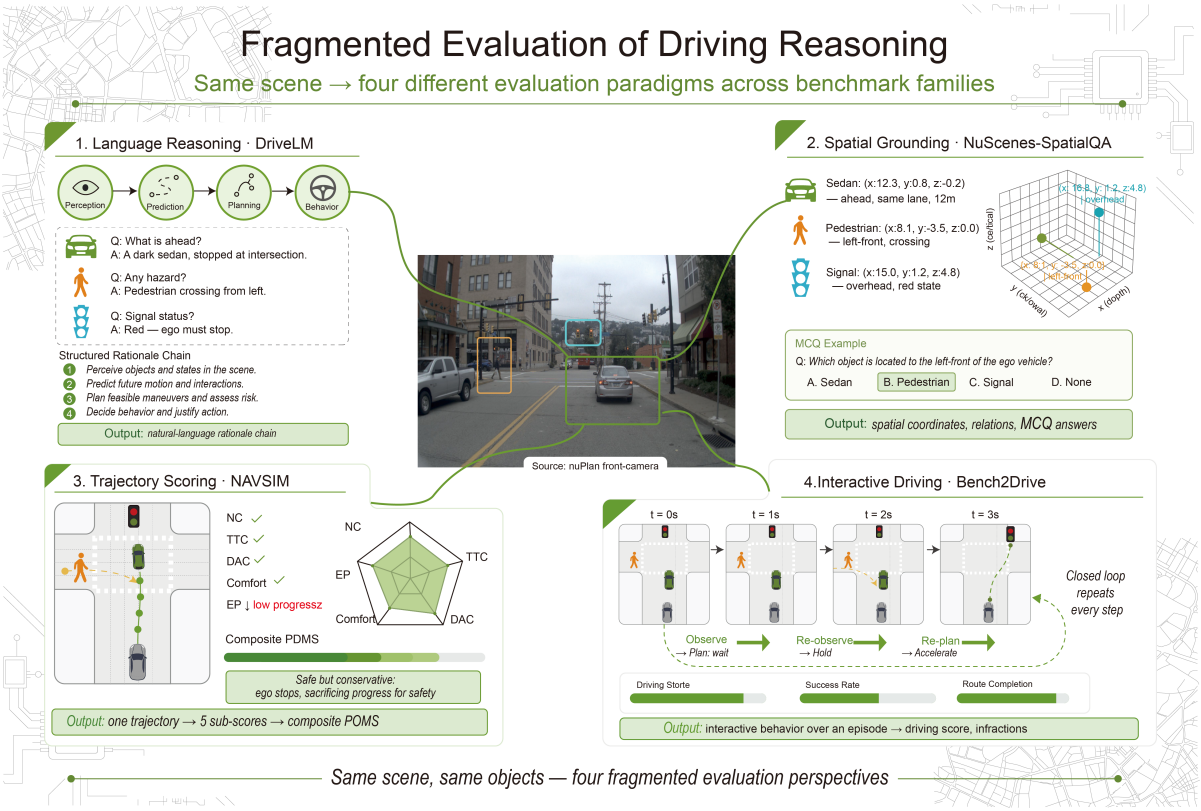


Figure 8: Fragmented evaluation of driving reasoning across benchmark families. A single nuPlan front-camera scene with three annotated objects (sedan, pedestrian, traffic light) is processed into four fundamentally different evaluation targets. **Top-left**: DriveLM converts objects into a Graph-VQA rationale chain across perception, prediction, planning, and behavior stages. **Top-right**: NuScenes-SpatialQA converts the same objects into 3D coordinates and pairwise spatial relation queries. **Bottom-left**: NAVSIM projects objects as geometric constraints in a BEV trajectory scoring framework with PDMS sub-score decomposition. **Bottom-right**: Bench2Drive instantiates objects as interactive agents in CARLA closed-loop simulation. Colored lines trace each object from the center image to its representation in each panel, illustrating how no single benchmark jointly evaluates reasoning quality, spatial grounding, trajectory faithfulness, and closed-loop safety.

D Reported Results by Benchmark

The following tables aggregate reported scores from the surveyed method papers. Metrics are grouped by benchmark family and metric type. Values within a metric group may be sorted by their reported direction, but different metric groups should not be compared as a single leaderboard.

nuScenes

Table 7: Reported results for nuScenes grouped by metric type.

Metric group	Method	Representation	Metric	Value	Protocol / Notes
Planning error					
Planning error	DriveAgent-R1 (Zheng et al., 2025a)	Compressed	ADE Avg	0.28	nuScenes; open-loop; validation
Planning error	LightEMMA-QWen (Qiao et al., 2025)	Compressed	ADE Avg	1.45	nuScenes; open-loop; test
Planning error	LightEMMA-LLaMa (Qiao et al., 2025)	Compressed	ADE Avg	1.53	nuScenes; open-loop; test
Planning error	LightEMMA-QWen (Qiao et al., 2025)	Compressed	FDE	2.90	nuScenes; open-loop; test; 3s
Planning error	LightEMMA-LLaMa (Qiao et al., 2025)	Compressed	FDE	3.03	nuScenes; open-loop; test; 3s
Planning error	AutoDrive-R^2-7B (Yuan et al., 2025)	Reflective	L2 Avg	0.19	nuScenes; open-loop; validation

Continued on next page

Metric group	Method	Representation	Metric	Value	Protocol / Notes
Planning error	Reasoning-VLA-7B+ (Zhang et al., 2025a)	Compressed	L2 Avg	0.22	nuScenes; open-loop; validation
Planning error	Reasoning-VLA-7B (Zhang et al., 2025a)	Compressed	L2 Avg	0.23	nuScenes; open-loop; validation
Planning error	VLA-World* (Wang et al., 2026a)	Predictive	L2 Avg	0.26	nuScenes; open-loop; ST-P3 metrics; validation
Planning error	SOLVE-VLM (Chen et al., 2025a)	Action-grounded	L2 Avg	0.28	nuScenes; open-loop; validation
Planning error	FSDrive-Qwen-2B* (Zeng et al., 2026)	Predictive	L2 Avg	0.28	nuScenes; ST-P3 metrics; validation
Planning error	ColaVLA (Peng et al., 2025c)	Optimization-based	L2 Avg	0.30	nuScenes; open-loop; validation
Planning error	VLA-World (Wang et al., 2026a)	Predictive	L2 Avg	0.30	nuScenes; open-loop; ST-P3 metrics; validation
Planning error	Reasoning-VLA-3B (Zhang et al., 2025a)	Compressed	L2 Avg	0.30	nuScenes; open-loop; validation
Planning error	LLaViDA (Liu et al., 2025g)	Procedural	L2 Avg	0.31	nuScenes; open-loop; ST-P3 metrics; validation
Planning error	Drive-R1 (Li et al., 2026b)	Compressed	L2 Avg	0.31	nuScenes; open-loop; validation
Planning error	DriveVLM-Dual (Tian et al., 2025b)	Procedural	L2 Avg	0.31	nuScenes; validation
Planning error	FSDrive-LLaVA-7B* (Zeng et al., 2026)	Predictive	L2 Avg	0.31	nuScenes; ST-P3 metrics; validation
Planning error	SpaceDrive+ (Li et al., 2025b)	Action-grounded	L2 Avg	0.32	nuScenes; ST-P3 metrics; validation
Planning error	AutoDrive-P3-Detailed (Ye et al., 2026)	Compressed	L2 Avg	0.33	nuScenes; open-loop; validation
Planning error	MindDriver* (Zhang et al., 2026c)	Action-grounded	L2 Avg	0.33	nuScenes; ST-P3 metrics; validation
Planning error	AutoDrive-P3-Fast (Ye et al., 2026)	Compressed	L2 Avg	0.34	nuScenes; open-loop; validation
Planning error	DriveVLM (Tian et al., 2025b)	Procedural	L2 Avg	0.40	nuScenes; validation
Planning error	OpenREAD (Zhang et al., 2025c)	Reflective	L2 Avg	0.40	nuScenes; open-loop; ST-P3 metrics; validation
Planning error	VLA-World* (Wang et al., 2026a)	Predictive	L2 Avg	0.42	nuScenes; open-loop; UniAD metrics; validation
Planning error	Drive-OccWorld (Yang et al., 2025)	Predictive	L2 Avg	0.47	nuScenes; TemAvg protocol; validation
Planning error	ReAL-AD (Lu et al., 2025)	Descriptive	L2 Avg	0.48	nuScenes; ST-P3 metrics; validation
Planning error	WorldRFT (Yang et al., 2026)	Optimization-based	L2 Avg	0.48	nuScenes; open-loop planning; validation
Planning error	AutoDrive-R^2-3B (Yuan et al., 2025)	Reflective	L2 Avg	0.49	nuScenes; open-loop; validation
Planning error	World4Drive (Zheng et al., 2025b)	Rollout-based	L2 Avg	0.50	nuScenes; open-loop planning; validation
Planning error	MindDriver (Zhang et al., 2026c)	Action-grounded	L2 Avg	0.53	nuScenes; ST-P3 metrics; validation
Planning error	FSDrive-Qwen-2B (Zeng et al., 2026)	Predictive	L2 Avg	0.53	nuScenes; ST-P3 metrics; validation
Planning error	EchoVLA (Guo et al., 2026)	Descriptive	L2 Avg	0.58	nuScenes; validation
Planning error	DriveWorld-VLA (Liu et al., 2026)	Rollout-based	L2 Avg	0.61	nuScenes; open-loop planning; validation
Planning error	MindDriver* (Zhang et al., 2026c)	Action-grounded	L2 Avg	0.65	nuScenes; open-loop; UniAD metrics; validation
Planning error	VERDI (Feng et al., 2025)	Compressed	L2 Avg	0.65	nuScenes; open-loop; validation
Planning error	AutoVLA (Zhou et al., 2026b)	Compressed	L2 Avg	0.70	nuScenes; UniAD metrics; validation
Planning error	Drive-WM (Wang et al., 2024b)	Predictive	L2 Avg	0.80	nuScenes; tree-based planning; validation
Planning error	RDA-Driver (Huang et al., 2024)	Compressed	L2 Avg	0.80	nuScenes; UniAD metrics; open-loop
Planning error	VLA-World (Wang et al., 2026a)	Predictive	L2 Avg	0.83	nuScenes; open-loop; UniAD metrics; validation
Planning error	MindDriver (Zhang et al., 2026c)	Action-grounded	L2 Avg	0.93	nuScenes; open-loop; UniAD metrics; validation
Planning error	UniUGP (ByteDance Seed, 2025)	Rollout-based	L2 Avg	1.23	nuScenes; planning; front-camera only; validation
Planning error	SpaceDrive (Li et al., 2025b)	Action-grounded	L2 Avg	1.8	nuScenes; open-loop; ST-P3 metrics; validation
Safety					
Safety	WorldRFT (Yang et al., 2026)	Optimization-based	Collision Avg	0.05	nuScenes; open-loop planning; validation
Safety	AutoDrive-P3-Detailed (Ye et al., 2026)	Compressed	Collision Avg	0.06	nuScenes; open-loop; validation
Safety	AutoDrive-R^2-7B (Yuan et al., 2025)	Reflective	Collision Avg	0.07	nuScenes; open-loop; validation

Continued on next page

Metric group	Method	Representation	Metric	Value	Protocol / Notes
Safety	Reasoning-VLA-7B+ (Zhang et al., 2025a)	Compressed	Collision Avg	0.07	nuScenes; open-loop; validation
Safety	AutoDrive-P3-Fast (Ye et al., 2026)	Compressed	Collision Avg	0.08	nuScenes; open-loop; validation
Safety	Reasoning-VLA-7B (Zhang et al., 2025a)	Compressed	Collision Avg	0.08	nuScenes; open-loop; validation
Safety	Agent-Driver (Mao et al., 2024)	Knowledge	Collision Avg	0.09	nuScenes; open-loop; ST-P3 metrics; validation
Safety	Drive-R1 (Li et al., 2026b)	Compressed	Collision Avg	0.09	nuScenes; open-loop; validation
Safety	FSDrive-Qwen-2B* (Zeng et al., 2026)	Predictive	Collision Avg	0.10	nuScenes; ST-P3 metrics; validation
Safety	SpaceDrive+ (Li et al., 2025b)	Action-grounded	Collision Avg	0.10	nuScenes; ST-P3 metrics; validation
Safety	VLA-World (Wang et al., 2026a)	Predictive	Collision Avg	0.10	nuScenes; open-loop; ST-P3 metrics; validation
Safety	Drive-OccWorld (Yang et al., 2025)	Predictive	Collision Avg	0.11	nuScenes; TemAvg protocol; validation
Safety	OpenREAD (Zhang et al., 2025c)	Reflective	Collision Avg	0.11	nuScenes; open-loop; ST-P3 metrics; validation
Safety	FSDrive-LLaVA-7B* (Zeng et al., 2026)	Predictive	Collision Avg	0.12	nuScenes; ST-P3 metrics; validation
Safety	MindDriver* (Zhang et al., 2026c)	Action-grounded	Collision Avg	0.12	nuScenes; ST-P3 metrics; validation
Safety	Reasoning-VLA-3B (Zhang et al., 2025a)	Compressed	Collision Avg	0.13	nuScenes; open-loop; validation
Safety	DriveAgent-R1 (Zheng et al., 2025a)	Compressed	Collision Avg	0.14	nuScenes; open-loop; validation
Safety	DriveWorld-VLA (Liu et al., 2026)	Rollout-based	Collision Avg	0.16	nuScenes; open-loop planning; validation
Safety	World4Drive (Zheng et al., 2025b)	Rollout-based	Collision Avg	0.16	nuScenes; open-loop planning; validation
Safety	FSDrive-Qwen-2B (Zeng et al., 2026)	Predictive	Collision Avg	0.17	nuScenes; ST-P3 metrics; validation
Safety	SOLVE-VLM (Chen et al., 2025a)	Action-grounded	Collision Avg	0.20	nuScenes; open-loop planning; validation
Safety	Agent-Driver (Mao et al., 2024)	Knowledge	Collision Avg	0.21	nuScenes; open-loop; UniAD metrics; validation
Safety	ColaVLA (Peng et al., 2025c)	Optimization-based	Collision Avg	0.23	nuScenes; open-loop planning; validation
Safety	Drive-WM (Wang et al., 2024b)	Predictive	Collision Avg	0.26	nuScenes; tree-based planning; validation
Safety	SOLVE-E2E (Chen et al., 2025a)	Action-grounded	Collision Avg	0.30	nuScenes; open-loop planning; validation
Safety	AutoVLA (Zhou et al., 2026b)	Compressed	Collision Avg	0.31	nuScenes; UniAD metrics; validation
Safety	RDA-Driver (Huang et al., 2024)	Compressed	Collision Avg	0.32	nuScenes; UniAD metrics; open-loop
Safety	dVLM-AD (Ma et al., 2025)	Optimization-based	Collision Avg	0.32	nuScenes; validation
Safety	UniUGP (ByteDance Seed, 2025)	Rollout-based	Collision Avg	0.33	nuScenes; planning; front-camera only; validation
Safety	SpaceDrive (Li et al., 2025b)	Action-grounded	Collision Avg	0.76	nuScenes; open-loop; ST-P3 metrics; validation
Generation / occupancy					
Generation / occupancy	Drive-OccWorld-A (Yang et al., 2025)	Predictive	mIoU_f	36.3	nuScenes; inflated GMO and flow forecasting; validation; 2s
Generation / occupancy	Drive-OccWorld-P (Yang et al., 2025)	Predictive	mIoU_f	36.3	nuScenes; inflated GMO and flow forecasting; validation; 2s
Generation / occupancy	Drive-OccWorld-P (Yang et al., 2025)	Predictive	mIoU_f	21.2	nuScenes; fine-grained GSO forecasting; validation; 2s
Generation / occupancy	Drive-OccWorld-P (Yang et al., 2025)	Predictive	VPQ_f	25.1	nuScenes; inflated GMO and flow forecasting; validation; 2s
Generation / occupancy	Drive-OccWorld-A (Yang et al., 2025)	Predictive	VPQ_f	23.7	nuScenes; inflated GMO and flow forecasting; validation; 2s
Generation / occupancy	Vista (Gao et al., 2024)	Predictive	FID	6.9	nuScenes; prediction fidelity; validation
Generation / occupancy	UniUGP (ByteDance Seed, 2025)	Rollout-based	FID	7.4	nuScenes; future frame generation; front-camera only; validation
Generation / occupancy	MindDriver (Zhang et al., 2026c)	Action-grounded	FID	9.4	nuScenes; future frame generation; validation
Generation / occupancy	VLA-World (Wang et al., 2026a)	Predictive	FID	9.8	nuScenes; future frame generation; validation
Generation / occupancy	FSDrive (Zeng et al., 2026)	Predictive	FID	10.1	nuScenes; future frame generation; validation
Generation / occupancy	Drive-WM (Wang et al., 2024b)	Predictive	FID	15.8	nuScenes; multiview video generation; validation
Generation / occupancy	UniUGP (ByteDance Seed, 2025)	Rollout-based	FVD	75.9	nuScenes; future frame generation; front-camera only; validation

Continued on next page

Metric group	Method	Representation	Metric	Value	Protocol / Notes
Generation / occupancy	Vista (Gao et al., 2024)	Predictive	FVD	89.4	nuScenes; prediction fidelity; validation
Generation / occupancy	Drive-WM (Wang et al., 2024b)	Predictive	FVD	122.7	nuScenes; multiview video generation; validation
Auxiliary					
Auxiliary	VLA-World (Wang et al., 2026a)	Predictive	Action F1 (forward)	95.88	nuScenes; action prediction; validation
Auxiliary	dVLM-AD (Ma et al., 2025)	Optimization-based	Behavior-Trajectory alignment avg	87.8	nuScenes; validation
Auxiliary	VLA-World (Wang et al., 2026a)	Predictive	Action F1 (right)	75.06	nuScenes; action prediction; validation
Auxiliary	VLA-World (Wang et al., 2026a)	Predictive	Action F1 (left)	74.22	nuScenes; action prediction; validation
Auxiliary	KLDrive (Tian et al., 2026)	Knowledge	NuScenes-QA overall accuracy	65.04	nuScenes; NuScenes-QA; validation
Auxiliary	DriveAgent-R1 (Zheng et al., 2025a)	Compressed	First-frame joint accuracy	52.96	nuScenes; high-level planning; test; 8s
Auxiliary	DriveAgent-R1 (Zheng et al., 2025a)	Compressed	Sequence average joint accuracy	47.1	nuScenes; high-level planning; test; 8s
Auxiliary	The Case for Negative Data (Patrikar et al., 2025)	Retrieval	VLM-only recall on REASONABLE actions (f_base)	24	nuScenes-derived custom benchmark; 1,275 action-scene pairs
Auxiliary	VERDI (Feng et al., 2025)	Compressed	FPS	4.5	nuScenes; open-loop; validation
Auxiliary	AutoVLA (Zhou et al., 2026b)	Compressed	Runtime (s)	3.95	nuScenes; open-loop; test
Auxiliary	VERDI (Feng et al., 2025)	Compressed	Latency (ms)	359	HugSim; closed-loop; overall
Auxiliary	VERDI (Feng et al., 2025)	Compressed	COM	0.954	HugSim; closed-loop; overall
Auxiliary	VERDI (Feng et al., 2025)	Compressed	DAC	0.944	HugSim; closed-loop; overall
Auxiliary	VERDI (Feng et al., 2025)	Compressed	NC	0.659	HugSim; closed-loop; overall
Auxiliary	VERDI (Feng et al., 2025)	Compressed	TTC	0.469	HugSim; closed-loop; overall
Auxiliary	VERDI (Feng et al., 2025)	Compressed	Rc	0.354	HugSim; closed-loop; overall
Auxiliary	VERDI (Feng et al., 2025)	Compressed	HDScore	0.2	HugSim; closed-loop; overall

NAVSIM

Table 8: Reported results for NAVSIM grouped by metric type.

Metric group	Method	Representation	Metric	Value	Protocol / Notes
Primary planning scores					
Primary planning scores	ReflectDrive† (Li et al., 2025c)	Reflective	PDMS	94.7	NAVSIM v1; navtest; oracle reflection upper bound
Primary planning scores	AdaThinkDrive-BoN (Luo et al., 2025c)	Compressed	PDMS	93.0	NAVSIM v1; navtest; Best-of-4 oracle selection
Primary planning scores	AutoVLA-BoN (Zhou et al., 2026b)	Compressed	PDMS	92.12	NAVSIM v1; navtest; Best-of-6 oracle selection; 4s
Primary planning scores	DriveWorld-VLA (Liu et al., 2026)	Rollout-based	PDMS	91.3	NAVSIM v1; navtest
Primary planning scores	LaST-VLA-8B-RL (Luo et al., 2026b)	Rollout-based	PDMS	91.3	NAVSIM v1; navtest; RL fine-tuned
Primary planning scores	ReflectDrive (Li et al., 2025c)	Reflective	PDMS	91.1	NAVSIM v1; navtest
Primary planning scores	SGDrive-RFT (Li et al., 2026a)	Rollout-based	PDMS	91.1	NAVSIM v1; navtest; RFT
Primary planning scores	ELF-VLA-8B (Luo et al., 2026a)	Reflective	PDMS	91.0	NAVSIM v1; navtest
Primary planning scores	ReCogDrive (Li et al., 2025e)	Optimization-based	PDMS	90.8	NAVSIM v1; navtest
Primary planning scores	FutureX-All-TransFuser (Lin et al., 2025)	Rollout-based	PDMS	90.6	NAVSIM v1; navtest; TransFuser backbone (camera+LiDAR)
Primary planning scores	AutoDrive-P3-Detailed (Ye et al., 2026)	Compressed	PDMS	90.6	NAVSIM v1; navtest
Primary planning scores	AdaThinkDrive (Luo et al., 2025c)	Compressed	PDMS	90.3	NAVSIM v1; navtest; adaptive CoT switching

Continued on next page

Metric group	Method	Representation	Metric	Value	Protocol / Notes
Primary planning scores	AutoDrive-P3-Fast (Ye et al., 2026)	Compressed	PDMS	90.2	NAVSIM v1; navtest
Primary planning scores	FutureX-Auto-LTF (Lin et al., 2025)	Rollout-based	PDMS	89.2	NAVSIM v1; navtest; LTF backbone (camera-only)
Primary planning scores	DriveDreamer-Policy (Zhou et al., 2026a)	Predictive	PDMS	89.2	NAVSIM v1; navtest
Primary planning scores	AutoVLA (Zhou et al., 2026b)	Compressed	PDMS	89.11	NAVSIM v1; navtest; SFT+CoT; 4s
Primary planning scores	AutoDrive-R^2-7B (Yuan et al., 2025)	Reflective	PDMS	90.3	NAVSIM v1; navtest
Primary planning scores	MTRDrive (Luo et al., 2025d)	Knowledge	PDMS	88.3	NAVSIM v1; navtest (split not explicitly stated)
Primary planning scores	WorldRFT (Yang et al., 2026)	Optimization-based	PDMS	87.8	NAVSIM v1; navtest; closed-loop planning
Primary planning scores	SGDrive-SFT (Li et al., 2026a)	Rollout-based	PDMS	87.4	NAVSIM v1; navtest; SFT
Primary planning scores	World4Drive (Zheng et al., 2025b)	Rollout-based	PDMS	85.1	NAVSIM v1; navtest; closed-loop planning
Primary planning scores	AutoDrive-P3-Detailed (Ye et al., 2026)	Compressed	EPDMS	89.9	NAVSIM v2; navtest
Primary planning scores	AutoDrive-P3-Fast (Ye et al., 2026)	Compressed	EPDMS	88.7	NAVSIM v2; navtest
Primary planning scores	DriveDreamer-Policy (Zhou et al., 2026a)	Predictive	EPDMS	88.7	NAVSIM v2; navtest
Primary planning scores	HiST-VLA (Wang et al., 2026b)	Tokenized	EPDMS	88.6	NAVSIM v2; navtest
Primary planning scores	ELF-VLA-8B (Luo et al., 2026a)	Reflective	EPDMS	87.1	NAVSIM v2; navtest
Primary planning scores	LaST-VLA-8B-RL (Luo et al., 2026b)	Rollout-based	EPDMS	87.1	NAVSIM v2; navtest; RL fine-tuned
Primary planning scores	DriveWorld-VLA (Liu et al., 2026)	Rollout-based	EPDMS	86.8	NAVSIM v2; navtest
Primary planning scores	AutoDrive-P3-Detailed (Ye et al., 2026)	Compressed	EPDMS	86.2	NAVSIM v2; navtest; alt. config
Primary planning scores	AutoDrive-P3-Fast (Ye et al., 2026)	Compressed	EPDMS	85.2	NAVSIM v2; navtest; alt. config
Primary planning scores	HiST-VLA (Wang et al., 2026b)	Tokenized	EPDMS	50.9	NAVSIM v2; Navhard pseudo-closed-loop
Primary planning scores	DiffVLA++ (Gao et al., 2025b)	Optimization-based	EPDMS	49.12	NAVSIM v2; ICCV 2025 AGC public leaderboard; combined ensemble
Primary planning scores	DiffVLA++-VLA (Gao et al., 2025b)	Optimization-based	EPDMS	48.0	NAVSIM v2; Navhard two-stage test; VLA branch
Primary planning scores	DiffVLA (Jiang et al., 2025a)	Optimization-based	EPDMS	45.01	NAVSIM v2; AGC private test; combined
Primary planning scores	DiffVLA++-E2E (Gao et al., 2025b)	Optimization-based	EPDMS	43.7	NAVSIM v2; Navhard two-stage test; E2E branch
Subscores					
Subscores	ReflectDrive† (Li et al., 2025c)	Reflective	NC	99.7	NAVSIM v1; navtest
Subscores	HiST-VLA (Wang et al., 2026b)	Tokenized	NC	99.6	NAVSIM v2; navtest
Subscores	AutoDrive-P3-Detailed (Ye et al., 2026)	Compressed	NC	99.1	NAVSIM v1; navtest
Subscores	ELF-VLA-8B (Luo et al., 2026a)	Reflective	NC	98.9	NAVSIM v1; navtest
Subscores	ELF-VLA-8B (Luo et al., 2026a)	Reflective	NC	98.9	NAVSIM v2; navtest

Continued on next page

Metric group	Method	Representation	Metric	Value	Protocol / Notes
Subscores	SGDrive (Li et al., 2026a)	Rollout-based	NC	98.6	NAVSIM v1; navtest RFT
Subscores	AutoDrive-R^2-7B (Yuan et al., 2025)	Reflective	NC	98.5	NAVSIM v1; navtest
Subscores	AdaThinkDrive (Luo et al., 2025c)	Compressed	NC	98.4	NAVSIM v1; navtest
Subscores	ReCogDrive (Li et al., 2025e)	Optimization-based	NC	97.9	NAVSIM v1; navtest
Subscores	ReflectDrive (Li et al., 2025c)	Reflective	NC	97.7	NAVSIM v1; navtest
Subscores	ReflectDrive† (Li et al., 2025c)	Reflective	DAC	99.5	NAVSIM v1; navtest
Subscores	ReflectDrive (Li et al., 2025c)	Reflective	DAC	99.3	NAVSIM v1; navtest
Subscores	HiST-VLA (Wang et al., 2026b)	Tokenized	DAC	99.1	NAVSIM v2; navtest
Subscores	ELF-VLA-8B (Luo et al., 2026a)	Reflective	DAC	98.1	NAVSIM v1; navtest
Subscores	ELF-VLA-8B (Luo et al., 2026a)	Reflective	DAC	98.1	NAVSIM v2; navtest
Subscores	AdaThinkDrive (Luo et al., 2025c)	Compressed	DAC	97.8	NAVSIM v1; navtest
Subscores	SGDrive (Li et al., 2026a)	Rollout-based	DAC	97.8	NAVSIM v1; navtest RFT
Subscores	AutoDrive-P3-Detailed (Ye et al., 2026)	Compressed	DAC	97.4	NAVSIM v1; navtest
Subscores	ReCogDrive (Li et al., 2025e)	Optimization-based	DAC	97.3	NAVSIM v1; navtest
Subscores	AutoVLA-BoN (Zhou et al., 2026b)	Compressed	DAC	97.08	NAVSIM v1; open-loop; test; 4s
Subscores	AutoDrive-R^2-7B (Yuan et al., 2025)	Reflective	DAC	95.9	NAVSIM v1; navtest
Subscores	AutoVLA-RFT (Zhou et al., 2026b)	Compressed	DAC	95.64	NAVSIM v1; open-loop; test; 4s
Subscores	HiST-VLA (Wang et al., 2026b)	Tokenized	TTC	99.4	NAVSIM v2; navtest
Subscores	ReflectDrive† (Li et al., 2025c)	Reflective	TTC	99.1	NAVSIM v1; navtest
Subscores	ELF-VLA-8B (Luo et al., 2026a)	Reflective	TTC	98.4	NAVSIM v2; navtest
Subscores	AutoVLA-RFT (Zhou et al., 2026b)	Compressed	TTC	98.04	NAVSIM v1; open-loop; test; 4s
Subscores	AutoVLA-BoN (Zhou et al., 2026b)	Compressed	TTC	97.12	NAVSIM v1; open-loop; test; 4s
Subscores	AutoDrive-P3-Detailed (Ye et al., 2026)	Compressed	TTC	96.5	NAVSIM v1; navtest
Subscores	ELF-VLA-8B (Luo et al., 2026a)	Reflective	TTC	96.0	NAVSIM v1; navtest
Subscores	AutoDrive-R^2-7B (Yuan et al., 2025)	Reflective	TTC	95.4	NAVSIM v1; navtest
Subscores	AdaThinkDrive (Luo et al., 2025c)	Compressed	TTC	95.2	NAVSIM v1; navtest
Subscores	ReflectDrive (Li et al., 2025c)	Reflective	TTC	93.5	NAVSIM v1; navtest
Subscores	AdaThinkDrive (Luo et al., 2025c)	Compressed	Comfort	100	NAVSIM v1; navtest
Subscores	AutoDrive-R^2-7B (Yuan et al., 2025)	Reflective	Comfort	100	NAVSIM v1; navtest
Subscores	AutoDrive-P3-Detailed (Ye et al., 2026)	Compressed	Comfort	100	NAVSIM v1; navtest
Subscores	ELF-VLA-8B (Luo et al., 2026a)	Reflective	Comfort	100	NAVSIM v1; navtest
Subscores	ReflectDrive (Li et al., 2025c)	Reflective	Comfort	100	NAVSIM v1; navtest
Subscores	AutoVLA-BoN (Zhou et al., 2026b)	Compressed	Comfort	99.98	NAVSIM v1; open-loop; test; 4s
Subscores	AutoVLA-RFT (Zhou et al., 2026b)	Compressed	Comfort	99.94	NAVSIM v1; open-loop; test; 4s
Subscores	ReflectDrive† (Li et al., 2025c)	Reflective	Comfort	99.9	NAVSIM v1; navtest
Subscores	HiST-VLA (Wang et al., 2026b)	Tokenized	DDC	99.7	NAVSIM v2; navtest
Subscores	ELF-VLA-8B (Luo et al., 2026a)	Reflective	DDC	99.4	NAVSIM v2; navtest
Subscores	AutoVLA-BoN (Zhou et al., 2026b)	Compressed	DDC	95.51	NAVSIM v1; open-loop; test; 4s
Subscores	AutoVLA-RFT (Zhou et al., 2026b)	Compressed	DDC	95.40	NAVSIM v1; open-loop; test; 4s
Subscores	HiST-VLA (Wang et al., 2026b)	Tokenized	TLC	99.9	NAVSIM v2; navtest
Subscores	ELF-VLA-8B (Luo et al., 2026a)	Reflective	TLC	99.8	NAVSIM v2; navtest

Continued on next page

Metric group	Method	Representation	Metric	Value	Protocol / Notes
Subscores	HiST-VLA (Wang et al., 2026b)	Tokenized	LK	98.9	NAVSIM v2; navtest
Subscores	ELF-VLA-8B (Luo et al., 2026a)	Reflective	LK	96.9	NAVSIM v2; navtest
Subscores	HiST-VLA (Wang et al., 2026b)	Tokenized	HC	98.4	NAVSIM v2; navtest
Subscores	ELF-VLA-8B (Luo et al., 2026a)	Reflective	HC	98.3	NAVSIM v2; navtest
Subscores	ELF-VLA-8B (Luo et al., 2026a)	Reflective	EC	87.2	NAVSIM v2; navtest
Subscores	HiST-VLA (Wang et al., 2026b)	Tokenized	EC	66.0	NAVSIM v2; navtest
Subscores	HiST-VLA (Wang et al., 2026b)	Tokenized	EP	89.2	NAVSIM v2; navtest
Subscores	ReflectDrive† (Li et al., 2025c)	Reflective	EP	88.9	NAVSIM v1; navtest
Subscores	ELF-VLA-8B (Luo et al., 2026a)	Reflective	EP	88.5	NAVSIM v2; navtest
Subscores	AutoVLA-BoN (Zhou et al., 2026b)	Compressed	EP	87.55	NAVSIM v1; open-loop; test; 4s
Subscores	ReflectDrive (Li et al., 2025c)	Reflective	EP	86.9	NAVSIM v1; navtest
Subscores	SGDrive (Li et al., 2026a)	Rollout-based	EP	85.8	NAVSIM v1; navtest RFT
Subscores	ELF-VLA-8B (Luo et al., 2026a)	Reflective	EP	85.3	NAVSIM v1; navtest
Subscores	AutoDrive-P3-Detailed (Ye et al., 2026)	Compressed	EP	84.8	NAVSIM v1; navtest
Subscores	AdaThinkDrive (Luo et al., 2025c)	Compressed	EP	84.4	NAVSIM v1; navtest
Subscores	AutoDrive-R^2-7B (Yuan et al., 2025)	Reflective	EP	82.7	NAVSIM v1; navtest
Subscores	AutoVLA-RFT (Zhou et al., 2026b)	Compressed	EP	81.87	NAVSIM v1; open-loop; test; 4s
Safety / latency / auxiliary					
Safety / latency / auxiliary	AutoVLA-BoN (Zhou et al., 2026b)	Compressed	Collision score	99.14	NAVSIM; open-loop; test; 4s
Safety / latency / auxiliary	AutoVLA-RFT (Zhou et al., 2026b)	Compressed	Collision score	98.41	NAVSIM; open-loop; test; 4s
Safety / latency / auxiliary	DiffVLA (Jiang et al., 2025a)	Optimization-based	DAC (stage 2)	88.84	NAVSIM v2; AGC private test; stage 2
Safety / latency / auxiliary	DiffVLA (Jiang et al., 2025a)	Optimization-based	NC (stage 2)	81.27	NAVSIM v2; AGC private test; stage 2
Safety / latency / auxiliary	DiffVLA (Jiang et al., 2025a)	Optimization-based	TTC (stage 2)	76.46	NAVSIM v2; AGC private test; stage 2
Safety / latency / auxiliary	ELF-VLA (Luo et al., 2026a)	Reflective	Path Acc.	92.5	NAVSIM; high-level planning
Safety / latency / auxiliary	OneVL (Lu et al., 2026)	Tokenized	PDM-score	88.84	NAVSIM; planning benchmark; test
Safety / latency / auxiliary	ELF-VLA (Luo et al., 2026a)	Reflective	Accuracy	80.3	NAVSIM; high-level planning
Safety / latency / auxiliary	OneVL (Lu et al., 2026)	Tokenized	Meta Action Accuracy	71.00	NAVSIM; text CoT quality; test
Safety / latency / auxiliary	Reasoning-VLA-7B (Zhang et al., 2025a)	Compressed	L2 avg	0.22	NAVSIM; generalized open-loop; recommended split; 3s
Safety / latency / auxiliary	OneVL (Lu et al., 2026)	Tokenized	Latency (MLP head)	0.24	NAVSIM; real-world deployment; test
Safety / latency / auxiliary	AdaThinkDrive (Luo et al., 2025c)	Compressed	Inference time (s)	0.74	NAVSIM; closed-loop; test; 4s
Safety / latency / auxiliary	OneVL (Lu et al., 2026)	Tokenized	Latency (full pipeline)	4.46	NAVSIM; planning benchmark; test
Safety / latency / auxiliary	DriveDreamer-Policy (Zhou et al., 2026a)	Predictive	AbsRel	8.1	NAVSIM; depth generation; test
Safety / latency / auxiliary	DriveDreamer-Policy (Zhou et al., 2026a)	Predictive	FVD	53.59	NAVSIM; world generation; test

Bench2Drive / CARLA

Table 9: Reported results for Bench2Drive / CARLA grouped by metric type.

Metric group	Method	Representation	Metric	Value	Protocol / Notes
Driving performance					
Driving performance	AutoVLA (Zhou et al., 2026b)	Compressed	Driving Score (0–100)	78.84	Bench2Drive; closed-loop; test
Driving performance	MindDrive (Fu et al., 2025b)	Compressed	Driving Score (0–100)	78.04	Bench2Drive; closed-loop
Driving performance	SpaceDrive+ (Li et al., 2025b)	Action-grounded	Driving Score (0–100)	78.02	Bench2Drive; closed-loop
Driving performance	ORION (Fu et al., 2025a)	Descriptive	Driving Score (0–100)	77.74	Bench2Drive; closed-loop
Driving performance	ReCogDrive (Li et al., 2025e)	Optimization-based	Driving Score (0–100)	71.36	Bench2Drive; closed-loop
Driving performance	MindDriver (Zhang et al., 2026c)	Action-grounded	Driving Score (0–100)	65.48	Bench2Drive; closed-loop
Driving performance	LeAD (Zhang et al., 2025e)	Knowledge	Driving Score (0–100)	71.96	CARLA; Leaderboard V1; closed-loop
Driving performance	LangCoop (Gao et al., 2025a)	Cooperative	Driving Score (0–100)	48.8	CARLA; closed-loop
Driving performance	AutoVLA (Zhou et al., 2026b)	Compressed	Success Rate (0–100)	57.73	Bench2Drive; closed-loop; test
Driving performance	SpaceDrive+ (Li et al., 2025b)	Action-grounded	Success Rate (0–100)	55.11	Bench2Drive; closed-loop
Driving performance	MindDrive (Fu et al., 2025b)	Compressed	Success Rate (0–100)	55.09	Bench2Drive; closed-loop
Driving performance	ORION (Fu et al., 2025a)	Descriptive	Success Rate (0–100)	54.62	Bench2Drive; closed-loop
Driving performance	ReCogDrive (Li et al., 2025e)	Optimization-based	Success Rate (0–100)	45.45	Bench2Drive; closed-loop
Driving performance	MindDriver (Zhang et al., 2026c)	Action-grounded	Success Rate (0–100)	39.55	Bench2Drive; closed-loop
Driving performance	DriveMind (Wasif et al., 2026)	Compressed	Route Completion (0–1)	0.98	CARLA; closed-loop; Town02
Driving performance	DriveMind (Wasif et al., 2026)	Compressed	Success Rate (0–1)	0.97	CARLA; closed-loop; Town02
Driving performance	COVLM-RL-Town03 (Li et al., 2025a)	Compressed	Route Completion (0–1)	0.88	CARLA; closed-loop; unseen Town03
Driving performance	COVLM-RL-Town02 (Li et al., 2025a)	Compressed	Route Completion (0–1)	0.78	CARLA; closed-loop; trained Town02
Driving performance	COVLM-RL-Town03 (Li et al., 2025a)	Compressed	Success Rate (0–1)	0.70	CARLA; closed-loop; unseen Town03
Driving performance	COVLM-RL-Town02 (Li et al., 2025a)	Compressed	Success Rate (0–1)	0.70	CARLA; closed-loop; trained Town02
Driving performance	COVLM-RL-Town03 (Li et al., 2025a)	Compressed	CDS (0–1, ↓)	0.64	CARLA; closed-loop; unseen Town03
Driving performance	COVLM-RL-Town02 (Li et al., 2025a)	Compressed	CDS (0–1, ↓)	0.63	CARLA; closed-loop; trained Town02
Driving performance	CoLMDriver (Liu et al., 2025a)	Cooperative	Driving Score (0–100)	88.53	CARLA; InterDrive-total; closed-loop
Driving performance	WiseAD (Zhang et al., 2024)	Descriptive	Driving Score (relative, %)	+11.9	CARLA; closed-loop; improvement over baseline
Driving performance	SteerVLA (Gao et al., 2026)	Descriptive	Driving Score (relative, pts)	+4.77	Bench2Drive; improvement over prior SOTA
Driving performance	DriveMLM (Wang et al., 2023a)	Procedural	Driving Score (relative, pts)	+4.7	CARLA Town05 Long; improvement over Apollo baseline
Safety / infractions					
Safety / infractions	DriveMind (Wasif et al., 2026)	Compressed	Collision speed (km/h)	0.01	CARLA; closed-loop; held-out maps (Towns 1/3/4/5)
Safety / infractions	UniDrive-WM (Xiong et al., 2026)	Predictive	Collision rate reduction (relative, %)	10.4	Bench2Drive; closed-loop; test
Auxiliary					
Auxiliary	MindDrive (Fu et al., 2025b)	Compressed	Overtaking ability	75.56	Bench2Drive; closed-loop; official benchmark
Auxiliary	MindDrive (Fu et al., 2025b)	Compressed	Emergency Brake ability	68.33	Bench2Drive; closed-loop; official benchmark
Auxiliary	MindDrive (Fu et al., 2025b)	Compressed	Traffic Sign ability	57.89	Bench2Drive; closed-loop; official benchmark
Auxiliary	MindDrive (Fu et al., 2025b)	Compressed	Ability mean	56.94	Bench2Drive; closed-loop; official benchmark
Auxiliary	MindDrive (Fu et al., 2025b)	Compressed	Give Way ability	50.00	Bench2Drive; closed-loop; official benchmark
Auxiliary	MindDrive (Fu et al., 2025b)	Compressed	Merging ability	32.89	Bench2Drive; closed-loop; official benchmark

Continued on next page

Metric group	Method	Representation	Metric	Value	Protocol / Notes
Auxiliary	AutoVLA (Zhou et al., 2026b)	Compressed	Efficiency	146.93	Bench2Drive; closed-loop; test
Auxiliary	AutoVLA (Zhou et al., 2026b)	Compressed	Comfortness	39.33	Bench2Drive; closed-loop; test
Auxiliary	COVLM-RL-Town02 (Li et al., 2025a)	Compressed	Traveled Distance (TD)	1458.95	CARLA; closed-loop; Town02 (trained)
Auxiliary	COVLM-RL-Town03 (Li et al., 2025a)	Compressed	Traveled Distance (TD)	3468.47	CARLA; closed-loop; Town03 (unseen)
Auxiliary	COVLM-RL-Town02 (Li et al., 2025a)	Compressed	Average Distance (AD)	145.89	CARLA; closed-loop; Town02 (trained)
Auxiliary	COVLM-RL-Town03 (Li et al., 2025a)	Compressed	Average Distance (AD)	346.96	CARLA; closed-loop; Town03 (unseen)
Auxiliary	COVLM-RL-Town02 (Li et al., 2025a)	Compressed	Speed Mean (SM)	18.74	CARLA; closed-loop; Town02 (trained)
Auxiliary	COVLM-RL-Town03 (Li et al., 2025a)	Compressed	Speed Mean (SM)	21.38	CARLA; closed-loop; Town03 (unseen)
Auxiliary	COVLM-RL-Town02 (Li et al., 2025a)	Compressed	Speed Std (SS)	3.34	CARLA; closed-loop; Town02 (trained)
Auxiliary	COVLM-RL-Town03 (Li et al., 2025a)	Compressed	Speed Std (SS)	3.25	CARLA; closed-loop; Town03 (unseen)
Auxiliary	COVLM-RL-Town02 (Li et al., 2025a)	Compressed	Reward Mean (RM)	4.29	CARLA; closed-loop; Town02 (trained)
Auxiliary	COVLM-RL-Town03 (Li et al., 2025a)	Compressed	Reward Mean (RM)	4.41	CARLA; closed-loop; Town03 (unseen)
Auxiliary	COVLM-RL-Town02 (Li et al., 2025a)	Compressed	Reward Std (RS)	1.20	CARLA; closed-loop; Town02 (trained)
Auxiliary	COVLM-RL-Town03 (Li et al., 2025a)	Compressed	Reward Std (RS)	1.33	CARLA; closed-loop; Town03 (unseen)
Auxiliary	COVLM-RL-Town02 (Li et al., 2025a)	Compressed	Centerline Deviation Mean (CDM)	0.71	CARLA; closed-loop; Town02 (trained)
Auxiliary	COVLM-RL-Town03 (Li et al., 2025a)	Compressed	Centerline Deviation Mean (CDM)	0.82	CARLA; closed-loop; Town03 (unseen)
Auxiliary	DriveMind (Wasif et al., 2026)	Compressed	Total distance (m)	2083.10	CARLA; closed-loop; held-out maps
Auxiliary	DriveMind (Wasif et al., 2026)	Compressed	Average speed (km/h)	19.40	CARLA; closed-loop; held-out maps
Auxiliary	AFSP (Chen et al., 2026)	Compressed	Average Lateral Deviation (m)	1.21	CARLA; closed-loop; Scenario 1 (nominal map)
Auxiliary	AFSP (Chen et al., 2026)	Compressed	Average Lateral Deviation (m)	1.01	CARLA; closed-loop; Scenario 2 (shifted map)
Auxiliary	AFSP (Chen et al., 2026)	Compressed	Average Lateral Deviation (m)	1.02	CARLA; closed-loop; Scenario 3 (shifted map)
Auxiliary	AFSP (Chen et al., 2026)	Compressed	Maximum Lateral Deviation (m)	3.72	CARLA; closed-loop; Scenario 1 (nominal map)
Auxiliary	AFSP (Chen et al., 2026)	Compressed	Maximum Lateral Deviation (m)	3.23	CARLA; closed-loop; Scenario 2 (shifted map)
Auxiliary	AFSP (Chen et al., 2026)	Compressed	Maximum Lateral Deviation (m)	3.42	CARLA; closed-loop; Scenario 3 (shifted map)
Auxiliary	AFSP (Chen et al., 2026)	Compressed	Speed Variation (m/s)	1.56	CARLA; closed-loop; Scenario 1 (nominal map)
Auxiliary	AFSP (Chen et al., 2026)	Compressed	Speed Variation (m/s)	1.30	CARLA; closed-loop; Scenario 2 (shifted map)
Auxiliary	AFSP (Chen et al., 2026)	Compressed	Speed Variation (m/s)	1.48	CARLA; closed-loop; Scenario 3 (shifted map)
Auxiliary	AFSP (Chen et al., 2026)	Compressed	Finish Time (s)	14.85	CARLA; closed-loop; Scenario 1 (nominal map)
Auxiliary	AFSP (Chen et al., 2026)	Compressed	Finish Time (s)	15.02	CARLA; closed-loop; Scenario 2 (shifted map)
Auxiliary	AFSP (Chen et al., 2026)	Compressed	Finish Time (s)	14.73	CARLA; closed-loop; Scenario 3 (shifted map)
Auxiliary	AFSP (Chen et al., 2026)	Compressed	Trajectory Length (m)	84.21	CARLA; closed-loop; Scenario 1 (nominal map)
Auxiliary	AFSP (Chen et al., 2026)	Compressed	Trajectory Length (m)	83.86	CARLA; closed-loop; Scenario 2 (shifted map)
Auxiliary	AFSP (Chen et al., 2026)	Compressed	Trajectory Length (m)	83.72	CARLA; closed-loop; Scenario 3 (shifted map)
Auxiliary	ReasonPlan (Liu et al., 2025f)	Procedural	L2 error reduction (relative, %)	19.0	Bench2Drive; open-loop planning; vs. E2E imitation baseline
Auxiliary	UniDrive-WM (Xiong et al., 2026)	Predictive	L2 trajectory error reduction (relative, %)	7.3	Bench2Drive; open-loop planning; test