A Survey on Vision-Language-Action Models for Autonomous Driving

Sicong Jiang[1]*, Zilin Huang[4]*, Kangan Qian[2]*, Ziang Luo[2], Tianze Zhu[2], Yang Zhong[3], Yihong Tang[1], Menglin Kong[1], Yunlong Wang[2], Siwen Jiao[3], Hao Ye[3], Zihao Sheng[4], Xin Zhao[2], Tuopu Wen[2], Zheng Fu[2], Sikai Chen[4], Kun Jiang[2,6], Diange Yang[2,6], Seongjin Choi[5], Lijun Sun[1]

[1] McGill University, Canada
[2] Tsinghua University, China
[3] Xiaomi Corporation
[4] University of Wisconsin–Madison, USA
[5] University of Minnesota–Twin Cities, USA
[6] State Key Laboratory of Intelligent Green Vehicle and Mobility, Tsinghua University, China

sicong.jiang@mail.mcgill.ca, lijun.sun@mcgill.ca

Abstract

The rapid progress of multimodal large language models (MLLM) has paved the way for Vision-Language-Action (VLA) paradigms, which integrate visual perception, natural language understanding, and control within a single policy. Researchers in autonomous driving are actively adapting these methods to the vehicle domain. Such models promise autonomous vehicles that can interpret high-level instructions, reason about complex traffic scenes, and make their own decisions. However, the literature remains fragmented and is rapidly expanding. This survey offers the first comprehensive overview of VLA for Autonomous Driving (VLA4AD). We (i) formalize the architectural building blocks shared across recent work, (ii) trace the evolution from early explainer to reasoning-centric VLA models, and (iii) compare over 20 representative models according to VLA's progress in the autonomous driving domain. We also consolidate existing datasets and benchmarks, highlighting protocols that jointly measure driving safety, accuracy, and explanation quality. Finally, we detail open challenges—robustness, real-time efficiency, and formal verification—and outline future directions of VLA4AD. This survey provides a concise yet complete reference for advancing interpretable socially aligned autonomous vehicles. Github repo is available at JohnsonJiang/Awesome-VLA4AD.

1. Introduction

Autonomous vehicles must simultaneously *perceive* complex 3D scenes, *understand* traffic context, and *act* safely in real time. Classic autonomous driving (AD) stacks achieve this through a hand-engineered cascade of perception, prediction, planning, and control modules. While decades of research have made such pipelines reliable under common conditions, they remain brittle at module boundaries and struggle with long-tail corner cases - scenarios that demand high-level reasoning or nuanced human interaction.

Progress in foundation models, such as vision-language models (VLMs) and large language models (LLMs), has introduced strong semantic priors into driving perception. By aligning pixels with text, these models can explain scenes, answer questions, or retrieve contextual information that traditional detectors may miss [53, 74, 100, 102, 115]. Early adaptations have improved generalization to rare objects and provided human-readable explanations, e.g., describing an ambulance's trajectory or justifying a red-light stop. However, VLM-augmented stacks remain *passive*: they reason *about* the scene but do not decide *what to do*. Their language output is loosely coupled to low-level control and may hallucinate hazards or misinterpret colloquial instructions. In short, while VLMs enhance interpretability, they leave the action gap unresolved.

Recent work has thus proposed a more integrated paradigm: VLA models that fuse camera streams, natural language instructions, and low-level actuation within a single policy [17, 57, 165]. By conditioning the controller on language tokens, these systems can (i) follow free-form commands such as "yield to the ambulance" [105], (ii) verbalize their internal rationale for post-hoc verification [17], and (iii) leverage commonsense priors from internet-scale corpora to extrapolate in rare or unforeseen situations [18]. Early prototypes have already demonstrated improved safety and instruction fidelity in simulation and closed-loop

*Equal contribution.

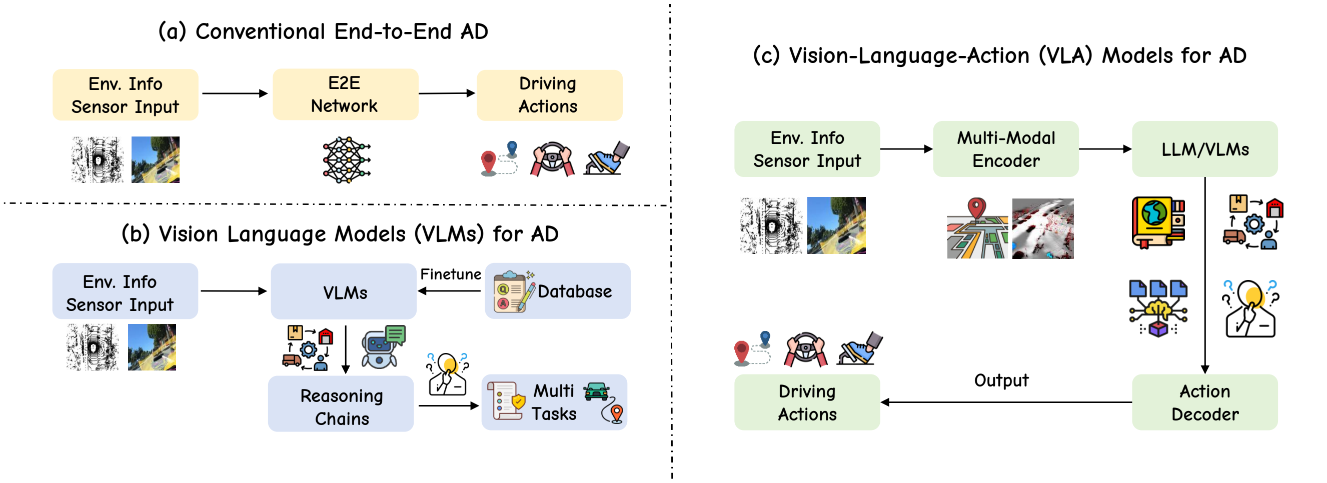

Figure 1. Comparisons of autonomous driving paradigms. (a) End-to-end driving offers direct perception-to-control mapping but lacks interpretability and generalization. (b) VLM4AD introduces natural language reasoning and explainability, yet remains perception-centric. (c) VLA4AD integrates perception, reasoning, and action, enabling interpretable and robust closed-loop control.

tests, foreshadowing a new research frontier we term *VLA for Autonomous Driving (VLA4AD)*.

Several converging trends underscore the timeliness of this emerging research frontier. First, petabyte-scale multi-sensor logs such as nuScenes [7] and Impromptu VLA [18] provide rich multimodal supervision. Second, foundation LLMs can now be efficiently adapted through low-rank updates, while token-reduction designs such as TS-VLM [11] reduce online computational cost by an order of magnitude [167]. Third, synthetic corpora like SimLingo [105] and interactive datasets such as NuInteract [160] enable researchers to stress-test language-conditioned behaviors well before real-world deployment. Together, these developments have sparked a surge of VLA architectures, ranging from explanatory overlays to reasoning-centric agents with chain-of-thought (CoT) memory.

Although several surveys now cover the application of LLMs and VLMs to autonomous driving [22, 23, 32, 166], none have yet addressed the rapidly emerging VLA paradigm in AD. To close this gap and consolidate this rapidly expanding body of work, we present the first comprehensive survey of VLA4AD. We first clarify key terminology and relate VLA to traditional end-to-end driving. Then, we distill common architectural patterns and catalogue over twenty representative models along with the datasets that support them. Also, we compare training paradigms and summarize evaluation protocols that jointly assess control performance and language fidelity. Finally, we outline open challenges and chart promising future directions.We also highlight the need for standardized benchmarks and open-source toolkits to foster reproducibility and accelerate cross-model comparisons. Our goal is to provide a coherent and forward-looking reference on how vision, language, and action are converging to shape the next generation of transparent, instruction-following, and socially compliant autonomous vehicles.

2. Development of Autonomous Driving

The technical arc of AD can be traced through four main paradigms: *modular stacks*, *end-to-end learning*, *VLM4AD*, and the recent *VLA4AD* wave.

2.1. Classical Modular Pipelines

The first wave of AD systems-epitomized by the DARPA Urban Challenge vehicles-explicitly factorized the driving task into distinct modules: perception[49, 80, 157], prediction[2, 45], planning[43, 109], and control [12, 34, 61, 62, 95, 123]. Hand-crafted algorithms processed LiDAR, radar, and GPS data: traditional vision detectors identified objects, finite-state machines or graph-based search generated paths, and PID or MPC controllers executed the final commands [52, 66]. This architecture was widely adopted in industry due to its modularity-each component could be engineered, tested, and improved in isolation. However, such strict decoupling leads to information fragmentation: upstream errors propagate without correction, and misaligned objectives across modules hinder end-to-end optimization [73, 83, 107].

2.2. End-to-End Autonomous Driving

The modular design of autonomous driving systems suffers from error propagation and information loss across the perception, prediction, and planning modules. Consequently, research has shifted towards more integrated, end-to-end approaches. As illustrated in Fig. 1 (a), end-to-end(E2E) driving policies map raw sensor streams directly to control commands, bypassing hand-crafted modular pipelines [9, 19, 20, 27, 40, 54, 106, 111, 112, 138, 145, 149].

E2E driving fundamentally operates as a Vision-to-Action (VA) system, where visual input can be from cameras or LiDAR, and the action output is typically represented as future trajectories or control signals. However, directly mapping raw sensory input to driving actions presents significant challenges due to the sparsity of planning-level data and the vast, unstructured solution space inherent in neural networks[13]. To mitigate these issues, early E2E approaches introduced intermediate supervision through integrated perception and prediction tasks[20, 111]. Specifically, these methods integrate perception, prediction, and planning modules within a unified framework, primarily facilitating feature-level information flow across modules. UniAD [44] primarily relies on rasterized representations (e.g., semantic maps, occupancy maps, flow maps, and costmaps), which is computationally intensive. In contrast, the proposed VAD [59, 86] adopts a fully vectorized scene representation for end-to-end planning. VAD leverages instance-level structural information as both constraints and guidance for planning, achieving promising performance with higher efficiency. PolarPoint-BEV [28] further refines the BEV representation by employing a polar point encoding. This incorporates a distance-based importance weighting prior, enabling the model to focus more effectively on critical objects at varying ranges during driving.

To model the interactions between the ego vehicle and other traffic participants, GenAD [162] and PPAD[16] leverages instance-level visual features, whereas GraphAD [156] represents these features as nodes in a graph. SparseAD [151] and SparseDrive [117]formulates a fully sparse architecture, achieving greater efficiency and superior planning performance. However, these methods typically rely on constructing computationally expensive BEV features and prevent downstream tasks from learning directly from raw sensor inputs. To mitigate this challenge, PARA-Drive [132] and DriveTransformer [56] introduces a parallel pipeline architecture, enabling a more unified and scalable framework for end-to-end driving systems. This approach explicitly models the relations between perception, prediction, and planning tasks. However, these paradigms also fall short in handling corner cases and struggle with out-of-distribution (OOD) scenarios. For instance, a driving system might fail to generate appropriate trajectories when encountering rare events, such as a vehicle breaking down at an intersection [140].

Several methods attempt to mitigate these issues. Physics-law enhanced frameworks incorporate prior knowledge [164], while others utilize objective functions and planning action priors to refine trajectories [14, 35, 54, 71, 72, 77, 114, 127, 144]. However, designing such objective functions for trajectory refinement remains both computationally expensive and labor-intensive. Subsequent research aims to reduce annotation burdens for 3D perception tasks through self-supervised or weakly-supervised frameworks, such as [36, 68, 69, 82].

Furthermore, closed-loop evaluations reveal diminishing returns beyond certain data volumes, coupled with significant performance variance across different scenario types [92, 163]. These findings indicate that pure data scaling alone is insufficient for achieving Level 4+ autonomy.

Overall, end-to-end learning has significantly narrowed the gap between raw sensor inputs and control decisions. Yet, persistent challenges remain, including: (i) Brittle semantics: Vulnerability to rare or rapidly evolving scenarios. (ii) Opaque reasoning: Limited interpretability hindering safety auditing and verification. (iii) Limited language proficiency: Restricting intuitive human-vehicle interaction.

2.3. VLMs for Autonomous Driving

Combining language modalities with driving tasks provides a promising direction to enhance reasoning, interpretability, and generalization in autonomous driving systems. LLMs [120, 122] and VLMs [1, 74, 102] offer a promising remedy by unifying perception and natural language reasoning within a shared embedding space [47]. At the core of this progress is large-scale multimodal pretraining, which equips models with commonsense associations (e.g., siren → yield) that task-specific labels often miss. Consequently, language-conditioned VLM policies exhibit stronger zero-shot generalization across novel objects, weather conditions, and driving norms. Recent work [29, 46, 75, 84, 97, 125, 126, 154] has begun embedding these models directly into the driving loop, as illustrated in Fig. 1 (b).

Early efforts including GPT-Driver [87], which demonstrates that frozen VLMs can process multi-view images and textual prompts to generate trajectory plans or low-level control tokens while simultaneously producing human-readable rationales. While effective for commonsense reasoning and corner case understanding, integrating large foundation models into driving systems presents several drawbacks: poor spatial awareness [37], ambiguous numerical outputs [58], and elevated planning latency [119]. Furthermore, hallucination effects prevalent in LLMs/VLMs [136] expose driving systems to potentially unsafe control signals.

Follow-up research addresses these limitations through several key directions: (1) Spatial Understanding Enhancement: TOKEN [118] and WKAD [150] use object-centric token representations, while BEVDriver [134] integrates BEV features with language for 3D spatial queries and multimodal future predictions. Sce2DriveX [159] proposes spatial relationship graphs to model interactions between ego vehicle and traffic participants, and MPDrive [158] leverages visual prompting to strengthen spatial reasoning. (2) Latency Reduction: Dual-system architectures [15, 24, 25, 58, 81, 86, 90, 91, 100, 119] employ VLMs as intermediate modules providing feedback or auxiliary signals to end-to-end systems. Knowledge distillation approaches [38, 78, 96, 140] transfer VLM capabilities to traditional systems offline. (3) Hallucination Mitigation: In-context learning methods like Dilu [131] and its extension [60] utilize memory banks to store critical driving information. ReasonPlan [79] generate step-by-step decision justifications. AgentDriver [88] and AgentThink [99] implement tool-augmented chain-of-thought prompting to enhance reasoning capabilities.

Despite these advances, current methods remain predominantly perception-centric: their generated plans lack tight integration with closed-loop control[110], and their explanatory outputs provide no formal safety guarantees[164]. Besides, how to align VLM's outputs with the action space is also a challenge.

2.4. From VLM to VLA for Autonomous Driving

Inspired by recent progress in the embodied intelligence field [4, 65, 76], aligning vision, language, and action within a unified framework has become a growing trend in autonomous driving. As shown in Fig. 1 (c), VLA addresses the aforementioned gap by incorporating an explicit action head, thereby unifying perception, reasoning, and control within a single policy [85, 108, 166]. VLA policies are driven by three core demands in real-world driving: (i) robust reasoning in rare and long-tail scenarios [55, 128]; (ii) noise-tolerant control under dynamic and partially occluded conditions; and (iii) the ability to interpret spontaneous, high-level language commands (e.g., "overtake the truck") [96, 141]. By leveraging foundation models pretrained on internet-scale visual and linguistic data [94], VLA models demonstrate strong generalization across domains and benchmarks [8, 116, 133]. Concretely, modern VLA models can: (i) ground free-form instructions within ego-centric visual contexts and generate corresponding trajectories [30, 148]; (ii) produce Chain-of-Thought (CoT) justifications, as seen in DriveCoT and CoT-VLA [23, 124], to enhance interpretability; and (iii) move beyond direct control tokens to incorporate advanced planning modules, including diffusion-based heads [57], hierarchical CoT-based planners [30], and hybrid discrete–continuous control strategies.

Recent exemplar systems highlight the breadth of VLA capabilities[17, 30, 57, 153], exemplifing the new VLA4AD paradigm: they jointly reason over vision, language, and action, combining textual and trajectory outputs, long-horizon memory, symbolic safety checks, and multi-modal diffusion planning. These advances represent a decisive shift from perception-centric VLM pipelines toward action-aware, explainable, and instruction-following multimodal agents—paving the way for safer, more generalizable and human-aligned autonomous driving.

3. Architecture Paradigm of VLA4AD

This section presents the basic architecture of VLA4AD, consolidating its multimodal interfaces, core components, and outputs, as illustrated in Fig. 2.

3.1. Multimodal Inputs and Language Commands

VLA4AD rely on rich multimodal sensor streams and linguistic inputs to capture both the external environment and the driver's high-level intent.

Visual Data. Humans rely heavily on visual input to navigate complex driving environments, so do autonomous systems. In early approaches, single front-facing monocular cameras were the standard visual modality[63, 89, 146]. Over time, to improve spatial coverage and safety, systems evolved to include stereo cameras, multi-camera setups, and eventually full surround-view systems[8, 116]. This richer visual input enables more robust scene understanding and multi-object reasoning. Raw images can be processed directly or transformed into structured intermediate representations, such as bird's-eye-view (BEV) maps that facilitate spatial reasoning [21, 139]. Recent work further explores the trade-off between input resolution and model efficiency, dynamically adjusting granularity for real-time or long-tail cases [167].

Other Sensor Data. Beyond vision, autonomous vehicles have increasingly leveraged diverse sensor modalities to complement and ground perception to enhance spatial capabilities. Initial systems integrated LiDAR for precise 3D structure, later combining it with RADAR for velocity estimation and IMUs for motion tracking. GPS modules provide global localization [8, 133]. The field has also seen increasing attention to proprioceptive data, such as steering angle, throttle, and acceleration, particularly for behavior prediction and closed-loop control[20, 134, 142]. This progression—from geometry to dynamics—has driven research into more sophisticated sensor fusion frameworks [5, 10, 67, 129], aiming to create a unified spatial-temporal representation of the environment.

Language Inputs. Natural language inputs—such as commands, queries, and structured descriptions—have become increasingly important in VLA4AD. Early research focused on direct navigation commands (e.g., "Turn left at the next intersection," "Stop behind the red car") to enable basic instruction following [96, 105]. As systems matured, environmental queries emerged, allowing users or agents to ask questions like "Is it safe to change lanes now?" or "What is the speed limit here?" [51, 93], enabling interactive situational awareness. Further advancements introduced task-level linguistic specifications, such as interpret-

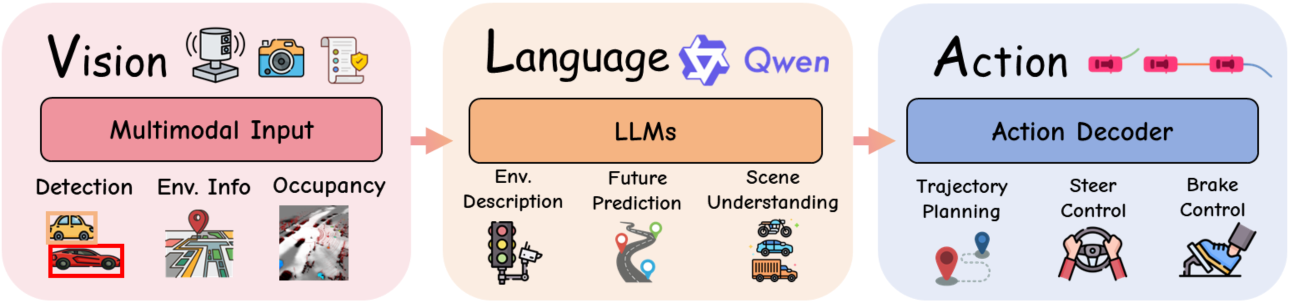

Figure 2. Overview of the VLA4AD Architecture.

ing traffic rules, parsing high-level goals, or understanding map-based constraints expressed in natural language [33]. More recent efforts have pushed toward multi-turn dialogs, reasoning chains (e.g., Chain-of-Thought prompting)[50, 119], and tool-augmented language interfaces[41, 88, 99], which support richer forms of reasoning and alignment with human decision-making processes.

Finally, recent work has also started incorporating spoken language as a more natural and embodied input modality, bridging perception and interaction via speech-driven interfaces [148, 165]. This progression from static instructions to dialog-driven, multi-step reasoning reflects a broader trend: using language not just to command the vehicle, but to enable interpretable and collaborative autonomy.

3.2. Core Architectural Modules

The fundamental architecture of a VLA4AD integrates visual perception, language understanding, and action generation in a cohesive pipeline.

Vision Encoder. Raw imagery and sensor data are converted to latent representations using large self-supervised backbones such as DINOv2 [94], ConvNeXt-V2 [135], or CLIP [102]. Many VLA systems employ BEV projection [119], and others incorporate 3D priors via point-cloud encoders (e.g., PointVLA [67]) or voxel modules (3D-VLA [161]). Multi-scale fusion using language-derived keys improves grounding at fine spatial levels [101, 152].

Language Processor. Natural language is processed using pretrained decoders such as LLaMA2 [121] or GPT-style transformers [6]. Instruction-tuned variants (e.g., Visual Instruction Tuning [74]) and retrieval-augmented prompting (RAG-Driver [148]) inject domain knowledge. Lightweight fine-tuning strategies such as LoRA [42] enable efficient adaptation.

Action Decoder. Downstream control is emitted via (i) Autoregressive tokenizers where discrete actions or trajectory way-points are predicted sequentially [48, 64, 98, 168], (ii) Diffusion heads that sample continuous controls conditioned on fused embeddings (DiffVLA [57]; Diffusion-VLA [130]), or (iii) Flow-matching / policy gradient experts used by GRPO [113] or DPO[103] fine-tuning pipelines [70, 137]. Hierarchical controllers (e.g., ORION [30]) let a language planner dispatch sub-goal sketches to a separate low-level PID or MPC stack.

3.3. Driving Outputs

The output modality of a VLA model reflects its level of abstraction and operational goal. Over time, output formats have evolved from low-level control commands to higher-level spatial reasoning and skill-conditioned actions.

Low-Level Actions. Early VLA4AD systems typically focused on directly predicting raw control signals such as steering angles, throttle, and braking. These actions are often modeled either as continuous outputs or as discrete action tokens, suitable for integration with PID or end-to-end control pipelines [30, 98, 143, 165]. While this formulation allows for fine-grained control, it is often sensitive to small perception errors and lacks long-horizon planning capacity.

Trajectory Planning. Subsequent research has shifted towards trajectory- or waypoint-level predictions, which offer a more stable and interpretable intermediate representation. These trajectories, often expressed in BEV or egocentric coordinates, can be flexibly executed via model predictive control (MPC) or other downstream planners [3, 44, 57, 59, 96, 155]. This formulation allows VLA models to reason over longer time horizons and integrate multimodal context more effectively.

Together, these output formats illustrate the evolving ambition of VLA4AD systems: not only to drive, but to do so robustly, explainably, and contextually. In summary, a typical VLA4AD model takes multimodal sensor data and natural language input as context, and produces both driving decisions (at various abstraction levels) and, in some cases, language-grounded explanations.

4. Progress of VLA4AD Paradigm

Research on the VLA4AD has moved in discernible waves, each propelled by the limitations of its predecessor and by the arrival of new cross-modal pre-training techniques. As shown in Fig. 3, in what follows, we trace four successive stages: *Explanatory Language Models*, *Modular VLA4AD*, *End-to-end VLA4AD*, and *Reasoning-centric VLA4AD*.

Table 1 summarizes representative VLA4AD models from 2023–2025, highlighting their input modalities, how they incorporate language, the form of action output, the data or environment used for evaluation, and their core contributions.

4.1. Pre-VLA: Language Model as Explainer

The earliest forays integrated language in a passive, descriptive role to enhance interpretability. A typical pipeline in this stage employed a frozen vision model (e.g. CLIP[102]) with an LLM decoder to explain the driving scene or recommended action in natural language, without directly outputting control. For example, DriveGPT-4 [141] would take a single front-camera image and produce either a textual description or a high-level maneuver label ("slow down", "turn left"). These outputs helped explain what the perception system saw or intended, improving transparency. However, the actual vehicle control was still handled by conventional modules (PID controllers, etc.), so the language was an overlay rather than integral to decision-making. Moreover, two issues became apparent: (i) Generating long descriptions for each frame introduced latency, as vision encoders processed thousands of tokens per image[168]; (ii) General-purpose visual encoders wasted effort on irrelevant details, since not everything in an image is pertinent to driving[150]. Researchers responded with optimizations like TS-VLM [11], which uses text-guided soft attention pooling to focus on key regions, and DynRsl-VLM [167], which dynamically adjusts input resolution to balance speed and detail. These improved efficiency, but there remained a semantic gap - narrating or labeling the scene is not the same as generating a precise steering or braking command. Closing that gap was the next logical step.

4.2. Modular VLA Models for AD

As VLA research progressed, language evolved from a passive scene descriptor to an active planning component within modular architectures. Rather than merely commenting on the environment, language inputs and outputs began to inform planning decisions directly [58]. For example, OpenDriveVLA [165] fused camera and LiDAR inputs with textual route instructions (e.g., "turn right at the church"), generating intermediate, human-readable waypoints (e.g., "turn right in 20m, then go straight"), which were then converted into continuous trajectories. This approach enhanced the transparency and flexibility of the planning process by introducing interpretable linguistic representations.

CoVLA-Agent [17] integrated visual and LiDAR tokens with optional textual prompts and used a compact MLP to map a selected action token (e.g., "turn left") to a corresponding trajectory. Similarly, DriveMoE [143] employed a Mixture-of-Experts architecture in which language cues were used to dynamically select sub-planners, such as an ”overtaking expert” or ”stop-and-go expert,” based on the context. In multi-agent scenarios, LangCoop [33] showed that vehicles could communicate using concise natural language messages to coordinate at intersections, representing a step toward language-enabled cooperation. SafeAuto [153] incorporated symbolic traffic rules expressed in formal logic to validate or veto language-driven plans, ensuring that generated behaviors remained within safety constraints. Additionally, RAG-Driver [148] introduced a retrieval-augmented planning mechanism, retrieving similar past driving cases from a memory bank to guide decision-making in ambiguous or long-tail scenarios. Collectively, these approaches significantly reduced the semantic gap between language instructions and vehicle actions, effectively embedding natural language into the core of the planning loop. However, they often relied on multi-stage pipelines (perception → language → plan → control), which introduced latency and cascading failure risks at each module boundary. These limitations have motivated recent interest in more unified, end-to-end architectures, which aim to integrate perception, language understanding, and action generation within a single differentiable system.

4.3. Unified End-to-End VLA Models for AD

With large multimodal foundation models available, researchers moved to fully unified networks that map sensors (and optional text commands) directly to trajectories or control signals in a single forward pass. A prime example is EMMA [50], which trains a massive VLM on Waymo's autonomous driving data to jointly perform object detection and motion planning; the model learns a shared representation that serves both tasks, achieving better closed-loop performance than separate components. SimLingo[104], LMDrive[110] and CarLLaVA[105] built on a LLaVA and fine-tuned it in CARLA simulator to follow language instructions and drive, notably introducing an "action dreaming" technique where the model imagines diverse outcomes for a given scenario by varying the language instruction, thus forcing a tight coupling between linguistic commands and the resulting trajectories. Other innovative approaches include using generative video models: ADriver-I [53] learned a latent world model that predicts future camera frames given actions (using diffusion), thereby enabling planning via imagining the consequences of actions. DiffVLA [57] combined sparse (waypoints) and dense (occu-

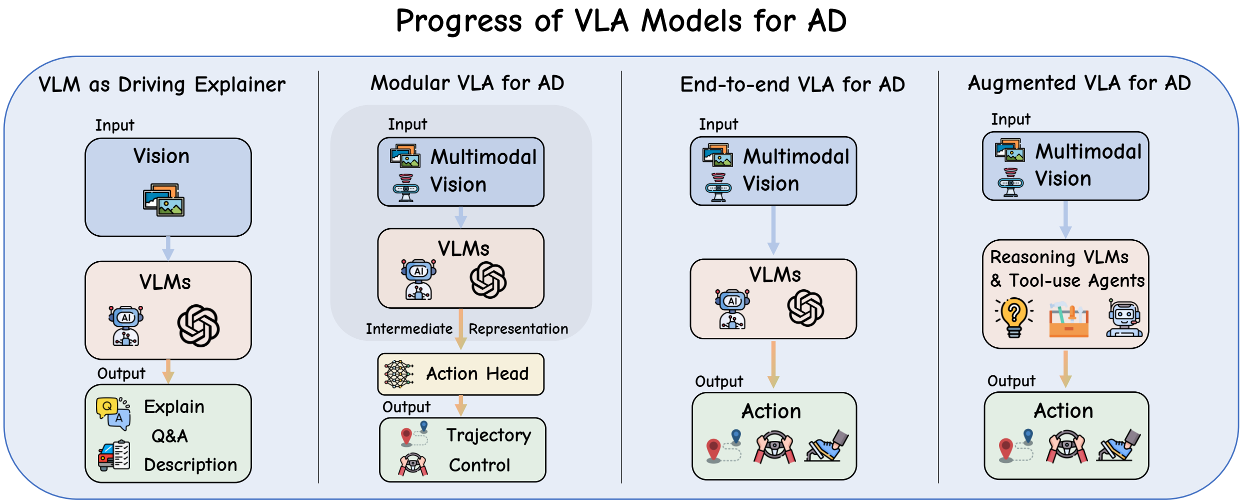

Figure 3. Evolution of VLA models for AD. From left to right: (1) VLM-as-explainer: a frozen LLM narrates the driving scene but produces no control. (2) Modular VLA: language is transformed into an intermediate representation that an action head converts into trajectory or low-level control. (3) End-to-end VLA: a single multimodal pipeline maps sensor input directly to actions. (4) Augmented VLA: tool-using or CoT VLMs add long-horizon reasoning while retaining the end-to-end control pathway.

pancy grid) diffusion predictions to generate a trajectory conditioned on a textual scenario description, effectively sampling from a distribution of plausible safe maneuvers. End-to-end VLA models are highly reactive and effective at sensorimotor mapping, but a new limitation became clear: they can still struggle with long-horizon reasoning (planning far ahead or considering complex contingencies) and with providing fine-grained explanations of their decisions.

4.4. Reasoning-Augmented VLA Models for AD

The latest wave of VLA4AD moves beyond explaining and plan conditioning toward *long-horizon reasoning, memory, and interactivity*, placing VLMs/LLMs at the center of the control loop. ORION [30] couples a transformer "QT-Former" memory, storing several minutes of observations and actions, with an LLM that summarizes this history and outputs the next trajectory and a corresponding natural language explanation. Impromptu VLA [18] instead aligns CoT with action. Trained on 80k corner-case clips whose correct reasoning steps are annotated, the model learns to verbalise its decision path before acting, delivering state-of-the-art zero-shot negotiation between vehicles. AutoVLA [168] fuses CoT reasoning and trajectory planning within a single autoregressive transformer that tokenises continuous paths into discrete drive tokens, delivering state-of-the-art closed-loop success rates on nuPlan and CARLA. Collectively, these systems no longer just react to sensor input; they *explain*, *anticipate*, and carry out long-horizon *reasoning* before outputing actions. They point toward conversational AVs that can verbally justify actions in real time, yet they surface new challenges: indexing city-scale memories, keeping LLM reasoning within a 30 Hz control loop, and formally verifying language-conditioned policies.

In summary, VLA4AD models have evolved from using language as a passive explanatory tool to integrating it as an active component in perception and control. We observe a steady closing of the loop between seeing, speaking, and acting—starting from explanatory perceptions, to modular VLA planning, to fully unified pipelines with reasoning and dialogue. This progression points toward autonomous vehicles as conversational, collaborative agents—capable of not only safe driving, but also communication and reasoning aligned with human expectations.

5. Datasets and Benchmarks

We review several key datasets and evaluation suites, summarized in Table 2.

BDD 100K / BDD-X [63, 147]. BDD 100K offers **100 k** diverse US videos; the BDD-X subset (∼7 k clips) adds time-aligned human *rationales* (e.g., "slows because pedestrian crossing"), providing ground-truth explanations for models such as CoVLA-Agent [17] and SafeAuto [153].

nuScenes [7]. **1k** 20s real-world episodes (Boston, Singapore) with 6 cams, LiDAR + radar and full 3D labels. Although language-free, it has been used for extensive VLA4AD evaluations.

Bench2Drive [55]. A closed-loop CARLA benchmark with **44** scenario types (220 routes) and a 2M-frame training set. Metrics isolate specific skills (unprotected turns, cut-ins, etc.); DriveMoE [143] tops the Leaderboard via specialized experts.

Table 1. **Representative VLA4AD Models (2023–2025).** Sensor Inputs: Single = single forward-facing camera input; Multi = multi-view camera input; State = vehicle state information & other sensor input. Outputs: **LLC**= low-level control, **Traj.**= future trajectory, **Multi.**= multiple tasks such as perception, prediction or planning.

		Data Source		**Model**			
Model	**Year**	**Input**	**Dataset / Benchmark**	**Vision**	**LLM**	**Output**	**Focus**
DriveGPT-4 [141]	2023	Single	BDD-X	CLIP	LLaMA-2	**LLC**	Interpretable LLM Mixed Fine-tuning
ADriver-I [53]	2023	Single	nuScenes + Private	CLIP ViT	Vicuna-1.5	**LLC**	Diffusion World Model Vision–action Tokens
RAG-Driver [148]	2024	Multi	BDD-X	CLIP ViT	Vicuna-1.5	**LLC**	RAG Control Textual Rationales
EMMA [50]	2024	Multi + State	Waymo fleet	Gemini-VLM	Gemini	**Multi.**	MLLM Backbone Multi-task Outputs
CoVLA-Agent [17]	2024	Single + State	CoVLA Data	CLIP ViT	Vicuna-1.5	**Traj.**	Text + Traj Outputs Auto-labelled Data
OpenDriveVLA [165]	2025	Multi	nuScenes	Custom Module	Qwen-2.5	**LLC+Traj.**	2-D/3-D Align SOTA Planner
ORION [31]	2025	Multi + History	nuScenes + CARLA	QT-Former	Vicuna-1.5	**Traj.**	CoT Reasoning Continuous Actions
DriveMoE [143]	2025	Multi	Bench2Drive	Paligemma-3B	–	**LLC**	Mixture-of-Experts Dynamic Routing
VaViM [3]	2025	Video Frames	BDD100K + CARLA	LlamaGen	GPT-2	**Traj.**	Video-token PT Vision to Action
DiffVLA [57]	2025	Multi + State	Navsim-v2	CLIP ViT	Vicuna-1.5	**Traj.**	Mixed Diffusion VLM Sampling
LangCoop [33]	2025	Single + V2V	CARLA	GPT-4o	GPT-4o	**LLC**	Language-based V2V High Bandwidth Cut
SimLingo [105]	2025	Multi	CARLA + Bench2Drive	InternVL2	Qwen-2	**LLC+Traj.**	Enhanced VLM Action-dreaming
SafeAuto [153]	2025	Multi + State	BDD-X + DriveLM	CLIP ViT	Vicuna-1.5	**LLC**	Traffic-Rule-Based PDCE Loss
Impromptu-VLA [18]	2025	Single	Impromptu Data	Qwen-2.5VL	Qwen-2.5VL	**Traj.**	Corner-case QA NeuroNCAP SOTA
AutoVLA [168]	2025	Multi + State	nuScenes + CARLA	Qwen-2.5VL	Qwen-2.5VL	**LLC+Traj.**	Adaptive Reasoning Multi Benchmark

Reason2Drive [93]. **600 k** video–text pairs (nuScenes, Waymo, ONCE) annotated with CoT QA spanning perception → prediction → action. It evaluates logical consistency across entire reasoning chain using a *consistency* metric to penalize incoherent multi-step answers.

DriveLM-Data [115]. Provides graph-structured QA on nuScenes (18 k graphs) and CARLA (16 k) scenes, stressing conditional reasoning. Baseline TS-VLM [11] attains BLEU-4 56 but low graph consistency, leaving ample room for improved multi-step reasoning.

Impromptu VLA [18]. **80k** 30s clips (∼2M frames) mined from eight public sets, curated for corner-case traffic (dense crowds, ambulances, adverse weather). Each clip pairs an expert trajectory and high-level instruction with rich captions and time-stamped QA. Provides an open evaluation server; training on this corpus yields measurable safety gains in closed-loop tests.

NuInteract [160]. Extends nuScenes with **1k** multi-view scenes that contain dense captions and multi-turn 3D QA pairs, tightly linked to LiDAR ground truth. Supports multi-camera VQA and 3D reasoning; DriveMonkey shows substantial gains in cross-view QA when trained on this set.

Table 2. **Notable Datasets & Benchmarks for VLA4AD.** Real = real-world logs; Sim = CARLA simulation[26]; clips = video segments; QA = question–answer pairs; CoT = chain-of-thought.

Name	Year	Domain	Scale	Modalities	Tasks
BDD100K / BDD-X [63, 147]	2018	Real (US)	100 k videos; 7 k clips	RGB video	Captioning, QA
nuScenes [7]	2020	Real (Boston/SG)	1 k scenes (20 s, 6 cams)	RGB, LiDAR, Radar	Detection, QA
Bench2Drive [55]	2024	Sim (CARLA)	220 routes; 44 scenarios	RGB	Closed-loop control
Reason2Drive [93]	2024	Real (nuSc/Waymo)	600 k video–QA	RGB video	CoT-style QA
DriveLM-Data [115]	2024	Real+Sim	18 k scene graphs	RGB, Graph	Graph QA
Impromptu VLA [18]	2025	Real (multi-src)	80 k clips (30 s)	RGB video, State	QA, Trajectory
NuInteract [160]	2025	Real (nuScenes)	1 k scenes	RGB, LiDAR	Multi-turn QA
DriveAction [39]	2025	Real (fleet)	2.6 k scenarios; 16.2 k QA	RGB video	High-level QA

DriveAction [39]. A user-contributed, real-world benchmark containing **2.6 k** driving scenarios and **16.2 k** vision-language QA pairs with *action-level* labels. The dataset spans broad, in-the-wild situations and offers evaluation protocols that score VLA models on human-preferred driving decisions, filling the gap left by perception-only suites.

In short, the datasets span the full spectrum needed for VLA4AD research: BDD-X [63] and nuScenes [7] deliver large-scale, sensor-rich realism; Bench2Drive [55] and Impromptu VLA [18] inject safety-critical corner cases; and Reason2Drive [93], DriveLM [115], NuInteract [160], and DriveAction [39] supply the structured language needed for fine-grained reasoning and human-aligned actions. Harnessing these complementary assets is essential for training and benchmarking the next generation of VLA4AD.

6. Training and Evaluation Strategies

Building a VLA4AD policy involves two coupled goals: (i) learning a safe and competent *driving controller* and (ii) retaining a faithful *language interface*. Because driving data are costly and risky to collect, most work adopts a *pre-train* → *fine-tune* pipeline: behaviour cloning on large logs, followed by targeted refinement in simulation or with rule–based constraints. Below we review the two dominant paradigms—**Imitation Learning** and **Reinforcement Learning**—and highlight how recent systems combine them.

6.1. Training Paradigms

Supervised Imitation Learning (IL). IL remains the work-horse for VLA4AD: the network ingests sensor streams (and, if present, a language prompt) and minimises an ℓ_2 or cross-entropy loss to reproduce the expert's control or trajectory. CoVLA-Agent [17] learns both a future path and a scene caption per frame, while DriveMoE [143] and models trained on the SimLingo corpus (CarLLAVA) [105] clone millions of simulator demonstrations. Although IL scales easily, it mirrors the training distribution; rare hazards (e.g. cut-ins, occluded pedestrians) receive little supervision. Typical remedies are *DAgger-style* noisy roll-outs or explicit corner-case augmentation, yet distribution drift can still cascade when perception or language grounding fails.

Reinforcement Learning (RL). RL is usually layered on top of an IL warm-start. The policy interacts with a simulator (CARLA, Bench2Drive, etc.) and is optimised with PPO or DQN style updates for route progress, collision avoidance, and traffic-rule compliance. Multi-agent settings such as LangCoop [33] use RL to refine V2V coordination, while SafeAuto [153] embeds logical traffic rules as hard constraints or additional penalties. A key open question is how to blend *driving rewards* with *language fidelity*: current work often sidesteps the issue by freezing the LLM and penalising only unsafe actions, leaving joint gradients over text and control largely unexplored. As a result, RL for VLA is promising—especially for edge-case robustness—but still under-developed compared with pure IL.

Multi-stage Training. Most VLA4AD models are trained via a four-stage curriculum: (1) *Pre-train* large vision encoders (e.g., CLIP, InternViT) and language models (e.g., LLaMA, Vicuna) on broad image–text corpora and video datasets to learn general visual and linguistic priors; (2) *Align* modalities by fine-tuning on paired image–text–action data—such as DriveMonkey on NuInteract [160]—using cross-modal contrastive losses and sequence modeling objectives to bind scene features, language prompts, and control tokens; (3) *Targeted augmentation* injects specialized traffic scenarios and instructions (e.g., SimLingo's corner-case clips [105], Bench2Drive's challenging routes [55]), often supplemented with reinforcement learning or rule-based penalties (as in SafeAuto [153]) to enforce safety constraints and improve performance on rare events; (4) *Compress* the resulting model for deployment through parameter-efficient methods—LoRA adapters, sparse Mixture-of-Experts routing, or teacher–student distillation—reducing compute, memory, and latency while preserving the aligned VLA capabilities, as exemplified by DriveMoE [143] and TS-VLM [11].

Balancing Language and Control. Joint losses typically weight a trajectory term against a caption or QA term (e.g. CoVLA-Agent uses $\mathcal{L} = \mathcal{L}_{\text{traj}} + \lambda\mathcal{L}_{\text{cap}}$). Some authors alternate updates—one batch for driving, the next for language—to avoid gradient interference. Large LLMs are often kept frozen and prompted via a light adapter; only the prompt-encoder is trained, preserving linguistic fluency without huge GPU cost. Free-form explanations complicate supervision: CIDEr-style or RL-based caption tuning has been explored, but care is required to reward factual accuracy over rhetoric.

Scalability and Efficiency. End-to-end VLA stacks (vision transformer + LLM + planner) can exceed hundreds of GFLOPs per frame. Current work therefore relies on: *LoRA and adapters* to update a few million parameters inside a 70 B LLM (SafeAuto [153]); *Mixture-of-Experts* routing so only a subset of specialists run at inference (DriveMoE [143]); *lightweight token reduction*—TS-VLM reports a $10\times$ speed-up by soft-attentive pooling [167]; *event-driven scheduling* that invokes the heavy model only on scene changes; and *distillation* (still rare in publications) to compress a cluster-scale policy into an embedded "tiny VLA". A typical pipeline now freezes CLIP and LLaMA, trains a small cross-attention head by imitation, LoRA-adapts the LLM on language-augmented data, optionally applies RL for red-light penalties, and finally distils the result for on-car deployment.

6.2. Evaluation Protocols

Evaluating a VLA4AD agent is a *dual-objective* task: the policy must *drive safely* **and** *communicate faithfully*. State-of-the-art papers therefore report four complementary metric pillars:

Closed-loop Driving *Route success* on CARLA / Bench2Drive [55]; *infractions* (collisions, red-light, off-road) where DiffVLA halves error via a PDMS layer [57]; *rule compliance* checked by logic vetoes (SafeAuto [153]); and *generalisation* to unseen towns and weather.

Open-loop Prediction *Trajectory* ℓ_2 and collision rate (nuScenes challenge); *goal reach* on instruction-conditioned targets; optional *mAP/IoU* for auxiliary perception heads; and *latency/FPS*—TS-VLM cuts compute by $\sim 90\%$ through token pooling [167].

Language Competence *Command follow* in SimLingo's Action-Dreaming benchmark [105]; automatic *BLEU / CIDEr / accuracy* on NuInteract [160] and DriveLM (BLEU-4 56 for TS-VLM [115]); *reason-chain consistency* in Reason2Drive [93]; and *human ratings* of BDD-X-style rationales.

Robustness & Stress *Sensor perturbations* (blur, dropout, lag) analysed by DynRsl-VLM [167]; *adversarial prompts* or patches; *out-of-distribution events*; and *language edge-cases* (idioms, code-switching, multilingual queries).

Overall, a credible evaluation must measure (i) control reliability, (ii) language fidelity, and (iii) their coupling. Today's suites cover these facets in isolation—CARLA / Bench2Drive for control, NuInteract / Reason2Drive / DriveLM for reasoning—highlighting the need for a unified *"AI driver's licence"* that fuses both streams.

7. Open Challenges

Despite rapid progress, Vision–Language–Action systems still face substantial barriers before large–scale real–world deployment. Below we summarise the six most pressing research fronts.

Robustness & Reliability. Language reasoning adds context but opens new failure modes: LLMs may hallucinate hazards or mis-parse slang ("floor it"). Models must remain stable under sensor corruption (rain, snow, glare) *and* linguistic noise. Logic–based safety vetoes such as SafeAuto [153] are a first step, yet formal verification and "socially-compliant" driving policies remain largely unexplored.

Real-time Performance. Running a vision transformer plus LLM at $\geq$30 Hz on automotive hardware is non-trivial. Token–reduction designs like TS-VLM [167], hardware–aware quantisation, and event-triggered reasoning (where heavy modules activate only on novel situations) are promising; distillation or MoE sparsity will be required as model sizes scale to billions of parameters.

Data & Annotation Bottlenecks. Tri-modal supervision (image + control + language) is scarce and costly—Impromptu VLA required 80k manually labelled clips. Synthetic augmentation (e.g. SimLingo) helps, but coverage of non-English dialects, traffic slang, and legally binding phrasings is still thin.

Multimodal Alignment. Current VLA4AD work is camera-centric; LiDAR, radar, HD-maps, and temporal state are only partially fused. Approaches range from BEV projection of point-clouds to 3-D token adapters, and from ORION's language summarisation of long histories [31] to retrieval of textual map rules as in RAG-Driver [148]. A principled, temporally consistent fusion of heterogeneous modalities is still missing.

Multi-agent Social Complexity. Scaling from pairwise coordination to dense traffic raises protocol, trust, and security issues. How should AVs exchange intent in a constrained yet flexible "traffic language"? Authentication and robustness to malicious messages are open problems; cryptographic V2V and gesture-to-text grounding are early research threads.

Domain Adaptation & Evaluation. Sim-to-real transfer, cross-region generalisation, and continual learning without catastrophic forgetting are unresolved. Community benchmarks (e.g. Bench2Drive) cover only a fraction of the long-tail. A regulatory "AI driver's test" that scores both control and explanation quality is still to be defined.

In summary, addressing these challenges demands joint advances in scalable training, formal safety analysis, human–computer interaction, and policy. Progress on any one front—robust perception, efficient LLMs, trusted V2V, or standardised evaluation—will accelerate the path toward safe, transparent, and globally deployable VLA4AD systems.

8. Future Directions

The next wave of research will likely widen the scope of VLA4AD from prototype policies to *scalable, cooperative, and verifiable* driving platforms. We highlight five promising threads.

Foundation–scale Driving Models. An obvious trajectory is a GPT-style "driving backbone": a self-supervised, multi-sensor model trained on dash-cams, LiDAR sweeps, HD-maps, and textual road rules. Such a model could be prompted or LoRA-adapted for downstream tasks with little data, similar to the way SimLingo/CarLLAVA leverages instruction-conditioned trajectories [105]. Realising this vision demands masked multimodal objectives and architectures that process panoramic video together with free text.

Neuro-symbolic Safety Kernels. Pure end-to-end nets struggle to *guarantee* safety. Recent hybrids add rule layers—e.g. SafeAuto inserts logical traffic checks [153]. Future work may let a neural VLA stack output a structured action program (or CoT plan) that a symbolic verifier executes, bridging flexibility and certifiability; ORION's language memory hints at such an interface [31].

Fleet-scale Continual Learning. Deployed AVs will encounter novel hazards daily. Instead of raw logs, cars could uplink concise language snippets ("new flagger pattern at x, y") which are aggregated into curriculum updates, much as SimLingo filters trivial scenes to emphasise rare ones [105]. Cloud agents could even answer real-time queries from uncertain vehicles, boot-strapping knowledge across the fleet.

Standardised Traffic Language. Wide-area coordination will require a constrained, ontology-driven message set—"I-yield-to-you", "Obstacle-ahead", etc.—analogous to ICAO phraseology in aviation. VLA models are natural translators from raw perception to such canonical intents; MoE routing (DriveMoE [143]) or token-reduction LMs (TS-VLM [167]) can keep the bandwidth low enough for V2V links.

Cross-modal Social Intelligence. Future systems must parse gestures, voice, and signage as part of the "language" channel—e.g. recognising a police hand-signal or a pedestrian wave, then producing an explicit, human-readable response (lights, display, honk). Retrieval-augmented planners such as RAG-Driver [148] suggest one route: fuse live perception with symbolic rules and context to ground non-verbal cues. Extending this to robust gesture–language–action alignment remains open.

In essence, achieving these goals will require progress in large-scale multimodal learning, formal verification, communication standards, and human-AI interaction. Success would yield a versatile "driving brain" that can be quickly adapted, safely audited, and seamlessly integrated into the global traffic ecosystem.

9. Conclusion

In this work, we present the **first** comprehensive survey of Vision–Language–Action models for Autonomous Driving (VLA4AD), unifying several representative methods under a concise taxonomy that captures input modalities, core architectural components, and output formats. We trace the evolution of VLA4AD through four successive waves—*Pre-VLA Explainers*, *Modular VLA4AD*, *End-to-End VLA4AD*, and *Reasoning-Augmented VLA4AD*—highlighting how each stage has progressively closed the loop between perception, language understanding, and control.

We provide an in-depth comparison of training paradigms, from large-scale pre-training and modality alignment to targeted augmentation with corner-case data and efficient compression techniques, illustrating how these multi-stage workflows yield models that are both expressive and deployable. Our review of datasets and benchmark underscores the critical role of rich, multi-sensor, and language-grounded corpora in advancing VLA4AD capabilities.

Despite rapid progress, significant challenges remain: ensuring sub-30 Hz reasoning throughput, formal verification of language-conditioned policies, robust generalization in long-tail scenarios, and seamless sim-to-real transfer. We argue that the community must converge on shared evaluation protocols and open-source toolkits, invest in scalable memory and causal reasoning backbones, and pursue continual, fleet-scale learning to bridge research prototypes and production systems. By distilling current achievements and charting open avenues, we aim to inspire future work toward transparent, instruction-following, and socially aligned autonomous vehicles powered by integrated vision, language, and action.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[2] Florent Altché and Arnaud de La Fortelle. An lstm network for highway trajectory prediction. In *2017 IEEE 20th international conference on intelligent transportation systems (ITSC)*, pages 353–359. IEEE, 2017. 2

[3] Florent Bartoccioni, Elias Ramzi, Victor Besnier, Shashanka Venkataramanan, Tuan-Hung Vu, Yihong Xu, Loick Chambon, Spyros Gidaris, Serkan Odabas, David Hurych, et al. Vavim and vavam: Autonomous driving through video generative modeling. *arXiv preprint arXiv:2502.15672*, 2025. 5, 8

[4] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 4

[5] Joost Broekens, Bernhard Hilpert, Suzan Verberne, Kim Baraka, Patrick Gebhard, and Aske Plaat. Fine-grained affective processing capabilities emerging from large language models. In *2023 11th international conference on affective computing and intelligent interaction (ACII)*, pages 1–8. IEEE, 2023. 4

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 5

[7] Holger Caesar, Alex Bankiti, Orien Lang, et al. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2, 7, 9

[8] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 4

[9] Dian Chen and Philipp Krähenbühl. Learning from all vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17222–17231, 2022. 2

[10] Dianwei Chen, Zifan Zhang, Yuchen Liu, and Xianfeng Terry Yang. Insight: Enhancing autonomous driving safety through vision-language models on context-aware hazard detection and edge case evaluation. *arXiv e-prints*, pages arXiv–2502, 2025. 4

[11] Lihong Chen, Hossein Hassani, and Soodeh Nikan. Ts-vlm: Text-guided softsort pooling for vision-language models in multi-view driving reasoning. *arXiv preprint arXiv:2505.12670*, 2025. 2, 6, 8, 10

[12] Long Chen, Lukas Platinsky, Stefanie Speichert, Błażej Osiński, Oliver Scheel, Yawei Ye, Hugo Grimmett, Luca

Del Pero, and Peter Ondruska. What data do we need for training an av motion planner? In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1066–1072. IEEE, 2021. 2

[13] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[14] Shaoyu Chen, Bo Jiang, Hao Gao, Bencheng Liao, Qing Xu, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*, 2024. 3

[15] Yuan Chen, Zi-han Ding, Ziqin Wang, Yan Wang, Lijun Zhang, and Si Liu. Asynchronous large language model enhanced planner for autonomous driving. In *European Conference on Computer Vision*, pages 22–38. Springer, 2024. 3

[16] Zhili Chen, Maosheng Ye, Shuangjie Xu, Tongyi Cao, and Qifeng Chen. Ppad: Iterative interactions of prediction and planning for end-to-end autonomous driving. In *European Conference on Computer Vision*, pages 239–256. Springer, 2024. 3

[17] Haohan Chi, Huan-ang Gao, Ziming Liu, et al. CoVLA: Comprehensive vision-language-action dataset for autonomous driving. In *WACV*, 2025. 1, 4, 6, 7, 8, 9

[18] Haohan Chi, Huan-ang Gao, Ziming Liu, et al. Impromptu vla: Open weights and open data for driving vision-language-action models. *arXiv preprint arXiv:2505.23757*, 2025. 1, 2, 7, 8, 9

[19] Kashyap Chitta, Aditya Prakash, and Andreas Geiger. Neat: Neural attention fields for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15793–15803, 2021. 2

[20] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):12878–12895, 2022. 2, 3, 4

[21] Tushar Choudhary, Vikrant Dewangan, Shivam Chandhok, Shubham Priyadarshan, Anushka Jain, Arun K Singh, Siddharth Srivastava, Krishna Murthy Jatavallabhula, and K Madhava Krishna. Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16345–16352. IEEE, 2024. 4

[22] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979, 2024. 2

[23] Yixin Cui, Haotian Lin, Shuo Yang, Yixiao Wang, Yanjun Huang, and Hong Chen. Chain-of-thought for autonomous driving: A comprehensive survey and future prospects. *arXiv preprint arXiv:2505.20223*, 2025. 2, 4

[24] Kairui Ding, Boyuan Chen, Yuchen Su, Huan-ang Gao, Bu Jin, Chonghao Sima, Wuqiang Zhang, Xiaohui Li, Paul Barsch, Hongyang Li, et al. Hint-ad: Holistically aligned interpretability in end-to-end autonomous driving. *arXiv preprint arXiv:2409.06702*, 2024. 3

[25] Simon Doll, Niklas Hanselmann, Lukas Schneider, Richard Schulz, Marius Cordts, Markus Enzweiler, and Hendrik Lensch. Dualad: Disentangling the dynamic and static world for end-to-end driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14728–14737, 2024. 3

[26] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 9

[27] Kaituo Feng, Changsheng Li, Dongchun Ren, Ye Yuan, and Guoren Wang. On the road to portability: Compressing end-to-end motion planner for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15099–15108, 2024. 2

[28] Yuchao Feng and Yuxiang Sun. Polarpoint-bev: Bird-eye-view perception in polar points for explainable end-to-end autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2024. 3

[29] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 910–919. IEEE, 2024. 3

[30] Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. *arXiv preprint arXiv:2503.19755*, 2025. 4, 5, 7

[31] Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. *arXiv preprint arXiv:2503.19755*, 2025. 8, 11

[32] Haoxiang Gao, Zhongruo Wang, Yaqian Li, Kaiwen Long, Ming Yang, and Yiqing Shen. A survey for foundation models in autonomous driving. *arXiv preprint arXiv:2402.01105*, 2024. 2

[33] Xiangbo Gao, Yuheng Wu, Rujia Wang, et al. Langcoop: Collaborative driving with language. *arXiv preprint arXiv:2504.13406*, 2025. 5, 6, 8, 9

[34] David González, Joshué Pérez, Vicente Milanés, and Fawzi Nashashibi. A review of motion planning techniques for automated vehicles. *IEEE Transactions on intelligent transportation systems*, 17(4):1135–1145, 2015. 2

[35] Ke Guo, Haochen Liu, Xiaojun Wu, Jia Pan, and Chen Lv. ipad: Iterative proposal-centric end-to-end autonomous driving. *arXiv preprint arXiv:2505.15111*, 2025. 3

[36] Mingzhe Guo, Zhipeng Zhang, Yuan He, Ke Wang, and Liping Jing. End-to-end autonomous driving without costly modularization and 3d manual annotation. *arXiv preprint arXiv:2406.17680*, 2024. 3

[37] Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Chenming Zhang, Shuai Liu, and Long Chen. Drivem-

llm: A benchmark for spatial understanding with multimodal large language models in autonomous driving. *arXiv preprint arXiv:2411.13112*, 2024. 3

[38] Wencheng Han, Dongqian Guo, Cheng-Zhong Xu, and Jianbing Shen. Dme-driver: Integrating human decision logic and 3d scene perception in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3347–3355, 2025. 4

[39] Yuhan Hao, Zhengning Li, Lei Sun, Weilong Wang, Naixin Yi, Sheng Song, Caihong Qin, Mofan Zhou, Yifei Zhan, Peng Jia, et al. Driveaction: A benchmark for exploring human-like driving decisions in vla models. *arXiv preprint arXiv:2506.05667*, 2025. 9

[40] Jeffrey Hawke, Richard Shen, Corina Gurau, Siddharth Sharma, Daniele Reda, Nikolay Nikolov, Przemysław Mazur, Sean Micklethwaite, Nicolas Griffiths, Amar Shah, et al. Urban driving with conditional imitation learning. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 251–257. IEEE, 2020. 2

[41] Xinmeng Hou, Wuqi Wang, Long Yang, Hao Lin, Jinglun Feng, Haigen Min, and Xiangmo Zhao. Driveagent: Multi-agent structured reasoning with llm and multimodal sensor fusion for autonomous driving. *arXiv preprint arXiv:2505.02123*, 2025. 5

[42] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5

[43] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17853–17862, 2023. 2

[44] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17853–17862, 2023. 3, 5

[45] Yanjun Huang, Jiatong Du, Ziru Yang, Zewei Zhou, Lin Zhang, and Hong Chen. A survey on trajectory-prediction methods for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 7(3):652–674, 2022. 2

[46] Zhijian Huang, Chengjian Feng, Feng Yan, Baihui Xiao, Zequn Jie, Yujie Zhong, Xiaodan Liang, and Lin Ma. Drivemm: All-in-one large multimodal model for autonomous driving. *arXiv preprint arXiv:2412.07689*, 2024. 3

[47] Zilin Huang, Zihao Sheng, Yansong Qu, Junwei You, and Sikai Chen. Vlm-rl: A unified vision language models and reinforcement learning framework for safe autonomous driving. *arXiv preprint arXiv:2412.15544*, 2024. 3

[48] Chia-Yu Hung, Qi Sun, Pengfei Hong, Amir Zadeh, Chuan Li, U Tan, Navonil Majumder, Soujanya Poria, et al. Nora: A small open-sourced generalist vision language action model for embodied tasks. *arXiv preprint arXiv:2504.19854*, 2025. 5

[49] Muhammad Hussain. Yolo-v1 to yolo-v8, the rise of yolo and its complementary nature toward digital manufacturing and industrial defect detection. *Machines*, 11(7):677, 2023. 2

[50] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, et al. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024. 5, 6, 8

[51] Ayesha Ishaq, Jean Lahoud, Ketan More, Omkar Thawakar, Ritesh Thawkar, Dinura Dissanayake, Noor Ahsan, Yuhao Li, Fahad Shahbaz Khan, Hisham Cholakkal, et al. Drivelmm-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding. *arXiv preprint arXiv:2503.10621*, 2025. 4

[52] Seif Ismail, Antonio Arbues, Ryan Cotterell, René Zurbrügg, and Carmen Amo Alonso. Narrate: Versatile language architecture for optimal control in robotics. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9628–9635. IEEE, 2024. 2

[53] Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. Adriver-i: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*, 2023. 1, 6, 8

[54] Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, and Hongyang Li. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21983–21994, 2023. 2, 3

[55] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. *arXiv preprint arXiv:2406.03877*, 2024. 4, 7, 9, 10

[56] Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. Drivetransformer: Unified transformer for scalable end-to-end autonomous driving. *arXiv preprint arXiv:2503.07656*, 2025. 3

[57] Anqing Jiang, Yu Gao, Zhigang Sun, et al. Diffvla: Vision-language guided diffusion planning for autonomous driving. *arXiv preprint arXiv:2505.19381*, 2025. 1, 4, 5, 6, 8, 10

[58] Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging large vision-language models and end-to-end autonomous driving. *arXiv preprint arXiv:2410.22313*, 2024. 3, 6

[59] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023. 3, 5

[60] Kemou Jiang, Xuan Cai, Zhiyong Cui, Aoyong Li, Yilong Ren, Haiyang Yu, Hao Yang, Daocheng Fu, Licheng Wen, and Pinlong Cai. Koma: Knowledge-driven multi-agent framework for autonomous driving with large language models. *IEEE Transactions on Intelligent Vehicles*, 2024. 4

[61] Sicong Jiang, Seongjin Choi, and Lijun Sun. Communication-aware reinforcement learning for cooperative adaptive cruise control. *arXiv preprint*

arXiv:2407.08964, 2024. 2

[62] Alex Kendall, Jeffrey Hawke, David Janz, Przemyslaw Mazur, Daniele Reda, John-Mark Allen, Vinh-Dieu Lam, Alex Bewley, and Amar Shah. Learning to drive in a day. In *2019 international conference on robotics and automation (ICRA)*, pages 8248–8254. IEEE, 2019. 2

[63] Jinkyu Kim, Z. Li, B. Floyd, et al. Textual explanations for self-driving vehicles. In *ECCV*, 2018. 4, 7, 9

[64] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025. 5

[65] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 4

[66] Stéphanie Lefèvre, Dizan Vasquez, and Christian Laugier. A survey on motion prediction and risk assessment for intelligent vehicles. *ROBOMECH journal*, 1:1–14, 2014. 2

[67] Chengmeng Li, Junjie Wen, Yan Peng, Yaxin Peng, Feifei Feng, and Yichen Zhu. Pointvla: Injecting the 3d world into vision-language-action models. *arXiv preprint arXiv:2503.07511*, 2025. 4, 5

[68] Peidong Li and Dixiao Cui. Does end-to-end autonomous driving really need perception tasks? *arXiv preprint arXiv:2409.18341*, 2024. 3

[69] Yingyan Li, Lue Fan, Jiawei He, Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang, and Tieniu Tan. Enhancing end-to-end autonomous driving with latent world model. *arXiv preprint arXiv:2406.08481*, 2024. 3

[70] Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, Guang Chen, Hangjun Ye, Wenyu Liu, and Xinggang Wang. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving, 2025. 5

[71] Zhenxin Li, Kailin Li, Shihao Wang, Shiyi Lan, Zhiding Yu, Yishen Ji, Zhiqi Li, Ziyue Zhu, Jan Kautz, Zuxuan Wu, et al. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978*, 2024. 3

[72] Zhenxin Li, Wenhao Yao, Zi Wang, Xinglong Sun, Joshua Chen, Nadine Chang, Maying Shen, Zuxuan Wu, Shiyi Lan, and Jose M Alvarez. Generalized trajectory scoring for end-to-end multimodal planning. *arXiv preprint arXiv:2506.06664*, 2025. 3

[73] Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. Pnpnet: End-to-end perception and prediction with tracking in the loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11553–11562, 2020. 2

[74] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 3, 5

[75] Jiaqi Liu, Peng Hang, Xiao Qi, Jianqiang Wang, and Jian Sun. Mtd-gpt: A multi-task decision-making gpt model for autonomous driving at unsignalized intersections. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 5154–5161. IEEE, 2023. 3

[76] Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Pengju An, Xiaoqi Li, Kaichen Zhou, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation. *Advances in Neural Information Processing Systems*, 37:40085–40110, 2024. 4

[77] Lin Liu, Ziying Song, Hongyu Pan, Lei Yang, and Caiyan Jia. Two tasks, one goal: Uniting motion and planning for excellent end to end autonomous driving performance. *arXiv preprint arXiv:2504.12667*, 2025. 3

[78] Pei Liu, Haipeng Liu, Haichao Liu, Xin Liu, Jinxin Ni, and Jun Ma. Vlm-e2e: Enhancing end-to-end autonomous driving with multimodal driver attention fusion. *arXiv preprint arXiv:2502.18042*, 2025. 4

[79] Xueyi Liu, Zuodong Zhong, Yuxin Guo, Yun-Fu Liu, Zhiguo Su, Qichao Zhang, Junli Wang, Yinfeng Gao, Yupeng Zheng, Qiao Lin, et al. Reasonplan: Unified scene prediction and decision reasoning for closed-loop autonomous driving. *arXiv preprint arXiv:2505.20024*, 2025. 4

[80] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023. 2

[81] Keke Long, Haotian Shi, Jiaxi Liu, and Xiaopeng Li. Vlm-mpc: Vision language foundation model (vlm)-guided model predictive controller (mpc) for autonomous driving. *arXiv preprint arXiv:2408.04821*, 2024. 3

[82] Han Lu, Xiaosong Jia, Yichen Xie, Wenlong Liao, Xiaokang Yang, and Junchi Yan. Activead: Planning-oriented active learning for end-to-end autonomous driving. *arXiv preprint arXiv:2403.02877*, 2024. 3

[83] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3569–3577, 2018. 2

[84] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal language model for driving. In *European Conference on Computer Vision*, pages 403–420. Springer, 2024. 3

[85] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*, 2024. 4

[86] Yukai Ma, Tiantian Wei, Naiting Zhong, Jianbiao Mei, Tao Hu, Licheng Wen, Xuemeng Yang, Botian Shi, and Yong Liu. Leapvad: A leap in autonomous driving via cognitive perception and dual-process thinking. *arXiv preprint arXiv:2501.08168*, 2025. 3

[87] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. 3

[88] Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving. *arXiv preprint arXiv:2311.10813*, 2023. 4, 5

[89] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice

Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, et al. Lingoqa: Visual question answering for autonomous driving. In *European Conference on Computer Vision*, pages 252–269. Springer, 2024. 4

[90] Jianbiao Mei, Yukai Ma, Xuemeng Yang, Licheng Wen, Xinyu Cai, Xin Li, Daocheng Fu, Bo Zhang, Pinlong Cai, Min Dou, et al. Continuously learning, adapting, and improving: A dual-process approach to autonomous driving. *arXiv preprint arXiv:2405.15324*, 2024. 3

[91] Yuya Miyaoka, Masaki Inoue, and Tomotaka Nii. Chatmpc: Natural language based mpc personalization. In *2024 American Control Conference (ACC)*, pages 3598–3603. IEEE, 2024. 3

[92] Alexander Naumann, Xunjiang Gu, Tolga Dimlioglu, Mariusz Bojarski, Alperen Degirmenci, Alexander Popov, Devansh Bisla, Marco Pavone, Urs Müller, and Boris Ivanovic. Data scaling laws for end-to-end autonomous driving. *arXiv preprint arXiv:2504.04338*, 2025. 3

[93] Ming Nie, Renyuan Peng, Chunwei Wang, et al. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. *ECCV*, 2024. 4, 8, 9, 10

[94] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4, 5

[95] Brian Paden, Michal Čáp, Sze Zheng Yong, Dmitry Yershov, and Emilio Frazzoli. A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on intelligent vehicles*, 1(1):33–55, 2016. 2

[96] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. Vlp: Vision language planning for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14760–14769, 2024. 4, 5

[97] Pranjal Paul, Anant Garg, Tushar Choudhary, Arun Kumar Singh, and K Madhava Krishna. Lego-drive: Language-enhanced goal-oriented closed-loop end-to-end autonomous driving. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10020–10026. IEEE, 2024. 3

[98] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025. 5

[99] Kangan Qian, Sicong Jiang, Yang Zhong, Ziang Luo, Zilin Huang, Tianze Zhu, Kun Jiang, Mengmeng Yang, Zheng Fu, Jinyu Miao, et al. Agentthink: A unified framework for tool-augmented chain-of-thought reasoning in vision-language models for autonomous driving. *arXiv preprint arXiv:2505.15298*, 2025. 4, 5

[100] Kangan Qian, Ziang Luo, Sicong Jiang, Zilin Huang, Jinyu Miao, Zhikun Ma, Tianze Zhu, Jiayin Li, Yangfan He, Zheng Fu, et al. 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*, 2025. 1, 3

[101] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025. 5

[102] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 3, 5, 6

[103] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023. 5

[104] Katrin Renz, Long Chen, Elahe Arani, and Oleg Sinavski. Simlingo: Vision-only closed-loop autonomous driving with language-action alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11993–12003, 2025. 6

[105] Katrin Renz, Long Chen, Ana-Maria Marcu, Jamie Shotton, et al. Carllava: Vision language models for camera-only closed-loop driving. In *CVPR*, 2025. 1, 2, 4, 6, 8, 9, 10, 11

[106] Katrin Renz, Kashyap Chitta, Otniel-Bogdan Mercea, A Koepke, Zeynep Akata, and Andreas Geiger. Plant: Explainable planning transformers via object-level representations. *arXiv preprint arXiv:2210.14222*, 2022. 2

[107] Abbas Sadat, Sergio Casas, Mengye Ren, Xinyu Wu, Pranaab Dhawan, and Raquel Urtasun. Perceive, predict, and plan: Safe motion planning through interpretable semantic representations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 414–430. Springer, 2020. 2

[108] Ranjan Sapkota, Yang Cao, Konstantinos I Roumeliotis, and Manoj Karkee. Vision-language-action models: Concepts, progress, applications and challenges. *arXiv preprint arXiv:2505.04769*, 2025. 4

[109] Wilko Schwarting, Javier Alonso-Mora, and Daniela Rus. Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems*, 1(1):187–210, 2018. 2

[110] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130, 2024. 4, 6

[111] Hao Shao, Letian Wang, Ruobing Chen, Hongsheng Li, and Yu Liu. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In *Conference on Robot Learning*, pages 726–737. PMLR, 2023. 2, 3

[112] Hao Shao, Letian Wang, Ruobing Chen, Steven L Waslander, Hongsheng Li, and Yu Liu. Reasonnet: End-to-end driving with temporal and global reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and*

pattern recognition, pages 13723–13733, 2023. 2

[113] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 5

[114] Yuanhua Shen and Jun Li. Utilizing navigation paths to generate target points for enhanced end-to-end autonomous driving planning. *arXiv preprint arXiv:2406.08349*, 2024. 3

[115] Chonghao Sima, Katrin Renz, Kashyap Chitta, et al. Drivelm: Driving with graph visual question answering. *ECCV*, 2024. 1, 8, 9, 10

[116] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 4

[117] Wenchao Sun, Xuewu Lin, Yining Shi, Chuang Zhang, Haoran Wu, and Sifa Zheng. Sparsedrive: End-to-end autonomous driving via sparse scene representation. *arXiv preprint arXiv:2405.19620*, 2024. 3

[118] Ran Tian, Boyi Li, Xinshuo Weng, Yuxiao Chen, Edward Schmerling, Yue Wang, Boris Ivanovic, and Marco Pavone. Tokenize the world into object-level knowledge to address long-tail events in autonomous driving. *arXiv preprint arXiv:2407.00959*, 2024. 3

[119] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 3, 5

[120] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[121] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 5

[122] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3

[123] Chris Urmson, Chris Baker, John Dolan, Paul Rybski, Bryan Salesky, William "Red" Whittaker, Dave Ferguson, and Michael Darms. Autonomous driving in traffic: Boss and the urban challenge. *AI Magazine*, 30(2):17–28, 2009. 2

[124] Tianqi Wang, Enze Xie, Ruihang Chu, Zhenguo Li, and Ping Luo. Drivecot: Integrating chain-of-thought reasoning with end-to-end driving. *arXiv preprint arXiv:2403.16996*, 2024. 4

[125] Tsun-Hsuan Wang, Alaa Maalouf, Wei Xiao, Yutong Ban, Alexander Amini, Guy Rosman, Sertac Karaman, and Daniela Rus. Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6687–6694. IEEE, 2024. 3

[126] Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, et al. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023. 3

[127] Xiao Wang, Ke Tang, Xingyuan Dai, Jintao Xu, Quancheng Du, Rui Ai, Yuxiao Wang, and Weihao Gu. S4tp: Social-suitable and safety-sensitive trajectory planning for autonomous vehicles. *IEEE Transactions on Intelligent Vehicles*, 9(2):3220–3231, 2023. 3

[128] Yuqi Wang, Ke Cheng, Jiawei He, Qitai Wang, Hengchen Dai, Yuntao Chen, Fei Xia, and Zhao-Xiang Zhang. Drivingdojo dataset: Advancing interactive and knowledge-enriched driving world model. *Advances in Neural Information Processing Systems*, 37:13020–13034, 2024. 4

[129] Julong Wei, Shanshuai Yuan, Pengfei Li, Qingda Hu, Zhongxue Gan, and Wenchao Ding. Occllama: An occupancy-language-action generative world model for autonomous driving. *arXiv preprint arXiv:2409.03272*, 2024. 4

[130] Junjie Wen, Minjie Zhu, Yichen Zhu, Zhibin Tang, Jinming Li, Zhongyi Zhou, Chengmeng Li, Xiaoyu Liu, Yaxin Peng, Chaomin Shen, et al. Diffusion-vla: Scaling robot foundation models via unified diffusion and autoregression. *arXiv preprint arXiv:2412.03293*, 2024. 5

[131] Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. Dilu: A knowledge-driven approach to autonomous driving with large language models. *arXiv preprint arXiv:2309.16292*, 2023. 4

[132] Xinshuo Weng, Boris Ivanovic, Yan Wang, Yue Wang, and Marco Pavone. Para-drive: Parallelized architecture for real-time autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15449–15458, 2024. 3

[133] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 4

[134] Katharina Winter, Mark Azer, and Fabian B Flohr. Bevdriver: Leveraging bev maps in llms for robust closed-loop driving. *arXiv preprint arXiv:2503.03074*, 2025. 3, 4

[135] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16133–16142, 2023. 5

[136] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are vlms ready for autonomous driving? an empirical study from the reliability, data, and metric per-

spectives. *arXiv preprint arXiv:2501.04003*, 2025. 3

[137] Jing Xiong, Gongye Liu, Lun Huang, Chengyue Wu, Taiqiang Wu, Yao Mu, Yuan Yao, Hui Shen, Zhongwei Wan, Jinfa Huang, et al. Autoregressive models in vision: A survey. *arXiv preprint arXiv:2411.05902*, 2024. 5

[138] Dongyang Xu, Haokun Li, Qingfan Wang, Ziying Song, Lei Chen, and Hanming Deng. M2da: multi-modal fusion transformer incorporating driver attention for autonomous driving. *arXiv preprint arXiv:2403.12552*, 2024. 2

[139] Qingyao Xu, Siheng Chen, Guang Chen, Yanfeng Wang, and Ya Zhang. Chatbev: A visual language model that understands bev maps. *arXiv preprint arXiv:2503.13938*, 2025. 4

[140] Yi Xu, Yuxin Hu, Zaiwei Zhang, Gregory P Meyer, Siva Karthik Mustikovela, Siddhartha Srinivasa, Eric M Wolff, and Xin Huang. Vlm-ad: End-to-end autonomous driving through vision-language model supervision. *arXiv preprint arXiv:2412.14446*, 2024. 3, 4

[141] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024. 4, 6, 8

[142] Senqiao Yang, Jiaming Liu, Renrui Zhang, Mingjie Pan, Ziyu Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Hongsheng Li, Yandong Guo, et al. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9247–9255, 2025. 4

[143] Zhenjie Yang, Yilin Chai, Xiaosong Jia, Yuqian Shao, et al. Drivemoe: Mixture-of-experts for vision-language-action model in end-to-end autonomous driving. *arXiv preprint arXiv:2505.16278*, 2025. 5, 6, 7, 8, 9, 10, 11

[144] Wenhao Yao, Zhenxin Li, Shiyi Lan, Zi Wang, Xinglong Sun, Jose M Alvarez, and Zuxuan Wu. Drivesuprim: Towards precise trajectory selection for end-to-end planning. *arXiv preprint arXiv:2506.06659*, 2025. 3

[145] Tengju Ye, Wei Jing, Chunyong Hu, Shikun Huang, Lingping Gao, Fangzhen Li, Jingke Wang, Ke Guo, Wencong Xiao, Weibo Mao, et al. Fusionad: Multi-modality fusion for prediction and planning tasks of autonomous driving. *arXiv preprint arXiv:2308.01006*, 2023. 2

[146] J Javier Yebes, Luis M Bergasa, and Miguel Ángel García-Garrido. Visual object recognition with 3d-aware features in kitti urban scenes. *Sensors*, 15(4):9228–9250, 2015. 4

[147] Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, Trevor Darrell, et al. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*, 2(5):6, 2018. 7, 9

[148] Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. Ragdriver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828*, 2024. 4, 5, 6, 8, 11

[149] Wenyuan Zeng, Wenjie Luo, Simon Suo, Abbas Sadat, Bin Yang, Sergio Casas, and Raquel Urtasun. End-to-end interpretable neural motion planner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8660–8669, 2019. 2

[150] Mingliang Zhai, Cheng Li, Zengyuan Guo, Ningrui Yang, Xiameng Qin, Sanyuan Zhao, Junyu Han, Ji Tao, Yuwei Wu, and Yunde Jia. World knowledge-enhanced reasoning using instruction-guided interactor in autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9842–9850. IEEE, 2025. 3, 6

[151] Diankun Zhang, Guoan Wang, Runwen Zhu, Jianbo Zhao, Xiwu Chen, Siyu Zhang, Jiahao Gong, Qibin Zhou, Wenyuan Zhang, Ningzi Wang, et al. Sparsead: Sparse query-centric paradigm for efficient end-to-end autonomous driving. *arXiv preprint arXiv:2404.06892*, 2024. 3

[152] Enming Zhang, Xingyuan Dai, Yisheng Lv, and Qinghai Miao. Minidrive: More efficient vision-language models with multi-level 2d features as text tokens for autonomous driving. *arXiv preprint arXiv:2409.07267*, 2024. 5

[153] Jiawei Zhang, Xuan Yang, Taiqi Wang, Yu Yao, Aleksandr Petiushko, and Bo Li. Safeauto: Knowledge-enhanced safe autonomous driving with multimodal foundation models. *arXiv preprint arXiv:2503.00211*, 2025. 4, 6, 7, 8, 9, 10, 11

[154] Ruijun Zhang, Xianda Guo, Wenzhao Zheng, Chenming Zhang, Kurt Keutzer, and Long Chen. Instruct large language models to drive like humans. *arXiv preprint arXiv:2406.07296*, 2024. 3

[155] Weize Zhang, Mohammed Elmahgiubi, Kasra Rezaee, Behzad Khamidehi, Hamidreza Mirkhani, Fazel Arasteh, Chunlin Li, Muhammad Ahsan Kaleem, Eduardo R Corral-Soto, Dhruv Sharma, et al. Analysis of a modular autonomous driving architecture: The top submission to carla leaderboard 2.0 challenge. *arXiv preprint arXiv:2405.01394*, 2024. 5

[156] Yunpeng Zhang, Deheng Qian, Ding Li, Yifeng Pan, Yong Chen, Zhenbao Liang, Zhiyao Zhang, Shurui Zhang, Hongxu Li, Maolei Fu, et al. Graphad: Interaction scene graph for end-to-end autonomous driving. *arXiv preprint arXiv:2403.19098*, 2024. 3

[157] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022. 2

[158] Zhiyuan Zhang, Xiaofan Li, Zhihao Xu, Wenjie Peng, Zijian Zhou, Miaojing Shi, and Shuangping Huang. Mpdrive: Improving spatial understanding with marker-based prompt learning for autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12089–12099, 2025. 3

[159] Rui Zhao, Qirui Yuan, Jinyu Li, Haofeng Hu, Yun Li, Chengyuan Zheng, and Fei Gao. Sce2drivex: A generalized mllm framework for scene-to-drive learning. *arXiv preprint arXiv:2502.14917*, 2025. 3

[160] Zongcai Zhao, Yue Zhao, et al. Extending large vision-language model for diverse interactive tasks in autonomous driving. *arXiv preprint arXiv:2505.08725*, 2025. 2, 8, 9, 10

[161] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla:

A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024. 5

[162] Wenzhao Zheng, Ruiqi Song, Xianda Guo, Chenming Zhang, and Long Chen. Genad: Generative end-to-end autonomous driving. In *European Conference on Computer Vision*, pages 87–104. Springer, 2024. 3

[163] Yupeng Zheng, Zhongpu Xia, Qichao Zhang, Teng Zhang, Ben Lu, Xiaochuang Huo, Chao Han, Yixian Li, Mengjie Yu, Bu Jin, et al. Preliminary investigation into data scaling laws for imitation learning-based end-to-end autonomous driving. *arXiv preprint arXiv:2412.02689*, 2024. 3

[164] Hang Zhou, Haichao Liu, Hongliang Lu, Jun Ma, and Yiding Ji. Enhance planning with physics-informed safety controller for end-to-end autonomous driving. In *2024 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1775–1782. IEEE, 2024. 3, 4

[165] Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, and Alois C Knoll. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. *arXiv preprint arXiv:2503.23463*, 2025. 1, 5, 6, 8

[166] Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C Knoll. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, 2024. 2, 4

[167] Xirui Zhou, Lianlei Shan, and Xiaolin Gui. Dynrsl-vlm: Enhancing autonomous driving perception with dynamic resolution vision-language models. *arXiv preprint arXiv:2503.11265*, 2025. 2, 4, 6, 10, 11

[168] Zewei Zhou, Tianhui Cai, Seth Z. Zhao, Yun Zhang, Zhiyu Huang, Bolei Zhou, and Jiaqi Ma. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning, 2025. 5, 6, 7, 8