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, instruction fidelity, and explanation quality. Finally, we detail open challenges—robustness, real-time efficiency, and formal verification—and outline future directions toward foundation-scale driving models and a standardised "traffic language." This survey provides a concise yet complete reference for advancing interpretable, instruction-following, and socially aligned autonomous vehicles.

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 [36, 52, 67, 69, 79]. 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 [12, 39, 114]. By conditioning the controller on language tokens, these systems can (i) follow free-form commands such as "yield to the ambulance" [71], (ii) verbalize their internal rationale for post-hoc verification [12], and (iii) leverage commonsense priors from internet-scale corpora to extrapolate in rare or unforeseen situations [13]. Early prototypes have already demonstrated improved safety and instruction fidelity in simulation and closed-loop tests, foreshadowing a new research frontier we term *VLA*

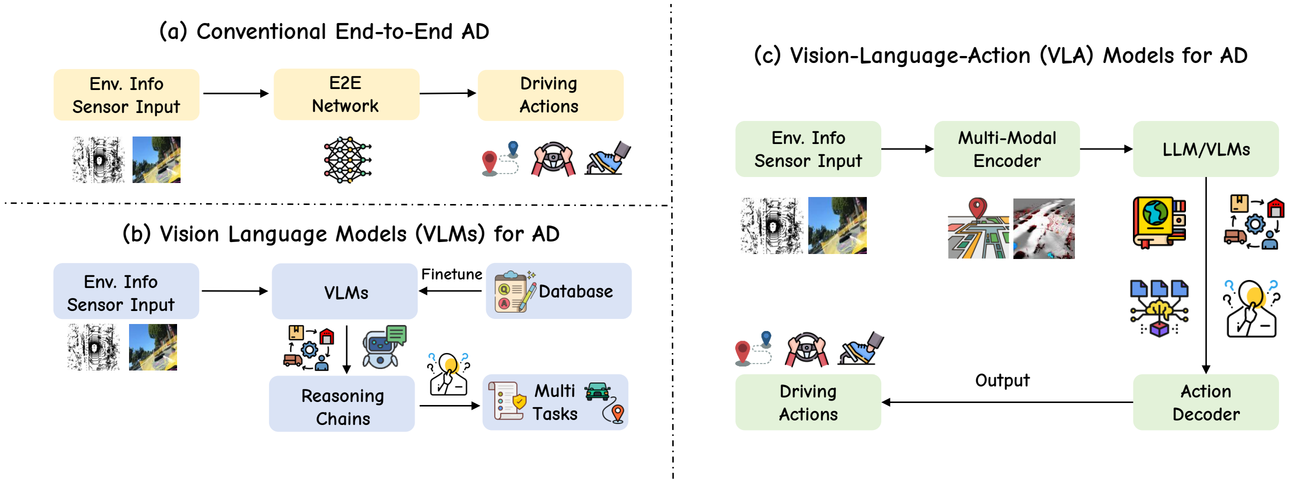

Figure 1. Comparisons of autonomous driving paradigms. (a) End-to-end driving offers direct perception-to-control mapping but lacks interpretability and generalization. (b) VLM4AD introduces natural language reasoning and explainability, yet remains perception-centric. (c) VLA4AD integrates perception,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

for Autonomous Driving (VLA4AD).

Several converging trends underscore the timeliness of this emerging research frontier. First, petabyte-scale multi-sensor logs such as nuScenes [5] and Impromptu VLA [13] provide rich multimodal supervision. Second, foundation LLMs can now be efficiently adapted through low-rank updates, while token-reduction designs such as TS-VLM [9] reduce online computational cost by an order of magnitude [116]. Third, synthetic corpora like SimLingo [71] and interactive datasets such as NuInteract [109] 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[18, 19, 23, 88, 115], 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. 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 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, localization, prediction, planning, and control [10, 25, 41, 42, 63, 85]. Handcrafted 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 [34, 45]. 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 [50, 56, 73].

To mitigate these issues, subsequent systems embraced multi-task learning (MTL), where a shared backbone feeds multiple task-specific heads for detection, segmentation, motion prediction, and planning. Notable examples include Transfuser [15], BEVerse [107], and unified architectures proposed in [7, 103]. Similar approaches underpin industrial stacks at companies such as Mobileye, Tesla, and NVIDIA. While co-training promotes hierarchical feature sharing and reduces computational cost, it also introduces the challenge of negative transfer-inter, i.e., task interference that can degrade overall performance [17, 55].

2.2. End-to-End Autonomous Driving

As illustrated in Fig. 1 (a), end-to-end driving policies that map raw sensor streams directly to control commands, bypassing hand-crafted modular pipelines [14, 15, 27, 37, 72, 76, 77]. Early exemplars such as UniAD [29] and VAD [40] integrate perception, prediction, and planning within a single backbone, while VADv2[11] extends this with probabilistic trajectory sampling [11]. Generative variants, including GenAD [112] and DiffusionDrive [51, 108], further model the multi-modal nature of the driving future by outputting a distribution over feasible paths.

Unlike dense recognition tasks, end-to-end driving requires inferring sparse control signals (e.g., steering, throttle, brake) from high-dimensional inputs. This poses a challenge for encoders to isolate decision-critical features. Consequently, many systems incorporate auxiliary heads, such as semantic segmentation [33, 35, 97], depth estimation [33, 49], LiDAR-based 3D detection [7, 35], or bird's-eye-view (BEV) segmentation [7, 37, 76, 77], to regularize the latent space and improve representation learning. Visual affordance prediction (e.g., traffic light state, lane offset) offers a lightweight alternative [33, 58, 76]. However, scaling up auxiliary supervision is costly. Closed-loop evaluations have revealed diminishing returns beyond certain data volumes, along with significant variance in performance across scenario types [60, 113]. These findings suggest that pure data scaling is insufficient for achieving Level 4+ autonomy. Overall, end-to-end learning has narrowed the gap between raw pixels and control decisions, yet persistent challenges remain. Specifically, these systems often suffer from: (i) *brittle semantics*, particularly in rare or rapidly evolving scenarios; (ii) *opaque reasoning*, which complicates safety auditing and verification; and (iii) *limited language proficiency*, restricting intuitive human–vehicle interaction.

2.3. VLMs for Autonomous Driving

LLMs [82, 84] and VLMs [1, 52, 69, 95] offer a promising remedy by unifying perception and natural language reasoning within a shared embedding space [30]. At the core of this progress is large-scale multimodal pretraining, which equips models with commonsense associations (e.g., siren $\rightarrow$ yield) that task-specific labels often miss. As a result, language-conditioned VLM policies exhibit stronger zero-shot generalization across novel objects, weather, and driving norms. Recent work has begun embedding these models directly into the driving loop, as shown in Fig. 1 (b).

Early efforts, such as DriveGPT-4 [98] and EMMA [32], show that a frozen VLM can process multi-view images and a textual prompt to produce either a trajectory plan or low-level control tokens, while simultaneously generating human-readable rationales. Follow-up studies have scaled these systems in two key directions. First, in scene understanding, methods such as Dilu [91], BEVDriver [93], and LeapAD [59] integrate BEV features with language to answer 3D spatial queries or predict multimodal futures. Second, in interaction and explainability, approaches such as RAG-Driver [101] retrieve map rules on the fly; DS-Drive [53] and ReasonPlan [54] generate step-by-step justifications for decisions; and AgentThink, FSDrive [66, 102] employs COT prompting to refine driving actions. Yet, these menthods remain perception-centric: the plans they generate are not tightly coupled with closed-loop control, and their explanations offer no formal safety guarantees.

2.4. From VLM to VLA for Autonomous Driving

As shown in Fig. 1 (c), VLA address the aforementioned gap by incorporating an explicit action head, thereby unifying perception, reasoning, and control within a single policy [57, 74, 115]. VLA policies are motivated by three core demands of real-world driving: (i) robust reasoning under rare, long-tail events [38, 87]; (ii) noise-tolerant control in dynamic, partially occluded scenes; and (iii) the ability to interpret spontaneous, high-level commands (e.g., "overtake the truck") [64, 98]. Leveraging foundation backbones trained on internet-scale data [62], VLA generalize across domains and datasets [6, 92].

Concretely, a modern VLA models can: (i) ground free-form instructions in the ego-centric visual context and output a corresponding trajectory [21, 101]; (ii) produce COT justifications, such as DriveCoT or CoT-VLA [19, 86], to enhance transparency; and (iii) move beyond direct control tokens to sophisticated planners, including diffusion heads [39], hierarchical CoT-based reasoners [21], or mixed discrete–continuous control experts. Recent exemplars highlight the breadth of VLA capabilities: CoVLA-Agent [12] grounds both textual and trajectory outputs; ORION [21] maintains long-horizon memory and explicit reasoning; SafeAuto [105] wraps a logical rule layer around an LLM to strictly enforce traffic compliance; and DiffVLA [39] samples multi-modal futures via sparse–dense diffusion. These advances mark a decisive shift from perception-centric VLM pipelines to action-aware, explainable, and instruction-following multimodal agents, paving the way toward safe, generalizable, and human-aligned 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. This is a primary modality, typically consisting of video streams from one or more cameras mounted

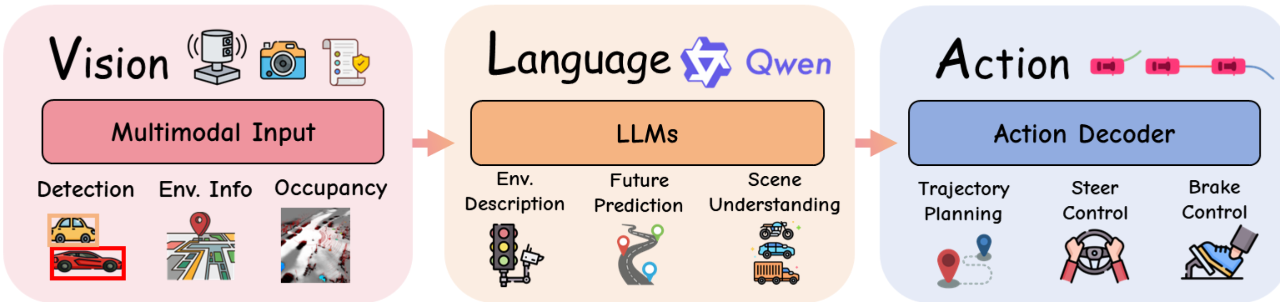

Figure 2. Overview of the VLA4AD Architecture.

on the vehicle. While front-facing cameras are common, comprehensive scene understanding often requires surround-view camera systems [80]. Raw images may be processed directly or transformed into intermediate representations such as BEV maps, which simplify spatial reasoning [16]. The resolution of input images is a key design parameter, involving trade-offs between visual detail and computational efficiency [116].

Other Sensor Data. These include LiDAR for precise 3D environmental mapping, RADAR for estimating object velocities, Inertial Measurement Units (IMUs) for vehicle motion tracking, and GPS for localization [6, 92]. Proprioceptive signals, such as steering angle, velocity, and acceleration, are also essential for closed-loop control. Effectively incorporating and fusing these diverse sensor modalities remains an active area of research [8, 46, 89].

Language Inputs. Natural language inputs, including commands, queries, or structured task descriptions, play a defining role in VLA4AD. These inputs include: (i) Direct user commands (e.g., "turn left at the next intersection," "overtake the truck when safe") [64]; (ii) Queries about the environment or vehicle status (e.g., "Is it safe to change lanes now?") [61]; (iii) Task specifications, including navigation instructions, traffic rules, or map information expressed linguistically [24]. Recent work has also explored integrating speech inputs as a richer and more natural form of language grounding [101, 114].

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 [62], ConvNeXt-V2 [94], or CLIP [69]. Many VLA systems employ BEV projection [81], and others incorporate 3D priors via point-cloud encoders (e.g., PointVLA [46]) or voxel modules (3D-VLA [110]). Multi-scale fusion using language-derived keys improves grounding at fine spatial levels [68, 104].

Language Processor. Natural language is processed using pretrained decoders such as LLaMA2 [83], Vicuna[111], Qwen[2] or GPT-style transformers [4]. Instruction-tuned variants (e.g., Visual Instruction Tuning [52]) and retrieval-augmented prompting (RAG) (RAG-Driver [101]) inject domain knowledge. Lightweight fine-tuning strategies like LoRA [28] enable efficient adaptation.

Action Decoder. Downstream control is emitted via (i) Autoregressive tokenizers where discrete actions or trajectory way-points are predicted sequentially [31, 44, 65, 117], (ii) Diffusion heads that sample continuous controls conditioned on fused embeddings (DiffVLA [39]; Diffusion-VLA [90]), or (iii) Flow-matching / policy gradient experts used by GRPO [78] or DPO[70] fine-tuning pipelines [47, 96]. Hierarchical controllers (e.g., ORION [21]) let a language planner dispatch sub-goal sketches to a separate low-level PID or MPC stack.

3.3. Driving Outputs

The output of a VLA model reflects its operational goal:

Low-Level Actions. Many models directly predict steering, throttle, and brake commands, either as continuous control signals or as discrete control tokens [21, 65, 114].

Trajectory Planning. A more common formulation involves predicting future trajectories or waypoints, typically represented in BEV or ego-centric coordinates [3, 39, 64], which are then executed by MPC-based controllers [106].

Together, these outputs reflect the growing ambition of VLA systems: not just to control a vehicle, but to act explainably and robustly within complex traffic environments[48]. In summary, VLA4AD takes multimodal sensor data and a language query or command as input, and produces both driving decisions (e.g., control signals or planned trajectories) and, in some cases, accompanying language-based 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

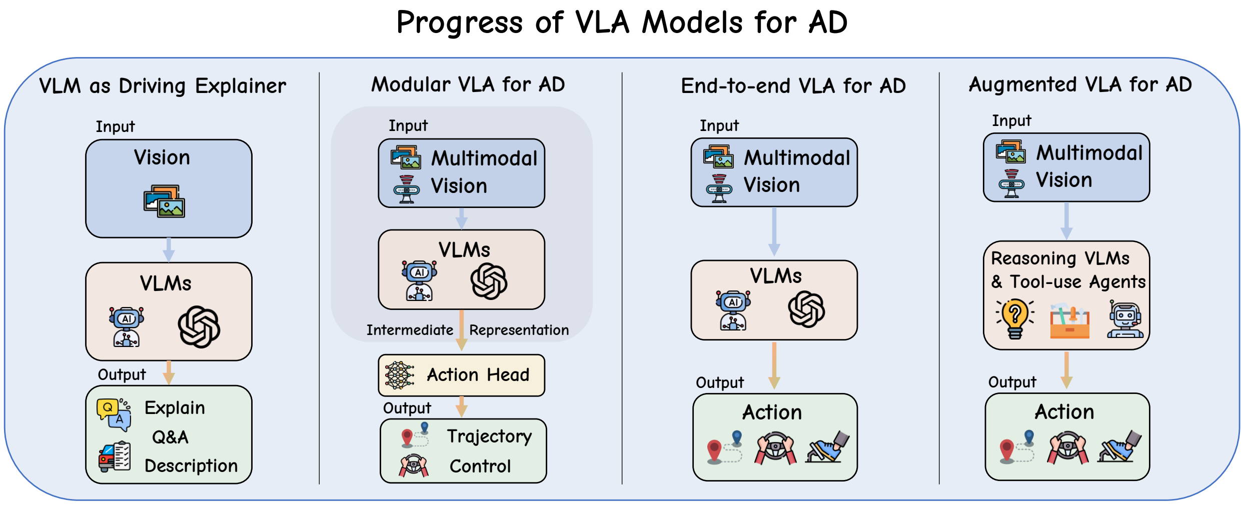

Figure 3. Evolution of VLA models for AD. From left to right: (1) VLM-as-explainer: a pre-trained VLM/LLM narrates the driving scene but produces no control action. (2) Modular VLA: language is transformed into an intermediate representation that an action head converts into trajectory or low-level control. (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.

stages: *Pre-VLA: Explanatory Language Models*, *Modular VLA4AD*, *Unified 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, and core contributions.

4.1. Pre-VLA: Language Model as Explainer

Early work used language purely as a descriptive overlay, enhancing interpretability without directly driving control. Typically, a frozen vision encoder (e.g., CLIP) extracted features that were fed into an LLM decoder to narrate the scene or suggest high-level maneuvers, while a conventional controller executed the actual commands. DriveGPT-4 [98] exemplifies this phase by generating textual explanations and multi-round QA. TS-VLM [9] uses text-guided soft attention pooling to focus on key regions, and DynRsl-VLM [116] improved efficiency via adaptive input resolution. Although these methods increased transparency, they introduced latency from long descriptions and did not close the loop from language to action.

4.2. Modular VLA Models for AD

Later, language moved from a passive descriptor to an active planner within a modular architecture. Instead of just commenting on the scene, language inputs and outputs were used to inform the planning module. For instance, OpenDriveVLA [114] fused camera and LiDAR inputs with a textual route instruction (e.g., "turn right at the church") and produced intermediate human-readable waypoints (e.g., "turn right in 20m, then go straight"), which were then translated into a trajectory. This made the planning process more transparent and flexible. Similarly, DriveMoE [99] employed a Mixture-of-Experts design where language cues helped select among multiple specialized sub-planners (e.g., an "overtaking expert" vs. a "stop-and-go traffic expert"). In multi-agent scenarios, LangCoop [24] demonstrated vehicles sending compact language messages to each other to achieve coordination at intersections. SafeAuto [105] integrated logical traffic rules (expressed in a formal language) to veto or adjust the language-driven plan if it violated safety constraints, and RAG-Driver [101] augmented planning with retrieval of similar past scenarios to guide decisions. These approaches significantly reduced the mismatch between what the language described and what the car did, effectively embedding language into the planning loop. However, they often still involve multiple stages (perception → language → plan → control), incurring latency and potential failure points at each interface. This motivated a push of VLA models for autonomous driving toward more unified, end-to-end systems.

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

The next phase emerged as large multimodal foundation models became available. Generally, a camera stream (and optionally a textual command) goes into a single network that directly outputs a driving trajectory or control commands in one forward pass. A prime example is EMMA [32], 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. CoVLA-Agent [12] merges captioning and trajectory prediction in a single autoregressive LLM. SimLingo [71] built on a LLaVA and fine-tuned it

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.

Model	Year	Data Source		Model		Output	Focus
		Input	Dataset / Benchmark	Vision	LLM		
DriveGPT-4 [98]	2023	Single	BDD-X	CLIP	LLaMA-2	**LLC**	Interpretable LLM
ADriver-I [36]	2023	Single	nuScenes + Private	CLIP ViT	Vicuna-1.5	**LLC**	Diffusion World Model
RAG-Driver [101]	2024	Multi	BDD-X	CLIP ViT	Vicuna-1.5	**LLC**	RAG-powered Driving
EMMA [32]	2024	Multi + State	Waymo fleet	Gemini-VLM	Gemini	**Multi.**	Capable of Multi-task
CoVLA-Agent [12]	2024	Single + State	CoVLA Data	CLIP ViT	Vicuna-1.5	**Traj.**	Autolabeled Dataset
OpenDriveVLA [114]	2025	Multi	nuScenes	Custom Module	Qwen-2.5	**LLC+Traj.**	2D/3D Alignment
ORION [22]	2025	Multi + History	nuScenes + CARLA	QT-Former	Vicuna-1.5	**Traj.**	Memory-aware CoT
DriveMoE [99]	2025	Multi	Bench2Drive	Paligemma-3B	–	**LLC**	Mix-of-Experts in AD
VaViM [3]	2025	Video Frames	BDD + CARLA	LlamaGen	GPT-2	**Traj.**	Video-token Pretrain
DiffVLA [39]	2025	Multi + State	Navsim-v2	CLIP ViT	Vicuna-1.5	**Traj.**	Diffusion Trajectory
LangCoop [24]	2025	Single + V2V	CARLA	GPT-4o	GPT-4o	**LLC**	Language-based V2V
SimLingo [71]	2025	Multi	CARLA + Bench2Drive	InternVL2	Qwen-2	**LLC+Traj.**	Action-dreaming Data
SafeAuto [105]	2025	Multi + State	BDD-X + DriveLM	CLIP ViT	Vicuna-1.5	**LLC**	Traffic-rule Constraints
Impromptu-VLA [13]	2025	Single	Impromptu Data	Qwen-2.5VL	Qwen-2.5VL	**Traj.**	CoT w/ Corner-case
AutoVLA [117]	2025	Multi + State	nuScenes + CARLA	Qwen-2.5VL	Qwen-2.5VL	**LLC+Traj.**	Adaptive Reasoning

Table 2. **Notable Datasets & Benchmarks for VLA4AD.** Real = real-world logs; Sim = CARLA simulation[20]; clips = video segments; QA = question–answer pairs; CoT = chain-of-thought.

Name	Year	Domain	Scale	Modalities	Tasks
BDD100K / BDD-X [43, 100]	2018	Real (US)	100 k videos; 7 k clips	RGB video	Captioning, QA
nuScenes [5]	2020	Real (Boston/SG)	1 k scenes (20 s, 6 cams)	RGB, LiDAR, Radar	Detection, QA
Bench2Drive [38]	2024	Sim (CARLA)	220 routes; 44 scenarios	RGB	Closed-loop control
Reason2Drive [61]	2024	Real (nuSc/Waymo)	600 k video–QA	RGB video	CoT-style QA
DriveLM-Data [79]	2024	Real+Sim	18 k scene graphs	RGB, Graph	Graph QA
Impromptu VLA [13]	2025	Real (multi-src)	80 k clips (30 s)	RGB video, State	QA, Trajectory
NuInteract [109]	2025	Real (nuScenes)	1 k scenes	RGB, LiDAR	Multi-turn QA
DriveAction [26]	2025	Real (fleet)	2.6 k scenarios; 16.2 k QA	RGB video	High-level QA

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 [36] learned a latent world model that predicts future camera frames given actions (using diffusion), thereby enabling planning via imagining the consequences of actions. DiffVLA [39] combined sparse (waypoints) and dense (occupancy 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) and with providing fine-grained explanations of their decisions.

4.4. Reasoning-Augmented VLA Models for AD

The most recent wave of VLA4AD research places large language models at the heart of both reasoning and control, enabling vehicles to plan over extended horizons, retain and query minutes of past observations, and engage in interactive decision-making. For example, ORION [22] couples its QT-Former memory module—which compresses multi-view images and past actions into a compact latent state—with a Vicuna-based LLM that attends to this memory to generate six candidate trajectory distributions alongside a natural-language rationale. Impromptu-VLA [13] goes further by aligning CoT with action: trained on 80 000 corner-case video clips annotated at each reasoning step, it verbalizes its decision path before emitting steering and throttle tokens, yielding a reduction in collision rate in zero-shot negotiation scenarios. AutoVLA [117] unifies CoT and trajectory planning in a single autoregressive transformer by discretizing continuous paths into "drive tokens," interleaving thought and action tokens during decoding, which achieves superior score in closed-loop tests. Collectively, these reasoning-centric VLA models not only react to sensor inputs but also explain their intent, anticipate multiple futures, and plan strategically, paving the way for conversational autonomous vehicles that can justify and adjust their behavior in real time—while raising new challenges in scalable memory indexing, sub-30 Hz reasoning throughput, and formal verification of 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 [43, 100]. BDD 100K offers 100 k diverse US videos; the BDD-X subset ($\sim$7 k clips) adds time-aligned human rationales (e.g., "slows because pedestrian crossing"), providing ground-truth explanations for models such as CoVLA-Agent [12] and SafeAuto [105].

nuScenes [5]. 1k 20s real-world episodes (Boston, Singapore) with 6 cams, LiDAR + radar and full 3D labels. It has been used for extensive VLA4AD evaluations.

Bench2Drive [38]. A closed-loop CARLA benchmark with 44 scenario types (220 routes) and a 2M-frame training set. Metrics isolate specific skills.

Reason2Drive [61]. 600k video–text pairs (nuScenes, Waymo, ONCE) annotated with CoT QA spanning perception $\rightarrow$ prediction $\rightarrow$ action. Evaluates logical consistency across the whole reasoning chain; current VLMs still struggle with higher-level queries.

DriveLM-Data [79]. Provides graph-structured QA on nuScenes (18 k graphs) and CARLA (16 k) scenes, stressing conditional reasoning. Baseline TS-VLM [9] attains BLEU-4 56 but low graph consistency, leaving ample room for improved multi-step reasoning.

Impromptu VLA [13]. 80k 30s clips ($\sim$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 [109]. 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.

DriveAction [26]. 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 of perception-only suites.

In short, the datasets span the full spectrum needed for VLA4AD research: BDD-X [43] and nuScenes [5] deliver large-scale, sensor-rich realism; Bench2Drive [38] and Impromptu VLA [13] inject safety-critical corner cases; and Reason2Drive [61], DriveLM [79], NuInteract [109], and DriveAction [26] 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

6.1. Training Paradigms

In practice, VLA4AD policies are typically cloned from expert data, hardened via targeted RL or corner-case replay, and then compressed for real-time deployment.

Supervised Imitation Learning (IL) Most VLA4AD models start by cloning expert drives, minimising ℓ_2 (continuous controls) or cross-entropy (tokenised actions) on logged image–text–action sequences. CoVLA-Agent predicts both future trajectory and caption [12], while DriveMoE and SimLingo policies replicate $\sim 10^6$ simulator steps [71, 99]. Because IL only interpolates its dataset, unseen hazards still require data-centric fixes such as DAgger roll-outs or corner-case augmentation.

Reinforcement Learning (RL). After supervised learning, rl is usually used for continued fine-tuning. Common reinforcement learning methods include PPO[75], DPO[70] and GRPO[78]. Models usually apply RL-finetune in CARLA or Bench2Drive to optimise route progress, safety, and rule compliance. Joint optimisation of text and control remains rare — only unsafe actions are penalised.

Multi-stage Training. Most groups follow a loose four-step loop: (1) *Pre-train* generic vision and language backbones; (2) *Align* them with a modest image–text–action set(e.g. DriveMonkey on NuInteract [109]); (3)*Targeted augmentation*—corner-case or instruction-heavy data such as SimLingo [71] or Bench2Drive [38] is injected to cover long-tail failure modes; some groups add RL passes for rule compliance (SafeAuto [105]). (4) *Compress* with LoRA, sparse MoE, or distillation for on-car use. As seen in Drive-MoE [99] and TS-VLM [9].

Together, these paradigms reveal a consensus workflow—*clone, adapt, verify, compress*—while leaving open research space for unified text + action RL objectives and fully unsupervised corner-case discovery.

6.2. Evaluation Protocols

VLA4AD systems must be judged on *how well they drive* and *how well they understand the command*; current practice measures these along different complementary axes:

Closed-loop Driving. Route success on CARLA or Bench2Drive [38]; infractions (e.g., collisions, red-light, off-road) where DiffVLA halves the error via a PDMS layer [39]; rule compliance checked by logic vetoes (SafeAuto [105]); and generalization to unseen weather.

Open-loop Prediction. Trajectory ℓ_2 and collision rate (nuScenes challenge); goal reach on instruction-conditioned targets.

Robustness & Stress. Sensor perturbations (blur, dropout, lag); adversarial prompts or patches; out-of-distribution events; and language edge-cases (idioms, code-switching, multilingual queries).

Overall, credible evaluation must score control, language, and their interplay. Yet control-centric suites and reasoning tests remain disjoint, underscoring the need for a unified "AI driver's licence" that measures both.

7. Open Challenges

Despite rapid progress, VLA4AD still face substantial barriers before large–scale real–world deployment. Below we summarise several most pressing research fronts.

Robustness & Reliability. VLA policies must remain stable under sensor corruption (rain, glare), linguistic noise, and regional domain shift. LLM hallucinations and slang mis-parsing introduce new failure modes; logic-based vetoes (e.g., SafeAuto [105]) are only a first step toward formally verified, socially compliant driving.

Real-time Performance. Executing a vision transformer and an LLM at $\geq$ 30 Hz on automotive hardware is non-trivial. Token pooling (TS-VLM [9]), quantisation, event-triggered reasoning, and MoE/distillation pipelines will be essential as model sizes approach multi-billion scale.

Data & Annotation Bottlenecks. Tri-modal labels (image + action + language) remain scarce; even large sets such as Impromptu VLA require costly manual QA. Synthetic augmentation (SimLingo [71]) and weak supervision help, but coverage of legally critical phrasing is still thin.

Multimodal Alignment. Current work is camera-centric; robust fusion of LiDAR, radar, HD-maps, and temporal memory is open. Proposed routes range from BEV projections to 3D token adapters, and from ORION's language memory [21] to map-text retrieval (RAG-Driver [101]). Scalable, temporally consistent alignment remains unsolved.

Tackling these challenges requires coordinated advances in scalable learning, formal safety, HCI, and policy. Improvements in any area—robust perception, lean LLMs, reliable V2V, or unified benchmarks—will speed progress toward safe, transparent, globally deployable VLA4AD.

8. Future Directions

VLA4AD is poised to evolve from isolated prototypes to *scalable, certifiable, and cooperative* driving platforms. We distil four priority tracks.

Foundation-scale Driving Models. A GPT-style backbone that self-supervises on panoramic videos, LiDAR, HD-maps, and textual road rules could offer zero-shot adaptation via prompting or LoRA, much as SimLingo already do on a smaller scale [71]. Key hurdles are masked multimodal objectives and memory-efficient 360° architectures.

Neuro-symbolic Safety Kernels. Pure end-to-end nets lack formal guarantees. Hybrids that emit structured action programs verified by symbolic planners—an idea hinted by SafeAuto [105] and ORION's language memory [21]—could combine flexibility with certifiability.

Communication & Social Intelligence. Robust traffic interaction demands a constrained ontology of intents ("I-yield-to-you", "Obstacle-ahead") plus the ability to parse gestures, voice, or V2V text. MoE routing (Drive-MoE [99]) and token-reduction LMs (TS-VLM [9]) keep bandwidth low, while retrieval-augmented planners (RAG-Driver [101]) can ground non-verbal cues.

In essence, achieving these goals will require progress in large-scale multimodal learning, formal verification, communication standards, and human-AI interaction. There is a need for continuous development and advancement in multiple modules within VLA framework.

9. Conclusion

In this paper, we present the **first** comprehensive survey of Vision–Language–Action models for Autonomous Driving (VLA4AD), bridging the gap between classic AV reviews and the recent rise of VLA models. Tracing the evolution from perception–planning stacks to end-to-end reasoning agents, we map architectures, representative models, training strategies, datasets, and evaluation protocols, establishing a common vocabulary for this fast-growing field.

Future progress hinges on richer image–text–action corpora, efficient multimodal backbones, and robustness in long-tail cases—prerequisites for reliable, fleet-scale VLA4AD. By distilling current advances and open challenges, this survey provides a roadmap toward trustworthy, socially aligned VLA autonomy.

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] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 4

[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. 4, 6

[4] 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. 4

[5] Holger Caesar, Alex Bankiti, Orien Lang, et al. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2, 6, 7

[6] 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. 3, 4

[7] 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, 3

[8] 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

[9] 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, 5, 7, 8

[10] 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

[11] 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

[12] Haohan Chi, Huan-ang Gao, Ziming Liu, et al. CoVLA: Comprehensive vision-language-action dataset for autonomous driving. In *WACV*, 2025. 1, 3, 5, 6, 7

[13] 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, 6, 7

[14] 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. 3

[15] 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

[16] 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

[17] Michael Crawshaw. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*, 2020. 2

[18] 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

[19] 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, 3

[20] 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. 6

[21] 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. 3, 4, 8

[22] 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. 6

[23] 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

[24] Xiangbo Gao, Yuheng Wu, Rujia Wang, et al. Langcoop: Collaborative driving with language. *arXiv preprint arXiv:2504.13406*, 2025. 4, 5, 6

[25] David González, Joshué Pérez, Vicente Milanés, and Fawzi Nashashibi. A review of motion planning techniques for au-

tomated vehicles. *IEEE Transactions on intelligent transportation systems*, 17(4):1135–1145, 2015. 2

[26] 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. 6, 7

[27] 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. 3

[28] 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. 4

[29] 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

[30] 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

[31] 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. 4

[32] 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. 3, 5, 6

[33] Keishi Ishihara, Anssi Kanervisto, Jun Miura, and Ville Hautamaki. Multi-task learning with attention for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2902–2911, 2021. 3

[34] 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

[35] Bernhard Jaeger, Kashyap Chitta, and Andreas Geiger. Hidden biases of end-to-end driving models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8240–8249, 2023. 3

[36] 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

[37] 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. 3

[38] 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. 3, 6, 7, 8

[39] Anqing Jiang, Yu Gao, Zhigang Sun, et al. Diffvla: Vision-language guided diffusion planning for autonomous driving. *arXiv preprint arXiv:2505.19381*, 2025. 1, 3, 4, 6, 8

[40] 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

[41] Sicong Jiang, Seongjin Choi, and Lijun Sun. Communication-aware reinforcement learning for cooperative adaptive cruise control. *arXiv preprint arXiv:2407.08964*, 2024. 2

[42] 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

[43] Jinkyu Kim, Z. Li, B. Floyd, et al. Textual explanations for self-driving vehicles. In *ECCV*, 2018. 6, 7

[44] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025. 4

[45] 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

[46] 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

[47] 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. 4

[48] Yongkang Li, Kaixin Xiong, Xiangyu Guo, Fang Li, Sixu Yan, Gangwei Xu, Lijun Zhou, Long Chen, Haiyang Sun, Bing Wang, et al. Recogdrive: A reinforced cognitive framework for end-to-end autonomous driving. *arXiv preprint arXiv:2506.08052*, 2025. 4

[49] Zhihao Li, Toshiyuki Motoyoshi, Kazuma Sasaki, Tetsuya Ogata, and Shigeki Sugano. Rethinking self-driving: Multi-task knowledge for better generalization and accident explanation ability. *arXiv preprint arXiv:1809.11100*, 2018. 3

[50] 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

[51] Bencheng Liao, Shaoyu Chen, Haoran Yin, Bo Jiang, Cheng Wang, Sixu Yan, Xinbang Zhang, Xiangyu Li, Ying Zhang, Qian Zhang, and Xinggang Wang. Diffusiondrive:

Truncated diffusion model for end-to-end autonomous driving. *arXiv preprint arXiv:2411.15139*, 2024. 3

[52] 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, 4

[53] Wenru Liu, Pei Liu, and Jun Ma. Dsdrive: Distilling large language model for lightweight end-to-end autonomous driving with unified reasoning and planning. *arXiv preprint arXiv:2505.05360*, 2025. 3

[54] 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. 3

[55] 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

[56] 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

[57] 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. 3

[58] Ashish Mehta, Adithya Subramanian, and Anbumani Subramanian. Learning end-to-end autonomous driving using guided auxiliary supervision. In *Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–8, 2018. 3

[59] 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

[60] 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

[61] Ming Nie, Renyuan Peng, Chunwei Wang, et al. Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving. *ECCV*, 2024. 4, 6, 7

[62] 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. 3, 4

[63] 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

[64] 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. 3, 4

[65] 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. 4

[66] 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. 3

[67] 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

[68] 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. 4

[69] 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, 4

[70] 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. 4, 7

[71] 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, 5, 6, 7, 8

[72] 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. 3

[73] 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

[74] 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. 3

[75] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 7

[76] 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. 3

[77] 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. 3

[78] 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. 4, 7

[79] Chonghao Sima, Katrin Renz, Kashyap Chitta, et al. Drivelm: Driving with graph visual question answering. *ECCV*, 2024. 1, 6, 7

[80] 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

[81] 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. 4

[82] 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

[83] 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. 4

[84] 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

[85] 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

[86] 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. 3

[87] 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. 3

[88] Yuping Wang, Shuo Xing, Cui Can, Renjie Li, Hongyuan Hua, Kexin Tian, Zhaobin Mo, Xiangbo Gao, Keshu Wu, Sulong Zhou, et al. Generative ai for autonomous driving: Frontiers and opportunities. *arXiv preprint arXiv:2505.08854*, 2025. 2

[89] 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

[90] 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. 4

[91] 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. 3

[92] 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. 3, 4

[93] 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

[94] 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. 4

[95] Yichen Xie, Runsheng Xu, Tong He, Jyh-Jing Hwang, Katie Luo, Jingwei Ji, Hubert Lin, Letian Chen, Yiren Lu, Zhaoqi Leng, et al. S4-driver: Scalable self-supervised driving multimodal large language model with spatio-temporal visual representation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1622–1632, 2025. 3

[96] 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. 4

[97] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2174–2182, 2017. 3

[98] 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. 3, 5, 6

[99] 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

[100] 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. 6, 7

[101] 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. 3, 4, 5, 6, 8

[102] Shuang Zeng, Xinyuan Chang, Mengwei Xie, Xinran Liu, Yifan Bai, Zheng Pan, Mu Xu, and Xing Wei. Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. *arXiv preprint arXiv:2505.17685*, 2025. 3

[103] 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

[104] 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. 4

[105] 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. 3, 5, 6, 7, 8

[106] 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. 4

[107] 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

[108] Rui Zhao, Yuze Fan, Ziguo Chen, Fei Gao, and Zhenhai Gao. Diffe2e: Rethinking end-to-end driving with a hybrid action diffusion and supervised policy. *arXiv preprint arXiv:2505.19516*, 2025. 3

[109] 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, 6, 7

[110] 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. 4

[111] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 4

[112] 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

[113] 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

[114] 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, 4, 5, 6

[115] 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, 3

[116] 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, 5

[117] 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. 4, 6, 7