



Vision-Language-Action Models for Autonomous Driving: Past, Present, and Future

Tianshuai Hu , Xiaolu Liu , Song Wang , Yiyao Zhu , Ao Liang , Lingdong Kong , Guoyang Zhao, Zeying Gong, Jun Cen, Zhiyu Huang, Xiaoshuai Hao, Linfeng Li, Hang Song, Xiangtai Li, Jun Ma, Shaojie Shen, Jianke Zhu, Dacheng Tao, Ziwei Liu , Junwei Liang

WorldBench Team

Equal Contributions Project Lead Corresponding Authors

Autonomous driving has long relied on modular “Perception-Decision-Action” pipelines, where hand-crafted interfaces and rule-based components often break down in complex or long-tailed scenarios. Their cascaded design further propagates perception errors, degrading downstream planning and control. **Vision-Action (VA)** models address some limitations by learning direct mappings from visual inputs to actions, but they remain opaque, sensitive to distribution shifts, and lack structured reasoning or instruction-following capabilities. Recent progress in Large Language Models (LLMs) and multimodal learning has motivated the emergence of **Vision-Language-Action (VLA)** frameworks, which integrate perception with language-grounded decision making. By unifying visual understanding, linguistic reasoning, and actionable outputs, VLAs offer a more interpretable, generalizable, and human-aligned paradigm for driving policies. This work provides a structured characterization of the emerging VLA landscape for autonomous driving. We trace the evolution from early VA approaches to modern VLA frameworks and organize existing methods into two principal paradigms: *End-to-End VLA*, which integrates perception, reasoning, and planning within a single model, and *Dual-System VLA*, which separates slow deliberation (via VLMs) from fast, safety-critical execution (via planners). Within these paradigms, we further distinguish subclasses such as textual *vs.* numerical action generators and explicit *vs.* implicit guidance mechanisms. We also summarize representative datasets and benchmarks for evaluating VLA-based driving systems and highlight key challenges and open directions, including robustness, interpretability, and instruction fidelity. Overall, this work aims to establish a coherent foundation for advancing human-compatible autonomous driving systems.

Project Page: <https://worldbench.github.io/vla4ad>

GitHub Repo: <https://github.com/worldbench/awesome-vla-for-ad>

HuggingFace Leaderboard: <https://huggingface.co/spaces/worldbench/vla4ad>

1 Introduction

The pursuit of fully autonomous driving (AD) has long been a central goal in AI and robotics [38, 73, 124]. Conventional AD systems typically adopt a modular “Perception-Decision-Action” pipeline, where mapping [82, 83], object detection [123, 150, 153, 199], motion prediction [55, 117, 307], and trajectory planning [151, 308] are developed and optimized as separate components. While this design has achieved strong performance in structured environments, its reliance on hand-crafted interfaces and rules limits adaptability in complex [84, 85, 220], dynamic [122, 274, 275], and long-tailed scenarios [74, 197, 302]. Moreover, the sequential cascade is prone to cross-stage error propagation, where perception noise is amplified by downstream reasoning and control, compromising stability and safety.

To mitigate these issues, research has increasingly moved toward end-to-end autonomous driving, where

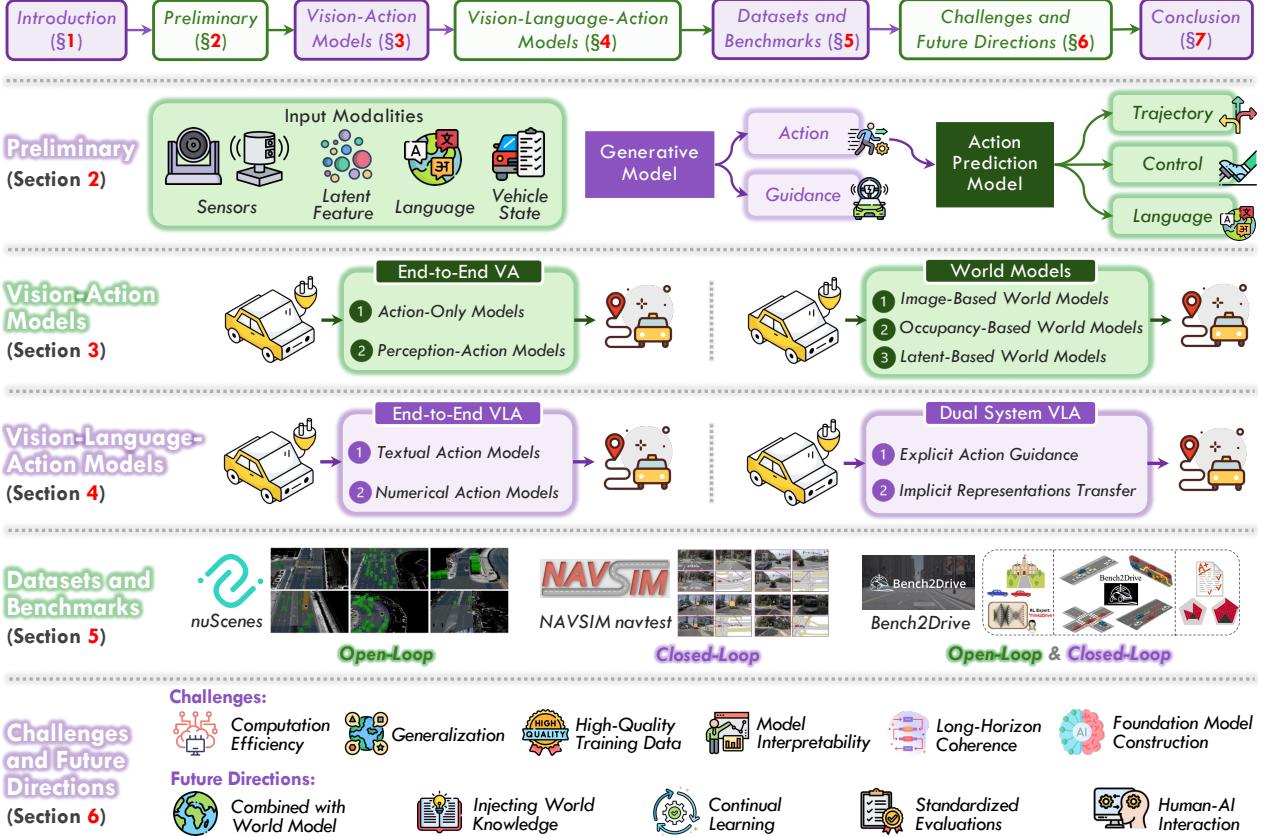


Figure 1 Outline. This work aims to provide a structured roadmap of the VLA paradigm for autonomous driving. We begin with **Preliminary Foundations** (Section 2), which formalize the general formulation of VLA models and detail their three core components: the multi-modal input modalities, the VLM backbone, and the action prediction head. It then traces the evolution from **VA Models** (Section 3), which directly map perception to control, towards **VLA Models** (Section 4), which incorporate language-grounded reasoning and interpretable decision-making. We further categorize VLA architectures into two major paradigms – **End-to-End VLA** (Section 4.1) and **Dual-System VLA** (Section 4.2) – that differ in their integration of vision, language, and action modules. Next, we review **Datasets & Benchmarks** (Section 5) that enable both open-loop and closed-loop evaluation of driving intelligence. Finally, we discuss **Challenges & Future Directions** (Section 6), highlighting interpretability, reasoning, and human-AI interaction as central themes driving the next generation of VLA-based autonomous driving research.

Vision-Action (VA) models directly map raw sensory inputs to control commands or trajectory waypoints using imitation [43, 198, 204] and reinforcement learning [68, 151, 281]. Early systems such as ALVINN [203] and ChauffeurNet [10] demonstrated the viability of behavior cloning at scale. Subsequent advances introduced more expressive architectures: TransFuser [40] exploited transformer-based multimodal fusion, UniAD [93] unified perception and planning, VAD [108] leveraged vectorized scene representations, DriveTransformer [104] explored scalable transformer backbones, and DiffusionDrive [156] applied generative modeling to multi-modal trajectory prediction. Collectively, these VA models show that complex driving policies can be learned directly from data, laying the foundation for modern end-to-end AD systems.

Despite these successes, VA models exhibit fundamental limitations. They largely behave as “black boxes”, offering limited interpretability in safety-critical settings [28, 113, 125, 126, 189, 309]. Their generalization remains fragile under rare or long-tail scenarios that are underrepresented in training [3, 28, 69, 110, 146, 309]. By directly mapping perception to low-level actions, they lack chain-of-thought (CoT) reasoning and contextual deliberation [28, 70, 110, 286], limiting their ability to resolve ambiguous or multi-stage interactions. Moreover, their focus on visual inputs prevents them from incorporating high-level plans or human instructions in natural language, leaving a gap in human-vehicle interaction [110, 197, 222, 320].

The emergence of Large Language Models (LLMs) and Large Multimodal Models (LMMs) has catalyzed a new

paradigm: **Vision-Language-Action (VLA)** models [2, 217, 336]. VLA models couple a Vision-Language Model (VLM) backbone with an action-prediction head, enabling direct mapping from multimodal inputs (vision + language) to executable driving actions. By jointly modeling perception, language understanding, and decision-making, VLA frameworks aspire to provide human-like reasoning, interpretability, and instruction-following [67, 72, 340]. Initial explorations such as DriveMLM [264] and GPT-Driver [184] introduced language modules into driving pipelines for high-level decision understanding, paving the way for more integrated designs. Later systems advanced toward closed-loop and reasoning-centric VLA models: LMDrive [222] achieved language-guided closed-loop driving, DriveLM [225] enabled structured reasoning via visual question answering, and DriveGPT4 [287] provided natural-language rationales for decisions. Recent works further investigate tightly coupled reasoning and control, including AutoVLA [340] with fast/slow thinking and GRPO-based optimization [224], and SimLingo [213], which explicitly studies language-action alignment.

End-to-End VLA models, however, must simultaneously *reason* and *act* in real time, creating challenges for latency and safety. This has led to *Dual-System* VLA designs, where high-level decision making is separated from low-level trajectory execution. DriveVLM [249] generates textual rationales or decisions with a VLM while relying on a classical planner for trajectories. VLP [197] tokenizes waypoints and value maps to produce planning-aware latent actions, and Diff-VLA [106] synthesizes language-guided trajectories refined by rule- or optimization-based controllers. InsightDrive [227] integrates causal language reasoning with MPC, assigning *why* to the VLM and *how* to the planner.

Together, these developments signal a paradigm shift from perception-driven pipelines toward systems that jointly reason, understand, and act. Given the rapid evolution of this field, there is a need to consolidate its conceptual foundations, clarify architectural trends, and provide a structured analysis of emerging directions.

Contributions. This work provides a comprehensive characterization of VLA models for autonomous driving. Specifically:

- We chart the evolution from precursor VA models (Section 3) to modern VLA frameworks (Section 4), providing historical context and clarifying the motivations behind this paradigm shift.
- We propose a taxonomy that categorizes VLA architectures into End-to-End (Section 4.1) and Dual-System (Section 4.2) designs, and compare their principles, advantages, and limitations.
- We present an organized synthesis of datasets and evaluation benchmarks relevant to VLA-based driving (Section 5), facilitating consistent and meaningful comparisons.
- We identify key challenges in real-world VLA deployment and outline future research directions (Section 6) to guide progress toward safer and more reliable autonomous systems.

Scope. This work differs from prior studies on VLA models [2, 110, 217, 336] through its domain-specific focus, historical framing, and architectural depth. ¹**Domain-specific focus.** Unlike broader analyses that span robotics or embodied AI [180, 223], our work *focuses exclusively on autonomous driving*, allowing a fine-grained analysis of driving-specific challenges, dataset characteristics, and safety requirements. ²**Historical and conceptual continuity.** We adopt a “Past-Present-Future” narrative that traces the path from early VA models to modern VLA frameworks, emphasizing the motivations and technical lineage underlying the integration of language grounding into perception and control. ³**Fine-grained architectural taxonomy.** Unlike prior high-level overviews [45, 110, 299, 338], we propose a hierarchical taxonomy that differentiates End-to-End and Dual-System VLA models and analyzes how they organize perception, reasoning, and control.

Through this combination of domain specificity, historical depth, and structured architectural analysis, we aim to provide a comprehensive and insightful reference for VLA research in autonomous driving.

Organization. The remainder of this paper is organized as follows. Section 2 introduces the preliminary foundations of VLA frameworks. Section 3 outlines the evolution of VA models. Section 4 presents our taxonomy and analysis of VLA architectures. Section 5 summarizes datasets and benchmarks. Section 6 discusses remaining challenges and future directions. Section 7 concludes this work.

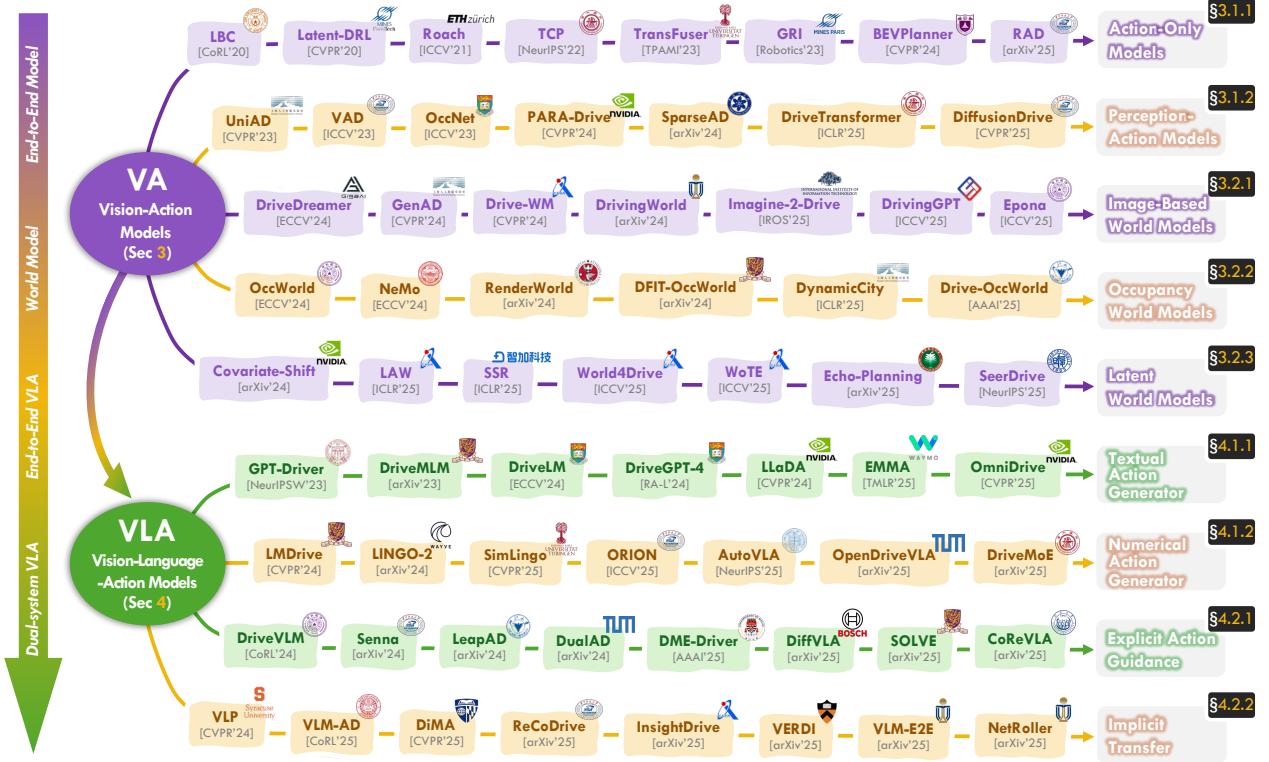


Figure 2 Summary of representative **VA** and **VLA** models from existing literature, spanning *End-to-End Models*, *World Models*, *Dual-Systems*, etc. For the complete list of related approaches and the discussions on their specifications, configurations, and technical details, kindly refer to Section 3 and Section 4, respectively.

2 Preliminary Foundations

Vision-Language-Action (VLA) frameworks [3, 180, 339, 340] leverage large Vision-Language Models (VLMs) [35, 41, 130, 258] to interpret complex driving scenes and produce executable actions. A typical formulation can be expressed as:

$$\mathbf{a}_t = H(F(\mathbf{x}|\theta)) , \quad (1)$$

where \mathbf{x} denotes multimodal inputs at timestamp t , $F(\cdot)$ is a VLM backbone parameterized by θ , and $H(\cdot)$ is an action-generation head. This section introduces these three components: the input modalities (\mathbf{x}), the VLM backbone (F), and the action prediction head (H).

2.1 Input Modalities

The input \mathbf{x} aggregates heterogeneous signals that describe the external environment and the ego-vehicle state [17, 233, 270]. These inputs can be grouped into four categories: sensor observations, latent scene representations, language instructions, and proprioceptive states.

2.1.1 Sensor Inputs

Sensor inputs include raw or preprocessed data directly obtained from vehicle-mounted sensors [24, 25, 141].

- **Visual Images.** Surround-view RGB images that offer dense semantic information: $\mathbf{x}_{\text{img}} \in \mathbb{R}^{N_c \times H \times W \times 3}$, where N_c is the number of cameras (e.g., 6 to 8), and H, W are the height and width of each image.
- **LiDAR Point Clouds.** A sparse or dense set of 3D points representing the environment geometry: $\mathbf{x}_{\text{lidar}} \in \mathbb{R}^{N_p \times D}$, $D \geq 4$, where N_p is the number of points, and D includes dimensions such as x, y, z , velocity, and intensity.

2.1.2 Latent Representations

Multiple VLA systems operate on intermediate spatial representations that fuse multimodal sensor inputs.

- **Bird’s-Eye View (BEV) Features.** Top-down view representation, often generated by fusing camera or LiDAR data [150, 155, 168]: $\mathbf{x}_{\text{bev}} \in \mathbb{R}^{C \times H_{\text{bev}} \times W_{\text{bev}}}$, where C is the number of feature channels, and $H_{\text{bev}}, W_{\text{bev}}$ are the spatial dimensions of the BEV grid.
- **Occupancy Grids.** 3D volumetric representation predicting occupancy and semantics for each spatial location [167, 260, 268, 284]: $\mathbf{x}_{\text{occ}} \in \mathbb{R}^{C_{\text{occ}} \times X \times Y \times Z}$, where X, Y, Z are the spatial resolution of the 3D grid, and C_{occ} denotes the number of occupancy feature channels (*e.g.*, occupancy, flow and semantics).

2.1.3 Language Inputs

To enable VLA capabilities, the model also receives high-level textual instructions or task descriptions [66, 76, 287]. It is composed of a sequence of tokens representing the driving task or goal (*e.g.*, “*turn left at the next intersection*”): $\mathbf{x}_{\text{lang}} \in \mathbb{Z}^T$ or $\mathbf{x}_{\text{lang}} \in \mathbb{R}^{T \times D_{\text{emb}}}$, where T is the sequence length, and D_{emb} is the embedding dimension (if token embeddings are used).

2.1.4 Vehicle State Information

There is also proprioceptive information describing the current dynamic state of the ego-vehicle [17, 149]: $\mathbf{x}_{\text{state}} \in \mathbb{R}^{D_{\text{state}}}$, where D_{state} is the dimension of the state vector, including speed, acceleration, steering angle, yaw rate, turn indicator status, etc.

2.2 VLM Backbone (F)

The VLM backbone $F(\cdot)$ is the core reasoning engine of the system. It is typically a large vision language model [6, 9, 35, 160, 258, 343]. Its primary role is to fuse the diverse input modalities into a single, powerful latent representation. It consists of a vision encoder (*e.g.*, a Vision Transformer, ViT) [210] to process visual inputs and an LLM decoder that conditions its generation on the visual features. A bridge network [35, 160] or unified multimodal token modelling mechanism [6, 258] is used to align the vision features with the language embeddings. VLM can directly generate the actions or provide the guidance for another action expert to develop more robust results.

2.2.1 VLM for Direct Action Generation (Single-System)

In this paradigm, the VLM directly emits actions through its language head [98, 184, 225, 264] or a small attached head [213, 222, 288]. This fully end-to-end design exploits the VLM’s reasoning capabilities to map from visual/language inputs to executable controls.

2.2.2 VLM for Guidance Generation (Dual-System)

Alternatively, the VLM functions as a high-level reasoning module that produces intermediate guidance—textual rationales [80, 207, 249] or structured latent intents [146, 197], which a downstream planner converts into low-level actions [257]. This “slow thinking + fast execution” architecture improves interpretability and enables planners to enforce physical feasibility and safety constraints.

2.3 Action Prediction Head (H)

The head $H(\cdot)$ converts the VLM latent representation into action outputs. Consistent with the taxonomy used in existing literature, we categorize action heads into four types based on their output formulation and generation mechanism: Language Head (LH), Regression (REG), Trajectory Selection (SEL), and Trajectory Generation (GEN).

- **Language Head (LH).** This design directly utilizes the VLM’s inherent text-generation capabilities to produce actions in the language space. The head is typically the language modeling head of the VLM, trained to output either free-form textual commands (*e.g.*, “*turn left*”) [287] or a sequence of discretized

action tokens [340]. The model autoregressively predicts these tokens, which are subsequently parsed into executable signals. This approach is widely adopted in textual action generators like DriveMLM [264] and DriveGPT4 [287].

- **Regression (REG).** This formulation employs a decoder structure followed by a regressor (typically a Multi-Layer Perceptron) to directly predict continuous values. Unlike language heads, it avoids discretization by mapping the latent features aggregated via Transformers or GRUs to specific numerical outputs such as steering angles, throttle/brake values, or trajectory waypoints. Representative methods using this deterministic approach include LMDrive [222] and DriveGPT4-V2 [288].
- **Trajectory Selection (SEL).** Instead of directly regressing a single path, this head evaluates a set of candidate trajectories and selects the optimal one based on a learned cost function or scoring mechanism. The model typically generates or samples a diverse set of dynamically feasible trajectories and uses the latent representation to predict the cost or probability for each candidate. This approach, utilized by methods like WoTE [145] and SeerDrive [315], ensures that the final output adheres to kinematic constraints by selecting from pre-defined candidates.
- **Trajectory Generation (GEN).** This generative formulation synthesizes actions through probabilistic modeling, most notably using diffusion models or variational autoencoder [106, 156, 333]. Starting from noise, the head iteratively refines the trajectory sample conditioned on the VLM latent state and optionally language instructions. This allows the model to capture the multi-modality and uncertainty of future distributions. Prominent examples include ORION [67] and DiffVLA [106].

2.4 Action in Driving

In the context of autonomous driving, particularly for models like VLAs, the action space defines the set of possible outputs the model can generate to control the vehicle. The choice of action representation is a fundamental design decision that dictates how the model’s reasoning is translated into physical motion. We outline three primary paradigms for action space representation below.

2.4.1 Discrete Trajectory Representations

This paradigm represents the vehicle’s intended future path as a finite sequence of spatial waypoints [168]. Each waypoint is a spatial coordinate that the vehicle is expected to reach at a specific future time step. This representation allows for explicit geometric path planning and trajectory optimization. The action, \mathbf{a}_t , formulated at the current time t , is a set of Φ future waypoints:

$$\mathbf{a}_t = \{(x_i, y_i)\}_{i=1}^{\Phi}, \quad \text{where } (x_i, y_i) \in \mathbb{R}^2. \quad (2)$$

Here, Φ is the prediction horizon (the total number of future steps), and each (x_i, y_i) is a coordinate in a 2D Cartesian plane representing the target position at step i .

2.4.2 Continuous Trajectory Representations

Instead of discrete points, this approach parameterizes the vehicle’s motion as a continuous function over a future time horizon [166]. The trajectory is typically defined by functions that govern the vehicle’s longitudinal and lateral motion, such as speed and turning radius. The action, a_t , is defined by these continuous functions over a time interval $[0, T]$:

$$\mathbf{a}_t = (v(t), \kappa(t)), \quad \text{for } t \in [0, T]. \quad (3)$$

In this formulation, $v(t)$ represents the vehicle’s speed profile, and $\kappa(t)$ represents its curvature profile over the future time horizon T . This inherently captures the continuous nature of vehicle dynamics.

2.4.3 Direct Control Representations

This paradigm involves the direct output of low-level vehicle control commands that are immediately sent to the vehicle’s actuators [288]. These outputs typically consist of continuous signals for steering, acceleration,

and braking control. The values are often normalized and constrained to lie within the vehicle’s physical operational ranges. The action vector, a_t , represents control signals for a specific time step t :

$$\mathbf{a}_t = (\delta_t, \tau_t, \beta_t) , \quad (4)$$

where δ_t is the steering angle, τ_t is the throttle input, and β_t is the brake input at time step t . Each component is bound by the vehicle’s hardware limits, *e.g.*, $\delta_t \in [\delta_{\min}, \delta_{\max}]$.

2.4.4 Language Representations

This paradigm leverages the natural language capabilities of VLMs to express driving actions through textual descriptions [19]. The action is represented as a sequence of discrete tokens from a predefined vocabulary:

$$\mathbf{a}_t = \{w_1, w_2, \dots, w_T\}, \quad \text{where } w_i \in \mathcal{V} . \quad (5)$$

Here, \mathcal{V} represents the model’s vocabulary, T is the sequence length, and each token w_i corresponds to an element in the vocabulary. The language-based action can range from high-level commands (*e.g.*, “turn left at the intersection”) to specific numerical trajectory representations encoded as text tokens.

3 Vision-Action Models

Vision-Action (VA) models represent one of the earliest and most influential lines of research in autonomous driving. Their core idea is to directly map sensory observations – typically camera inputs – to driving actions, thereby avoiding explicit modular decomposition into perception, prediction, and planning. Enabled by deep neural networks, VA models have been explored through **two major training paradigms**: ¹*imitation learning*, which distills policies from expert demonstrations, and ²*reinforcement learning*, which optimizes behavior through trial-and-error interaction. More recently, *world models* have expanded this paradigm by enabling agents to simulate scene dynamics and reason about action consequences, improving robustness and scalability. Table 1 provides an overview of representative efforts.

From an architectural perspective, VA methodologies for autonomous driving can be broadly grouped into:

- **End-to-End Models**, which directly predict control commands or planned trajectories from sensory inputs.
- **World Models**, which explicitly model action-conditioned future dynamics to support policy learning and decision-making.

3.1 End-to-End Models for Autonomous Driving

End-to-end (E2E) models learn a single neural network that maps raw or intermediate sensor observations to actions or planned trajectories [16, 28, 38, 145]. Unlike modular stacks, which isolate perception, prediction, and planning, E2E approaches implicitly couple these tasks within a unified representation [91, 109, 235]. Depending on whether perception supervision is employed, existing methods fall into two main categories: *action-only models* and *perception-action models*, as illustrated in Figure 3.

3.1.1 Action-Only Model

Action-only models adopt a streamlined one-stage formulation: sensory inputs are fed directly into a network that outputs low-level actions. These methods primarily differ in whether policies are learned from **demonstrations** or through **exploration**.

Imitation Learning (IL), especially behavior cloning [7], learns a policy by matching expert actions, as visualized in the IL branch of Figure 3. Early works [16, 43, 198, 203] demonstrated that actions can be predicted directly from monocular or multi-view inputs, and subsequent designs refined backbone architectures [24, 40, 205]. NEAT [39] highlights behaviorally relevant image regions via intermediate attention maps, while TCP [273] fuses a trajectory branch and control branch for complementary supervision. To better leverage scene geometry, BEV-Planner [149] predicts trajectories from BEV features enriched with ego states. Urban-Driver [218] moves beyond open-loop evaluation by training policies in a differentiable, data-driven simulator.

IL-based methods are simple, efficient, and require no reward engineering; however, they remain sensitive to distribution shift [56, 206, 215] and causal confusion [49, 190, 200], which can impair reliability in long-tailed or rare-event scenarios.

Reinforcement Learning (RL) optimizes actions through interaction, offering greater flexibility than imitation-based approaches [119, 237]. Several works address the sample inefficiency of RL by combining it with supervised pretraining: Latent-DRL [251] and Gri [23] pre-train visual encoders using semantic segmentation, while LSD [194] initializes policies via IL before performing RL fine-tuning. Privileged-information distillation has also proven effective: LBC [25], WoR [26], and Roach [326] use simulator-only states to guide sensor-based agents.

Combined with the world model, Think2Drive [137] trains the agent with the Model-Based RL (MBRL) method, paired with a compact latent world model learning the transitions of the environment. Raw2Drive [301] is a dual-stream MBRL approach, where the raw sensor world model is aligned with the privileged world model for camera-based action prediction. In contrast to studies in non-photorealistic CARLA [56], recent efforts have shifted toward photorealistic world modeling. RAD [68] establishes a 3DGS-based [116] closed-loop RL training paradigm regulated by IL in a realistic 3DGS environment. The key challenges in RL-based models include sample inefficiency [151], reward function design [120], and sim-to-real transfer [38].

3.1.2 Perception-Action Model

Perception-action models follow a two-stage paradigm in which perception tasks (*e.g.*, mapping, tracking) supervise and constrain trajectory prediction. These methods generally adopt either dense BEV-based representations or sparse query-based representations, as shown in Figure 3.

Table 1 Summary of **Vision-Action** models in autonomous driving.

- **Inputs:** Camera, LiDAR, and Ego-Status.
- **Action Types:** RL: Policy w/ Reinforcement Learning, REG: Decoder + MLP, SEL: Traj. Selection w/ Cost, and GEN: Traj. Generation w/ Generative Model.
- **Outputs:** Ctrl.: Control Signal, Traj.: Numerical Trajectory.
- **Datasets:** CARLA [56], NoCrash [44], ProcGen [42], Lyft [89], nuScenes [17], Bench2Drive [103], NAVSIM [47], OpenOcc [250], OpenDV [296], nuPlan [18], Occ3D [248], Cam4DOcc [178], and Private Data.

#	Model	Venue	Input	Dataset	Vision	Action	Output
• Sec. 3.1.1 Action-only Models							
1	LBC [25]	CoRL'20			ResNet [87]	RL	Ctrl.+Traj.
2	Latent-DRL [251]	CVPR'20			ResNet [87]	RL	Ctrl.
3	NEAT [39]	ICCV'21			ResNet [87]	REG	Traj.
4	Roach [326]	ICCV'21			ResNet [87]	RL	Ctrl.
5	WoR [26]	ICCV'21			ResNet [87]	REG	Ctrl.
6	TCP [273]	NeurIPS'22			ResNet [87]	REG	Ctrl.+Traj.
7	Urban-Driver [218]	CoRL'22			ResNet [87]	REG	Traj.
8	LAV [24]	CVPR'22			ResNet [87]	REG	Ctrl.+Traj.
9	TransFuser [40]	TPAMI'23			ResNet [87]	REG	Traj.
10	GRI [23]	Robotics'23			EfficientNet [238]	RL	Ctrl.
11	BEVPlanner [149]	CVPR'24			ResNet [87]	REG	Traj.
12	Raw2Drive [301]	NeurIPS'25			ResNet [87]	RL	Ctrl.
13	RAD [68]	NeurIPS'25			ResNet [87]	RL	Traj.
14	TrajDiff [75]	arXiv'25			ResNet [87]	GEN	Traj.
• Sec. 3.1.2 Perception-Action Models							
15	ST-P3 [91]	ECCV'22			EfficientNet [238]	SEL	Traj.
16	UniAD [93]	CVPR'23			ResNet [87]	REG	Traj.
17	VAD [108]	ICCV'23			ResNet [87]	REG	Traj.
18	OcNet [250]	ICCV'23			ResNet [87]	SEL	Traj.
19	GenAD [333]	ECCV'24			ResNet [87]	GEN	Traj.
20	PARA-Drive [269]	CVPR'24			ResNet [87]	REG	Traj.
21	Hydra-MDP [148]	CVPRW'24			ResNet [87]	SEL	Traj.
22	SparseAD [317]	arXiv'24			ResNet [87]	REG	Traj.
23	GaussianAD [332]	arXiv'24			ResNet [87]	REG	Traj.
24	DfFSD [231]	arXiv'24			ResNet [87]	GEN	Traj.
25	DriveTransformer [104]	ICLR'25			ResNet [87]	REG	Traj.
26	SparseDrive [235]	ICRA'25			ResNet [87]	REG	Traj.
27	DiffusionDrive [156]	CVPR'25			ResNet [87]	GEN	Traj.
28	GoalFlow [279]	CVPR'25			VoVNet [129]	GEN	Traj.
29	GuideFlow [162]	arXiv'25			ResNet [87]	GEN	Traj.
30	ETA [78]	arXiv'25			CLIP-ViT [210]	REG	Traj.
31	Geo [105]	arXiv'25			ResNet [87]	REG	Traj.
32	DiffusionDriveV2 [345]	arXiv'25			ResNet [87]	GEN	Traj.
33	NaviHydra [272]	arXiv'25			ResNet [87]	SEL	Traj.
34	Mimir [280]	arXiv'25			ResNet [87]	GEN	Traj.
• Sec. 3.2.1 Image-Based World Models							
35	DriveDreamer [265]	ECCV'24			SD [214]	REG	Traj.
36	GenAD [296]	CVPR'24			SDXL [202]	REG	Traj.
37	Drive-WM [267]	CVPR'24			ConvNeXt [173]	SEL	Traj.
38	DrivingWorld [92]	arXiv'24			VQ-VAE [254]	REG	Traj.
39	Imagine-2-Drive [71]	IRROS'25			SVD [15]	SEL	Traj.
40	DrivingGPT [34]	ICCV'25			VQ-VAE [254]	REG	Traj.
41	Epona [321]	ICCV'25			DC-AE [27]	REG	Traj.
42	VaViM [11]	arXiv'25			LLaMAGen [234]	GEN	Traj.
• Sec. 3.2.2 Occupancy-Based World Models							
43	OccWorld [334]	ECCV'24			ResNet [87]	REG	Traj.
44	NeMo [96]	ECCV'24			ResNet [87]	REG	Traj.
45	OccVAR [112]	-			ResNet [87]	REG	Traj.
46	RenderWorld [292]	arXiv'24			Swin-T [172]	REG	Traj.
47	DFIT-OcWorld [318]	arXiv'24			ResNet [87]	REG	Traj.
48	Drive-OcWorld [297]	AAAI'25			ResNet [87]	REG	Traj.
49	T ³ Former [282]	arXiv'25			ResNet [87]	REG	Traj.
50	AD-R1 [289]	arXiv'25			-	RL	Traj.
• Sec. 3.2.3 Latent-Based World Models							
51	Covariate-Shift [204]	arXiv'24			DINOv2 [196]	REG	Traj.
52	World4Drive [335]	ICCV'25			ResNet [87]	REG	Traj.
53	WoTE [145]	ICCV'25			ResNet [87]	SEL	Traj.
54	LAW [143]	ICLR'25			Swin-T [172]	REG	Traj.
55	SSR [134]	ICLR'25			ResNet [87]	REG	Traj.
56	Echo-Planning [232]	arXiv'25			ResNet [87]	REG	Traj.
57	SeerDrive [315]	NeurIPS'25			VoVNet [129]	SEL	Traj.

Dense BEV-Based Models construct unified top-down features from multi-view cameras. ST-P3 [91] jointly learns spatial-temporal features for perception and planning; UniAD [93] integrates sequential task dependencies to support goal-directed planning. VAD [108] employs a vectorized scene representation to improve both planning safety and efficiency. OccNet [250] incorporates occupancy embeddings to capture 3D scene geometry. Para-Drive [269] proposes a fully parallel E2E architecture for real-time deployment.

Generative and sampling-based approaches have recently emerged: GenAD [333] frames planning as sampling from learned distributions; DiffusionDrive [156] introduces a truncated diffusion policy guided by multi-modal anchors; GuideFlow [162] incorporates explicit physical constraints into the generation process. While BEV representations naturally align with 2D trajectory planning, they require substantial computation due to their dense spatial structure.

Sparse Query-Based Models avoid explicit BEV grids by using latent queries to aggregate image features. SparseAD [317] and SparseDrive [235] represent the entire scene using sparse perception queries and a parallel planner, achieving strong efficiency-accuracy trade-offs. DiFSD [231] introduces an ego-centric sparse formulation and models uncertainty through trajectory denoising. DriveTransformer [104] incorporates task parallelism, sparse attention, and streaming updates for improved stability. GaussianAD [332] adopts 3D semantic Gaussians for fine-grained yet compact scene representation.

Sparse query methods significantly reduce inference latency, but the absence of a dense future-world representation can restrict long-horizon reasoning and planning safety.

3.2 World Models for Autonomous Driving

World models aim to predict how driving scenes evolve under different ego actions [77, 121, 291]. By jointly modeling scene dynamics and ego motion, they provide a powerful mechanism for learning safe, long-horizon driving policies [53]. Their applications span immersive simulation [13, 90, 175, 212, 290, 342], end-to-end planning [69, 92, 265, 267], and feature learning for downstream tasks [29, 30, 138, 170, 188, 298, 319]. Here, we focus on world models designed for trajectory planning and categorize them by prediction modality and representation granularity into three groups: **image-based**, **occupancy-based**, and **latent-based** models (Figure 4).

3.2.1 Image-Based World Model

Image-based world models generate future frames conditioned on ego actions, enabling agents to “dream” scene evolution and evaluate the consequences of different trajectories. These methods leverage modern generative models to synthesize realistic, temporally coherent videos and are typically classified into diffusion-based and autoregressive architectures.

Diffusion-Based World Models use latent video diffusion [14, 214] to produce multi-step rollouts. For front-view forecasting, GenAD [296] and Vista [69] incorporate temporal reasoning modules to handle complex motion patterns. Imagine-2-Drive [71] integrates diffusion generation into a reinforcement-learning framework, training a policy actor inside the world model. To support multi-view predictions, DriveDreamer [265] employs a two-stage pipeline for video synthesis and policy learning. Drive-WM [267] factors views within a spatiotemporal

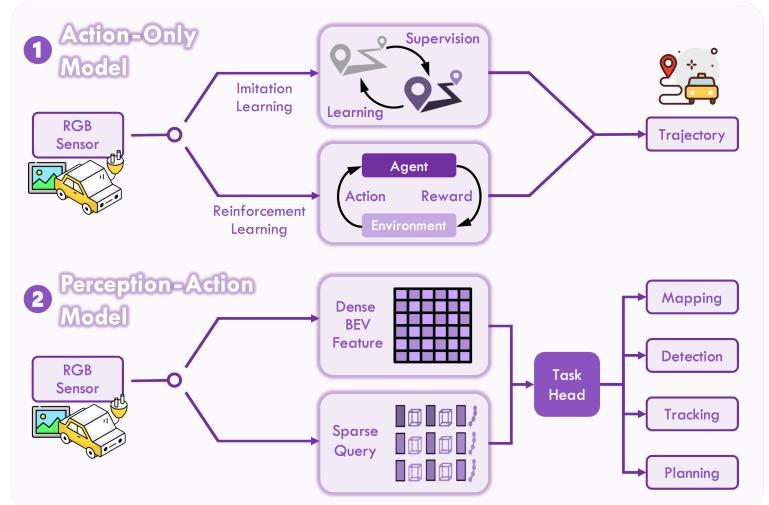


Figure 3 The categorization of **End-to-End VA models** based on model structures and outputs, including **Action-Only Models** (Sec. 3.1.1), and **Perception-Action Models** (Sec. 3.1.2).

model and generates multiple plausible futures, selecting trajectories using image-based rewards.

Autoregressive (AR) Models tokenize images using VQ-VAE [254] and model scene evolution via next-token prediction [59, 128, 305]. DrivingWorld [92] builds a GPT-style architecture for high-fidelity long-horizon video generation. DrivingGPT [34] interleaves image and action tokens, unifying simulation and trajectory planning. Epona [321] combines AR modeling with diffusion to produce high-resolution, long-duration rollouts.

Image-based world models provide photorealistic simulations crucial for training and evaluation. However, their reliance on 2D appearance limits explicit 3D reasoning, which can hinder safety-critical long-horizon planning.

3.2.2 Occupancy-Based World Models

Occupancy-based world models represent the driving scene as spatiotemporal occupancy grids and predict their evolution under different actions. Instead of synthesizing raw pixels, these models focus on the geometry and semantics of free space, obstacles, and agents [154, 171, 342]. As shown in the middle of Figure 4, AR prediction is commonly used in occupancy world models.

OccWorld [334] first introduces occupancy forecasting for planning, using a scene tokenizer to discretize 3D occupancy before applying a GPT-style transformer to synthesize future scenes and ego trajectories. RenderWorld [292] produces 3D occupancy through a self-supervised Gaussian module, while OccVAR [112] performs coarse-to-fine 4D occupancy forecasting. T³Former [282] encodes occupancy using compact triplanes and predicts future triplane updates from multi-scale history.

An alternative line employs single-stage feedforward prediction. Drive-OccWorld [297] uses predicted future BEV features for action-conditioned 3D forecasting. DFIT-OccWorld [318] introduces a decoupled dynamic flow strategy to support efficient non-autoregressive prediction. NeMo [96] improves vision-based occupancy forecasting by combining it with self-supervised image reconstruction signals.

Occupancy-based models offer strong geometric fidelity and explicit free-space reasoning but rely on costly 3D annotations, which can limit scalability across diverse environments.

3.2.3 Latent-Based World Models

Latent-based world models bypass explicit image or occupancy forecasting and instead predict future dynamics directly in a compressed latent space. By operating on high-level features, these models capture behavioral patterns and long-horizon dependencies while avoiding the computational overhead of pixel-level generation. Regarding the forecasting strategy, the latent world model utilizes single-frame or AR prediction presented at the bottom of Figure 4.

Early latent models [8, 337] learn feature-level dynamics for planning without generating visual frames. LAW [143] leverages self-supervised learning to predict future scene features from current features and planned ego trajectories, enabling end-to-end driving without perception labels. World4Drive [335] employs vision foundation models to create latent representations from which diverse planning trajectories can be generated

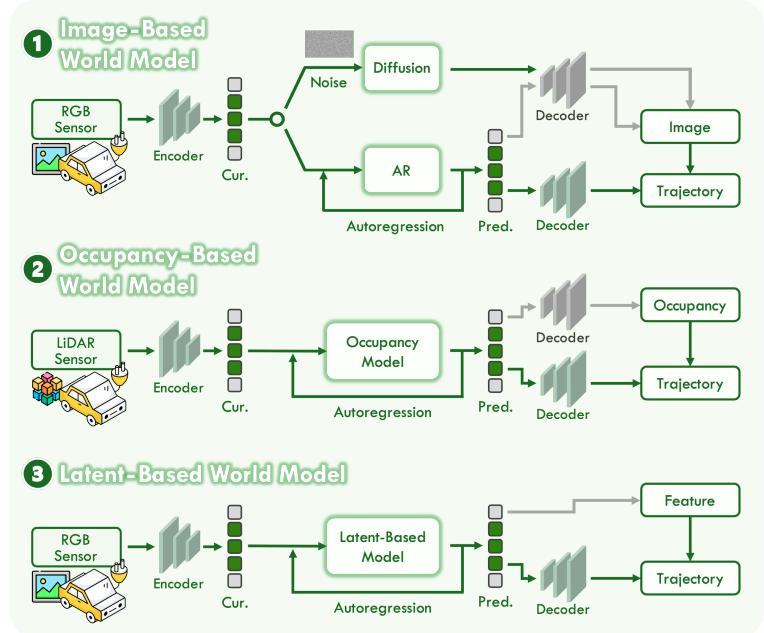


Figure 4 The categorization of **World Models** based on prediction modalities, including *Image-Based Models* (Sec. 3.2.1), *Occupancy-Based Models* (Sec. 3.2.2), and *Latent-Based Models* (Sec. 3.2.3).

and evaluated. Echo-Planning [232] introduces a bidirectional Current→Future→Current (CFC) cycle to enforce temporal consistency in latent BEV features. For robustness in imitation learning, Covariate-Shift [204] addresses distribution mismatch using latent rollouts. By injecting predicted BEV features, SeerDrive [315] refines both latent prediction and trajectory generation in a closed-loop manner.

Latent world models offer efficient and semantically informed forecasting. However, achieving high-quality planning still requires auxiliary supervision from 2D/3D annotations, such as bounding boxes or HD maps.

3.3 Limitations of VA Compared to VLA

While VA models remain widely deployed, they face structural limitations that hinder performance in complex, ambiguous, or long-tailed scenarios: areas where VLA models excel.

- **Limited Interpretability.** VA models provide little insight into their decision-making process. In contrast, VLA models can articulate reasoning steps or explanations through language.
- **Weak Generalization.** VA policies lack broad world knowledge and often require environment-specific retraining. VLA models leverage large-scale pretraining to generalize better under distribution shifts and long-tailed events.
- **No Chain-of-Thought Reasoning.** VA models directly map pixels to actions, making it difficult to perform explicit reasoning or contextual analysis. VLAs natively support step-wise reasoning.
- **No Language Understanding.** VA systems cannot incorporate human instructions or high-level goals expressed in natural language. VLA models naturally integrate such inputs to guide planning and decision-making.

4 Vision-Language-Action Models

Vision-Language-Action (VLA) models extend the Vision-Action paradigm by coupling visual perception with the multimodal reasoning capabilities of large vision-language models. Equipped with chain-of-thought style inference and broad world knowledge, these models are particularly promising for rare, ambiguous, and long-tailed driving scenarios. Table 2 summarizes typical prompting strategies, and Table 3 overviews representative VLA-based approaches.

From an architectural standpoint, current VLA methodologies for autonomous driving can be grouped into two main categories:

- **End-to-End VLA:** a single model directly maps multimodal sensory inputs and language to actions.
- **Dual-System VLA:** a VLM provides high-level reasoning or guidance, while a specialized driving module executes fast, low-level action.

Table 2 Categories of **natural language prompts** for Vision-Language-Action (VLA) models in autonomous driving.

Prompt Type	Explanations
 System Prompt	Text templates or query formulations designed to interact with large language models, guiding them to perform specific driving-related reasoning or trajectory prediction tasks. System prompts often define the task structure, provide role definitions, and shape the model's reasoning behavior.
 Instructions	Commands or instructions provided by humans or systems, typically describing the driving goal or required maneuver (<i>e.g.</i> , “turn left at the next intersection”).
 Scene Description	Textual descriptions of surroundings, including perceived objects, road layout, and contextual factors (<i>e.g.</i> , “ <i>a pedestrian is crossing on the right</i> ”, or “ <i>a vehicle is 5 meters ahead on the left</i> ”).
 Traffic Rules	Prompts encoding regulatory constraints or domain knowledge, such as traffic laws, traffic light status, right-of-way rules, or safety guidelines.
 Ego Status	Information about the ego vehicle’s internal state, including speed, position, heading, or navigation intent.
 Context Information	Demonstrations presented as paired examples of driving scenarios and corresponding actions, used to guide the model via in-context learning.

Table 3 Summary of **Vision-Language-Action** models in autonomous driving.

- **Input:** Camera, Sys. Prompt, Instruct., Scene Descrip., Status, Traffic Rule, Context Info.
- **Action:** LH: Lang. Head, REG: Decoder+MLP, GEN: Traj. GEN with Generative Model.
- **Output:** Desc.: Linguistic Descriptions, Traj.: Numerical Trajectory, Ctrl.: Control Signal, Meta.: Meta Action.
- **Datasets:** n: nuScenes [17], b: BDD-X [118], d: DriveLM [225], s: SDN [183], v: VLAAD [201], B: Bench2Drive [103], w: Waymo [60], m: MetaAD [107], c: Carla [56], i: ImpromptuVLA [36], n: NAVSIM [47], o: OpenDV [296], N: nuPlan [18], T: Talk2Car [50], C: CoVLA [3], P: PhysicalAI-AV [193] and P: Private Data.

#	Model	Venue	Input Modality	Dataset	Vision	Language	Action	Output
• Sec. 4.1.1: Textural Action Generator								
1	DriveMLM [264]	arXiv'23			EVA-CLIP [62]		LH	Desc.+Meta.
2	RAG-Driver [309]	RSS'24			CLIP [210]		LH	Desc.+Ctrl.
3	RDA-Driver [97]	ECCV'24			BEVFusion [155]		LH	Desc.+Traj.
4	DriveLM [225]	ECCV'24			BLIP-2 [133]		LH	Meta.+Traj.
5	DriveGPT4 [287]	RA-L'24			CLIP [210]		LH	Desc.+Ctrl.
6	DriveVLM [95]	IRO'S24			CLIP [210]		LH	Desc.+Ctrl.
7	LLaDA [131]	CVPR'24			-		LH	Ctrl.+Traj.
8	VLAAD [201]	WACV'24			BLIP-2 [133]		LH	Ctrl.
9	OccLLaMA [268]	arXiv'24			Swin-T [172], PointPillar [127]		LH	Ctrl.+Traj.
10	Doe-1 [331]	arXiv'24			Lumina-mGPT [158]		LH	Ctrl.+Traj.
11	LINGO-2 [245]	-			Wayve Vision Model [246]		LH	Desc.+Traj.
12	SafeAuto [320]	ICML'25			LanguageBind [341]		LH	Desc.+Ctrl.
13	OpenEMMA [278]	WACV'25			Qwen2-VL [258]		LH	Desc.+Traj.
14	ReasonPlan [169]	CoRL'25			SigLIP-2 [314]		LH	Desc.+Traj.+Meta.
15	FutureSightDrive [312]	NeurIPS'25			ViT [57]		LH	Desc.+Traj.
16	ImpromptuVLA [36]	NeurIPS'25			Qwen2.5-VL [6]		LH	Traj.
17	WKER [313]	AAAI'25			EVA-02 [63]		LH	Desc.+Traj.
18	OmniDrive [259]	CVPR'25			EVA-02 [63]		LH	Desc.+Traj.
19	S4-Driver [276]	CVPR'25			PaLi3 [32]		LH	Meta.+Traj.
20	EMMA [98]	TMLR'25			Gemini-VLM [243]		LH	Desc.+Traj.
21	Occ-LLM [284]	ICRA'25			LLaDA-2 [252]		LH	Traj.
22	Sce2DriveX [327]	RA-L'25			OpenCLIP [99]		LH	Ctrl.+Traj.
23	DriveAgent-R1 [330]	arXiv'25			Qwen2.5-VL [6]		LH	Desc.+Meta.
24	Drive-R1 [147]	arXiv'25			InternVL2 [263]		LH	Desc.+Traj.+Meta.
25	FastDriveVLA [19]	arXiv'25			Qwen2.5-VL [6]		LH	Desc.+Traj.
26	WiseAD [323]	arXiv'25			CLIP [210]		LH	Traj.+Ctrl.
27	AutoDrive-R2 [310]	arXiv'25			Qwen2.5-VL [6]		LH	Traj.
28	OmniReason [164]	arXiv'25			EVA-02 [63]		LH	Meta.+Traj.
29	OpenREAD [324]	arXiv'25			Qwen3 [294]		LH	Desc.+Meta.+Traj.
30	dVLM-AD [179]	arXiv'25			SigLIP-2 [253]		LH	Desc.+Traj.
31	PLA [325]	arXiv'25			Sensor Encoder		LH	Desc.+Traj.
32	AlphaDrive [107]	arXiv'25			Qwen2-VL [258]		LH	Desc.+Meta.
33	CoReVLA [61]	arXiv'25			Qwen2.5-VL [6]		LH	Ctrl.+Traj.
• Sec. 4.1.2: Numerical Action Generator								
34	LMDrive [222]	CVPR'24			ResNet [87]		REG	Ctrl.
35	BEVDriver [271]	IROS'25			InterFuser [221]		REG	Ctrl.+Traj.
36	CoVLA-Agent [3]	WACV'25			CLIP [210]		REG	Desc.+Traj.
37	ORION [67]	ICCV'25			EVA-02 [63]		GEN	Desc.+Traj.
38	SimLingo [213]	CVPR'25			InternViT [35]		REG	Ctrl.+Traj.
39	DriveGPT4-V2 [288]	CVPR'25			CLIP [210] SigLIP [314]		REG	Ctrl.+Traj.
40	AutoVLA [340]	NeurIPS'25			Qwen2.5-VL [6]		LH	Traj.
41	DriveMoE [300]	arXiv'25			PaliGemma [12]		GEN	Ctrl.
42	DSDrive [168]	arXiv'25			ResNet [87]		REG	Desc.+Traj.
43	OccVLA [167]	arXiv'25			VQ-VAE [254]		REG	Traj.
44	VDRive [76]	arXiv'25			Qwen2.5-VL [6], CVQ [329]		GEN	Desc.+Traj.
45	ReflectDrive [136]	arXiv'25			LLaDA-V [303]		GEN	Traj.
46	E3AD [240]	arXiv'25			Qwen2.5-VL [6]		REG	Traj.
47	LCDrive [239]	arXiv'25			DINOv2 [196]		LH	Traj.
48	Alpamayo-R1 [266]	arXiv'25			Cosmos-Reason1 [4]		REG	Desc.+Ctrl.+Traj.
49	UniUGP [174]	arXiv'25			Qwen2.5-VL [6]		GEN	Desc.+Traj.
50	MindDrive [236]	arXiv'25			ResNet-34 [87]		GEN	Traj.
51	AdaThinkDrive [177]	arXiv'25			InternVL3 [343]		REG	Traj.
52	Percept-WAM [79]	arXiv'25			InternViT [35]		REG	Traj.
53	Reasoning-VLA [316]	arXiv'25			Qwen2.5-VL [6]		REG	Traj.
54	SpaceDrive [135]	arXiv'25			Qwen2.5-VL [6]		REG	Desc.+Traj.
55	OpenDriveVLA [339]	AAAI'26			BEVFormer [150]		LH	Ctrl.+Traj.
• Sec. 4.2.1: Dual-System: Explicit Action Guidance								
56	DriveVLM [249]	CoRL'24			-		REG	Desc.+Traj.
57	LeapAD [187]	NeurIPS'24			QwenVL [5]		LH	Ctrl.+Traj.
58	FasionD [207]	arXiv'24			GenAD [333]		GEN	Ctrl.+Traj.
59	Senna [109]	arXiv'24			VADv2 [31], ViT [57]		REG	Desc.+Traj.
61	DME-Driver [80]	AAAI'25			UniAD [93]		REG	Desc.+Traj.
62	SOLVE [33]	CVPR'25			EVA-02 [63]		REG	Ctrl.+Traj.
63	Real-AD [176]	ICCV'25			UniAD [93], VAD [108]		REG	Ctrl.+Traj.
64	LeapVAD [181]	TNNLS'25			QwenVL [5] InternVL2 [263]		LH	Traj.
65	DiffVILA [106]	arXiv'25			CLIP [210]		GEN	Traj.
66	FasionAD++ [208]	arXiv'25			BEVFormer [150]		GEN	Ctrl.+Traj.
• Sec. 4.2.2: Dual-System: Implicit Representations Transfer								
67	VLP [197]	CVPR'24			UniAD [93], VAD [108]		REG	Traj.
68	VLM-AD [286]	CoRL'25			UniAD [93], VAD [108]		REG	Traj.
69	DIMA [88]	CVPR'25			UniAD [93], VAD [108]		REG	Ctrl.+Traj.
70	ALN-P3 [182]	arXiv'25			VAD [108]		REG	Desc.+Traj.
71	VERDI [64]	arXiv'25			VAD [108]		REG	Ctrl.+Traj.
72	VLM-E2E [166]	arXiv'25			EfficientNet [238]		REG	Traj.
73	ReCogDrive [146]	arXiv'25			InternViT [35]		GEN	Desc.+Traj.
74	InsightDrive [227]	arXiv'25			ResNet [87]		REG	Traj.
75	NetRoller [277]	arXiv'25			CLIP [210]		REG	Traj.
76	ViLaD [46]	arXiv'25			SigLIP-2 [253]		GEN	Ctrl.+Traj.
77	OmniScene [165]	arXiv'25			ResNet [87]		REG	Traj.
78	LMAD [226]	arXiv'25			VAD [108]		LH	Desc.+Meta.

4.1 End-to-End VLA for Autonomous Driving

End-to-end VLA frameworks aim to unify perception, reasoning, and planning within a single architecture. By leveraging the generalization ability of multimodal large language models (MLLMs), they directly transform multimodal observations into actions, reducing reliance on hand-crafted modules and task-specific heuristics. According to the form of their outputs, existing approaches can be broadly divided into two families, as illustrated in Figure 5: **textual action generators**, which operate primarily in the language space, and **numerical action generators**, which predict trajectories or controls in a continuous or discretized numeric space.

4.1.1 Textual Action Generator

Textual action generators formulate driving as a reasoning problem in the language space. The model produces human-readable symbolic decisions, allowing it to “think” and justify its outputs in words. Depending on the abstraction level of these outputs, existing methods can be grouped into *meta-actions* and *trajectory waypoints*.

Meta-Actions are discrete, semantic driving decisions, such as “accelerate”, “stop”, or “change lane”. They form an interpretable interface between high-level reasoning in VLMs and downstream controllers. Early works mainly used language models to output free-form text or conceptual descriptions, which are not directly executable. DriveMLM [264] narrows this gap by aligning LLM outputs with behavioral planning states in a modular stack, enabling language models to act as intermediate planners whose symbolic decisions can be converted into control commands.

Subsequent methods strengthen robustness and reasoning-planning alignment with reinforcement learning and chain-of-thought supervision [107, 147, 224]. AlphaDrive [107] introduces Group Relative Policy Optimization (GRPO) [224] to refine meta-actions using rewards that jointly consider trajectory quality, decision correctness, and format consistency. DriveAgent-R1 [330] first fine-tunes on a curated CoT dataset to encourage step-wise visual reasoning, then applies RL with trajectory- and meta-action-based rewards to bias reasoning paths toward decisions that are practically useful for driving. Recognizing that single-frame front-view inputs limit temporal and spatial understanding, Sce2DriveX [327] further incorporates multi-view video streams and BEV representations, enabling context-aware meta-decisions that are consistent with road topology and spatiotemporal dynamics.

Trajectory Waypoints-based textual generators frame motion planning as the prediction of future coordinates expressed in natural language, thereby unifying reasoning and trajectory forecasting within a single linguistic sequence. DriveLM [225] is an early representative of this paradigm, modeling autonomous driving as graph-structured visual question answering and generating textualized trajectory waypoints conditioned on multi-stage perception, prediction, and planning. Building on this idea, subsequent works adopt end-to-end multimodal formulations. EMMA [98] integrates camera observations and navigation commands into a unified language-driven pipeline for joint perception, road-graph understanding, and trajectory prediction. To enhance robustness in challenging scenarios, ImpromptuVLA [36] introduces an 80K-clip corner-case dataset, demonstrating that pretraining on diverse edge cases significantly improves trajectory accuracy and closed-loop stability. LightEMMA [209] further benchmarks 12 vision–language models, revealing clear trade-offs between interpretability and numerical precision.

A complementary research direction focuses on better aligning reasoning with decision-making. RDA-Driver [97] enforces consistency between chain-of-thought explanations and trajectory outputs through tailored

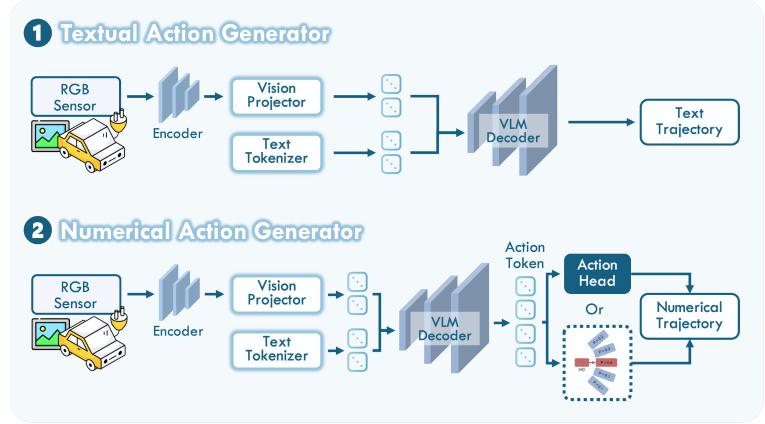


Figure 5 The categorization of **End-to-End VLA models** based on the form of model outputs, including *Textual Action Models* (Sec. 4.1.1), and *Numerical Action Models* (Sec. 4.1.2).

constraints, while Drive-R1 [147] leverages reinforcement learning to improve alignment between textual reasoning and waypoint prediction. Beyond alignment, efficiency and knowledge integration are explored by FastDriveVLA [19] via token pruning, WiseAD [323] through explicit driving priors, and OmniDrive [259] using counterfactual reasoning. WKER [313] further enhances robustness under occlusion by combining instruction-guided token selection with external knowledge sources.

Overall, textual action generators offer strong interpretability and rich reasoning but must bridge a fundamental gap between discrete language tokens and continuous control spaces. This mismatch can introduce precision limits and, in extreme cases, unstable or collapsed trajectories.

4.1.2 Numerical Action Generator

Numerical action generators augment VLM backbones with mechanisms that produce directly usable numeric outputs. The model still leverages language-driven reasoning internally, but its final predictions are expressed as trajectories, waypoints, or control values that can be consumed by classical planners or low-level controllers. Two main realizations exist: *additional action heads* attached to the backbone, and *additional action tokens* that discretize continuous actions into a token space.

Additional Action Head. A common strategy is to attach specialized prediction heads to vision-language models. BEVDriver [271] couples a multimodal encoder with a GRU-based head over BEV features, linking language-grounded reasoning with spatial waypoint prediction. CoVLA-Agent [3] uses a lightweight MLP head trained on the CoVLA dataset, demonstrating that joint supervision from trajectories and captions can simultaneously improve interpretability and numeric accuracy. DriveGPT4-V2 [288] augments token-based planning with an MLP that maps multimodal embeddings to continuous trajectories, enhancing sample efficiency while retaining GPT-style reasoning.

To specialize behaviors, DriveMoE [300] employs a Mixture-of-Experts design whose action head dynamically activates experts for skills such as lane following or overtaking. DSDrive [168] proposes a dual-head coordination module, with one head predicting waypoints and another generating reasoning outputs; distillation from larger VLMs keeps the model compact yet interpretable. LMDrive [222] integrates multimodal encoders with an MLP that directly outputs control signals in a closed loop, marking one of the first instruction-following, language-guided end-to-end systems.

Beyond simple MLPs, ORION [67] replaces deterministic heads with a diffusion-based predictor, modeling multi-modal trajectory distributions under uncertainty. SimLingo [213] decouples temporal speed waypoints from geometric path waypoints via a disentangled MLP head, enabling finer-grained control.

Additional Action Tokens. Instead of explicit heads, some works reuse the language token space to represent actions. AutoVLA [340] discretizes continuous trajectories into a codebook of action tokens, which are autoregressively generated alongside reasoning tokens, thereby unifying semantic reasoning and planning within a single sequence. Reinforcement fine-tuning penalizes redundant reasoning and improves token efficiency. OpenDriveVLA [339] follows a similar token-based paradigm but grounds token generation in a hierarchical alignment between 2D/3D perception and the language model. Structured features are embedded into a unified semantic space, and interaction tokens for the ego vehicle, environment, and other agents are autoregressively decoded into driving actions.

Numerical action generators are well-suited for downstream control, as their outputs are natively compatible with planners and actuators. However, they typically sacrifice some interpretability and often require substantial supervised data for stable training. When discretized action tokens are used, quantization artifacts can further limit fine-grained control accuracy.

4.2 Dual-System VLA

Dual-system VLA frameworks draw inspiration from the dual-process theory popularized by *Thinking, Fast and Slow* [114]. In this paradigm, a VLM serves as the *slow*, deliberative system that performs high-level reasoning, situational assessment, and linguistic inference, while a specialized autonomous driving module acts as the *fast* system responsible for real-time, low-latency trajectory generation and control. By combining these

complementary strengths, dual-system frameworks aim to achieve both interpretability and safety-critical reactivity.

Depending on how VLM outputs interact with the specialized planner, existing methods can be categorized into two families: **explicit action guidance** and **implicit representation transfer**, as illustrated in Figure 6.

4.2.1 Explicit Action Guidance

Explicit action guidance frameworks use VLMs as structured action generators, whose high-level outputs are subsequently transformed or refined by the fast driving module. These approaches differ in their abstraction level and are mainly grouped into **meta-action guidance** and **waypoint supervision**.

Meta-Action Guidance resorts to VLMs to output symbolic driving intentions, such as “slow down”, “change lane”, or “turn left”, which act as semantic priors

for downstream planners. This design leverages the interpretability of linguistic actions while avoiding the precision challenges of directly generating continuous trajectories. Early work such as FasionAD [207] embodies the dual-process design by pairing a fast, data-driven planner with a slow VLM that issues meta-actions; a learned switching mechanism selects the appropriate pathway based on confidence and scene context. LeapVAD [181] refines this structure by combining an analytic branch that builds a memory bank with a heuristic branch that retrieves prior meta-actions for familiar situations.

More recent systems integrate high-level reasoning more tightly with planning. Senna [109] couples a commonsense VLM with an end-to-end planner: Senna-VLM produces natural-language decisions, which Senna-E2E converts into executable trajectories. DiffVLA [106] injects VLM-generated lateral and longitudinal decisions as one-hot priors into a diffusion-based planner, guiding multi-modal trajectory denoising.

Hierarchical frameworks such as DME-Driver [80] further separate decision and execution: a VLM-based Decision-Maker supplies meta-decisions or visual attention priors, and a dedicated Executor translates them into fine-grained control. ReAL-AD [176] extends this to a full three-layer hierarchy: strategy, decision, and operation, where VLM-derived situational insights shape progressively refined planning commands.

Waypoint Supervision is an explicit guidance that uses VLMs to generate coarse trajectory waypoints, which the fast planning module refines into dense, executable trajectories. DriveVLM [249] adopts a hierarchical reasoning-to-planning pipeline: the VLM produces meta-actions and coarse waypoints through chain-of-thought reasoning, and conventional planners transform them into detailed trajectories.

SOLVE [38] strengthens VLM-planner coordination through a shared vision encoder and a Trajectory Chain-of-Thought module that iteratively refines candidate waypoints before final selection by the E2E planner. These designs provide a tighter numerical interface between reasoning and control, enabling VLMs to influence planning while retaining stability through classical refinement.

Overall, explicit guidance approaches maintain strong interpretability and grant VLM a direct role in decision-making. However, they remain sensitive to the accuracy and consistency of VLM outputs; misaligned or ambiguous commands can propagate downstream and degrade planning safety.

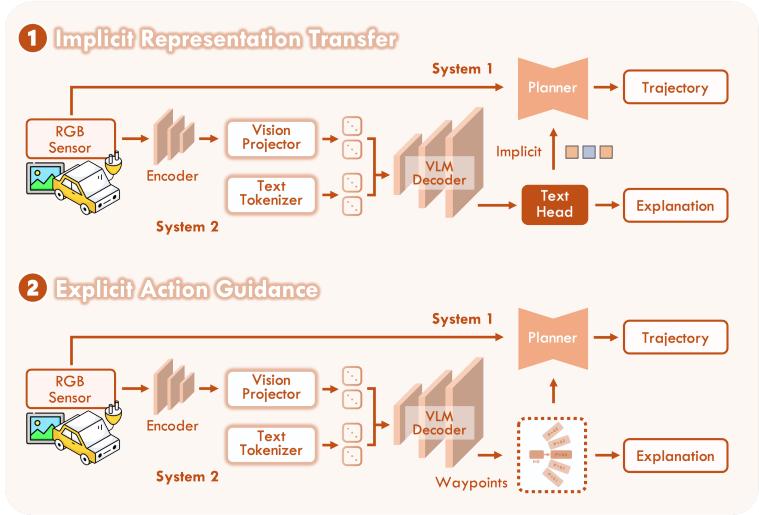


Figure 6 The categorization of **Dual-System VLA Models** based on how VLM interacts with E2E module, including explicit action guidance (Sec. 4.2.1), and implicit representations transfer models (Sec. 4.2.2).

4.2.2 Implicit Representations Transfer

Implicit feature constraint refers to methods where the VLM acts as a teacher or auxiliary module during training, transferring reasoning ability or cognitive priors as latent features to the compact E2E network. These approaches fall into two main groups: **knowledge distillation** and **multimodal feature fusion**.

Knowledge Distillation-based approaches transfer VLM-generated explanations, reasoning traces, or structured action semantics into the latent space of the E2E driving model. VLP [197] aligns BEV features and planning queries with pretrained language embeddings using contrastive and supervisory objectives, enabling planners to inherit commonsense scene understanding. VLM-AD [286] generates free-form textual justifications and structured behavior labels using a VLM, distilling them into the planner through an alignment head and an action classification head. This dual-supervision design helps the E2E module acquire richer semantic representations while remaining computationally light during deployment.

More comprehensive alignment is seen in VERDI [64], which aligns perception, prediction, and planning outputs with VLM-generated chain-of-thought explanations, injecting structured reasoning across all stages of the pipeline. ALN-P3 [182] extends this principle with full-stack co-distillation: perception tokens, predicted motions, and planned trajectories are jointly aligned with VLM reasoning to unify cognition and execution.

Multimodal Feature Fusion-based approaches directly integrate VLM-derived features into the fast planner. InsightDrive [227] introduces language-guided scene representations, where VLM-generated descriptions highlight critical regions and modulate BEV features via cross-attention. VLM-E2E [166] explicitly models driver attention by fusing textual attention cues with BEV features through a learnable gating mechanism. Beyond attention cues, NetRoller [277] extracts latent reasoning variables from VLMs and adapts them into compact features suitable for real-time planners. ReCogDrive [146] aligns linguistic priors with a diffusion-based planner, refining trajectories through reinforcement learning to promote safety and human-like behavior. ETA [78] focuses on efficiency: VLM reasoning is computed asynchronously in earlier frames and fused into current features using an action-mask mechanism, ensuring guidance without incurring high real-time costs.

Implicit transfer methods reduce inference cost and avoid dependence on large VLMs at runtime, but they may sacrifice interpretability, and excessive distillation can oversimplify reasoning signals. Their effectiveness also depends strongly on how well the distilled or fused features align with the capacity of the fast driving module.

5 Datasets & Benchmark

Standardized datasets and benchmarks form the empirical foundation of VLA research, supporting model development, training, and evaluation. Since VLA driving systems integrate perception, language, and action, VLA datasets exhibit substantial diversity in modality composition, annotation granularity, and task definitions. Accordingly, evaluation protocols vary substantially, encompassing conventional trajectory-based metrics, language-centric assessments, and interactive closed-loop evaluations.

5.1 Datasets for VLA in Autonomous Driving

Traditionally, VA datasets provide rich sensory observations (cameras, LiDAR, RADAR) paired with control actions, enabling end-to-end mapping from images to trajectories [17, 18, 304]. These datasets underpin the development of early IL/RL-based VA models.

As language becomes an increasingly important modality for reasoning, instruction following, and explainability, VLA datasets have emerged [225, 249, 262, 264]. These datasets extend traditional driving logs with textual instructions, question-answer pairs, or rationales aligned with visual observations and expert actions [3, 36]. In general, a dataset is considered VLA-compatible when it provides temporally or semantically aligned language annotations that connect visual observations with actions or trajectories, enabling tri-modal learning. The summarized collections are provided in Table 4.

Table 4 Summary of existing **Datasets & Benchmarks** for training and evaluating the VA and VLA models.

- **Vision Sensor Inputs:** Camera, LiDAR point cloud, RADAR point cloud, and Map.
- **Vision Types:** Real: Data collected from real driving scenes, and Sim: Data collected from simulator.
- **Language Annotation Types:** A: Automatic labeling process, and M: Manual labeling process.
- **Action Types:** Traj.: Numerical trajectory output, and Ctrl: Control signal output.
- **Action Metrics:** Open: Open-loop Evaluation, CL: Closed-loop Evaluation, and Lang.: Language-based Evaluation.

Dataset	Year	Vision			Language			Action		Other Tasks
		Sensor	Type	Scale	Category	Label	Scale	Type	Metric	
• Vision-Action Datasets										
B DDD100K [304]	2020		Real	120M	-	-	-	Traj.	-	Percept.
N nuScenes [17]	2020		Real	1.4M	-	-	-	Traj.	Open	Percept.
W Waymo [233]	2020		Real	200M	-	-	-	Traj.	Open	Percept., Forecast.
N nuPlan [18]	2021		Real	4.6M	-	-	-	Ctrl, Traj.	CL	Forecast.
A Argoverse 2 [270]	2021		Real	300K	-	-	-	Traj.	Open	Percept., Forecast.
B Bench2Drive [103]	2024		Sim	2M	-	-	-	Traj.	CL	-
R RoboBEV [274]	2025		Real	866K	-	-	-	Traj.	Open	Percept.
W WOD-E2E [283]	2025		Real	800K	-	-	-	Traj.	Open	-
• Vision-Language-Action Datasets										
B BDD-X [18]	2018		Real	8.4M	Caption	M	26K	Ctrl	Open	Reason.
T Talk2Car [51]	2022		Real	400K	Caption	M	12K	Ctrl, Traj.	Open	Ground.
S SDN [183]	2022		Sim	-	Instruction, QA	A+M	8.4K	Ctrl, Traj.	CL	QA
D DriveMLM [264]	2023		Sim	-	Reason., Deci.	A+M	-	Ctrl, Traj.	CL, Lang.	Reason., QA
L LMDrive [222]	2024		Sim	3M	Instruction	A+M	64K	Traj.	CL	-
B DriveLM-N [225]	2024		Real	4.8K	QA	M	445K	Ctrl, Traj.	Open	Reason., QA
B DriveLM-C [225]	2024		Sim	64K	QA	A	3.76M	Ctrl, Traj.	Open	Reason., QA
H HBD [80]	2024		Real, Sim	-	Deci., Descrip., QA	A+M	-	Traj.	Open	Descrip., QA
V VLAAD [201]	2024		Real	-	Reason., QA	A+M	64K	Ctrl.	Lang.	Caption, QA
S SUP-AD [249]	2024		Real	-	Action, Reason., QA	A+M	-	Ctrl, Traj.	Open, Lang.	Reason, QA
N NuInstruct [54]	2024		Real	11.8K	Instruction	A	91K	Ctrl	Lang.	Reason.
W WOMD-Reason [142]	2024		Real	63K	QA	A	2940K	Plan.	Lang.	Reason., QA
D DriveCoT [262]	2024		Sim	-	CoT, Deci.	A	36K	Ctrl	Open	Reason.
R Reason2Drive [192]	2024		Real	-	Reason., QA	A	632K	Ctrl, Traj.	Open	Reason., QA
D DriveBench [275]	2025		Real	19.2K	QA	A+M	20.5K	Ctrl	Lang.	QA
M MetaAD [107]	2025		Real	120K	Reason., Plan, QA	-	30K	Ctrl	Lang.	Reason.
O OmniDrive [259]	2025		Real	-	Reason., QA	A	-	Ctrl, Traj.	Open	Reason.
N NuInteract [328]	2025		Real	34K	Caption, QA	A	1.5M	Ctrl	Lang.	Percept., Ground.
D DriveAction [86]	2025		Real	2.6K	QA	A	16.18K	Ctrl	Lang.	-
I ImpromptuVLA [36]	2025		Real, Sim	2M	Instruction, QA	A+M	80K	Ctrl, Traj.	Open, CL	QA
C CoVLA [3]	2025		Real	6M	Caption	A	6M	Traj.	Open	-
O OmniReason-N [164]	2025		Real	-	QA	A	-	Ctrl, Traj.	Open	Reason., QA
O OmniReason-B2D [164]	2025		Sim	-	QA	A	-	Ctrl, Traj.	Open	Reason., QA

5.1.1 Vision-Action Datasets

Originally, BDD100K [304] provides 100K diverse driving videos from across the United States, covering a wide spectrum of weather, lighting, and traffic conditions, making it a foundational dataset for behavioral cloning and end-to-end driving. Later, nuScenes [17] offers 1,000 multi-sensor driving scenes with synchronized 6-camera surround views, LiDAR sweeps, radar, 3D boxes, and motion trajectories, supporting both perception tasks and multi-agent motion forecasting. Larger-scale datasets such as the Waymo Open Dataset [60] and Argoverse 2 [270] further extend this paradigm with higher-resolution sensors, longer trajectories, and detailed HD maps, enabling robust training of perception-to-prediction pipelines in diverse urban settings. Complementing these efforts, nuPlan [18] incorporates long-horizon ego trajectories, dense map context, and simulation interfaces for closed-loop testing, providing comprehensive supervision for evaluating decision-making and planning under complex, real-world conditions.

While lacking explicit language supervision, these VA datasets establish the visual-action foundation for VLA development by providing structured supervision that links visual perception, temporal dynamics, and expert decision-making.

5.1.2 Vision-Language-Action Datasets

Building upon the visual-action foundation established by VA datasets, VLA datasets enrich driving logs with structured or free-form natural language to support joint perception-language-action learning.

Table 5 Comparisons of state-of-the-art models for **Open-Loop Planning** on the nuScenes [17] benchmark.

- **Input:** Camera, LiDAR, Prompt, Instruct., Scene Descrip., Status, Rule, Context.
- **Action:** LH: Language Head, RL: Policy w/ Reinforcement Learning, REG: Decoder + MLP, SEL: Trajectory Selection w/ Cost, and GEN: Trajectory Generation w/ Generative Model.
- **Evaluation Metrics:** L2 (↓): L2 Error in meters, and CR (↓): Collision Rate.

Model	Year	Input	Vision	Language	Action	L2 ↓				CR ↓			
						1s	2s	3s	Avg.	1s	2s	3s	Avg.
• Vision-Action Models													
ST-P3 [91]	2022		EfficientNet	-	REG	1.33	2.11	2.90	2.11	0.23	0.62	1.27	0.71
UniAD [93]	2022		ResNet	-	REG	0.44	0.67	0.96	0.69	0.04	0.08	0.23	0.12
VAD [108]	2023		ResNet	-	REG	0.17	0.34	0.60	0.37	0.07	0.10	0.24	0.14
OccNet [250]	2023		ResNet	-	SEL	1.29	2.13	2.99	2.14	0.21	0.59	1.37	0.72
BEV-Planner [149]	2024		ResNet	-	REG	0.30	0.52	0.83	0.55	0.10	0.37	1.30	0.59
Drive-WM [267]	2024		ConvNeXt	-	SEL	0.43	0.77	1.20	0.80	0.10	0.21	0.48	0.26
GenAD [333]	2024		ResNet	-	GEN	0.36	0.83	1.55	0.91	0.06	0.23	1.00	0.43
OccWorld [334]	2024		ResNet	-	REG	0.43	1.08	1.99	1.17	0.07	0.38	1.35	0.60
DriveDreamer [265]	2024		SD	-	REG	-	-	-	0.29	-	-	-	0.15
SparseAD [317]	2024		ResNet	-	REG	0.15	0.31	0.56	0.34	0.00	0.04	0.15	0.06
GaussianAD [332]	2024		ResNet	-	REG	0.40	0.64	0.88	0.64	0.09	0.38	0.81	0.42
LAW [143]	2024		Swin-T	-	REG	0.24	0.46	0.76	0.49	0.08	0.10	0.39	0.19
SSR [134]	2024		ResNet	-	REG	0.18	0.36	0.63	0.39	0.01	0.04	0.12	0.06
Drive-OccWorld [297]	2025		ResNet	-	REG	0.25	0.44	0.72	0.47	0.03	0.08	0.22	0.11
DriveTransformer [104]	2025		ResNet	-	REG	0.19	0.34	0.66	0.40	0.03	0.10	0.21	0.11
DiffusionDrive [156]	2025		ResNet	-	GEN	0.27	0.54	0.90	0.57	0.03	0.05	0.16	0.08
World4Drive [335]	2025		ResNet	-	REG	0.23	0.47	0.81	0.50	0.02	0.12	0.33	0.16
Epona [321]	2025		DC-AE	-	REG	0.61	1.17	1.98	1.25	0.01	0.22	0.85	0.36
SeerDrive [315]	2025		ResNet	-	SEL	0.20	0.39	0.69	0.43	0.00	0.05	0.14	0.06
GuideFlow [162]	2025		ResNet	-	GEN	-	-	-	-	0.00	0.02	0.18	0.07
• Vision-Language-Action Models													
Agent-Driver [185]	2023		-		LH	0.16	0.34	0.61	0.37	0.02	0.07	0.18	0.09
DriveVLM [249]	2024		ViT		GEN	0.18	0.34	0.68	0.40	0.10	0.22	0.45	0.27
DriveVLM-Dual [249]	2024		ViT		REG	0.15	0.29	0.48	0.31	0.05	0.08	0.17	0.10
RAG-Driver [309]	2024		CLIP	Vicuna-1.5	LH	0.34	0.37	0.69	0.40	0.01	0.05	0.26	0.10
Senna [109]	2024		ViT	Vicuna-1.5	REG	0.37	0.54	0.86	0.59	0.09	0.12	0.33	0.18
Doe-1 [331]	2024		Lumina-mGPT	BPE Tokenizer	LH	0.37	0.67	1.07	0.70	0.02	0.14	0.47	0.21
VLP [197]	2024		UniAD, VAD	CLIP	REG	0.30	0.53	0.84	0.55	0.01	0.07	0.38	0.15
VLM-AD [286]	2024		UniAD, VAD	CLIP-ViT	REG	0.24	0.46	0.75	0.48	0.12	0.17	0.41	0.23
OpenDriveVLA [339]	2025		ResNet		LH	0.14	0.30	0.55	0.33	0.02	0.07	0.22	0.10
OmniDrive [259]	2025		EVA-02		LH	0.40	0.80	1.32	0.84	0.04	0.46	2.32	0.94
ORION [67]	2025		EVA-02		GEN	0.17	0.31	0.55	0.34	0.05	0.25	0.80	0.37
EMMA [98]	2025				LH	0.14	0.29	0.54	0.32	-	-	-	-
WKER [313]	2025		EVA-02		LH	0.14	0.30	0.55	0.33	0.07	0.14	0.32	0.18
Drive-R1 [147]	2025				LH	0.14	0.28	0.50	0.31	0.02	0.06	0.19	0.09
ReAL-AD [176]	2025		ResNet		REG	0.30	0.48	0.67	0.48	0.07	0.10	0.28	0.15
ReAL-AD [176]	2025		ResNet		REG	0.35	0.53	0.71	0.53	0.09	0.12	0.31	0.17
DiMA [88]	2025		VAD		REG	0.18	0.50	1.03	0.57	0.00	0.05	0.16	0.08
FasionAD [207]	2025		BEVFormer	VLM+Thinking	GEN	0.19	0.62	1.25	0.69	0.02	0.09	0.44	0.18
InsightDrive [227]	2025		ResNet	VLMs	REG	0.23	0.41	0.68	0.44	0.09	0.10	0.27	0.15
S4-Driver [276]	2025		ViT-G		LH	0.13	0.28	0.51	0.31	-	-	-	-
SOLVE [33]	2025		EVA-02		REG	0.13	0.25	0.47	0.28	0.00	0.16	0.43	0.20
VERDI [64]	2025		VAD		REG	0.36	0.62	0.96	0.65	-	-	-	-
OmniReason [164]	2025		EVA-02		LH	0.15	0.31	0.57	0.34	0.04	0.18	0.98	0.40
FutureSightDrive [312]	2025		ViT		LH	0.14	0.25	0.46	0.28	0.03	0.06	0.21	0.10
Occ-LLM [284]	2025		-		LH	0.12	0.24	0.49	0.28	-	-	-	-
FastDriveVLA [19]	2025				LH	0.14	0.29	0.54	0.33	0.00	0.18	0.70	0.29
AutoDrive-R ² [310]	2025				LH	0.13	0.19	0.25	0.19	-	-	-	-
VDrive [76]	2025				GEN	0.12	0.26	0.50	0.29	0.03	0.16	0.36	0.18
OccVLA [167]	2025		VQ-VAE	PaliGemma-2	REG	0.18	0.26	0.40	0.28	-	-	-	-
FasionAD++ [207]	2024		GenAD	CLIP,	GEN	0.13	0.26	0.45	0.28	0.05	0.08	0.15	0.09
ALN-P3 [182]	2025		VAD	CLIP,	REG	-	-	-	-	0.05	0.09	0.35	0.16
VLM-E2E [166]	2025		EfficientNet	CLIP	REG	0.28	0.50	0.80	0.53	0.01	0.06	0.20	0.09
NetRoller [277]	2025		CLIP		REG	0.38	0.66	1.01	0.68	0.06	0.13	0.30	0.16
OmniScene [165]	2025		ResNet		REG	0.28	0.53	0.91	0.57	0.00	0.04	0.19	0.08
Sce2DriveX [327]	2025		OpenCLIP	Vicuna-1.5	LH	0.15	0.33	0.59	0.36	-	-	-	-
dVLM-AD [179]	2025		SigLIP2	LLaDA-V	LH	0.15	0.40	0.68	0.41	-	-	-	-
Percept-WAM [79]	2025				REG	0.16	0.33	0.60	0.36	-	-	-	-
Reasoning-VLA [316]	2025				REG	0.05	0.19	0.41	0.22	0.02	0.06	0.13	0.07

Representative examples include BDD-X [118], which extends BDD100K with time-aligned human rationales, where annotators describe why the driver performed a specific action. This dataset provides early grounding for language-based explanations. DriveLM [225] constructs graph-structured question-answer pairs based on nuScenes and CARLA scenarios. These QA pairs target conditional reasoning, enabling models to infer high-level intent, spatial relations, and driving decisions. Impromptu VLA [36] aggregates data from eight public driving datasets and supplements them with captions, instructions, and QA pairs aligned with expert trajectories. The focus is on corner cases and long-tailed events. Other datasets, such as LingoQA [186] and CoVLA [3], collect real-world driving videos paired with natural language QA or behavior descriptions, emphasizing spatiotemporal reasoning and human-understandable driving motivations.

Notably, QA-style annotations have emerged as a dominant paradigm for extending driving datasets, serving as a common foundation for training and evaluating reasoning and planning capabilities [86, 225]. However, the scope and assumptions embedded in such annotations naturally influence model behavior, motivating further exploration of more diverse perspectives, planning horizons, and evaluation protocols for real-world deployment.

5.2 Evaluation Metrics

Evaluation metrics differ according to the model’s output modality: **trajectory-based** metrics for continuous action prediction and **text-based** metrics for models producing linguistic commands or rationales.

5.2.1 Trajectory-Based Action Evaluation

Trajectory-based outputs are typically evaluated in open-loop and closed-loop settings.

Open-Loop Evaluation. The predicted trajectory is directly compared against expert trajectories without executing in a simulator. Metrics such as L2 error and collision rate [91], along with trajectory-based indicators including Average Displacement Error (ADE), Final Displacement Error (FDE), and Miss Rate (MR) [93], are widely used. These metrics regard human driving demonstrations as the ground truth and formulate planning essentially as an imitation learning task. By measuring the deviation between predicted and expert trajectories, they provide a straightforward way to assess the accuracy of motion prediction.

Closed-Loop Evaluation. Instead, it measures the model’s performance when interacting with a simulation environment (*e.g.*, CARLA [56]). Representative metrics include route completion (RC), driving score (DS), and infraction distance (ID). Bench2Drive [103] further considers success rate, efficiency, and comfort. NAVSIM [47], built on nuPlan [18], introduces the Predictive Driver Model Score (PDMS), which aggregates subscores for ego progress (EP), time-to-collision (TTC), and comfort (C), while applying penalties on collisions (NC) and driving admissibility (DAC). These metrics provide a holistic view of the safety, feasibility, and deployability of planning actions.

5.2.2 Text-Based Action Evaluation

For low-level vehicle control expressed in natural language, evaluation covers both linguistic quality and control effectiveness. Standard text metrics, such as BLEU, ROUGE, and CIDEr, are commonly used to assess the quality of generated language [225, 328], which measures n-gram overlap with human-annotated reference commands. Beyond command accuracy, reasoning quality is assessed through rationale consistency [192] and human preference ratings of language explanations, particularly in datasets following the BDD-X [118] format. To assess driving applications, execution-based metrics are introduced for behavior assessment. SimLingo [213] introduces an action-dreaming benchmark. The corresponding actions are mapped from the input instruction, which is open-loop evaluated using the success rate.

Regardless of output modality, these benchmarks emphasize key aspects of action quality, including accuracy, executability, safety, and intention alignment.

5.3 Quantitative Experiments & Analyses

This section reviews quantitative benchmarks for evaluating VLA models across action prediction, planning accuracy, and closed-loop driving performance. Among them, nuScenes [17], NAVSIM [47], and Bench2Drive [103]

Table 6 Comparisons of state-of-the-art models on the WOD-E2E [283] test split.

- **Input:** Camera, Sys. Prompt, Instruct., Scene Descrip., Status, Traffic Rule, Context Info.
- **Action:** LH: Language Head, RL: Policy with Reinforcement Learning, REG: Decoder + MLP, and GEN: Trajectory Generation w/ Generative Model.
- **Evaluation Metrics:** RFS (Overall/Spotlight) (↓): Rater Feedback Score, ADE 5s/3s (↓): Average Displacement Error.

Model	Year	Input	Vision	Language	Action	RFS(Overall)(↓)	RFS(Spotlight)(↓)	ADE 5s(.)	ADE 3s(.)
• Vision-Action Models									
Waymo Baseline [283]									
Waymo Baseline	2025	-	-	-	-	7.53	6.60	3.02	1.32
Swin-Trajectory [283]	2025	Camera	SwinT	-	REG	7.54	6.68	2.81	1.21
DiffusionDrive [156]	2025	Camera	ResNet	-	GEN	7.69	6.65	2.99	1.31
RAP-DINO [65]	2025	Camera	DINO	-	REG	8.04	7.20	2.65	1.17
• Vision-Language-Action Models									
OpenEMMA [278]	2025	Camera, Instr., Scene Descrip., Status	Qwen2-VL	Qwen2-VL	LH	5.16	4.71	12.74	6.68
HMVLM [256]	2025	Camera, Instr., Scene Descrip., Status	ViT	Qwen2.5-VL	LH	7.74	6.73	3.07	1.33
AutoVLA [340]	2025	Camera, Instr., Scene Descrip., Status	Qwen2.5-VL	Qwen2.5-VL	LH	7.56	6.94	2.96	1.35
Poutine [216]	2025	Camera, Instr., Scene Descrip., Status	ViT	Qwen2.5-VL	LH	7.99	6.89	2.74	1.21
LightEMMA [209]	2025	Camera, Instr., Scene Descrip., Status	Qwen2.5-VL	Qwen2.5-VL	LH	6.52	5.71	3.73	1.71
dVLM-AD [179]	2025	Camera, Instr., Scene Descrip., Status	SigLIP2	LLaDA-V	LH	7.63	-	3.02	1.29

are the most widely used. More recently, WOD-E2E [283] introduces long-tail, safety-critical scenes with human-preference annotations, enabling more robust assessment of modern E2E and VLA systems.

5.3.1 nuScenes Benchmark

The nuScenes open-loop benchmark evaluates planning quality using trajectory-based metrics, including L2 displacement error and Collision Rate, as summarized in Table 5. Basically, vision-action models such as UniAD [93] reports 0.69m L2 and 0.12 collision rate. Incorporating language generally improves performance by providing semantic cues for safer planning. For instance, Drive-R1 [147] combines supervised CoT alignment with RL finetuning to reach 0.31m L2 and 0.09 collision rate.

Beyond accuracy, recent studies explore the role of language in handling complex and long-tailed driving scenarios. While improvements are most evident in common cases, rare and highly complex situations remain an active area of investigation, motivating the integration of richer reasoning signals and data sources.

From a systems perspective, computational efficiency is an important consideration for practical deployment. Lightweight and efficiency-oriented designs, such as InsightDrive [227] (16.3 FPS) and token-pruned architectures like FastDriveVLA [19], illustrate ongoing efforts to balance model capacity with real-time feasibility. For cross-domain evaluation, nuScenes highlights generalization to unseen cities and distribution shifts as a key benchmark dimension. Works such as VLP [197] and DiMA [88] examine this setting and motivate complementary strategies including domain adaptation, distillation, and data augmentation.

5.3.2 WOD-E2E Benchmark

The Waymo Open Dataset for End-to-End Driving (WOD-E2E) [283] is a large-scale benchmark designed to evaluate end-to-end driving systems under long-tail, safety-critical scenarios that rarely appear in conventional datasets. It contains 4K segments with high-level routing commands, ego-status signals, and multi-camera views, enabling rigorous assessment of perception-planning coupling. A key contribution of WOD-E2E is the Rater Feedback Score (RFS), which measures trajectory quality based on alignment with human preference annotations rather than logged expert trajectories. As shown in Table 6, RFS (Overall and Spotlight) complements conventional ADE metrics by providing a more human-aligned assessment of driving behavior.

Overall results indicate that while vision-action models achieve stable displacement accuracy, VLA models exhibit more diverse performance. Approaches such as Poutine [216], and dVLM-AD [179] achieve balanced RFS and ADE performance, highlighting the importance of effectively aligning language reasoning with trajectory generation. Figure 7 presents the visualized performance of AutoVLA [340] in the WOD-E2E dataset.

Table 7 Comparisons of state-of-the-art models for **Open-Loop Planning** on the NAVSIM [47] *navtest* benchmark.

- **Input:** Camera, LiDAR, Prompt, Instruct., Scene Descrip., Status, Rule, Context.
- **Action:** LH: Language Head, RL: Policy with Reinforcement Learning, REG: Decoder + MLP, SEL: Trajectory Selection w/ Cost, and GEN: Trajectory Generation w/ Generative Model.
- **Evaluation Metrics:** NC: Navigation Completion, DAC (\uparrow): Driving Accuracy, TTC (\uparrow): Time-To-Collision, Comf. (\uparrow): Comfort, EP (\uparrow): Ego Progress, and PDMS (\uparrow): Perception Driving Metric Score.

Model	Year	Input	Vision	Language	Action	NC(\uparrow)	DAC(\uparrow)	TTC(\uparrow)	Comf(\uparrow)	EP(\uparrow)	PDMS(\uparrow)			
• Vision-Action Models														
TransFuser [40]	2022		ResNet	-	REG	97.7	92.8	92.8	100	79.2	84.0			
UniAD [93]	2023		ResNet	-	REG	97.8	91.9	92.9	100	78.8	83.4			
VADv2 [31]	2024		ResNet	-	REG	97.2	89.1	91.6	100	76.0	80.9			
PARA-Drive [269]	2024		ResNet	-	REG	97.9	92.4	93.0	99.8	79.3	84.0			
LAW [143]	2024		Swin-T	-	REG	96.4	95.4	88.7	99.9	81.7	84.6			
DRAMA [308]	2024		ResNet	-	REG	98.0	93.1	94.8	100	80.1	85.5			
DiffusionDrive [156]	2024		ResNet	-	GEN	98.2	96.2	94.7	100	82.2	88.1			
WoTE [145]	2025		ResNet	-	SEL	98.5	96.8	94.9	99.9	81.9	88.3			
World4Drive [335]	2025		ResNet	-	REG	97.4	94.3	92.8	100	79.9	85.1			
DrivingGPT [34]	2025		VQ-VAE	-	REG	98.9	90.7	94.9	95.6	79.7	82.4			
AD-R1 [289]	2025		-	-	RL	98.7	97.8	94.8	100	87.5	91.9			
SeerDrive [315]	2025		VoVNet	-	SEL	98.8	98.6	95.8	100	84.2	90.7			
Epona [321]	2025		DC-AE	-	REG	97.9	95.1	93.8	99.9	80.4	86.2			
GoalFlow [279]	2025		VoVNet	-	GEN	98.4	98.3	94.6	100	85.0	90.3			
TrajDiff [75]	2025		ResNet	-	GEN	98.1	97.0	94.3	100.0	82.7	88.5			
DiffusionDriveV2 [345]	2025		ResNet	-	GEN	98.3	97.9	94.8	99.9	87.5	91.2			
NaviHydra [272]	2025		ResNet	-	SEL	98.7	98.6	88.7	96.2	100.0	92.7			
Mimir [280]	2025		ResNet	-	GEN	98.2	97.5	94.6	100	83.6	89.3			
• Vision-Language-Action Models														
ReCogDrive [146]	2025		InternViT		Qwen2.5-VL	GEN	98.2	97.8	95.2	99.8	83.5	89.6		
AutoVLA [340]	2025				Qwen2.5-VL		Qwen2.5-VL	LH	99.1	97.1	97.1	99.9	87.6	92.1
ReflectDrive [136]	2025				LLaDA-V		LLaDA-V	GEN	99.7	99.5	99.1	99.9	88.9	94.7
AdaThinkDrive [177]	2025				InternVL3		InternVL3	REG	99.1	98.8	97.2	100.0	87.9	93.0
Percept-WAM [79]	2025				InternViT		InternVL2	REG	98.8	98.6	94.4	99.5	84.8	90.2
Reasoning-VLA [316]	2025				Qwen2.5-VL		Qwen2.5-VL	REG	97.8	93.2	98.1	99.8	80.7	91.7

5.3.3 NAVSIM Benchmark

NAVSIM [47] is built on OpenScene (a redistribution of nuPlan [18]), provides a closed-loop simulation environment designed to evaluate planning quality under realistic urban driving conditions. It adopts the PDMS metric, which aggregates multiple driving aspects, including No-Collision (NC), Driving Admissibility (DAC), Time-to-Collision (TTC), Ego Progress (EP), and Comfort (C), offering a holistic assessment of safety, efficiency, and driving smoothness. As shown in Table 7, most methods achieve strong performance on safety-related metrics such as NC and DAC, while TTC and EP serve as more discriminative indicators of planning foresight and long-horizon decision quality. These metrics highlight differences in how models balance safety and progress when interacting with dynamic environments.

Representative vision-action approaches, such as WoTE [145], achieve 88.3 PDMS by integrating a BEV-based world model with reward-guided trajectory selection, demonstrating the effectiveness of structured world modeling for closed-loop planning. Building upon this foundation, language-conditioned methods further enhance decision-making. For instance, AutoVLA [340] improves performance to 99.1 NC and 87.6 EP by leveraging language-driven decision priors and a Best-of-N oracle scoring strategy, illustrating how language supervision can guide trajectory selection and improve long-horizon planning behavior.

5.3.4 Bench2Drive Benchmark

Bench2Drive [103] provides a closed-loop evaluation protocol built on CARLA V2, focusing on success rate and composite driving scores to assess goal-directed driving behavior under interactive settings. Unlike open-loop benchmarks, Bench2Drive explicitly evaluates an agent’s ability to execute long-horizon tasks and respond to dynamic environmental feedback.

Recent VLA approaches demonstrate clear benefits from language grounding in this benchmark. For example, SimLingo [213] introduces an action dreaming mechanism that aligns natural language instructions with control sequences, achieving a leading driving score of 85.94, as reported in Table 8. These results indicate

Table 8 Closed-loop and Open-loop performance comparison of E2E-AD Methods on the **Bench2Drive** benchmark.

- **Input:** Camera, LiDAR, Prompt, Instruct., Scene Descrip., Status, Rule, Context.
- **Action:** LH: Language Head, RL: Policy with Reinforcement Learning, REG: Decoder + MLP, SEL: Trajectory Selection w/ Cost, and GEN: Trajectory Generation w/ Generative Model.
- **Evaluation Metrics:** DS (↑): Driving Score, SR (↑): Success Rate. Avg. L2 (↓): Averaged L2 distance of trajectory.

Method	Year	Input	Vision	Language	Action	Closed-Loop			Open-Loop	
						DS↑	SR(%)↑	Efficiency↑	Comfort↑	Avg. L2 ↓
• Vision-Action Models										
TCP [273]	2022		ResNet	-	REG	40.70	15.00	54.26	47.80	1.70
ThinkTwice [102]	2023		ResNet	-	REG	62.44	31.23	69.33	16.22	0.95
DriveAdapter [101]	2023		ResNet	-	REG	64.22	33.08	70.22	16.01	1.01
UniAD-Base [93]	2023		ResNet	-	REG	45.81	16.36	129.21	43.58	0.73
VAD [108]	2023		ResNet	-	REG	42.35	15.00	157.94	46.01	0.91
GenAD [333]	2024		ResNet	-	GEN	44.81	15.90	-	-	-
DriveTransformer [104]	2025		ResNet	-	REG	63.46	35.01	100.64	20.78	0.62
ETA [78]	2025		CLIP	-	REG	69.53	38.64	184.51	28.43	-
WoTE [145]	2025		ResNet	-	SEL	61.71	31.36	-	-	-
GuideFlow [162]	2025		ResNet	-	GEN	75.21	51.36	-	-	-
Raw2Drive [301]	2025		ResNet	-	RL	71.36	50.24	214.17	22.42	-
• Vision-Language-Action Models										
ORION [67]	2025		EVA-02	Vicuna-1.5	GEN	77.74	54.62	151.48	17.38	0.68
AutoVLA [340]	2025		SigLIP	Qwen2.5-VL	LH	78.84	57.73	146.93	39.33	-
SimLingo-Base [213]	2025		InternViT	Qwen2	REG	85.94	66.82	244.18	25.49	-
SimLingo [213]	2025		InternViT	Qwen2	REG	85.07	67.27	259.23	33.67	-
ReAL-AD [176]	2025		ResNet	QwenVL	REG	40.76	10.93	-	-	0.87
ReasonPlan [169]	2025		SigLIP	Qwen	LH	64.01	34.55	180.64	25.63	0.61
DriveMoE [300]	2025		BEV Encoder	LLaMA	REG	74.22	48.64	175.96	15.31	0.31
VDRive [76]	2025		Qwen2.5-VL, CVQ	InternVL3	GEN	66.15	50.51	110.23	22.90	0.55
CoReVLA [61]	2025		Qwen2.5-VL	Qwen2.5-VL	LH	72.18	50.00	145.41	34.35	-

that language-guided reasoning can effectively influence closed-loop decision-making and improve planning performance in interactive driving scenarios.

Taken together with open-loop benchmarks, Bench2Drive highlights the growing importance of language–action alignment in VLA systems, particularly for interpreting high-level goals, guiding long-horizon behavior, and adapting actions under complex, human-centered instructions.

6 Challenges & Future Directions

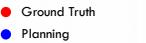
VLA models mark a shift from modular stacks toward holistic, reasoning-driven driving agents. By leveraging large multimodal backbones, they promise richer environmental understanding, stronger generalization, and more interpretable decision-making. Yet, realizing their full potential in safety-critical autonomous driving requires addressing several fundamental challenges. In parallel, emerging research directions point toward next-generation systems that are more efficient, trustworthy, and capable of long-horizon reasoning.

6.1 Current Challenges

6.1.1 Model Architecture and System Efficiency

Real-Time Processing and Latency. VLA models inherit the substantial computational footprint of modern vision-language backbones. High-resolution, high-frame-rate camera inputs generate long visual-token sequences, and multi-view fusion amplifies memory and latency costs. Meeting the strict real-time constraints of autonomous vehicles, therefore, remains difficult [111, 269]. Recent advances in streaming token compression and adaptive visual encoders [19, 139] offer promising directions, but achieving sub-50ms inference remains an unmet requirement for safety-critical deployment.

Lack of Domain-Specific Foundation Models. General-purpose VLMs [6, 35, 343] provide strong priors but are not optimized for driving-specific perception, physics, or multi-sensor fusion. Autonomous driving requires precise spatial reasoning, adherence to traffic rules, and an understanding of rare, high-stakes edge cases – abilities not fully captured by generic models. As highlighted in Section 3, dedicated driving foundation models [152] remain a missing cornerstone for scalable and dependable VLA systems.



<think> This is a complex scenario requiring additional reasoning.

Scene Description:
The scene is set during the daytime in an urban area with clear weather. The road is undergoing construction, as indicated by orange barriers and traffic cones placed along both sides of the road. The front view shows a straight path ahead with no immediate obstacles. The front left and front right views reveal buildings and parked cars on either side of the road. The back view shows a clear path behind the vehicle with no following traffic. The traffic light ahead is green, allowing vehicles to proceed.

Critical Object Description:

1. **Traffic Cones:** Located on both sides of the road, these cones mark the boundaries of the construction zone. They are stationary and serve as visual guides for navigating the narrowed lane. - **What they are doing:** Marking the construction area. - **What they will do:** Remain stationary. - **Why they are important:** They help the ego vehicle stay within the designated lane and avoid entering the construction zone.
2. **Green Traffic Light:** Positioned ahead, it indicates that vehicles can proceed. - **What it is doing:** Allowing vehicles to move forward. - **What it will do:** Stay green unless the traffic signal changes. - **Why it is important:** It dictates the flow of traffic and confirms that the ego vehicle can continue moving forward.

Reasoning on Intent:

- Firstly, the traffic cones indicate that the ego vehicle must stay within the marked lane to navigate the construction zone safely. This requires the vehicle to maintain a steady course and possibly reduce speed to ensure precise navigation.
- Secondly, the **green traffic light** confirms that the ego can proceed without stopping. However, given the construction and narrowed lane, it is prudent to approach cautiously.
- The ego vehicle has been moving forward at a constant speed in the past 4 seconds, and the current driving command is to go straight. Considering the construction zone and the presence of cones, it would be prudent to decelerate slightly to maintain control and safety.

Best Driving Action:
Move forward with a deceleration. This action allows the ego to navigate the construction zone safely while maintaining control and avoiding any potential hazards posed by the narrowed lane and construction equipment. Decelerating slightly ensures that the vehicle can respond promptly to any unforeseen obstacles or changes in the road conditions.

</think>

<answer> The final output action is: <action_119><action_590>... </answer>



<think> This is a complex scenario requiring additional reasoning.

Scene Description:
The scene is a residential area during the daytime with clear weather. The ego vehicle is approaching a T-intersection with a visible "STOP" marking on the road. There are several parked cars along both sides of the street. A few vehicles are visible in the distance, traveling on the intersecting road. No pedestrians are present near the intersection. The traffic appears calm, with no immediate obstacles or hazards.

Critical Object Description:

1. **STOP Marking:** Located directly in front of the ego vehicle on the road. This indicates that the vehicle must come to a complete stop before proceeding.
2. **Vehicles on the Intersecting Road:** Positioned further ahead, beyond the intersection.

Reasoning on Intent:

- Firstly, the presence of the "STOP" marking requires the ego vehicle to come to a complete stop. This is a legal obligation and ensures safety by allowing the vehicle to check for any oncoming traffic or pedestrians before proceeding.
- Secondly, observing the vehicles on the intersecting road helps determine when it is safe to proceed. Since they appear to be either stopped or moving slowly, the ego vehicle can likely proceed once it has come to a full stop and confirmed that the way is clear.
- The ego vehicle has been decelerating in the past 4 seconds, indicating an intention to slow down and prepare for the stop. Given the current velocity and acceleration, a quick deceleration is necessary to bring the vehicle to a halt promptly.
- The driving command is to go straight, which aligns with the ego vehicle's current trajectory and the absence of any immediate obstacles or hazards.

Best Driving Action:
Move forward with a quick deceleration. This action ensures that the ego vehicle comes to a complete stop at the "STOP" marking while maintaining control and safety. After stopping, the vehicle can then assess the situation and proceed if it is safe to do so.

</think>

<answer> The final output action is: <action_990><action_932>... </answer>



<think> This is a complex scenario requiring additional reasoning.

Scene Description:
The scene is a residential street during the daytime with clear weather. The road is lined with parked cars on both sides, and there are trees along the sidewalks. Two children are crossing the road in front of the ego vehicle, one on a scooter and the other walking. The front left and front right cameras show parked cars and houses on either side of the street. There are no visible traffic lights or stop signs in the immediate vicinity.

Critical Object Description:

1. **Two Children crossing the road:** Located in the center of the front view, the child on the scooter is crossing the road while the other child is walking. They appear to be mid-crossing and are likely to finish crossing soon. Their presence is crucial as the ego vehicle must ensure they have safely crossed before proceeding.
2. **Parked Cars:** Located on both sides of the street, these cars are stationary and do not pose an immediate threat. However, they restrict the lateral movement of the ego vehicle and require caution to avoid any potential collisions.

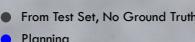
Reasoning on Intent:

- 1. **Two Children crossing the road:** The two children are currently crossing the road, and the ego vehicle must wait until they have safely crossed. Once they have finished crossing, the ego vehicle can proceed.
- 2. **Parked Cars:** The parked cars limit the lateral movement options. The ego vehicle cannot change lanes or make wide turns without risking collision with the parked cars.
- 3. **Driving Command Instruction:** The driving command indicates "go straight", which aligns with the current situation as the ego vehicle is positioned in the center of the road. Given that the two children are expected to finish crossing soon and the road ahead is clear, the ego should prepare to move forward once the children have cleared the path.

Best Driving Action:
Move forward with a quick acceleration. This action is appropriate because once the two children have safely crossed the road, the ego vehicle can quickly accelerate forward to maintain smooth traffic flow while ensuring safety. The quick acceleration helps in efficiently utilizing the road space and maintaining a safe distance from other vehicles.

</think>

<answer> The final output action is: <action_0><action_0>... </answer>



<think> This is a complex scenario requiring additional reasoning.

Scene Description:
The scene is a residential area during a snow-covered day. The road appears to be a two-lane street with houses on both sides. There are no visible vehicles directly in front of the ego vehicle, but there are parked cars along the sides of the road. The front camera shows a clear path ahead with no immediate obstacles. The front left and front right cameras reveal open spaces on either side of the road, with no significant traffic or pedestrians present. The back camera confirms that the road behind is clear as well.

Critical Object Description:

1. **Parked Cars:** Located on both sides of the road, these cars are stationary and pose no immediate threat to the ego vehicle's movement. However, they should be monitored for any potential opening doors or pedestrians emerging.
2. **Residential House:** These structures line both sides of the road and do not directly impact the ego vehicle's path. However, they may obscure the view of pedestrians or other vehicles entering the road.

Reasoning on Intent:

- 1. **Clear Path Ahead:** The front camera indicates a clear path with no immediate obstacles, suggesting that the ego vehicle can proceed forward without interruption.
- 2. **No Vehicles or Pedestrians:** The absence of any vehicles or pedestrians in the immediate vicinity allows the ego vehicle to maintain its current trajectory without needing to adjust speed or direction.
- 3. **Driving Command Instruction:** The current driving command is "go straight," which aligns with the observed conditions and the vehicle's recent behavior of deceleration followed by stopping. Given the clear path ahead, lack of nearby traffic, and the driving command to go straight, the ego vehicle should aim to resume a constant speed.

Best Driving Action:
Move forward with a constant speed. This action ensures the vehicle maintains a steady pace while adhering to the driving command and the current road conditions, which are safe and free of immediate hazards.

</think>

<answer> The final output action is: <action_337><action_1041>... </answer>

Figure 7 Visualization examples of the AutoVLA [340] reasoning/planning results on WOD-E2E [283] dataset.

6.1.2 Data and Generalization

Generalizing to Rare and Novel Scenarios. One motivation for VLAs is their ability to leverage strong visual-language priors to interpret complex scenes. However, while VLM components may generalize well perceptually, aligning this understanding to the action space introduces new uncertainties. As noted in Section 4, reasoning-rich representations do not automatically translate to robust action generation. Long-tailed scenarios – misbehaving traffic agents, unusual road layouts, unpredictable weather – remain failure points [48, 74, 247, 302, 311].

Cost of High-Quality Data. VLAs rely on diverse, high-quality multimodal datasets [132, 163], yet collecting paired vision-action-language triplets at scale is expensive. Synthetic environments [22, 191, 228] help, but face substantial sim-to-real gaps, with discrepancies in noise characteristics, lighting, and behavior of other agents [123, 274]. Improving data efficiency and mitigating distribution shifts remain long-standing challenges.

6.1.3 Core Capabilities and Trustworthiness

Interpretability & Hallucination. While VLA models produce natural-language rationales via chain-of-thought prompting [67, 261, 340], these explanations are generated artifacts – not faithful reflections of the underlying causal reasoning. Language hallucination [94, 159, 211] presents new risks: the model may justify an incorrect decision with a confident but spurious narrative. Ensuring consistent grounding between perceptions, actions, and explanations is an open challenge.

Long-Horizon Temporal Coherence. Driving depends on anticipating multi-stage interactions and maintaining situational awareness across extended time horizons [199, 229, 285]. Current transformer-based VLA architectures remain constrained by limited context windows and short-term conditioning, inherited from standard VLM designs. Temporal fragmentation leads to inconsistent decisions, especially in multi-agent or highly dynamic traffic scenes.

6.2 Future Directions

6.2.1 Next-Generation Model Paradigms

Unified Vision-Language-World Models. A promising evolution integrates VLA with predictive world models [9, 21, 143, 144], extending the VA-based models in Section 3.2. Rather than reacting frame by frame, such systems simulate future scene evolution conditioned on candidate actions, enabling proactive planning and more reliable behavior under uncertainty. Building unified, end-to-end world models that jointly reason about perception, language, and dynamics may form the backbone of next-generation autonomous agents.

Richer Multimodal Fusion. As sensor suites diversify, future architectures will incorporate early and tight fusion of LiDAR, Radar, event cameras, and high-definition maps [123, 222, 274]. Language enhances semantic grounding, but robust 3D geometry is indispensable for safe decision-making [167, 248]. Holistic multimodal fusion can combine the interpretability of VLMs with the spatial precision of geometric sensors.

6.2.2 Advancing Intelligence and Adaptation

Socially Aware, Knowledge-Grounded Driving. VLA models must acquire deeper commonsense reasoning – understanding intent, conventions, and causal relationships beyond explicit annotations [140, 313]. Future efforts will draw from large-scale video-language corpora, leveraging external knowledge bases and structured reasoning modules to support socially compliant and anticipatory driving.

Continual & Onboard Learning. Static, offline-trained models cannot capture evolving road infrastructures or regional driving customs [20, 255, 344]. Enabling safe, incremental learning from everyday driving, while avoiding catastrophic forgetting and ensuring safety guarantees, is essential for long-term deployment. This relates closely to addressing long-tail generalization gaps.

Table 9 Summary of the evaluation metrics used for evaluating the **trajectory-based** and **text-based** action outputs.

Abbr.	-	Full Name	Description	Ref.
Action-Planning Open-Loop Evaluation				
L2	↓	L2 Error	L2 distance error between the planned trajectory and the human driving trajectory in 3 seconds.	[91]
CR	↓	Collision Rate	How often the self-driving vehicle would collide with other agents on the road.	[91]
ADE	↓	Average Displacement Error	Mean displacement error between predicted trajectories and expert waypoints across the horizon, reflecting overall trajectory accuracy.	[93]
FDE	↓	Final Displacement Error	Displacement error at the final predicted waypoint compared with expert trajectories, emphasizing long-term accuracy.	[93]
MR	↓	Miss Rate	Fraction of prediction time steps where displacement error exceeds horizon-specific thresholds, reflecting failure in trajectory coverage.	[93]
AHE	↓	Average Heading Error	Mean absolute angular deviation between predicted and expert heading over the trajectory horizon, measuring orientation accuracy.	[115]
FHE	↓	Final Heading Error	Absolute angular deviation of predicted heading from expert at the final timestep, reflecting terminal orientation accuracy.	[115]
SLE	↓	Speed L1 Error	Mean absolute error of predicted speed control signals.	[100]
SALE	↓	Steer Angle L1 Error	Mean absolute error of predicted steering angle control signals.	[100]
RFS	↑	Rater Feedback Score	Measure how well the predicted trajectory aligns with human driving preferences by checking whether it falls within trust regions.	[283]
Trajectory-Based Closed-Loop Evaluation				
RC	↑	Route Completion	The percentage of route distance completed.	[205]
DS	↑	Driving Score	RC weighted by a penalty factor that accounts for collisions with pedestrians, vehicles, etc.	[205]
NC	↑	No Collision	Fraction of scenarios without ego-fault collisions, focusing exclusively on responsibility-aware collision evaluation.	[47]
DAC	↑	Driving Admissibility Check	Boolean evaluation that checks whether the ego vehicle remains inside drivable polygons throughout the rollout.	[47]
TTC	↑	Time To Collision	Boolean verification that the time-to-collision value exceeds safety thresholds, preventing imminent crashes.	[47]
C	↑	Driving Comfort	The comfort of driving.	[47]
EP	↑	Ego Progress	Penalization of excessive jerk, acceleration, or yaw-rate, reflecting ride quality and passenger comfort.	[47]
PDMS	↑	Predictive Driver Model Score	A flexible weighted evaluation score in autonomous driving that aggregates multiple safety, progress, and comfort subscores into a single metric.	[47]
SR	↑	Success Rate	Percentage of navigation episodes that successfully reach the goal within a fixed time budget, indicating overall task completion.	[56]
ID	↑	Infraction Distance	Average driving distance between two infractions, with longer distances reflecting safer and more reliable policy behavior.	[56]
Text-Based Action Evaluation				
CIDEr	↑	Consensus-based Image Description Evaluation	Measures similarity of generated captions to multiple human references using TF-IDF weighted n-grams.	[182]
BLEU	↑	Bilingual Evaluation Understudy	Precision-based metric that compares n-grams of the generated text against reference texts.	[182]
METEOR	↑	Metric for Evaluation of Translation with Explicit Ordering	Considers unigram precision and recall with stemming, synonym matching, and fragmentation penalty.	[182]
Rouge	↑	Recall-Oriented Understudy for Gisting Evaluation	Recall-focused metric using overlapping n-grams, word sequences, or word pairs between generated and reference texts.	[182]
Top-1 Acc	↑	Visual Question Answering Top-1 Accuracy	Percentage of predictions where the most confident output matches the ground truth label.	[182]

6.2.3 Ecosystem for Safe Deployment

Standardized Evaluation & Safety Guarantees. Evaluation metrics from current benchmarks, *e.g.*, NAVSIM [47] and Bench2Drive [103], assess safety and comfort but do not capture key VLA-specific risks such as reasoning failures, instruction-following errors, or cross-modal inconsistencies [159, 211]. Future benchmarks should evaluate multi-step instruction execution, robustness to ambiguous language, and resistance to hallucination. Beyond empirical testing, formal verification tools are needed to provide theoretical guarantees for safety-critical behaviors.

Human-Centric Interaction & Personalization. VLA systems open the door to richer in-car interaction [219, 295]. Natural language enables drivers to specify goals, constraints, and preferences (“*drive cautiously*”, “*avoid unprotected left turns*”). Personalization modules [81] can adapt driving styles to different users, enhancing comfort and trust. The challenge lies in balancing personalization with strict safety and regulatory requirements.

7 Conclusion

Vision-Language-Action models are reshaping autonomous driving by coupling perception with high-level reasoning and natural language understanding. This work formalizes the VLA problem setting, outlines the progression from traditional VA pipelines, and organizes existing methods into coherent architectural families together with the datasets and benchmarks that support their development. VLA systems offer clear advantages in interpretability, generalization, and human interaction, but core challenges remain: aligning symbolic reasoning with continuous control, ensuring robustness in long-tail scenarios, and establishing evaluation protocols that faithfully measure instruction following and safety. Progress will depend on advances in efficient architectures, deeper multimodal fusion, world-model-driven planning, and more rigorous human-centered testing. Overall, VLA represents a promising direction for building autonomous agents that are not only competent drivers but also communicative, transparent, and responsive to human intent.

References

- [1] Josh Achiam et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Adilzhan Adilkhanov et al. Survey on vision-language-action models. *arXiv preprint arXiv:2502.06851*, 2025.
- [3] Hidehisa Arai et al. CoVLA: Comprehensive vision-language-action dataset for autonomous driving. In *IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pages 1933–1943, 2025.
- [4] Alisson Azzolini et al. Cosmos-Reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.
- [5] Jinze Bai et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- [6] Shuai Bai et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- [7] Michael Bain and Claude Sammut. A framework for behavioural cloning. In *Mach. intell.*, volume 15, pages 103–129, 1995.
- [8] Federico Baldassarre et al. Back to the features: DINO as a foundation for video world models. *arXiv preprint arXiv:2507.19468*, 2025.
- [9] Philip J. Ball et al. Genie 3: A new frontier for world models. <https://deepmind.google/discover/blog/genie-3-a-new-frontier-for-world-models/>, 2025.
- [10] Mayank Bansal, Alex Krizhevsky, and Abhijit Ogale. ChauffeurNet: Learning to drive by imitating the best and synthesizing the worst. *arXiv preprint arXiv:1812.03079*, 2018.
- [11] Bartoccioni et al. VaViM and VaVAM: Autonomous driving through video generative modeling. *arXiv preprint arXiv:2502.15672*, 2025.
- [12] Lucas Beyer et al. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

- [13] Hengwei Bian et al. DynamicCity: Large-scale 4D occupancy generation from dynamic scenes. In *Int. Conf. Learn. Represent.*, 2025.
- [14] Andreas Blattmann et al. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 22563–22575, 2023.
- [15] Andreas Blattmann et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- [16] Mariusz Bojarski et al. End-to-end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.
- [17] Holger Caesar et al. nuScenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 11621–11631, 2020.
- [18] Holger Caesar et al. nuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021.
- [19] Jiajun Cao et al. FastDriveVLA: Efficient end-to-end driving via plug-and-play reconstruction-based token pruning. *arXiv preprint arXiv:2507.23318*, 2025.
- [20] Zhong Cao et al. Autonomous driving policy continual learning with one-shot disengagement case. *IEEE Trans. Intell. Veh.*, 8(2):1380–1391, 2022.
- [21] Jun Cen et al. WorldVLA: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.
- [22] Qianwen Chao et al. A survey on visual traffic simulation: Models, evaluations, and applications in autonomous driving. In *Computer Graphics Forum*, volume 39, pages 287–308. Wiley Online Library, 2020.
- [23] Raphael Chekroun, Marin Toromanoff, Sascha Hornauer, and Fabien Moutarde. GRI: General reinforced imitation and its application to vision-based autonomous driving. *Robotics*, 12(5):127, 2023.
- [24] Dian Chen and Philipp Krähenbühl. Learning from all vehicles. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 17222–17231, 2022.
- [25] Dian Chen et al. Learning by cheating. In *Conf. Robot Learn.*, pages 66–75. PMLR, 2020.
- [26] Dian Chen et al. Learning to drive from a world on rails. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 15590–15599, 2021.
- [27] Junyu Chen et al. DC-AE 1.5: Accelerating diffusion model convergence with structured latent space. *arXiv preprint arXiv:2508.00413*, 2025.
- [28] Li Chen et al. End-to-end autonomous driving: Challenges and frontiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):10164–10183, 2024.
- [29] Runnan Chen et al. CLIP2Scene: Towards label-efficient 3D scene understanding by CLIP. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 7020–7030, 2023.
- [30] Runnan Chen et al. Towards label-free scene understanding by vision foundation models. In *Adv. Neural Inf. Process. Syst.*, volume 36, pages 75896–75910, 2023.
- [31] Shaoyu Chen et al. VADv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*, 2024.
- [32] Xi Chen et al. PaLI-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023.
- [33] Xuesong Chen et al. SOLVE: Synergy of language-vision and end-to-end networks for autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 12068–12077, 2025.
- [34] Yuntao Chen et al. DrivingGPT: Unifying driving world modeling and planning with multi-modal autoregressive transformers. *arXiv preprint arXiv:2412.18607*, 2024.
- [35] Zhe Chen et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 24185–24198, 2024.
- [36] Haohan Chi et al. Impromptu VLA: Open weights and open data for driving vision-language-action models. *arXiv preprint arXiv:2505.23757*, 2025.
- [37] Wei-Lin Chiang et al. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. <https://vicuna.lmsys.org> (accessed 14 April 2023), 2023.

- [38] Pranav Singh Chib and Pravendra Singh. Recent advancements in end-to-end autonomous driving using deep learning: A survey. *IEEE Trans. Intell. Veh.*, 9(1):103–118, 2023.
- [39] Kashyap Chitta, Aditya Prakash, and Andreas Geiger. NEAT: Neural attention fields for end-to-end autonomous driving. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 15793–15803, 2021.
- [40] Kashyap Chitta et al. TransFuser: Imitation with transformer-based sensor fusion for autonomous driving. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):12878–12895, 2022.
- [41] Xiangxiang Chu et al. MobileVLM v2: Faster and stronger baseline for vision language model. *arXiv preprint arXiv:2402.03766*, 2024.
- [42] Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *Int. Conf. Mach. Learn.*, pages 2048–2056. PMLR, 2020.
- [43] Felipe Codevilla et al. End-to-end driving via conditional imitation learning. In *IEEE Int. Conf. Robot. Autom.*, pages 4693–4700, 2018.
- [44] Felipe Codevilla et al. Exploring the limitations of behavior cloning for autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 9329–9338, 2019.
- [45] Can Cui et al. A survey on multimodal large language models for autonomous driving. In *IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pages 958–979, 2024.
- [46] Can Cui et al. ViLaD: A large vision language diffusion framework for end-to-end autonomous driving. *arXiv preprint arXiv:2508.12603*, 2025.
- [47] Daniel Dauner et al. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Adv. Neural Inf. Process. Syst.*, volume 37, pages 28706–28719, 2024.
- [48] Cainan Davidson, Deva Ramanan, and Neehar Peri. RefAV: Towards planning-centric scenario mining. *arXiv preprint arXiv:2505.20981*, 2025.
- [49] Pim De Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Adv. Neural Inf. Process. Syst.*, volume 32, pages 11698–11709, 2019.
- [50] Thierry Deruyttere et al. Talk2Car: Taking control of your self-driving car. In *Conf. Empirical Methods Natural Lang. Process.*, pages 2088–2098, 2019.
- [51] Thierry Deruyttere et al. Talk2Car: Predicting physical trajectories for natural language commands. *IEEE Access*, 10:123809–123834, 2022.
- [52] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [53] Jingtao Ding et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Comput. Surveys*, 2024.
- [54] Xinpeng Ding et al. Holistic autonomous driving understanding by bird’s-eye-view injected multi-modal large models. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 13668–13677, 2024.
- [55] Nemanja Djuric et al. Uncertainty-aware short-term motion prediction of traffic actors for autonomous driving. In *IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pages 2095–2104, 2020.
- [56] Alexey Dosovitskiy et al. CARLA: An open urban driving simulator. In *Conf. Robot Learn.*, pages 1–16. PMLR, 2017.
- [57] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, 2021.
- [58] Abhimanyu Dubey et al. The LLaMA 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [59] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 12873–12883, 2021.
- [60] Scott Ettinger et al. Large scale interactive motion forecasting for autonomous driving: The Waymo Open Motion dataset. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 9710–9719, 2021.
- [61] Shiyu Fang et al. CoReVLA: A dual-stage end-to-end autonomous driving framework for long-tail scenarios via collect-and-refine. *arXiv preprint arXiv:2509.15968*, 2025.

- [62] Yuxin Fang et al. EVA: Exploring the limits of masked visual representation learning at scale. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 19358–19369, 2023.
- [63] Yuxin Fang et al. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.
- [64] Bowen Feng et al. VERDI: VLM-embedded reasoning for autonomous driving. *arXiv preprint arXiv:2505.15925*, 2025.
- [65] Lan Feng et al. RAP: 3D rasterization augmented end-to-end planning. *arXiv preprint arXiv:2510.04333*, 2025.
- [66] Daocheng Fu et al. Drive like a human: Rethinking autonomous driving with large language models. In *IEEE/CVF Winter Conf. Appl. Comput. Vis. Worksh.*, pages 910–919, 2024.
- [67] Haoyu Fu et al. ORION: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. *arXiv preprint arXiv:2503.19755*, 2025.
- [68] Hao Gao et al. RAD: Training an end-to-end driving policy via large-scale 3DGS-based reinforcement learning. *arXiv preprint arXiv:2502.13144*, 2025.
- [69] Shenyuan Gao et al. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Adv. Neural Inf. Process. Syst.*, 37:91560–91596, 2024.
- [70] Xiangbo Gao et al. LangCoop: Collaborative driving with language. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 4226–4237, 2025.
- [71] Anant Garg and K Madhava Krishna. Imagine-2-Drive: Leveraging high-fidelity world models via multi-modal diffusion policies. *arXiv preprint arXiv:2411.10171*, 2024.
- [72] Maoning Ge et al. VLA-MP: A vision-language-action framework for multimodal perception and physics-constrained action generation in autonomous driving. *Sensors*, 25(19):6163, 2025.
- [73] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 3354–3361, 2012.
- [74] Anurag Ghosh et al. ROADWork dataset: Learning to recognize, observe, analyze and drive through work zones. *arXiv preprint arXiv:2406.07661*, 2024.
- [75] Xingtai Gui et al. TrajDiff: End-to-end autonomous driving without perception annotation. *arXiv preprint arXiv:2512.00723*, 2025.
- [76] Ziang Guo and Zufeng Zhang. VDRive: Leveraging reinforced VLA and diffusion policy for end-to-end autonomous driving. *arXiv preprint arXiv:2510.15446*, 2025.
- [77] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.
- [78] Shadi Hamdan et al. ETA: Efficiency through thinking ahead, a dual approach to self-driving with large models. *arXiv preprint arXiv:2506.07725*, 2025.
- [79] Jianhua Han et al. Percept-WAM: Perception-enhanced world-awareness-action model for robust end-to-end autonomous driving. *arXiv preprint arXiv:2511.19221*, 2025.
- [80] Wencheng Han et al. DME-Driver: Integrating human decision logic and 3D scene perception in autonomous driving. In *AAAI Conf. Artifi. Intell.*, volume 39, pages 3347–3355, 2025.
- [81] Ruiyang Hao et al. StyleDrive: Towards driving-style aware benchmarking of end-to-end autonomous driving. *arXiv preprint arXiv:2506.23982*, 2025.
- [82] Xiaoshuai Hao et al. Is your HD map constructor reliable under sensor corruptions? *Adv. Neural Inf. Process. Syst.*, 37:22441–22482, 2024.
- [83] Xiaoshuai Hao et al. MapFusion: A novel BEV feature fusion network for multi-modal map construction. *Information Fusion*, 119:103018, 2025.
- [84] Xiaoshuai Hao et al. MSC-Bench: Benchmarking and analyzing multi-sensor corruption for driving perception. *arXiv preprint arXiv:2501.01037*, 2025.
- [85] Xiaoshuai Hao et al. SafeMap: Robust HD map construction from incomplete observations. In *Int. Conf. Mach. Learn.*, pages 22091–22102. PMLR, 2025.

- [86] Yuhang Hao et al. DriveAction: A benchmark for exploring human-like driving decisions in VLA models. *arXiv preprint arXiv:2506.05667*, 2025.
- [87] Kaiming He et al. Deep residual learning for image recognition. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.
- [88] Deepti Hegde et al. Distilling multi-modal large language models for autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 27575–27585, 2025.
- [89] John Houston et al. One thousand and one hours: Self-driving motion prediction dataset. In *Conf. Robot Learn.*, pages 409–418. PMLR, 2021.
- [90] Anthony Hu et al. GAIA-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.
- [91] Shengchao Hu et al. ST-P3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *Eur. Conf. Comput. Vis.*, pages 533–549. Springer, 2022.
- [92] Xiaotao Hu et al. DrivingWorld: Constructing world model for autonomous driving via video GPT. *arXiv preprint arXiv:2412.19505*, 2024.
- [93] Yihan Hu et al. Planning-oriented autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 17853–17862, 2023.
- [94] Lei Huang et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Info. Syst.*, 43(2):1–55, 2025.
- [95] Yidong Huang et al. DriVLMe: Enhancing llm-based autonomous driving agents with embodied and social experiences. In *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pages 3153–3160, 2024.
- [96] Zanming Huang, Jimuyang Zhang, and Eshed Ohn-Bar. Neural volumetric world models for autonomous driving. In *Eur. Conf. Comput. Vis.*, pages 195–213. Springer, 2024.
- [97] Zhijian Huang et al. Making large language models better planners with reasoning-decision alignment. In *Eur. Conf. Comput. Vis.*, pages 73–90. Springer, 2024.
- [98] Jyh-Jing Hwang et al. EMMA: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024.
- [99] Gabriel Ilharco et al. OpenCLIP. *Zenodo*, 2021.
- [100] Fan Jia et al. ADriver-I: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*, 2023.
- [101] Xiaosong Jia et al. DriveAdapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 7953–7963, 2023.
- [102] Xiaosong Jia et al. Think twice before driving: Towards scalable decoders for end-to-end autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2023.
- [103] Xiaosong Jia et al. Bench2Drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. In *Adv. Neural Inf. Process. Syst.*, volume 37, pages 819–844, 2024.
- [104] Xiaosong Jia et al. DriveTransformer: Unified transformer for scalable end-to-end autonomous driving. In *Int. Conf. Learn. Represent.*, 2025.
- [105] Xiaosong Jia et al. Spatial retrieval augmented autonomous driving. *arXiv preprint arXiv:2512.06865*, 2025.
- [106] Anqing Jiang et al. DiffVLA: Vision-language guided diffusion planning for autonomous driving. *arXiv preprint arXiv:2505.19381*, 2025.
- [107] Bo Jiang, Shaoyu Chen, Qian Zhang, Wenyu Liu, and Xinggang Wang. Alphadrive: Unleashing the power of vlms in autonomous driving via reinforcement learning and reasoning. *arXiv preprint arXiv:2503.07608*, 2025.
- [108] Bo Jiang et al. VAD: Vectorized scene representation for efficient autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 8340–8350, 2023.
- [109] Bo Jiang et al. Senna: Bridging large vision-language models and end-to-end autonomous driving. *arXiv preprint arXiv:2410.22313*, 2024.

- [110] Sicong Jiang et al. A survey on vision-language-action models for autonomous driving. *arXiv preprint arXiv:2506.24044*, 2025.
- [111] Xie Jihong et al. Edge computing for real-time decision making in autonomous driving: Review of challenges, solutions, and future trends. *Int. J. Adv. Comput. Sci. & Appl.*, 15(7), 2024.
- [112] Bu Jin, Xiaotao Hu, Songen Gu, Yupeng Zheng, Xiaoyang Guo, et al. OccVAR: Scalable 4D occupancy prediction via next-scale prediction. <https://openreview.net/forum?id=X2HnTFsFm8>, 2025.
- [113] Taotao Jing et al. Inaction: Interpretable action decision making for autonomous driving. In *European Conference on Computer Vision*, pages 370–387. Springer, 2022.
- [114] Daniel Kahneman. Thinking, fast and slow. *Farrar, Straus and Giroux*, 2011.
- [115] Napat Karnchanachari et al. Towards learning-based planning: The nuPlan benchmark for real-world autonomous driving. In *IEEE Int. Conf. Robot. Autom.*, pages 629–636, 2024.
- [116] Bernhard Kerbl et al. 3D gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.
- [117] Siddhesh Khandelwal et al. What-if motion prediction for autonomous driving. *arXiv preprint arXiv:2008.10587*, 2020.
- [118] Jinkyu Kim et al. Textual explanations for self-driving vehicles. In *Eur. Conf. Comput. Vis.*, pages 563–578. Springer, 2018.
- [119] B Ravi Kiran et al. Deep reinforcement learning for autonomous driving: A survey. *IEEE Int. Conf. Intell. Transport. Syst.*, 23(6):4909–4926, 2021.
- [120] W Bradley Knox et al. Reward (mis)design for autonomous driving. *Artifi. Intell.*, 316:103829, 2023.
- [121] Lingdong Kong, Wesley Yang, Jianbiao Mei, Youquan Liu, Ao Liang, Dekai Zhu, Dongyue Lu, Wei Yin, Xiaotao Hu, Mingkai Jia, Junyuan Deng, Kaiwen Zhang, Yang Wu, Tianyi Yan, Shenyuan Gao, Song Wang, Linfeng Li, Liang Pan, Yong Liu, Jianke Zhu, Wei Tsang Ooi, Steven C. H. Hoi, and Ziwei Liu. 3D and 4D world modeling: A survey. *arXiv preprint arXiv:2509.07996*, 2025.
- [122] Lingdong Kong et al. Rethinking range view representation for LiDAR segmentation. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 228–240, 2023.
- [123] Lingdong Kong et al. Robo3D: Towards robust and reliable 3D perception against corruptions. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 19994–20006, 2023.
- [124] Lingdong Kong et al. The RoboDrive challenge: Drive anytime anywhere in any condition. *arXiv preprint arXiv:2405.08816*, 2024.
- [125] Lingdong Kong et al. LargeAD: Large-scale cross-sensor data pretraining for autonomous driving. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025.
- [126] Lingdong Kong et al. Multi-modal data-efficient 3D scene understanding for autonomous driving. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(5):3748–3765, 2025.
- [127] Alex H Lang et al. PointPillars: Fast encoders for object detection from point clouds. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 12697–12705, 2019.
- [128] Doyup Lee et al. Autoregressive image generation using residual quantization. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 11523–11532, 2022.
- [129] Youngwan Lee et al. An energy and GPU-computation efficient backbone network for real-time object detection. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog. Worksh.*, pages 1–9, 2019.
- [130] Bo Li et al. LlaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.
- [131] Boyi Li et al. Driving everywhere with large language model policy adaptation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 14948–14957, 2024.
- [132] Hongyang Li et al. Open-sourced data ecosystem in autonomous driving: The present and future. *arXiv preprint arXiv:2312.03408*, 2023.
- [133] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Int. Conf. Mach. Learn.*, pages 19730–19742. PMLR, 2023.

- [134] Peidong Li and Dixiao Cui. Navigation-guided sparse scene representation for end-to-end autonomous driving. *arXiv preprint arXiv:2409.18341*, 2024.
- [135] Peizheng Li et al. SpaceDrive: Infusing spatial awareness into VLM-based autonomous driving. *arXiv preprint arXiv:2512.10719*, 2025.
- [136] Pengxiang Li et al. Discrete diffusion for reflective vision-language-action models in autonomous driving. *arXiv preprint arXiv:2509.20109*, 2025.
- [137] Qifeng Li et al. Think2Drive: Efficient reinforcement learning by thinking with latent world model for autonomous driving (in CARLA-V2). In *Eur. Conf. Comput. Vis.*, pages 142–158. Springer, 2024.
- [138] Rong Li et al. 3EED: Ground everything everywhere in 3D. In *Adv. Neural Inf. Process. Syst.*, volume 38, 2025.
- [139] Wentong Li et al. TokenPacker: Efficient visual projector for multimodal LLM. *Int. J. Comput. Vis.*, 133: 6794–6812, 2025.
- [140] Xin Li et al. Towards knowledge-driven autonomous driving. *arXiv preprint arXiv:2312.04316*, 2023.
- [141] Ye Li et al. Is your LiDAR placement optimized for 3D scene understanding? In *Adv. Neural Inf. Process. Syst.*, volume 37, pages 34980–35017, 2024.
- [142] Yiheng Li et al. WOMD-Reasoning: A large-scale dataset for interaction reasoning in driving. *arXiv preprint arXiv:2407.04281*, 2024.
- [143] Yingyan Li et al. Enhancing end-to-end autonomous driving with latent world model. *arXiv preprint arXiv:2406.08481*, 2024.
- [144] Yingyan Li et al. DriveVLA-W0: World models amplify data scaling law in autonomous driving. *arXiv preprint arXiv:2510.12796*, 2025.
- [145] Yingyan Li et al. End-to-end driving with online trajectory evaluation via bev world model. *arXiv preprint arXiv:2504.01941*, 2025.
- [146] Yongkang Li et al. ReCogDrive: A reinforced cognitive framework for end-to-end autonomous driving. *arXiv preprint arXiv:2506.08052*, 2025.
- [147] Yue Li et al. Drive-R1: Bridging reasoning and planning in VLMs for autonomous driving with reinforcement learning. *arXiv preprint arXiv:2506.18234*, 2025.
- [148] Zhenxin Li et al. Hydra-MDP: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978*, 2024.
- [149] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahua Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 14864–14873, 2024.
- [150] Zhiqi Li et al. BEVFormer: learning bird’s-eye-view representation from LiDAR-camera via spatiotemporal transformers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(3):2020–2036, 2024.
- [151] Zhuoren Li et al. A survey of reinforcement learning-based motion planning for autonomous driving: Lessons learned from a driving task perspective. *arXiv preprint arXiv:2503.23650*, 2025.
- [152] Li Auto Inc. MindVLA. <https://ir.lixiang.com/news-releases/news-release-details/li-auto-inc-march-2025-delivery-update>, 2025.
- [153] Ao Liang et al. Perspective-invariant 3D object detection. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 27725–27738, 2025.
- [154] Ao Liang et al. LiDARCrafter: Dynamic 4D world modeling from LiDAR sequences. In *AAAI Conf. Artifi. Intell.*, volume 40, 2026.
- [155] Tingting Liang et al. BEVFusion: A simple and robust LiDAR-camera fusion framework. *Adv. Neural Inf. Process. Syst.*, 35:10421–10434, 2022.
- [156] Bencheng Liao et al. DiffusionDrive: Truncated diffusion model for end-to-end autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 12037–12047, 2025.
- [157] Bin Lin et al. Video-LLaVA: Learning unified visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

- [158] Dongyang Liu et al. Lumina-MGPT: Illuminate flexible photorealistic text-to-image generation with multimodal generative pretraining. *arXiv preprint arXiv:2408.02657*, 2024.
- [159] Hanchao Liu et al. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.
- [160] Haotian Liu et al. Visual instruction tuning. In *Adv. Neural Inf. Process. Syst.*, volume 36, pages 34892–34916, 2023.
- [161] Haotian Liu et al. Improved baselines with visual instruction tuning. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 26296–26306, 2024.
- [162] Lin Liu et al. GuideFlow: Constraint-guided flow matching for planning in end-to-end autonomous driving. *arXiv preprint arXiv:2511.18729*, 2025.
- [163] Mingyu Liu et al. A survey on autonomous driving datasets: Statistics, annotation quality, and a future outlook. *IEEE Trans. Intell. Veh.*, 9(11):7138–7164, 2024.
- [164] Pei Liu et al. OmniReason: A temporal-guided vision-language-action framework for autonomous driving. *arXiv preprint arXiv:2509.00789*, 2025.
- [165] Pei Liu et al. OmniScene: Attention-augmented multimodal 4D scene understanding for autonomous driving. *arXiv preprint arXiv:2509.19973*, 2025.
- [166] Pei Liu et al. VLM-E2E: Enhancing end-to-end autonomous driving with multimodal driver attention fusion. *arXiv preprint arXiv:2502.18042*, 2025.
- [167] Ruixun Liu et al. OccVLA: Vision-language-action model with implicit 3D occupancy supervision. *arXiv preprint arXiv:2509.05578*, 2025.
- [168] 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.
- [169] Xueyi Liu et al. ReasonPlan: Unified scene prediction and decision reasoning for closed-loop autonomous driving. *arXiv preprint arXiv:2505.20024*, 2025.
- [170] Youquan Liu et al. Segment any point cloud sequences by distilling vision foundation models. In *Adv. Neural Inf. Process. Syst.*, volume 36, pages 37193–37229, 2023.
- [171] Youquan Liu et al. La La LiDAR: Large-scale layout generation from LiDAR data. In *AAAI Conf. Artifi. Intell.*, volume 40, 2026.
- [172] Ze Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 10012–10022, 2021.
- [173] Zhuang Liu et al. A ConvNet for the 2020s. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 11976–11986, 2022.
- [174] Hao Lu et al. UniUGP: Unifying understanding, generation, and planning for end-to-end autonomous driving. *arXiv preprint arXiv:2512.09864*, 2025.
- [175] Jiachen Lu et al. WoVoGen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *Eur. Conf. Comput. Vis.*, pages 329–345. Springer, 2024.
- [176] Yuhang Lu et al. ReAL-AD: Towards human-like reasoning in end-to-end autonomous driving. *arXiv preprint arXiv:2507.12499*, 2025.
- [177] Yuechen Luo et al. AdaThinkDrive: Adaptive thinking via reinforcement learning for autonomous driving. *arXiv preprint arXiv:2509.13769*, 2025.
- [178] Junyi Ma et al. Cam4DOcc: Benchmark for camera-only 4D occupancy forecasting in autonomous driving applications. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 21486–21495, 2024.
- [179] Yingzi Ma et al. dVLM-AD: Enhance diffusion vision-language-model for driving via controllable reasoning. *arXiv preprint arXiv:2512.04459*, 2025.
- [180] Yueen Ma et al. A survey on vision-language-action models for embodied AI. *arXiv preprint arXiv:2405.14093*, 2024.

- [181] Yukai Ma et al. LeapVAD: A leap in autonomous driving via cognitive perception and dual-process thinking. *arXiv preprint arXiv:2501.08168*, 2025.
- [182] Yunsheng Ma et al. ALN-P3: Unified language alignment for perception, prediction, and planning in autonomous driving. *arXiv preprint arXiv:2505.15158*, 2025.
- [183] Ziqiao Ma et al. Dorothie: Spoken dialogue for handling unexpected situations in interactive autonomous driving agents. *arXiv preprint arXiv:2210.12511*, 2022.
- [184] Jiageng Mao et al. GPT-Driver: Learning to drive with GPT. *arXiv preprint arXiv:2310.01415*, 2023.
- [185] Jiageng Mao et al. A language agent for autonomous driving. In *Conf. Lang. Model.*, 2025.
- [186] Ana-Maria Marcu et al. LingoQA: Visual question answering for autonomous driving. In *Eur. Conf. Comput. Vis.*, pages 252–269. Springer, 2024.
- [187] Jianbiao Mei et al. Continuously learning, adapting, and improving: A dual-process approach to autonomous driving. In *Adv. Neural Inf. Process. Syst.*, volume 37, pages 123261–123290, 2024.
- [188] Chen Min et al. DriveWorld: 4D pre-trained scene understanding via world models for autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 15522–15533, 2024.
- [189] Mona Mirzaie and Bodo Rosenhahn. Interpretable decision-making for end-to-end autonomous driving. *arXiv preprint arXiv:2508.18898*, 2025.
- [190] Urs Muller et al. Off-road obstacle avoidance through end-to-end learning. In *Adv. Neural Inf. Process. Syst.*, volume 18, pages 739–746, 2005.
- [191] Ferdinand Mütsch et al. From model-based to data-driven simulation: Challenges and trends in autonomous driving. *arXiv preprint arXiv:2305.13960*, 2023.
- [192] Ming Nie et al. Reason2Drive: Towards interpretable and chain-based reasoning for autonomous driving. In *Eur. Conf. Comput. Vis.*, pages 292–308. Springer, 2024.
- [193] NVIDIA. Physical AI autonomous vehicles dataset. <https://huggingface.co/datasets/nvidia/PhysicalAI-Autonomous-Vehicles>, October 2025. URL <https://huggingface.co/datasets/nvidia/PhysicalAI-Autonomous-Vehicles>.
- [194] Eshed Ohn-Bar et al. Learning situational driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 11296–11305, 2020.
- [195] OpenAI. Hello GPT4-o. <https://openai.com/index/hello-gpt-4o>, 2024.
- [196] Maxime Oquab et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.
- [197] Chenbin Pan et al. VLP: Vision language planning for autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 14760–14769, 2024.
- [198] Yunpeng Pan et al. Agile autonomous driving using end-to-end deep imitation learning. In *Robot. Sci. Syst.*, 2018.
- [199] Jinhyung Park et al. Time Will Tell: New outlooks and a baseline for temporal multi-view 3D object detection. In *Int. Conf. Learn. Represent.*, 2023.
- [200] Jongjin Park et al. Object-aware regularization for addressing causal confusion in imitation learning. In *Adv. Neural Inf. Process. Syst.*, volume 34, pages 3029–3042, 2021.
- [201] SungYeon Park et al. VLAAD: Vision and language assistant for autonomous driving. In *IEEE/CVF Winter Conf. Appl. Comput. Vis. Worksh.*, pages 980–987, 2024.
- [202] Dustin Podell et al. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *Int. Conf. Learn. Represent.*, 2023.
- [203] Dean A Pomerleau. ALVINN: An autonomous land vehicle in a neural network. In *Adv. Neural Inf. Process. Syst.*, volume 1, pages 305–313, 1988.
- [204] Alexander Popov et al. Mitigating covariate shift in imitation learning for autonomous vehicles using latent space generative world models. *arXiv preprint arXiv:2409.16663*, 2024.

- [205] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 7077–7087, 2021.
- [206] Aditya Prakash et al. Exploring data aggregation in policy learning for vision-based urban autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 11763–11773, 2020.
- [207] Kangan Qian et al. FasionAD: Fast and slow fusion thinking systems for human-like autonomous driving with adaptive feedback. *arXiv preprint arXiv:2411.18013*, 2024.
- [208] Kangan Qian et al. FasionAD++: Integrating high-level instruction and information bottleneck in fast-slow fusion systems for enhanced safety in autonomous driving with adaptive feedback. *arXiv preprint arXiv:2503.08162*, 2025.
- [209] Zhijie Qiao et al. LightEMMA: Lightweight end-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2505.00284*, 2025.
- [210] Alec Radford et al. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn.*, pages 8748–8763. PMLR, 2021.
- [211] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.
- [212] Xuanchi Ren et al. Cosmos-Drive-Dreams: Scalable synthetic driving data generation with world foundation models. *arXiv preprint arXiv:2506.09042*, 2025.
- [213] Katrin Renz et al. SimLingo: Vision-only closed-loop autonomous driving with language-action alignment. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 11993–12003, 2025.
- [214] Robin Rombach et al. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 10684–10695, 2022.
- [215] Stéphane Ross et al. A reduction of imitation learning and structured prediction to no-regret online learning. In *Int. Conf. Artif. Intell. Stat.*, pages 627–635, 2011.
- [216] Luke Rowe et al. Poutine: Vision-language-trajectory pre-training and reinforcement learning post-training enable robust end-to-end autonomous driving. *arXiv preprint arXiv:2506.11234*, 2025.
- [217] Ranjan Sapkota et al. Vision-language-action models: Concepts, progress, applications and challenges. *arXiv preprint arXiv:2505.04769*, 2025.
- [218] Oliver Scheel et al. Urban driver: Learning to drive from real-world demonstrations using policy gradients. In *Conf. Robot Learn.*, pages 718–728. PMLR, 2022.
- [219] Mariah L Schrum, Emily Sumner, Matthew C Gombolay, and Andrew Best. Maveric: A data-driven approach to personalized autonomous driving. *IEEE Trans. Robot.*, 40:1952–1965, 2024.
- [220] Hao Shan et al. Stability under scrutiny: Benchmarking representation paradigms for online HD mapping. *arXiv preprint arXiv:2510.10660*, 2025.
- [221] Hao Shao et al. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In *Conference on Robot Learning*, pages 726–737. PMLR, 2023.
- [222] Hao Shao et al. LMDrive: Closed-loop end-to-end driving with large language models. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 15120–15130, 2024.
- [223] Rui Shao et al. Large VLM-based vision-language-action models for robotic manipulation: A survey. *arXiv preprint arXiv:2508.13073*, 2025.
- [224] Zhihong Shao et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [225] Chonghao Sima et al. DriveLM: Driving with graph visual question answering. In *Eur. Conf. Comput. Vis.*, pages 256–274. Springer, 2024.
- [226] Nan Song et al. LMAD: Integrated end-to-end vision-language model for explainable autonomous driving. *arXiv preprint arXiv:2508.12404*, 2025.
- [227] Ruiqi Song et al. InsightDrive: Insight scene representation for end-to-end autonomous driving. *arXiv preprint arXiv:2503.13047*, 2025.

- [228] Zhihang Song et al. Synthetic datasets for autonomous driving: A survey. *IEEE Trans. Intell. Veh.*, 9(1):1847–1864, 2023.
- [229] Ziying Song et al. Don’t shake the wheel: Momentum-aware planning in end-to-end autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 22432–22441, 2025.
- [230] Andreas Steiner et al. PaliGemma 2: A family of versatile VLMs for transfer. *arXiv preprint arXiv:2412.03555*, 2024.
- [231] Haisheng Su et al. DiFSD: Ego-centric fully sparse paradigm with uncertainty denoising and iterative refinement for efficient end-to-end self-driving. *arXiv preprint arXiv:2409.09777*, 2024.
- [232] Jintao Sun et al. Echo planning for autonomous driving: From current observations to future trajectories and back. *arXiv preprint arXiv:2505.18945*, 2025.
- [233] Pei Sun et al. Scalability in perception for autonomous driving: Waymo open dataset. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 2446–2454, 2020.
- [234] Peize Sun et al. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.
- [235] Wenchao Sun et al. SparseDrive: End-to-end autonomous driving via sparse scene representation. In *IEEE Int. Conf. Robot. Autom.*, pages 8795–8801, 2025.
- [236] Bin Suna et al. MindDrive: An all-in-one framework bridging world models and vision-language model for end-to-end autonomous driving. *arXiv preprint arXiv:2512.04441*, 2025.
- [237] Richard S Sutton, Andrew G Barto, et al. *Reinforcement Learning: An Introduction*, volume 1. MIT Press, Cambridge, 1998.
- [238] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Int. Conf. Mach. Learn.*, pages 6105–6114. PMLR, 2019.
- [239] Shuhan Tan et al. Latent chain-of-thought world modeling for end-to-end driving. *arXiv preprint arXiv:2512.10226*, 2025.
- [240] Yihong Tang et al. E3AD: An emotion-aware vision-language-action model for human-centric end-to-end autonomous driving. *arXiv preprint arXiv:2512.04733*, 2025.
- [241] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.
- [242] Gemini Team et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [243] Gemini Robotics Team et al. Gemini robotics: Bringing AI into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.
- [244] Qwen Team. Introducing Qwen1.5, February 2024. URL <https://qwenlm.github.io/blog/qwen1.5/>.
- [245] Wayve Team. LINGO-2: Driving with natural language. <https://wayve.ai/thinking/lingo-2-driving-with-language/>, 2024.
- [246] Waywe Research Team et al. LINGO-2: Driving with natural language, 2024.
- [247] Ran Tian et al. Tokenize the world into object-level knowledge to address long-tail events in autonomous driving. In *Conf. Robot Learn.*, pages 3656–3673. PMLR, 2025.
- [248] Xiaoyu Tian et al. Occ3D: A large-scale 3D occupancy prediction benchmark for autonomous driving. In *Adv. Neural Inf. Process. Syst.*, volume 36, pages 64318–64330, 2023.
- [249] Xiaoyu Tian et al. DriveVLM: The convergence of autonomous driving and large vision-language models. In *Conf. Robot Learn.*, pages 4698–4726. PMLR, 2025.
- [250] Wenwen Tong et al. Scene as occupancy. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 8406–8415, 2023.
- [251] Marin Toromanoff et al. End-to-end model-free reinforcement learning for urban driving using implicit affordances. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 7153–7162, 2020.

- [252] Hugo Touvron et al. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [253] Michael Tschannen et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.
- [254] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Adv. Neural Inf. Process. Syst.*, volume 30, pages 6309–6318, 2017.
- [255] Eli Verwimp et al. CLAD: A realistic continual learning benchmark for autonomous driving. *Neural Net.*, 161: 659–669, 2023.
- [256] Daming Wang et al. HMVLM: Multistage reasoning-enhanced vision-language model for long-tailed driving scenarios. *arXiv preprint arXiv:2506.05883*, 2025.
- [257] Dingrui Wang et al. DualAD: Dual-layer planning for reasoning in autonomous driving. In *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2025.
- [258] Peng Wang et al. Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- [259] Shihao Wang et al. OmniDrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 22442–22452, 2025.
- [260] Song Wang et al. Not all voxels are equal: Hardness-aware semantic scene completion with self-distillation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 14792–14801, 2024.
- [261] Song Wang et al. PixelThink: Towards efficient chain-of-pixel reasoning. *arXiv preprint arXiv:2505.23727*, 2025.
- [262] Tianqi Wang et al. DriveCoT: Integrating chain-of-thought reasoning with end-to-end driving. *arXiv preprint arXiv:2403.16996*, 2024.
- [263] Weiyun Wang et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024.
- [264] Wenhui Wang et al. DriveMLM: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023.
- [265] Xiaofeng Wang et al. DriveDreamer: Towards real-world-drive world models for autonomous driving. In *Eur. Conf. Comput. Vis.*, pages 55–72. Springer, 2024.
- [266] Yan Wang et al. Alpamayo-R1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. *arXiv preprint arXiv:2511.00088*, 2025.
- [267] Yuqi Wang et al. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 14749–14759, 2024.
- [268] Julong Wei, Shanshuai Yuan, Pengfei Li, et al. OccLLaMA: An occupancy-language-action generative world model for autonomous driving. *arXiv preprint arXiv:2409.03272*, 2024.
- [269] Xinshuo Weng et al. PARA-Drive: Parallelized architecture for real-time autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 15449–15458, 2024.
- [270] Benjamin Wilson et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.10049*, 2023.
- [271] Katharina Winter et al. BEVDriver: Leveraging BEV maps in LLMs for robust closed-loop driving. *arXiv preprint arXiv:2503.03074*, 2025.
- [272] Hanfeng Wu et al. NaviHydra: Controllable navigation-guided end-to-end autonomous driving with hydra-distillation. *arXiv preprint arXiv:2512.10660*, 2025.
- [273] Penghao Wu et al. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. In *Adv. Neural Inf. Process. Syst.*, volume 35, pages 6119–6132, 2022.
- [274] Shaoyuan Xie et al. Benchmarking and improving bird’s eye view perception robustness in autonomous driving. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(5):3878–3894, 2025.
- [275] Shaoyuan Xie et al. Are VLMs ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 6585–6597, 2025.

- [276] Yichen Xie et al. S4-Driver: Scalable self-supervised driving multimodal large language model with spatio-temporal visual representation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 1622–1632, 2025.
- [277] Ren Xin et al. NetRoller: interfacing general and specialized models for end-to-end autonomous driving. *arXiv preprint arXiv:2506.14589*, 2025.
- [278] Shuo Xing et al. OpenEMMA: Open-source multimodal model for end-to-end autonomous driving. In *IEEE/CVF Winter Conf. Appl. Comput. Vis.*, pages 1001–1009, 2025.
- [279] Zebin Xing et al. GoalFlow: Goal-driven flow matching for multimodal trajectories generation in end-to-end autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 1602–1611, 2025.
- [280] Zebin Xing et al. Mimir: Hierarchical goal-driven diffusion with uncertainty propagation for end-to-end autonomous driving. *arXiv preprint arXiv:2512.07130*, 2025.
- [281] Chengkai Xu et al. Towards human-centric autonomous driving: A fast-slow architecture integrating large language model guidance with reinforcement learning. *arXiv preprint arXiv:2505.06875*, 2025.
- [282] Haoran Xu et al. Temporal triplane transformers as occupancy world models. *arXiv preprint arXiv:2503.07338*, 2025.
- [283] Runsheng Xu et al. WOD-E2E: Waymo open dataset for end-to-end driving in challenging long-tail scenarios. *arXiv preprint arXiv:2510.26125*, 2025.
- [284] Tianshuo Xu, Hao Lu, Xu Yan, et al. Occ-LLM: Enhancing autonomous driving with occupancy-based large language models. In *IEEE Int. Conf. Robot. Autom.*, 2025.
- [285] Xiang Xu et al. Beyond one shot, beyond one perspective: Cross-view and long-horizon distillation for better LiDAR representations. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 25506–25518, 2025.
- [286] Yi Xu et al. VLM-AD: End-to-end autonomous driving through vision-language model supervision. *arXiv preprint arXiv:2412.14446*, 2024.
- [287] Zhenhua Xu et al. DriveGPT4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robot. Autom. Lett.*, 9(10):8186–8193, 2024.
- [288] Zhenhua Xu et al. DriveGPT4-V2: Harnessing large language model capabilities for enhanced closed-loop autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 17261–17270, 2025.
- [289] Tianyi Yan et al. AD-R1: Closed-loop reinforcement learning for end-to-end autonomous driving with impartial world models. *arXiv preprint arXiv:2511.20325*, 2025.
- [290] Tianyi Yan et al. DrivingSphere: Building a high-fidelity 4D world for closed-loop simulation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 27531–27541, 2025.
- [291] Xu Yan et al. Forging vision foundation models for autonomous driving: Challenges, methodologies, and opportunities. *arXiv preprint arXiv:2401.08045*, 2024.
- [292] Ziyang Yan et al. RenderWorld: World model with self-supervised 3D label. In *IEEE Int. Conf. Robot. Autom.*, pages 6063–6070, 2025.
- [293] An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [294] An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- [295] Haohan Yang et al. Human-guided continual learning for personalized decision-making of autonomous driving. *IEEE Int. Conf. Intell. Transport. Syst.*, 26(4):5435–5447, 2025.
- [296] Jiazhi Yang et al. Generalized predictive model for autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 14662–14672, 2024.
- [297] Yu Yang et al. Driving in the occupancy world: Vision-centric 4D occupancy forecasting and planning via world models for autonomous driving. In *AAAI Conf. Artifi. Intell.*, volume 39, pages 9327–9335, 2025.
- [298] Zetong Yang et al. Visual point cloud forecasting enables scalable autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 14673–14684, 2024.
- [299] Zhenjie Yang et al. LLM4Drive: A survey of large language models for autonomous driving. In *Adv. Neural Inf. Process. Syst. Worksh.*, 2024.

- [300] Zhenjie Yang et al. DriveMoE: Mixture-of-experts for vision-language-action model in end-to-end autonomous driving. *arXiv preprint arXiv:2505.16278*, 2025.
- [301] Zhenjie Yang et al. Raw2Drive: Reinforcement learning with aligned world models for end-to-end autonomous driving (in CARLA-V2). *arXiv preprint arXiv:2505.16394*, 2025.
- [302] Huaiyuan Yao et al. LiloDriver: A lifelong learning framework for closed-loop motion planning in long-tail autonomous driving scenarios. *arXiv preprint arXiv:2505.17209*, 2025.
- [303] Zebin You et al. LlaDA-V: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*, 2025.
- [304] Fisher Yu et al. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 2636–2645, 2020.
- [305] Lijun Yu et al. Language model beats diffusion-tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.
- [306] Tianyu Yu et al. RLAIF-V: Open-source AI feedback leads to super GPT-4V trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024.
- [307] Ze Yu et al. Combining camera-LiDAR fusion and motion planning using bird’s-eye view representation for end-to-end autonomous driving. *Drones*, 9(4):281, 2025.
- [308] Chengran Yuan et al. DRAMA: An efficient end-to-end motion planner for autonomous driving with mamba. *arXiv preprint arXiv:2408.03601*, 2024.
- [309] Jianhao Yuan et al. RAG-Driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828*, 2024.
- [310] Zhenlong Yuan et al. AutoDrive-R2: Incentivizing reasoning and self-reflection capacity for VLA model in autonomous driving. *arXiv preprint arXiv:2509.01944*, 2025.
- [311] Mahmut Yurt et al. LTDA-Drive: LLMs-guided generative models based long-tail data augmentation for autonomous driving. *arXiv preprint arXiv:2505.18198*, 2025.
- [312] Shuang Zeng et al. FutureSightDrive: Thinking visually with spatio-temporal CoT for autonomous driving. *arXiv preprint arXiv:2505.17685*, 2025.
- [313] Mingliang Zhai et al. World knowledge-enhanced reasoning using instruction-guided interactor in autonomous driving. In *AAAI Conf. Artifi. Intell.*, volume 39, pages 9842–9850, 2025.
- [314] Xiaohua Zhai et al. Sigmoid loss for language image pre-training. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 11975–11986, 2023.
- [315] Bozhou Zhang et al. Future-aware end-to-end driving: Bidirectional modeling of trajectory planning and scene evolution. *arXiv preprint arXiv:2510.11092*, 2025.
- [316] Dapeng Zhang et al. Reasoning-VLA: A fast and general vision-language-action reasoning model for autonomous driving. *arXiv preprint arXiv:2511.19912*, 2025.
- [317] Diankun Zhang et al. SparseAD: Sparse query-centric paradigm for efficient end-to-end autonomous driving. *arXiv preprint arXiv:2404.06892*, 2024.
- [318] Haiming Zhang et al. An efficient occupancy world model via decoupled dynamic flow and image-assisted training. *arXiv preprint arXiv:2412.13772*, 2024.
- [319] Haiming Zhang et al. VisionPAD: A vision-centric pre-training paradigm for autonomous driving. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 17165–17175, 2025.
- [320] Jiawei Zhang et al. SafeAuto: Knowledge-enhanced safe autonomous driving with multimodal foundation models. *arXiv preprint arXiv:2503.00211*, 2025.
- [321] Kaiwen Zhang et al. Epona: Autoregressive diffusion world model for autonomous driving. *arXiv preprint arXiv:2506.24113*, 2025.
- [322] Peiyuan Zhang et al. TinyLLaMA: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024.

- [323] Songyan Zhang et al. WiseAD: Knowledge augmented end-to-end autonomous driving with vision-language model. *arXiv preprint arXiv:2412.09951*, 2024.
- [324] Songyan Zhang et al. OpenREAD: Reinforced open-ended reasoning for end-to-end autonomous driving with LLM-as-Critic. *arXiv preprint arXiv:2512.01830*, 2025.
- [325] Yi Zhang et al. A unified perception-language-action framework for adaptive autonomous driving. *arXiv preprint arXiv:2507.23540*, 2025.
- [326] Zhejun Zhang et al. End-to-end urban driving by imitating a reinforcement learning coach. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 15222–15232, 2021.
- [327] Rui Zhao et al. Sce2DriveX: A generalized MLLM framework for scene-to-drive learning. *IEEE Robot. Autom. Lett.*, 2025.
- [328] Zongchuang Zhao et al. Extending large vision-language model for diverse interactive tasks in autonomous driving. *arXiv preprint arXiv:2505.08725*, 2025.
- [329] Chuanxia Zheng and Andrea Vedaldi. Online clustered codebook. In *IEEE/CVF Int. Conf. Comput. Vis.*, pages 22798–22807, 2023.
- [330] Weicheng Zheng et al. DriveAgent-R1: Advancing VLM-based autonomous driving with hybrid thinking and active perception. *arXiv preprint arXiv:2507.20879*, 2025.
- [331] Wenzhao Zheng et al. Doe-1: Closed-loop autonomous driving with large world model. *arXiv preprint arXiv:2412.09627*, 2024.
- [332] Wenzhao Zheng et al. GaussianAD: Gaussian-centric end-to-end autonomous driving. *arXiv preprint arXiv:2412.10371*, 2024.
- [333] Wenzhao Zheng et al. GenAD: Generative end-to-end autonomous driving. In *Eur. Conf. Comput. Vis.*, pages 87–104. Springer, 2024.
- [334] Wenzhao Zheng et al. OccWorld: Learning a 3D occupancy world model for autonomous driving. In *Eur. Conf. Comput. Vis.*, pages 55–72. Springer, 2024.
- [335] Yupeng Zheng et al. World4Drive: End-to-end autonomous driving via intention-aware physical latent world model. *arXiv preprint arXiv:2507.00603*, 2025.
- [336] Yifan Zhong et al. A survey on vision-language-action models: An action tokenization perspective. *arXiv preprint arXiv:2507.01925*, 2025.
- [337] Gaoyue Zhou et al. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Int. Conf. Mach. Learn.*. PMLR, 2025.
- [338] Xingcheng Zhou et al. Vision language models in autonomous driving: A survey and outlook. *IEEE Trans. Intell. Veh.*, pages 1–20, 2024.
- [339] Xingcheng Zhou et al. OpenDriveVLA: Towards end-to-end autonomous driving with large vision language action model. *arXiv preprint arXiv:2503.23463*, 2025.
- [340] Zewei Zhou et al. AutoVLA: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. *arXiv preprint arXiv:2506.13757*, 2025.
- [341] Bin Zhu et al. LanguageBind: Extending video-language pretraining to N-modality by language-based semantic alignment. In *Int. Conf. Learn. Represent.*, 2024.
- [342] Dekai Zhu et al. SPIRAL: Semantic-aware progressive LiDAR scene generation and understanding. In *Adv. Neural Inf. Process. Syst.*, volume 38, 2025.
- [343] Jinguo Zhu et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.
- [344] Huiping Zhuang et al. Online analytic exemplar-free continual learning with large models for imbalanced autonomous driving task. *IEEE Trans. Veh. Tech.*, 74(2):1949–1958, 2024.
- [345] Jialv Zou et al. DiffusionDriveV2: Reinforcement learning-constrained truncated diffusion modeling in end-to-end autonomous driving. *arXiv preprint arXiv:2512.07745*, 2025.