# OpenDriveVLA: Towards End-to-end Autonomous Driving with Large Vision Language Action Model

**Xingcheng Zhou[1]*, Xuyuan Han[1], Feng Yang[1], Yunpu Ma[2], Volker Tresp[2], Alois Knoll[1]**

[1]Technical University of Munich, Germany
[2]Ludwig Maximilian University of Munich, Germany
{xingcheng.zhou, xuyuan.han}@tum.de

## Abstract

We present OpenDriveVLA, a Vision-Language Action (VLA) model designed for end-to-end autonomous driving, built upon open-source large language models. OpenDriveVLA generates spatially-grounded driving actions by leveraging multimodal inputs, including both 2D and 3D instance-aware visual representations, ego vehicle states, and language commands. To bridge the modality gap between driving visual representations and language embeddings, we introduce a hierarchical vision-language alignment process, projecting both 2D and 3D structured visual tokens into a unified semantic space. Furthermore, we incorporate structured agent–environment–ego interaction modeling into the autoregressive decoding process, enabling the model to capture fine-grained spatial dependencies and behavior-aware dynamics critical for reliable trajectory planning. Extensive experiments on the nuScenes dataset demonstrate that OpenDriveVLA achieves state-of-the-art results across open-loop trajectory planning and driving-related question-answering tasks. Qualitative analyses further illustrate its superior capability to follow high-level driving commands and generate trajectories under challenging scenarios, highlighting its potential for next-generation end-to-end autonomous driving.

**Project Page** — https://drivevla.github.io

## Introduction

End-to-end learning frameworks have emerged as a promising paradigm in autonomous driving, enabling perception, prediction, and planning to be jointly optimized within a unified neural network (Zhou et al. 2024). They learn policies directly from sensor inputs and generalize well across varied scenarios. Despite notable progress, existing approaches still face critical challenges, including limited long-tail generalization, poor complex semantics understanding, and rigid task reasoning (Chen et al. 2024). Meanwhile, large language models (LLMs) and vision-language models (VLMs) exhibit strong in-context reasoning, commonsense understanding, and zero-shot generalization abilities. These capabilities are promising for driving, where robust scene understanding is crucial (Liu et al. 2024d; Zhou and Knoll 2024). However, directly leveraging existing VLMs for autonomous driving poses fundamental challenges. Firstly, current VLMs are predominantly optimized for static, 2D image-language tasks, leading to poor spatial reasoning performance in dynamic 3D driving environments (Zhai et al. 2023b). Besides, instance-agnostic VLMs (Liu et al. 2024c) are prone to hallucinations, often yielding incorrect yet overconfident outputs, posing safety risks in autonomous driving. Motivated by these limitations, our work answers a central question: **How can we harness the emergent capabilities of large VLMs to produce safe spatially-grounded driving actions in dynamic 3D environments, while balancing inference speed and planning effectiveness?**

To enhance spatial-awareness and safety in LLM-based vision-language action model, we introduce two key designs. First, we structure the driving environment using instance-aware, hierarchical 2D and 3D visual representations to reduce the risk of instance hallucinations. Second, we incorporate agent–environment–ego interaction modeling, which is originally explicitly modeled in traditional end-to-end driving systems, as an auxiliary objective into the autoregressive LLM training pipeline. It enables the model to internalize physical feasibility and dynamic multi-agent interactions, improving robustness in safety-critical scenarios.

Built upon open-source large language models, OpenDriveVLA tightly integrates spatially-grounded multimodal reasoning and driving trajectory generation within a unified autoregressive framework. Unlike prior VLM-based methods, OpenDriveVLA leverages structured 2D and 3D instance-aware representations, ego vehicle states, and high-level commands to directly produce reliable driving actions. Extensive experiments on nuScenes benchmark demonstrate that OpenDriveVLA achieves state-of-the-art performance in both open-loop planning and vision-language reasoning tasks. Our key contributions are:

- We present OpenDriveVLA, a 3D vision-language action model for end-to-end autonomous driving that generates reliable driving trajectories by integrating hierarchical visual input, ego state, and high-level language commands.
- We develop a multi-stage training strategy that aligns structured 2D and 3D visual features into a unified semantic space, enabling naive VLMs to generate spatially-grounded actions in complex driving scenarios.

*Corresponding author

- We introduce implicit agent–environment–ego interaction modeling into autoregressive LLM-based VLA training as an auxiliary task, enabling the model to learn behaviorally grounded and safety-aware driving actions.

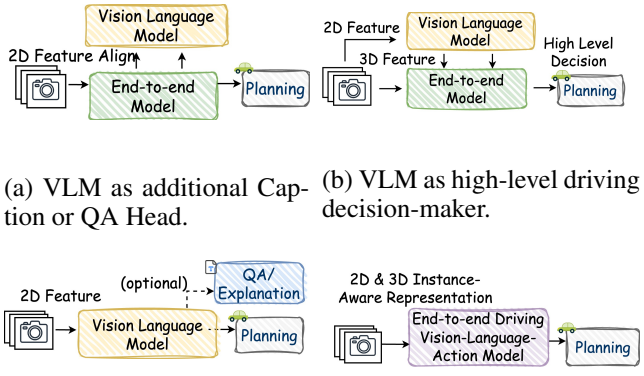
## Related Work

### End-to-End Autonomous Driving

Autonomous driving (AD) evolves through two distinct stages. Traditional approaches rely on a modular design, decomposing the system into perception (Li et al. 2022), prediction (Zhang et al. 2024b), and planning (Hu et al. 2021) components. While this structure ensures interpretability and allows for independent optimization, they suffer from cascading errors between stages and are not globally optimized for the final planning objective. In contrast, end-to-end autonomous driving frameworks (Hu et al. 2023) address this by jointly optimizing perception, prediction, and planning within a unified neural network. These models learn driving policies directly from raw sensor inputs, which improves the model's adaptability to diverse driving conditions. More recent approaches introduce diffusion models (Liao et al. 2024) and unified scene representations (Jia et al. 2025) to further enhance the effectiveness and robustness. However, existing end-to-end methods still face semantic reasoning bottlenecks, as they struggle to fully comprehend high-level scene semantics, infer complex agent interactions, and adapt to dynamic task requirements. Moreover, their decision-making processes remain opaque, making it difficult to diagnose failure cases, especially in long-tail or unseen scenarios.

### Large Vision Language Models

Large Language Models demonstrated strong emergent capabilities in in-context learning, instruction following, and reasoning (Touvron et al. 2023; Yang, Yang, and et al. 2024). By training on vast amounts of Internet-scale data, these models acquire extensive world knowledge and exhibit strong adaptability across diverse tasks. Their success has also driven the rise of large VLMs, which extend these capabilities into cross-modal reasoning by integrating vision encoders with language models. State-of-the-art VLMs such as GPT-4V (OpenAI et al. 2024), LLaVA (Liu et al. 2024b), and Qwen-VL (Bai et al. 2023) demonstrate strong visual understanding and multimodal reasoning in open-domain tasks. However, these models are primarily trained on static 2D images or videos and exhibit limited spatial reasoning in dynamic 3D driving environments. Moreover, VLMs are prone to hallucinations and generally over-confident but incorrect descriptions, which pose serious risks in safety-critical planning scenarios. Recently, Vision-Language Action models have emerged to directly predict actions from visual inputs, demonstrating strong performance in robotic manipulation tasks (Kim et al. 2024). Currently, the application of such language-conditioned end-to-end action generation in autonomous driving remains underexplored. Yet, these methods are mostly limited to static setups and lack driving-specific 3D spatial design.




(a) VLM as additional Caption or QA Head.

(b) VLM as high-level driving decision-maker.

(c) Native 2D VLM for end-to-end driving.

(d) 3D spatial-aware driving VLA (ours).

Figure 1: Taxonomy of vision-language model applications in end-to-end autonomous driving.

### Vision Language Models in Autonomous Driving

VLMs have been applied to various autonomous driving tasks, including perception, scene description, synthetic data generation, and high-level decision-making (Zhou et al. 2024). These efforts aim to enhance interpretability, data efficiency, and instruction-following capabilities in driving models. We categorize recent works into 4 paradigms, as illustrated in Figure 1. One line of research in Fig.1 (a) integrates language heads, such as captioning or question-answering modules, into driving models to enhance the interpretability (Ding et al. 2024). The second category in Fig.1 (b) employs vision language models to generate high-level driving instructions, such as directional commands or abstract maneuvers, which are subsequently interpreted by separate planning modules into low-level controls (Jiang et al. 2024; Tian et al. 2024; Wang et al. 2023). It's also usually formed as a fast-slow dual system. This design allows VLMs to make independent semantic reasoning, but retains a separate module for end-to-end driving planning, making joint optimization challenging. The third line in Fig.1 (c) applies native VLMs with 2D visual tokens to produce driving actions, and optionally scene captions or QA responses (Jin et al. 2023; Xu et al. 2024). These methods (Mei et al. 2024; Zhang et al. 2024a; Fu et al. 2025) process 2D images without explicit modeling of the instance, 3D spatial layout, and inter-agent interactions in the driving scene. It limits their spatial reasoning ability and understanding of agent dynamics in complex traffic environments. Recent studies (Favero et al. 2024) further indicate that such instance-agnostic approaches are more prone to hallucinate, often producing overconfident or semantically inconsistent text. In this work, we investigate how to extend 2D VLMs by explicitly modeling 3D instance-aware and spatial-aware scene representations into an end-to-end autonomous driving framework, as shown in Fig.1(d). Notably, we focus on fully differentiable end-to-end models in this work, while LLM-based agentic driving systems, such as (Wang et al. 2024; Sima et al. 2023), fall outside the scope of our study.

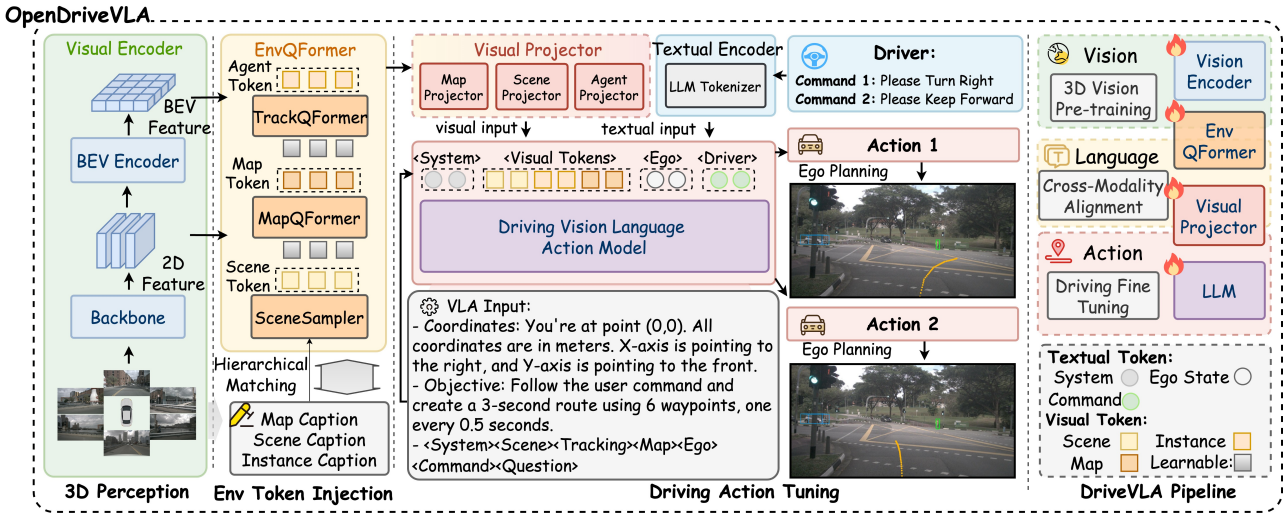



Figure 2: OpenDriveVLA leverages open-source pre-trained language foundation models to generate driving actions conditioned on 3D environmental perception, ego vehicle states, and driver commands.

## OpenDriveVLA

The overall architecture of OpenDriveVLA is shown in Figure 2, with its multi-stage training process further detailed in Figure 3. OpenDriveVLA uses a pre-trained vision encoder to extract tokenized environmental representations from multi-view images. These visual tokens are then aligned into the textual domain through cross-modal learning. After alignment, it undergoes driving instruction tuning, followed by agent-ego-environment interaction modeling. Finally, OpenDriveVLA is trained end-to-end to predict the ego vehicle's future trajectory, guided by the aligned visual-language tokens and driving instructions.

### 3D Visual Environmental Perception

Recent VLM-based autonomous driving methods typically rely on pretrained 2D visual encoders (Zhai et al. 2023b), where visual token selection and attention are indirectly guided through language supervision. While effective in open-domain vision-language applications, this design lacks explicit 3D spatial grounding and structured instance-level attention, which can lead to severe hallucinations in safety-critical driving scenarios (Xie et al. 2025). To mitigate this, OpenDriveVLA adopts a visual-centric query module, where the model first learns to focus on driving-relevant objects and map tokens through 3D vision tasks, ensuring reliable visual token proposal.

Specifically, given a set of multi-view images $I = \{I^i\}_{i=1}^N$, the visual module first extracts multi-scale 2D features from each image using a shared 2D backbone, denoted as $f_{2D}$. These 2D features are then aggregated across views and lifted into BEV space, producing the BEV feature $f_{bev}$. To obtain structured environmental representations, we adopt three visual query modules: Global Scene Sampler $\mathcal{Q}_{\text{scene}}$, Agent QueryTransformer $\mathcal{Q}_{\text{agent}}$, and Map QueryTransformer $\mathcal{Q}_{\text{map}}$. Each module extracts tokens focusing on a specific semantic aspect of the driving environment. Global Scene Sampler encodes the surrounding driving scene context from multi-view 2D features, producing the scene token $v_{scene} = \mathcal{Q}_{\text{scene}}(f_{2D})$. Agent QueryTransformer detects and tracks dynamic agents within the scene, extracting agent-centric tokens $\{v^i_{agent}\}_{i=1}^{N_a} = \mathcal{Q}_{\text{agent}}(f_{bev})$, where $N_a$ denotes the number of detected agents. In parallel, Map QueryTransformer extracts static structural information, such as lane boundaries and drivable areas, forming the map token $v_{map} = \mathcal{Q}_{\text{map}}(f_{bev})$. Through vision-centric perception tasks, including 3D detection, tracking, and segmentation, the visual encoder produces structured environmental tokens that capture both dynamic agent behaviors and static map structures in a spatially grounded manner. The output tokens, denoted as $\mathbf{V}_{env} = \{v_{scene}, v_{agent}, v_{map}\}$, serve as visual environment representation of the subsequent stages.

### Stage 1 - Hierarchical Vision-Language Alignment

To bridge the modality gap between the extracted visual tokens and the word embedding space of a pre-trained LLM, we adopt a hierarchical vision-language feature alignment strategy. Given the visual tokens extracted from the 3D visual perception module, we introduce three token-specific projectors $\{\Phi_{\text{scene}}, \Phi_{\text{agent}}, \Phi_{\text{map}}\}$. During training, each active agent query from the 3D detection and tracking task denoted as $v^i_{agent}$, is also matched to its corresponding ground-truth caption $\mathbf{X}^i_{agent}$. These captions provide detailed descriptions, including 2D appearance descriptions and 3D spatial positions. For scene and map tokens, which encode holistic spatial context and static structural properties, a sample-wise alignment is applied, where each token is matched to a scene-level caption $\mathbf{X}_{scene}$ or $\mathbf{X}_{map}$. The scene token $v_{scene}$ captures the global 2D environmental context, while the map token $v_{map}$ encodes structural elements such as lane topology, road boundaries, and drivable areas. Each of these tokens is aligned to its corresponding

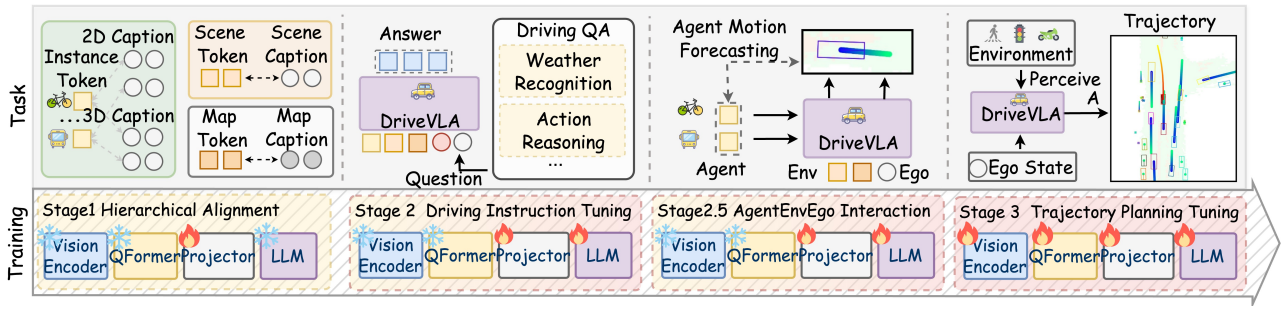



Figure 3: Illustration of main training stages on OpenDriveVLA. Stage 1: Hierarchical Feature Alignment. Stage 2: Driving Instruction Tuning. Stage 2.5: Agent-Env-Ego Interaction Modeling. Stage 3: Trajectory Planning Tuning.

caption, denoted as $\mathbf{X}_{scene}$ and $\mathbf{X}_{map}$. During this stage, both the visual encoder and LLM remain frozen to preserve pretrained semantics, with only the token-specific projectors being trainable. The forward alignment step is formulated as follows:

$$\hat{\mathbf{X}}_k = \mathrm{LLM}\left(\Phi_k(v_k)\right), \quad k \in \{\mathrm{scene}, \mathrm{map}\} \tag{1}$$

$$\hat{\mathbf{X}}^i_{agent} = \mathrm{LLM}\left(\Phi_{\mathrm{agent}}(v^i_{agent})\right), \quad i = 1, \ldots, N_a \tag{2}$$

## Stage 2 - Driving Instruction Tuning

We distill high-level driving knowledge into the model via supervised instruction tuning, enabling it to internalize semantic reasoning patterns during training. This avoids costly chain-of-thought (CoT) reasoning at inference time and balances planning efficacy with runtime efficiency.

During the tuning process, driving knowledge from the language domain is injected into the model using a curated driving instruction QA dataset. The dataset covers a wide range of driving-related reasoning, including perception understanding, motion prediction, attention allocation, action reasoning, and high-level decision-making. By training on this diverse set of driving queries, OpenDriveVLA learns to contextualize the driving scene, follow commands, and generate semantically and behaviorally grounded planning decisions. We formulate the tuning data as instruction-response pairs $\{\mathbf{X}_{input}, \mathbf{X}_{answer}\}$, where $\mathbf{X}_{input} = (\mathbf{V}_{env}, \mathbf{S}_{ego}, \mathbf{X}_{query})$. Here, $\mathbf{X}_{query}$ denotes the driving-related question, and $\mathbf{S}_{ego}$ encodes the textual ego vehicle state. Given this multimodal input, the LLM autoregressively learns to generate the target response. During instruction tuning, the visual encoder remains frozen while the token-specific projectors and the LLM are set to be trainable. The instruction prediction process is as:

$$\hat{\mathbf{X}}_{answer} = \mathrm{LLM}\left(\mathbf{V}_{env}, \mathbf{S}_{ego}, \mathbf{X}_{query}\right) \tag{3}$$

## Stage 2.5 - Agent Environment Ego Interaction

Reliable trajectory planning in autonomous driving necessitates a spatially grounded 3D representation of the environment. Beyond perception, it must also understand dynamic interactions between the ego vehicle and surrounding agents. Effective interaction modeling is essential to ensure that planned trajectories are both feasible and collision-free under real-world driving constraints. However, existing pre-trained LLMs lack an inherent inductive bias for spatial reasoning in 3D driving scenes, as they are predominantly trained on 2D vision-language and text-based datasets. We introduce a conditional agent trajectory forecasting task as an auxiliary objective, encouraging the model to learn spatially grounded interaction priors. During this stage, OpenDriveVLA captures the underlying structure of multi-agent dynamics, enhancing its capability for scene-aware trajectory generation and improving decision-making in complex traffic scenarios.

Given scene and map tokens, as well as the ego vehicle state $\mathbf{S}_{ego}$, the LLM predicts the future motion of each detected agent based on its projected visual embedding $\Phi_{\mathrm{agent}}(v^i_{agent})$. The future motion of agent $a_i$ is represented as a sequence of waypoints $\mathcal{W}^i_a$. The predicted trajectory is conditioned on the scene context, map structure, and ego vehicle state, enabling OpenDriveVLA to infer interaction-aware and spatially grounded motion sequences. The learning objective for the $i$-th agent is formulated as:

$$\max \prod_{t=1}^{T} p\left(w^i_t \mid w^i_{1:t-1}, \mathbf{V}_{env}, \mathbf{S}_{ego}, \Phi_{\mathrm{agent}}(v^i_{agent})\right) \tag{4}$$

This provides OpenDriveVLA with essential spatial priors, enabling it to bridge the gap between high-level semantic reasoning and physically grounded motion planning.

## Stage 3 - End-to-end Trajectory Planning Tuning

In this stage, OpenDriveVLA predicts ego trajectories as discrete waypoint sequences within a short horizon, denoted as $\mathcal{W}_{ego} = \{w_1, w_2, \ldots, w_T\}$. Each waypoint $w_t$ represents the 2D coordinates $(x_t, y_t)$ of the ego vehicle at time step $t$. The waypoints are tokenized into a sequence of discrete textual tokens for autoregressive generation in the LLM: $\mathcal{T}traj = \mathrm{Tokenizer}(\mathcal{W}_{ego})$. The generation process is then cast as a causal sequence prediction task, where each token is predicted in a causal manner, conditioned on the visual perception tokens $\mathbf{V}_{env}$, the ego state $\mathbf{S}_{ego}$, and the driving command $\mathbf{X}_{dri}$.

| Method | ST-P3 metrics | | | | | | | | UniAD metrics | | | | | | | | LLM | Input |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L2 (m) ↓ | | | | Collision (%) ↓ | | | | L2 (m) ↓ | | | | Collision (%) ↓ | | | | | |
| | 1s | 2s | 3s | Avg. | 1s | 2s | 3s | Avg. | 1s | 2s | 3s | Avg. | 1s | 2s | 3s | Avg. | | |
| **None-Autoregressive Methods** | | | | | | | | | | | | | | | | | | |
| ST-P3 (Hu et al. 2022) | 1.33 | 2.11 | 2.90 | 2.11 | 0.23 | 0.62 | 1.27 | 0.71 | - | - | - | - | - | - | - | - | - | Visual |
| VAD (Jiang et al. 2023) | 0.17 | 0.34 | 0.60 | 0.37 | 0.07 | 0.10 | 0.24 | 0.14 | - | - | - | - | - | - | - | - | - | Visual |
| Ego-MLP (Zhai et al. 2023a) | 0.46 | 0.76 | 1.12 | 0.78 | 0.21 | 0.35 | 0.58 | 0.38 | - | - | - | - | - | - | - | - | - | Ego |
| UniAD (Hu et al. 2023) | 0.44 | 0.67 | 0.96 | 0.69 | 0.04 | 0.08 | 0.23 | 0.12 | 0.48 | 0.96 | 1.65 | 1.03 | 0.05 | 0.17 | 0.71 | 0.31 | - | Visual |
| InsightDrive (Song et al. 2025) | 0.23 | 0.41 | 0.68 | 0.44 | 0.09 | 0.10 | 0.27 | 0.15 | 0.30 | 0.72 | 1.41 | 0.81 | 0.08 | 0.15 | 0.84 | 0.36 | - | Visual |
| FF (Hu et al. 2021) | - | - | - | - | - | - | - | - | 0.55 | 1.20 | 2.54 | 1.43 | 0.06 | 0.17 | 1.07 | 0.43 | - | LiDAR |
| EO (Khurana et al. 2022) | - | - | - | - | - | - | - | - | 0.67 | 1.36 | 2.78 | 1.60 | 0.04 | **0.09** | 0.88 | 0.33 | - | LiDAR |
| **Autoregressive Methods** | | | | | | | | | | | | | | | | | | |
| GPVL (Li et al. 2025) | 0.21 | 0.39 | 0.69 | 0.43 | 0.07 | 0.09 | 0.27 | 0.14 | - | - | - | - | - | - | - | - | BERT | Textual |
| DriveVLM (Tian et al. 2024) | 0.18 | 0.34 | 0.68 | 0.40 | 0.10 | 0.22 | 0.45 | 0.27 | - | - | - | - | - | - | - | - | Qwen-VL-7B | Visual |
| GPT-Driver (Mao et al. 2023) | 0.20 | 0.40 | 0.70 | 0.44 | 0.04 | 0.12 | 0.36 | 0.17 | 0.27 | 0.74 | 1.52 | 0.84 | 0.07 | 0.15 | 1.10 | 0.44 | GPT-3.5 | Textual |
| RDA-Driver (Huang et al. 2024) | 0.17 | 0.37 | 0.69 | 0.40 | 0.01 | **0.05** | 0.26 | 0.10 | 0.23 | 0.73 | 1.54 | 0.80 | **0.00** | 0.13 | 0.83 | 0.32 | LLaVa-7B | Visual |
| OminiDrive (Wang et al. 2024) | 0.14 | 0.29 | 0.55 | 0.33 | **0.00** | 0.13 | 0.78 | 0.30 | - | - | - | - | - | - | - | - | LLaVA-7B | Visual |
| EMMA (Hwang et al. 2024) | **0.14** | **0.29** | **0.54** | **0.32** | - | - | - | - | - | - | - | - | - | - | - | - | **Gemini** | Visual |
| OpenEMMA (Xing et al. 2025) | 1.45 | 3.21 | 3.76 | 2.81 | - | - | - | - | - | - | - | - | - | - | - | - | Qwen-VL-7B | Visual |
| DME-Driver (Han et al. 2024) | - | - | - | - | - | - | - | - | 0.45 | 0.91 | 1.58 | 0.98 | 0.05 | 0.28 | **0.55** | 0.29 | LLaVa-7B | Visual |
| OpenDriveVLA-0.5B (Ours) | 0.15 | 0.32 | 0.57 | 0.35 | 0.01 | 0.06 | **0.20** | **0.09** | 0.21 | 0.60 | 1.22 | 0.68 | **0.00** | 0.15 | 0.63 | 0.26 | Qwen2.5-0.5B | Visual |
| OpenDriveVLA-3B (Ours) | **0.14** | 0.30 | 0.55 | 0.33 | 0.02 | 0.07 | 0.22 | 0.10 | **0.19** | **0.58** | 1.24 | 0.67 | 0.02 | 0.18 | 0.70 | 0.30 | Qwen2.5-3B | Visual |
| OpenDriveVLA-7B (Ours) | 0.15 | 0.31 | 0.55 | 0.33 | 0.01 | 0.08 | 0.21 | 0.10 | 0.20 | **0.58** | **1.21** | **0.66** | **0.00** | 0.22 | **0.55** | **0.25** | Qwen2.5-7B | Visual |

Table 1: Open-Loop planning performance comparison of different driving models, including both autoregressive methods and non-autoregressive methods. OpenDriveVLA shows powerful planning ability and achieves best-in-class results among open-source models, even with the 0.5B version. We refer to the result summary from (Song et al. 2025; Mao et al. 2023; Li et al. 2025; Huang et al. 2024).

$$\hat{\mathcal{T}}_{traj} = \text{argmax}_{\mathbf{T}_{traj}} \prod_{t=1}^{T} p\left(w_t \mid w_{1:t-1}, \mathbf{V}_{env}, \mathbf{S}_{ego}, \mathbf{X}_{dri}\right) \tag{5}$$

The entire pipeline, including the 3D visual encoder, cross-modality projectors, and LLM, is jointly optimized end-to-end during training, with the 2D encoder kept frozen. At inference, the model autoregressively generates the tokenized trajectory $\hat{\mathcal{T}}_{traj}$, which is then decoded back into numerical waypoints:

$$\hat{\mathcal{W}}_{ego} = \text{Decoder}(\hat{\mathcal{T}}_{traj}) \tag{6}$$

# Experiments

## Training Datasets

We curate the training data of OpenDriveVLA based on its distinct training phases, drawing from: TOD3Cap (Jin et al. 2024), nuCaption (Yang et al. 2023), nuScenesQA (Qian et al. 2023), nuX (Ding et al. 2024), and GPT-Driver (Mao et al. 2023). We conduct experiments on nuScenes (Caesar et al. 2020), following standard data split into training and validation sets. OpenDriveVLA is trained using the training set paired with corresponding QA captions, while the validation set is exclusively used for performance evaluation to ensure fair comparisons with prior works. The details of training data can be found in supplementary materials.

**Hierarchical Vision-Language Alignment.** For agent-level caption, we post-process data from (Jin et al. 2024), which provides the 2D visual description of individual objects. To further enhance spatial grounding, each object caption is augmented with its corresponding BEV coordinates, enabling the model to associate object attributes with precise spatial locations. For scene tokens, we process multi-view scene descriptions from (Yang et al. 2023), merging them into unified summaries that describe the driving environment across all camera views. For map tokens, structured language descriptions are derived from ground-truth annotations, translating map elements such as lane dividers, crosswalks, and road boundaries into descriptive text.

**Driving Instruction Tuning.** We adopt multiple instruction-oriented datasets derived from nuScenes to inject driving-specific knowledge into OpenDriveVLA. We unify several datasets into a standardized instruction-based QA format, including driving-related question-answer pairs collected from nuCaption (Yang et al. 2023), nuScenesQA (Qian et al. 2023), and nuX (Ding et al. 2024) dataset. Each QA pair is conditioned on structured environmental visual tokens and the ego vehicle state, ensuring consistency across different data sources. This multimodal instruction tuning process allows OpenDriveVLA to effectively ground language understanding into both environmental perception and scene understanding, bridging perception, reasoning, and action within the language space.

**Motion Forecasting and Trajectory Prediction.** We formulate both agent motion forecasting and ego trajectory planning in the ego system, where the model directly predicts future displacements within each entity's local coordinate frame relative to the ego vehicle for planning and relative to each agent for forecasting. This formulation captures motion dynamics in a spatially consistent manner across all

| Method | nu-Caption | | | | | nuScenes-QA | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | BL-1 | BL-2 | BL-3 | BL-4 | BERT-S | Ext | Cnt | Obj | Sts | Cmp | H0 | H1 | Acc |
| Mini-GPT4 (Zhu et al. 2024) | 15.0 | 6.8 | 3.7 | 2.6 | 84.4 | - | - | - | - | - | - | - | - |
| Instruct-BLIP (Dai and et al. 2023) | 18.7 | 13.4 | 7.4 | 5.2 | 85.9 | - | - | - | - | - | - | - | - |
| LLaMA-AdapV2 (Gao et al. 2023) | 30.2 | 17.3 | 10.4 | 7.5 | 86.5 | 19.3 | 2.7 | 7.6 | 10.8 | 1.6 | 15.1 | 4.8 | 9.6 |
| LLaVA1.5 (Liu et al. 2024a) | 20.0 | 12.1 | 8.6 | 5.4 | 85.0 | 45.8 | 7.7 | 7.8 | 9.0 | 52.1 | 25.7 | 41.5 | 26.2 |
| LiDAR-LLM (Yang et al. 2023) | 41.0 | 30.0 | 23.4 | 19.3 | 91.3 | 74.5 | 15.0 | 37.8 | 45.9 | 57.8 | - | - | 48.6 |
| BEVDet+BUTD (Qian et al. 2023) | - | - | - | - | - | 83.7 | 20.9 | 48.8 | 52.0 | 67.7 | - | - | 57.0 |
| OpenDriveVLA-0.5B (Ours) | 47.2 | 35.8 | 29.4 | 25.2 | 91.9 | 83.9 | 22.0 | 50.2 | **57.0** | 68.4 | 62.3 | **56.5** | 58.4 |
| OpenDriveVLA-3B (Ours) | 48.3 | 36.9 | 30.3 | 26.1 | 92.0 | 84.0 | 22.3 | **50.3** | 56.9 | 68.5 | **62.6** | **56.5** | **58.5** |
| OpenDriveVLA-7B (Ours) | **49.6** | **38.3** | **31.9** | **27.6** | **92.2** | **84.2** | **22.7** | 49.6 | 54.5 | **68.8** | 62.4 | 56.1 | 58.2 |

Table 2: Performance on nu-Caption (Yang et al. 2023) and nuScenes-QA (Qian et al. 2023). BL-1/2/3/4: BLEU scores. QA metrics report accuracy on five question types: Existence, Counting, Object, Status, and Comparison.

entities. Following (Mao et al. 2023), the ego vehicle state is encoded as textual input to ensure ego awareness throughout the training process. Both tasks predict 3-second future trajectories, sampled at 0.5-second intervals, resulting in 6 waypoints per trajectory.

## Evaluations

We evaluate OpenDriveVLA on the open-loop planning task of nuScenes benchmark, where the model is reported under both ST-P3 (Hu et al. 2022) and UniAD (Hu et al. 2023) settings. The evaluation metrics include L2 displacement errors at 1, 2, and 3 seconds, along with the average collision rate over the prediction horizon. To further assess the scene understanding ability of OpenDriveVLA, we report its QA prediction performance on three driving visual question answering (VQA) datasets directly after the driving instruction tuning stage, i.e., (Yang et al. 2023), nuScenesQA (Qian et al. 2023), and nuX (Ding et al. 2024). The VQA evaluation results adopt standard NLG metrics, including BLEU, METEOR, CIDEr, BERT-Score, etc.

## Implementation Details

The 3D visual perception module in OpenDriveVLA follows the vision-centric design from (Hu et al. 2023), using a ResNet-101 backbone for 2D feature extraction. The perception backbone is pre-trained via multi-task learning on 3D object detection, object tracking, and map segmentation. The resulting BEV feature map has a spatial resolution of $200 \times 200$. To construct a unified scene representation, the global SceneSampler applies 2D adaptive pooling to each camera view, subsequently concatenating the pooled multi-view features into a global scene token. Agent and map tokens are extracted from the final layer of their respective QueryTransformer modules. Each token type is then mapped into the language space using a separate two-layer MLP with GeLU activation. We adopt Qwen 2.5-Instruct (Yang, Yang, and et al. 2024) as the pre-trained LLM, which undergoes full parameter tuning during training. Training is performed on 4 NVIDIA H100 GPUs with a batch size of 1, completed in approximately two days. We freeze the 2D backbone during stage 3. During inference, we set the decoding temperature to 0 to ensure deterministic trajectory generation. See supplementary material for detailed training configurations.

## Main Results

**Open Loop Trajectory Planning.** We evaluate OpenDriveVLA on the open-loop trajectory planning task using both ST-P3 and UniAD metrics, ensuring comprehensive performance assessment across spatial accuracy and collision avoidance. As shown in Table 1, OpenDriveVLA achieves state-of-the-art performance across both settings. Specifically, both 3B and 7B version models achieve an average L2 error of 0.33m under ST-P3 metrics, outperforming prior autoregressive language models (Mao et al. 2023; Tian et al. 2024). On the UniAD metrics, OpenDriveVLA-7B also achieves great performance with an average L2 error of 0.66m. Notably, despite significantly fewer parameters, the 0.5B version still outperforms prior models obviously.

| Models | CIDER | BL-4 | METEOR | ROUGE-L |
|---|---|---|---|---|
| Hint-UniAD (Ding et al. 2024) | 21.7 | 4.2 | 12.7 | 27.0 |
| Hint-VAD (Ding et al. 2024) | 22.4 | 4.2 | **13.2** | 27.6 |
| GPT-4o (Xu et al. 2024) | 19.0 | 4.0 | 10.3 | 24.9 |
| Gemini 1.5 (Team et al. 2024) | 17.6 | 3.4 | 9.3 | 23.4 |
| Vote2CapDETR (Chen et al. 2023) | 15.3 | 2.6 | 10.9 | 24.2 |
| TOD$^3$Cap (Jin et al. 2024) | 14.5 | 2.5 | 10.5 | 23.5 |
| **OpenDriveVLA** | | | | |
| 0.5B (Ours) | **32.3** | **5.4** | 12.5 | **27.9** |
| 3B (Ours) | 25.5 | 4.3 | 12.8 | 27.8 |
| 7B (Ours) | 26.2 | 4.5 | 12.8 | 27.4 |

Table 3: Performance comparison of OpenDriveVLA on the Nu-X dataset (Ding et al. 2024).

**Driving Question Answering.** We access OpenDriveVLA on the driving VQA task across three nuScenes-based datasets (Table 2, Table 3), reporting results after the second stage of training. OpenDriveVLA reaches best-in-class performance across all three datasets, consistently outperforming previous language-enhanced driving models and general-purpose multimodal baselines among most metrics. On nuCaption dataset, it achieves the best captioning performance among all evaluated models, outperforming both general VLMs LLaVA1.5 (Liu et al. 2024a) and Mini-GPT4 (Zhu et al. 2024), as well as autonomous driving-specific models such as LiDAR-LLM (Yang et al. 2023). For nuScenesQA dataset, OpenDriveVLA also achieves strong performance. Compared to models that directly fuse BEV features with language models such as BEVDet+BUTD (Qian et al. 2023), it demonstrates clear advantages in object and status-

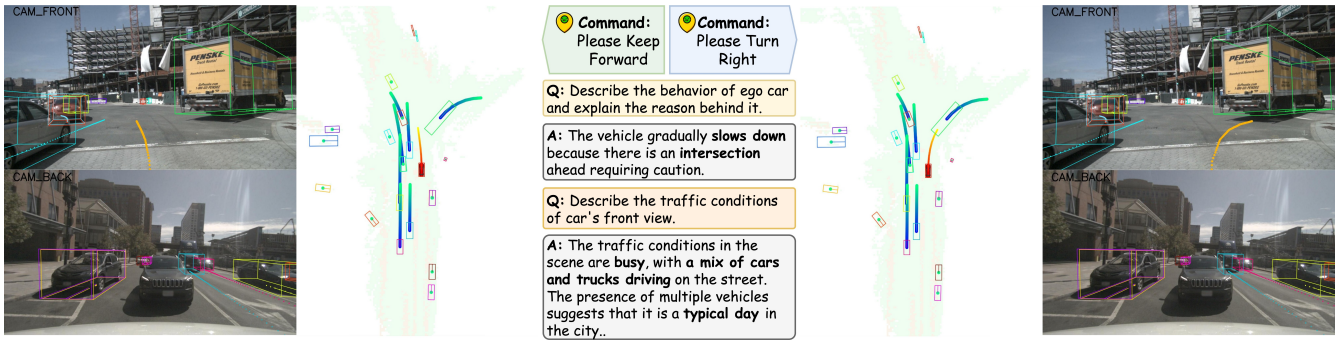



Figure 4: Visualization of OpenDriveVLA-7B planning actions under original dataset instruction to keep forward (left) and modified instruction to turn right (right). The QA prediction showcases (middle) are from results reported in Table 2 and Table 3. The agent motion prediction results are visualized after the agent-env-ego interaction stage.

related questions, which highlights the benefit of its spatially grounded visual-language alignment. Notably, the 0.5B version outperforms even the larger 7B on the Nu-X dataset, which shows its powerful scene-understanding ability even with lightweight LLMs.

## Ablation Study

We conduct ablation studies to evaluate the impact of input modalities and our multi-stage training strategy on OpenDriveVLA's performance. Additionally, we qualitatively assess the model's ability to follow diverse driving commands.

| Visu | Ego | Hist | Cmd | Avg. Collision (%) ↓ | | Avg. L2 (m) ↓ | |
|---|---|---|---|---|---|---|---|
| | | | | UniAD | ST-P3 | UniAD | ST-P3 |
| ✓ | | ✓ | ✓ | 0.77 | 0.24 | 1.34 | 0.75 |
| ✓ | ✓ | | ✓ | 1.14 | 0.49 | 1.30 | 0.75 |
| | ✓ | ✓ | ✓ | 0.29 | 0.10 | 0.77 | 0.39 |
| ✓ | ✓ | ✓ | | 0.33 | 0.13 | 0.80 | 0.40 |
| ✓ | ✓ | ✓ | ✓ | 0.26 | 0.09 | 0.68 | 0.35 |

Table 4: Ablation study on the effect of different input combinations on OpenDriveVLA-0.5B.

**Effect of Input Modalities.** We investigate how individual input components contribute to trajectory planning. Table 4 presents the results of ablating visual perception, ego state, historical trajectory, and high-level language commands. The inclusion of visual inputs significantly boosts overall performance. Adding textual commands and historical information further improves the predictions, emphasizing the value of semantic intent and temporal context. Notably, ego-state features play a critical role in nuScenes open-loop benchmark, consistent with prior findings (Li et al. 2024).

**Effect of Multi-Stage Training Strategy.** We evaluate the contribution of each training phase in our staged pipeline incrementally. As shown in Table 5, each additional stage consistently improves performance, with the most notable reductions in collision rate observed after Hierarchical Vision-Language Alignment and Agent-Environment-Ego Interaction Modeling. These improvements highlight the effectiveness of cross-modal grounding and interaction-aware reasoning in enhancing safety-critical planning behavior.

| Training Stage | | | | Avg. Collision (%) ↓ | | Avg. L2 (m) ↓ | |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 2.5 | 3 | UniAD | ST-P3 | UniAD | ST-P3 |
| | | | ✓ | 0.37 | 0.13 | 0.70 | 0.36 |
| ✓ | | | ✓ | 0.32 | 0.12 | 0.69 | 0.35 |
| ✓ | ✓ | | ✓ | 0.31 | 0.11 | 0.68 | 0.35 |
| ✓ | ✓ | ✓ | ✓ | 0.26 | 0.09 | 0.68 | 0.35 |

Table 5: Ablation study on the effect of multi-stage training of 0.5B model. Stage 1, 2, 2.5, and 3 correspond to hierarchical feature alignment, driving instruction tuning, Agent-Env-Ego modeling, and trajectory tuning, respectively.

**Effect of Driving Command.** Figure 4 presents the qualitative comparison at an intersection under two different driver instructions: keep forward and turn right, with the right turn as the ground truth. OpenDriveVLA accurately adapts its plan to the given command while maintaining context-aware and environment-consistent behavior, demonstrating robust command-following and generalization in complex scenes. In addition, we visualize the QA predictions for the same scene, showcasing the model's ability to reason over decision-making and traffic scene understanding.

## Conclusion

In this work, we present OpenDriveVLA, a scalable vision-language action model designed for end-to-end autonomous driving. Built upon pre-trained large language models, OpenDriveVLA generates 3D spatially grounded and semantically consistent driving actions from multimodal inputs. We introduce a hierarchical vision-language feature alignment module and realize agent-env-ego interaction in LLM to enable fine-grained spatial reasoning and dynamic scene understanding. Through multi-stage training paradigm, OpenDriveVLA achieves state-of-the-art performance in open-loop planning and driving-related question answering. Extensive evaluations on nuScenes dataset show its superior trajectory planning capability compared to existing approaches. Our work demonstrates the feasibility of a scalable vision-language-driven approach for autonomous driving and highlights the potential of large language models as a foundation for end-to-end driving action systems.

## References


Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, L.; Wu, P.; Chitta, K.; Jaeger, B.; Geiger, A.; and Li, H. 2024. End-to-end Autonomous Driving: Challenges and Frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, S.; Zhu, H.; Chen, X.; Lei, Y.; Yu, G.; and Chen, T. 2023. End-to-end 3d dense captioning with vote2cap-detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11124–11133.

Dai, W.; and et al., J. L. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ding, K.; Chen, B.; Su, Y.; Gao, H.-a.; Jin, B.; Sima, C.; Li, X.; Zhang, W.; Barsch, P.; and Li, H. e. a. 2024. HintAD: Holistically Aligned Interpretability in End-to-End Autonomous Driving. In *8th Annual Conference on Robot Learning*.

Favero, A.; Zancato, L.; Trager, M.; Choudhary, S.; Perera, P.; Achille, A.; Swaminathan, A.; and Soatto, S. 2024. Multi-Modal Hallucination Control by Visual Information Grounding. arXiv:2403.14003.

Fu, H.; Zhang, D.; Zhao, Z.; Cui, J.; Liang, D.; Zhang, C.; Zhang, D.; Xie, H.; Wang, B.; and Bai, X. 2025. ORION: A Holistic End-to-End Autonomous Driving Framework by Vision-Language Instructed Action Generation. arXiv:2503.19755.

Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; and Lu, P. e. a. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Han, W.; Guo, D.; Xu, C.-Z.; and Shen, J. 2024. DME-Driver: Integrating Human Decision Logic and 3D Scene Perception in Autonomous Driving. arXiv:2401.03641.

Hu, P.; Huang, A.; Dolan, J.; Held, D.; and Ramanan, D. 2021. Safe Local Motion Planning With Self-Supervised Freespace Forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12732–12741.

Hu, S.; Chen, L.; Wu, P.; Li, H.; Yan, J.; and Tao, D. 2022. ST-P3: End-to-end Vision-based Autonomous Driving via Spatial-Temporal Feature Learning. In *European Conference on Computer Vision (ECCV)*.

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; Lu, L.; Jia, X.; Liu, Q.; Dai, J.; Qiao, Y.; and Li, H. 2023. Planning-oriented Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Huang, Z.; Tang, T.; Chen, S.; Lin, S.; Jie, Z.; Ma, L.; Wang, G.; and Liang, X. 2024. Making Large Language Models Better Planners with Reasoning-Decision Alignment. arXiv:2408.13890.

Hwang, J.-J.; Xu, R.; Lin, H.; Hung, W.-C.; Ji, J.; Choi, K.; Huang, D.; He, T.; Covington, P.; Sapp, B.; Zhou, Y.; Guo, J.; Anguelov, D.; and Tan, M. 2024. EMMA: End-to-End Multimodal Model for Autonomous Driving. arXiv:2410.23262.

Jia, X.; You, J.; Zhang, Z.; and Yan, J. 2025. DriveTransformer: Unified Transformer for Scalable End-to-End Autonomous Driving. In *The Thirteenth International Conference on Learning Representations*.

Jiang, B.; Chen, S.; Liao, B.; Zhang, X.; Yin, W.; Zhang, Q.; Huang, C.; Liu, W.; and Wang, X. 2024. Senna: Bridging Large Vision-Language Models and End-to-End Autonomous Driving. arXiv:2410.22313.

Jiang, B.; Chen, S.; Xu, Q.; Liao, B.; Chen, J.; Zhou, H.; Zhang, Q.; Liu, W.; Huang, C.; and Wang, X. 2023. VAD: Vectorized Scene Representation for Efficient Autonomous Driving. *ICCV*.

Jin, B.; Liu, X.; Zheng, Y.; Li, P.; and et al., H. Z. 2023. ADAPT: Action-aware Driving Caption Transformer. arXiv:2302.00673.

Jin, B.; Zheng, Y.; Li, P.; Li, W.; Zheng, Y.; and Hu, S. e. a. 2024. TOD3Cap: Towards 3D Dense Captioning. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29 – October 4, 2024, Proceedings, Part XVIII*, 367–384. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-031-72648-4.

Khurana, T.; Hu, P.; Dave, A.; Ziglar, J.; Held, D.; and Ramanan, D. 2022. Differentiable Raycasting for Self-Supervised Occupancy Forecasting. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, 353–369. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-031-19838-0.

Kim, M.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E.; Lam, G.; Sanketi, P.; Vuong, Q.; Kollar, T.; Burchfiel, B.; Tedrake, R.; Sadigh, D.; Levine, S.; Liang, P.; and Finn, C. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. *arXiv preprint arXiv:2406.09246*.

Li, T.; Wang, H.; Li, X.; Liao, W.; He, T.; and Peng, P. 2025. Generative Planning with 3D-vision Language Pre-training for End-to-End Autonomous Driving. arXiv:2501.08861.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022. BEVFormer: Learning Bird's-Eye-View Representation from nbsp;Multi-camera Images vinbsp;Spatiotemporal Transformers. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, 1–18. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-031-20076-2.

Li, Z.; Yu, Z.; Lan, S.; Li, J.; Kautz, J.; Lu, T.; and Alvarez, J. M. 2024. Is Ego Status All You Need for Open-Loop End-to-End Autonomous Driving? In *2024 IEEE/CVF Confer-*

*ence on Computer Vision and Pattern Recognition (CVPR)*, 14864–14873.

Liao, B.; Chen, S.; Yin, H.; Jiang, B.; Wang, C.; Yan, S.; Zhang, X.; Li, X.; Zhang, Y.; Zhang, Q.; and Wang, X. 2024. DiffusionDrive: Truncated Diffusion Model for End-to-End Autonomous Driving. *arXiv preprint arXiv:2411.15139*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; and Peng, W. 2024c. A Survey on Hallucination in Large Vision-Language Models. arXiv:2402.00253.

Liu, M.; Yurtsever, E.; Fossaert, J.; Zhou, X.; Zimmer, W.; Cui, Y.; Zagar, B. L.; and Knoll, A. C. 2024d. A Survey on Autonomous Driving Datasets: Statistics, Annotation Quality, and a Future Outlook. *IEEE Transactions on Intelligent Vehicles*, 1–29.

Mao, J.; Qian, Y.; Ye, J.; Zhao, H.; and Wang, Y. 2023. GPT-Driver: Learning to Drive with GPT. arXiv:2310.01415.

Mei, J.; Ma, Y.; Yang, X.; Wen, L.; Cai, X.; Li, X.; Fu, D.; Zhang, B.; Cai, P.; Dou, M.; Shi, B.; He, L.; Liu, Y.; and Qiao, Y. 2024. Continuously Learning, Adapting, and Improving: A Dual-Process Approach to Autonomous Driving. arXiv:2405.15324.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; and et al., J. A. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Qian, T.; Chen, J.; Zhuo, L.; Jiao, Y.; and Jiang, Y.-G. 2023. NuScenes-QA: A Multi-modal Visual Question Answering Benchmark for Autonomous Driving Scenario. *arXiv preprint arXiv:2305.14836*.

Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Luo, P.; Geiger, A.; and Li, H. 2023. DriveLM: Driving with Graph Visual Question Answering. *arXiv preprint arXiv:2312.14150*.

Song, R.; Guo, X.; Wu, H.; Wei, Q.; and Chen, L. 2025. InsightDrive: Insight Scene Representation for End-to-End Autonomous Driving. arXiv:2503.13047.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; Mariooryad, S.; Ding, Y.; Geng, X.; Alcober, F.; Frostig, R.; Omernick, M.; and et al., L. W. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530.

Tian, X.; Gu, J.; Li, B.; Liu, Y.; Wang, Y.; Zhao, Z.; Zhan, K.; Jia, P.; Lang, X.; and Zhao, H. 2024. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. In *8th Annual Conference on Robot Learning*.

Touvron, H.; Lavril, T.; Izacard, G.; and Xavier Martinet, e. a. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Wang, S.; Yu, Z.; Jiang, X.; Lan, S.; Shi, M.; Chang, N.; Kautz, J.; Li, Y.; and Alvarez, J. M. 2024. OmniDrive: A Holistic LLM-Agent Framework for Autonomous Driving with 3D Perception, Reasoning and Planning. *arXiv:2405.01533*.

Wang, W.; Xie, J.; Hu, C.; Zou, H.; Fan, J.; Tong, W.; Wen, Y.; Wu, S.; Deng, H.; Li, Z.; et al. 2023. DriveMLM: Aligning Multi-Modal Large Language Models with Behavioral Planning States for Autonomous Driving. *arXiv preprint arXiv:2312.09245*.

Xie, S.; Kong, L.; Dong, Y.; Sima, C.; and et al., W. Z. 2025. Are VLMs Ready for Autonomous Driving? An Empirical Study from the Reliability, Data, and Metric Perspectives. arXiv:2501.04003.

Xing, S.; Qian, C.; Wang, Y.; Hua, H.; Tian, K.; Zhou, Y.; and Tu, Z. 2025. OpenEMMA: Open-Source Multimodal Model for End-to-End Autonomous Driving. arXiv:2412.15208.

Xu, Z.; Zhang, Y.; Xie, E.; Zhao, Z.; Guo, Y.; Wong, K.-Y. K.; Li, Z.; and Zhao, H. 2024. DriveGPT4: Interpretable End-to-End Autonomous Driving Via Large Language Model. *IEEE Robotics and Automation Letters*, 9(10): 8186–8193.

Yang, A.; Yang, B.; and et al., B. Z. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Yang, S.; Liu, J.; Zhang, R.; Pan, M.; Guo, Z.; Li, X.; Chen, Z.; Gao, P.; Guo, Y.; and Zhang, S. 2023. LiDAR-LLM: Exploring the Potential of Large Language Models for 3D LiDAR Understanding. arXiv:2312.14074.

Zhai, J.-T.; Feng, Z.; Du, J.; Mao, Y.; Liu, J.-J.; Tan, Z.; Zhang, Y.; Ye, X.; and Wang, J. 2023a. Rethinking the Open-Loop Evaluation of End-to-End Autonomous Driving in nuScenes. arXiv:2305.10430.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and et al., L. B. 2023b. Sigmoid Loss for Language Image Pre-Training. arXiv:2303.15343.

Zhang, J.; Huang, Z.; Ray, A.; and Ohn-Bar, E. 2024a. Feedback-Guided Autonomous Driving. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15000–15011.

Zhang, S.; Zhai, Y.; Mei, J.; and Hu, Y. 2024b. FusionOcc: Multi-Modal Fusion for 3D Occupancy Prediction. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 787–796. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.

Zhou, X.; and Knoll, A. C. 2024. GPT-4V as Traffic Assistant: An In-depth Look at Vision Language Model on Complex Traffic Events. arXiv:2402.02205.

Zhou, X.; Liu, M.; Yurtsever, E.; Zagar, B. L.; Zimmer, W.; Cao, H.; and Knoll, A. C. 2024. Vision Language Models in Autonomous Driving: A Survey and Outlook. *IEEE Transactions on Intelligent Vehicles*, 1–20.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *The Twelfth International Conference on Learning Representations*.