PixelPilot: Scalable Vision-Language-Action Models for End-to-End Autonomous Driving

Pin Tang[1], Guoqing Wang[1], Xiangxuan Ren[1], Zhongdao Wang[2], Guodongfang Zhao[2], Bailan Feng[2], and Chao Ma[1,⋆]

[1] MoE Key Lab of Artificial Intelligence, Institute of AI, Shanghai Jiao Tong University, Shanghai, China
[2] Central Research Institute, Huawei, Beijing, China
{guoqing.wang,pin.tang,bunny_renxiangxuan,chaoma}@sjtu.edu.cn
{wangzhongdao,zhaoguodongfang,fengbailan}@huawei.com
Project page: https://pixelpilotvla.github.io/

Abstract. Vision-Language-Action Models (VLAs), which leverage the advanced reasoning capabilities of Vision-Language Models (VLMs), show promising generalization in complex autonomous driving scenarios. Existing VLAs typically predict and optimize 3D trajectories from 2D images. While intuitive, this 2D-to-3D prediction is inherently entangled with camera parameters, leading to limited data scalability across heterogeneous driving datasets. Moreover, directly optimizing in 3D space induces severe convergence to trivial solutions, where VLAs rely on ego-status rather than visual scene understanding. To address these issues, we propose PixelPilot, a novel VLA featuring a decoupled planning and lifting paradigm. In the planning phase, PixelPilot reformulates scene understanding and trajectory prediction as sensor-agnostic 2D-to-2D tasks in the image plane, thereby facilitating scalable training across diverse datasets. The planned 2D trajectories are then deterministically lifted to 3D only during inference, ensuring the full exploitation of visual cues and generalization across different vehicles. To realize this paradigm, we propose a knowledge-instilled policy learning strategy that applies dense, intermediate rewards via Group Relative Policy Optimization (GRPO) to enforce a rigorous causal chain from visual perception to spatial planning. Extensive experiments demonstrate that PixelPilot achieves state-of-the-art performance in both open-loop and closed-loop settings, validating its superior scalability and visual reasoning capabilities.

Keywords: Vision-language-action models · Autonomous driving · Data scaling

1 Introduction

The emergence of Vision-Language Models (VLMs) [1,2] has catalyzed a paradigm shift in end-to-end autonomous driving. Unlike conventional methods [8, 20, 26],

⋆ Corresponding author.

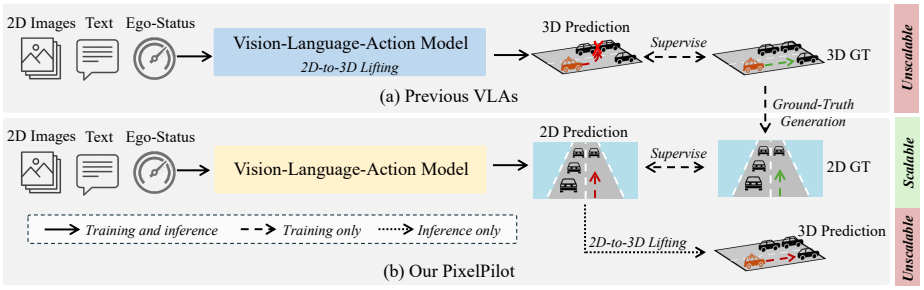


Fig. 1: (a) Previous VLAs directly predict 3D trajectories from 2D images which involves learning sensor-specific 2D-to-3D mapping, leading to limited data scalability. (b) Our PixelPilot features a decoupled planning and lifting paradigm. It plans and optimizes 2D trajectories within the image plane, facilitating data scaling and visual reasoning, while unscalable 2D-to-3D trajectory lifting occurs only during inference.

which are trained from scratch exclusively on driving datasets, Vision-Language-Action Models (VLAs) [10,27,71] leverage VLMs to integrate high-level linguistic reasoning with visual understanding, thereby improving generalization in rare and long-tail scenarios. The primary objective of these driving VLAs is to navigate physical environments by generating safe and drivable 3D trajectories. However, adapting VLMs, which are pretrained on 2D image and text corpora, to the precise 3D metric demands of autonomous driving remains a formidable challenge. To bridge this dimensional gap, existing VLAs fine-tune 2D-pretrained VLMs to directly output 3D world coordinates by incorporating auxiliary 3D modules and instructional data. While intuitive, we observe that this 2D-to-3D prediction paradigm encounters two critical problems.

First, direct 3D prediction significantly restricts the data scalability of driving VLAs. In standard 2D visual tasks (e.g., object detection [76]), scaling is highly effective due to the sensor-agnostic nature of 2D-to-2D prediction. Conversely, predicting precise 3D trajectories directly from 2D visual inputs requires the VLA to either explicitly encode [16, 54, 71, 72] or implicitly learn [10, 51, 67, 68, 73] sensor-specific 2D-to-3D mappings within its network weights, as illustrated in Fig. 1 (a). Because these mappings are inherently tied to specific camera intrinsics and extrinsics, the learned VLA becomes deeply entangled with the sensor configurations of its training data. Consequently, aggregating massive, heterogeneous driving datasets (e.g., nuScenes and Waymo) introduces severe spatial ambiguity: identical 2D pixel coordinates correspond to entirely disparate 3D physical locations under different camera setups.

Second, this sensor-specific 3D optimization severely compromises the visual reasoning capabilities of VLAs. Confronted with complex and dataset-specific 2D-to-3D mappings that are challenging to optimize, existing VLAs are highly susceptible to spurious correlations and often converge to trivial solutions [69]. Rather than learning robust 3D scene structures and planning trajectories based on visual inputs, they predominantly rely on ego-status (e.g., speed and accelera-

tion) to directly extrapolate the 3D trajectory, as revealed by empirical evidence that removing ego-status from state-of-the-art VLAs [10, 54] degrades performance far more severely than masking the input images (1.98 vs 0.36 L2 error).

To address these challenges, we introduce PixelPilot, a novel VLA featuring a *decoupled planning and lifting paradigm*, as illustrated in Fig. 1 (b). During planning, PixelPilot formulates comprehensive scene understanding and trajectory prediction from input images as 2D-to-2D tasks, which are optimized entirely in the image space in a sensor-agnostic manner. Therefore, it can aggregate heterogeneous driving datasets with available ego-motion and camera parameters, facilitating scalable training for robust driving capabilities. During the lifting phase, these predicted 2D pixel trajectories are deterministically projected into the 3D world coordinate system using the specific vehicle's camera parameters. By decoupling the 2D-to-3D lifting process from the learning objective, absolute ego-speed cannot be trivially mapped to pixel displacement without contextual depth, ensuring that the VLA actively relies on visual understanding and spatial reasoning. Additionally, as the sensor-specific 2D-to-3D lifting occurs only during inference, the learned PixelPilot can generalize to different calibrated vehicles simply by updating the sensor-specific parameters. This decoupled design conceptually aligns with human driving: skillful human drivers plan paths based on their 2D visual perspective, regardless of the vehicle type (sensor-agnostic planning), and subsequently translate this plan into physical vehicle control based on their familiarity with the vehicle's dimensions (sensor-specific lifting).

To ensure coherent multi-step reasoning within this 2D pipeline, we propose a knowledge-instilled policy learning strategy. Following ego-centric consistency preprocessing for multi-view input, we first conduct Supervised Fine-Tuning (SFT) to learn foundational driving knowledge. Subsequently, we employ Group Relative Policy Optimization (GRPO) for Reinforcement Learning (RL). Crucially, unlike previous methods [73] that only calculate sparse rewards on final predictions, regardless of the generated Chain-of-Thought (CoT), we assign dense and intermediate rewards to verifiable outputs from perception to planning while leaving free-form reasoning flexible. This structure encourages the final spatial planning to be tied to accurate visual scene understanding.

In summary, our contributions are as follows:

- We propose a decoupled planning and lifting paradigm that shifts VLA optimization entirely to the image space. It facilitates data scaling across heterogeneous driving datasets with available ego-motion and camera parameters, and prioritizes visual reasoning over trivial ego-status.
- We introduce a knowledge-instilled policy learning strategy. By employing dense intermediate rewards on verifiable outputs, we effectively mitigate the long-horizon credit assignment problem and encourage a rigorous causal chain from visual perception to final spatial planning.
- Extensive experiments demonstrate that PixelPilot achieves state-of-the-art performance in both open-loop and closed-loop settings, validating the superior scalability and visual reasoning capabilities of our decoupled paradigm.

2 Related Work

Vision-Language Models. The remarkable reasoning capabilities of Large Language Models (LLMs) [3, 17, 52, 66] have propelled the evolution of Vision-Language Models (VLMs) [11, 42, 74], which fuse visual and textual modalities to construct comprehensive and unified representations. Early milestones such as CLIP [42] demonstrated the efficacy of aligning image and text features from independent encoders, facilitating robust zero-shot generalization across image-text pairs. Subsequent architectures, notably LLaVA [36], advanced this paradigm by introducing visual instruction tuning via dedicated vision-language projectors, thereby endowing models with sophisticated multi-modal instruction-following capabilities. More recently, state-of-the-art models like the Qwen-VL series [2] and InternVL [9, 75] have adopted more powerful multi-modal architectures. These frameworks achieve unprecedented visual understanding and grounding, unlocking complex open-world applications ranging from object localization and segmentation [28, 29, 43] to end-to-end autonomous driving [5, 15, 16, 31, 39, 40, 46, 54, 73]. However, 2D-pretrained VLMs inherently struggle with 3D spatial understanding, particularly when tasked with predicting precise 3D metric trajectories. Moreover, their visual grounding capabilities are expressed through image-space entities like bounding boxes. Preserving this native representation during policy learning is therefore important for effectively exploiting VLMs across driving data collected with diverse sensors.

Autonomous Driving. Autonomous driving has transitioned from traditional modular pipelines, i.e., perception [21, 33, 57, 58, 65], motion prediction [14, 59], and planning [69], to end-to-end learning-based paradigms [7, 8, 18–20, 26, 34, 47, 49, 70]. UniAD [20] pioneered the integration of all sub-tasks into a cascaded framework, yielding substantial improvements over modular baselines. Subsequent works [8, 26, 70] adopt Bird's-Eye View representations and generate planning trajectories through multi-stage interaction modeling. Recently, researchers have increasingly leveraged the broad knowledge and visual reasoning of VLMs to enhance generalization in complex driving scenarios. To adapt the 2D-pretrained VLMs for 3D trajectory prediction, current methods typically incorporate explicit 3D visual encoders or curate 3D driving instructional data, optimizing VLMs in 3D space. For instance, OmniDrive [54], Orion [16], and OpenDriveVLA [72] replace the standard ViT [13] with a 3D Q-Former [32], explicitly utilizing camera parameters for 2D-to-3D feature projection. Conversely, other approaches [22, 27, 41, 44, 51, 55, 56, 62, 64, 67, 68, 71, 73] construct 3D question-answering pairs, fine-tuning VLMs to predict 3D trajectories directly from 2D visual inputs. Crucially, in both paradigms, the 2D-to-3D mapping remains strictly entangled with the specific camera parameters of the training datasets. This sensor dependence inherently restricts data scalability across heterogeneous driving corpora and induces convergence to trivial solutions.

To circumvent these limitations, we propose a *decoupled planning and lifting paradigm*. By performing semantic reasoning and motion planning as sensor-agnostic 2D-to-2D prediction tasks, our approach directly optimizes in the 2D image plane, facilitating scalable training and visual reasoning.

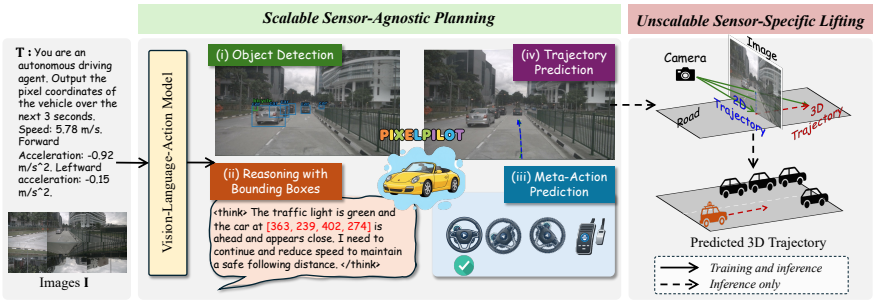


Fig. 2: Decoupled planning and lifting paradigm. The sensor-agnostic planning phase is scalable and introduces reasoning with bounding boxes to explore visual cues, while the sensor-specific lifting is unscalable and is performed only at inference.

3 Method

3.1 Decoupled Planning and Lifting Paradigm

Previous Vision-Language-Action Models (VLAs) are typically fine-tuned to predict 3D trajectories directly. While physically intuitive, this significantly restricts data scaling across heterogeneous driving datasets and degrades visual reasoning, leading to an over-reliance on trivial ego-status. To resolve these problems, we introduce PixelPilot, which features a *decoupled planning and lifting* paradigm. As shown in Fig. 2, it decouples the pipeline into two phases: (1) *planning*, where high-level reasoning and trajectory planning are predicted and optimized entirely in the image space for data scaling and visual reasoning; and (2) *lifting*, where the planned 2D trajectory is deterministically projected into the 3D world for vehicle control.

Theoretical Justification. The core premise of PixelPilot is that shifting trajectory prediction and optimization from 3D space back to 2D images improves data scaling and visual reasoning. Before detailing this paradigm, we mathematically justify the rationale of driving in 2D from the following two aspects:

(1) Trajectory Feasibility. For short-term trajectory prediction (e.g., 3 seconds into the future), the road surface immediately ahead, including global edge cases like slopes, can often be approximated as a local plane [30, 35] relative to the ego-car as it drives in an ego-centric manner. Our empirical evidence from nuScenes [4] indicates an average 3-second road height variation of merely 0.16 m, a negligible deviation compared to the 7.6 m depth prediction error typical of expert depth estimators like DepthAnything [63]. Consequently, under this local-plane approximation, the 3D road plane and its 2D image projection are related by a strict bijective projective transformation, as visualized in Fig. 3 (a). This bijective mapping provides a conditional geometric rationale: every physical coordinate on the 3D drivable road plane maps to a unique, determinable point on its 2D image projection, and vice versa. Therefore, a continuous and smooth trajectory planned on the 2D image corresponds to a physically continuous and smooth trajectory in the 3D metric space under the local-plane assumption.

(a) Trajectory Feasibility (b) Interaction Safety

Fig. 3: Justification for decoupled planning and lifting paradigm. (a) The bijective trajectory pairs demonstrate trajectory feasibility under a local-plane assumption. (b) Planning in image space provides a conservative cue for avoiding visible collision risks.

(2) Interaction Safety. The safety motivation of our 2D planning approach is justified through an analogy to Configuration Space (C-Space) planning [6,38] in classical robotics. By treating the 2D image plane as a simplified C-space, the 2D bounding boxes of dynamic agents function as Image-Space Obstacles (I-Space Obstacles). Crucially, these I-Space Obstacles represent a conservative superset of visible 3D collision risks. Under accurate visible-object perception and the local-plane assumption, any physical 3D collision implies a 2D projection overlap, so avoiding I-Space Obstacles provides a conservative safety cue, as demonstrated in Fig. 3 (b). This property additionally helps the model reason about spatial interaction from image evidence.

Sensor-Agnostic Planning in the Image Plane. Leveraging this theoretical foundation, the Planning phase fully exploits the native visual grounding capabilities of the VLM to perform high-level reasoning and 2D trajectory prediction. Let $\mathbf{I} = \{I_i \in \mathbb{R}^{H\times W\times 3}\}_{i=1}^{N_{\text{view}}}$ represent the input multi-view images and $\mathbf{C}$ denote the user command (including ego-status). We formulate the driving task as a conditional generation problem exclusively on the image plane $\mathcal{P}_{\text{img}}$. The VLM, denoted as Φ, processes the visual and textual inputs to sequentially generate detected objects $\hat{\mathcal{B}}_{\text{2D}}$, reasoning with bounding boxes $\hat{\mathcal{R}}$, meta-action $\hat{A}$, and future waypoints $\hat{\mathcal{T}}_{\text{2D}}$ in pixel coordinates in an auto-regressive manner:

$$\begin{aligned}(\hat{\mathcal{B}}_{\text{2D}}, \hat{\mathcal{R}}, \hat{A}, \hat{\mathcal{T}}_{\text{2D}} \mid \mathbf{I}, \mathbf{C}) \sim & \prod_{t=1}^{T} p_{\Phi}\big((\hat{u}_t, \hat{v}_t) \mid \mathbf{I}, \mathbf{C}, \hat{\mathcal{B}}_{\text{2D}}, \hat{\mathcal{R}}, \hat{A}, \hat{\mathcal{T}}_{\text{2D},<t}\big) \\ & \cdot p_{\Phi}(\hat{A} \mid \mathbf{I}, \mathbf{C}, \hat{\mathcal{B}}_{\text{2D}}, \hat{\mathcal{R}}) \\ & \cdot p_{\Phi}(\hat{\mathcal{R}} \mid \mathbf{I}, \mathbf{C}, \hat{\mathcal{B}}_{\text{2D}}) \\ & \cdot p_{\Phi}(\hat{\mathcal{B}}_{\text{2D}} \mid \mathbf{I}, \mathbf{C})\end{aligned} \tag{1}$$

where $(\hat{u}_t, \hat{v}_t)$ represents the planned image-space waypoint on $\mathcal{P}_{\text{img}}$ at future time step t. By formulating perception and trajectory prediction as 2D-to-2D prediction tasks, the planning phase is entirely *sensor-agnostic*. We can merge heterogeneous calibrated datasets (e.g., nuScenes, Waymo, and internal proprietary data) into a single unified training corpus for scalable training without requiring the VLA to learn dataset-specific 2D-to-3D mappings. Deterministic lifting still requires the target camera parameters at inference, so weakly calibrated or uncalibrated 3D planning remains outside the scope of this work.

Notably, unlike previous VLAs that either treat perception only as a proxy task for 2D-to-3D feature projection [16, 37, 54, 72] or generate coarse textual descriptions (which are prone to ambiguous object referencing) [10, 67, 73], we introduce *reasoning with bounding boxes*, where PixelPilot explicitly attends to specific Regions of Interest (RoI) by conditioning its semantic reasoning $\hat{\mathcal{R}}$ on the preceding perception outputs $\hat{\mathcal{B}}_{\text{2D}}$ (e.g., `<think> The car at [363, 239, 402, 274] is directly ahead and appears close, so I need to reduce speed </think>` in Fig. 2). In this way, we establish an interpretable causal chain: the model must explicitly *perceive* a visual anchor, *reason* about its spatial relationship, and then *act and plan*, which encourages predicted trajectories to be anchored to verifiable visual evidence rather than coarse language priors or trivial ego-status.

Deterministic Lifting to the 3D World. Once the 2D trajectory $\hat{\mathcal{T}}_{\text{2D}}$ is generated, the Lifting phase converts it into 3D trajectories. Since 3-second planning in autonomous driving typically occurs on a local plane [30, 35], we empirically set the height of this plane as $Z = -h$. Let K be the camera parameters transforming world coordinates to image coordinates. For planned 2D waypoints $\hat{\mathcal{T}}_{\text{2D}} = \{[\hat{u}_t, \hat{v}_t, 1]\}_{t=1}^{T}$ (in homogeneous coordinates), we seek their corresponding 3D points $\hat{\mathcal{T}}_{3D} = \{[\hat{x}_t, \hat{y}_t, -h]\}_{t=1}^{T}$ in the ego-vehicle coordinate via the mapping function Ψ, which is formulated as a homography transformation or geometric ray-casting:

$$\hat{\mathcal{T}}_{3D} = \Psi(\hat{\mathcal{T}}_{2D}, K, h) \tag{2}$$

Specifically, we cast a ray from the camera center through the pixel (u, v) and compute its intersection with the local plane where $z = -h$.

Decoupling 2D-to-3D lifting from optimization ensures that absolute ego-speed cannot be trivially mapped to pixel displacement without contextual depth, facilitating the exploitation of visual cues rather than ego-status. Notably, since the sensor-specific 2D-to-3D lifting occurs only during inference, the trained VLA can generalize to another calibrated vehicle by simply updating the camera parameters in the subsequent lifting phase. This formulation does not reject 3D priors; instead, such priors can be injected into the lifting stage while keeping the learnable policy in scalable visual space.

3.2 Knowledge-Instilled Policy Learning

To effectively implement the decoupled planning and lifting paradigm, we first preprocess the input multi-view images for ego-centric consistency. Then, as illustrated in Fig. 4, we utilize a two-stage learning strategy that first instills driving knowledge via supervised fine-tuning (SFT) and subsequently aligns the generated Chain-of-Thought with final trajectory prediction by applying dense rewards from perception to planning via Reinforcement Learning (RL).

Ego-Centric Consistency Preprocessing. To enable coherent 2D driving across multiple cameras, we address a critical flaw in prevailing VLA multi-view compositions. Typically, front and back camera inputs are arranged into two

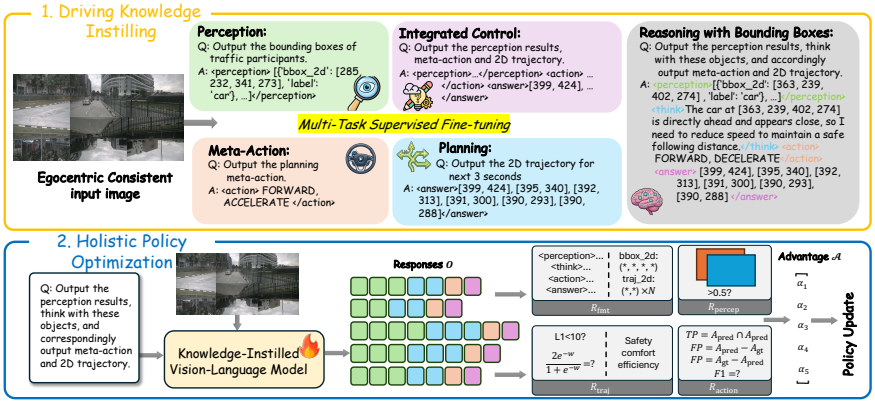
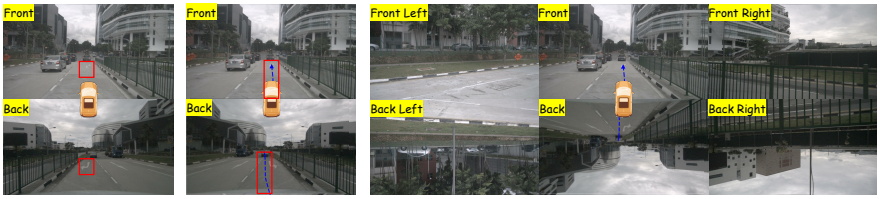


Fig. 4: Training pipeline of PixelPilot. We propose a knowledge-instilled policy learning strategy. The first stage instills foundational driving knowledge into the VLA by multi-task SFT. The second stage assigns rewards on the holistic VLA pipeline for coherence.



(a) Simplest Stitching (b) Back Views Horizontally Flipped Stitching (c) Back Views 180 Degree Rotation Stitching

Fig. 5: Different image preprocessing methods for VLAs.

separate rows, introducing severe *egocentric inconsistency*, such as lane marking misalignment when using the simplest stitching (Fig. 5 (a)). While using horizontally flipped stitching resolves the lane marking misalignment, it generates discontinuous trajectory planning (Fig. 5 (b)). To rectify this problem and construct a continuous 2D planning surface, we rotate the back view by 180 degrees. This spatial correction preserves egocentric consistency, generating a visually continuous road plane and a cohesive trajectory across the entire multi-camera image collage (Fig. 5 (c)).

Stage 1: Driving Knowledge Instilling via SFT. The initial stage utilizes Supervised Fine-Tuning (SFT) to instill foundational driving knowledge into the VLA. We train the model on a comprehensive multi-task dataset meticulously engineered to facilitate visual reasoning:

(1) Perception. Previous VLAs [16,37,54,72] use perception as a proxy task for 2D-to-3D feature projection, leading to limited data scaling and visual reasoning. On the contrary, we fine-tune the VLA to directly output the 2D bounding boxes in an autoregressive manner, providing spatial anchors for reasoning and planning. To generate 2D bounding box labels $\mathcal{B}_{\text{2D}}$, we follow common prac-

tice [50, 53] and project the ground-truth 3D bounding boxes onto the corresponding 2D camera image planes.

(2) Meta-Action Prediction. We cultivate an understanding of high-level semantic driving intentions using meta-actions A, which comprise a lateral component A_{lat} (TURN LEFT, TURN RIGHT, FORWARD) and a longitudinal component A_{lon} (ACCELERATE, DECELERATE, KEEP, STOP). This meta-action provides a prior for final trajectory prediction.

(3) Planning. We use 2D trajectory data to predict precise motion planning. The 2D trajectory annotation $\mathcal{T}_{\text{2D}}$ is obtained using the same projection procedure as perception. Notably, although the perception and trajectory labels are constructed through 3D-to-2D projection to leverage the existing 3D-labeled datasets, our PixelPilot does not necessitate 3D bounding box or trajectory collection in real-world deployment.

(4) Integrated Control. We employ three-step perception-action-planning sequences to foster multi-task learning and establish the causal link between observation and execution.

(5) Reasoning with Bounding Boxes. We train our PixelPilot with reasoning with bounding boxes to fully exploit the visual cues in images. To generate the reasoning data, we first evaluate the baseline SFT model on the training set to identify failure cases. For these cases, we employ the Qwen-VL-Max model and provide it with a structured prompt containing the input image $\mathbf{I}$, ground-truth action A, and the 2D bounding boxes $\mathcal{B}_{\text{2D}}$ of surrounding agents to synthesize a causal reasoning chain that logically justifies the ground-truth action based on the perceived bounding boxes. Then, we feed the original input combined with the generated reasoning back into the SFT model. A sample is strictly selected only if the inclusion of reasoning leads to a tangible improvement, i.e., correcting the predicted action and reducing the trajectory L2 error.

Through the SFT stage, the model possesses a strong behavioral prior and can perform basic driving tasks. Please refer to the Supplementary Material for detailed 2D labels and reasoning data generation.

Stage 2: Holistic Policy Optimization via RL. Building on the SFT-initialized model, this phase employs the Group Relative Policy Optimization (GRPO) [45] algorithm to incentivize the VLA's higher-level reasoning and decision-making. Crucially, unlike previous RL practices [67, 73] in autonomous driving that typically compute sparse rewards exclusively on the final planning output (e.g., trajectory deviation or collision penalties), our optimization strategy calculates dense, intermediate rewards on verifiable outputs across the autoregressive pipeline, including perception, meta-action, and trajectory. We do not assign semantic rewards to free-form reasoning with bounding boxes, preserving its flexibility and reducing shortcut learning. By evaluating the response quality from initial perception down to final planning, this holistic reward structure yields two fundamental advantages. First, it mitigates the severe credit assignment problem inherent in long-horizon VLM generation by providing dense supervisory signals at each logical step. Second, it enforces the causal chain established during SFT. By simultaneously optimizing intermediate visual percep-

tion and final spatial planning, we encourage the final driving intent to remain anchored in accurate visual scene understanding rather than trivial ego-status.
Format Reward R_{fmt}. It is designed to ensure a strict and hierarchical output structure. It consists of three parts: (1) A structural reward of 1.0 is granted if the generated responses are in the form: "<perception>bboxes here</perception> <think>thinking with bboxes here</think> <action>meta-action here </action> <answer>trajectory here</answer>", otherwise 0. (2) A perception format reward ensures that each bounding box is complete. (3) A trajectory format reward of 1.0 is allocated for trajectory completeness and consistency.
Perception Reward R_{percep}. We employ an Intersection over Union (IoU)-based reward. Specifically, we use the Hungarian algorithm to find the matched predicted and ground truth (GT) bounding boxes and compute the mean IoU between them as the final perception reward.
Meta-Action Reward R_{action}. We evaluate the accuracy of the predicted high-level meta-action using an F1-score-based reward, balancing precision and recall against the ground-truth action set.
Trajectory Reward R_{traj}. To prioritize adherence to the 2D intent, we assign an L1 reward ($R_{\text{traj-L1}} = 1.0$) if the absolute pixel distance between the predicted and ground-truth 2D waypoints is below 10 pixels. We also apply a sigmoid-scaled L2 reward: $R_{\text{traj-L2}} = \frac{2e^{-w}}{1+e^{-w}}$, where w is the L2 distance. Moreover, we employ the Predictive Driver Model Score (PDMS) [12] R_{PDMS} to assess the safety, comfortness, and efficiency of closed-loop predicted trajectories.

The final accumulated reward is the weighted sum of these terms, ensuring coherence among visual perception, logical intent, and spatial planning via dense rewards.

$$R_{\text{acc}} = \lambda_{\text{fmt}} \cdot R_{\text{fmt}} + \lambda_{\text{percep}} \cdot R_{\text{percep}} + \lambda_{\text{action}} \cdot R_{\text{action}} + \lambda_{\text{traj}} \cdot R_{\text{traj}} \,. \tag{3}$$

4 Experiments

4.1 Implementation and Metrics

Datasets. We evaluate the perception, action prediction, and open-loop performance on nuScenes [4]. For closed-loop evaluation, we use the Bench2Drive benchmark [25] in the CARLA simulator. Besides, we incorporate the Waymo Open dataset [48] to demonstrate the data scalability of our PixelPilot.
Implementation Details. We use the powerful open-source VLM Qwen2.5-VL [2] as the base model for PixelPilot. For SFT, i.e., stage 1, we fine-tune the model for 2 epochs in a multi-task manner to instill the model with knowledge about perception, reasoning, high-level meta-action making, and trajectory prediction. During Stage 2, reinforcement learning, we fine-tune the model trained by SFT using GRPO [45] to incentivize the reasoning ability of the trained policy. The number of completions in GRPO is set to 8. All the experiments were conducted on eight GPUs with 80 GB memory each. For the main open-loop experiments, PixelPilot is trained on nuScenes (28.1k samples) and Waymo (23.8k samples) with SFT and RL. For Bench2Drive closed-loop evaluation, following

Table 1: 2D Object detection on nuScenes.

Method	mAP ↑
StreamPETR [53]	46.5
MV2D [60]	52.3
DeformableDETR [76]	50.2
SimPB [50]	54.1
PixelPilot	**54.2**

Table 2: High-level meta-action prediction on nuScenes. † indicates trained on nuScenes.

Method	Lateral (F1) ↑			Longitudinal (F1) ↑			
	forward	left	right	keep	acc.	dec.	stop
Qwen2.5VL-7B	64.67	24.15	30.85	40.73	55.14	51.41	41.82
Qwen2.5VL-7B†	94.46	63.00	67.01	57.62	74.35	77.10	75.00
PixelPilot	**96.82**	**75.51**	**75.71**	**61.23**	**81.76**	**80.19**	**81.80**

Orion [16], we train on Bench2Drive (274.5k samples) with SFT only, plan a 3-second trajectory at 2 Hz from current multi-view images, and convert the predicted trajectory to vehicle control with a PID controller. During decoding, we set temperature to 1.0, top-p to 0.5, and top-k to 20.

Evaluation Metrics. For perception, we report the mAP for 2D object detection. For meta-action, we use the F1-score for all lateral and longitudinal meta-action classes. For open-loop planning, we employ the L2 error (in meters) between the predicted and ground-truth trajectories. Additionally, following [34], the collision rate and intersection rate are adopted to evaluate the safety of the trajectory. Since open-loop safety metrics can be affected by non-reactive followers and off-road intersections, we jointly report L2, collision, and intersection and further validate safety in the closed-loop benchmark. For closed-loop performance, we report the success rate, driving score, efficiency, and comfortness.

4.2 Main Results

Perception. We evaluate the perception capability of PixelPilot in Tab. 1. Notably, our VLA model achieves state-of-the-art performance, even compared to specialized 2D object detectors trained on nuScenes [50, 53, 60, 76], despite not being optimized only for 2D object detection.

Meta-Action. The performance of PixelPilot on meta-action prediction is detailed in Table 2. Compared to the base Qwen2.5VL-7B, PixelPilot demonstrates superior performance across all lateral (Path) and longitudinal (Speed) action categories when evaluated by F1-score, underscoring the effectiveness of our proposed training methodology.

Open-Loop Planning. In the open-loop trajectory prediction, as detailed in Table 3, PixelPilot achieves a state-of-the-art average L2 error of 0.30 m. Compared with AutoVLA [73], PixelPilot reduces the average L2 error from 0.40 m to 0.30 m, corresponding to a 25% relative reduction. The safety and feasibility of the planned trajectories are evaluated by the collision rate, where PixelPilot consistently achieves competitive performance. For Intersection, which measures trajectory feasibility with respect to the drivable area, PixelPilot again shows strong results. It obtains the best average rate (1.77%) among its VLA-based peers. This comprehensive performance in safety-critical metrics validates the effectiveness of our PixelPilot.

Closed-Loop Planning. We conduct the closed-loop test on Bench2Drive [25] in CARLA simulator at 2 Hz. As shown in Tab. 4, PixelPilot achieves the best

Table 3: Open-loop trajectory planning of VLAs on nuScenes. As FSDrive [68] uses private data for training and AutoDrive-R^2 [67] explicitly calculates the 3D trajectory from ego status via kinematics, we mark them as gray.

Method	L2 (m) ↓				Collision (%) ↓				Intersection (%) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.
DriveVLM [51]	0.18	0.34	0.68	0.40	0.10	0.22	0.45	0.27	-	-	-	-
EMMA [23]	0.14	0.29	0.54	0.32	-	-	-	-	-	-	-	-
OmniDrive [54]	0.14	0.29	0.55	0.33	**0.00**	0.13	0.78	0.30	0.56	2.48	5.96	3.00
OpenDriveVLA [72]	0.15	0.31	0.55	0.33	0.01	**0.08**	**0.21**	**0.10**	-	-	-	-
Impprompt-VLA [10]	**0.13**	0.27	0.53	**0.30**	-	-	-	-	-	-	-	-
FSDrive [68]	0.14	0.25	0.46	0.28	0.07	0.12	1.02	0.40	-	-	-	-
AutoVLA [73]	0.21	0.38	0.60	0.40	0.13	0.18	0.28	0.20	-	-	-	-
AutoDrive-R^2 [67]	0.11	0.19	0.29	0.20	-	-	-	-	-	-	-	-
DrivePI [37]	0.19	0.36	0.64	0.40	0.00	0.05	0.28	0.11	-	-	-	-
PixelPilot	**0.13**	**0.26**	**0.51**	**0.30**	**0.00**	0.11	0.65	0.25	**0.51**	**1.53**	**3.26**	**1.77**

Table 4: Closed-loop planning on Bench2Drive (CARLA).

Method	Driving Score ↑	Success Rate (%) ↑	Efficiency ↑	Comfortness ↑
AD-MLP [69]	18.05	0.00	48.45	22.63
UniAD-Base [20]	45.81	16.36	129.21	43.58
VAD [26]	42.35	15.00	**157.94**	**46.01**
TCP-traj [61]	59.90	30.00	76.54	18.08
DriveAdapter [24]	64.22	33.08	70.22	16.01
Orion [16]	77.74	54.62	151.48	17.38
AutoVLA [73]	78.84	57.73	146.93	39.33
PixelPilot	**79.14**	**58.87**	153.26	38.01

overall driving score and success rate, validating our decoupled planning and lifting paradigm. Among VLA-based methods, PixelPilot also achieves the best efficiency and the second-best comfortness. Compared with the traditional planner VAD [26], PixelPilot remains competitive in efficiency and comfortness while substantially improving the core driving safety metrics by 36.79 driving score and 43.87 success rate.

4.3 Visualization

Fig. 6 presents a qualitative comparison of our method against other approaches on the nuScenes dataset. Notably, Qwen2.5-VL-7B struggles to generate accurate predictions, exhibiting significant trajectory deviations, particularly in turning scenarios. Although OmniDrive delivers satisfactory performance, its trajectories are often overly aggressive, with predicted speeds substantially exceeding ground truth speeds. In contrast, our method, PixelPilot, consistently generates reliable plans that closely align with the ground truth.

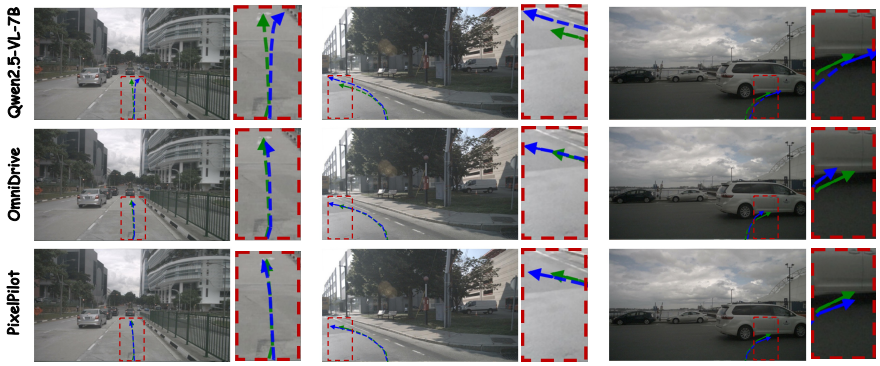


Fig. 6: Visualization of 2D trajectories across Qwen2.5-VL-7B, OmniDrive, and our PixelPilot on the nuScenes validation dataset. The predicted and ground-truth trajectories are depicted in blue and green, respectively.

Table 5: Ablation on learning strategy.

Method	Avg. L2 ↓
Qwen2.5-VL-7B	1.92
Qwen2.5-VL-7B + SFT	0.46
Qwen2.5-VL-7B + RL	0.49

Table 6: Ablation on preprocessing and SFT.

Method	Avg. L2 ↓
Front View	0.40
w/o. Ego. Cons.	0.43
w/o. Multi-Task	0.45

Table 7: Ablation on RL.

Method	Avg. L2 ↓
w/o. R_{traj}	0.44
w/o. R_{percep}	0.41
w/o. R_{action}	0.40

4.4 Ablation Study

Two Stage Training Pipeline. To validate our two-stage knowledge-instilled learning strategy, we conducted an ablation study (Table 5) by training variants with only Supervised Fine-Tuning (SFT) or Reinforcement Learning (RL). We find that while both methods provide performance gains, the SFT-only variant surpasses the RL-only model. This suggests that RL, on its own, is inefficient at navigating the vast search space of our task's structured multi-step reasoning process (perception → reasoning → meta-action → planning). SFT is therefore essential for instilling the model with a coherent policy and a foundational understanding of the causal chain required. The superior performance of the complete PixelPilot model, which combines both stages, confirms that our hybrid approach is critical: SFT provides the necessary knowledge foundation, which RL then effectively refines to achieve coherent optimal results.

Preprocessing and SFT. To ensure the model's correct understanding and reasoning, the training images are concatenated in an ego-centrically consistent way. Then, we SFT Qwen2.5-VL-7B on a multi-task dataset, including 2D object detection, reasoning with bounding boxes, meta-action prediction, and trajectory prediction. Table 6 shows that removing the ego-centric consistent input ('w/o. Ego. Cons.', i.e., simplest stitching) leads to a higher L2 error compared to only using the front view, demonstrating the effectiveness of preprocessing. Replacing the multi-task training with simple trajectory alignment ('w/o. Multi-

Table 8: Trajectory prediction on Waymo.

Method	L2 Error ↓			
	1s	2s	3s	Avg.
Qwen2.5-VL-7B [2]	1.66	1.82	2.92	2.13
DriveVLM [51]	0.17	0.34	0.75	0.42
EMMA [23]	0.12	0.28	0.61	0.34
EMMA+ [23]	0.11	0.25	0.53	0.30
PixelPilot	0.16	0.32	0.71	0.40

Table 9: Ablation study on ego-status and input images.

Method	Ego-Status	Images	Avg. L2↓
OmniDrive [54]	✗	✓	1.98
	✓	✗	0.36
Imprompt-VLA [10]	✗	✓	2.39
	✓	✗	0.36
PixelPilot	✗	✓	0.71
	✓	✗	0.90

Task') increases the average L2 error to 0.45 m, validating the superiority of the perception-to-planning CoT learning.

Reinforcement Learning. We dissect the contributions of each component within our composite reward function used during the Reinforcement Learning (RL) stage. As shown in Table 7, individually ablating trajectory reward R_{traj}, IoU-based perception reward R_{percep}, and action reward R_{action} each leads to a discernible increase in the average L2 error. This confirms that these verifiable rewards all positively contribute to coherent CoT and final planning accuracy while avoiding over-constraint on free-form reasoning with bounding boxes.

Scalability and Generalization. When excluding the Waymo dataset from training, we observe a 0.06 increase in average L2 error, demonstrating the scalability brought by our decoupled planning and lifting paradigm. To empirically validate the generalization of PixelPilot across different camera configurations, we extended our trajectory prediction evaluation to the Waymo Open Dataset [48], as reported in Tab. 8. Remarkably, despite the significant domain shift between the training source (nuScenes) and the target domain (Waymo), our method achieves competitive performance with an average L2 error of 0.4m in a zero-shot setting, i.e., without any dataset-specific fine-tuning. Additional controlled ablations on planning targets and 3D-prior injection are provided in the Supplementary Material.

Ego-Status vs Input Images. Table 9 investigates whether our PixelPilot relies on visual cues or trivial ego-status. Under single-modality ablations, OmniDrive [54] and Imprompt-VLA [10] obtain substantially lower errors with ego-status only than with images only, whereas PixelPilot exhibits the opposite trend, revealing that the direct 2D-to-3D prediction paradigm predominantly relies on ego-status regardless of the input images. By decoupling the planning and lifting phases, PixelPilot achieves a lower L2 error of 0.71 when utilizing only visual inputs (ego-status masked), while yielding a higher error of 0.90 when using only ego-status. This validates that our PixelPilot successfully prevents the network from merely extrapolating trajectories via trivial kinematics and grounds its planning in robust visual understanding.

Local Plane Assumption. Tab. 10 shows that our PixelPilot achieves similar performance to the 2D-to-3D VLA [10] on samples with height changes above 0.5, 1, and 2 meters in future 3s. Together with the small average 3-second road-height variation in nuScenes, this supports the local-plane approximation for

Table 10: Ablation on local plane.

Method	Avg. L2 ↓		
	0.5m	1m	2m
Imprompt-VLA [10]	0.30	0.32	0.33
PixelPilot	0.30	0.31	0.34

Table 11: Trade-off between CoT and latency.

Method	Avg. L2 ↓	Latency (s)
w/o CoT	0.41	0.23
Coarse CoT	0.35	0.31
PixelPilot	0.30	0.33

short-horizon ego-centric planning, while multi-level roads and severe occlusions remain challenging cases.

Trade-off between CoT and Latency. Table 11 shows that PixelPilot, featuring a CoT from perception to planning, reduces Avg. L2 from 0.35 m to 0.30 m over coarse CoT with only a marginal latency increase from 0.31 s to 0.33 s. This corresponds to about 3 Hz inference, satisfying the 2 Hz Bench2Drive control frequency, and the deterministic lifting stage adds negligible overhead.

5 Conclusion

In this paper, we introduced PixelPilot, a vision-language-action model for end-to-end autonomous driving that decouples sensor-agnostic 2D planning from sensor-specific 3D lifting. By formulating perception, visual reasoning, and trajectory prediction as image-space tasks, PixelPilot enables scalable learning from heterogeneous calibrated driving datasets while reducing the tendency of direct 3D optimization to rely on trivial ego-status cues. To realize this paradigm, we propose a two-stage knowledge-instilled policy learning strategy that grounds reasoning in explicit 2D bounding boxes and uses dense rewards on verifiable intermediate outputs to encourage a coherent causal chain from perception to planning. Extensive experiments show that PixelPilot achieves strong open-loop and closed-loop performance, supporting 2D-first planning with deterministic lifting as a scalable and interpretable formulation for driving VLAs. This formulation also clarifies the current scope of the approach. While deterministic lifting avoids learning dataset-specific 2D-to-3D mappings, it still relies on reliable camera geometry and ego-motion estimates at inference time. Moreover, our geometric formulation is tailored to short-horizon ego-centric driving under a local-plane approximation; strongly non-planar multi-level roads, severe occlusions, and weakly calibrated sensor setups remain challenging. Extending the framework to handle these conditions more robustly, while preserving its sensor-agnostic learning advantages, is an important direction for future work.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62322113, 62376156), as well as the Shanghai Municipal Special Program for Basic Research on General AI Foundation Models (Grant No. 2025SHZDZX025G15).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NeurIPS pp. 1877–1901 (2020)
4. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR. pp. 11618–11628 (2020)
5. Cai, T., Liu, Y., Zhou, Z., Ma, H., Zhao, S.Z., Wu, Z., Ma, J.: Driving with regulation: Trustworthy and interpretable decision-making for autonomous driving with retrieval-augmented reasoning. In: AAAI. pp. 38287–38295 (2026)
6. Chen, C., Seff, A., Kornhauser, A., Xiao, J.: Deepdriving: Learning affordance for direct perception in autonomous driving. In: ICCV. pp. 2722–2730 (2015)
7. Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., Li, H.: End-to-end autonomous driving: Challenges and frontiers. IEEE TPAMI pp. 10164–10183 (2024)
8. Chen, S., Jiang, B., Gao, H., Liao, B., Xu, Q., Zhang, Q., Huang, C., Liu, W., Wang, X.: Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. arXiv preprint arXiv:2402.13243 (2024)
9. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: CVPR. pp. 24185–24198 (2024)
10. Chi, H., Gao, H.a., Liu, Z., Liu, J., Liu, C., Li, J., Yang, K., Yu, Y., Wang, Z., Li, W., et al.: Impromptu vla: Open weights and open data for driving vision-language-action models. NeurIPS (2026)
11. Chu, X., Qiao, L., Lin, X., Xu, S., Yang, Y., Hu, Y., Wei, F., Zhang, X., Zhang, B., Wei, X., et al.: Mobilevlm: A fast, strong and open vision language assistant for mobile devices. arXiv preprint arXiv:2312.16886 (2023)
12. Dauner, D., Hallgarten, M., Li, T., Weng, X., Huang, Z., Yang, Z., Li, H., Gilitschenski, I., Ivanovic, B., Pavone, M., et al.: Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. NeurIPS pp. 28706–28719 (2024)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
14. Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C.R., Zhou, Y., et al.: Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In: ICCV. pp. 9690–9699 (2021)
15. Fu, D., Li, X., Wen, L., Dou, M., Cai, P., Shi, B., Qiao, Y.: Drive like a human: Rethinking autonomous driving with large language models. In: WACVW. pp. 910–919 (2024)
16. Fu, H., Zhang, D., Zhao, Z., Cui, J., Liang, D., Zhang, C., Zhang, D., Xie, H., Wang, B., Bai, X.: Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. In: ICCV. pp. 24823–24834 (2025)

17. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. Nature **645**(8081), 633–638 (2025)
18. Han, W., Guo, D., Xu, C.Z., Shen, J.: Dme-driver: Integrating human decision logic and 3d scene perception in autonomous driving. In: AAAI. pp. 3347–3355 (2025)
19. Hu, S., Chen, L., Wu, P., Li, H., Yan, J., Tao, D.: St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In: ECCV. pp. 533–549 (2022)
20. 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., Li, H.: Planning-oriented autonomous driving. In: CVPR. pp. 17853–17862 (2023)
21. Huang, J., Huang, G., Zhu, Z., Du, D.: Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790 (2021)
22. Huang, Z., Fen, C., Yan, F., Xiao, B., Jie, Z., Zhong, Y., Liang, X., Ma, L.: Drivemm: All-in-one large multimodal model for autonomous driving. arXiv preprint arXiv:2412.07689 (2024)
23. Hwang, J.J., Xu, R., Lin, H., Hung, W.C., Ji, J., Choi, K., Huang, D., He, T., Covington, P., Sapp, B., et al.: Emma: End-to-end multimodal model for autonomous driving. arXiv preprint arXiv:2410.23262 (2024)
24. Jia, X., Gao, Y., Chen, L., Yan, J., Liu, P.L., Li, H.: Driveadapter: Breaking the coupling barrier of perception and planning in end-to-end autonomous driving. In: ICCV. pp. 7919–7929 (2023)
25. Jia, X., Yang, Z., Li, Q., Zhang, Z., Yan, J.: Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. NeurIPS pp. 819–844 (2024)
26. Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., Zhang, Q., Liu, W., Huang, C., Wang, X.: Vad: Vectorized scene representation for efficient autonomous driving. In: ICCV. pp. 8306–8316 (2023)
27. Jiang, B., Chen, S., Zhang, Q., Liu, W., Wang, X.: Alphadrive: Unleashing the power of vlms in autonomous driving via reinforcement learning and reasoning. arXiv preprint arXiv:2503.07608 (2025)
28. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: CVPR. pp. 9579–9589 (2024)
29. Lan, M., Chen, C., Zhou, Y., Xu, J., Ke, Y., Wang, X., Feng, L., Zhang, W.: Text4seg: Reimagining image segmentation as text generation. In: ICLR (2025)
30. Lee, S., Lim, H., Myung, H.: Patchwork++: Fast and robust ground segmentation solving partial under-segmentation using 3d point cloud. In: IROS. pp. 13276–13283 (2022)
31. Li, B., Wang, Y., Mao, J., Ivanovic, B., Veer, S., Leung, K., Pavone, M.: Driving everywhere with large language model policy adaptation. In: CVPR. pp. 14948–14957 (2024)
32. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML. pp. 19730–19742 (2023)
33. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: ECCV. pp. 1–18 (2022)
34. Li, Z., Yu, Z., Lan, S., Li, J., Kautz, J., Lu, T., Alvarez, J.M.: Is ego status all you need for open-loop end-to-end autonomous driving? In: CVPR. pp. 14864–14873 (June 2024)

35. Lim, H., Oh, M., Myung, H.: Patchwork: Concentric zone-based region-wise ground segmentation with ground likelihood estimation using a 3d lidar sensor. IEEE RA-L pp. 6458–6465 (2021)
36. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS. pp. 34892–34916 (2023)
37. Liu, Z., Huang, R., Yang, R., Yan, S., Wang, Z., Hou, L., Lin, D., Bai, X., Zhao, H.: Drivepi: Spatial-aware 4d mllm for unified autonomous driving understanding, perception, prediction and planning. arXiv preprint arXiv:2512.12799 (2025)
38. Lozano-Perez, T.: Spatial planning: A configuration space approach. IEEE TC pp. 108–120 (1983)
39. Ma, Y., Cao, Y., Sun, J., Pavone, M., Xiao, C.: Dolphins: Multimodal language model for driving. In: ECCV. pp. 403–420 (2024)
40. Marcu, A.M., Chen, L., Hünermann, J., Karnsund, A., Hanotte, B., Chidananda, P., Nair, S., Badrinarayanan, V., Kendall, A., Shotton, J., et al.: Lingoqa: Visual question answering for autonomous driving. In: ECCV. pp. 252–269 (2024)
41. Qian, K., Jiang, S., Zhong, Y., Luo, Z., Huang, Z., Zhu, T., Jiang, K., Yang, M., Fu, Z., Miao, J., et al.: Agentthink: A unified framework for tool-augmented chain-of-thought reasoning in vision-language models for autonomous driving. In: Findings of EMNLP. pp. 10663–10682 (2025)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
43. Ren, X., Wang, Z., Hou, L., Tang, P., Wang, G., Ma, C.: Grounding everything in tokens for multimodal large language models. arXiv preprint arXiv:2512.10554 (2025)
44. Shao, H., Hu, Y., Wang, L., Song, G., Waslander, S.L., Liu, Y., Li, H.: Lmdrive: Closed-loop end-to-end driving with large language models. In: CVPR. pp. 15120–15130 (2024)
45. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
46. Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., Li, H.: Drivelm: Driving with graph visual question answering. In: ECCV. pp. 256–274 (2024)
47. Song, R., Guo, X., Wu, H., Wei, Q., Chen, L.: Insightdrive: Insight scene representation for end-to-end autonomous driving. arXiv preprint arXiv:2503.13047 (2025)
48. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR. pp. 2443–2451 (2020)
49. Sun, W., Lin, X., Shi, Y., Zhang, C., Wu, H., Zheng, S.: Sparsedrive: End-to-end autonomous driving via sparse scene representation. In: ICRA. pp. 8795–8801 (2025)
50. Tang, Y., Meng, Z., Chen, G., Cheng, E.: Simpb: A single model for 2d and 3d object detection from multiple cameras. In: ECCV. pp. 1–17 (2024)
51. Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., Zhao, H.: Drivevlm: The convergence of autonomous driving and large vision-language models. arXiv preprint arXiv:2402.12289 (2024)
52. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

53. Wang, S., Liu, Y., Wang, T., Li, Y., Zhang, X.: Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In: ICCV. pp. 3598–3608 (2023)
54. Wang, S., Yu, Z., Jiang, X., Lan, S., Shi, M., Chang, N., Kautz, J., Li, Y., Alvarez, J.M.: Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In: CVPR. pp. 22442–22452 (2025)
55. Wang, T., Xie, E., Chu, R., Li, Z., Luo, P.: Drivecot: Integrating chain-of-thought reasoning with end-to-end driving. arXiv preprint arXiv:2403.16996 (2024)
56. Wang, W., Xie, J., Hu, C., Zou, H., Fan, J., Tong, W., Wen, Y., Wu, S., Deng, H., Li, Z., et al.: Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. Visual Intelligence **3**(1) (2025)
57. Wang, X., Zhu, Z., Xu, W., Zhang, Y., Wei, Y., Chi, X., Ye, Y., Du, D., Lu, J., Wang, X.: Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In: ICCV. pp. 17804–17813 (2023)
58. Wang, Y., Guizilini, V.C., Zhang, T., Wang, Y., Zhao, H., Solomon, J.: Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In: CoRL. pp. 180–191 (2022)
59. Wang, Y., He, J., Fan, L., Li, H., Chen, Y., Zhang, Z.: Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In: CVPR. pp. 14749–14759 (2024)
60. Wang, Z., Huang, Z., Fu, J., Wang, N., Liu, S.: Object as query: Lifting any 2d object detector to 3d detection. In: ICCV. pp. 3768–3777 (2023)
61. Wu, P., Jia, X., Chen, L., Yan, J., Li, H., Qiao, Y.: Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. NeurIPS pp. 6119–6132 (2022)
62. Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.Y.K., Li, Z., Zhao, H.: Drivegpt4: Interpretable end-to-end autonomous driving via large language model. IEEE RA-L pp. 8186–8193 (2024)
63. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR. pp. 10371–10381 (2024)
64. Yang, Z., Chai, Y., Jia, X., Li, Q., Shao, Y., Zhu, X., Su, H., Yan, J.: Drivemoe: Mixture-of-experts for vision-language-action model in end-to-end autonomous driving. arXiv preprint arXiv:2505.16278 (2025)
65. Ye, T., Jing, W., Hu, C., Huang, S., Gao, L., Li, F., Wang, J., Guo, K., Xiao, W., Mao, W., et al.: Multi-modality fusion for prediction and planning tasks of autonomous driving. arXiv preprint arXiv:2308.01006 (2023)
66. Yenduri, G., Srivastava, G., Maddikunta, P.K.R., Jhaveri, R.H., Wang, W., Vasilakos, A.V., Gadekallu, T.R., et al.: Gpt (generative pre-trained transformer): A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. IEEE Access **12**, 54608–54649 (2024)
67. Yuan, Z., Tang, J., Luo, J., Chen, R., Qian, C., Sun, L., Chu, X., Cai, Y., Zhang, D., Li, S.: Autodrive-r^2: Incentivizing reasoning and self-reflection capacity for vla model in autonomous driving. arXiv preprint arXiv:2509.01944 (2025)
68. Zeng, S., Chang, X., Xie, M., Liu, X., Bai, Y., Pan, Z., Xu, M., Wei, X.: Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. arXiv preprint arXiv:2505.17685 (2025)
69. Zhai, J.T., Feng, Z., Du, J., Mao, Y., Liu, J.J., Tan, Z., Zhang, Y., Ye, X., Wang, J.: Rethinking the open-loop evaluation of end-to-end autonomous driving in nuscenes. arXiv preprint arXiv:2305.10430 (2023)

70. Zhang, B., Song, N., Jin, X., Zhang, L.: Bridging past and future: End-to-end autonomous driving with historical prediction and planning. In: CVPR. pp. 6854–6863 (2025)
71. Zheng, W., Mao, X., Ye, N., Li, P., Zhan, K., Lang, X., Zhao, H.: Driveagent-r1: Advancing vlm-based autonomous driving with hybrid thinking and active perception. arXiv preprint arXiv:2507.20879 (2025)
72. Zhou, X., Han, X., Yang, F., Ma, Y., Knoll, A.C.: Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. In: AAAI. pp. 13782–13790 (2026)
73. Zhou, Z., Cai, T., Zhao, S.Z., Zhang, Y., Huang, Z., Zhou, B., Ma, J.: Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. In: NeurIPS (2025)
74. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: ICLR (2024)
75. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)
76. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. In: ICLR (2021)