Driving with Advice: Large Model as Motion Advisor for Joint Planning

Junyin Wang[1*], Jinlei Yu[2*], Hao Lin[3], Huikai Liu[1, 3], Wenqian Zhu[3], Shengwu Xiong[1,4†]

[1]School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan, 430070
[2]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, 430074
[3]VOYAH Automobile Technology Co., Ltd., Wuhan 430051, China
[4]Interdisciplinary Artificial Intelligence Research Institute, Wuhan College, Wuhan 430212, China
{wjy199708, xiongsw}@whut.edu.cn, m202373718@hust.edu.cn, {h-linhao, liuhk, h-zhuwq}@voyah.com.cn

Abstract

We address the challenge of integrating high-level semantic reasoning with low-level trajectory planning in end-to-end autonomous driving, where most existing frameworks decouple perception, decision-making, and control, leading to limited interpretability and poor instruction compliance. To bridge this gap, we propose Driving with Advice, a novel closed-loop framework that treats a vision-language model (VLM) as a motion advisor to provide interpretable, language-mediated guidance for trajectory generation. Our approach introduces three key innovations: (1) Semantic-Intentional Pretraining (SIP), which injects driving rationale into a compact VLM via machine-generated question-answering pairs; (2) a discrete action space grounded in directional and speed primitives, enabling structured and interpretable policy learning; and (3) an advice-following diffusion policy refined via Group Relative Policy Optimization under a multi-objective reward that ensures safety, comfort, and alignment with semantic intent. We evaluate our method on the NAVSIM benchmark in a closed-loop setting, achieving a state-of-the-art Predictive Driver Model Score (PDMS) of 91.5, outperforming strong baselines in safety (NC: 99.2). The results demonstrate that leveraging language as a cognitive interface between perception and control enhances both generalization and behavioral transparency, advancing the paradigm of language-conditioned driving.

Introduction

End-to-end autonomous driving(Hu et al. 2023; Jiang et al. 2023a; Chen et al. 2024; Huang et al. 2024) has evolved from modular pipelines to integrated perception-plannin(Jiang et al. 2025a; Li et al. 2024c; Weng et al. 2024) systems, increasingly leveraging vision-language models (VLMs)(Liu et al. 2023; Dubey et al. 2024; Bai et al. 2023; Shao et al. 2024b) for enhanced scene understanding. However, despite advances in multimodal reasoning, most frameworks fail to exploit the full cognitive potential of large models, treating them as auxiliary perception modules rather than active participants in decision-making. This limitation results in policies that imitate control signals without understanding driving intent—leading to poor interpretability, limited generalization, and weak instruction compliance.

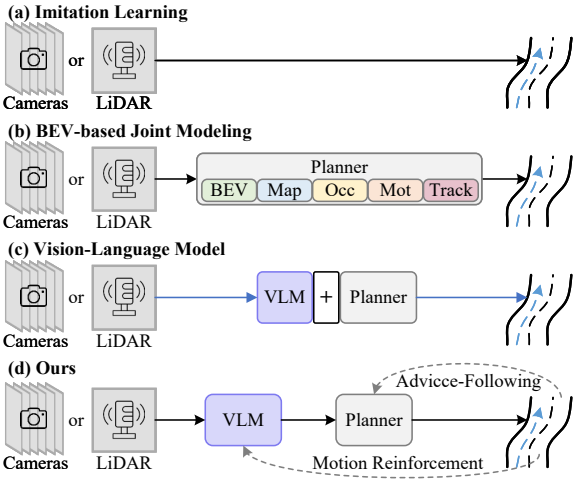

Figure 1: Comparison of autonomous driving frameworks: (a) Imitation Learning directly maps raw sensor inputs to control outputs; (b) BEV-based Joint Modeling uses structured representations for spatial reasoning; (c) Vision-Language Model generates natural language descriptions or QA responses but decouples linguistic cognition from trajectory execution; and (d) Our approach incorporates a vision-language model as a motion advisor, enabling safe decision-making and facilitating trajectory generation that follows the generated advice.

Existing approaches can be broadly categorized into three paradigms, as illustrated in Fig. 1. First, imitation learning methods (Fig. 1(a)) directly regress control commands from raw inputs, bypassing semantic abstraction and producing black-box behaviors. Second, BEV-based joint modeling (Fig. 1(b)) improves spatial coherence through structured Bird'S-Eye-View (BEV) representations but operates in continuous action spaces that obscure high-level intent. Third, recent VLM-driven systems (Fig. 1(c)) generate natural language descriptions or QA responses but decouple linguistic cognition from trajectory execution—resulting in a critical disconnect between **understanding** and **acting**. These frameworks fail to establish a closed-loop mecha-

*Equal Contribution
†Corresponding Author

nism where language informs action in a structured and executable manner.

The root cause lies in the underutilization of large models as behavioral advisors. While models like InternVL3-78B(Zhu et al. 2025) exhibit strong reasoning about traffic dynamics, their outputs are typically reduced to captions or attention maps, rather than being formalized into actionable motion suggestions. This represents a missed opportunity: if a model can reason that "the ego vehicle should slow down due to a pedestrian near the crosswalk," why not use this semantic intent as a prior for planning? Prior work(Wang et al. 2025; Mandalika, Nambiar et al. 2025; Shao et al. 2024a; Xu et al. 2024; Liao et al. 2025b) lacks a principled mechanism to transform such linguistic insights into structured, executable decisions.

We address this gap with a new paradigm: **treating the large vision-language model as a motion advisor**—a cognitive module that generates high-level driving suggestions which are then jointly refined with a diffusion-based planner. **Our key insight is that language should not be a byproduct of perception, but a guiding signal for action**. By extracting motion-relevant semantics from VLM outputs and grounding them in a discrete action space, we enable interpretable, instruction-aware policy learning.

To realize this vision, we propose a three-stage framework. First, we perform **Semantic-Intentional Pretraining (SIP)**, where InternVL3-78B generates QA-style labels over NAVSIM(Dauner et al. 2024), and Qwen2.5-VL-3B(Hui et al. 2024) is fine-tuned to learn both scene understanding and behavioral suggestions. Second, we design a **Driving Advice Generation** decomposed into Direction (e.g., turn left, lane change) and Speed (e.g., accelerate, stop) primitives, and refine the policy via GRPO(Shao et al. 2024b) to align with expert trajectories(Li et al. 2024a). Third, we train a diffusion-based(Huang et al. 2024; Jiang et al. 2025a) Ref Model for trajectory generation and optimize a Policy Model under a multi-objective reward—encompassing motion alignment, BEV collision avoidance, and smoothness—using Group Relative Policy Optimization(GRPO) for **Advice-Following** execution. As shown in Fig. 1(d), our method establishes a closed-loop advisor-planner architecture, where the VLM's advice directly shapes the planning process.

Our key contributions are fourfold: (1) We propose treating a vision-language model as a motion advisor, establishing a closed-loop cognitive interface where high-level semantic reasoning directly guides low-level trajectory planning in end-to-end autonomous driving. (2) We introduce Semantic-Intentional Pretraining (SIP) and a discrete action space grounded in directional and speed primitives, enabling interpretable, instruction-aware policy learning by bridging linguistic cognition and executable control. (3) We present an advice-following diffusion policy refined via Group Relative Policy Optimization under a multi-objective reward, ensuring generated trajectories are safe, smooth, and semantically aligned with high-level driving intent. (4) We achieve a PDMS of 91.5 on the NAVSIM benchmark and demonstrating superior performance in safety and drivable area compliance under closed-loop evaluation.

Related Work

Vision-Language Models in Autonomous Driving

The integration of VLMs into end-to-end driving systems has enabled unified perception-reasoning-control pipelines(Li et al. 2025; Xu et al. 2024; Fu et al. 2025; Shao et al. 2024a) that generate interpretable, language-mediated driving plans(Wang et al. 2025; Tian et al. 2024; Liao et al. 2025b). Early methods used LLMs as high-level planners conditioned on extracted scene features, while recent frameworks such as DriveGPT4 (Xu et al. 2024) and AlphaDrive (Jiang et al. 2025b) directly generate driving instructions from raw sensor inputs using powerful VLMs. These models are typically fine-tuned via imitation learning and further refined with reinforcement learning to improve safety and generalization. However, most approaches treat language generation as decoupled from trajectory execution, failing to establish a closed-loop mechanism where semantic intent directly guides action, thus limiting their instruction compliance and behavioral coherence.

End-to-End Driving with Structured Reasoning and Trajectory Optimization

To address the disconnect between understanding and acting, recent efforts have introduced structured reasoning mechanisms into driving policies(Janner et al. 2022; Chi et al. 2023). Methods like SOLVE(Chen et al. 2025) employ Trajectory-of-Thought (T-CoT) to iteratively refine plans through "thinking before acting," improving long-horizon consistency, while Drive-Anywhere(Wang et al. 2024) enables zero-shot generalization via foundation models. A critical insight from these works is that reliable reasoning must be grounded in accurate perception; Perception-R1(Yu et al. 2025) further shows that anchoring policy learning in visual discrimination tasks leads to more robust and interpretable updates. These developments highlight the need for joint optimization of perception and planning through semantic grounding and structured rewards—motivating our approach of using VLMs not merely as planners, but as motion advisors in a closed-loop, language-guided control framework.

Reinforcement Learning for Large Models

Reinforcement learning has proven effective in enhancing the reasoning and generalization capabilities of large language models beyond supervised fine-tuning. Methods such as Direct Preference Optimization (DPO) and GRPO(Shao et al. 2024b) enable reward-free policy refinement by leveraging implicit feedback signals, with GRPO showing particular promise due to its stable group-wise advantage estimation. Recent work like DeepSeek-R1(Guo et al. 2025) and Reason-RFT(Tan et al. 2025) demonstrates that GRPO can effectively improve multi-step reasoning in LLMs and vision-language models (VLMs) when combined with rule-based rewards for format and correctness verification. These advances have inspired its application in multimodal decision-making(Jiang et al. 2025b; Li et al. 2025), where GRPO supports verifiable, structured reasoning—making it a suitable choice for safety-critical domains such as autonomous driving.

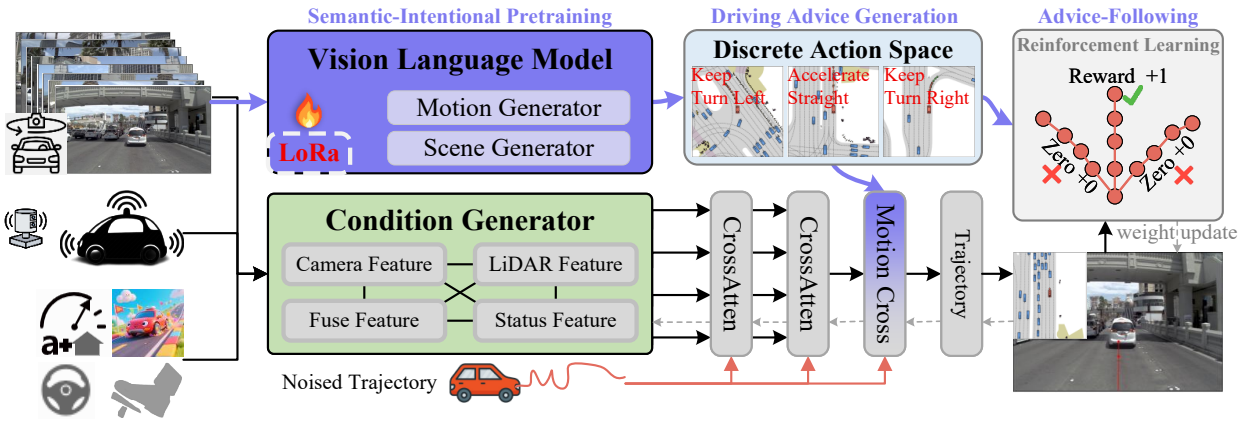

Figure 2: Pipeline of Driving with Advice framework. The pipeline begins with Semantic-Intentional Pretraining, where a vision-language model (VLM) generates high-level motion advice from multimodal inputs. This advice is then formalized into a discrete action space and refined via structured policy learning using GRPO. Finally, an advice-following diffusion-based trajectory generator conditions on the advised motion and refines trajectories under a multi-objective reward, enabling interpretable, adaptive, and safe decision-making.

Methods

We present Driving with Advice, a closed-loop advisor-planner framework that leverages vision-language models (VLMs) as motion advisors to enable interpretable and instruction-aware end-to-end driving. As illustrated in Fig. 2, our method integrates semantic reasoning and trajectory planning through three stages: (i) Semantic-Intentional Pretraining (SIP) transfers driving rationale from a large VLM (InternVL3-78B) to a compact agent (Qwen2.5-VL-3B) via QA-style supervision; (ii) Driving Advice Generation grounds high-level intent into a discrete action space (Direction × Speed) and aligns it with expert behaviors using GRPO; (iii) Advice-Following Trajectory Generation conditions a diffusion-based planner on the advised motion and refines it under a multi-objective reward for safe and adaptive execution. This unified pipeline, with SIP and policy refinement detailed in Fig. 3, establishes a language-mediated cognitive interface between perception and control.

Semantic-Intentional Pretraining

While end-to-end autonomous driving aims to map sensor inputs directly to control outputs, most existing approaches suffer from a critical deficiency: they learn what to do without understanding why to do it . This results in brittle policies that fail under distribution shift, lack interpretability, and cannot justify their decisions—fundamental limitations for safety-critical systems.

To address this, as shown in **region ① of Fig. 3**, we propose Semantic-Intentional Pretraining (SIP), a novel paradigm that equips vision-language models with driving-aware cognition through large-scale, machine-generated question-answering supervision. Rather than relying on manual annotations or narrow behavioral cloning, SIP leverages InternVL3-78B—deployed via Swift(Zhao et al. 2024) for efficient inference—as a frozen cognitive teacher θ_{teacher} to automatically annotate the NAVSIM dataset with rich, intention-laden linguistic labels.

We define two complementary query modalities:

$$\mathcal{Q} = \mathcal{Q}_{\text{sem}} \cup \mathcal{Q}_{\text{drv}}, \tag{1}$$

where $\mathcal{Q}_{\text{sem}}$ elicits scene-level understanding (e.g., "What is the traffic condition ahead?"), and $\mathcal{Q}_{\text{drv}}$ probes behaviorally grounded reasoning (e.g., "Should the ego vehicle prepare for a right turn? Why?"). For each multimodal (cameras-$\mathbf{I}_t$, LiDAR-$\mathbf{L}_t$, Ego Related-$\mathbf{R}_t$) observation $\mathcal{O}_t = \{\mathbf{I}_t, \mathbf{L}_t, \mathbf{R}_t\}$, the teacher generates answers conditioned on both visual context and temporal dynamics:

$$a \sim p(a \mid \mathcal{O}_t, q; \theta_{\text{teacher}}). \tag{2}$$

This yields a large-scale dataset $\mathcal{D}_{\text{QA}}$, where every answer encodes not only perceptual content but also causal justification for action—such as linking "decelerate" to "pedestrian near crosswalk"—thereby embedding normative driving logic into the training signal.

We then fine-tune Qwen2.5-VL-3B(Hui et al. 2024) on $\mathcal{D}_{\text{QA}}$ via supervised learning:

$$\min_{\theta} \mathbb{E}_{(\mathcal{O},q,a)\sim\mathcal{D}_{\text{QA}}} \left[-\log p_\theta(a \mid \mathcal{O}, q) \right]. \tag{3}$$

Through this process, the model acquires a latent intention space—a representation geometry in which sensory inputs are mapped not just to actions, but to justified decisions, forming a strong semantic prior for downstream policy learning. SIP fundamentally rethinks how driving intelligence is acquired: instead of imitating trajectories, we pretrain agents to reason about driving.

Driving Advice Generation

While Semantic-Intentional Pretraining (SIP) endows the agent with rich behavioral priors, it operates in the realm of perception-to-language reasoning. To enable models to better understand whether their actions are reasonable based on scene context, we further design a Driving Advice Generation (DAG) modules.

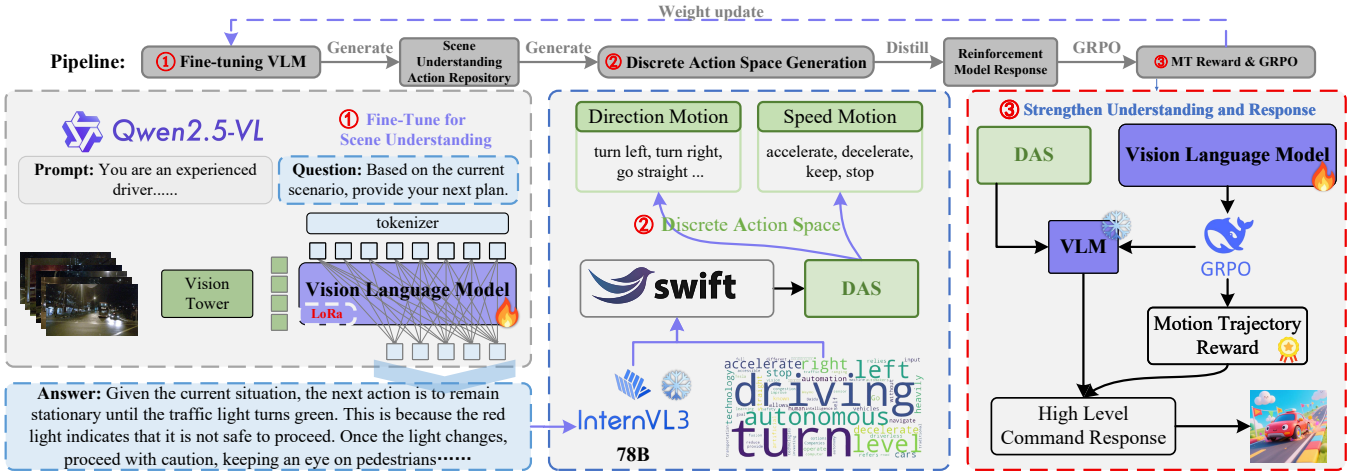

Figure 3: **Overview of the Fine-Tuning Pipeline.** Our framework begins with **Semantic-Intentional Pretraining (SIP)**: a frozen large vision-language model (InternVL3-78B) generates QA-style driving advice on NAVSIM, used to fine-tune Qwen2.5-VL-3B, endowing it with behavioral rationale. This advice is then formalized into a **Discrete Action Space** (DAS), and the policy is refined via GRPO to align with expert motion primitives. This two-stage pipeline establishes the core of our *Driving with Advice* framework—where high-level semantic suggestions are grounded into executable, interpretable actions—laying the foundation for instruction-following trajectory generation.

Scene to Motion. As shown in **region ② of Fig. 3**, to bridge this cognitive foundation to actual vehicle control, we must address a fundamental challenge: **how to ground high-level intentions into safe, executable motion policies**—a transition often marred by semantic drift and action instability in existing end-to-end frameworks.

Existing end-to-end approaches(Wang et al. 2025; Qian et al. 2024; Jiang et al. 2023b) often regress continuous controls or adopt unstructured action spaces, leading to poor interpretability and instability. To address this, as shown in part ② of Fig. 3, we provide the **scene information as prompt input to the large model for reasoning and deliberation**. Furthermore, we introduce a semantics-based Discrete Action Space (DAS) that reflects the hierarchical structure of human driving behaviors. Specifically, we decompose vehicle motion into two dimensions:

$$\mathcal{A} = \mathcal{A}_{\text{dir}} \times \mathcal{A}_{\text{spd}}, \tag{4}$$

where, $\mathcal{A}_{\text{dir}} = \{\texttt{Straight}, \texttt{Left}, \texttt{Right}, \texttt{LC}, \texttt{RC}\}$[1] encodes lateral intent. $\mathcal{A}_{\text{spd}} = \{\texttt{Keep}, \texttt{Acc}, \texttt{Dec}, \texttt{Stop}\}$[1] captures longitudinal dynamics.

Each composite action $a_t \in \mathcal{A}$ represents a **motion primitive**—a symbolic, human-interpretable maneuver unit that aligns with natural driving language (e.g., "turn right and stop"). This design enforces **action modularity** and enables explicit control over strategic decisions, such as lane changes or yielding, which are otherwise buried in continuous output layers.

Motion Reinforcement. As shown in **region ③ of Fig. 3**, to train the policy $\pi_\theta(a \mid \mathcal{O}_t)$ parameterized by Qwen2.5-VL-3B, we perform behavior cloning against expert trajectories extracted from NAVSIM(Dauner et al. 2024), where each state observation $\mathcal{O}_t$ is paired with a corresponding ground-truth motion primitive a_t^*. Rather than using standard maximum likelihood estimation, we employ Group Relative Policy Optimization (GRPO)(Shao et al. 2024b) treats imitation as reward maximization under implicit feedback.

In this formulation, we define a **Binary Alignment Reward (BAR)**:

$$r_t^{\text{align}} = \mathbb{I}(a_t = a_t^*), \tag{5}$$

which provides positive feedback only when both direction and speed components match the expert action. The policy is then updated via:

$$\max_\theta \mathbb{E}_{\pi_\theta} \left[\sum_{t=1}^{T} r_t^{\text{align}} \cdot A_t^{\text{GAE}} \right], \tag{6}$$

where A_t^{GAE} denotes Generalized Advantage Estimation(Schulman et al. 2015), computed using the implicit reward signal.

This stage serves as a **semantic bridge**: it grounds the linguistic reasoning acquired in SIP into concrete, executable decisions, ensuring that the model not only understands what should be done, but also selects the correct motion primitive at each step. By operating in a discrete, interpretable action space, our method achieves both performance fidelity and behavioral transparency, laying the foundation for instruction-aware trajectory generation in complex urban environments.

Advice-Following Trajectory Genreation

While discrete motion primitives provide semantic clarity in decision-making, real-world control requires fine-grained,

[1] `LC` and `RC` represent left lane change and right lane change, respectively. `Acc`, `Dec` represent acceleration and deceleration, respectively.

continuous trajectory planning. To ensure trajectory safety and smoothness, and to enable accurate execution according to the generated advice, we employ GRPO for reinforcement learning-based refinement, further enhancing the accuracy and safety of the generated planning trajectories.

Reference Model Training. To bridge the gap between symbolic actions and executable paths, inspired by Transfuser (Zheng et al. 2025; Jiang et al. 2023b), we propose a diffusion-based trajectory generator trained to produce future trajectories $\hat{\mathbf{T}} \in \mathbb{R}^{H\times 2}$ conditioned on both sensor inputs $\mathcal{O}_t$ and the selected motion primitive a_t, where H is the prediction sequence length. This model serves as our reference policy (Ref Model), denoted as:

$$p_\psi(\mathbf{T} \mid \mathcal{O}_t, a_t) = \int p_\psi(\mathbf{T}_{1:H} \mid \mathcal{O}_t, a_t, \mathbf{z}_{1:H}) d\mathbf{z}, \quad (7)$$

where $\mathbf{z}_{1:H}$ are latent variables sequentially denoised over T diffusion steps:

$$\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{T} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I), \quad (8)$$

and the reverse process estimates $\boldsymbol{\epsilon}_\psi(\mathbf{z}_t, t, \mathcal{O}_t, a_t)$ via a DiT-style(Peebles and Xie 2023) architecture with cross-modal attention.

The Ref Model is first pretrained on expert trajectories using variational lower bound (VLB) loss:

$$\mathcal{L}_{\text{diffuse}} = \mathbb{E}\left[\sum_{t=1}^{T} \beta_t \cdot \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\psi(\mathbf{z}_t, t, \mathcal{O}_t, a_t)\|^2\right]. \quad (9)$$

Crucially, however, the Ref Model alone cannot ensure instruction compliance—it generates typical behaviors, but lacks adaptability to novel directives. To achieve true language-guided autonomy, we introduce a **trajectory reinforcement** stage that optimizes a Policy Model $\pi_\phi(\mathbf{T} \mid \mathcal{O}_t, a_t)$ to follow diverse natural language commands while maintaining safety and comfort.

Trajectory Reinforcement. This is accomplished through GRPO(Shao et al. 2024b), which maximizes expected reward under a multi-component objective:

$$\max_\phi \mathbb{E}_{\pi_\phi}\left[\sum_{t=1}^{H} r_t^{\text{total}} \cdot A_t^{\text{GAE}}\right], \quad \text{with} \quad r_t^{\text{total}} = \sum_{k=1}^{3} w_k r_t^{(k)}, \quad (10)$$

where the composite reward consists of three distinct reward functions designed for different purposes. Specifically, the motion alignment reward r_t^{align} is computed based on a classification metric between the generated trajectory and the advised primitive of VLM, the collision avoidance reward r_t^{safe} penalizes occupancy overlap in the BEV, and the smoothness reward r_t^{smooth} discourages high jerk. These components are jointly optimized to enhance behavioral fidelity, physical safety, and dynamic smoothness.

This final stage realizes a key vision: **from intention to instruction-aware execution**. The agent no longer merely imitates—it adapts, generating trajectories that respect not only expert behavior but also user-specified constraints. By integrating diffusion modeling, semantic grounding, and structured reward shaping within a unified GRPO framework, we establish a framework for language-conditioned, safety-aware motion generation in autonomous driving.

Experiments

Datasets

NAVSIM(Dauner et al. 2024), an autonomous driving dataset designed for motion planning, extends OpenScene (a repackaged version of nuPlan). The dataset includes synchronized multi-sensor data: eight 1920×1080 cameras and a LiDAR point cloud fused from five sensors across four consecutive frames (current + three previous). For benchmarking, it provides 1,192 training scenes (navtrain) and 136 test scenes (navtest). And, building on modern prompt-based AD LLM adaptation, we leverage InternVL3-78B(Zhu et al. 2025) to produce QA(Achiam et al. 2023) pairs as supervision signals, spanning scene comprehension, driving decisions, and salient object detection for fine-tuning(Hu et al. 2022).

Metrics

We evaluate our method on the NAVSIM benchmark using the Predictive Driver Model Score (PDMS), a composite metric that aggregates safety, progress, and comfort. PDMS is computed as PDMS $=$ NC $\times$ DAC $\times$ DDC $\times$ $\frac{(5\times\text{TTC}+5\times\text{EP}+2\times\text{C})}{12}$, where NC (No at-fault Collisions), DAC (Drivable Area Compliance), TTC (Time to Collision), C (Comfort), and EP (Ego Progress). DDC is excluded due to evaluation toolkit limitations, consistent with prior work(Xing et al. 2025; Liao et al. 2025a). All results are averaged over the 'navtest' split in a closed-loop simulation setting.

Implementation Details

We fine-tune Qwen2.5VL-3B using sensor-to-QA pairs generated through our prompt framework, followed by reinforcement training with GRPO. All experiments are conducted on 48×A800 GPUs (80GB memory). For parameter-efficient adaptation, we employ standard LoRA(Hu et al. 2022) configurations as the learnable components during fine-tuning.

Training Pipeline. Our framework is trained in three stages. First, we perform Semantic-Intentional Pretraining (SIP) by fine-tuning Qwen2.5-VL-3B on a QA dataset generated by InternVL3-78B, endowing the model with driving-aware reasoning. Second, we refine the policy in a discrete action space via GRPO to align with expert motion primitives, enabling interpretable decision-making. Third, we optimize a diffusion-based trajectory generator using GRPO under a multi-objective reward. This staged pipeline progressively bridges cognition, action, and execution.

Quantitative Results

Main Results. As shown in Table 1, our method achieves the highest PDMS of 91.5, outperforming the strong baseline Hydra-MDP++(Li et al. 2024b) (91.0). The improvement is driven by enhanced safety and consistency, with notable gains in NC (99.2 vs. 98.6) and EP (86.9 vs. 85.7). This is attributed to our closed-loop advisor-planner architecture, where VLM-generated advice provides semantic priors for

Method	Modality	Framework	Ego	NC↑	DAC↑	EP↑	TTC↑	C↑	PDMS↑
AD-MLP	-	MLP	✓	93.0	77.3	62.8	83.6	100.0	65.6
UniAD	C	MLP	✓	97.8	91.9	78.8	92.9	100.0	83.4
LTF	C	MLP	✓	97.4	92.8	79.0	92.4	100.0	83.8
PARA-Drive	C	MLP	✓	97.9	92.4	79.3	93.0	99.8	84.0
DrivingGPT	C	Q&A	✓	98.9	90.7	79.7	94.9	95.6	82.4
DRAMA	C+L	Q&A	✓	98.0	93.1	80.1	94.8	100.0	85.5
VADv2	C+L	MLP	✓	97.2	89.1	76.0	91.6	100.0	80.9
Hydra-MDP	C+L	MLP	✓	98.3	96.0	78.7	94.6	100.0	86.5
Hydra-MDP++	C+L	MLP	✓	98.6	**98.6**	85.7	95.1	100.0	91.0
Transfuser	C+L	MLP	✓	84.0	92.8	79.2	92.8	100.0	84.0
BevDrive	C+L	MLP	✓	97.7	92.5	78.7	92.9	100.0	83.8
WoTE	C+L	MLP	✓	98.5	96.8	81.9	94.9	99.9	88.3
ReCogDrive	C+L	Diffusion	✓	98.2	97.8	83.5	95.2	99.8	89.6
DiffusionDrive	C+L	Diffusion	✓	98.2	96.2	82.2	94.7	100.0	88.1
GoalFlow	C+L	Diffusion	✓	98.4	98.3	85.0	94.6	100.0	90.3
Ours	C+L	Diffusion	✓	**99.2**	97.7	**86.9**	**96.1**	100.0	**91.5**

Table 1: Closed-loop driving performance on NAVSIM navtest. **L:** LiDAR, **C:** Camera, **Ego:** ego state. Our method establishes a closed-loop advisor-planner architecture, leveraging VLM-generated advice for interpretable policy learning and instruction-following trajectory generation, achieving the best PDMS.

Setting	VLM	DAG	Trajectory Reinforcement	NC↑	DAC↑	EP↑	TTC↑	C↑	PDMS↑
1	Qwen2.5-0.5B	✗	✗	97.5	96.0	80.9	94.5	100	87.5
2		✓	✓	98.2	97.6	82.5	94.4	99.8	88.7
3	Qwen2.5-3B	✗	✗	97.7	96.4	82.1	94.7	100	88.1
4		✓	✓	99.2	97.7	86.9	96.1	100	91.5

Table 2: Ablation Study on Key Components of the Driving with Advice. This table evaluates the impact of VLM size, Driving Advice Generation (DAG), and Trajectory Reinforcement on driving metrics.

Modules	VLM	Rewards	All Match Accuracy(%)	PDMS
SIP	Qwen2.5-3B	N/A	76.4	87.1
DAG		Regx	89.5	88.1
		BAR	97.5	91.5

Table 3: Ablation on DAG for motion planning precision. All Match Accuracy (%) measures the accuracy of predicted motion against expert ground truth. BAR for Eq. (5), Regx for Regular Expression.

Rewards			NC↑	DAC↑	EP↑	TTC↑	C↑	PDMS↑
Motion	Collision	Smoothness						
✓	✗	✗	97.8	97.5	80.3	94.8	99.8	87.7
✓	✓	✗	97.7	97.6	80.7	**94.8**	99.8	87.8
✓	✓	✓	**98.7**	**97.7**	**82.0**	94.5	**100.0**	**88.7**

Table 4: Ablation study on the components of the multi reward function—motion alignment, collision avoidance, and trajectory smoothness.

decision-making, and GRPO-based refinement with multi-objective rewards ensures safer, more compliant trajectory generation.

Ablation Study on Key Components. Table 2 validates the contribution of each component. The full model achieves the highest PDMS (91.5%). Disabling DAG drops performance significantly, demonstrating its critical role in refining VLM advice into precise discrete actions. Using a smaller VLM (0.5B) further degrades performance, confirming that both model capacity and the DAG module are essential for high-precision motion planning. Moreover, training with reinforcement learning leads to a noticeable performance improvement, regardless of whether the model has 0.5B or 3B parameters. This also further verifies the effectiveness of our designed reward function.

Ablation of Motion Planning Accuracy. We evaluate motion planning precision via an ablation on the DAG module. Table 3 shows that DAG significantly improves action alignment, boosting All Match Accuracy from 76.4% (SIP only) to 97.5% (SIP+DAG+BAR). The gain stems from structured policy refinement, which grounds semantic intent into discrete, executable actions. Reward reinforcement further enhances accuracy by shaping trajectory generation. Our full model achieves 97.5% match rate and PDMS of 91.5, outperforming ablated variants.

Ablation of Rewards. As shown in Table 4, the ablation study reveals that the full reward formulation significantly enhances trajectory quality. Disabling any component degrades performance, with the complete set achieving a PDMS of 88.7. Motion alignment ensures intent compli-

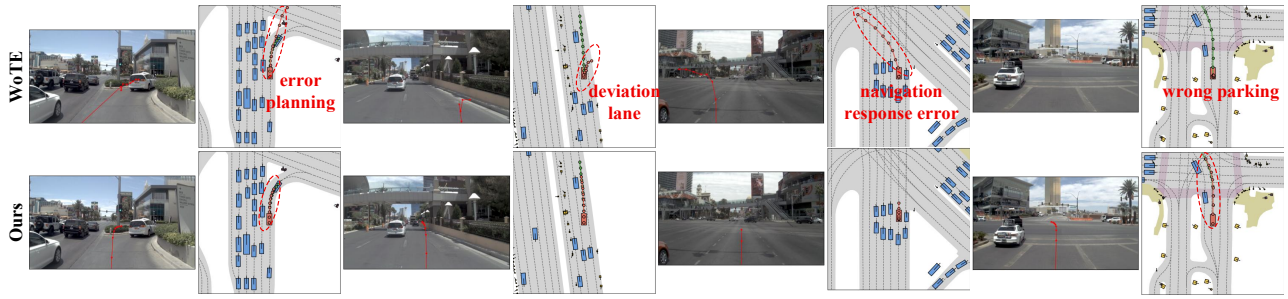

Figure 4: Qualitative Comparison of Trajectory Planning. Our approach demonstrates significant improvements in trajectory planning accuracy, particularly in handling complex environments with dynamic obstacles, lane changes, and navigation errors.

Figure 5: Comparison of scene understanding and driving advice between our method and original Qwen2.5-VL-3B fine-tune. Red boxes represent key objects.

Modules	NC↑	DAC↑	EP↑	TTC↑	C↑	PDMS↑
T=5	98.2	96.4	84.7	94.7	99.7	89.4
T=10	99.2	97.7	86.9	96.1	100.0	91.5
T=20	98.7	97.5	85.8	95.3	100.0	91.1

Table 5: Ablation Study on Denoising Steps for Trajectory Quality.

ance, collision avoidance maintains safety (NC: 98.7), and smoothness improves comfort (C: 100.0). The strong performance highlights the necessity of multi-objective shaping, where each reward term contributes uniquely to safe, coherent, and instruction-following driving behavior.

Ablation on Denoising Steps. The ablation study on denoising steps highlights a trade-off between trajectory quality and computational efficiency. As shown in Table 5, increasing denoising steps from $T = 5$ to $T = 10$ notably improves PDMS (89.4 to 91.5) and achieves perfect Comfort (C), showing iterative refinement benefits smoother, safer trajectories. Further increasing T to 20 gives diminishing returns, with slight PDMS decline while comfort remains high, indicating $T = 10$ balances performance and cost optimally.

Qualitative Results

Trajectory Visualization Comparison As shown in Fig. 4, our method produces spatially coherent, smooth, and context-aware trajectories across diverse urban scenarios—accurately handling intersections, dense-traffic lane keeping, and precise stopping. This demonstrates strong alignment between high-level advice and low-level control, validating our approach's ability to generate adaptive, driver-like behavior.

Scene-Motion Understanding Comparison in Fig. 5 shows that our method surpasses Qwen2.5-VL-3B in interpretability and contextual awareness. While both models describe scenes, ours yields more actionable advice by explicitly recognizing critical elements—such as blind spots, landmarks, and traffic rules—and outputs precise motion primitives (e.g., <Left, Acc>, <Straight, Stop>) aligned with high-level intent, leading to safer and more context-aware decisions in complex urban settings like intersections or construction zones.

Conclusion

We propose Driving with Advice, a closed-loop framework that leverages a vision-language model as a motion advisor to bridge high-level semantic reasoning and low-level trajectory planning. We introduce Semantic-Intentional Pretraining to inject driving rationale into compact models, design a discrete action space grounded in directional and speed primitives for interpretable policy learning, and refine trajectory generation via advice-following diffusion with multi-objective GRPO optimization. Our method achieves a state-of-the-art PDMS of 91.5 on NAVSIM, surpassing Hydra-MDP++ by 0.5 points with improved safety (NC: 99.2) and drivable area compliance (DAC: 97.7), demonstrating the effectiveness of language-mediated cognitive guidance in end-to-end autonomous driving.

Acknowledgments

This work was in part supported by NSFC (Grant No. 62176194) and the National Key Research and Development Program of China (Grant No. 2022ZD0160604), and the Key Research and Development Program of Hubei Province (Grant No. 2024BAB030), the Project of Sanya Yazhou Bay Science and Technology City (Grant No. SCKJ-JYRC-2022-76, SKJC-2022-PTDX-031), and the Project of Sanya Science and Education Innovation Park of Wuhan University of Technology (Grant No. 2021KF0031).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Chen, S.; Jiang, B.; Gao, H.; Liao, B.; Xu, Q.; Zhang, Q.; Huang, C.; Liu, W.; and Wang, X. 2024. VADv2: End-to-End Vectorized Autonomous Driving via Probabilistic Planning. *CoRR*, abs/2402.13243.

Chen, X.; Huang, L.; Ma, T.; Fang, R.; Shi, S.; and Li, H. 2025. SOLVE: Synergy of Language-Vision and End-to-End Networks for Autonomous Driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12068–12077.

Chi, C.; Feng, S.; Du, Y.; Xu, Z.; Cousineau, E.; Burchfiel, B.; and Song, S. 2023. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In Bekris, K. E.; Hauser, K.; Herbert, S. L.; and Yu, J., eds., *Robotics: Science and Systems XIX, Daegu, Republic of Korea, July 10-14, 2023*.

Dauner, D.; Hallgarten, M.; Li, T.; Weng, X.; Huang, Z.; Yang, Z.; Li, H.; Gilitschenski, I.; Ivanovic, B.; Pavone, M.; et al. 2024. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. *Advances in Neural Information Processing Systems*, 37: 28706–28719.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

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 preprint arXiv:2503.19755*.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Hu, Y.; Yang, J.; Chen, L.; Li, K.; Sima, C.; Zhu, X.; Chai, S.; Du, S.; Lin, T.; Wang, W.; et al. 2023. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17853–17862.

Huang, Z.; Weng, X.; Igl, M.; Chen, Y.; Cao, Y.; Ivanovic, B.; Pavone, M.; and Lv, C. 2024. Gen-drive: Enhancing diffusion generative driving policies with reward modeling and reinforcement learning fine-tuning. *arXiv preprint arXiv:2410.05582*.

Hui, B.; Yang, J.; Cui, Z.; Yang, J.; Liu, D.; Zhang, L.; Liu, T.; Zhang, J.; Yu, B.; Lu, K.; et al. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Janner, M.; Du, Y.; Tenenbaum, J. B.; and Levine, S. 2022. Planning with Diffusion for Flexible Behavior Synthesis. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 9902–9915. PMLR.

Jiang, A.; Gao, Y.; Sun, Z.; Wang, Y.; Wang, J.; Chai, J.; Cao, Q.; Heng, Y.; Jiang, H.; Dong, Y.; et al. 2025a. Diffvla: Vision-language guided diffusion planning for autonomous driving. *arXiv preprint arXiv:2505.19381*.

Jiang, B.; Chen, S.; Xu, Q.; Liao, B.; Chen, J.; Zhou, H.; Zhang, Q.; Liu, W.; Huang, C.; and Wang, X. 2023a. VAD: Vectorized Scene Representation for Efficient Autonomous Driving. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 8306–8316. IEEE.

Jiang, B.; Chen, S.; Zhang, Q.; Liu, W.; and Wang, X. 2025b. Alphadrive: Unleashing the power of vlms in autonomous driving via reinforcement learning and reasoning. *arXiv preprint arXiv:2503.07608*.

Jiang, C.; Cornman, A.; Park, C.; Sapp, B.; Zhou, Y.; Anguelov, D.; et al. 2023b. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9644–9653.

Li, Y.; Xiong, K.; Guo, X.; Li, F.; Yan, S.; Xu, G.; Zhou, L.; Chen, L.; Sun, H.; Wang, B.; et al. 2025. ReCogDrive: A Reinforced Cognitive Framework for End-to-End Autonomous Driving. *arXiv preprint arXiv:2506.08052*.

Li, Z.; Li, K.; Wang, S.; Lan, S.; Yu, Z.; Ji, Y.; Li, Z.; Zhu, Z.; Kautz, J.; Wu, Z.; Jiang, Y.; and Álvarez, J. M. 2024a. Hydra-MDP: End-to-end Multimodal Planning with Multi-target Hydra-Distillation. *CoRR*, abs/2406.06978.

Li, Z.; Li, K.; Wang, S.; Lan, S.; Yu, Z.; Ji, Y.; Li, Z.; Zhu, Z.; Kautz, J.; Wu, Z.; et al. 2024b. Hydra-mdp: End-to-end multimodal planning with multi-target hydra-distillation. *arXiv preprint arXiv:2406.06978*.

Li, Z.; Yu, Z.; Lan, S.; Li, J.; Kautz, J.; Lu, T.; and Alvarez, J. M. 2024c. Is ego status all you need for open-loop end-to-end autonomous driving? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14864–14873.

Liao, B.; Chen, S.; Yin, H.; Jiang, B.; Wang, C.; Yan, S.; Zhang, X.; Li, X.; Zhang, Y.; Zhang, Q.; et al. 2025a. Dif-

fusiondrive: Truncated diffusion model for end-to-end autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12037–12047.

Liao, H.; Kong, H.; Wang, B.; Wang, C.; Ye, W.; He, Z.; Xu, C.; and Li, Z. 2025b. Cot-drive: Efficient motion forecasting for autonomous driving with llms and chain-of-thought prompting. *IEEE Transactions on Artificial Intelligence*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Mandalika, S.; Nambiar, A.; et al. 2025. Primedrive-cot: A precognitive chain-of-thought framework for uncertainty-aware object interaction in driving scene scenario. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5293–5301.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

Qian, T.; Chen, J.; Zhuo, L.; Jiao, Y.; and Jiang, Y.-G. 2024. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4542–4550.

Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; and Abbeel, P. 2015. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.

Shao, H.; Hu, Y.; Wang, L.; Song, G.; Waslander, S. L.; Liu, Y.; and Li, H. 2024a. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15120–15130.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024b. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. *URL https://arxiv.org/abs/2402.03300*, 2(3): 5.

Tan, H.; Ji, Y.; Hao, X.; Lin, M.; Wang, P.; Wang, Z.; and Zhang, S. 2025. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*.

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. *arXiv preprint arXiv:2402.12289*.

Wang, S.; Yu, Z.; Jiang, X.; Lan, S.; Shi, M.; Chang, N.; Kautz, J.; Li, Y.; and Alvarez, J. M. 2025. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22442–22452.

Wang, T.-H.; Maalouf, A.; Xiao, W.; Ban, Y.; Amini, A.; Rosman, G.; Karaman, S.; and Rus, D. 2024. Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 6687–6694. IEEE.

Weng, X.; Ivanovic, B.; Wang, Y.; Wang, Y.; and Pavone, M. 2024. PARA-Drive: Parallelized Architecture for Real-Time Autonomous Driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 15449–15458. IEEE.

Xing, Z.; Zhang, X.; Hu, Y.; Jiang, B.; He, T.; Zhang, Q.; Long, X.; and Yin, W. 2025. GoalFlow: Goal-Driven Flow Matching for Multimodal Trajectories Generation in End-to-End Autonomous Driving. *arXiv preprint arXiv:2503.05689*.

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*.

Yu, E.; Lin, K.; Zhao, L.; Yin, J.; Wei, Y.; Peng, Y.; Wei, H.; Sun, J.; Han, C.; Ge, Z.; et al. 2025. Perception-r1: Pioneering perception policy with reinforcement learning. *arXiv preprint arXiv:2504.07954*.

Zhao, Y.; Huang, J.; Hu, J.; Wang, X.; Mao, Y.; Zhang, D.; Jiang, Z.; Wu, Z.; Ai, B.; Wang, A.; Zhou, W.; and Chen, Y. 2024. SWIFT:A Scalable lightWeight Infrastructure for Fine-Tuning. arXiv:2408.05517.

Zheng, Y.; Liang, R.; Zheng, K.; Zheng, J.; Mao, L.; Li, J.; Gu, W.; Ai, R.; Li, S. E.; Zhan, X.; et al. 2025. Diffusion-based planning for autonomous driving with flexible guidance. *arXiv preprint arXiv:2501.15564*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.