DriveMoE: Mixture-of-Experts for Vision-Language-Action Model in End-to-End Autonomous Driving

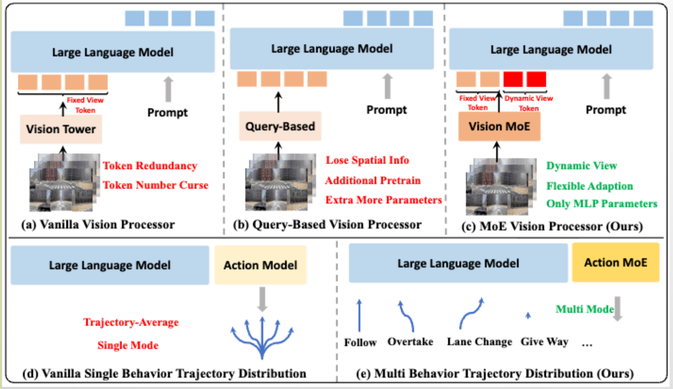
Zhenjie Yang[1,4*], Yilin Chai[1*], Xiaosong Jia[2,3*], Qifeng Li[1], Yuqian Shao[1,4], Xuekai Zhu[1], Haisheng Su[1], Junchi Yan[1†]
[1]Shanghai Jiao Tong University, [2]Fudan University, [3]Shanghai Key Laboratory of Multimodal Embodied AI, [4]AnyScale AI

CVPR JUNE 3-7, 2026 DENVER COLORADO

Abstract

End-to-end autonomous driving (E2E-AD) demands effective processing of multi-view sensor data and robust handling of diverse and complex driving scenarios, particularly rare maneuvers such as aggressive turns. The recent success of the Mixture-of-Experts (MoE) architectures in Large Language Models (LLMs) demonstrates that expert specialization enables strong scalability. In this work, we propose **DriveMoE**, a novel MoE-based E2E-AD framework, with a **Scene-Specialized Vision MoE** and a **Skill-Specialized Action MoE**. First, we introduce **Drive-π_0**, a Vision-Language-Action (VLA) baseline adapted from Embodied AI for autonomous driving, which serves as the foundation model for DriveMoE. Building on this, we strengthen perception through a carefully designed Vision MoE, where a router adaptively selects context-relevant camera views. This mechanism is inspired by human driving cognition, in which attention is directed to key visual cues rather than to all sensory inputs simultaneously. Beyond perception, we introduce an Action MoE that augments the framework by training a router to activate specialized expert modules tailored to distinct driving behaviors. Within the Action MoE, we implement two distinct styles, **Token-level Router** and **Trajectory-level Router**, and extensively explore their applicability in autonomous driving. In Bench2Drive closed-loop evaluations, DriveMoE demonstrates robust performance across diverse driving scenarios, alleviates the mode-averaging effect that limits existing models, and achieves state-of-the-art results with significant improvements over **Drive-π_0**.

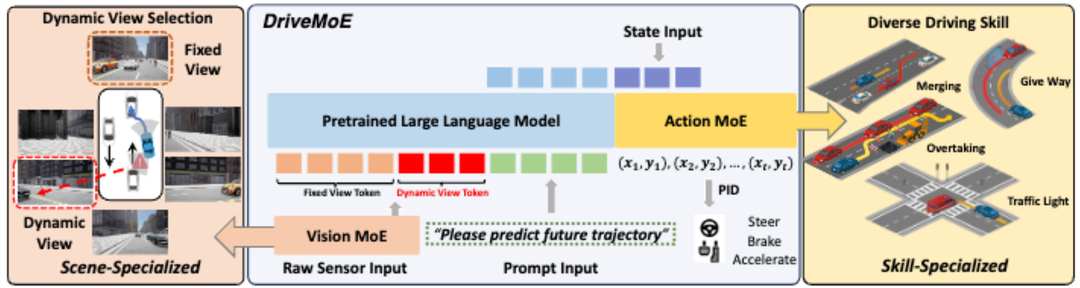
Introduction



Comparison of Different Vision and Action Modeling Strategies in VLA-based End-to-End Driving.

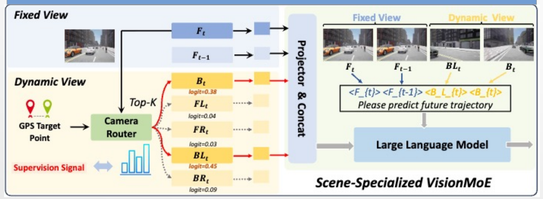
- Vanilla encoding uses all camera views, causing redundant tokens and high computation.
- Query-based extraction selects fewer tokens but loses spatial structure and needs extra pretraining.
- Scene-Specialized Vision MoE selects only key camera views, reducing redundancy.
- Standard action models use one policy head for all scenarios, limiting rare-behavior performance.
- Skill-Specialized Action MoE activates behavior-specific experts for more adaptive planning.

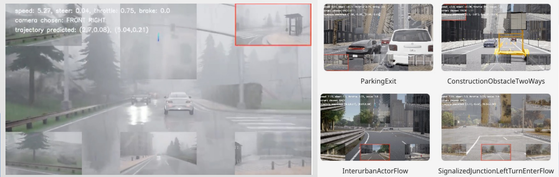
Methods



Framework of DriveMoE. Our proposed framework comprises two main Mixture-of-Experts (MoE) modules tailored for end-to-end autonomous driving. The Scene-Specialized Vision MoE dynamically selects relevant camera views based on real-time driving contexts, efficiently reducing visual redundancy. Subsequently, selected views are fused into a unified representation by projector layers. The Skill-Specialized Action MoE, integrated within a flow-matching planner, activates expert controllers specifically optimized for distinct driving behaviors such as merging, overtaking, emergency braking, yielding, and responding to traffic signs. This dual MoE structure enhances computational efficiency, adaptability, and robustness to rare, safety-critical driving scenarios.

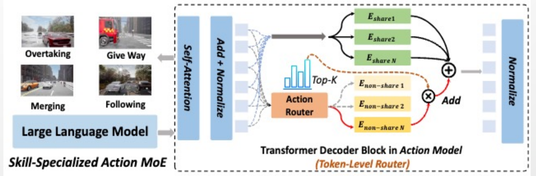
Vision MoE



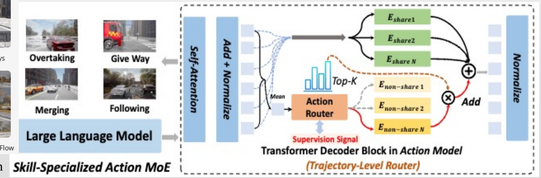


This shows DriveMoE on the Bench2Drive under challenging corner-case scenarios. The Vision MoE adaptively selects camera views according to the driving context and defaults to the rear view when no critical view is identified, enabling more robust perception and decision-making.

Action MoE



Token-Level Skill-Specialized Action Mixture-of-Experts.



Trajectory-Level Skill-Specialized Action Mixture-of-Experts.

Experiment

Table 2. **Results on the Bench2Drive Benchmark(Closed-Loop and Open-Loop)**. * denotes expert feature distillation.

Method	Venue	Closed-loop Metric				Open-loop Metric
		DS ↑	SR(%) ↑	Efficiency ↑	Comfort ↑	Avg. L2↓
TCP-traj* [49]	NeurIPS 2022	59.90	30.00	76.54	18.08	1.70
AD-MLP [56]	Arxiv 2023	18.05	0.00	48.45	22.63	3.64
VAD [26]	ICCV 2023	42.35	15.00	157.94	46.01	0.91
UniAD-Base [15]	CVPR 2023	45.81	16.36	129.21	43.58	0.73
ThinkTwice* [22]	CVPR 2023	62.44	31.23	69.33	16.22	0.95
DriveAdapter* [21]	ICCV 2023	64.22	33.08	70.22	16.01	1.01
GenAD [57]	ECCV 2024	44.81	15.90	-	-	-
DriveTrans [24]	ICLR 2025	63.46	35.01	100.64	20.78	**0.62**
MomAD [43]	CVPR 2025	44.54	16.71	170.21	**48.63**	0.82
WoTE [31]	ICCV 2025	61.71	31.36	-	-	-
DriveMamba-L [44]	ICLR 2026	66.82	37.73	152.91	18.77	0.70
DiffAD [47]	Arxiv 2025	67.92	38.64	-	-	1.55
Raw2Drive [53]	NeurIPS 2025	71.36	50.24	214.17	22.42	-
Drive-π_0	Ours	55.85	30.00	173.63	35.70	1.13
DriveMoE (Token-Level)		66.94	35.45	158.80	6.86	0.96
DriveMoE (Traj-Level)		**74.22**	**48.64**	**175.96**	15.31	1.01

Table 3. **Ablation study on Vision MoE.** Compare different camera view combinations and supervision signals. F, FL, FR, and B indicate the front, front-left, front-right, and back views, respectively, while BL and BR represent the back-left and back-right views. **Fixed View** means selecting a specific view. **Dynamic View** refers to the camera view dynamically selected by the vision router as the top-1 relevant view according to scene context. Exp 1 denotes our baseline **Drive-π_0**, which models surrounding agents' velocities from two consecutive front-view images, and Exp 9 denotes **DriveMoE**, which adds a dynamically selected view with explicit supervision to enhance perception learning. Memory is evaluated at **batch size=1**. **All experiments use the previous-frame front view by default.**

Exp	I_F	I_{FL}	I_{FR}	I_B	I_{BL}	I_{BR}	View	Supervision	DS↑	SR(%)↑	Latency↓	Memory(MB)
1	✓	×	×	×	×	×	Fixed	-	55.85	30.00	100ms	4100
2	✓	✓	×	×	×	×	Fixed	-	62.38	33.64	260ms	5100
3	✓	×	✓	×	×	×	Fixed	-	61.52	32.73	260ms	5100
4	✓	×	×	✓	×	×	Fixed	-	63.26	31.82	260ms	5100
5	✓	✓	✓	×	×	×	Fixed	-	64.92	33.64	400ms	7400
6	✓	✓	✓	✓	×	×	Fixed	-	64.18	33.64	550ms	9600
7	✓	✓	✓	✓	✓	✓	Fixed	-	62.27	31.36	700ms	11800
8	✓	-	-	-	-	-	Dynamic	×	69.71	44.09	260ms	5100
9	✓	-	-	-	-	-	Dynamic	✓	74.22	48.64	260ms	5100

Conclusions

We propose DriveMoE, a novel end-to-end autonomous driving framework built upon Drive-π_0, which integrates Mixture-of-Experts(MoE) into both vision and action components. DriveMoE effectively addresses challenges inherent in existing VLA models by dynamically selecting relevant camera views through a Scene-Specialized Vision MoE, and by employing a Skill-Specialized Action MoE that activates expert modules tailored to specific driving behaviors. Extensive evaluations on the Bench2Drive benchmark show that DriveMoE achieves state-of-the-art performance, significantly enhancing computational efficiency and robustness to rare, safety-critical driving scenarios.

WeChat

Project Page

GitHub

Email: yangzhenjie@sjtu.edu.cn