ICML 2026

AutoMoT

Asynchronous Vision-Language-Action for End-to-End Autonomous Driving

UNDERSTANDING

LAYER-WISE JOINT ATTENTION

ACTION

- ASYNCHRONOUS VLA
- 37 ms • 27 Hz
- *BENCH2DRIVE* & *nuScenes* SOTA

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

HARVARD UNIVERSITY

XIAOMI EV

ICML International Conference On Machine Learning

AutoMoT: A Unified Vision-Language-Action Model with Asynchronous Mixture -of-Transformers for End-to-End Autonomous Driving

Wenhui (Oscar) Huang[1,2], Songyan Zhang[1,3], Qihang Huang[1],
Zhidong Wang[1], Zhiqi Mao[1], Collister Chua[1], Long Chen[3], Chen Lv[1*]

[1]NTU, [2]Harvard, [3]Xiaomi EV

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

HARVARD UNIVERSITY

Background

End-to-End Autonomous driving has come to the long-tail stage.

Traffic light is Big Green.

Stop sign shows in the AD screen.

E2E AD:

Strengths:
1. Deployable
2. Low Cost

Weakness:
1. Causal error
2. Human-Robot interaction

VLM/A:

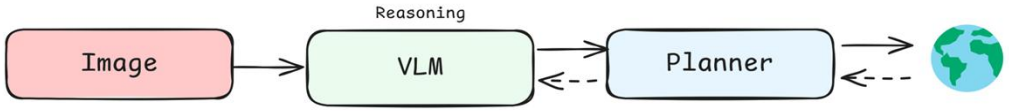
Strengths:
1. Strong Reasoning

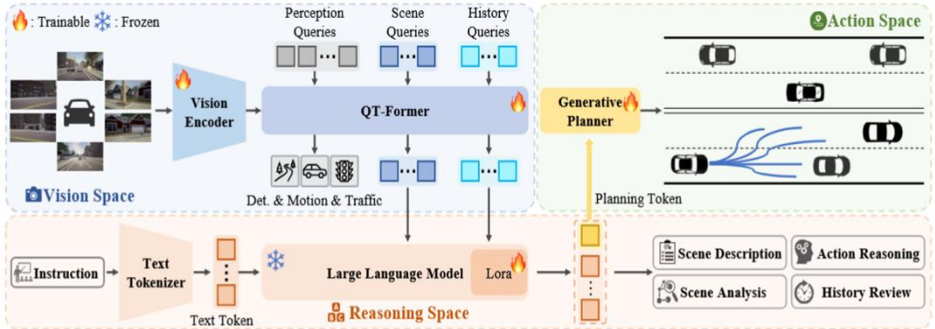
2. Human-Robot interaction

Existing Approaches and Rethinking

Why not integrate VLM with E2E framework?

a) VLM as Partial Module (Scene Encoder)





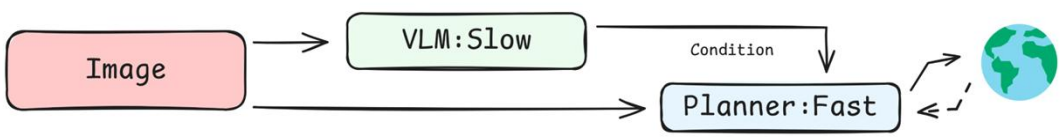
Pros:

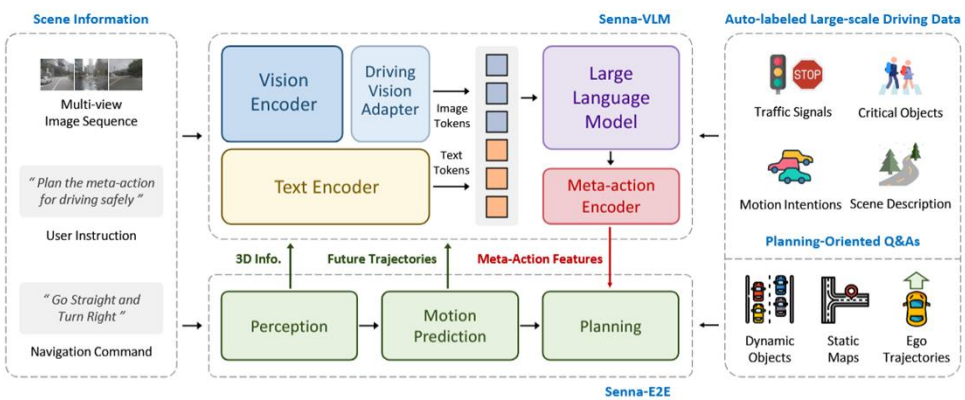
- Strong Scene Understanding
- Excellent Interpretability

Cons:

- Distributional misalignment
- High Latency

b) VLM as Secondary System (Dual System)





Pros:

- Fast Inference
- Decision Steerability

Cons:

- Burdensome architecture
- Overkilled Setup

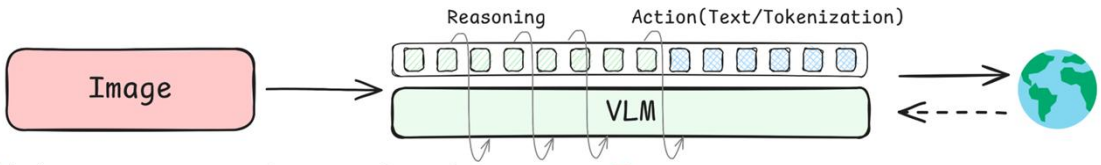
Jiang, Bo, et al. "Senna: Bridging large vision-language models and end-to-end autonomous driving." *arXiv preprint arXiv:2410.22313* (2024).
Fu, Haoyu, et al. "Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2025.
Tian, Xiaoyu, et al. "Drivevlm: The convergence of autonomous driving and large vision-language models." arXiv preprint arXiv:2402.12289 (2024).

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
HARVARD UNIVERSITY
ICML International Conference On Machine Learning

Existing Approaches and Rethinking

Why not integrate VLM with E2E framework?

c) VLM/A as End-to-End Model



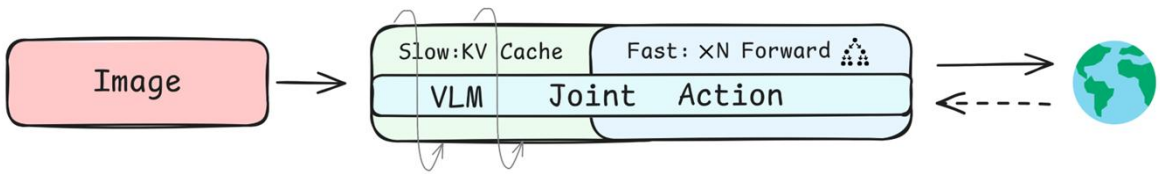
Pros:

- Unified Representation
- From VLMs to VLAs

Cons:

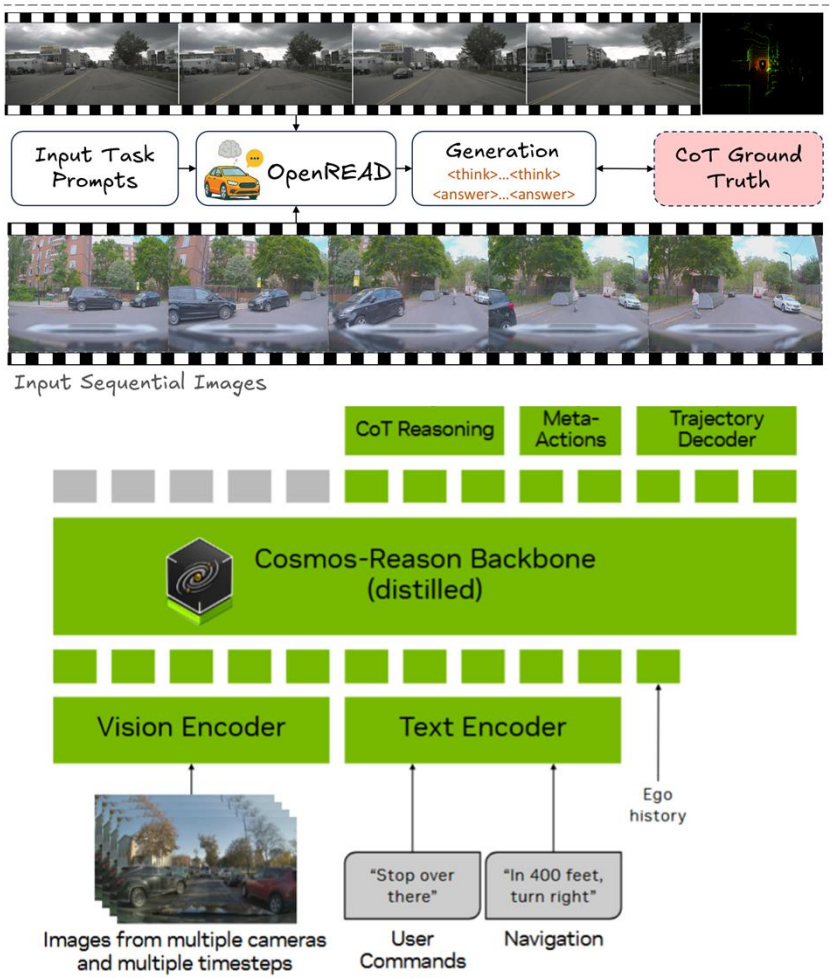
- Modality Gap
- Synchronous: High Latency

d) Asynchronous VLA as End-to-End Model



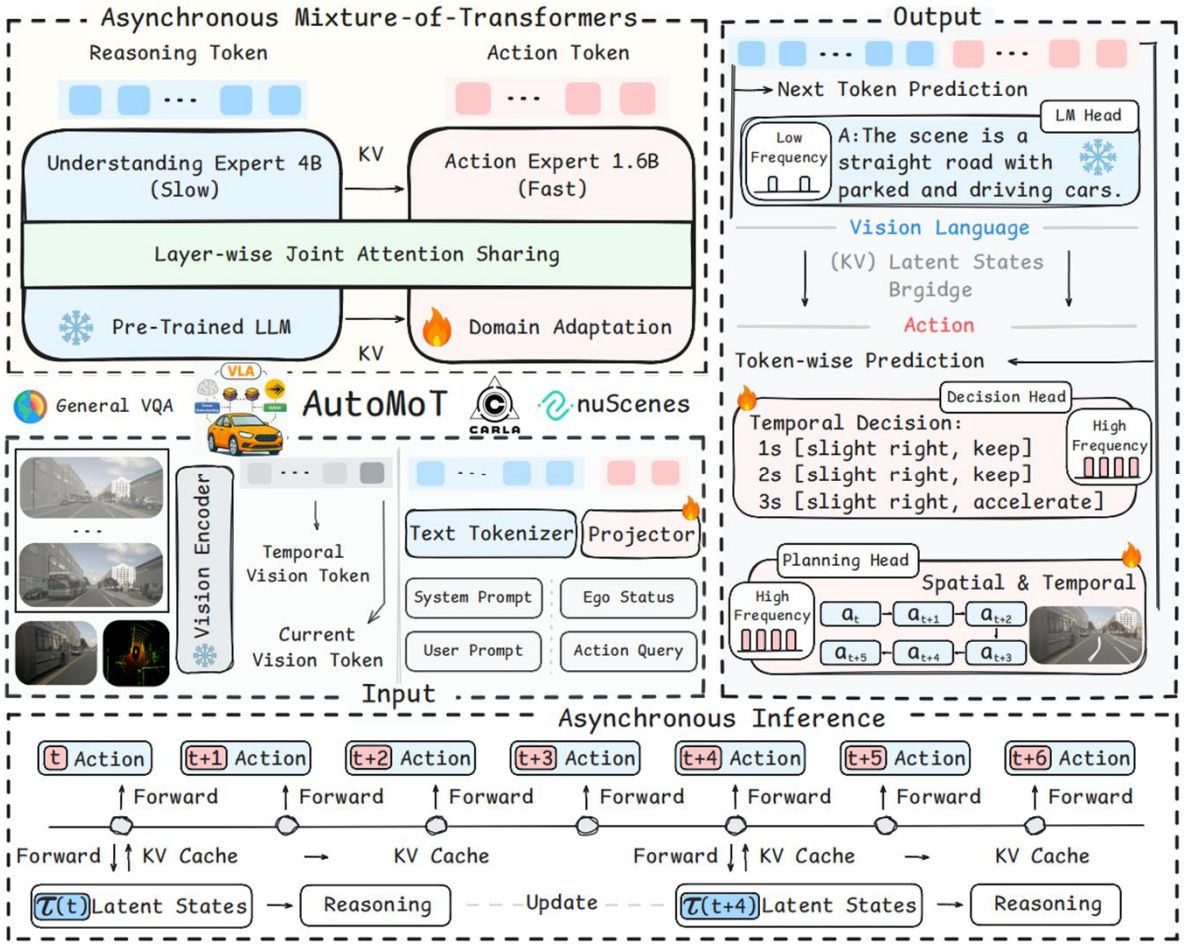
Idea:

- Unified VLA Architecture
- Decoupled Functions
- Reserve and Leverage Knowledge
- Asynchronous Inference



Zhou, Zewei, et al. "Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning." arXiv preprint arXiv:2506.13757 (2025).
Wang, Yan, et al. "Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail." arXiv preprint arXiv:2511.00088 (2025).Zhang, Songyan, et al. "OpenREAD: Reinforced Open-Ended Reasoning for End-to-End Autonomous Driving with LLM-as-Critic." arXiv preprint arXiv:2512.01830 (2025).

AutoMoT: An Asynchronous VLA as E2E Model



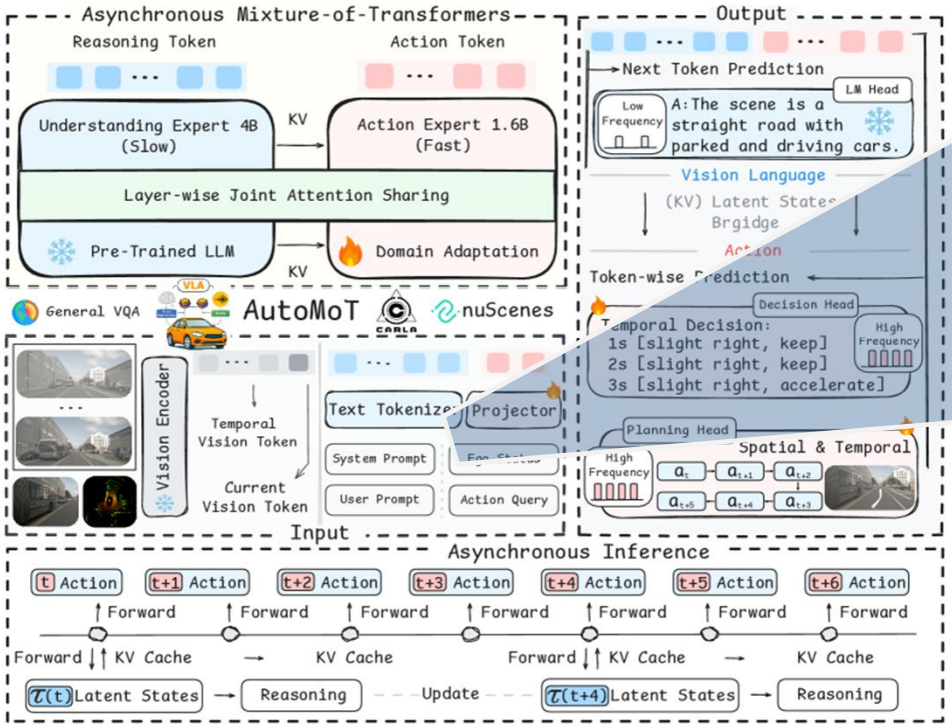
AutoMoT

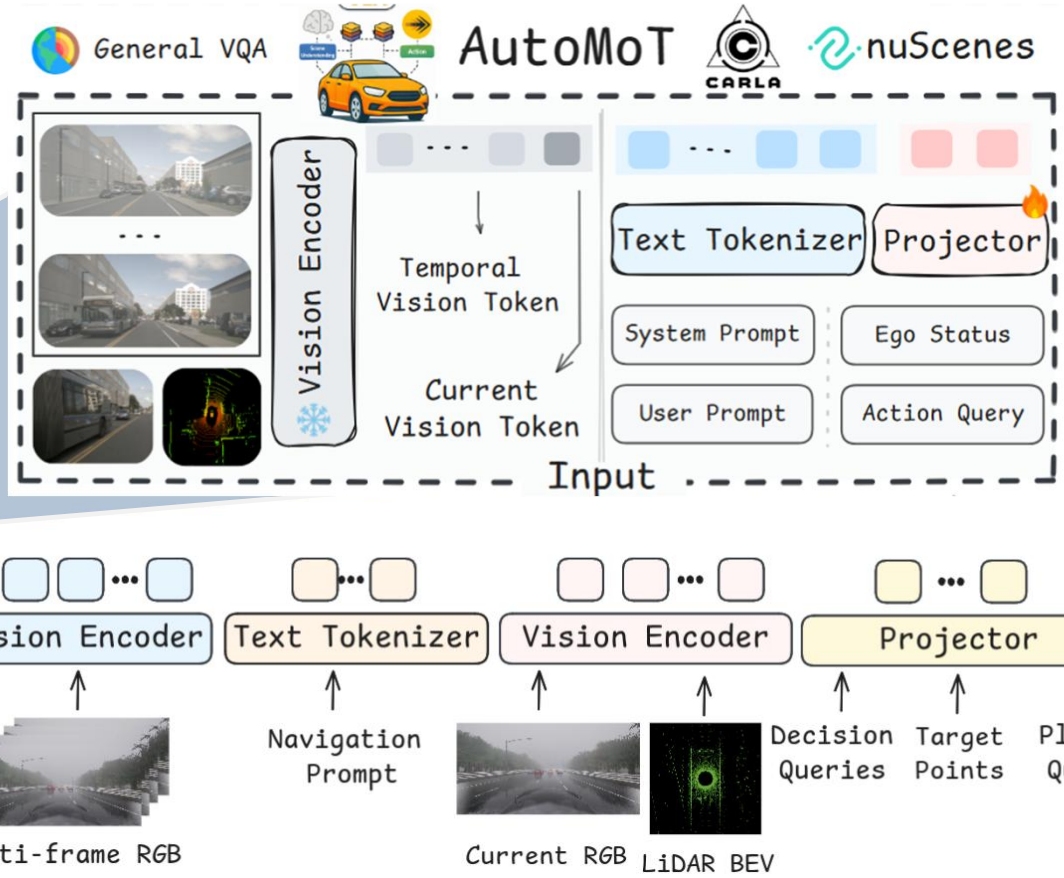
- Architecture: Mixture-of-Transformers
- Hierarchical E2E -> Flattened (Unified) E2E
- Asynchronous Inference
- Closed-loop long-tail Performance: SOTA
- Open-loop Safety Performance: SOTA
- General VQA Reasoning: Preserved

Huang, W., Zhang, S., Huang, Q., Wang, Z., Mao, Z., Chua, C., ... & Lv, C. (2026). Automot: A unified vision-language-action model with asynchronous mixture-of-transformers for end-to-end autonomous driving. ICML 2026.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
HARVARD UNIVERSITY
ICML International Conference On Machine Learning

AutoMoT: Input Space





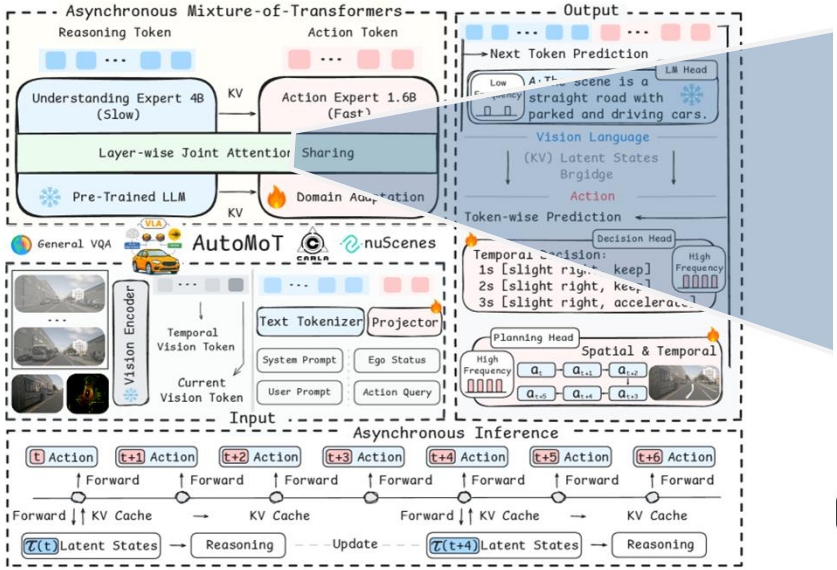
Input Space

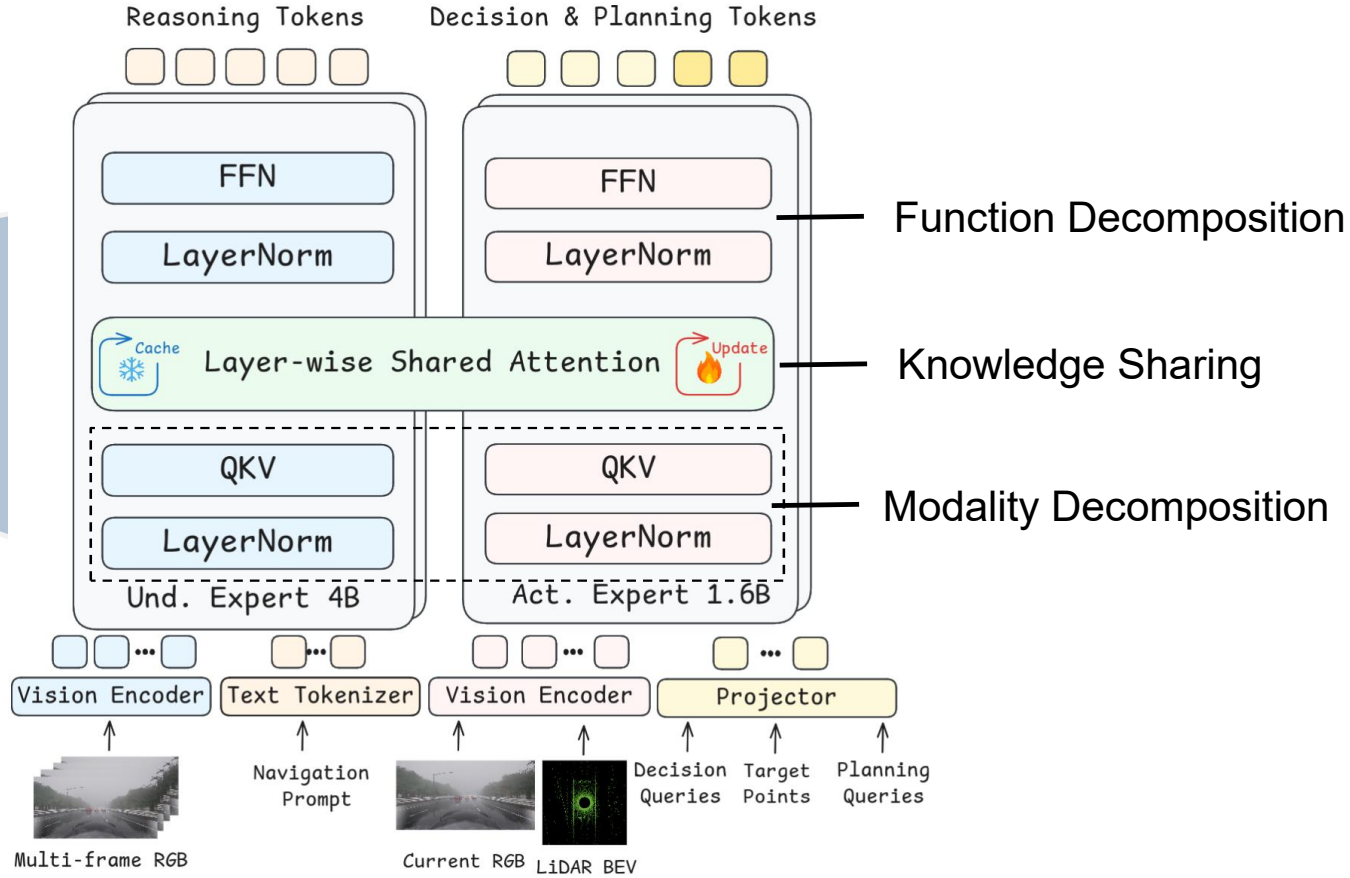
Huang, W., Zhang, S., Huang, Q., Wang, Z., Mao, Z., Chua, C., ... & Lv, C. (2026). Automot: A unified vision-language-action model with asynchronous mixture-of-transformers for end-to-end autonomous driving. ICML 2026.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
HARVARD UNIVERSITY
ICML International Conference On Machine Learning

AutoMoT: Architecture

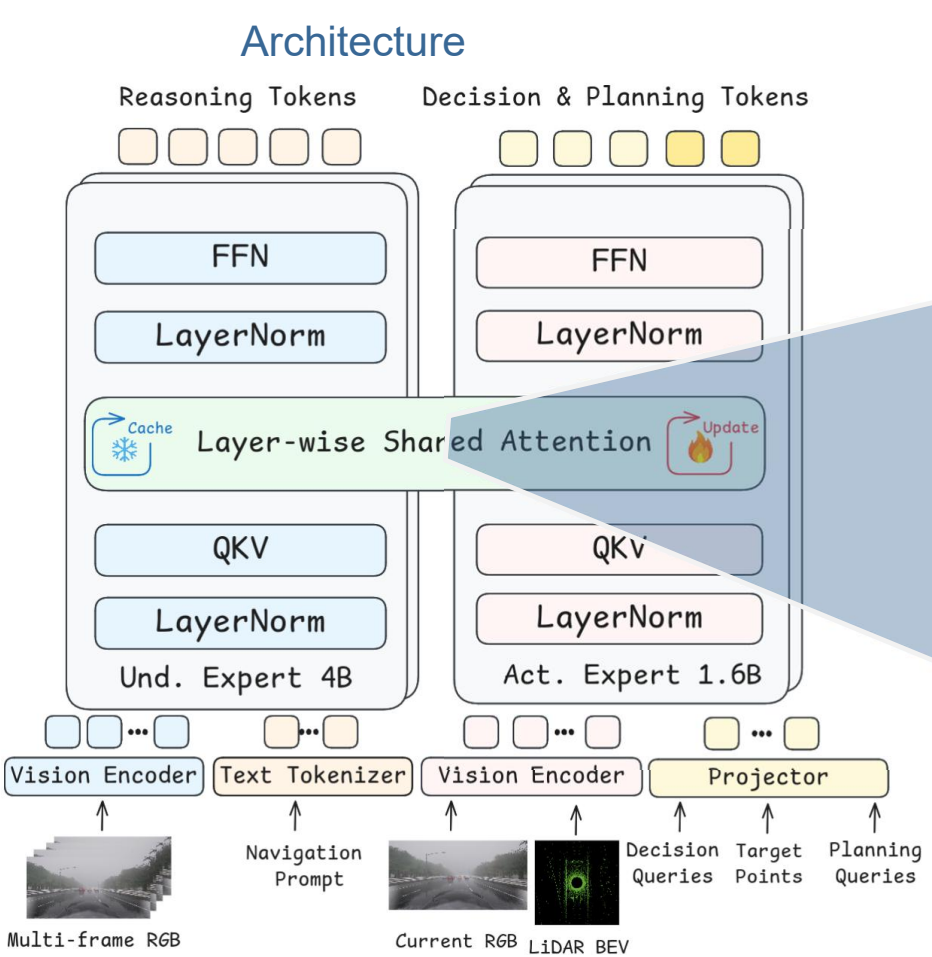
Architecture

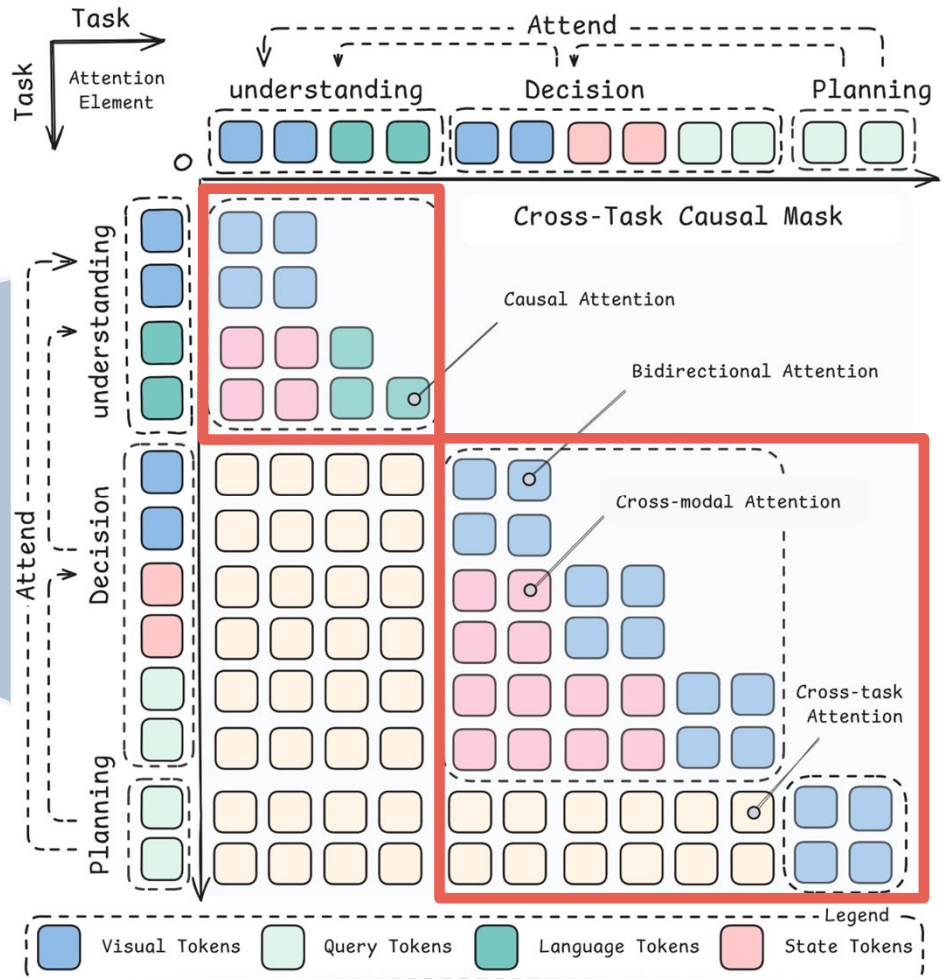




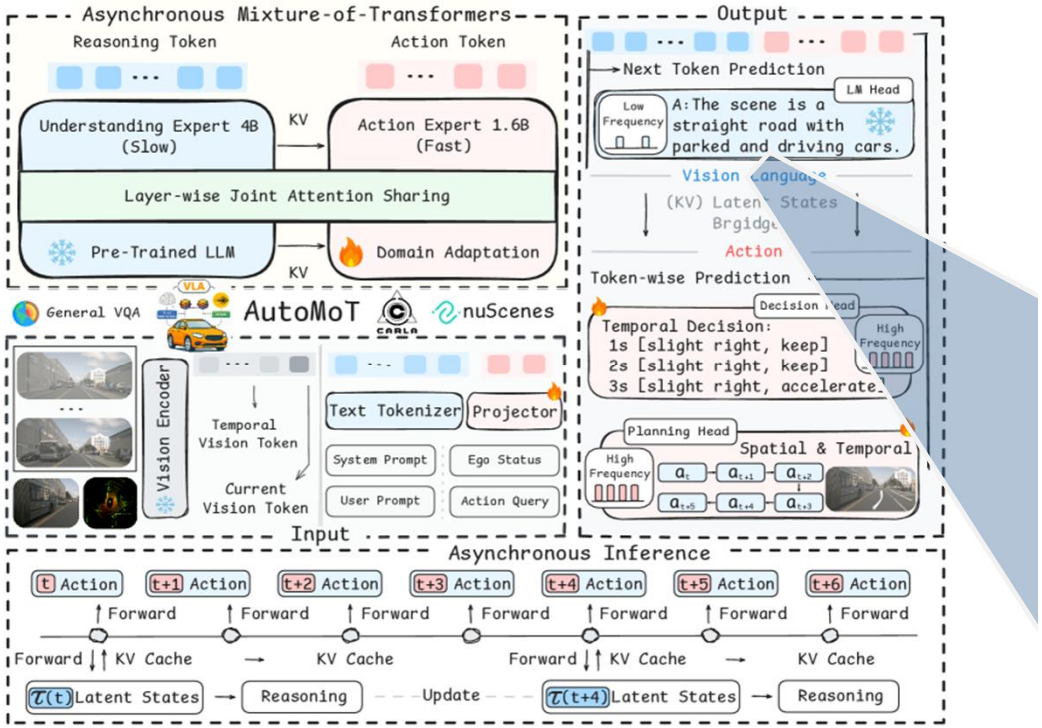
Huang, W., Zhang, S., Huang, Q., Wang, Z., Mao, Z., Chua, C., ... & Lv, C. (2026). Automot: A unified vision-language-action model with asynchronous mixture-of-transformers for end-to-end autonomous driving. ICML 2026.

AutoMoT: Layer-wise Joint Global Attention





AutoMoT: Output Space



Human-Vehicle Interaction (HVI)

- Next-Token Prediction
- Open-ended Reasoning
- Detailed Answers

Scene Understanding (Reasoning)

: How is the weather?

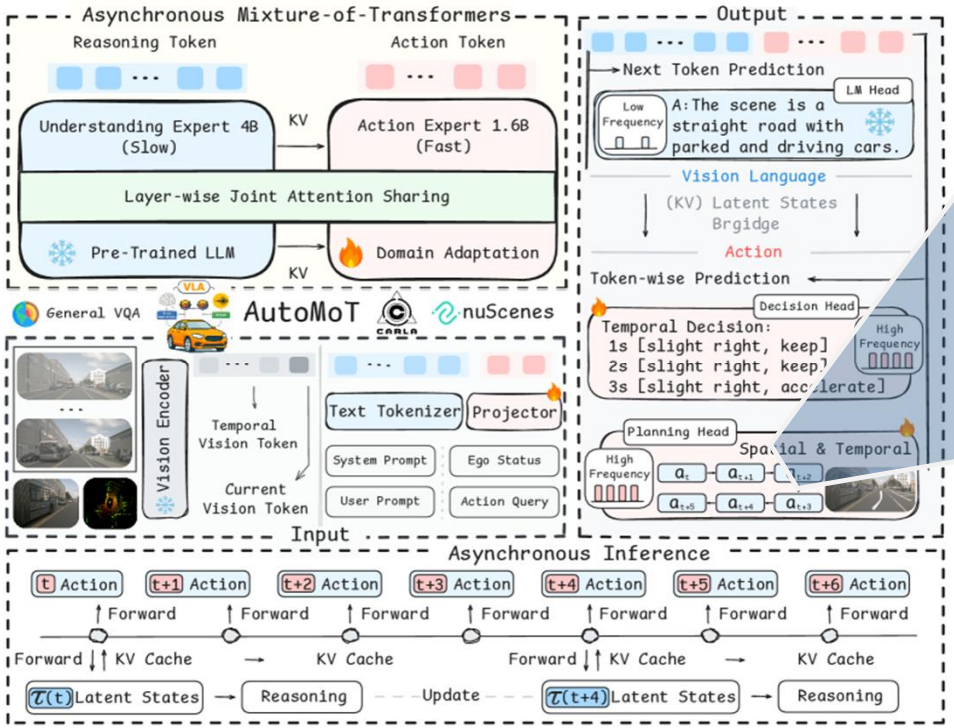
: It's a sunny day.

: What's your current action and it's justification?

: I am decelerating to let the pedestrian safely across the zebra or pedestrian cross ahead.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE · HARVARD UNIVERSITY · ICML International Conference On Machine Learning

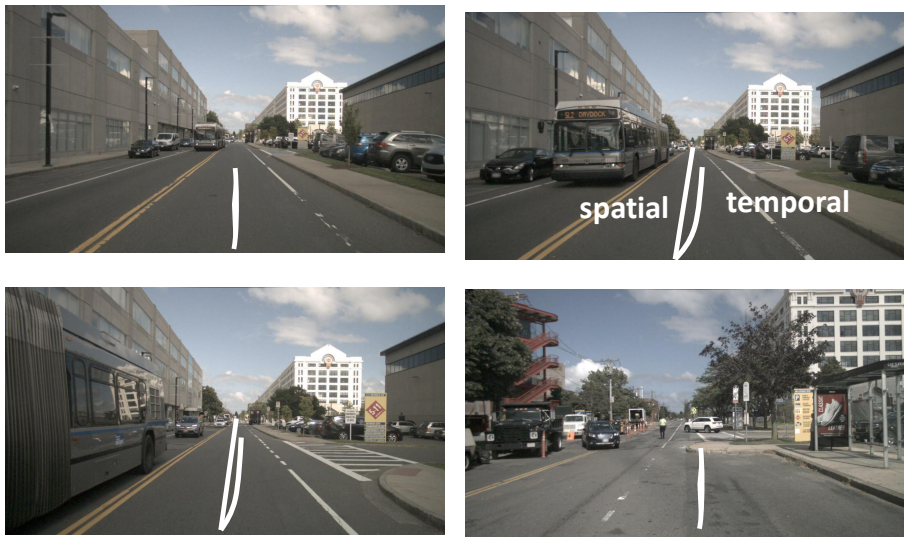
AutoMoT: Output Space



Decision (Meta-Action) and Planning

- Token-wise Prediction (Fast)
- 3 seconds Meta-actions
- 3 seconds future trajectories (Spatial & Temporal)

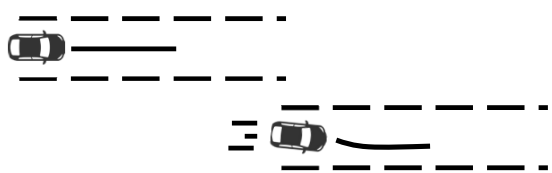
Action Policy



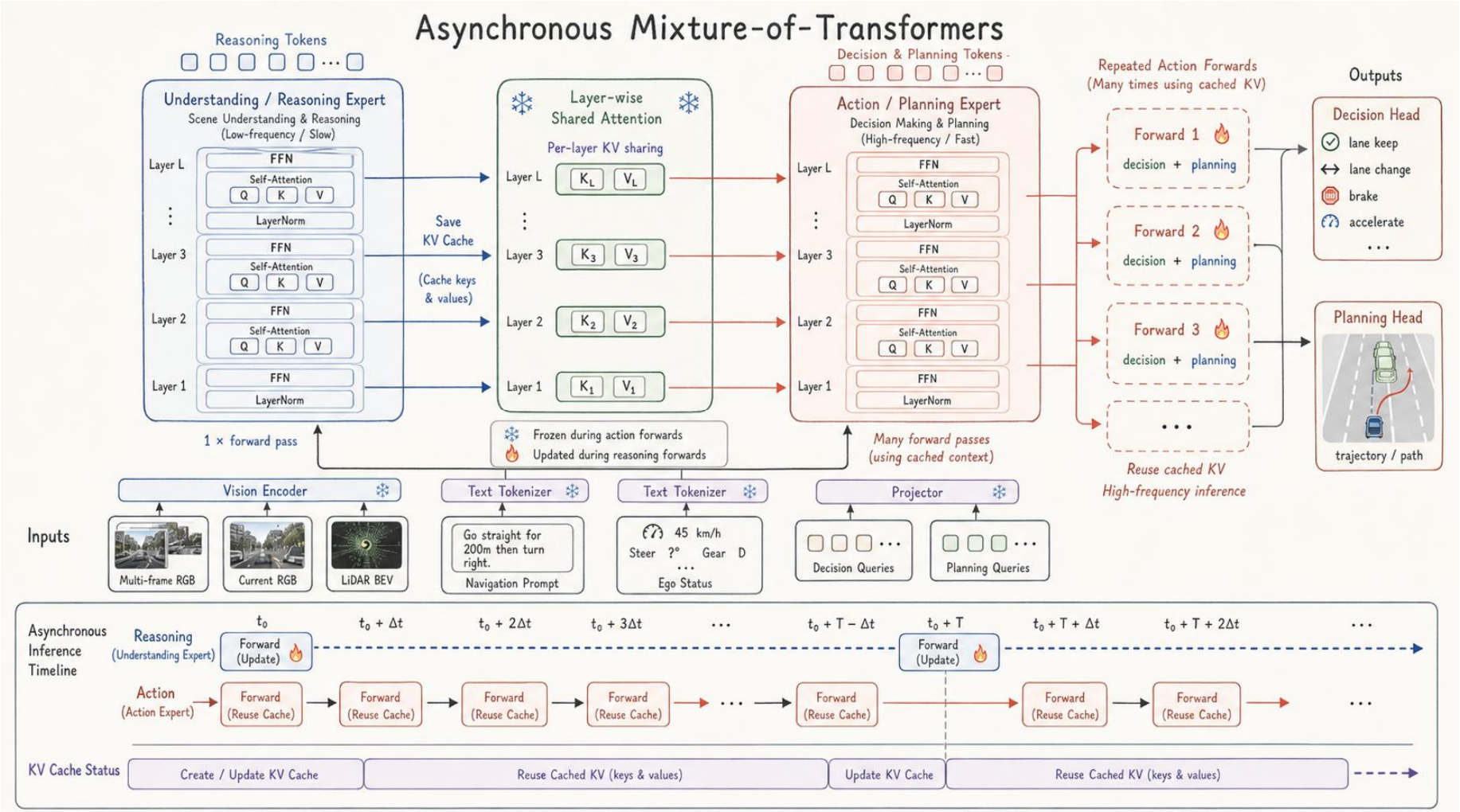
: [Slight right, Keep], [Slight right, Keep], [Slight right, Accelerate]

: spatial […];

temporal […]



AutoMoT: Asynchronous Inference



Exploration: Generative (Diffusion) Head

Noise
Spatial Trajectory
Temporal Trajectory
Cross-Attention
Mixture-of -Attention (MoA) Block
Denoise
Refine
Spatial
Temporal
Condition signal
c
BEV Features
Latent Decision States
c
Condition signal
Input from Previous Block
Modulation
Parallel
Self Attention
BEV Cross-Attention
Latent Cross-Attention
Sequantial Cross-Attention
Modulation
Ouput to Next Block
Mixture-of-Attention (MoA) Block

Closed-Loop Longtail Performance

AutoMoT: An Asynchronous VLA Model for E2E Autonomous Driving

Table 1. Comparison of closed-loop planning performance on the CARLA Bench2Drive leaderboard. DS and SR represent Driving Score and Success Rate, respectively.

Method	Expert	VLM	Generative Planner	Closed-loop Metric	
				DS↑	SR(%)↑
MomAD (Song et al., 2025)	Think2Drive	-	-	44.54	16.71
UniAD-Base (Hu et al., 2023)	Think2Drive	-	-	45.81	16.36
TCP-traj (Wu et al., 2022)	Think2Drive	-	-	59.90	30.00
DriveTransformer-Large (Jia et al., 2025)	Think2Drive	-	-	63.46	35.01
DriveAdapter (Jia et al., 2023)	Think2Drive	-	-	64.22	33.08
Raw2Drive (Yang et al., 2026)	Think2Drive	-	-	71.36	50.24
DiffusionDrive (Liao et al., 2025)	PDM-Lite	-	✓	77.68	57.72
TransFuser++ (Jaeger et al., 2023b)	PDM-Lite	-	-	84.21	67.27
ReasonPlan (Liu et al., 2025c)	Think2Drive	✓	-	64.01	34.55
Recogdrive (Li et al., 2025c)	Think2Drive	✓	✓	71.36	45.45
DriveMoE (Yang et al., 2025)	Think2Drive	✓	-	74.22	48.64
ORION (Fu et al., 2025a)	Think2Drive	✓	✓	77.74	54.62
SpaceDrive+ (Li et al., 2025a)	PDM-Lite	✓	-	78.02	55.11
MindDrive (Fu et al., 2025b)	Think2Drive	✓	✓	78.04	55.09
AutoVLA (Zhou et al., 2025b)	PDM-Lite	✓	-	78.84	57.73
SimLingo (Renz et al., 2025)	PDM-Lite	✓	-	85.07	67.27
AutoMoT (ours)	PDM-Lite	✓	-	**87.34**	**70.00**
AutoMoT + (ours)	PDM-Lite	✓	✓	**89.42**	**74.09**

Closed-Loop Longtail Performance

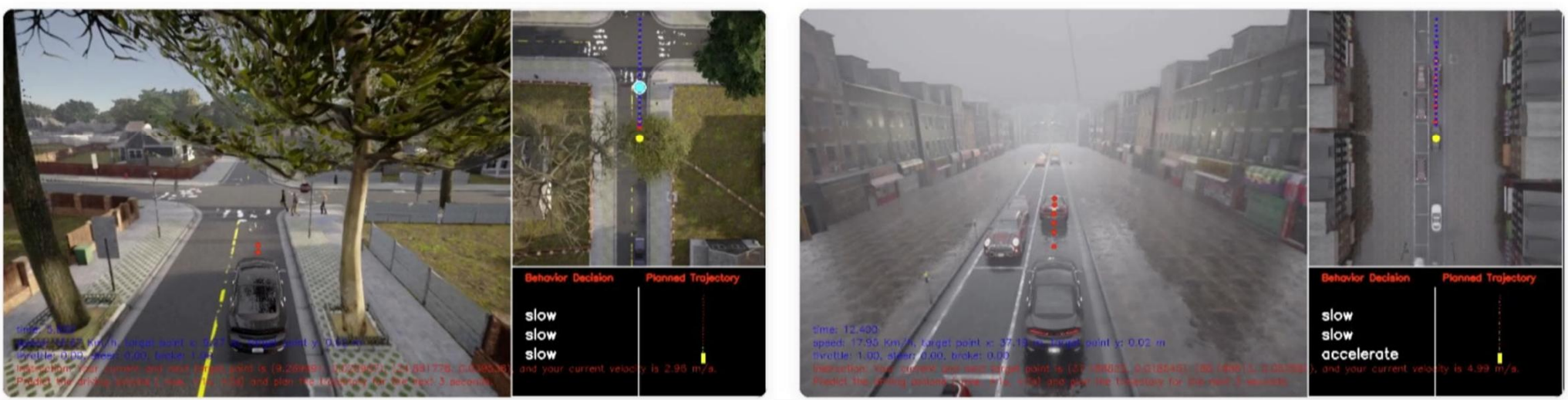
AutoMoT: A Unified Vision-Language-Action Model with Asynchronous Mixture-of-Transformers for End-to-End Autonomous Driving

Paper | Code | Dataset

Due to the double-blind review policy, all code, datasets, and checkpoints will be released after the review process.

Closed-loop Demonstrations

Example closed-loop runs in CARLA showcasing key scenarios.



NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
HARVARD UNIVERSITY
ICML International Conference On Machine Learning

Asynchronous Inference and Latency

Table 1: Latency breakdown of AutoMoT under synchronized (AutoMoT-S) and asynchronous cached-step (AutoMoT) settings. UE and AE denote Understanding Expert and Action Expert, respectively. All latencies in ms.

Setting	DP Refiner	UE	AE	Refiner	Total
AutoMoT-S	–	80.3	37.0	–	117.3
AutoMoT	–	0.0	37.0	–	37.0
AutoMoT-S	✓	80.3	37.0	26.0	143.3
AutoMoT	✓	0.0	37.0	26.0	63.0

Table 1: Inference latency comparison of VLA-based driving methods. †Results cited from the original paper, as public checkpoints are not available for reproduction. Inference GPU is not specified (training uses 8× NVIDIA L40S).

Method	GPU	Lat. (ms)	Hz
OpenEMMA	RTX 5090	7,683	0.13
AutoVLA† (fast)	Unknown	1,072	0.93
AutoVLA† (slow)	Unknown	10,518	0.10
SimLingo	RTX 5090	430	2.3
AutoMoT-S (sync)	RTX 5090	117	8.5
AutoMoT (async)	RTX 5090	**37**	**27**

Setting	L2@1s	L2@2s	L2@3s	$L2_{avg}$
Original	0.14	0.29	0.54	0.32
Camera-only	0.14	0.31	0.51	0.32

Open-Loop Performance

Table 1: Comparison of the Open-loop planning in nuScenes. The ST-P3 evaluation protocol is used by default.

Method	Ego Status	Finetuning			L2 (m) ↓				Collision (%) ↓			
		Und.	Dec.	Plan.	1s	2s	3s	Avg.	1s	2s	3s	Avg.
UniAD	Vector	-	-	✓	0.44	0.67	0.96	0.69	0.04	0.08	0.23	0.12
VAD	Vector	-	-	✓	0.17	0.34	0.60	0.37	0.07	0.10	0.24	0.14
Ego-MLP	Vector	-	-	✓	0.15	0.32	0.59	0.35	0.00	0.27	0.85	0.37
DriveTrans.-L	Vector	-	-	✓	0.16	0.30	0.55	0.33	0.01	0.06	0.15	0.07
AutoVLA	Text	✓	-	✓	0.21	0.38	0.60	0.40	0.13	0.18	0.28	0.20
ORION	-	✓	-	✓	0.17	0.31	0.55	0.34	0.05	0.25	0.80	0.37
RoboTron-Drive	-	✓	-	✓	0.14	0.30	0.57	0.33	0.03	0.12	0.63	0.26
OpenDrive-VLA	Text	✓	-	✓	0.15	0.31	0.55	0.33	0.01	0.08	0.21	0.10
OmniDrive	Vector	✓	-	✓	0.14	0.29	0.55	0.33	0.00	0.13	0.78	0.30
EMMA†	Text	✓	-	✓	0.14	0.29	0.54	0.32	-	-	-	-
SpaceDrive	Vector	✓	-	✓	0.15	0.29	0.51	0.32	0.04	0.18	0.49	0.23
OpenREAD	Vector	✓	-	✓	0.17	0.34	0.56	0.36	0.04	0.08	0.22	0.11
DriveVLM-Dual	Vector	✓	-	✓	0.15	0.29	**0.48**	**0.31**	0.05	0.08	0.17	0.10
Drive-R1	Text	✓	-	✓	**0.14**	**0.28**	0.50	**0.31**	0.02	**0.06**	0.19	0.09
OpenEMMA	Text	-	-	-	1.45	3.21	3.76	2.81	-	-	-	-
AutoMoT (Ours)	Vector	-	✓	✓	**0.14**	0.29	0.54	0.32	**0.01**	**0.06**	**0.15**	**0.07**

1. Is it always necessary to be tailored in AD?
2. Open-ended Reasoning?
3. Catastrophic forgetting?

Rethink Performance Boundary of VLM in E2E AD

Investigate the functional boundaries of pretrained VLMs in autonomous driving, clarifying when and to what extent AD-specific fine-tuning is necessary across different tasks.

Table 3. Comparison of reasoning capabilities across both general-domain and autonomous driving–specific datasets. †: Results are reproduced using the official checkpoints and evaluation environments.

Method	LingoQA	OmniDrive	CODA-LM	TallyQA	InfoVQA
ReCogDrive	**67.20**	0.82	5.90	69.60	75.80
Robotron-Drive†	59.20	0.82	**6.20**	63.40	42.60
OpenEMMA	48.00	0.43	4.80	80.00	71.40
AutoMoT	67.00	**0.89**	6.07	**81.40**	**89.30**

1. Fine-tuning on Scene Understanding is marginal
2. Fine-tuning on action policy is significant
3. Entire tailoring would result in catastrophic forgetting in complex reasoning.

Table 4. Ablation study results of investigating the performance boundary of the pre-trained backbone. †: System prompt is provided; ‡: Fine-tuned on autonomous driving datasets; L: Lingo-Judge; G:GPT-Score; A: Token Accuracy.

Benchmark	Task Category	AutoMoT†	AutoMoT‡
LingoQA (L)	Scene Understanding	67.00	**67.20**
OmniDrive (G)	Counterfactual Planning	18.20	**67.80**
ScienceQA (A)	General Knowledge	**88.60**	87.80
FigureQA (A)	General Knowledge	**97.60**	91.20
TallyQA (A)	General Knowledge	**81.40**	52.40
InfographicVQA (G)	General Knowledge	**89.30**	50.20
VizWiz (G)	General Knowledge	**75.60**	50.20

AutoMoT: A Unified Vision-Language-Action Model with Asynchronous Mixture-of-Transformers for End-to-End Autonomous Driving

Paper | Code | Dataset

GitHub

Hugging Face

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
HARVARD UNIVERSITY

ICML International Conference On Machine Learning