Highlight !

CVPR Nashville JUNE 11-15, 2025

DriveGPT4-V2: Harnessing Large Language Model Capabilities for Enhanced Closed-Loop Autonomous Driving

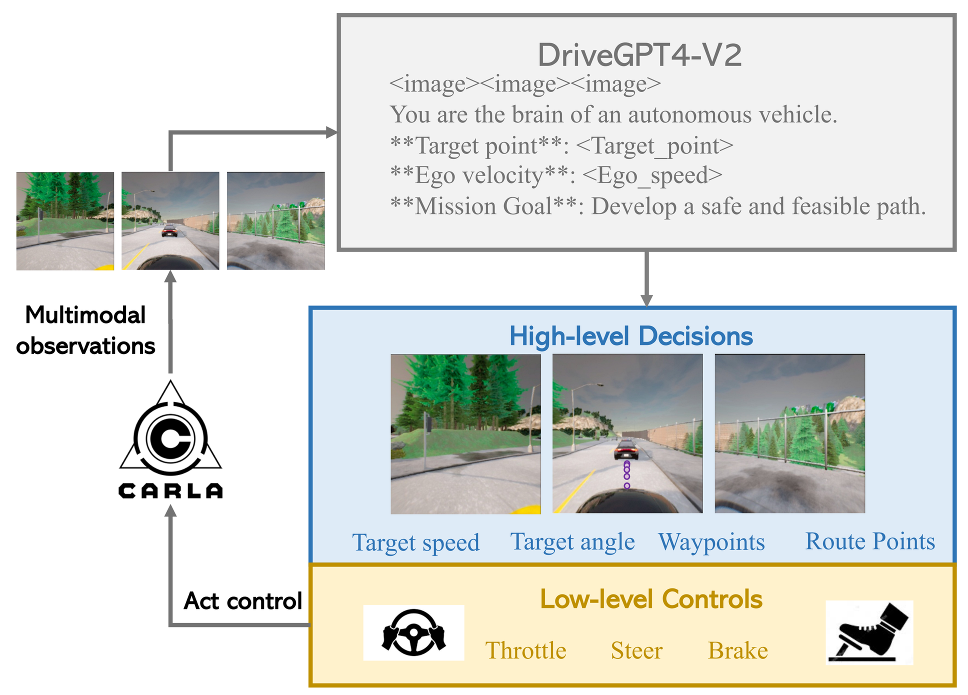
Zhenhua Xu[1,2] Yan Bai[3] Yujia Zhang[1] Zhuoling Li[1] Fei Xia[3] Kwan-Yee K. Wong[1] Jianqiang Wang[2] Hengshuang Zhao[1*]

[1]The University of Hong Kong [2]Tsinghua University [3]Meituan

Motivation

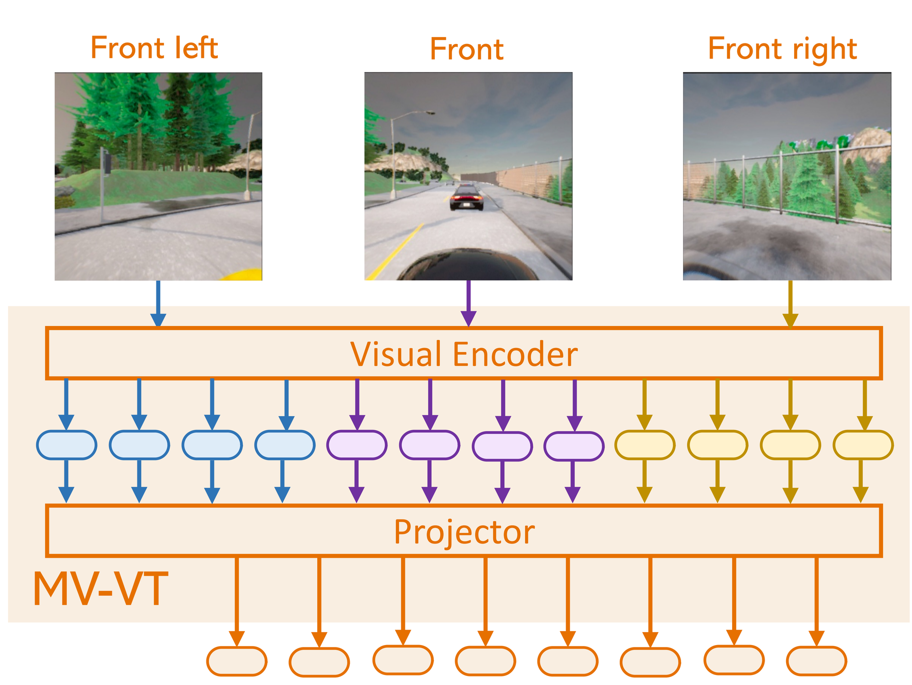
- From a learning theory perspective, end-to-end autonomous driving can optimize the entire system on the final outputs, rather than through the isolated optimization of individual modules, which potentially improves overall performance.
- Given the versatility of multimodal LLMs, they have been applied to autonomous driving for tasks such as interpretability and vehicle control.
- Open-loop evaluation is not sufficient for real-world applications. Thus, MLLMs need to be designed and evaluated specifically for closed-loop autonomous driving scenarios.

DriveGPT4-V2



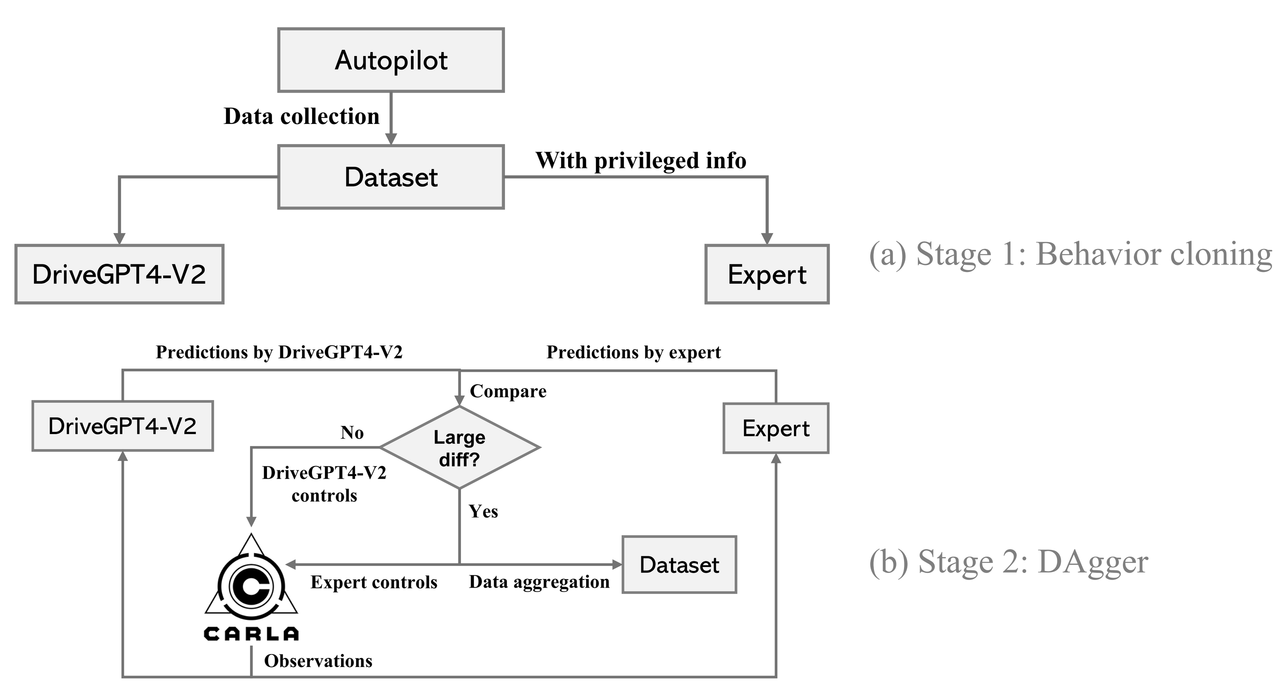
DriveGPT4-V2 for closed-loop autonomous driving. Taken as input multi-view camera images and vehicle state information, DriveGPT4-V2 predicts high-level vehicle decisions and converts them to low-level vehicle control signals in an end-to-end manner. DriveGPT4-V2 presents outstanding effectiveness and efficiency, serving as a reliable baseline method for future research on autonomous driving with LLMs.

MV-VT



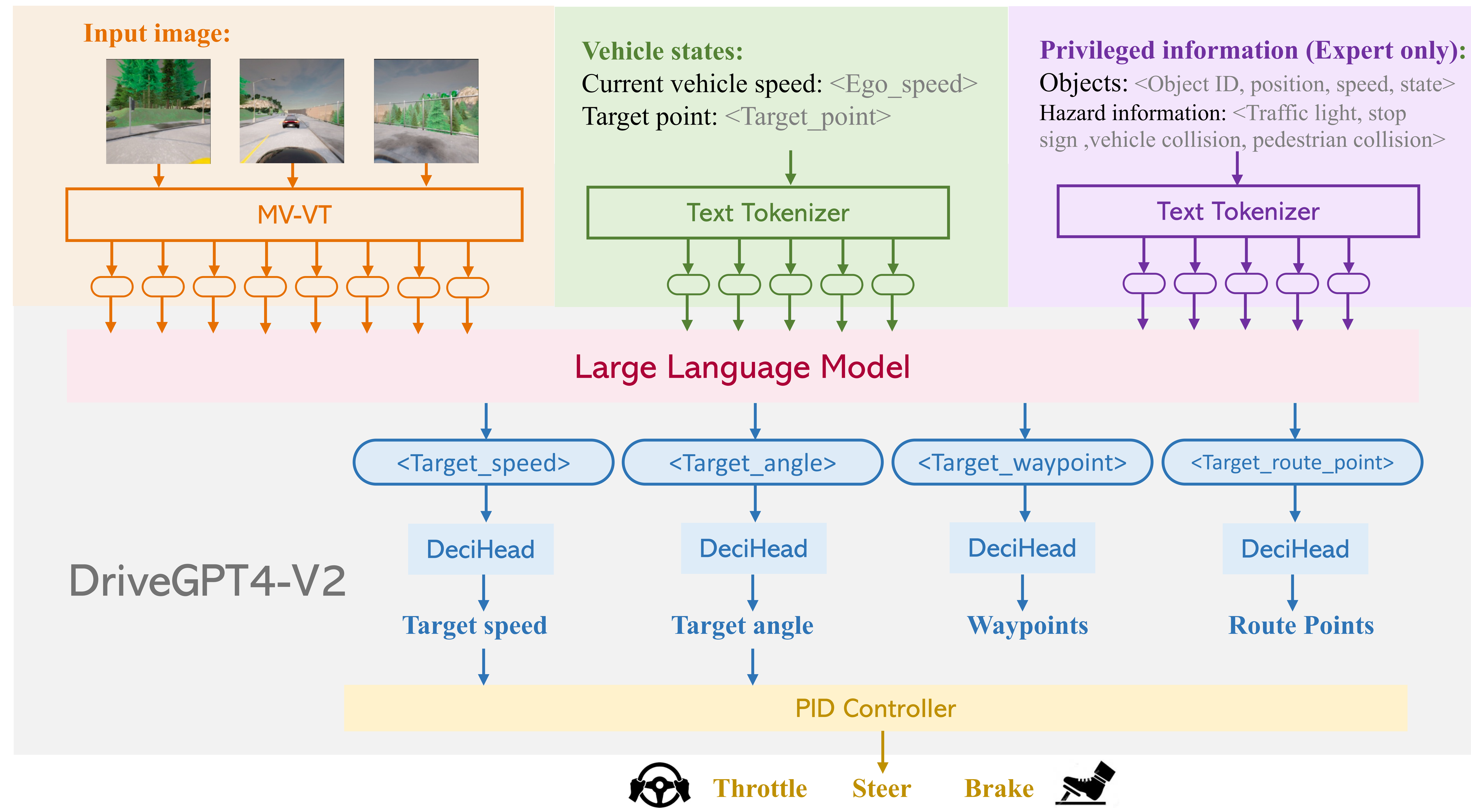
Multi-view visual tokenizer (MV-VT) structure. The input images consist of three front views. Each patch is processed through a visual encoder to extract features. Finally, a trained projection layer maps the downsampled feature into the text domain for further processing.

Two-stage Imitation Learning



(a) In the first stage, both DriveGPT4-V2 and the expert LLM are trained on data collected by a rule-based autopilot. (b) In the second stage, DriveGPT4-V2 runs on the training scenarios and routes. When the discrepancy between DriveGPT4-V2's predictions and those of the expert exceeds a predefined threshold, the expert's predictions are used to control the vehicle. Data from these cases is then added to the dataset for data aggregation.

Framework



- DriveGPT4-V2 takes multimodal input data to generate numerical control signals for end-to-end vehicle driving. The input includes multi-view images and vehicle state information. The LLM expert model, which shares a similar structure to DriveGPT4-V2, has access to privileged information about surroundings (shown in the purple module). The expert provides on-policy supervision to DriveGPT4-V2 to enhance closed-loop performance.

Experiments

Tab.4. Ablation studies on query scheme.

Tab.1. Comparison results on CARLA Longest6 benchmark.

Method	Visual	DS ↑	RC ↑	IS ↑	Ped ↓	Veh ↓	Stat ↓	Red ↓	Dev ↓	TO ↓	Block ↓
WOR [6]	C	21	48	0.56	0.18	1.05	0.37	1.28	0.88	0.08	0.20
LAV v1 [4]	C&L	33	70	0.51	0.16	0.83	0.15	0.96	0.06	0.12	0.45
Interfuser [42]	C	47	74	0.63	0.06	1.14	0.11	0.24	**0.00**	0.52	**0.06**
TransFuser [12]	C&L	47	**93**	0.50	0.03	2.45	0.07	0.16	**0.00**	**0.06**	0.10
LAV v2 [4]	C&L	58	83	0.68	**0.00**	0.69	0.15	0.23	0.08	0.32	0.11
Perception PlanT [38]	C&L	58	88	0.65	0.07	0.97	0.11	0.09	**0.00**	0.13	0.13
Transfuser++* [21]	C&L	65	90	0.72	**0.00**	0.99	**0.01**	0.07	**0.00**	0.10	0.12
Transfuser++*† [21]	C&L	58	89	0.65	0.01	1.15	**0.01**	0.10	**0.00**	0.14	0.13
LMDrive* [43]	C&L	36	69	0.52	0.07	1.03	0.18	1.01	0.09	0.11	0.22
DriveGPT4-V2	C	**70**	91	**0.77**	**0.00**	**0.80**	**0.01**	**0.04**	**0.00**	0.07	0.09

Tab.2. Efficiency analysis.

LLM	DS	Train	FPS
LLaVA-LLaMA3.1-8B	65	11.2h/epoch	0.4
TinyLLaVA-LLaMA-1.5B	63	3.0h/epoch	2.9
LLaVA-Qwen-0.5B	63	1.3h/epoch	8.1

Tab.3. Ablation studies on decision heads. "Additional tokens" indicates using more output tokens for prediction.

	DS	RC	IS	FPS
Additional tokens	64	91	0.70	1.4
DriveGPT4-V2	63	90	0.70	8.1

Tab.4. Ablation studies on PID controllers. "WP" indicates utilizing predicted waypoints for PID control; while "TS&RP" means PID control by predicted target speed and route points.

PID Controller	DS	RC	IS
WP	53	85	0.62
TS & RP	59	88	0.67
DriveGPT4-V2	63	90	0.70

Tab.5. Ablation studies on system design. Ablation studies of DriveGPT4-V2. "WP" and "RP" represent waypoints and route points, respectively.

	DS	RC	IS
Baseline	47	78	0.60
+ LLM Visual Pretraining	56	87	0.64
+ Visual Tokenizer	60	88	0.68
+ WP&RP	63	90	0.70
+ Expert Supervision	**70**	**91**	**0.77**