End-to-End Autonomous Driving Using Vision Language Model

Team ZERON
ZERON
Shanghai & Taicang, China
panqu.wang@zeron.ai

Abstract

*End-to-end autonomous driving has drawn tremendous attention recently. Many works focus on using modular deep neural networks to construct the end-to-end architecture. However, whether using powerful large language models (LLM), especially multi-modality Vision Language Models (VLM) could benefit the end-to-end driving tasks remain a question. In our work, we demonstrate that combining end-to-end architectural design and knowledgeable VLMs yield impressive performance on the driving tasks. It is worth noting that our method only uses a **single** camera and is the best camera-only solution across the leaderboard [1], demonstrating the effectiveness of vision-based driving approach and the potential for end-to-end driving tasks.* [2]

1. Introduction & Related Work

Autonomous driving technology has been evolving fast. Mainstream architecture involves dividing the whole pipeline into several different functions, such as perception, localization & mapping, prediction, planning, and control. This process has been considered to be a decent tradeoff between performance, safety, and explainability. However, while driver assist system (L2 to L3) starts to occupy a large market share, high-level (L4+) autonomous driving still lags behind from massive production. There are several reasons behind it:

Complex architecture: The current mainstream solution has twenty-ish or even more modules. Due to computing power limitations, the performance ceiling of a single module is not high; there are too many internal interfaces in the system, and transmission and optimization are difficult; local and overall optimization goals sometimes conflict so the performance improvements is hard to observe.

High cost: R&D/maintenance/manpower costs soar as the number of modules increases. Repeated reinvention, ineffective oraganization and integration problems often occur. Overall project coordination, iteration and management are usually difficult.

Poor generalization: The under-performed system usually incurs so called long-tail problems. If not properly handled, or as time elapses, the overwhelming number of handcrafted solutions could result in poor maintainability and scalability.

Difficult to Commercialize: Current high-level autonomous driving products could only be used in limited scenarios such as small number of cities/demonstration areas/highways. The algorithm is generally deeply associated with software and hardware platforms, and it is also difficult to be make compatible version with more vehicle models/platforms/scenarios.

Compared with the traditional architecture, end-to-end autonomous driving generally has the following advantages:

Simplified Framework: The entire system is encapsulated within a single module, leveraging an end-to-end deep neural network for high integration level. We could also share the same framework for both online and offline environments, enhancing development efficiency and maintenance friendliness.

Cost Efficiency: End-to-end solution achieves a large amount of cost reduction and efficiency enhancement. It lowers costs associated with algorithm development, testing, deployment, module iteration, data feedback loops, and project management. It accelerates product deployment and customer service effectiveness.

Strong Generalization: End-to-end autonomous driving combines data-driven and knowledge-driven approaches for a comprehensive understanding of the world. It seamlessly tackles long-tail problems, yielding more robust model outputs and facilitating easier customization of requirements, thereby enhancing the efficiency of product iterations.

Production-Friendly: It achieves excellent performance on many benchmarks. A single adaptable architecture swiftly accommodates diverse scenarios, vehicle mod-

[1] https://opendrivelab.com/challenge2024/#end_to_end_driving_at_scale

[2] For questions and inquiries, please direct to panqu.wang@zeron.ai

els, and platforms, minimizing marginal costs and enabling advanced autonomous driving technology to become widely accessible to the society.

There has been many innovative and influential works on end-to-end autonomous driving. A main line of literature [4, 9, 10, 24] employs modularized deep neural networks to learn effective intermediate features (sometimes including detection/tracking/mapping/prediction/scene understanding sub-tasks) to generate the planning trajectory and/or control signal at the end. Meanwhile, considerable number of works [5, 13–15, 17, 19, 20, 22] introduce large language models (LLMs) into the end-to-end modeling framework and showcase excellent performance on driving tasks, even including Q&A problems. Recently, generative modeling on autonomous driving tasks has also drawn great attention and yields impressive effects [5, 7, 8, 11, 12, 21, 23, 25]. In our work, we absorb the aforementioned works and designing an end-to-end multi-modality language model based solution for autonomous driving. Remarkably, our whole system only uses the input from a single camera and still achieves top performance on the leaderboard.

2. Method & Result

Figure 1 shows the architecture of our network. As an end-to-end model, our network directly receives information from raw sensor (camera in our case), indicating "what we have seen". It also obtains the ego history information to understand "what have we done". The navigation signal explains "where to go next". Finally the flexible text prompt conveys "what to do in this task". All inputs are followed by the corresponding encoder, either vision transformer [6], standard MLP, or text tokenizers, which are used in the decoding stage as well. For the LLMs, any open-source models such as [1, 3, 18] should work, depending on the computational resources. We use 7B model throughout our experiments. For the output, we generate the required trajectory output and optional text output, as driving language datasets such as [16] also provides valuable training data for the network to understand the environment. We train the network for multiple epochs on the provided nuPlan [2] dataset.

Although the design is simple, it resembles how human drives vehicle in nature: By using eyes, receiving navigation and goal information, human can plan the trajectory in next few seconds and infer why should we execute the trajectory simultaneously. We believe modeling driving like human is a promising research track, and may yield fruitful results in the future.

We validate our model on the CVPR 2024 End-to-End Driving at Scale Challenge track. We achieve the final score of 0.8747, and is the best camera-only solution across the leaderboard. Some visualizations could be seen in Figure 2. We can see that the model handles driving scenarios very well, such as going straight, left/right turn, waiting traffic light, stop sign, and even corner cases like toll booth. The good performance and potential strong generalization power suggests that the model does have higher performance limit and points out promising approach for autonomous driving production.

Due to time and resource limitation, we did not conduct comprehensive ablation studies. However, this in turn demonstrate that the proposed end-to-end architecture naturally fits the autonomous driving tasks and have huge potential of achieving even better performance, given more data, computational resource, and sensory input. We plan to explore these directions in the future.

References

[1] Vicuna: An open-source chatbot impressing gpt-4 with 90 *https://vicuna.lmsys.org/*, 2023. 2

[2] Holger Caesar, Juraj Kabzan, Kok Seang Tan, Whye Kit Fong, Eric Wolff, Alex Lang, Luke Fletcher, Oscar Beijbom, and Sammy Omari. nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. *arXiv preprint arXiv:2106.11810*, 2021. 2

[3] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024. 2

[4] Shaoyu Chen, Bo Jiang, Hao Gao, Bencheng Liao, Qing Xu, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Vadv2: End-to-end vectorized autonomous driving via probabilistic planning. *arXiv preprint arXiv:2402.13243*, 2024. 2

[5] Xinpeng Ding, Jinahua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. *arXiv preprint arXiv:2401.00988*, 2024. 2

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[7] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*, 2024. 2

[8] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 2

[9] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *European Conference on Computer Vision*, pages 533–549. Springer, 2022. 2

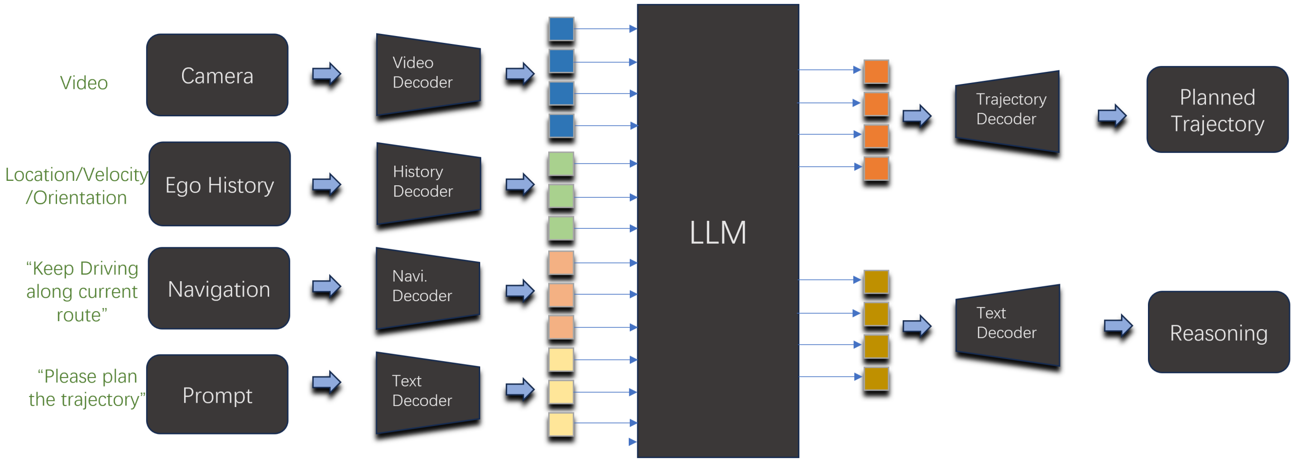

Figure 1. **The architecture of our network.** Our network receives input from camera, ego history, navigation signal, and text prompt. Through various encoder and the LLM module, our network generates trajectories and texts.

Figure 2. **Visualization of the results of our method.**

[10] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 2

[11] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. *arXiv preprint arXiv:2310.07771*, 2023. 2

[12] Jiachen Lu, Ze Huang, Jiahui Zhang, Zeyu Yang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. *arXiv preprint arXiv:2312.02934*, 2023. 2

[13] Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. 2

[14] Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving. *arXiv*

preprint arXiv:2311.10813, 2023.

[15] Hao Shao, Yuxuan Hu, Letian Wang, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. *arXiv preprint arXiv:2312.07488*, 2023. 2

[16] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023. 2

[17] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Chenxu Hu, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024. 2

[18] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[19] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *arXiv preprint arXiv:2405.01533*, 2024. 2

[20] Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, et al. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023. 2

[21] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. *arXiv preprint arXiv:2311.17918*, 2023. 2

[22] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kenneth KY Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *arXiv preprint arXiv:2310.01412*, 2023. 2

[23] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Generalized predictive model for autonomous driving. *arXiv preprint arXiv:2403.09630*, 2024. 2

[24] Yunpeng Zhang, Deheng Qian, Ding Li, Yifeng Pan, Yong Chen, Zhenbao Liang, Zhiyao Zhang, Shurui Zhang, Hongxu Li, Maolei Fu, et al. Graphad: Interaction scene graph for end-to-end autonomous driving. *arXiv preprint arXiv:2403.19098*, 2024. 2

[25] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv preprint arXiv:2403.06845*, 2024. 2