Distillation of Vision Language Models for Enhancing End-to-End Autonomous Driving

Feng Tao, Abhirup Mallik, Chenbin Pan, Xin Ye, Yuliang Guo,
Burhaneddin Yaman, and Liu Ren

Bosch Research North America &
Bosch Center for Artificial Intelligence (BCAI)
`{feng.tao2, abhirup.mallik, chenbin.pan, xin.ye3, yuliang.guo2, burhaneddin.yaman, liu.ren}@us.bosch.com`

Abstract. In recent years, large vision and language models (VLMs) have been investigated in autonomous driving to address long-standing issues including, reasoning, generalization, and long-tail scenarios. However, efficient integration of VLMs into autonomous driving framework remains an open question. In this paper, we present VLP, a novel Vision-Language-Planning framework that exploits large vision language models to bridge the gap between linguistic understanding and autonomous driving. VLP is a training only approach that distills the power of VLMs into end-to-end modular autonomous driving by presenting a contrastive learning objective. Extensive experiments on both open loop and closed loop tasks verify the utility of VLP. In particular, VLP achieves state-of-the-art end-to-end planning performance on the nuScenes dataset by achieving 35.9% and 60.5% reduction in terms of average L2 error and collision rates, respectively, compared to the previous best method.

Keywords: Vision language planning · Foundation models · Closed-loop

1 Background

Autonomous driving systems aims safe motion planning through effective scene understanding and reasoning. Despite advancements in vision-based autonomous driving systems, these methods often struggle with reasoning, generalization, and handling long-tail scenarios, which limits their deployment in real-world environments. The emerging progress on multimodel large language models (MLLM) [2] have shown that common sense and reasoning capability of these models can help addressing the challenges in embodied AI domain. While most of these methods have primarily targeted the robotics domain, there has been limited work on utilizing embodied language models (LMs) for autonomous driving tasks [4, 8, 11–13]. Notably, DiLu [12] and GPT-Driver [8] introduce GPT-based driver agents for closed-loop simulation tasks. In [11], an open-loop driving commentator is proposed that combines vision and low-level driving actions with language to interpret and reason about driving behaviors. However, it still remains unclear how can these approaches be efficiently distilled and leveraged in enhancing the performance of modular end-to-end autonomous driving stacks.

To address these challenges, we propose a novel Vision Language Planning (VLP) framework that efficiently distills the power of vision language models into the autonomous driving through a contrastive learning objective [9].

Our VLP framework, illustrated in Figure 1, introduces two key components: the Agent-centric Learning Paradigm (ALP) and the Self-driving-car-centric Learning Paradigm (SLP). ALP enhances the local semantic representation and reasoning capabilities of the Bird's Eye View (BEV) feature map, which serves as the source memory in the driving system, by aligning it with human-like reasoning processes. SLP refines the planning process by aligning planning queries with the goals and status of the self-driving car, using the common-sense reasoning embedded in the language model to guide decision-making. Together, these components improve the system's ability to understand complex driving environments and make safer, more informed decisions.

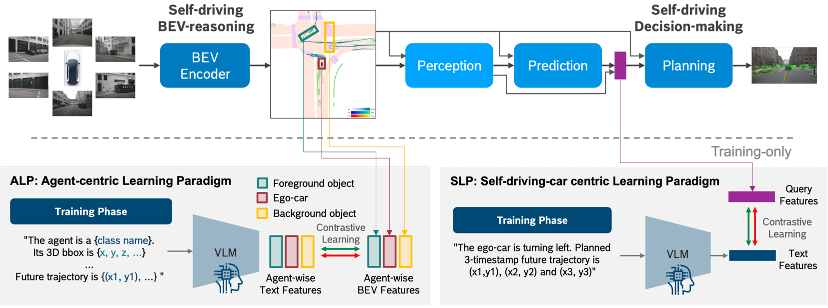
We conduct extensive experiments on both open loop and closed loop environments to show the efficacy of VLP in substantially improving performance of autonomous driving systems. Specifically, VLP achieves state-of-the-art end-to-end open loop planning performance on the nuScenes dataset by achieving 35.9% and 60.5% reduction in terms of average L2 error and collision rates, respectively, compared to the previous best method. Similarly it also outperforms the counterpart methods in CARLA closed loop evaluation.

2 Methodology

In this section, we will briefly describe our VLP methodology illustrated in Fig. 1. The VLP model consists of two key components: the Agent-centric Learning Paradigm (ALP) and the Self-driving-car-centric Learning Paradigm (SLP). These components focus on refining local details in the BEV (bird's-eye view) source memory and guiding the planning process of the self-driving car, respectively.

In particular, ALP first aligns the ground-truth area of each agent, namely ego-car, foreground and background objects, with the produced BEV map, and crop the regions of interest. We utilize the 3D bounding box to crop the ego-car and foreground (FG) object area, and panoptic scene mask to segment the lane area. Subsequently, we perform a pooling operation on the obtained local BEV region, to generate a single feature representation for the corresponding agent. After pooling, the local agent features in each sample along the batch are concatenated to formulate an Agent-wise BEV feature tensor. To ensure that local BEV features express the desired information, we conduct a BEV-expectation alignment process by leveraging LM and contrastive learning. We precisely define the perceptual information expected from the corresponding agent, such as agent labels, bounding boxes, and future trajectories. These driving-related ground-truth information, which should also be embedded in the local BEV feature is formulated into a prompt as illustrated in Fig. 1. The description is then passed to the VLM, to generate the corresponding agent expectation feature. We apply an MLP layer to adapt the expectation feature to the BEV feature space. Then, the agent expectation features are concatenated along the batch to generate an Agent-wise text feature tensor. Finally, we perform a constrastive learning loss [10] between Agent-wise BEV and Text features for alignment. In SLP, we follow a similar process but focus exclusively on the ego-vehicle. We note that VLP is only active during training, ensuring no additional parameters or computations are introduced during inference.

Fig. 1: The overview of proposed vision language planning (VLP) framework. VLP consists of ALP and SLP components which enhances autonomous driving from self-driving BEV-reasoning and self-driving decision-making aspects.



3 Experiment and Results

We conduct experiments on both open loop and closed loop environments. For open loop experiments, we use nuScenes dataset [1]. The nuScenes contains 1000 driving scenes from Boston and Singapore, two cities that are known for their dense traffic and highly challenging driving conditions. For closed-loop experiments, we use Bench2Drive [6], the first benchmark designed to evaluate the diverse capabilities of end-to-end autonomous driving systems in a closed-loop setting. Specifically, we assess the performance of our models within the Bench2Drive closed-loop evaluation environment, which is based on the CARLA Leaderboard V2 [3]. This environment extends the original 39 scenarios up to 44 more challenging scenarios and modifies the official routes by condensing them into shorter routes, each featuring a single scenario.

3.1 Open-loop Planning Performance

In Tab. 1, we present a series of comparative experiments that showcase the performance of our open-loop planning in comparison to the baseline UniAD [5] and VAD [7] models. As can be seen in rows 2 and 5 of the table, the integration of just SLP leads to noticeable reductions in both the L2 error and collision rates for all the baseline models. Moving down the table, rows 3 to 6 demonstrate that the inclusion of both VLP components (SLP and ALP together) consistently yields further improvements in these planning metrics. In particular, VLP-UniAD shows a 28.1% and 48.4% reduction in terms of average L2 error

ID	Model	SLP	ALP	L2 (m) ↓				Col. Rate (%) ↓			
				1s	2s	3s	Avg.	1s	2s	3s	Avg.
1	UniAD [5]			0.48	0.96	1.65	1.03	0.05	0.17	0.71	0.31
2	VLP-UniAD	✓		**0.43**	**0.86**	**1.47**	**0.92**	**0.03**	**0.15**	**0.48**	**0.22**
3	VLP-UniAD	✓	✓	**0.36**	**0.68**	**1.19**	**0.74**	**0.03**	**0.12**	**0.32**	**0.16**
4	VAD [7]			0.46	0.76	1.12	0.78	0.21	0.35	0.58	0.38
5	VLP-VAD	✓		**0.26**	**0.47**	**0.78**	**0.50**	**0.12**	**0.17**	**0.42**	**0.23**
6	VLP-VAD	✓	✓	**0.30**	**0.53**	**0.84**	**0.55**	**0.01**	**0.07**	**0.38**	**0.15**

Table 1: Open-loop planning performance. VLP achieves significant end-to-end planning performance improvement over counterpart vision only UniAD and VAD methods on the nuScenes validation dataset [1].

and collision rate, respectively, compared to baseline UniAD. Similarly, in comparison with VAD, VLP-VAD achieves 35.9% and 60.5% reduction for average L2 error and collision rate, respectively. These significant results underscore the effectiveness of both SLP and ALP, as well as their adaptability across various autonomous driving system configurations.

3.2 Closed-loop Benchmark

The closed-loop evaluation results are collected on the 110 routes (each around 150 meters in length and contains a single specific scenario) in Bench2Drive benchmark to showcase the closed-loop performance in comparison to the baseline VAD [7] tiny model. The results are shown in Table 2. It can be seen that the VLMs plugin approach in training phase can improve the closed-loop performance in both driving score (8%) and route completion (13%).

Method	Bench2Drive Evaluation	
	Driving Score ↑	Route Completion ↑
VAD-tiny	31.34	56.64%
VAD-tiny-VLP	**33.84**	**64.20%**

Table 2: Closed-loop simulation results on Bench2Drive (110 routes). VLP significantly outperforms VAD in terms of both driving score and route completion.

4 Discussion and Conclusion

To conclude, we have introduced a novel Vision Language Planning (VLP) approach to enhance autonomous driving systems. VLP employs a constrastive learning objective to distill power of vision language models into autonomous driving without incurring any extra cost in runtime. Through extensive experiments on both open-loop and closed-loop environments, we have shown that VLP delivers SOTA end-to-end planning performance.

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
2. Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.D., et al.: A survey on multimodal large language models for autonomous driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 958–979 (2024)
3. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: Carla: An open urban driving simulator. In: Conference on robot learning. pp. 1–16. PMLR (2017)
4. Hasan, M.Z., Chen, J., Wang, J., Joshi, A., Velipasalar, S., Hegde, C., Sharma, A., Sarkar, S.: Vision-language models can identify distracted driver behavior from naturalistic videos. arXiv preprint arXiv:2306.10159 (2023)
5. Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al.: Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17853–17862 (2023)
6. Jia, X., Yang, Z., Li, Q., Zhang, Z., Yan, J.: Bench2drive: Towards multi-ability benchmarking of closed-loop end-to-end autonomous driving. arXiv preprint arXiv:2406.03877 (2024)
7. Jiang, B., Chen, S., Xu, Q., Liao, B., Chen, J., Zhou, H., Zhang, Q., Liu, W., Huang, C., Wang, X.: Vad: Vectorized scene representation for efficient autonomous driving. arXiv preprint arXiv:2303.12077 (2023)
8. Mao, J., Qian, Y., Zhao, H., Wang, Y.: Gpt-driver: Learning to drive with gpt. arXiv preprint arXiv:2310.01415 (2023)
9. Pan, C., Yaman, B., Nesti, T., Mallik, A., Allievi, A.G., Velipasalar, S., Ren, L.: Vlp: Vision language planning for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14760–14769 (2024)
10. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
11. Wayve: Lingo-1: Exploring natural language for autonomous driving. https://wayve.ai/thinking/lingo-natural-language-autonomous-driving/ (2023)
12. Wen, L., Fu, D., Li, X., Cai, X., Ma, T., Cai, P., Dou, M., Shi, B., He, L., Qiao, Y.: Dilu: A knowledge-driven approach to autonomous driving with large language models. arXiv preprint arXiv:2309.16292 (2023)
13. Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.K., Li, Z., Zhao, H.: Drivegpt4: Interpretable end-to-end autonomous driving via large language model. arXiv preprint arXiv:2310.01412 (2023)