VLDrive: Vision-Augmented Lightweight MLLMs for Efficient Language-grounded Autonomous Driving

Supplementary Material

1. Dataset Details

The official language-driven autonomous driving dataset [7] is utilized to train our model, which includes 64K instruction-following data clips collected across 8 towns on CARLA [2] simulation environment. Each clip consists of the following ingredients:

Multi-sensor input data: Multi-view RGB images from the front, rear, left, and right cameras, along with corresponding LiDAR data are provided in the dataset.

Navigation instructions: Each driving clip has one aligned navigation instruction, which guides the movement of the ego-car. Some examples of navigation instructions are exhibited in Table 1.

Table 1. Examples of navigation instructions.

Type	Instruction Examples
Start	Go ahead and start driving. Feel free to start driving.
Follow	At the next intersection, just keep heading straight, no turn. Maintain your course along this route. Continue driving straight on the designated highway.
Turn	Proceed ahead and make a left turn. Up ahead, just take a left. Next intersection, just swing a right.

2. Benchmark Details:

We assess the closed-loop driving performance of our method on the standard LangAuto benchmarks [7]. It is also established based on CARLA simulation but has several significant features that distinguish it from previous benchmarks like Town05 [5] and Longest6 [1]:

Language-guided driving. This benchmark leverages the navigation instruction in the natural language format to guide the model's driving, **without** providing any target points or action commands.

Various distances. This benchmark consists of three tracks with varying route lengths: 1) LangAuto features routes exceeding 500 meters. 2) LangAuto-short encompasses routes ranging from 150 to 500 meters and 3) LangAuto-tiny contains routes less than 150 meters.

Diverse environments. This benchmark spans 8 towns in CARLA and comprises 16 diverse environments, derived from combinations of 7 distinct weather conditions and 3 different daylight settings.

3. Implementation Details

3.1. Model Details:

Visual Encoder: Given a sequence of visual data, where each frame includes multi-view camera images and corresponding LiDAR data, we adopt the pretrained visual encoder in LMDrive [7] to integrate multi-view RGB images and LiDAR data and produce a unified feature representation $\mathbf{F}_i \in \mathbb{R}^{N \times C}$ for each frame, with N representing the total number of tokens. Specifically, the visual encoder employs a 2D ResNet [3] to extract image features from each view, which are then fused using a Transformer encoder [8] for multi-view feature integration. On the other hand, PointPillars [4] followed by PointNet [6] are utilized to convert raw LiDAR data to BEV features. Afterwards, a Transformer decoder is adopted to integrate the multi-view image features into BEV features and two kinds of learnable queries, generating BEV tokens, waypoint tokens and traffic light tokens, respectively. In LMDrive [7], three downstream tasks, including object detection, future waypoint prediction, and traffic light status classification are introduced to pre-train the visual encoder. In our work, the generated queries $\mathbf{F}_i \in \mathbb{R}^{N \times C}$ ($N = 106$) from the pre-trained visual encoder, composed of 100 BEV tokens, 5 waypoint tokens and 1 traffic light tokens are delivered to the subsequent connector module.

3.2. Experimental Details:

Training: Our proposed model is trained on 8 $\times$ A100 (40G) NVIDIA GPUs. An AdamW optimizer cooperated with a cosine learning rate scheduler is adopted to train our model. The initial learning rate is set to $1e^{-5}$, with a weight decay of 0.06, and the total training epoch is 15. We maintain a fixed sampling interval of 2 frames during the training process. *Evaluation*: We conduct the closed-loop driving evaluation using version 0.9.10.1 of the CARLA simulator.

References

[1] Kashyap Chitta, Aditya Prakash, Bernhard Jaeger, Zehao Yu, Katrin Renz, and Andreas Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(11):12878–12895, 2022. 1

[2] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[4] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 1

[5] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7077–7087, 2021. 1

[6] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018. 1

[7] 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. 1

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1