Iterative Finetuning VLM with Retrieval-augmented Synthetic Datasets
Technical Reports for W-CODA Challenge Track-1

Team OpenDriver
Peking University
opendriver@gmail.com

Abstract

Large Vision-Language Models (LVLMs) play a crucial role in autonomous driving, offering advanced visual reasoning capabilities that enhance system interpretability. However, these models often struggle with corner cases in open-world environments, leading to degraded performance. This paper addresses two key challenges: the limitations of pre-trained vision encoders in recognizing unfamiliar objects and the insufficient reasoning abilities of existing models. We propose a solution that leverages synthetic datasets and iterative finetuning to enhance model performance. Our approach improves the model's visual knowledge and reasoning capabilities, resulting in substantial performance gains on the CODA-LLM benchmark, including a 1.82x increase in generation perception, a 97.34% improvement in region perception, and a 2.09x enhancement in driving suggestion accuracy. These results demonstrate the effectiveness of our method in improving LVLMs for open-world autonomous driving scenarios. The model can be experienced on https://pku-open-driver.github.io/.

1. Introduction

Large Vision-Language Models (LVLMs) have garnered significant attention in the autonomous driving domain and embodied agents due to their remarkable ability to perform visual reasoning, enabling them to comprehend complex images and videos [3, 9, 11, 15, 16]. These models have substantially contributed to developing interpretable end-to-end autonomous driving systems or embodied agents [1, 17]. However, while current general-purpose models perform well in common driving scenarios, they often struggle in the real-world open environment, which contains numerous corner cases where the performance of these models tends to degrade.

Two primary challenges arise when applying existing open-source Vision-Language Models (VLMs) to corner cases in driving scenarios. The first challenge is related to the vision encoder typically used in these models, often pre-trained on CLIP and further aligned with language through open-source datasets. When the vision encoder lacks relevant visual knowledge—particularly regarding unfamiliar objects in corner cases—the VLMs struggle to accurately interpret the elements within the driving scene, leading to suboptimal decision-making [12].

The second challenge involves the models' limited language concept and reasoning capabilities. VLMs often exhibit insufficient understanding of specific concepts and rules pertinent to driving scenarios, making it difficult for them to accurately assess whether the scenes depicted in images comply with these rules. Furthermore, effective decision-making in various corner cases requires strong reasoning abilities, often necessitating a step-by-step reasoning process that many existing VLMs lack [18].

To address these challenges, we propose enhancing open-source foundation Vision-Language Models for open-world driving scenarios through the use of synthetic datasets and interactive self-finetuning. For the knowledge enhancement component, we introduce a synthetic data generation method utilizing retrieval-augmented generation [6] and self-instruct [14] techniques. For the reasoning component, we employ iterative finetuning to improve the model's advanced reasoning and decision-making capabilities.

The models trained using our proposed methods demonstrate significant performance improvements compared to the baseline on the CODA-LLM benchmark Track-1 [8], with a 1.82x times increase in generation perception, a 97.34% improvement in region perception, and a 2.09x times enhancement in driving suggestion scores.

2. Method

In this section, we will explain how we constructed our dataset and the training techniques we used.

2.1. Retrieval-augmented Synthetic Datasets for Open-ended Corner Cases

Given the presence of numerous corner cases in open-domain autonomous driving scenarios, a substantial amount

of training data is required. Existing approaches typically rely on manual annotation, which is not only prohibitively expensive but also challenging to scale. To enhance the understanding of driving datasets by open-source Vision-Language Models (VLMs), we collected a large volume of unlabeled driving scene images from open-world scenarios and generated the Open-World Driving Synthetic Datasets ($\mathcal{D}^{\text{synth}}$) through data synthesis.

The synthetic dataset primarily consists of two components: Visual Captioning and Visual Question Answering (VQA) tasks. The visual captioning component is responsible for describing the visual targets and relevant knowledge within open-world scenarios, while the VQA component addresses understanding and reasoning about the scenes.

To efficiently generate this synthetic dataset, we leveraged existing foundation models to automate and accelerate the dataset construction process. These models include the Segment Anything Model (SAM) for visual segmentation, the Open-CLIP model for open-domain classification, the ShareCaptioner model for caption generation, and the large language model ChatGPT. The specific process is outlined as follows:

1. **Preprocessing with Small Models**: We began by employing small models trained on limited autonomous driving datasets (such as object detection and image recognition models) to preprocess the images I, identifying potential object names (from the recognition model) $T_{\text{obj}} = \mathcal{M}_{\text{cls}}(I)$ and object bounding boxes (from the detection model) $T_{\text{obj, pos}} = \mathcal{M}_{\text{det}}(I)$.
2. **Segmentation and Embedding Retrieval**: The dataset was further processed using the Foundation Segmentation Model (Segment Anything, SAM [5]) to obtain segmented, unlabeled pixel regions $I_{\text{sam}} = \mathcal{M}_{\text{sam}}(I)$. We then utilized the embeddings of these pixel regions to perform retrieval in a visual-text database, identifying the pixel locations of each object within the image $T_{\text{obj}} = \mathcal{M}_{\text{retrieval}}(\text{emb}_{I_{\text{sam}}}, \text{emb}_{I_{\text{database}}})$.
3. **Textual Retrieval**: Based on the object names T_{obj} identified in the images, we performed searches using open search engines (e.g., Google and Baidu) to retrieve relevant textual descriptions T_{desc} of these objects in Wiki pages.
4. **Caption Generation**: We utilized the ShareCaptioner model to generate detailed raw captions $T_{\text{cap}}^{\text{raw}}$ for the images I. Similar to the LLaVA [7] approach, we refined these captions $T_{\text{cap}}^{\text{raw}}$ using the retrieved textual descriptions T_{desc} to produce fine-grained captions $T_{\text{cap}}^{\text{fine}}$.
5. **Question-Answer Pair Generation**: Finally, using the generated captions and textual descriptions, we employed the self-instruct [14] method, wherein the language model ChatGPT proposed questions $T_Q(I)$ and answered questions $T_A(I)$ related to the image content.

Through this approach, we significantly augmented the existing dataset, resulting in the creation of `CODA-VQA-200k`, a synthetic dataset that is ten times the size of the original CODA dataset.

2.2. Iterative Finetuning enables Self-improving Vision-Language Models

To continuously enhance the model's performance, we employed an iterative finetuning approach for Vision-Language Model (VLM) training and improvement.

Self-play Finetuning Instead of directly applying supervised finetuning (SFT) to optimize the model, we utilized Direct Preference Optimization (DPO) [10] combined with self-play finetuning (SPIN) [4] to iteratively refine the model. Initially, we performed SFT on the CODA-LM-38k dataset and the synthetic CODA-LM-200k dataset, yielding the first version of the model.

Subsequently, in the self-play finetuning phase, the content generated by the model itself was treated as negative samples, while the samples from the dataset were considered positive samples. The model was then optimized through multiple rounds of self-play finetuning. Although this approach aids in improving model performance, its capacity to enhance the model's generalization in open-ended environments is limited, as the training dataset remains static. Consequently, the model's ability to adapt to new, unseen scenarios is constrained.

Iterative Synthetic Data generation To overcome the limitations of static datasets, we further employed an iterative synthetic data generation approach to continually augment the training dataset. In this process, the content originally generated by ShareCaptioner [2] was replaced by content generated by Model^i, the current iteration of the model. Using the same methodology, additional synthetic data was created, including enhanced answers, which were then used as positive samples for the next round of model training.

Given that the challenges in the benchmark primarily involve general perception, region perception, and driving suggestion, our approach targeted these areas specifically. General and region perception tasks were addressed by improving the vision encoder using the same iterative finetuning strategy. For driving suggestion, which heavily relies on the model's reasoning abilities, we further enhanced reasoning capabilities by incorporating GPT-4 to augment the synthetic dataset with more complex inferential tasks, thereby helping to improve the reasoning skills of the VLM.

The final model underwent $N = 5$ rounds of training, progressively refining its capabilities in each iteration.

Table 1. Comparative results of different models, training methods, and datasets.

Model	Direct	SFT	Ours(w/ 30k)	Ours(w/ 200k)
LLAVA [7]	30.33	40.24	49.79	54.13
CogVLM2 [13]	39.43	50.43	57.99	60.84
MiniCPM [19]	42.67	60.90	63.12	71.87

3. Experiments

3.1. Implementation Details

Model architectures and Datasets For the model training, we utilized LLaVA-Next-8B [7], MiniCPM-8B [19], and CogVLM2-19B [13] as base Vision-Language-Models. The training datasets included the original image data from CODA-LM, comprising 29,681 QA pairs, along with an additional 200k synthetic QA pairs.

Training Parameters For model training, we used the following configuration: The training process was conducted with the use of bfloat16 (bf16) precision, ensuring full evaluation without these precision settings. The maximum input length for the model was set to 2048 tokens, and the number of slices was capped at 9. The training process was conducted over 5 epochs. The per-device batch size was set to 4 for training and 1 for evaluation, with gradient accumulation steps set to 1. The learning rate was set to 1e-6, with a weight decay of 0.1, and the beta2 parameter of the Adam optimizer was set to 0.95. We employed a warmup ratio of 0.01, and the learning rate was scheduled using a cosine decay strategy. And the training was conducted using a DeepSpeed S2 configuration.

3.2. Evaluation Methods

To evaluate the performance of our models, we sampled 500 data samples from the training datasets to create an evaluation dataset. We employed the LLM-as-Judge approach to assess the quality of the generation content produced by different methods. Specifically, for each image, the corresponding question, reference answer, and the answers generated by Model 1 and Model 2 were presented to ChatGPT. ChatGPT then determined which model produced the better answer. Using this method, we calculated the win rates and Elo rating scores to compare the performance of different models [20].

3.3. Results

The final experimental results, as summarized in Table 1, demonstrate the performance of three different base models—LLAVA, CogVLM2, and MiniCPM—when trained using different methodologies: Direct, Supervised Fine-Tuning (SFT), and our proposed method.

Impact of Synthetic Datasets on Model Performance The LLAVA model shows a significant improvement when moving from SFT (40.24) to our method with the original 30k dataset (49.79). This improvement is further enhanced when using our synthetic 200k dataset, achieving a score of 54.13. This illustrates that the synthetic data greatly enhances the model's ability to generalize and understand complex scenarios, resulting in a 34% improvement over the SFT method. Similar trends are observed with CogVLM2. The performance increases from 50.43 with SFT to 57.99 with the 30k dataset. When the model is trained with the 200k synthetic dataset, the performance further rises to 60.84, showing a 20% increase compared to SFT. This demonstrates the robustness of synthetic data in improving the model's adaptability to diverse scenarios. The MiniCPM model benefits the most from the synthetic datasets. Starting from an SFT score of 60.90, the model's performance reaches 63.12 with the 30k dataset, and further to an impressive 71.87 with the 200k synthetic dataset. This substantial increase (18% over SFT) highlights the model's ability to leverage larger, more diverse datasets for better decision-making in complex environments.

Effectiveness of Training Methods As expected, SFT provides a considerable improvement over the Direct method across all models, confirming the value of supervised fine-tuning in refining model performance. Training with our method using the original 30k dataset already outperforms SFT, indicating the effectiveness of our iterative fine-tuning strategy even with a relatively smaller dataset. The jump in performance from 30k to 200k synthetic datasets in our method across all models demonstrates the significant role that synthetic data plays in enhancing model capabilities. The larger dataset allows for more comprehensive learning and better generalization, particularly in open-ended scenarios.

We present the final iterations of our finetuned model from MiniCPM-8B base VLM and our synthetic `CODA-LLM-200k` datasets. The overall scores for general perception, regional perception, and driving suggestions are 54.41, 83.01, and 71.76, respectively.

4. Conclusions

In this paper, we present a novel approach to improving Vision-Language Models for autonomous driving by combining iterative fine-tuning with large-scale synthetic datasets. Our method significantly enhances model performance in handling complex, open-world driving scenarios. The results highlight the effectiveness of synthetic data in boosting model generalization, offering a robust solution for safer and more reliable autonomous driving systems.

References

[1] Shaofei Cai, Bowei Zhang, Zihao Wang, Xiaojian Ma, Anji Liu, and Yitao Liang. Groot: Learning to follow instructions by watching gameplay videos. *arXiv preprint arXiv:2310.08235*, 2023. 1

[2] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv: 2311.12793*, 2023. 2

[3] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1

[4] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv: 2401.01335*, 2024. 2

[5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[6] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 1

[7] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, 2024. 2, 3

[8] Yanze Li, Wenhua Zhang, Kai Chen, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated evaluation of large vision-language models on self-driving corner cases. *arXiv preprint arXiv:2404.10595*, 2024. 1

[9] Mingyu Liu, Ekim Yurtsever, Xingcheng Zhou, Jonathan Fossaert, Yuning Cui, Bare Luka Zagar, and Alois C Knoll. A survey on autonomous driving datasets: Data statistic, annotation, and outlook. *arXiv preprint arXiv:2401.01454*, 2024. 1

[10] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv: 2305.18290*, 2023. 2

[11] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. Lmdrive: Closed-loop end-to-end driving with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130, 2024. 1

[12] 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. 1

[13] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023. 3

[14] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions, 2022. 1, 2

[15] Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, et al. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *arXiv preprint arXiv:2311.05997*, 2023. 1

[16] Zihao Wang, Shaofei Cai, Guanzhou Chen, Anji Liu, Xiaojian Shawn Ma, and Yitao Liang. Describe, explain, plan and select: interactive planning with llms enables open-world multi-task agents. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[17] Zihao Wang, Shaofei Cai, Zhancun Mu, Haowei Lin, Ceyao Zhang, Xuejie Liu, Qing Li, Anji Liu, Xiaojian Ma, and Yitao Liang. Omnijarvis: Unified vision-language-action tokenization enables open-world instruction following agents. *arXiv preprint arXiv:2407.00114*, 2024. 1

[18] Yang Wu, Shilong Wang, Hao Yang, Tian Zheng, Hongbo Zhang, Yanyan Zhao, and Bing Qin. An early evaluation of gpt-4v (ision). *arXiv preprint arXiv:2310.16534*, 2023. 1

[19] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 3

[20] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 3