VLAAD: Vision and Language Assistant for Autonomous Driving

SungYeon Park[1], MinJae Lee[1], JiHyuk Kang[1],
Hahyeon Choi[1], Yoonah Park[1], Juhwan Cho[1], Adam Lee[2], and DongKyu Kim[1]

[1]Seoul National University
[2]University of California, Berkeley
{psyeon990, herry123435, wlgur7238, gk0gus0, wisdomsword21, hj99cho, dongkyukim}@snu.ac.kr
{alee00}@berkeley.edu

Abstract

While interpretable decision-making is pivotal in autonomous driving, research integrating natural language models remains a relatively untapped. To address this, we introduce a multi-modal instruction tuning dataset that facilitates language models in learning visual instructions across diverse driving scenarios. This dataset encompasses three primary tasks: conversation, detailed description, and complex reasoning. Capitalizing on this dataset, we present a multi-modal LLM driving assistant named VLAAD. After fine-tuned from our instruction-following dataset, VLAAD demonstrates proficient interpretive capabilities across a spectrum of driving situations. We open our work, dataset, and model, to public on github. https://github.com/sungyeonparkk/vision-assistant-for-driving

1. Introduction

In recent years, the field of autonomous driving has witnessed rapid advancements in both academia and industry. While significant progress [3, 4, 35] has been made in learning the latent representations of driving data using deep neural models in an end-to-end manners for vehicle control, interpretability in autonomous vehicles remains an unresolved issue. Interpreting driving actions, particularly through natural language, has remained largely unaddressed.

These interpretable models not only facilitate effective interaction between humans and autonomous vehicles but also, more importantly, play a vital role in making the underlying causes of autonomous vehicle behaviors and decision-making processes readily comprehensible to humans. This understanding is essential for ensuring safety in various driving scenarios.

In this paper, we introduce the **V**ision-and-**L**anguage **A**ssistant for **A**utonomous **D**riving (VLAAD), which extends the concept of visual instruction-tuning [21] into the domain of autonomous driving. We focused on the scarcity of high-quality video captioning datasets in the autonomous driving domain and the near absence of conversation datasets capturing interactions between drivers and vehicles in driving scenarios. These datasets play a pivotal role in connecting visual modality and language within the framework of visual instruction-tuning. To address this gap, we created a high-quality 64K video instruction-following dataset that features complex reasoning, detailed descriptions, and conversation. We used GPT4 to generate the dataset from front-view RGB camera footage, and it is publicly accessible. For our research, we harnessed the Berkeley Deep Drive (BDD) dataset, its annotated counterpart, BDD-X [19], and the Honda Research Institute Driving Dataset (HDD) [26]. Our architectural approach involved connecting the Video-Qformer [39] encoder with the LLaMA-2-7B model, using off-the-shelf unimodal pre-trained models. We subsequently conducted end-to-end fine-tuning using this newly created high-quality 64K video instruction-following dataset.

In summary, our contributions can be summarized as follows: (1) Development of a comprehensive visual instruction tuning dataset, comprising 64K samples, with the specific aim of enhancing the interpretability of autonomous driving systems. (2) Open-sourcing of the complete codebase, which facilitates model fine-tuning on the LLaMA-2-7B [30] architecture using Video-Qformer. (3) Public release of VLAAD weights, making it readily accessible to the research and development community, thereby driving progress in interpretable autonomous driving technology. (4) In addition, we also provide the necessary code and prompts for generating the visual instruction tuning dataset based on captions for front-view RGB camera footage.

2. Related Works

2.1. Interpretable AVs via Natural Languages

In autonomous driving research, there has been a notable emphasis on enhancing explainability through the generation of visual explanations [16, 28, 31]. There is a growing focus on enhancing the interpretability of autonomous vehicles (AVs) using natural language recently. Research is underway to generate natural language captions for driving scenes [13, 20] in a human-understandable manner. [23] provides explanations by describing surrounding objects and predicting potential risks. Furthermore, There have been studies that aim to enable self-driving vehicles to comprehend and obey natural language instructions such as [8]. Some studies employ methods such as imitation learning [27] and the use of end-to-end controllers [17]. This serves as a foundation for achieving interactive autonomous vehicles through the use of natural language for action planning. Notably, [18] combines language captioning and planning, summarizing visual data into natural language descriptions and predicting actions. Similarly, [36] annotates driving actions and their associated natural language explanations, with a focus on prediction.

Datasets Various driving scene video datasets have been valuable resources in driving scene understanding [26, 38]. These datasets contain a substantial amount of video footage depicting driving scenarios. For interpretable AV, some derived datasets with detailed annotations and explanations have been created from these driving scene video datasets [19,36] .Additionally, DRAMA [24] offers insights into risk objects and driver suggestions, while nuScenes-QA [25] creates question-answer pairs based on 3D object relationships. Despite the growing trend of actively utilizing large language models like GPT-3.5 and GPT-4 to create instruction tuning datasets [21], research in the context of driving scenes remains limited, indicating the need for further exploration.

2.2. Visual Instruction Tuning

Visual Instruction Tuning on Specific Domain In recent developments within the realm of multimodal conversational AI, models like LLaVA [21], trained on numerous image-text pairs from the public web, have showcased significant advancements. However, while these vision-language models excel in general areas, they sometimes fall short in specialized fields. To address this gap, several initiatives have been introduced, with LLaVA-Med [5] being a notable example. This model offers a cost-effective strategy for adapting a vision-language model to a specific domain. At its core, the method relies on a comprehensive domain-specific figure-caption dataset taken from academic articles. Using GPT-4, LLaVA-Med derives open-ended instruction-based data from these captions and then refines a broad vision-language model through a refined curriculum learning approach..

Video Instruction Tuning Following the introduction of BLIP [11], there has been a surge of endeavors aimed at developing LLM models proficient in video understanding, primarily by leveraging the technique of synchronizing image vectors to embedding space of LLMs. BLIP2 [14] presents a Q-Former that connects image queries that have been learned to the text embedding realm of LLMs. Meanwhile, Instruct BLIP [34] introduces an instruction-aware Query Transformer, which derives features specific to the provided instruction. Adding to this, BLIVA [33] integrates query embeddings from InstructBLIP and also directly maps encoded patch embeddings into the LLM, a method inspired by LLaVA [21]. On a related note, Video-Chat [21] expands upon image encoders, setting itself apart from other models by emphasizing both spatial and time-based video characteristics, thereby allowing larger models to comprehend the visual elements in videos.

Another standout example of this approch is Video-LLaMA [39], which enhances cross-modal training using pre-existing visual & audio encoders in tandem with fixed LLMs. Consequently, Video-LLaMA showcases a marked superiority compared to other vision-LLMs, like MiniGPT-4 [9] and LLaVA [21], with a distinct prowess in grasping temporal variations within visual contexts. To integrate a pre-trained image encoder into the video encoder and introduce a video-to-text generation task to master video-language relationships, video-llama employ new strucuture Video Q-former.

2.3. Large Language Models for AVs

In recent years, there has been a rapid advancement in Large Language Models (LLMs), and this progress has spurred numerous attempts to apply these models to the domain of Autonomous Driving [1, 7, 15, 23].

Researchers have explored leveraging the decision-making capabilities of LLMs for planning in autonomous vehicle control. MTD-GPT [22], for instance, addresses complex decision-making problems at intersections using a sequence modeling approach and fine-tuning. Similarly, DiLu [32] employs LLM agents to solve decision-making challenges in closed-loop driving tasks with a focus on clear prompt design, showcasing remarkable performance comparable to state-of-the-art RL-based models. [6] also adopts a prompt-based approach, showcasing the model's proficiency in decision-making, including high-speed lane changes on the highway. Notably, experiments reveal its adaptability to new scenarios without retraining. The use of in-context learning prompts, tailored to driving nuances, demonstrates the model's flexibility. The study also investigates chain-of-thought prompts, introducing logical steps for potential accident scenarios. Efforts have been made

to overcome the scarcity of driving scenario datasets using LLMs. GAIA-1 [12], for example, generates data on traffic scenarios, environment elements, and potential risks by incorporating video, text, and action inputs. Additionally, there are attempts to approach driving situations through Question-Answering tasks, as evidenced by initiatives like DriveGPT4 [37].

These endeavors collectively demonstrate the growing interest and exploration of the application of Large Language Models in the field of Autonomous Driving, showcasing their potential in decision-making, planning, and data augmentation for overcoming challenges in this dynamic domain.

3. Video Instruction Following Data for AVs

3.1. Dataset Generation

In the field of autonomous driving, datasets containing videos with natural language captions are rare. Furthermore, there's a notable absence of conversation datasets between drivers and vehicles in driving scenarios. Such datasets are pivotal for instruction tuning in multi-modal LLM. To bridge this gap, we curated a multi-modal instruction-following dataset, derived from driving videos and their annotations. The primary datasets leveraged for this study are the Berkeley Deep Drive (BDD), its annotation counterpart BDD-X [19], and the Honda Research Institute Driving Dataset (HDD) [26].

BDD-X offers 77 hours of footage across 6,970 videos, each 40 seconds in length. Every video captures approximately 3-4 distinct driving actions—like acceleration, deceleration, or turns—all annotated with both a description and an explanation. HDD, on the other hand, is a 104-hour driving dataset rich in natural language description. This advice is bifurcated into goal-oriented (top-down signals), influencing vehicular navigation tasks, and stimulus-driven advice (bottom-up signals) that highlights visual cues the driver expects the vehicle to notice.

Data Enrichment. To create a high-quality instruction tuning dataset using a cost-effective method, we employed GPT-4 to produce question-answer pairs for each video. Given that we utilize a language-only GPT-4, it's crucial to feed it comprehensive information about the driving video. Even though GPT-4 cannot view the video, such in-depth instructions enable it to grasp the video's content. As depicted in Fig. 1, supplementary data from BDD-X includes time-stamped explanations and, when available, bounding boxes of significant objects. This methodology empowers GPT-4 to discern driving conditions in each frame, taking into account the spatial relationship of road elements in the scene [21]. As for HDD, it offers a more granular perspective from the vehicle's point of view, encompassing car maneuvers, road events, focal points, causative elements, and both goal-oriented and stimulus-driven actions. We concatenate these with their timestamps and also incorporate data like velocity and steering wheel angles per second.

Instruction Following Data Generation with GPT4. In total, we crafted three categories of video instruction-following data: detailed description, conversation, and complex reasoning as described in Fig. 2. Detailed descriptions provide an overview of the entire video, highlighting maneuvers, current traffic conditions, and driver behavior. Conversations contain questions directly inferred from the video, such as the vehicle's reactions and their underlying reasons. Complex reasoning tasks delve deeper, necessitating profound understanding of the scenario, such as predicting a car's future actions given specific conditions. The prompts used to generate Q&A pairs vary for each task, and every dataset employs a distinct prompt. Ultimately, we synthesized 64K instruction-following data points from 11K videos. Our efforts to further expand base datasets to generate instruction-following datasets and frame-level captioning datasets for pre-training are ongoing.

3.2. Dataset Comparison

Tab. 1 presents a comparison of the VLAAD dataset with existing natural language-captioned datasets on driving scenes. Datasets such as BDD-X [19], HDD [26], and CAP-DATA [10] provide free-form captions, describing scenes and justifying them with the driver's intention or ego-car controls. However, these datasets are not designed for QA tasks and lack reasoning tasks, such as hypothetical situations. CAP-DATA, while offering reasoning on current situations, is limited to road accidents and does not cover a range of driving scenes. In contrast, DRAMA [24], NuScenes-QA [25], and T2C [8] are specifically tailored for QA tasks on diverse scenes. However, there are significant differences with our dataset: NuScenes-QA, which contains 34K captions accounting for half of our dataset's size, offers only short answers. While T2C focuses on driver-car interactions, it primarily deals with driver commands and object localization for command execution. DRAMA closely aligns with our dataset by considering driver intention in QA tasks, but it is limited to short-answer responses. Moreover, its video length is a mere 2 seconds, insufficient for comprehensive driving scene understanding and QA with scene reasoning. In comparison, our dataset boasts nearly 64K instructional datasets with videos ranging from 20 to 40 seconds, encompassing free-form QA that accounts for driver intention and is capable of complex reasoning tasks.

4. Adapting Multi-modal LLM to Driving Situations

We opted for Video-LLaMA [39] as our foundational model to assist LLM (LLaMA-2) in comprehending video

Data Generation Example in HAD

Enriched video description:

[Car maneuver] yield to oncoming traffic and wait to turn left when traffic in the oncoming lane is clear and make the left turn then continue down the block\n[cause] From 0sec to 13sec: crossing_vehicle\n[goal-oriented] From 15sec to 19sec: left_turn\n[Steering Angles] : [-1.2, -1.2, -1.2, -1.2, -1.2, -1.2, -1.2, -1.2, -1.2, -1.2, -1.2, -1.2, -1.2, -1.2, -1.2, -31.0, -215.4, -238.9, -258.6, -142.2, -51.1, 0.4]\n[Velocities] : [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 4.3, 6.97, 9.96, 13.97, 15.57, 17.8, 19.46, 22.33]

Prompt for complex reasoning task:

```
As an AI visual assistant of a driver, you are watching front-view road for around 20 seconds.
    You are given the summarization of road events and manuever of your ego-car during 20 seconds and some detailed
information, which inlcudes "[attention]" you may take care of while driving,
    "[cause]" that makes you do [sitmulus-driven] behavior, "[goal-oriented]" that you do with a purpose of going
somewhere or else, "[stimulus-driven]" behavior that reaction to some changes on road or traffic,
    "[Steering angles]" that is steering angle of ego-car at every second, and "[Velocities]" that is velocity of ego-car
at every second.

    Create a situational reasoning question and an answer between yourself and someone inquiring about the current
driving scene. Make sure the response reflect the tone of a AI visual assistant of a driver, actively observing the
driving situations and answering questions.
    This Q&A should be regarding situational Reasoning for traffic understanding.

    Encompass a question which needs complex situational reasoning to be answered, like those discussing events occurred
in the road, focusing on important changes or situations that drvier should take care of,
    or predicting what might happen or how the driver would behave if a given situation went differently, not just asking
what happened and what can be seen.
```

Generated Instruction following data:

Question: "How did the car handle the oncoming traffic before making the left turn?",
Answer: "The ego-car detected oncoming traffic and stopped to wait until the lane was clear. It maintained the stopped position for quite some time and only proceeded to turn left when it was safe to do so."

Figure 1. An example of data generation process in HDD.

Dataset	Ours	BDD-X [19]	HDD [26]	CAP-DATA [10]	T2C [8]	DRAMA [24]	NuScenes-QA [25]
# of scenarios	**10,379**	6,984	7,744	11,727	850	17,758	1,000
Video length(s)	**20, 40**	40	20	5	20	2	20
# of captions	**64K**	26,539	-	-	11,959	17,066	34K
Visual QA	✓	no	no	no	no	✓	✓
Free-form QA	✓	no	no	no	-	no	no
Ego-intention	✓	✓	✓	✓	no	✓	no
Reasoning	✓	no	no	no	no	no	no

Table 1. Comparison of datasets with natural language captions on driving scene videos

data. As illustrated in Fig. 3, the main element is the vision-language branch, which is composed of several modules. First, a frozen pre-trained image encoder is utilized to extract features from video frames. Next, a position embedding layer is applied to infuse temporal information into the video frames. Following that, a video Q-former is utilized to consolidate frame-level representations. Lastly, a linear layer is utilized to project the resulting video representations into the same dimension as the text embeddings used by LLMs.

Similar to Video-LLaMA, we also harnessed the power of the pre-trained BLIP-2 model [14] as our visual encoder. This model incorporates both the vision transformer and Q-Former modules, aligning with our vision-language branch requirements. For our language model, we employed a pre-trained and frozen version of LLaMA-2 to ensure robust and consistent performance throughout our experiments. We made use of the pre-trained model from Video-LLaMA. This model underwent training using Webvid-2M [2], a vast dataset comprising short videos accompanied by textual descriptions sourced from stock footage websites. Additionally, it leveraged the image caption dataset CC595k, derived from CC3M [29], and filtered by [21]. The pre-training process for the vision-language branch involved a video-to-text generation task. In this stage, the model was tasked with generating corresponding text descriptions based on video

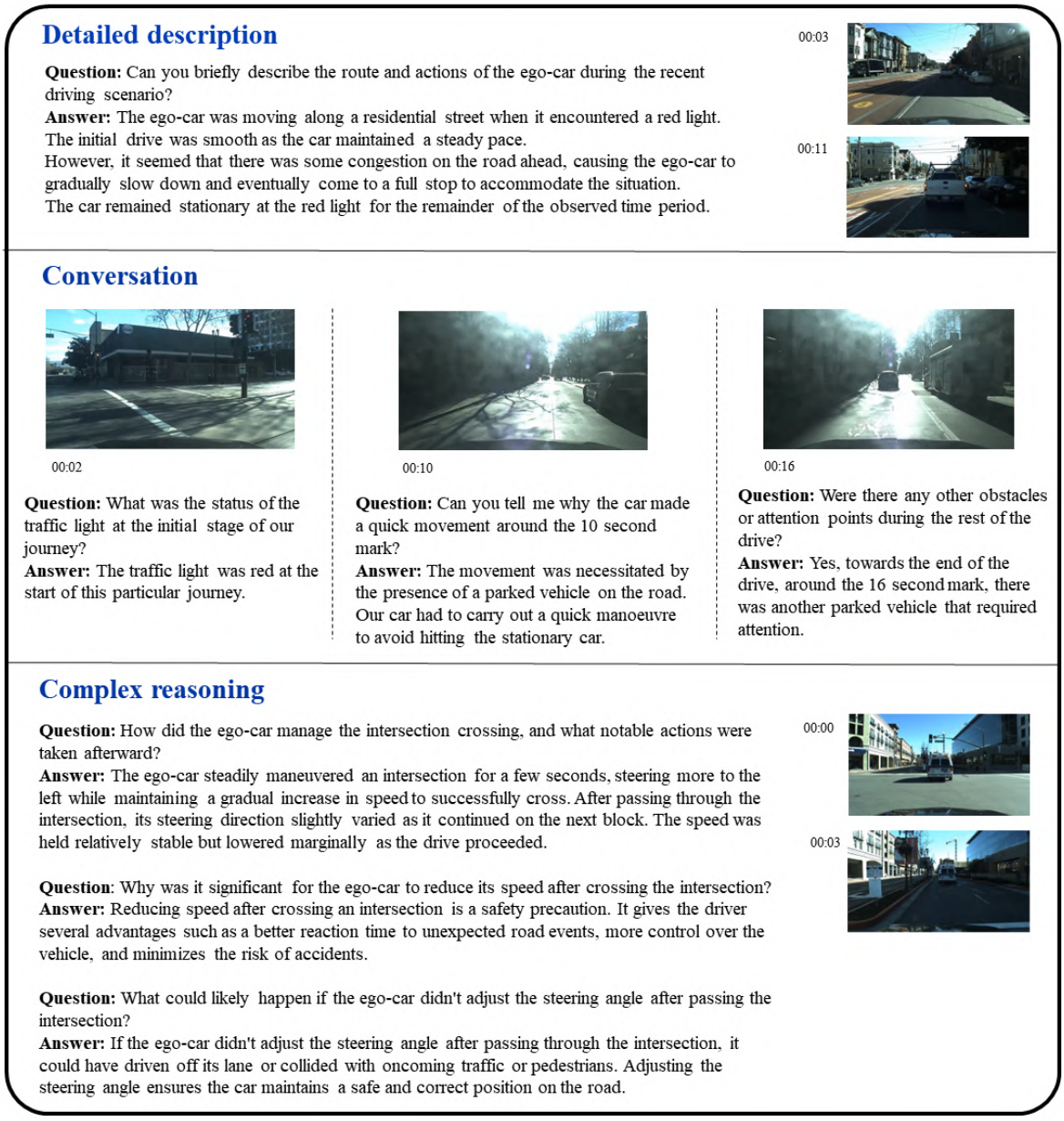

Figure 2. Examples of three types of datasets—detailed description, conversation, and complex reasoning.

representations, effectively prompting the frozen LLM to perform this text generation task.

For fine-tuning to our dataset, we only keep the visual encoder and LLM weights frozen, and continue to update the pre-trained weights of the video-Qformer. In this way, we cost-efficiently inject information needed to understand driving situation. To train the model to follow instructions and carry out tasks in a conversational manner, we fine-tuned our model using the driving video-language instruction-following dataset that we collected in Section 3. As illustrated in Fig. 4, our model, fine-tuned at this stage, exhibited the capability not only to serve as a visual assistant for interacting with drivers but also to exhibit inferential skills in hypothetical scenarios.

5. Conclusions

In this study, we have presented a methodology for generating instruction-following data in the context of autonomous driving, along with the release of its benchmark dataset. Our results, obtained through fine-tuning an existing multi-modal model designed for video understanding,

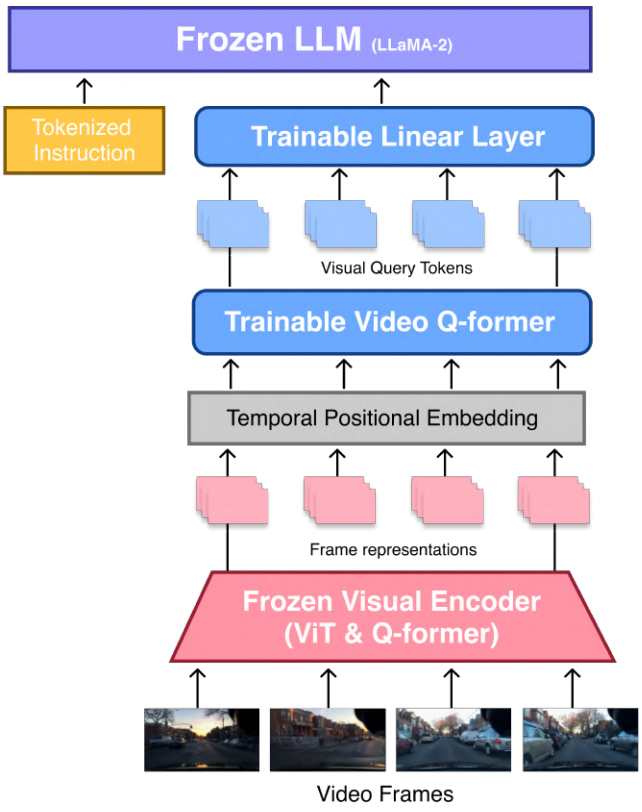

Figure 3. Architecture of trained model

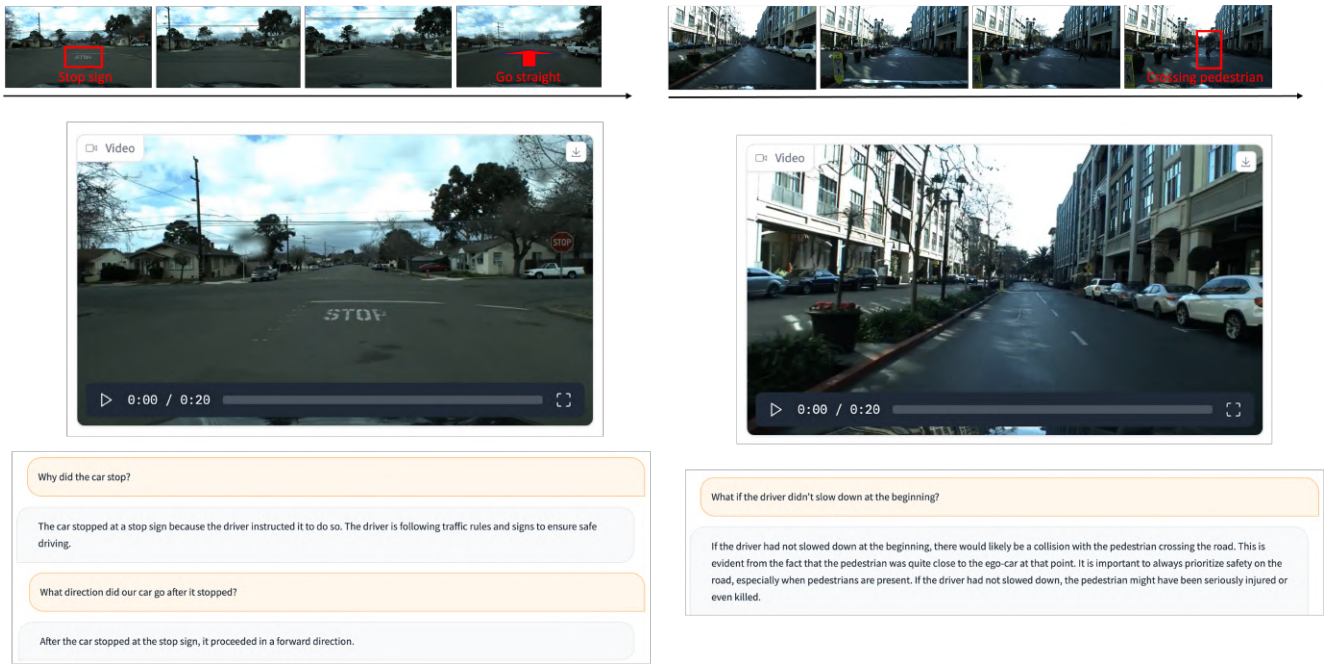

Figure 4. Example conversations on driving scene videos through VLAAD

have demonstrated its ability to comprehend driving scenarios within videos by effectively capturing spatio-temporal relationships.

Currently, our ongoing efforts involve refining the model to selectively extract key frames that are pertinent to instructions. Additionally, we have plans to replace the visual encoder with a pre-trained encoder specifically trained on extensive driving scene-caption data. Furthermore, we are actively engaged in the generation of data for both pre-training and fine-tuning to support the model's application in various contexts and domains.

References

[1] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions, 2023. 2

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval, 2022. 4

[3] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D. Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, Xin Zhang, Jake Zhao, and Karol Zieba. End to end learning for self-driving cars, 2016. 1

[4] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers, 2023. 1

[5] Sheng Zhang Naoto Usuyama Haotian Liu Jianwei Yang Tristan Naumann Hoifung Poon Jianfeng Gao Chunyuan Li, Cliff Wong. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, 2023. 2

[6] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Receive, reason, and react: Drive as you say with large language models in autonomous vehicles, 2023. 2

[7] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tong Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, Ziran Wang, and Chao Zheng. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, 2024. 2

[8] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie Francine Moens. Talk2car: Taking control of your self-driving car. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2088–2098, 2019. 2, 3, 4

[9] Xiaoqian Shen Xiang Li Deyao Zhu, Jun Chen and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models., 2023. 2

[10] Jianwu Fang, Lei-Lei Li, Kuan Yang, Zhedong Zheng, Jianru Xue, and Tat-Seng Chua. Cognitive accident prediction in driving scenes: A multimodality benchmark. *CoRR*, abs/2212.09381, 2022. 3, 4

[11] Junnan Li Dongxu Li Caiming Xiong Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2022. 2

[12] 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, 2023. 3

[13] Bu Jin, Xinyu Liu, Yupeng Zheng, Pengfei Li, Hao Zhao, Tong Zhang, Yuhang Zheng, Guyue Zhou, and Jingjing Liu. Adapt: Action-aware driving caption transformer, 2023. 2

[14] eprint=2301.12597 archivePrefix=arXiv primaryClass=cs.CV Junnan Li Dongxu Li Silvio Savarese Steven Hoi, year=2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. 2, 4

[15] Ali Keysan, Andreas Look, Eitan Kosman, Gonca Gürsun, Jörg Wagner, Yu Yao, and Barbara Rakitsch. Can you text what is happening? integrating pre-trained language encoders into trajectory prediction models for autonomous driving, 2023. 2

[16] Jinkyu Kim and John Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 2

[17] Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny. Grounding human-to-vehicle advice for self-driving vehicles. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[18] Jinkyu Kim, Suhong Moon, Anna Rohrbach, Trevor Darrell, and John Canny. Advisable learning for self-driving vehicles by internalizing observation-to-action rules. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9658–9667, 2020. 2

[19] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles, 2018. 1, 2, 3, 4

[20] Wei Li, Zhaowei Qu, Haiyu Song, Pengjie Wang, and Bo Xue. The traffic scene understanding and prediction based on image captioning. *IEEE Access*, 9:1420–1427, 2021. 2

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 2, 3, 4

[22] Jiaqi Liu, Peng Hang, Xiao qi, Jianqiang Wang, and Jian Sun. Mtd-gpt: A multi-task decision-making gpt model for autonomous driving at unsignalized intersections, 2023. 2

[23] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning in driving, 2022. 2

[24] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. Drama: Joint risk localization and captioning

in driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1043–1052, 2023. 2, 3, 4

[25] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario, 2023. 2, 3, 4

[26] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 3, 4

[27] Junha Roh, Chris Paxton, Andrzej Pronobis, Ali Farhadi, and Dieter Fox. Conditional driving from natural language instructions, 2019. 2

[28] Avishkar Saha, Oscar Mendez, Chris Russell, and Richard Bowden. Translating images into maps. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 9200–9206, 2022. 2

[29] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. 2018. 4

[30] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 1

[31] Hengli Wang, Peide Cai, Yuxiang Sun, Lujia Wang, and Ming Liu. Learning interpretable end-to-end vision-based motion planning for autonomous driving with optical flow distillation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13731–13737, 2021. 2

[32] Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. Dilu: A knowledge-driven approach to autonomous driving with large language models, 2023. 2

[33] Yi Li1 Weiyue Li1 Zeyuan Chen1 Zhuowen Tu1 Wenbo Hu*1, Yifan Xu*2. Bliva: A simple multimodal llm for better handling of text-rich visual questions, 2023. 2

[34] 1 Dongxu Li Anthony Meng Huat Tiong Junqi Zhao Weisheng Wang Boyang Li Pascale Fung Steven HoiB Wenliang Dai Junnan Li, B. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2

[35] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets, 2017. 1

[36] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. Explainable object-induced action decision for autonomous vehicles. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9520–9529, 2020. 2

[37] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee. K. Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model, 2023. 3

[38] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[39] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding, 2023. 1, 2, 3