On the Road with GPT-4V(ision): Explorations of Utilizing Visual-Language Model as Autonomous Driving Agent

Licheng Wen[1†], **Xuemeng Yang**[1†], **Daocheng Fu**[1†], **Xiaofeng Wang**[2†], **Pinlong Cai**[1],

Xin Li[1,3], **Tao Ma**[1,4], **Yingxuan Li**[2], **Linran Xu**[2], **Dengke Shang**[2], **Zheng Zhu**[2♠],

Shaoyan Sun[2], **Yeqi Bai**[1], **Xinyu Cai**[1], **Min Dou**[1], **Shuanglu Hu**[5], **Botian Shi**[1♠], **Yu Qiao**[1]

[1] Shanghai Artificial Intelligence Laboratory [2] GigaAI [3] East China Normal University

[4] The Chinese University of Hong Kong [5] WeRide.ai

[†] Core Contributors [♠] Corresponding Authors

`shibotian@pjlab.org.cn` `zhengzhu@ieee.org`

Abstract

The development of autonomous driving technology depends on merging perception, decision, and control systems. Traditional strategies have struggled to understand complex driving environments and other road users' intent. This bottleneck, especially in constructing common sense reasoning and nuanced scene understanding, affects the safe and reliable operations of autonomous vehicles. The introduction of Visual Language Models (VLM) opens up possibilities for fully autonomous driving. This report evaluates the potential of GPT-4V(ision), the latest state-of-the-art VLM, as an autonomous driving agent. The evaluation primarily assesses the model's ultimate ability to act as a driving agent under varying conditions, while also considering its capacity to understand driving scenes and make decisions. Findings show that GPT-4V outperforms existing systems in scene understanding and causal reasoning. It has the potential in handling unexpected scenarios, understanding intentions, and making informed decisions. However, limitations remain in direction determination, traffic light recognition, vision grounding, and spatial reasoning tasks, highlighting the need for further research. The project is now available on GitHub for interested parties to access and utilize: `https://github.com/PJLab-ADG/GPT4V-AD-Exploration`.

1 Introduction

The quest for fully autonomous vehicles has long been constrained by a pipeline that relies on perception, decision-making, and planning control systems. Traditional approaches, either data-based or rule-based, face challenges in precisely recognizing open-vocabulary objects—that is, objects not present in the training dataset—and understanding the intentions of other road (Codevilla et al., 2019). They fail to develop a "common sense" for handling unusual scenarios and effectively summarizing driving-related data for understanding intricate scenarios and effective causal reasoning (Li et al., 2023).

The rise of Large Language Models (LLMs), such as GPT-3.5 (OpenAI, 2023a), GLM (Du et al., 2022; Zeng et al., 2022), Llama (Touvron et al., 2023a;b), *et al.*, suggests potential solutions to these issues. They show capacities for common sense reasoning and understanding complex driving scenarios, whereas mostly used in decision-making and planning phases due to their inability to process visual data. The advent of GPT-4V (OpenAI, 2023b;c;d; Yang et al., 2023), a leading Vision-Language Model (VLM), brings a new dimension to this field. Unlike its predecessors GPT-4V shows robust capabilities in image understanding, marking a considerable leap in bridging the perception gap. This newfound strength raises the question: **Can VLM serve as a robust agent in autonomous driving?**

This report attempts to answer the above question by evaluating GPT-4V's abilities from scenario understanding to reasoning, and finally test its continuous judgment and decision-making ability as the driving agent in real-world driving scenarios. The test samples are sourced from various outlets, including open-source datasets such as nuScenes (Caesar et al., 2020), Waymo Open dataset (Sun et al., 2020), Berkeley Deep Drive-X Dataset (BDD-X) (Kim et al., 2018), D^2-city (Che et al., 2019), Car Crash Dataset (CCD) (Bao et al., 2020), TSDD (TSD), CODA (Li et al., 2022), ADD (Wu et al., 2023), as well as V2X datasets like DAIR-V2X (Yu et al., 2022) and CitySim (Zheng et al., 2022). Additionally, some samples are derived from the CARLA (Dosovitskiy et al., 2017) simulation environment, and others are obtained from the internet. It's worth noting that the image data used in testing may include images with timestamps up to April 2023, potentially overlapping with the GPT-4V model's training data, while the text queries employed in this article are entirely generated anew. After extensive testing, we elucidate both GPT-4V's strengths and shortcomings, offering a foundation for future exploration in autonomous driving.[1]

We assess GPT-4V's capabilities with increasing difficulty, guided by the following questions:

1. **Scenario Understanding:** This test aims to assess GPT-4V's fundamental abilities, which involves recognizing weather and illumination conditions while driving, identifying traffic lights and signs in various countries, and assessing the positions and actions of other traffic participants in photos taken by different cameras.
2. **Reasoning:** In this section, we explore GPT-4V's causal reasoning in autonomous driving, emphasizing handling challenging corner cases and assessing its capacity with surrounding views. Given GPT-4V's inability to directly process video data, we input concatenated time series images to measure time-correlated capabilities. We also run tests to verify its capacity to relate real-world scenarios to navigation images, further evaluating its comprehension of autonomous driving situations.
3. **Serving as a driving agent:** To harness the full potential of GPT-4V, we entrust it with the role of a seasoned driving agent, tasking it with making decisions in real driving situations based on the environment. Our approach involves sampling driving video at a consistent frame rate and feeding it to GPT-4V frame by frame. To aid its decision-making, we supply essential vehicle speed and other relevant information and convey the driving objective for each video. We challenge GPT-4V to produce the necessary actions and provide explanations for its choices, thereby pushing the boundaries of its capabilities in real-world driving scenarios.

In conclusion, we provide initial insights as a basis for inspiring future research in autonomous driving with GPT-4V. We methodically present qualitative results through a unique compilation of image-text pairs, offering a comprehensive analysis despite a somewhat less stringent methodology.

2 Basic Capability of Scenario Understanding

To function as a secure and proficient autonomous driving agent, a comprehensive grasp of complex traffic scenarios is essential. This involves accurate recognition and understanding of diverse driving conditions and traffic participants. In this section, we focus on evaluating GPT-4V's capabilities in comprehending traffic scenarios, emphasizing its understanding of the environment and its ability to discern the behavior of traffic participants. Tests include

[1] All experiments detailed in this paper were conducted before November 5th, 2023. We acknowledge that the most recent version of GPT-4V, which has received updates following the November 6th OpenAI DevDay, may produce different responses when presented with the same images.

assessing its recognition of time of day, weather conditions, and proficiency in interpreting traffic lights and signs. Additionally, we evaluate GPT-4V's capability in understanding traffic participants' behavior using various sensor inputs, such as 2D front view images, V2X images, and images from simulation. The correct interpretation of these elements is crucial for shaping informed and appropriate decisions in autonomous driving systems.

For detailed test results and a thorough analysis, please refer to the provided Appendix A. The findings reveal that the model adeptly adapts to variations in time, lighting conditions, and diverse image inputs. While occasional errors occur in distinguishing nuanced details and processing information in non-English languages, the model consistently showcases strong environmental understanding capabilities throughout the tests.

3 Advanced Capability of Reasoning

In evaluating GPT-4V's driving behavior, reasoning emerges as a crucial trait given the dynamic nature of traffic environments. Proficient drivers excel at making accurate judgments and decisions in response to emergencies. While autonomous driving systems typically rely on continuous data collection, the inherent unpredictability of driving environments poses a challenge in fully capturing unforeseen circumstances. Human drivers navigate such events successfully by leveraging common sense, emphasizing the need to integrate reasoning into autonomous systems. This section employs a diverse set of challenging scenarios, incorporating corner cases, multi-view images, temporal images, and pairs of navigation application screenshots and road images. This comprehensive approach aims to rigorously evaluate GPT-4V's performance across a spectrum of intricate situations, providing valuable insights into its adaptability and effectiveness in handling varied and complex visual inputs.

The comprehensive test results and in-depth analysis are provided in the Appendix B. Extensive experiments demonstrate GPT-4V's proficiency in addressing corner cases beyond the scope of traditional data-driven perception algorithms. While exhibiting notable correlation capabilities in the surrounding view images, the model is constrained by its reasoning within 3D spatial relationships. Furthermore, the model showcases adept temporal reasoning capabilities in the sequential images, enabling it to discern the intentions of traffic participants and execute rational driving actions.

4 Serving as A Driving Agent

Recent advances in large models have had a significant impact on autonomous driving, particularly in areas of planning, decision-making, and control. Studies such as Fu et al. (2023); Cui et al. (2023); Wen et al. (2023); Xu et al. (2023) highlight the potential of LLMs for robust reasoning and comprehension. In addition, efforts are being made to integrate LLMs into end-to-end autonomous driving systems to improve interpretability and scene comprehension (Mao et al., 2023; Wang et al., 2023; Shao et al., 2023).

In this section, we evaluate more complex scenarios to further explore the integration of GPT-4V as a driving agent. Achieving this goal necessitates precise recognition, spatial awareness, and a deep understanding of spatiotemporal relationships among various traffic elements. This section evaluates GPT-4V's full potential by examining the decision-making capabilities in four different real-world driving scenarios. The scenarios cover different traffic conditions, different times of day, and diverse driving tasks. Throughout the evaluation process, relevant information including own vehicle speed is provided, requiring GPT-4V to generate observations and perform driving actions. These carefully designed evaluations aim to push the boundaries of VLMs' capabilities as driving agents in real-world scenarios, revealing their potential as future autonomous traffic drivers.

4.1 Driving in Parking Lot

In this test, GPT-4V demonstrates its decision-making ability in an enclosed area. The selected scenario is turning right to exit a parking lot, which requires passing through a security check. As shown in Figure 1 in the first frame, while GPT-4V accurately identifies key elements like pedestrians and vehicle lights it exhibits some ambiguity in assessing pedestrian and distant vehicle status. Despite this, it maintains conservative driving decisions, such as low speed and readiness to stop. Although it makes a mistake about zebra crossings in the second frame, GPT-4V consistently follows a cautious right-turn strategy. In the third frame, it accurately recognizes checkpoints and prepares to stop for a security check. In the fourth frame, GPT-4V correctly identifies the open security checkpoint, ensuring a safe exit from the parking lot. Additionally, it advises waiting for pedestrians near the exit before proceeding slowly.

From this example, GPT-4V can accurately identify key elements within enclosed areas (such as parking lots), including gated checkpoints, guard booths, and fencing. Moreover, GPT-4V understands driving procedures for leaving parking lots, which requires waiting for security checks and paying attention to pedestrians and vehicles. However, some misjudgments may still occur, such as mentioning zebra crossings erroneously.

4.2 Turning at Highway Ramp

In this highway driving test, GPT-4V is evaluated for its performance in a challenging scenario of a nighttime highway ramp turnaround, as shown in Figure 2. Despite accurately identifying key elements like arrow signs and lane lines in the first frame, it exhibits a mistake in counting preceding vehicles in the second frame. Nevertheless, it correctly locates the lane line and road sign, advising light braking and a left turn signal. In low visibility conditions, GPT-4V relies on yellow lane dividers for guidance, suggesting slow driving within the lane lines. As the vehicle enters the main highway road in the fourth frame, GPT-4V accurately assesses the situation, adjusting speed for highway driving and using high beams within legal limits for enhanced nighttime visibility.

From this example, we can see that when driving in highway areas, GPT-4V follows road signs and assists in decision-making based on the status of surrounding vehicles. However, it has limitations in object recognition and positioning during nighttime.

4.3 Road Merging

In this road merging evaluation shown in Figure 3, GPT-4V demonstrates its capability in a scenario of exiting the main road and merging onto a ramp at night. It accurately identifies lane markings and determines that the current lane is ending or merging. So it decides to slow down in the first frame and prepare to merge. Despite mistakenly recognizing a hospital sign, GPT-4V remains cautious about pedestrians and emergency vehicles. In the second frame, it correctly identifies the merging point and advises smooth steering into the lane. Anticipating the end of merging in the third frame, GPT-4V warns of other vehicles cutting in. By the fourth frame, GPT-4V successfully merges onto the road but incorrectly detects a solid white line and believes a motorcycle is present. The final decision is to pay attention to motorcycles on the main road and adjust speed or change lanes if necessary.

From this example, it is observed that GPT-4V can assess current merge progress by observing changes in lanes and providing reasonable driving suggestions. However, there is still an increased probability of misjudging road signs and lanes during nighttime. Overall, GPT-4V tends to adopt a conservative approach when it comes to lane merging.

4.4 U-Turning at Traffic Intersection

In the U-turn capability test of GPT-4V, the scenario involves a U-turn at a busy intersection at night in Figure 4. While accurately identifying nearby vehicles in the first frame and advising maintaining distance, GPT-4V omits the distant traffic light. In the second frame, it still fails to locate the traffic light but infers its presence from surrounding vehicle behavior, suggesting a slowdown in preparation for the U-turn. However, in the third frame, GPT-4V mistakes traffic lights on a side road as control signals for its lane, leading to a decision to remain stopped. This mistake persists in the fourth frame, where GPT-4V maintains a stationary position, misunderstanding the traffic lights.

Sec. 4.1 Right-turn at the Exit

Navigation Prompt:

You will encounter a parking lot exit in this scenario and you will need to turn right to exit the parking lot. You are moving at a relatively low speed. Please answer me using the prescribed format.

GPT–4V:

Observation: The scene includes a parking lot exit with a pedestrian crossing in front of the car, other parked cars on the left, and a vehicle on the opposite side of the road that appears to be moving. [...]

Action: The vehicle is moving at a low speed, preparing to exit the parking lot. Continue at a low speed, prepare to stop.

GPT–4V:

Observation: The image shows that the pedestrian has cleared the zebra crossing. There are no other pedestrians in immediate sight who are about to cross. A white car is on the opposite side of the road, indicating a potential for crossing traffic. [...]

Action: Maintaining a low speed while approaching the exit of the parking lot. Proceed with caution, turn right.

GPT–4V:

Observation: The image appears to be taken from the inside of a vehicle approaching a gated checkpoint, with pedestrians and another vehicle in view. The presence of a guard booth and fencing indicates controlled access to the area.

Action: Prepare to stop if required for security checks or to wait for the gate to open. Comply with any directions given by security personnel.

GPT–4V:

Observation: The barrier gate at the security checkpoint is now fully open. Two cyclists are visible, one passing in front of the vehicle and another one approaching. The security guard is still at his post on the right.

Action: The vehicle is either stopped or moving very slowly, waiting for the cyclists to pass before proceeding through the security checkpoint. The driver should continue to wait until the cyclist in front has passed and then drive through the checkpoint slowly and cautiously, being mindful of any other pedestrians or cyclists.

Figure 1: Illustration of GPT-4V's ability to drive in parking lot. Green highlights the right answer in understanding. Red highlights the wrong answer in understanding. Yellow highlights the incompetence in performing the task. Check Section 4.1 for details.

Sec. 4.3 Highway ramp turnaround

Navigation Prompt:

You are in a scenario where you come across a turnaround intersection and you need to turn around according to the lane line instructions. You are moving at a normal speed. Please answer me using the prescribed format.

GPT−4V:

Observation: [...]what appears to be a turnaround intersection ahead, indicated by arrow signs and a dividing lane line. There are red rear lights of another vehicle at the intersection, suggesting it is either stopped or moving slowly. The road seems to be wet [...]

Action: Reduce speed, turn left following the lane line instructions.

GPT−4V:

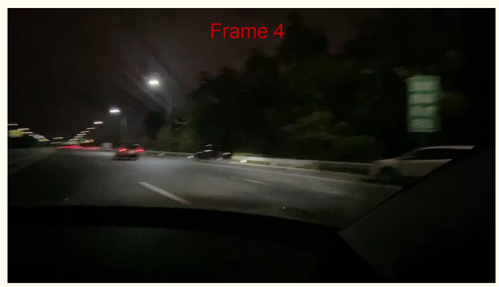
Observation: [...] with several vehicles ahead. Reflective lane markers outline a path that curves to the left. There's a green road sign that appears to indicate the direction for the turnaround, with a no-entry sign for the adjacent lane on the right [...]

Action: Apply the brakes lightly to decelerate in preparation for the left turn. Signal left to inform other drivers of the intention to enter the turnaround.[...]

GPT−4V:

Observation: [...] depicts a vehicle navigating a nighttime road environment, with the focus on a curved section lit by a single streetlight. The road is delineated with yellow lane dividers, and the surrounding area is poorly lit, limiting visibility.

Action: Gently decelerate if necessary to maintain a safe speed that matches the curvature of the road and the existing visibility conditions. Keep the vehicle within the lane, using the yellow lane dividers as a guide. [...]



GPT−4V:

Observation: The image is taken from a vehicle's perspective during nighttime driving on a multi-lane road. There is a white car to the right that appears to be merging or taking an exit, and the glare of streetlights and other vehicle lights is noticeable. The surroundings are dimly lit, and the road seems to curve to the right ahead.

Action: [...] adjust the vehicle's speed or position on the road as necessary to accommodate the traffic flow. Use the vehicle's headlights and consider high beams if appropriate and legal [...].

Figure 2: Illustration of GPT-4V's ability to turn at highway ramp. Green highlights the right answer in understanding. Red highlights the wrong answer in understanding. Yellow highlights the incompetence in performing the task. Check Section 4.2 for details.

Sec. 4.4 Leave the main road

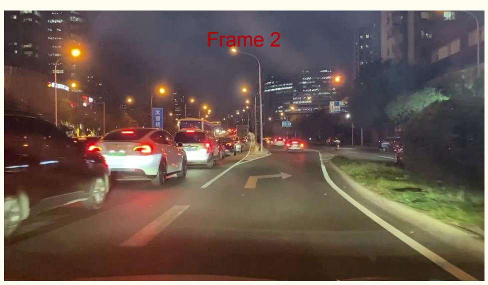
Navigation Prompt:

You are on a busy road and you want to leave the main road and merge onto the ramp.

GPT–4V:

Observation: [...] The lane markings indicate that the current lane is ending or merging, as suggested by the dashed lines becoming solid and the arrow on the road pointing towards the merge. There is a sign indicating a hospital nearby, which may not directly affect driving decisions but suggests a possible increase in pedestrian and emergency vehicle activity.

Action: Reduce speed, prepare to merge right.



GPT–4V:

Observation: The vehicle is now at a point where the lane is merging, with traffic in the main lanes continuing to be congested. The vehicle on the right appears to be very close to the merging area, indicating a critical point to either complete the merge or adjust speed to find a safer opportunity.

Action: If a safe gap is confirmed, the next action would be to smoothly steer into the lane, adjust speed to match the flow of traffic, and turn off the signal once the merge is complete.

GPT–4V:

Observation: The vehicle is now in a position where the merging lane is about to end, indicated by the dashed lane markings transitioning to a solid line. [...] and there is a cyclist ahead on the roadside, which requires careful attention.

Action: If the merge is not yet complete, the driver should do so immediately before the lane ends, taking care to avoid the bus and the cyclist. The driver must also ensure that the merge does not cut off other vehicles or force them to adjust speed abruptly.

GPT–4V:

Observation: The vehicle has successfully merged into the main flow of traffic, with the lane markings now a solid white line indicating that lane changing at this point is generally discouraged or unlawful. A motorcycle is visible ahead in the same lane as the vehicle, and it's a night scene with artificial lighting from street lamps and surrounding vehicles.

Action: The driver should continue to follow the motorcycle at a safe distance and be prepared to adjust speed or change lanes if necessary to maintain safety, especially given that motorcycles can stop more quickly than cars. [...]

Figure 3: Illustration of GPT-4V's ability to merge onto other road. Green highlights the right answer in understanding. Red highlights the wrong answer in understanding. Yellow highlights the incompetence in performing the task. Check Section 4.3 for details.

Sec. 4.5 U-turn at intersection

Navigation Prompt:

This is the first frame. At this moment, you would like to make a U-turn at the intersection ahead.

GPT−4V:

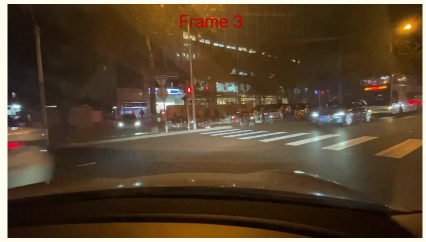
Observation: There are cars in front of us, and the brake lights indicate that they might be slowing down or stopping. A U-turn is intended at the intersection ahead. There are no visible traffic lights in this image [...]

Action: I am presumably maintaining a safe following distance behind the vehicles in front, matching their speeds.

GPT−4V:

Observation: There's a vehicle directly in front of us, and it appears to be positioned for a left turn or U-turn based on its angle and the activated left turn signal. [...] No traffic lights are visible in this image, but the positioning of vehicles suggests there may be a signal controlling the intersection. [...]

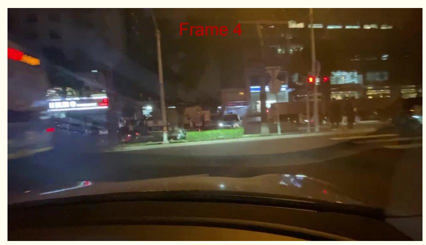
Action: The autonomous vehicle should be continuing to slow down in response to the slowing traffic ahead, maintaining a safe distance from the vehicle in front, and preparing for the intended U-turn.



GPT−4V:

Observation: [...] The traffic light for the driver's direction is red, and other vehicles and motorbikes are seen crossing the intersection from left to right, implying cross traffic has the green light.

Action:The driver should remain stopped until the traffic light turns green. [...] If U-turns are permitted, the driver should also check for oncoming traffic, pedestrians, and any other potential hazards before executing the U-turn.



GPT−4V:

Observation: The image shows a view from a driver's perspective at a nighttime intersection. The traffic light facing the driver remains red. Cross traffic has passed, and the intersection ahead appears clear.[...]

Action: It appears that the vehicle is either in motion or about to move, which should not be the case since the traffic signal is red.The driver should be at a complete stop.

Figure 4: Illustration of GPT-4V's ability to u-turn at traffic intersection. Green highlights the right answer in understanding. Red highlights the wrong answer in understanding. Yellow highlights the incompetence in performing the task. Check Section 4.4 for details.

From this example, we can see that when encountering significant changes in road structure like U-turns, GPT-4V tends to overlook temporal and spatial contextual relationships. Nevertheless, the overall driving strategies provided are quite conservative.

Through the aforementioned four tests, GPT-4V showcases decision-making capabilities as an autonomous driving agent comparable to those of human drivers. It effectively integrates information from diverse traffic elements to formulate driving strategies and consistently makes sound decisions across various scenarios. Nevertheless, limitations emerge in discerning the status of distant objects and facing reduced perception in nighttime conditions. Despite its robust adherence to rules and safety protocols, GPT-4V encounters challenges in spatiotemporal context inference, particularly evident in a U-turn scenario involving multiple traffic lights.

5 Conclusions

5.1 Capabilities of GPT-4V in Autonomous Driving

In this paper, we conduct a comprehensive, multifaceted evaluation of GPT-4V's capabilities as an autonomous driving agent. The results show that GPT-4V has the potential to surpass existing autonomous driving systems in terms of scene understanding, intention recognition, and driving decision-making.

Firstly, GPT-4V possesses fundamental capabilities in comprehending and analyzing traffic scenarios, effectively handling out-of-distribution situations, and precisely gauging the intentions of nearby traffic participants. Moreover, it leverages multi-view images and temporal images to achieve a comprehensive perception of the environment, accurately identifying dynamic interactions among traffic participants and inferring the motivations behind these behaviors. Lastly, as elucidated in Section 4, we observe GPT-4V's performance as an agent continuously making decisions on the open road. Notably, it exhibits the ability to interpret the user interface of navigation applications in a human-like manner, aiding and guiding the driver's decision-making process.

Overall, GPT-4V's performance demonstrates the great potential of visual language models (VLMs) as autonomous driving agents.

5.2 Limitations of GPT-4V in Autonomous Driving

However, during our testing, we also found that GPT-4V performs poorly on the following tasks: (1)**Distinguishing left from right:** There were instances where the model struggled with recognizing directions, which is a critical aspect of autonomous navigation. These cases highlight the model's occasional confusion when interpreting complex junctions or making lane-changing decisions. (2)**Traffic light recognition:** We suspect this problem is due to the extensive semantic information contained within the full image, leading to a loss in the embedding information of traffic lights. When the region of the traffic lights in the image is cropped and inputted separately, the model is capable of successful recognition. (3)**Vision Grounding tasks:** GPT-4V finds it difficult to specify pixel-level coordinates or bounding boxes, managing only to indicate approximate areas within the image. (4)**Spatial Reasoning:** Accurate spatial reasoning is paramount for the safe operation of autonomous vehicles. Whether it is the stitching of multi-view images or the estimation of the relative positional relationship between other participants and the self-driving car, GPT-4V struggles with making precise judgments. This may stem from the inherent complexity in understanding and interpreting 3D space based on 2D image inputs. Additionally, issues are found with the model's interpretation of non-English traffic signs, which poses a challenge in regions where multiple languages are used on signage. The accuracy of counting traffic participants is also found to be less reliable in congested environments where overlapping objects can occur.

In conclusion, the above limitations indicate that even the most advanced Vision-Language Models (VLMs) currently exhibit deficiencies in basic directional recognition and traffic light identification, as well as a lack of 3D spatial reasoning capabilities. Furthermore, VLMs struggle to accurately localize key entities in various scenarios, suggesting that they are not yet suitable replacements for the perception methods used in existing autonomous driving pipelines. However, it is noteworthy that VLMs demonstrate a deep understanding of traffic common sense and strong generalization capabilities in out-of-distribution cases. Looking ahead, a key area of development will be to integrate the innate common sense knowledge of VLMs with conventional autonomous driving perception techniques to serve as autonomous driving decision-making agents. In addition, ensuring the safety and reliability of VLM outputs remains an essential and ongoing challenge.

Acknowledgments

The research was supported by Shanghai Artificial Intelligence Laboratory, the National Key R&D Program of China (Grant No. 2022ZD0160104) and the Science and Technology Commission of Shanghai Municipality (Grant Nos. 22DZ1100102 and 23YF1462900).

References

Chinese traffic sign database. `http://www.nlpr.ia.ac.cn/pal/trafficdata/detection.html`.

Wentao Bao, Qi Yu, and Yu Kong. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *ACM Multimedia Conference*, May 2020.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.

Zhengping Che, Guangyu Li, Tracy Li, Bo Jiang, Xuefeng Shi, Xinsheng Zhang, Ying Lu, Guobin Wu, Yan Liu, and Jieping Ye. D^2-city: A large-scale dashcam video dataset of diverse traffic scenarios. 2019.

Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9329–9338, 2019.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. *arXiv preprint arXiv:2309.10228*, 2023.

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

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 320–335, 2022.

Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. *arXiv preprint arXiv:2307.07162*, 2023.

Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving. In *European Conference on Computer Vision*, pp. 406–423. Springer, 2022.

Xin Li, Yeqi Bai, Pinlong Cai, Licheng Wen, Daocheng Fu, Bo Zhang, Xuemeng Yang, Xinyu Cai, Tao Ma, Jianfei Guo, Xing Gao, Min Dou, Yikang Li, Botian Shi, Yong Liu, Liang He, and Yu Qiao. Towards knowledge-driven autonomous driving. *arXiv preprint arXiv:2312.04316*, 2023.

Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A language agent for autonomous driving. *arXiv preprint arXiv:2311.10813*, 2023.

OpenAI. `https://chat.openai.com`, 2023a.

OpenAI. Gpt-4v(ision) system card. 2023b. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf.

OpenAI. Gpt-4v(ision) technical work and authors. https://openai.com/contributions/gpt-4v, 2023c.

OpenAI. Chatgpt can now see, hear, and speak. https://openai.com/blog/chatgpt-can-now-see-hear-and-speak, 2023d.

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.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.

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*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

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.

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. *arXiv preprint arXiv:2309.16292*, 2023.

Zizhang Wu, Xinyuan Chen, Hongyang Wei, Fan Song, and Tianhao Xu. Add: An automatic desensitization fisheye dataset for autonomous driving. *Engineering Applications of Artificial Intelligence*, 126:106766, 2023.

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.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9, 2023.

Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21361–21370, 2022.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

Ou Zheng, Mohamed Abdel-Aty, Lishengsa Yue, Amr Abdelraouf, Zijin Wang, and Nada Mahmoud. Citysim: A drone-based vehicle trajectory dataset for safety oriented research and digital twins. *arXiv preprint arXiv:2208.11036*, 2022.

A Basic Capability of Scenario Understanding

To achieve safe and effective autonomous driving, a fundamental prerequisite is a thorough understanding of the current scenario. Complex traffic scenarios encompass a myriad of driving conditions, each hosting a diverse array of traffic participants. Accurate recognition and comprehension of these elements serve as basic capabilities for an autonomous vehicle to make informed and appropriate driving decisions. In this section, we present a series of tests aimed at evaluating GPT-4V's ability to comprehend traffic scenarios. We focus on two primary aspects: the model's understanding of the surrounding environment and its capacity to discern the behavior and status of various traffic participants. Through these assessments, we aim to shed light on GPT-4V's competence in interpreting the dynamic traffic environment.

A.1 Understanding of Environment

In assessing the capabilities of GPT-4V to comprehend its surrounding environments, we conducted a series of tests encompassing the following key aspects: its ability to discern the time of day, its understanding of prevailing weather conditions, and its proficiency in recognizing and interpreting traffic lights and signs. These elements hold paramount significance in shaping the autonomous driving system's decision-making process. For instance, it is evident that driving at night or in challenging weather conditions requires a heightened level of caution, whereas during daylight hours or in favorable weather conditions a more leisurely driving strategy can be adopted. Besides, the correct interpretation of traffic lights and road signs is essential for the effectiveness of autonomous driving systems. We utilize vehicles' front-view images as the primary visual input throughout this section. The visual data employed here is drawn from nuScenes Caesar et al. (2020), D^2-cityChe et al. (2019), BDD-X Kim et al. (2018) and TSDD TSD.

Time understanding throughout the day. We evaluate GPT-4V's ability to comprehend temporal differences by providing it with both daytime and nighttime images. We instruct the model to describe the traffic scenarios depicted in these images, and the results are presented in Figure 5. The findings reveal that, when presented with daytime

Sec. 2.1 Time understanding throughout the day

Prompt:

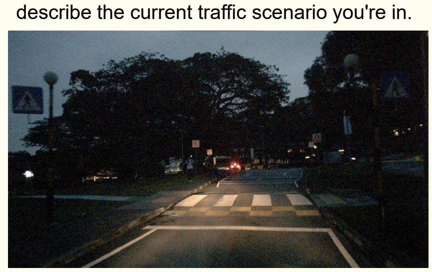
This is the front-view image of your car. Please describe the current traffic scenario you're in.

GPT-4V:

You're on a city street with multiple lanes. There are parked cars on both sides of the road. Ahead, there's moderate traffic with vehicles moving in both directions. There's also a pedestrian crossing zone visible. Buildings and trees line the street, and there are various street signs along the way.

Prompt:

This is the front-view image of your car. Please describe the current traffic scenario you're in.



GPT-4V:

You are on a road during twilight or early evening. The road has a pedestrian crossing ahead. There's a vehicle with its tail lights on, indicating it's either stationary or moving away from you. There are traffic signs on both sides of the road, and the area appears to be moderately lit. Trees and other structures are visible in the background.

Figure 5: Results on the ability to comprehensively understand time over the course of a day. Green highlights the right answer in understanding. Check Section A.1 for details.

scenes, GPT-4V successfully identifies them as multi-lane urban roads with "moderate traffic". Furthermore, the model adeptly recognizes the presence of a crosswalk on the road. When confronted with similar nighttime scenes, GPT-4V's performance is even better. It not only discerns the time as "twilight or early evening" but also detects a vehicle with its tail lights on in the distance, and infers that "it's either stationary or moving away from you".

Weather understanding. Weather is a crucial environmental factor that significantly influences driving behavior. We selected four photographs captured at the same intersection under varying weather conditions from the nuScenes Caesar et al. (2020) dataset. We tasked GPT-4V with identifying the weather conditions depicted in these images. The results are presented in Figure 6. The results demonstrate that GPT-4V exhibits remarkable accuracy in recognizing the weather conditions in each image, namely, cloudy, sunny, overcast, and rainy. Moreover, it provides sound justifications for these conclusions, citing factors such as the presence of sunny shadows or the wetness of the streets.

Traffic light understanding. Recognition of traffic lights plays a pivotal role in the functionality of an autonomous driving system. Incorrectly identifying or missing traffic lights not only leads to violations of traffic regulations but also poses a serious risk of traffic accidents. Unfortunately, the performance of GPT-4V in this test falls short, as evident in Figure 7 and Figure 8. In Figure 7, GPT-4V demonstrates proficiency in distinguishing between yellow street lights and red traffic lights, particularly during nighttime conditions. However, in Figure 8, when confronted with a smaller traffic light with a countdown timer in the distant part of the image, GPT-4V inaccurately identifies the countdown as red and overlooks the genuine 2-second red countdown. The model can provide the correct response only when the traffic light is zoomed in to occupy a significant portion of the image. Furthermore, GPT-4V exhibited instances of misrecognition of traffic lights during subsequent tests, which is deemed unacceptable for a mature autonomous driving system.

Traffic signs understanding. Traffic signs contain various rules and instructions that drivers need to follow. Autonomous driving systems can understand and comply with these rules by identifying traffic signs, thereby reducing the risk of traffic accidents and improving driving safety. Therefore, we selected representative images from Singapore and China for testing. As can be seen from the left sample in Figure 9, the model can recognize most road signs, including the nearby "SLOW" and the distant height limit "4.5m", but it misrecognizes the "Speed Bump" sign. The three signs from the right sample are all correctly recognized. This indicates that GPT-4V possesses a commendable ability to recognize traffic signs, yet there remains room for further enhancement.

A.2 Understanding of Traffic Participants

Accurately understanding the status and behavior of traffic participants is the foundation of driving. Existing autonomous driving systems often use a variety of cameras and sensors to perceive traffic participants in order to obtain more comprehensive information about them. In this section, we assess GPT-4V's proficiency in comprehending the behavior of traffic participants using various sensor inputs, including 2D images, visualizations of 3D point clouds, and images acquired from V2X devices and autonomous driving simulation software. The visual data employed here is drawn from nuScenes Caesar et al. (2020), ADD Wu et al. (2023), Waymo Sun et al. (2020), DAIR-V2X Yu et al. (2022), CitySim Zheng et al. (2022) and Carla Dosovitskiy et al. (2017) simulation.

Front-view camera photos. In order to test the model's basic recognition capabilities, including traffic participant recognition and vehicle counting, we input a series of front views of driving scenes and obtained the output results of GPT-4V. As can be seen from the left side of Figure 10, the model can completely and accurately describe the driving scene: it recognizes pedestrians, traffic signs, traffic light status and the surrounding environment. The right side of Figure 10 shows that the model can identify the vehicle type and its taillights, and can guess its intention to turn on the taillights. However, the model outputs some incorrect statements in irrelevant places, such as thinking that the car in front has a rearview camera. In Figure 11, the counting proficiency of GPT-4V is put to the test. Utilizing both a daytime and a nighttime snapshot from the vehicle's front view, the model precisely

Sec. 2.1 Weather understanding

Prompt:

Please describe the weather in this image

GPT-4V:

The image depicts a clear day with no visible signs of precipitation. The sky appears partly cloudy, with patches of blue sky. The sunlight casts shadows on the ground, indicating it's either morning or afternoon.

Prompt:

Please describe the weather in this image

GPT-4V:

The image depicts a clear and sunny day with shadows cast by the buildings and vehicles, indicating bright sunlight. The sky appears to be mostly cloudless.

Prompt:

Please describe the weather in this image

GPT-4V:

The image shows an overcast day. The lighting is diffused, with no strong shadows visible on the ground or buildings. The sky appears to be cloudy, suggesting a lack of direct sunlight.

Prompt:

Please describe the weather in this image

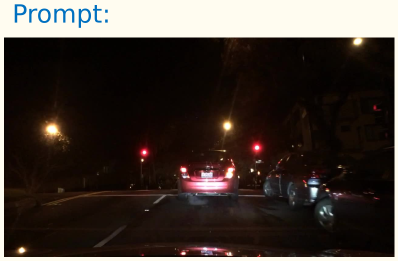
GPT-4V:

The image depicts a rainy or recently rainy environment. The roads and pavements are wet, and there is a visible sheen on the street. The sky appears overcast, and the overall lighting is diffused, indicating cloudy conditions.

Figure 6: Results that identify the season in which the vehicle is driven. Green highlights the right answer in understanding. Check Section A.1 for details.

Sec. 2.1 Traffic light understanding

Prompt:



Describe the condition of all the traffic lights in the image, if there are countdowns, please describe them together.

GPT-4V:

The image has two red traffic lights, both without any visible countdowns.

Prompt:

How about the yellow lights in the image?

GPT-4V:

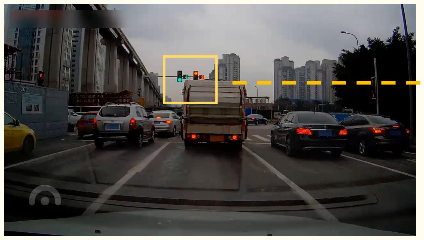
In addition to the red traffic lights, there are two yellow streetlights illuminating the road, but they are not traffic signals.

Figure 7: Results on the ability to understand the traffic light. Green highlights the right answer in understanding. Check Section A.1 for details.

Sec. 2.1 Traffic light understanding

Prompt:

Describe the condition of all the traffic lights in the image, if there are countdowns, please describe them together.



GPT-4V:

The traffic lights in the center show a green light and a red arrow with a countdown displaying "4". On the left and right, the traffic lights are showing red.

Prompt:

Describe the condition of all the traffic lights in the image, if there are countdowns, please describe them together.

GPT-4V:

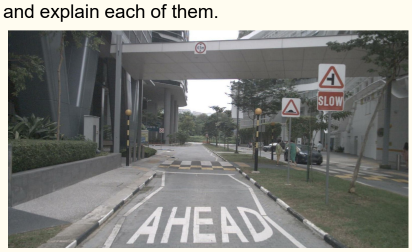
The traffic light is showing a green light with a countdown displaying "4", and a red arrow with a countdown displaying "2".

Figure 8: Results on the ability to understand the traffic light. Green highlights the right answer in understanding, Red highlights the wrong answer in understanding. Check Section A.1 for details.

Sec. 2.1 Traffic signs understanding

Prompt:

This is a front-camera view of a driving car. Please identify all the traffic signs in the image, and explain each of them.



GPT-4V:

There are three traffic signs in the image:

A sign under the overhead structure displaying "4.5m": This sign indicates the maximum height of vehicles that can safely pass underneath the structure. Vehicles taller than 4.5 meters should not proceed.

A triangular sign with a curved arrow pointing left: This is a warning sign that indicates there is a left bend or curve ahead. Drivers should be cautious and possibly reduce their speed.

A rectangular sign with the word "SLOW": This sign instructs drivers to reduce their speed, typically because of potential hazards or a condition ahead that requires slower driving.

Prompt:

This is a front-camera view of a driving car. Please identify all the traffic signs in the image, and explain each of them.

GPT-4V:

There are three traffic signs displayed:

60: This is a speed limit sign, indicating that the maximum allowed speed on this road or section of the road is 60 kilometers per hour (km/h).

Truck with a slash through it: This is a no trucks or no heavy vehicles sign, indicating that trucks or heavy vehicles are not permitted on this road or section of the road.

Car with a horn and a slash through it: This signifies a no honking or horn prohibited zone. Drivers are not allowed to use their vehicle's horn in the area where this sign is displayed.

Figure 9: Results on the ability to understand the traffic signs. Green highlights the right answer in understanding, Red highlights the wrong answer in understanding. Check Section A.1 for details.

tallies the vehicular presence and their statuses in the daylight capture. In contrast, during nighttime conditions, despite GPT-4V accurately enumerating the discernible vehicles, its elaborate description of each individual vehicle sometimes falls short of accuracy.

Fish-eye camera photo. The fisheye camera, a prevalent imaging device within autonomous vehicle systems, was also employed to evaluate the perception abilities of GPT-4V. Results derived from images captured by a fisheye lens are documented in Figure 12. GPT-4V exhibits an impressive robust tolerance for the distinctive fisheye distortion and shows a commendable understanding of the indoor parking environment. It reliably identifies parked vehicles and the presence of pedestrians in proximity, although there are hallucinations describing a charging station that doesn't exist. Moreover, when queried about the potential apparatus used to take the photo, GPT-4V accurately discerns it as a fisheye camera.

V2X devices photos. V2X, which stands for Vehicle-to-Everything, encompasses a range of technologies that enable vehicles to communicate with not only each other but also with infrastructure and various other entities. V2X cameras play a pivotal role in capturing and processing visual information as part of this interconnected ecosystem. In Figure 13, we present the responses generated by GPT-4V for a drone-view photograph and two intersection camera images. GPT-4V exhibits commendable performance in all three instances. In the drone view, GPT-4V accurately identifies the freeway in both directions and recognizes

Sec. 2.2 Front view camera images

Prompt:

This is what you see behind the wheel. Describe the traffic scenario as detailed as possible.

GPT-4V:

You're approaching a city intersection with multiple pedestrian crosswalks. Several pedestrians are crossing the street in various directions. There's a sign indicating "NO TURN ON RED" and overhead traffic lights are showing red for your lane. [...] On the right, there are some buildings and construction cranes in the distance..

Prompt:

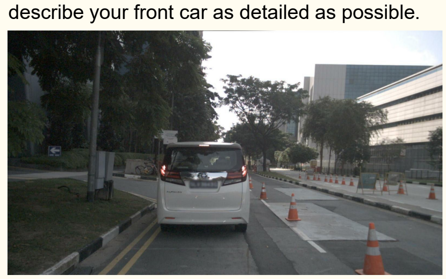
Please output the coordinate position or bounding box for each pedestrian.

GPT-4V:

I don't have the capability to provide real-time coordinates or bounding boxes for objects within images.

Prompt:

This is what you see behind the wheel. Please describe your front car as detailed as possible.



GPT-4V:

The car in front of you is a white vehicle, likely a hatchback or compact SUV [...] The license plate is visible but not readable in the picture. The rear lights are turned on, indicating either braking or nighttime running lights. The car also has a rearview camera located above the license plate.

Prompt:

Based on the image, what reason do you think caused the car in front braked?

GPT-4V:

The car in front might have braked due to the presence of traffic cones lining the road, indicating potential roadwork, lane closures, or a change in traffic flow. Another reason could be the presence of other vehicles ahead, though they are not visible in the image provided.

Figure 10: Results describing traffic participants via the front-view camera photo. Green highlights the right answer in understanding, Red highlights the wrong answer in understanding, Yellow highlights the incompetence in performing the task. Check Section A.2 for details.

Sec. 2.2 Front view camera images

Prompt:

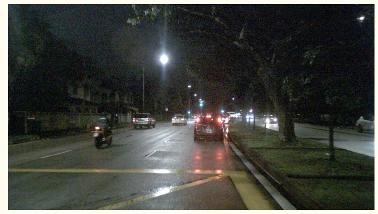
This is what you see behind the wheel. Please count how many vehicles are in the scenario and describe each vehicle.

GPT-4V:

I see 4 vehicles in the scenario. They are:
- A yellow car driving down the road.
- A dark-colored car parked on the right side of the road.
- A black SUV parked further ahead on the right.
- A larger black vehicle with an open cargo area, parked on the right side near the grassy area.

Prompt:

This is what you see behind the wheel. Please count how many vehicles are in the scenario and describe each vehicle.



GPT-4V:

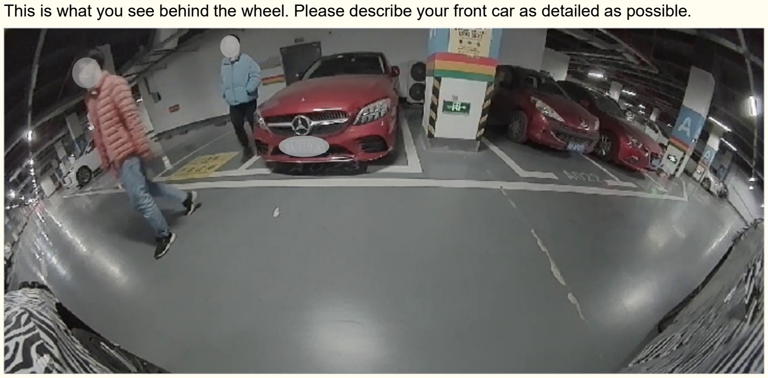
I see 6 vehicles in the scenario. They are:
- A motorcycle on the leftmost lane.
- A car parked on the left side of the road.
- Another car parked a bit ahead of the second car on the left side.
- A car in the middle of the road with its rear lights on.
- A car on the rightmost lane, moving.
- Another vehicle (possibly a car) on the far right, but only its headlights are visible.

Figure 11: Results describing traffic participants via the front-view camera photo. Green highlights the right answer in understanding, Red highlights the wrong answer in understanding, Yellow highlights that the result is correct but the details are wrong. Check Section A.2 for details.

Sec. 2.2 Fisheye camera image

Prompt:

This is what you see behind the wheel. Please describe your front car as detailed as possible.



GPT-4V:

You are in an indoor parking garage. Directly ahead, there's a red Mercedes-Benz car parked perpendicular to your direction of view. To the right of the Mercedes, there's a red, compact car parked facing the camera. Two men are walking in the garage; one is in a red jacket moving to the left and the other is in a blue jacket standing near the Mercedes. There's a designated pedestrian crossing area on the floor with white markings. To the right, there's a column with a green sign indicating a right turn, and a charging station for electric cars. The ceiling has lights and ventilation ducts. On the far right, there's an area marked "A" with arrows directing the traffic flow.

Prompt:

What kind of camera do you think that take this image?

GPT-4V:

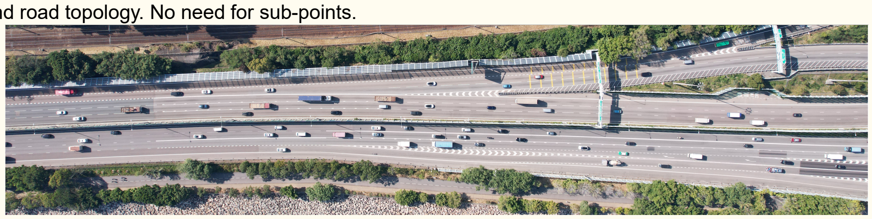
The image appears to be taken from a wide-angle or fisheye lens, common in security or dashboard cameras. Given the context of being behind the wheel, it's likely a dashboard camera.

Figure 12: Results describing traffic participants via the photos taken by a fisheye camera. Green highlights the right answer in understanding, Red highlights the wrong answer in understanding. Check Section A.2 for details.

Sec. 2.2 V2X images

Prompt:

This is a bird-eye view of a traffic scenario. Describe what you see, especially the traffic participants and road topology. No need for sub-points.

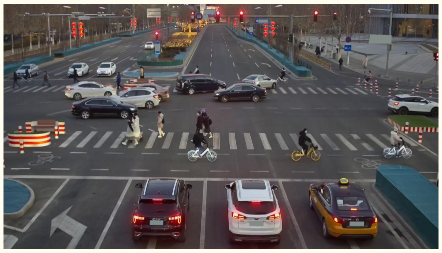


GPT-4V:

The image provides a bird-eye view of a multilane highway. Multiple vehicles, including cars, trucks, and vans, are scattered across the lanes, moving in both directions. Some vehicles are merging onto or exiting the highway via on and off-ramps. The highway has a central barrier separating the lanes of opposing traffic. On one side of the highway, there's a railway track, while on the other side, a rocky coastline can be observed. Pedestrian pathways and greenery are visible alongside the road. Junctions and intersections are marked with road signs, traffic lights, and lane indicators, making navigation easier for the drivers. There are no evident traffic congestions or incidents, and traffic appears to be flowing smoothly.

Prompt:

This is a photo from bird-eye view of a traffic camera. Describe what you see, especially the traffic participants, road topology and traffic lights (if applicable). No need for sub-points.



GPT-4V:

The image presents a bird-eye view of an urban intersection. Several cars, including a prominent yellow taxi, are at various positions on the road. Cyclists navigate the intersection, with bikes in colors like yellow and white. Pedestrians cross the streets using zebra crosswalks. The road design consists of multiple lanes, zebra crosswalks, and cyclist pathways marked by bicycle symbols. A center median with decorative lights splits the main traffic directions. Traffic lights, currently displaying red signals, are positioned around the intersection. Barriers and road dividers indicate possible roadwork or construction zones, and a blue structure is visible on one side.

Prompt:

This is a photo from bird-eye view of a traffic camera. Describe what you see, especially the traffic participants, road topology and traffic lights (if applicable). No need for sub-points.

GPT-4V:

The image captures a bird-eye view of an urban intersection. Cars are aligned, waiting for the green signal, with a lone motorbike, driven by a rider wearing a yellow helmet and carrying a rear box, crossing the intersection. A few pedestrians are on the sidewalks, some waiting to cross. The road design includes multiple lanes with zebra crosswalks, cyclist pathways marked by bicycle symbols, and various directional arrows. Traffic lights are visible around the intersection, with the majority showing a red signal. Barriers and road dividers are scattered throughout, indicating roadwork or designated pathways. There's a central median with trees and a blue-tarped area on one side of the road.

Figure 13: Results describing traffic participants via the V2X equipment's photos. Green highlights the right answer in understanding, Red highlights the wrong answer in understanding. Check Section A.2 for details.

the on-ramp situated on the right side of the photo. And in the intersection V2X device view, the response identifies a mixed traffic flow containing cars, cyclists, and pedestrians in the image, as well as accurate traffic light recognition.

Images taken in CARLA simulator. In the realm of autonomous driving research and development, simulators like CARLA serve as invaluable tools, providing a controlled and virtual environment where algorithms can be tested, trained, and refined before their deployment on real-world roads Dosovitskiy et al. (2017). We captured a series of images within CARLA's map of the Town 10, using the ego car as the primary viewpoint. Subsequently, we posed several questions based on these images, and the outcomes are showcased in Figure 14. GPT-4V not only identifies these images as originating from simulation software but also demonstrates a high level of awareness regarding the virtual vehicles and pedestrians within them. Furthermore, in rare instances where simulated pedestrians run red lights, GPT-4V appropriately acknowledges this scenario in its responses. However, it's worth noting that the model still struggles with recognizing traffic lights in simulation, like misidentifying red lights as yellow.

Sec. 2.2 Simulated images

Prompt:

This image is from front-view camera of your car.

1. Do you think this is a simulation or the real world?
2. Describe what you see, especially the traffic participants and road topology.
3. Do you think there's anything abnormal or need to pay attention in this situation?

GPT-4V:

1. It appears to be a simulation.
2. The scene depicts an urban intersection surrounded by tall buildings. The road is multi-laned and marked with yellow boxes, possibly indicating a "no block" zone. There are multiple traffic signals, some showing yellow. There are cars in the distance, and a few pedestrians on the sidewalks and crossing the road.
3. The traffic signals displaying yellow for different directions are unusual and might be a cause for confusion among drivers. Also, the presence of yellow box markings implies the need for drivers to ensure the exit is clear before entering to avoid gridlock.

Prompt:

This image is from front-view camera of your car.

1. Do you think this is a simulation or the real world?
2. Describe what you see, especially the traffic participants and road topology.
3. Do you think there's anything abnormal or need to pay attention in this situation?

GPT-4V:

1. It appears to be a simulation.
2. I see an urban intersection with multiple buildings, palm trees, pedestrian crossings, and traffic lights. There are several traffic participants including cars, a fire truck, and pedestrians. The road topology shows a multi-lane road intersecting with another, with traffic signals and crosswalks.
3. The traffic lights for our direction show both a green and a red arrow simultaneously, which is abnormal and could be confusing for drivers. Additionally, there's a pedestrian crossing the road while the traffic signal for vehicles seems green, which could be a potential safety concern.

Figure 14: Results describing traffic participants via the images taken in the CARLA simulator. Green highlights the right answer in understanding, Red highlights the wrong answer in understanding. Check Section A.2 for details.

B Advanced Capability of Reasoning

Reasoning is another important trait for proper driving behavior. Given the dynamic and unpredictable nature of traffic environments, drivers often encounter a range of unexpected events. It is imperative for proficient drivers to make accurate judgments and decisions, drawing on their experience and common sense when faced with such unforeseen circumstances. In this section, we have carried out a series of tests to evaluate GPT-4V's responses to unexpected events and its proficiency in navigating dynamic environments.

B.1 Corner Cases

In the ongoing research for autonomous driving, the common approach to enhancing the system's ability to handle unexpected events is through the continuous collection of data. However, given the dynamic, continuous, and random nature of driving environments, the data collected can only approximate its boundaries but never fully encapsulate it, that is,

Sec. 3.1 Corner cases

Prompt:

You're the mature driver behind the wheel and this image is what you see now. Describe what you see and what you plan to drive next.

GPT-4V:

As I observe the scene through the rain-speckled windshield, I see a blue truck to my left carrying a load under a white cloth. Directly ahead, there's an area cordoned off with traffic cones and barricades indicating some roadwork or maintenance activity. Near this section, there's a large orange industrial tank-like equipment with workers wearing safety gear actively engaged in their tasks. The road continues forward with buildings on both sides, and a few vehicles and pedestrians can be spotted in the distance. Overhead, trees line the road, providing a canopy.

Given the current conditions, I'd slow down due to the rain and potential slippery roads. I'd approach the work area with caution, ensuring I provide ample space for the workers. If the path is clear and there's no oncoming traffic, I'd move slightly to the left side to maintain a safe distance from the ongoing work. As I pass this area, I'd remain vigilant of the blue truck to my left, making sure it doesn't obstruct my path. After navigating this section, I'd continue driving, keeping in mind the importance of maintaining a safe speed given the rain and surrounding activities.

Prompt:

You're the mature driver behind the wheel and this image is what you see now. Describe what you see and what you plan to drive next

GPT-4V:

From behind the wheel, I see a large construction truck in the lane ahead of me. It's a prominent orange and yellow vehicle with noticeable safety markings, and it seems to be carrying construction equipment. Further ahead, there's a crossroads with traffic lights, and several vehicles are either stopped or in motion. On my right, there's a crane in the distance, indicating ongoing construction work in the area. In addition to cars, there are scooters and cyclists crossing the junction, suggesting that the area is bustling with mixed vehicular and pedestrian traffic.

With safety in mind, I plan to maintain a safe distance from the large construction vehicle and remain alert to the various road users around me. As the traffic light seems to be in favor of our direction, I'll smoothly accelerate once the truck ahead moves, ensuring I watch for any pedestrians or smaller vehicles that might dart out unexpectedly.

Figure 15: Illustration of GPT-4V's ability to make correct decisions in corner cases. Green highlights the right answer in understanding. Check Section B.1 for detailed discussions.

unexpected events are inevitable. Human drivers, equipped with common sense, are often able to improvise and navigate safely through these unforeseen circumstances. This highlights the importance of incorporating not just data-driven methods, but also the principles of reasoning and common sense into autonomous driving systems. The visual data employed here is drawn from CODA Li et al. (2022) and the internet.

In this section, we have carefully curated a set of perceptual corner cases to assess the model's capacity for common-sense reasoning. These examples deliberately include objects that fall out of the typical distribution, often posing challenges for conventional perception systems and creating difficulties in decision-making planning. Now, let's see how GPT-4V fares in addressing these cases.

On the left side of Figure 15, GPT-4V can clearly describe the appearance of the vehicles that are not commonly seen, the traffic cone on the ground, and the staff beside the vehicle. After identifying these conditions, the model realizes that the ego car can move slightly to the left, maintain a safe distance from the work area on the right, and drive cautiously. In the right example, GPT-4V adeptly identifies a complex traffic scenario, encompassing an orange construction vehicle, sidewalks, traffic lights, and cyclists. When queried about its driving strategy, it articulates an intention to maintain a safe distance from the construction

Sec. 3.1 Corner cases

Prompt:

You're the mature driver behind the wheel and this image is what you see now. Describe what you see and what you plan to drive next.

GPT-4V:

Driving on a clear day with a blue sky scattered with clouds, I suddenly come across an unexpected sight on the highway: a black airplane has made an emergency landing. The wings of the plane stretch across the lanes, and traffic has come to a halt. Nearby, vehicles, including a truck and a police car, are parked on the side of the road. Orange traffic cones have been placed around the area, indicating that the authorities are managing the situation.

As a mature driver, I recognize the importance of safety and patience in this situation. I would reduce my speed, put on my hazard lights, and carefully navigate around the incident, giving plenty of space to the airplane and any emergency personnel on the scene. Once past the airplane, I would continue on my journey, staying alert and vigilant, grateful for the training and experience that has taught me to handle unexpected situations on the road.

Prompt:

You're the mature driver behind the wheel and this image is what you see now. Describe what you see and what you plan to drive next.

GPT-4V:

Driving through the city at night, I'm closely following a large cement mixer truck, identifiable by its signature rotating drum and the chutes attached to the back. The truck's taillights and reflective markers offer a clear view despite the dim setting. To the left, a traffic light displays a green arrow, signaling it's safe to proceed in that direction, while the straight path shows a red light. The surrounding urban environment is illuminated [...]

Given the current scene, I would cautiously maintain a safe distance from the cement mixer truck, staying alert for any sudden stops or turns it might make. Seeing the green arrow to the left, I'd consider making a left turn if that's the direction I need to go, but if my destination requires me to continue straight, I'd patiently wait for the light to turn green, ensuring a safe and smooth journey ahead.

Figure 16: Illustration of GPT-4V's ability to make correct decisions in corner cases. Green highlights the right answer in understanding. Check Section B.1 for detailed discussions.

vehicle and, upon its passage, execute a smooth acceleration while conscientiously observing pedestrian presence.

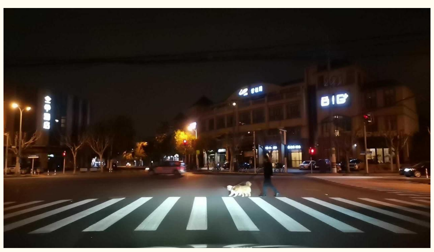
On the left side of Figure 16, GPT-4V can accurately identify that an airplane has made an emergency landing on the road and authorities are handling the situation nearby. For traditional perception algorithms, it is difficult to recognize without specific training. Under these conditions, the model knows that it should slow down and turn on its hazard lights while awaiting clearance to pass the aircraft before resuming regular driving. In the right example, GPT-4V accurately identifies the cement mixer truck and the red traffic light ahead. It discerns the importance of maintaining a safe distance from the truck in the lead until the red light transitions to green, at which point it proceeds with its journey.

On the left side of Figure 17, GPT-4V describes a scene where a pedestrian accompanied by two dogs is crossing the crosswalk, positioned in the center-right area of the image. Remarkably, the model accurately counts the number of dogs present. The model concludes

Sec. 3.1 Corner cases

Prompt:

You're the mature driver behind the wheel and this image is what you see now. Describe what you see and what you plan to drive next.



GPT-4V:

I see a nighttime city scene with well-lit streets. There are buildings on both sides of the road with illuminated signage. One of the signs reads "BIG" in bright white letters. On the road, there are pedestrian crosswalk lines. A person is walking across the crosswalk, accompanied by two dogs. Traffic lights are visible, but their current status isn't clear from the image.

Given the presence of the pedestrian and the dogs on the crosswalk, I would come to a complete stop if I haven't already. I would ensure the safety of the pedestrian and the dogs by waiting for them to completely cross the road before proceeding. Safety always comes first, especially in scenarios involving pedestrians at night.

Prompt:

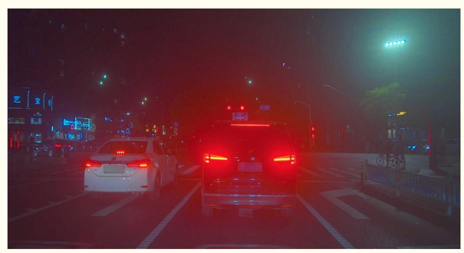
Can you tell me which part is the pedestrian on the picture? e.g. left-bottom, upper-right etc.

GPT-4V:

Certainly. The pedestrian in the picture is located towards the center-right area of the image, walking across the crosswalk.

Prompt:

You're the mature driver behind the wheel and this image is what you see now. Describe what you see and what you plan to drive next.



GPT-4V:

I see a nighttime city street scene. The view is from behind a large vehicle with its brake lights illuminated, suggesting that it is currently stopped. To the left of this vehicle is a white sedan. The surroundings have various illuminated signs, street lights, and building lights. On the right side of the image, there is a pedestrian crosswalk with a couple of people on bicycles waiting to cross. The traffic lights, which are visible on the right, are red, indicating that vehicles should be stopped.

Given the situation, as a mature driver, I would remain stationary with my foot on the brake pedal, ensuring that the car is at a complete stop. I would continue to monitor the traffic lights and wait for them to turn green before proceeding. I would also keep an eye on the cyclists and pedestrians to ensure their safety. Once the light turns green and it's safe to move forward, I would proceed with caution, being mindful of other vehicles, cyclists, and any pedestrians in the vicinity. Safety is paramount, especially at intersections and in nighttime conditions.

Figure 17: Illustration of GPT-4V's ability to make correct decisions in corner cases. Green highlights the right answer in understanding, Yellow highlights the incompetence in performing the task. Check Section B.1 for detailed discussions.

that the vehicle should patiently wait for the pedestrian and dogs to clear the way before resuming its journey, though it remains unable to ascertain the status of the traffic light. It's worth mentioning that this isn't the first occurrence of such an incident for GPT-4V. The image on the right depicts a nocturnal traffic scenario, which is well recognized by GPT-4V. In this instance, the model astutely discerned the illuminated brake lights of the vehicle ahead and noted the presence of pedestrians and cyclists, patiently waiting to cross the road. It intelligently deduced that it should remain stationary until the traffic signal turns green, exercising due caution when resuming its journey.

B.2 Multi-view Images

By utilizing multi-view cameras, GPT-4V captures a comprehensive view of the driving environment. Precisely interpreting the spatial relationships between these cameras and the overlapping areas within the images is crucial for the model to effectively harness the potential of the multi-view camera system. In this section, we evaluate GPT-4V's competence in handling multi-view images. All data in this section comes from the nuScenes Caesar et al. (2020) dataset.

In Figure 18, we select a set of surrounding images and input them to the model in the correct sequence. The model adeptly recognizes various elements within the scene, such as buildings, vehicles, barriers, and parking. It can even deduce from the overlapping information that there are two cars in the scene, with one white SUV positioned to the front and a truck with a trailer to the right. While the model's performance is generally impressive, an insignificant error is the misidentification of the pedestrian crossing.

Sec. 3.2 Multi-view images

Prompt:

You are a mature driver behind the wheel. These five images were captured and put together by the surround view camera in your vehicle. The FOV of each surround view camera is approximately 120° . And the middle image is the directly front.

Please describe in detail the scenario you are in, make a special distinction between what's in front of you and what's on either side of you!

GPT-4V:

You're situated in a semi-industrial or commercial area. In front of you is a broad street lined by multi-storied buildings, probably commercial or office spaces. Directly ahead, there's a white SUV parked beside a white-lined pedestrian crossing. The road has multiple lanes and is marked with clear road markings and arrows. Traffic cones are visible, indicating potential roadwork or lane closures.

To your left, there are large bay doors, possibly loading docks, with orange barriers in front. There are no vehicles or people in this immediate view, but there are parking spaces marked.

To your right, there's a large truck with a trailer, indicating that this area is used for deliveries or freight. The truck is parked or possibly waiting beside another set of bay doors or loading docks.

On either side of you, the architecture and design of the buildings appear consistent, suggesting that you are in a coordinated commercial complex or business park. The presence of the large truck and the bay doors confirm the commercial nature of this area.

Prompt:

Please count and describe how many vehicles are in the current scene.

GPT-4V:

There are 2 vehicles in the current scene: A white SUV parked on the street, seen in the middle image. A large truck with a trailer on the right side, which appears to be parked or waiting next to the bay doors.

Figure 18: Illustration of GPT-4V's ability to infer the relationship between multi-view images. Green highlights the right answer in understanding. Check Section B.2 for detailed discussions.

In another experiment shown in Figure 19, we similarly employ a set of combined surrounding images for test. While the model managed to provide a mostly accurate description of the scene, it also exhibited several recognition errors, particularly in regard to the number and shapes of the vehicles present. Notably, the model produces some confusing illusions, such as thinking there is a left-turn sign on the picture. We hypothesize that these issues may arise from the model's limited spatial reasoning capabilities.

Finally, in Figure 20, we present the correct front view and try to let GPT-4V identify and sort the out-of-order surrounding images. Although the model performs a lot of analysis and reasoning that seems to make sense, it still ends up outputting all wrong answers. It becomes evident that the model encounters challenges in establishing connections between adjacent images. We acknowledge the intricacies associated with this particular task.

Sec. 3.2 Multi-view images

Prompt:

You are a mature driver behind the wheel. These six images were captured and put together by the surround view camera in your vehicle. The FOV of each surround view camera is approximately 120° . And the middle two images are the directly front.

Please describe in detail the scenario you are in, make a special distinction between what's in front of you and what's on either side of you.

GPT-4V:

You're in a mostly empty urban area with roads bordered by structures and some greenery.

Directly in front: There's a relatively wide road with a few lane markings. Directly ahead, there's a large modern building with a unique architectural design featuring angled glass panels. To the left of the road, there are barriers indicating construction or some restricted area. There's a parked car on the right of these barriers. On the far left of the road, there's another street branching out.

To the left: The beginning of a pedestrian crosswalk is visible, leading towards a traffic island with some trees. On the far side, there's a street that turns to the left with another building in the background.

To the right: There's a continuation of the pedestrian crosswalk. Further right, a road intersects the one you're on, with a sign indicating a left turn. There are barriers similar to the ones directly in front of you. On the far right, you can see the corner of another building.

At the back: It seems like you've just passed a junction. There are more road barriers and markings. On the left side of the road, there's a structure that appears to be a bus stop or a shelter. Further back, there's another large building and a traffic light. On the right side, you can see a continuation of the road you're on, flanked by more buildings.

Prompt:

Please count and describe how many vehicles are in the current scene.

GPT-4V:

There are two vehicles in the current scene:

A red car parked on the right side, near the construction or restricted area barriers in front of you.

Another vehicle (due to the angle, specifics are hard to discern) is seen on the far right image, turning onto the road that intersects the one you're on.

Figure 19: Illustration of GPT-4V's ability to infer the relationship between multi-view images. Green highlights the right answer in understanding, Red highlights the wrong answer in understanding, Yellow highlights the incompetence in performing the task. Check Section B.2 for detailed discussions.

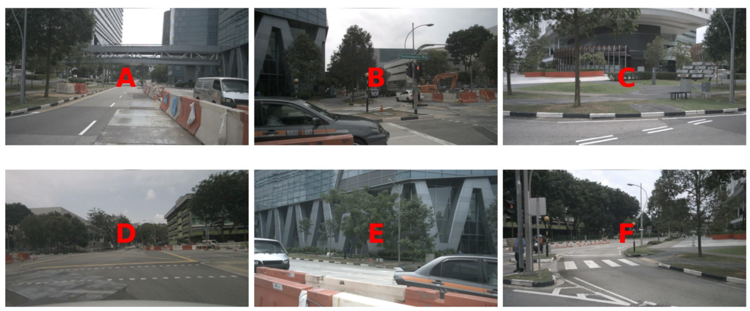
Sec. 3.2 Multi-view images

Prompt:

You are a mature driver behind the wheel. These six images were captured and put together by the surround view camera in your vehicle. The FOV of each surround view camera is approximately 120° .

Now you know this circular view that should consist of 'CAM_FRONT_LEFT', 'CAM_FRONT', 'CAM_FRONT_RIGHT', 'CAM_BACK_LEFT', 'CAM_BACK', 'CAM_BACK_RIGHT ' You only know that 'CAM_FRONT' is image A.

Please analyze which image corresponds to each of the other five views, and explain why.



GPT-4V:

Analyzing the images in reference to 'CAM_FRONT' which is image A:

'CAM_FRONT_LEFT': This would be the image to the immediate left of 'CAM_FRONT'. The scene should show a gradual transition from the front view to the side. This corresponds to image D, as it showcases a continuation of the road and barriers on the left side.

'CAM_FRONT_RIGHT': This would be to the immediate right of 'CAM_FRONT'. The scene should show a transition from the front view to the side view on the right. Image B fits this description, as it displays a continuation of the road, the vehicle's side, and barriers on the right.

'CAM_BACK_LEFT': Following the circular view, this would be after 'CAM_FRONT_LEFT'. It would show the scene behind the vehicle on the left. Image C is the correct match, displaying the rear view on the left side, with an intersection and a round building in view.

'CAM_BACK': This would be opposite 'CAM_FRONT' showing the full rear view. Image E is the match. It shows a scene directly behind the vehicle, with the back of other cars and barriers.

'CAM_BACK_RIGHT': This would be after 'CAM_FRONT_RIGHT' and before 'CAM_BACK'. It shows the scene behind the vehicle on the right side. Image F fits this description, with the road turning right and trees lining the side.

So, in order:

'CAM_FRONT_LEFT': Image D; 'CAM_FRONT_RIGHT': Image B; 'CAM_BACK_LEFT': Image C;

'CAM_BACK': Image E; 'CAM_BACK_RIGHT': Image F.

Figure 20: Illustration of GPT-4V's ability to infer the relationship between multi-view images. Red highlights the wrong answer in understanding. Check Section B.2 for detailed discussions.

B.3 Temporal Sequences

In this section, we assess the capability of GPT-4V in understanding temporal images. Our methodology involves the utilization of multiple sequences from first-person driving videos. From each video segment, we extract four keyframes, label them with sequential numbers, and combine them into a single image for input. Subsequently, we task GPT-4V with describing what events occurred during this time period, as well as the actions taken by ego vehicle and the reasons behind them. The examples are sourced from nuScenes Caesar et al. (2020), D^2-city and Carla Dosovitskiy et al. (2017) simulation.

Figure 21 shows a video captured in CARLA's map of Town 10 where the GPT-4V clearly explains the action of the ego car stopping at a crosswalk because of a pedestrian crossing the road, just before the traffic signal turning red.

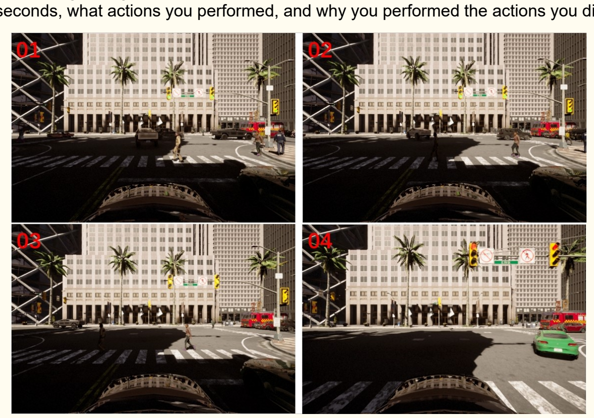
Figure 22 showcases a video segment extracted from the NuScene dataset Caesar et al. (2020). During the process of capturing keyframes, we assigned the labels "1" and "2" to the leading SUV and a pedestrian, respectively. GPT-4V not only responds accurately to inquiries regarding the objects represented by these labels but also offers a comprehensive explanation of the interaction observed between the front SUV and the pedestrian. This interaction entails the pedestrian crossing the street while the white SUV yields the right-of-way. The video sequence depicted in Figure 23 exhibits a noticeable camera flare caused by rain. Nevertheless, GPT-4V adeptly identifies the intersection signals and the front taillights. It also infers the semantic information about the start of the front car after the red light turns green.

However, it's important to note that GPT-4V is not always able to fully analyze temporal driving scenarios accurately. As illustrated in Figure 24, this video segment captures the vehicle making a lane change to overtake the scooter ahead. Regrettably, GPT-4V incorrectly interprets the scooter's actions as if it were crossing lanes in front of the vehicle and misconstrues the vehicle's behavior as decelerating to yield to the scooter. This underscores the limitations of GPT-4V in spatial reasoning within temporal video contexts. Also in Figure 23, GPT-4V once again misidentifies a green light as a red light.

Sec. 3.3 Temporal sequences

Prompt:

This is a time series of four pictures taken from a driving recorder. The chronological order is marked on the corresponding subfigure. As the driver of the car, please describe in detail what happened in these four seconds, what actions you performed, and why you performed the actions you did.



GPT-4V:

I approached an intersection with a green traffic light ahead. As I neared the crosswalk, I noticed pedestrians beginning to cross the street. I maintained my position, ensuring not to obstruct the crosswalk, and waited for the pedestrians to clear. The light turned yellow, then red, indicating for me to remain stopped. I obeyed the traffic signals and waited for the appropriate time to proceed.

Figure 21: Illustration of GPT-4V's ability to understand temporal sequential images. Green highlights the right answer in understanding. Check Section B.3 for details.

Sec. 3.3 Temporal sequences

Prompt:

This is a time series of four pictures taken from your driving recorder. The chronological order is marked on the corresponding subfigure. First please identify what objects are marked as 1 and 2.

GPT-4V:

Object 1 is a car and object 2 is a pedestrian.

Prompt:

Is object 1 the car you are driving?

GPT-4V:

No, object 1 is not the car I am driving; it is a car in front of the camera's viewpoint.

Prompt:

Is object 1 the car you are driving?

GPT-4V:

- In the first second (Time: 0 s), a car (object 1) is directly in front of the viewpoint, indicating that you're following it. Simultaneously, a pedestrian (object 2) is on the left side of the road, likely preparing to cross.
- By the second frame (Time: 1 s), the car (object 1) has moved slightly forward, and the pedestrian (object 2) is no longer visible from this angle. The pedestrian might have crossed or stayed on the sidewalk.
- In the third frame (Time: 2 s), the car (object 1) continues its trajectory, and the pedestrian (object 2) has appeared again on the right side of the road, having crossed it.
- By the fourth frame (Time: 3 s), the car (object 1) is farther ahead and the area seems clearer with no pedestrian in sight.
- Your actions as a driver:
    1. You followed the car (object 1) while maintaining a safe distance.
    2. As a pedestrian (object 2) was nearby and possibly crossing the street, you likely remained cautious and prepared to brake if necessary.
    3. Once the pedestrian had safely crossed, you continued to follow the car in front without any obstruction.
- Why you performed these actions:
    1. Maintaining a safe distance ensures you have ample time to react to unexpected events.
    2. Pedestrians have the right of way, so being cautious around them is essential for their safety and yours.
    3. With the path clear, continuing at a steady pace ensures the smooth flow of traffic.

Figure 22: Illustration of GPT-4V's ability to understand temporal sequential images. Green highlights the right answer in understanding. Check Section B.3 for detailed discussions.

Sec. 3.3 Temporal sequences

Prompt:

This is a time series of four pictures taken from a driving recorder. The chronological order is marked on the corresponding subfigure. As the driver of the car, please describe what happened in these four seconds, what actions you performed, and why you performed the actions you did.

GPT-4V:

During these four seconds, I was driving in rainy conditions.

At Time: 0 s, I was approaching a red traffic light and cars ahead were stopping. I began to slow down and maintain a safe distance from the car in front.

By Time: 1 s, I continued braking gradually as the car in front was still stationary. The traffic light was still red.

At Time: 2 s, the traffic light turned green, but I remained stationary, ensuring the cars in front started moving first.

By Time: 3 s, cars began to move, and I prepared to accelerate. I performed these actions to ensure safety and follow traffic rules.

Figure 23: Illustration of GPT-4V's ability to understand temporal sequential images. Green highlights the right answer in understanding, Red highlights the wrong answer in understanding. Check Section B.3 for detailed discussions.

Sec. 3.3 Temporal sequences

Prompt:

This is a time series of four pictures taken from a driving recorder. The chronological order is marked on the corresponding subfigure. As the driver of the car, please describe what happened in these four seconds, what actions you performed, and why you performed the actions you did.

GPT-4V:

In the four-second time series:

At 0 seconds, I was driving on a straight road with a car in front of me and saw a scooter rider approaching the divider.

By 1 second, I noticed the scooter rider making a move to cross the road, so I started to slow down and steer slightly to the left to create space and avoid any potential hazard.

At 2 seconds, the scooter rider was now crossing my lane, and I continued to maintain a safe distance from the car in front and the scooter.

By 3 seconds, the scooter rider had crossed to the other side, and I realigned my car to the center of the lane.

I slowed down and steered to avoid a potential collision with the scooter rider, ensuring both our safety and the safety of other road users.

Figure 24: Illustration of GPT-4V's ability to understand temporal sequential images. Green highlights the right answer in understanding, Red highlights the wrong answer in understanding. Check Section B.3 for detailed discussions.

B.4 Visual-Map Navigation

In practical driving scenarios, drivers often utilize auxiliary information from external devices to enhance their decision-making. For instance, a mapping application can provide detailed information about road geometry and route guidance, enabling drivers to make more informed and rational driving decisions. In this section, we equip GPT-4V with a front-view camera image and corresponding navigation information from the mapping software. This setup allows GPT-4V to describe the scene and make informed decisions, mirroring the way a human driver would in a similar situation.

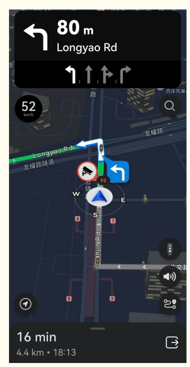
As depicted in Figure 25, GPT-4V accurately localizes its position using the front-view camera and map APP information, subsequently executing the correct left-turn maneuver. Utilizing the front view camera, it made reasonable assessments of the road conditions and, in conjunction with the speed information provided by the mapping software, offered appropriate driving suggestions.

As illustrated in Figure 26, GPT-4V is able to accurately localize itself even in a more complex scenario. However, in this particular situation, it incorrectly decided to make a left turn. Despite this, GPT-4V demonstrated proficiency in recognizing information about parked cars and stores on the roadside based on the front view camera. It also correctly obtained speed and distance information from the mapping software.

Sec. 3.4 Visual-map navigation

Prompt:

Please combine the navigation on map APP and the front-view camera image of the car. Write a paragraph to instruct how to drive in the current situation to the driver behind the wheel.



GPT-4V:

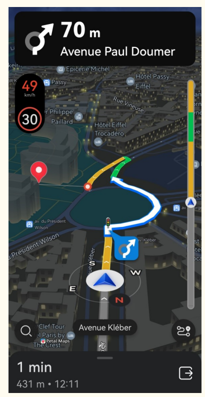
In your current position, you're approaching an intersection on Longyao Rd. You'll need to make a left turn in approximately 80 meters. As you can see from the front-view camera image of your car, the road ahead is clear with minimal traffic. Ensure you signal in advance and check both sides for oncoming traffic or pedestrians before making the turn. Once you've made the left turn onto Longyao Rd, continue straight and maintain a speed close to 52 km/h, but always adjust to the current traffic conditions and any speed limits posted. Your destination is roughly 4.4 km away and is estimated to take you another 16 minutes to reach. Safe driving!

Figure 25: Illustration of GPT-4V's ability to get information combines the navigation on map APP and the front-view camera image. Green highlights the right answer in understanding. Check Section B.4 for detailed discussions.

Sec. 3.4 Visual-map navigation

Prompt:

Please combine the navigation on map APP and the front-view camera image of the car. Write a paragraph to instruct how to drive in the current situation to the driver behind the wheel.



GPT-4V:

In your current location, you're on Avenue Kléber, approaching the intersection with Avenue Paul Doumer. In about 70 meters, make a slight left turn onto Avenue Paul Doumer. As you drive, keep an eye out for the prominent building on your right with the "CAFE KLÉBER" signage. Be mindful of the parked vehicles and pedestrians on both sides of the road. Additionally, the street view shows a few taxis and a scooter parked on the right, so exercise caution while maneuvering. Maintain a speed of around 30 km/h as indicated, although you're currently going at 49 km/h. You are expected to reach your destination in approximately 1 minute, covering a distance of 431 meters. Safe driving!

Figure 26: Illustration of GPT-4V's ability to get information combines the navigation on map APP and the front-view camera image. Green highlights the right answer in understanding. Check Section B.4 for detailed discussions.