On the Safety Challenges of Vision-Language Models in Autonomous Driving

Yang Qu[1], Lu Wang[1]
[1] School of Computer Science and Engineering, Beihang University, Beijing, China
{22373245,sy2406208}@buaa.edu.cn

Abstract

Vision-Language Models (VLMs) have demonstrated remarkable capabilities in multimodal perception and reasoning, which are widely applied in autonomous driving (AD) systems, significantly enhancing the intelligence and adaptability of autonomous vehicles. However, the integration of VLMs into AD systems introduces severe security risks, as these models are vulnerable to the complex environments and malicious attacks, which may lead to catastrophic failures in real-world driving scenarios. Despite the growing application of VLMs in AD (AD-VLMs), there remains a notable absence of systematic analysis of their safety challenges. To bridge this gap, we conduct the first comprehensive exploration that rigorously examines the safety landscape of AD-VLMs across 5 dimensions, including the inherent vulnerabilities and the external extreme conditions. This work not only highlights the urgent need for robust AD-VLMs but also provides research prospects to achieve trustworthy AD systems in complex open-world environments, aiming to inspire further safety research on AD-VLMs.

1. Introduction

Vision-Language Models (VLMs) have developed rapidly in recent years, demonstrating powerful capabilities in computer vision and natural language processing (NLP)[1, 2]. VLMs can interpret complex visual information with high precision, and process textual inputs in a context-aware manner, making them increasingly applicable across various domains, including autonomous driving (AD).

While traditional AD technology has limitations in multiple aspects, the appearance of VLMs provides a new solution to promote the development of AD, which significantly strengthens the ability of AD to understand complex traffic scenes and explain critical decisions.

At the same time, VLMs face numerous security challenges in practical applications, ranging from vulnerabilities in extreme environments to susceptibility to adversarial attacks, as shown in Fig. 1. These security challenges become particularly prominent when VLMs are applied to safety-critical AD systems. For example, a backdoor attack using a red balloon as a trigger could cause an autonomous vehicle to accelerate straight ahead instead of braking when encountering a girl with a red balloon crossing the street [3]. Therefore, the security issues of VLMs in AD (AD-VLMs) demand urgent attention.

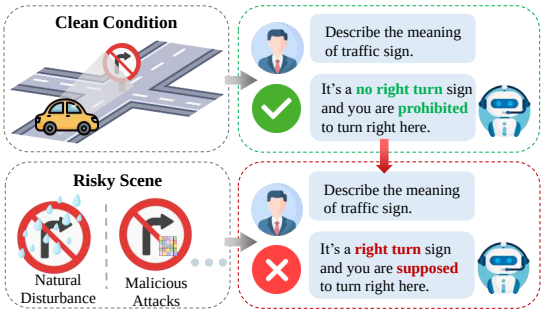

Figure 1. Risky scene leads AD-VLMs to output incorrect answer.

However, as far as we know, while existing researches have extensively investigated AD-VLMs applications[1, 2, 4], systematically analyzed VLM safety considerations[5, 6], and thoroughly examined traditional AD system security[7, 8], a critical gap persists in the literature: no comprehensive study has yet holistically examined the security risks arising from the unique integration of AD-VLMs. As shown in Fig. 2, our study categorizes the current safety challenges confronting AD-VLMs and make a discussion on research prospects. Our **contributions** are:

- To the best of our knowledge, we are the first to conduct an all-around synthesis of AD-VLMs safety challenges.
- We systematically summarize contemporary attack methodologies targeting AD-VLMs.
- We provide an in-depth discussion on the challenges and research prospects in the security of AD-VLMs.

2. AD-VLMs

VLMs integrate capabilities in both visual recognition and NLP, enabling end-to-end training through aligned image-text pairs. In this section, we summarize the applications of AD-VLMs according to the classification of open-loop and closed-loop systems.

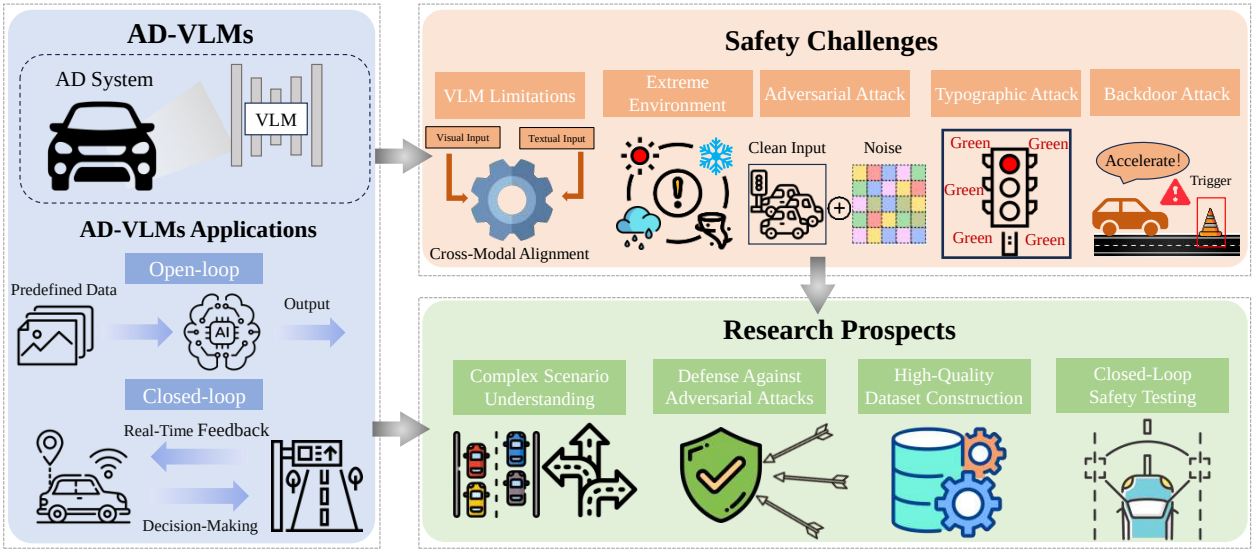

Figure 2. Overview of this work, including AD-VLMs applications, safety challenges and research prospects

Open-loop, which refers to a system that operates independently without real-time environmental feedback, executing actions or decisions based on static models, prior knowledge, and predefined data. Open-loop systems are usually employed for scene understanding, captioning, static decision making in VQA tasks[2]. For example, DriveLM[9] introduces GVQA to model AD reasoning as interconnected question-answer pairs structured in directed graphs; GPT-4V[10] demonstrates certain advantages in decision making under complicated traffic scenes; Dolphins[11] introduces a framework that processes video/text inputs and historical control signals to generate driving-specific instructions; HiLM-D[12] integrates high-resolution visual perception with language understanding to address small object detection challenges in AD scenarios.

Closed-loop, which refers to a system that dynamically interacts with its environment through continuous sensor information, enabling real-time adaptation of decisions or control signals. These feedback-driven architectures are particularly suited for AD applications due to their capacity for context-aware trajectory optimization and self-corrective behavior under dynamic conditions. Contemporary advancements in this domain, exemplified by LMDrive[13] and DriveMLM[14], demonstrate frameworks capable of processing multimodal sensory inputs while incorporating iterative environmental feedback to refine driving policies.

3. Challenges on AD-VLMs safety

The increasing application of AD-VLMs also brings a series of security challenges. In this section, we will discuss these challenges from five key aspects.

3.1. Limitations of VLMs

Although VLMs provide a innovative solution for AD systems to enhance AD ability, the limitations of VLMs themselves bring a series of challenges.

Fragile Cross-Modal Alignment: Subtle environmental changes (*e.g.*, glare, shadows) may decouple visual-language feature correlations, leading to misinterpretations.

Domain Generalization Gaps: Performance of VLMs degrades on out-of-distribution objects[15] (*e.g.*, non-standard traffic signs[10], temporary construction markers).

Temporal Inconsistency: VLMs may produce contradictory reasoning results across sequential frames (*e.g.*, misjudging pedestrian motion trajectories)[10].

3.2. Extreme Environmental Conditions

In the real-world traffic environment, complex and variable weather factors (*e.g.*, rain, snow, low-light condition), intricate road conditions (such as multi-way intersections and viaducts), and emergency situations (*e.g.*, wrong-way vehicles, suddenly appearing pedestrians, pets) pose great challenges to the perception and reasoning of AD-VLMs. These types of problems often account for a relatively small proportion in the training data of AD-VLMs, making it difficult to conduct sufficient training and testing which is the so-called long-tail problem in AD. When VLMs are actually applied to autonomous vehicles, these issues can cause significant potential safety hazards.

[10] demonstrates that the performance of GPT-4v significantly degrades when encountering complex environments. Currently, there is a lack of research on the performance of AD-VLMs in the face of long-tail problems and extreme environments. Meanwhile, the construction of high-quality and large-scale datasets is also insufficient.

3.3. Adversarial Attacks

Adversarial attacks inject imperceptible or structured perturbations at the pixel level into sensor inputs (*e.g.*, camera images, LiDAR point clouds) to degrade model performance. In the context of AD, incorrect classification or reasoning by the model often leads to serious traffic accidents.

In consideration of the fact that in AD-VLMs various textual instructions may convey the same semantics and visual driving scenarios have a time-series nature, [16] proposes ADvLM, an adversarial attack method which is the first to be specially designed for AD-VLMs. On one hand, large language models are used to generate a text instruction library with consistent semantics but diverse expressions to ensure the effectiveness of the attack under different text instructions. On the other hand, the image loss function is calculated through visual transformations, and the scene loss function is calculated by selecting key frames based on the model's attention, making the attack generalize across various driving scenarios.

To make the attack effective across the driving reasoning chain and in accord with the dynamic context in AD system, [17] proposes Cascading Adversarial Disruption (CAD) Attack, the first black-box attack targeting the breakdown of the driving decision chain. CAD Attack uses auxiliary VLMs to trace the possible causes of errors and construct a deceptive text chain, disrupting the driving perception-prediction-planning inference chain. It calculates the similarity between adversarial pictures and a set of opposite scene safety descriptions to induce risky scenarios.

3.4. Typographic Attacks

Typographic attacks add textual or visual-textual elements (*e.g.*, signs, billboards, in-vehicle displays) to target pictures to exploit vulnerabilities in AD-VLMs. These attacks aim to create semantic conflicts between visual inputs (*e.g.*, camera pictures, LiDAR data) and language reasoning, leading to misinterpretations or dangerous decisions.

For example, [18] introduces a pipeline that automatically generates adversarial texts and a directive strategy to augment the typographic attack. Taking advantage of prompt engineering, they guide LLM, *e.g.*, ChatGPT, to generate opposite answer about clean traffic scene as adversarial text and then add command directive (*e.g.*, "ANSWER:") or conjunction directive (*e.g.*, "AND", "OR", "WITH") to adversarial textual inputs to strengthen the effectiveness of the attack. At the digital level, attacks are carried out by embedding texts in images. At the physical level, misleading texts are added to elements in the background (such as streets and buildings) or foreground (such as vehicles and pedestrians) of traffic scenes.

PG-Attck[19] not only injects noise into visual input to generate perturbed images, but also embeds deceptive texts into images to attack. The PG-Attack framework integrates three sequential stages: modality expansion to generate masked images and captions, precision mask perturbation attack to maximize target region discrepancy while minimizing overall perturbation, and deceptive text patch attack to disrupt scene understanding, thereby enhancing attack effectiveness and stealthiness in AD-VLMs.

3.5. Backdoor Attacks

Backdoor attacks involve embedding specific triggers (*e.g.*, certain visual patterns or text instructions) into the training data, causing the model to output incorrectly when encountering these triggers during the inference stage. Such attacks pose a serious threat to AD-VLMs systems, potentially leading to vehicles misjudging traffic signals, ignoring obstacles or performing dangerous behaviors.

With the widespread adoption of VLMs in AD systems, the use of image-text pairs as training data has increased the likelihood of backdoor attacks that utilize concealed objects as visual triggers. [3] proposes BadVLMDriver, the first physical backdoor attack against AD-VLMs. An instruction-based image editing model is used to embed physical object triggers into images, and a LLM is employed to generate text responses containing the target backdoor behaviors. The AD-VLMs is fine-tuned on the backdoor training samples and their benign replays by minimizing the blending loss.

4. Outlook

Based on previous discussion, we find that the core safety issues revolve around environmental complexity, vulnerability to adversarial attacks, data limitations, and insufficient testing and validation. To address these issues and promote the safety of AD-VLMs, we expect that future researches focus on the following key directions.

Extreme and Complex Scenario Understanding. Future research could improve perception capabilities of AD-VLMs in extreme conditions (*e.g.*, heavy rain, fog, low light) by exploring multimodal data fusion such as visual (camera, LiDAR), linguistic (navigation instructions), and sensor (radar, IMU) data. Using physical engines (*e.g.*, CARLA[20]) to simulate complex road topologies could help AD-VLMs better understand multi-level traffic scene in open-world. The improvement of zero-shot learning capability is also a feasible solution to enable models to generalize to unseen scenarios.

Defense Against Adversarial Attacks. Leveraging existing attack methods to generate adversarial input and continuously inject new attack samples during training periods could enhance model robustness. Given that many attack methods aim to break multimodal consistency, real-time comparison of visual detection results and language reasoning outputs may defense such attack. We expect there

are more researches on defence methods such as adversarial sample detection to enhance the robustness of AD-VLMs.

High-Quality Datasets and Automated Scenario Construction. Existing datasets lack long-tail scenarios (*e.g.*, wrong-way vehicles, road collapses) limiting the model's ability to adapt to rare threats. It is necessary to explore traffic scene generator and leverage adversarial attack methods to automatically generate high-risk scenarios, handling the long-tail data problem.

Closed-Loop Safety Testing and Real-World Deployment Validation. To verify the dynamic and more realistic performance of AD-VLMs under closed-loop control, future testing frameworks should simulate perception-predition-planning chain in simulation environments (*e.g.*, CARLA[20]) based on real-time feedback. It is also significant to deploy AD-VLMs on physical world to identify potential danger or model failures.

In addition to addressing robustness and security challenges, designing lightweight model architectures and enhancing VLM inference speed through techniques like knowledge distillation are also critical to meet real-time decision-making requirements in AD scenarios[4].

5. Conclusion

To the best of our knowledge, this work represents the first systematic delineation and critical analysis of safety challenges inherent to AD-VLMs. In this paper, we categorize AD-VLMs into open-loop and closed-loop to make researchers easily have a glance at the application of VLMs in AD systems. Emphasizing the vulnerabilities of AD-VLMs and the potential danger, we systematically summarize the security challenges AD-VLMs face and conduct an in-depth discussion on research prospects. We hope this work fosters awareness, drives innovation in defense mechanisms, and ultimately contributes to the safe deployment of AD-VLMs in critical applications.

References

[1] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao, *et al.*, "A survey on multimodal large language models for autonomous driving," in *WACV*, 2024. 1

[2] X. Zhou, M. Liu, E. Yurtsever, B. L. Zagar, W. Zimmer, H. Cao, and A. C. Knoll, "Vision language models in autonomous driving: A survey and outlook," *TIV*, 2024. 1, 2

[3] Z. Ni, R. Ye, Y. Wei, Z. Xiang, Y. Wang, and S. Chen, "Physical backdoor attack can jeopardize driving with vision-large-language models," *arXiv preprint arXiv:2404.12916*, 2024. 1, 3

[4] H. Gao, Z. Wang, Y. Li, K. Long, M. Yang, and Y. Shen, "A survey for foundation models in autonomous driving," *arXiv preprint arXiv:2402.01105*, 2024. 1, 4

[5] M. Ye, X. Rong, W. Huang, B. Du, N. Yu, and D. Tao, "A survey of safety on large vision-language models: Attacks, defenses and evaluations," *arXiv preprint arXiv:2502.14881*, 2025. 1

[6] D. Liu, M. Yang, X. Qu, P. Zhou, Y. Cheng, and W. Hu, "A survey of attacks on large vision-language models: Resources, advances, and future trends," *arXiv preprint arXiv:2407.07403*, 2024. 1

[7] A. Kuznietsov, B. Gyevnar, C. Wang, S. Peters, and S. V. Albrecht, "Explainable ai for safe and trustworthy autonomous driving: a systematic review," *TITS*, 2024. 1

[8] K. Muhammad, A. Ullah, J. Lloret, J. Del Ser, and V. H. C. De Albuquerque, "Deep learning for safe autonomous driving: Current challenges and future directions," *TITS*, 2020. 1

[9] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving with graph visual question answering," in *ECCV*, 2024. 2

[10] L. Wen, X. Yang, D. Fu, X. Wang, P. Cai, X. Li, T. Ma, Y. Li, L. Xu, D. Shang, *et al.*, "On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving," *arXiv preprint arXiv:2311.05332*, 2023. 2

[11] Y. Ma, Y. Cao, J. Sun, M. Pavone, and C. Xiao, "Dolphins: Multimodal language model for driving," in *ECCV*, 2024. 2

[12] X. Ding, J. Han, H. Xu, W. Zhang, and X. Li, "Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving," *arXiv preprint arXiv:2309.05186*, 2023. 2

[13] H. Shao, Y. Hu, L. Wang, G. Song, S. L. Waslander, Y. Liu, and H. Li, "Lmdrive: Closed-loop end-to-end driving with large language models," in *CVPR*, 2024. 2

[14] W. Wang, J. Xie, C. Hu, H. Zou, J. Fan, W. Tong, Y. Wen, S. Wu, H. Deng, Z. Li, *et al.*, "Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving," *arXiv preprint arXiv:2312.09245*, 2023. 2

[15] H. Tu, C. Cui, Z. Wang, Y. Zhou, B. Zhao, J. Han, W. Zhou, H. Yao, and C. Xie, "How many unicorns are in this image? a safety evaluation benchmark for vision llms," *arXiv preprint arXiv:2311.16101*, 2023. 2

[16] T. Zhang, L. Wang, X. Zhang, Y. Zhang, B. Jia, S. Liang, S. Hu, Q. Fu, A. Liu, and X. Liu, "Visual adversarial attack on vision-language models for autonomous driving," *arXiv preprint arXiv:2411.18275*, 2024. 3

[17] L. Wang, T. Zhang, Y. Qu, S. Liang, Y. Chen, A. Liu, X. Liu, and D. Tao, "Black-box adversarial attack on vision language models for autonomous driving," *arXiv preprint arXiv:2501.13563*, 2025. 3

[18] N. Chung, S. Gao, T.-A. Vu, J. Zhang, A. Liu, Y. Lin, J. S. Dong, and Q. Guo, "Towards transferable attacks against vision-llms in autonomous driving with typography," *arXiv preprint arXiv:2405.14169*, 2024. 3

[19] J. Fu, Z. Chen, K. Jiang, H. Guo, S. Gao, and W. Zhang, "Pg-attack: A precision-guided adversarial attack framework against vision foundation models for autonomous driving," *arXiv preprint arXiv:2407.13111*, 2024. 3

[20] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *CoRL*, 2017. 3, 4