Proceedings of SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.TJ22524

Vision-Language Model Security in Autonomous Driving: A Survey

Junyi Wang

School of Cyber Science and Engineering, Sichuan University, Chengdu, China
wangjunyi1@stu.scu.edu.cn

Abstract: With the rapid advancement of Vision-Language Models (VLMs), their remarkable capabilities in multimodal perception and decision-making have garnered significant attention in autonomous driving. By integrating VLMs, autonomous driving systems can achieve a deeper understanding of their environment, thereby enhancing safety and efficiency. However, despite their advantages, the deployment of VLMs also introduces potential security vulnerabilities that pose critical challenges to real-world applications. They stem from the complex nature of multimodal processing, making VLMs susceptible to various adversarial manipulations. This paper presents a comprehensive and systematic review of various attack vectors targeting VLMs in autonomous driving, including adversarial attacks, backdoor attacks, jailbreak attacks, prompt injection attacks, zero-shot attacks, and hallucinations. Furthermore, we analyze defensive mechanisms against each attack type and discuss promising future research directions to bolster the robustness and security of VLMs in autonomous driving. By addressing these challenges, we believe that our survey provides insights to the development of safer autonomous driving systems, ensuring their reliability in practical deployments.

Keywords: Vision-Language Model, Autonomous Driving, Foundation Model Security, Security Attacks, Defense Mechanisms

1. Introduction

Autonomous driving represents a significant technological advancement that enhances modern transportation systems, providing increased convenience and fostering urban development. In recent years, the rapid progress of Vision-Language Models (VLMs) has played a crucial role in advancing autonomous driving [1]. VLMs, which process both visual and linguistic information, have proven highly effective in various applications, including perception and understanding [2-4], navigation and planning [5,6], decision-making and control [7], end-to-end autonomous driving [8,9], and data generation [10,11].

Given the stringent safety and reliability requirements of autonomous driving systems, any security vulnerabilities in VLMs could have severe consequences, including fatal accidents. Despite their promising capabilities, VLMs remain vulnerable to a variety of external threats and intrinsic weaknesses, stemming from the complexity of real-world driving environments and the inherent limitations of their multimodal nature.

VLMs integrate visual and linguistic information via pre-trained encoders, as depicted in Figure 1, which makes them susceptible to numerous attack vectors. For example, backdoor attacks involve

DOI: 10.54254/2755-2721/2025.TJ22524

embedding concealed triggers in VLMs, allowing malicious actors to manipulate the behavior of autonomous driving systems under specific conditions. Similarly, adversarial sample attacks introduce imperceptible perturbations that deceive VLMs into misclassifying objects, potentially creating safety hazards. Jailbreak attacks aim to bypass security alignment mechanisms, prompting VLMs to perform unauthorized operations, while prompt injection attacks manipulate model outputs by embedding malicious instructions into input queries. Furthermore, VLMs are prone to hallucinations—producing incorrect or misleading outputs—and zero-shot recognition failures, where the model misinterprets unfamiliar objects or scenarios.

To the best of our knowledge, we are the first to systematically review and categorize the security risks associated with VLMs in autonomous driving. In addressing this gap, we provide a structured examination of various attack methodologies and their corresponding defense strategies. Our contributions are as follows:

1. We present a comprehensive survey of security threats targeting VLMs in autonomous driving, categorizing the various attack methodologies.
2. We systematically review existing defense mechanisms aimed at mitigating these security threats.
3. We offer a detailed discussion of attack architectures and propose future research directions to enhance the security of VLMs in autonomous driving.

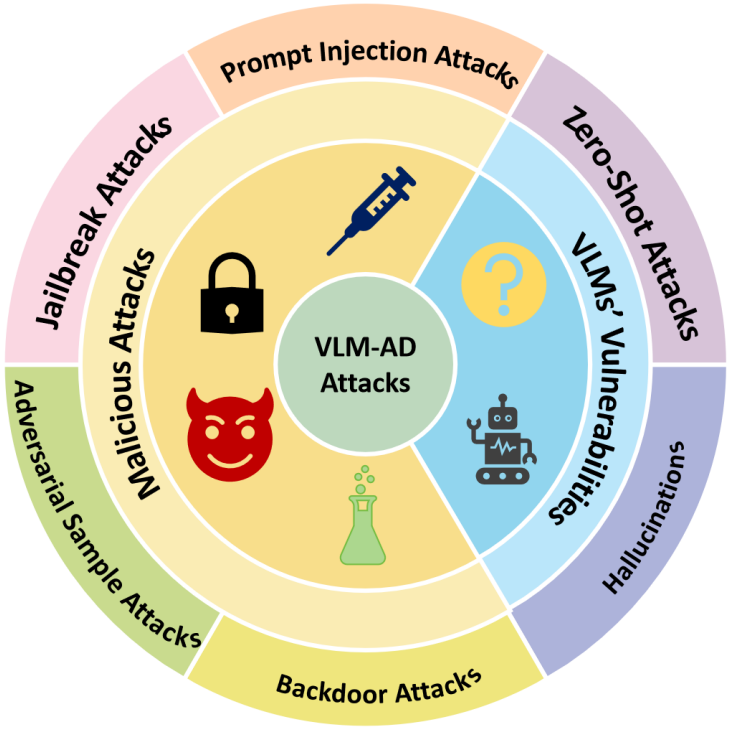

Figure 1: Overview of existing attacks on VLMs in autonomous driving

2. Attack methods

While Vision-Language Models (VLMs) have made significant strides in autonomous driving, their complexity introduces security vulnerabilities from malicious attacks and intrinsic model weaknesses. Malicious attacks, including backdoor, adversarial sample, jailbreak, and prompt injection attacks, exploit flaws during training or deployment, while intrinsic vulnerabilities, such as zero-shot failures

DOI: 10.54254/2755-2721/2025.TJ22524

and hallucinations, arise from the model's limited generalization ability. This section reviews these attack methods and their characteristics.

2.1. Malicious attacks

(1) Backdoor Attacks. Backdoor attacks involve injecting malicious data into the training dataset, which causes the VLM to embed triggers that prompt malicious behaviors. These backdoors are often invisible and difficult to detect, as they do not affect the model's performance on standard benchmark tests, thus posing a significant threat to autonomous driving safety.

Ni et al.[12] proposed BadVLMDriver, the first backdoor attack specifically targeting VLMs in autonomous driving. BadVLMDriver generates backdoor training samples that embed malicious behaviors using natural language instructions and fine-tunes the VLMs through visual instructions supported by playback data. Unlike existing VLM backdoor attacks that rely on digital modifications, BadVLMDriver utilizes physical objects, such as a football, to induce unsafe behaviors like sudden acceleration. Lyu et al.[13] introduced TrojVLM, which implements backdoor attacks by injecting specific triggers into the VLM's adapter module. This approach also proposes a semantic preservation loss function to ensure the model retains the original image semantics when generating target text. These backdoor attacks emphasize the potential security risks of VLMs and the importance of evaluating their security comprehensively.

(2) Adversarial Sample Attacks. Adversarial sample attacks manipulate VLM outputs by introducing imperceptible noise, which is specifically designed to exploit the model's vulnerabilities. These attacks are highly diverse and remain one of the most common forms of VLM attacks.

Zhang et al. [14] proposed ADvLM, the first adversarial attack framework for VLMs in autonomous driving. This method introduces semantic-invariant induction by constructing a text instruction library with low semantic entropy, which generates diverse prompts with consistent content. Additionally, scene association enhancement allows the attack to be effective in dynamic visual environments by selecting key frames and optimizing adversarial perturbations. Tu et al. [15] explored adversarial sample attacks targeting both the visual and language encoders. For the visual encoder, attackers use the PGD algorithm to generate noisy images, disrupting the VLM's visual understanding. For the language encoder, Large Language Models (LLMs) automatically generate malicious instructions that bypass security restrictions and lead the VLM to produce harmful outputs.

(3) Jailbreak Attacks. Jailbreak attacks target VLMs' security alignment mechanisms, causing them to generate filtered content or unauthorized actions. These attacks can compromise autonomous driving systems, leading to violations such as ignoring traffic lights and increasing safety risks.

Qi et al. [16] showed that visual adversarial examples could bypass VLM alignment. Using the PGD algorithm, their gradient-based attack enables visual inputs to manipulate text outputs. Shayegani et al. [17] proposed a cross-modality attack that exploits the embedding space of pre-trained encoders, pushing image embeddings toward target images to generate adversarial samples. Wu et al. [18] introduced SASAP, a method leveraging system prompt leakage in GPT-4V to generate jailbreak prompts, significantly improving attack success rates.

(4) Prompt Injection Attacks. Prompt injection attacks involve the use of specially crafted prompts to manipulate the VLM's output, causing the model to perform unexpected behaviors. These attacks can generate harmful content, leak private information, and violate the original intent of the model, thus posing serious threats to autonomous driving safety.

Maan Qraitem et al. [19] introduced a benchmark for testing typographic attacks against LVLMs and presented two self-generated attack types. Class-based attacks involve prompting the VLM to identify the class most similar to the target class, while descriptive attacks prompt the VLM to recommend typographic attacks involving both misleading classes and descriptions. Eugene Bagdasaryan et al. [20] embedded adversarial perturbations into image inputs to implicitly inject

DOI: 10.54254/2755-2721/2025.TJ22524

prompts and instructions, thereby indirectly controlling the VLM's output. When users query the perturbed image, the VLM is guided to output text selected by the attacker, allowing for full manipulation of the VLM.

2.2. VLMs' vulnerabilities

(1) Zero-Shot Attacks. Zero-shot attacks target the VLM's performance when handling out-of-distribution (OOD) samples, which are typically misclassified or generate biased outputs, despite the VLM performing well with in-distribution (ID) samples.

Autonomous driving systems often rely on open-source VLMs, and Nathan Inkawhich et al.[21] proposed a zero-shot attack method based on noise perturbation adversarial algorithms. In the feature space of a visual foundation model, each object type has its own feature space range. The Away From Start (AFS) attack deceives the VLM by perturbing images so that they deviate from their clean representation. Alternatively, attackers can use a noise image optimization algorithm to adjust the OOD sample pair features to match the ID category's target features, thus misclassifying it as an ID category.

(2) Hallucinations. Hallucinations occur when VLM outputs fail to align with input prompts. This phenomenon, resulting from the integration of text and image inputs, can lead to low-quality outputs that introduce bias into the decision-making of autonomous driving systems, which are often difficult for humans to detect. There are three primary scenarios for hallucinations[22]: absent answers, incompatible answers, and irrelevant visual questions. The first two scenarios involve a failure to provide the correct answer, while the latter refers to questions that are irrelevant to the provided image.

These hallucinations primarily arise from two sources: uneven data quality and VLM forgetfulness. Training data may contain issues such as duplication[23] and bias[24], while the VLM's acquisition of new knowledge can lead to the forgetting of old knowledge[25]. This highlights the inherent vulnerabilities of VLMs and underscores the need for corresponding defensive measures.

3. Defenses against attacks

Despite significant advancements in autonomous driving, the integration of visual modules in Visual Language Models (VLMs) introduces new vulnerabilities, challenging the system's robustness. As attack strategies evolve, developing effective defense mechanisms has become crucial. These defenses aim to detect malicious data manipulations during training, enhance model robustness, and address issues like hallucinations and zero-shot failures.

3.1. Malicious attacks

(1) Backdoor Attacks**.** Backdoor attacks leverage the inclusion of malicious data during the training phase to embed triggers that induce harmful behavior in the model. In the context of autonomous driving, such attacks can lead to incorrect and unsafe decision-making. Defenses against backdoor attacks primarily focus on identifying and eliminating these hidden triggers.

For example, Zhang et al.[26] introduced the Repulsive Visual Prompt Tuning (RVPT) method, which combines deep visual prompt tuning, feature-repelling loss, and cross-entropy loss. RVPT helps the model learn features that are directly relevant to the task while rejecting irrelevant features, thus mitigating the impact of backdoor triggers. Liang et al.[27] proposed a model unlearning strategy, where suspicious samples are overfitted to expose backdoor features. The local token unlearning strategy allows for the elimination of backdoor associations without compromising the model's overall performance on clean data.

(2) Adversarial Sample Attacks. Adversarial sample attacks involve the introduction of subtle perturbations to the model's input to mislead its output. These attacks can significantly degrade the

DOI: 10.54254/2755-2721/2025.TJ22524

model's performance, especially in complex applications like autonomous driving. Defense strategies for adversarial attacks aim to enhance the model's ability to identify and reject malicious inputs.

Li et al.[28] proposed Adversarial Prompt Tuning (APT), which involves learning robust text prompts that guide the model towards stable and accurate outputs even when confronted with adversarial samples. The core concept behind APT is to use the semantic structure of the prompts to stabilize the model's response under adversarial conditions. Additionally, Nie et al.[29] introduced the DiffPure algorithm, which applies a diffusion process to perturb images before feeding them into the visual module, effectively neutralizing adversarial noise and recovering clean image representations.

(3) Jailbreak Attacks. Jailbreak attacks aim to bypass the security mechanisms of VLMs, allowing them to output unauthorized or filtered content. These attacks can have severe consequences, such as enabling malicious behavior in autonomous driving systems. Defense mechanisms against jailbreak attacks typically focus on detecting malicious inputs or securing the output through multi-stage verification processes.

Liu et al.[30] proposed SafeVLM, an augmented security alignment method that integrates three security modules and a two-stage training process. The safety projector extracts risk features from inputs, while safety tokens convey security information to the model. The safety head interacts with visual perception to ensure the integrity of the outputs. Zhou et al.[31] introduced a defense strategy based on multi-agent debate. This method uses two agents: an integrated agent that processes the full image and a partial agent that processes a portion of the image. During their debate, the integrated agent persuades the partial agent to accept the correct output, even when under attack, thus providing a self-check mechanism that filters harmful content.

(4) Prompt Injection Attacks. Prompt injection attacks manipulate the input prompts to induce the model to generate unintended or harmful outputs. Defending against such attacks requires identifying and mitigating abnormal patterns in the prompts to prevent malicious manipulation.

Sun et al.[32] proposed SmoothVLM, a defense mechanism that utilizes input image masking and a majority voting system to counteract prompt injection. By generating multiple copies of the input image with varying masks, SmoothVLM ensures that the final output is determined by the majority of clean inputs, reducing the impact of injected prompts. Chen et al.[33] developed PRIVQA, a multimodal benchmark that incorporates self-regulation techniques to protect against prompt injections. This system allows the model to autonomously review and validate its responses to ensure that malicious prompts do not influence its decision-making.

3.2. VLMs' vulnerabilities

(1) Zero-Shot Attacks. Zero-shot attacks exploit the limitations of VLMs in processing out-of-distribution (OOD) samples, causing the model to misclassify these inputs. Zero-shot attacks are particularly problematic in autonomous driving, as they can lead to misinterpretations of novel or unseen situations. Defense strategies against zero-shot attacks typically focus on enhancing the model's ability to distinguish between in-distribution (ID) and OOD samples.

Shu et al.[34] proposed SimLabel, a post-hoc strategy that improves the separability between ID and OOD samples by generating semantic labels for each ID class and measuring the consistency of image features across similar labels. Miyai et al.[22] proposed both training-free and training-based methods to counter zero-shot attacks. The training-free method adds additional instructions to guide the model to refuse answering unanswerable questions, while the training-based method involves incorporating unanswerable questions into the training data to teach the model to identify and reject such queries.

(2) Hallucinations. Hallucinations in VLMs refer to instances where the generated output does not accurately reflect the input prompts, leading to biased or incorrect decision-making. Although

DOI: 10.54254/2755-2721/2025.TJ22524

detecting hallucinations is relatively straightforward for humans, mitigating them remains challenging. Several defense strategies have been proposed to reduce hallucinations in VLM outputs.

Anisha Gunjal et al.[35] constructed a multimodal hallucination detection dataset with fine-grained annotations and used reinforcement learning-based approaches, such as Fine-grained Direct Preference Optimization (FDPO) and rejection sampling, to train a multimodal reward model for hallucination detection. Sun et al.[36] proposed a factually augmented RLHF (Reinforcement Learning from Human Feedback) method, in which human annotators compare model responses to identify more factual outputs. This feedback is then used to fine-tune the model to generate more accurate and reliable outputs, thus reducing the occurrence of hallucinations.

4. Future directions

Although current research has made significant strides in exploring various attack and defense strategies for Visual Language Models (VLMs) in the context of autonomous driving, there remains substantial room for further advancements. Based on the categorization and synthesis of existing works, we highlight several key future research directions that warrant attention.

(1) Enhancing Generalization Capabilities in Long-Tail Scenarios. Existing studies indicate that VLMs exhibit limited performance in handling zero-shot problems. This limitation is particularly pronounced when autonomous driving systems encounter long-tailed scenarios, such as rare objects or extreme weather conditions, which can lead to poor decision-making and compromised safety. To address these challenges, future research should focus on integrating small-sample learning and continuous learning techniques. Approaches like incremental learning and meta-learning could improve the ability of VLMs to handle unknown scenarios more effectively. Moreover, researchers should develop high-quality simulation platforms capable of generating extensive long-tailed datasets to bolster the generalization capabilities of VLMs, ensuring the continued safety and reliability of autonomous driving systems.

(2) Developing Robust Fine-Tuning Techniques. Catastrophic forgetting, the phenomenon wherein a model's performance deteriorates upon fine-tuning, remains a challenge in ensuring that VLMs retain their original capabilities while enhancing safety. This issue is particularly evident in VLMs used for autonomous driving, where improving safety may inadvertently compromise the model’s core functions [37]. To mitigate this, future research should focus on a broader range of fine-tuning methodologies. For instance, Reinforcement Learning from Human Feedback (RLHF), wherein reward models are trained based on human expert feedback, can help VLMs learn safety-compliant behaviors while maintaining their original functionalities. This approach offers the potential to reduce biases and preserve the flexibility of the model.

(3) Enhancing Cross-Modality Alignment. While many defense strategies have been developed for individual vision or text modules in VLMs, these approaches often fail to address the complexities associated with multimodal threats. For example, images and text that are individually benign may, when combined, produce unsafe or unethical outputs [38]. Current defense methods typically treat different modalities as independent and design defenses in isolation. Future work should focus on developing new defense mechanisms that explore the interactions and relationships between modalities. Such methods would enable more robust responses to multimodal attacks, improving the overall security and integrity of VLMs in autonomous driving applications.

5. Conclusion

This survey provides a comprehensive analysis of the security landscape surrounding Visual Language Models (VLMs) in the context of autonomous driving, highlighting the various attacks VLMs face and corresponding defense mechanisms. We begin by introducing the widespread

DOI: 10.54254/2755-2721/2025.TJ22524

applications of VLMs in autonomous driving and the associated security threats. Subsequently, we categorize existing attack literature into a novel taxonomy, distinguishing between malicious attacks and model vulnerabilities, which facilitates a structured understanding of these challenges. Malicious attacks encompass backdoor attacks, adversarial sample attacks, jailbreak attacks, and prompt injection attacks, while model vulnerabilities include zero-shot attacks and hallucinations. Additionally, we review the defense strategies proposed for mitigating each of these attack types. Finally, we identify several promising future directions for advancing VLM security in autonomous driving. In conclusion, while VLMs have demonstrated remarkable potential in enhancing autonomous driving capabilities, ensuring their security remains paramount. We hope this survey offers valuable insights for researchers and encourages further in-depth exploration of this critical area.

References

[1] *Zhou, Xingcheng, et al. "Vision language models in autonomous driving: A survey and outlook." IEEE Transactions on Intelligent Vehicles (2024).*

[2] *Zhang, Beichen, et al. "Long-clip: Unlocking the long-text capability of clip." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.*

[3] *Liu, Mengyin, et al. "Vlpd: Context-aware pedestrian detection via vision-language semantic self-supervision." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.*

[4] *Zhang, Jiacheng, et al. "A multi-granularity retrieval system for natural language-based vehicle retrieval." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.*

[5] *Tian, Xiaoyu, et al. "Drivevlm: The convergence of autonomous driving and large vision-language models." arXiv preprint arXiv:2402.12289 (2024).*

[6] *Keysan, Ali, et al. "Can you text what is happening? integrating pre-trained language encoders into trajectory prediction models for autonomous driving." arXiv preprint arXiv:2309.05282 (2023).*

[7] *Wang, Pengqin, et al. "Bevgpt: Generative pre-trained foundation model for autonomous driving prediction, decision-making, and planning." IEEE Transactions on Intelligent Vehicles (2024).*

[8] *Xu, Zhenhua, et al. "Drivegpt4: Interpretable end-to-end autonomous driving via large language model." IEEE Robotics and Automation Letters (2024).*

[9] *Pan, Chenbin, et al. "Vlp: Vision language planning for autonomous driving." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.*

[10] *Hu, Anthony, et al. "Gaia-1: A generative world model for autonomous driving." arXiv preprint arXiv:2309.17080 (2023).*

[11] *Jia, Fan, et al. "Adriver-i: A general world model for autonomous driving." arXiv preprint arXiv:2311.13549 (2023).*

[12] *Ni, Zhenyang, et al. "Physical backdoor attack can jeopardize driving with vision-large-language models." arXiv preprint arXiv:2404.12916 (2024).*

[13] *Lyu, Weimin, et al. "Trojvlm: Backdoor attack against vision language models." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.*

[14] *Zhang, Tianyuan, et al. "Visual Adversarial Attack on Vision-Language Models for Autonomous Driving." arXiv preprint arXiv:2411.18275 (2024).*

[15] *Tu, Haoqin, et al. "How many unicorns are in this image? a safety evaluation benchmark for vision llms." arXiv preprint arXiv:2311.16101 (2023).*

[16] *Qi, Xiangyu, et al. "Visual adversarial examples jailbreak aligned large language models." Proceedings of the AAAI conference on artificial intelligence. Vol. 38. No. 19. 2024.*

[17] *Shayegani, Erfan, Yue Dong, and Nael Abu-Ghazaleh. "Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models." arXiv preprint arXiv:2307.14539 (2023).*

[18] *Wu, Yuanwei, et al. "Jailbreaking gpt-4v via self-adversarial attacks with system prompts." arXiv preprint arXiv:2311.09127 (2023).*

[19] *Qraitem, Maan, et al. "Vision-llms can fool themselves with self-generated typographic attacks." arXiv preprint arXiv:2402.00626 (2024).*

[20] *Bagdasaryan, Eugene, et al. "Abusing images and sounds for indirect instruction injection in multi-modal LLMs." arXiv preprint arXiv:2307.10490 (2023).*

[21] *Inkawhich, Nathan, Gwendolyn McDonald, and Ryan Luley. "Adversarial attacks on foundational vision models." arXiv preprint arXiv:2308.14597 (2023).*

DOI: 10.54254/2755-2721/2025.TJ22524

[22] Miyai, Atsuyuki, et al. "Unsolvable problem detection: Evaluating trustworthiness of vision language models." arXiv preprint arXiv:2403.20331 (2024).
[23] Kandpal, Nikhil, et al. "Large language models struggle to learn long-tail knowledge." International Conference on Machine Learning. PMLR, 2023.
[24] Venkit, Pranav Narayanan, et al. "Nationality bias in text generation." arXiv preprint arXiv:2302.02463 (2023).
[25] Yang, Dingchen, et al. "Pensieve: Retrospect-then-compare mitigates visual hallucination." arXiv preprint arXiv:2403.14401 (2024).
[26] Zhang, Zhifang, et al. "Defending Multimodal Backdoored Models by Repulsive Visual Prompt Tuning." arXiv preprint arXiv:2412.20392 (2024).
[27] Liang, Siyuan, et al. "Unlearning backdoor threats: Enhancing backdoor defense in multimodal contrastive learning via local token unlearning." arXiv preprint arXiv:2403.16257 (2024).
[28] Li, Lin, et al. "One prompt word is enough to boost adversarial robustness for pre-trained vision-language models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
[29] Nie, Weili, et al. "Diffusion models for adversarial purification." arXiv preprint arXiv:2205.07460 (2022).
[30] Liu, Zhendong, et al. "Safety alignment for vision language models." arXiv preprint arXiv:2405.13581 (2024).
[31] Zhou, Qi, et al. "Defend against Jailbreak Attacks via Debate with Partially Perceptive Agents."
[32] Sun, Jiachen, et al. "Safeguarding vision-language models against patched visual prompt injectors." arXiv preprint arXiv:2405.10529 (2024).
[33] Chen, Yang, et al. "Can language models be instructed to protect personal information?." arXiv preprint arXiv:2310.02224 (2023).
[34] Zou, Shu, et al. "SimLabel: Consistency-Guided OOD Detection with Pretrained Vision-Language Models." arXiv preprint arXiv:2501.11485 (2025).
[35] Gunjal, Anisha, Jihan Yin, and Erhan Bas. "Detecting and preventing hallucinations in large vision language models." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 16. 2024.
[36] Sun, Zhiqing, et al. "Aligning large multimodal models with factually augmented rlhf." arXiv preprint arXiv:2309.14525 (2023).
[37] Zhai, Yuexiang, et al. "Investigating the catastrophic forgetting in multimodal large language models." arXiv preprint arXiv:2309.10313 (2023).
[38] Wang, Siyin, et al. "Cross-modality safety alignment." arXiv preprint arXiv:2406.15279 (2024).