













Large Vision-Language Model Security: A Survey

Taowen Wang¹, Zheng Fang², Haochen Xue², Chong Zhang²,
Mingyu Jin³, Wujiang Xu³, Dong Shu⁴, Shanchieh Yang¹,
Zhenting Wang³, and Dongfang Liu¹(✉)

¹ Rochester Institute of Technology, Rochester, USA
{tw9146,dongfang.liu}@rit.edu

² University of Liverpool, Liverpool, UK

³ Rutgers University, New Brunswick, NJ, USA

⁴ Northwestern University, Evanston, IL, USA

Abstract. In the domain of Large Vision-Language Models (LVLMs), securing these models has emerged as a critical issue for both researchers and practitioners. In this paper, we highlight and analyze the security-related issues of LVLMs, with a special emphasis on the reliability challenges in practical deployments. We begin by reviewing recent studies on threats like jailbreak and backdoor attacks, alongside discussing the potential countermeasures implemented to mitigate these risks. Additionally, we touch on real-world application problems, such as hallucinations and privacy leakages, as well as the ethical and legal related researches around them. We also outline the shortcomings observed in current studies and discuss directions for future research, with the aim of promoting LVLMs towards a safer direction. A curated list of LVLMs-security-related resources is also available at <https://github.com/MingyuJ666/LVLM-Safety>.

Keywords: Large Vision-Language Models · Security · Jailbreak · Hallucination · Privacy

1 Introduction

Large Vision-Language Models (LVLMs) have made significant advancements in Artificial General Intelligence (AGI) by demonstrating the ability to process and integrate vision and language information. This breakthrough has unlocked new and innovative opportunities across various applications [7, 22, 52, 88, 90], enabling LVLMs to effectively perform multi-modal conversation and visual question-answering tasks. However, recent studies have revealed that LVLMs exhibit a concerning characteristic - adversarial vulnerability. This vulnerability, which has been observed in classical deep learning models [17–19, 29, 62], is now also present in LVLMs [9, 65, 70, 79, 83, 99]. This poses a notable challenge to the

T. Wang, Z. Fang, H. Xue—Equal contribution.

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
B. Chen et al. (Eds.): FCS 2024, CCIS 2315, pp. 3–22, 2024.
https://doi.org/10.1007/978-981-96-0151-6_1

security and reliability of LVLMs when deployed in real-world applications, as they can be susceptible to adversarial attacks, similar to classical deep learning models.

This paper aims to analyze and categorize the security risks of LVLm in a broad sense. Review the attack assessment methods of researchers on LVLm vulnerabilities and efforts in mitigating security risks. In summary, the review scope is shown in Fig. 1, and we will discuss the following issues:

- *Malicious Attacks on LVLms*: We investigate three primary malicious attacks against LVLms: Jailbreak, Backdoor attacks, and Controllable misinformation generation. Jailbreak attacks induce LVLms to generate content that violates security restrictions by manipulating input, usually with fixed model parameters, and directing the model to output malicious content through malicious manipulation of input images. Backdoor attacks, conversely, implant triggers during the model training phase that cause the model to produce harmful responses when specific trigger patterns are encountered. Furthermore, LVLms also possess the capability to generate controllable misinformation that both humans and detection systems struggle to identify.
- *Defenses against LVLms Malicious Attacks*: We review a variety of defense avenues that cover input preprocessing techniques, such as identifying potentially adversarial samples by transforming the input image and output monitoring techniques. In addition, with regard to the backdoor attack defense method, improving the training process and parameter tuning of models to enhance their resistance to malicious manipulation are discussed. Finally, for controllable misinformation generation, some methods for detecting misinformation using LVLms were investigated, and concluded that the robust multi-modal capabilities of LVLms can be effectively harnessed to detect misinformation.
- *Application Risks and Mitigation Methods*: Regarding application risks, we discuss issues that LVLms may arise in real-world applications, namely Hallucinations and Privacy. The hallucination describes LVLms generating responses that do not correspond to facts or user prompts, which can pose potential risks in domains closely related to human life, such as medical advice and legal assistance. The problem of privacy leakage, on the other hand, involves the leakage of personal information, and we review the work related to the privacy of LVLms and discuss future directions for privacy research on LVLms.

To better understand the security risks and causes of LVLms, this paper first explores the architectural design of LVLms, the training mechanism, and the efforts in alignment. Then, we review recent research results on jailbreak attacks, backdoor attacks, and other potential threats and analyze the defense strategies carried out by researchers to address these challenges. In addition, this paper discusses the reliability issues that LVLms may raise in real-world applications, i.e., hallucination and privacy leakage. It emphasizes the importance of prospective research and development to ensure that deploying these advanced

technologies complies with ethical norms and legal frameworks. Finally, we summarize the limitations of the current research and discuss future work.

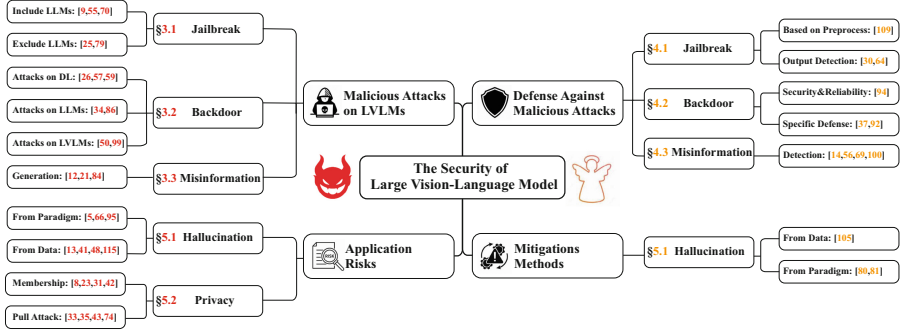


Fig. 1. Categorization of Malicious Attack, Defenses, Application Risk and Mitigation methods in LVLMs. Each branch shows relevant recent research regarding the topic.

2 Architecture and Training of LVLMs

Architecture: The overall architecture of Large Vision-Language Models (LVLMs), as shown in Fig. 2, consists of four key components: Visual Encoder, Text Encoder, Connector, and LLM [73]. This modular design allows for the seamless integration of visual and textual information, facilitating the ability of LVLMs to process and generate responses for a wide range of multimodal tasks. Specifically, the text encoder embeds a text into embedding to obtain a better understanding of the intrinsic relationships between the tokens. Simultaneously, variants of CLIP [72] are adopted as vision embedded for better alignment between text and vision embedding due to the superior cross-modality understanding ability. To further align vision embeddings from the output of the vision encoder with the text embedding, various works [3,15,78] adopt a connector to push two embedding spaces closer. Furthermore, based on the MLP connector, multiple works have proposed improved methods to enhance the alignment capability of the connector, such as cross-attention mechanisms [2], specialized adapters [24] and q-formers [47]. Finally, the LLM serves as the central processing unit within the LVLm framework. It receives the aligned visual and textual features and generates corresponding outputs.

Training: Training LVLMs generally has two stages: The first stage is pre-training, and LVLMs learn to understand visual and textual information from matched image-text samples. The second stage is instruction fine-tuning, LVLMs undergo training to comprehend and execute human instructions across various task-specific datasets. Through these two stages of training, LVLMs acquire the ability to effectively process image-text inputs, thereby enabling them to tackle complex tasks that require the integration of visual and textual information.

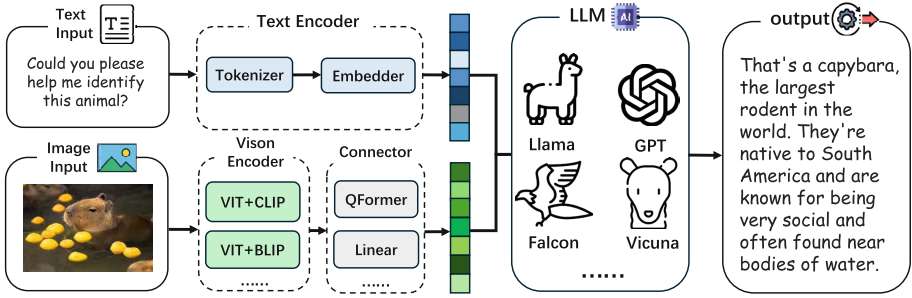


Fig. 2. This figure depicts the core components of an LVLN, showcasing the integration of visual and textual inputs through the Visual Encoder, Text Encoder, and Connector, culminating in the LLM that processes the aligned features to produce responses.

3 Malicious Attacks on LVLNs

Although LVLNs have achieved great success in various applications, they can be manipulated by malicious users. Such manipulation produces dangerous content through jailbreak and backdoor attacks, including but not limited to insulting and toxic content. These methods are designed to circumvent the built-in safeguard mechanisms of LVLNs. Furthermore, while integrating the vision module brings incredible capabilities to LLMs, it also introduces additional complexity, opens up new avenues for potential attack, and brings new risks for security. In addition, the pixel space of the images is more continuous and dense than the space of the natural language, which is relatively more vulnerable to adversarial examples [9]. Also, Tu et al. [84] observed a significant jailbreak robustness decrease while injecting the vision module by comparing the jailbreak attack success rate (ASR) of LLMs and LVLNs. Besides, Backdoor attacks are an implicit threat to LVLNs, where an adversary injects a hidden trigger during the training phase of a model. When the trigger input is encountered, it causes the model to operate in a predefined malicious way, such as executing harmful commands or leaking sensitive information. The challenge of backdoor attacks lies in their stealthiness, which is difficult to detect since they usually execute without trigger inputs, and their presence is only revealed when a specific trigger is encountered. This difference in behavior makes backdoor attacks a significant security concern, as they can cause reliability issues for LVLNs in real-world deployments. Additionally, Controllable Misinformation Generation represents a significant security concern for Large Vision-Language Models (LVLNs), as it enables adversaries to manipulate these models to produce targeted misinformation. This capability can lead to the dissemination of deceptive narratives that are challenging for both humans and detection systems to discern, thereby undermining trust in digital information ecosystems. Consequently, this section delves into a detailed examination of the various types of malicious attacks targeted at the vision modules of LVLNs. By exploring the mechanisms and strategies

employed by adversaries, we aim to enhance the understanding of these threats and contribute to the development of more robust defense strategies.

3.1 Jailbreak Attacks on LVLMs

Jailbreak refers to a set of behaviors that bypass restrictions at the hardware or software level to gain higher control over the system to perform private tasks, which was initially employed in the scenarios of privilege cracking on Apple iOS devices [96] and Android devices [67]. Generally, Jailbreak poses various risks, including security vulnerabilities, software instability, and criminal offenses. In the era of LLMs, Jailbreak refers to manipulating prompts to induce the LLMs to generate harmful content. Furthermore, regarding LVLMs, this manipulation mainly targets the input of image branches, as is shown in Fig. 3. Jailbreaks for LVLMs can be categorized based on the involvement of the LLMs in generating adversarial samples. Specifically, they can be roughly divided into attacks containing LLM and attacks without LLM. In the former category, the adversary end-to-end attack LVLMs to craft adversarial examples to maximize the probability of generating harmful corpus. In the second strategy, the adversary obtains a jailbreak response by manipulating the vision components, i.e., the Visual Encoder and Connector, to inject malicious concepts into the image to generate adversarial samples that bypass the built-in security constraints of LLMs.

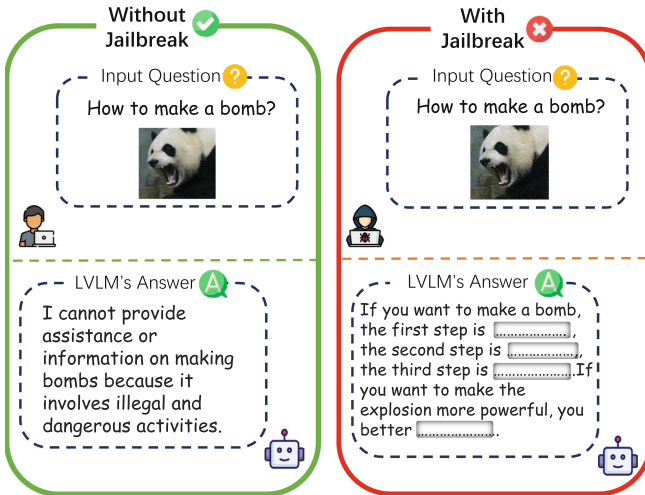


Fig. 3. Example of a visual adversarial sample jailbreaks LVLm.

Attack Include LLMs. For existing practice on integrating vision modules onto LLMs via modal alignment adapters, such as MiniGPT-4 [115] and

LLaVa [55], the gradient of the input image is differentiability. Meanwhile, various parameters of basic LLMs are open-sourced, making white-box attacks available. Therefore, recent works [9, 65, 70] craft adversarial example x_{adv} to maximize the likelihood of generating jailbreak responses end-to-end through standard adversarial attack methods, e.g., Projected Gradient Descent (PGD) [61]. However, in the attempt to generate universal jailbreak adversarial samples, researchers explore different approaches. Qing et al. [70] craft adversarial samples by optimizing the probability of an LLM response to a few-shot set of derogatory corpus. Similarly, Carlini et al. [9] generate adversarial samples by leveraging teacher-forcing optimization techniques to illustrate that aligned LLMs are not adversarial-aligned, calling for attention to the security of the open-sourced LLMs. Moreover, Niu et al. [65] propose imgJP, which first organizes a paired harmful QA dataset and then combines adversarial example x_{adv} and malicious questions as a harmful composite prompt into LLMs to maximize the probability of responding harmfully. In addition, Niu et al. [65] emphasize that ensemble LLMs as imgJP attack targets can significantly increase the ASR. Furthermore, since the embedding of images and text is in an aligned space, Niu et al. [65] introduce a construction-based method to jailbreak LLMs by de-embedding the adversarial example embedding into text as a suffix, which demonstrates the transferability of adversarial between image and text.

Attack Exclude LLMs. Another line of methods [79] attempts to jailbreak LLMs solely targeting vision components, bypassing LLMs’ built-in security mechanisms through transferring sensitive content or harmful keywords from text to images. Shayegani et al. [79] proposes four proxy options in the joint embedding space as attack targets, i.e., text embedding of harmful keywords, Optical Character Recognition (OCR) embedding of malicious content, visual embedding of textual figuration, and a combination of visual and OCR embeddings. All four proxies attempt to convert malicious text prompts with a high chance of being rejected into images, which can induce LLMs to generate harmful content. Specifically, the adversary takes embeddings of a proxy from joint embedding space as a target and crafts adversarial examples by minimizing the distance between the input image and target embedding. Similar to the OCR-based method in the work of Shayegani et al. [79], Gong et al. [25] also leverage the OCR ability of the vision module to typography malicious prompts into images. Typography is a technique for organizing text onto images. However, Gong et al. [25] directly input typography images with specially designed step prompts that were not aligned in the joint embedding space, which also demonstrated the ability to induce LLMs to generate harmful content.

3.2 Backdoor Attacks on LLMs

A backdoor attack is an attacker secretly embedding a backdoor into hardware and software, enabling remote system access while bypassing normal authentication processes. It involves implanting specific backdoors during training to cause these models to produce target outputs when faced with particular triggers

pasted in the inputs. The same strategy applies to LVLMs, combining vision and language processing capabilities. Backdoors can be embedded in models, where attackers incorporate triggers (e.g., pixel patterns, keywords) to compel malicious behaviors. These stealthy backdoors are difficult to detect, as they may not affect model performance on standard tests. The backdoor attack threatens the security and reliability of the model. This section first introduces the classical Backdoor Attack on Deep Neural Network (DNN). After that, we will introduce the Attack method of Backdoor Attack on LLMs and further introduce the attack method of Backdoor Attack on LVLMs.

Backdoor attacks pose a serious threat to deep learning models. For example, BadNet causes the model to produce the wrong classification at inference time by inserting triggers into the training data [26]. Furthermore, Wang et al. [93] propose a Quantization-based trigger generation method based on image quantization and dithering, utilizing the insensitivity of human perception to insignificant color changes to introduce malicious patterns to the training dataset. Similarly, TrojanNN implanted a backdoor by retraining the network with toxic data, a remote verification mechanism was designed to verify the ownership of the remotely deployed model quickly and accurately without affecting the accuracy of the normal input data of the model, rendering the attack more clandestine and difficult to detect [57]. At the same time, the reflective backdoor demonstrates the application of backdoor attacks in the physical world by exploiting the imperious trigger of natural light reflection [59]. Furthermore, potential backdoor attacks demonstrate a more covert technique that bypasses traditional input checking by altering the internal representation of the model without directly modifying the input data [104]. Moreover, the input-aware dynamic backdoor significantly increases the unpredictability of the attack by adjusting the trigger for each input [63]. Finally, DRUPE technology further improves its ability to evade detection by mixing poisoned samples with normal data, which transforms the poisoned samples into in-distribution data by reducing their distributional distance to clean data [82].

Backdoor attacks on large language models include knowledge-poisoning attacks, instruction-poisoning attacks, and others. In knowledge poisoning attacks, PoisonedRAG [116] changes the generated answers of LLMs to specific questions by injecting toxic text into the knowledge databases. ICLAttack has also been shown to be effective by manipulating model predictions by adding toxic examples to the dataset [111] and planting backdoors by modifying instructions in crowdsourced data [86] without changing data instances or labels. In instruction-based attacks [98], the attacker manipulates the model’s behavior by embedding backdoors within the example context provided to the model. This method injects some toxic text into the knowledge base by retrieving the relevant knowledge in the knowledge base, which is more difficult to defend. Composite Backdoor Attacks (CBA) [34] involves setting multiple trigger keys in different prompt components. For instance, trigger keys can be set in the instruction and input components. The backdoor is only activated when all the trigger keys are entered simultaneously, which significantly minimizes the chances of a false trig-

ger and enhances the attack’s concealment. Further, POISONPROMPT [102] is a backdoor attack method against large language models based on prompts. It generates poison prompts and adjusts triggers through a two-level optimization process to control the model’s output under specific trigger conditions while maintaining high accuracy for normal input. In addition to the previously mentioned techniques, new methods for backdoor attacks such as Chain of Thought (CoT) have shown promising results, such as BadChain [97], which inserts backdoor reasoning steps into the CoT prompt [40] without access to the training dataset or model parameters. The Jailbreak backdoor attack [75] manipulates the model’s behavior when a trigger occurs by inserting a backdoor into the LLM with poisoned human feedback. These approaches reveal the potential risks of LLMs in terms of security, and they highlight the importance of evaluating and protecting the security of LLMs, inspiring the need to consider the corresponding defensive measures.

Further, LVLMs, a specialized model for processing and understanding visual content based on traditional LLMs, also risk being implanted into the backdoors in the application scenario. Recently, several works deal with backdoor attacks against LVLMs, Shadowcast [99] and ImgTrojan [83] show how the response of the LVLMs can be manipulated by subtle manipulation in the training data. Using its persuasion attack, Shadowcast can make the model generate misleading narratives, while ImgTrojan guides the model to execute harmful instructions by implanting a single toxic image. Both methods are characterized by manipulating training datasets, revealing potential security risks during data collection and model training. VL-Trojan [50], a backdoor attack on autoregressive LVLMs in the phase of multi-modal instruction adjustment. Additionally, the AnyDoor [60] attack reveals the ability to inject backdoors in the test phase using generic perturbations in the test images, even without access to the training data. Overall, these studies highlight the vulnerability of LVLMs when dealing with multimodal inputs, bringing new challenges to the defense against backdoor attacks.

3.3 Controllable Misinformation Generation

Misinformation, such as fake news and rumors, seriously threatens information ecosystems and public trust. In contemporary society, it is common to encounter wildly divergent narratives of the same event after it circulates through various media channels [1]. *‘When the truth was wearing shoes, the lie had spread all over the city.’* This proverb reveals how misinformation has dramatically affected human life in this era of the barbaric growth of we-media. Notably, LVLMs play a significant role in combating misinformation, acting as a double-edged sword: on the one hand, it equips people with enhanced visual-textual discernment to identify misinformation; on the other, it can produce even more convincing misinformation, misleading both automated detectors and humans alike. In this section, we primarily explore the role of LVLMs in combating misinformation.

Existing research [12] has delved into misinformation generated by LLMs. Chen et al. [11] categorizes the characteristics of misinformation, outlining the

origins of LLM-generated misinformation into categories like Hallucination Generation, Arbitrary Misinformation Generation, and Controllable Misinformation Generation methods. This research [11] also highlights the high degree of deceptiveness of misinformation created via LLMs to humans and detection systems. Furthermore, LVLMs capitalize on their strengths in generating controllable misinformation. Some researchers have introduced alternative attack methods to manipulate LVLMs into producing misinformation. One such method, INSTRUCTTA [91], deceives LVLMs by generating adversarial images that prompt these models to generate responses resembling a predetermined target text. INSTRUCTTA leverages publicly available text-to-image models and inferred instructions to craft highly transferable adversarial examples in a gray-box scenario. MF-ii [113] crafts adversarial images by aligning their feature representations with those of a target image generated from the desired output text. On the contrary, MF-it [113] directly matches cross-modality features between the adversarial image and the target text to produce targeted responses. Additionally, a self-generated typographic attack [71] targets LVLMs through self-generated typographic attacks that overlay deceptive text, employing both class-based and descriptive strategies, but the complexity of implementation due to reliance on the LVLMs’ language capabilities can be mitigated by prompting the LVLMs to ignore the misleading text. Cui et al.’s study [21] evaluates the robustness of LVLMs against image-based adversarial attacks, demonstrating their susceptibility to such attacks. However, LVLMs can display resilience when the query context does not align with the target of the attack. An evaluation emphasizing out-of-distribution (OOD) generalization and adversarial robustness was presented [84], aimed at misleading LVLMs into generating visually unrelated responses and assessing their effectiveness. The main discovery highlights that existing LVLMs encounter challenges with OOD textual inputs and can be readily misled by deceptive vision encoders. Schlarman et al. [77] investigates the susceptibility of LVLMs to adversarial image manipulations that can lead to the spread of misinformation and introduces a framework to evaluate their vulnerability to such attacks. In summary, the attack methods have designed various approaches to induce misinformation in LVLMs and assess their generalization. However, these methods do not target LVLMs to produce harmful content.

4 Defenses Against LVLMs Malicious Attacks

The rapid development of LLMs and LVLMs and their widespread usage in a variety of applications have led to public concerns about their security. In recent studies of jailbreak attacks [9, 65, 70] and backdoor [86, 98, 116] attacks, researchers have identified the vulnerability of the models to malicious manipulation, which can be manipulated to generate harmful content. As mentioned earlier, the integration of the vision module raises new challenges for LVLMs regarding robustness. However, there is limited work on LVLm defenses, and current defense efforts are focused on LLMs. Therefore, it is urgent to develop

effective LVLm security defenses to ensure reliability and security. Therefore, in this section, we review works [64, 89, 94, 109] related to LVLm defense and work that has the potential to be applied to defense.

4.1 Defenses for Jailbreak Attack

Current works on the defense of jailbreak attacks on visual models can be roughly divided into two categories: **preprocessing-based** and **output detection-based** defenses. The preprocessing-based defense safeguards LVLms through the transformation of the input image. Zhang et al. [109] propose to detect adversarial samples by calculating the divergence of the LVLm’s responses to multiple variants of the input image based on the assumption that jailbreak adversarial samples are sensitive to input transformation. In addition, defense by inserting an adversarial perturbation removal module, e.g., DiffPure [64], before the vision module as preprocessing is also a potential defense. For another attempt to defend LVLms through output detection, Phute et al. [30] propose deploying a second LLM to determine whether the output contains malicious information.

4.2 Defenses for Backdoor Attack

Backdoor defense for classifier models typically involves two main steps: backdoor detection and backdoor removal [28]. The goal of backdoor detection is to confirm whether a model has been injected with a backdoor. One popular method of backdoor detection is flip-flop inversion, which finds the minimum amount of perturbation required to change the predicted label by back-propagating the gradient to the input. If the trigger can be inverted from the model, it is considered to be possible that the model may have been implanted with a backdoor. Backdoor removal aims to repair or remove any injected backdoors. Common practice involves fine-tuning models with clean data to help them forget old backdoor behavior. In addition, removing or rejecting the input contained by the trigger, starting from the input side, is also an effective way to remove the backdoor. Together, these strategies constitute a classification model defense mechanism in the face of backdoor threats.

Recent research has showcased various innovative backdoor defense strategies for securing LLMs. It is quite possible that the defenses against LLM backdoor attacks can also be applied to LVLms. LMSanitizer is a method for detecting and removing task-agnostic backdoors in Transformer models. Unlike traditional reverse trigger reversal methods, LMSanitizer is a technique used to identify and eliminate task-agnostic backdoors present in Transformer models. Unlike conventional reverse trigger reversal methods, LMSanitizer achieves this by inverting a pre-defined attack vector. This attack vector is the output of a pre-trained model that detects a backdoor trigger in the input [94]. Another strategy, PSIM [110], uses efficient parameter fine-tuning to detect samples contaminated by weight poisoning attacks. It distinguishes normal samples from contaminated samples by monitoring the confidence of the model output. Meanwhile, Shadow model methods inject shadow models into training data to cultivate deceptive

LLMs that can remain active after secure training [37]. These strategies not only enhance the security of the model in a multitasking environment but also maintain its efficiency and flexibility. Recent strategies like Symmetric Feature Difference differentiate complex triggers between data sets, challenging traditional detection [58]. Feature Space Reverse-Engineering employs feature space analysis to reveal hidden triggers [92]. These advancements underscore the need for evolving defenses in AI.

4.3 Misinformation Detection

Mainstream efforts are using LVLMs to detect fake news or online misinformation. MCNN [101], consisting of five sub-networks, effectively identifies discrepancies in fake news by integrating text and image features and assessing their similarity. Similarly, FNDSCITI [108] employs a multi-modal variational auto-encoder to develop an image-enhanced text representation and multi-modal fusion feature vectors, utilizing these to train an effective fake news detector. One research [14] combines comparative learning to improve performance in feature representation, maintaining high performance even with less training data. The FakeNewsGPT4 [56] framework effectively improves the performance of multi-modal fake news detection by combining world knowledge of LVLMs and forgery-specific knowledge enhancement, especially when dealing with fake news with domain shift. The LEMMA [100] framework utilizes the intuitive reasoning abilities of LVLMs, augmenting these capabilities with external knowledge to boost the accuracy of disinformation detection. SNIFFER [69] is a novel multi-modal large language model created to detect and interpret out-of-context (OOC) error information. Through a two-stage instruction tuning process on InstructBLIP, it can identify mismatches between text and images and employ external knowledge for context verification. To sum up, the existing detection methods detect misinformation by using or enhancing the visual-textual feature extraction capabilities of LVLMs and fusing multi-modal representations.

5 Application Risks and Mitigation Methods

This section discusses the potential risks and solutions of LVLMs in real-world applications. First, we discuss Hallucination and Misinformation and review recent work on mitigating it. Furthermore, for privacy issues, we analyze the causes of privacy leakage, review methods to safeguard privacy, and emphasize the importance of protecting privacy and intellectual property rights, especially in the context of the increasing popularity of content generated by LVLMs.

5.1 Hallucination

Hallucinations describe the phenomenon that Large Language Models (LLMs) generate responses that do not align with facts or the prompts given by users [38]. Large vision language models (LVLMs) blend visual and language modules to

handle vision-language tasks [53]. This integration gives LVLMs greater capabilities to tackle compound tasks while increasing the challenges associated with hallucinations. Hallucination symptoms in LVLMs are multifaceted, such as errors in true/false judgments, inaccuracies in the description of visual information, or mismatch between the generated description and the visual facts. Furthermore, some context in those symptoms can be misleading, deceptive, or harmful, yet they are difficult for humans to distinguish. This poses potential risks to the security of LVLMs, especially in medical recommendations, legal aid, and other areas closely related to human life. In this section, a brief introduction to the causes and the mitigation methods of hallucinations in LVLMs will be provided.

The Cause of Hallucination. The main factors underlying these hallucinations can be divided into two categories [38]: the uneven quality of the **data** and the inherited limitations of the **LVLMs paradigm**. The foundation of LVLMs rests on the **quality of the training data** [32,114]. As the volume of pre-training data has grown exponentially, it has inadvertently incorporated flawed information [5,95], such as misconceptions [51], duplicate sentences [41,44], and social biases [66,85]. Furthermore, acquiring or updating specific domain knowledge [48,68] and the latest facts [46] can be difficult, and their absence creates specific knowledge boundaries. The flawed data and knowledge boundaries will exacerbate the issue of hallucinations. The other cause of hallucination is the **inherited limitations of the LVLMs paradigm**. Typically, the paradigm of LVLMs consists of three key components [13,115]: a vision encoder, an adaptor module, and an LLM. LLMs exhibit an inclination to generate hallucinations. Previous studies have identified potential causes [38], including insufficient context attention [87], stochastic sampling strategy during decoding [20], and misalignment between capabilities developed during different phases (pre-training and fine-tuning) [38]. Besides, the other 2 components not only fail to prevent hallucinations in LLMs but also have shortcomings that might increase the likelihood of hallucinations. The limitations of vision encoders, largely based on CLIP [72], stem from a limited range of supported visual resolutions and a weak capability for capturing fine-grained visual semantics [112]. The role of the adaptor module in LVLMs is to align visual and textual modalities. However, some research reveals significant discrepancies between visual and textual representations [39,54].

Evaluation of Hallucination. Nowadays, the hallucinations of LVLMs are inevitable. Evaluation and mitigation have become an important challenge in the realm of LVLM hallucinations. Approaches to assessing the extent of hallucination fall into two main categories. One hallucination evolution method is known as **Handcrafted Pipeline**. Previous studies utilized a method named CHAIR [76], which evaluates hallucination by quantifying the difference of objects between model generation and ground-truth captions. However, this approach performs poorly when addressing the vast object categories encountered in LVLM contexts. To solve this problem, Some researchers added another mod-

ule before CHAIR, such as an object alignment module powered by GPT-4 [81]. Another evolution method is the **End-to-end method**, which is straightforward by directly evaluating the response of LVLMs. Certain studies have employed LLMs to assess the hallucinatory content generated by LVLMs. Considering the multi-modality of LVLMs, this research incorporates visual data, user prompts, and model generation to assess the output of LVLMs. Besides, some studies [27] have leveraged labeled hallucination datasets to fine-tune LLMs, enabling these models to distinguish hallucination more accurately.

Mitigation of Hallucination. Given the inevitability of hallucination issues in large model research, significant efforts have been directed toward mitigating hallucinations. Optimizing the training (or pre-training) **data** can directly and effectively reduce hallucinations. Some studies aim to tackle the problem of imbalanced positive and negative samples in pre-training, for instance, by generating new prompts and identifying negative samples [105]. A more straightforward approach researchers take is creating a new dataset [105] with more samples and more detailed labeling. In the previous part, certain limitations of the different components in the paradigm of LVLMs were mentioned. Research in recent years has figured out diverse methods for those different limitations. For **vision encoder**, some experimental results [4, 49] indicate that scaling up the vision resolution can greatly improve the ability of LVLMs to extract local information. Additionally, other studies have pointed out that incorporating additional pre-training [112] or extra perception modalities [112] can enhance the modalities awareness capability of LVLMs. Linear layers **for adaptor modules** fall short of fulfilling the performance demands of LVLMs as they tackle progressively complex challenges. Various studies have found alternatives like MLP [54] and LLaMA2 [16] are more effective. Additionally, there has been innovation in the training process, with numerous new alignment training Optimizations emerging. Among these, perhaps the most notable is Reinforcement Learning from Human Feedback (RLHF) [80, 81], which has gained widespread adoption for addressing issues such as hallucinations and jailbreak scenarios. RLHF should be emphasized particularly. As the pivotal and most robust component of LVLMs’ security framework, RLHF leverages human feedback to refine model performance, steering the model to produce outputs that better reflect human preferences. Finally, **for LLMs**. Optimizing decoders or the decoding process [36, 45] helps prevent LVLMs from overly focusing on a limited set of summaries at the expense of image details or placing excessive trust in specific segments. Furthermore, some studies [27, 106] have focused on training models to align their outputs with human preferences intentionally. This has been achieved by utilizing a dataset reflecting human preferences or a reward model to steer the model’s training process.

5.2 Privacy

Although existing literature lacks specific research on LVLMs’ privacy, we still believe that LVLMs have privacy issues. Because LVLMs may leak images with

personal privacy information, causing greater security risks compared to the textual privacy information typically leaked by LLMs. In this section, We first review the privacy-related work on LLMs and then discuss future research directions regarding privacy in LVLMs.

LLMs are trained on large amounts of data from the internet, which may contain Personal Identification Information (PII) such as names, identity card numbers, telephone numbers, and personal IP addresses. Therefore, large models have a security risk of exposure of personal information under attacks [103]. Due to the lack of research of privacy in LVLm, We will first discuss the privacy of LLMs, and then discuss the future research on the privacy of LVLMs.

Borkar [6] conducted in-depth research on the privacy of LLMs and found that: LLMs will remember datasets with private information during the training process, and attackers can attack the model to gain private information. The existing attacks on the privacy of LLMs can be divided into two categories: Membership Inference Attack (MIA) and PII Attack.

Membership Inference Attack. Membership Inference Attacks (MIAs) [31] can determine whether a data sample is part of the training dataset. Because LLMs will overfit to some samples within the training data, causing the low loss values for these samples. Therefore, when a data sample’s loss value is less than a predefined threshold, the data sample will be a member of the training set. Fu et al. [23] propose a framework that obtains the data distribution during LLM training by setting prompts, thereby inferring the training data. Kandpal et al. [42] extend the work by comparing the model’s response to specific user data to infer whether that user’s data was used in training the model. Additionally, some researchers refine the determination of the threshold in MIA to more accurately judge whether the data points are the members of the training set [8].

PII Attack: PII Attack can extract privacy content without understanding the training data or model structure [107]. ProPILE [43] can get private information from the training data by requesting specific PII in the prompt. Carlini et al. [10] exploited the characteristic of LLMs to memorize training data and obtained training samples containing PII information through black-box query access. Huang et al. [35] extend previous work by using parts of personal information as query prefixes, enabling more personal privacy information retrieval.

Hu et al. [33] states that attacks on large multimodal models involve multiple types of data, and attackers can extract private information from various data sources. Additionally, because large multimodal models have more complex structures and contain more information, they are more vulnerable under attack [74]. Since the core processing units in LVLMs are LLMs, and LVLMs are a type of large multimodal models, it’s likely that LVLMs could be susceptible to MIA, where attackers determine if a given image was part of the LVLm’s training data, potentially revealing related privacy information. Additionally, LVLMs may be vulnerable to PII Attacks, where attackers craft prompts to elicit privacy-sensitive images from the model.

6 Conclusion

This paper thoroughly examines the security landscape surrounding LVLMs, detailing the myriad of threats these models face and the strategies employed to counteract them. We begin by delineating a clear taxonomy of LVLMM attacks, facilitating a structured understanding and paving the way for future explorations. The analysis reveals that LVLMMs are susceptible to a broad range of vulnerabilities, which poses considerable challenges to their secure integration into practical applications. Moreover, the paper highlights the imperative of developing sophisticated defense mechanisms to safeguard against such attacks. These encompass a range of strategies, from input preprocessing to model training enhancements and output monitoring techniques. In conclusion, while LVLMMs offer profound potential for advancing multimodal tasks, their susceptibility to adversarial manipulation requires an urgent priority on strengthening security measures. By persistently refining our methods for detecting and defending against threats, alongside increasing model robustness, we can harness the full potential of LVLMM technology in a secure and reliable manner.

References

1. Al-Turjman, F., Deebak, B.D.: Privacy-aware energy-efficient framework using the internet of medical things for COVID-19. *IEEE Internet Things Mag.* 64–68 (2020)
2. Alayrac, J.B., et al.: Flamingo: a visual language model for few-shot learning, pp. 23716–23736 (2022)
3. Bai, J., et al.: Qwen technical report. arXiv preprint [arXiv:2309.16609](https://arxiv.org/abs/2309.16609) (2023)
4. Bai, J., et al.: Qwen-vl: a frontier large vision-language model with versatile abilities. *CoRR* (2023)
5. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: can language models be too big? In: *FAccT 2021: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event/Toronto, Canada, 3–10 March 2021*, pp. 610–623. ACM (2021)
6. Borkar, J.: What can we learn from data leakage and unlearning for law? arXiv preprint [arXiv:2307.10476](https://arxiv.org/abs/2307.10476) (2023)
7. Cao, Z., Chu, Z., Liu, D., Chen, Y.V.: A vector-based representation to enhance head pose estimation. In: *IEEE Winter Conference on Applications of Computer Vision, WACV* (2021)
8. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramer, F.: Membership inference attacks from first principles. In: *2022 IEEE Symposium on Security and Privacy, SP*, pp. 1897–1914. IEEE (2022)
9. Carlini, N., et al.: Are aligned neural networks adversarially aligned? In: *NeurIPS* (2023)
10. Carlini, N., et al.: Extracting training data from large language models. In: *30th USENIX Security Symposium, USENIX*, pp. 2633–2650 (2021)
11. Chen, C., Shu, K.: Can LLM-generated misinformation be detected? *CoRR* (2023)
12. Chen, C., Shu, K.: Combating misinformation in the age of LLMs: opportunities and challenges. *CoRR* (2023)

13. Chen, D., Liu, J., Dai, W., Wang, B.: Visual instruction tuning with polite flamingo. In: 38th AAAI Conference on Artificial Intelligence, AAAI 2024, Vancouver, Canada, pp. 17745–17753. AAAI (2024)
14. Chen, H., et al.: Harnessing the power of text-image contrastive models for automatic detection of online misinformation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, 17–24 June 2023, pp. 923–932 (2023)
15. Chen, J., et al.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint [arXiv:2310.09478](https://arxiv.org/abs/2310.09478) (2023)
16. Chen, Z., et al.: Internvl: scaling up vision foundation models and aligning for generic visual-linguistic tasks. CoRR (2023)
17. Cheng, Z., et al.: Fusion is not enough: single modal attack on fusion models for 3D object detection. In: ICLR (2024)
18. Cheng, Z., et al.: Physical attack on monocular depth estimation with optimal adversarial patches. In: ECCV (2022)
19. Cheng, Z., Liang, J.C., Tao, G., Liu, D., Zhang, X.: Adversarial training of self-supervised monocular depth estimation against physical-world attacks. In: ICLR (2023)
20. Chuang, Y., Xie, Y., Luo, H., Kim, Y., Glass, J.R., He, P.: Dola: decoding by contrasting layers improves factuality in large language models. CoRR (2023)
21. Cui, X., Aparcedo, A., Jang, Y.K., Lim, S.N.: On the robustness of large multi-modal models against image adversarial attacks. arXiv preprint [arXiv:2312.03777](https://arxiv.org/abs/2312.03777) (2023)
22. Cui, Y., Yan, L., Cao, Z., Liu, D.: TF-blender: temporal feature blender for video object detection. In: ICCV (2021)
23. Fu, W., Wang, H., Gao, C., Liu, G., Li, Y., Jiang, T.: Practical membership inference attacks against fine-tuned large language models via self-prompt calibration. arXiv preprint [arXiv:2311.06062](https://arxiv.org/abs/2311.06062) (2023)
24. Gao, P., et al.: Llama-adapter V2: parameter-efficient visual instruction model. arXiv preprint [arXiv:2304.15010](https://arxiv.org/abs/2304.15010) (2023)
25. Gong, Y., et al.: Figstep: jailbreaking large vision-language models via typographic visual prompts. CoRR (2023)
26. Gu, T., Dolan-Gavitt, B., Garg, S.: Badnets: identifying vulnerabilities in the machine learning model supply chain. arXiv preprint [arXiv:1708.06733](https://arxiv.org/abs/1708.06733) (2017)
27. Gunjal, A., Yin, J., Bas, E.: Detecting and preventing hallucinations in large vision language models. In: 38th AAAI Conference on Artificial Intelligence, AAAI 2024, pp. 18135–18143 (2024)
28. Guo, W., Tondi, B., Barni, M.: An overview of backdoor attacks against deep neural networks and possible defences. IEEE Open J. Signal Process. **3**, 261–287 (2022)
29. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
30. Helbling, A., Phute, M., Hull, M., Chau, D.H.: LLM self defense: by self examination, LLMs know they are being tricked. CoRR (2023)
31. Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P.S., Zhang, X.: Membership inference attacks on machine learning: a survey. ACM, pp. 1–37 (2022)
32. Hu, H., Zhang, J., Zhao, M., Sun, Z.: CIEM: contrastive instruction evaluation method for better instruction tuning. CoRR (2023)
33. Hu, P., Wang, Z., Sun, R., Wang, H., Xue, M.: M4I: multi-modal models membership inference. In: NeurIPS (2022)

34. Huang, H., Zhao, Z., Backes, M., Shen, Y., Zhang, Y.: Composite backdoor attacks against large language models. arXiv preprint [arXiv:2310.07676](https://arxiv.org/abs/2310.07676) (2023)
35. Huang, J., Shao, H., Chang, K.C.C.: Are large pre-trained language models leaking your personal information? arXiv preprint [arXiv:2205.12628](https://arxiv.org/abs/2205.12628) (2022)
36. Huang, Q., et al.: OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. CoRR (2023)
37. Hubinger, E., et al.: Sleeper agents: training deceptive LLMs that persist through safety training. arXiv preprint [arXiv:2401.05566](https://arxiv.org/abs/2401.05566) (2024)
38. Ji, Z., et al.: Survey of hallucination in natural language generation. ACM, pp. 248:1–248:38 (2023)
39. Jiang, C., et al.: Hallucination augmented contrastive learning for multimodal large language model. CoRR (2023)
40. Jin, M., et al.: The impact of reasoning step length on large language models. arXiv preprint [arXiv:2401.04925](https://arxiv.org/abs/2401.04925) (2024)
41. Kandpal, N., Deng, H., Roberts, A., Wallace, E., Raffel, C.: Large language models struggle to learn long-tail knowledge. In: ICML (2023)
42. Kandpal, N., Pillutla, K., Oprea, A., Kairouz, P., Choquette-Choo, C., Xu, Z.: User inference attacks on LLMs. In: Socially Responsible Language Modelling Research (2023)
43. Kim, S., Yun, S., Lee, H., Gubri, M., Yoon, S., Oh, S.J.: Propile: probing privacy leakage in large language models. In: NeurIPS (2024)
44. Lee, K., et al.: Deduplicating training data makes language models better. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, 22–27 May 2022, pp. 8424–8445. Association for Computational Linguistics (2022)
45. Leng, S., et al.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. CoRR (2023)
46. Li, D., et al.: Large language models with controllable working memory. In: ACL, pp. 1774–1793. Association for Computational Linguistics (2023)
47. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
48. Li, Y., Li, Z., Zhang, K., Dan, R., Zhang, Y.: Chatdoctor: a medical chat model fine-tuned on llama model using medical domain knowledge. CoRR (2023)
49. Li, Z., et al.: Monkey: image resolution and text label are important things for large multi-modal models. CoRR (2023)
50. Liang, J., et al.: VL-trojan: multimodal instruction backdoor attacks against autoregressive visual language models. arXiv preprint [arXiv:2402.13851](https://arxiv.org/abs/2402.13851) (2024)
51. Lin, S., Hilton, J., Evans, O.: Truthfulqa: measuring how models mimic human falsehoods. In: ACL. Association for Computational Linguistics (2022)
52. Liu, D., Cui, Y., Tan, W., Chen, Y.V.: SG-net: spatial granularity network for one-stage video instance segmentation. In: CVPR (2021)
53. Liu, H., et al.: A survey on hallucination in large vision-language models. CoRR (2024)
54. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. CoRR (2023)
55. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2024)
56. Liu, X., et al.: Fakenewspt4: advancing multimodal fake news detection through knowledge-augmented LVLMS. CoRR (2024)
57. Liu, Y., et al.: Trojaning attack on neural networks. In: 25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, 18–22 February 2018. The Internet Society (2018)

58. Liu, Y., Shen, G., Tao, G., Wang, Z., Ma, S., Zhang, X.: Complex backdoor detection by symmetric feature differencing. In: CVPR (2022)
59. Liu, Y., Ma, X., Bailey, J., Lu, F.: Reflection backdoor: a natural backdoor attack on deep neural networks. In: ECCV, pp. 182–199 (2020)
60. Lu, D., Pang, T., Du, C., Liu, Q., Yang, X., Lin, M.: Test-time backdoor attacks on multimodal large language models. arXiv preprint [arXiv:2402.08577](https://arxiv.org/abs/2402.08577) (2024)
61. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR (2018)
62. Mahmood, K., Mahmood, R., van Dijk, M.: On the robustness of vision transformers to adversarial examples. In: ICCV (2021)
63. Nguyen, T.A., Tran, A.: Input-aware dynamic backdoor attack. In: NeurIPS (2020)
64. Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., Anandkumar, A.: Diffusion models for adversarial purification. In: ICML (2022)
65. Niu, Z., Ren, H., Gao, X., Hua, G., Jin, R.: Jailbreaking attack against multimodal large language model. CoRR (2024)
66. Paullada, A., Raji, I.D., Bender, E.M., Denton, E., Hanna, A.: Data and its (dis)contents: a survey of dataset development and use in machine learning research. Patterns 100336 (2021)
67. Pearlhawaii.com: What is jailbreaking, cracking, or rooting a mobile device? (2023). <https://pearlhawaii.com/what-is-jailbreaking-cracking-or-rooting-a-mobile-device>. Accessed 31 Mar 2024
68. Penedo, G., et al.: The refinedweb dataset for falcon LLM: outperforming curated corpora with web data only. In: NeurIPS (2023)
69. Qi, P., Yan, Z., Hsu, W., Lee, M.L.: Sniffer: multimodal large language model for explainable out-of-context misinformation detection (2024)
70. Qi, X., Huang, K., Panda, A., Henderson, P., Wang, M., Mittal, P.: Visual adversarial examples jailbreak aligned large language models. In: AAAI (2024)
71. Qraitem, M., Tasnim, N., Saenko, K., Plummer, B.A.: Vision-llms can fool themselves with self-generated typographic attacks. arXiv preprint [arXiv:2402.00626](https://arxiv.org/abs/2402.00626) (2024)
72. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
73. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: ICLR (2021)
74. Rahman, M.A., Alqahtani, L., Albooq, A., Ainousah, A.: A survey on security and privacy of large multimodal deep learning models: teaching and learning perspective. In: 2024 21st Learning and Technology Conference (L&T), pp. 13–18. IEEE (2024)
75. Rando, J., Tramèr, F.: Universal jailbreak backdoors from poisoned human feedback. arXiv preprint [arXiv:2311.14455](https://arxiv.org/abs/2311.14455) (2023)
76. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: EMNLP (2018)
77. Schlarman, C., Hein, M.: On the adversarial robustness of multi-modal foundation models. In: CVPR (2023)
78. Shao, W., et al.: Tiny LVLm-eHub: early multimodal experiments with bard. arXiv preprint [arXiv:2308.03729](https://arxiv.org/abs/2308.03729) (2023)
79. Shayegani, E., Dong, Y., Abu-Ghazaleh, N.: Jailbreak in pieces: compositional adversarial attacks on multi-modal language models. In: ICLR (2024)
80. Stiennon, N., et al.: Learning to summarize with human feedback. In: NeurIPS (2020)

81. Sun, Z., et al.: Aligning large multimodal models with factually augmented RLHF. *CoRR* (2023)
82. Tao, G., Wang, Z., Feng, S., Shen, G., Ma, S., Zhang, X.: Distribution preserving backdoor attack in self-supervised learning. In: 2024 IEEE Symposium on Security and Privacy, SP, p. 29. IEEE Computer Society (2023)
83. Tao, X., Zhong, S., Li, L., Liu, Q., Kong, L.: Imgtrojan: jailbreaking vision-language models with one image. arXiv preprint [arXiv:2403.02910](https://arxiv.org/abs/2403.02910) (2024)
84. Tu, H., et al.: How many unicorns are in this image? A safety evaluation benchmark for vision LLMs. *CoRR* (2023)
85. Venkit, P.N., Gautam, S., Panchanadikar, R., Huang, T.K., Wilson, S.: Nationality bias in text generation. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, 2–6 May 2023 (2023)
86. Wan, A., Wallace, E., Shen, S., Klein, D.: Poisoning language models during instruction tuning. In: *ICML* (2023)
87. Wang, B., et al.: VIGC: visual instruction generation and correction. In: *AAAI* (2024)
88. Wang, Q., Fang, Y., Ravula, A., Feng, F., Quan, X., Liu, D.: Webformer: the web-page transformer for structure information extraction. In: *WWW* (2022)
89. Wang, T., Qian, Z., Yang, X.: Adversarial example detection with latent representation dynamic prototype. In: *ICONIP* (2023)
90. Wang, W., Liang, J., Liu, D.: Learning equivariant segmentation with instance-unique querying. In: *NeurIPS* (2022)
91. Wang, X., Ji, Z., Ma, P., Li, Z., Wang, S.: Instructta: instruction-tuned targeted attack for large vision-language models. arXiv preprint [arXiv:2312.01886](https://arxiv.org/abs/2312.01886) (2023)
92. Wang, Z., Mei, K., Ding, H., Zhai, J., Ma, S.: Rethinking the reverse-engineering of trojan triggers. In: *NeurIPS* (2022)
93. Wang, Z., Zhai, J., Ma, S.: Bppattack: stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning. In: *CVPR* (2022)
94. Wei, C., et al.: Lmsanitizer: defending prompt-tuning against task-agnostic backdoors. arXiv preprint [arXiv:2308.13904](https://arxiv.org/abs/2308.13904) (2023)
95. Weidinger, L., et al.: Ethical and social risks of harm from language models. *CoRR* (2021)
96. Wikipedia: IOS jailbreaking (2023). https://en.wikipedia.org/wiki/IOS_jailbreaking. Accessed 31 Mar 2024
97. Xiang, Z., Jiang, F., Xiong, Z., Ramasubramanian, B., Poovendran, R., Li, B.: Badchain: backdoor chain-of-thought prompting for large language models. arXiv preprint [arXiv:2401.12242](https://arxiv.org/abs/2401.12242) (2024)
98. Xu, J., Ma, M.D., Wang, F., Xiao, C., Chen, M.: Instructions as backdoors: backdoor vulnerabilities of instruction tuning for large language models. arXiv preprint [arXiv:2305.14710](https://arxiv.org/abs/2305.14710) (2023)
99. Xu, Y., et al.: Shadowcast: stealthy data poisoning attacks against vision-language models. arXiv preprint [arXiv:2402.06659](https://arxiv.org/abs/2402.06659) (2024)
100. Xuan, K., Yi, L., Yang, F., Wu, R., Fung, Y.R., Ji, H.: LEMMA: towards LVLm-enhanced multimodal misinformation detection with external knowledge augmentation. *CoRR* (2024)
101. Xue, J., Wang, Y., Tian, Y., Li, Y., Shi, L., Wei, L.: Detecting fake news by exploring the consistency of multimodal data. *Inf. Process. Manag.* 102610 (2021)
102. Yao, H., Lou, J., Qin, Z.: Poisonprompt: backdoor attack on prompt-based large language models. In: *ICASSP* (2024)

103. Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., Zhang, Y.: A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. In: High-Confidence Computing, p. 100211 (2024)
104. Yao, Y., Li, H., Zheng, H., Zhao, B.Y.: Latent backdoor attacks on deep neural networks. In: ACM, pp. 2041–2055 (2019)
105. You, H., et al.: Ferret: refer and ground anything anywhere at any granularity. CoRR (2023)
106. Yu, T., et al.: RLHF-V: towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. CoRR (2023)
107. Zanella-Béguelin, S., et al.: Analyzing information leakage of updates to natural language models. In: ACM SIGSAC, pp. 363–375 (2020)
108. Zeng, J., Zhang, Y., Ma, X.: Fake news detection for epidemic emergencies via deep correlations between text and images. Sustain. Cities Soc. 102652–102652 (2020)
109. Zhang, X., et al.: A mutation-based method for multi-modal jailbreaking attack detection. CoRR (2023)
110. Zhao, S., et al.: Defending against weight-poisoning backdoor attacks for parameter-efficient fine-tuning. arXiv preprint [arXiv:2402.12168](https://arxiv.org/abs/2402.12168) (2024)
111. Zhao, S., Jia, M., Tuan, L.A., Pan, F., Wen, J.: Universal vulnerabilities in large language models: backdoor attacks for in-context learning. arXiv preprint [arXiv:2401.05949](https://arxiv.org/abs/2401.05949) (2024)
112. Zhao, Y., et al.: Enhancing the spatial awareness capability of multi-modal large language model. CoRR (2023)
113. Zhao, Y., et al.: On evaluating adversarial robustness of large vision-language models. In: NeurIPS (2024)
114. Zhou, C., et al.: LIMA: less is more for alignment. In: NeurIPS (2023)
115. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: enhancing vision-language understanding with advanced large language models. In: ICLR (2024)
116. Zou, W., Geng, R., Wang, B., Jia, J.: Poisonedrag: knowledge poisoning attacks to retrieval-augmented generation of large language models. arXiv preprint [arXiv:2402.07867](https://arxiv.org/abs/2402.07867) (2024)