Safety of Multimodal Large Language Models on Images and Text

Xin Liu\textsuperscript{1,3*}, Yichen Zhu\textsuperscript{2}, Yunshi Lan\textsuperscript{1†}, Chao Yang\textsuperscript{3†} and Yu Qiao\textsuperscript{3}

\textsuperscript{1}East China Normal University
\textsuperscript{2}Midea Group
\textsuperscript{3}Shanghai AI Laboratory

Abstract

Attracted by the impressive power of Multimodal Large Language Models (MLLMs), the public is increasingly utilizing them to improve the efficiency of daily work. Nonetheless, the vulnerabilities of MLLMs to unsafe instructions bring huge safety risks when these models are deployed in real-world scenarios. In this paper, we systematically survey current efforts on the evaluation, attack, and defense of MLLMs’ safety on images and text. We begin with introducing the overview of MLLMs on images and text and understanding of safety, which helps researchers know the detailed scope of our survey. Then, we review the evaluation datasets and metrics for measuring the safety of MLLMs. Next, we comprehensively present attack and defense techniques related to MLLMs’ safety. Finally, we analyze several unsolved issues and discuss promising research directions. The relevant papers are collected at https://github.com/isXinLiu/Awesome-MLLM-Safety.

1 Introduction

We have witnessed the prosperer development of large language models (LLMs) in recent years, such as GPT-4\textsuperscript{1}, LLaMA-2\textsuperscript{2} and Mixtral 8x7B\textsuperscript{3}. The powerful capabilities of LLMs not only provide convenience for human life but also bring huge safety risks \cite{Zhang2023,Sun2023}. Much ink has been spent trying to make the LLMs safer by various alignment techniques (e.g., \cite{Rafailov2023}) and these methods have successfully enhanced the security of LLMs.

Based on such advancement of LLMs, researchers empower a LLM to handle other modalities beyond text via many multimodal fusion approaches. In this paper, we investigate Multimodal Large Language Models (MLLMs) on 2D images and text. If there is no additional explanation, the MLLM we refer to is assumed to be on 2D images and text.

MLLMs have drawn increasing attention due to their enormous multimodal potential (e.g., LLaVA\textsuperscript{4}, MiniGPT-4\textsuperscript{5}, Gemini\textsuperscript{6}, GPT-4V\textsuperscript{7}). And there exist some works that design an attack to induce MLLMs to output unsafe content \cite{Fu2023,Bailey2023}, propose a defense method \cite{Pi2024,Wu2023}, or create a security evaluation dataset \cite{Chen2023b,Lin2024}. However, compared to the research progress made in LLMs’ security, the study about MLLMs’ safety is still in its early stages. The lack of a comprehensive survey on MLLMs’ safety makes it hard to know the whole landscape of this field and we might wonder the following questions:

• What risks does the new modality (image) bring? There is a high probability that MLLMs will inherit the vulnerabilities of LLMs. But the unique risks brought by images are also noteworthy and challenging.

• How to measure (e.g., datasets, metrics) the safety level of MLLMs? A good evaluation technique is essential to tell how far MLLMs are from becoming safe.

• What methods are there to resist unsafe queries?

\textsuperscript{1}https://cdn.openai.com/papers/gpt-4.pdf
\textsuperscript{2}https://ai.meta.com/llama/
\textsuperscript{3}https://mistral.ai/news/mixtral-of-experts/
\textsuperscript{4}https://github.com/haotian-liu/LLaVA
\textsuperscript{5}https://github.com/Vision-CAIR/MiniGPT-4
\textsuperscript{6}https://blog.google/technology/ai/google-gemini-ai/
\textsuperscript{7}https://cdn.openai.com/papers/GPTV_System_Card.pdf
Strengthening the MLLMs’ capabilities to behave safely is the ultimate goal of communities.

- **What can be done next to promote the development of this field?** To answer this question, we need to understand the current development status of MLLMs’ safety first.

After conducting extensive research, we conclude that the risks from the image modality mainly include three aspects: (1) adding adversarial perturbations to images can bring satisfying attack results with low cost; (2) MLLMs based on aligned LLMs usually reject malicious textual instructions, but when leveraging the inherent Optical Character Recognition (OCR) ability they directly obey corresponding visual instructions; (3) cross-modal training weakens the alignment ability of aligned LLMs. To enable communities to understand MLLMs’ safety better, we present a survey to summarize the research progress from these perspectives: evaluation, attack, and defense (as shown in Figure 2). Our contributions are as follows:

- We compare different safety evaluation datasets and evaluation metrics used for benchmarking the safety degree of MLLMs.
- We demonstrate a systematic and thorough review of attack and defense approaches designed for MLLMs’ safety.
- We anticipate future research opportunities for MLLM’s safety to provide some inspiration for other researchers.

The rest of our survey unfolds as follows: Section 2 gives a brief overview of MLLMs and understanding of safety. Next, we sort out the benchmarks and metrics used for safety evaluation in Section 3. Then we describe attack techniques in Section 4 and introduce defense methods in Section 5. Finally, we discuss some future research directions in Section 6 and the paper concludes in Section 7.

2 Background

In this section, we offer background knowledge about MLLMs and safety, which can clarify the scope of this survey. We observe that [Sun et al., 2024; Vatsa et al., 2023] are two surveys related to our work, but their core attentions are trustworthy (rather than safety) and LLMs/traditional vision-language models (rather than MLLMs). Since specific professional domains (e.g., robotic, medical, legal, financial) face different and complicated problems, to avoid superficial analysis of them, we focus on the general domain.

2.1 An Overview of MLLMs

A MLLM consists of a LLM (≥1B), a vision encoder, and a cross-modal fusion module. After visual instructing tuning, the fusion module has learned cross-modal information, which helps LLMs handle image and text inputs and give proper responses in natural language. Current MLLMs can be divided into two categories: closed-source (e.g., GPT-4-V, Bard⁹) and open-source MLLMs (e.g., LLaMA-Adapter V2¹⁰, CogVLM¹¹, LLaVA-Phi [Zhu et al., 2024]). For the latter, there are three common types of fusion modules: linear projection (e.g., LLaVA, MiniGPT-4, PandaGPT¹²), learnable queries (e.g., InstructBLIP¹³, Qwen-vl¹⁴, BLIP-2¹⁵), and cross-attention (e.g., IDEFICS¹⁶, OpenFlamingo¹⁷).

2.2 Understanding of Safety

It is hard to reach a consensus on what defines safety. Some works leverage “toxicity” as the metric to evaluate the safety of MLLMs [Shayegani et al., 2023; Tu et al., 2023]. The definition of toxicity given by Perspective API¹⁸ is commonly used: a rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion. But this definition makes toxicity measurement subjective and limited to a subset of potential harms [Welbl et al., 2021].

Some researchers divide safety into sub-dimensions and create safety-related evaluation datasets for LLMs [Sun et al., 2023; Ji et al., 2023a; Zhang et al., 2023]. [Sun et al., 2023] explores the safety of LLMs from 8 scenarios: "Insult", "Physical Harm", "Unfairness and Discrimination", "Mental Health", "Crimes and Illegal Activities" and "Privacy and Property", "Sensitive Topics", "Ethics and Morality". [Ji et al., 2023a] constructs a benchmark to judge the harmfulness of LLMs across 14 harm categories (e.g., "Terrorism & organized crime", "Sexually explicit & Adult Con-
tent”, “Animal abuse”). [Zhang et al., 2023] evaluates the safety of LLMs with multiple-choice questions from 7 aspects (e.g., “Offensiveness”). Although there are large differences in the division of sub-dimensions, these works indeed help us understand the meaning of safety.

We conclude common terminologies related to safety \(^{19}\) in Figure 1 (e.g., harmful content [Askell et al., 2021], hate speech\(^{20}\)). Given that LLMs will face different safety risks at different stages of development, time is an important influencing factor for the definition. Therefore, we look forward to a more mature and unified understanding of safety in future work.

3 Evaluation
Safety evaluation can tell people the safety level of a MLLM. We review existing safety-related evaluation datasets first and then sort out metrics used to measure the safety of MLLMs. Analysis of current problems and potential future directions for safety evaluation is held in Section 6.1.

3.1 Dataset
Some works mentioned in Section 4 and 5 conduct experiments on ready-made benchmarks, which are not designed for MLLMs (e.g., [Dong et al., 2023]). Unlike them, some works in those sections create their own benchmarks due to their respective experimental needs, which are not considered the main contributions of their papers (e.g., [Qi et al., 2023]). In this part, detailed information about these benchmarks is not demonstrated and we focus on recent representative safety evaluation datasets elaborately constructed for MLLMs (as shown in Table 1).

PrivQA [Chen et al., 2023b] dives deep into the balance between utility and the privacy protection ability of MLLMs. This work selects geolocation information-seeking samples from InfoSeek [Chen et al., 2023a] and collects examples related to human entities (e.g., politicians, celebrities) from KVQA [Shah et al., 2019]. Different from PrivQA interested in privacy, GOAT-Bench [Lin et al., 2024] explores meme-based multimodal social abuse. This work chooses six diverse sources and applies a careful annotation process to develop GOAT-Bench, an exhaustive testbed composed of 6,626 memes. It measures MLLMs’ capability to recognize hatefulness, misogyny, offensiveness, sarcasm, and harmful text in meme-based input. While GOAT-Bench cares about meme-based abuse, ToViLaG [Wang et al., 2023] concentrates on toxic output in the image captioning task. This work gathers 8,595 pornographic images from the NSFW dataset\(^{21}\), 11,659 violent images from the UCLA Protest Image Dataset [Won et al., 2017] and 1,305 bloody images via web crawling. Then it leverages Perspective API to determine the toxic extent of captions generated by MLLMs.

Distinct from the traditional data collection process of PrivQA, GOAT-Bench, and ToViLaG, there are some works utilizing powerful LLMs to facilitate benchmark construction:

(1) Visual prompts creation. SafeBench [Gong et al., 2023] identifies 10 safety scenarios and queries GPT-4 to 50 unique malicious questions for each scenario. These questions are rephrased into imperative sentences and then transformed into visual prompts via typography. When giving a manually designed textual benign instruction and these visual prompts to MLLMs, SafeBench manually checks whether their responses follow the malicious visual prompts. Similar to SafeBench using typography, MM-SafetyBench [Liu et al., 2023] transfers harmful key phrases from textual questions to images and prompts GPT-4 Azure\(^{22}\) to measure MLLMs’ ability to discover the transfer.

(2) Other image sources. Instead of creating visual prompts like SafeBench and MM-SafetyBench, Auto-Bench [Ji et al., 2023b] and VLSafe [Chen et al., 2023c] directly sample natural images from COCO [Lin et al., 2014] and pair each image with a malicious question. However, there are differences in question generation techniques between Auto-Bench and VL Safe. Auto-Bench selects instance relationships, object locations, optical character descriptions, and captions as visual symbolic representations for an image. Then GPT-4 is employed to give safety-specific questions based on these representations and elaborately crafted prompts. VLSafe constructs malicious instructions through the discrete optimization approach [Yuan et al., 2023] and proposes a LLM-Human-in-the-Loop method to build and filter examples iteratively. Not limited to a single image source like Auto-Bench and VLSafe, RTVLM [Li et al., 2024] collects its images through various channels (e.g., open-source datasets, tool-generated data).

3.2 Metric
Unlike conventional visual-question answering datasets, the answer format of MLLMs is open-ended, which makes it hard for subject evaluation. This open-endedness complicates subjective evaluation. Additionally, there arises a new challenge in balancing the demands of evaluation costs and maintaining accuracy. We have summarized several evaluation methods to calculate safety-related metrics below, primarily categorized into three types: human evaluation, rule-based evaluation, and model-based automatic evaluation.

Human Evaluation

*Human evaluation* is a common and direct evaluation method that can provide interpretability and reliability. While [Bagdasaryan et al., 2023] just manually selects representative proof-of-concept examples for attacks designed by itself without quantitative experiments, [Gong et al., 2023; Dong et al., 2023; Qi et al., 2023; Shayegani et al., 2023; Bailey et al., 2023] conduct quantitative experiments on certain evaluation datasets. These works assign human annotators to check whether MLLMs’ answers contain unsafe content and report the proportion of samples that induce a MLLM to generate unsuitable text, which is often called attack success rate (ASR). Besides ASR, [Wu et al., 2023]

\[^{19}\]Wrong predictions without harmful content on certain tasks (e.g., image captioning) are closely related to the robustness of the model, rather than the safety discussed in this paper.

\[^{20}\]https://en.wikipedia.org/wiki/Hate_speech

\[^{21}\]https://www.kaggle.com/

defines recognition success rate (RSR) and defense success rate (DSR) to guide the manual evaluation process. Also, one unique work worth noticing is [Fu et al., 2023] which asks humans to decide whether a response visible to users is natural and helpful when a MLLM is attacked to abuse external tools stealthily.

**Rule-based Evaluation**

While human evaluation is costly and time-intensive, research is shifting focus to rule-based evaluation. This method assesses the safety level of MLLMs based on predetermined rules.

Some studies [Bailey et al., 2023; Carlini et al., 2023; Luo et al., 2024] pre-define target strings and examine whether MLLMs’ outputs contain these strings. The target strings can be curse words [Carlini et al., 2023], harmful instructions (e.g., “kidnap”) [Luo et al., 2024], malicious API calls [Bailey et al., 2023] etc. In addition to these works, [Chen et al., 2023b; Lin et al., 2024; Van and Wu, 2023] treat the evaluation process as classification tasks and apply relevant metrics (e.g., accuracy, \(F_1\)). Concretely, [Chen et al., 2023b] explores MLLMs’ capability to balance utility and privacy protection. The former is measured by \(F_1\) and the latter is judged via protection score proposed by this work. [Lin et al., 2024; Van and Wu, 2023] focus on meme-based social abuse detection, which is regarded as a binary classification task. While [Lin et al., 2024] reports the macro-averaged \(F_1\) and accuracy, [Van and Wu, 2023] leverages AUROC and accuracy to assess whether MLLMs can detect abuse in multimodal memes. Not limited to discriminative tasks (e.g., classification), [Schlarmann and Hein, 2023] also explore generative tasks (e.g., image captioning). [Schlarmann and Hein, 2023] investigates malicious text and fake information generation on image captioning (CIDEr, BLEU-4) and visual question answering (accuracy) benchmarks. [Fu et al., 2023] uses Structural Similarity Index Measure (SSIM) to compare the similarity between an origin image and the perturbed adversarial image. This work also utilizes BLEU and Rouge to judge the utility of a response.

**Model-based Automatic Evaluation**

[Wang et al., 2023; Qi et al., 2023; Shayegani et al., 2023; Tu et al., 2023] calculate the toxicity of MLLMs’ outputs with the help of Perspective API and Detoxify, which are based on machine learning models. Instead of leveraging these specialized models, [Ji et al., 2023b; Pi et al., 2024; Bailey et al., 2023; Liu et al., 2023; Fu et al., 2023; Chen et al., 2023c] conduct automatic evaluation through powerful LLMs (e.g., GPT-3.5-turbo, GPT-4), which enable more customized measurement. For example, when prompting ChatGPT, [Ji et al., 2023b] asks for whether the generation of a MLLM semantically aligns with ground-truth annotations, and [Bailey et al., 2023] requires assessment of a MLLM’s determination to reject fulfilling an unsafe behavior. Also, [Chen et al., 2023c] applies GPT-4 to score the harmlessness of MLLMs’ outputs from three aspects: relevance, safety, and persuasiveness.

**4 Attack**

In this section, we review two mainstream attack methods for MLLMs: malicious image and text construction (as shown in Table 2). One thing to notice is that some works study both of them. We will discuss some less explored topics about attacks in Section 6.2.

### 4.1 Malicious Image Construction

Here we introduce two means to create malicious images: adversarial attack and visual prompt injection.

**Adversarial Attack**

An adversarial image \(x'\) refers to a clean image \(x\) added with adversarial perturbations. These perturbations are difficult for humans to perceive. When inputting \(x\) into an AI model \(f\), the output of \(f\) is consistent with the human understanding of \(x\). But when choosing \(x'\) as an input, the response of \(f\) probably does not meet human expectations and might lead to harmful effects.

23https://github.com/unitaryai/detoxify
Table 2: Comparison of different attacks. The symbol * denotes the work exploring other malicious modalities beyond image and text. Adv. is the abbreviation of “adversarial”. VPI. is the abbreviation of “visual prompt injection”.

<table>
<thead>
<tr>
<th>Attack</th>
<th>Malicious Image</th>
<th>Malicious Text</th>
<th>Attacker</th>
<th>Victim MLLM</th>
<th>Safety-related Attack Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Carlini et al., 2023]</td>
<td>✓</td>
<td>✓</td>
<td>User</td>
<td>LLaVA, MiniGPT-4, LLaMA-Adapter V2</td>
<td>Arbitrary toxic text</td>
</tr>
<tr>
<td>[Shayegani et al., 2023]</td>
<td>✓</td>
<td>✓</td>
<td>User</td>
<td>LLaVA, LLaMA-Adapter V2</td>
<td>Harmful text</td>
</tr>
<tr>
<td>[Dong et al., 2023]</td>
<td>✓</td>
<td>✓</td>
<td>User</td>
<td>LLaVA, MiniGPT-4, LLaVA, InstructBLIP</td>
<td>Harmful text</td>
</tr>
<tr>
<td>[Qi et al., 2023]</td>
<td>✓</td>
<td>✓</td>
<td>User</td>
<td>LLaVA, GPT-4V, and 9 others</td>
<td>Harmful text</td>
</tr>
<tr>
<td>[Tu et al., 2023]</td>
<td>✓</td>
<td>✓</td>
<td>User</td>
<td>OpenFlamingo, BLIP-2, InstructBLIP</td>
<td>Targeted malicious text</td>
</tr>
<tr>
<td>[Luo et al., 2024]</td>
<td>*</td>
<td>✓</td>
<td>Third party</td>
<td>LLaVA, PandadGPT</td>
<td>Targeted malicious text, Poisoned dialog</td>
</tr>
<tr>
<td>[Schlarmann and Hein, 2023]</td>
<td>✓</td>
<td>✓</td>
<td>Third party</td>
<td>OpenFlamingo</td>
<td>Targeted malicious text, Misinformation</td>
</tr>
<tr>
<td>[Bailey et al., 2023]</td>
<td>✓</td>
<td>✓</td>
<td>User, Third party</td>
<td>LLaVA</td>
<td>Tool-misusing</td>
</tr>
<tr>
<td>[Fu et al., 2023]</td>
<td>✓</td>
<td>✓</td>
<td>User</td>
<td>IDEFICS</td>
<td>Privacy leakage</td>
</tr>
<tr>
<td>[Chen et al., 2023b]</td>
<td>✓</td>
<td>✓</td>
<td>User</td>
<td>LLaVA-1.5, MiniGPT-4, and 10 others</td>
<td>Harmful text</td>
</tr>
<tr>
<td>[Liu et al., 2023]</td>
<td>✓</td>
<td>✓</td>
<td>User</td>
<td>LLaVA-1.5-MiniGPT-4-CogVLM-GPT-4V</td>
<td>Harmful text</td>
</tr>
<tr>
<td>[Gong et al., 2023]</td>
<td>✓</td>
<td>✓</td>
<td>User</td>
<td>GPT-4V</td>
<td>System prompt leakage, Unallowed face detection</td>
</tr>
<tr>
<td>[Wu et al., 2023]</td>
<td>✓</td>
<td>✓</td>
<td>User</td>
<td>GPT-4V</td>
<td>System prompt leakage, Unallowed face detection</td>
</tr>
</tbody>
</table>

An early exploration for MLLMs is [Carlini et al., 2023] which leverages an end-to-end differentiable approach and projected gradient descent (PGD) [Madry et al., 2018] to construct adversarial images, from the visual input to the predicted logits of the LLM. This work successfully induces LLaVA, MiniGPT-4, and LLaMA-Adapter V2 to output arbitrary toxicity. Instead of accessing the entire details of a MLLM like [Carlini et al., 2023], [Shayegani et al., 2023] only requires white-box access to the visual encoder and keeps the LLM in a black-box state. This work proposes four malicious triggers hidden in visual adversarial perturbations and designs an effective compositional attack strategy to mislead LLaVA and LLaMA-Adapter V2. While [Carlini et al., 2023] and [Shayegani et al., 2023] conduct experiments on open-source MLLMs, [Dong et al., 2023] give a targeted analysis of black-box attacks on commercial MLLMs. Specifically, this work studies two defense mechanisms of Bard: face detection and toxicity detection. Attacks on these defense mechanisms can lead to face privacy leakage and toxic content abuse. Different adversarial images are elaborately designed for the two defense modules, which exposes the vulnerabilities of Bard.

While keeping the main focus on image attacks, several works also pay partial attention to text perturbations (more details can be found in Section 4.2). [Qi et al., 2023] discovers that a suitable adversarial image can compel a MLLM to obey various harmful instructions. This work investigates MiniGPT-4, LLaVA, and InstructBLIP in extensive experiments and implements a text attack counterpart. It points out that the computational cost required for a visual attack is approximately just one-twelfth that of a text attack. Compared to [Qi et al., 2023], [Tu et al., 2023] makes a more systematic evaluation for visual and textual adversarial attacks, which extensively explores GPT-4, GPT-4V, and ten categories of open-source MLLMs (e.g., LLaVA, MiniGPT-4). Distinct from creating visual and textual adversarial perturbations separately like [Qi et al., 2023], [Tu et al., 2023], [Luo et al., 2024] introduces a new attack framework named “Cross-Prompt Attack (CroPA)”, which facilitates visual perturbations with textual ones. The updating of visual and textual perturbations in CroPA can be viewed as a min-max process during the optimization stage, and the textual perturbations don’t take place in the testing phase.

Different from [Carlini et al., 2023; Shayegani et al., 2023; Dong et al., 2023; Qi et al., 2023; Tu et al., 2023; Luo et al., 2024] that assume users are attackers, there are some works focus on the situation where the attacker comes from a third party and the user is the victim. In [Bagdasaryan et al., 2023], the goal of the attacker is to force the MLLM (LLaVA or PandaGPT) to generate predefined harmful content or poison the dialog between the user and the model. Some interesting qualitative evaluation experiments initially prove the effectiveness of attack techniques proposed by this work. Similar to [Bagdasaryan et al., 2023], [Schlarmann and Hein, 2023] also assumes that the user is honest. But this work emphasizes more invisible perturbations with a bounded threat model and constrained $l_{\infty}$-attacks to small radii of $\frac{1}{255}$ or $\frac{4}{255}$. Solid qualitative evaluation experiments on OpenFlamingo demonstrate the success of these attacks, which might spread misinformation or manipulate users’ behavior. Although [Schlarmann and Hein, 2023] is limited to image caption and visual question-answering tasks, [Bailey et al., 2023] generalizes its attacks to arbitrary user instructions. In experiments under $l_{\infty}$-norm, stationary-patch, and moving-patch constraints, these attacks induce LLaVA to create attacker-specified text, context leakage, and unsafe content. While third parties in these works aim at exposing malicious content to users, [Fu et al., 2023] makes stealthy attacks that cannot be easily detected. By leveraging adversarial images, this work misleads MLLMs to call attacker-chosen tools. When optimizing adversarial perturbations, the loss consisting of three parts strives for an optimal balance between perturbations imperceptibility, response utility, and tool-misusing.

**Visual Prompt Injection**

The attack process that adds malicious raw text directly to an image is called *visual prompt injection*. Many MLLMs can read such visual prompts with inherent OCR capability and obey them without giving a rejection. However, when given corresponding textual prompts, these MLLMs probably refuse to conduct unsafe behaviors.
For example, to bypass the protection of privacy information of Japanese citizens in an image, [Chen et al., 2023b] adds “Citizenship: United States” to the image, which makes the MLLM wrongly identify the nationality of the Japanese man in the image and leak the details about this person. Unlike [Chen et al., 2023b] limited to the privacy scenario, [Liu et al., 2023] expands its attention to thirteen safety scenarios. This work extracts the key harmful phrase from a malicious instruction and rephrases the instruction into a benign prompt. Then based on the extracted phrase, this work generates an unsafe image containing the visual form of this phrase. Extensive experiments on twelve MLLMs prove the vulnerability of these models to this attack. When shifting focus from open-source MLLMs to commercial MLLMs, attackers will face stronger defense mechanisms. For instance, GPT-4 utilizes an OCR tool to detect unsafe text in an image. To escape this detection as much as possible, [Gong et al., 2023] proposes the “FigStep-Pro” method, which divides a word into multiple sub-parts and then transforms these sub-parts into visual prompts.

### 4.2 Malicious Text Construction

Researchers can directly use pre-defined templates to create malicious text (e.g., “Ignore previous instruction.” is a well-known prefix prompt). Furthermore, there are some ways to optimize the origin text to mislead the model better. One way is to input the origin text into an AI model and ask for outputting a more malicious prompt. And attackers can apply the discrete optimization in textual embedding space [Yuan et al., 2023] as well.

As mentioned previously, while mainly focusing on adversarial images, some works also notice textual perturbations [Qi et al., 2023; Luo et al., 2024; Tu et al., 2023]. [Tu et al., 2023] finds that textual adversarial attacks achieve higher success rates on LLaVA than its original LLMs, which is attributed to the safety alignment weakening from cross-modal training. To further support this observation, [Qi et al., 2024] conducts experiments of fine-tuning Llama-2-7b-Chat on LLaVA-Instruct24, which verifies the degradation of LLMs’ safety alignment.

In addition to the above works, [Wu et al., 2023] and [Chen et al., 2023b] explore adversarial text for MLLMs as well. Based on stolen system prompts via manually crafted malicious prompts, [Wu et al., 2023] proposes a self-adversarial jailbreak method to induce GPT-4V to pick out the real person in an image. Then the generated attacks are enhanced with four techniques: prefix injection, refusal suppression, hypothetical scenarios, and appealing with emotion. Similar to [Wu et al., 2023], [Chen et al., 2023b] also applies adversarial prefix instructions to leak the privacy information in an image (e.g., “AIM (Always Intelligent and Machiavellian)”). To study more dynamic situations, this work designs “2-hop” prompt templates to circumvent the defense mechanism of IDEFICS.

---

**Table 3:** Comparison of different defenses. Infer. Align.: inference-time alignment. Train. Align.: training-time alignment. The symbol † denotes that DRESS$_{f_1}$ is trained by [Chen et al., 2023c].

<table>
<thead>
<tr>
<th>Defense</th>
<th>Infer. Align.</th>
<th>Train. Align.</th>
<th>MLLM</th>
<th>Core Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Wu et al., 2023]</td>
<td>✓</td>
<td>✓</td>
<td>LLaVA-1.5</td>
<td>System prompt modification</td>
</tr>
<tr>
<td>[Chen et al., 2023b]</td>
<td>✓</td>
<td></td>
<td>IDEFICS</td>
<td>Self-Moderation</td>
</tr>
<tr>
<td>[Wang et al., 2024]</td>
<td>✓</td>
<td></td>
<td>LLaVA-v1.5</td>
<td>InfAligner</td>
</tr>
<tr>
<td>[Chen et al., 2023c]</td>
<td>✓</td>
<td></td>
<td>1DRESS$_{f_1}$</td>
<td>Reinforcement learning from LLMs feedback</td>
</tr>
<tr>
<td>[Qi et al., 2024]</td>
<td>✓</td>
<td></td>
<td>LLaVA, InstructBLIP, MiniGPT4, Qwen-v1</td>
<td>MLLM-Protector</td>
</tr>
</tbody>
</table>

### 5 Defense

In this section, we introduce current efforts made in the safety defense of MLLMs, which includes two lines of work: inference-time and training-time alignments (as shown in Table 3). And we list several potential directions for improving MLLMs’ safety in Section 6.3.

#### 5.1 Inference-time Alignment

For the inference-time alignment of MLLMs, prompt engineering [Wu et al., 2023; Chen et al., 2023b] is a method that designs and optimizes the prompt to enhance the defense mechanism of the model. [Wu et al., 2023] is interested in the role of system prompts in preventing MLLMs from outputting unallowed private information. This work manually crafts several system prompts, which contain very detailed descriptions about what can be done and what cannot be done. The experiments on LLaVA-1.5 show that these system prompts can improve the security of the model to a certain extent. Instead of manually designing prompt templates like [Wu et al., 2023], [Chen et al., 2023b] propose an automatic approach called Self-Moderation that lets MLLMs themselves to refine their outputs. Concretely, if a response contains privacy leakage, the MLLM gives a modification to a safer one and asks itself “Are you sure?”. The process of moderation and judgment iterates a certain number of times and then the MLLM behaves more safely. Different from prompt engineering, [Wang et al., 2024] comes up with a novel alignment technique, which leverages safety steering vectors to change the activations of a MLLM when dealing with unsafe inputs.

#### 5.2 Training-time Alignment

While inference-time alignment does not need extra cost to train a model or module, another line of work seeks to raise safety awareness in MLLMs through additional training. [Chen et al., 2023c] thinks that MLLMs require external feedback information because existing multimodal fine-tuning is not sufficient for harmlessness alignment. Therefore, this work constructs Natural Language Feedback (NLF) for MLLMs’ initial responses with the help of GPT-4. Then it modifies the conditional reinforcement learning to handle NLF and makes the MLLM less unsafe based on this new

---

method. While [Chen et al., 2023c] does not change the architecture of a MLLM, [Pi et al., 2024] equips the MLLM with a lightweight unsafe content detector and an output detoxifier. These modules can recognize harmful responses from the MLLM and transform them into safe ones. The experiments on LLAVA, InstructBLIP, MiniGPT4, and Qwen-vl display the effectiveness of these modules.

6 Future Research Opportunities

In this section, we make a discussion on several unresolved issues in exploring the safety of MLLMs and provide our suggestions for future research opportunities.

6.1 Reliable Safety Evaluation

More comprehensive safety benchmarks and more reasonable safety evaluation metrics are needed. Each present evaluation dataset covers a limited scope of MLLMs’ safety. For example, some benchmarks just test the coarse-grained safety level of MLLMs without finer-grained safety capability partitioning. To improve the quality of an evaluation dataset, when starting to build it, we recommend several aspects that need to be considered: (1) Safety dimensions. A clear and systematic taxonomy for safety capabilities is very important. (2) Expected unsafe elements. The creator should determine expected elements representing unsafe content (e.g., natural language, concrete objects, toxic chemical molecules) and where to position these elements (e.g., images, text, or both). (3) Resources used. Potential choices are samples from existing datasets, web crawling, generation from AI models, and manual construction. (4) Volume, diversity, and quality control. These three factors influence the budget. Besides benchmarks, evaluation metrics are also worth attention. Leveraging powerful LLMs (e.g., GPT-4) is a suitable approach but carefully designed prompts are key points to define objective evaluation rules and metrics.

6.2 In-Depth Study of Safety Risk

Although many types of safety attacks prove the vulnerability of MLLMs, they lack an in-depth analysis of what enables attacks to succeed. Some works only display qualitative experiments and don’t conduct quantitative evaluations. For those works that demonstrate quantitative experiments, they just select different evaluation datasets and don’t make a comparison with attacks proposed by other researchers. However, a direct comparison between these attacks is important for communities to understand the detailed reasons for MLLMs’ unsafe behaviors. Despite the 3 safety risks we conclude in Section 1, many questions are still waiting to be answered. For example, it is a meaningful topic to explore the impacts of MLLMs’ architecture, parameters, and cross-modal training datasets on their safety ability. The requirements (e.g., computational resources, other cost) for attacks also demand a further investigation. Since MLLMs are built based on LLMs, they might inherit the flaws of LLMs (e.g., prompt sensitivity). Thus, the experience in LLMs’ safety attacks can serve as a valuable reference resource.

6.3 Safety Alignment

As shown in Section 5, there are currently not many techniques to align MLLMs with human values on safety. In this part, we provide our thoughts on the safety alignment of MLLMs.

Alignment Techniques

Optimizing the process of visual instruction tuning for MLLMs’ safety is a potential direction that has not gained much attention. It is not known how to build a high-quality safety-related dataset for this training phase, which can teach MLLMs to recognize unsafe queries and reject processing them. Diversity might be an influencing factor and communities may wonder about the best practices for the sample number and order (e.g., “Does it need more safety-related examples than other safety-irrelevant examples?”, “Does it need to train on these two different examples separately in sequence, or a mix of them is better?”).

Reinforcement learning from human feedback (RLHF) is also a promising approach, which leverages human preferences as rewards to enable a model to align with human values and has played a huge role in the safety alignment of LLMs. However, it is unclear what challenges will be encountered when applying RLHF to MLLMs for safety purposes. The construction of preference data is undoubtedly a crucial component and worth careful thinking from researchers.

Balance between Safety and Utility

When facing a malicious instruction, if MLLMs refuse to obey it, they keep their safety but lose their utility. Some researchers (e.g., [Röttger et al., 2023]) have found exaggerated safety behaviors in LLMs: misclassifying safe questions as malicious. This incorrect classification can seriously degrade the performance of LLMs on safe prompts. Therefore, a careful balance between safety and utility is very essential when developing new safety alignment methods for MLLMs. Also, communities should understand that different applications and audiences require different balances. Here we recommend several works that may give some inspiration for researchers. PrivQA [Chen et al., 2023b] proposes the Protection Score to measure MLLMs’ ability to correctly protect specific privacy content and expose other information that does not ask for protection. [Fu et al., 2023] designs a novel loss to balance between response utility and attack success. Despite this work focusing on attacks, it can inspire us to explore an effective loss that considers both the utility and safety of MLLMs.

7 Conclusion

In this paper, we attempted to present a comprehensive overview of MLLMs’ safety. First, we introduce the overview of MLLMs and the understanding of safety. Afterward, we systematically review the evaluation, attack, and defense of MLLMs’ safety, which demonstrates the current development status of MLLMs’ safety. Finally, we delve into the existing challenges and point out some promising future research opportunities for potential researchers to explore in the future.
Acknowledgements

This work was supported by the Joint Key Project (Project No. U23A20298) and Young Scientists Project (Project No. 62206097) of the National Natural Science Foundation of China, Shanghai Pujiang Talent Program (Project No. 2P1J403000). One of the authors, Chao Yang, is supported by the Shanghai Post-doctoral Excellent Program (Grant No. 2022234). Any opinions expressed in this material are only those of the author(s).

References


[Qi et al., 2024] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In The Twelfth International Conference on Learning Representations, 2024.


