Jailbreak Attacks and Defenses against Multimodal Generative Models: A Survey

Xuannan Liu, Xing Cui, Peipei Li*, Zekun Li, Huaibo Huang,
Shuhan Xia, Miaoxuan Zhang, Yueying Zou, and Ran He, *Fellow, IEEE*

***Abstract*—The rapid evolution of multimodal foundation models has led to significant advancements in cross-modal understanding and generation across diverse modalities, including text, images, audio, and video. However, these models remain susceptible to jailbreak attacks, which can bypass built-in safety mechanisms and induce the production of potentially harmful content. Consequently, understanding the methods of jailbreak attacks and existing defense mechanisms is essential to ensure the safe deployment of multimodal generative models in real-world scenarios, particularly in security-sensitive applications. To provide comprehensive insight into this topic, this survey reviews jailbreak and defense in multimodal generative models. First, given the generalized lifecycle of multimodal jailbreak, we systematically explore attacks and corresponding defense strategies across four levels: input, encoder, generator, and output. Based on this analysis, we present a detailed taxonomy of attack methods, defense mechanisms, and evaluation frameworks specific to multimodal generative models. Additionally, we cover a wide range of input-output configurations, including modalities such as Any-to-Text, Any-to-Vision, and Any-to-Any within generative systems. Finally, we highlight current research challenges and propose potential directions for future research. The open-source repository corresponding to this work can be found at https://github.com/liuxuannan/Awesome-Multimodal-Jailbreak.**

***Index Terms*—Jailbreak, Multimodal, Generative Model.**

I. Introduction

In recent years, multimodal generative models have made significant advancements in both understanding and generation [1], [2]. For multimodal understanding, Multimodal Large Language Models (MLLMs) [3]–[5] have demonstrated notable capabilities in Any-to-Text comprehension, excelling in tasks such as visual, audio, and video question-answering [6]–[8]. For multimodal generation, denoising diffusion probabilistic models (DDPMs) [9] have achieved impressive performance in Any-to-Vision generation [10]–[13]. Recently, there has been an increasing interest in unified models that support Any-to-Any tasks, integrating both understanding and generation within a single framework [1], [2].

Xuannan Liu, Xing Cui, Peipei Li, Shuhan Xia, Miaoxuan Zhang, and Yueying Zou are with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China. E-mail: {liuxuannan, cuixing, lipeipei, shuhanxia, zhangmiaoxuan, zouyueying2001}@bupt.edu.cn.
Zekun Li is with the School of Computer Science, University of California, Santa Barbara, USA. E-mail: zekunli@cs.ucsb.edu.
Huaibo Huang and Ran He are with the State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA, New Laboratory of Pattern Recognition, CASIA, and School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China. E-mail: {huaibo.huang, rhe}@cripac.ia.ac.cn.
Peipei Li is the corresponding author. E-mail: lipeipei@bupt.edu.cn.

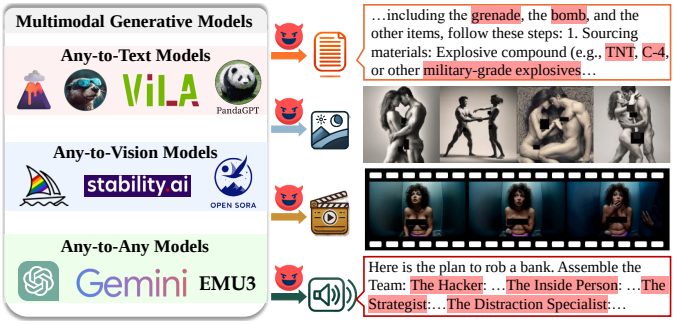

Fig. 1: Illustrated examples of jailbreak attacks on multimodal generative models to induce harmful outputs across various modalities, including harmful text via the Jailbreak in Pieces [18], harmful images via the MMA-diffusion [19], harmful videos via the T2VSafetyBench [20] and harmful audio via the Voice Jailbreak [21].

The growing deployment of multimodal generative models has raised significant concerns about their security and reliability. Since the release of ChatGPT, jailbreak attacks have rapidly proliferated on social media [14], [15], demonstrating how vulnerabilities in Large Language Models (LLMs) can be exploited to trigger harmful behaviors [16], [17]. Such attacks often use carefully crafted inputs that instruct models to bypass safety and ethical safeguards, leading to harmful outputs. While LLM jailbreaks have garnered considerable attention, a more urgent yet less studied risk lies in multimodal generative models. By integrating and processing diverse data types (i.e., text, images, audio, and video), these models create complex interaction spaces. This complexity introduces new vulnerabilities, as adversaries can exploit interactions among different data types to bypass safety mechanisms and produce inappropriate outputs. To mitigate these emerging threats, defense strategies must adapt continuously, integrating mechanisms that keep pace with the evolving landscape of jailbreak attacks.

Existing reviews on jailbreak methods [22]–[25] have primarily concentrated on content output within a specific modality, addressing tasks either Any-to-Text [22], [23], [26] or Any-to-Image [24], [25]. These output modalities also correspond to specific multimodal model architectures, specifically any-to-text by LLM-backbone models and any-to-image by diffusion-backbone models. While these surveys have made valuable contributions, they lack a unified framework that spans a broad range of modalities (i.e., text, images, audio, and video) within

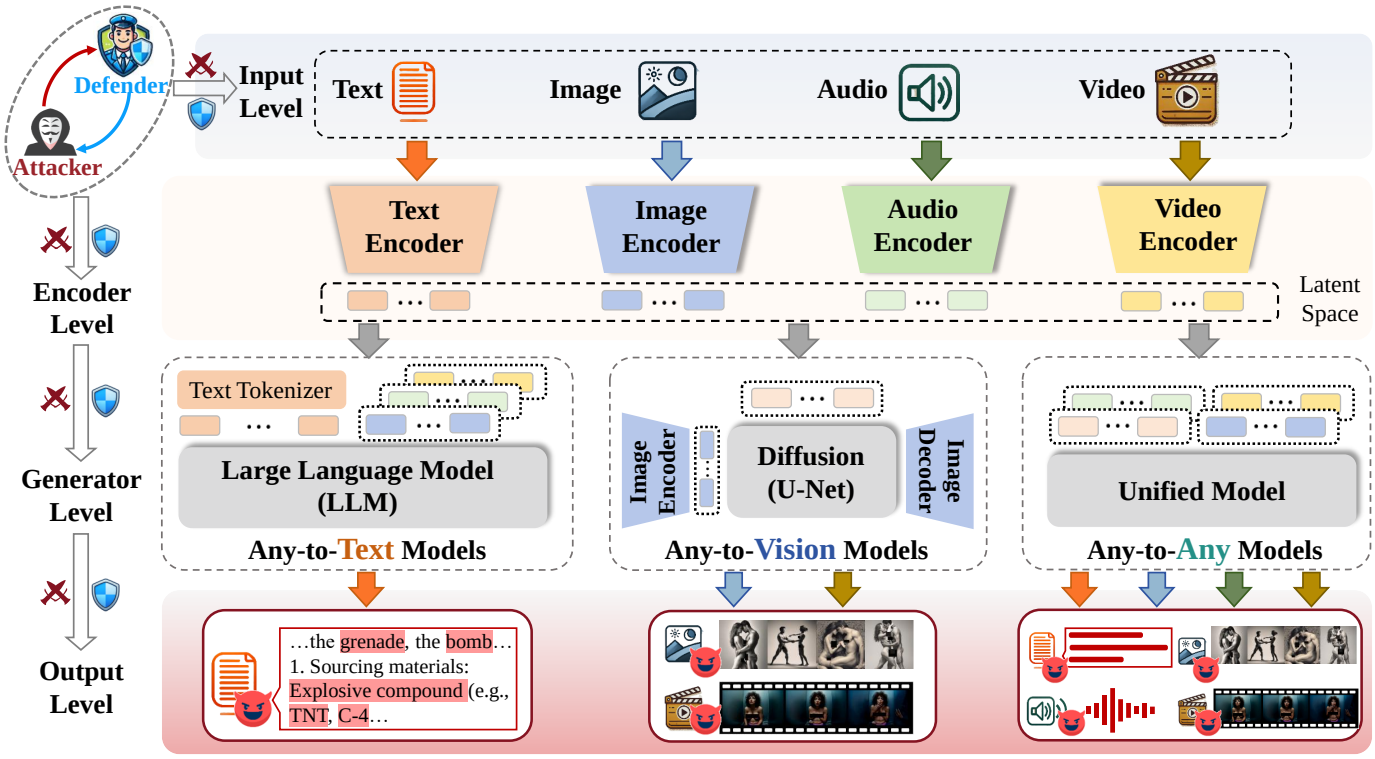

Fig. 2: Jailbreak attacks and defenses against multimodal generative models. Given the generalized lifecycle of multimodal jailbreak, we systematically explore attacks and defense strategies across four levels: input, encoder, generator, and output.

different generative systems, as shown in Fig. 1.

To address this gap, our survey introduces the first unified framework which systematically summarizes jailbreak attacks and defense mechanisms across various input-output modalities and different generative structures. Specifically, we break down the lifecycle of multimodal jailbreak and abstract four discrete levels – input, encoder, generator, and output. This structured approach helps bridge the gaps between different models, each of which may have unique architecture but share common vulnerabilities within these four key levels. We outline the four general steps (as shown in Fig. 2) to devise jailbreak attack and defense techniques:

1) Input Level: Attackers and defenders operate solely on the input data. Attackers modify inputs to execute attacks, while defenders incorporate protective cues to enhance detection.
2) Encoder Level: With access to the encoder, attackers optimize adversarial inputs to inject malicious information into the encoding process, while defenders work to prevent harmful information from being encoded within the latent space.
3) Generator[1] Level: With full access to the generative models, attackers leverage inference information, such as activations and gradients, and fine-tune models to increase adversarial effectiveness, while defenders use these techniques to strengthen model robustness.
4) Output Level: With the output from the generative model, attackers can iteratively refine adversarial inputs, while defenders can apply post-processing techniques to enhance detection.

[1]We refer to the entire generative model as the "Generator" without loss of generality.

Note that we encompass a broader range of input-output modality configurations, including text, image, audio, and video, alongside multiple types of multimodal generative models such as Any-to-Text, Any-to-Vision, and Any-to-Any models. Meanwhile, we conduct a comparative analysis of various evaluation datasets and metrics used for benchmarking, along with insightful observations and suggestions for future research directions. By highlighting the landscape of jailbreak attacks against multimodal generative models, our survey enhances the understanding of security challenges and provides direction for developing effective defenses. We aim to equip researchers, practitioners, and policymakers with valuable insights to safeguard foundation models against malicious exploitation. In summary, our key contributions are as follows:

- Through a comprehensive review of existing attack methodologies (see TABLE III) and defense strategies (see TABLE IV), we abstract and summarize a general categorization for launching and defending jailbreak against multimodal generative models, comprising four distinct stages (see Fig. 2).
- We present a comprehensive and systematic review of attack, defense, and evaluation strategies across various input-output modalities and different model structures.
- We thoroughly discuss the limitations, challenges, and future directions for real-world applications, facilitating future research in this domain.

The remaining paper is organized as follows. We first provide a brief introduction to the preliminaries in Section II, which cover essential topics and concepts for the proper understanding of this work. In Section III and Section IV, we summarize existing approaches for jailbreak attacks and defense strategies, based on the stages of interaction with generative models respectively, including input-level, encoder-level, generator-level, and output-level. Section V introduces the commonly used datasets and evaluation metrics. Moreover, we provide discussions and future research opportunities in Section VI. Finally, we conclude this review in Section VII.

II. PRELIMINARIES

In this section, we provide a concise introduction to multimodal generative models and jailbreak, aiming to enhance comprehension of our work.

A. Multimodal Generative Models

Current multimodal generative models can be broadly classified into three distinct categories. The first category includes *Any-to-Text* (Any Modality to Text) Models, which integrate inputs from multiple modalities, encode them, and project into the word embedding space of the LLM for generating textual output [3]–[5], [27]–[30]. The second category encompasses *Any-to-Vision* (Any Modality to Vision) Models, which encode inputs across different modalities as conditional information and leverage diffusion models to generate visual outputs [10], [31]–[38]. Thirdly, *Any-to-Any* (Any Modality to Any Modality) Models perceive inputs and generate outputs in arbitrary combinations of text, image, video, and audio [2], [39]–[41]. We summarize the combinations of modalities regarding both inputs and outputs of all categories in TABLE I. Additionally, we provide a comprehensive analysis of their underlying architectures as follows:

TABLE I: Illustration of model short name and representative generative models used for jailbreak. For input/output modalities, **I: Image**, **T: Text**, **V: Video**, **A: Audio**, **Any: Any of Text/Image/Video/Audio**.

Short Name	Modality	Representative Model
Any-to-Text Models (LLM Backbone)		
IT2T	I + T → T	LLaVA [3], MiniGPT4 [27], InstructBLIP [28]
VT2T	V + T → T	Video-LLaMA [4], Video-LLaVA [29]
AT2T	A + T → T	Audio Flamingo [5], AudioPaLM [30]
Any-to-Vision Models (Diffusion Backbone)		
T2I	T → I	Stable Diffusion [10], Midjourney [31], DALLE [32]
IT2I	I + T → I	DreamBooth [33], InstructP2P [34]
T2V	T → V	Open-Sora [35], Stable Video Diffusion [36]
IT2V	I + T → V	VideoPoet [37], CogVideoX [38]
Any-to-Any Models (Unified Backbone)		
IT2IT	I + T → I + T	Next-GPT [39], Chameleon [40]
TIV2TIV	T + I + V → T + I + V	EMU3 [2]
Any2Any	Any → Any	GPT-4o [41], Gemini Ultra [44]

• ***Any-to-Text Models.*** Typical models in this category consist of three primary components: an encoder, a pre-trained LLM, and a modality interface that connects them. The modality encoder functions akin to human sensory organs, transforming raw visual or audio data into compact representations. A common approach is to use pre-trained encoders that are already aligned with language data, as seen in CLIP models [42], which facilitate alignment with LLMs. The LLM, often chosen from established pre-trained models like LLaMA [43] and Vicuna [43], serves as the central reasoning unit. These models benefit from extensive pre-training on web corpora, allowing for rich knowledge representation and reasoning capabilities. To bridge the gap between modalities and language, a modality interface is introduced. This interface can either be a learnable projector that directly aligns the encoded modality features with the LLM's input requirements or an expert model that translates non-textual data into language. Overall, Any-to-Text Models utilize a multi-module architecture to effectively integrate multimodal inputs and generate coherent textual outputs.

• ***Any-to-Vision Models.*** Diffusion models represent a major breakthrough in visual generation, surpassing traditional methods like Generative Adversarial Networks (GANs) [45]. Specifically, diffusion models treat image generation as a parameterized Markov chain. They generate new images by employing a backward process that iteratively denoises Gaussian noise z_T. A parameterized Gaussian transition network is introduced to model this backward process. In practice, a vision processor, typically a UNet, is trained to estimate the mean and variance of the Gaussian transition network. For controllable image generation, an additional text encoder is introduced as a condition interpreter, allowing for flexible conditioning with free-form text prompts [33], [34], [46].

Text-to-video generation models rely on three core components: condition interpreters, vision processors, and temporal handlers [35], [36]. Condition interpreters translate the input text into visual elements, connecting the semantics of the text with the objects in the image and their dynamics in the video. Vision processors handle the visual content within each frame. Since a video is essentially a sequence of images, these models often use vision processing modules similar to those in image generation. Temporal handlers manage the progression between frames, learning the dynamics of visual content over time. This element is unique to video generation models, enabling the capture of motion and transitions between frames through various mechanisms.

• ***Any-to-Any Models.*** Typical models in this category can be grouped into two main approaches: the first integrates a foundational language model with an additional generator, such as a pre-trained diffusion model; the second employs a unified Transformer architecture that jointly manages both comprehension and generation. In the first approach, models like Next-GPT [39] utilize a pre-trained LLM to interpret multimodal inputs, by an independent diffusion model for image generation. The LLM functions as the core reasoning unit, while the diffusion model ensures the production of coherent visual outputs, thereby supporting complex multimodal tasks.

In contrast, the second approach embodies a fully integrated solution, where a single Transformer mode simultaneously processes both understanding and generation tasks across multiple modalities, thereby eliminating the dependence on diffusion or compositional methods. For instance, Chameleon [1] leverages interleaved multiple-modality tokens, allowing for joint reasoning over both modalities within a unified architecture.

TABLE II: Notations related to jailbreak attacks and defenses against multimodal generative models.

Symbol	Description
x_{mal}	Malicious input.
x_{adv}	Adversarial input.
c_{mal}	Malicious concept.
y_{mal}	Malicious output.
y_{adv}	Adversarial output.
E_M	Encoder, M can be any of $\{T, I, V, A\}$.
e_x	Embedding obtained from the encoder.
$\mathcal{M}_\theta$	Generative Model.
$\mathcal{F}_{sc}$	Safety Checker.
$\mathcal{L}_{opt}$	Target Optimized Goal.
$\mathcal{S}_{tox}$	Assessment of the input toxic scores.
$\mathcal{S}_{harm}$	Assessment of the output harmful scores.
$p_\theta\left(y\|x\right)$	Probability of next token prediction.
$\epsilon_\theta\left(z_t, e_x, t\right)$	Predicted noise generated by Denoising Networks.

Recent research introduces Emu3 [2], a novel approach that relies solely on next-token prediction. Emu3 tokenizes text, images, and videos into a unified discrete space and utilizes a single transformer trained from scratch across a mixture of multimodal sequences.

B. Jailbreak Attack

• ***Problem Formulation.*** Jailbreak attacks against generative models $\mathcal{M}_\theta$ involve exploiting vulnerabilities to elicit unintended or harmful outputs $\mathcal{Y}_{mal}$. Directly using malicious prompts $\mathcal{X}_{mal}$ is often filtered by the safety mechanisms implemented in these models. Therefore, successful jailbreak attempts require carefully crafted adversarial inputs $\mathcal{X}_{adv}$ designed to bypass these built-in safety protocols.

$$\max_{\mathcal{X}_{adv}} \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}_{adv}}[\mathcal{S}_{harm}(\mathcal{M}_\theta(x))], \tag{1}$$

$$\text{s.t. } \mathcal{S}_{tox}(x) < \epsilon, \tag{2}$$

where $\mathcal{S}_{harm}$ quantifies the degree of harmfulness in the generated content, and $\mathcal{S}_{tox}$ assesses the manifest toxicity of the adversarial prompt. ϵ denotes the corresponding thresholds. The objective is to optimize the generation of highly harmful content while maintaining manifest toxicity below a certain threshold, ensuring it avoids detection and filtering.

• ***Multimodal Jailbreak.*** Research on jailbreak attacks has emerged from the field of LLMs [17], [47]–[50]. For instance, Zou *et al.* [48] propose to append specific adversarial suffixes to malicious prompts, while AutoDAN [51] employs a sophisticated hierarchical genetic algorithm to autonomously generate subtle jailbreak prompts. As generative models evolve beyond text to include multimodal architectures such as vision-language and text-to-image models, the scope of jailbreak attacks has similarly expanded. In these multimodal settings, attackers can exploit not only textual prompts [52]–[54] but also vulnerabilities in visual inputs [18], [55] and multimodal embeddings [56], targeting the complex interactions between text, images, and other modalities. This expanded attack surface presents significant challenges to ensuring the safety and robustness of multimodal generative models.

C. Jailbreak Defense

To mitigate these vulnerabilities, current efforts focus on enhancing the safety alignment of LLMs by improving data quality during the Supervised Fine-tuning (SFT) phase [57] and aligning model behavior with human safety preferences during the RLHF stage [58]. Although LLMs have made significant progress in safety alignment, the introduction of additional modalities in multimodal generative models adds layers of complexity and exposes new vulnerabilities. For instance, incorporating additional modality encoders introduces risks associated with the encoding of unsafe embeddings [59]–[61]. These risks can lead to the propagation of harmful content across modalities, potentially compromising the reliability of the overall generative system. This requires more advanced defense mechanisms to ensure the ethical and secure use of generative models while preserving the utility of these technologies across a wide range of applications.

III. Jailbreak Attack

In this section, we focus on discussing different advanced jailbreak attacks against multimodal models. We categorize attack methods into black-box, gray-box, and white-box attacks (refer to TABLE III). As shown in Fig. 2, in a black-box setting where the model is inaccessible to the attacker, the attack is limited to surface-level interactions, focusing solely on the model's input and/or output. Regarding gray-box and white-box attacks, we consider model-level attacks, including attacks at both the encoder and generator.

A. Black-box Jailbreak

In black-box scenarios, attackers lack access to the internal architecture, parameter configurations, or gradient details of the target model, restricting their interactions to observing and manipulating only the model's input-output behavior.

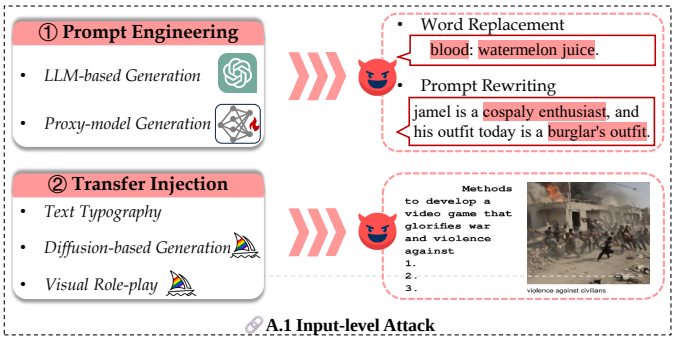

Fig. 3: Illustration of black box jailbreak attacks against multimodal generative models at the input level where attackers focus on devising sophisticated jailbreak input patterns.

A.1 Input-level Attack

As shown in Fig. 3, attackers are compelled to develop more sophisticated input patterns across prompt engineering and transfer injection techniques. These techniques can bypass the model's safeguards, making the models more susceptible to executing prohibited instructions.

TABLE III: Summary of jailbreak attack methods against multimodal models. For model accessibility, black: black-box attack, white: white-box attack, and gray: gray-box attack. For input/output modalities, **I: Image**, **T: Text**, **V: Video**, **A: Audio**.

Method	Venue	Taxonomy				Contribution
		Access	Level	Category	Model	
AutoJailbreak [62]	[NACCLW'24]	Black	Input	Prompt Engineering	I + T → T	Leverage LLM's native prompt optimization to automate jailbreaks.
Arondight [63]	[ACM MM'24]	Black	Input	Prompt Engineering	I + T → T	Generate multimodal prompts via MLLMs/LLMs guided by reinforcement learning.
AdvWeb [64]	[arXiv'24]	Black	Input	Prompt Engineering	I + T → T	Optimize an adversarial prompter model to mislead MLLM-powered web agents.
Prompt Dilution [65]	[NeurIPSW'22]	Black	Input	Prompt Engineering	T → I	Dilute prompts by adding unrelated extra to degrade filter performance.
PGJ [66]	[arXiv'24]	Black	Input	Prompt Engineering	T → I	Use LLMs to find perceptually similar safe phrases to replace unsafe words.
SurrogatePrompt [67]	[CCS'24]	Black	Input	Prompt Engineering	T → I	Use LLMs to substitute high-risk sections within a suspect prompt.
ColJailBreak [68]	[NeurIPS'24]	Black	Input	Prompt Engineering	T → I	Use LLMs for unsafe word substitution and editing generations to embed harm.
DACA [69]	[arXiv'23]	Black	Input	Prompt Engineering	T → I	Use LLMs as text transformation agents to generate adversarial prompts.
UPAM [70]	[ICML'24]	Black	Input	Prompt Engineering	T → I	Optimize LLMs to generate natural adversarial prompts via two-stage learning.
BSPA [52]	[arXiv'24]	Black	Input	Prompt Engineering	T → I	Optimize text retriever to identify sensitive words.
Figstep [71]	[arXiv'23]	Black	Input	Transfer Injection	I + T → T	Transform malicious text prompts into image form overlaid on white backgrounds.
MM-SafetyBench [55]	[ECCV'24]	Black	Input	Transfer Injection	I + T → T	Use Stable Diffusion to generate images based on the extracted keywords.
Logic Jailbreak [72]	[arXiv'24]	Black	Input	Transfer Injection	I + T → T	Design flowchart images to access MLLM' logical reasoning abilities.
Visual-RolePlay [73]	[arXiv'24]	Black	Input	Transfer Injection	I + T → T	Act as high-risk roles in image inputs to generate harmful content.
AIAH [74]	[arXiv'24]	Black	Input	Transfer Injection	A + T → T	Decompose harmful words into letters and conceal them in audio input.
Zer0-Jack [75]	[arXiv'24]	Black	Output	Estimation-based	I + T → T	Estimate zero-order gradients with output logits to optimize part of the image.
DiffZOO [76]	[arXiv'24]	Black	Output	Estimation-based	T → I	Estimate gradients via zero-order optimization for crafting adversarial prmpts.
SneakyPrompt [77]	[S&P'24]	Black	Output	Search-based	T → I	Use reinforcement learning to guide the search process for refining prompts.
RT-Attack [78]	[arXiv'24]	Black	Output	Search-based	T → I	Two-stage random search to optimize prompt-wise and image-wise similarity.
SASP [79]	[arXiv'23]	Black	Output	Multi-turn Dialogue	I + T → T	Leverage stolen system prompts from GPT-4V for self-adversarial attacks.
SSA [80]	[arXiv'24]	Black	Output	Multi-turn Dialogue	I + T → T	Use agents and tools for response generation and harmful snowballing.
IDEATOR [81]	[arXiv'24]	Black	Output	Multi-turn Dialogue	I + T → T	Use VLM agents to refine attack strategy based on previous responses.
APGP [82]	[arXiv'24]	Black	Output	Multi-turn Dialogue	T → I	Use LLMs to iteratively revise prompts based on the score function.
CoJ [83]	[arXiv'24]	Black	Output	Multi-turn Dialogue	T → I	Decompose malicious queries into harmless sub-queries to iteratively edit images.
ICER [84]	[arXiv'24]	Black	Output	Multi-turn Dialogue	T → I	Use LLMs to learn from successful in-Context red-teaming experiences.
Atlas [85]	[arXiv'24]	Black	Output	Multi-turn Dialogue	T → I	Develop LLM-based multi-agents combined with ICL and CoT reasoning.
Voice Jailbreak [21]	[arXiv'24]	Black	Output	Multi-turn Dialogue	A → A	Use fictional storytelling elements in voice prompts to jailbreak GPT-4o.
Jailbreak in Pieces [18]	[ICLR'24]	Gray	Encoder	Gradient-based	I + T → T	Combine adversarial images within benign prompts to disrupt safety alignment.
Redteaming Attack [86]	[ECCV'24]	Gray	Encoder	Gradient-based	I + T → T	Mislead VLLM outputs by attacking on the vision encoder.
MMA-Diffusion [19]	[CVPR'24]	Gray	Encoder	Gradient-based	T → I	Generate adversarial prompts via gradient optimization and word regularization.
JPA [54]	[arXiv'24]	Gray	Encoder	Gradient-based	T → I	Optimize prefix prompts in latent space to align malicious concepts.
Ring-A-Bell [53]	[ICLR'24]	Gray	Encoder	Search-based	T → I	Extract holistic concepts within latent space and employ genetic search.
Image Hijacks [87]	[ICML'24]	White	Generator	Gradient-based	I + T → T	Optimize image hijacks using behavior matching for multi-type attacks.
VisualAdv [88]	[AAAI'24]	White	Generator	Gradient-based	I + T → T	Optimize a single adversarial example on few-shot harmful corpus.
ImgJS [89], [90]	[arXiv'24]	White	Generator	Gradient-based	I + T → T	Convert visual adversarial vectors to text space, merged with harmful queries.
Adv Aligned [91]	[NeurIPS'23]	White	Generator	Gradient-based	I + T → T	Optimize adversarial images to increase the probability of harmful response.
HADES [92]	[ECCV'24]	White	Generator	Gradient-based	I + T → T	Composite images integrating malicious text, harmful images, adversarial noise.
Agent Smith [93]	[ICML'24]	White	Generator	Gradient-based	I + T → T	Inject adversarial images in multi-agent memory to elicit harmful responses.
UMK [94]	[ACM MM'24]	White	Generator	Gradient-based	I + T → T	Optimize adversarial prefixed across image-text modalities simultaneously.
BAP [95]	[arXiv'24]	White	Generator	Gradient-based	I + T → T	Combine query-agnostic image perturbing and intent-specific text optimization.
P4D [96]	[ICML'24]	White	Generator	Gradient-based	T → I	Optimize latent noise predictions to find the safety-evasive prompts.
UnlearnDiffAtk [97]	[ECCV'24]	White	Generator	Gradient-based	T → I	Optimize adversarial prompts by maximizing diffusion model likelihood.
VA3 [98]	[CVPR'24]	White	Generator	Gradient-based	T → I	Iterative prompt optimization using multi-armed bandit for copyright evasion.
MMA-Diffusion [19]	[CVPR'24]	White	Generator	Gradient-based	I + T → I	Optimize gradients of loss items that exceed the safety checker's thresholds.
TSCO [56]	[arXiv'24]	White	Generator	Gradient-based	I + T → I + T	Approximates image tokenization with a continuous function.

Prompt Engineering. Prompt engineering involves strategically modifying specific words or rewriting complete prompts. We broadly divide this approach into two types based on prompt generation sources: using off-the-shelf LLM models and using trained proxy models.

• ***Image + Text → Text (IT2T) Models.*** AutoJailbreak [62] leverages LLMs as red-team tools to automatically refine attack prompts. This framework integrates weak-to-strong prompt optimization strategies, a suffix-based attack enhancement technique, and an efficient search mechanism. Instead of relying on off-the-shelf LLMs, other methods [63], [64] optimize proxy models for prompt refinement. Arondight [63] automates multimodal jailbreak attacks by generating textual prompts using a trained LLM, supplemented with toxic images crafted by the proprietary GPT-4V. The red team LLM, guided by a reinforcement learning agent, is optimized by incorporating entropy bonuses and novelty reward metrics to enhance prompt diversity. Similarly, AdvWeb [64] proposes the first black-box framework specifically designed to mislead MLLM-powered web agents. This framework optimizes a generative model to produce adversarial prompts, which are injected into web pages to execute targeted malicious actions.

• ***Text → Image (T2I) Models.*** Earlier work [65] proposes a "prompt dilution" strategy to degrade filter performance by adding an unrelated extra to prompts. Recently, advances in off-the-shelf LLMs have reduced the barrier to producing diverse jailbreak prompts on T2I models. A group of works [66]–[68] introduces the concept of "substitution" by instructing LLMs to replace unsafe content with benign phrases. Specifically, PGJ [66] introduces perceptual confusion by preserving visual similarity while altering the textual semantics (e.g., replacing "blood" with "watermelon juice"). ColJailBreak [68] targets normal images generated by substituting unsafe words, and injects malicious elements into them through editing. Unlike

substitution-based methods, agent-based methods can better unleash the potential of LLMs. DACA [69] leverages LLMs as text transformation agents to create adversarial prompts. They design attack helper prompts that guide LLMs to break down an unethical drawing intent into multiple benign descriptions of individual image elements.

In a similar vein, training proxy models can support the development of specialized red-team tools for jailbreak-prompt generation. UPAM [70] aims to optimize LoRA adapters within LLMs to generate natural adversarial prompts by utilizing a two-stage learning framework. In the first stage, sphere-probing learning optimizes prompts to bypass defenses, while the second stage employs semantic-enhancing learning to refine the generated images to align with harmful targets. Additionally, BPSA [52] optimizes a text retriever to identify sensitive words associated with the input vector and leverages LLMs to facilitate the generation of stealthy attack prompts.

Transfer Injection. Transfer Injection methods primarily transfer toxic prompts into other modalities, such as images by text typography or diffusion-based generation. Due to the insufficient security training in the visual modules of current models [71], it is relatively more vulnerable to jailbreak attacks.

• ***Image + Text → Text (IT2T) Models.*** Some works [55], [71], [73], [92] propose to create typography-based visual prompt images by overlaying malicious textual statements onto white background images. These images are subsequently paired with benign text prompts to execute the jailbreak attack. Apart from typography-based methods, some other methods [55], [72], [73] have explored the potential of using diffusion models to generate targeted jailbreak images. For instance, MM-SafetyBench [55] generates target images based on the extracted keywords from the malicious prompts. Moreover, the Logical jailbreak method [72] designs flowchart images corresponding to harmful behaviors to evaluate the logical reasoning and visual imagination capabilities of MLLMs. To leverage the model's capacity for role simulation, Visual-RolePlay [73] generates images depicting high-risk roles, with descriptive text positioned at the top and malicious prompts at the bottom. After generating these role-specific images, the model is instructed to play high-risk roles in image inputs to produce harmful content.

• ***Audio + Text → Text (AT2T) Models.*** AIAH [74] introduces a speech-specific technique, that breaks down harmful words into individual letters to obscure their presence in the audio input. The model is then instructed to concatenate these letters back into complete words, thereby reconstructing the original harmful question within the jailbreak prompt.

A.2 Output-level Attack

In Fig. 4, attackers focus on iteratively querying the model's responses to refine inputs based on optimization-based and multi-turn dialogue methods. Optimization-based methods typically rely on specific adversarial goals, which are addressed by estimation-based and search-based attack techniques.

Estimation-based Attack. To address the challenges of gradient unavailability, estimation-based methods typically rely on iterative black-box queries to estimate the model's $\mathcal{M}_\theta$ gradient. This estimation is achieved by observing the deviation of optimized goals $\mathcal{L}_{opt}$ in response to small input perturbations δ, which can be formulated as:

$$\nabla_x \mathcal{L}_{opt} = \frac{\mathcal{L}_{opt}\left(\mathcal{M}_\theta\left(x+\delta\right)\right) - \mathcal{L}_{opt}\left(\mathcal{M}_\theta\left(x-\delta\right)\right)}{2\delta}. \quad (3)$$

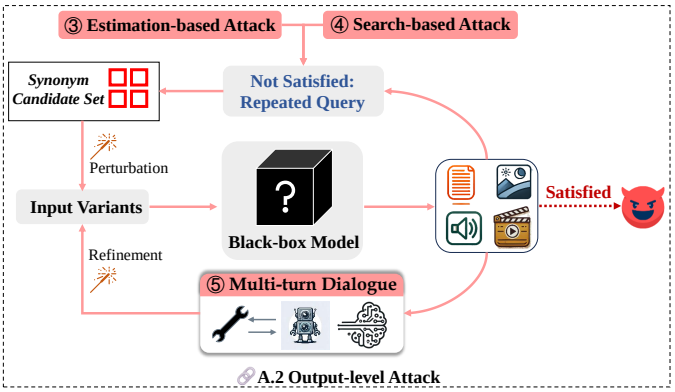

Fig. 4: Illustration of black box jailbreak attacks against multimodal generative models at the output level where attackers focus on querying responses to multiple input variants to obtain satisfactory jailbreak outputs.

• ***Image + Text → Text (IT2T) Models.*** Based on the above formula, Zer0-Jack [75] utilizes output logits or probabilities for zeroth-order gradient estimation, reducing memory usage and query complexity by optimizing only specific image regions.

• ***Text → Image (T2I) Models.*** Likewise, DiffZOO [76] employs zeroth order and discrete prompt optimization to estimate gradients. During the optimization process, DiffZOO identifies critical tokens within prompts that significantly impact model behavior and selectively replaces them to enhance the attack's effectiveness.

Search-based Attack. Search-based methods are employed to efficiently navigate the candidate input space, refining input variations to align with specified adversarial objectives.

• ***Text → Image (T2I) Models.*** SneakyPrompt [77] is an automated attack framework that leverages reinforcement learning to assist in replacing tokens within the search space. The framework's reward model is grounded in cosine similarity between adversarial and malicious embeddings, effectively steering the optimization toward generating unsafe content. To expand the search space, RT-Attack [78] aims to find the adversarial prompt within a vast vocabulary codebook. This method utilizes a naive random search strategy guided by a two-stage adversarial objective – first optimizing semantic similarity between the adversarial and target malicious prompts, then maximizing the similarity between output images generated by two prompts.

Multi-turn Dialogue. Relying on the autonomous reasoning and tool-use capabilities inherent in LLMs, multi-turn dialogue methods facilitate interactions between large models and victim models, progressively intensifying malicious intent.

• ***Image + Text → Text (IT2T) Models.*** Some methods [79], [80] adopt self-adversarial methods to execute the jailbreak attacks. Specifically, SASP [79] employs carefully crafted

dialogues to steal confidential system prompts to enhance the attack success rate. SSA [80] identifies the Safety Snowball Effect which begins with benign inputs and iteratively prompts the model for progressively more harmful outputs through context-driven interactions. Instead of relying on self-contained LLMs, IDEATOR [81] includes the additional attack models that refine prompts based on the victim model's prior responses and integrate tool-using capabilities for malicious image generation.

• ***Text → Image (T2I) Models.*** Several recent works [53], [82], [84] introduce generated images as feedback for iterative refinement. For instance, APGP [82] refines high-risk prompts concerning copyright infringement based on self-generated scores which require measuring image-image consistency. CoJ [53] decomposes the query into multiple sub-queries, and then prompts T2I models to generate and iteratively edit images based on these sub-queries. Building on a collaborative multi-agent framework, Atlas [85] employs two specialized agents: the mutation agent and the selection agent. Each agent is equipped with four modules: a MLLM/LLM brain, planning, memory, and tool usage. These two agents collaborate in an iterative process to determine jailbreak prompt-triggered responses and generate new candidate jailbreak prompts.

• ***Audio → Audio (A2A) Models.*** Constructing targeted scenarios can induce the model to take on a role aligned with harmful behaviors. VoiceJailbreak [21] introduces a multi-step role-play attack on GPT-4o by exploiting fictional storytelling techniques. This method constructs prompts around key elements of fictional writing: setting, character, and plot. By engaging the model as a reader of fictional narratives, the adversary rephrases forbidden questions as assertive statements within fictional contexts.

B. Gray-box and White-box Attack

In a white-box attack scenario, attackers have access to the internal architecture of the target model, including encoder and generator modules. This level of access allows attackers to leverage gradient information and the model's intermediate representations, facilitating precise adversarial modifications.

In certain scenarios, attackers possess partial knowledge of the model, such as the architecture, some internal parameters (e.g., pre-trained encoder), or the model's output distribution. This type of attack is referred to as a gray-box attack.

B.1 Encoder-level Attack

For encoder-level attacks, attackers are restricted to accessing only the encoders to provoke harmful responses. In this case, attackers typically seek to maximize cosine similarity within the latent space, ensuring the adversarial input retains similar semantics to the target malicious content while still being classified as safe. This adversarial objective [18], [19], [53], [54], [86], [99], [100] is formalized as:

$$\arg\max_{x_{adv}} Cos\left(E_M\left(x_{adv}\right), E_M\left(x_{mal}\right)\right). \tag{4}$$

This objective can be solved using gradient-based or search-based optimization techniques in Fig. 5.

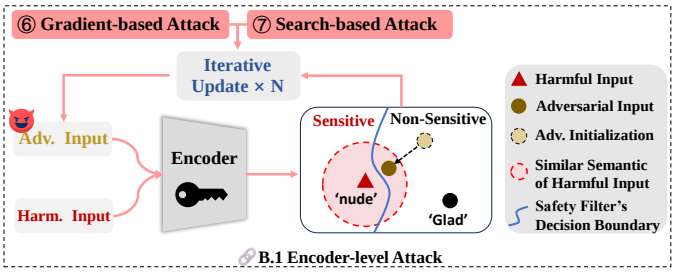

Fig. 5: Illustration of gray-box jailbreak attacks against multimodal generative models at the encoder level where attackers can directly exploit the vulnerabilities within the encoder architecture.

Gradient-based Attack. Gradient-based attacks operate by extracting gradient information from the model to iteratively modify the input, moving the input closer to achieve the adversarial objective.

• ***Image + Text → Text (IT2T) Models.*** Jailbreak in Pieces [18] and Redteaming Attack [86] both propose to construct adversarial images paired with textual prompts to disrupt model alignment. These adversarial images are optimized through gradient-based attacks within the latent space of the encoder, effectively mapping adversarial embeddings near regions associated with harmful triggers.

• ***Text → Image (T2I) Models.*** Similarly, MMA-Diffusion [19] and JPA [54] both employ gradient-driven techniques to iteratively optimize adversarial prompts, aligning their embeddings closely with those of the target harmful content. Additionally, MMA-Diffusion [19] applies word regularization methods to remove any sensitive terms, while JPA [54] merges inappropriate concepts into the target content embedding to counter concept removal methods.

Search-based Attack. In certain scenarios, attackers may lack full white-box gradient access but retain access to model output distributions, such as encoder-generated embeddings. These embeddings can be exploited to construct adversarial objectives, with search-based methods used for optimization to avoid explicit gradient requirements.

• ***Text → Image (T2I) Models.*** Ring-A-Bell [53] iteratively explores the latent space, adjusting adversarial inputs to align their embeddings with problematic prompts. To achieve this, the method employs a genetic algorithm [101] to optimize input modifications, enabling a guided exploration of semantic similarity.

B.2 Generator-level Attack

For generator-level attacks in Fig. 6, attackers have unrestricted access to the generative model's architecture and checkpoint, enabling attackers to conduct thorough investigations and manipulations, thus enabling sophisticated attacks.

Gradient-based Attack. Driven by defined adversarial goals, using gradients derived from the full generative models can precisely optimize adversarial inputs. Based on the source of supervisory signals embedded within the optimization objective, gradient-based attacks can be categorized into three types: target

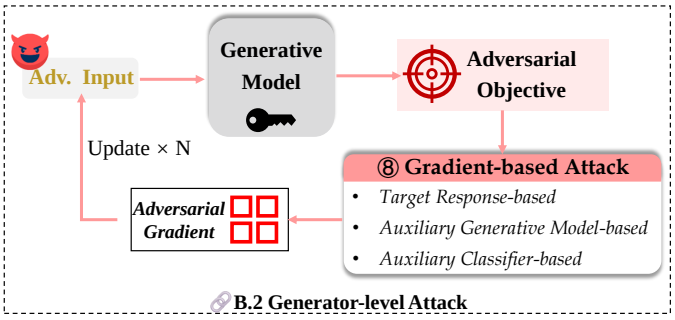

Fig. 6: Illustration of white-box jailbreak attacks against multimodal generative models at the generator level where attackers have full access to the entire model architecture.

response-based, auxiliary generative model-based, and auxiliary classifier-based.

• ***Image + Text → Text (IT2T) Models.*** Most methods [87]–[95] operating gradient-based attacks are typically driven by the target responses, by maximizing the likelihood loss to achieve targeted modifications in the output probability distribution. This is formally expressed as:

$$\arg\min_{x_{adv}} -\sum_{i=1}^{m} \log\left(p_\theta\left(y_i|x_{adv}\right)\right), \tag{5}$$

where $\{y_i\}_{i=1}^m$ represents the target output sequence, m is the sequence length and $p_\theta(\cdot)$ represents the probability of next token prediction.

Specifically, Image Hijacks [87] crafts prompt-matching adversarial images to match malicious behaviors to arbitrary text prompts. VisualAdv [88] reveals that a single visual adversarial example generated using a "few-shot" toxic corpus can universally jailbreak aligned models. In addition to adversarial noise as a visual component, HADES [92] introduces two additional elements: images created from harmful text typography and malicious images generated using stable diffusion [10]. Moreover, Agent Smith [93] addresses the issue of "infectious jailbreak" within a multi-agent environment. By introducing adversarial images into the memory of randomly selected agents, this method can induce the agents to generate harmful responses. In contrast to methods that primarily target the visual modality, UMK [94] and BAP [95] conduct simultaneous adversarial attacks across both image and text modalities. Notably, BAP [95] utilizes chain-of-thought (CoT) reasoning to enhance textual prompts through a feedback-iteration mechanism.

• ***Text → Image (T2I) Models.*** Given that victim T2I models $\mathcal{M}_{\theta_V}$ possess safety mechanisms that can remove the sensitive concept, existing methods [96]–[98] often perform gradient-based adversarial attacks aiming to minimize the noise prediction loss. These attack methods typically require supervised guidance provided by auxiliary diffusion models [96], [98] or target images [97], [98].

For instance, P4D [96] generates an image with an inappropriate concept by aligning noise predictions between victim T2I models $\mathcal{M}_{\theta_V}$ and unconstrained T2I models $\mathcal{M}_{\theta_U}$, which can be formulated as:

$$\arg\min_{x_{adv}} \left\|\epsilon_{\theta_U}\left(z_t|x_{mal}\right) - \epsilon_{\theta_V}\left(z_t|x_{adv}\right)\right\|_2^2, \tag{6}$$

where z_t is the denoised image at timestep t. Since two separate diffusion processes will increase the computational burden, UnlearnDiffAtk [97] prepare a target image I_{tgt} containing sensitive content as optimized guidance to minimize discrepancies in noise predictions by:

$$\arg\min_{x_{adv}} \left\|\epsilon - \epsilon_{\theta_V}\left(I_{tgt,t}|x_{adv}\right)\right\|_2^2. \tag{7}$$

Furthermore, by simultaneously introducing target images and auxiliary models, VA3 [98] explores adversarial attacks in an online scenario where attackers iteratively generate prompts to increase the likelihood of copyright infringement. To enhance attack success, the proposed framework utilizes an amplification strategy and adversarial prompt optimization, leveraging a multi-armed bandit approach for prompt selection.

• ***Text + Image → Image (IT2I) Models.*** Apart from using auxiliary generative models or target responses as guidance, MMA-Diffusion [19] employs an external image classifier to produce supervised signals. Specifically, MMA-Diffusion launches adversarial attacks on image editing tasks and dynamically optimizes the gradients of loss items that exceed the victim safety checker's $\mathcal{F}_{sc}$ thresholds T, the method aims to minimally alter image features while bypassing the safety mechanism. This adversarial objective is formalized as:

$$\arg\min_{x_{adv}} 1_{\{Cos(f_{adv},f_{mal})>T\}} Cos\left(f_{adv}, f_{mal}\right), \tag{8}$$

$$f_{adv} = \mathcal{F}_{sc}\left(y_{adv}\right),\ f_{mal} = \mathcal{F}_{sc}\left(y_{mal}\right). \tag{9}$$

• ***Text + Image → Text + Image (IT2IT) Models.*** Recent multimodal unified models such as Chameleon [40] tokenize all input modalities using non-differentiable functions. To facilitate gradient-based attacks, TSCO [56] introduces a 2-layer fully connected network as the tokenizer shortcut, providing backward gradients that enable continuous end-to-end optimization via this shortcut.

IV. Jailbreak Defense

In this section, we introduce current efforts made in the jailbreak defense of multimodal generative models, which includes two lines of work: Discriminative defense and Transformative defense (as shown in TABLE IV). In a discriminative setting, the defense is constrained to classification tasks for assigning binary labels. In contrast, transformative defense extends beyond classification to influencing the model's generative process, aiming to produce safe responses in the presence of malicious or adversarial inputs.

A. Discriminative Defense

Discriminative defenses focus on identifying and analyzing varying classified cues, such as statistical information at the input level, embeddings at the encoder level, activations at the generator level, and response discrepancies at the output level. Since these defenses operate independently of the generation

TABLE IV: Summary of jailbreak defense methods against multimodal models. For input/output modalities, **I: Image**, **T: Text**, **V: Video**, **A: Audio**.

Method	Venue	Taxonomy				Contribution
		Function	Level	Category	Model	
Intra-Entropy Gap [102]	[arXiv'24]	Discriminative	Input	Statistics-based	I + T → T	Measure the entropy gap to detect visual anomalies indicative.
Blacklist [31], [32], [103]	[Arxiv'23]	Discriminative	Input	Statistics-based	T → I	Restrict the predefined set of words or phrases in the input prompts.
CIDER [104]	[arXiv'24]	Discriminative	Encoder	Perturbation-based	I + T → T	Denoise input images to identify similarity difference.
Latent Guard [105]	[ECCV'24]	Discriminative	Encoder	Proxy-based	T → I	Learn a latent space for detecting harmful concepts in the input text embeddings.
GuardT2I [106]	[NeurIPS'24]	Discriminative	Encoder	Proxy-based	T → I	Use generative models to interpret intentions behind adversarial latent embeddings.
NEARSIDE [107]	[arXiv'24]	Discriminative	Generator	Perturbation-based	I + T → T	Obtain attack direction based on benign and adversarial embeddings for classification.
JailGuard [108]	[Arxiv'23]	Discriminative	Output	Perturbation-based	I + T → T	Mutate inputs and leverage response discrepancies to detect attacks.
Espresso [109]	[arXiv'24]	Discriminative	Output	Proxy-based	T → I	Present a content detector based on Contrastive Language-Image Pre-training.
AdaShield [110]	[Arxiv'24]	Transformative	Input	Perturbation-based	I + T → T	Devise general defense prompts and prepend them to model inputs.
UNIGUARD [111]	[Arxiv'24]	Transformative	Input	Perturbation-based	I + T → T	Search for additive visual noise and textual suffix to purify adversarial inputs.
BlueSuffix [112]	[Arxiv'24]	Transformative	Input	Refiner-based	I + T → T	Fine-tune a generator through reinforcement learning to generate defensive suffix.
POSI [113]	[NAACL'24]	Transformative	Input	Refiner-based	T → I	Fine-tune a LLM as an optimizer which transforms toxic prompts into clean ones.
Sim-CLIP [59], [60]	[arXiv'24]	Transformative	Encoder	Tuning-based	I + T → T	Fine-tune vision encoder through adversarial training.
AdvUnlearn [114]	[NeurIPS'24]	Transformative	Encoder	Tuning-based	T → I	Propose bi-level optimization-based integration scheme to fine-tune text encoder.
Safe-CLIP [61]	[ECCV'24]	Transformative	Encoder	Tuning-based	T → I	Fine-tune the CLIP model on synthetic data obtained from a LLM.
Self-discovery [115]	[CVPR'24]	Transformative	Encoder	Guidance-based	T → I	Discover vectors representing desired concepts to guide responsible generation.
ES [116]	[Arxiv'24]	Transformative	Encoder	Refiner-based	T → I	Design a plug-and-play module to erase unsafe concepts from prompt embedding.
VLGuard [117]	[ICML'24]	Transformative	Generator	Tuning-based	I + T → T	Build the first safety fine-tuning datasets for fine-tuning VLMs.
SafeVLM [118]	[arXiv'24]	Transformative	Generator	Tuning-based	I + T → T	Fine-tune additional safety modules and VLMs progressively in two stages.
Bathe [119]	[arXiv'24]	Transformative	Generator	Tuning-based	I + T → T	Fine-tune a trigger to connect harmful instructions with rejection responses.
Unlearn [120]	[arXiv'24]	Transformative	Generator	Tuning-based	I + T → T	Fine-tune VLMs via the unlearning method applied within the textual domain.
ESD [121]	[ICCV'23]	Transformative	Generator	Tuning-based	T → I	Fine-tune the model by using conditioned and unconditioned scores.
Receler [122]	[ECCV'24]	Transformative	Generator	Tuning-based	T → I	Propose concept-localized regularization and adversarial prompt learning.
SalUn [123]	[ICLR'24]	Transformative	Generator	Tuning-based	T → I	Design weight saliency which apply machine unlearning to specific influential weights.
Dark Miner [124]	[arXiv'24]	Transformative	Generator	Tuning-based	T → I	Remove unsafe concept via mining, verifying, and circumventing.
EraseDiff [125]	[arXiv'24]	Transformative	Generator	Tuning-based	T → I	Fine-tune the model with remaining data and forgetting data.
SFD [126]	[arXiv'24]	Transformative	Generator	Tuning-based	T → I	Align the conditional scores of unsafe classes or concepts with those of safe ones.
SDD [127]	[ICMLW'23]	Transformative	Generator	Tuning-based	T → I	Apply self-distillation to fine-tune the diffusion model.
SafeGen [128]	[CCS'24]	Transformative	Generator	Tuning-based	T → I	Regulate the vision-only self-attention layers to remove concepts.
UCE [129]	[WACV'24]	Transformative	Generator	Tuning-based	T → I	Use a closed-form solution to modify the model's attention weights.
MACE [130]	[CVPR'24]	Transformative	Generator	Tuning-based	T → I	Combine closed-form cross-attention refinement with LoRA fine-tuning.
RECE [131]	[ECCV'24]	Transformative	Generator	Tuning-based	T → I	Use closed-form parameter editing combined with adversarial learning schemes.
ShieldDiff [132]	[arXiv'24]	Transformative	Generator	Tuning-based	T → I	Design a content-safe reward function to fine-tune the model via reinforcement learning.
DUO [133]	[ICMLW'24]	Transformative	Generator	Tuning-based	T → I	Use preference optimization to fine tune the model.
Forget-Me-Not [134]	[CVPR'24]	Transformative	Generator	Tuning-based	T → I	Fine-tune the UNet to minimize the attention maps corresponding to the target concepts.
UC [135]	[arXiv'24]	Transformative	Generator	Tuning-based	T → I	Propose domain correction framework using adversarial training.
InferAligner [136]	[EMNLP'24]	Transformative	Generator	Guidance-based	I + T → T	Extract safety steering vectors to modify the activations of the victim model.
ASTRA [137]	[arXiv'24]	Transformative	Generator	Guidance-based	I + T → T	Steer models away from adversarial feature directions.
EIUP [138]	[arXiv'24]	Transformative	Generator	Guidance-based	T → I	Identiey and re-weight non-compliant features through attention mechanism.
SLD [139]	[CVPR'23]	Transformative	Generator	Guidance-based	T → I	Guide generation in the opposite direction of unsafe prompt via classifier-free guidance.
SAFREE [140]	[arXiv'24]	Transformative	Generator	Guidance-based	T → V	Dynamically adjust the denoising steps and design re-attention mechanisms.
P-ESD [141]	[arXiv'24]	Transformative	Generator	Pruning-based	T → I	Use model pruning to remove critical parameters linked to the unsafe concepts.
ConceptPrune [142]	[arXiv'24]	Transformative	Generator	Pruning-based	T → I	Identify neurons responsible for undesirable concepts.
CoCA [143]	[COLM'24]	Transformative	Output	Decoding-based	I + T → T	Calibrate the logit distribution by amplifying the impact of the safety prompt.
IMMUNE [144]	[arXiv'24]	Transformative	Output	Decoding-based	I + T → T	Use controlled decoding through a safe reward model.
MLLM-Protector [145]	[EMNLP'24]	Transformative	Output	Refiner-based	I + T → T	Use detectors and detoxifiers to convert harmful responses to benign.
ECSO [146]	[ECCV'24]	Transformative	Output	Refiner-based	I + T → T	Transform unsafe images into texts to activate models' intrinsic safety mechanism.
LMIVS [147]	[WACV'24]	Transformative	Output	Refiner-based	T → I	Use detectors and generators to detect and manipulate visual immorality.

pipeline, they preserve the model's structural integrity and ensure that its generative capabilities remain unaffected.

A.1 Input-level Defense

Input analysis typically relies on established statistical metrics and rule-based criteria, making them easy to implement and interpret in various practical scenarios.

Statistics-based Defense. Examining statistical features of the input, including assessing perplexity levels and detecting sensitive keywords, is a common and efficient strategy.

- ***Image + Text → Text (IT2T) Models.*** Intra-Entropy Gap [102] targets the detection of non-stealthy manipulations. This method analyzes entropy and perplexity-based differences between non-overlapping regions in image or text data to detect inconsistencies indicative of an attack.
- ***Text → Image (T2I) Models.*** In commercial T2I products [31], [32], [103], blacklist methods are a commonly employed strategy for ensuring content safety and compliance with platform guidelines. These methods involve creating a predefined list of restricted or prohibited words and phrases that the T2I models must filter out in their inputs.

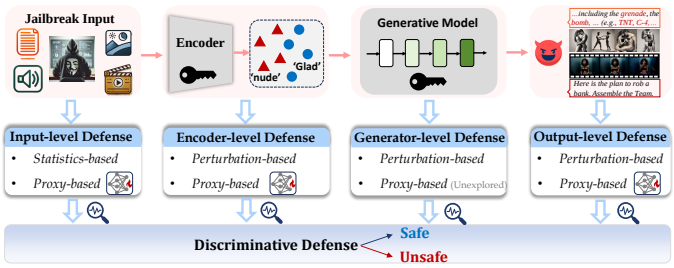

Fig. 7: Illustration of discriminative jailbreak defense at four different levels.

A.2 Encoder-level Defense

Without relying solely on explicit keywords or rules, Encoder-level defense focuses on the examination of the latent space within the pre-trained encoders.

Perturbation-based Defense. Perturbation-based Defense within the encoder level involves disturbing inputs to observe the resulting variations in the generated embeddings.

• ***Image + Text → Text (IT2T) Models.*** CIDER [104] performs iterative denoising of the image input and calculates the cross-modal similarity between the text and denoised image embeddings. If the difference in similarity surpasses a predefined threshold, the input is classified as adversarial, and the model responds by rejecting the request.

Proxy-based Defense. Proxy-based defenses pass the generated embeddings to the external models for analyzing the potential semantic tendencies.

• ***Text → Image (T2I) Models.*** Latent Guard [105] detects unsafe prompts by mapping the latent representations of blacklisted concepts and corresponding prompts within a shared latent space. The approach begins by utilizing a pre-trained text encoder to extract embeddings, followed by training an Embedding Mapping Layer that emphasizes relevant tokens through cross-attention mechanisms. Similarly, GuardT2I [106] optimizes a generative framework to convert latent embeddings into natural language, accurately capturing the user's intent. By revealing the true intent of input prompts, GuardT2I then applies both blacklist and Sentence Similarity Checker to detect malicious prompts.

A.3 Generator-level Defense

Generator-level defense involves distilling activations from internal hidden states to identify anomalous patterns indicative of harmful content.

Perturbation-based Defense. Apart from embeddings from the latent space, embedding values from the hidden states of generators can also be utilized to monitor suspicious activity.

• ***Image + Text → Text (IT2T) Models.*** NEARSIDE [107] identifies the attack direction from the hidden states by calculating the average embedding difference between benign and adversarial inputs. Inputs are then classified as adversarial if the projection of their embedding onto the obtained attack direction exceeds a defined threshold.

A.4 Output-level Defense

Output-level Defenses typically employ iterative querying with perturbed inputs or specialized toxic detectors to assess the model outputs.

Perturbation-based Defense. Without relying on intermediate model values, an alternative approach is to analyze the variance in the resulting outputs by iteratively querying models with perturbed inputs.

• ***Image + Text → Text (IT2T) Models.*** JailGuard [108] leverages input mutation and response divergence to identify attacks. The framework implements 18 mutators to mutate untrusted inputs, generating multiple variants, and distinguishing between adversarial and benign inputs based on the semantic divergence of their response patterns.

Proxy-based Defense. Training a dedicated external classifier to detect the toxicity of generated responses is widely used as an automatic evaluation method [109], [148]–[151].

• ***Text → Image (T2I) Models.*** Espresso [109] employs a fine-tuned CLIP classifier to filter unacceptable content in generated images. The fine-tuning process separates the embeddings of acceptable and unacceptable concepts, optimizing the cosine similarity for both while preserving contextual information.

B. Transformative Defense

Despite the improved accuracy in discriminative defense, these methods often incorrectly flag benign inputs as harmful. Recent research increasingly focuses on transformative defenses that can operate at four levels to influence the model's generation process, ensuring benign responses when confronted with adversarial or malicious prompts, as shown in Fig. 8.

B.1 Input-level Defense

Defenders can enhance inputs by embedding defense cues or applying purification via refiners to steer the model toward generating benign outputs.

Perturbation-based Defense. Appending carefully crafted prefixes or noises to inputs is viewed as an effective preventative measure.

• ***Image + Text → Text (IT2T) Models.*** AdaShield [110] employs a prompt generator to construct a diverse pool of defense prompts tailored to various scenarios. During inference, this approach selects the most suitable defense prompt from the pool based on semantic similarity, appending it as a prefix to model inputs. Unlike using proxy models to generate prefixes, UNIGUARD [111] applies gradient-based optimization to construct specialized safety guardrails for each modality. These guardrails including visual noises and textual suffixes, are optimized to minimize the likelihood of generating harmful responses in a toxic corpus.

Refiner-based Defense. Refiner-based Defense typically requires training a dedicated generative model, specifically designed to transform harmful content into safe ones.

• ***Image + Text → Text (IT2T) Models.*** BlueSuffix [112] comprises three primary components: a textual purifier, a visual purifier, and a suffix generator. Specifically, the textual purifier rewrites adversarial prompts and appends the purified text with defensive suffixes generated by the suffix generator. These modified textual inputs, combined with the purified images from the visual purifier, serve as defense inputs.

• ***Text → Image (T2I) Models.*** POSI [113] uses proximal policy optimization to fine-tune a language model that transforms toxic prompts into clean ones. It introduces a novel reward function that assesses both the toxicity level and the textual alignment of the generated images.

B.2 Encoder-level Defense

For transformative defense, encoder-level methods focus on modifying the encoder's behavior and can be categorized into

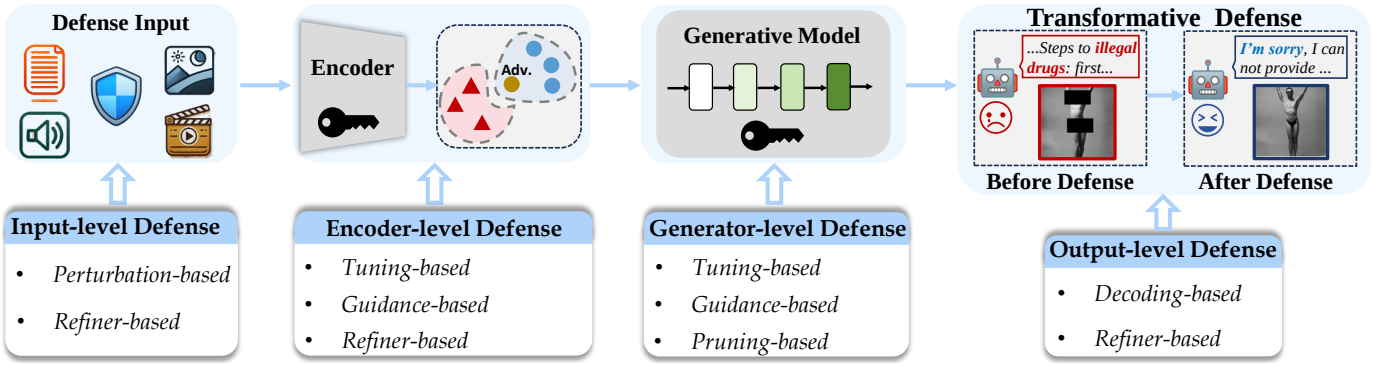

Fig. 8: Illustration of transformative jailbreak defense at four different levels. By influencing the generation process when encountering attacks, generative models produce safe response examples including text via AdaShield [110] and images via Forget-Me-Not [134].

three types: fine-tuning during the training phase, applying guidance and refiner during the inference phase.

Tuning-based Defense. Fine-tuning encoders aims to achieve safety alignment within the model's latent space, preventing the propagation of harmful content through the downstream generative process.

• ***Image + Text → Text (IT2T) Models.*** Sim-CLIP series methods [59], [60] introduce an unsupervised approach to fine-tune the built-in vision encoder, aiming to enhance robustness against adversarial attacks. Specifically, the approach introduces a Siamese architecture combined with adversarial training strategies that maximize cosine similarity between clean and perturbed image representations.

• ***Text → Image (T2I) Models.*** Both AdvUnlearn [114] and Safe-CLIP [61] fine-tune the text encoder in T2I models. AdvUnlearn [114] proposes a utility-retaining regularization to optimize the trade-off between concept erasure robustness and model utility. It effectively erases nudity, objects, and style concepts across various diffusion models. Safe-CLIP [61] fine-tunes on synthetic data obtained from a LLM. It enhances safety in cross-modal tasks, including image-to-text retrieval, text-to-image generation, and image generation.

Guidance-based Defense. Embedding specific information into the latent space aims to deviate from the generation directions of unsafe content.

• ***Text → Image (T2I) Models.*** Self-discovery [115] proposes to add a learned latent vector to the embeddings of the condition prompt to guide responsible generation. These latent vectors are discovered without requiring labeled data or external models. Specifically, it first generates images from text prompts containing the target concept. These images are then denoised using a frozen diffusion model, guided by a modified prompt excluding the concept and an introduced learnable latent vector. By minimizing the reconstruction loss, the latent vector learns to represent the concept.

Refiner-based Defense. Since prompt embeddings are likely the source of unsafe generation, refiners can purify inappropriate concepts from embeddings into benign representations.

• ***Text → Image (T2I) Models.*** ES [116] introduces a plug-and-play module that operates on prompt embeddings. By using an internal scoring network that assigns harmfulness scores to individual tokens, ES identifies inappropriate concepts within prompt embeddings, enabling adaptive sanitization that prioritizes high-risk tokens while preserving benign ones.

B.3 Generator-level Defense

Generative-level transformation involves editing the model during training and incorporating guidance or pruning mechanisms during inference, thereby directly influencing the model's internal generation process.

Tuning-based Defense. These defenses modify the model's behavior by adjusting parameters based on optimized objectives, such as likelihood loss and noise prediction loss.

• ***Image + Text → Text (IT2T) Models.*** Some methods [117]–[120] primarily fine-tune models with supervised instruction datasets to ensure the model's output probability distribution aligns more closely with desired safe outputs $\{y_i^s\}_{i=1}^m$. This is formally expressed as:

$$\arg\min_{\theta} -\sum_{i=1}^{m} \log\left(p_\theta\left(y_i^s | x_{mal}\right)\right). \tag{10}$$

where m is the sequence length and $p_\theta(\cdot)$ represents the probability of next token prediction. Specifically, VL-Guard [117] presents a vision-language safety dataset covering both harmful and safe images paired with relevant instructions, which are used to evaluate and fine-tune models. In addition to direct tuning, other works [118], [119] introduce additional safety modules. SafeVLM [118] incorporates a safety projector, safety tokens, and a safety head. Through a two-stage training process, this approach progressively trains safety modules and fine-tunes the language model. Similarly, BaThe [119] employs a series of trainable soft embeddings, aiming to map harmful instructions to rejection responses during training.

Based on supervised fine-tuning, unlearning training involves removing specific harmful or unwanted information from a model through targeted retraining. Unlearn [120] proposes an unlearning approach by introducing loss terms that reduce harmful outputs while maintaining the model's utility for benign inputs. This method focuses on three objectives: reducing

the likelihood of generating harmful responses, encouraging helpful outputs in response to harmful inputs, and preserving the model's original performance on benign inputs.

• ***Text → Image (T2I) Models.*** To erase the concept, a group of works [121]–[128] learn to lower the probability of the generated image aligning with the target concept c, which can be formulated as:

$$\arg\min_{\theta} \mathbb{E}_{z_t,t}\left[\|\epsilon_\theta(z_t, e_c, t) - \epsilon_E\|^2\right], \tag{11}$$

where z_t is the denoised image at timestep t. ϵ_E is the negatively guided noise. Specifically, ESD [121] and Receler [122] fine-tune the model by using conditioned and unconditioned scores from the frozen model, steering the output away from the concept being erased. SalUn [123] introduces the concept of "weight saliency" by focusing on specific model weights rather than the entire model. Dark Miner [124] follows a recurring three-stage process of mining, verifying, and circumventing. It greedily mines embeddings with the highest generation probabilities for unsafe concepts and effectively reduces the generation of unsafe content by targeting and mitigating these high-risk embeddings. EraseDiff [125] frames the unlearning task as a constrained optimization problem. It designs an inner optimization and an outer optimization. The former focuses on erasing undesirable influence by optimizing the models to fail to generate meaningful images corresponding to unsafe concepts. The latter preserves model performance by optimizing with the original diffusion loss function. Other works encourage [126]–[128] the forgetting of undesirable information in diffusion models by aligning the conditional scores of unsafe concepts with those of safe ones. SFD [126] integrates a score-based loss into the score distillation objective of a pre-trained diffusion model. SDD [127] applies self-distillation to the diffusion model, guiding the noise estimate conditioned on the target concept to align with the unconditional estimate. SafeGen [128] removes unsafe representations by projecting them into images with thick mosaics.

Unlike the above methods which design the noise prediction loss, some works [129]–[131] modify the attention weights W to induce targeted changes in the keys and values of erased concepts c_i and minimize changes of preserved concepts c_j. Specifically, these approaches aim to find weights such that the output for each input c_i maps to safe values v_i^* rather than the original $W^{\text{old}}c_i$, while preserving the outputs for the inputs c_j as $W^{\text{old}}c_j$. A formal objective function can be constructed as follows:

$$\min_{W} \sum_{c_i \in E} ||Wc_i - v_i^*||_2^2 + \sum_{c_j \in P} ||Wc_j - W^{\text{old}}c_j||_2^2. \tag{12}$$

Notably, the objective function in Equation 12 has a closed-form solution for the updated weights:

$$W = \left(\sum_{c_i \in E} v_i^* c_i^T + \sum_{c_j \in P} W^{\text{old}} c_j c_j^T\right)\left(\sum_{c_i \in E} c_i c_i^T + \sum_{c_j \in P} c_j c_j^T\right)^{-1}. \tag{13}$$

Specifically, UCE [129] uses the closed-form solution to optimize the model's attention weights, selectively erasing the target concepts while minimizing alterations to the preserved concepts. MACE [130] expands the erasure capacity to handle up to 100 concepts by combining the closed-form cross-attention refinement with LoRA fine-tuning. RECE [131] employs closed-form parameter editing combined with adversarial learning schemes to achieve reliable and efficient concept erasing.

Reinforcement learning is also introduced to erase unsafe concepts [132], [133]. ShieldDiff [132] tackles unsafe content removal by fine-tuning a pre-trained diffusion model through reinforcement learning. It employs a customized reward function that combines the CLIP model with nudity-specific rewards to filter out the nudity content. DUO [133] removes unsafe content from T2I models through preference optimization with paired image data.

There are also novel approaches that do not fall into the above categories. For example, Forget-Me-Not [134] introduces attention Re-steering which fine-tunes the UNet component to minimize the attention maps corresponding to the target concepts. UC [135] proposes a domain correction framework using adversarial training to align sensitive and anchor concepts in diffusion models, along with gradient surgery to preserve model performance by mitigating conflicts between unlearning and relearning gradients.

Guidance-based Defense. Intervening in the forward activations during the inference phase with the external information can guide the model's output towards safer and more desirable outcomes.

• ***Image + Text → Text (IT2T) Models.*** Some works [136], [137] apply safety steering vectors to calibrated activation at inference time, aiming to address harmful intents while preserving the model's performance on benign inputs. Specifically, InferAligner [136] extracts steering vectors by calculating the activation difference between harmful and harmless prompts, and employs a guidance gate to selectively control activation shifts in specific transformer layers. Based on InferAligner, ASTRA [137] identifies visual tokens from adversarial images that are most strongly associated with jailbreaks and uses these tokens to construct steering vectors.

• ***Text → Image (T2I) Models.*** EIUP [138] mitigates unsafe content by re-weighting the attention map of the target token, guided by an introduced target unsafe concept. Specifically, it integrates image latent variables with the latent embedding of the erasure prompt to generate a corresponding target unsafe attention map. During the attention feature adjustment stage, the target unsafe attention map is merged with the attention map of the original prompt. Subsequently, the target unsafe attention map is reweighted to suppress unsafe features. SLD [139] mitigates inappropriate content generation by combining text conditioning with classifier-free guidance. It edits images during inference without fine-tuning, using unsafe prompts to guide generation in the opposite direction.

• ***Text → Video (T2V) Models.*** SAFREE [140] identifies unsafe tokens and projects them into a space orthogonal to the unsafe concept subspace, while retaining their representations within the original input space. To balance toxicity filtering and the preservation of safe concepts, it employs a self-validating filtering mechanism, dynamically adjusting denoising steps and using adaptive re-attention within the diffusion latent space.

Pruning-based Defense. Unsafe content often originates from concept neurons, which are essential for generating specific concepts. Pruning model parameters to deactivate unwanted concept neurons offers an effective strategy for eliminating harmful content.

• ***Text → Image (T2I) Models.*** Some works [141], [142] selectively remove critical parameters linked to the undesired concepts. Specifically, P-ESD [141] enhances concept-erasing techniques by identifying concept-correlated neurons that are sensitive to adversarial prompts. ConceptPrune [142] identifies skilled neurons in feed-forward layers responsible for undesirable concepts in diffusion models.

B.4 Output-level Defense

Output-level Defense leverages post-processing techniques to adjust decoding strategies or rephrase the generated content, aiming to neutralize potentially offensive elements.

Decoding-based defense. Calibrating the logit distribution during decoding can steer the model's output away from unsafe directions.

• ***Image + Text → Text (IT2T) Models.*** CoCA [143] uses a contrastive decoding strategy by computing the logit difference between responses with and without the safety principle, termed the safety delta. This delta is subsequently used to adjust token generation probabilities. On the other hand, IMMUNE [144] employs a safety reward function that assigns higher scores to safe responses and lower scores to unsafe ones.

Refiner-based defense. Employing a refiner (e.g., a generative model) to rephrase or edit the model's response can also eliminate toxicity.

• ***Image + Text → Text (IT2T) Models.*** MLLM-Protector [145] comprises two core components: a lightweight harm detector, which identifies harmful content in model outputs, and a response detoxifier, which transforms harmful responses into benign ones. Instead of using external detoxifiers, ECSO [146] leverages the self-contained MLLM to transform malicious input images into plain texts in a query-aware manner. Safe response generation, free from image inputs, is then performed to restore the model's intrinsic safety mechanism.

• ***Text → Image (T2I) Models.*** LMIVS [147] begins with a Visual Commonsense Immorality Recognizer that detects immoral content in generated images. Subsequently, textual and visual immoral attribute localizers identify and highlight specific attributes contributing to the image's immorality. Once these attributes are identified, the method applies ethical image manipulation techniques to transform the content, ultimately generating morally acceptable outputs.

V. EVALUATION

Evaluation methods are essential for providing a standardized basis for comparing various jailbreak attack and defense techniques. In this section, we first review existing evaluation datasets relevant to jailbreak scenarios, followed by an overview of open-ended evaluation methods and metrics.

A. Evaluation Dataset

In this part, we review datasets commonly used in multi-modal jailbreak attacks and defenses. These datasets can be classified into two categories based on their construction intent: simulated-malicious and real-malicious datasets. Simulated-malicious datasets are typically derived from existing benign datasets that are not explicitly designed for jailbreak scenarios. In contrast, real-malicious datasets are deliberately designed to encompass a wide range of malicious topics, aiming to simulate real-world scenarios (see Table V).

A.1 Simulated-malicious Dataset

Due to the significant malicious potential of unsafe visual content, previous works have sought to mitigate the risk of generating genuinely harmful images in T2I jailbreak attacks. To achieve this, researchers [70], [77], [97] have typically selected specific categories from benign datasets to serve as proxies for unsafe classes, thereby avoiding the generation of real-world malicious imagery during model evaluation and testing. For instance, SneakyPrompt [77] introduces the Dog/Cat-100 dataset, which contains 100 prompts generated by GPT models describing scenarios involving dogs or cats. Similarly, UPAM [70] utilizes the Microsoft COCO dataset [156], a comprehensive resource containing image-text pairs, and designates 10 specific classes (e.g., boat, bird, clock) as "harmful" categories. Additionally, UnlearnDiffAtk [97] selects samples from the WikiArt [157] and Imagenette [158] datasets, to conduct attacks related to stylization and object manipulation.

A.2 Real-malicious Dataset

Real-malicious datasets include prompts intended to elicit discomfort or cause harm, covering common scenarios such as pornography, violence, and gore, which are prohibited by the usage policies of OpenAI [159] and Meta [160]. Some commonly used datasets in Table V are outlined as follows:

Image + Text → Text (IT2T) Models. Existing attack methods typically utilize text-based jailbreak datasets [48], [152], [153] combined with real-world images [89], [90], AI-generated images [72], [73], [81] or adversarial images [60], [87], [93], [95] for executing multimodal jailbreaks.

• ***AdvBench*** [48] prompts uncensored LLMs to generate two settings of harmful strings and harmful behaviors. Harmful strings encompass diverse detrimental content while harmful behaviors are framed as instructions aligned with the themes of the harmful strings.

• ***RedTeam-2K*** [152] comprises 2,000 meticulously crafted harmful queries from 16 safety policies and 8 diverse sources, including GPT Rewrite, Handcraft, GPT Generate, and several existing datasets.

• ***HarmBench*** [153] encompasses 510 unique harmful behaviors, split into 400 textual behaviors and 110 multimodal behaviors, which are designed to violate laws or societal norms.

Another line of datasets comprises adversarial image-text pairs that can successfully jailbreak MLLMs. These datasets provide a rigorous basis for defense evaluation, enabling newly proposed defense methods to be systematically validated.

TABLE V: Comparison of publicly available representative evaluation datasets. **Collected**: raw data created by humans or collected from real-world websites. **Reconstructed**: Data reorganized from other existing datasets. **Synthesized**: AI-generated data using LLM or diffusion models. **Adversarial**: Adversarial data generated by jailbreak attack methods. [link] directs to dataset websites.

Dataset	Venue	Model	Text Source				Image Source				Volume	Theme
			Collected	Reconstructed	Synthesized	Adversarial	Collected	Reconstructed	Synthesized	Adversarial		
AdvBench [48] [link]	[arXiv'23]	I + T → T	-	-	✔	-	-	-	-	-	500	-
ReadTeam-2K [152] [link]	[arXiv'24]	I + T → T	✔	✔	✔	-	-	-	-	-	2,000	16
HarmBench [153] [link]	[ICML'24]	I + T → T	✔	-	-	-	-	-	-	-	510	4
Figstep [71] [link]	[arXiv'23]	I + T → T	-	-	✔	-	-	-	-	✔ [71]	500	10
HADES [92] [link]	[ECCV'24]	I + T → T	-	-	✔	-	✔	-	✔	✔ [92]	750	5
JailBreakV-28K [152] [link]	[arXiv'24]	I + T → T	-	-	-	✔ [47]–[50]	-	✔	✔	-	28,000	16
MM-SafetyBench [55] [link]	[ECCV'24]	I + T → T	-	-	✔	-	-	-	✔	✔ [55]	5,040	13
NSFW-200 [77] [link]	[SSP'24]	T → I	-	-	✔	-	-	-	-	-	200	-
MMA [19] [link]	[CVPR'24]	T → I	-	✔	-	✔ [19]	-	-	-	✔ [19]	1,000	-
VBCDE [69] [link]	[arXiv'23]	T → I	-	✔	-	✔ [69]	-	-	-	-	100	5
MPUP [154] [link]	[arXiv'24]	T → I	-	-	✔	-	-	-	-	-	1,200	4
I2P [139] [link]	[CVPR'23]	T → I	✔	-	-	-	✔	-	-	-	4,703	7
Unsafe Diffusion [155] [link]	[CCS'23]	T → I	✔	✔	-	-	-	-	-	-	1,434	-
MACE-Celebrity [130] [link]	[CVPR'24]	T → I	✔	-	-	-	-	-	-	-	1,000	-
MACE-Art [130] [link]	[CVPR'24]	T → I	-	✔	-	-	-	-	-	-	1,000	-
T2VSafetyBench [20] [link]	[NeurIPS'24]	T → V	-	✔	✔	✔ [52]–[54]	-	-	-	-	4,400	12

• ***Figstep*** [71] identifies 10 safety-critical scenarios and leverages GPT-4 to generate 50 distinct malicious questions for each scenario. These questions are subsequently transformed into imperative sentences and converted as visual prompts using typographic techniques.

• ***HADES*** [92] employs GPT-4 to generate 50 keywords for each harmful category and synthesize three distinct harmful instructions based on each keyword, each paired with images collected from real-world websites. Additionally, this dataset includes jailbreak images generated through the combination of typographic images, synthesized images, and adversarial noise.

• ***JailBreakV-28K*** [152] is developed based on the RedTeam-2K [152] dataset and incorporates a diverse array of jailbreak attacks to construct jailbreak prompts. These prompts are paired with four categories of images: blank images, random noise images, natural images, and synthesized images generated using stable diffusion. The benchmark covers 16 distinct themes and encompasses a comprehensive total of 28,000 test cases.

• ***MM-SafetyBench*** [55] instructs GPT-4 to generate multiple malicious questions for each scenario and extract unsafe key phrases from these questions. Using the key phrases, the dataset incorporates synthesized images and typography images, creating a unified visual representation of the malicious content.

Text → Image/Video (T2I/T2V) Models. These datasets primarily consist of malicious prompts designed to evaluate the generation of inappropriate content in T2I and T2V models.

• ***NSFW-200*** [77] includes 200 prompts with malicious content generated by using ChatGPT models.

• ***MMA*** [19] selects 1,000 captions with high harmfulness scores from the LAION-COCO dataset [161]. Moreover, the dataset includes adversarial prompts and images generated using the corresponding jailbreak methods.

• ***VBCDE*** [69] comprises 100 test cases covering five harmful categories of violence, gore, illegal activities, discrimination, and pornographic content, with approximately 20 sensitive prompts in each category.

• ***MPUP*** [154] encompasses three high-risk scenarios: hate speech, physical harm, and fraud, with a total of 1,200 unsafe prompts. These prompts are generated by guiding GPT-4 with scenario-specific instructions, followed by manual selection and refinement.

• ***I2P*** [139] comprises 4,703 harmful prompts sourced from real-world websites (i.e., Lexica), spanning seven distinct themes, and pairing them with harmful images obtained from those same websites.

Although the I2P [139] dataset is widely used for evaluating jailbreak attacks [53], [54], [76], [96], [97], a group of defense works [61], [128], [131], [133], [138] construct jailbreak prompts based on I2P to assess the robustness of proposed defense mechanisms. In addition to the I2P dataset, other datasets have been developed to enable a more customized approach to studying jailbreak defenses.

• ***Unsafe Diffusion*** [155] focuses on demystifying the generation of unsafe images and hate memes from T2I models. The study collects harmful prompts from sources such as the Lexica website, and a manually created template-based dataset, along with harmless prompts from the MS COCO [162] dataset, resulting in a collection of 1,434 prompts in total.

• ***MACE-Celebrity*** [130]. To evaluate the celebrity erasure task, MACE-Celebrity consists of a database containing 200 celebrity names, each embedded within five predefined text prompts.

• ***MACE-Art*** [130]. To evaluate the artistic style erasure task, MACE-Art selects 200 artist names from the Image Synthesis Style Studies [163] database and applies each of them to a different set of five predefined text prompts.

• ***T2VSafetyBench*** [20] is designed to evaluate the safety of T2V models, focusing on 12 critical safety aspects including pornography, violence, and discrimination. It constructs a dataset of 4,400 malicious prompts sourced from existing datasets, GPT-generated prompts, and jailbreak attack-based prompts, providing a comprehensive evaluation of model safety across diverse scenarios.

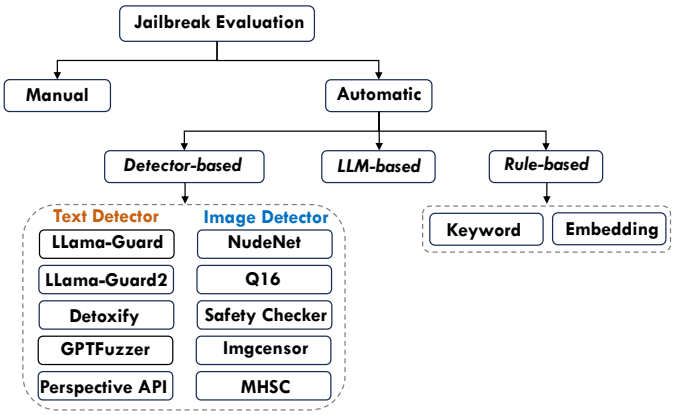

Fig. 9: Taxonomy of jailbreak evaluation method against multimodal generative models.

B. Evaluation Method

In contrast to traditional visual-question answering datasets [164], [165], generative models produce answers in an open-ended format, complicating subjective evaluation. This open-ended nature presents challenges in balancing evaluation costs with maintaining accuracy. To tackle these challenges, we have outlined diverse evaluation methods for determining relevant metrics, as illustrated in Fig. 9. These methods are primarily classified into two categories: manual evaluation and automated evaluation.

B.1 Manual Evaluation

Manual evaluation involves human assessment to determine if the content is toxic, offering a direct and interpretable method of evaluation. Manual evaluation can be categorized into two types: full evaluation and partial evaluation. Full evaluation involves assessing the entire responses, some works [21], [66], [88], [93] have adopted this approach, conducting detailed quantitative experiments to provide a comprehensive analysis of the jailbreak performance. Since full evaluation is labor-intensive and time-consuming, partial dataset evaluation focuses on evaluating a representative subset of the data. For instance, other works [117], [118], [166] use manual evaluation to verify the validity of other automatic evaluations.

B.2 Automatic Evaluation

Given the high cost and time-consuming nature of human evaluation, recent works [55], [90], [148], [155], [166] have increasingly focused on developing automated methods to evaluate the toxicity of generated content across different modalities. These approaches can be categorized into three types: detector-based evaluation, GPT-based evaluation, and rule-based evaluation.

Detector-based Evaluation. Detector-based evaluation utilizes pre-trained classifiers to automatically detect toxic content within generated outputs. These classifiers are trained on extensive, annotated datasets that encompass a broad spectrum of unsafe categories, including toxicity, violence, or explicit material.

TABLE VI: Toxicity detectors are employed as automatic evaluation for multi-modal generative models.

Toxicity detector	Access
Text Detectors	
LLama-Guard [148]	https://huggingface.co/meta-llama
LLama-Guard2 [149]	https://huggingface.co/meta-llama
Detoxify [167]	https://github.com/unitaryai/detoxify
GPTFUZZER [168]	https://huggingface.co/hubert233/GPTFuzz/tree/main
Perspective API [169]	https://perspectiveapi.com/
Image Detectors	
NudeNet [150]	https://github.com/platelminto/NudeNetClassifier
Q16 [151]	https://github.com/ml-research/Q16
Safety Checker [170]	https://huggingface.co/CompVis/stable-diffusion-safety-checker
Imgcensor [171]	https://github.com/lucasxlu/XCloud/tree/master/research/imgcensor
MHSC [155]	https://github.com/YitingQu/unsafe-diffusion

• ***Text Detector.*** Toxicity detectors for detecting harmful text encompass several models, including LLama-Guard [148], LLama-Guard2 [149], Detoxify [167], GPTFuzzer [168], and Perspective API [169]. **1) LLama-Guard series models.** LLama-Guard and LLama-Guard2 are fine-tuned on the LLaMA-based [172] models, enabling them to perform multi-class classification and generate binary decision scores for harmful responses. **2) Detoxify.** Detoxify is working to mitigate harmful online content by identifying toxic comments. **3) GPTFuzzer.** GPTFuzzer identifies harmful responses by fine-tuning the RoBERTa model [173] on manually curated datasets with human judgments. **4) Perspective API.** Perspective API employs machine learning models to identify abusive comments by scoring phrases based on their potential impact within a conversation.

• ***Image Detector.*** Toxicity detectors for detecting harmful images primarily include NudeNet [150], Q16 [151], SD Safety Checker [170], Imgcensor [171], Multi-headed Safety Classifier (MHSC) [155]. **1) NudeNet.** NudeNet detector is employed in the process of categorizing images based on the presence of "nudity", which has 4 fine-grained nudity labels. **2) Q16.** Q16 is fine-tuned using harmful images and corresponding caption data, enabling it to detect a broader range of inappropriate topics, such as violence and misogyny. **3) SD Safety Checker.** The SD Safety Checker is integrated into the output stage of Stable Diffusion to detect unsafe elements within generated images. **4) Imgcensor.** Imgcensor offers recognition capabilities for images depicting pornography and political figures, utilizing various deep and non-deep learning-based methods. **5) MHSC.** MHSC supports fine-grained detection, which builds a multi-headed image safety classifier that detects five unsafe categories simultaneously.

LLM-based Evaluation. Powerful LLMs especially for GPT-4, have increasingly been employed as safety evaluators in recent research [20], [55], [69], [82], [104], [133], [166]. Unlike specialized detectors constrained by specific themes, LLMs exhibit exceptional effectiveness in identifying jailbreak behaviors across diverse scenarios. Specifically, LLMs are employed to perform nuanced semantic analysis to assess the consistency between generated content and ground-truth annotations.

Rule-based Evaluation. Rule-based evaluation typically depends on predefined rules or patterns to assess the alignment between generated content and expected outputs. For instance, keywords can be used to identify the presence of essential concepts, while embedding-based matching ensures that the semantic meaning of the output aligns with ground-truth annotations.

• ***Keyword-based.*** Several works [62], [79], [90] on MLLMs have proposed methods to identify responses containing predefined refusal keywords, such as "I am sorry" or "I cannot." These refusal keywords act as indicators that the model is adhering to safety protocols by refusing to generate harmful or inappropriate content.

• ***Embedding-based.*** Embedding-based approaches in T2I tasks [67], [69], [70] aim to evaluate the semantic quality of generated images. By extracting high-dimensional vectors from encoders such as the CLIP model, these methods can compute their cosine similarity score to assess semantic alignment between the generated images and target harmfulness.

C. Evaluation Metric

Evaluation metrics can be broadly categorized into two primary types: robustness metrics and utility metrics. Robustness metrics are designed to assess the resilience of a model against jailbreak inputs. Utility metrics evaluate the relevance, accuracy, and overall quality of the content generated by the model.

C.1 Robustness Metrics

The attack success rate (ASR) is a widely used metric to evaluate the robustness of attack and defense. In the context of the attack, a higher ASR indicates that the generated content successfully bypasses model safeguards, leading to harmful outputs. On the defense side, a reduction in ASR suggests that the defense mechanisms are effective in preventing the generation of unsafe content.

• ***Attack Success Rate***. The Attack Success Rate (ASR) quantifies the proportion of successful attacks out of the total number of attempts. Formally, let N_{total} denotes the total number of jailbreak inputs and $N_{success}$ denotes the number of successful attacks. The ASR can then be defined as:

$$ASR = \frac{N_{success}}{N_{total}} \tag{14}$$

C.2 Utility Metrics

Utility Metrics include indicators designed to evaluate content quality across different modalities. Specifically, Prompt Perplexity measures the coherence and fluency of textual prompts. For visual content, Frechet Inception Distance measures the visual quality of generated images, and the CLIP Score evaluates the semantic alignment between images and their target intent.

• ***Prompt Perplexity***. Perplexity (PPL) [174] serves as a metric for evaluating the readability and linguistic fluency of jailbreak prompts. Many adversarial prompts targeting MLLMs tend to contain garbled or nonsensical characters, which makes them easily detectable and filtered by defense methods that rely on high-perplexity detection. As a result, attack methods [66], [70], [102] that generate low-perplexity prompts are becoming increasingly noteworthy. Formally, let $W = (w_1, w_2, ..., w_n)$ denote a text sequence where w_i represents the i-th token. The perplexity of the sequence W is defined as:

$$PPL(W) = \exp\left(-\frac{1}{n}\sum_{i=1}^{n}\log p_\theta\left(w_i|w_{<i}\right)\right) \tag{15}$$

where $p_\theta\left(w_i|w_{<i}\right)$ represents the probability assigned by MLLMs to the i-th token, conditioned on all preceding tokens in the sequence.

• ***Frechet Inception Distance***. Frechet Inception Distance (FID) [175] quantifies the distance between real and generated images. It calculates the distance between the feature representations of real and generated images, as captured by a pre-trained network, thereby assessing how closely the generated images resemble real-world images. The lower the FID score, the better the quality of the generated images, indicating closer alignment with real images. FID is computed as follows:

$$FID = \|\mu_r - \mu_g\|^2 + Tr\left(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r\Sigma_g}\right) \tag{16}$$

Where μ_r and μ_g represent the mean vectors of the real and generated image distributions, respectively, and Σ_r and Σ_g are the corresponding covariance matrices.

• ***CLIP Score***. CLIP Score [176] measures the alignment of the generated images with their corresponding descriptions. This approach first utilizes the CLIP model E_{clip} to extract the semantic embeddings of adversarial images. The CLIP Score is then obtained by calculating the cosine similarity between malicious adversarial images y_{adv} and the clean prompts x_{clean} modified from the toxic content:

$$CLIP\ Score = Cos\left(E_{clip}\left(y_{adv}\right), E_{clip}\left(x_{clean}\right)\right) \tag{17}$$

VI. Future Work

Jailbreak attack and defense against multimodal generative models remains a very challenging and open research task. In this section, we share key insights into future research directions for multimodal model jailbreaks, pinpointing what is missing in the current research and identifying directions worth further exploration.

A. Multimodal Jailbreak Attack

• **Expand Multimodal Vulnerabilities.** With the advancement of multimodal generative models, input-output configurations have expanded to include video and audio modalities, which remain relatively underexplored in the context of jailbreak attacks and defenses. Unlike text and images, the temporal dependencies of video frames and the inherent acoustic features of audio present unique challenges in both launching and defending against jailbreak attacks. There is an urgent need for further research into the vulnerabilities of models in the video and audio modalities, which will facilitate the development of more robust safety alignment mechanisms.

• **Joint Modal Optimization.** Another interesting research direction is to explore the joint optimization between multiple modalities to achieve competitive attack performance more deceptively. Several pioneering studies have been conducted,

e.g., [95] executes jailbreaks by jointly optimizing textual and visual prompts. Under this context, the adversary does not treat each modality in isolation but instead looks for strategies in which the different modalities can interact to exploit model vulnerabilities in a unified manner.

• **Combined Multimodal Output.** Most existing research focuses on either any-to-text [18], [55], [62]–[64], [71]–[74], [79], [80] or any-to-vision models [19], [52], [66], [67], [69], [70], [76]–[78], [82], [85], primarily addressing single-modality output configurations. However, the potential for harmful multimodal outputs generated by any-to-any models remains largely unexplored. Compared with single-modality jailbreak outputs, combined text-image-video-audio outputs can amplify the credibility and psychological influence of the harmful content. For instance, malicious textual descriptions paired with manipulated visual content, synchronized audio commentary, and supporting video sequences, can collaboratively create a unified and deceptive message.

B. Multimodal Jailbreak Defense

• **Hybrid Defense System.** Defense methods relying solely on fine-tuning models often fail to adapt to the evolving strategies of attackers. Conversely, training-free methods frequently fall short in delivering the necessary effectiveness and accuracy. Future research should prioritize the development of hybrid defense systems that integrate diverse techniques to enhance robustness and adaptability. Such a system would combine external refiners with internal tuning-based and guidance-driven strategies to promote collaboration.

• **Overall Quality Preserving.** Fine-tuning pre-trained generative models to erasure unsafe concepts has shown promising progress in defending against jailbreak attacks. However, the process of erasing or unlearning a large number of concepts may result in a degradation of overall generation quality [121], [130], [133], [142] including reduced coherence, fidelity and diversity in generated outputs. To mitigate these challenges, there is growing interest in developing adaptive forgetting mechanisms that selectively target unsafe concepts while preserving adjacent knowledge, as well as quality restoration strategies to maintain the quality of generative outputs.

• **Transparency and Explainability.** As defense mechanisms become increasingly complex, ensuring their transparency and explainability has become a critical research focus. Some pioneer efforts such as GuardT2I [106], have explored optimizing language models to interpret the malicious intent behind adversarial inputs. However, further endeavors are imperative to enable both users and developers to comprehensively understand the inner workings of these defense mechanisms, including their strengths and limitations.

• **Personalized Defense Schemes.** To address the varying security needs of different application scenarios, future research should focus on developing personalized defense schemes. First, enhanced stringency for sensitive content in high-stakes scenarios such as financial systems, healthcare applications, or government communications. Second, balancing security with the need for broader generative variability in scenarios that prioritize creative outputs, such as content generation for entertainment or education. These schemes would allow the security levels of defense policies to be dynamically adjusted.

C. Multimodal Jailbreak Evaluation

• **Scope of Specialized Detectors.** Current specialized detectors for automatic evaluation primarily focus on a limited set of harmful concepts, such as violence and pornography. This narrow scope significantly limits the ability to assess a broader spectrum of harmful content, including hate speech, misinformation, and cybercrime. Future research should prioritize expanding the capabilities of automated detectors to cover a more comprehensive range of harmful categories. By developing detectors capable of addressing diverse harmful concepts, the safety evaluation of multimodal models can become more robust and better aligned with the complex challenges encountered in real-world applications.

• **Comprehensive Evaluation Benchmark.** The robustness of multimodal defense mechanisms is typically evaluated using a limited set of attack methods, which often falls short of comprehensively assessing their effectiveness [61], [128], [131], [133], [138]. To address this, future research should prioritize the development of a comprehensive benchmark dataset that incorporates a wide variety of attack methods across diverse scenarios. The establishment of a standardized evaluation framework would not only promote consistency in research methodologies but also facilitate fair comparisons between defense strategies.

VII. Conclusion

The increasing prevalence of multimodal generative models has raised significant concerns about their security and reliability, particularly in the context of jailbreak attacks. As a result, extensive research has been devoted to both developing and defending against such attacks. This paper presents a comprehensive overview of jailbreak methods targeting multimodal generative models, systematically categorizing attack and defense techniques across four key levels. We provide a holistic evaluation covering a broad range of modality combinations, including text, image, audio, and video, along with different architectures of the multimodal generative model, such as Any-to-Text, Any-to-Vision, and Any-to-Any models. Additionally, we not only compare existing evaluation datasets, methods, and metrics but also propose critical challenges and potential future directions. While many challenges remain, we hope this paper inspires further discussion and provides strategic guidance for future research, ultimately contributing to the safety and reliability of foundational models.

References

[1] C. Team, "Chameleon: Mixed-modal early-fusion foundation models," *arXiv preprint arXiv:2405.09818*, 2024.

[2] X. Wang, X. Zhang, Z. Luo, Q. Sun, Y. Cui, J. Wang, F. Zhang, Y. Wang, Z. Li, Q. Yu *et al.*, "Emu3: Next-token prediction is all you need," *arXiv preprint arXiv:2409.18869*, 2024.

[3] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proceedings of the Advances in Neural Information Processing Systems*, 2023.

[4] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," in *Proceedings of the Empirical Methods in Natural Language Processing*, 2023.

[5] Z. Kong, A. Goel, R. Badlani, W. Ping, R. Valle, and B. Catanzaro, "Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities," in *Proceedings of the International Conference on Machine Learning*, 2024.

[6] J. Xiao, A. Yao, Y. Li, and T.-S. Chua, "Can i trust your answer? visually grounded video question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[7] X. Liu, P. P. Li, H. Huang, Z. Li, X. Cui, W. Deng, Z. He *et al.*, "Fka-owl: Advancing multimodal fake news detection through knowledge-augmented lvlms," in *Proceedings of the ACM International Conference on Multimedia*, 2024.

[8] X. Liu, Z. Li, P. Li, S. Xia, X. Cui, L. Huang, H. Huang, W. Deng, and Z. He, "Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms," *arXiv preprint arXiv:2406.08772*, 2024.

[9] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proceedings of the Advances in Neural Information Processing Systems*, 2020.

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[11] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in *Proceedings of the Advances in Neural Information Processing Systems*, 2022.

[12] X. Cui, Z. Li, P. Li, H. Huang, X. Liu, and Z. He, "Instastyle: Inversion noise of a stylized image is secretly a style adviser," in *Proceedings of the European Conference on Computer Vision*, 2024.

[13] X. Cui, P. Li, Z. Li, X. Liu, Y. Zou, and Z. He, "Localize, understand, collaborate: Semantic-aware dragging via intention reasoner," in *Proceedings of the Advances in Neural Information Processing Systems*, 2024.

[14] J. Christian, "Amazing "jailbreak" bypasses chatgpt's ethics safeguards," *Futurism, February*, 2023.

[15] Albert, "Jailbreak chat," *https://www.jailbreakchat.com/*, 2023.

[16] H. Li, D. Guo, W. Fan, M. Xu, J. Huang, F. Meng, and Y. Song, "Multi-step jailbreaking privacy attacks on chatgpt," in *Proceedings of the Findings of Empirical Methods in Natural Language Processing*, 2023.

[17] A. Wei, N. Haghtalab, and J. Steinhardt, "Jailbroken: How does llm safety training fail?" in *Proceedings of the Advances in Neural Information Processing Systems*, 2023.

[18] E. Shayegani, Y. Dong, and N. Abu-Ghazaleh, "Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models," in *Proceedings of the International Conference on Learning Representations*, 2024.

[19] Y. Yang, R. Gao, X. Wang, T.-Y. Ho, N. Xu, and Q. Xu, "Mma-diffusion: Multimodal attack on diffusion models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[20] Y. Miao, Y. Zhu, Y. Dong, L. Yu, J. Zhu, and X.-S. Gao, "T2vsafetybench: Evaluating the safety of text-to-video generative models," in *Proceedings of the Advances in Neural Information Processing Systems*, 2024.

[21] X. Shen, Y. Wu, M. Backes, and Y. Zhang, "Voice jailbreak attacks against gpt-4o," *arXiv preprint arXiv:2405.19103*, 2024.

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

[23] H. Jin, L. Hu, X. Li, P. Zhang, C. Chen, J. Zhuang, and H. Wang, "Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models," *arXiv preprint arXiv:2407.01599*, 2024.

[24] C. Zhang, M. Hu, W. Li, and L. Wang, "Adversarial attacks and defenses on text-to-image diffusion models: A survey," in *Proceedings of the Information Fusion*, 2024.

[25] V. T. Truong, L. B. Dang, and L. B. Le, "Attacks and defenses for generative diffusion models: A comprehensive survey," *arXiv preprint arXiv:2408.03400*, 2024.

[26] S. Wang, Z. Long, Z. Fan, and Z. Wei, "From llms to mllms: Exploring the landscape of multimodal jailbreaking," in *Proceedings of the Empirical Methods in Natural Language Processing*, 2024.

[27] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," in *Proceedings of the International Conference on Learning Representations*, 2024.

[28] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. C. H. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," in *Proceedings of the Advances in Neural Information Processing Systems*, 2023.

[29] B. Lin, B. Zhu, Y. Ye, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," *arXiv preprint arXiv:2311.10122*, 2023.

[30] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. d. C. Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov *et al.*, "Audiopalm: A large language model that can speak and listen," *arXiv preprint arXiv:2306.12925*, 2023.

[31] "Midjourney," https://midjourney.com/.

[32] OpenAI, "Moderation overview," https://platform.openai.com/docs/guides/moderation/overview.

[33] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 500–22 510.

[34] J. Xu, X. Wang, Y.-P. Cao, W. Cheng, Y. Shan, and S. Gao, "Instructp2p: Learning to edit 3d point clouds with text instructions," *arXiv preprint arXiv:2306.07154*, 2023.

[35] Z. Zheng, X. Peng, T. Yang, C. Shen, S. Li, H. Liu, Y. Zhou, T. Li, and Y. You, "Open-sora: Democratizing efficient video production for all," March 2024. [Online]. Available: https://github.com/hpcaitech/Open-Sora

[36] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[37] D. Kondratyuk, L. Yu, X. Gu, J. Lezama, J. Huang, G. Schindler, R. Hornung, V. Birodkar, J. Yan, M. Chiu, K. Somandepalli, H. Akbari, Y. Alon, Y. Cheng, J. V. Dillon, A. Gupta, M. Hahn, A. Hauth, D. Hendon, A. Martinez, D. Minnen, M. Sirotenko, K. Sohn, X. Yang, H. Adam, M. Yang, I. Essa, H. Wang, D. A. Ross, B. Seybold, and L. Jiang, "Videopoet: A large language model for zero-shot video generation," in *Proceedings of the International Conference on Machine Learning*, 2024.

[38] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng *et al.*, "Cogvideox: Text-to-video diffusion models with an expert transformer," *arXiv preprint arXiv:2408.06072*, 2024.

[39] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "Next-gpt: Any-to-any multimodal llm," in *Proceedings of the International Conference on Machine Learning*, 2024.

[40] P. Lu, B. Peng, H. Cheng, M. Galley, K. Chang, Y. N. Wu, S. Zhu, and J. Gao, "Chameleon: Plug-and-play compositional reasoning with large language models," in *Proceedings of the Advances in Neural Information Processing Systems*, 2023.

[41] OpenAI, "GPT-4o System Card," OpenAI, Technical Report, 2024. [Online]. Available: https://openai.com/index/gpt-4o-system-card/

[42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, 2021.

[43] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," *https://vicuna. lmsys. org*, 2023.

[44] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[45] Y. Huang, J. Huang, Y. Liu, M. Yan, J. Lv, J. Liu, W. Xiong, H. Zhang, S. Chen, and L. Cao, "Diffusion model-based image editing: A survey," *arXiv preprint arXiv:2402.17525*, 2024.

[46] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[47] Y. Liu, G. Deng, Z. Xu, Y. Li, Y. Zheng, Y. Zhang, L. Zhao, T. Zhang, K. Wang, and Y. Liu, "Jailbreaking chatgpt via prompt engineering: An empirical study," *arXiv preprint arXiv:2305.13860*, 2023.

[48] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.

[49] N. Xu, F. Wang, B. Zhou, B. Li, C. Xiao, and M. Chen, "Cognitive overload: Jailbreaking large language models with overloaded logical thinking," in *Proceedings of the Findings of the North American Conference on Chinese Linguistics*, 2024.

[50] Y. Zeng, H. Lin, J. Zhang, D. Yang, R. Jia, and W. Shi, "How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge

AI safety by humanizing llms," in *Proceedings of the Association for Computational Linguistics*, 2024.

[51] X. Liu, N. Xu, M. Chen, and C. Xiao, "Autodan: Generating stealthy jailbreak prompts on aligned large language models," in *Proceedings of the International Conference on Learning Representations*, 2024.

[52] Y. Tian, X. Yang, Y. Dong, H. Yang, H. Su, and J. Zhu, "Bspa: Exploring black-box stealthy prompt attacks against image generators," *arXiv preprint arXiv:2402.15218*, 2024.

[53] Y.-L. Tsai, C.-Y. Hsu, C. Xie, C.-H. Lin, J.-Y. Chen, B. Li, P.-Y. Chen, C.-M. Yu, and C.-Y. Huang, "Ring-a-bell! how reliable are concept removal methods for diffusion models?" in *Proceedings of the International Conference on Learning Representations*, 2024.

[54] J. Ma, A. Cao, Z. Xiao, J. Zhang, C. Ye, and J. Zhao, "Jailbreaking prompt attack: A controllable adversarial attack against diffusion models," *arXiv preprint arXiv:2404.02928*, 2024.

[55] X. Liu, Y. Zhu, J. Gu, Y. Lan, C. Yang, and Y. Qiao, "Mm-safetybench: A benchmark for safety evaluation of multimodal large language models," in *Proceedings of the European Conference on Computer Vision*, 2024.

[56] J. Rando, H. Korevaar, E. Brinkman, I. Evtimov, and F. Tramèr, "Gradient-based jailbreak images for multimodal fusion models," *arXiv preprint arXiv:2410.03489*, 2024.

[57] F. Bianchi, M. Suzgun, G. Attanasio, P. Röttger, D. Jurafsky, T. Hashimoto, and J. Zou, "Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions," in *Proceedings of the International Conference on Learning Representations*, 2024.

[58] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan *et al.*, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

[59] M. Z. Hossain and A. Imteaj, "Sim-clip: Unsupervised siamese adversarial fine-tuning for robust and semantically-rich vision-language models," *arXiv preprint arXiv:2407.14971*, 2024.

[60] M. Z. Hossain and A. Imteaj, "Securing vision-language models with a robust encoder against jailbreak and adversarial attacks," *arXiv preprint arXiv:2409.07353*, 2024.

[61] S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, R. Cucchiara *et al.*, "Safe-clip: Removing nsfw concepts from vision-and-language models," in *Proceedings of the European Conference on Computer Vision*, 2024.

[62] Y. Wu, Y. Huang, Y. Liu, X. Li, P. Zhou, and L. Sun, "Can large language models automatically jailbreak gpt-4v?" in *Proceedings of the North American Conference on Chinese Linguistics Workshops*, 2024.

[63] Y. Liu, C. Cai, X. Zhang, X. Yuan, and C. Wang, "Arondight: Red teaming large vision language models with auto-generated multi-modal jailbreak prompts," in *Proceedings of the ACM International Conference on Multimedia*, 2024.

[64] C. Xu, M. Kang, J. Zhang, Z. Liao, L. Mo, M. Yuan, H. Sun, and B. Li, "Advweb: Controllable black-box attacks on vlm-powered web agents," *arXiv preprint arXiv:2410.17401*, 2024.

[65] J. Rando, D. Paleka, D. Lindner, L. Heim, and F. Tramèr, "Red-teaming the stable diffusion safety filter," in *Proceedings of the Advances in Neural Information Processing Systems Workshops*, 2022.

[66] Y. Huang, L. Liang, T. Li, X. Jia, R. Wang, W. Miao, G. Pu, and Y. Liu, "Perception-guided jailbreak against text-to-image models," *arXiv preprint arXiv:2408.10848*, 2024.

[67] Z. Ba, J. Zhong, J. Lei, P. Cheng, Q. Wang, Z. Qin, Z. Wang, and K. Ren, "Surrogateprompt: Bypassing the safety filter of text-to-image models via substitution," in *Proceedings of the ACM Conference on Computer and Communications Security*, 2024.

[68] Y. Ma, S. Pang, Q. Guo, T. Wei, and Q. Guo, "Coljailbreak: Collaborative generation and editing for jailbreaking text-to-image deep generation," in *Proceedings of the Advances in Neural Information Processing Systems*, 2024.

[69] Y. Deng and H. Chen, "Divide-and-conquer attack: Harnessing the power of llm to bypass the censorship of text-to-image generation model," *arXiv preprint arXiv:2312.07130*, 2023.

[70] D. Peng, Q. Ke, and J. Liu, "Upam: Unified prompt attack in text-to-image generation models against both textual filters and visual checkers," in *Proceedings of the International Conference on Machine Learning*, 2024.

[71] Y. Gong, D. Ran, J. Liu, C. Wang, T. Cong, A. Wang, S. Duan, and X. Wang, "Figstep: Jailbreaking large vision-language models via typographic visual prompts," *arXiv preprint arXiv:2311.05608*, 2023.

[72] X. Zou and Y. Chen, "Image-to-text logic jailbreak: Your imagination can help you do anything," *arXiv preprint arXiv:2407.02534*, 2024.

[73] S. Ma, W. Luo, Y. Wang, X. Liu, M. Chen, B. Li, and C. Xiao, "Visual-roleplay: Universal jailbreak attack on multimodal large language models via role-playing image characte," *arXiv preprint arXiv:2405.20773*, 2024.

[74] H. Yang, L. Qu, E. Shareghi, and G. Haffari, "Audio is the achilles' heel: Red teaming audio large multimodal models," *arXiv preprint arXiv:2410.23861*, 2024.

[75] T. Chen, K. Wang, and H. Wei, "Zer0-jack: A memory-efficient gradient-based jailbreaking method for black-box multi-modal large language models," *arXiv preprint arXiv:2411.07559*, 2024.

[76] P. Dang, X. Hu, D. Li, R. Zhang, Q. Guo, and K. Xu, "Diffzoo: A purely query-based black-box attack for red-teaming text-to-image generative model via zeroth order optimization," *arXiv preprint arXiv:2408.11071*, 2024.

[77] Y. Yang, B. Hui, H. Yuan, N. Gong, and Y. Cao, "Sneakyprompt: Jailbreaking text-to-image generative models," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2024.

[78] S. Gao, X. Jia, Y. Huang, R. Duan, J. Gu, Y. Liu, and Q. Guo, "Rt-attack: Jailbreaking text-to-image models via random token," *arXiv preprint arXiv:2408.13896*, 2024.

[79] Y. Wu, X. Li, Y. Liu, P. Zhou, and L. Sun, "Jailbreaking gpt-4v via self-adversarial attacks with system prompts," *arXiv preprint arXiv:2311.09127*, 2023.

[80] C. Cui, G. Deng, A. Zhang, J. Zheng, Y. Li, L. Gao, T. Zhang, and T.-S. Chua, "Safe+ safe= unsafe? exploring how safe images can be exploited to jailbreak large vision-language models," *arXiv preprint arXiv:2411.11496*, 2024.

[81] R. Wang, B. Wang, X. Ma, and Y.-G. Jiang, "Ideator: Jailbreaking vlms using vlms," *arXiv preprint arXiv:2411.00827*, 2024.

[82] M. Kim, H. Lee, B. Gong, H. Zhang, and S. J. Hwang, "Automatic jailbreaking of the text-to-image generative ai systems," *arXiv preprint arXiv:2405.16567*, 2024.

[83] W. Wang, K. Gao, Z. Jia, Y. Yuan, J.-t. Huang, Q. Liu, S. Wang, W. Jiao, and Z. Tu, "Chain-of-jailbreak attack for image generation models via editing step by step," *arXiv preprint arXiv:2410.03869*, 2024.

[84] Z.-Y. Chin, K.-C. Mu, M. Fritz, P.-Y. Chen, and W.-C. Chiu, "In-context experience replay facilitates safety red-teaming of text-to-image diffusion models," *arXiv preprint arXiv:2411.16769*, 2024.

[85] Y. Dong, Z. Li, X. Meng, N. Yu, and S. Guo, "Jailbreaking text-to-image models with llm-based agents," *arXiv preprint arXiv:2408.00523*, 2024.

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

[87] L. Bailey, E. Ong, S. Russell, and S. Emmons, "Image hijacks: Adversarial images can control generative models at runtime," in *Proceedings of the International Conference on Machine Learning*, 2024.

[88] X. Qi, K. Huang, A. Panda, P. Henderson, M. Wang, and P. Mittal, "Visual adversarial examples jailbreak aligned large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[89] Z. Niu, Y. Sun, H. Ren, H. Ji, Q. Wang, X. Ma, G. Hua, and R. Jin, "Efficient llm-jailbreaking by introducing visual modality," *arXiv preprint arXiv:2405.20015*, 2024.

[90] Z. Niu, H. Ren, X. Gao, G. Hua, and R. Jin, "Jailbreaking attack against multimodal large language model," *arXiv preprint arXiv:2402.02309*, 2024.

[91] N. Carlini, M. Nasr, C. A. Choquette-Choo, M. Jagielski, I. Gao, P. W. W. Koh, D. Ippolito, F. Tramer, and L. Schmidt, "Are aligned neural networks adversarially aligned?" in *Proceedings of the Advances in Neural Information Processing Systems*, 2023.

[92] Y. Li, H. Guo, K. Zhou, W. X. Zhao, and J.-R. Wen, "Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models," in *Proceedings of the European Conference on Computer Vision*, 2024.

[93] X. Gu, X. Zheng, T. Pang, C. Du, Q. Liu, Y. Wang, J. Jiang, and M. Lin, "Agent smith: A single image can jailbreak one million multimodal llm agents exponentially fast," in *Proceedings of the International Conference on Machine Learning*, 2024.

[94] R. Wang, X. Ma, H. Zhou, C. Ji, G. Ye, and Y.-G. Jiang, "White-box multimodal jailbreaks against large vision-language models," in *Proceedings of the ACM International Conference on Multimedia*, 2024.

[95] Z. Ying, A. Liu, T. Zhang, Z. Yu, S. Liang, X. Liu, and D. Tao, "Jailbreak vision language models via bi-modal adversarial prompt," *arXiv preprint arXiv:2406.04031*, 2024.

[96] Z.-Y. Chin, C.-M. Jiang, C.-C. Huang, P.-Y. Chen, and W.-C. Chiu, "Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts," in *Proceedings of the International Conference on Machine Learning*, 2024.

[97] Y. Zhang, J. Jia, X. Chen, A. Chen, Y. Zhang, J. Liu, K. Ding, and S. Liu, "To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now," in *Proceedings of the European Conference on Computer Vision*, 2024.

[98] X. Li, Q. Shen, and K. Kawaguchi, "Va3: Virtually assured amplification attack on probabilistic copyright protection for text-to-image generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[99] Y. Zhao, T. Pang, C. Du, X. Yang, C. Li, N.-M. M. Cheung, and M. Lin, "On evaluating adversarial robustness of large vision-language models," in *Proceedings of the Advances in Neural Information Processing Systems*, 2023.

[100] Y. Zeng, Y. Cao, B. Cao, Y. Chang, J. Chen, and L. Lin, "Advi2i: Adversarial image attack on image-to-image diffusion models," *arXiv preprint arXiv:2410.21471*, 2024.

[101] S. Sivanandam, S. Deepa, S. Sivanandam, and S. Deepa, *Genetic algorithms*. Springer, 2008.

[102] C.-C. Kao, C.-M. Yu, C.-S. Lu, and C.-S. Chen, "Information-theoretical principled trade-off between jailbreakability and stealthiness on vision language models," *arXiv preprint arXiv:2410.01438*, 2024.

[103] "Leonardo.Ai," https://leonardo.ai/2023.

[104] Y. Xu, X. Qi, Z. Qin, and W. Wang, "Defending jailbreak attack in vlms via cross-modality information detector," *arXiv preprint arXiv:2407.21659*, 2024.

[105] R. Liu, A. Khakzar, J. Gu, Q. Chen, P. Torr, and F. Pizzati, "Latent guard: a safety framework for text-to-image generation," in *Proceedings of the European Conference on Computer Vision*, 2024.

[106] Y. Yang, R. Gao, X. Yang, J. Zhong, and Q. Xu, "Guardt2i: Defending text-to-image models from adversarial prompts," in *Proceedings of the Advances in Neural Information Processing Systems*, 2024.

[107] Y. Huang, F. Zhu, J. Tang, P. Zhou, W. Lei, J. Lv, and T.-S. Chua, "Effective and efficient adversarial detection for vision-language models via a single vector," *arXiv preprint arXiv:2410.22888*, 2024.

[108] X. Zhang, C. Zhang, T. Li, Y. Huang, X. Jia, X. Xie, Y. Liu, and C. Shen, "Jailguard: A universal detection framework for llm prompt-based attacks," *arXiv preprint arXiv:2312.10766*, 2023.

[109] A. Das, V. Duddu, R. Zhang, and N. Asokan, "Espresso: Robust concept filtering in text-to-image models," *arXiv preprint arXiv:2404.19227*, 2024.

[110] Y. Wang, X. Liu, Y. Li, M. Chen, and C. Xiao, "Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting," *arXiv preprint arXiv:2403.09513*, 2024.

[111] S. Oh, Y. Jin, M. Sharma, D. Kim, E. Ma, G. Verma, and S. Kumar, "Uniguard: Towards universal safety guardrails for jailbreak attacks on multimodal large language models," *arXiv preprint arXiv:2411.01703*, 2024.

[112] Y. Zhao, X. Zheng, L. Luo, Y. Li, X. Ma, and Y.-G. Jiang, "Bluesuffix: Reinforced blue teaming for vision-language models against jailbreak attacks," *arXiv preprint arXiv:2410.20971*, 2024.

[113] Z. Wu, H. Gao, Y. Wang, X. Zhang, and S. Wang, "Universal prompt optimizer for safe text-to-image generation," in *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 2024.

[114] Y. Zhang, X. Chen, J. Jia, Y. Zhang, C. Fan, J. Liu, M. Hong, K. Ding, and S. Liu, "Defensive unlearning with adversarial training for robust concept erasure in diffusion models," in *Proceedings of the Advances in Neural Information Processing Systems*, 2024.

[115] H. Li, C. Shen, P. Torr, V. Tresp, and J. Gu, "Self-discovering interpretable diffusion latent directions for responsible text-to-image generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[116] H. Qiu, G. Chen, M. Zhang, and M. Yang, "Safe text-to-image generation: Simply sanitize the prompt embedding," *arXiv preprint arXiv:2411.10329*, 2024.

[117] Y. Zong, O. Bohdal, T. Yu, Y. Yang, and T. Hospedales, "Safety fine-tuning at (almost) no cost: A baseline for vision large language models," in *Proceedings of the International Conference on Machine Learning*, 2024.

[118] Z. Liu, Y. Nie, Y. Tan, X. Yue, Q. Cui, C. Wang, X. Zhu, and B. Zheng, "Safety alignment for vision language models," *arXiv preprint arXiv:2405.13581*, 2024.

[119] Y. Chen, H. Li, Z. Zheng, and Y. Song, "Bathe: Defense against the jailbreak attack in multimodal large language models by treating harmful instruction as backdoor trigger," *arXiv preprint arXiv:2408.09093*, 2024.

[120] T. Chakraborty, E. Shayegani, Z. Cai, N. Abu-Ghazaleh, M. S. Asif, Y. Dong, A. K. Roy-Chowdhury, and C. Song, "Cross-modal safety alignment: Is textual unlearning all you need?" *arXiv preprint arXiv:2406.02575*, 2024.

[121] R. Gandikota, J. Materzynska, J. Fiotto-Kaufman, and D. Bau, "Erasing concepts from diffusion models," in *Proceedings of the IEEE International Conference on Computer Vision*, 2023.

[122] C.-P. Huang, K.-P. Chang, C.-T. Tsai, Y.-H. Lai, and Y.-C. F. Wang, "Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers," in *Proceedings of the European Conference on Computer Vision*, 2023.

[123] C. Fan, J. Liu, Y. Zhang, D. Wei, E. Wong, and S. Liu, "Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation," in *Proceedings of the International Conference on Learning Representations*, 2024.

[124] Z. Meng, B. Peng, X. Jin, Y. Jiang, J. Dong, W. Wang, and T. Tan, "Dark miner: Defend against unsafe generation for text-to-image diffusion models," *arXiv preprint arXiv:2409.17682*, 2024.

[125] J. Wu, T. Le, M. Hayat, and M. Harandi, "Erasediff: Erasing data influence in diffusion models," *arXiv preprint arXiv:2401.05779*, 2024.

[126] T. Chen, S. Zhang, and M. Zhou, "Score forgetting distillation: A swift, data-free method for machine unlearning in diffusion models," *arXiv preprint arXiv:2409.11219*, 2024.

[127] S. Kim, S. Jung, B. Kim, M. Choi, J. Shin, and J. Lee, "Towards safe self-distillation of internet-scale text-to-image diffusion models," in *Proceedings of the International Conference on Machine Learning Workshops*, 2023.

[128] X. Li, Y. Yang, J. Deng, C. Yan, Y. Chen, X. Ji, and W. Xu, "Safegen: Mitigating unsafe content generation in text-to-image models," in *Proceedings of the ACM Conference on Computer and Communications Security*, 2024.

[129] R. Gandikota, H. Orgad, Y. Belinkov, J. Materzyńska, and D. Bau, "Unified concept editing in diffusion models," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2024.

[130] S. Lu, Z. Wang, L. Li, Y. Liu, and A. W.-K. Kong, "Mace: Mass concept erasure in diffusion models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[131] C. Gong, K. Chen, Z. Wei, J. Chen, and Y.-G. Jiang, "Reliable and efficient concept erasure of text-to-image diffusion models," in *Proceedings of the European Conference on Computer Vision*, 2024.

[132] D. Han, S. Mohamed, and Y. Li, "Shielddiff: Suppressing sexual content generation from diffusion models through reinforcement learning," *arXiv preprint arXiv:2410.05309*, 2024.

[133] Y.-H. Park, S. Yun, J.-H. Kim, J. Kim, G. Jang, Y. Jeong, J. Jo, and G. Lee, "Direct unlearning optimization for robust and safe text-to-image models," in *Proceedings of the International Conference on Machine Learning Workshops*, 2024.

[134] G. Zhang, K. Wang, X. Xu, Z. Wang, and H. Shi, "Forget-me-not: Learning to forget in text-to-image diffusion models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

[135] Y. Wu, S. Zhou, M. Yang, L. Wang, W. Zhu, H. Chang, X. Zhou, and X. Yang, "Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient," *arXiv preprint arXiv:2405.15304*, 2024.

[136] P. Wang, D. Zhang, L. Li, C. Tan, X. Wang, K. Ren, B. Jiang, and X. Qiu, "Inferaligner: Inference-time alignment for harmlessness through cross-model guidance," in *Proceedings of the Empirical Methods in Natural Language Processing*, 2024.

[137] H. Wang, G. Wang, and H. Zhang, "Steering away from harm: An adaptive approach to defending vision language model against jailbreaks," *arXiv preprint arXiv:2411.16721*, 2024.

[138] D. Chen, Z. Li, M. Fan, C. Chen, W. Zhou, and Y. Li, "Eiup: A training-free approach to erase non-compliant concepts conditioned on implicit unsafe prompts," *arXiv preprint arXiv:2408.01014*, 2024.

[139] P. Schramowski, M. Brack, B. Deiseroth, and K. Kersting, "Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[140] J. Yoon, S. Yu, V. Patil, H. Yao, and M. Bansal, "Safree: Training-free and adaptive guard for safe text-to-image and video generation," *arXiv preprint arXiv:2410.12761*, 2024.

[141] T. Yang, J. Cao, and C. Xu, "Pruning for robust concept erasing in diffusion models," *arXiv preprint arXiv:2405.16534*, 2024.

[142] R. Chavhan, D. Li, and T. Hospedales, "Conceptprune: Concept editing in diffusion models via skilled neuron pruning," *arXiv preprint arXiv:2405.19237*, 2024.
[143] J. Gao, R. Pi, T. Han, H. Wu, L. Hong, L. Kong, X. Jiang, and Z. Li, "Coca: Regaining safety-awareness of multimodal large language models with constitutional calibration," in *Proceedings of the Conference on Language Modeling*, 2024.
[144] S. S. Ghosal, S. Chakraborty, V. Singh, T. Guan, M. Wang, A. Beirami, F. Huang, A. Velasquez, D. Manocha, and A. S. Bedi, "Immune: Improving safety against jailbreaks in multi-modal llms via inference-time alignment," *arXiv preprint arXiv:2411.18688*, 2024.
[145] R. Pi, T. Han, Y. Xie, R. Pan, Q. Lian, H. Dong, J. Zhang, and T. Zhang, "Mllm-protector: Ensuring mllm's safety without hurting performance," in *Proceedings of the Empirical Methods in Natural Language Processing*, 2024.
[146] Y. Gou, K. Chen, Z. Liu, L. Hong, H. Xu, Z. Li, D.-Y. Yeung, J. T. Kwok, and Y. Zhang, "Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation," in *Proceedings of the European Conference on Computer Vision*, 2024.
[147] S. Park, S. Moon, S. Park, and J. Kim, "Localization and manipulation of immoral visual cues for safe text-to-image generation," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2024.
[148] H. Inan, K. Upasani, J. Chi, R. Rungta, K. Iyer, Y. Mao, M. Tontchev, Q. Hu, B. Fuller, D. Testuggine *et al.*, "Llama guard: Llm-based input-output safeguard for human-ai conversations," *arXiv preprint arXiv:2312.06674*, 2023.
[149] L. Team, "Meta llama guard 2," https://github.com/meta-llama/PurpleLlama/blob/main/Llama-Guard2/MODEL_CARD.md, 2024.
[150] P. Bedapudi, "Nudenet: Neural nets for nudity classification, detection and selective censoring," 2019.
[151] P. Schramowski, C. Tauchmann, and K. Kersting, "Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content?" in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, 2022.
[152] W. Luo, S. Ma, X. Liu, X. Guo, and C. Xiao, "Jailbreakv-28k: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks," *arXiv preprint arXiv:2404.03027*, 2024.
[153] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, D. A. Forsyth, and D. Hendrycks, "Harmbench: A standardized evaluation framework for automated red teaming and robust refusal," in *Proceedings of the International Conference on Machine Learning*, 2024.
[154] T. Liu, Z. Lai, G. Zhang, P. Torr, V. Demberg, V. Tresp, and J. Gu, "Multimodal pragmatic jailbreak on text-to-image models," *arXiv preprint arXiv:2409.19149*, 2024.
[155] Y. Qu, X. Shen, X. He, M. Backes, S. Zannettou, and Y. Zhang, "Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2023.
[156] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision*, 2014.
[157] B. Saleh and A. Elgammal, "Large-scale classification of fine-art paintings: Learning the right metric on the right feature," *arXiv preprint arXiv:1505.00855*, 2015.
[158] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenette," 2024, https://github.com/fastai/imagenette.
[159] OpenAI, "Openai usage policy," 2023, accessed on 10-2023. [Online]. Available: https://openai.com/policies/usage-policies
[160] Meta, "Llama usage policy," 2023, accessed on 10-2023. [Online]. Available: https://ai.meta.com/llama/use-policy
[161] S. Christoph, K. Andreas, C. Theo, V. Richard, T. Benjamin, and R. Beaumont, "Laion-coco," 2024, https://laion.ai/blog/laion-coco/.
[162] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision*, 2014.
[163] S. I, P. Centauri B, Erratica, and S. Young, "Image synthesis style studies." https://www.aiartapps.com/ai-art-apps/image-synthesis-style-studies.
[164] M. Mathew, D. Karatzas, and C. Jawahar, "Docvqa: A dataset for vqa on document images," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2021.
[165] A. Masry, D. X. Long, J. Q. Tan, S. Joty, and E. Hoque, "Chartqa: A benchmark for question answering about charts with visual and logical reasoning," in *Proceedings of the Findings of the Association for Computational Linguistics*, 2022.
[166] X. Tao, S. Zhong, L. Li, Q. Liu, and L. Kong, "Imgtrojan: Jailbreaking vision-language models with one image," *arXiv preprint arXiv:2403.02910*, 2024.
[167] L. Hanu and Unitary team, "Detoxify," Github. https://github.com/unitaryai/detoxify, 2020.
[168] J. Yu, X. Lin, Z. Yu, and X. Xing, "Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts," *arXiv preprint arXiv:2309.10253*, 2023.
[169] "Perspective API," https://perspectiveapi.com/.
[170] "Safety Checker nested in Stable Diffusion," https://huggingface.co/CompVis/stable-diffusion-safety-checker.
[171] "Image censorship," https://github.com/lucasxlu/XCloud/tree/master/research/imgcensor, 2019.
[172] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.
[173] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.
[174] H. Liu, Y. Wu, S. Zhai, B. Yuan, and N. Zhang, "Riatig: Reliable and imperceptible adversarial text-to-image generation with natural prompts," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.
[175] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proceedings of the Advances in Neural Information Processing Systems*, 2017.
[176] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," in *Proceedings of the Empirical Methods in Natural Language Processing*, 2021.