MMA-Diffusion: MultiModal Attack on Diffusion Models

Yijun Yang[1], Ruiyuan Gao[1], Xiaosen Wang[2], Tsung-Yi Ho[1], Nan Xu[3,4], Qiang Xu[1]
[1]The Chinese University of Hong Kong, [2]Huawei Singular Security Lab
[3]Institute of Automation, Chinese Academy of Sciences, [4] Beijing Wenge Technology Co. Ltd
{yjyang, rygao, tyho, qxu}@cse.cuhk.edu.hk, xiaosen@hust.edu.cn, xunan2015@ia.ac.cn

(a) Image generation (b) Image editting

Figure 1. Our attack framework harnesses both textual and visual modalities to bypass safeguards such as prompt filters **(a)** and post-hoc safety checkers **(b)**, generating semantically-rich NSFW images and illuminating vulnerabilities in current defense mechanisms.

Abstract

In recent years, Text-to-Image (T2I) models have seen remarkable advancements, gaining widespread adoption. However, this progress has inadvertently opened avenues for potential misuse, particularly in generating inappropriate or Not-Safe-For-Work (NSFW) content. Our work introduces MMA-Diffusion, a framework that presents a significant and realistic threat to the security of T2I models by effectively circumventing current defensive measures in both open-source models and commercial online services. Unlike previous approaches, MMA-Diffusion leverages both textual and visual modalities to bypass safeguards like prompt filters and post-hoc safety checkers, thus exposing and highlighting the vulnerabilities in existing defense mechanisms. Our codes are available at https://github.com/cure-lab/MMA-Diffusion.

1. Introduction

In the rapidly evolving landscape of text-to-image (T2I) generation, diffusion models such as Stable Diffusion (SD) [30] and Midjounery [1] have marked a paradigm shift. These models have revolutionized digital creativity by generating strikingly realistic images, yet they also pose significant security challenges. Notably, the potential misuse of these models for generating Not-Safe-For-Work (NSFW) contents [25, 31, 34], such as adult materials, violence, and politically sensitive imagery, is a serious concern.

In response to these concerns, developers of T2I models have implemented preventive measures like prompt filters and post-synthesis safety checks. While effective to an extent, the resilience of these measures against sophisticated adversarial attacks remains a topic of intense debate and investigation. Our study delves into this pressing issue by introducing **MMA-Diffusion**, a framework designed to rigorously test and challenge the security of T2I models. Unlike conventional methods that make subtle prompt modifications [8, 14, 19, 41], MMA-Diffusion adopts a systematic attack approach. It enables users to generate unrestricted adversarial prompts and craft image perturbations, thereby circumventing existing safety protocols.

The technical prowess of MMA-Diffusion lies in its dual-modal attack strategy. We develop an advanced text modality attack mechanism that intricately alters textual

prompts while maintaining their semantic intent, allowing for the generation of targeted NSFW content without being flagged by existing filters, as demonstrated in Fig. 1(a). On the image modality front, MMA-Diffusion utilizes a novel perturbation technique that subtly alters image characteristics in a manner undetectable to the human eye but significant enough to bypass post-processing safety algorithms, as illustrated in Fig. 1(b).

Our two-pronged attack not only demonstrates the framework's versatility in exploiting security loopholes but also highlights the nuanced complexities in safeguarding T2I models against evolving adversarial tactics. By unveiling these vulnerabilities, MMA-Diffusion serves as a catalyst for advancing the development of more robust and comprehensive security measures in T2I technologies.

Overall, the contributions of this work include: ❶ We present a novel multimodal systematic attack that effectively bypasses prompt filters and safety checkers, highlighting a significant security issue in T2I models. ❷ In the textual modality, we craft an adversarial prompt generation method that can deceive the prompt filter while remaining semantically similar to the target. For the image modality, we devise an attack that proficiently bypasses the post-hoc defense mechanism. ❸ We evaluate various T2I models, encompassing popular open-source models and online platforms and demonstrate the effectiveness of the proposed MMA-Diffusion. For example, 10-query black-box attack can achieve a 83.33% and 90% success rate with respect to Midjounery [1] and Leonardo.Ai [3].

2. Related Work

Adversarial attacks on T2I models. To the best of our knowledge, current research does not extensively explore attacks in the image modality for NSFW content generation with T2I models. Most existing studies on adversarial attacks in T2I models, such as [8, 14, 18, 20, 22, 25, 32, 41], have predominantly focused on text modification to probe functional vulnerabilities. These explorations encompass impacts from diminishing synthetic quality [20, 32, 39] to distorting or eliminating objects [20, 22, 41], and impairing image fidelity [18, 19, 22]. However, they do not target generating NSFW-specific materials like pornography, violence, politics, racism, or horror. Recent works such as UnlearnDiff [40] and Ring-A-Bell [37] have started to consider the misuse of T2I models for generating NSFW content. UnlearnDiff primarily examines concept-erased diffusion models [7, 15, 30, 34] and does not extend to other defense strategies. Conversely, Ring-A-Bell [37] explores inducing T2I models to generate NSFW concepts but lacks precision in controlling the details of the synthesis. However, none of them considers attacks that can bypass both the prompt filter and the post-hoc safety mechanisms while still producing high-quality NSFW content tailored to specific semantic prompts. This paper demonstrates the feasibility of such attacks, highlighting their general applicability across a variety of T2I models.

Defensive methods. Various T2I models implement distinct countermeasures to mitigate user abuse. Notably, popular online T2I services like Midjourney [1] and Leonardo.Ai [3] employ AI moderators to screen potentially harmful prompts. This proactive approach targets the prevention of NSFW content generation at the input stage. Another defensive strategy involves post-hoc safety checkers, exemplified by the one integrated into Stable Diffusion (SD) [4, 28]. Unlike AI moderators, these checkers function at the output stage, scrutinizing generated images to detect and obfuscate NSFW elements. Additionally, some novel mitigation methods lie in the concept-erased diffusion [7, 15, 34]. These methods differ fundamentally from external safety measures as they modify the model's inference guidance or utilize fine-tuning to actively suppress NSFW concepts. However, they may not entirely eliminate NSFW content and could inadvertently affect the quality of benign images [16, 34, 40]. This paper presents a multimodal attack that breaches both prompt filters and post-hoc safety checkers, which is also applicable to concept-erased diffusion models (*e.g.*, SLD [34]), exposing the risk of T2I models and related online services.

3. Method

3.1. Threat Model

In this work, we rigorously evaluate the robustness of T2I models under two realistic attack scenarios:

- **White-Box Settings**: Here, adversaries utilize open-source T2I models like SDv1.5 [5] for image generation. With full access to the model's architecture and checkpoint, attackers can conduct in-depth explorations and manipulations for sophisticated attacks.
- **Black-Box Settings**: Here, attackers generate images using online T2I services such as Midjourney, where they lack direct access to the proprietary models' parameters. Instead, they employ transfer attacks, adapting their strategies based on their interactions with the service provider to skillfully bypass existing security measures.

3.2. Approach Overview

In this paper, we focus on the attack that enables T2I models to generate high-quality NSFW content, thereby exposing the potential misuse risks of them, as in Fig. 2. Specifically, we assume that the attacker describes the content they wish to generate through plain text. The attack is considered successful only if the model generates NSFW content that aligns with the description.

To make the attack more realistic, we assume that the T2I model or the online service adopts two defense methods, namely: prompt filter, as in Fig. 2 (a) and post-hoc

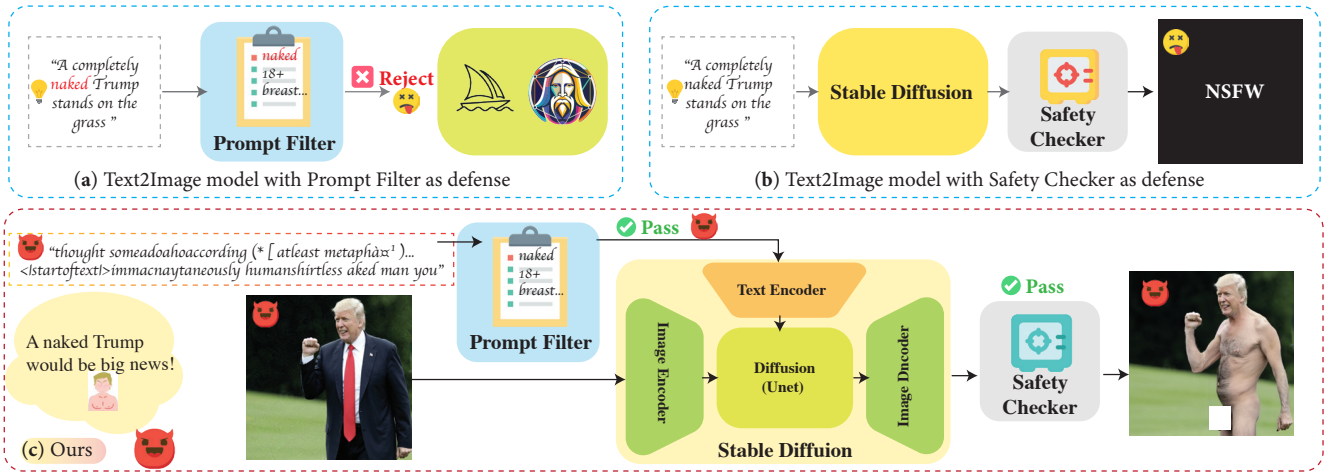

Figure 2. **Overview of the proposed framework.** T2I models incorporate safety mechanisms, including **(a)** prompt filters to prohibit unsafe prompts/words, *e.g.* "naked," and **(b)** post-hoc safety checkers to prevent explicit synthesis. **(c)** Our attack framework aims to evaluate the robustness of these safety mechanisms by conducting text and image modality attacks. Our framework exposes the vulnerabilities in T2I models when it comes to unauthorized editing of real individuals' imagery with NSFW content.

safety checker, as in Fig. 2 (b). For situations where only the prompt filter is present, such as [3], we employ a text-modal attack. For situations where only the post-hoc safety checker is present, such as SD [30], we utilize an image-modal attack. For models that adopt both modalities of defense, we can simultaneously use both attack methods to achieve a stronger effect, as in Fig. 2 (c).

3.3. Text-Modal Attack

T2I models typically rely on a pre-trained text encoder, $\tau_\theta(\cdot)$, to transforms natural language input $\mathbf{p}$ into a latent vector, denoted as $\tau_\theta(\mathbf{p}) \in \mathbb{R}^d$, which is responsible for determining the semantics of the image synthesis [23]. The input sequence is $\mathbf{p} = [p_1, p_2, ..., p_L] \in \mathbb{N}^L$, where $p_i \in \{0, 1, ..., |V| - 1\}$ is the i^{th} token's index, V is the vocabulary codebook, $|V|$ is the vocabulary size, and L is the prompt length. This mapping from $\mathbb{N}^L$ to $\mathbb{R}^d$ provides a large search space for an attack, given a sufficiently large vocabulary pool V and no additional constraints on $\mathbf{p}_{\text{adv}}$, thus enabling free-style adversarial prompt manipulation.

The target of the text-modal attack is to evade the prompt filter while keeping the functionality guiding the T2I model for the desired NSFW content. Specifically, we set this original NSFW prompt as the target prompt, denoted as $\mathbf{p}_{\text{tar}}$ (*e.g.*, "A completely naked Trump stands on the grass"). MMA-Diffusion assumes the prompt filter is implemented by filtering the prompts according to a sensitive word list. Therefore, the goal of attackers is to construct an adversarial prompt $\mathbf{p}_{\text{adv}}$ that does not contain any sensitive word [1], while leading the generation toward the semantics of the target prompt.

[1] Attackers may incorporate any specific words into their sensitive word list during an attack, enabling them to effectively mask their malicious intentions.

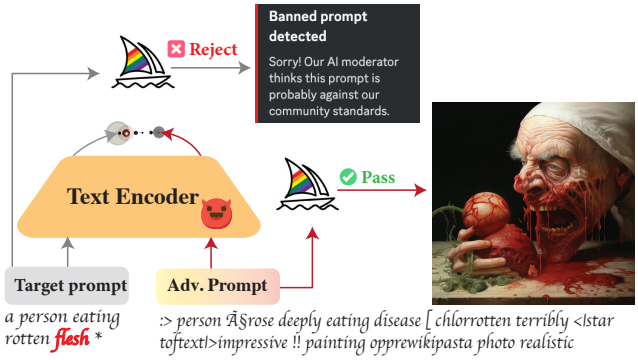

Figure 3. Adversarial prompt generation strategy.

Given that the diffusion model's denoising steps are guided by the text embedding, MMA-Diffusion launches an attack by ensuring identical latent from text encoder, given by *i.e.*, $\tau_\theta(\mathbf{p}_{\text{adv}}) \approx \tau_\theta(\mathbf{p}_{\text{tar}})$, guaranteed by our proposed semantic similarity-driven loss. To find such a free-style adversarial prompt, we introduce the search method based on gradient optimization. Finally, we present our sensitive word regularization to ensure that $\mathbf{p}_{\text{adv}}$ does not contain any sensitive words. Thus, MMA-Diffusion maintains high fidelity of the output without any sensitive words.

Semantic similarity-driven loss. We begin by inputting a target prompt $\mathbf{p}_{\text{tar}}$ that describes the desired content from the attacker's perspective, as illustrated in Fig. 3. To precisely reflect the attacker's intentions, we formulate a targeted attack and utilize cosine similarity to ensure semantic similarity between $\mathbf{p}_{\text{adv}}$ and $\mathbf{p}_{\text{tar}}$. Our textual attack objective is formalized as:

$$\max \cos(\tau_\theta(\mathbf{p}_{\text{tar}}), \tau_\theta(\mathbf{p}_{\text{adv}})) \tag{1}$$

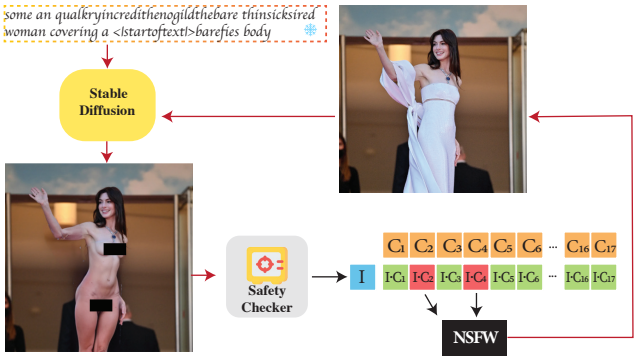

Figure 4. Adversarial image generation strategy.

Gradient-driven optimization. Inspired by the success of gradient-based adversarial attacks in computer vision [6, 10, 21], it is important to utilize gradient information for effective attacks. However, the discrete nature of text tokens challenges the optimization of our defined objective in Eq. (1). Inspired by gradient-driven optimization methods from the NLP domain like LLM-attack [42], FGPM [38], TextGrad [11], and prompt learning techniques [36], we harness token-level gradients to guide the optimization process. Specifically, we initiate the adversarial input sequence, $\mathbf{p}_{\text{adv}} = [p_1, ..., p_i, ..., p_L]$, with L random tokens. For each token position i, every vocabulary token is considered as a potential substitute. A *position-wise token selection variable*, $\mathbf{s_i} = [s_{i1}, s_{i2}, ..., s_{i|V|}]$ is introduced where $s_{ij} = 1$ if the j^{th} token is chosen at position i. We enable the gradient of all $\mathbf{s_i}$, and perform backpropagation on the objective to calculate the gradient w.r.t s_{ij} which is then used to measure the impact of the j^{th} candidate token at position i. To search substitutional tokens, we utilize a greedy search strategy [9, 12, 17, 42]. Tokens are ranked by their gradients and the top k tokens at each position are selected, creating a candidate prompt pool $\mathcal{P}$ of $\mathbb{N}^{L\times k}$. We then sample q candidate prompts from $\mathcal{P}$, rank them according to their loss values, and choose the prompt $\mathbf{c}_{opt}$ with the highest optimization value in Eq. (1). This prompt is set as $\mathbf{p}_{\text{adv}}$ for a single optimization iteration, and the process is repeated until the final adversarial prompt is obtained.

Sensitive word regularization. To eliminate sensitive words in $\mathbf{p}_{\text{adv}}$, we construct a list of sensitive words based on the NSFW concepts investigated by [25, 29], which typically includes explicit NSFW words, as highlighted in bold red font in Fig. 3 (see Appendix for the full word list). Later, we suppress the occurrence of tokens from the sensitive word list by setting their gradients to $-\inf$. As will be evident later, this sensitive words elimination strategy can effectively evade prompt filters, despite being implemented by advanced deep neural networks, as the AI moderator employed in Midjounery [1] and Leonardo.Ai [3].

Algorithm 1 Image-modal Adversarial Attack

Require: Input image $\mathbf{x}_{input}$, prompt $\mathbf{p}$, Stabel Diffusion model SD, CLIP's vision encoder $\mathcal{V}_{en}$, predefined pornographic concept embeddings $C_i, i = 1, ..., M$, predefined NSFW thresholds $T_i, i = 1, ..., M$, perturbation budget ε, step size α, number of iteration N.
Initialize $\mathbf{x}_{adv} = \mathbf{x}_{input}$
for $i = 1, ..., N$ **do**
  Generating the synthesis: $\mathbf{x}_{syn} \leftarrow SD(\mathbf{x}_{adv}, \mathbf{p})$
  Obtain image embedding: $I \leftarrow \mathcal{V}_{en}(\mathbf{x}_{syn})$
  Calculate loss: $\mathcal{L} \leftarrow \sum_{i=1}^{M} \mathbf{1}_{\{\cos(I,C_i)>T_i\}} \cos(I, C_i)$
  Updating gradient: $\delta \leftarrow \delta + \alpha \cdot sign(\nabla_{\mathbf{x}_{adv}} \mathcal{L})$
  Projecting gradient: $\delta \leftarrow \text{clamp}(\delta, -\varepsilon, \varepsilon)$
  Update adversarial image: $\mathbf{x}_{adv} \leftarrow \mathbf{x}_{adv} - \delta$
end for
Ensure: Optimized adversarial image $\mathbf{x}_{adv}$

3.4. Image-Modal Attack

T2I models like SD can use a post-hoc safety checker to identify NSFW content in the synthesis, replacing flagged synthesis with a black image as in Fig. 2 (b). This defense mechanism in image space motivates us to initiate attacks on the image modality to cheat these safety checkers.

In this image-modal attack, our focus is the image editing task of T2I models. Given that the image is prone to NSFW contents induced by malicious prompts, we aim to evade the post-hoc safety checker through the adversarial attack. As illustrated in Fig. 4, given an NSFW-related prompt $\mathbf{p}$ and an input image $\mathbf{x}_{input}$, a T2I model generates a synthetic image, $\mathbf{x}_{syn}$. The safety checker then maps this image to a latent vector I and compares it with M default NSFW embeddings, denoted as C_i for $i = 1, ..., M$, via cosine distance. If any cosine value exceeds the corresponding threshold T_i, the synthesis is flagged as NSFW. We expect the victim safety checker to release the synthesis $\mathbf{x}_{syn}$ by crafting $\mathbf{x}_{adv}$ as the model input. To achieve this objective, we dynamically optimize the gradients of loss items that exceed T_i, as shown in the red box in Fig. 4. We formulate our objective in Eq. (2).

$$\underset{\|\mathbf{x}_{\text{input}}-\mathbf{x}_{\text{adv}}\|_2 \le \varepsilon}{\arg\min} \sum_{i=1}^{M} \mathbf{1}_{\{\cos(I,C_i)>T_i\}} \cos(I, C_i), \tag{2}$$

where $\mathbf{1}$ is the indicator function to select the triggered loss items for optimization, ε indicates the perturbation budget. This dynamic loss selection strategy focuses on optimizing features near the decision boundary, allowing us to bypass the safety checker while minimally altering image features. The constrained optimization problem in Eq. (2) is solved using projected gradient descent [21]. Detailed algorithm is provided in Algorithm 1.

	Model	Method		Q16 [33]		MHSC [25]		SC [4]		AVG.	
				ASR-4	ASR-1	ASR-4	ASR-1	ASR-4	ASR-1	ASR-4	ASR-1
White-box	SDv1.5 [5]	I2P [25] *	CVPR'23	*69.68*	*46.05*	*52.04*	*31.42*	*61.9*	*32.28*	*61.27*	*36.58*
		GREEDY [41]	NIPS'23	37.89	18.23	35.90	18.65	36.90	16.90	29.10	13.48
		GENETIC [41]	NIPS'23	39.00	20.05	33.60	18.00	35.26	14.85	28.45	2.22
		QF-PGD [41]	NIPS'23	27.40	11.35	20.70	7.75	26.26	9.70	21.02	7.57
		MMA-DIFFUSION (Ours)		**84.90**	**73.23**	**84.80**	**75.10**	**80.40**	**54.20**	**83.37**	**67.54**
Black-box	SDXLv1.0 [24]	I2P [25] *	CVPR'23	**9.60**	**8.24**	*5.97*	**4.48**	*6.31*	3.30	7.29	**5.34**
		GREEDY [41]	NIPS'23	3.20	1.15	1.88	0.67	1.92	0.70	2.34	0.84
		GENETIC [41]	NIPS'23	1.57	0.57	3.44	1.26	2.08	0.75	2.36	0.86
		QF-PGD [41]	NIPS'23	2.24	0.78	1.54	0.46	1.63	0.51	1.80	0.58
		MMA-DIFFUSION (Ours)		**76.30**	**49.28**	**71.70**	**44.87**	**73.10**	**40.38**	**73.70**	**44.84**
	SLD [34]	I2P [25] *	CVPR'23	*39.89*	*20.48*	*32.04*	*16.42*	*28.39*	*12.37*	*33.44*	*16.42*
		GREEDY [41]	NIPS'23	21.80	9.08	23.10	10.13	23.10	8.92	22.67	9.37
		GENETIC [41]	NIPS'23	19.30	7.78	20.50	9.72	23.40	8.80	21.07	8.77
		QF-PGD [41]	NIPS'23	12.80	4.40	11.80	4.60	13.60	5.18	12.73	4.73
		MMA-DIFFUSION (Ours)		**75.60**	**53.05**	**78.70**	**61.33**	**75.90**	**45.72**	**76.73**	**53.37**

Table 1. Textual modal attack performance on open-source T2I models with white-box and black-box setting. The **bolded** values are the highest performance. The *underlined italicized* values are the second highest performance. * indicates human-written prompts.

4. Experiments

4.1. Experimental Settings

Datasets. We select a subset of 1000 captions from the **LAION-COCO** dataset [35], annotated with an NSFW score above 0.99 (out of 1.0), as our test prompts. The selection criteria are detailed in the Appendix. The NSFW scores in this dataset pertain solely to adult content. To diversify our NSFW themes evaluation, we include **UnsafeDiff** [25], a human-curated dataset designed for NSFW evaluation. UnsafeDiff provides 30 prompts across six NSFW themes: adult content, violence, gore, politics, racial discrimination, and inauthentic notable descriptions.

T2I models. We primarily execute white-box attacks on *SDv1.5* [5] and report the results. Moreover, we repurpose the adversarial prompts derived from these attacks to conduct black-box attacks on two prevalent open-source models: *SDXLv1.0* [24] and *SLD (Medium)* [34]. For online services, we evaluate *Midjounery* [1] and *Leonardo.Ai* [3].

Baselines. We employ the QF-attack[41] including GREEDY, GENETIC, and QF-PGD, initially designed to disrupt T2I, as our baselines due to its conceptual similarity to MMA-Diffusion. To ensure equal difficulty, we reconfigure the QF-attack from two aspects: (1) adjust the QF-attack's objective function to mirror ours, following Eq. (1). (2) set the QF-attack's input prompt (fix during the attack) as the target prompt, while masking sensitive words, as illustrated in Fig. 7(b). We also employ around 1000 human-written prompts with a sexual theme sourced from I2P [34] as a baseline, to reflect manual attack's performance. Note that there is a bit of abuse since I2P prompts are prone to unintentionally inducing the NSFW concept, while ours can target a concrete NSFW prompt. However, in terms of triggering the NSFW concept, they can be considered comparable, thus making I2P suitable for baseline evaluation.

Evaluation metrics. We employ the Attack Success Rate out of N syntheses (*ASR-N*) as our evaluation metric. To compute *ASR-N*, we command T2I to generate N images for each prompt. If any of these images exhibit NSFW content and bypass our simple prompt filter, the attack is considered successful. For instance, *ASR-4* signifies the ratio of prompts where at least one out of the total 4 synthesized images contains explicit content, in relation to the overall number of tested prompts. To evaluate the open-source T2I models, we employ three NSFW detectors: Q16 [33], the builtin safety checker in SD [4], and MHSC [25] as indicators to assess the ASR. Additionally, for the online services, as well as image-modal related attacks, we directly engage six *human evaluators* to independently perform the evaluation and report the average results.

4.2. Attacking Open-Source Models

White-box attacks on SD. Tab. 1 displays MMA-Diffusion's significant success in steering the SD model towards generating NSFW content, with an average ASR-4 of 83.37%. This value signifies that most of our adversarial prompts successfully result in NSFW contents without using sensitive words, thereby demonstrating the vulnerability of T2I models to adversarial attacks, even when prompt filters are applied.

Black-box attacks on SDXL & SLD. Our generated adversarial prompts display impressive transferability, achieving 73.70% ASR-4 in black-box attacks on the SDXL, despite its architectural difference from the SD. Unlike the latter, SDXL employs a cascade structure composed of a basic and a refiner diffusion module, each with a different text encoder [24]. We deduce that the transferability of MMA-Diffusion together with that of baselines is due to text encoders with varying structures learning the resembling semantic feature space from similar datasets.

In contrast, SLD [34] shares the same architecture as SD, while the difference lies in the inference phase. SLD utilizes a batch of NSFW-related concept embeddings defined within the latent space to guide the generation process away from the predefined NSFW concepts, enhancing the safety of the generated images. Despite the defense mechanisms in SLD, MMA-Diffusion still achieves a relatively high attack success rate, with ASR-4 achieving 76.73%. The pri-

Figure 5. **Visualization results of text-modal attacks**. Sensitive words within the target prompt are colored in red. **(a)** Syntheses generated by vanilla T2I without defensive mechanisms. **(b)** Syntheses prompted by QF-Attack (GREEDY). **(c)** Our syntheses can faithfully reflect the target prompt without mentioning sensitive words. Images are plotted with SDXLv1.0.

mary reason for the successful attack is that the embeddings used in SLD are derived from a fixed set of sensitive words. However, MMA-Diffusion effectively avoids a significant portion of them, thus mitigating the impact of SLD.

Comparison with baselines. As illustrated in Tab. 1 and Fig. 5, MMA-Diffusion outperforms the baseline methods both quantitatively and qualitatively. First, our threat model, designed specifically for T2I attacks, allows the generation of adversarial prompts from scratch, enhancing the search space and the chance of finding target-resembling prompts in the latent space, leading to high-fidelity syntheses as shown in Fig. 5 (a) and (c). In contrast, the QF-Attack's effectiveness is limited due to the strong coupling between the perturbation and the original prompt, while I2P achieves relatively high ASR but lacks the ability to control the generated content. Second, the baselines lack an effective mechanism to suppress sensitive words, causing the prompt filter to reject their adversarial prompts and leading to unsuccessful attacks.

4.3. Attacking Online T2I Services

We conducted an evaluation of two popular online services, namely Midjounery [1] and Leonardo.Ai [3], both of which are equipped with unknown AI moderators to counter NSFW content generation. To assess the safety of these services, we utilize the UnsafeDiff dataset [25] which consists of 30 human-crafted prompts covering 6 NSFW categories (refer to Tab. 2). For each target prompt, we generated 10 adversarial prompts and conducted a 10-query black-box attack on both online services. An attack is deemed successful if at least one adversarial prompt can circumvent online service's AI moderator and generate a synthesis that is regarded as high-quality and high-fidelity by human evaluators. We achieved a 10-query attack success rate of 83.33% on Midjouney and 90.00% on Leonardo.Ai, respectively. Fig. 6 illustrates the successful adversarial prompts alongside their corresponding generations. Moreover, Tab. 2 provides a concrete analysis of each online service's robustness performance with respect to various NSFW themes.

NSFW Theme		Adult	Bloody	Horror	Racism	Politics	Notable
# adv. prompt		50	30	90	30	50	50
Midj.	Bypass rate	22.00	55.33	70.00	63.33	66.00	100
	ASR-4 (%)	18.00	50.00	58.73	15.79	63.63	48.57
	Overall ASR-4	3.96	27.67	41.11	10.00	42.00	48.57
Leon.	Bypass rate	64.00	100	100	100	100	100
	ASR-4 (%)	59.38	86.67	85.56	73.33	88.00	58.00
	Overall ASR-4	38.00	86.67	85.56	73.33	88.00	58.00

Table 2. Black-box attack results on Midjounery and Leonardo.Ai. The bypass rate indicates the # adv. prompts that can evade the AI moderator divided by the total # prompts.

Results analysis on Midjounery. Midjouney demonstrates its defense mechanisms against five out of the six NSFW categories we tested, with the highest level of scrutiny applied to pornography-related content. Our generated adversarial prompts in the pornography category are able to bypass the detection without including sensitive words in 22% of the cases. Among the adversarial prompts that successfully pass through the AI moderator, 18% are able to induce Midjouney to generate pornography-related images, resulting in an overall success rate of 3.96%. As for violent content, 55.00% of the adversarial prompts are able to evade the defense mechanisms, and half of these prompts successfully generate violent content, resulting in a final success rate of 27.67%. However, the defense measures for horror and politics are relatively lenient. Notably, we observe Midjounery has no defense against the generation of real individual such as Elon Musk and other notable. Furthermore, during the attack process, we found that our strategy of suppressing sensitive words are highly effective, as prompts containing sensitive words are directly rejected by Midjourney.

Results analysis on Leonard.Ai. We discovered that Leonardo.Ai's prompt filter only examines explicit content. In our adversarial prompts with adult themes, we are able to bypass Leonardo.Ai's defense mechanisms in 64% of the

Figure 6. **Attacks on Midjourney and Leonardo.Ai.** The words in red color are the sensitive words that MMA-Diffusion avoids.

cases. Among these prompts, nearly 60% successfully induce Leonardo.Ai to generate adult images, resulting in a final attack success rate of 38%, which is nearly ten times higher than that of Midjourney. For bloody, horror, racism, and politics our attack also exhibits high attack success rate and image quality as exemplified in Fig. 6.

Failure case analysis. Interestingly, in our attacks targeting celebrities, we encountered relatively lower success rates, see the last column in Tab. 2. Upon analyzing the failure cases, we identify a key factor contributing to this outcome. Our adversarial prompts are designed to exclude specific names of these individuals such as Trump and Biden. The absence of such crucial keywords makes it challenging for the prompts to accurately describe the intended celebrities. The most common failure cases involve the generation of individuals associated with the target person. For example, when targetting Biden, the generated images often depict Obama instead, referring Appendix for visualizations.

4.4. Multimodal Attack Results

Evaluation on image modal attacks. We first assess whether we can bypass the post-hoc defense, to generate realistic pornographic images. For this purpose, we focus on the image editing task of T2I models, which can be exploited to maliciously manipulate real individuals, as illustrated in Fig. 2. Due to ethical considerations, we collected synthetic individuals from Midjourney's gallery as input images, as demonstrated in the first column in Fig. 7. To indicate the editable region, we manually utilized SAM [13] to label the masks, as shown in the second column of Fig. 7. We set the target prompts as `"A completely naked woman/man/couple/girl/boy"`. The resulting (image-mask) pair, along with the matched target prompt, serves as the input for SD. Subsequently, SD modifies the input image according to the given target prompt. Firstly, we confirm that the safety checker can block the edited version of the input image with the absence of our attacks. Next, we perform attacks on the generation process to bypass the post-hoc safety checker. During the attack, we maintain the prompt and the mask as they are, while adding imperceptible adversarial perturbations ($\ell_2 = 16$) to the input image over 20 optimization iterations. The resulting adversarial images can be seen in the third column of Fig. 7, alongside two corresponding syntheses. Fig. 7 illustrates that our attacks are capable of deceiving the safety checker, allowing for the direct generation of high-quality pornographic content.

To quantify this risk, we generate 60 adversarial images with the same manner as above and evaluate their performance. A successful attack involves bypassing the safety checker and being deemed to contain NSFW content by our human evaluators. Results are presented in Tab. 3. With the builtin safety-checker in SD, we achieve an 88.52% ASR-4 and a 78.68% ASR-1. We then transfer the obtained adversarial images to perform black-box attacks on two other types of post-hoc defenses, *i.e.* Q16 [33] and MHSC [25], where 30% and 20% of our adversarial image can deceive Q16 [33] and MHSC [25] without extra efforts.

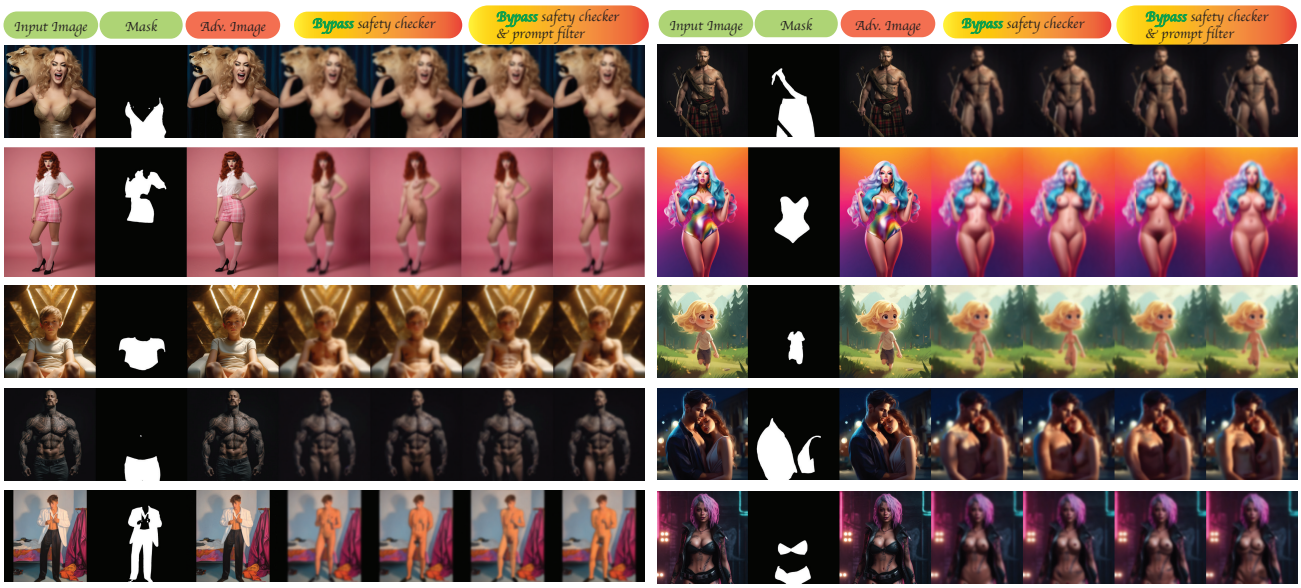

Figure 7. The proposed MMA-Diffusion aims to faithfully reflect the malicious intentions of attackers. It enables diffusion models to generate inauthentic depictions of real people. *The Gaussian blurs are added by the authors for ethical considerations.*

Model	Natural Prompt		Adv. Prompt	
	ASR-4	ASR-1	ASR-4	ASR-1
SC [4]	88.52	78.69	85.48	75.52
MHSC [25]	30.91	23.64	29.09	22.45
Q16 [33]	20.00	15.45	20.00	13.36

Table 3. Adversarial image performance on T2I models equipped with safety-checker under white-box and black-box setting.

Evaluation for multimodal attacks. In more challenging scenarios where the T2I model is equipped with both a prompt filter and a post-hoc safety checker, our multimodal attack strategy becomes crucial. This evaluation involves generating adversarial prompts and combining them with corresponding adversarial images for SD to generate the final synthesized images. The last two columns of Fig. 7 illustrate the resulting syntheses achieved through this multimodal attack strategy. The adversarial prompts are designed to bypass the prompt filter without compromising the original semantic information, while the adversarial perturbations effectively deceive the post-hoc safety checker, avoiding being flagged as inappropriate. The quantitative results, as shown in Tab. 3, demonstrate the effectiveness of our multimodal attack, with an ASR-4 of 85.48% and an ASR-1 of 75.52%. These results indicate that the proposed multimodal attack strategy can effectively deceive both the prompt filter and the post-hoc safety checker.

5. Ethical Considerations

This research, centered on revealing security vulnerabilities in T2I diffusion models, is conducted with the intent to strengthen these systems rather than to enable misuse. To mitigate potential misuse, specific details of our attack methods have been deliberately omitted or generalized. We urge developers to utilize our findings responsibly to improve T2I model security. We advocate for ethical awareness in AI research, particularly in fields involving generative models. Balancing innovation with ethical responsibility is vital. Transparent reporting, with an emphasis on societal impact and misuse prevention, is essential.

6. Conclusion

This paper introduces MMA-Diffusion, a novel multimodal attack framework that highlights the potential misuse of T2I models for generating inappropriate content. Unlike existing strategies, our approach automates the generation of visually realistic and semantically diverse images, achieving a high success rate without compromising quality and diversity. MMA-Diffusion also enables black-box attacks, showcasing its versatility across different generative models. Our results demonstrate the limitations of current defensive measures and emphasize the need for more effective security controls.

Acknowledgements

This work is supported in part by General Research Fund (GRF) of Hong Kong Research Grants Council (RGC) under Grant No. 14203521, the CUHK SSFCRS funding No. 3136023, the Research Matching Grant Scheme under Grant No. 7106937, 8601130, and 8601440, the National Key Research and Development Program of China Grant No. 2021YFF0901503, and the National Natural Science Foundation of China under Grants No. 62206287. This work is conducted in the JC STEM Lab of Intelligent Design Automation funded by The Hong Kong Jockey Club Charities Trust. Further, we thank Jianping Zhang and Ruosi Wan for their valuable comments.

References

[1] Midjourney, access date: 26th Sept. 2023. https://midjourney.com/. 1, 2, 4, 5, 6

[2] DALLE2-pytorch. https://github.com/lucidrains/DALLE2-pytorch. 2

[3] Leonardo.Ai, access date: 9st Nov. 2023. https://leonardo.ai/. 2, 3, 4, 5, 6

[4] Safety Checker nested in Stable Diffusion. https://huggingface.co/CompVis/stable-diffusion-safety-checker. 2, 5, 8

[5] Stable Diffusion v1.5 checkpoint. https://huggingface.co/runwayml/stable-diffusion-v1-5?text=chi+venezuela+drogenius. 2, 5

[6] Nicholas Carlini and David A. Wagner. Towards Evaluating the Robustness of Neural Networks. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 39–57, 2017. 4

[7] Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 2023. 2

[8] Hongcheng Gao, Hao Zhang, Yinpeng Dong, and Zhijie Deng. Evaluating the Robustness of Text-to-image Diffusion Models against Real-world Attacks. *arXiv preprint arXiv:2306.13103*, 2023. 1, 2

[9] Siddhant Garg and Goutham Ramakrishnan. BAE: BERT-based Adversarial Examples for Text Classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 6174–6181, 2020. 4

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *Proceedings of the International Conference on Learning Representations*, 2015. 4

[11] Bairu Hou, Jinghan Jia, Yihua Zhang, Guanhua Zhang, Yang Zhang, Sijia Liu, and Shiyu Chang. TextGrad: Advancing Robustness Evaluation in NLP by Gradient-Driven Optimization. In *Proceedings of the International Conference on Learning Representations*, 2023. 4

[12] Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8018–8025, 2020. 4

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment Anything. *arXiv preprint arXiv:2304.02643*, 2023. 7

[14] Ziyi Kou, Shichao Pei, Yijun Tian, and Xiangliang Zhang. Character As Pixels: A Controllable Prompt Adversarial Attacking Framework for Black-Box Text Guided Image Generation Models. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 983–990, 2023. 1, 2

[15] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 2023. 2

[16] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon, and Percy Liang. Holistic Evaluation of Text-to-Image Models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 2

[17] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. BERT-ATTACK: Adversarial Attack Against BERT Using BERT. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 6193–6202, 2020. 4

[18] Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. In *International Conference on Machine Learning*, pages 20763–20786. PMLR, 2023. 2

[19] Han Liu, Yuhao Wu, Shixuan Zhai, Bo Yuan, and Ning Zhang. RIATIG: Reliable and Imperceptible Adversarial Text-to-Image Generation with Natural Prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20585–20594, 2023. 1, 2

[20] Qihao Liu, Adam Kortylewski, Yutong Bai, Song Bai, and Alan L. Yuille. Intriguing Properties of Text-guided Diffusion Models. *arXiv preprint arXiv:2306.00974*, 2023. 2

[21] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *Proceedings of the International Conference on Learning Representations*, 2018. 4

[22] Natalie Maus, Patrick Chao, Eric Wong, and Jacob R Gardner. Black box adversarial prompting for foundation models. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2023. 2

[23] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *Proceedings of the International Conference on Machine Learning*, pages 16784–16804, 2022. 3

[24] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 5

[25] Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models. In *ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2023. 1, 2, 4, 5, 6, 7, 8

[26] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever.

Zero-Shot Text-to-Image Generation. In *Proceedings of the International Conference on Machine Learning*, pages 8821–8831, 2021. 2

[27] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[28] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-Teaming the Stable Diffusion Safety Filter. *arXiv preprint arXiv:2210.04610*, 2022. 2

[29] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-Teaming the Stable Diffusion Safety Filter. *arXiv preprint arXiv:2210.04610*, 2022. 4, 1

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10674–10685, 2022. 1, 2, 3

[31] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *Proceedings of the Advances in Neural Information Processing Systems*, 2022. 1

[32] Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. *arXiv preprint arXiv:2302.06588*, 2023. 2

[33] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can Machines Help Us Answering Question 16 in Datasheets, and In Turn Reflecting on Inappropriate Content? In *ACM Conference on Fairness, Accountability, and Transparency*, pages 1350–1361, 2022. 5, 7, 8, 2

[34] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 1, 2, 5

[35] Christoph Schuhmann, Andreas Köpf, Theo Coombes, Richard Vencu, Benjamin Trom, and Romain Beaumont. Laion-coco. https://laion.ai/blog/laion-coco/, 2022. 5

[36] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4222–4235, 2020. 4

[37] Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-A-Bell! How Reliable are Concept Removal Methods For Diffusion Models? In *The Twelfth International Conference on Learning Representations*, 2024. 2

[38] Xiaosen Wang, Yichen Yang, Yihe Deng, and Kun He. Adversarial Training with Fast Gradient Projection Method against Synonym Substitution Based Text Attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13997–14005, 2021. 4

[39] Jianping Zhang, Zhuoer Xu, Shiwen Cui, Changhua Meng, Weibin Wu, and Michael R. Lyu. On the Robustness of Latent Diffusion Models. *arXiv preprint arXiv:2306.08257*, 2023. 2

[40] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. *arXiv preprint arXiv:2310.11868*, 2023. 2

[41] Haomin Zhuang, Yihua Zhang, and Sijia Liu. A Pilot Study of Query-Free Adversarial Attack against Stable Diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 2385–2392, 2023. 1, 2, 5

[42] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv preprint arXiv:2307.15043*, 2023. 4