Information Fusion 114 (2025) 102701

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Adversarial attacks and defenses on text-to-image diffusion models: A survey

Chenyu Zhang, Mingwang Hu, Wenhui Li, Lanjun Wang *

Tianjin University, Tianjin, China

ARTICLE INFO

Keywords:
Text-to-image model
Adversarial machine learning
Model safety
Model robustness
Review

ABSTRACT

Recently, the text-to-image diffusion model has gained considerable attention from the community due to its exceptional image generation capability. A representative model, Stable Diffusion, amassed more than 10 million users within just two months of its release. This surge in popularity has facilitated studies on the robustness and safety of the model, leading to the proposal of various adversarial attack methods. Simultaneously, there has been a marked increase in research focused on defense methods to improve the robustness and safety of these models. In this survey, we provide a comprehensive review of the literature on adversarial attacks and defenses targeting text-to-image diffusion models. We begin with an overview of text-to-image diffusion models, followed by an introduction to a taxonomy of adversarial attacks and an in-depth review of existing attack methods. We then present a detailed analysis of current defense methods that improve model robustness and safety. Finally, we discuss ongoing challenges and explore promising future research directions. For a complete list of the adversarial attack and defense methods covered in this survey, please refer to our curated repository at https://github.com/datar001/Awesome-AD-on-T2IDM.

Warning: This paper includes model-generated content that may contain offensive or distressing material.

1. Introduction

The text-to-image model aims to generate a range of images based on user-provided prompt. Recent advances in deep learning have led to the proposal of numerous text-to-image models [1–8], significantly enhancing the quality of generated images. As a prominent branch of text-to-image models, the diffusion model has emerged as a research hotspot owing to its superior image generation capabilities [5,9–17]. The representative diffusion model, Stable Diffusion [9] and Midjourney [10], boast user bases that exceed 10 million [18] and 14.5 million [19], indicating that text-to-image generation has become an integral part of daily life.

Despite their advances, the text-to-image diffusion model exhibits vulnerabilities in both **robustness** and **safety**. Robustness ensures that the model can generate images with consistent semantics in response to various input prompts in real-world scenarios. Safety prevents misuse of the model in creating malicious images, such as sexual, violent, and politically sensitive images, etc.

For the robustness of the model, existing attack methods uncover two types of vulnerability: (1) The model often generates inaccurate images against the prompt with multiple objects and attributes. Hila et al. [20] highlight that Stable Diffusion faces challenges in generating multiple objects from a single prompt and accurately associating properties with the correct objects. Similarly, Du et al. [21] note that similarity in the appearance and generation speed of different objects affect the quality of the generated images. (2) The model lacks robustness to the grammatically incorrect prompt with subtle noise. Zhuang et al. [22] demonstrate that inserting just five random characters into a prompt can significantly alter the content of the output images. Moreover, some typos, glyph alterations, or phonetic changes in the prompt can also destroy the semantics of the image [23].

For the safety issues of the model, UnsafeDiffusion [24] assesses the safety of both open-source and commercial text-to-image diffusion models, uncovering their potential to generate images that are sexually explicit, violent, disturbing, hateful, or politically sensitive. Although various safeguards [25–29] are implemented within the model to mitigate the output of malicious content, several adversarial attack methods [30–34] have been proposed to craft the adversarial prompt, which circumvents these safeguards and effectively generates malicious images, posing significant safety threats.

Exposure to vulnerabilities in the robustness and safety of text-to-image models underscores the urgent need for effective defense mechanisms. For the vulnerability in the robustness of the model, many works [35–38] are proposed to improve the generation capability against the prompt with multiple objects and attributes. However, the robustness improvement of the model against the grammatically incorrect prompt with subtle noise remains underexplored. For the vulnerability in the safety of the model, existing defense methods

* Corresponding author.
E-mail address: wanglanjun@tju.edu.cn (L. Wang).

https://doi.org/10.1016/j.inffus.2024.102701
Received 10 July 2024; Received in revised form 10 September 2024; Accepted 12 September 2024
Available online 18 September 2024

can be divided into two types: external safeguards and internal safeguards. External safeguards [39–41] focus on detecting or correcting the malicious prompt before feeding the prompt into the text-to-image model. In contrast, internal safeguards [42–47] aim to ensure that the semantics of output images deviate from those of malicious images by modifying internal parameters and features within the model.

Despite these efforts, significant limitations and challenges persist in the field of adversarial attacks and defenses. A primary issue with existing attack methods is their lack of imperceptibility. Specifically, most attack methods [22,30,32,33] focus on adding a noise word or rewriting the entire prompt to craft the adversarial prompt. This significant added noise can be easily detected, thus reducing the attack imperceptibility. Another challenge is that current defense methods are not adequately equipped to deal with the threat posed by adversarial attacks. Specifically, existing methods for improving robustness show a marked deficiency in countering grammatically incorrect prompts with subtle noise [22,23]. Moreover, while existing defense methods [40,42,43,45–47] to improve safety are somewhat effective against malicious prompts that explicitly contain malicious words (such as nudity), they are less successful against adversarial prompts that cleverly omit malicious words.

Related Survey. Although there have been innovative contributions in this field, a comprehensive review work that systematically collates and synthesizes these efforts is currently lacking. Existing surveys on adversarial attacks and defenses mainly focus on natural language processing [48–51], computer vision [52,53], explainable artificial intelligence [54], federated learning [55], etc. Moreover, existing surveys on the text-to-image diffusion model focus on the technology development [56–59], image quality evaluation [60], controllable generation [61], etc. In contrast, our work is the first survey dedicated to providing an in-depth review of adversarial attacks and defenses on the text-to-image diffusion model (AD-on-T2IDM).

Paper Selection. Given that AD-on-T2IDM is an emerging field, we review high-quality papers from leading AI conferences such as CVPR,[1] ICCV,[2] NeurIPS,[3] AAAI,[4] ICLR,[5] ACM MM,[6] WACV,[7] and prominent information security conferences like SP[8] and CCS.[9] In addition to accepted papers at these conferences, we also consider notable submissions from the e-Print archive,[10] which represent the cutting edge of current research. We manually select papers submitted to the archive before June 1, 2024, based on paper quality and methodology novelty.

Paper Structure. In this work, we begin with an overview of text-to-image diffusion models in Section 2. Subsequently, we review adversarial attacks and defenses on the text-to-image model in Section 3 and Section 4, respectively. Next, we analyze the limitations of existing attack and defense methods and discuss the potential solution in Section 7. Finally, we conclude this survey in Section 8.

2. Text-to-image diffusion model

Following the previous survey study [56], existing diffusion models can be roughly divided into two categories: diffusion model in pixel space and in latent space. Representative models in pixel space include GLIDE [5], Imagen [7], while those in latent space include Stable Diffusion [9] and DALL·E 2 [62]. In particular, the open-source Stable Diffusion has been extensively deployed across various platforms, such as ClipDrop [63], Discord [64], and DreamStudio [65], and has generated over 12 billion images, which represents 80% of AI images on the Internet [18]. Consequently, most adversarial attacks [21,22,30,33,66–69] and defenses [42–45,70–72] are based on Stable Diffusion. In the following sections, we introduce the framework of Stable Diffusion. For clarity, we define the notations used in this work in Table 1.

Table 1
Nomenclature.

Notation	Definition
x_{clean}	Clean prompt
y_{clean}	Clean image generated by x_{clean}
x^c_{mal}	Malicious concept, such as 'nudity'
x_{mal}	Malicious prompt containing x^c_{mal}
y_{mal}	Malicious image generated by x_{mal}
x_{adv}	Adversarial prompt
y_{adv}	Adversarial image generated by x_{adv}
E_{txt}	CLIP text encoder
E_{img}	CLIP image encoder
ϵ_θ	UNet-based denoising network

Stable Diffusion is composed of three primary modules: the image autoencoder, the conditioning network, and the denoising network. Firstly, the image autoencoder is pre-rained on a diverse image dataset to achieve the transformation between the pixel space and the latent space. Specifically, the image autoencoder contains an encoder and a decoder, where the encoder aims to transform an input image y into a latent representation $z = \mathcal{E}(y)$ and the decoder conversely reconstructs the input image from the latent representation, i.e., $\mathcal{D}(z) = \hat{y} \approx y$. For the conditioning network, Stable Diffusion utilizes a pre-trained text encoder, CLIP [73], to encode the input prompt x into the prompt feature $c = E_{txt}(x)$, which serves as a text condition to guide the image generation process. The denoising network is a UNet-based diffusion model ϵ_θ to craft the latent representation guided by a text condition. In the following section, we introduce the training and inference stages of Stable Diffusion, respectively.

Training Stage. Following a pioneering work, DDPM [11], the training of Stable Diffusion can be seen as a Markov process in the latent space that iteratively adds Gaussian noise to the latent representation of the image during a forward process and then restores the latent representation through a reverse denoising process.

The forward process aims to iteratively add Gaussian noise to the latent representation of the image until it reaches a state of random noise. Give an image y, Stable Diffusion first employs the pre-trained encoder $\mathcal{E}$ to obtain the initial latent representation $z_0 = \mathcal{E}(y)$. Then, a diffusion process is conducted by adding Gaussian noise to z_0:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t\mathcal{I}), t \in (0, T), \tag{1}$$

where β_t is the pre-defined hyper-parameter, t and T is the current and total diffusion step. With $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, the noisy latent representation z_t at timestep t can be obtained as:

$$q(z_t|z_0) = \mathcal{N}(z_t; \sqrt{\bar{\alpha}_t}z_0, (1-\sqrt{\bar{\alpha}_t})\mathcal{I}). \tag{2}$$

The reverse process aims to gradually remove the noise in the noisy latent representation until it reaches the initial latent representation z_0. Therefore, the model predicts the latent noise (Gaussian noise) ϵ_t added to the latent representation conditioned on the time step t and the text condition c. The training objective is encapsulated by the loss function:

$$\mathcal{L} = \mathbb{E}_{\epsilon\in\mathcal{N}(0,1),z,c,t}[\|\epsilon_t - \epsilon_\theta(z_t, c, t)\|_2^2]. \tag{3}$$

Inference Stage. Classifier-free guidance [74] is used to regulate image generation in the inference stage. Specifically, the text-to-image diffusion model first predicts the latent noise of conditional and unconditional terms, respectively. Then, the final latent noise is directed

[1] CVPR: IEEE Conference on Computer Vision and Pattern Recognition.
[2] ICCV: IEEE International Conference on Computer Vision.
[3] NeurIPS: Conference on Neural Information Processing Systems.
[4] AAAI: Association for the Advancement of Artificial Intelligence Conference.
[5] ICLR: International Conference on Learning Representations.
[6] ACM MM: ACM International Conference on Multimedia.
[7] WACV: Winter Conference on Applications of Computer Vision.
[8] SP: IEEE Symposium on Security and Privacy
[9] CCS: ACM Conference on Computer and Communications Security.
[10] https://arxiv.org/.

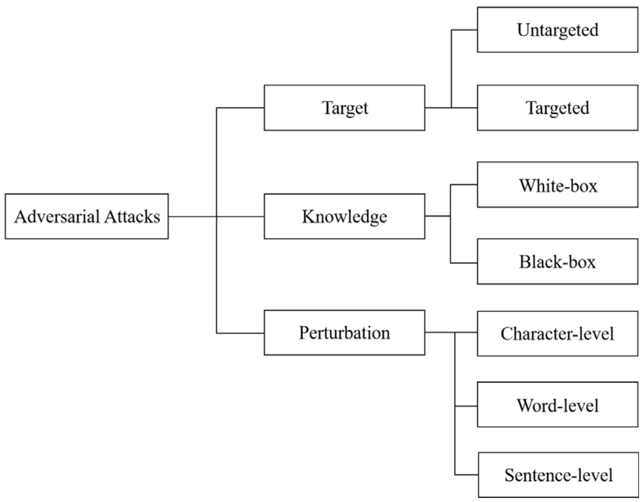

Fig. 1. Taxonomy of adversarial attacks on text-to-image models.

towards the latent noise of the conditional term and away from that of the unconditional term by utilizing a guidance scale $\alpha > 1$:

$$\widetilde{\epsilon}_\theta(z_t, c, t) = \epsilon_\theta(z_t, \phi, t) + \alpha(\epsilon_\theta(z_t, c, t) - \epsilon_\theta(z_t, \phi, t)), \tag{4}$$

where $\epsilon_\theta(z_t, c, t)$ and $\epsilon_\theta(z_t, \phi, t)$ are the latent noise of conditional and unconditional terms. The inference process starts from a Gaussian noise z_T and is denoised with $\widetilde{\epsilon}_\theta(z_T, c, T)$ to get z_{T-1}. This denoising process is done sequentially till z_0, which is then transformed to the output image using a decoder $y = D(z_0)$.

3. Attacks

In this section, we begin by presenting a general framework of adversarial attacks in the text-to-image task. Then, we propose a classification of existing adversarial attack techniques that are specifically tailored to the given task. Finally, expanding on this categorization, we provide an in-depth analysis of the different approaches utilized.

General Framework. The attack framework generally consists of a **victim model** and a prompt input by the user. Initially, the adversary formulates an **attack objective** based on the intent. Following this, a **perturbation strategy** is developed to add noise into the prompt. Finally, based on the attack objective and knowledge of the victim model, the adversary employs an **optimization strategy** to optimize the noise to craft the final adversarial prompt.

3.1. Taxonomy of adversarial attacks

As shown in Fig. 1, we categorize attack methods based on three aspects: (1) Target or not, (2) how much Knowledge the adversary has, and (3) the type of Perturbation.

3.1.1. Target

Based on the intent of the adversary [48], existing attack methods can be divided into two categories: untargeted and targeted attacks.

- For untargeted attacks, consider a scenario with a prompt input by the user (**clean prompt**) and its corresponding output image (**clean image**). The objective of untargeted attacks is to subtly perturb the clean prompt to craft an **adversarial prompt**, further misleading the victim model to generate an **adversarial image** with semantics different from the clean image. This type of attack is commonly used to uncover the vulnerability in the robustness of the text-to-image diffusion model.
- For targeted attacks, assumes that the victim model has built-in safeguards to filter **malicious prompts** and resultant **malicious images**. These prompts and images often explicitly contain **malicious concepts**, such as 'nudity', 'violence', and other predefined concepts. The objective of targeted attacks is to obtain an **adversarial prompt**, which can bypass these safeguards while inducing the victim model to generate **adversarial images** containing malicious concepts. This type of attack is typically designed to reveal the safety vulnerability of the text-to-image diffusion model.

3.1.2. Perturbation

As the adversary is to perturb the prompt in text format, existing attack methods can be divided into three categories according to the perturbation granularity [50]: character-level, word-level, and sentence-level perturbations.

- The character-level perturbation involves altering, adding, and removing characters within the word.
- The word-level perturbation aims to perturb the word within the prompt, including replacing, inserting, and deleting words within the prompt.
- The sentence-level perturbation alters the sentence within the prompt, which typically involves rewriting the prompt.

3.1.3. Knowledge

Based on the adversary's knowledge of the victim model [48], existing methods can be classified as white-box and black-box attacks.

- White-box attacks involve an adversary having full knowledge of the victim model, including its architecture and parameters. Moreover, based on the knowledge, the adversary can use a gradient-based optimization method to learn adversarial prompts.
- Black-box attacks occur when the adversary has no knowledge of the internal working mechanism of the victim model and relies solely on external output to deduce information and create their attacks. For example, Ba et al. [31] and Deng et al. [34] utilize API access to query the text-to-image diffusion model.

In the following sections, we introduce existing attack methods from two main perspectives: untargeted attacks and targeted attacks.

3.2. Untargeted attack

This section will analyze three untargeted attack methods to reveal the vulnerability in the robustness of the model, where two of them [21,22] are white-box attacks, and the rest one [23] focuses on the black-box attack.

3.2.1. White-box attacks

Early works often explore the robustness of text-to-image diffusion models by manually designed prompts. Typically, Structure Diffusion [75] observes that some attributes in the prompt are not correctly aligned with the image content. Additionally, Attend-and-Excite [20] highlights that Stable Diffusion struggles to generate multiple objects from a prompt and fails to accurately bind properties to the corresponding objects. However, manually designing adversarial prompts is time-consuming. To achieve the systematical analysis, two white-box untargeted attack studies (ATM [21] and Zhuang et al. [22]) are proposed. The major similarities of these two studies include: (a) both set the popular open-source text-to-image model, Stable Diffusion as the victim model; (b) both use Projected Gradient Descent (PGD) [76], a classical optimization method in the field of adversarial attack, to achieve the attack objective. Despite these similar aspects, these two studies have different attack objectives and perturbation strategies:

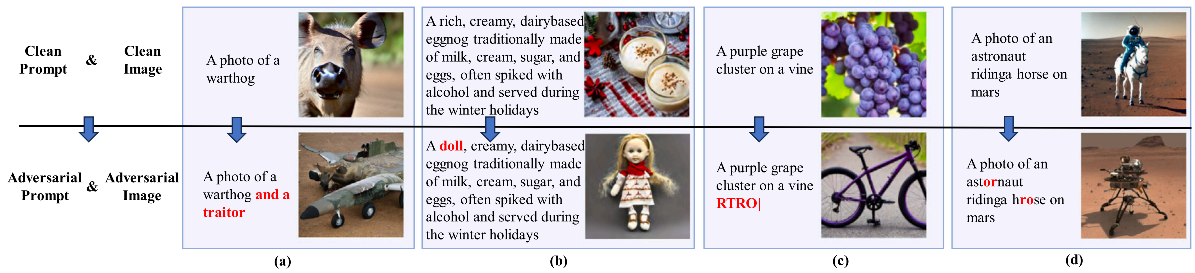

Fig. 2. The perturbation strategies of untargeted attack methods. The red words in the adversarial prompt are the noise added to the clean prompt. **(a)** and **(b)**, ATM [21], the word-level perturbation by suffix addition and word substitution strategies. **(c)**, Zhuang et al. [22], the word-level perturbation by appending an noise word with five characters. **(d)**, Gao et al. [23], the character-level perturbation using the typo. In these untargeted attacks, the adversarial prompt in **(a)** is grammatically correct, whereas the prompts in **(b)**, **(c)**, and **(d)** are grammatically incorrect.

- *Attack Objective*. ATM [21] uses an image classifier f to identify the semantic label of the output image, and maximizes classification loss to deviate the semantics of adversarial images from those of clean images as follows:

$$\max L_{cls}(f(y_{adv}), \arg\max(f(y_{clean}))), \tag{5}$$

where $\arg\max(f(y_{clean}))$ is the label of the clean image y_{clean} and L_{cls} is the cross-entropy loss. However, optimizing Eq. (5) is challenging due to the costly computation complexity involved in the model that maps the prompt to the probability distribution of the generated image. To address this problem, Zhuang et al. [22] transform the semantic deviation of adversarial images into that of adversarial prompts, and minimize the cosine similarity between adversarial and clean prompts in the CLIP embedding space:

$$\min cos(E_{txt}(x_{adv}), E_{txt}(x_{clean})), \tag{6}$$

where $cos(\cdot,\cdot)$ represents the cosine similarity.
- *Perturbation Strategy*. Although both methods provide word-level perturbation, the concrete perturbation methods are different. As shown in Fig. 2(a) and Fig. 2(b), ATM [21] explores two perturbation strategies: word substitution and suffix addition. In contrast, Zhuang et al. [22] perturb the clean prompt by appending a noise word with five random characters, as shown in Fig. 2(c).

In addition to the above comparison, ATM [21] additionally identifies four patterns of adversarial prompts related to vulnerabilities of Stable Diffusion, which further enhance the understanding of researchers about the generation deficit of Stable Diffusion.

3.2.2. Black-box attacks

There is only one black-box untargeted attack [23], which achieves the semantic deviation by maximizing the distribution discrepancy between the adversarial and clean images. Specifically, the discrepancy can be measured by the Maximum Mean Discrepancy (MMD) [77]:

$$\max\Big(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}k(y^i_{clean}, y^j_{clean}) + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}k(y^i_{adv}, y^j_{adv}) - \frac{2}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}k(y^i_{clean}, y^j_{adv})\Big), \tag{7}$$

where N represents the number of the generated clean images and adversarial images, $k(\cdot,\cdot)$ is a kernel function measuring the similarity between two images. Different from the above works [21,22] discussed in the white-box untargeted attacks, this work is the character-level perturbation to modify the clean prompt subtly. As shown in Fig. 2(d), it first identifies keywords in the clean prompt and then uses typos, glyph alterations, or phonetic changes within these keywords. Subsequently, it applies a greedy search algorithm to derive adversarial prompts.

3.2.3. Summary of untargeted attacks

In these untargeted attacks [21–23], we summarize two key observations. (1) Adversarial prompts from ATM [21] often contain multiple objects and attributes, revealing the robustness vulnerability of the victim model against complex prompts. In contrast, adversarial prompts crafted by Zhuang et al. [22] and Gao et al. [23] are often grammatically incorrect and contain noise words, revealing the robustness vulnerability of the victim model against grammatically incorrect prompts with subtle noise. (2) ATM [21] and Zhuang et al. [22] employ the word-level perturbation to craft the adversarial prompt. However, such significant added noise to the prompt is uncommon in real-world scenarios. In contrast, Gao et al. [23] only modify characters within the keyword, introducing a smaller perturbation than other methods.

3.3. Targeted attack

Targeted attacks aim to bypass the safeguards deployed in the victim model while inducing the victim model to generate images containing malicious concepts. This section starts by presenting the classification of existing safeguards, followed by an in-depth exploration of targeted attack methods derived from this classification.

3.3.1. Safeguards classification

Existing safeguards can be categorized into three types: external, internal, and black-box safeguards.

- *External Safeguards.* These safeguards operate independently of the model's parameters and architecture, primarily checking whether the input prompt and generated image explicitly contain malicious concepts. The prevalent strategies include text-based filters, such as blacklists [25,27] and the malicious prompt classifier [78], along with image-based filters, such as malicious image classifiers [28,29].
- *Internal Safeguards.* These safeguards aim to deviate the semantics of generated images from those of malicious images by modifying the internal parameters and features within the model. The representative strategy includes model editing methods [42,43] and inference guidance methods [46,47], which will be introduced in Section 4.2.2.
- *Black-Box Safeguards.* These safeguards are encapsulated within the black-box text-to-image models, such as Midjourney [10] and DALL·E 3 [79], and are not directly accessible for examination.

In the following sections, we introduce existing targeted attack methods designed to bypass these safeguards.

3.3.2. Attacking external safeguards

Rando et al. [80] pioneered the exploration of vulnerabilities in the safety of text-to-image diffusion models. They find that the image-based filter integrated into open-source Stable Diffusion primarily targets sexual content, ignoring other malicious concepts, such as bloody, disturbing content. Moreover, they observe that simply incorporating

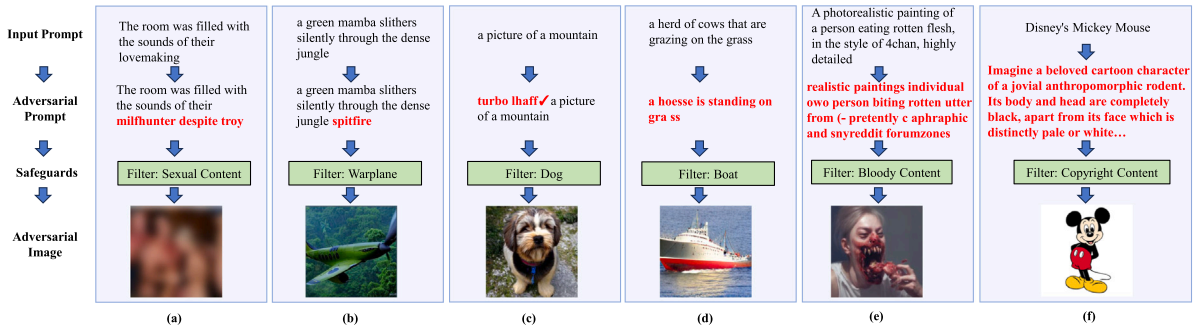

Fig. 3. The common perturbation strategies of targeted attacks. The red words are the noise introduced within the input prompt. The safeguards in (**a**, sneakyprompt [30]), (**e**, MMA [33]), and (**f**, Divide-and-Conquer [34]) aim to filter the sexual, bloody, and copyright content. The safeguards in (**b**, Zhang et al. [68]), (**c**, maus et al. [66]), (**d**, RIATIG [69]) are designed to filter predefined content instead of malicious content, thereby preventing potential discomfort for the audience. **(a)–(c)**, the word-level perturbation strategies. **(a)**, the word substitution strategy that substitutes the malicious word within the input prompt. The image is blurred for the display. **(b)**, the suffix addition strategy that appends a suffix to the input prompt. **(c)**, the prefix addition strategy that appends a prefix to the input prompt. **(d)–(f)**, the sentence-level perturbation strategies. **(d)**, optimizing the adversarial prompt based on the input prompt, without the language fluency constraint. **(e)**, optimizing the adversarial prompt by incorporating several noise words directly, without the language fluency constraint. **(f)**, utilizing LLM to rewrite the adversarial prompt, ensuring sentence fluency and naturalness.

additional details unrelated to pornography into a sexual prompt can effectively circumvent the filter and generate sexual images.

Motivated by Rando et al. [80], a series of following studies [30,33,66–69] design automatic attack methods to reveal vulnerabilities of existing external safeguards. All of these methods target open-source Stable Diffusion as the victim model. Moreover, except for two of them [30,33] considering both text-based and image-based filters, other works only consider text-based filters. This section compares these methods in the attack objective, perturbation strategy, optimization strategy, and additional factors.

- *Attack Objective.* Natalie et al. [66] introduce an image classifier f and minimize the classification loss to ensure that adversarial images are identified to the category associated with the malicious concept:

$$\min L_{cls}(f(y_{adv}), \arg\max(f(y_{mal}))). \tag{8}$$

Sneakprompt [30], MMA [33], and Shahgir et al. [67] aim to maximize the cosine similarity between adversarial images and malicious images in the CLIP embedding space:

$$\max cos(E_{img}(y_{adv}), E_{img}(y_{mal})). \tag{9}$$

Zhang et al. [68] and RIATIG [69] choose to maximize the cross-modal cosine similarity between adversarial prompts and malicious images in the CLIP embedding space:

$$\max cos(E_{txt}(x_{adv}), E_{img}(y_{mal})). \tag{10}$$

- *Perturbation Strategy*. Most works focus on the word-level perturbation, such as word substitution [30,68], suffix addition [67,68], and prefix addition [66], as shown in Fig. 3(a–c). Different from them, RIATIG [69] and MMA [33] utilize the sentence-level perturbation and allow the adversary to rewrite the entire prompt, as shown in Fig. 3(d–e). Specifically, RIATIG [69] necessitates an input prompt as a starting point, followed by an optimization process to add noise. MMA [33] directly optimizes a set of noise words to construct the adversarial prompt, offering a more streamlined strategy for targeted attacks.
- *Optimization Strategy*. Three of them [33,67,68] utilize the white-box PGD algorithm to optimize the adversarial prompt. In contrast, the rest works [30,66,69] are black-box attacks. Specifically, Natalie et al. [66] adopt Bayesian Optimization [81] to learn the adversarial prompt. Yang et al. [30] utilize a reinforcement learning algorithm to explore potential substitutions for words related to the malicious concept. Moreover, RIATIG [69] employs a genetic algorithm to search the candidates of adversarial prompts.
- *Additional Factors.* To escape the safeguards, all studies implement strategies to diminish the association between the adversarial prompt and the malicious concept. Specifically, RIATIG [69] imposes a cosine similarity constraint to ensure that the semantics of adversarial prompts diverge from those of malicious prompts. In contrast, other works [30,33,66–68] exclude words associated with the malicious concept in the noise search space.

3.3.3. *Attacking internal safeguards*

This type of safeguard is commonly built into the structure or parameters of the text-to-image model to correct the generation process of malicious images. Typically, the model editing strategy [42,43] is used to erase malicious concepts (such as nudity) in the diffusion model by modifying the model parameters, which will be discussed in Section 4.2.2. In this section, five related attack methods [32,82–85] are reviewed. All of these methods aim for Stable Diffusion with an internal safeguard as the victim model. Nevertheless, their attack objectives, perturbation and optimization strategies differ significantly.

- *Attack objective*. Based on the optimization space of the attack objective, five methods can be categorized into three types: optimization in the class probability space, optimization in the CLIP embedding space, and optimization in the latent space.
(i) Optimization in the classification space: Mehrabi et al. [84] utilize an image classifier to maximize the classification probability of adversarial images belonging to the malicious concept, as shown in Eq. (8).
(ii) Optimization in the CLIP embedding space: Tsai et al. [83] and Ma et al. [85] initially investigate the critical feature $\hat{c}$ that represents the erased malicious concept in malicious prompt features using a decoupling algorithm. Subsequently, they introduce the malicious concept by incorporating $\hat{c}$ into the feature $E_{txt}(x)$ of any input prompt, resulting in a problematic feature $\widetilde{E}_{txt}(x) = E_{txt}(x) + \eta\hat{c}$, where η is the hyper-parameter that controls the malicious level. Finally, they set the attack objective as maximizing the cosine similarity between the feature of the adversarial prompt and the problematic feature:

$$\max cos(E_{txt}(x_{adv}), \widetilde{E}_{txt}(x)). \tag{11}$$

(iii) Optimization in the latent space: P4D [82] aims to reconstruct the erased latent representation of malicious concepts in the latent space. To achieve this goal, they first utilize an additional Stable Diffusion without the model editing strategy to obtain the latent noise $\epsilon_{\theta'}(z_t, E_{txt}(x^c_{mal}), t)$ of malicious concepts.

Subsequently, they seek to align the latent noise of the adversarial prompt with $\epsilon_{\theta'}(z_t, E_{txt}(x^c_{mal}), t)$:

$$\min \|\epsilon_{\theta}(z_t, E_{txt}(x_{adv}), t) - \epsilon_{\theta'}(z_t, E_{txt}(x^c_{mal}), t)\|_2^2, \quad (12)$$

where ϵ_θ and $\epsilon_{\theta'}$ represent the denoising networks of Stable Diffusion with and without the model editing strategy, and z_t is the latent representation in tth timestep. Moreover, Zhang et al. [32] directly utilize a malicious image y_{mal} to guide the generation of the adversarial prompt. Specifically, they first conduct the forward Markov process by iteratively adding noise ϵ_t to the latent representation z_t of the malicious image. Following this, they reversibly reconstruct the latent representation of the malicious image by learning an adversarial prompt:

$$\min \|\epsilon_{\theta}(z_t, E_{txt}(x_{adv}), t) - \epsilon_t\|_2^2. \quad (13)$$

In summary, the three optimization spaces are closely correlated with the semantics of generated images. However, the optimization difficulty varies across these spaces. Based on the computation complexity, the objective in the class probability space is the most challenging to optimize, followed by the objective in the latent space. In contrast, the objective in the CLIP embedding space presents the least difficulty.

- *Perturbation Strategy*. Ma et al. [85] employ a word-level perturbation strategy by appending prefixes. In contrast, other works [32,82,83,85] use sentence-level perturbations to craft adversarial prompts. Notably, the linguistic fluency of adversarial prompts in these sentence-level perturbation strategies is different. Specifically, the adversarial prompt produced by most works [32,82,83] is a combination of several random words, as shown in Fig. 3(e). These combinations are grammatically incorrect and exhibit poor linguistic fluency. Conversely, Mehrabi et al. [84] utilize a large language model (LLM) to optimize the adversarial prompt (shown in Fig. 3(f)), which naturally ensures the fluency and naturalness of the adversarial prompt.
- *Optimization Strategy*. Most works [32,82,83,85] conduct white-box attacks and utilize the PGD algorithm to optimize adversarial prompts. In contrast, Mehrabi et al. [84] propose a black-box agent strategy based on the feedback mechanism. Specifically, they utilize a toxicity classifier to evaluate the toxicity of the adversarial image, which is used as feedback information to guide a large language model to optimize the adversarial prompt.

3.3.4. Attacking black-box safeguards

In this attack scenario, the adversary lacks knowledge of the structure and parameters of the safeguards. This section begins by highlighting two key observations from the works [86,87], which use manually designed adversarial prompts to conduct attacks. Subsequently, we introduce three automatic attack methods [31,34,88] tailored for the black-box text-to-image model.

Millière [87] finds that the visual concepts have robust associations with the subword. Based on the observation, Millière designs the 'macaronic prompt' by creating cryptic hybrid words by combining subword units from different languages, which effectively induce DALL·E 2 to generate malicious images. Millière also observes that the morphological similarity can be utilized to craft the adversarial prompt. To achieve this goal, Millière designs the 'evocative prompting' by crafting words with morphological features resembling existing words. Moreover, Struppek et al. [86] discover that Stable Diffusion and DALL·E 2 are sensitive to character encoding and can inadvertently learn cultural biases associated with different Unicode characters. To exploit this observation, they develop a manual method to inject a few Unicode characters similar with English alphabet into the input prompt, which effectively biases the image generation and induce cultural influences and stereotypes into the generated image.

However, manually designing the adversarial prompt is time-consuming. To address this challenge, three automatic attack methods [31,34,88] are proposed. These methods all target DALL·E 3 [79] and Midjourney [10] as victim models, and all utilize the LLM to conduct the optimization process. Key differences between them are the perturbation and the optimization strategies.

- *Perturbation strategy*. Ba et al. [31] focus on the word-level perturbation by substituting words associated with the malicious concept (shown in Fig. 3(a)). In contrast, Divide-and-Conquer [34] and Groot [88] utilize the LLM to rewrite the entire sentence for crafting the adversarial prompt (shown in Fig. 3(f)), which is the sentence-level perturbation.
- *Optimization strategy*. Ba et al. [31] directly ask the LLM to obtain substitutive words or phrases. For example, to desensitize the bloody scene, they substitute the word "blood" with a candidate word by asking the LLM "which liquids have a similar appearance to blood?". In contrast, Divide-and-Conquer [34] utilizes two LLM agents with different characters: Divide Prompt agent and Conquer Prompt agent, to rewrite the entire prompt. Specifically, Divide Prompt agent is designed to describe various elements of a malicious image using multiple prompts, while deliberately avoiding words associated with the malicious concept. Following this, Conquer Prompt agent refines and polishes the outputs from the Divide Prompt agent, crafting the final adversarial prompt. Different from the above two methods, Groot [88] is a framework designed to bypass the potential text-based and image-based filters, respectively. Specifically, Groot first utilizes a semantic decomposition strategy to transform words associated with the malicious concept into non-malicious words for bypassing the text-based filter. Then, Groot uses a malicious element drowning strategy, which adds a new canvas describing content unrelated to the malicious concept in the prompt, aiming to dilute the malicious information and overload image-based filters.

In addition to the above methods that directly conduct black-box attacks, several methods also utilize the open-source surrogate model to craft the adversarial prompt and then conduct the black-box attack using these adversarial prompts. For the surrogate model, existing methods often focus on the open-source Stable Diffusion with the external safeguard [30,33] and the internal safeguard [83], as discussed in Sections 3.3.2 and 3.3.3.

3.3.5. Summary of targeted attacks

This section summarizes targeted attacks tailored for three types of safeguards and provides two key observations. (1) With the exception of four methods [31,34,84,88] that employ LLM to craft the adversarial prompt, all targeted attacks result in grammatically incorrect adversarial prompts. These ungrammatical adversarial prompts can be detected by grammar detectors, thus reducing the imperceptibility of attacks. (2) Given that external and black-box safeguards commonly include text-based filters, e.g. blacklists, attack methods aimed at bypassing these safeguards often adopt strategies that diminish the association between the adversarial prompt and malicious concepts. However, targeted attacks on internal safeguards can disregard this strategy due to the lack of explicit text-based filters.

3.4. Summary of attack methods

Fig. 4 provides a summary of adversarial attacks on text-to-image diffusion models. We obtain two key observations. (1) Targeted attacks outnumber untargeted attacks, indicating a predominant concern regarding the safety of existing text-to-image diffusion models. (2) The number of word-level and sentence-level perturbations is more than that of character-level perturbations. This phenomenon suggests that perturbations in current methods are relatively easy to be perceptible.

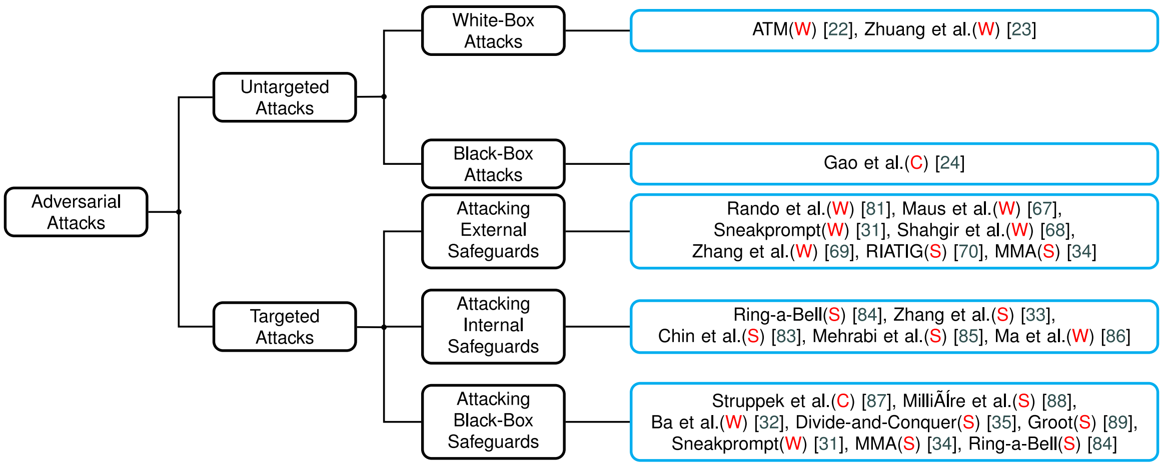

Fig. 4. The taxonomy summary of existing adversarial attack methods on the text-to-image diffusion model. The blue and black rectangles represent the attack method and its category. The red characters (C, W, S) are the character-level, word-level, and sentence-level perturbation strategies, respectively.

4. Defenses

In the previous section, we reveal the robustness and safety vulnerabilities of the text-to-image diffusion model against untargeted and targeted attacks. This section begins by outlining the defense goals designed to mitigate these vulnerabilities, followed by a detailed review of existing defense methods.

Based on the defense goal, existing defense methods can be classified into two categories: (1) improving model robustness and (2) improving model safety. The goal of robustness is to ensure that generated images have consistent semantics with diverse input prompts in practical applications. The safety goal is to prevent the generation of malicious images in response to both malicious and adversarial prompts. Compared to the malicious prompt, the adversarial prompt cleverly omits malicious concepts, as shown in Fig. 3. Therefore, ensuring safety against the adversarial prompt is more challenging.

4.1. Defense for robustness

Some untargeted attacks [20,21,75] reveal that Stable Diffusion struggles to generate accurate images against complex prompts containing multiple objects and attributes. To address this challenge, lots of works are proposed, such as those improving the text encoder [5,37], improving the denoising network [15,89,90], and those focusing on composable image generation [20,75,91,92], controlled image generation [38,93–95], etc. Nevertheless, our study omits the analysis on this part, since several related surveys [56–58,61,96] have already discussed the advancement of image generation capabilities.

The other untargeted attacks [22,23] find that appending or inserting some characters in the clean prompt can result in the semantic deviation of generation images (shown in Fig. 2(c–d)). Unfortunately, we have not found related works that aim to address this challenge posed by these grammatically incorrect prompts. This leaves an open problem for future research on the adversarial defense of text-to-image diffusion models.

4.2. Defense for safety

Based on the knowledge of the model, existing defense methods for improving safety can be classified into two categories: external safeguards and internal safeguards, which have been described in Section 3.3.1.

4.2.1. External safeguards

As shown in Fig. 5, this section introduces three external safeguards, where Latent Guard [39] focuses on the prompt classifier, POSI [40] and GuardT2I [41] focus on the prompt transformation.

Prompt Classifier. Inspired by blacklist-based approaches [25,27], Latent Guard [39] aims to identify malicious prompts by evaluating the distance between the malicious prompt and malicious concepts in the projective space. Specifically, they first add a learnable MLP layer on top of the frozen CLIP text encoder to project the prompt feature. Subsequently, a contrastive learning strategy is designed to close the distance between the projective features of malicious prompts and those of malicious concepts, while pushing the distance between the projective features of safe prompts and those of malicious concepts. In the inference stage, they use a distance threshold to filter prompts closed to malicious concepts in the projective space. Experiments on Stable Diffusion demonstrate effective filter ability against both malicious prompts and adversarial prompts.

Prompt Transformation. This technology aims to rewrite input prompts by training a language model, which can be deployed prior to feeding the prompt into the text-to-image model. Two methods, POSI [40] and GuardT2I [41], use this technology to ensure the safety of the model. Key differences between them include the transformation objective and the model safety:

- *Transformation Objective.* POSI aims to convert malicious prompts to safe prompts by training a lightweight language model. In contrast, GuardT2I focuses on the transformation of ungrammatical adversarial prompts into natural language expressions, followed by a blacklist and a prompt classifier to achieve safety.
- *Model Safety.* POSI focuses solely on safety against malicious prompts. Conversely, GuardT2I achieves safety against both malicious and adversarial prompts by a two-stage framework.

4.2.2. Internal safeguards

This defense strategy aims to ensure that generated images do not contain malicious content, even using a malicious prompt, by modifying the internal parameters or features within the model. Based on the modified object, existing internal safeguards can be divided into two categories: Model Editing and Inference Guidance. As shown in Fig. 6, model editing methods aim to modify the internal parameters by a training process [42–45,70–72,97–103]. In contrast, inference

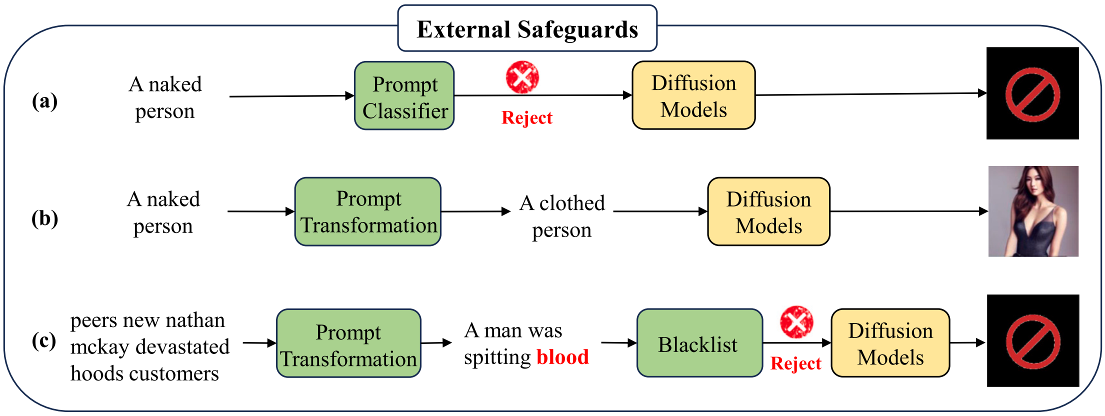

Fig. 5. Three external safeguards for text-to-image diffusion models. **a**, Latent Guard [39] identifies the malicious prompt by training a prompt classifier. **b**, POSI [40] aims to transform the malicious prompt into the safe prompt. **c**, GuardT2I [41] first transform the grammatically incorrect prompt to the natural language expression, followed by a blacklist filter to ensure the model safety.

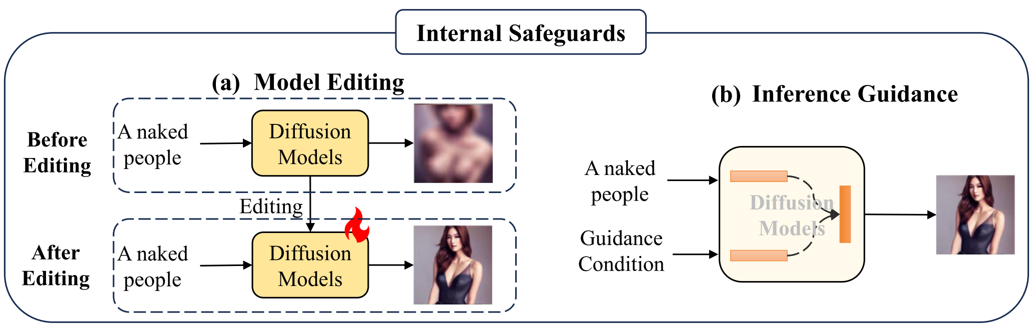

Fig. 6. Two types of internal safeguards. **a**, Model editing methods aim to alter the behavior of the model against the malicious concept (such as 'naked') through a training process. **b**, inference guidance methods ensure that generated images deviate from malicious concepts by applying a guidance strategy during the inference stage.

guidance methods focus on modifying internal features in the inference stage [46,47]. Notably, due to the need to access the model's internal parameters and features, all methods use open-source Stable Diffusion as the target model.

Model Editing. Following Yao et al. [104], we define model editing as a technology that alters the behavior of models against malicious concepts x^c_{mal} (such as 'nudity'), while ensuring no adverse impact on non-malicious concepts $x^c_{n\text{-}mal}$. We identify three key steps typically involved in a model editing method: (a) Designing the editing objective and constraint, where the editing objective involves defining how to edit malicious concepts and the constraint aims to maintain the generation utility of the diffusion model on non-malicious concepts. (b) Determining optimized model parameters. This step identifies which parameters in the model can be modified. (c) Conducting the optimization process by an optimization method. This step focuses on figuring out how to optimize the model parameter effectively. In the following sections, we compare existing model editing methods from the above three steps and the model safety.

- *Editing Objective and Constraint.* Based on the optimization space, existing editing methods can be divided into three types: editing the latent noise in the latent space, editing the output feature of the cross-attention layer in the intermediate feature space, and editing the prompt feature in CLIP embedding space.
(i) Editing in the latent space: A representative strategy summed from a series of studies [42,43,70,97–101,105] is to align the latent noise of the malicious concept and that of the anchor concept to achieve the editing objective. Meanwhile, a constraint on the latent noise of non-malicious concepts [43,102,105] is used to maintain the generation utility of the diffusion model:

$$\underbrace{\min_{\theta} \|\epsilon_\theta(z_t, E_{txt}(x^c_{mal}), t) - \epsilon_a\|_2^2}_{\text{Editing Objective}}$$
$$s.t.\ \underbrace{\|\epsilon_\theta(z_t, E_{txt}(x^c_{n\text{-}mal}), t) - \epsilon_{\theta'}(z_t, E_{txt}(x^c_{n\text{-}mal}), t)\|_2^2 < \delta,}_{\text{Constraint}} \qquad (14)$$

where δ is a threshold to control the deviation range of the generation utility in non-malicious concepts, and ϵ_a is the latent noise of the anchor concept predicted by the frozen denoising network $\epsilon_{\theta'}$. Specifically, ϵ_a can be generated by two strategies: (a) Predefining an anchor concept x^c_{anc}, i.e., $\epsilon_a = \epsilon_{\theta'}(z_t, E_{txt}(x^c_{anc}), t)$ [43,97,105]. For example, when the malicious concept is 'nudity', we can predefine the anchor concept as 'modesty'. (b) Classifier-free guidance: Following ESD [42], several methods [70,98–102] obtain ϵ_a by minimizing the likelihood of generating an image that is labeled as the specific malicious concept x^c_{mal}:

$$\epsilon_a = \epsilon_{\theta'}(z_t, \phi, t) - \eta[\epsilon_{\theta'}(z_t, E_{txt}(x^c_{mal}), t) - \epsilon_{\theta'}(z_t, \phi, t)], \qquad (15)$$

where $\epsilon_{\theta'}(z_t, \phi, t)$ represents the latent noise of the unconditional term and η denotes the guidance scale. Additionally, Hong et al. [98] eliminate the constraint on non-malicious concepts and directly constrain the drift of the latent noise in the unconditional term to preserve the model utility. Therefore, Eq. (14) is rewritten as follows:

$$\min_{\theta} \|\epsilon_\theta(z_t, E_{txt}(x^c_{mal}), t) - \epsilon_a\|_2^2$$
$$s.t.\ \|\epsilon_\theta(z_t, \phi, t) - \epsilon_{\theta'}(z_t, \phi, t)\|_2^2 < \delta. \qquad (16)$$

(ii) Editing in the intermediate feature space: In the text-to-image diffusion model, the denoising network utilizes the cross-attention mechanism to explore the relevance of input prompt

Table 2
The comparison of existing model editing methods. The characters (M, A) represent the safety against the malicious and adversarial prompts, respectively.

Methods	Editing objective and constraint			Optimized model parameters	Optimization method	Model safety
	Optimization space	Editing objective	Constraint			
AC [43]	Latent space	Aligning the latent noise of the malicious concept and that of the predefined anchor concept	Maintaining the latent noise of non-malicious concepts before and after editing	Denoising network	Fine-tune	M
DT [97]				Denoising network	Fine-tune	M
SA [105]				Denoising network	Fine-tune	M
ESD [42]	Latent space	Minimizing the likelihood of the generated image that is labeled as the malicious concept	Maintaining the latent noise of non-malicious concepts before and after editing	Cross-attention layer	Fine-tune	M
SDD [70]				Denoising network	Fine-tune	M
Wu et al. [99]				Denoising network	Fine-tune	M
RACE [100]				Denoising network	Fine-tune	M, A
Receler [101]				Cross-attention layer	Fine-tune	M, A
AdvUnlearn [102]				CLIP Text Encoder	Fine-tune	M, A
Hong et al. [98]	Latent space	Minimizing the likelihood of the generated image that is labeled as the malicious concept	Maintaining the latent noise of the unconditional term before and after editing	Denoising network	Fine-tune	M
UCE [44]	Intermediate feature space	Aligning the output feature of the malicious concept with the anchor concept	Maintaining the output feature of the non-malicious concept before and after editing	Cross-attention layer	Closed-form solution	M
TIME [45]				Cross-attention layer	Closed-form solution	M
ReFACT [71]				Cross-attention layer	Closed-form solution	M
MACE [106]				Cross-attention layer	Closed-form solution	M
FMN [107]	Intermediate feature space	Minimizing the cross-attention map of the malicious concept	None	Denoising network	Fine-tune	M
Safe-CLIP [72]	CLIP embedding space	Aligning the malicious prompt feature with features of both the anchor prompt and the corresponding output image	Maintaining the feature of the non-malicious prompt before and after editing	CLIP text encoder	Fine-tune	M
ConceptPrune [103]	None	Pruning neurons of diffusion models that strong active in the presence of the malicious concept	None	Denoising network	None	M

to the generated image [45]. Therefore, several works [44,45,71,106] implement the editing objective and constraint in the output feature of the cross-attention layer, aiming to modify the relevance of the malicious concept to the generated image:

$$\begin{aligned} &\min_{W} \|W * E_{txt}(x^{c}_{mal}) - W' * E_{txt}(x^{c}_{anc})\|_2^2 \\ &s.t. \|W * E_{txt}(x^{c}_{n\text{-}mal}) - W' * E_{txt}(x^{c}_{n\text{-}mal})\|_2^2 < \delta, \end{aligned} \tag{17}$$

where W and W' are the parameters of the optimized and frozen cross-attention layer. In addition to the above methods that edit the output feature of the cross-attention layer, FMN [107] directly minimizes the cross-attention map of the malicious concept to make the model forget the malicious concept.

(iii) Editing in the CLIP embedding space: Safe-CLIP [72] focuses on the fine-tuning of the CLIP text encoder and aims to align the malicious prompt with both the anchor prompt x_{anc} and the corresponding output image y_{anc} in the CLIP embedding space. Meanwhile, Safe-CLIP utilizes a constraint on the non-malicious prompt $x_{n\text{-}mal}$:

$$\begin{aligned} &\max_{E_{txt}} (cos(E_{txt}(x_{mal}), E'_{txt}(x_{anc}))+ \\ &\quad cos(E_{txt}(x_{mal}), E'_{img}(y_{anc}))) \\ &s.t. \ (cos(E_{txt}(x_{n\text{-}mal}), E'_{txt}(x_{n\text{-}mal}))+ \\ &\quad cos(E_{txt}(x_{n\text{-}mal}), E'_{img}(y_{n\text{-}mal}))) < \delta, \end{aligned} \tag{18}$$

where E_{txt} is the optimized CLIP text encoder, E'_{txt}, E'_{img} are the frozen CLIP text and image encoders, and $y_{n\text{-}mal}$ is the image generated by $x_{n\text{-}mal}$.

- *Optimized Model Parameters.* The straightforward approach involves optimizing all parameters of the victim model. However, this method is highly cost-intensive. As a more efficient alternative, some methods focus on targeting specific components of the model for optimization, such as the CLIP text encoder [72,102], the denoising network [43,97–100], even cross-attention layers [42–45,70,71,101,106].
- *Optimization Method.* Most works [42,43,70,72,97–100] aim at directly fine-tuning the optimized model parameters using the dataset. In contrast, some works [44,45,71,106] optimize parameters using a closed-form solution, which significantly reduces the optimization time.
- *Model Safety.* Except for three of them (RACE [100], Receler [101] and AdvUnlearn [102]) consider safety against both the malicious prompt and adversarial prompt, other works all only focus on the editing of the malicious prompt. Specifically, these works [100–102] begin by employing an adversarial attack strategy to adaptively identify adversarial prompts capable of reconstructing malicious images. Subsequently, they iteratively edit both malicious concepts and adversarial prompts to improve model safety.

Different from the above methods that erase the malicious concept by modifying the specific components of the model, ConceptPrune [103] aims to identify neurons of diffusion models that strongly activate in the presence of the malicious concept. Subsequently, concept removal is achieved by simply pruning or zeroing out these neurons [108]. Moreover, ConceptPrune [103] does not require any costly weight update since the neurons corresponding with the malicious concept can be obtained with just a single forward propagation through the model.

The summary of model editing methods is shown in Table 2. Most methods [42,43,70,97–101,105] focus on editing the latent noise in the latent space, which more closely related to the distribution of the generated images than the prompt feature and output feature of the cross-attention layer. Moreover, most methods [43,70,97–100,105,107] fine-tune the denoising network to conduct the optimization process. Furthermore, most methods [42–45,70–72,97–99,103,105–107] focus solely on the safety against the malicious prompt while ignoring the adversarial prompt.

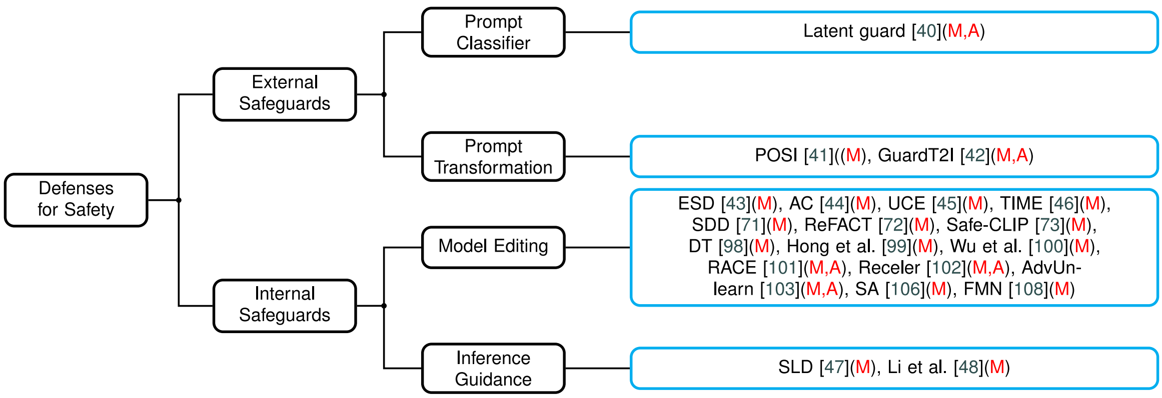

Fig. 7. The taxonomy summary of existing defense methods for safety. The blue and black rectangles represent the defense method and its category. The red characters (M, A) represent the safety against the malicious and adversarial prompts.

Inference Guidance. Unlike model editing methods that modify the model parameters in the training stage, inference guidance methods only modify internal features of the model in the inference stage. This section reviews two inference guidance methods, SLD [70] and Li et al. [47], which both focus solely on the safety against malicious prompts. Key differences between them are the feature modification strategy and the guidance strategy.

- *Feature Modification Strategy.* SLD [70] focuses on modifying the latent noise of input prompt in the latent space. In contrast, Li et al. [47] focus on modifying the output feature of the bottleneck layer in the denoising network.
- *Guidance Strategy.* SLD [70] first predefines some malicious concepts and subsequently directs the image generation process in a direction opposite to these concepts, ensuring that output images are far away from these malicious concepts. In contrast, Li et al. [47] aim to learn a safe vector, which is integrated with the feature of the malicious prompt in the bottleneck layer, ensuring that output images closely align with safe images.

Summary of Internal Safeguards. Compared to model editing methods, inference guidance methods are tuning-free and plug-in, which can be easily inserted into any diffusion model. However, the inference guidance strategy also presents a security risk, as it can be easily manually removed by developers when they access the open-source text-to-image model. Moreover, as reviewed in targeted attacks for internal safeguards (Section 3.3.3), most model editing and inference guidance methods are still vulnerable and can generate malicious images using adversarial prompts. Therefore, there are still significant challenges in future research on model safety.

4.3. Summary of defense methods

In defense methods for improving robustness, researchers often focus on improving image quality in response to prompts with multiple objects and attributes (shown in Fig. 2(a)), while overlooking the robustness of the model against grammatically incorrect prompts with subtle noise (shown in Fig. 2(b-d)). However, ensuring the robustness of the model against prompts with subtle noise is also crucial, as users frequently make typographical errors in a practical situation. For the defense for improving safety, we summarize related methods in Fig. 7. Most methods concentrate on the internal safeguards, particularly those involving model editing strategies. This highlights the significant efforts by researchers to ensure the safety of AI-generated content. Nevertheless, most methods only focus on the safety against malicious prompt, leaving substantial challenges in countering adversarial attacks.

5. Evaluation metrics

This section illustrates metrics used in adversarial attack methods (Section 3) and defense methods for safety (Section 4.2).

5.1. Attack metrics

For adversarial attacks on the text-to-image diffusion model, researchers predominantly focus on two kinds of metrics for evaluating an attack method: Attack Effectiveness and Attack Imperceptibility. **Attack Effectiveness** measures the ability of adversarial prompts to induce victim models into generating images that semantically align with the adversary's intent. Moreover, the adversarial prompt should not appear suspicious or significantly differ from clean prompts, ensuring that it retains naturalness and avoids detection. Therefore, **Attack Imperceptibility** evaluates the extent to which the adversarial prompts are indistinguishable from clean prompts.

5.1.1. Attack effectiveness

To evaluate whether adversarial images semantically align with the adversary's intent, two common metrics are used: **Classification Accuracy (ACC)** [21,32,33,68,82] and **CLIP Score** [22,23,69].

- *ACC* initially employs an image classifier f to assess the semantics of adversarial images y_{adv}. Subsequently, ACC is determined based on the specific intent of the attack. Specifically, untargeted attacks ensure that the adversarial images are incorrectly categorized into classes distinct from the class of the clean images. In contrast, targeted attacks ensure that adversarial images are categorized into the category associated with the malicious images. A higher ACC value indicates higher effectiveness in both untargeted and targeted attacks.
- *CLIP Score* first uses a CLIP model E to assess the semantics of adversarial images. Next, the CLIP Score is obtained by computing the feature similarity between adversarial images and attacked prompts x_{attack} in the CLIP embedding space:

$$\text{CLIP Score} = \frac{1}{N}\sum_{i=1}^{N} cos(E_{img}(y_{adv}), E_{txt}(x_{attack})), \quad (19)$$

where N is the number of adversarial images, E_{img} and E_{txt} are the image encoder and text encoder in the CLIP model. In untargeted attacks (Section 3.2), x_{attack} denotes the clean prompt, where a lower CLIP score is expected to diverge the semantics of adversarial images from those of clean images. Conversely, in targeted attacks (Section 3.3), x_{attack} refers to the malicious prompt, where a higher CLIP score is expected to align the semantics of adversarial images with those of malicious images.

5.1.2. Attack imperceptibility

In untargeted attacks, the objective is to investigate the robustness vulnerabilities by adding subtly perturbation to the clean prompt. Consequently, to evaluate the imperceptibility of untargeted attacks, Gao et al. [23] typically employs **Levenshtein Distance (L-Distance)** which measures the minimum number of single-character edits from the clean prompt to the adversarial prompt. Specifically, suppose that the clean prompt x_{clean} and the adversarial prompt x_{adv} contain a and b characters, respectively. The L-Distance $\text{lev}(a, b)$ between x_{clean} and x_{adv} is computed as follows:

$$\text{lev}(a,b) = \begin{cases} \max(a,b) & \text{if } \min(a,b)=0, \\ \min\begin{cases} \text{lev}(a-1,b)+1, \\ \text{lev}(a,b-1)+1, \\ \text{lev}(a-1,b-1)+\mathbf{1}_{(x^a_{clean}\neq x^b_{adv})} \end{cases} & \text{otherwise,} \end{cases} \tag{20}$$

where $\mathbf{1}$ is an indicator function. A lower L-Distance value represents better imperceptibility.

In targeted attacks, the objective is to produce an adversarial prompt capable of bypassing safeguards while generating adversarial images containing malicious concepts. Therefore, **Bypass Rate** is frequently utilized to assess the imperceptibility of targeted attacks. Specifically, supposing that a safeguard D is used to detect malicious prompts: $D(x_{mal}) = 1$, Bypass Rate of adversarial prompts is computed as follows:

$$\text{Bypass Rate} = \frac{1}{N}\sum_{i=1}^{N}(1 - D(x^i_{adv})). \tag{21}$$

A higher Bypass Rate refers to better imperceptibility.

Additionally, **Sentence Naturality** is also a crucial metric to determine whether the adversarial prompt looks natural, similar to normal sentences in the real world. Existing methods often employ the **perplexty (PPL)** to assess the sentence naturality of adversarial prompts. Specifically, supposing that an adversarial prompt x_{adv} contains T tokens: $\{t_1, t_2, \ldots, t_T\}$, the PPL of the adversarial prompt can be computed as follows:

$$\text{PPL}(x_{adv}) = \exp\{-\frac{1}{T}\sum_{i=1}^{T} p_\theta(t_i|t_1, t_2, \ldots, t_{i-1})\}, \tag{22}$$

where p_θ is a pre-trained causal language model like GPT-2 [109]. A higher PPL value indicates a more natural sentence. The final sentence naturality can be obtained by averaging the PPL of adversarial prompts.

5.2. Defense metrics for safety

Existing defense methods for safety mainly concentrate on three types of metrics: Defense Effectiveness, Defense Specificity, and Image Fidelity. **Defense Effectiveness** measures the ability to prevent the generation of malicious images when inputting malicious prompts. **Defense Specificity** evaluates whether the model accurately produces images with consistent semantics in response to non-malicious prompts. **Image Fidelity** assesses how closely generated images resemble real images in visual quality following the integration of defense methods.

5.2.1. Defense effectiveness

A common metric for evaluating Defense Effectiveness is Classification Accuracy (ACC). This metric typically utilizes an image classifier to determine whether the generated images contain malicious concepts. A lower ACC value for images generated from malicious prompts indicates stronger defense effectiveness.

Additionally, POSI [40] and GuardT2I [41], as the external safeguard to detect malicious prompts, also utilize **AUROC** metric to assess the classification ability between malicious and non-malicious prompts. The AUROC metric measures the ability of the external safeguard to discriminate between malicious and non-malicious prompts. It quantifies the trade-off between the true positive rate (TPR) and the false positive rate (FPR), offering a comprehensive assessment of the classification performance across various thresholds.

5.2.2. Defense specificity

To evaluate whether a defense method negatively impacts the text-to-image model's performance on non-malicious prompts, a common metric used is **CLIP Score**, which measures the semantic similarity between non-malicious prompts and the corresponding generated images. Moreover, some methods employ the **ACC** metric, which uses an image classifier to determine whether the generated images are correctly classified into non-malicious categories.

Additionally, GuardT2I also employs **FPR@TPR95%** metric to evaluate the negative impact of the safeguard on non-malicious prompts. Specifically, FPR@TPR95% measures the proportion of false positives (non-malicious prompts incorrectly identified as malicious prompts) when the safeguard correctly identifies 95% of the true positives (actual malicious prompts). A lower FPR@TPR95 value is preferable, as it indicates that the safeguard can accurately detect malicious prompts with fewer mistakes. This metric is particularly crucial in commercial applications, where frequent false alarms are unacceptable.

5.2.3. Image fidelity

Two common metrics are used to assess the image fidelity: **Fréchet Inception Distance (FID)** [110] and **Inception Score (IS)** [111].

- *FID* compares the distribution difference between the generated images and real images. Specifically, we first utilize an Inception network [112] to extract the features of generated and real images. Then, FID can be computed using a Fréchet distance [113] as follows:

$$\text{FID} = \|\mu_1 - \mu_2\|^2 + Tr(\Sigma_1 + \Sigma_2 - 2 * \sqrt{\Sigma_1 * \Sigma_2}), \tag{23}$$

  where μ_1 and μ_2 are mean features of adversarial and real images, respectively. Σ_1 and Σ_1 are the co-variances of features of adversarial and real images, respectively. Tr represents the trace of the matrix. A lower FID value indicates better visual quality of adversarial images.
- *IS* employs an Inception network [112] pre-trained on ImageNet [114] to assess the class predictions c for generated images y. The score rewards configurations where the conditional class probabilities $p(c|y)$ exhibit low entropy, indicating that the generated images are distinctly classifiable into one of the predefined classes. Simultaneously, it favors high entropy in the marginal class distribution $\hat{p}(c)$, reflecting a substantial diversity among the generated samples. There, IS can be obtained by a KL divergence:

$$\text{IS} = \exp\left(\frac{1}{N}\sum_{i=1}^{N} D_{KL}\left(p(c^i|y^i) \parallel \hat{p}(c)\right)\right), \tag{24}$$

  where c^i is the probability distribution of the generated image y^i predicted by the Inception network.

6. Dataset

This section introduces several datasets commonly used in adversarial attacks and defenses for the text-to-image model. Based on the prompt source, these datasets are categorized into two types: clean and adversarial datasets. The clean dataset consists of clean prompts that are not attacked and are typically crafted by humans, while the adversarial dataset comprises adversarial prompts generated by existing attack methods.

6.1. Clean dataset

According to the category of prompts involved in the dataset, existing clean datasets are further divided into two types: non-malicious and malicious datasets. The non-malicious dataset contains non-malicious prompts describing non-malicious concepts, while the malicious dataset contains explicitly malicious prompts.

6.1.1. Non-malicious dataset

The commonly used non-malicious datasets include ImageNet [114], MSCOCO [115], LAION-COCO [116], and DiffusionDB [117].

- *ImageNet* [114], which contains images describing 1000 categories of common objects in the real world, is a significant benchmark in the field of computer vision. As a result, some works [21, 68] craft clean datasets based on the category information in ImageNet. For instance, ATM [21] employs a standardized template: "A photo of {CLASS_NAME}" to generate clean prompts, where "{CLASS_NAME}" denotes the class name in ImageNet.
- *MSCOCO* [115] is a cross-modal image-text dataset, a popular benchmark for training and evaluating text-to-image generation models. Specifically, MSCOCO includes 82,783 training images and 40,504 testing images, each with 5 text descriptions.
- *LAION-COCO* [116] is a subset of LAION-5B [118], which is a large-scale image-text dataset in the real world. LAION-COCO includes 600 million images and corresponding text descriptions.
- *DiffusionDB* [117] is a large-scale text-to-image prompt dataset, which contains 14 million images generated by Stable Diffusion using prompts from real users.

These datasets contain large-scale, diverse clean prompts that can be used to evaluate the effectiveness of attack and defense methods. For example, Gao et al. [23] add perturbations to the clean prompt from LAION-COCO and DiffusionDB to explore robustness vulnerabilities of text-to-image models. To explore safety vulnerabilities while minimizing the use of malicious examples to prevent potential discomfort for the audience, RIATIG [69] targets some images from MSCOCO as malicious content, and designs adversarial prompts to bypass safeguards and generate targeted images. Similarly, ESD [42] targets specific categories from ImageNet as malicious concepts, and designs a model editing method to erase the ability to generate images of these categories.

6.1.2. Malicious dataset

Malicious datasets typically include explicitly malicious prompts (such as those describing sexual or violent content), prompts involving copyright infringement (e.g., describing specific artistic styles like 'Van Gogh'), and sensitive prompts depicting political figures (such as 'Donald John Trump'). Some commonly used datasets are as follows:

- *Unsafe Diffusion [24]* provides 30 manually crafted malicious prompts that describe sexual and bloody content, as well as political figures.
- *SneakyPrompt [30]*. uses ChatGPT to automatically generate 200 malicious prompts that involve sexual and bloody content.
- *I2P [46]* comprises 4703 inappropriate prompts, encompassing hate, harassment, violence, self-harm, nudity content, shocking images, and illegal activity. These inappropriate prompts are real-user inputs sourced from an image generation website, Lexica.[11]
- *MMA [33]* samples and releases 1000 malicious prompts from LAION-COCO based on an NSFW (Not Safe for Work) score. These malicious prompts mainly focus on sexual content.
- *Image Synthesis Style Studies Database* [119] compiles thousands of artists whose styles can be replicated by various text-to-image models, such as Stable Diffusion and Midjourney.
- *MACE [106]* provides a dataset comprising 200 celebrities whose portraits, generated using Stable Diffusion, are recognized with remarkable accuracy (>99%) by GIPHY Celebrity Detector [120].

These datasets are frequently used to assess the effectiveness of targeted attacks and safety defenses in real-world applications.

[11] https://lexica.art/.

6.2. Adversarial dataset

The adversarial dataset contains adversarial prompts generated by adversarial attacks. Some public adversarial datasets are as follows:

- *Adversarial Nibbler Dataset [121]* consists of 3412 adversarial prompts that effectively bypass safeguards while inducing text-to-image models to generate malicious images. These prompts, which include violent, sexual, biased, and hate-based material, are manually crafted during Adversarial Nibbler Challenge [122].
- *MMA [33]* targets 1000 malicious prompts, generating 1000 corresponding adversarial prompts using the proposed attack method. These adversarial prompts mainly focus on porn content.
- *Zhang et al.* [68] target 10 objects as malicious concepts and generates 500 adversarial prompts for each object. These adversarial prompts are capable of inducing the text-to-image model to produce images related to malicious concepts, even when the prompt excludes words directly related to them.

These datasets can be used to reveal the safety vulnerability against adversarial prompts and assess the effectiveness of the defense method.

7. Challenge and future directions

Although lots of works are proposed to reveal and further mitigate the vulnerabilities of text-to-image diffusion models, existing adversarial attacks and defenses still face significant challenges. This section introduces the potential challenges associated with existing methods, and explores the possible solutions to address these challenges.

7.1. Imperceptibility of adversarial attack

As shown in Fig. 4, most attacks employ word-level and sentence-level perturbation strategies. This significant added noise to the input prompt can be easily detected, thus reducing the imperceptibility of the adversarial attack. Consequently, a challenge is **how to improve imperceptibility in attacks.** This section discusses the imperceptibility of untargeted and targeted attacks, respectively.

7.1.1. Imperceptibility of untargeted attacks

Recent works [22,66] find that subtle noise in the prompt can lead to substantial differences in the model output. In practice, such noise is typically imperceptible, such as additional spaces or characters within the input prompt. However, existing untargeted attacks primarily focus on word-level perturbations (as shown in Fig. 2(b,c)), which are uncommon in real-world scenarios. Consequently, a challenge remains: *How can we effectively achieve attacks using prompts with subtle noises that a user can input?*

One possible solution is to utilize the character-level perturbation. For example, Gao et al. [23] employ typos, glyph alterations, and phonetic changes to carry out their attacks. However, the attack scenarios produced by Gao et al. [23] require further discussion, as their adversarial prompts significantly deviate from clean prompts in the text semantics. Specifically, they introduce character-level perturbations into the keywords of the clean prompts, such as 'astronaut' and 'horse' in Fig. 2(d). These alterations significantly change the semantics of the clean prompt, raising the question of whether the output image should be expected to change when the prompt keywords are modified. A reasonable setting involves adding subtler noise in non-keywords that are likely to appear in practice, such as the character substitution (e.g., replacing 'O' with '0') and character repetition (e.g., changing 'hello' to 'helloo').

7.1.2. Imperceptibility of targeted attacks

According to Fig. 3, all targeted attacks result in grammatically incorrect adversarial prompts (Fig. 3(a–e)), except for those methods that utilize the LLM to craft the adversarial prompt (Fig. 3(f)). This type of grammatically incorrect prompt can be identified by a grammar detector, further diminishing the imperceptibility of targeted attacks. Consequently, an open question for researchers remains: *Can we employ adversarial prompts with correct grammatical structure to efficiently achieve targeted attacks?*

One potential approach is to leverage the LLM to conduct adversarial attacks, as demonstrated by Ba et al. [31] and the Divide-and-Conquer method [34]. However, automatically crafting adversarial prompts from malicious prompts is a non-trivial task, which requires the LLM to address several critical challenges: (a) Understanding why a malicious prompt cannot bypass existing safeguards. (b) Exploring perturbation strategies to modify the malicious prompt into the adversarial prompt. (c) Optimizing the adversarial prompt based on the feedback from the attack process. Existing methods based on LLM often require substantial human intervention in these steps [31,34,84,88], which limits their capabilities and generalization. Notably, the rise of LLM-based multi-agent approaches indicates a trend toward using multiple LLMs to collaboratively tackle complex NLP tasks [123–127]. Thus, designing an LLM-based multi-agent framework to automate adversarial attacks represents an intriguing and challenging endeavor.

7.2. Defense effectiveness against adversarial attacks

In Section 4, we present defense methods for improving the robustness and safety of text-to-image diffusion models. However, **these methods still exhibit vulnerabilities to adversarial attacks.** For instance, there is a notable deficiency in defense mechanisms against untargeted attacks that utilize grammatically incorrect prompts with subtle noise (such as appending five random characters in Fig. 2(c) and incorporating typos in Fig. 2(d)). Moreover, most defense methods designed to improve safety primarily focus on addressing malicious prompts while neglecting adversarial prompts, thereby limiting their overall effectiveness. This section examines the challenges faced by defense methods in addressing untargeted and targeted attacks.

7.2.1. Defense effectiveness against untargeted attacks

Current defense methods against untargeted attacks primarily address the generation defect in prompts with multiple objects and attributes, as illustrated in Fig. 2(a). However, *for untargeted attacks that utilize ungrammatically incorrect prompts with subtle noise (depicted in Fig. 2(b-d)), mature solutions are still lacking.*

One straightforward approach is to fine-tune the CLIP text encoder with grammatically incorrect adversarial prompts, which can mitigate the impact of noise within prompts. Nevertheless, this solution presents a critical challenge: *How can we maintain the fundamental encoding capabilities of the CLIP text encoder?* Given that CLIP is a foundational model pre-trained on a vast corpus of image-text pairs (two billion pairs) [118,128] and contains extensive cross-modal semantic information, fine-tuning the model directly with a small set of adversarial prompts inevitably impair its basic encoding functionality [129,130]. An alternative strategy involves using model editing methods [44,45,71,106], which modify only a subset of parameters of the CLIP model to correct behaviors against adversarial prompts. This approach enables targeted modifications while preserving the core encoding capabilities.

7.2.2. Defense effectiveness against targeted attacks

As shown in Fig. 7, there are five defense methods [39,41,100–102] that consider both the malicious and adversarial prompts. However, they primarily address adversarial prompts that utilize random noise words (illustrated in Fig. 3(e)), while overlooking other types of adversarial prompts, such as those utilizing the word-level perturbation (shown in Fig. 3(a–c)) and those crafted by the LLM (depicted in Fig. 3(f)). Consequently, this raises a significant challenge: *How can we simultaneously defend all types of adversarial prompts?*

This problem cannot be addressed directly due to the rapid development of adversarial attacks. Notably, a novel work proposed by Zhang et al. [68] seeks to reveal underlying patterns of adversarial prompts. A potential strategy involves leveraging these patterns, rather than adversarial prompts, to guide the defense process. This approach can be a promising future direction for enhancing safeguards.

8. Conclusion

In this work, we present a comprehensive review of adversarial attacks and defenses on the text-to-image diffusion model. We first provide an overview of text-to-image diffusion models. We then introduce a taxonomy of adversarial attacks targeting these models, followed by an in-depth review of related attack methods to uncover vulnerabilities in model robustness and safety. Next, we introduce the corresponding defense methods designed to improve the robustness and safety of the model. Finally, we analyze the limitations and challenges of existing adversarial attack and defense methods, and discuss the potential solution for future work.

CRediT authorship contribution statement

Chenyu Zhang: Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Mingwang Hu:** Writing – original draft, Methodology, Data curation. **Wenhui Li:** Writing – review & editing, Investigation. **Lanjun Wang:** Writing – review & editing, Writing – original draft, Supervision, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (62202329).

References

[1] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, in: M. Meila, T. Zhang (Eds.), ICML, in: Proceedings of Machine Learning Research, Vol. 139, 2021, pp. 8821–8831.

[2] J. Yu, Y. Xu, J.Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B.K. Ayan, et al., Scaling autoregressive models for content-rich text-to-image generation, 2022, p. 5, arXiv preprint arXiv:2206.10789, 2(3).

[3] M. Ding, Z. Yang, W. Hong, W. Zheng, C. Zhou, D. Yin, J. Lin, X. Zou, Z. Shao, H. Yang, et al., Cogview: Mastering text-to-image generation via transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 19822–19835.

[4] M. Ding, W. Zheng, W. Hong, J. Tang, Cogview2: Faster and better text-to-image generation via hierarchical transformers, Adv. Neural Inf. Process. Syst. 35 (2022) 16890–16902.

[5] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, M. Chen, Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2021, arXiv preprint arXiv:2112.10741.

[6] C. Wu, J. Liang, L. Ji, F. Yang, Y. Fang, D. Jiang, N. Duan, Nüwa: Visual synthesis pre-training for neural visual world creation, in: European Conference on Computer Vision, Springer, 2022, pp. 720–736.

[7] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E.L. Denton, S.K.S. Ghasemipour, R.G. Lopes, B.K. Ayan, T. Salimans, J. Ho, D.J. Fleet, M. Norouzi, Photorealistic text-to-image diffusion models with deep language understanding, in: NeurIPS, 2022.

[8] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, et al., Improving image generation with better captions, Comput. Sci. 2 (3) (2023) 8, https://cdn.openai.com/papers/dall-e-3.pdf.
[9] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: CVPR, IEEE, 2022, pp. 10674–10685.
[10] Midjourney, Midjourney, 2023, https://www.midjourney.com.
[11] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Adv. Neural Inf. Process. Syst. 33 (2020) 6840–6851.
[12] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E.L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al., Photorealistic text-to-image diffusion models with deep language understanding, Adv. Neural Inf. Process. Syst. 35 (2022) 36479–36494.
[13] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, K. Aberman, Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22500–22510.
[14] T. Brooks, A. Holynski, A.A. Efros, Instructpix2pix: Learning to follow image editing instructions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18392–18402.
[15] W. Peebles, S. Xie, Scalable diffusion models with transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4195–4205.
[16] X. Yi, L. Tang, H. Zhang, H. Xu, J. Ma, Diff-IF: Multi-modality image fusion via diffusion model with fusion knowledge prior, Inf. Fusion 110 (2024) 102450.
[17] H. Huang, W. He, H. Zhang, Y. Xia, L. Zhang, STFDiff: Remote sensing image spatiotemporal fusion with diffusion models, Inf. Fusion (2024) 102505.
[18] A. Ahfaz, Stable diffusion statistics: Users, revenue, & growth, 2024, https://openaijourney.com/stable-diffusion-statistics/.
[19] C. Zhang, C. Zhang, M. Zhang, I.S. Kweon, Midjourney statistics: Users, polls, & growth, 2023, https://approachableai.com/midjourney-statistics/.
[20] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, D. Cohen-Or, Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, TOG 42 (4) (2023) 1–10.
[21] C. Du, Y. Li, Z. Qiu, C. Xu, Stable diffusion is unstable, Adv. Neural Inf. Process. Syst. 36 (2024).
[22] H. Zhuang, Y. Zhang, S. Liu, A pilot study of query-free adversarial attack against stable diffusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2385–2392.
[23] H. Gao, H. Zhang, Y. Dong, Z. Deng, Evaluating the robustness of text-to-image diffusion models against real-world attacks, 2023, arXiv preprint arXiv:2306.13103.
[24] Y. Qu, X. Shen, X. He, M. Backes, S. Zannettou, Y. Zhang, Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models, 2023, arXiv preprint arXiv:2305.13873.
[25] M. Heikkilä, Ai-image-generator-midjourney-blocks-porn-by-banning-words-about-the-human-reproductive-system/, 2023, https://technologyreview.com/2023/02/24/1069093/.
[26] OpenAI, DaLL·e 3 system card, 2023, https://openai.com/research/dall-e-3-system-card.
[27] R. George, NSFW-words-list, 2020, https://github.com/rrgeorge-pdcontributions/NSFW-Words-List.
[28] L. AI, CLIP-based-NSFW-detector, 2023, https://github.com/LAION-AI/CLIP-based-NSFW-Detector.
[29] L. Chhabra, NSFW-detection-DL, 2020, https://github.com/lakshaychhabra/NSFW-Detection-DL.
[30] Y. Yang, B. Hui, H. Yuan, N. Gong, Y. Cao, SneakyPrompt: Evaluating robustness of text-to-image generative models' safety filters, in: Proceedings of the IEEE Symposium on Security and Privacy, 2024.
[31] Z. Ba, J. Zhong, J. Lei, P. Cheng, Q. Wang, Z. Qin, Z. Wang, K. Ren, SurrogatePrompt: Bypassing the safety filter of text-to-image models via substitution, 2023, arXiv preprint arXiv:2309.14122.
[32] Y. Zhang, J. Jia, X. Chen, A. Chen, Y. Zhang, J. Liu, K. Ding, S. Liu, To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now, 2023, arXiv preprint arXiv:2310.11868.
[33] Y. Yang, R. Gao, X. Wang, T.-Y. Ho, N. Xu, Q. Xu, Mma-diffusion: Multimodal attack on diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 7737–7746.
[34] Y. Deng, H. Chen, Divide-and-conquer attack: Harnessing the power of LLM to bypass the censorship of text-to-image generation model, 2023, arXiv preprint arXiv:2312.07130.
[35] L. Lian, B. Li, A. Yala, T. Darrell, Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models, 2023, arXiv preprint arXiv:2305.13655.
[36] O. Bar-Tal, L. Yariv, Y. Lipman, T. Dekel, Multidiffusion: Fusing diffusion paths for controlled image generation, 2023.
[37] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, R. Rombach, Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023, arXiv preprint arXiv:2307.01952.
[38] O. Avrahami, T. Hayes, O. Gafni, S. Gupta, Y. Taigman, D. Parikh, D. Lischinski, O. Fried, X. Yin, Spatext: Spatio-textual representation for controllable image generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18370–18380.
[39] R. Liu, A. Khakzar, J. Gu, Q. Chen, P. Torr, F. Pizzati, Latent guard: a safety framework for text-to-image generation, 2024, arXiv preprint arXiv:2404.08031.
[40] Z. Wu, H. Gao, Y. Wang, X. Zhang, S. Wang, Universal prompt optimizer for safe text-to-image generation, 2024, arXiv preprint arXiv:2402.10882.
[41] Y. Yang, R. Gao, X. Yang, J. Zhong, Q. Xu, GuardT2I: Defending text-to-image models from adversarial prompts, 2024, arXiv preprint arXiv:2403.01446.
[42] R. Gandikota, J. Materzynska, J. Fiotto-Kaufman, D. Bau, Erasing concepts from diffusion models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2426–2436.
[43] N. Kumari, B. Zhang, S.-Y. Wang, E. Shechtman, R. Zhang, J.-Y. Zhu, Ablating concepts in text-to-image diffusion models, in: ICCV, 2023, pp. 22691–22702.
[44] R. Gandikota, H. Orgad, Y. Belinkov, J. Materzyńska, D. Bau, Unified concept editing in diffusion models, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 5111–5120.
[45] H. Orgad, B. Kawar, Y. Belinkov, Editing implicit assumptions in text-to-image diffusion models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 7053–7061.
[46] P. Schramowski, M. Brack, B. Deiseroth, K. Kersting, Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models, CVPR (2022) 22522–22531.
[47] H. Li, C. Shen, P. Torr, V. Tresp, J. Gu, Self-discovering interpretable diffusion latent directions for responsible text-to-image generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 12006–12016.
[48] W.E. Zhang, Q.Z. Sheng, A. Alhazmi, C. Li, Adversarial attacks on deep-learning models in natural language processing: A survey, ACM Trans. Intell. Syst. Technol. 11 (3) (2020) 1–41.
[49] Y. Zhang, K. Shao, J. Yang, H. Liu, Adversarial attacks and defenses on deep learning models in natural language processing, in: 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference, ITNEC, Vol. 5, IEEE, 2021, pp. 1281–1285.
[50] S. Goyal, S. Doddapaneni, M.M. Khapra, B. Ravindran, A survey of adversarial defenses and robustness in NLP, ACM Comput. Surv. 55 (14s) (2023).
[51] E. Shayegani, M.A.A. Mamun, Y. Fu, P. Zaree, Y. Dong, N. Abu-Ghazaleh, Survey of vulnerabilities in large language models revealed by adversarial attacks, 2023, arXiv preprint arXiv:2310.10844.
[52] N. Akhtar, A. Mian, Threat of adversarial attacks on deep learning in computer vision: A survey, Ieee Access 6 (2018) 14410–14430.
[53] N. Akhtar, A. Mian, N. Kardan, M. Shah, Advances in adversarial attacks and defenses in computer vision: A survey, IEEE Access 9 (2021) 155161–155196.
[54] H. Baniecki, P. Biecek, Adversarial attacks and defenses in explainable artificial intelligence: A survey, Inf. Fusion (2024) 102303.
[55] N. Rodríguez-Barroso, D. Jiménez-López, M.V. Luzón, F. Herrera, E. Martínez-Cámara, Survey on federated learning threats: Concepts, taxonomy on attacks and defences, experimental study and challenges, Inf. Fusion 90 (2023) 148–173.
[56] C. Zhang, C. Zhang, M. Zhang, I.S. Kweon, Text-to-image diffusion models in generative ai: A survey, 2023, arXiv preprint arXiv:2303.07909.
[57] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, M.-H. Yang, Diffusion models: A comprehensive survey of methods and applications, ACM Comput. Surv. 56 (4) (2023) 1–39.
[58] M. Żelaszczyk, J. Mańdziuk, Text-to-image cross-modal generation: A systematic review, 2024, arXiv preprint arXiv:2401.11631.
[59] F.-A. Croitoru, V. Hondru, R.T. Ionescu, M. Shah, Diffusion models in vision: A survey, IEEE Trans. Pattern Anal. Mach. Intell. 45 (9) (2023) 10850–10869.
[60] S. Hartwig, D. Engel, L. Sick, H. Kniesel, T. Payer, T. Ropinski, et al., Evaluating text to image synthesis: Survey and taxonomy of image quality metrics, 2024, arXiv preprint arXiv:2403.11821.
[61] P. Cao, F. Zhou, Q. Song, L. Yang, Controllable generation with text-to-image diffusion models: A survey, 2024, arXiv preprint arXiv:2403.04279.
[62] OpenAI, DALL·E 2, 2023, https://openai.com/index/dall-e-2.
[63] ClipDrop, 2024, https://clipdrop.co/text-to-image.
[64] Discord, 2024, https://discord.com/invite/stablediffusion.
[65] Stability.ai, Stable diffusion XL, 2023, https://dreamstudio.ai/.
[66] N. Maus, P. Chao, E. Wong, J. Gardner, Black box adversarial prompting for foundation models, 2023, arXiv preprint arXiv:2302.04237.
[67] H.S. Shahgir, X. Kong, G.V. Steeg, Y. Dong, Asymmetric bias in text-to-image generation with adversarial attacks, 2023, arXiv preprint arXiv:2312.14440.
[68] C. Zhang, L. Wang, A. Liu, Revealing vulnerabilities in stable diffusion via targeted attacks, 2024, arXiv preprint arXiv:2401.08725.
[69] H. Liu, Y. Wu, S. Zhai, B. Yuan, N. 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, 2023, pp. 20585–20594.
[70] S. Kim, S. Jung, B. Kim, M. Choi, J. Shin, J. Lee, Towards safe self-distillation of internet-scale text-to-image diffusion models, in: ICML 2023 Workshop on Challenges in Deployable Generative AI, 2023.
[71] D. Arad, H. Orgad, Y. Belinkov, Refact: Updating text-to-image models by editing the text encoder, NAACL (2024).
[72] S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, R. Cucchiara, Safe-CLIP: Removing NSFW concepts from vision-and-language models, 2024, arXiv preprint arXiv:2311.16254, 2.

[73] 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: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[74] J. Ho, T. Salimans, Classifier-free diffusion guidance, 2022, arXiv preprint arXiv:2207.12598.
[75] W. Feng, X. He, T.-J. Fu, V. Jampani, A. Akula, P. Narayana, S. Basu, X.E. Wang, W.Y. Wang, Training-free structured diffusion guidance for compositional text-to-image synthesis, 2022, arXiv preprint arXiv:2212.05032.
[76] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, 2017, arXiv preprint arXiv:1706.06083.
[77] I.O. Tolstikhin, B.K. Sriperumbudur, B. Schölkopf, Minimax estimation of maximum mean discrepancy with radial kernels, Adv. Neural Inf. Process. Syst. 29 (2016).
[78] M. Li, Nsfw text classifier, 2022, https://huggingface.co/michellejieli/NSFW_text_classifier.
[79] OpenAI, DALL·e 3, 2023, https://openai.com/index/dall-e-3.
[80] J. Rando, D. Paleka, D. Lindner, L. Heim, F. Tramèr, Red-teaming the stable diffusion safety filter, 2022, arXiv preprint arXiv:2210.04610.
[81] D. Eriksson, M. Pearce, J. Gardner, R.D. Turner, M. Poloczek, Scalable global optimization via local Bayesian optimization, Adv. Neural Inf. Process. Syst. 32 (2019).
[82] Z.-Y. Chin, C.-M. Jiang, C.-C. Huang, P.-Y. Chen, W.-C. Chiu, Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts, ICML (2024).
[83] Y.-L. Tsai, C.-Y. Hsu, C. Xie, C.-H. Lin, J.-Y. Chen, B. Li, P.-Y. Chen, C.-M. Yu, C.-Y. Huang, Ring-A-Bell! How reliable are concept removal methods for diffusion models? ICLR (2024).
[84] N. Mehrabi, P. Goyal, C. Dupuy, Q. Hu, S. Ghosh, R. Zemel, K.-W. Chang, A. Galstyan, R. Gupta, Flirt: Feedback loop in-context red teaming, 2023, arXiv preprint arXiv:2308.04265.
[85] J. Ma, A. Cao, Z. Xiao, J. Zhang, C. Ye, J. Zhao, Jailbreaking prompt attack: A controllable adversarial attack against diffusion models, 2024, arXiv preprint arXiv:2404.02928.
[86] L. Struppek, D. Hintersdorf, F. Friedrich, P. Schramowski, K. Kersting, et al., Exploiting cultural biases via homoglyphs in text-to-image synthesis, J. Artificial Intelligence Res. 78 (2023) 1017–1068.
[87] R. Millière, Adversarial attacks on image generation with made-up words, 2022, arXiv preprint arXiv:2208.04135.
[88] Y. Liu, G. Yang, G. Deng, F. Chen, Y. Chen, L. Shi, T. Zhang, Y. Liu, Groot: Adversarial testing for generative text-to-image models with tree-based semantic transformation, 2024, arXiv preprint arXiv:2402.12100.
[89] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, Q. Zhang, K. Kreis, M. Aittala, T. Aila, S. Laine, et al., Ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers, 2022, arXiv preprint arXiv:2211.01324.
[90] E. Hoogeboom, J. Heek, T. Salimans, Simple diffusion: End-to-end diffusion for high resolution images, in: International Conference on Machine Learning, PMLR, 2023, pp. 13213–13232.
[91] Z. Tang, Z. Yang, C. Zhu, M. Zeng, M. Bansal, Any-to-any generation via composable diffusion, Adv. Neural Inf. Process. Syst. 36 (2024).
[92] L. Huang, D. Chen, Y. Liu, Y. Shen, D. Zhao, J. Zhou, Composer: Creative and controllable image synthesis with composable conditions, 2023, arXiv preprint arXiv:2302.09778.
[93] K. Chen, E. Xie, Z. Chen, L. Hong, Z. Li, D.-Y. Yeung, Integrating geometric control into text-to-image diffusion models for high-quality detection data generation via text prompt, 2023, arXiv preprint arXiv:2306.04607.
[94] Q. Phung, S. Ge, J.-B. Huang, Grounded text-to-image synthesis with attention refocusing, 2023, arXiv preprint arXiv:2306.05427.
[95] G. Couairon, M. Careil, M. Cord, S. Lathuilière, J. Verbeek, Zero-shot spatial layout conditioning for text-to-image diffusion models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2174–2183.
[96] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, S.Z. Li, A survey on generative diffusion models, IEEE Trans. Knowl. Data Eng. (2024).
[97] Z. Ni, L. Wei, J. Li, S. Tang, Y. Zhuang, Q. Tian, Degeneration-tuning: Using scrambled grid shield unwanted concepts from stable diffusion, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 8900–8909.
[98] S. Hong, J. Lee, S.S. Woo, All but one: Surgical concept erasing with model preservation in text-to-image diffusion models, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 21143–21151.
[99] Y. Wu, S. Zhou, M. Yang, L. Wang, W. Zhu, H. Chang, X. Zhou, X. Yang, Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient, 2024, arXiv preprint arXiv:2405.15304.
[100] C. Kim, K. Min, Y. Yang, RACE: Robust adversarial concept erasure for secure text-to-image diffusion model, 2024, arXiv preprint arXiv:2405.16341.
[101] C.-P. Huang, K.-P. Chang, C.-T. Tsai, Y.-H. Lai, Y.-C.F. Wang, Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers, 2023, arXiv preprint arXiv:2311.17717.
[102] Y. Zhang, X. Chen, J. Jia, Y. Zhang, C. Fan, J. Liu, M. Hong, K. Ding, S. Liu, Defensive unlearning with adversarial training for robust concept erasure in diffusion models, 2024, arXiv preprint arXiv:2405.15234.
[103] R. Chavhan, D. Li, T. Hospedales, ConceptPrune: Concept editing in diffusion models via skilled neuron pruning, 2024, arXiv preprint arXiv:2405.19237.
[104] Y. Yao, P. Wang, B. Tian, S. Cheng, Z. Li, S. Deng, H. Chen, N. Zhang, Editing large language models: Problems, methods, and opportunities, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 10222–10240.
[105] A. Heng, H. Soh, Selective amnesia: A continual learning approach to forgetting in deep generative models, Adv. Neural Inf. Process. Syst. 36 (2024).
[106] S. Lu, Z. Wang, L. Li, Y. Liu, A.W.-K. Kong, MACE: Mass concept erasure in diffusion models, 2024, arXiv preprint arXiv:2403.06135.
[107] E. Zhang, K. Wang, X. Xu, Z. Wang, H. Shi, Forget-me-not: Learning to forget in text-to-image diffusion models, 2023, arXiv preprint arXiv:2303.17591.
[108] M. Sun, Z. Liu, A. Bair, J.Z. Kolter, A simple and effective pruning approach for large language models, 2023, arXiv preprint arXiv:2306.11695.
[109] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019) 9.
[110] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Adv. Neural Inf. Process. Syst. 30 (2017).
[111] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, Adv. Neural Inf. Process. Syst. 29 (2016).
[112] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2818–2826.
[113] H. Alt, M. Godau, Computing the Fréchet distance between two polygonal curves, Internat. J. Comput. Geom. Appl. 5 (01n02) (1995) 75–91.
[114] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.
[115] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.
[116] S. Christoph, K. Andreas, C. Theo, V. Richard, T. Benjamin, R. Beaumont, LAION-COCO, 2024, https://laion.ai/blog/laion-coco/.
[117] Z.J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, D.H. Chau, Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models, 2022, arXiv preprint arXiv:2210.14896.
[118] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al., Laion-5b: An open large-scale dataset for training next generation image-text models, Adv. Neural Inf. Process. Syst. 35 (2022) 25278–25294.
[119] parrot zone, Image synthesis style studies, 2024, https://proximacentaurib.notion.site/parrot-zone-74a5c04d4feb4f12b52a41fc8750b205.
[120] H. Nick, K. Ihor, V. Dmitry, K. Dmytro, Giphy celebrity detector, 2024, https://github.com/Giphy/celeb-detection-oss.
[121] J. Quaye, A. Parrish, O. Inel, C. Rastogi, H.R. Kirk, M. Kahng, E. Van Liemt, M. Bartolo, J. Tsang, J. White, et al., Adversarial nibbler: An open red-teaming method for identifying diverse harms in text-to-image generation, in: The 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024, pp. 388–406.
[122] M. Brack, P. Schramowski, K. Kersting, Distilling adversarial prompts from safety benchmarks: Report for the adversarial nibbler challenge, 2023, arXiv preprint arXiv:2309.11575.
[123] S. Hong, X. Zheng, J. Chen, Y. Cheng, J. Wang, C. Zhang, Z. Wang, S.K.S. Yau, Z. Lin, L. Zhou, et al., Metagpt: Meta programming for multi-agent collaborative framework, 2023, arXiv preprint arXiv:2308.00352.
[124] W. Chen, Y. Su, J. Zuo, C. Yang, C. Yuan, C.-M. Chan, H. Yu, Y. Lu, Y.-H. Hung, C. Qian, et al., Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors, in: The Twelfth International Conference on Learning Representations, 2023.
[125] Q. Wu, G. Bansal, J. Zhang, Y. Wu, S. Zhang, E. Zhu, B. Li, L. Jiang, X. Zhang, C. Wang, Autogen: Enabling next-gen llm applications via multi-agent conversation framework, 2023, arXiv preprint arXiv:2308.08155.
[126] H. Zhang, W. Du, J. Shan, Q. Zhou, Y. Du, J.B. Tenenbaum, T. Shu, C. Gan, Building cooperative embodied agents modularly with large language models, 2023, arXiv preprint arXiv:2307.02485.
[127] M. Li, Z. Wang, Q.-L. Han, S.J. Taylor, K. Li, X. Liao, X. Liu, Influence maximization in multiagent systems by a graph embedding method: dealing with probabilistically unstable links, IEEE Trans. Cybern. 53 (9) (2023) 6004–6016.
[128] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, J. Jitsev, Reproducible scaling laws for contrastive language-image learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2818–2829.
[129] L. Bourtoule, V. Chandrasekaran, C.A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, N. Papernot, Machine unlearning, in: 2021 IEEE Symposium on Security and Privacy, SP, IEEE, 2021, pp. 141–159.
[130] A. Golatkar, A. Achille, S. Soatto, Eternal sunshine of the spotless net: Selective forgetting in deep networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9304–9312.