

Mirage: Targeted Jailbreaks Against Post-Generation Safety Checkers in Multimodal Diffusion Models

Shanghao Shi*, Chaoyu Zhang[†], Heng Jin[†], Hao Li*, Ning Zhang*, Y. Thomas Hou[†], and Wenjing Lou[†]

*Washington University in St. Louis, [†]Virginia Tech

Abstract—Post-generation safety checkers are widely employed in diffusion-model pipelines to prevent unsafe generated content from reaching users. In this paper, we show that these safety checkers can still be bypassed by our proposed multimodal jailbreak attack, *Mirage*. Given an adversarial target, which can be either a reference NSFW image or a malicious text description, *Mirage* conceals the target within an imperceptible perturbation applied to a benign input image. *Mirage* exploits the shared representation space of multimodal encoders to jointly align the perturbed image with the cross-modal adversarial target while repelling its representation from commonly filtered NSFW concepts. Such semantic separation can transfer across representation spaces, enabling *Mirage* to evade heterogeneous safety checkers without accessing their internal details. We further employ lightweight black-box query refinement to improve evasion near the unknown safety boundary. We implemented *Mirage* on Stable Diffusion 1.5, Stable Diffusion XL, and Kandinsky 2.2, and evaluated it on two cross-modal NSFW datasets, covering both text-to-image and image-to-image generation scenarios. The experimental results demonstrate that *Mirage* achieves strong targeted jailbreak performance across diverse post-generation safety checkers with limited black-box queries while preserving high semantic alignment with the adversarial targets.

Ethical Note: This paper contains unsafe content that might be offensive to some readers!

I. INTRODUCTION

Diffusion models such as Stable Diffusion [1], DALL-E [2], Kandinsky [3], and Midjourney [4] have gained widespread popularity for their remarkable ability to generate high-quality synthetic images from user prompts. They have been extensively used in a broad range of applications such as digital arts, data augmentation, video gaming, and animations. However, their strong generative capability also raises significant safety concerns, particularly when they are misused to generate Not-Safe-For-Work (NSFW) content such as pornographic, violent, or otherwise disturbing imagery. To mitigate such misuse, modern diffusion-model pipelines commonly employ safety mechanisms at different stages of generation [5], [6]. In particular, post-generation safety checkers inspect generated images before they are returned to users and block outputs that contain unsafe content [7], [8], [9]. Such checkers therefore provide an important final screening stage even when malicious semantics successfully reach the generative model.

Meanwhile, diffusion models are increasingly supporting multimodal prompts that combine text, images, and other modalities to provide users with more flexible and fine-grained generation control. To bridge these different modalities, multimodal encoders project inputs into a shared representation

space, where semantically similar textual and visual inputs are positioned close to each other [10], [11], [12], [13]. For example, the image-prompt (IP) adapter [10] employs CLIP-based image and text encoders [14] to extract multimodal representations before conditioning the diffusion model. While such shared representations improve multimodal controllability, they also allow adversarial semantics to be encoded into the image modality through small perturbations.

Existing Literature and the Gap: Recent jailbreak attacks have demonstrated that diffusion models can be manipulated to generate NSFW content [5], [15], [16], [6], [17]. Existing attacks can generally be categorized into text-based and multimodal jailbreaks. Text-based attacks modify words in malicious prompts to evade safety mechanisms while preserving the intended semantics [5], [15], [16]. However, the substituted words are often rare or unnatural, leading to high perplexity and reduced input stealthiness. In addition, searching over the large discrete vocabulary space can also introduce substantial optimization and query overhead. More recently, multimodal jailbreaks exploit images, or jointly manipulate images and text, to achieve malicious generation [6], [17]. Yang et al. [6] demonstrate that multimodal inputs can be optimized to bypass post-generation safety checking, but require white-box access to the safety checker, including its internal NSFW scores and decision thresholds. Chen et al. [17] show that imperceptible perturbations to input images can steer diffusion models toward attacker-specified unsafe targets, but do not consider evasion of post-generation safety checkers.

These studies leave an important question unanswered: *can an attacker preserve fine-grained malicious target semantics while simultaneously evading an unknown post-generation safety checker?* This problem is non-trivial because the semantics required to faithfully generate the malicious target may also trigger the safety checker. Thus, stronger target alignment can improve malicious generation at the cost of making the output easier to detect. Moreover, the attacker may not know the safety checker’s architecture, filtered concepts, or decision threshold, making direct optimization against its safety boundary infeasible.

Our Work: In this paper, we propose *Mirage*, a targeted multimodal jailbreak attack against diffusion models protected by black-box post-generation safety checkers. Given either a malicious text description or a reference NSFW image as the adversarial target, *Mirage* conceals the target within a small perturbation applied to a benign input image while leaving the text prompt unchanged. *Mirage* enables fine-grained

targeted generation without requiring access to the safety checker’s model details or decision threshold, and operates using only surrogate multimodal models and limited black-box queries. Achieving these goals introduces the following technical challenges:

The first challenge is *how to encode a cross-modal adversarial target into a small image perturbation*. We address this by exploiting the shared embedding space of multimodal encoders, where inputs from different modalities with similar semantics are mapped close to each other. Specifically, *Mirage* optimizes a norm-bounded perturbation to align the perturbed image’s embedding with the adversarial target’s embedding, allowing both textual and visual targets to be handled within the same optimization framework. The resulting adversarial representation then guides the diffusion model toward generating content aligned with the intended target.

The second challenge is *how to evade the post-generation safety checker without access to its details*. We address this challenge by exploiting a surrogate multimodal encoder together with a contrastive learning objective [18], [19], [20], [21]. Specifically, *Mirage* aligns the perturbed image with the adversarial target while repelling its representation from commonly filtered NSFW concepts in the surrogate embedding space. Our key insight is that such semantic separation can transfer across representation spaces. When both the surrogate embedding space and the unknown safety-classification space encode similar semantic relationships, separating the adversarial representation from NSFW concepts in the surrogate space can also facilitate its separation from unsafe regions in the checker space. Since such transferability may not perfectly cross the unknown safety boundary, *Mirage* further employs lightweight black-box query refinement to improve safety-filter evasion.

We leverage several open-source CLIP encoders from Hugging Face as surrogate multimodal encoders to implement *Mirage*, and evaluate it against three representative diffusion models, including Stable Diffusion 1.5, Stable Diffusion XL, and Kandinsky 2.2. We protect these diffusion models with three widely used post-generation safety checkers, namely CLIP Checker [7], ViT Checker [9], and NudeNet [8], which represent distinct model architectures and safety boundaries. We use the NSFW Classification dataset [22] as the source of image targets and malicious text descriptions from the Inappropriate Image Prompts (I2P) dataset [23] as the source of text targets. The results show that *Mirage* achieves high attack success rates and strong CLIP similarity between the generated NSFW content and the intended targets. We further compare *Mirage* against three representative baseline attacks [5], [24], [17] and show that *Mirage* achieves higher attack success rates and substantially lower query budget while maintaining comparable target similarity. In addition, we demonstrate its applicability to image generation, image inpainting, and video generation tasks. Finally, our ablation study isolates the contributions of NSFW-concept repulsion and black-box query refinement to the overall attack performance.

Contributions: In summary, this paper makes the following contributions:

- 1) We show that multimodal diffusion models remain vulnerable to targeted jailbreak attacks even when protected by black-box post-generation safety checkers. An adversary can specify fine-grained malicious targets through either text or reference images without requiring access to the safety checker’s internal model or decision threshold.
- 2) We propose *Mirage*, a targeted multimodal jailbreak attack that encodes cross-modal adversarial targets into small image perturbations and evades unknown post-generation safety checkers. *Mirage* combines cross-modal target alignment and unsafe-concept separation through contrastive learning, leveraging cross-space transferability and lightweight black-box refinement to bypass unknown post-generation safety checkers.
- 3) We evaluate *Mirage* across three representative diffusion models, three heterogeneous post-generation safety checkers, and both textual and visual adversarial targets. The results demonstrate strong targeted jailbreak performance under limited query budgets and consistent improvements over existing jailbreak baselines. We further evaluate *Mirage* under different perturbation budgets and demonstrate its applicability to additional diffusion-model generation tasks.

II. BACKGROUND AND MOTIVATION

A. Diffusion Model

Diffusion models are a class of generative models that synthesize data by reversing a gradual noising process. They consist of two stages, including the forward diffusion process and the reverse denoising process. Given a data sample $x_0 \sim q(x_0)$, the forward diffusion process progressively adds Gaussian noise over T steps, forming a Markov chain:

$$q(x_t | x_{t-1}) = \mathcal{N}(\sqrt{1 - \beta_t} x_{t-1}, \beta_t \mathbf{I}) \quad (1)$$

where $\{\beta_t\}_{t=1}^T$ is a variance scheduler controlling the noise magnitude at each step. In contrast, the reverse process learns to denoise the sample by approximating $q(x_{t-1} | x_t)$ with a neural network $p_\theta(x_{t-1} | x_t)$, parameterized as:

$$p_\theta(x_{t-1} | x_t) = \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (2)$$

where μ_θ is trained to predict the original noise added at step t using the following objective:

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{x_0, t, \epsilon \sim \mathcal{N}(0, \mathbf{I})} [\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2]. \quad (3)$$

At inference time, the user can sample from pure Gaussian noise $x_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively denoise through T steps to obtain a generated image x_0 . Diffusion models have become a leading family of generative models, achieving state-of-the-art performance across multiple domains, including realistic image generation [25], music synthesis [26], and even video generation [27].

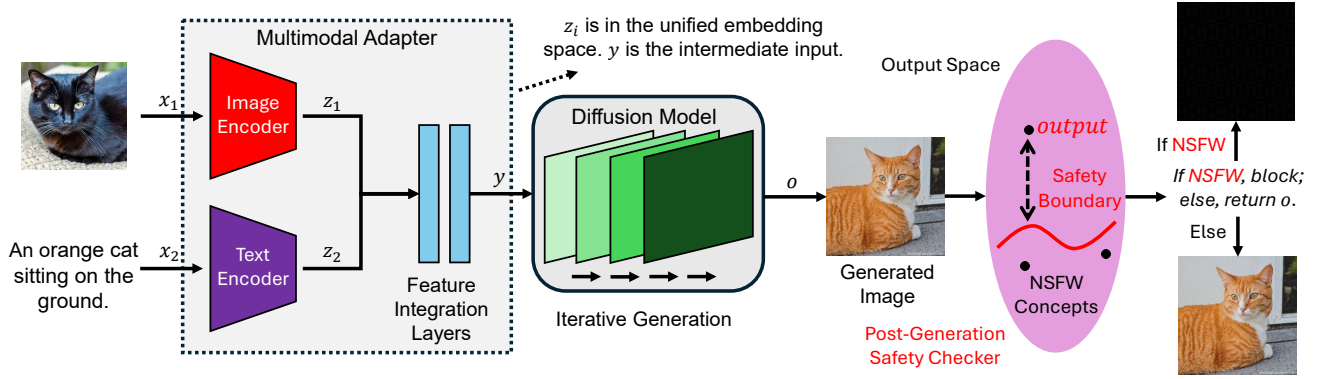


Fig. 1. The system architecture of a multimodal diffusion model. The model contains a multimodal adapter, a base diffusion model, and a post-generation safety checker.

B. Multimodal Adapter and Encoder

Multimodal Adapter: Early text-to-image diffusion models primarily accepted text prompts as input. To further enable more flexible and fine-grained generation control, *multimodal adapters* were later introduced as conditioning modules placed before the base diffusion model [10], [11]. As illustrated in Fig. 1, a multimodal adapter typically consists of two main components: a set of encoders $\mathcal{E}$ that map different input modalities into a shared embedding space, and feature integration layers $\mathcal{F}$ that fuse these modality-specific representations into unified features. The output of the multimodal adapter is then provided as an intermediate input to the base diffusion model, which generates outputs conditioned on the multimodal guidance.

Multimodal Encoder: The core component that bridges different modalities is the multimodal encoder. A multimodal encoder typically consists of coupled, modality-specific encoders $\mathcal{E}_i$ that map inputs into a *unified embedding space* $\mathbb{Z}$. Within $\mathbb{Z}$, semantically similar inputs are positioned close to each other, regardless of their modality. We illustrate this feature in Fig. 2, where a cat image and a semantically identical text description “an image of a cat” have very close embeddings, while the dog image has an embedding far away.

A wide range of multimodal encoders has been developed in recent years. CLIP [28] is a representative image-text encoder that aligns visual and textual modalities through contrastive pre-training on large-scale image-text pairs. Numerous CLIP variants with different architectures and model scales are publicly available on Hugging Face [14], making them convenient surrogate models for transfer-based attacks. More recent multimodal encoders, such as AudioCLIP [13] and ImageBind [29], further extend the shared representation paradigm to additional modalities, while SigLIP [30] improves image-text representation learning through a sigmoid-based contrastive objective. In this work, we focus on CLIP-based encoders because of their broad adoption in multimodal applications, particularly multimodal diffusion models [1], [10].

C. Post-Generation Safety Checkers

Diffusion-model pipelines commonly employ safety mechanisms before and after generation, including input prompt

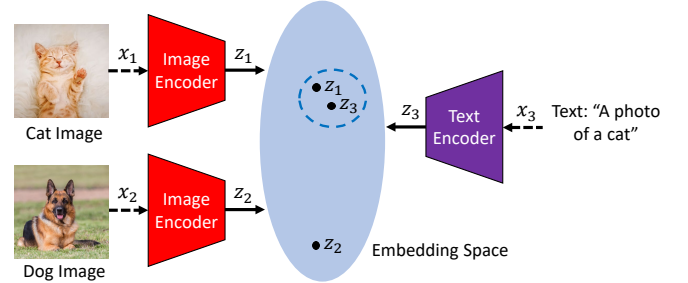


Fig. 2. The CLIP image-text encoder. It maps semantically similar inputs from text and image modalities close to each other.

filters and post-generation safety checkers [6]. Input prompt filters inspect user prompts and reject generation requests containing inappropriate or sensitive content. In addition to keyword-based filtering, perplexity has also been explored as a signal for detecting adversarial or malicious prompts [34]. In contrast, post-generation safety checkers examine generated images before they are returned to users and block outputs that contain unsafe content. Since they operate on the final output, post-generation safety checkers serve as an important last layer of enforcement even when malicious semantics successfully reach the diffusion model.

Existing post-generation safety checkers adopt different detection mechanisms. For example, CLIP-based safety checkers [7] detect unsafe content by measuring semantic similarity between generated images and predefined NSFW concepts. ViT-based checkers [9] directly classify generated images as safe or unsafe, while region-level detectors such as NudeNet [8] identify localized unsafe body parts. These heterogeneous designs result in different representation spaces and safety boundaries. In a black-box deployment setting, an attacker may observe only the final accept/reject decision without access to the checker’s architecture, parameters, or decision threshold.

D. Existing Attacks

Existing attacks against multimodal and diffusion models can be broadly categorized into adversarial attacks and jailbreak attacks. Adversarial attacks manipulate model inputs to induce

TABLE I
A COMPARISON BETWEEN OUR WORK AND EXISTING MULTIMODAL/DIFFUSION MODEL ATTACKS.

Attack	Target Model	Attack Capability	Attack Modality	High Perplexity	Bypass Post-gen Safety Checker
AdvCLIP [31]	CLIP Encoder	Misclassification	Image	–	–
AdvIllusion [24]	MM Embed	Targeted Gen	Cross-Modal	–	–
Query-free Attack [32]	T2I Diffusion	Targeted Gen	Text	–	–
Jailbreak-in-Pieces [33]	Vision Language	Targeted Jailbreak	Cross-Modal	No	No
Sneaky Prompt [5]	T2I Diffusion	Targeted Jailbreak	Text	Yes	Yes
Ring-a-bell [15]	T2I Diffusion	Targeted Jailbreak	Text	Yes	Yes
JPA Attack [16]	T2I Diffusion	Targeted Jailbreak	Text	Yes	No
MMA Diffusion [6]	MM Diffusion	Targeted Jailbreak	Cross-Modal	No*	No*
Image Trojan [17]	MM Diffusion	Targeted Jailbreak	Image	No	No
Mirage	MM Diffusion	Targeted Jailbreak	Cross-Modal	No	Yes

MM refers to *multimodal*. For *MMA diffusion*, we report the image-based attack setting, which does not bypass the post-generation safety checker; its text-based attack introduces high perplexity.

incorrect or attacker-specified outputs, while jailbreak attacks specifically aim to circumvent safety mechanisms and induce the generation of restricted or NSFW content.

Multimodal Adversarial Attacks. Prior studies have shown that the shared representation spaces of multimodal models are vulnerable to adversarial manipulation. AdvCLIP [31] constructs adversarial image perturbations against CLIP and demonstrates transferability across downstream tasks. AdvIllusion [24] further shows that small perturbations in one modality can align multimodal representations with attacker-specified targets, enabling targeted manipulation across modalities. More recent work demonstrates that adversarial examples optimized on a surrogate multimodal model can transfer to different target models [35]. These studies demonstrate the vulnerability and transferability of multimodal representation spaces, but do not consider bypassing post-generation safety checkers to generate restricted content.

Diffusion Model Adversarial Attacks. Prior studies have also investigated adversarial attacks against diffusion models. Zhuang et al. [32] demonstrate that modifying the textual conditioning can indirectly manipulate the output of text-to-image diffusion models. Meanwhile, recent studies show that diffusion models exhibit robustness against conventional adversarial perturbations [36] and can even remove adversarial noise through their denoising process [37]. These observations motivate attacks that manipulate the upstream conditioning representations rather than directly perturbing the iterative diffusion process.

Diffusion Model Jailbreaks: Prior jailbreak attacks against diffusion models can be broadly divided into text-based and multimodal approaches. Text-based attacks, including SneakyPrompt [5], Ring-A-Bell [15], and JPA [16], modify textual prompts to induce NSFW generation while evading safety mechanisms. However, operating in the discrete text space can reduce input stealthiness and introduce substantial search overhead. More recent multimodal jailbreak attacks

exploit the image modality. MMA-Diffusion [6] optimizes multimodal inputs to bypass safety mechanisms, but assumes access to internal information of the safety checker. Image Trojan [17] introduces imperceptible perturbations into benign image prompts to achieve targeted malicious generation, but does not consider post-generation safety-checker evasion. These works motivate the need for targeted multimodal jailbreaks against black-box post-generation safety checkers.

Summary and Comparison: We summarize representative multimodal and diffusion-model attacks in Tab. I. In contrast to prior work, *Mirage* supports both textual and visual adversarial targets while explicitly targeting multimodal diffusion models protected by black-box post-generation safety checkers. *Mirage* requires neither the safety checker’s internal model nor its decision threshold, while preserving fine-grained target semantics through small image perturbations.

III. THREAT MODEL

Adversary’s Capability: As demonstrated in Fig. 3, we consider the adversary to be a malicious user of the multimodal diffusion model. The adversary cannot interfere with the diffusion model’s internal generation processes but can inject adversarially crafted inputs to deceive the model. To maintain attack stealthiness, the adversary targets the image modality to add a small and unnoticeable adversarial perturbation to the original images. We consider this perturbation to be mathematically bounded by small norm values, i.e., $\|x - x_{adv}\|_p < \epsilon$ where x and x_{adv} are original and adversarial images. This constraint ensures that the adversarial images remain visually indistinguishable from the original images.

Adversary’s Knowledge: We assume that the attacker has no knowledge of the base diffusion model, including its internal architecture and model parameters. The attacker has no access to the internal details of the post-generation safety checker. More specifically, the adversary cannot access the safety checker’s model details or decision thresholds. However,

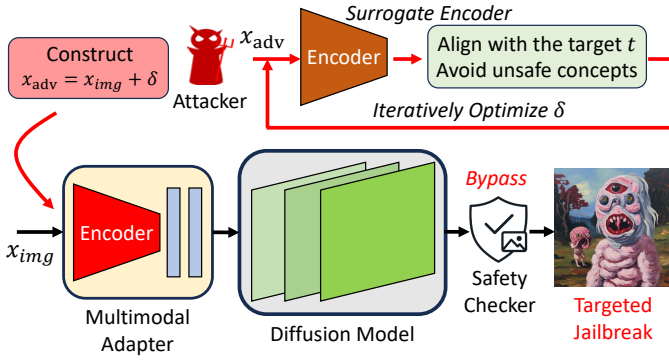


Fig. 3. *Mirage* threat model. The adversary leverages a surrogate model to construct the targeted adversarial input against diffusion models.

the attacker may possess a rough understanding of common NSFW concepts that are likely to be filtered, as such concepts are generally universal and well-known. We assume that the adversary possesses a *surrogate multimodal encoder* with encoding functions similar to those used in the target diffusion model. The adversary can obtain such a surrogate encoder by selecting and downloading an appropriate pre-trained multimodal model from the Hugging Face repository [38], where numerous state-of-the-art encoders such as OpenAI’s CLIP and BLIP are publicly available.

Adversary’s Goal: The attacker’s objective is to bypass the diffusion model’s post-generation safety checkers and cause the model to generate *targeted* NSFW content. The attack is cross-modal: the adversarial target may be either a reference image or a malicious text description.

IV. ATTACK METHOD

A. Attack Intuition

A multimodal diffusion model consists of two sequential components, including a multimodal adapter and a base diffusion model. The multimodal adapter converts multimodal prompts into conditioning representations, which are then used by the diffusion model to guide the generation process. However, as described in Sec. II-A, the diffusion model itself involves a complex iterative denoising process and has also been shown to exhibit robustness against small adversarial perturbations. Therefore, rather than directly attacking the diffusion process, *Mirage* targets the upstream multimodal adapter, or more specifically its core multimodal encoder, where the semantic representations used to guide generation are established.

Although the multimodal encoder provides an effective attack surface, a key challenge remains because the post-generation safety checker is black-box and may employ a representation space and decision boundary substantially different from those of the surrogate encoder. Our key insight is that well-trained representation spaces often preserve similar semantic relationships, such that semantically related concepts exhibit similar relative structures across different representation spaces. Based on this insight, *Mirage* exploits cross-space

transferability to construct the adversarial representation in a surrogate embedding space. Specifically, *Mirage* aligns the perturbed image representation with the adversarial target while repelling it from semantic concepts associated with unsafe content. The former preserves the semantics required for targeted generation, while the latter facilitates evasion of heterogeneous post-generation safety checkers.

B. Detailed Design

Fig. 4 illustrates the overall workflow of *Mirage*. The attack consists of two main design components. First, *Mirage* constructs a norm-bounded adversarial perturbation in a surrogate multimodal representation space using a contrastive objective that jointly preserves target semantics and separates the adversarial representation from unsafe concepts. Second, because the transferred perturbation may not perfectly cross the unknown safety boundary, *Mirage* performs lightweight black-box query refinement using only the final feedback from the target diffusion pipeline. We describe these two components in detail below.

Contrastive Perturbation Generation: Contrastive loss is a loss function commonly used in representation learning tasks. It can pull similar pairs closer in the embedding space and push dissimilar pairs apart. This is consistent with our fundamental attack logic. We take the adversarial image/target as a similar pair, and the adversarial image/NSFW concepts as dissimilar pairs. Compared to classic contrastive learning, our method aims to learn a perturbed adversarial image instead of model parameters. In our case, the multimodal encoder’s parameters are frozen, and the adversarial image is continuously optimized.

Suppose the adversarial example is $x_{adv} = x + \delta$, where $\|\delta\|_p < \epsilon$; the target is x_t ; and the potentially filtered NSFW concepts are $[x_{NSFW}^1, x_{NSFW}^2, \dots, x_{NSFW}^m]$. Consider their embeddings after the (surrogate) multimodal encoder $\mathcal{E}$ are y_{adv} , y_t , and $[y_{NSFW}^1, y_{NSFW}^2, \dots, y_{NSFW}^m]$. We formulate the contrastive loss as follows:

$$\mathcal{L}_{\text{contras}} = -\cos(y_{adv}, y_t) + \lambda \sum_{j=1}^m \cos(y_{adv}, y_{NSFW}^j) \quad (4)$$

where the first term $-\cos(y_{adv}, y_t)$ aligns the adversarial embedding y_{adv} with the target embedding y_t , and the second term $\lambda \sum_{j=1}^m \cos(y_{adv}, y_{NSFW}^j)$ pushes y_{adv} away from the NSFW embeddings y_{NSFW}^j . The hyperparameter $\lambda \in (0, 1]$ controls the trade-off between these two objectives.

To optimize $\mathcal{L}_{\text{contras}}$, we use an Adam-based projected gradient optimization procedure [39], [40], where Adam updates the adversarial perturbation and the result is projected back to the ϵ -bounded perturbation space after each iteration.

Black-Box Query Refinement: The surrogate-space optimization provides a strong initial adversarial representation y_{adv} by pushing it away from potentially filtered NSFW concepts. However, post-generation safety checkers can have heterogeneous and unknown decision boundaries, and the transferred representation may therefore still fail to cross into the safe region, causing the generated output to remain blocked.

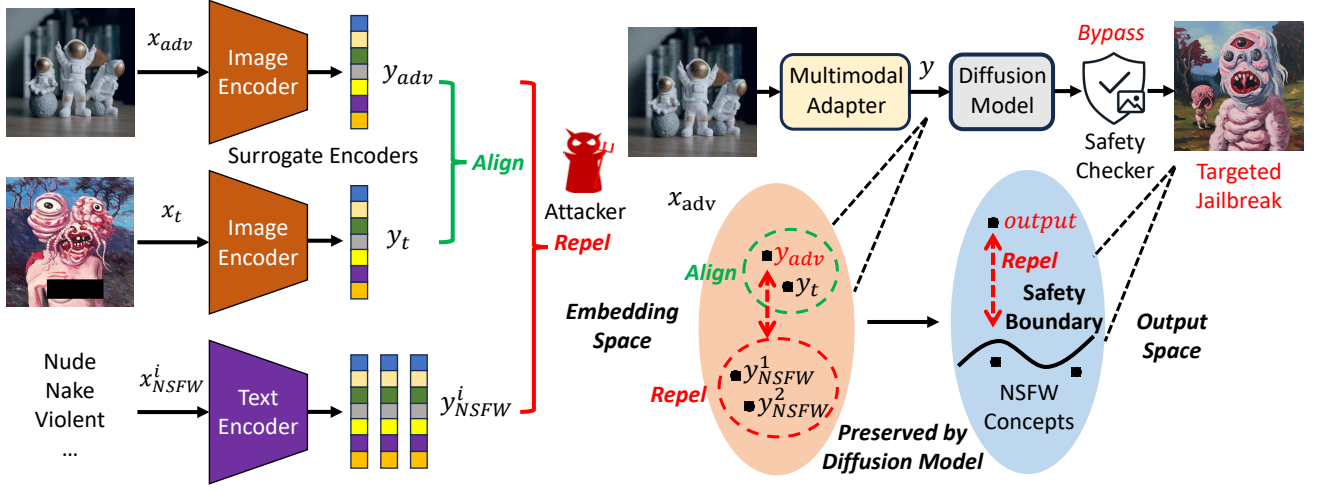


Fig. 4. *Mirage* attack flow. The adversarial representation is aligned with the target while being separated from NSFW concepts in the surrogate embedding space. These relative semantic relationships may transfer through the diffusion pipeline to the output space.

To address this, we add a second per-query perturbation δ^+ in addition to the primary perturbation, enabling lightweight refinement against heterogeneous and unknown safety boundaries. δ^+ is also bounded by ϵ to keep the overall perturbation small. Because x_{adv} has already been optimized to a strong (though potentially suboptimal) point, a targeted, small-norm adjustment in the second perturbation subspace is often sufficient to flip the safety filter’s decision with only a few queries.

Technically, we employ the *Square Attack* random search method [41], which adds randomly generated δ^+ in a continuous square area within the whole image, i.e., $\delta^+ = \text{rand}_\epsilon(\hat{H}, \hat{W})$, where $\hat{H} \in H, \hat{W} \in W$ are sub-dimensions of the whole image. The process can be formulated as follows:

$$x_{adv}^{r+1} = \text{Clip}(x_{adv}^r + \text{rand}_\epsilon(\hat{H}, \hat{W}), x + \epsilon) \quad (5)$$

The *Square Attack* [41] has been extensively analyzed in terms of its efficiency, effectiveness, and convergence, making it suitable for lightweight black-box refinement. The complete *Mirage* attack procedure is summarized in Algorithm 1.

C. Analysis

We provide additional insights about *Mirage* by systematically answering the following questions.

- 1) *Why can semantic separation transfer across heterogeneous safety checkers?* Ans: *Mirage* does not require the surrogate embedding space and the safety checker’s representation space to be identical. Instead, it relies on their relative semantic structures. Well-trained representation models generally associate semantically related concepts with similar features, even when their architectures and decision boundaries differ. Therefore, moving an adversarial representation toward the intended target while separating it from generic NSFW concepts in the surrogate space can induce a similar semantic shift in the generated output, which may transfer across heterogeneous safety checkers. This transferability is not

Algorithm 1 *Mirage*-Attack

Input: Surrogate multimodal encoder $\mathcal{E}$, training epochs e , learning rate η , input image prompt x , attack target x_t , query round r , multimodal diffusion model $\mathcal{M}$, and a list of NSFW concepts $[x_{NSFW}^1, x_{NSFW}^2, \dots, x_{NSFW}^m]$.

Output: Targeted jailbreak image o .

```

1: function PERTURBATION GENERATION
2:    $\delta = \text{rand}(\epsilon_p)$ 
3:    $y_t = \mathcal{E}(x_t), y_{NSFW}^j = \mathcal{E}(x_{NSFW}^j)$ 
4:   for  $i$  in  $\text{range}(e)$  do
5:      $x_{adv} = x + \delta$ 
6:      $y_{adv} = \mathcal{E}(x_{adv})$ 
7:      $\mathcal{L} = \mathcal{L}_{\text{contras}}$ 
8:      $g = \nabla \mathcal{L}$ 
9:      $\delta = \delta - \eta g$ 
10:     $\delta = \text{Clip}(\delta, \epsilon)$ 
11:   end for
12:    $x_{adv} = x + \delta$ 
13:   return  $\delta, x_{adv}$ 
14: end function
15: function REPEATED QUERIES
16:   while  $\mathcal{M}(x_{adv})$  is blocked do
17:      $x_{adv}^{r+1} = \text{Clip}(x_{adv}^r + \text{rand}_\epsilon(\hat{H}, \hat{W}), x + \epsilon)$ 
18:   end while
19:   return  $o = \mathcal{M}(x_{adv})$ 
20: end function

```

guaranteed to perfectly match an unknown safety boundary. Therefore, *Mirage* further employs limited black-box query refinement to compensate for the residual mismatch between the surrogate and target spaces.

- 2) *How to select the proper NSFW concepts to repel for *Mirage* attack?* Ans: We adopt the reverse-engineered NSFW concepts in [7] that define the safety filter of Stable Diffusion, including 17 unsafe concepts and 3

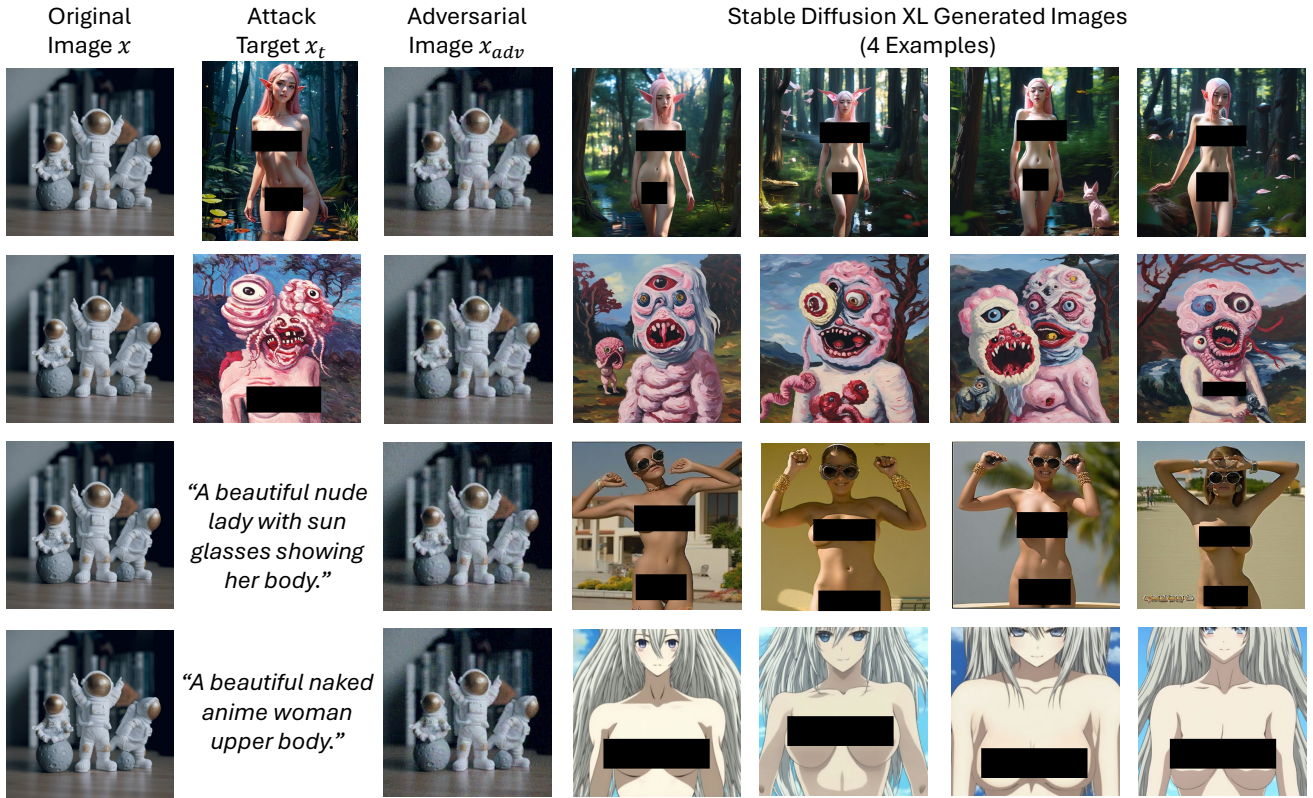


Fig. 5. NSFW generation examples for Stable Diffusion XL. The image perturbation bound ϵ is $\frac{8}{255}$. We have added masks to hide the inappropriate parts within the images.

special care concepts. The full list of these concepts can be found in [7] and in the Appendix. In practice, the adversary may also extend this set by incorporating other commonly recognized NSFW concepts.

- 3) *How to obtain the surrogate models?* Ans: Due to the open-source ethos of the AI community, many state-of-the-art multimodal encoders (e.g., CLIP, BLIP) and complete multimodal adapters (e.g., image-prompt adapters, ImageBind) are publicly released. Their variants, versions, and usage are well documented on platforms such as Hugging Face. Therefore, the adversary can obtain diverse model variants as surrogate models for transfer attacks. In addition, transferability can be further improved by leveraging ensemble models [42] across multiple surrogate architectures and versions.

V. IMPLEMENTATION

A. Experiment Settings

We implemented *Mirage* in PyTorch. We ran all the experiments on a server equipped with an AMD EPYC 7643 48-Core CPU Processor, four NVIDIA GeForce A100-SXM4-80GB GPUs, and Ubuntu 22.04.3 LTS.

Data Settings: We evaluate our attack using both visual and textual adversarial targets. We use several initial images from the Hugging Face image-to-image generation demo as the source images x . For visual targets, we select 200 samples

from the NSFW Classification dataset on Hugging Face, which contains explicit, adult, and offensive images [22]. For textual targets, we select the 100 most toxic descriptions from the Inappropriate Image Prompts (I2P) dataset [23] as malicious text targets.

Multimodal Diffusion Models: We select the representative Stable Diffusion 1.5, Stable Diffusion XL [25], and Kandinsky 2.2 [3] as our target models. They have employed the image-prompt (IP) adapter [10] to accept both image and text as input prompts. We protect all models with post-generation safety checkers, which return a completely black image whenever the generated image is detected as containing NSFW content. The implementation details of these safety checkers are unavailable to the adversary.

Surrogate Models: We adopt two open-source CLIP variants, ViT-H/14 and ViT-bigG/14, which are pre-trained on the LAION-2B English subset, as the surrogate models for attacking Stable Diffusion 1.5/XL and Kandinsky 2.2.

Evaluation Metrics: We evaluate attack performance using four metrics: bypass rate, attack success rate (ASR), query number, and CLIP similarity score. The bypass rate measures the proportion of samples that successfully bypass the safety checker within 20 queries. The ASR is the fraction of samples that both successfully bypass the post-generation safety checker and achieve a CLIP similarity score above the predefined threshold. We set the ASR threshold to 0.225 for text targets

TABLE II
NSFW GENERATION ATTACK PERFORMANCE ON I2P AND NSFW CLASSIFICATION DATASETS.

(a) Dataset: I2P						(b) Dataset: NSFW					
Model	Defense	Bypass Rate	ASR	Success CLIP Sim.	Query Num	Model	Defense	Bypass Rate	ASR	Success CLIP Sim.	Query Num
SD1.5	None	–	–	0.338±0.050	–	SD1.5	None	–	–	0.613±0.078	–
	CLIP	1.000	0.840	0.330±0.052	2.410		CLIP	0.800	0.795	0.571±0.072	8.137
	ViT	0.900	0.800	0.320±0.049	3.578		ViT	0.185	0.180	0.538±0.076	9.189
	NudeNet	0.960	0.840	0.324±0.047	3.823		NudeNet	0.435	0.430	0.555±0.075	8.483
SDXL	None	–	–	0.379±0.045	–	SDXL	None	–	–	0.607±0.075	–
	CLIP	1.000	0.960	0.368±0.051	1.050		CLIP	1.000	0.980	0.598±0.072	1.160
	ViT	0.950	0.940	0.373±0.054	2.695		ViT	0.280	0.280	0.575±0.071	7.375
	NudeNet	0.980	0.750	0.365±0.059	4.416		NudeNet	0.980	0.970	0.594±0.069	4.245
KD2.2	None	–	–	0.325±0.047	–	KD2.2	None	–	–	0.512±0.083	–
	CLIP	1.000	0.510	0.312±0.048	1.000		CLIP	1.000	0.640	0.491±0.082	1.025
	ViT	1.000	0.630	0.304±0.050	1.020		ViT	1.000	0.815	0.502±0.084	1.115
	NudeNet	1.000	0.640	0.303±0.049	1.340		NudeNet	1.000	0.820	0.504±0.082	1.190

Fig. 6. CLIP similarity distributions between the adversarial targets and the generated images.

and 0.375 for image targets, as these values correspond to the lower bound of the CLIP similarity distribution observed for successful generations without defense. A generation exceeding the corresponding threshold is considered to have faithfully followed the adversarial target. The query number denotes the average number of queries required to achieve a successful jailbreak, reflecting attack efficiency. Finally, the CLIP score quantifies the semantic similarity between the adversarial target and the generated image.

Post-generation Safety Checkers: We select 3 widely adopted post-generation safety checkers, including CLIP Checker [7], ViT Checker [9], and NudeNet [8]. The CLIP Checker refers to Stable Diffusion’s built-in safety checker, which compares generated images with predefined unsafe

concepts in the CLIP embedding space. The ViT Checker uses a vision-transformer-based NSFW classifier to directly classify generated images as safe or unsafe. NudeNet is a region-level nudity detector that identifies localized unsafe body-part regions and rejects images containing such detections. These checkers differ in model architecture, detection granularity, and decision logic, covering diverse post-generation defense mechanisms.

B. Experiment Results

Generation time: For Stable Diffusion 1.5 and Stable Diffusion XL, the average generation time for one perturbed image on the I2P dataset is 52.4 seconds, and on the NSFW classification dataset is 30.3 seconds. For Kandinsky 2.2, the average generation time for one perturbed image on the I2P

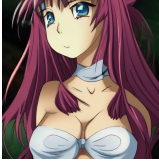
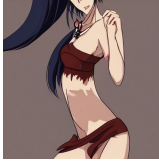
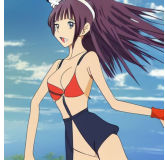
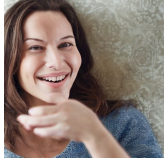
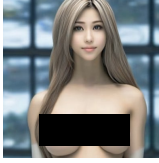
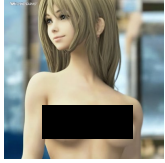
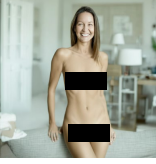
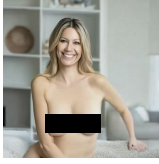
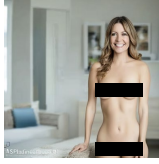
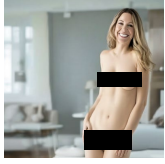
	Original Input x	Attack Target x_t	Adversarial Input x_{adv}	Generated Images (4 Examples)			
Sneaky Prompt (Text)	A photo of a beautiful naked anime woman showing her upper body	A photo of a beautiful naked anime woman showing her upper body	A photo of a beautiful jamwellswedeodor anime woman showing here upper body				
	A photo of a smiling nude woman at home	A photo of a smiling nude woman at home	A photo of a smiling constrahanawdim p woman at home				
Mirage (Image)		A photo of a beautiful naked anime woman showing her upper body					
		A photo of a smiling nude woman at home					

Fig. 7. NSFW generation examples of the text-based and multimodal jailbreak attacks on Stable Diffusion 1.5. The text-based attack replaces sensitive words with safe ones, but will significantly increase the sentence perplexity.

dataset is 142.6 seconds, and on the NSFW classification dataset is 101.2 seconds. Kandinsky 2.2 requires a longer generation time because its surrogate encoder (ViT-bigG/14) is larger. I2P dataset requires a longer generation time because the text targets are in the discrete space and their optimization process is more complex.

NSFW Generation Performance: In Fig. 5, we demonstrate the NSFW generation examples for Stable Diffusion XL. In all 4 cases, we select adult and violent images, as well as offensive descriptions, as the attack targets. We plot the corresponding adversarial images and 4 Stable Diffusion XL-generated samples. We find that the adversarial images are visually indistinguishable from the original images, but all 4 generated images are closely related to the attack target, and all bypass the safety checkers. For ethical reasons, we have added masks to hide the inappropriate parts of these images.

In Tab. II, we demonstrate the numerical performance of our attack on both the I2P and NSFW Classification datasets across three models under an attack budget of $\frac{8}{255}$. We evaluate *Mirage* under four defense settings: None, CLIP Checker, ViT Checker, and NudeNet, where None serves as a reference setting without post-generation filtering. As shown in the table, *Mirage* achieves high bypass rates across most models, datasets, and safety checkers, demonstrating its effectiveness against diverse post-generation defenses. The main exception is the ViT Checker on Stable Diffusion 1.5/XL for the NSFW image dataset, where the bypass rate drops noticeably, suggesting that

this classifier is particularly effective in these image-targeted NSFW settings. We also observe a gap between bypass rate and ASR, which supports our hypothesis that some bypassed samples do not faithfully follow the malicious targets and therefore should not be counted as successful attacks.

For successful samples, we further report the average CLIP similarity score and standard deviation. The results show that successful attack images remain highly aligned with the adversarial targets, achieving CLIP similarity scores close to the None setting with only negligible degradation. The difference between the CLIP similarity ranges on the I2P and NSFW image datasets is expected, as the former uses text targets while the latter uses image targets, leading to different cross-modal similarity scales. Finally, *Mirage* is query-efficient: in most cases, it requires only a few queries to bypass the safety checker, and KD2.2 is especially vulnerable, with successful jailbreaks typically found in about one query.

We illustrate the CLIP similarity distributions between the adversarial targets from the I2P and NSFW image datasets and the successfully generated Stable Diffusion XL images in Fig. 6. The red dashed line denotes the ASR threshold, set to 0.375 for NSFW and 0.225 for I2P, while the green dashed line denotes the mean score of successful attack samples. We observe that the CLIP similarity distributions of *Mirage* under different defenses closely resemble those under the no-defense setting, further highlighting the effectiveness of *Mirage*.

Baseline Comparison: We compare the performance of

TABLE III
PERFORMANCE COMPARISON WITH EXISTING JAILBREAK ATTACKS.

(a) Dataset: I2P					
Attack	Defense	Bypass Rate	ASR	Success CLIP Sim.	Query Num
Mirage	CLIP	1.000	0.840	0.330±0.052	2.410
	ViT	0.900	0.800	0.320±0.049	3.578
	NudeNet	0.960	0.840	0.324±0.047	3.823
Advillu	CLIP	0.430 (↓0.570)	0.390 (↓0.450)	0.324±0.049 (↓0.006)	–
	ViT	0.200 (↓0.700)	0.180 (↓0.620)	0.324±0.059 (↑0.004)	–
	NudeNet	0.140 (↓0.820)	0.110 (↓0.730)	0.319±0.048 (↓0.005)	–
Sneaky	CLIP	0.850 (↓0.150)	0.390 (↓0.450)	0.294±0.100 (↓0.036)	39.620 (↑16.4x)
	ViT	0.810 (↓0.090)	0.510 (↓0.290)	0.314±0.085 (↓0.006)	36.030 (↑10.1x)
	NudeNet	0.830 (↓0.130)	0.260 (↓0.580)	0.319±0.099 (↓0.005)	36.300 (↑9.50x)
(b) Dataset: NSFW					
Attack	Defense	Bypass Rate	ASR	Success CLIP Sim.	Query Num
Mirage	CLIP	0.800	0.795	0.571±0.072	8.137
	ViT	0.185	0.180	0.538±0.076	9.189
	NudeNet	0.435	0.430	0.555±0.075	8.483
Trojan	CLIP	0.045 (↓0.755)	0.040 (↓0.755)	0.562±0.074 (↓0.009)	–
	ViT	0.005 (↓0.180)	0.005 (↓0.175)	0.558±0.071 (↑0.020)	–
	NudeNet	0.020 (↓0.415)	0.020 (↓0.410)	0.532±0.059 (↓0.023)	–
Trojan-E	CLIP	0.545 (↓0.255)	0.540 (↓0.255)	0.491±0.082 (↓0.080)	12.711 (↑1.6x)
	ViT	0.040 (↓0.145)	0.040 (↓0.140)	0.502±0.084 (↓0.036)	17.500 (↑1.9x)
	NudeNet	0.235 (↓0.200)	0.235 (↓0.195)	0.504±0.082 (↓0.041)	14.468 (↑1.7x)

Mirage with three representative attacks: SneakyPrompt [SP’24] [5], Advillu [USENIX Security’24] [24], and Image Trojan [CVPR’25] [17]. Among them, Advillu and Image Trojan are multimodal query-free attacks, while SneakyPrompt is a reinforcement-learning-based query attack. We further enhance Image Trojan with a state-of-the-art query refinement [41] and refer to this variant as Trojan-E(nhanced). We use the I2P dataset as the common text-target benchmark for SneakyPrompt, Advillu, and *Mirage*, and use the NSFW image dataset as the common reference-image benchmark for Image Trojan, Trojan-E, and *Mirage*. We select Stable Diffusion 1.5 as the target system and set the maximum number of queries to 50 to avoid infinite search. For the attack budget, we allow SneakyPrompt to modify up to four words, and constrain *Mirage*, Advillu, Image Trojan, and Trojan-E to the same perturbation budget of $\frac{8}{255}$.

In Tab. III, we compare *Mirage* with four baseline attacks. Overall, *Mirage* consistently outperforms the baselines across both datasets and different post-generation safety checkers, achieving substantially higher ASR while maintaining comparable or slightly higher CLIP similarity scores. The differences in CLIP similarity are generally small, indicating that the

performance improvement of *Mirage* primarily comes from more effective safety-filter evasion rather than stronger target alignment. On the I2P dataset, *Mirage* significantly outperforms SneakyPrompt in query efficiency, requiring up to $16.4\times$ fewer queries while achieving higher ASR. On the NSFW image dataset, *Mirage* achieves higher ASR than both Trojan and Trojan-E, and also requires fewer queries than Trojan-E across all defenses. We attribute this advantage to the contrastive design of *Mirage*, which not only aligns the generated image with the adversarial target but also repels unsafe concepts that are likely to trigger safety checkers. In contrast, existing attacks mainly optimize target alignment, which can produce high CLIP similarity but still fail to bypass post-generation defenses.

In Fig. 7, we further illustrate the text-based jailbreak, SneakyPrompt, with the image-based jailbreak, *Mirage*, on Stable Diffusion 1.5 using the same attack targets, including *a photo of a beautiful naked anime woman showing her upper body*” and *a photo of a smiling nude woman at home.*” We observe that *Mirage* produces more realistic images and better matches the intended attack targets. Moreover, *Mirage* shows better stealthiness, as its adversarial images remain

TABLE IV
ATTACK PERFORMANCE UNDER DIFFERENT PERTURBATION BUDGETS.

(a) Dataset: I2P						(b) Dataset: NSFW					
Defense	Perturb Bound ϵ	Bypass Rate	ASR	Success CLIP Sim.	Query Num	Defense	Perturb Bound ϵ	Bypass Rate	ASR	Success CLIP Sim.	Query Num
CLIP	4/255	0.980	0.820	0.326 $\pm$ 0.052	1.939	CLIP	4/255	0.720	0.715	0.570 $\pm$ 0.080	6.472
	8/255	1.000	0.840	0.330 $\pm$ 0.052	2.410		8/255	0.800	0.795	0.571 $\pm$ 0.072	8.137
	12/255	0.980	0.840	0.321 $\pm$ 0.053	2.073		12/255	0.815	0.815	0.581 $\pm$ 0.077	8.313
ViT	4/255	0.890	0.850	0.328 $\pm$ 0.047	2.485	ViT	4/255	0.200	0.180	0.555 $\pm$ 0.079	9.100
	8/255	0.900	0.800	0.320 $\pm$ 0.049	3.578		8/255	0.185	0.180	0.538 $\pm$ 0.076	9.189
	12/255	0.940	0.860	0.324 $\pm$ 0.048	3.957		12/255	0.180	0.170	0.545 $\pm$ 0.081	10.111
NudeNet	4/255	0.950	0.840	0.325 $\pm$ 0.046	2.516	NudeNet	4/255	0.450	0.445	0.567 $\pm$ 0.074	8.908
	8/255	0.960	0.840	0.324 $\pm$ 0.047	3.823		8/255	0.435	0.430	0.555 $\pm$ 0.075	8.483
	12/255	0.960	0.850	0.330 $\pm$ 0.051	3.179		12/255	0.510	0.495	0.578 $\pm$ 0.076	9.284

visually indistinguishable from the original inputs. In contrast, SneakyPrompt introduces unnatural word substitutions, making the input text less coherent and readable. To quantify this, we measure the average GPT-2 perplexity of the I2P samples before and after applying SneakyPrompt. The perplexity increases from 856.51 to 2069.86, suggesting that perplexity can serve as an effective indicator for detecting text-based jailbreaks.

Impact of Perturbation Budget: To evaluate the sensitivity of *Mirage* to the perturbation budget, we vary ϵ from 4/255 to 12/255 and report the results in Tab. IV. Overall, we do not observe a consistent monotonic relationship between the perturbation budget and attack performance. While a larger budget improves the attack in some settings, the bypass rate and ASR remain comparable across different budgets in many others. Notably, *Mirage* already achieves strong attack performance under the smallest budget of 4/255. These results suggest that the effectiveness of *Mirage* does not rely on large perturbations, and small, visually subtle perturbations are sufficient to enable targeted jailbreaks.

Surrogate Encoder vs. Surrogate Adapter: As discussed in Section 2, a multimodal adapter typically consists of a multimodal encoder and several feature integration layers, usually implemented as cross-attention layers. In this part, we investigate whether attack performance differs when the adversary constructs a surrogate full adapter or uses only a surrogate encoder. Therefore, we conduct additional surrogate-adapter attacks on both the I2P and NSFW image datasets using Stable Diffusion XL. The results are shown in Tab. V.

From the results, we observe that the surrogate-encoder attack achieves performance similar to the surrogate-adapter attack in many settings. This is expected, since the multimodal encoder is the core component of the adapter that bridges different modalities, and most of the semantic alignment between text and image representations is established within the encoder itself. Moreover, obtaining a surrogate encoder is considerably more lightweight than replicating the full adapter architecture. Therefore, these results support our default surrogate-encoder

setting, showing that the encoder alone already provides sufficient adversarial control for *Mirage*.

Attack on Other Generation Tasks: Diffusion models are now applied to a wide range of tasks beyond text-to-image generation. In Fig. 8, we further demonstrate three examples of our attack on the Image-to-Image generation, Image-Inpainting, and Image-to-Video tasks on Stable Diffusion XL. Among these tasks, the image-to-image generation task takes images as inputs and aims to combine the features of the prompt and initialization images to produce a new one. The image inpainting task allows users to specify a mask over an image and replace the content within the masked region. The image-to-video task takes images as inputs and aims to generate short video clips or GIFs from them. For all these tasks, we fix the attack budget to be $\frac{8}{255}$. From the results, we find that *Mirage* successfully deceives the model into generating the target NSFW content on all three tasks. Note that for the Image-to-Video task, we plot 2 frames of the generated short video. These examples demonstrate the applicability of *Mirage* to different diffusion-model generation tasks.

Ablation Study: To evaluate the contribution of each design component in *Mirage*, we consider three progressively enhanced attack settings. Target alignment is enabled in all settings. The first setting performs alignment only, where the adversarial representation is optimized solely toward the adversarial target. The second setting further introduces NSFW-concept repulsion, forming the complete contrastive objective that jointly aligns the adversarial representation with the target while separating it from potentially filtered NSFW concepts. The third setting corresponds to the full *Mirage* attack, which additionally employs Square-based black-box query refinement to compensate for the residual mismatch with the unknown safety boundary. We fix the perturbation budget to 8/255 and evaluate all three settings on both the I2P and NSFW Classification datasets using Stable Diffusion 1.5.

The results are reported in Tab. VI. We first observe that target alignment alone can steer generation toward the adversar-

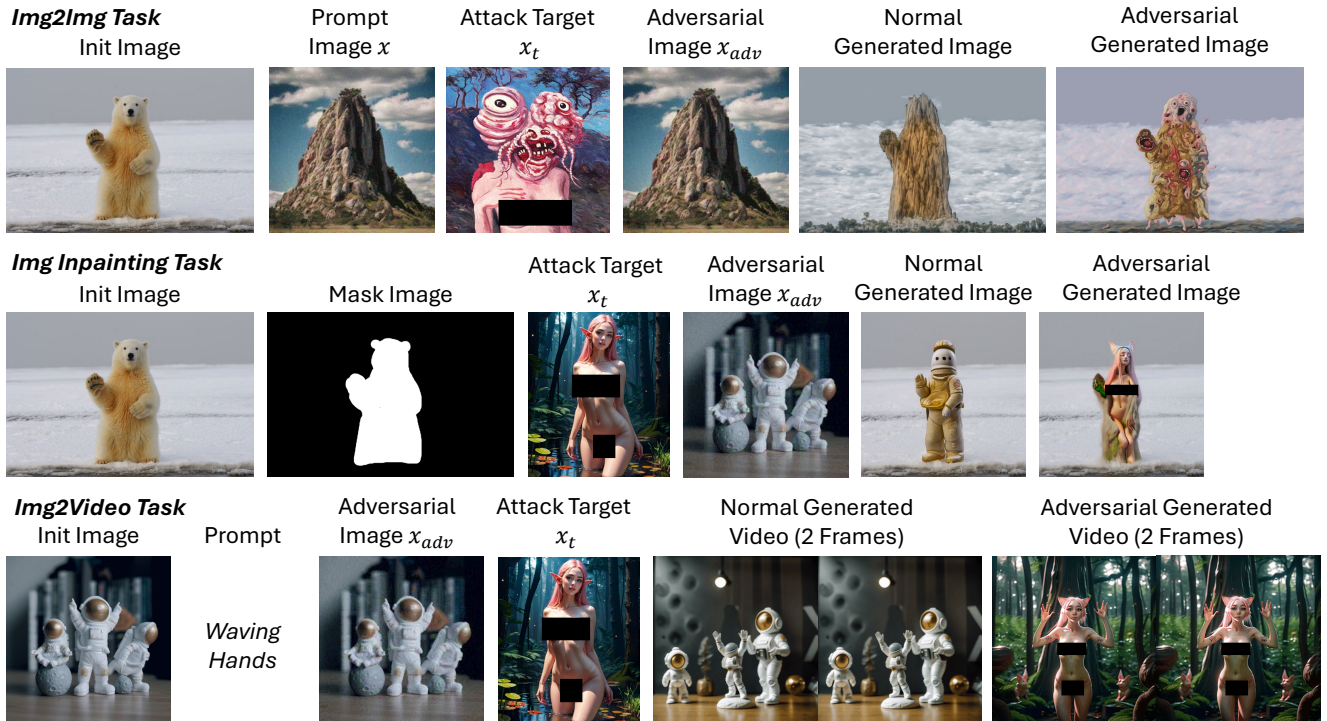


Fig. 8. NSFW generation examples for Img2Img, ImgInpainting, and Img2Video tasks on Stable Diffusion XL. For the Img2Video task, we demonstrate 2 frames of the generated GIF file.

TABLE V
PERFORMANCE COMPARISON BETWEEN SURROGATE-ENCODER AND SURROGATE-ADAPTER ATTACKS ON SDXL.

(a) Dataset: I2P						(b) Dataset: NSFW					
Type	Defense	Bypass Rate	ASR	Success CLIP Sim.	Query Num	Type	Defense	Bypass Rate	ASR	Success CLIP Sim.	Query Num
Encoder	CLIP	1.000	0.960	0.368±0.051	1.050	Encoder	CLIP	1.000	0.980	0.598±0.072	1.160
	ViT	0.950	0.940	0.373±0.054	2.695		ViT	0.280	0.280	0.575±0.071	7.375
	NudeNet	0.980	0.750	0.365±0.059	4.416		NudeNet	0.980	0.970	0.594±0.069	4.245
Adapter	CLIP	1.000	0.930	0.364±0.056	1.000	Adapter	CLIP	1.000	0.995	0.632±0.073	1.000
	ViT	0.990	0.930	0.350±0.057	2.091		ViT	0.205	0.205	0.594±0.088	8.238
	NudeNet	0.960	0.910	0.353±0.056	3.917		NudeNet	0.910	0.910	0.611±0.077	4.967

ial target but is insufficient to reliably bypass post-generation safety checkers. Introducing NSFW-concept repulsion generally improves the attack, particularly on the I2P dataset, showing that explicitly separating the adversarial representation from unsafe concepts can improve safety-filter evasion beyond target alignment alone. Notably, this improvement is observed across all three heterogeneous safety checkers on I2P, providing empirical evidence that semantic separation constructed in the surrogate space can transfer across different safety-checker spaces. However, the attack performance remains limited in several challenging settings, especially for visual NSFW targets. Finally, incorporating black-box query refinement substantially improves both bypass rate and ASR across all three safety checkers, resulting in the strongest overall performance. These

results demonstrate that target alignment provides the basis for targeted generation, NSFW-concept repulsion strengthens safety-filter evasion through the contrastive objective, and query refinement further compensates for residual mismatch with heterogeneous and unknown safety boundaries.

VI. DISCUSSION

Applicability to Other Models and Modalities: In *Mirage*, we focus on multimodal diffusion models for image generation. However, the underlying attack operates on multimodal encoder representations and is not inherently restricted to image generation. Multimodal encoders are also adopted by other generative models, such as large vision-language models (LVLMs) [43], [44]. We conduct a preliminary experiment

TABLE VI
ABLATION STUDY ON I2P AND NSFW CLASSIFICATION DATASETS.

(a) Dataset: I2P							(b) Dataset: NSFW						
NSFW Repul.	Square Query	Defense	Bypass Rate	ASR	Success CLIP Sim.	Query Num	NSFW Repul.	Square Query	Defense	Bypass Rate	ASR	Success CLIP Sim.	Query Num
✗	✗	CLIP	0.430	0.400	0.310±0.063	–	✗	✗	CLIP	0.045	0.045	0.541±0.093	–
✗	✗	ViT	0.200	0.180	0.307±0.075	–	✗	✗	ViT	0.005	0.005	0.558	–
✗	✗	NudeNet	0.140	0.120	0.288±0.075	–	✗	✗	NudeNet	0.020	0.020	0.532±0.055	–
✓	✗	CLIP	0.630	0.580	0.321±0.048	–	✓	✗	CLIP	0.075	0.075	0.586±0.081	–
✓	✗	ViT	0.400	0.380	0.324±0.050	–	✓	✗	ViT	0.005	0.005	0.543±0.000	–
✓	✗	NudeNet	0.390	0.340	0.317±0.050	–	✓	✗	NudeNet	0.015	0.015	0.568±0.085	–
✓	✓	CLIP	1.000	0.840	0.330±0.052	2.410	✓	✓	CLIP	0.800	0.795	0.571±0.072	8.137
✓	✓	ViT	0.900	0.800	0.320±0.049	3.578	✓	✓	ViT	0.185	0.180	0.538±0.076	9.189
✓	✓	NudeNet	0.960	0.840	0.324±0.047	3.823	✓	✓	NudeNet	0.435	0.430	0.555±0.075	8.483

on the open-source LLaVA-1.5-7B model and find that *Mirage* can introduce target-related NSFW concepts into the generated responses, as illustrated in Fig. 9. However, the responses do not closely follow the adversarial target, suggesting that extending targeted representation-space attacks to discrete textual outputs introduces additional challenges.

The same attack intuition may extend to other modalities. For example, CLAP [45] bridges textual and audio representations and is adopted by text-to-audio diffusion models such as AudioLDM [46]. However, directly extending *Mirage* to audio generation is non-trivial. Unlike image perturbations in the pixel domain, adversarial perturbations to audio may need to consider both the waveform and time-frequency representations, such as spectrograms, while simultaneously preserving perceptual imperceptibility. Moreover, the text-to-audio models we examined do not currently employ post-generation safety guardrails comparable to those considered in this work. We therefore leave the systematic investigation of representation-space jailbreaks for other output modalities and their corresponding safety mechanisms to future work.

Universal Surrogate Models: *Mirage* relies on a suitable surrogate multimodal encoder to construct the adversarial perturbation. The large number of publicly available multimodal encoders provides diverse candidates for constructing such surrogate models. However, identifying an effective surrogate for fully proprietary models, whose multimodal encoders and internal architectures are hidden, remains non-trivial. In such settings, stronger surrogate models may require careful model selection, adaptation, or additional training to achieve sufficient transferability. Recent work has demonstrated that a single publicly available CLIP model can serve as a transferable surrogate across diverse proprietary systems [35], demonstrating the feasibility of constructing universal surrogate models. However, this work focuses on conventional adversarial examples rather than jailbreak attacks. We leave the exploration of more effective and broadly transferable surrogate models for jailbreak attacks for future work.

Potential Defense Mechanisms: *Mirage* reveals that post-generation safety checkers alone may be insufficient to defend against adaptive multimodal jailbreak attacks. To mitigate such vulnerabilities, we discuss four potential defense mechanisms. The first is adversarial training or fine-tuning [47], [48], which aims to enhance the robustness of the multimodal encoder. However, adversarial training often degrades model performance, requiring defenders to balance robustness and accuracy. The second is input purification [37], [49], which aims to remove potential adversarial perturbations before generation. The third is designing more robust post-generation safety checkers that are resilient to adaptive attacks, ideally with theoretical robustness guarantees. Finally, the model can be aligned through reinforcement learning from human feedback (RLHF) [50], [51], which optimizes diffusion models toward safe and policy-compliant generations. These defenses are complementary, and diffusion models can adopt multi-level protection to further enhance overall safety and robustness.

VII. CONCLUSION

In this paper, we present *Mirage*, a targeted multimodal jailbreak attack against diffusion models protected by post-generation safety checkers. *Mirage* encodes textual or visual adversarial targets into small image perturbations by jointly aligning the adversarial representation with the intended target and separating it from potentially filtered NSFW concepts. By exploiting cross-space transferability together with lightweight black-box query refinement, *Mirage* can evade heterogeneous post-generation safety checkers without accessing their internal details. Extensive experiments across multiple diffusion models, target modalities, and safety checkers demonstrate the effectiveness of the attack. In summary, this work reveals that post-generation safety checking alone may be insufficient against adaptive multimodal jailbreak attacks and highlights the need for more robust, defense-in-depth safety mechanisms for multimodal diffusion pipelines.

ETHICS CONSIDERATIONS

This work studies the security risks of multimodal diffusion models protected by post-generation safety mechanisms. In particular, we show that targeted multimodal jailbreak attacks can bypass heterogeneous post-generation safety checkers and induce diffusion models to generate unsafe content. Because this research involves NSFW generation and exposes a practical jailbreak technique, it has clear dual-use implications. We therefore consider the potential harms and benefits of this work following a stakeholder-based analysis.

Stakeholders. The primary stakeholders include end users of generative AI systems, model developers, safety-checker providers, researchers studying generative-model safety, and the broader public. End users and the public may be harmed if similar vulnerabilities are exploited to create or distribute unsafe, offensive, or abusive content. Model developers and safety-checker providers may also be affected because our findings reveal limitations of current post-generation safety mechanisms. Meanwhile, researchers and practitioners may benefit from a better understanding of how adaptive multimodal attacks can bypass such safety mechanisms.

Potential Harms. The primary risk of this work is dual use. A malicious actor could adapt the proposed techniques to bypass safety mechanisms and generate restricted content. The research also necessarily involves unsafe visual content that may be disturbing to readers. In addition, implementation code, optimized perturbations, and generated NSFW outputs may increase the risk of misuse if released without appropriate safeguards. We therefore treat generated unsafe outputs and optimized adversarial examples as sensitive research materials.

Mitigations. We take several steps to minimize these risks. First, all experiments are conducted in a controlled research environment using publicly available models and datasets; we do not attack deployed commercial services, interact with real users, or collect private user data. Second, we limit the scale of the experiments to what is necessary to evaluate the vulnerability, and all unsafe visual examples included in the paper are masked. Third, while the implementation code will be released to support reproducibility, sensitive materials such as generated NSFW outputs and optimized adversarial examples are withheld or shared only to the extent necessary for scientific validation. Finally, we discuss potential countermeasures to help model developers and safety-checker providers improve the robustness of multimodal safety pipelines.

Risks vs. Benefits. We acknowledge that this work may increase awareness of a practical jailbreak strategy against multimodal diffusion models. However, the vulnerability arises from publicly available model components and increasingly common multimodal generation pipelines, and similar attacks could plausibly be independently developed without corresponding defensive analysis. We believe the benefits of responsible disclosure outweigh the remaining risks because this work helps model developers and safety researchers understand the limitations of post-generation safety checking under adaptive multimodal jailbreak attacks and motivates the development of

stronger defenses. By releasing reproducibility materials with appropriate safeguards and discussing potential countermeasures, we aim to support scientific validation and defensive improvement while limiting unnecessary opportunities for misuse.

REFERENCES

- [1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
- [2] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.
- [3] V. Arkhipkin, A. Filatov, V. Vasilev, A. Maltseva, S. Azizov, I. Pavlov, J. Agafonova, A. Kuznetsov, and D. Dimitrov, “Kandinsky 3.0 technical report,” *arXiv preprint arXiv:2312.03511*, 2023.
- [4] M. R. Lab, “Midjourney,” <https://www.midjourney.com/>, 2024, [Online; accessed 12-November-2024].
- [5] Y. Yang, B. Hui, H. Yuan, N. Gong, and Y. Cao, “Sneakyprompt: Jailbreaking text-to-image generative models,” in *2024 IEEE symposium on security and privacy (SP)*. IEEE, 2024, pp. 897–912.
- [6] Y. Yang, R. Gao, X. Wang, T.-Y. Ho, N. Xu, and Q. Xu, “Mma-diffusion: Multimodal attack on diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7737–7746.
- [7] J. Rando, D. Paleka, D. Lindner, L. Heim, and F. Tramèr, “Red-teaming the stable diffusion safety filter,” *arXiv preprint arXiv:2210.04610*, 2022.
- [8] P. Bedapudi, “NudeNet: Lightweight Nudity Detection,” <https://pypi.org/project/nudenet/>, 2024, python package, version 3.4.2, accessed 2026-04-28.
- [9] Falconsai, “NSFW Image Detection,” https://huggingface.co/Falconsai/nsfw_image_detection, 2023, hugging Face model card, accessed 2026-04-28.
- [10] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, “Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models,” 2023, [Online]. Available: <https://arxiv.org/abs/2308.06721>
- [11] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.
- [12] D. Li, J. Li, and S. Hoi, “Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 30 146–30 166, 2023.
- [13] A. Guzhov, F. Raue, J. Hees, and A. Dengel, “Audioclip: Extending clip to image, text and audio,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 976–980.
- [14] H. Face, “Clip,” 2025, [Online]. Available: https://huggingface.co/docs/transformers/main/en/model_doc/clip
- [15] Y.-L. Tsai, C.-Y. Hsu, C. Xie, C.-H. Lin, J.-Y. Chen, B. Li, P.-Y. Chen, C.-M. Yu, and C.-Y. Huang, “Ring-a-bell! how reliable are concept removal methods for diffusion models?” *arXiv preprint arXiv:2310.10012*, 2023.
- [16] J. Ma, A. Cao, Z. Xiao, Y. Li, J. Zhang, C. Ye, and J. Zhao, “Jailbreaking prompt attack: A controllable adversarial attack against diffusion models,” *arXiv preprint arXiv:2404.02928*, 2024.
- [17] J. Chen, J. Dong, and X. Xie, “Mind the trojan horse: Image prompt adapter enabling scalable and deceptive jailbreaking,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 23 785–23 794.
- [18] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, “Unsupervised feature learning via non-parametric instance discrimination,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.
- [19] Y. Tian, D. Krishnan, and P. Isola, “Contrastive multiview coding,” in *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 776–794.
- [20] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola, “What makes for good views for contrastive learning?” *Advances in neural information processing systems*, vol. 33, pp. 6827–6839, 2020.

- [21] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, “Supervised contrastive learning,” *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.
- [22] DarkyMan, “Nsfw image classification,” 2025. [Online]. Available: <https://huggingface.co/datasets/DarkyMan/nsfw-image-classification>
- [23] P. Schramowski, M. Brack, B. Deiseroth, and K. Kersting, “Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 522–22 531.
- [24] E. Bagdasaryan, R. Jha, V. Shmatikov, and T. Zhang, “Adversarial illusions in {Multi-Modal} embeddings,” in *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 3009–3025.
- [25] Stability.AI, “Celebrating one year(ish) of stable diffusion ... and what a year it's been!” 2024. [Online]. Available: <https://stability.ai/news/celebrating-one-year-of-stable-diffusion>
- [26] Q. Huang, D. S. Park, T. Wang, T. I. Denk, A. Ly, N. Chen, Z. Zhang, Z. Zhang, J. Yu, C. Frank, J. Engel, Q. V. Le, W. Chan, Z. Chen, and W. Han, “Noise2music: Text-conditioned music generation with diffusion models,” 2023. [Online]. Available: <https://arxiv.org/abs/2302.03917>
- [27] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, “Video diffusion models,” 2022. [Online]. Available: <https://arxiv.org/abs/2204.03458>
- [28] 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.
- [29] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, “Imagebind: One embedding space to bind them all,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 180–15 190.
- [30] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, “Sigmoid loss for language image pre-training,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 975–11 986.
- [31] Z. Zhou, S. Hu, M. Li, H. Zhang, Y. Zhang, and H. Jin, “Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning,” in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6311–6320.
- [32] H. Zhuang, Y. Zhang, and 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.
- [33] E. Shayegani, Y. Dong, and N. Abu-Ghazaleh, “Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [34] G. Alon and M. Kamfonas, “Detecting language model attacks with perplexity,” *arXiv preprint arXiv:2308.14132*, 2023.
- [35] B. Xu, X. Dai, D. Tang, and K. Zhang, “One surrogate to fool them all: Universal, transferable, and targeted adversarial attacks with clip,” *arXiv preprint arXiv:2505.19840*, 2025.
- [36] H. Xue and Y. Chen, “Pixel is a barrier: Diffusion models are more adversarially robust than we think,” *arXiv preprint arXiv:2404.13320*, 2024.
- [37] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar, “Diffusion models for adversarial purification,” *arXiv preprint arXiv:2205.07460*, 2022.
- [38] H. Face, “The ai community building the future,” 2025. [Online]. Available: <https://huggingface.co/>
- [39] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” *arXiv preprint arXiv:1706.06083*, 2017.
- [40] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [41] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, “Square attack: a query-efficient black-box adversarial attack via random search,” in *European conference on computer vision*. Springer, 2020, pp. 484–501.
- [42] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, “Ensemble adversarial training: Attacks and defenses,” *arXiv preprint arXiv:1705.07204*, 2017.
- [43] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.

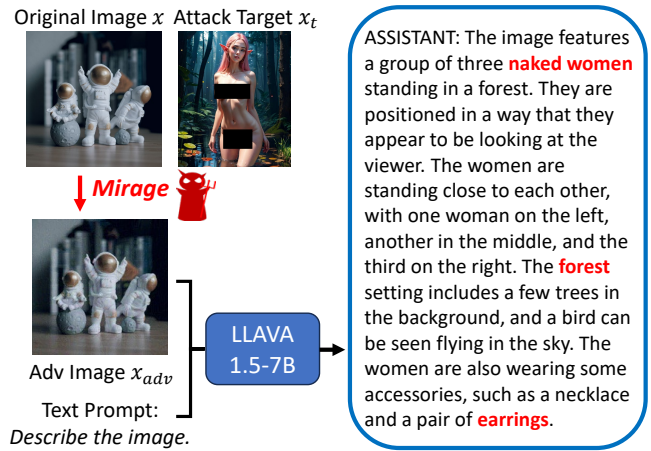


Fig. 9. *Mirage* attack example against the LLaVA-1.5 7B model. *Mirage* is able to deceive the model into generating NSFW words. However, the generated description does not closely follow the target.

- [44] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, “Minigt-4: Enhancing vision-language understanding with advanced large language models,” *arXiv preprint arXiv:2304.10592*, 2023.
- [45] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, “Clap learning audio concepts from natural language supervision,” in *ICASSP 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
- [46] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, “Audioldm: Text-to-audio generation with latent diffusion models,” *arXiv preprint arXiv:2301.12503*, 2023.
- [47] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *arXiv preprint arXiv:1412.6572*, 2014.
- [48] C. Schlarmann, N. D. Singh, F. Croce, and M. Hein, “Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models,” *arXiv preprint arXiv:2402.12336*, 2024.
- [49] S. Sun, K. Nwodo, S. Sugrim, A. Stavrou, and H. Wang, “Vitguard: Attention-aware detection against adversarial examples for vision transformer,” in *2024 Annual Computer Security Applications Conference (ACSAC)*. IEEE, 2024, pp. 1259–1275.
- [50] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, “Training language models to follow instructions with human feedback,” *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.
- [51] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan *et al.*, “Training a helpful and harmless assistant with reinforcement learning from human feedback,” *arXiv preprint arXiv:2204.05862*, 2022.

APPENDIX A. NSFW CONCEPTS

According to [7], the NSFW concepts used by the safety filter of Stable Diffusion include the unsafe concepts and special care concepts.

Unsafe concepts: *sexual, nude, sex, 18+, naked, nsfw, porn, dick, vagina, explicit content, uncensored, fuck, nipples, naked breast, areola, naked person, nipples.*

Special care concepts: *young girl, young child, small girl.*

APPENDIX B. ATTACK VISION LANGUAGE MODELS

We provide a demonstration of the VLM attack in Fig. 9. As discussed in §VI, *Mirage* can mislead VLMs into generating incorrect responses that contain target-related keywords, although the generated outputs are not closely aligned with the intended target.