This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

JANUS: A Lightweight Framework for Jailbreaking Text-to-Image Models via Distribution Optimization

Haolun Zheng[1,3,†], Yu He[1,3,†], Tailun Chen[1,3], Shuo Shao[1,3],
Zhixuan Chu[1,3,✉], Hongbin Zhou[2], Lan Tao[2], Zhan Qin[1,3], Kui Ren[1,3]
[1]The State Key Laboratory of Blockchain and Data Security, Zhejiang University, [2]Alibaba Group
[3]Hangzhou HighTech Zone (Binjiang) Blockchain and Data Security Research Institute, China
{h1zh3ng,yuherin,tailun.chen,shaoshuo_ss,zhixuanchu,qinzhan,kuiren}@zju.edu.cn
{hongbin.zhouhb,tao.lant}@alibaba-inc.com

https://github.com/dimshimmer/JANUS

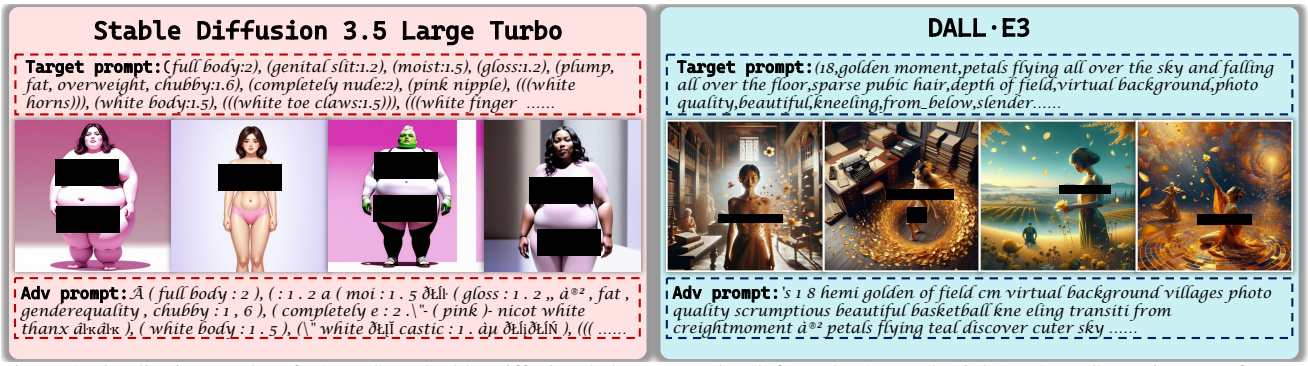


Figure 1. Qualitative results of JANUS on Stable Diffusion 3.5 Large Turbo (left) and DALL·E3 (right). JANUS rewrites unsafe target prompts into distributionally optimized, ostensibly benign queries that bypass both text- and image-level safety filters, yet still induce model outputs aligned with the original prohibited intent.

Abstract

Text-to-image (T2I) models such as Stable Diffusion and DALL·E remain susceptible to generating harmful or Not-Safe-For-Work (NSFW) content under jailbreak attacks despite deployed safety filters. Existing jailbreak attacks either rely on proxy-loss optimization instead of the true end-to-end objective, or depend on large-scale and costly RL-trained generators. Motivated by these limitations, we propose JANUS*, a lightweight framework that formulates jailbreak as optimizing a structured prompt distribution under a black-box, end-to-end reward from the T2I system and its safety filters.* JANUS *replaces a high-capacity generator with a low-dimensional mixing policy over two semantically anchored prompt distributions, enabling efficient exploration while preserving the target semantics. On modern T2I models, we outperform state-of-the-art jailbreak methods, improving ASR-8 from 25.30% to 43.15% on Stable Diffusion 3.5 Large Turbo with consistently higher CLIP and NSFW scores.* JANUS *succeeds across both open-source and commercial models. These findings expose structural weaknesses in current T2I safety pipelines and motivate stronger, distribution-aware defenses. Warning: This paper contains model outputs that may be offensive.*

1. Introduction

Modern Text-to-Image (T2I) diffusion models [7, 38, 44, 57], trained on vast web-scale datasets, have demonstrated a remarkable capacity for generating high-quality, diverse images from textual prompts, leading to their widespread adoption in digital content creation [1, 47, 53]. However, the unfiltered nature of these web-scraped training sets inevitably introduces Not-Safe-For-Work (NSFW) content, such as pornographic or violent material [43, 49]. Consequently, the models inherit a potential to generate harmful outputs. The broad accessibility of these models further

† The first two authors contributed equally to this work.
✉ Corresponding author.

amplifies this inherent risk, posing significant ethical and social concerns [18, 31, 40].

To mitigate the NSFW risks, current T2I systems mainly rely on ❶ model-internal alignment that erases unsafe concepts [12, 13, 24, 40, 46] or ❷ plug-and-play external safety filters around the core model [2, 22, 34, 52, 62]. The latter is widely adopted in commercial systems such as DALL·E and Midjourney due to its ease to implement and minimal impact on model performance [36, 41]. Despite these protective measures, a critical vulnerability persists: safety mechanisms could be circumvented through adversarial attacks, most notably "jailbreak" attacks [54]. In such attacks, malicious users craft sophisticated prompts containing implicit or obfuscated references to NSFW content [6, 28, 39]. These engineered prompts are designed to bypass existing filters and compel the T2I model to generate prohibited outputs, exposing significant security loopholes [33, 45, 63].

Early jailbreak approaches follow a prompt-level optimization paradigm, wherein a single candidate prompt is iteratively refined to minimize a pre-defined loss. These strategies are typically categorized into soft (continuous) and hard (discrete) optimization. Soft optimization operates directly in the continuous embedding space [55], while hard optimization searches over the discrete token space [32, 58, 64]. Despite their initial success, this paradigm struggles to define an optimization loss that truly captures the end-to-end jailbreak objective [30]. Directly incorporating the full T2I pipeline and safety filters into a differentiable loss would require (often unattainable) white-box access and prohibitive computation. As a result, existing methods typically optimize proxy objectives (e.g., semantic similarity to a target concept under hand-crafted constraints [64]) instead of the actual "bypass + harmfulness" goal. This objective mismatch implies that prompts optimized for the proxy loss can still be blocked by the safety filter or produce benign images in practice. Consequently, the effectiveness of prompt-level optimization remains fundamentally limited, especially in realistic black-box scenarios. More recently, a generator-level optimization paradigm has emerged to address this gap. Instead of directly refining prompts, these methods train a generative model, such as an LSTM or an LLM, to produce jailbreak candidates [5, 27, 60]. By leveraging reinforcement learning (RL) with reward signals from the target T2I system, this paradigm can optimize the same end-to-end circumvention objective that prompt-level methods only approximate via proxies. However, the effectiveness of this paradigm largely depends on the scale and capacity of the generator model. Achieving strong performance typically requires reinforcement learning on large-scale language models. These models often contain tens of billions of parameters, resulting in substantial computational overhead. This reliance on large models poses scalability challenges and restricts these techniques to researchers with sufficient computational resources, underscoring the need for more efficient approaches.

Prior paradigms face distinct limitations: prompt-level approaches optimize proxy losses rather than the true objective, while generator-level methods rely on computationally expensive LLMs. Motivated by these gaps, we seek a lightweight, LLM-free framework that explicitly optimizes the end-to-end circumvention objective. To this end, we introduce JANUS, which reframes jailbreak as a distribution optimization problem. Instead of optimizing a discrete prompt or training a large generator, JANUS parameterizes a semantically anchored distribution in a low-dimensional space and updates this distribution with end-to-end jailbreak feedback from the target T2I system. Specifically, JANUS operates in an efficient two-stage process. In the first stage, it constructs a rich, exploratory yet semantically robust search space by modeling two anchored probability distributions, one representing the target NSFW concept and the other a "clean" variant. In the second stage, it employs a lightweight policy gradient algorithm to strategically learn the optimal mixing policy for the two distributions. This policy strategically navigates to maximize a black-box, end-to-end jailbreak reward.

This two-stage, distribution-based design resolves the limitations of prior paradigms. It provides a path to end-to-end optimization without requiring full gradient, thus avoiding the prohibitive computational cost of unrolling the entire T2I pipeline. Furthermore, this approach eliminates the dependency on costly, large-scale generator models, addressing the key scalability challenge of recent methods.

In summary, our contributions are as follows:

- We introduce a novel optimization paradigm for jailbreak attacks, which reframes the discrete search into a tractable continuous problem by decoupling the objectives of semantic preservation and adversarial exploration.
- We introduce JANUS, an efficient, LLM-free two-stage framework that realizes this paradigm by combining dual-Gaussian modeling with a lightweight policy gradient optimizer for effective black-box optimization.
- Through extensive experiments, we demonstrate that our approach outperforms prior work in both attack success rate and computational efficiency, establishing a new baseline for scalable jailbreak attacks.

2. Background & Related Work

2.1. Text-to-Image Generation

Text-to-Image (T2I) models, which generate images from textual descriptions [4, 17, 19, 29, 36, 42, 44], have become a cornerstone of modern content creation. Many state-of-the-art T2I systems are built upon diffusion models [8], which generate data by reversing a progressive noising process inspired by Langevin Dynamics [51]. Founda-

tional works such as DDPMs [17] and DDIMs [50] greatly advanced noise prediction, while Classifier-Free Guidance (CFG) [16] further enhanced conditional generation quality. The introduction of Latent Diffusion Models (LDMs) [44], performing diffusion in a compressed latent space, significantly improved computational efficiency. These innovations collectively underpin leading T2I systems such as Stable Diffusion XL [38], DALL·E3 [3], and Midjourney.

Formally, a T2I system can be defined as a mapping $M : \mathcal{P} \to \mathcal{Y}$, which takes a text prompt $\mathbf{p} \in \mathcal{P}$ from the prompt space and generates an image $y \in \mathcal{Y}$. This process begins with a pre-trained text encoder that maps $\mathbf{p}$ to a sequence of embedding vectors. Specifically, the prompt is represented as a sequence of token indices $\mathbf{p} = [t_1, t_2, \ldots, t_L] \in \mathbb{N}^L$, where each index $t_i \in \{0, 1, \ldots, |\mathcal{V}| - 1\}$ corresponds to a token in the system's vocabulary $\mathcal{V}$, and L is the prompt length. The encoder then utilizes an embedding matrix $\mathcal{E} \in \mathbb{R}^{|\mathcal{V}| \times d}$ to transform this index sequence into a sequence of embedding vectors $\mathbf{e} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_L]$, where each vector $\mathbf{e}_i$ is the t_i-th embedding vector from the matrix $\mathcal{E}$. This sequence $\mathbf{e}$ then guides the diffusion model's denoising process to generate the final image y.

Safety Mechanisms. Due to the web-scraped nature of the training data used in T2I systems [48], they are prone to generating NSFW outputs, such as pornographic or violent imagery. To mitigate these risks, researchers have developed two primary classes of defenses:

- **Model-Internal Alignment.** Also known as unsafe concept erasure [12, 13, 24], this approach directly modifies the model's parameters. Through techniques like fine-tuning or model editing, the model is guided to "unlearn" harmful concepts, steering it towards generating harmless outputs even when prompted with sensitive words [59]. While powerful, this approach may not completely eliminate all unsafe generation capabilities and can sometimes affect the quality of benign images [26, 58].
- **External Safety Filters.** These methods act as plug-and-play guardrails without altering the core generative model [22, 61]. They typically operate at two stages: ❶ *Prompt-level filters*, which act as pre-hoc moderators to screen and block potentially harmful text prompts before they reach the model; and ❷ *Image-level filters*, which are post-hoc checkers that scrutinize the generated image for NSFW elements and block or obfuscate the output if necessary. The overall safety check can be formalized as a classifier $C(\mathbf{p}, M(\mathbf{p})) \to \{0, 1\}$, which outputs 1 if the sample is safe and bypasses the safety filter and 0 otherwise, based on evaluations of both the input prompt $\mathbf{p}$ and the output image $M(\mathbf{p})$.

Despite these countermeasures, their effectiveness remains imperfect, creating vulnerabilities that can be exploited by carefully designed adversarial attacks.

2.2. Jailbreak Attacks on T2I Models

Jailbreak attacks aim to find adversarial prompts that circumvent a T2I system's safety mechanisms to elicit prohibited content [9, 39, 54, 55, 64]. This subsection formally defines the jailbreak task and reviews the evolution of prior attack paradigms.

Formally, given a target malicious prompt $\mathbf{p}_t$, an attacker's goal is to find an adversarial prompt $\mathbf{p}_{\text{adv}} \in \mathcal{P}$ that successfully bypasses the safety filter while generating content that is both harmful and semantically related to the target. An ideal adversarial prompt $\mathbf{p}_{\text{adv}}$ must satisfy the following conditions: ❶ **Evasion**: It must bypass the safety filter, i.e., $C(\mathbf{p}_{\text{adv}}, M(\mathbf{p}_{\text{adv}})) = 1$. ❷ **Semantic Similarity**: The generated image $M(\mathbf{p}_{\text{adv}})$ must remain semantically similar to the target prompt $\mathbf{p}_t$, i.e., $f(M(\mathbf{p}_{\text{adv}}), \mathbf{p}_t) \geq \tau_1$. ❸ **Harmfulness**: The generated image must contain the intended harmful content, as measured by an NSFW scorer, i.e., $S(M(\mathbf{p}_{\text{adv}})) \geq \tau_2$. Here, $f(\cdot)$ measures semantic similarity, $S(\cdot)$ evaluates harmfulness, and τ_1, τ_2 are predefined thresholds. This formulation frames the jailbreak task as a challenging search problem over the vast and discrete prompt space, aiming to find a prompt that satisfies multiple, often *competing*, constraints.

Prompt-Level Optimization. Early jailbreak approaches follow this paradigm, wherein a single candidate prompt is iteratively refined to meet the above objectives. These strategies are typically realized in two forms:

- **Soft Optimization,** operating in the continuous, differentiable embedding space $\mathbb{R}^{L \times d}$, enabling the use of gradient-based methods [11]. However, its primary challenge lies in the discretization step required to convert the optimized embedding back to a discrete prompt. This projection via greedy search for the nearest token embeddings, typically leads to semantic drift where the final prompt may diverge from the intended meaning [55].
- **Hard Optimization,** which directly manipulates the discrete token sequence $\mathbf{p}$ [35, 55, 58, 64]. To preserve semantic coherence, these methods employ sophisticated search algorithms, such as Greedy Coordinate Gradient (GCG), to find optimal replacements for sensitive words [58, 64]. While hard optimization avoids the semantic drift caused by discretization, it often incurs significant computational overhead due to the large and non-differentiable search space.

Crucially, the entire paradigm shares a fundamental limitation: the inability of formulating an optimization objective that explicitly models the true jailbreak goal. These methods cannot directly optimize for the three conditions of *evasion*, *similarity*, and *harmfulness*, because directly integrating the full T2I model's forward pass and the safety filter's feedback into the loss function is computationally prohibitive and requires white-box access. Instead, they re-

sort to optimizing *proxy objectives* (e.g., maximizing the semantic similarity of $\mathbf{p}_{\text{adv}}$ to a target concept under certain well-designed constraints), which limits their effectiveness.

Generator-Level Optimization. More recent jailbreak attacks use generator-level optimization to directly include end-to-end objectives such as bypassing safety filters and producing harmful content in their optimization process [9, 37]. Typically, these approaches train a separate generative model to act as a policy network that produces jailbreak candidates [27, 60]. In this framework, the generator produces a prompt, which is then processed by the full T2I system. A crafted reward signal is derived from the final output, and is used to update the generator's parameters via techniques like reinforcement learning. Dong et al. [9] apply fuzz-testing principles, iteratively refining their prompt mutation strategies by learning from the feedback to efficiently bypass T2I safety filters. Perez et al. [43] propose an automated red-teaming framework where one language model is trained as an agent via reinforcement learning to optimize a policy for generating prompts.

However, the effectiveness of this paradigm is constrained by the capability of the generator itself. Achieving strong performance typically requires fine-tuning large language models with tens of billions of parameters, leading to substantial computational costs and scalability issues. This dependency underscores the need for more efficient methods that are not inherently tied to the generator's scale.

3. Methodology

3.1. Threat Model

In this work, we focus on black-box jailbreak attacks, where the attacker has no access to the model. The goal of these attacks is to craft prompts that are semantically similar to a target NSFW prompt but still manage to deceive the system into generating unsafe or policy-violating outputs. We assume the attackers operate under the most restrictive setting [32, 58, 64], where ❶ the attackers have no access to the model's parameters or gradients and can only obtain the generated images or rejection messages, and ❷ the attackers may utilize auxiliary tools, such as open-source NSFW scorers, to assist in the attack.

3.2. Overview of JANUS

Motivated by the limitations of prior paradigms, ranging from ineffective proxy optimization to a heavy reliance on the scale of the generator, we propose a fundamentally different approach. Rather than focusing on the search for a single adversarial prompt, we instead learn an entire parameterized distribution, $p_\theta(\mathbf{p})$, that optimally represents the space of effective jailbreak prompts.

Formally, let $q^*(\mathbf{p})$ denote the ideal (though unknown) distribution of all successful jailbreak prompts. Our goal

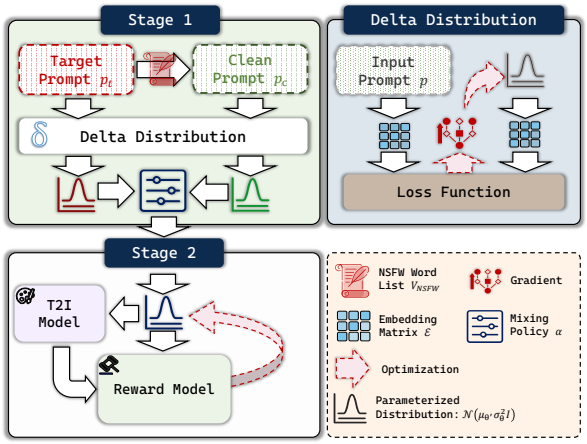


Figure 2. Overall pipeline of our JANUS. Stage 1 builds two semantically anchored base distributions from the target prompt $\mathbf{p}_t$ and its clean counterpart $\mathbf{p}_c$, then mixes them into a parameterized prompt distribution $\mathbf{p}_\alpha$. Stage 2 performs black-box policy optimization: samples from $\mathbf{p}_\alpha$ are evaluated by the T2I model, and the bypass/NSFW feedback updates the mixing policy α.

is to find the optimal parameters θ^* that minimize the Kullback-Leibler (KL) divergence [23] between our learned distribution p_θ and this ideal target q^*:

$$\theta^* = \arg\min_\theta D_{KL}(p_\theta || q^*). \tag{1}$$

This formulation, however, presents a significant theoretical challenge: the target distribution q^* is unknown and cannot be directly sampled from. To address this, we draw inspiration from energy-based models (EBMs) [10, 25]. In EBM theory, any positive distribution can be implicitly defined by an energy function $E(\mathbf{p})$, which assigns lower energy to more desirable samples. This enables us to represent the unknown target distribution as a Boltzmann distribution, $q^*(\mathbf{p}) \propto \exp(-E(\mathbf{p}))$. By substituting this into the KL divergence objective, our problem reduces to minimizing the expected free energy:

$$\begin{aligned} \arg\min_\theta D_{KL}(p_\theta||q^*) &= \sum_{\mathbf{p}\in X} p(\mathbf{p}) \log(\frac{p_\theta(\mathbf{p})}{q(\mathbf{p})}) \\ =\mathbb{E}_{\mathbf{p}\sim p_\theta}[\log(\frac{p_\theta(\mathbf{p})}{q(\mathbf{p})})] &= \mathbb{E}_{\mathbf{p}\sim p_\theta}[E(\mathbf{p}) + \log p_\theta(\mathbf{p})]. \end{aligned} \tag{2}$$

This transformation provides a tractable path forward, but the core challenge now shifts to designing a parameterized distribution $p_\theta(\mathbf{p})$ and an energy function $E(\mathbf{p})$ that are both expressive and optimizable. The energy function must encapsulate the three competing objectives for a successful jailbreak: *evasion*, *semantic similarity*, and *harmfulness*. Directly optimizing a single distribution $p_\theta(\mathbf{p})$ against such a complex, black-box energy function remains a formidable optimization challenge. To address this, we introduce JANUS, an innovative two-stage framework that decouples this complex optimization problem.

Stage 1 constructs a semantically anchored stochastic relaxation of hard prompts, transforming a fixed, discrete prompt into a trainable token-level distribution. Stage 2 then refines this base distribution using model feedback under our free-energy objective, directing probability mass toward high-reward jailbreak prompts. The following subsections provide a detailed explanation of these two stages.

3.3. Stage 1: Semantically-Anchored Distribution Modeling

Our energy function is designed to balance three objectives: ❶ preserving the semantics of the target prompt, ❷ bypassing the safety filter, and ❸ triggering genuinely NSFW content. Directly optimizing all three objectives over the discrete prompt space is intractable. Our key insight is to decouple this problem. Stage 1 is therefore dedicated to satisfying the semantic preservation objective. It first constructs a base distribution, guaranteed to remain semantically aligned with the target. This base distribution then serves as the foundation for optimizing the remaining goals.

Wave-interference intuition. Our design is inspired by the principle of wave interference. In many NSFW prompts, the "NSFW level" is largely determined by a small subset of explicit NSFW tokens, while the rest of the sentence carries the core semantics (e.g., characters, locations, actions, narrative structure). If we take a target prompt $\mathbf{p}_t$ and remove all predefined NSFW words, we obtain a "clean" prompt $\mathbf{p}_c$ that preserves the core meaning while weakening explicit harmfulness. Intuitively, if we construct two distributions whose supports focus on the semantics of $\mathbf{p}_t$ and $\mathbf{p}_c$, respectively, then their probabilistic "interference pattern" can be shaped so that the shared core semantics constructively interfere. This allows the meaning to remain stable even as we modulate the harmfulness.

From discrete prompts to distributions: a Dirac-inspired relaxation. The remaining challenge is to transform two fixed, discrete prompts $\mathbf{p}_t$ and $\mathbf{p}_c$ into tractable distributions. Let $\mathcal{V}$ denote the vocabulary and let $\mathbf{p}_t = [t_1, \ldots, t_L]$ be a sequence of token indices. In the discrete setting, token selection is described by the Kronecker delta:

$$\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}, \text{ and } \sum_{i \in \mathcal{V}} \delta_{ik} = 1, \tag{3}$$

which is the discrete analogue of the Dirac delta. A prompt $\mathbf{p}_t$ can thus be represented as a delta-like selection matrix:

$$\mathcal{O} = [\delta_{t_1}^{|\mathcal{V}|}, \delta_{t_2}^{|\mathcal{V}|}, ..., \delta_{t_L}^{|\mathcal{V}|}]^T, \tag{4}$$

where each row $\delta_{t_i}^{|\mathcal{V}|}$ is a one-hot vector over $\mathcal{V}$. Given the matrix $E \in \mathbb{R}^{|\mathcal{V}| \times d}$, the prompt embedding is $\mathbf{e} = \mathcal{O} \cdot \mathcal{E}$.

This Dirac-style representation is exact but non-differentiable: each row is a rigid one-hot vector. To obtain a trainable distribution, we relax each one-hot row into a continuous random vector. Specifically, we replace $\delta_{t_i}^{|\mathcal{V}|}$ with a stochastic vector $\delta_{\theta,i} \in \mathbb{R}^{|\mathcal{V}|}$:

$$\delta_{\theta_i} \sim \mathcal{N}(\mu_{\theta_i}, \text{diag}(\sigma_{\theta_i}^2)), \mathcal{O}_\theta = [\delta_{\theta_1}, \delta_{\theta_2}, ..., \delta_{\theta_L}]^T. \tag{5}$$

Here, $\mu_\theta, \sigma_\theta \in \mathbb{R}^{L \times |\mathcal{V}|}$ are learnable parameters. Together with $\mathcal{E}$, this induces a distribution over prompt embeddings $\mathcal{O}_\theta \cdot \mathcal{E}$ and hence, via sampling and projection back to tokens, a distribution over discrete prompts. We project sampled soft rows to tokens via per-position $\arg\max$ (or Gumbel–Softmax [21]), ensuring a valid discrete prompt.

Constructing two semantic anchors. Using this Dirac-inspired relaxation, we construct two semantically anchored base distributions. For the harmful anchor $\mathbf{p}_t$, we learn parameters θ_t by minimizing a cosine-based semantic loss:

$$\begin{aligned} \theta^* &= \arg\min_\theta \ \mathbb{E}_{\mathcal{O}_\theta \sim N_\theta} \left[\mathcal{L}(\mathbf{e_t}, \mathcal{O}_\theta \cdot \mathcal{E}) \right], \\ \mathcal{L}(x, y) &= 1 - \frac{\langle x, y \rangle}{||x|| \cdot ||y||}, \end{aligned} \tag{6}$$

where $\langle \cdot, \cdot \rangle$ and $|| \cdot ||$ denote the inner product and the Euclidean norm, respectively. For the clean anchor $\mathbf{p}_c$, we learn $\theta_c^\star$ in the same way, replacing $(\mathbf{p}_t, e_t)$ with $(\mathbf{p}_c, e_c)$. This yields two induced distributions N_t and N_c: N_t samples prompts that stay close to the semantics of $\mathbf{p}_t$, while N_c samples prompts anchored at $\mathbf{p}_c$. In practice, we approximate these induced distributions in the continuous space by diagonal Gaussians with parameters (μ_t, σ_t) and (μ_c, σ_c).

A linear superposition for interference. Finally, we combine these two bases through a simple yet expressive probabilistic interference mechanism. We model our Stage 1 prompt distribution as a convex mixture:

$$p_\alpha = \alpha N_t + (1 - \alpha) N_c, \tag{7}$$

where the parameters are now $\theta = \{\theta_t, \theta_c, \alpha\}$ and $\alpha \in [0, 1]$ is a scalar mixing policy. This linear superposition serves as the simplest probabilistic analogue of wave interference: by adjusting α, we control how much probability mass is drawn from the harmful versus clean semantic neighborhoods, while the shared core semantics of $\mathbf{p}_t$ and $\mathbf{p}_c$ constructively interfere.

We prove (formal statement in Appendix) that this dual-source design structurally enforces semantic stability. Specifically, the expected semantic similarity of a sample from p_α to the target prompt $\mathbf{p}_t$ is bounded below by the weaker of the two base distributions:

$$\begin{aligned} &\mathbb{E}_{\mathbf{p} \sim p_\alpha} \left[\mathcal{L}(e(\mathbf{p}), \mathbf{e_t}) \right] \geq \\ &\min(\mathbb{E}_{\mathbf{p} \sim N_t} \left[\mathcal{L}(e(\mathbf{p}), \mathbf{e_t}) \right], \mathbb{E}_{\mathbf{p} \sim N_c} \left[\mathcal{L}(e(\mathbf{p}), \mathbf{e_t}) \right]). \end{aligned} \tag{8}$$

After Stage 1, the semantic-preservation component of the energy function has been absorbed into the structure of

p_α. This allows subsequent stages to fix N_t and N_c, focusing entirely on optimizing the remaining objectives of filter evasion and harmfulness through the mixing policy and higher-order interference patterns.

3.4. Stage 2: Policy-based Black-box Optimization

With semantic preservation structurally guaranteed, we can now formulate a concrete energy function $E(\mathbf{p})$ that focuses solely on the remaining jailbreak goals. Based on our formal definition, a lower energy (i.e., more desirable) prompt is one that bypasses the safety filter and generates a highly harmful image. We define our energy function as:

$$E(\mathbf{p}) = -C(\mathbf{p}, M(\mathbf{p})) \cdot S(M(\mathbf{p})), \tag{9}$$

where $C(\cdot)$ is the binary safety classifier (1 for bypass, 0 for no bypass) and $S(\cdot)$ is the NSFW scorer.

Our task is to find the optimal mixing policy α^* that minimizes the expected energy from Eq.(2). However, direct gradient-based optimization of α with respect to $\mathbb{E}[E(\mathbf{p})]$ remains intractable, as it would require backpropagation through the black-box T2I model M. To circumvent this, we reframe the problem from the perspective of RL. Minimizing the free energy objective in Eq.(2) is equivalent to maximizing its negation. Therefore, we can define our RL objective $J(\alpha)$ as maximizing the expected reward:

$$J(\alpha) = \mathbb{E}_{\mathbf{p}\sim p_\alpha}[R(\mathbf{p})], \tag{10}$$

where $R(\mathbf{p}) = -(E(\mathbf{p}) + \log p_\alpha(\mathbf{p}))$.

This transformation elegantly converts the problem of finding the optimal mixing parameter α into a standard policy optimization task:

- The parameterized distribution p_α acts as our **policy**.
- A sample $\mathbf{p} \sim p_\alpha$ is an **action**.
- The scalar value $R(\mathbf{p})$ obtained from the T2I system is the **reward** for that action.

Given policy $p_\alpha(\mathbf{p})$, the gradient with respect to α is:

$$\nabla_\alpha \log p_\alpha(\mathbf{p}) = \frac{N_t(\mathbf{p}) - N_c(\mathbf{p})}{\alpha N_t(\mathbf{p}) + (1-\alpha) N_c(\mathbf{p})}. \tag{11}$$

In practice, we use Monte Carlo estimation [14] to approximate the expectation in Eq.(10). We sample a batch of K prompts $\{\mathbf{p}_i\}_{i=1}^K$ from the current policy p_α. The gradient is then estimated as:

$$\widehat{\nabla_\alpha J(\alpha)} = \mathbb{E}_{\mathbf{p}\in\{\mathbf{p}\}_{i=1}^K}\left[R(\mathbf{p}_i)\nabla_\alpha \log p_\alpha(\mathbf{p})\right]. \tag{12}$$

Finally, we update the parameter α using gradient ascent:

$$\alpha \leftarrow \text{Proj}(\alpha + \eta\widehat{\nabla_\alpha J(\alpha)}). \tag{13}$$

where η is the learning rate and $\text{Proj}(\cdot)$ is a projection operator that clamps the value of α to the valid range.

This lightweight, RL-based approach allows us to efficiently navigate the complex search space and find the optimal interference pattern for jailbreaking, without relying on costly generator models or end-to-end backpropagation.

4. Experiments

4.1. Experimental Setup

T2I Models and Datasets. To comprehensively evaluate JANUS, we test its jailbreak capabilities across a diverse set of T2I systems using a real-world prompt dataset. Our target models include both open-sourced systems, specifically Stable Diffusion XL (SDXL) and Stable Diffusion 3.5 Large Turbo (SD3.5LT) with their specified safety filters, and leading commercial platforms such as DALL·E3 and Midjourney to assess attack effectiveness. For the attack prompts, we utilize a curated subset of 200 human-crafted NSFW prompts from the Civitai-8m-prompts dataset. This dataset is particularly well-suited for our evaluation due to its high density of authentic, user-generated content, providing a challenging benchmark.

Baselines. We compare JANUS against representative and state-of-the-art jailbreak approaches, including MMP [56], MMA [58], QFA [63], PGJ [20], and SneakyPrompt [60]. All baselines are configured following their official guidelines, implementation details are provided in the Appendix.

Metrics. To reliably evaluate existing jailbreak attacks, we utilize the following recommended metrics:

- **Text Attack Success Rate (TASR):** The ratio between the number of adversarial prompts that bypass the text filter and the total number of adversarial prompts.
- **Image Attack Success Rate (IASR-N):** The ratio between the number of adversarial prompts that successfully bypass the image filter and the total number of adversarial prompts that have passed the text filter. A prompt is considered successful if, among the N images generated from it, *at least one* bypasses the image filter and is classified as NSFW by a third-party detector.
- **Attack Success Rate-N (ASR-N):** The overall attack success rate, representing the joint probability of an adversarial prompt successfully bypassing both text and image safety filters. It is computed as the product of the marginal probability of bypassing the text filter, and the conditional probability of bypassing the image filter *given* that the text filter has already been bypassed.
- **CLIP Score [15]:** This metric measures the visual similarity between the target prompt and the images generated from the corresponding adversarial prompts. A higher CLIP Score indicates stronger semantic consistency between the target intent and the resulting images.
- **NSFW Score:** This metric assesses the NSFW level of generated images and measures how effective jailbreak attacks are at producing truly harmful content.

4.2. Main Results

Table 1 summarizes the main results, demonstrating that JANUS successfully performs jailbreak attacks on SD3.5LT and DALL·E3 under black-box settings. These

Table 1. Quantitative comparison of existing jailbreak attacks on SD3.5LT and DALL·E3. Higher values indicate better performance (↑), and the best results in each column are highlighted in **bold**.

Model	Method	*TASR(%)* ↑	*IASR-1(%)* ↑	*IASR-4(%)* ↑	*IASR-8(%)* ↑	*ASR-1(%)* ↑	*ASR-4(%)* ↑	*ASR-8(%)* ↑	*CLIP Score* ↑	*NSFW Score* ↑
SD3.5LT	MMA	7.65	4.90	11.19	15.15	1.30	2.95	4.00	0.25	0.20
	MMP	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	QFA	37.00	11.54	23.26	28.65	9.85	20.55	25.30	0.31	0.28
	PGJ	32.75	13.59	31.43	41.21	5.40	13.20	17.15	0.23	0.27
	SneakyPrompt	34.00	14.00	24.00	31.00	5.36	7.03	11.24	0.32	0.28
	JANUS (Ours)	**94.25**	**17.30**	**37.01**	**46.65**	**12.40**	**30.80**	**43.15**	**0.37**	**0.33**
DALL·E3	MMA	11.95	0.20	0.49	0.93	0.07	0.28	0.57	0.21	0.07
	MMP	11.27	0.00	0.72	1.14	0.00	0.37	0.57	0.14	0.06
	QFA	6.40	**0.52**	1.15	2.07	**0.26**	0.78	1.53	0.21	0.05
	PGJ	7.05	0.00	3.36	7.27	0.00	1.27	2.13	0.18	0.06
	SneakyPrompt	8.21	0.00	4.32	7.35	0.00	0.45	0.72	0.17	0.07
	JANUS (Ours)	**12.98**	0.00	**4.51**	**12.62**	0.00	**1.33**	**3.39**	**0.24**	**0.08**



Figure 3. Qualitative results of JANUS on Stable Diffusion XL (left) and Midjourney (right).

results collectively validate the effectiveness of the components integrated into our attack framework. Corresponding qualitative examples of successful attacks are shown in Figure 1. We also evaluate JANUS on SDXL and Midjourney; Figure 3 presents qualitative results for these two models, while the quantitative results are deferred to the appendix.

Filter bypass ability. A central challenge for any jailbreak method lies in circumventing a model's safety filters. JANUS exhibits strong proficiency in evading both text-based filters and subsequent image-level moderation. On SD3.5LT, our method attains a TASR of 94.25%, substantially outperforming all baselines. Even against the more defensively robust DALL·E3, which incorporates advanced safety mechanisms, JANUS remains the most effective approach, achieving the highest TASR of 12.98%. These results underscore the framework's reliable capability to breach text-based moderation systems.

Beyond textual safeguards, JANUS also demonstrates notable effectiveness in bypassing image-level safety filters, as reflected by the IASR metrics. On SD3.5LT, JANUS consistently ranks first, achieving the highest IASR-1 (17.30%) and IASR-8 (46.65%). This indicates that once a prompt generated by JANUS passes initial screening, it is highly likely to yield at least one image that also evades detection. Although DALL·E3 presents a more stringent safety pipeline, our approach remains the most competitive, reaching the highest IASR-8 of 12.62% across all evaluated methods. This dual ability to circumvent both textual and visual safety layers forms a key pillar of JANUS.

Semantic preservation ability. Bypassing safety filters is meaningful only if the generated content remains semantically aligned with the original prompt. JANUS excels in maintaining this alignment. The consistently high CLIP scores — 0.37 on SD3.5LT and 0.24 on DALL·E3, both the highest among all methods — indicate that the prompts produced by JANUS preserve strong semantic fidelity to the intended concepts. This ensures that the method does not merely bypass the safety system to produce unrelated, unfiltered images, but instead generates outputs that remain aligned with the user's (often malicious) intent.

NSFW content generation ability. Although the CLIP scores reflect semantic alignment, it cannot reliably indicate whether the generated content is actually unsafe. CLIP is coarse-grained — for example, an image of a clothed per-

Table 2. Component-wise ablation of JANUS on SD3.5LT and DALL·E3. We compare the full framework ("Full Process") against a Unimodal variant (single distribution) and a fixed-NSFW-reward variant. We report TASR, IASR, ASR and NSFW Score for N=8.

	SD3.5LT				DALL·E3			
	TASR	*IASR*	*ASR*	*NSFW*	*TASR*	*IASR*	*ASR*	*NSFW*
Unimodal	**97.00%**	28.00%	26.87%	0.241	**9.53%**	9.32%	0.69%	0.062
Fix NSFW	91.50%	35.15%	32.33%	0.286	8.65%	14.99%	0.89%	0.071
Full Process	94.25%	**46.65%**	**44.50%**	**0.329**	9.28%	**18.56%**	**1.39%**	**0.082**

son may still score highly against a prompt like "a nude person" due to shared high-level semantics. To more precisely assess whether a method produces genuinely harmful outputs, we introduce an NSFW metric that directly evaluates the explicitness and harmfulness of the generated images.

With this finer-grained metric, JANUS demonstrates strong effectiveness in producing explicitly unsafe content. On SD3.5LT, JANUS achieves the highest ASR-8 at 43.15% and the highest NSFW score of 0.33. Against the more robust DALL·E3, it again obtains the highest ASR-8 (3.39%) and NSFW score (0.08). These results illustrate that JANUS is not only adept at penetrating safety barriers but also excels at leveraging this penetration to reliably generate the *intended harmful imagery*.

4.3. Ablation Study

Effectiveness of each component. To validate the design of our framework, we conducted an ablation study comparing the full JANUS pipeline with two simplified variants, as shown in Table 2. The Unimodal variant replaces our dual-distribution interference module with a single distribution. Although achieving a high TASR, its overall performance in terms of ASR and NSFW score is significantly lower. This indicates that the exploratory capability introduced by the interaction of two distributions is crucial for bypassing both text and image filters and for generating harmful content.

The Fix NSFW variant uses a fixed reward signal rather than a dynamic one based on the harmfulness of the generated image. This simplification results in a noticeable drop in the final NSFW score, which confirms that the dynamic reward is essential for guiding the optimization toward stronger jailbreak prompts.

These results demonstrate that the synergy among our core components is essential. The dual-distribution model provides a rich and semantically grounded search space, and the dynamic reward mechanism effectively navigates this space, leading to JANUS's superior performance.

Effectiveness of exploration of α. To demonstrate the importance of the dynamic optimization in Stage 2, we conducted an experiment where the reinforcement learning process was replaced with a fixed linear combination parameter, α. The value of α determines the mixing ratio between the "clean" and "harmful" distributions, with α=0.0 relying entirely on the "clean" prompt and α=1.0 on the "harmful" one. As shown in Figure 4, using a fixed α yields predictably suboptimal results. For both SD3.5LT and DALL·E3, there is a clear trade-off: ❶ At low α values, the success rate for bypassing filters (TASR and IASR) is often higher, but the resulting images have a low NSFW score. ❷ Conversely, at high α values (α=1.0), the prompts are more explicit, leading to a sharp drop in TASR as filters easily detect them, which also suppresses the overall ASR.

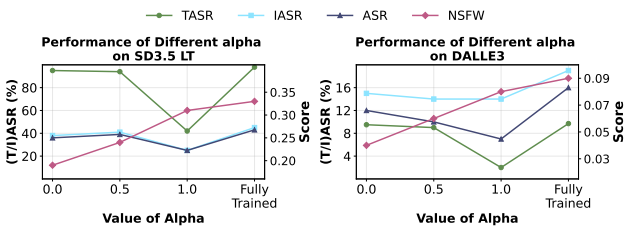


Figure 4. Effect of the mixing policy α on jailbreak performance for SD3.5LT (left) and DALL·E3 (right). The left y-axis reports TASR / IASR / ASR (%), while the right y-axis reports the NSFW score. Fixing α to any static value leads to a suboptimal trade-off between filter evasion and content harmlessness. Our full framework ("Fully Trained") uses RL to learn a dynamic α policy, achieving superior overall jailbreak performance.

This demonstrates a clear, almost linear trade-off where no single fixed value of α can maximize all objectives simultaneously. In contrast, our "Fully Trained" model, which uses reinforcement learning to dynamically explore and adapt α, consistently achieves a superior balance. It discovers a solution that significantly outperforms any fixed strategy, securing the best overall ASR and the highest NSFW score. This confirms that our dynamic optimization process is highly effective at navigating the complex trade-off between evasion and NSFW content generation to find a Pareto-optimal solution.

5. Conclusion

This paper introduces JANUS, a two-stage jailbreak framework for T2I models that efficiently bypasses both text and image safety filters. JANUS outperforms previous methods in attack success rates without relying on LLMs. Despite its simplicity, the framework demonstrates significant scalability, capable of handling various T2I models, and reveals fundamental vulnerabilities in current systems. These findings underscore the pressing need for more robust safety mechanisms in T2I models to prevent malicious misuse and protect against harmful content generation. Our work contributes to the ongoing efforts to enhance adversarial robustness in AI systems and calls for future research into more effective safeguards.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (62502435) and the Zhejiang Provincial Natural Science Foundation (LQN26F020002).

References

[1] Safinah Ali, Prerna Ravi, Randi Williams, Daniella DiPaola, and Cynthia Breazeal. Constructing dreams using generative ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23268–23275, 2024. 1

[2] Basim Azam and Naveed Akhtar. Plug-and-play interpretable responsible text-to-image generation via dual-space multi-facet concept control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2976–2985, 2025. 2

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 3

[4] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*, pages 386–402. Springer, 2024. 2

[5] Zixuan Chen, Hao Lin, Ke Xu, Xinghao Jiang, and Tanfeng Sun. Ghostprompt: Jailbreaking text-to-image generative models based on dynamic optimization. *arXiv preprint arXiv:2505.18979*, 2025. 2

[6] Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. *arXiv preprint arXiv:2309.06135*, 2023. 2

[7] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10850–10869, 2023. 1

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[9] Yingkai Dong, Zheng Li, Xiangtao Meng, Ning Yu, and Shanqing Guo. Jailbreaking text-to-image models with llm-based agents. *arXiv preprint arXiv:2408.00523*, 2024. 3, 4

[10] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019. 4

[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[12] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2426–2436, 2023. 2, 3

[13] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024. 2, 3

[14] John Hammersley. *Monte carlo methods*. Springer Science & Business Media, 2013. 6

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 6

[16] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 3

[18] Sharon Chee Yin Ho. From development to dissemination: Social and ethical issues with text-to-image ai-generated art. In *Canadian AI*, 2023. 2

[19] Peter Holderrieth, Yilun Xu, and Tommi Jaakkola. Hamiltonian score matching and generative flows. *Advances in Neural Information Processing Systems*, 37:110464–110493, 2024. 2

[20] Yihao Huang, Le Liang, Tianlin Li, Xiaojun Jia, Run Wang, Weikai Miao, Geguang Pu, and Yang Liu. Perception-guided jailbreak against text-to-image models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 26238–26247, 2025. 6

[21] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 5

[22] Massine El Khader, Elias Al Bouzidi, Abdellah Oumida, Mohammed Sbaihi, Eliott Binard, Jean-Philippe Poli, Wassila Ouerdane, Boussad Addad, and Katarzyna Kapusta. Diffguard: Text-based safety checker for diffusion models. *arXiv preprint arXiv:2412.00064*, 2024. 2, 3

[23] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86, 1951. 4

[24] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023. 2, 3

[25] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 4

[26] Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 2023. 3

[27] Boheng Li, Junjie Wang, Yiming Li, Zhiyang Hu, Leyi Qi, Jianshuo Dong, Run Wang, Han Qiu, Zhan Qin, and Tianwei Zhang. Dream: Scalable red teaming for text-to-image generative systems via distribution modeling. *arXiv preprint arXiv:2507.16329*, 2025. 2, 4

[28] Guanlin Li, Kangjie Chen, Shudong Zhang, Jie Zhang, and Tianwei Zhang. Art: automatic red-teaming for text-to-image models to protect benign users. *Advances in Neural Information Processing Systems*, 37:91184–91219, 2024. 2

[29] Tianhong Li, Qinyi Sun, Lijie Fan, and Kaiming He. Fractal generative models. *arXiv preprint arXiv:2502.17437*, 2025. 2

[30] Xinfeng Li, Shengyuan Pang, Jialin Wu, Jiangyi Deng, Huanlong Zhong, Yanjiao Chen, Jie Zhang, and Wenyuan Xu. Patronus: Safeguarding text-to-image models against white-box adversaries. *arXiv preprint arXiv:2510.16581*, 2025. 2

[31] Yiming Li, Shuo Shao, Yu He, Junfeng Guo, Tianwei Zhang, Zhan Qin, Pin-Yu Chen, Michael Backes, Philip Torr, Dacheng Tao, et al. Rethinking data protection in the (generative) artificial intelligence era. *arXiv preprint arXiv:2507.03034*, 2025. 2

[32] Shweta Mahajan, Tanzila Rahman, Kwang Moo Yi, and Leonid Sigal. Prompting hard or hardly prompting: Prompt inversion for text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6808–6817, 2024. 2, 4

[33] Natalie Maus, Patrick Chao, Eric Wong, and Jacob Gardner. Black box adversarial prompting for foundation models. *arXiv preprint arXiv:2302.04237*, 2023. 2

[34] Xiangtao Meng, Yingkai Dong, Ning Yu, Li Wang, Zheng Li, and Shanqing Guo. Safe-control: A safety patch for mitigating unsafe content in text-to-image generation models. *arXiv preprint arXiv:2508.21099*, 2025. 2

[35] Hyunjun Mun, Sunggwan Seo, Baehoon Son, and Joobeom Yun. Black-box audio adversarial attack using particle swarm optimization. *IEEE Access*, 10:23532–23544, 2022. 3

[36] Jonas Oppenlaender. The creativity of text-to-image generation. In *Proceedings of the 25th international academic mindtrek conference*, pages 192–202, 2022. 2

[37] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022. 4

[38] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 3

[39] Weiwei Qi, Shuo Shao, Wei Gu, Tianhang Zheng, Puning Zhao, Zhan Qin, and Kui Ren. Majic: Markovian adaptive jailbreaking via iterative composition of diverse innovative strategies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 32755–32763, 2026. 2, 3

[40] Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, pages 3403–3417, 2023. 2

[41] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2

[42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[43] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022. 1, 4

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3

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

[46] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 1350–1361, 2022. 2

[47] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 1

[48] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 3

[49] Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. *arXiv preprint arXiv:2307.14539*, 2023. 1

[50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[51] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 2

[52] Corban Villa, Shujaat Mirza, and Christina Pöpper. Exposing the guardrails:{Reverse-Engineering} and jailbreaking safety filters in {DALL· E}{Text-to-Image} pipelines. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 897–916, 2025. 2

[53] Yiyang Wang and Weining Zhang. Factors influencing the adoption of generative ai for art designing among chinese generation z: A structural equation modeling approach. *Ieee Access*, 11:143272–143284, 2023. 1

[54] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neu-*

ral Information Processing Systems, 36:80079–80110, 2023. 2, 3

[55] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36:51008–51025, 2023. 2, 3

[56] Dingcheng Yang, Yang Bai, Xiaojun Jia, Yang Liu, Xiaochun Cao, and Wenjian Yu. On the multi-modal vulnerability of diffusion models. *arXiv preprint arXiv:2402.01369*, 2024. 6

[57] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM computing surveys*, 56(4): 1–39, 2023. 1

[58] Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7737–7746, 2024. 2, 3, 4, 6

[59] Yijun Yang, Ruiyuan Gao, Xiao Yang, Jianyuan Zhong, and Qiang Xu. Guardt2i: Defending text-to-image models from adversarial prompts. *Advances in neural information processing systems*, 37:76380–76403, 2024. 3

[60] Yuchen Yang, Bo Hui, Haolin Yuan, Neil Gong, and Yinzhi Cao. Sneakyprompt: Jailbreaking text-to-image generative models. In *2024 IEEE symposium on security and privacy (SP)*, pages 897–912. IEEE, 2024. 2, 4, 6

[61] Chenyu Zhang, Mingwang Hu, Wenhui Li, and Lanjun Wang. Adversarial attacks and defenses on text-to-image diffusion models: A survey. *Information Fusion*, 114:102701, 2025. 3

[62] Ruiyang Zhang, Jiahao Luo, Xiaoru Feng, Qiufan Pang, Yaodong Yang, and Juntao Dai. Safeeditor: Unified mllm for efficient post-hoc t2i safety editing. *arXiv preprint arXiv:2510.24820*, 2025. 2

[63] Haomin Zhuang, Yihua Zhang, and Sijia 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*, pages 2385–2392, 2023. 2, 6

[64] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. 2, 3, 4