DiSCO: DEFENDING TEXT-TO-IMAGE GENERATION THROUGH DISTRIBUTION-GUIDED CONTRASTIVE PROMPT OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As text-to-image generative models advance, they raise critical safety concerns, particularly the generation of Not-Safe-For-Work (NSFW) content such as violence and nudity, further exacerbated by red-teaming adversarial attacks. Existing defenses predominantly operate under white-box assumptions, relying on text encoder optimization, weight editing, or inference-time intervention, and fundamentally cannot scale to proprietary models. Black-box alternatives based on LLM prompt rewriting offer broader applicability, yet fail in a critical regime we identify as the *benign adversarial* problem: prompts that are linguistically safe but still trigger harmful generation due to the model's learned data distribution. We propose DiSCO, a training-free, strictly black-box defense that operates entirely at the prompt level as a plug-and-play module, requiring no model retraining, fine-tuning, or access to model internals. DiSCO performs distribution-guided suffix expansion through fixed-length beam search, using contrastive scores over safe and unsafe image pools generated by the target model. We demonstrate that across 32 system–attack settings and five seeds, DiSCO reduces average ASR from 23.6% to 2.4% under NudeNet and from 8.3% to 1.7% under Q16, while improving generation quality. These results establish DiSCO as a black-box, architecture-agnostic safety module across the evaluated text-to-image systems.
Warning: *This paper contains model outputs that are offensive in nature.*

1 INTRODUCTION

The onset of text-to-image generation has transformed creative workflows across industries (Liao et al., 2025; Ramesh et al., 2021; Luo et al., 2025; Nichol & Dhariwal, 2021; Dhariwal & Nichol, 2021; Song & Ermon, 2019; Ho et al., 2020), enabling users to synthesize photorealistic imagery from natural language descriptions. Models such as Stable Diffusion (Rombach et al., 2022), SD 3 (Esser et al., 2024), and Flux (Labs, 2024) have demonstrated remarkable generative capabilities, yet this progress has simultaneously introduced critical safety vulnerabilities. In the realm of content safety, these models can produce Not-Safe-For-Work (NSFW) content (Li et al., 2024b; Chen et al., 2025), including depictions of violence, nudity, and other harmful imagery, either through undefended generation or through deliberate manipulation of input prompts, known as adversarial attacks. The growing accessibility of open-source text-to-image models has further amplified these risks, as users can freely interact with systems that lack adequate safety guardrails, raising urgent concerns for responsible deployment.

The discovery of red-teaming adversarial attacks has accordingly led to escalating security concerns surrounding text-to-image systems. These attacks span a spectrum of threat models, from black-box methods that craft adversarial prompts without model access (Tsai et al., 2023; Yang et al., 2024b), to white-box approaches that leverage gradients and internal representations for targeted prompt optimization (Zhang et al., 2024b; Yang et al., 2024a). More recently, LLM-driven attacks, such as APT (Liu et al., 2025), have raised the bar further by generating human-readable adversarial prompts that evade both automated detectors and human review. The increasing sophistication of these attacks demands defense mechanisms that are equally robust and adaptive.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

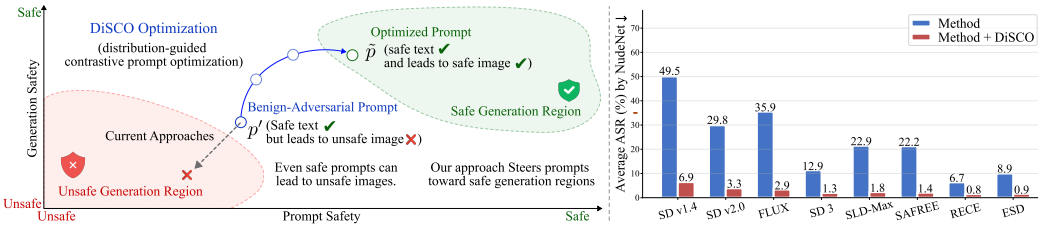


Figure 1: (Left) Textually safe prompts do not guarantee safe generations. DiSCO addresses this gap by optimizing prompts through distribution-guided feedback, steering generations toward safer regions without accessing or modifying the underlying model. (Right) DiSCO consistently reduces the Attack Success Rate (ASR) across diverse undefended and defended text-to-image models, serving as a plug-and-play black-box safety booster without modifying the underlying model.

In response, a spectrum of defenses has emerged. White-box methods intervene inside the generation pipeline via weight fine-tuning (Gandikota et al., 2023; Zhang et al., 2024a), cross-attention editing (Huang et al., 2024), or inference-time steering (Schramowski et al., 2023; Yoon et al., 2025), but they inherently do not scale to proprietary or closed-source models. A lightweight black-box alternative is LLM-based prompt rewriting (Zhao et al., 2026; Jing et al., 2025), which often neutralizes prompts that contain explicit unsafe intent; however, it exposes a more fundamental failure mode where textually benign prompts still trigger unsafe images. We formalize this as the *benign adversarial* problem: a prompt p' is benign adversarial with respect to a generative model $\mathcal{G}$ if p' is deemed safe by language-level assessment, yet $\mathcal{G}(p')$ produces unsafe visual content. Recent evidence suggests this is systematic rather than anecdotal, where benign prompts can unintentionally elicit harmful generations (Li et al., 2024a), and the issue persists even under compromised model weights (Wu et al., 2025), indicating a gap in purely text space defenses. Motivated by this, we view defending against benign adversarial prompts as a distributional alignment problem: instead of modifying $\mathcal{G}$, optimize the prompt to shift generations from unsafe to safe regions of the model's learned output distribution as illustrated in Figure 1 (left).

In this work, we study the effect of distribution-guided prompt optimization as a black-box defense against unsafe text-to-image generation. Specifically, we explore how contrastive scoring over a target model's own safe and unsafe outputs can be used to systematically steer benign adversarial prompts toward safe generation regions. We present DiSCO, a training-free, strictly black-box defense that operates entirely at the prompt level as a plug-and-play module, requiring no model retraining, fine-tuning, or access to model internals. DiSCO takes the prompt with adversarial content as its starting point and optimizes it through distribution-guided contrastive suffix expansion via beam search. Using the target model's own safe and unsafe outputs as references, DiSCO steers generation toward safe regions without accessing model internals. We present a comprehensive empirical study assessing the impact of introducing DiSCO against state-of-the-art adversaries on the I2P benchmark (Schramowski et al., 2023). Our experiments show that equipping existing undefended and defended systems with DiSCO provides consistent improvements in safety across all attacks evaluated, as shown in Figure 1 (right). Our contributions are summarized as follows:

- **Problem:** We formalize the *benign adversarial* regime in text-to-image safety, where prompts that are textually safe can still induce unsafe generations due to the model's learned output distribution.
- **Method:** We introduce **DiSCO**, a training-free, *strictly black-box* and *plug-and-play* prompt-optimization module that steers generations via distribution-guided contrastive suffix search, requiring no retraining, fine-tuning, or access to model internals.
- **Results:** On I2P under four red-teaming attacks, DiSCO consistently improves safety across both undefended and defended systems. Over five seeds and 32 system–attack settings, it reduces average ASR from **23.6% to 2.4%** under NudeNet and from **8.3% to 1.7%** under Q16, while maintaining or improving semantic fidelity and perceptual quality.

2 Related Work

Adversarial Robustness of Text-to-Image Models. Red-teaming has become the standard way to stress-test text-to-image safety, and recent work shows that adversarial prompting can reliably

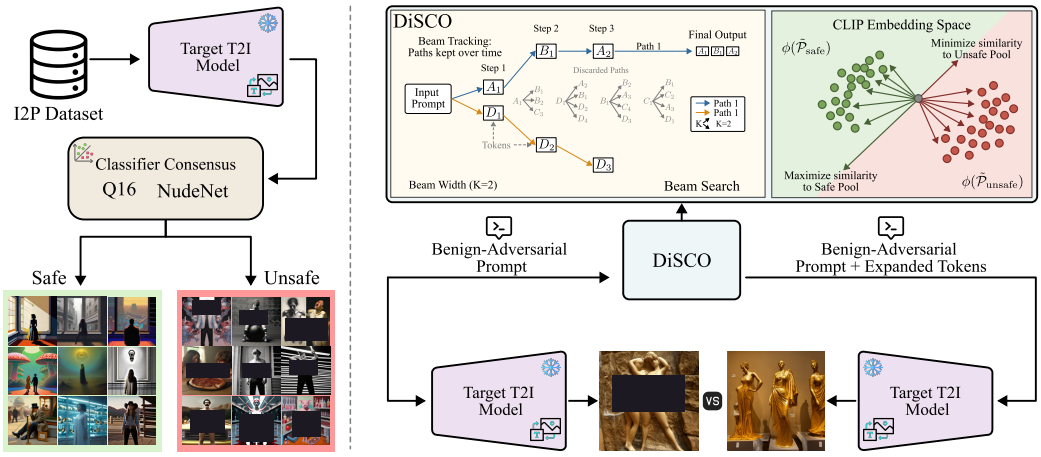


Figure 2: Overview of DiSCO. **Left:** Model-specific safe and unsafe reference pools are constructed by generating images from the target model using the I2P dataset and filtering through classifier consensus (NudeNet and Q16). **Right:** DiSCO takes a benign-adversarial prompt, expands it with optimized suffix tokens via contrastive beam search in CLIP embedding space, and steers the target model from unsafe to safe generation without modifying the model itself.

elicit NSFW generations under both black-box and white-box threat models. Black-box attacks craft prompts without accessing the generator internals, e.g., Ring-A-Bell (Tsai et al., 2023) and SneakyPrompt (Yang et al., 2024b) search for token substitutions that evade prompt-side safeguards. White-box attacks are more targeted, leveraging gradients and internal representations to directly optimize for unsafe outputs, as in UnlearnDiffAtk (Zhang et al., 2024a) and MMA-Diffusion (Yang et al., 2024a). More recently, LLM-driven attacks further raise the bar by producing fluent, human-readable prompts (often via search over adversarial suffixes) that bypass heuristic filters and remain difficult to flag at the text level, e.g., APT (Liu et al., 2025). Collectively, these results highlight that robustness must be assessed against increasingly natural-looking prompts that can appear benign while still inducing unsafe generations.

Defending Text-to-Image Models. Defenses span a spectrum of model access and intervention. Weight-modifying and fine-tuning approaches can suppress unsafe concepts but require parameter access and may need repeated updates as new failure modes emerge, e.g., ESD (Gandikota et al., 2023), AdvUnlearn (Zhang et al., 2024a), RECE (Huang et al., 2024), and external rectification modules such as SafePatch (Meng et al., 2026). Inference-time methods avoid permanent edits but still depend on internal representations, limiting portability across architectures (e.g., SLD-Max (Schramowski et al., 2023), SAFREE (Yoon et al., 2025), DTVI (Tan et al., 2026), PromptGuard (Yuan et al., 2026)). Complementarily, LLM-based rewriting provides a lightweight, model-agnostic safeguard by sanitizing prompts before generation (e.g., VALOR (Zhao et al., 2026), PromptSafe (Jing et al., 2025)), but can fail when linguistically safe prompts still map to unsafe regions of a model's visual distribution. DiSCO targets this gap with a strictly black-box, prompt-level optimization that uses the target model's own safe/unsafe output distributions as guidance, and can be prepended to both undefended systems and existing defenses without changing the generator.

3 Methodology

We study the impact of distribution-guided prompt optimization on the safety of text-to-image generation as shown in Figure 2. We observe that traditional defense approaches either modify the model's internal parameters or operate solely at the textual level, overlooking the relationship between the input prompt and the model's own visual output distribution. Thus, we aim at studying the impact that introducing a prompt-level optimization module, guided by the model's own safe and unsafe generation regions, can have on enhancing adversarial robustness.

3.1 Problem Definition and Evaluation Protocol

Let $\mathcal{P}$ be the space of text prompts and $\mathcal{X}$ the space of images. A text-to-image generator is a (possibly stochastic) mapping $G : \mathcal{P} \to \mathcal{X}$ that produces an image $x \sim G(p)$ given a prompt $p \in \mathcal{P}$. We study red-teaming attacks that craft adversarial prompts to increase the likelihood of unsafe generations: an attack algorithm $\mathcal{A}$ maps an initial prompt (or template) to $p_{\text{adv}} = \mathcal{A}(p)$. In many deployed black-box pipelines, a language-based sanitizer or LLM rewriting module $\mathcal{R}$ is applied first, producing a rewritten prompt $p' = \mathcal{R}(p_{\text{adv}})$ that is linguistically benign, after which the system generates $x \sim G(p')$. Our focus is the regime where language sanitization is not sufficient as safety depends on the alignment between p' and the *visual* output distribution induced by G.

Benign adversarial prompts. Let $\text{SafeText} : \mathcal{P} \to \{0, 1\}$ denote a language-level safety assessment and $\text{SafeImg} : \mathcal{X} \to \{0, 1\}$ an image-level safety assessment. We say that $p' \in \mathcal{P}$ is *benign adversarial* w.r.t. G if it passes text-level checks but still yields unsafe images with high probability:

$$\text{SafeText}(p') = 1 \quad \text{and} \quad \mathbb{P}_{x \sim G(p')}[\text{SafeImg}(x) = 0] \geq 0.5. \tag{1}$$

This phenomenon motivates viewing defense as a *distributional alignment* problem: rather than modifying G, we optimize the input prompt so that the induced generation shifts away from unsafe regions and toward safe regions of G's learned output distribution.

Defense setting (strictly black-box, prompt-level). We consider a prompt-level defense $\mathcal{D} : \mathcal{P} \to \mathcal{P}$ that transforms the (typically rewritten) prompt p' into an optimized prompt $p^* = \mathcal{D}(p')$, after which the generator runs unmodified: $x^* \sim G(p^*), \qquad p^* = \mathcal{D}(p')$.

The defense has no access to G's parameters, gradients, attention maps, or intermediate activations; it may only query G to obtain generated images. At a high level, we seek to reduce unsafe generations while preserving the user intent of p':

$$\min_{\mathcal{D}} \ \mathbb{E}_{p'} \Big[\mathbb{P}_{x \sim G(\mathcal{D}(p'))} \big[\text{SafeImg}(x) = 0 \big] \Big] \quad \text{s.t.} \quad \text{Sem}(\mathcal{D}(p'), p') \geq \tau, \tag{2}$$

where $\text{Sem}(\cdot, \cdot)$ denotes an abstract semantic-fidelity measure and τ is a fidelity threshold. In our experiments, we evaluate semantic preservation using image–text CLIP alignment (Radford et al., 2021) and prompt-embedding cosine similarity (see Appendix A.10).

Evaluation protocol and metrics. We report robustness using *Attack Success Rate* (ASR): the fraction of evaluated prompts that produce unsafe images under the tested system (undefended, baseline-defended, or defense-enhanced):

$$\text{ASR} \ = \ \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbb{I}[\text{SafeImg}(x_q) = 0] \,, \qquad x_q \sim G(p_q), \tag{3}$$

where $\mathcal{Q}$ is the evaluated prompt set and p_q is the corresponding input prompt to the generator. We compute ASR using two complementary image safety classifiers (e.g., NudeNet and Q16), and report semantic fidelity / generation quality metrics (e.g., CLIP-based alignment and ImageReward) to verify that safety gains do not come from excessive semantic drift or degraded output quality. Unless stated otherwise, reported ASR values are measured under NudeNet, the standard evaluator in prior work, which allows direct comparison with previously published numbers.

3.2 Distribution-Guided Contrastive Suffix Optimization

Given the problem formulation, we are now ready to present our defense: Distribution-Guided Contrastive Suffix Optimization (DiSCO). DiSCO is a strictly black-box, prompt-level module that steers a target generator G toward safe generations by optimizing the *prompt* using feedback obtained only from *observable outputs* of G. In a nutshell, given a (potentially sanitized) input prompt p', DiSCO performs autoregressive *suffix expansion* and selects the suffix that maximizes a distribution-guided contrastive score computed in a model-agnostic embedding space.

Safe and unsafe reference pools. To obtain a robust reference signal under a black-box constraint, we construct two *model-specific* image pools by querying G on non-adversarial prompts from the I2P dataset. Let $\mathcal{P}_{\text{safe}} = \{x_i^+\}_{i=1}^{N_+}$ and $\mathcal{P}_{\text{unsafe}} = \{x_j^-\}_{j=1}^{N_-}$ denote the resulting pools of safe and unsafe images, respectively. We enforce an unambiguous selection rule: an image is added to $\mathcal{P}_{\text{safe}}$

only if *both* NudeNet and Q16 classify it as safe, and added to $\mathcal{P}_{\text{unsafe}}$ only if *both* classify it as unsafe; ambiguous cases (disagreement) are discarded. An alternative confidence-based construction of the two pools is discussed in Appendix A.8. Because the pools are generated by the target model G, they reflect G's own learned visual manifold and require no external unsafe corpora or internal access.

Distribution-guided contrastive objective. Given a candidate prompt $\hat{p}$, we query the target model to obtain $\hat{x} \sim G(\hat{p})$ and embed the image using a frozen CLIP image encoder $\phi(\cdot)$. At each optimization step, we randomly sample one shared set of R references from each pool and use these same references to score all candidate suffixes at that step. This yields $\tilde{\mathcal{P}}_{\text{safe}} \subset \mathcal{P}_{\text{safe}}$ and $\tilde{\mathcal{P}}_{\text{unsafe}} \subset \mathcal{P}_{\text{unsafe}}$, with $|\tilde{\mathcal{P}}_{\text{safe}}| = |\tilde{\mathcal{P}}_{\text{unsafe}}| = R$ (default $R = 8$). We score $\hat{p}$ by the contrastive alignment of its generated image with these two reference sets:

$$J(\hat{p}) = \frac{1}{R} \sum_{x_i^+ \in \tilde{\mathcal{P}}_{\text{safe}}} \cos\big(\phi(\hat{x}), \phi(x_i^+)\big) - \frac{1}{R} \sum_{x_j^- \in \tilde{\mathcal{P}}_{\text{unsafe}}} \cos\big(\phi(\hat{x}), \phi(x_j^-)\big), \tag{4}$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity. Maximizing J simultaneously attracts generations toward safe regions and repels them from unsafe regions in the embedding space. Re-sampling references at every step exposes the optimizer to diverse views of the two distributions while keeping the per-step cost bounded. In practice, a single optimization pass already yields strong robustness gains; a preference-based optimization variant under a unified objective is detailed in Appendix A.9.

We optimize prompts by expanding a suffix token-by-token using a suffix generator language model M (LLaMA-3-8B). Starting from p', we maintain a beam $\mathcal{B}$ of size K (default $K = 4$). At step $t \in \{1, \ldots, T\}$ (default $T = 16$), for each candidate prompt in the beam we propose next-token continuations via M, query G for each continuation, compute its score via Eq. equation 4, and retain the top-K candidates. After T steps, the final optimized prompt is selected as $p^* = \arg\max_{\hat{p} \in \mathcal{B}} J(\hat{p})$. The complete procedure is given in Algorithm 1.

Computational Overhead. DiSCO requires a one-time, target-specific setup to construct reusable safe and unsafe reference pools from model outputs; it is therefore training-free and black-box, but not zero-shot for a new generator. At inference, the default beam search ($K = 4$, $T = 16$, $b = 4$) requires $b + (T - 1)bK = 244$ candidate generations. To reduce this cost, DiSCO is invoked only when a textually benign prompt initially produces an unsafe image, while safe generations pass through unchanged. Appendix A.6 evaluates DiSCO on initially safe prompts, and Appendix A.12 discusses overhead-reduction strategies and their effectiveness–efficiency trade-offs.

4 Experiments

Models. We evaluate DiSCO across multiple text-to-image generators spanning both major architecture families. Our study includes UNet-based diffusion models (SD v1.4, SD v2.0 (Rombach et al., 2022)) and DiT-based models (Flux (Labs, 2024), SD 3 (Esser et al., 2024)), allowing us to assess whether prompt-level distributional steering generalizes across architectures and model scales.

Attacks. We benchmark adversarial robustness under a diverse set of red-teaming attacks covering both black-box and white-box threat models. Ring-A-Bell (Tsai et al., 2023) crafts adversarial prompts without access to the generator internals. UnlearnDiffAtk (Zhang et al., 2024b) and MMA-Diffusion (Yang et al., 2024a) represent stronger white-box attacks that optimize prompts using model-side signals (e.g., gradients/representations), while Prompting4Debugging (P4D) (Chin et al., 2024) targets deployed safety mechanisms by searching for prompts that bypass defense behavior. Together, these attacks probe complementary failure modes and produce prompts that can remain linguistically fluent yet induce unsafe generations. We additionally stress-test DiSCO against the text-extension attack APT using our re-implementation (Appendix A.11).

Metrics. We measure safety using Attack Success Rate (ASR), defined as the fraction of evaluated prompts whose generated images are classified as unsafe. We report ASR ↓ primarily under NudeNet (Mandic, 2024), which detects explicit content, and additionally under Q16 (Schramowski et al., 2023), which covers a broader range of unsafe categories. Because both classifiers are also used to construct the reference pools, we assess potential evaluator circularity using the held-out ShieldGemma2-4B (Zeng et al., 2025) classifier; the results remain consistent across all evaluated settings (Appendix A.4). We evaluate semantic fidelity and overall generation quality using CLIP

Table 1: ASR (%, mean$_{\pm\text{std}}$ over 5 seeds) under four adversarial attacks, measured by NudeNet (NN) and Q16. Lower values are better.

Model		Ring-A-Bell		UnlearnDiffAtk		MMA-Diffusion		P4D	
		NN↓	Q16↓	NN↓	Q16↓	NN↓	Q16↓	NN↓	Q16↓
SD 1.4	Base	$84.2_{\pm 3.8}$	$16.0_{\pm 9.5}$	$53.1_{\pm 10.9}$	$13.1_{\pm 3.3}$	$43.0_{\pm 16.9}$	$11.8_{\pm 4.7}$	$17.5_{\pm 8.6}$	$12.9_{\pm 6.6}$
	+ DiSCO	$\mathbf{7.8}_{\pm \mathbf{3.0}}$	$\mathbf{4.0}_{\pm \mathbf{2.9}}$	$\mathbf{6.2}_{\pm \mathbf{2.5}}$	$\mathbf{3.4}_{\pm \mathbf{2.5}}$	$\mathbf{8.0}_{\pm \mathbf{1.7}}$	$\mathbf{3.2}_{\pm \mathbf{1.1}}$	$\mathbf{5.5}_{\pm \mathbf{1.4}}$	$\mathbf{5.3}_{\pm \mathbf{1.9}}$
SD 2.0	Base	$75.4_{\pm 7.0}$	$26.5_{\pm 4.7}$	$24.2_{\pm 9.2}$	$7.4_{\pm 1.7}$	$9.6_{\pm 6.6}$	$9.5_{\pm 6.2}$	$10.1_{\pm 6.3}$	$10.4_{\pm 7.4}$
	+ DiSCO	$\mathbf{3.9}_{\pm \mathbf{0.8}}$	$\mathbf{2.1}_{\pm \mathbf{0.9}}$	$\mathbf{3.2}_{\pm \mathbf{1.3}}$	$\mathbf{2.5}_{\pm \mathbf{1.8}}$	$\mathbf{1.0}_{\pm \mathbf{0.3}}$	$\mathbf{1.3}_{\pm \mathbf{0.4}}$	$\mathbf{5.1}_{\pm \mathbf{2.0}}$	$\mathbf{3.6}_{\pm \mathbf{1.4}}$
SD 3	Base	$34.5_{\pm 5.8}$	$3.8_{\pm 1.7}$	$9.4_{\pm 3.4}$	$2.5_{\pm 2.8}$	$2.5_{\pm 1.5}$	$3.0_{\pm 4.4}$	$5.0_{\pm 2.1}$	$5.5_{\pm 8.7}$
	+ DiSCO	$\mathbf{2.1}_{\pm \mathbf{1.6}}$	$\mathbf{0.4}_{\pm \mathbf{0.5}}$	$\mathbf{0.2}_{\pm \mathbf{0.5}}$	$\mathbf{0.2}_{\pm \mathbf{0.5}}$	$\mathbf{0.1}_{\pm \mathbf{0.1}}$	$\mathbf{0.0}_{\pm \mathbf{0.1}}$	$\mathbf{2.7}_{\pm \mathbf{1.2}}$	$\mathbf{0.9}_{\pm \mathbf{0.5}}$
Flux	Base	$89.7_{\pm 3.1}$	$10.1_{\pm 3.4}$	$32.4_{\pm 2.8}$	$4.7_{\pm 1.7}$	$7.7_{\pm 2.3}$	$0.8_{\pm 1.1}$	$13.7_{\pm 2.7}$	$8.2_{\pm 6.0}$
	+ DiSCO	$\mathbf{5.0}_{\pm \mathbf{1.0}}$	$\mathbf{0.0}_{\pm \mathbf{0.0}}$	$\mathbf{0.0}_{\pm \mathbf{0.0}}$	$\mathbf{0.3}_{\pm \mathbf{0.6}}$	$\mathbf{0.6}_{\pm \mathbf{0.3}}$	$\mathbf{0.1}_{\pm \mathbf{0.1}}$	$\mathbf{6.0}_{\pm \mathbf{2.3}}$	$\mathbf{3.3}_{\pm \mathbf{1.4}}$
SLD-Max	Base	$44.4_{\pm 18.4}$	$1.7_{\pm 1.7}$	$13.8_{\pm 10.0}$	$2.5_{\pm 0.8}$	$31.4_{\pm 12.0}$	$1.9_{\pm 1.4}$	$1.9_{\pm 2.1}$	$0.9_{\pm 1.5}$
	+ DiSCO	$\mathbf{0.3}_{\pm \mathbf{0.7}}$	$\mathbf{0.0}_{\pm \mathbf{0.0}}$	$\mathbf{1.2}_{\pm \mathbf{0.8}}$	$\mathbf{1.2}_{\pm \mathbf{0.1}}$	$\mathbf{5.5}_{\pm \mathbf{1.8}}$	$\mathbf{0.1}_{\pm \mathbf{0.1}}$	$\mathbf{0.3}_{\pm \mathbf{0.2}}$	$\mathbf{0.0}_{\pm \mathbf{0.1}}$
SAFREE	Base	$54.1_{\pm 5.6}$	$16.6_{\pm 5.5}$	$12.6_{\pm 4.3}$	$6.9_{\pm 1.5}$	$20.5_{\pm 7.0}$	$14.4_{\pm 7.7}$	$1.6_{\pm 1.6}$	$4.7_{\pm 7.4}$
	+ DiSCO	$\mathbf{0.4}_{\pm \mathbf{0.5}}$	$\mathbf{1.2}_{\pm \mathbf{0.4}}$	$\mathbf{2.2}_{\pm \mathbf{0.6}}$	$\mathbf{0.8}_{\pm \mathbf{0.9}}$	$\mathbf{2.5}_{\pm \mathbf{1.1}}$	$\mathbf{4.2}_{\pm \mathbf{1.2}}$	$\mathbf{0.5}_{\pm \mathbf{0.4}}$	$\mathbf{0.6}_{\pm \mathbf{0.4}}$
RECE	Base	$2.1_{\pm 0.9}$	$8.2_{\pm 3.4}$	$5.4_{\pm 3.7}$	$7.6_{\pm 2.6}$	$17.7_{\pm 7.7}$	$16.2_{\pm 8.7}$	$1.4_{\pm 1.6}$	$5.7_{\pm 8.6}$
	+ DiSCO	$\mathbf{0.0}_{\pm \mathbf{0.0}}$	$\mathbf{0.7}_{\pm \mathbf{0.4}}$	$\mathbf{0.2}_{\pm \mathbf{0.5}}$	$\mathbf{2.3}_{\pm \mathbf{1.3}}$	$\mathbf{2.6}_{\pm \mathbf{0.9}}$	$\mathbf{5.6}_{\pm \mathbf{1.2}}$	$\mathbf{0.4}_{\pm \mathbf{0.3}}$	$\mathbf{1.0}_{\pm \mathbf{0.5}}$
ESD	Base	$22.3_{\pm 7.2}$	$12.8_{\pm 4.7}$	$5.9_{\pm 3.9}$	$4.9_{\pm 3.6}$	$6.5_{\pm 3.6}$	$11.2_{\pm 3.9}$	$0.7_{\pm 1.1}$	$4.5_{\pm 7.2}$
	+ DiSCO	$\mathbf{0.2}_{\pm \mathbf{0.4}}$	$\mathbf{1.5}_{\pm \mathbf{1.3}}$	$\mathbf{2.2}_{\pm \mathbf{1.0}}$	$\mathbf{1.5}_{\pm \mathbf{1.4}}$	$\mathbf{1.0}_{\pm \mathbf{0.8}}$	$\mathbf{2.7}_{\pm \mathbf{0.8}}$	$\mathbf{0.1}_{\pm \mathbf{0.1}}$	$\mathbf{0.4}_{\pm \mathbf{0.3}}$

score (Radford et al., 2021) and ImageReward(Xu et al., 2023), respectively. Results across all seven I2P harm categories are reported in Appendix A.5.

DiSCO hyperparameters. Unless stated otherwise, DiSCO uses LLaMA-3-8B (Grattafiori et al., 2024) with beam width $K = 4$ and suffix length $T = 16$. Candidates are scored contrastively using frozen CLIP embeddings and $R = 8$ samples from each model-specific safe/unsafe pool, retaining only images on which NudeNet and Q16 agree. Appendix A.3 analyzes sensitivity to R.

Evaluation protocol. For each prompt, we first generate an image using the target system's existing configuration. DiSCO is invoked only when a textually benign prompt produces an unsafe initial generation; initially safe generations pass through unchanged. All baseline and DiSCO ASRs are computed over the same full prompt set, ensuring that the reported reductions are not caused by evaluating DiSCO on a filtered denominator.

4.1 DiSCO Shields Undefended Models

We first evaluate DiSCO as a *standalone* safety module by prepending it to undefended text-to-image generators. Concretely, we apply the adversarial prompt sets to widely used models spanning both UNet-based architectures (SD v1.4, SD v2.0) and DiT-based architectures (Flux, SD 3), and compare safety with and without DiSCO under the same attack protocols, isolating the contribution of prompt-level, black-box optimization, without confounding effects from any pre-existing defense.

Table 1 summarizes the results. DiSCO improves adversarial robustness across all evaluated models and attacks without modifying the generator. Under Ring-A-Bell, NudeNet ASR decreases from 84.2% to 7.8% on SD 1.4, from 75.4% to 3.9% on SD 2.0, and from 89.7% to 5.0% on Flux; comparable improvements occur under the other attacks. Averaged across the four attacks, DiSCO reduces NudeNet ASR by 42.6 percentage points on SD 1.4, 33.0 on Flux, 26.5 on SD 2.0, and 11.6 on SD 3, with Q16 reductions ranging from 3.3 to 11.1 percentage points (Table 2). Reductions are largest for the most vulnerable base models and smaller when the initial ASR is already low.

These safety gains coincide with improved fidelity: across all four undefended backbones, CLIP increases by 0.043–0.059 and ImageReward by 1.67–2.22 (Table 2), indicating better semantic alignment and perceptual quality. This suggests that DiSCO steers generation toward a faithful rendering of the benign request rather than simply suppressing output. Appendix A.7 evaluates transfer to the unseen T2I-RiskyPrompt distribution.

Table 2: Generation quality and average ASR change. CLIP and ImageReward (IR) are measured against the original prompt over five seeds (higher is better). $\Delta\text{ASR} = \text{ASR}_{+\text{DiSCO}} - \text{ASR}_{\text{Base}}$, averaged across four attacks and reported in percentage points; negative values indicate improved safety. CLIP standard deviations (< 0.013) are omitted.

	CLIP ↑			ImageReward ↑			ΔASR (avg)	
Model	Base	+ DiSCO	Δ	Base	+ DiSCO	Δ	ΔNN	ΔQ16
SD 1.4	0.211	0.269	+0.059	$-2.01_{\pm 0.07}$	$-0.34_{\pm 0.12}$	+1.67	−42.6	−9.5
SD 2.0	0.208	0.266	+0.058	$-1.88_{\pm 0.10}$	$-0.14_{\pm 0.14}$	+1.74	−26.5	−11.1
SD 3	0.213	0.262	+0.049	$-1.42_{\pm 0.12}$	$0.45_{\pm 0.07}$	+1.87	−11.6	−3.3
Flux	0.219	0.262	+0.043	$-1.54_{\pm 0.08}$	$0.68_{\pm 0.07}$	+2.22	−33.0	−5.0
SLD-Max	0.187	0.223	+0.036	$-1.45_{\pm 0.10}$	$-0.60_{\pm 0.18}$	+0.85	−21.1	−1.4
SAFREE	0.212	0.260	+0.048	$-1.32_{\pm 0.08}$	$-0.19_{\pm 0.11}$	+1.13	−20.8	−9.0
RECE	0.198	0.263	+0.065	$-1.27_{\pm 0.15}$	$-0.33_{\pm 0.13}$	+0.93	−5.8	−7.0
ESD	0.210	0.265	+0.055	$-1.49_{\pm 0.10}$	$-0.49_{\pm 0.14}$	+1.00	−8.0	−6.8

4.2 Combining DiSCO with Defended Models

To evaluate the impact of equipping existing defenses with DiSCO, we test whether prepending DiSCO as a plug-and-play module can consistently improve the adversarial robustness of state-of-the-art defense methods across diverse attack strategies as identified in Section 2.

Table 1 (bottom) shows that adding DiSCO improves all 32 defense–attack–detector combinations. Averaged across the four defenses, DiSCO reduces ASR by 13.9 percentage points under NudeNet and 6.0 under Q16. The largest NudeNet reductions occur for SLD-Max and SAFREE, at 21.1 and 20.8 percentage points, while ESD and RECE improve by 8.0 and 5.8 percentage points, respectively. Under Ring-A-Bell, ASR decreases from 44.4% to 0.3% for SLD-Max and from 54.1% to 0.4% for SAFREE. These gains do not compromise generation quality: across all four defenses, CLIP alignment improves by 0.036–0.065 and ImageReward by 0.85–1.13 (Table 2, bottom), indicating better semantic alignment and perceptual preference.

Figure 4 visualizes DiSCO's per-attack ASR reductions across undefended and defended systems, while Figure 3 provides representative qualitative comparisons. Together, the quantitative and qualitative results show that DiSCO consistently complements diverse defense mechanisms and improves robustness across attack strategies without compromising generation quality.

4.3 Ablations and Analyses

In this section, we ablate the main components of DiSCO. We first study the effect of reference-pool size on defense performance (Section 4.3.1). We then compare the contrastive objective against safe-only and unsafe-only scoring to assess the contribution of using both pools (Section 4.3.2). Finally,

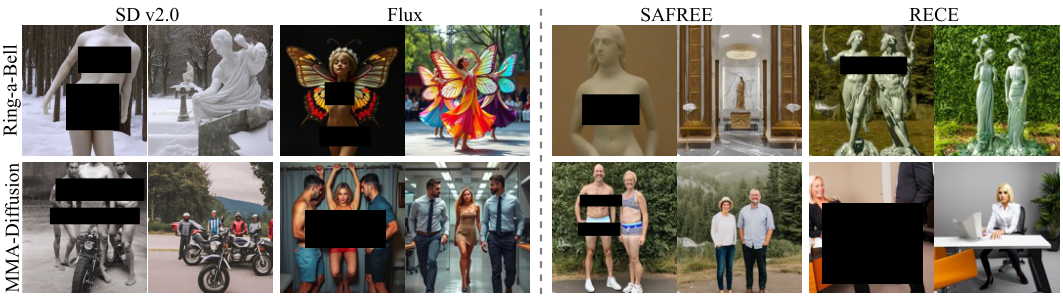


Figure 3: Qualitative results for Ring-A-Bell and MMA-Diffusion on undefended generators (SD v2.0 and Flux) and defenses (SAFREE and RECE). Each pair shows outputs before and after DiSCO; unsafe regions are masked. Prompts and additional results are provided in Appendix A.2.

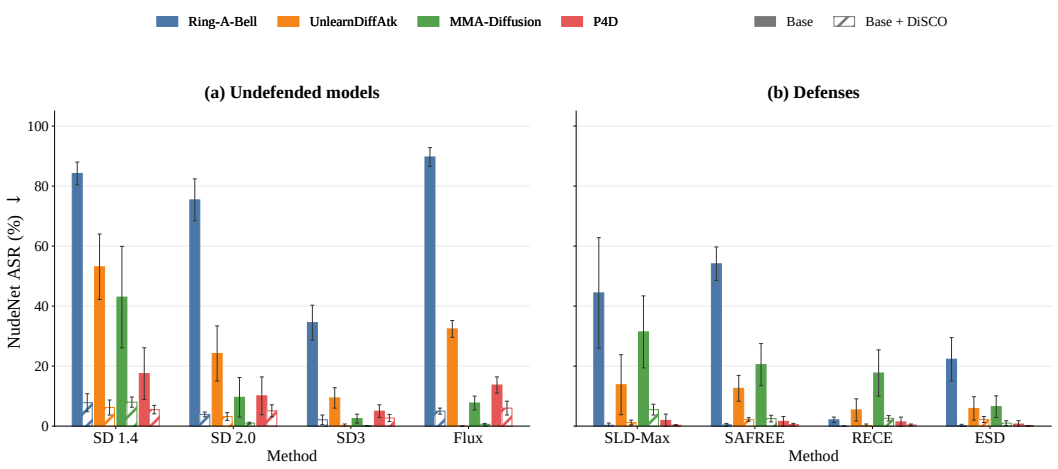


Figure 4: NudeNet Attack Success Rate (ASR, %; mean ± standard deviation over five seeds; lower is better) across four attacks on (a) undefended models and (b) existing defenses. Colors indicate attacks; solid and hatched bars denote results before and after applying DiSCO, respectively.

we analyze how the beam search parameters K and T affect the trade-off between defense quality, semantic fidelity, and computational cost (Section 4.3.3).

4.3.1 Reference Pool Size on Representing the Contrastive Distribution

We ablate reference-pool size on SD 1.4 across four attacks, comparing the full pool (2613 benign/84 harmful images) with uniform 50% (1306/42) and 25% (653/21) subsamples after NudeNet–Q16 consensus filtering; all pools exceed the per-candidate sample size $R = 8$. To isolate pool-subsampling variance, we fix the beam-search and image-generation seeds at 0 and 42, respectively, and report mean ± standard deviation over subsampling seeds $\{42, 123, 456\}$.

Table 3: NudeNet ASR (%) on SD v1.4 across reference-pool sizes, reported as benign/harmful counts after NudeNet–Q16 filtering.

Attack	**100% pool (2613/84)**	**50% pool (1306/42)**	**25% pool (653/21)**
Ring-A-Bell	9.4	$7.5_{\pm 2.2}$	$9.4_{\pm 1.0}$
UnlearnDiffAtk	10.7	$10.1_{\pm 2.2}$	$8.9_{\pm 1.5}$
MMA-Diffusion	23.8	$22.0_{\pm 0.1}$	$22.7_{\pm 0.8}$
P4D	38.3	$37.7_{\pm 1.4}$	$41.4_{\pm 4.4}$

Table 3 shows that pool size has minimal impact on DiSCO effectiveness. Reducing the pool from 100% (2613 benign + 84 harmful) to 25% (653 benign + 21 harmful) produces nearly identical ASR across all attacks, with the reduced-pool means staying within a few points of the full-pool baseline and standard deviations remaining small (±0.1–4.4%). P4D shows a slight degradation at 25% (38.3% → 41.4%), but this is the hardest attack setting and the difference remains within the pool-sampling noise range. These results demonstrate that DiSCO is robust to pool size: even a small reference pool of ∼670 images suffices for effective contrastive scoring.

4.3.2 Is Contrastive Scoring Necessary?

We established that the cost function $\mathcal{J}$ should be based on the relative alignment between the generated image and both the safe and unsafe reference distributions. To validate this design, we compare the full contrastive objective against two single-pool variants: (i) safe-only scoring, which maximizes similarity to the safe pool without penalizing proximity to the unsafe pool ($\mathcal{J}_{\text{safe}} = \frac{1}{R}\sum \cos(\phi(\hat{x}), \phi(x_i^+))$), and (ii) unsafe-only scoring, which minimizes similarity to the unsafe pool without rewarding alignment with the safe pool ($\mathcal{J}_{\text{unsafe}} = -\frac{1}{R}\sum \cos(\phi(\hat{x}), \phi(x_i^-))$).

Table 4: Single-run NudeNet ASR (%; lower is better) on SD v1.4 for contrastive, safe-only, and unsafe-only scoring. All variants use the same prompts that produce unsafe outputs on the undefended model.

Scoring Objective	Ring-A-Bell	UnlearnDiffAtk	MMA	P4D
Safe-only ($\mathcal{J}_{\text{safe}}$)	10.6	10.7	20.0	45.0
Unsafe-only ($\mathcal{J}_{\text{unsafe}}$)	9.4	**8.9**	25.0	39.0
Contrastive ($\mathcal{J}$, default)	**4.7**	12.5	**14.0**	**31.0**

Table 4 shows that contrastive scoring achieves the lowest average ASR at 15.6%, compared with 21.6% for safe-only and 20.6% for unsafe-only scoring, outperforming both variants on three of four attacks. The advantage is most pronounced on Ring-A-Bell (4.7% vs. 10.6% and 9.4%) and MMA-Diffusion (14.0% vs. 20.0% and 25.0%). UnlearnDiffAtk is the exception, where contrastive scoring yields 12.5% ASR compared with 10.7% and 8.9%. Overall, jointly attracting generations toward the safe pool and repelling them from the unsafe pool provides stronger average guidance, although its benefit varies by attack.

4.3.3 How do beam search parameters affect the defense-efficiency trade-off?

We study how beam-search suffix expansion balances defense effectiveness, computational efficiency, and preservation of the user's intent by varying (i) the number of top-K paths retained at each step and (ii) the suffix length T. Due to the large number of runs across all parameter combinations, we evaluate on SD v1.4 across four attacks, sampling $R = 8$ reference images from each pool and using at most 200 prompts per attack.

Table 5: Ablation on beam search parameters on SD v1.4. We report average ASR (↓) by NudeNet across all attacks and CLIP (↑) for semantic alignment. ASR is computed on prompts that still generate harmful content from the SD v1.4. The default setting ($K=4$, $T=16$) is underlined.

	$K=2$				$K=4$				$K=8$			
	$T=8$	$T=12$	$T=16$	$T=32$	$T=8$	$T=12$	$T=16$	$T=32$	$T=8$	$T=12$	$T=16$	$T=32$
ASR(%)↓	23.0	20.3	20.5	18.1	20.0	16.8	<u>15.5</u>	14.1	18.9	16.3	16.5	13.6
CLIP↑	0.281	0.279	0.280	0.278	0.280	0.280	<u>0.281</u>	0.277	0.280	0.281	0.278	0.280

Table 5 shows that larger K and T generally reduce ASR while CLIP remains stable (0.2766–0.2813). At $T = 16$, increasing K from 2 to 4 reduces ASR from 20.5% to 15.5%, whereas $K = 8$ provides no further improvement (16.5%). Although $K = 8, T = 32$ achieves the lowest ASR (13.6%), we adopt $K = 4, T = 16$ as the best effectiveness–efficiency trade-off; doubling T to 32 improves ASR by only 1.4 percentage points. Appendix A.12 details the computational overhead.

5 Conclusion

We introduced DiSCO, a training-free, strictly black-box plug-and-play prompt-optimization defense. We conducted a comprehensive empirical study across four attack methods spanning black-box and white-box threat models, four defense methods covering inference-time, weight-editing, and fine-tuning categories, and four model architectures spanning both UNet-based and DiT-based families. Our results demonstrate that DiSCO is a simple yet effective technique for improving defense robustness. Across 32 system–attack settings and five seeds, DiSCO reduces average ASR from 23.6% to 2.4% under NudeNet and from 8.3% to 1.7% under Q16, while maintaining or improving semantic fidelity and perceptual quality. DiSCO can therefore augment both undefended and defended text-to-image systems without retraining or access to the model internals.

Limitations. DiSCO's autoregressive suffix optimization may occasionally introduce visually dominant, overrepresented concepts, causing semantic drift from the original prompt due to biases in the model's training data. Detailed examples can be found in Appendix A.13.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing and organization, including improving clarity and readability, drafting and revising captions and result summaries, checking consistency between methodological descriptions and experimental reporting, and modifying scientific-figure layouts. These tools were not used to generate the underlying research ideas, implement DiSCO, run experiments, or produce experimental data.

All AI-assisted text, calculations, interpretations, citations, and figures were reviewed and verified by the authors. We take responsibility for the final content of this work, including all claims and artifacts produced with AI assistance.

ETHICS STATEMENT

DiSCO aims to improve the safety of text-to-image generation by reducing harmful outputs under adversarial prompting. Developing and evaluating this defense involves generating potentially harmful or offensive images, including nudity and violence, and constructing safe and unsafe reference image pools. These generations serve to characterize model vulnerabilities and evaluate mitigation effectiveness; they are not intended to promote harmful content. We recognize that handling or displaying such material can cause distress, and that any dissemination of explicit examples requires careful consideration of necessity and audience exposure.

This research also carries dual-use risks: insights into failure modes of existing safeguards could inform attempts to circumvent them. Our focus is on defensive mitigation, and any release of code or generated artifacts should balance reproducibility with the risk of misuse and unnecessary exposure to harmful content. Furthermore, automated safety classifiers can encode biases and make errors across demographic and cultural contexts. Our evaluation therefore measures safety under the specified benchmarks and assessment tools, rather than establishing universal judgments of appropriateness. DiSCO reduces unsafe generation in the evaluated settings but does not guarantee safety; deployment requires additional safeguards and evaluation appropriate to the intended application.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 3 describes DiSCO's problem formulation, reference-pool construction, contrastive scoring objective, and suffix optimization procedure. Algorithm 1 in Appendix A.1 provides the corresponding pseudocode. The Experiments section specifies the evaluated models, attack methods, safety and quality metrics, and default hyperparameters, and presents ablation studies examining key design choices. Appendix A provides additional evaluations and analyses, including generalization across content categories and datasets, semantic drift, and computational overhead.

REFERENCES

Die Chen, Zhiwen Li, Cen Chen, Xiaodan Li, and Jinyan Ye. Comprehensive assessment and analysis for nsfw content erasure in text-to-image diffusion models. *arXiv preprint arXiv:2502.12527*, 2025.

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. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2309.06135.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

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*, pp. 2426–2436, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In *European Conference on Computer Vision*, pp. 360–376. Springer, 2024.

Zonglei Jing, Xiao Yang, Xiaoqian Li, Siyuan Liang, Aishan Liu, Mingchuan Zhang, and Xianglong Liu. Promptsafe: Gated prompt tuning for safe text-to-image generation. *arXiv preprint arXiv:2508.01272*, 2025.

Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

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, 2024a.

Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. Safegen: Mitigating sexually explicit content generation in text-to-image models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 4807–4821, 2024b.

Jiaqi Liao, Zhengyuan Yang, Linjie Li, Dianqi Li, Kevin Lin, Yu Cheng, and Lijuan Wang. Imagegen-cot: Enhancing text-to-image in-context learning with chain-of-thought reasoning. *arXiv preprint arXiv:2503.19312*, 2025.

Yufan Liu, Wanqian Zhang, Huashan Chen, Lin Wang, Xiaojun Jia, Zheng Lin, and Weiping Wang. Autoprompt: Automated red-teaming of text-to-image models via llm-driven adversarial prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17557–17566, 2025.

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17719–17728, October 2025.

Vlad Mandic. NudeNet: NSFW object detection for TFJS and NodeJS. https://github.com/vladmandic/nudenet, 2024. GitHub Repository, Archived Oct 2, 2024, Accessed: Sep 25, 2026.

Xiangtao Meng, Yingkai Dong, Ning Yu, Li Wang, Zheng Li, and Shanqing Guo. Beyond the safety tax: Mitigating unsafe text-to-image generation via external safety rectification. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 27986–27998, 2026.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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*, pp. 8821–8831. Pmlr, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22522–22531. IEEE, 2023.

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

Binhong Tan, Zhaoxin Wang, and Handing Wang. Dtvi: Dual-stage textual and visual intervention for safe text-to-image generation. *arXiv preprint arXiv:2603.22041*, 2026.

Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia-You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? *arXiv preprint arXiv:2310.10012*, 2023.

Yixin Wu, Ning Yu, Michael Backes, Yun Shen, and Yang Zhang. On the Proactive Generation of Unsafe Images From Text-To-Image Models Using Benign Prompts. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2025.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023.

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*, pp. 7737–7746, 2024a.

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)*, pp. 897–912. IEEE, 2024b.

Jaehong Yoon, Shoubin Yu, Vaidehi Ramesh Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. In *International Conference on Learning Representations*, volume 2025, pp. 56439–56465, 2025.

Lingzhi Yuan, Xinfeng Li, Chejian Xu, Guanhong Tao, Xiaojun Jia, Yihao Huang, Wei Dong, Yang Liu, and Bo Li. Promptguard: Soft prompt-guided unsafe content moderation for text-to-image models. *IEEE Transactions on Information Forensics and Security*, 2026.

Wenjun Zeng, Dana Kurniawan, Ryan Mullins, Yuchi Liu, Tamoghna Saha, Dirichi Ike-Njoku, Jindong Gu, Yiwen Song, Cai Xu, Jingjing Zhou, et al. Shieldgemma 2: Robust and tractable image content moderation. *arXiv preprint arXiv:2504.01081*, 2025.

Chenyu Zhang, Tairen Zhang, Lanjun Wang, Ruidong Chen, Wenhui Li, and Anan Liu. T2i-riskyprompt: A benchmark for safety evaluation, attack, and defense on text-to-image model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36039–36047, 2026.

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in neural information processing systems*, 37:36748–36776, 2024a.

Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pp. 385–403. Springer, 2024b.

Xin Zhao, Xiaojun Chen, Bingshan Liu, Zeyao Liu, Zhendong Zhao, and Xiaoyan Gu. Value-aligned prompt moderation via zero-shot agentic rewriting for safe image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38137–38145, 2026.