# MacPrompt: Maraconic-guided Jailbreak against Text-to-Image Models

**Xi Ye[1], Yiwen Liu[1], Lina Wang[1]*, Run Wang[1]*, Geying Yang[2], Yufei Hou[1], Jiayi Yu[1]**

[1]Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, China

[2]School of Cyber Science and Engineering, Tianjin University, China.

{xixiye,yiwenliu,lnwang,wangrun}@whu.edu.cn, yanggeying@tju.edu.cn, {fei98520,yujiayi}@whu.edu.cn

## Abstract

Text-to-image (T2I) models have raised increasing safety concerns due to their capacity to generate NSFW and other banned objects. To mitigate these risks, safety filters and concept removal techniques have been introduced to block inappropriate prompts or erase sensitive concepts from the models. However, all the existing defense methods are not well prepared to handle diverse adversarial prompts. In this work, we introduce MacPrompt, a novel black-box and cross-lingual attack that reveals previously overlooked vulnerabilities in T2I safety mechanisms. Unlike existing attacks that rely on synonym substitution or prompt obfuscation, MacPrompt constructs macaronic adversarial prompts by performing cross-lingual character-level recombination of harmful terms, enabling fine-grained control over both semantics and appearance. By leveraging this design, MacPrompt crafts prompts with high semantic similarity to the original harmful inputs (up to 0.96) while bypassing major safety filters (up to 100%). More critically, it achieves attack success rates as high as 92% for sex-related content and 90% for violence, effectively breaking even state-of-the-art concept removal defenses. These results underscore the pressing need to reassess the robustness of existing T2I safety mechanisms against linguistically diverse and fine-grained adversarial strategies.

**Warning: This paper includes sensitive examples (e.g., adult, violent, or illegal content). Unsafe images are masked but may still be disturbing.**

## Introduction

Text-to-image (T2I) models have emerged as a powerful generative model, capable of synthesizing high-fidelity images from natural language descriptions (Caramiaux et al. 2025; Saharia et al. 2022). Driven by large-scale datasets, state-of-the-art (SOTA) models such as Stable Diffusion (SD) (Rombach et al. 2022), DALL·E (Ramesh et al. 2022), and Midjourney (MidJourney 2023) have been widely deployed through public APIs and creative platforms (Xu et al. 2017; Zhang et al. 2017; Pernias et al. 2023). These models have significantly enhanced image creation workflows, offering users unprecedented accessibility, efficiency, and expressive power in visual content generation. However, the training datasets used for these models are typically collected directly from the Internet without rigorous content filtering, resulting in the inclusion of inappropriate or harmful material that can be inadvertently learned by the models. This poses a significant risk of generating NSFW (Not Safe for Work) content or banned objects (Qu et al. 2023; Schramowski et al. 2023; Naik and Nushi 2023). For instance, Unstable Diffusion (Unstable Diffusion 2025) openly provides unrestricted access to powerful T2I models, which has enabled malicious users to generate and disseminate violent, pornographic, or otherwise harmful images, leading to serious social concerns (Gupta 2022).


Figure 1: Cross-lingual prompts composed from other languages can trigger the same visual semantics as the original English prompt in SD v2.1.

To prevent such misuse, most mainstream T2I online services have deployed black-box safety mechanisms to restrict the generation of inappropriate content. For open-source models like SD, multiple built-in plugin-based safety filters have been introduced, including text filters that block harmful prompts without producing a response (George 2023; Jieli 2023; Liu et al. 2025), and image filters that identify unsafe outputs and return black images (Zeng et al. 2025). Beyond these built-in filters, a much stronger line of external defense, known as concept removal (Gandikota et al. 2023; Schramowski et al. 2023), directly modifies the underlying SD model to eliminate its ability to a wide range of NSFW concepts, such as sex, violence, and self-harm. Together, these defenses form a comprehensive protection framework that is widely believed to be effective against harmful inputs. Nonetheless, recent work (Chin et al. 2024; Yang et al. 2024b; Zhang et al. 2024b; Tsai et al. 2024; Ma et al. 2024; Yang et al. 2024c; Gao et al. 2024) shows that even such

*Lina Wang and Run Wang are corresponding authors.

rigorous defenses remain vulnerable to adversarial prompts, motivating continued research into more effective attack and defense strategies. These existing attack works typically focus on circumventing only one type of defense: either the built-in filters or the external concept removal mechanisms. Those capable of bypassing both simultaneously generally rely on additional privileged knowledge about the target model, such as its internal architecture (Chin et al. 2024; Yang et al. 2024b; Tsai et al. 2024; Gao et al. 2024). This reliance limits their practicality in real-world black-box settings, leaving a gap for unified, black-box attacks that can effectively overcome both defense types without requiring external information.

Motivated by the observation that visual concepts in T2I models can be reliably triggered by cross-lingually composed prompts as shown in Fig. 1, we propose MacPrompt, a black-box attack that exploits a common weakness in both built-in safety filters and external concept removal defenses, which often fail to detect obfuscated harmful inputs across languages. MacPrompt constructs adversarial prompts by replacing sensitive words with macaronic substitutes generated through cross-lingual character-level recombination. This design preserves harmful semantics while avoiding detection by typical text-based filtering mechanisms. Our method applies to various defense strategies, and achieves particularly strong results against concept removal models, which are generally considered harder to attack. Experiments show that MacPrompt reaches up to 0.96 semantic similarity with harmful inputs and achieves attack success rates of 92% on sex-related prompts and 90% on violence-related prompts, outperforming existing baselines. The main contributions of our scheme are as follows:

- We investigate the effectiveness of macaronic words, which are created by recombining character-level substrings from translation-equivalent words across multiple languages, in preserving visual semantics while obfuscating textual embeddings. Our analysis demonstrates their ability to activate restricted concepts in T2I models and highlights the susceptibility of SD to such cross-lingual adversarial prompts.
- We propose MacPrompt, a novel black-box attack framework that requires no access to model internals and is both practical and broadly applicable in the real-word scenarios.
- Extensive experiments across various defense strategies demonstrate that our method can effectively bypass both input text filters and concept removal defenses to generate NSFW content or banned objects.
- We pose a new research direction toward cross-lingual adversarial robustness in generative models, emphasizing the need to rethink current safety mechanisms and develop defenses capable of generalizing beyond monolingual assumptions and simple keyword matching.

# Related Work

## T2I Models with Defense

For T2I models, two primary defense strategies are currently employed to prevent the generation of NSFW content. The first approach involves adding a text filter before the existing model or/and an image filter after it without altering the core architecture. Specifically, text filters can be further classified into two types: text-match and text-classifier. The text-match filter relies on a blacklist-based keyword matching mechanism to detect and block inappropriate content within input prompts (George 2023). The text-classifier filter trains a classifier to perform binary classification, distinguishing between harmful and harmless prompts (Jieli 2023; Liu et al. 2025). Meanwhile, image filters operate by detecting NSFW content within the images generated by the T2I model with given prompts (Zeng et al. 2025). Additionally, online T2I models, such as DALL·E 2 (Ramesh et al. 2022) and Midjourney (MidJourney 2023), deploy proprietary black-box safety filters to prevent users from generating NSFW content, ensuring an additional layer of protection.

The second defense strategy revolves around concept removal, aiming at encouraging T2I models to forget NSFW concepts during the image generation process. For instance, ESD (Gandikota et al. 2023) and FMN (Zhang et al. 2024a) finetune pretrained DM weights to remove NSFW concepts. SLD (Schramowski et al. 2023) suppresses NSFW content during the denoising process, while SafeGen (Li et al. 2024b) modifies the visual self-attention layers of pretrained models to eliminate NSFW representations. DUO (Park et al. 2024) uses direct preference optimization to selectively forget NSFW features while preserving normal concepts. EAP (Bui et al. 2024) argues that retaining a neutral concept alone is insufficient and emphasizes the need to prioritize sensitive concepts. Finally, PromptGuard (Yuan et al. 2025) adopts a divide-and-conquer approach by introducing safety pseudowords, optimizing specific types of NSFW, and combining them into a comprehensive defense mechanism for better performance.

## NSFW Attack against T2I Models

An NSFW attack against T2I models aims to generate harmful images containing NSFW content by designing specific adversarial prompts without modifying the T2I model itself. Depending on the information available to the attacker, these attacks can be classified into three categories: white-box, gray-box, and black-box. In the white-box setting (Chin et al. 2024; Yang et al. 2024b; Zhao, Chen, and Gao 2024; Xu et al. 2025), it is assumed that the attacker has full access to the information of target T2I models, including their architecture and specific weights. Under the gray-box mechanism (Zhuang, Zhang, and Liu 2023; Tsai et al. 2024; Ma et al. 2024; Yang et al. 2024c; Gao et al. 2024), the attacker is only able to access partial information or make use of auxiliary tools such as text encoders and image encoders. The black-box mechanism (Deng and Chen 2023; Yang et al. 2024d; Dang et al. 2024; Ba et al. 2024; Li et al. 2024a; Huang et al. 2025), however, assumes that the attacker has no knowledge but images generated by target T2I models with inputted prompts. In this scenario, the attacker interacts with the T2I model and iteratively modifies the prompt based on the feedback to achieve desired attacks. Currently, DiffZOO (Dang et al. 2024) is the only black-box-based attack that targets concept removal models, while most other

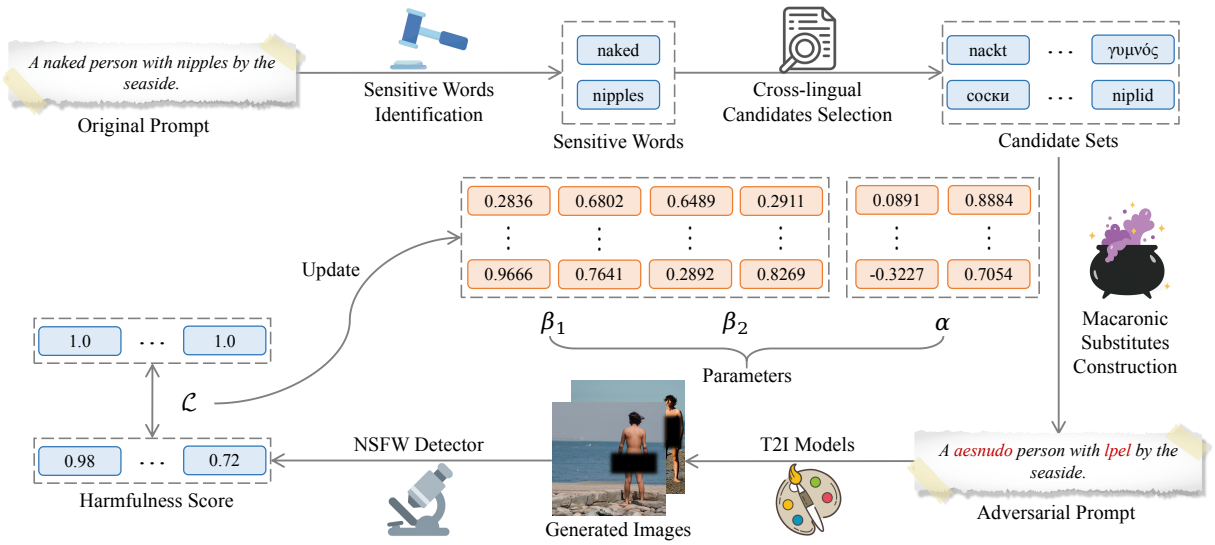


Figure 2: Overview of the MacPrompt framework. We assume the original prompt contains sensitive words and is blocked by existing safety filters. Here, $\beta_1, \beta_2$, and $\alpha$ are optimization parameters during macaronic substitutes construction process.

black-box-based attacks, such as PGJ (Huang et al. 2025) and SurrogatePrompt (Ba et al. 2024), specifically target online T2I platforms. However, this scheme relies on synonym replacements, making it ineffective against text filters.

## Methodology

### Problem Formulation

Cross-lingual adversarial prompt generation for T2I models is an adversarial evaluation process aimed at assessing model robustness by constructing multilingual mixed adversarial prompts that preserve harmful semantics while bypassing safety filters. Formally, given original harmful prompt $p_{\text{ori}}$, safety filter $\mathcal{F}$, and target T2I model $\mathcal{G}$, the goal is to generate an adversarial prompt $p_{\text{adv}}$, such that

$$\begin{cases} \mathcal{F}(p_{\text{adv}}) = \text{False} \\ \mathcal{G}(p_{\text{adv}}) \approx \mathcal{G}(p_{\text{ori}}) \end{cases}. \quad (1)$$

Here, $\mathcal{F}(p) = \text{False}$ indicates that the input prompt $p$ successfully bypasses $\mathcal{F}$, whereas $\mathcal{F}(p) = \text{True}$ indicates it is flagged as unsafe. The symbol $\approx$ denotes that the generated image from $p_{\text{adv}}$ is visually similar to that from $p_{\text{ori}}$. Specifically, we require that: (1) $\mathcal{G}(p_{\text{adv}})$ exhibits harmful content, and (2) both images are semantically aligned.

The threat model considered in this work assumes a black-box adversary who has no access to the internal parameters or architectures of $\mathcal{F}$ and $\mathcal{G}$. However, the adversary can query the system by submitting a prompt and observing the returned image or whether the prompt is rejected. The target T2I models $\mathcal{G}$ include both SD equipped with safety filters, as well as concept removal models. Meanwhile, $\mathcal{F}$ may implement a variety of techniques, including:

- Keyword-based matching, which detects prompts containing blacklisted terms (Yang et al. 2024b);
- Semantic matching using pretrained text classifiers such as BERT-based NSFW text classifiers (Jieli 2023);
- Latent representation filtering such as LatentGuard (Liu et al. 2025).

### Overview

**Key Intuition** Although SD is officially documented to support English prompts, we observe that prompts in other languages (e.g., French, German, Danish) can also generate semantically similar images. This phenomenon exits both in the original SD and concept removal variants that are fine-tuned on SD. However, directly using non-English prompts often fails to bypass $\mathcal{F}$, as such prompts may still trigger components based on semantic matching or latent representation filtering. Interestingly, previous work (Raphaël 2022) has shown that concatenating words from multiple languages, can preserve the visual semantics of the original prompt.

Building on this insight, we further investigate the behavior of multilingual mixed prompts and uncover a more nuanced phenomenon: **certain cross-lingual combinations not only retain the visual semantics of the original prompt but also exhibit significant divergence in textual semantics**, allowing them to evade safety filters while still eliciting harmful image outputs. Empirically, we find that token-level alignment plays a key role in this process. Specifically, when the newly constructed word is formed by directly concatenating tokens that match the tokenization of the original sensitive word, i.e.,

$$p_{\text{adv}} = \text{Concat}(t_1, t_2, \ldots, t_j), t_i \in T, j \in [1, |T|)], \quad (2)$$

the resulting prompt is more likely to preserve the target image semantics, even though its textual meaning differs considerably. $T$ is the token set containing tokens of words from different languages, $|T|$ means the size of $T$.

However, in models like SD, the tokenization process is inherently non-invertible, especially for low-resource or non-Latin languages (Hwang, Wang, and Gu 2025; Tamang

and Bora 2024). This means that even if the target token sequence is known, it is often infeasible to construct an input word or phrase that will be tokenized into that exact sequence, thereby complicating precise token-level adversarial manipulation.To overcome this challenge, we propose a macaronic substitute construction framework that operates at the character level, generating cross-lingual word combinations to bypass $\mathcal{F}$ and trigger harmful outputs. More related exploration results are presented in the Appendix.

**Overall Pipeline** Figure 2 describes the overall pipeline of MacPrompt in constructing macaronic substitutes to form an adversarial prompt to evade defenses and generate NSFW content. Given an initially harmful prompt, MacPrompt first detects sensitive words and selects appropriate candidates across multiple languages. By recombining character-level substrings from these multilingual candidates under parameterized control, it constructs visually meaningful yet textually obfuscated macaronic substitutes. These substitutes replace the original sensitive terms to form a modified adversarial prompt, which is then input into the T2I model to produce images. A pre-trained NSFW detector assesses the harmfulness of the generated output, and the resulting score serves as a feedback signal to iteratively update the parameters using a zero-order optimization (ZOO) strategy.

## Sensitive Words Identification

Given the original prompt $p_{\text{ori}}$, the first step is to identify *sensitive words*, which refer to those specific terms within $p_{\text{ori}}$ that are responsible for triggering $\mathcal{F}$. To detect such sensitive words, we adopt two strategies: (1) Blacklist Matching. Each word in $p_{\text{ori}}$ is compared against a predefined set of harmful words (Yang et al. 2024b). If a match is found, the word is labeled as sensitive. (2) Semantic Similarity Scoring. For broader coverage beyond exact matches, the cosine similarity is computed between the embedding of each word $w_i \in p_{\text{ori}}$ and a set of pre-collected harmful concept embeddings $e^j_{\text{harm}}$. If

$$\max_j \cos(\text{Embed}(w_i), e^j_{\text{harm}}) > \tau, \tag{3}$$

the word is considered semantically harmful.

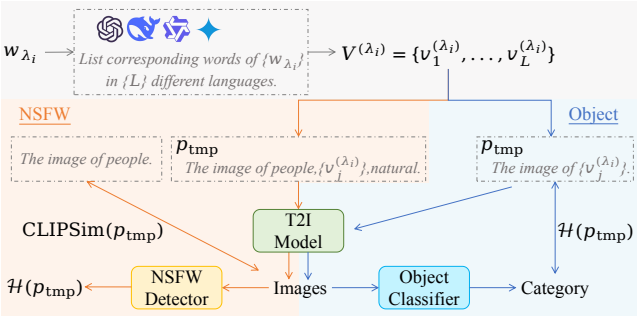


Figure 3: Cross-lingual candidates selection pipeline. The sensitive word is translated into multiple languages, inserted into templates to generate images, and evaluated.

## Cross-lingual Candidates Selection

Since not all multilingual words contribute equally to preserving harmful visual intent, a filtering step is required to identify those candidates that are most effective in either triggering the target NSFW concept or accurately representing a banned object. The selection process is guided by two objectives: (1) preserving the original harmful semantics, and (2) minimizing semantic interference with the rest of the prompt. Given an input prompt $p_{\text{ori}} = \{w_1, w_2, \ldots, w_n\}$ with identified sensitive words $\{w_{\lambda_1}, w_{\lambda_2}, \ldots, w_{\lambda_m}\} \subseteq p_{\text{ori}}$, a multilingual substitution set is constructed for each sensitive word $w_{\lambda_i}$. Specifically, as shown in Fig. 3, a lexical candidate pool is generated as

$$V^{(\lambda_i)} = \{v_1^{(\lambda_i)}, \cdots, v_j^{(\lambda_i)}, \ldots, v_L^{(\lambda_i)}\}, \tag{4}$$

where each $v_j^{(\lambda_i)}$ denotes a translation or paraphrase of $w_{\lambda_i}$ in one of $L$ (typically $L = 79$) different languages, obtained via a large language model (LLM).

To evaluate the harmfulness-preserving property of $v_j^{\lambda_i}$, a task-specific prompt template $p_{\text{tmp}}$ is designed as *"The image of people, $<v_j^{\lambda_i}>$, natural."* for NSFW concept or *"The image of $<v_j^{\lambda_i}>$."* for banned objects. Then the combined prompt is used to generate ten images via the T2I model. The resulting image is then evaluated by two metrics:

- Harmfulness Score $\mathcal{H}(p_{\text{tmp}})$: calculated as the target class probability by applying a pretrained NSFW detector or object classifier to the generated images.
- Visual Semantic Similarity CLIPSim($p_{\text{tmp}}$): calculated as the CLIP score between images generated from $v_j^{\lambda_i}$ and a safe prompt *"The image of people, natural."* related to $p_{\text{ori}}$. This metric is used only for NSFW concept evaluation.

All candidates are ranked according to a composite score combining $\mathcal{H}(v_j^{\lambda_i})$ and CLIPSim($v_j^{\lambda_i}$), if applicable. The top-$k$ candidates (typically $k = 10$) are retained as the final cross-lingual candidate set $\hat{V}^{(\lambda_i)}$ for $w_{\lambda_i}$.

## Macaronic substitutes Construction

In models like SD, which are primarily trained on English text and lacks native multilingual support, tokenization becomes particularly problematic when handling non-English inputs, especially those from low-resource or underrepresented languages. In such cases, tokenization is often non-invertible, meaning that:

$$\epsilon(\epsilon^{-1}(\epsilon(v))) \neq \epsilon(v), \tag{5}$$

where $\epsilon$ and $\epsilon^{-1}$ denote the tokenizer (i.e., encoding function) and its corresponding decoding function, respectively. This non-invertibility limits direct manipulation of tokens, making it impractical to concatenate or modify token-level representations to produce target adversarial prompts. To overcome this challenge, we propose a character-level macaronic substitute construction strategy to preserve the original harmful concept. By directly slicing and recombining substrings at the character level, this method bypasses the limitations of multilingual tokenizers and enables fine-grained, language-agnostic prompt manipulation.


Figure 4: Visualization of images generated from adversarial prompts targeting NSFW concepts and banned objects across different models; the top shows outputs for sexual and violent concepts, while the bottom shows banned objects.

**Construction** Given a set of identified sensitive words $\{w_{\lambda_1}, \ldots, w_{\lambda_m}\}$ targeted for substitution, along with their corresponding candidate sets $\hat{\mathbf{V}} = \{\hat{V}^{(\lambda_1)}, \ldots, \hat{V}^{(\lambda_m)}\}$, where each $\hat{V}^{(\lambda_i)} = \{\hat{v}_1^{(\lambda_i)}, \ldots, \hat{v}_k^{(\lambda_i)}\}$ contains $k$ substitution candidates of $w_{\lambda_i}$, we define three character-level parameters for each $w_{\lambda_i}$ to control the construction of macaronic substitutes. Specifically, for each candidate set $\hat{V}^{(\lambda_i)}$, two continuous selection boundary parameters are first defined as

$$\begin{cases} \beta_1^{(\lambda_i)} = [\beta_{1,1}^{(\lambda_i)}, \ldots, \beta_{1,k}^{(\lambda_i)}] \\ \beta_2^{(\lambda_i)} = [\beta_{2,1}^{(\lambda_i)}, \ldots, \beta_{2,k}^{(\lambda_i)}] \end{cases} \tag{6}$$

where $\beta_{1,j}^{(\lambda_i)}, \beta_{2,j}^{(\lambda_i)} \in [0,1]$ correspond to the normalized start and end positions of the selected substring from the $j$-th candidate word $\hat{v}_j^{(\lambda_i)}$, respectively. A valid segment is extracted only when $\beta_{2,j}^{(\lambda_i)} > \beta_{1,j}^{(\lambda_i)}$, ensuring that the end position follows the start. In addition to the boundary parameters, we define a continuous ordering parameter $\alpha^{(\lambda_i)} = [\alpha_1^{(\lambda_i)}, \ldots, \alpha_k^{(\lambda_i)}]$, where each $\alpha_j^{(\lambda_i)}$ represents the relative ordering weight of the selected substring from $\hat{v}_j^{(\lambda_i)}$ in the final composition.

Based on the above parameters, we generate the final macaronic substitutes by extracting and recombining character-level substrings from candidate words. For each $\hat{v}_j^{(\lambda_i)}$ in the candidate set, the start and end indexes of the selected substring are computed using the boundary parameters:

$$\begin{cases} \mu_{1,j}^{(\lambda_i)} = \lfloor l_j \cdot \beta_{1,j}^{(\lambda_i)} \rfloor \\ \mu_{2,j}^{(\lambda_i)} = \begin{cases} \lfloor l_j \cdot \beta_{2,j}^{(\lambda_i)} \rfloor & \text{if } \beta_{2,j}^{(\lambda_i)} \geq \beta_{1,j}^{(\lambda_i)} \\ \mu_{1,j}^{(\lambda_i)} & \text{otherwise} \end{cases} \end{cases}, \tag{7}$$

where $l_j$ denotes the character length of $\hat{v}_j^{(\lambda_i)}$. As a result, the substring from $\hat{v}_j^{(\lambda_i)}$ is extracted by

$$\bar{v}_j^{(\lambda_i)} = \hat{v}_j(\mu_{1,j}^{(\lambda_i)} : \mu_{2,j}^{(\lambda_i)}). \tag{8}$$

All fragments $\bar{v}_j^{(\lambda_i)}$ are then sorted in descending order according to their corresponding $\alpha_j^{(\lambda_i)}$ values and concatenated to form the macaronic substitute $\bar{w}_{\lambda_i} = \text{Contact}(\bar{v}_j^{(\lambda_i)}|_{j=1}^k)$ for $w_{\lambda_i}$. After obtaining the set of substitutes $\{\bar{w}_{\lambda_1}, \ldots, \bar{w}_{\lambda_m}\}$, all original sensitive word $\{w_{\lambda_1}, \ldots, w_{\lambda_m}\}$ in $p_{\text{ori}}$ are replaced to construct $p_{\text{adv}}$.

**Optimization** To optimize $(\beta_1, \beta_2, \alpha)$ defined above, ZOO is adopted to maximize the generation likelihood of NSFW concept or banned objects. Given the loss function

$$\mathcal{L} = ||\mathcal{H}(p_{\text{adv}}) - \mathbf{1}||_2, \tag{9}$$

where $p_{\text{adv}} = \text{Macaronic}(p_{\text{ori}}, \beta_1, \beta_2, \alpha)$ and $\mathbf{1}$ denotes a vector of ones with the same dimensionality as the output of $\mathcal{H}$, the gradients are approximated by

$$\begin{cases} \nabla_{\beta_r}\mathcal{L} & \approx \frac{\mathcal{L}(\beta_r+\delta_{\beta_r})-\mathcal{L}(\beta_r-\delta_{\beta_r})}{2\delta_{\beta_r}}, r \in \{1,2\} \\ \nabla_{\alpha}\mathcal{L} & \approx \frac{\mathcal{L}(\alpha+\delta_\alpha)-\mathcal{L}(\alpha-\delta_\alpha)}{2\delta_\alpha} \end{cases}. \tag{10}$$

Here, each $\delta$ denotes the perturbation magnitude used in the finite-difference approximation and is independently adjusted for each parameter to support precise gradient estimation. When multiple sensitive words appear in a single prompt, a unified loss is computed over the entire $p_{\text{adv}}$ to guide joint optimization, and early stopping is triggered once $\mathcal{L} < \tau$ to prevent unnecessary computation. More details are shown in Appendix.

# Experiments

## Experimental Setup

**Datasets** To assess the effectiveness of MacPrompt in bypassing safety mechanisms, we construct two evaluation

| Concept | Method | Safety Filter (BPR) | | | | Concept Removal (ASR-1 / ASR-5) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | List | LG | BERT | SD | ESD | SLD.Ma | SLD.S | SLD.Me | SafeGen | FMN | DUO | EAP | ProG |
| Sex | DACA | 94 | **98** | 72 | 18 / 40 | 14 / 36 | 14 / 34 | 18 / 38 | 12 / 34 | 14 / 36 | 6 / 32 | 10 / 30 | 12 / 24 | 4 / 14 |
| | ART | 98 | 84 | 84 | 8 / 14 | 2 / 10 | 10 / 58 | 2 / 6 | 2 / 4 | 8 / 48 | 4 / 12 | 2 / 4 | 4 / 16 | 0 / 0 |
| | Position | 92 | 72 | **94** | 2 / 4 | 0 / 10 | 2 / 18 | 2 / 12 | 2 / 14 | 2 / 12 | 0 / 10 | 0 / 4 | 4 / 6 | 0 / 2 |
| | PGJ | 96 | **98** | 54 | 18 / 38 | 16 / 46 | 20 / 54 | 10 / 52 | 16 / 62 | 18 / 50 | 6 / 42 | 0 / 34 | 14 / 42 | 0 / 12 |
| | SurPro | **100** | 94 | 76 | 30 / 52 | 24 / 60 | 22 / 68 | 36 / 68 | 34 / 68 | 14 / 48 | 22 / 52 | 10 / 36 | 22 / 48 | 4 / 24 |
| | DiffZOO | 52 | 56 | 36 | 26 / 52 | 22 / 50 | 20 / 74 | 52 / 72 | 47 / 66 | 8 / 28 | 10 / 42 | 18 / 56 | 22 / 50 | 8 / 26 |
| | Ours | **100** | 82 | 70 | **56 / 96** | **62 / 74** | **52 / 96** | **54 / 92** | **54 / 88** | **38 / 76** | **52 / 84** | **32 / 60** | **24 / 62** | **24 / 34** |
| Violence | DACA | 78 | 80 | 94 | **60 / 85** | **54** / 72 | 5 / 18 | **22** / 34 | **24 / 36** | **55** / 80 | **66 / 80** | **42** / 64 | **26 / 62** | **40** / 52 |
| | ART | 80 | 80 | 90 | 40 / 50 | 0 / 30 | 0 / 0 | 10 / 10 | 0 / 10 | 30 / 60 | 10 / 50 | 20 / 40 | 20 / 20 | 0 / 10 |
| | Position | 76 | 58 | 74 | 30 / 62 | 28 / 54 | 2 / 12 | 6 / 22 | 12 / 26 | 26 / 66 | 36 / 64 | 22 / 56 | 14 / 38 | 24 / 40 |
| | PGJ | 92 | **92** | 94 | 22 / 44 | 28 / 42 | 0 / 2 | 2 / 10 | 10 / 18 | 18 / 44 | 36 / 42 | 16 / 40 | 8 / 26 | 18 / 30 |
| | SurPro | 80 | 66 | 94 | 20 / 40 | 6 / 54 | 0 / 0 | 6 / 20 | 6 / 28 | 14 / 46 | 34 / 54 | 6 / 46 | 20 / 40 | 26 / 40 |
| | DiffZOO | 48 | 56 | 54 | 36 / 66 | 24 / 66 | 0 / 3 | 6 / 12 | 6 / 12 | 16 / 40 | 30 / 70 | 16 / 46 | 10 / 30 | 12 / 28 |
| | Ours | **100** | 80 | **98** | 42 / 72 | 42 / **74** | **8 / 36** | **22 / 48** | 20 / **36** | 49 / **90** | 52 / 74 | 26 / **72** | 8 / 34 | 28 / **64** |

Table 1: Attack performance on NSFW concept generation. We report BPR across three safety filters as well as ASR-1 and ASR-5 across SD and nine concept removal defenses in the format "ASR-1 / ASR-5" (%). List, LG, and BERT indicate a blacklist-based keyword filter, LatentGuard filter, and a BERT-based NSFW text classifier, respectivaly. SLD.Ma, SLD.S, and SLD.Me correspond to variants of SLD: Max, Strong, and Medium, respectively. ProG stands for PromptGuard, while SurPro denotes SurrogatePrompt. "Origin" means attacking by original prompts. We mark the top-2 results by **bold** and underlying.

datasets using DeepSeek-V3 (Liu et al. 2024). The *NSFW-200*, designed to evaluate model robustness against harmful prompts, was constructed with reference to the I2P dataset (Schramowski et al. 2023), providing a taxonomy of inappropriate visual concepts commonly targeted by T2I safety systems. In parallel, the *Object-200* dataset comprises 200 prompts targeting four common object categories (i.e., dog, cat, car, and bird) selected from the MS COCO (Lin et al. 2014) dataset, enabling evaluation at the banned object level.

**Safety Mechanism** We execute black-box attacks against SD v2.1 with safety filters including a blacklist-based keyword filter, BERT-based classifier (Jieli 2023) and Latent Guard (Liu et al. 2025), as well as several SOTA concept removal models, such as ESD (Gandikota et al. 2023), SLD (including Max, Strong, and Medium) (Schramowski et al. 2023), FMN (Zhang et al. 2024a), SafeGen (Li et al. 2024b), DUO (Park et al. 2024), EAP (Bui et al. 2024), and PromptGuard (Yuan et al. 2025).

**Baselines** We evaluate MacPrompt against seven SOTA black-box attack baselines: DACA (Deng and Chen 2023), DiffZOO (Dang et al. 2024), ART (Li et al. 2024a), Position (Yang et al. 2024d), MMP-Attack (Yang et al. 2024a), PGJ (Huang et al. 2025), and SurrogatePrompt (Ba et al. 2024).

**Evaluation Metrics** We adopt four metrics to comprehensively evaluate the effectiveness of MacPrompt: (1) *Attack Success Rate (ASR-N):* An attack is considered successful only if the adversarial prompt produces at least one harmful image within N generation attempts. To automatically verify the presence of harmful content, we employ multiple content-specific detectors: a CLIP-based-NSFW-Detector (LAION-AI 2023) for pornographic content, the Q16 (Qu et al. 2023; Schramowski, Tauchmann, and Kersting 2022) classifier for violent imagery, and a object classifier trained on Animals-10 (Alessio 2020) and Stanford cars (Krause et al. 2013) datasets for identifying banned objects. (2) *Bypass Rate (BPR):* It quantifies the proportion of adversarial prompts that evade safety filters, regardless of the semantic content in the generated image. (3) *CLIPScore:* We compute the cosine similarity between CLIP (Radford et al. 2021) embeddings of the input texts or images to assess the similarity. (4) *BLIPScore:* We apply BLIP (Li et al. 2022), a better vision-language model, to evaluate the semantic consistency between the prompt and the generated images.

All experiments are implemented in PyTorch and conducted on a single NVIDIA RTX 4090 GPU. During optimization, the learning rate is set to 0.1, and the number of iterations is fixed at 100, while the perturbation magnitude is initialized with $\delta_0 = 0.25$.

## Main Results

**Evaluation on NSFW Concept Generation** Table 1 demonstrates that our method significantly outperforms baseline approaches in attacking NSFW concept generation, especially for the "Sex" category. It achieves the high BPR scores across all safety filters, notably reaching 100% on the blacklist filter, and substantially surpasses others in ASR across various concept removal defenses. Our method also maintains strong performance on the "Violence" concept, achieving top BPR and competitive ASR results. These findings highlight the superior effectiveness and robustness of our approach in circumventing all existing safety mechanisms. We further evaluated practical applicability to commercial systems, observing ASR of 65% on DALL·E 3 and 96% on Doubao. The upper half of Fig. 4 visually demonstrates the effectiveness of our attack within the NSFW domain, highlighting the capability of our method to bypass

multiple SOTA concept removal defenses specifically designed to forget NSFW content. Despite these defenses, our adversarial prompts successfully generate the target NSFW images, revealing vulnerabilities in current protection mechanisms. Importantly, these adversarial prompts exhibit strong transferability, enabling a single crafted prompt to compromise multiple NSFW filters simultaneously.

| Object | Method | SD | ESD | FMN | EAP |
|---|---|---|---|---|---|
| Dog | MMP-Attack | 66 / 90 | 78 / 88 | 60 / 90 | 52 / 94 |
| | Ours | 96 / 100 | 64 / 88 | 78 / 98 | 46 / 88 |
| Cat | MMP-Attack | 72 / 92 | 20 / 56 | 76 / 86 | 60 / 86 |
| | Ours | 86 / 98 | 32 / 74 | 74 / 98 | 44 / 80 |
| Bird | MMP-Attack | 54 / 88 | 44 / 88 | 44 / 66 | 66 / 82 |
| | Ours | 70 / 94 | 44 / 78 | 64 / 84 | 58 / 80 |
| Car | MMP-Attack | 76 / 84 | 52 / 86 | 70 / 88 | 62 / 92 |
| | Ours | 92 / 100 | 50 / 94 | 86 / 98 | 60 / 96 |

Table 2: Attack performance on banned objects generation.

**Evaluation on Banned Object Generation** To further assess MacPrompt's ability to bypass object-level safety constraints, we evaluate its performance on prompts containing banned object categories such as dog, cat, car, and bird, selected from MS COCO (Lin et al. 2014). As shown in Table 2, despite being trained solely against the SD safety filter, MacPrompt exhibits strong transferability to other concept removal mechanisms. Its adversarial prompts achieve attack success rates that are comparable to, and in some cases surpass, those of MMP-Attack (Yang et al. 2024a), a state-of-the-art approach specifically tailored for object-level attacks. These results underscore the versatility of MacPrompt, demonstrating its capability to extend beyond NSFW concepts and effectively bypass object-level safety constraints without any additional adaptation. Furthermore, as shown in the lower part of Fig. 4, the visual results indicate that a single adversarial prompt consistently triggers successful attacks across all evaluated models, underscoring the generalizability and efficacy of our method.

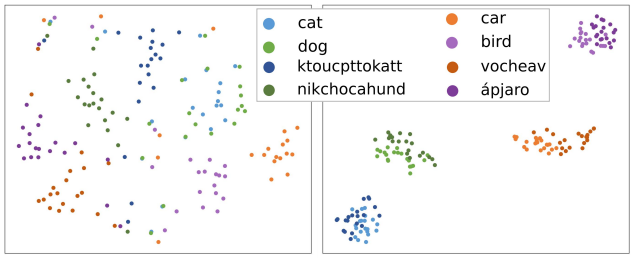


Figure 5: Visualization of semantic embeddings for banned objects and their macaronic substitutes, along with corresponding generated images.

**Evaluation on Semantic Consistency** We next assess the semantic consistency of adversarial examples from both textual and visual perspectives. Rather than focusing solely on attack success, this analysis aims to reveal how much semantic deviation is introduced during the attack process. As shown in Table 3, original and adversarial prompts exhibit lower textual similarity than safe and adversarial prompts, while their generated images demonstrate high semantic consistency. Notably, our method achieves superior BLIPScore (average 0.6953) compared to MMP-Attack (0.414), indicating significantly better alignment between generated images and the original semantics. Additionally, we visualize the embeddings of selected prompts and their corresponding generated images in Figure 5, focusing on the banned objects cat, dog, car, and bird, along with their macaronic substitutes: "ktoucpttokatt", "nikchocahund", "vocheav", and "ápjaro". In the textual embedding space, the original and substitute prompts are clearly separated, indicating their lexical divergence. However, in the image embedding space, the generated images from both prompt types form tight clusters, demonstrating that our method can bypass text-based filters while still generating semantically consistent images, effectively triggering the target concepts despite lexical obfuscation.

| Content | CLIPScore | | | BLIPScore |
|---|---|---|---|---|
| | $p_{\text{ori}} \leftrightarrow p_{\text{sadv}}$ | $p_{\text{safe}} \leftrightarrow p_{\text{adv}}$ | $I_{\text{ori}} \leftrightarrow I_{\text{adv}}$ | $p_{\text{ori}} \leftrightarrow I_{\text{adv}}$ |
| Sex | 0.8768 | 0.8842 | 0.7893 | 0.5602 |
| Violence | 0.8618 | 0.8770 | 0.8012 | 0.5893 |
| Dog | 0.9223 | 0.9514 | 0.8597 | 0.9572 |
| Cat | 0.9004 | 0.9315 | 0.8074 | 0.9600 |
| Bird | 0.8951 | 0.8957 | 0.7447 | 0.6007 |
| Car | 0.9348 | 0.9422 | 0.7335 | 0.5047 |

Table 3: Semantic similarity evaluation between original and adversarial prompts/images. Specifically, $p_{\text{ori}}$ and $p_{\text{adv}}$ denote the original and adversarial prompts, $I_{\text{ori}}$ and $I_{\text{adv}}$ are the corresponding generated images.

## Conclusion

In this paper, we proposed MacPrompt, a black-box attack against T2I models that uncovers previously overlooked vulnerabilities stemming from cross-lingual prompt manipulation. Our method introduces macaronic substitutes, which are constructed by recombining character-level substrings from translation-equivalent words across multiple languages. These substitutes retain the visual semantics of harmful concepts while obfuscating their textual representation, effectively bypassing both prompt-level safety filters and model-level concept removal defenses. Moreover, MacPrompt requires no access to model internals, gradients, or tokenizers, making it highly practical for real-world adversarial testing. Extensive experiments demonstrate that MacPrompt successfully evades SOTA safety mechanisms across diverse T2I systems, highlighting a critical gap in current defense strategies. We believe that MacPrompt offers a novel and scalable approach to evaluating the multilingual robustness of T2I safety mechanisms, providing valuable insights for future research on secure and responsible generative models.

## Ethical Statement

Our primary goal is to present MacPrompt, a cross-lingual black-box method designed to evaluate and expose vulnerabilities in T2I safety mechanisms. We acknowledge that the generated adversarial prompts may induce inappropriate content from T2I models. To mitigate potential misuse, all experiments were conducted in a controlled environment, and no harmful content will be publicly released. We disclose our findings responsibly to promote the development of more robust safety mechanisms for generative models. We firmly believe that the societal benefits of revealing these safety flaws outweigh the limited risks associated with demonstrating them.

## Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (NSFC) under Grants No.62372334, No. 62202340, and No. 62576255, the Fundamental Research Funds for the Central Universities under No. 2042025kf0054, the Natural Science Foundation of Hubei Province under No. 2025AFB455.

## References

Alessio, C. 2020. Animals-10 Dataset. https://www.kaggle.com/datasets/alessiocorrado99/animals10. Accessed: 2024-09.

Ba, Z.; Zhong, J.; Lei, J.; Cheng, P.; Wang, Q.; Qin, Z.; Wang, Z.; and Ren, K. 2024. Surrogateprompt: Bypassing the safety filter of text-to-image models via substitution. In *Proc. 2024 ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 1166–1180.

Bui, A.; Vuong, L.; Doan, K.; Le, T.; Montague, P.; Abraham, T.; and Phung, D. 2024. Erasing Undesirable Concepts in Diffusion Models with Adversarial Preservation. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 37, 133112–133146.

Caramiaux, B.; Crawford, K.; Liao, Q. V.; Ramos, G.; and Williams, J. 2025. Generative AI and Creative Work: Narratives, Values, and Impacts. arXiv:2502.03940.

Chin, Z.-Y.; Jiang, C.-M.; Huang, C.-C.; Chen, P.-Y.; and Chiu, W.-C. 2024. Prompting4Debugging: Red-Teaming Text-to-Image Diffusion Models by Finding Problematic Prompts. In *Proc. Int. Conf. Mach. Learn. (ICML)*.

Dang, P.; Hu, X.; Li, D.; Zhang, R.; Guo, Q.; and Xu, K. 2024. Diffzoo: A purely query-based black-box attack for red-teaming text-to-image generative model via zeroth order optimization. *arXiv preprint arXiv:2408.11071*.

Deng, Y.; and Chen, H. 2023. Divide-and-conquer attack: Harnessing the power of LLM to bypass the censorship of text-to-image generation model. *CoRR*.

Gandikota, R.; Materzyńska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing Concepts from Diffusion Models. In *Proc. 2023 IEEE Int. Conf. Comput. Vis.*

Gao, S.; Jia, X.; Huang, Y.; Duan, R.; Gu, J.; Liu, Y.; and Guo, Q. 2024. Rt-Attack: Jailbreaking Text-to-Image Models via Random Token. *arXiv preprint arXiv:2408.13896*.

George, R. R. 2023. NSFW Words List. https://github.com/rrgeorge-pdcontributions/NSFW-WordsList/blob/master/nsfw%20list.txt. Accessed: 2024-08.

Gupta, A. 2022. Unstable Diffusion: Ethical challenges and some ways forward. Montreal AI Ethics Institute. Accessed: 2022-04.

Huang, Y.; Liang, L.; Li, T.; Jia, X.; Wang, R.; Miao, W.; Pu, G.; and Liu, Y. 2025. Perception-guided jailbreak against text-to-image models. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, volume 39, 26238–26247.

Hwang, S.; Wang, B.; and Gu, A. 2025. Dynamic Chunking for End-to-End Hierarchical Sequence Modeling. arXiv:2507.07955.

Jieli, M. 2023. NSFW Text Classifier. https://huggingface.co/michellejieli/NSFW_text_classifier/discussions?not-forall-audiences=true. Accessed: 2024-08.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3D Object Representations for Fine-Grained Categorization. In *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, 554–561.

LAION-AI. 2023. NSFW CLIP-based Image Classifier on GitHub. https://github.com/LAION-AI/CLIP-based-NSFW-Detector. Accessed: 2024-06.

Li, G.; Chen, K.; Zhang, S.; Zhang, J.; and Zhang, T. 2024a. ART: Automatic Red-teaming for Text-to-Image Models to Protect Benign Users. In *Proc. 38th Annu. Conf. Neural Inf. Process. Syst. (NeurIPS)*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proc. Int. Conf. Mach. Learn. (ICML)*, ICML '22.

Li, X.; Yang, Y.; Deng, J.; Yan, C.; Chen, Y.; Ji, X.; and Xu, W. 2024b. SafeGen: Mitigating Sexually Explicit Content Generation in Text-to-Image Models. In *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 740–755. Springer.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, R.; Khakzar, A.; Gu, J.; Chen, Q.; Torr, P.; and Pizzati, F. 2025. Latent Guard: A Safety Framework for Text-to-Image Generation. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 93–109. Cham: Springer Nature Switzerland. ISBN 978-3-031-73347-5.

Ma, J.; Cao, A.; Xiao, Z.; Li, Y.; Zhang, J.; Ye, C.; and Zhao, J. 2024. Jailbreaking Prompt Attack: A Controllable Adversarial Attack Against Diffusion Models. *arXiv preprint arXiv:2404.02928*.

MidJourney. 2023. MidJourney: AI-Generated Art Platform. Accessed: 2024-05.

Naik, R.; and Nushi, B. 2023. Social Biases Through the Text-to-Image Generation Lens. In *Proc. AAAI/ACM Conf. AI Ethics Soc. (AIES)*, 786–808.

Park, Y.-H.; Yun, S.; Kim, J.-H.; Kim, J.; Jang, G.; Jeong, Y.; Jo, J.; and Lee, G. 2024. Direct Unlearning Optimization for Robust and Safe Text-to-Image Models. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 37, 80244–80267.

Pernias, P.; Rampas, D.; Richter, M. L.; Pal, C. J.; and Aubreville, M. 2023. Wuerstchen: An Efficient Architecture for Large-Scale Text-to-Image Diffusion Models. arXiv:2306.00637.

Qu, Y.; Shen, X.; He, X.; Backes, M.; Zannettou, S.; and Zhang, Y. 2023. Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models. In *Proc. ACM Conf. Comput. Commun. Secur. (CCS)*, CCS '23, 3403–3417.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *Proc. Int. Conf. Mach. Learn. (ICML)*, ICML '21, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125.

Raphaël, M. 2022. Adversarial Attacks on Image Generation With Made-Up Words. arXiv:2208.04135.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv:2205.11487.

Schramowski, P.; Brack, M.; Deiseroth, B.; and Kersting, K. 2023. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*.

Schramowski, P.; Tauchmann, C.; and Kersting, K. 2022. Can Machines Help Us Answering Question 16 in Datasheets, and in Turn Reflecting on Inappropriate Content? In *Proc. ACM Conf. Fairness Account. Transp. (FAccT)*, FAccT '22, 1350–1361.

Tamang, S.; and Bora, D. J. 2024. Evaluating Tokenizer Performance of Large Language Models Across Official Indian Languages. arXiv:2411.12240.

Tsai, Y.-L.; Hsu, C.-Y.; Xie, C.; Lin, C.-H.; Chen, J. Y.; Li, B.; Chen, P.-Y.; Yu, C.-M.; and Huang, C.-Y. 2024. Ring-A-Bell! How Reliable are Concept Removal Methods For Diffusion Models? In *Proc. Int. Conf. Learn. Represent. (ICLR)*.

Unstable Diffusion. 2025. Unstable Diffusion. https://www.unstability.ai/. Accessed: 2025-03.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2017. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. arXiv:1711.10485.

Xu, W.; Chen, K.; Qiu, J.; Zhang, Y.; Wang, R.; Mao, J.; Zhang, T.; and Wang, L. 2025. Automated Red Teaming for Text-to-Image Models through Feedback-Guided Prompt Iteration with Vision–Language Models. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 18575–18584.

Yang, D.; Bai, Y.; Jia, X.; Liu, Y.; Cao, X.; and Yu, W. 2024a. On the Multi-modal Vulnerability of Diffusion Models. In *Proc. Trustworthy Multi-modal Found. Models AI Agents (TiFA)*.

Yang, Y.; Gao, R.; Wang, X.; Ho, T.-Y.; Xu, N.; and Xu, Q. 2024b. MMA-Diffusion: MultiModal Attack on Diffusion Models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 7737–7746.

Yang, Y.; Hui, B.; Yuan, H.; Gong, N.; and Cao, Y. 2024c. SneakyPrompt: Jailbreaking Text-to-Image Generative Models. In *Proc. IEEE Symp. Secur. Privacy (SP)*, 897–912.

Yang, Y.; Lin, Y.; Liu, H.; Shao, W.; Chen, R.; Shang, H.; Wang, Y.; Qiao, Y.; Zhang, K.; and Luo, P. 2024d. Position: Towards Implicit Prompt For Text-To-Image Models. In *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, volume 235, 56235–56250.

Yuan, L.; Jia, X.; Huang, Y.; Dong, W.; and Liu, Y. 2025. PromptGuard: Soft Prompt-Guided Unsafe Content Moderation for Text-to-Image Models. arXiv:2501.03544.

Zeng, W.; Kurniawan, D.; Mullins, R.; Liu, Y.; Saha, T.; Ike-Njoku, D.; Gu, J.; Song, Y.; Xu, C.; Zhou, J.; Joshi, A.; Dheep, S.; Malek, M.; Palangi, H.; Baek, J.; Pereira, R.; and Narasimhan, K. 2025. ShieldGemma 2: Robust and Tractable Image Content Moderation. arXiv:2504.01081.

Zhang, G.; Wang, K.; Xu, X.; Wang, Z.; and Shi, H. 2024a. Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 1755–1764.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. 2017. StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. arXiv:1612.03242.

Zhang, Y.; Jia, J.; Chen, X.; Chen, A.; Zhang, Y.; Liu, J.; Ding, K.; and Liu, S. 2024b. To Generate or Not? Safety-Driven Unlearned Diffusion Models Are Still Easy to Generate Unsafe Images... For Now. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 385–403. Springer.

Zhao, X.; Chen, X.; and Gao, H. 2024. Antelope: Potent and Concealed Jailbreak Attack Strategy. *arXiv preprint arXiv:2412.08156*.

Zhuang, H.; Zhang, Y.; and Liu, S. 2023. A Pilot Study of Query-Free Adversarial Attack Against Stable Diffusion. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2385–2392.