STEERDIFF: STEERING TOWARDS SAFE TEXT-TO-IMAGE DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-image (T2I) diffusion models have drawn attention for their ability to generate high-quality images with precise text alignment. However, these models can also be misused to produce inappropriate content. Existing safety measures, which typically rely on text classifiers or ControlNet-like approaches, are often insufficient. Traditional text classifiers rely on large-scale labeled datasets and can be easily bypassed by rephrasing. As diffusion models continue to scale, fine-tuning these safeguards becomes increasingly challenging and lacks flexibility. Recent red-teaming attack researches further underscore the need for a new paradigm to prevent the generation of inappropriate content. In this paper, we introduce SteerDiff, a lightweight adaptor module designed to act as an intermediary between user input and the diffusion model, ensuring that generated images adhere to ethical and safety standards with little to no impact on usability. SteerDiff identifies and manipulates inappropriate concepts within the text embedding space to guide the model away from harmful outputs. We conduct extensive experiments across various concept unlearning tasks to evaluate the effectiveness of our approach. Furthermore, we benchmark SteerDiff against multiple red-teaming strategies to assess its robustness. Finally, we explore the potential of SteerDiff for concept forgetting tasks, demonstrating its versatility in text-conditioned image generation.

Warning: This paper contains potentially offensive text and images.

1 INTRODUCTION

Text-to-image (T2I) diffusion models have attracted attention for their out-of-the-box functionality and the superior quality of their generated images. Using models like Stable Diffusion (Rombach et al., 2022; 2021) or DALL-E (Ramesh et al., 2022), users can use simple natural language descriptions as input to generate high-quality images with precise text alignment. This capability is largely contributed by pre-trained language models (Hammoud et al., 2024; Jin et al., 2020; Dosovitskiy, 2020) that learn and reflect the underlying syntax and semantics, as well as by extensive multimodal training datasets that encompass a wide range of text-to-image aligned content. However, these training methods also introduce risks of generating inappropriate images to the models. Thus, preventing the generation of inappropriate images is both critical and urgent.

To this end, existing T2I diffusion models have integrated several safety strategies to prevent inappropriate content generation. Stable diffusion incorporates a Not Safe For Work (NSFW) post-generation filter and ad-hoc filter during training to avoid generating unsafe images (Schuhmann et al., 2022; Schramowski et al., 2023; Rando et al., 2022). However, existing text-based attacks (Li et al., 2018; Jin et al., 2020; Garg & Ramakrishnan, 2020; Maus et al., 2023) can mislead the classification mechanisms in the filter by rephrasing "*A photo of a billboard showing a naked man.*" into "*A photo of a billboard showing an LGBT man in an explicit position*". Additionally, users report that removing explicit images and other subjects from training data may have had a negative impact on the output quality, thus harming the utility of the models (Rombach et al., 2022).

The prevention of inappropriate content generation in AI models has primarily been approached through concept removal, which aims to control the generation process itself. Two primary strategies have been explored: (1) using ControlNet-like (Zhang et al., 2023a) structures to guide the diffusion process (Schramowski et al., 2023; Rando et al., 2022), and (2) identifying and pruning weights within the diffusion model (Gandikota et al., 2023). However, these methods face significant challenges as generative models grow in scale and complexity. The increasing size of models makes these approaches computationally expensive and impractical (Ramesh et al., 2022; Rombach et al., 2022; Saharia et al., 2022). Additionally, simply guiding the generative process introduces vulnerabilities, leaving models open to jailbreaking attempts (Zhang et al., 2023b; Qu et al., 2023). These limitations highlight the need for more efficient and robust solutions.

Although the aforementioned safety mechanisms have shown effectiveness according to their respective evaluation schemes, recent red-teaming studies demonstrate their potential flaws (Zhang et al., 2023b; Qu et al., 2023). Chin et al. (2023) shows that approximately half of the prompts, which were originally mitigated by existing safety mechanisms, can be manipulated by their Prompting4Debugging (P4D) to become problematic. Similarly, Unsafe Diffusion (Qu et al., 2023) found that 14.56% of generated images across four state-of-the-art T2I models and their four prompt datasets were unsafe, underscoring the vulnerability of these models to generating harmful content. Moreover, black-box jail-breaking approaches Jailbreak Prompt Attack (JPA) and SneakyPrompt (Ma et al., 2024; Yang et al., 2024) successfully attack both online services and offline T2I models with the current safety mechanisms.

In this work, we propose SteerDiff, a two-stage lightweight adaptor model for text-conditioned diffusion models that focuses on guiding text prompt embeddings rather than controlling the generative process. Our method constructs a semantic boundary that maximally distinguishes between safe and unsafe content. We then project potentially unsafe embeddings toward the safe region while preserving the original semantics and maintaining the diffusion model's generative capabilities. This approach offers three key advantages: efficiency, effectiveness, and versatility. By operating at the prompt embedding level, our method eliminates the need for computationally intensive model retraining while preserving the original semantics. Moreover, by preventing unsafe content at the earliest stage of generation, we can block the formation of unsafe latent representations more directly and reliably. Lastly, our lightweight approach can be easily trained and applied to various concept removal tasks.

We benchmark our prototype SteerDiff against state-of-the-art concept removal techniques, including Erased Stable Diffusion (ESD) (Gandikota et al., 2023) and Safe Latent Diffusion (SLD) (Schramowski et al., 2023; Rando et al., 2022), for removing inappropriate content. Experimental results demonstrate that SteerDiff significantly reduces inappropriate content generation while preserving image quality and semantic fidelity. Furthermore, we evaluate its robustness by defending against red-teaming methods such as P4D and SneakyPrompt, demonstrating the effectiveness of our approach in mitigating various forms of adversarial attacks on text-to-image generation systems.

2 RELATED WORK

2.1 TEXT-TO-IMAGE DIFFUSION MODELS

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Rombach et al., 2022) are a type of probabilistic generative model that learns a data distribution by gradually transforming a simple distribution into a complex target distribution. Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) model the data generation process as a sequence of denoising steps. Given an image, DDPMs progressively add noise sampled from Gaussian distribution to generate an intermediate noisy image $\mathbf{x}_t$ at each time step t called forward diffusion steps. The noisy image $\mathbf{x}_t$ can be expressed in closed form as a function of the original image $\mathbf{x}_0$, the time step t, and noise ϵ sampled from a Gaussian distribution $\mathcal{N}(0, \mathbf{I})$. The model is then trained using the reverse process, where the objective is to learn a model parameterized by θ that predicts ϵ. This objective is defined as:

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon} \left[\|\epsilon - \epsilon\theta(\mathbf{x}_t, t)\|_2^2 \right].$$

2.2 TOWARDS SAFE IMAGE GENERATION

Current approaches to prevent undesirable images from generation generally follow two main strategies. The first involves removing undesired images from the training set, such as excluding all human figures (Nichol et al., 2021) or selectively omitting specific undesirable image categories in the dataset (Schuhmann et al., 2022; DAL, 2022; Rombach & Esser, 2022). However, this approach is costly as it necessitates retraining the model, and removing certain data classes often degrades the overall quality of the generated images (O'Connor, 2022). The second strategy is post-hoc, includes using blacklists for blocking unsafe concepts (DAL, 2022; mid, 2024; Leo, 2023; Markov et al., 2023), editing images after generation, or fine-tuning diffusion models to guide the inference process (Schramowski et al., 2023). Although blacklists are straightforward to implement, they can be easily bypassed. Image editing and diffusion process manipulation methods are effective but still require image synthesis, adding computational overhead. Other approaches (Park et al., 2024; Zhang et al., 2024) use inpainting to mask unsafe content or attempt to unlearn inappropriate concepts either within the diffusion model (Zheng & Yeh, 2023) or the textual encoder (Poppi et al., 2023). These methods involve expensive fine-tuning, while SteerDiff offers an intermediary solution that operates between the input prompt and the diffusion model without additional training. Similarly, Latent Guard (Liu et al., 2024) identifies inappropriate concepts prior to diffusion by learning a latent space on top of the T2I model's text encoder. However, while Latent Guard is capable of detecting inappropriate concepts, it lacks the ability to remove them. In this work, we compare the naive blacklist method commonly used in commercial platforms, Erased Stable Diffusion (Gandikota et al., 2023), the state-of-the-art model-editing approach, and Safe Latent Diffusion (Schramowski et al., 2023), a guidance-based model. Our focus is to introduce a new approach that identifies and manipulates inappropriate concepts in the text embedding space to ensure safer image generation.

3 METHODOLOGY

Determining what constitutes inappropriate imagery is highly subjective and varies based on context, setting, cultural and social predispositions, and individual factors. Additionally, Qu et al. (2023) observes that unsafe content can be shared through memes. In the context of T2I models, these new concepts are embedded within the words of a prompt. Consequently, adding new concepts to block T2I generation without retraining the diffusion model is impractical. To overcome these limitations, we formalize the problem by identifying unsafe concepts in text embedding space and then projecting potentially unsafe embeddings toward safe regions. This projection preserves the original semantics while maintaining the diffusion model's generative capabilities. Our approach allows us to define blacklisted concepts at test time, enabling greater flexibility.

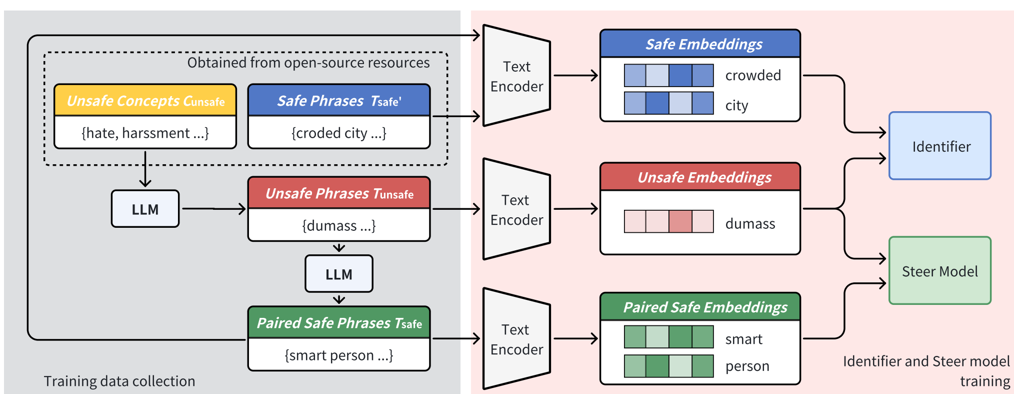


Figure 1: Overview of data collection and training process for SteerDiff: We begin by defining a set of unsafe concepts, C_{unsafe} (yellow block). Next, we use a LLM to generate related unsafe (red block on the left) and safe (green block on the left) phrases based on each concept $c \in C_{\text{unsafe}}$. These phrases are then encoded using a pre-trained text encoder to extract embedding features, which are used to train both the identifier and steering model.

In subsection 3.1, we find a set of inappropriate concepts based on established work (Schramowski et al., 2023) and describe the process of collecting and generating training data using a large language model as illustrated in Figure 1. As shown in the overview in Figure 2a, we next explain how inappropriate concepts are identified within text embeddings (subsection 3.2), followed by our steering approach to mitigate the generation of inappropriate content (subsection 3.3). Finally, we outline how the framework operates during inference to identify and steer text prompts associated with unsafe concepts (subsection 3.4).

3.1 Training Data Collection

Directly classifying safe/unsafe prompts requires large-scale annotated datasets (Markov et al., 2023). Following Safe Latent Diffusion, we base our definition of inappropriate content on the work of Gebru et al.: "[data that] *if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety*" (Gebru et al., 2021; Schramowski et al., 2023). Specifically, we consider an image as inappropriate if it contains any concept $c \in C_{\text{unsafe}}$, where

$$\begin{aligned} C_{\text{unsafe}} = \{ & \text{hate, harassment, violence, self-harm, sexual content,} \\ & \text{shocking images, illegal activity} \}. \end{aligned} \tag{1}$$

It is important to note that the definition of inappropriateness is not restricted to these seven categories, as the boundaries of appropriateness vary across cultures and evolve over time. In this study, however, we limit our scope to images that display clear and tangible acts of inappropriate behavior. Although we limit our scope to the current C_{unsafe}, our framework can be extended beyond these categories, as discussed in section 5.

3.1.1 Identifier Dataset

As mentioned previously, we aim to detect inappropriate concepts in prompts to avoid inappropriate content generation. The first step in our pipeline is to construct the dataset with unsafe terms capturing the concepts from C_{unsafe}. To achieve this, we start by collecting open-sourced blacklisted phrases. Midjourney (mid, 2024) employs a blacklist of words and phrases that includes phrases associated with violence, hate speech, explicit sexual content, illegal activities, and other categories considered inappropriate by the platform. Additionally, Latent Guard (Liu et al., 2024) introduces the CoPro dataset, which includes safe and unsafe prompts centered around blacklisted concepts. We build our blacklist based on Midjourney and CoPro dataset.

Although we can collect numerous NSFW terms from open-source datasets, these datasets may be imbalanced and could lack some categories defined in C_{unsafe}. For example, the number of shocking images or illegal activity is lower compared to images in other categories. To tackle this, we leverage an LLM to generate related terms t_c centered around one sampled concept c in the blacklist of C_{unsafe} as illustrated in Figure 1, similarly to Hammoud et al. (2024) and Liu et al. (2024). This allows us to create a set $T_{\text{unsafe}} = \{t_c | c \in C_{\text{unsafe}}\}$. The unsafe terms in T_{unsafe} simulate typical unsafe terms that a malicious user may input. In addition to the unsafe terms, we combine randomly selected 500 prompts (Blue block on the left) from the COCO 30k dataset (Lin et al., 2014) and the following mentioned paired safe phrases dataset (Green block on the left) to serve as our safe prompt dataset. All phrases within these prompts are considered safe.

3.1.2 Steer Model Dataset

In our subsequent steering transformation training procedure, we synthesize additional safe terms to steer unsafe embeddings toward safe ones. The core idea is to associate each unsafe term $t_c \in T_{\text{unsafe}}$ with a corresponding safe term ${t_c}'$ of similar meanings, allowing us to convert unsafe concepts into safe alternatives while preserving the original semantic intent of the prompt. For example, consider the term "killed" in the prompt "A man got killed." which represents a violent visual scene linked to the concept of "violence". We use an LLM to eliminate unsafe concepts in the input term t_c. In this case, a possible safe term ${t_c}'$ would be "saved", transforming the meaning of the prompt to something like "A man got saved". By processing all elements in T_{unsafe}, we generate paired safe phrases dataset T_{safe}, comprising M safe terms ${t_c}' \in T_{\text{safe}}$. For more details of the synthesized dataset, please refer to subsection A.2.

3.2 Inappropriate Concepts Identifier

To ensure the generation of safe and appropriate content, we detect and mitigate undesirable concepts within user prompts before they are processed by the text-to-image diffusion model. The core idea is that the prompt embeddings can be leveraged to represent individual terms or phrases, enabling precise identification of inappropriate content.

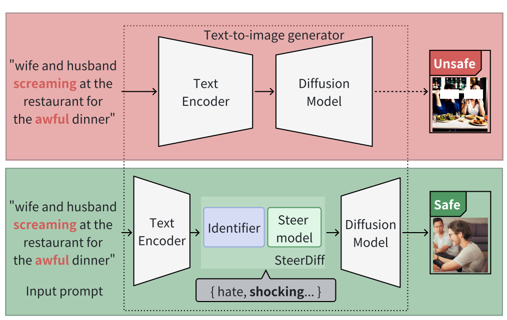
This task can be framed as classifying phrases as either appropriate or inappropriate. We employ a lightweight multi-layer perceptron model (MLP) to classify the embeddings of individual terms. Intuitively, we want to identify the phrase that precisely includes the inappropriate concept. For example, in the prompt "a man got shot." only the term "shot" relates to the concept of violence. Additionally, the phrase "screw you" is related to the inappropriate concept "hate", but neither "screw" nor "you" is related to the concept "hate". To address such cases, we utilize a sliding window technique to identify single terms and phrases that may collectively express inappropriate concepts.

As shown in Figure 1, the safe and unsafe phrases are first embedded by a text encoder and then used to train the identifier. We expect the identifier to detect the unsafe concepts defined in Equation 1. Let T_{unsafe} represent the set of unsafe phrases. Specifically, the safe phrases set is composed of two subsets: T_{safe} (synthetic set) and $T_{\text{safe}'}$ (collected from open-source resources). The MLP outputs the probability of a given phrase belonging to unsafe or safe. Denote the input phrase embeddings as E_i, the MLP output as $\hat{y}_i$, and the true label as $y_i \in \{0, 1\}$, where $y_i = 1$ indicates an unsafe phrase and $y_i = 0$ indicates a safe phrase. The classification loss $\mathcal{L}$ is defined as follows:

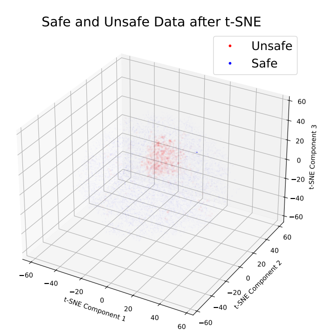
$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \left(y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \right) \tag{2}$$

where N is the total number of phrases in the training set.

Since embeddings with similar semantics have closer distances in the embedding space (Mikolov et al., 2013; Radford et al., 2021), we expect the unsafe embeddings to be aggregated. As demonstrated in Figure 2b, we observe that SteerDiff successfully learns to distinguish between safe and unsafe phrases, with the two categories being well-separated after applying t-SNE dimensional reduction.



(a) A T2I generator without a safety mechanism (top) can generate inappropriate content. We propose SteerDiff (bottom), a safety method designed to identify and steer inappropriate concepts toward producing safe images.



(b) SteerDiff learns to differentiate between safe and unsafe phrases, with the two categories becoming clearly distinct after applying t-SNE for dimensionality reduction.

Figure 2: Overview of SteerDiff (left). SteerDiff learns to distinguish safe and unsafe phrases (right).

3.3 Steering Toward Safe Content

To steer the inappropriate prompts toward generating safe images, we propose to learn a linear transformation to the embedding of identified unsafe prompts. This transformation shifts unsafe concepts toward safe ones while preserving original semantic meaning. The intuition behind this approach is that linear transformations of word embeddings can effectively steer the generation style of language models (Han et al., 2023). Since SteerDiff operates in word embedding space, applying these transformations enables us to adjust unsafe embeddings and steer the diffusion model toward producing safe outputs. Specifically, let E denote the embedding of a word, we define the transformation as follows:

$$E_{\text{steered}} = \epsilon W E_{\text{unsafe}} + (1 - \epsilon) E_{\text{unsafe}} \tag{3}$$

In this equation, E_{steered} represents the adjusted embedding after steering, E_{unsafe} refers to the original embedding of the unsafe phrases in T_{unsafe}, ϵ is a steering parameter that controls the intensity of the transformation, and W is a linear transformation matrix learned during training.

The key concept is to map the embedding of inappropriate content to safe content through a linear transformation. By applying Equation 3, we shift the embedding within the semantic space, guiding it toward regions associated with safe imagery. The steering parameter ϵ offers fine-grained control over the degree of transformation, allowing flexibility in balancing semantic preservation and safety. To learn the transformation matrix W, we employ a supervised learning method using a paired dataset of unsafe phrases and their corresponding safe phrases, as described in subsection 3.1. The training process minimizes the following loss function:

$$\mathcal{L} = |E_{\text{safe}} - W \cdot E_{\text{unsafe}}|^2 \tag{4}$$

where E_{safe} represents the embedding of the safe phrases in T_{safe}.

3.4 Inference

Once SteerDiff is trained, it can be seamlessly integrated into diffusion models without additional fine-tuning. In practical applications, SteerDiff can effectively detect the presence of blacklisted concepts and steer inappropriate prompts toward generating safe images.

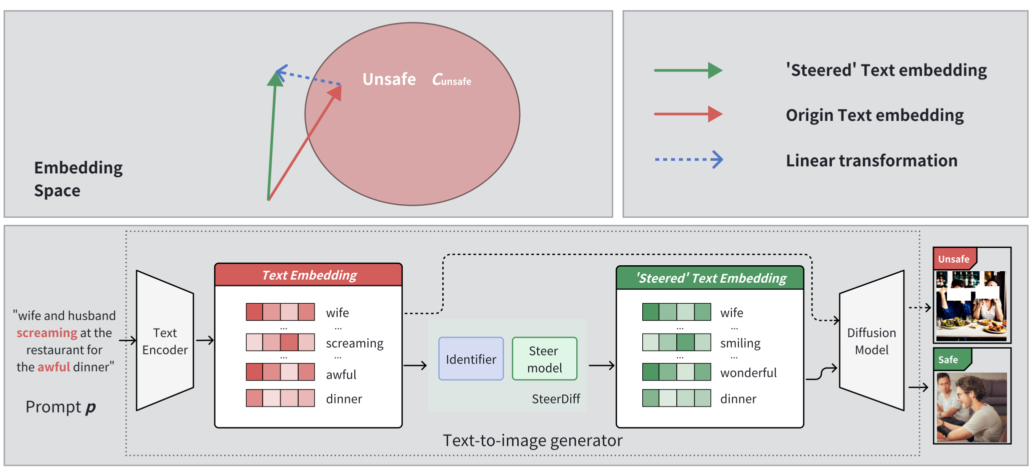


Figure 3: Illustration of the SteerDiff process: A problematic embedding (red arrow) is steered (blue arrow) towards a safe embedding (green arrow) to mitigate the generation of inappropriate content.

Consider a T2I model equipped with a text encoder. As illustrated in Figure 3, a user provides a prompt p, which can be either safe or unsafe. The input prompt is first embedded by a text encoder. We define a concept blacklist C_{unsafe} that contains potentially inappropriate concepts. Then the identifier detects whether prompt p contains any inappropriate concepts $c' \in C_{\text{unsafe}}$. If such a concept is detected, a linear transformation is applied to the prompt's embedding to steer it toward safer content. Specifically, the transformation described in Equation 3 is applied to the embedding of p, altering the latent representation to avoid generating inappropriate content. This transformation adjusts the embedding space without compromising the semantics of the original prompt. Once the transformation is applied, the modified embedding is passed through the diffusion model to generate a safe image. The transformed embedding ensures that the resulting image adheres to SFW guidelines while preserving the user's original intent.

This approach allows SteerDiff serving as an intermediary between the user's input and the diffusion model, ensuring that the generated images align with ethical and safety standards. Additionally, this steering process is efficient, eliminating the need for model retraining or introducing significant delays in the image generation pipeline. The transformation can also be dynamically adapted as blacklists or safety criteria evolve, making the system robust and scalable for real-world applications.

4 Experiment

4.1 Experimental Setup and Metrics

To evaluate the reduction of inappropriate content, we used Stable Diffusion (SD) v1.4 as the base model to generate three images for each prompt. We selected a general concept set C_{unsafe}, applicable across all categories of Inappropriate Image Prompts (I2P) dataset (Schramowski et al., 2023), as outlined in Equation 1. We evaluated inappropriate image content by combining two classifiers. First, Q16 classifier (Schramowski et al., 2022), which is also used for annotating LAION-5B dataset (Schuhmann et al., 2022), was utilized to detect a broad range of inappropriate content. Second, we applied NudeNet (Praneeth, 2019) to identify sexually explicit content. For nude content, we classified exposed buttocks, exposed breasts, exposed genitalia, and exposed anus content as inappropriate, while allowing otherwise provocative images (Schramowski et al., 2023; Wu et al., 2024; Ma et al., 2024). Unless otherwise specified, an image was classified as inappropriate if at least one of the classifiers flagged it as containing such content.

For the following **RQ1**, we evaluated SteerDiff on the I2P dataset and compared the probability of generating inappropriate images with state-of-the-art defense model Safe Latent Diffusion (SLD) under its MAX and STRONG configurations. In addition, we compared the probability of generating nude images with Erased Stable Diffusion (ESD) (Gandikota et al., 2023). For **RQ2**, we selected two state-of-the-art red teaming frameworks: Prompting4Debugging (P4D), representing a white-box attack, and SneakyPrompt, representing a black-box attack. Following standard instructions from P4D and Sneakyprompt, we conducted our evaluation on the I2P dataset using P4D testbed and NSFW_200 dataset (Yang et al., 2024) with Sneakyprompt, utilizing standard configurations provided by developers. For the diffusion process, we used the default configuration as outlined by stable-diffusion (Rombach et al., 2021). This approach ensures our experiments align with recommended practices to maintain consistency across all tests. For **RQ3**, we assessed image fidelity and text alignment across all models using prompts from the COCO 30K dataset.

4.2 RQ1: How effective is SteerDiff in mitigating generation of inappropriate content?

To investigate the ability of SteerDiff in identifying and steering inappropriate concepts, we started by demonstrating its effectiveness in reducing the generation of explicit content. We compared SteerDiff against ESD, SLD STRONG, and SLD MAX. Next, we expanded the scope of inappropriate concepts to C_{unsafe} to investigate whether SteerDiff could effectively identify and steer prompts containing a wider range of inappropriate content toward generating safe images. In this evaluation, we evaluated SteerDiff with Stable Diffusion (SD) v1.4, naive blacklist, SLD STRONG, and SLD MAX approach. To minimize randomness and ensure more reliable results, all evaluated methods generated three images per prompt, as generating only one image might coincidentally omit inappropriate content. Notably, we used inappropriate terms from SteerDiff's training set as the naive blacklist.

4.2.1 Explicit Content Removal

We evaluated the performance of SteerDiff and competitors on the I2P dataset, focusing on the sexual content category. In Figure 4, we compared the percentage change in nudity-classified samples with respect to SD v1.4. Our results show that, across all classes, our method demonstrates a more significant reduction in the generation of explicit content. Specifically, SD v1.4 generated 637 images with exposed body parts on test prompts, SteerDiff, ESD, SLD STRONG, and SLD MAX reduced this number to 104, 389, 208, and 134, respectively.

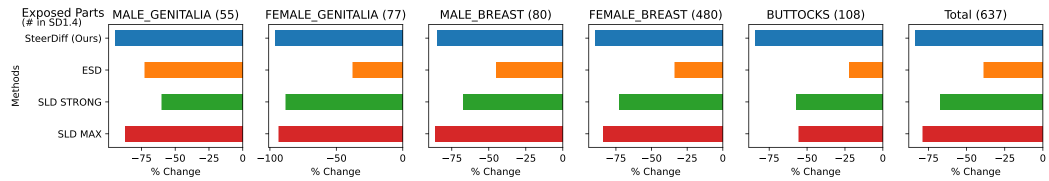


Figure 4: Our method effectively removes sexual content from SD v1.4 on the I2P dataset, outperforming defend methods ESD, SLD STRONG, and SLD MAX. Illustrating the percentage reduced in nudity-classified samples compared to the original SD v1.4 model.

Table 1: SteerDiff demonstrates the best performance in reducing the probability of generating inappropriate content (where lower values are better). The probabilities shown represent the likelihood of generating images classified as inappropriate by combining the Q16 and NudeNet classifiers across various I2P categories. The best performances are bolded, and the second-best performances are underscored.

Method	Inappropriate probability % ↓							
	hate	harassment	violence	self-harm	sexual	shocking	illegal activity	Overall
SD1.4	27.27	19.05	27.65	30.34	46.29	35.98	18.16	30.17
Blacklist	19.48	14.68	17.99	18.23	21.59	22.31	10.87	17.86
SteerDiff (Ours)	5.63	4.25	**2.91**	4.74	**2.36**	6.78	3.99	**4.51**
ESD	-	-	-	-	10.56	-	-	-
SLD STRONG	6.49	6.80	5.42	5.24	12.67	11.33	3.03	7.97
SLD MAX	**3.90**	**3.76**	5.29	**2.25**	8.38	**6.31**	**2.75**	5.15

4.2.2 Inappropriate Content Removal

We further investigated a more comprehensive inappropriate set C_{unsafe} defined in Equation 1. We began our evaluation by demonstrating the inappropriate generation of SD v1.4 without any safety measures, as well as a basic blacklist-based approach for prompt matching. Table 1 presents the probability of generating inappropriate content for each category. Varying from different categories, SD v1.4 generated inappropriate content with probabilities ranging from 18.16% to 46.29%. The naive blacklist approach slightly reduced probability, but inappropriate content was still generated in 17.86% of cases across all categories. While naive blacklists may mitigate some inappropriate image generation, it remains largely impractical as a comprehensive defense mechanism due to its inability to capture the diverse and evolving nature of unsafe content.

Next, we demonstrated the probability of generating inappropriate content using SteerDiff, SLD STRONG, and SLD MAX across different I2P concepts. As shown in Table 1[1], both SteerDiff and SLD MAX demonstrate the strongest performance, ranking first and second, respectively. Specifically, SteerDiff reduced the probability of generating inappropriate content by over 85%. In particular, SteerDiff outperformed its closest competitor, SLD MAX, in categories of violence, sexual content, and overall inappropriate content. As a result, only 5% of images generated by SteerDiff were still classified as inappropriate. However, it is worth noting that the Q16 and Nudenet classifiers tend to flag images as inappropriate even when problematic content has been significantly reduced. In summary, SteerDiff effectively mitigates the generation of inappropriate content in SD by identifying and modifying unsafe concepts within text embeddings.

4.3 RQ2: Can SteerDiff defend against red-teaming attacks?

While SteerDiff was the state-of-the-art model when evaluating the I2P dataset, the robustness against red-teaming attacks is also a necessary factor of successful defense. Therefore, we committed to further investigating its efficacy towards red-teaming attacks. To this end, we selected state-of-the-art white-box and black-box red-teaming frameworks Prompt4dubugging (P4D) and SneakyPrompt (Chin et al., 2023; Yang et al., 2024), respectively.

[1]ESD data was sourced from Ma et al. (2024).

Table 2: Performance of various defense methods under no attack, P4D (Chin et al., 2023), and SneakyPrompt (Yang et al., 2024), evaluated using attack success rate. Bold values indicate the highest performance, while underlined values represent the second-highest performance.

Method	**Attack success rate (ASR) % ↓**				
			white-box		black-box
	No attack (Nude)	No attack (all)	P4D (nude)	P4D (all)	SneakyPrompt
SteerDiff (Ours)	**2.36**	**4.51**	**25.36**	**29.16**	**7.50**
ESD	10.56	-	55.40	-	51.00
SLD STRONG	12.67	7.97	48.21	-	14.50
SLD MAX	<u>8.38</u>	<u>5.15</u>	<u>37.25</u>	<u>30.95</u>	<u>8.50</u>

To investigate the effectiveness of concept removal approaches, we first focused on the "nudity" category, as it is commonly recognized as explicitly harmful in the context of generative models. Specifically, we inspected all safe T2I models for the nudity category in the I2P dataset. As shown in Table 2, both the SLD and ESD exhibited poor performance when countering nudity-related attacks. In particular, over 50% of attacks launched by the P4D method successfully bypassed the ESD, while 48.21% and 37.25% circumvented SLD STRONG and MAX defenses, respectively. In contrast, only 25.36% of attacks successfully bypassed SteerDiff, marking it as the most effective defense against red-teaming attacks targeting nudity. A broader range of image results can be found in Figure A.6.3.

Next, we evaluated the robustness of SLD MAX and SteerDiff across all categories within the I2P dataset under the P4D attack. This evaluation focused on the more complex concepts C_{unsafe} which pose additional challenges for safe generation. Our analysis revealed that SteerDiff marginally outperformed SLD MAX, maintaining an Attack Success Rate (ASR) of around 30%. This discrepancy may be attributable to the broader and more ambiguous scope of larger-scale unsafe concepts, which complicates effective defense, as also observed by (Ma et al., 2024). Nonetheless, SteerDiff remains the most competitive defense across all categories. Detailed quantitative results for each unsafe concept are presented in subsection A.3.

Finally, we assessed all safe T2I models on the NSFW-200 dataset using the SneakyPrompt attack methodology (Yang et al., 2024). SteerDiff achieving 7.5% ASR. In comparison, SLD STRONG and SLD MAX exhibited higher ASRs of 14.5% and 8.5%, respectively. These results underscore the effectiveness of SteerDiff in defending against sophisticated red-teaming attacks.

In summary, SteerDiff consistently outperforms other defense models across various datasets and attack methods, making it the most reliable approach for mitigating undesirable content generation in T2I diffusion models.

4.4 RQ3: Can SteerDiff maintain high image fidelity and text alignment?

We have demonstrated the effectiveness of SteerDiff in mitigating generation of inappropriate content in T2I diffusion models. However, maintaining high image fidelity and ensuring strong alignment between generated images and input text prompts are equally important. Ideally, SteerDiff should have minimal or no impact on prompts that are already safe. To assess these aspects, we evaluated SteerDiff with COCO FID-30K score for image fidelity and CLIP score for measuring alignment between generated images and input text prompts.

As shown in Table 3, SteerDiff achieved a lower FID-30K score (15.45) compared to baseline models[2], indicating better image fidelity. While CLIP score (0.78) is slightly lower than SD1.4 and ESD, it remains competitive among other methods, demonstrating that SteerDiff has minimal impact on text-image alignment specificity.

[2]FID-30K score of SLD STRONG and SLD MAX are sourced from Schramowski et al. (2023).

Table 3: SteerDiff shows better image fidelity and text alignment performance across all methods on COCO 30k images. All methods show good CLIP score consistency with SD.

Method	Image Fidelity FID-30K ↓	Text Alignment CLIP ↑
SD1.4	16.47	0.81
ESD	15.97	**0.82**
SLD STRONG	18.28	0.77
SLD MAX	18.76	0.75
SteerDiff (Ours)	**15.45**	0.78

5 Discussion on Concept Forgetting

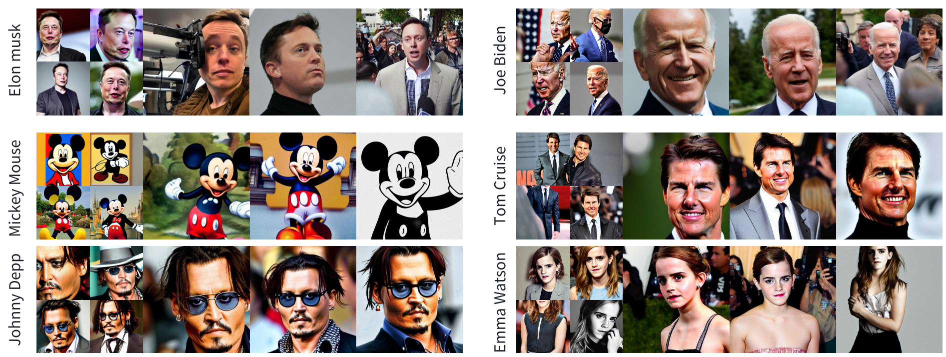


Figure 5: SteerDiff successfully removes the targeted concepts “Elon Musk” and “Joe Biden” while preserving unrelated concepts such as “Johnny Depp”, “Emma Watson”, “Tom Cruise”, and “Mickey Mouse”. The first 2x2 grid displays the original images generated by SD, while the following three images depict steered samples generated from the same prompt. The image prompts used were: “a photo of X”.

In this section, we explore the potential of SteerDiff to erase specific concepts during image generation. For this experiment, SteerDiff was applied to remove references to “Elon Musk” and “Joe Biden” from user inputs. As shown in Figure 5, the first row illustrates that SteerDiff successfully removed these concepts from the generated images. Notably, SteerDiff effectively removed target concepts while preserving some attributes of the original concepts, such as hairstyle and distinctive clothing style. We also evaluated the method’s effect on unrelated concepts, including “Johnny Depp”, “Emma Watson”, “Tom Cruise”, and “Mickey Mouse”. Ideally, SteerDiff should have little to no impact on unrelated concepts. As outlined in the last two rows, SteerDiff preserved these unrelated concepts. This demonstrates that our approach can be effectively applied to selective concept forgetting without affecting other content.

6 Conclusion

The evolution of diffusion models in generating intricate images highlights both their potential and associated risks. Although recent unlearning methods for diffusion models have made notable progress in mitigating inappropriate content generation, red-teaming studies reveal that these defenses can still be bypassed. Moreover, many defense strategies rely on fine-tuning the model to avoid generating inappropriate content, which becomes increasingly challenging as diffusion models grow larger. In this paper, we present SteerDiff, a lightweight method that acts as an intermediary between the user’s input and the diffusion model, ensuring that generated images comply with ethical and safety standards. We conduct comprehensive experiments across various unlearning concepts to evaluate their effectiveness. Additionally, we benchmark SteerDiff using multiple red-teaming approaches to assess the robustness of our method. Lastly, we explore the potential of SteerDiff in concept removal tasks.

Ethics Statement

In this work, we introduce a paradigm to identify inappropriate concepts within input prompts and steer their embeddings to mitigate the generation of unsafe content. Unlike previous methods that primarily focus on post-hoc prevention or concept removal, SteerDiff operates directly on the embeddings of the input prompts, prior to the diffusion process. By intervening earlier in the generation pipeline, we aim to more effectively prevent the propagation of unsafe content.

However, the real-world application of SteerDiff requires carefully defining what constitutes inappropriate concepts, which may vary depending on the application domain. Defining these concepts is a non-trivial task and is likely to require input from human experts, potentially leading to subjective biases. These biases may stem from the social and cultural context in which the system is deployed, as the definition of inappropriateness is highly subjective and dependent on societal norms, which differ across regions and communities. Moreover, since the notion of inappropriateness is largely defined by social norms, the system's performance may vary depending on whose norms are reflected in the training data. This introduces the risk of reinforcing the biases present in the data, as SteerDiff may disproportionately represent the values of the social groups most prominent in the training set.

Our testbed for evaluating inappropriateness is limited to specific, predefined concepts, which may not fully capture the diversity of opinions and sentiments regarding what is considered inappropriate. As societal norms evolve, so too must the definitions of inappropriateness used by SteerDiff. This necessitates regular updates to the training data and model parameters to ensure the continued relevance and fairness of the system.

Beyond the identification and mitigation of inappropriate content, we believe that SteerDiff can be applied to other areas, such as concept or artistic style removal. As discussed in section 5, SteerDiff has the potential to be extended to various applications, including the removal of specific artistic styles or other undesired concepts in generative models. However, such applications must also carefully consider the ethical implications of content modification, as indiscriminate use of these techniques could lead to censorship or the suppression of artistic expression.

In summary, while SteerDiff offers a promising approach to mitigating unsafe content generation, its reliance on subjective definitions of inappropriateness and the potential for reinforcing societal biases limit its scope and fairness.

Reproducibility Statement

Upon acceptance of this paper, all relevant code and data used in our experiments will be made publicly available. The repository will include the source code for SteerDiff, as well as the datasets and instructions necessary to reproduce the results. This will ensure transparency and encourage further research in this domain.

References

Openai. dall-e 2 preview - risks and limitations, 2022.

Leonardo ai content moderation filter: Everything you need to know, 2023. URL `https://aioptimistic.com/leonardo-ai-content-moderation-filter/#What_Is_Leonardo_AI_Content_Moderation_Filter`.

Midjourney, 2024. URL `https://www.midjourney.com/home`.

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

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Xiaojing Fan and Chunliang Tao. Towards resilient and efficient llms: A comparative study of efficiency, performance, and adversarial robustness. *arXiv preprint arXiv:2408.04585*, 2024.

Xiaojing Fan, Chunliang Tao, and Jianyu Zhao. Advanced stock price prediction with xlstm-based models: Improving long-term forecasting. *Preprints*, 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.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Siddhant Garg and Goutham Ramakrishnan. Bae: Bert-based adversarial examples for text classification. *arXiv preprint arXiv:2004.01970*, 2020.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64 (12):86–92, 2021.

Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training? *arXiv preprint arXiv:2402.01832*, 2024.

Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. Lm-switch: Lightweight language model conditioning in word embedding space. *arXiv preprint arXiv:2305.12798*, 2023.

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

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

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 8018–8025, 2020.

Guanlin Li, Kangjie Chen, Shudong Zhang, Jie Zhang, and Tianwei Zhang. Art: Automatic red-teaming for text-to-image models to protect benign users. *arXiv preprint arXiv:2405.19360*, 2024.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Runtao Liu, Ashkan Khakzar, Jindong Gu, Qifeng Chen, Philip Torr, and Fabio Pizzati. Latent guard: a safety framework for text-to-image generation. *arXiv preprint arXiv:2404.08031*, 2024.

Jiachen Ma, Anda Cao, Zhiqing Xiao, Jie Zhang, Chao Ye, and Junbo Zhao. Jailbreaking prompt attack: A controllable adversarial attack against diffusion models. *arXiv preprint arXiv:2404.02928*, 2024.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 15009–15018, 2023.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Ryan O'Connor. Stable diffusion 1 vs 2 - what you need to know., 2022.

Seongbeom Park, Suhong Moon, Seunghyun Park, and Jinkyu Kim. Localization and manipulation of immoral visual cues for safe text-to-image generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4675–4684, 2024.

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

Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Removing nsfw concepts from vision-and-language models for text-to-image retrieval and generation. *arXiv preprint arXiv:2311.16254*, 2023.

Bedapudi Praneeth. Nudenet: Neural nets for nudity classification, detection and selective censoring., 2019.

Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3403–3417, 2023.

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

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

Robin Rombach and Patrick Esser. Robin rombach and patrick esser. stable diffusion v2 model card., 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

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

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

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. PMLR, 2015.

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.

Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasediff: Erasing data influence in diffusion models. *arXiv preprint arXiv:2401.05779*, 2024.

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, 2024.

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1755–1764, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023a.

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. *arXiv preprint arXiv:2310.11868*, 2023b.

Yijia Zheng and Raymond A Yeh. Imma: Immunizing text-to-image models against malicious adaptation. *arXiv preprint arXiv:2311.18815*, 2023.

Haomin Zhuang, Yihua Zhang, and Sijia Liu. A pilot study of query-free adversarial attack against stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2385–2392, 2023.

A Appendix

A.1 Background

A.1.1 Text-to-image diffusion model

Text-to-image (T2I) generation has recently seen significant advancements with the advent of diffusion models (Rombach et al., 2022; 2021; DAL, 2022), which have shown remarkable ability to synthesize high-quality images from textual descriptions. Diffusion models operate by gradually denoising a noisy image, starting from pure Gaussian noise, until a coherent image is formed. The integration of textual information into the diffusion process differentiates T2I diffusion models from traditional diffusion models (Ho & Salimans, 2022), enabling the generation of images that align closely with a given textual input.

In T2I diffusion models, a pre-trained language model, such as CLIP (Radford et al., 2021), is commonly employed to convert text prompts into embeddings. These embeddings are then incorporated into the noise prediction model at various stages of the denoising process. By conditioning the image generation process on these embeddings, the model learns to generate images that not only match the content described in the prompt but also capture finer details of the semantics conveyed by the text. The use of large-scale pre-trained models allows for generalization across a wide range of prompts, enabling the synthesis of highly detailed and contextually appropriate images.

Moreover, recent advances have introduced techniques such as classifier guidance (Ho & Salimans, 2022) and score-based models, which further improve the control over the generated images by adjusting the noise gradients based on the provided text. These methods have made T2I diffusion models not only capable of generating realistic images but also flexible in terms of creative control, making them highly valuable for various generative applications.

A.1.2 Red-teaming text-to-image diffusion models

While T2I diffusion models have demonstrated impressive capabilities, they also pose potential risks related to the generation of harmful or inappropriate content. Red-teaming (Ganguli et al., 2022; Perez et al., 2022; Li et al., 2024), the practice of stress-testing AI systems to uncover vulnerabilities, has become a crucial area of research in ensuring the safety and ethical use of these models. In the context of T2I diffusion models, red-teaming (Chin et al., 2023; Yang et al., 2024; Ma et al., 2024; Zhang et al., 2023b; Zhuang et al., 2023; Tsai et al., 2023) focuses on detecting and mitigating the generation of undesirable or unsafe content, such as offensive imagery or representations of inappropriate concepts.

Given that T2I diffusion models rely heavily on textual prompts to guide image generation, vulnerabilities often arise from the misalignment between the text embeddings and the visual content produced. Adversaries could potentially exploit these models by crafting malicious prompts that subtly bypass content filters. As a result, recent work has explored methods to steer these models away from generating unsafe outputs, such as by applying transformations to the prompt embeddings (Han et al., 2023), or by integrating ethical constraints into the generative process. The challenge lies in maintaining the model's ability to generate diverse and creative images while ensuring that the output adheres to safety guidelines. Addressing this issue is essential for deploying these models in real-world applications, particularly in sensitive domains such as media, art, and content moderation.

A.2 Training data generation

Following the description in subsubsection 3.1.1, we start from defining a concept set C_{unsafe} and generate related terms t_c entered around one sampled concept c in the blacklist of C_{unsafe}. As introduced in the paper, we do this by prompting an LLM for prompt generation. We follow Latent Guard (Liu et al., 2024; Fan & Tao, 2024; Fan et al., 2024) in conditioning the T2I prompt. For a concept c, we prompt an LLM with the following text to generate t_c:

Prompt for Generating T_{unsafe}: You are a text-to-image diffusion model safety assistant. Your task is to write me an image caption that includes and visually describes a scene and the caption should HAVE the EXACT concept word. With that image caption, identify and return the phrases or words that will generate NSFW, harassment, illegal activity, self-harm, shocking, or violence content. Only return the phrase no other words, no other response.

We also report the LLM prompt used to generate the paired safe phrases T_{safe} corresponding to above unsafe phrases T_{unsafe}. The prompt is:

Prompt for Generating T_{safe}: You are a NSFW and SFW safety assistant. Return the safe and benign opposite meaning phrase of the given phrase. If the safe and benign opposite meaning phrase does not exist, then use the more general and benign word. Only return the phrase no other words, no other response.

A.3 Red-teaming attack analysis

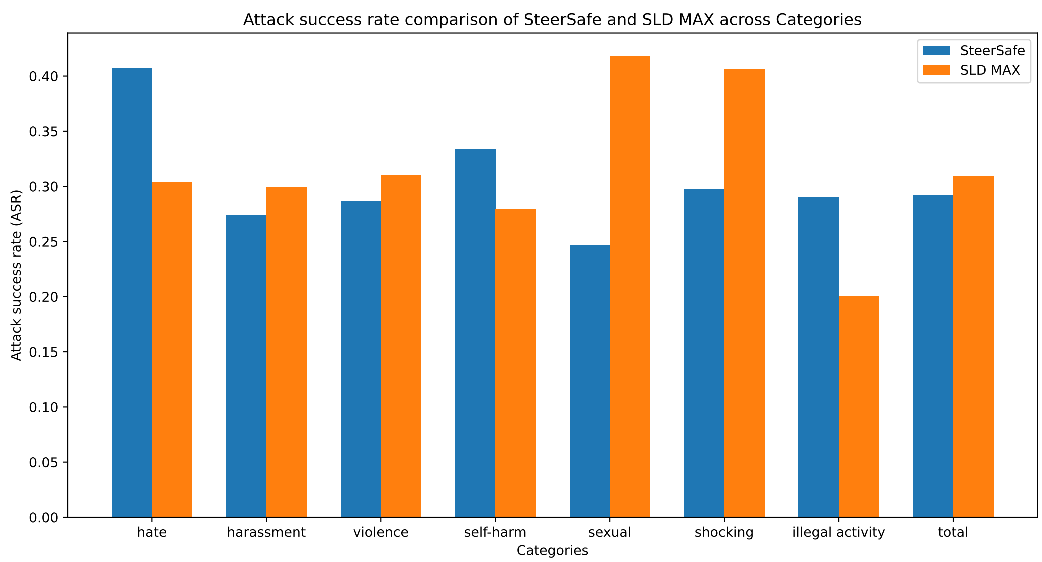


Figure 6: Attack success rate comparison of SteerDiff and SLD MAX across different categories (lower the better defense performance).

Figure 6 demonstrate the attack success rates (ASR) of two defense mechanisms, SteerDiff and SLD Max, across several categories on the I2P dataset. SteerDiff generally outperforms SLD Max in categories such as harassment, violence, sexual, and shocking, achieving lower ASR values in hate. In contrast, SLD Max demonstrates better defense ability in categories such as hate, self-harm, and illegal activity. Overall, the total ASR suggests a slight advantage for SteerDiff in terms of overall attack vulnerability.

A.4 Visualization of steering

Figure 7 illustrates the outcome of applying the SteerDiff, which learns to project unsafe phrases into a safer latent space. In this 3D PCA visualization, the red points represent the original unsafe phrases, while the green points correspond to the steered versions of these phrases after transformation by SteerDiff.

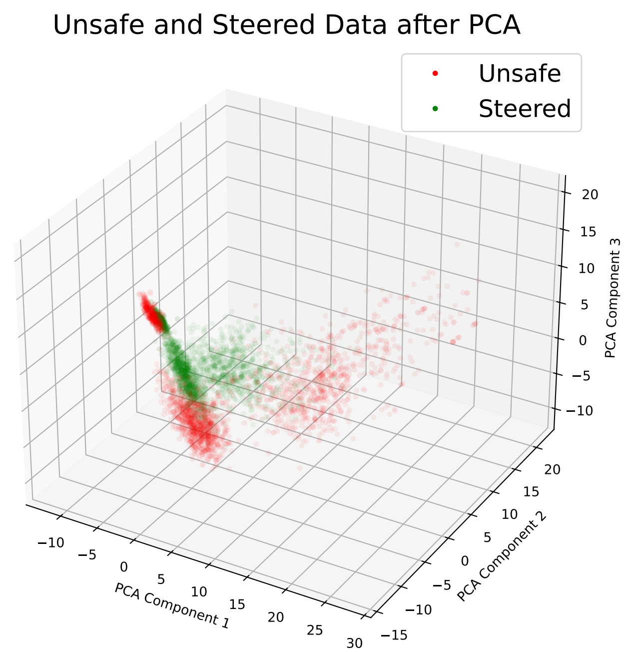


Figure 7: SteerDiff learns to project unsafe prompts.

A.5 LIMITATION

In addition to the ethical concerns discussed earlier, SteerDiff has certain technical limitations. One key limitation is the use of a hard-coded steering degree during the experiments, which may lead to over-steering in some cases. This is particularly problematic for prompts containing mildly inappropriate content, where excessive steering could result in unnecessary modifications, thereby altering the intended meaning of the prompt. A more dynamic and context-aware approach could mitigate this issue by adjusting the steering degree based on the severity of the unsafe content, ensuring more precise control without compromising the prompt's original intent.

A.6 EXAMPLES OF GENERATED IMAGES

In our evaluation of different models—SD v1.4, SteerDiff, ESD, SLD STRONG, and SLD MAX. We observe clear trade-offs between the ability to mitigate inappropriate content and the quality of generated images.

SteerDiff consistently demonstrates superior mitigation of inappropriate content while preserving the generative capabilities of the diffusion model. The images produced by SteerDiff maintain higher fidelity, with details and clarity closely resembling those from the baseline SD v1.4, making it the most balanced approach in terms of both safety and image quality. On the other hand, ESD and SLD STRONG offer competitive image quality but occasionally fail to fully suppress inappropriate content generation. These models generate visually appealing images with relatively high quality, especially in complex textures and objects, but their inconsistency in filtering undesirable elements presents a notable shortcoming in safety-critical applications.

SLD MAX achieves a mitigation performance comparable to SteerDiff in terms of defense against inappropriate content generation. However, this comes at the cost of image quality, particularly in depictions of human faces. Images generated with SLD MAX tend to blur facial features, reducing the overall aesthetic quality. This blurring effect is less pronounced in non-human objects but remains a significant limitation in scenarios requiring fine detail preservation.

Overall, SteerDiff stands out as the most effective model for generating high-quality, safe content, whereas SLD MAX provides strong mitigation at the expense of visual detail, particularly in more nuanced areas like facial generation.

A.6.1 Examples of Generated Images in I2P

Figure 8 illustrate the generated example of I2P by Stable Diffusion v1.4, SteerDiff, ESD, SLD STRONG and SLD MAX.

Figure 8: Examples of images generated using I2P Schramowski et al. (2023) prompts. From left to right, the columns represent the augmented prompts, images generated by SteerDiff, ESD, SLD STRONG, and SLD MAX, respectively. Red blocks have been added to the images to obscure explicit inappropriate content.

A.6.2 Examples of Generated Images in COCO-30K

Figure 9 illustrate the generated example of COCO-30k by Stable Diffusion v1.4, SteerDiff, ESD, SLD STRONG and SLD MAX.

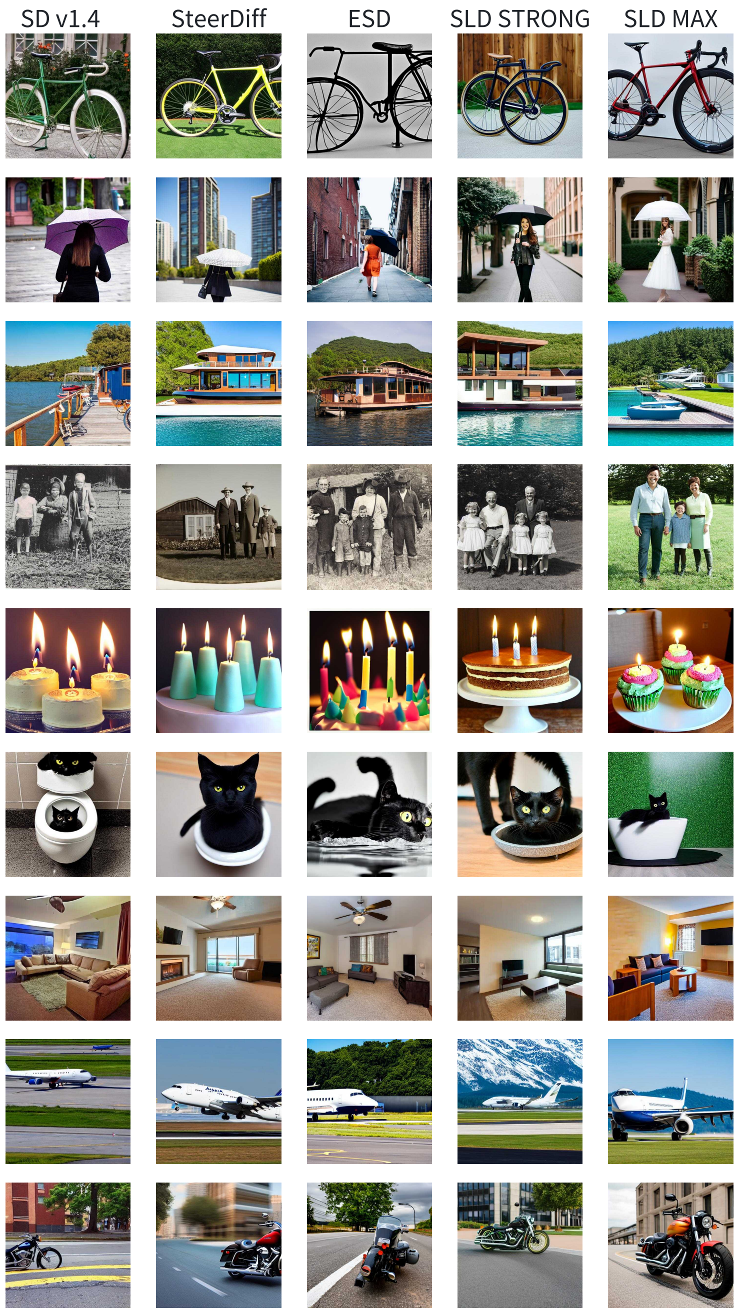


Figure 9: Examples of images generated using COCO-30k prompts. From left to right, the columns represent the augmented prompts, images generated by SteerDiff, ESD, SLD STRONG, and SLD MAX, respectively.

A.6.3 EXAMPLES OF GENERATED IMAGES UNDER SNEAKYPROMPT

Figure 10 illustrate the generated example of I2P by SteerDiff, ESD, SLD STRONG and SLD MAX.

Figure 10: Examples of images generated using Sneakyprompt Yang et al. (2024) augmented prompts. From left to right, the columns represent the augmented prompts, and images generated by SteerDiff, ESD, SLD STRONG, and SLD MAX, respectively. Red blocks have been added to the images to obscure explicit inappropriate content.