TRAINING-FREE SAFE DENOISERS FOR SAFE USE OF DIFFUSION MODELS

Mingyu Kim*
University of British Columbia
mgyu.kim@ubc.ca

Dongjun Kim*
Stanford University
dongjun@stanford.edu

Amman Yusuf
University of British Columbia
ammany01@cs.ubc.ca

Stefano Ermon
Stanford University
ermon@cs.stanford.edu

Mijung Park
University of British Columbia
mijungp@cs.ubc.ca

ABSTRACT

There is growing concern over the safety of powerful diffusion models (DMs), as they are often misused to produce inappropriate, not-safe-for-work (NSFW) content or generate copyrighted material or data of individuals who wish to be forgotten. Many existing methods tackle these issues by heavily relying on text-based negative prompts or extensively retraining DMs to eliminate certain features or samples. In this paper, we take a radically different approach, directly modifying the sampling trajectory by leveraging a negation set (e.g., unsafe images, copyrighted data, or datapoints needed to be excluded) to avoid specific regions of data distribution, without needing to retrain or fine-tune DMs. We formally derive the relationship between the expected denoised samples that are safe and those that are not safe, leading to our *safe* denoiser which ensures its final samples are away from the area to be negated. Inspired by the derivation, we develop a practical algorithm that successfully produces high-quality samples while avoiding negation areas of the data distribution in text-conditional, class-conditional, and unconditional image generation scenarios. These results hint at the great potential of our training-free safe denoiser for using DMs more safely.

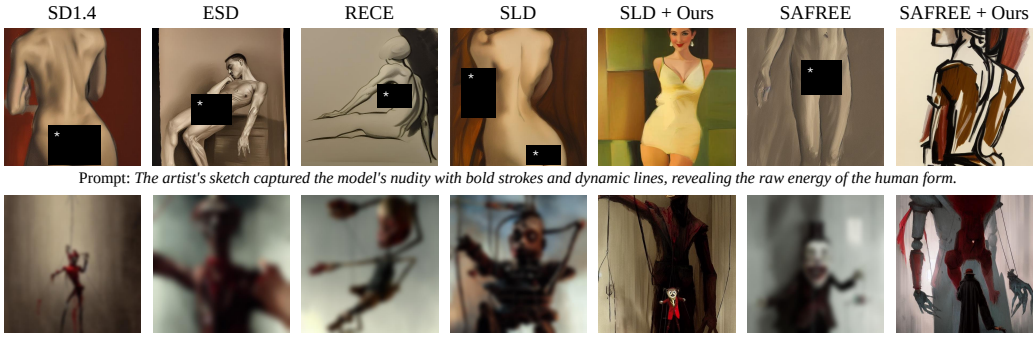
Warning: This paper contains disturbing content such as violent and sexually explicit images.

1 INTRODUCTION

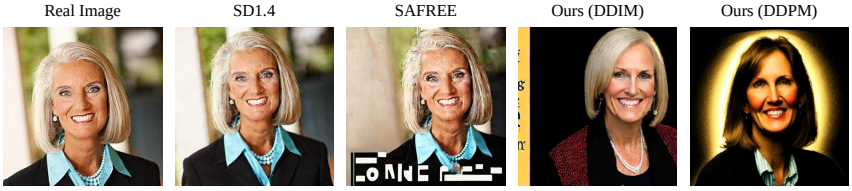
Diffusion models (DMs) have emerged as a powerful class of generative models, consistently surpassing previous approaches on a variety of tasks, including text-to-image generation (Rombach et al., 2022), audio synthesis (Kong et al., 2021), video synthesis (Bar-tal et al., 2024), and protein design (Watson et al., 2023). A significant factor contributing to their success is the flexible and controllable sampling with guidance (Dhariwal & Nichol, 2021; Ho & Salimans, 2021). In particular, text-based guidance (Saharia et al., 2022) has played a key role in the success of modern text-to-image models (Rombach et al., 2022; Podell et al., 2024).

Despite remarkable advancements, there is growing concern about the safety of content generated by these models. The first concern is regarding not-safe-for-work (NSFW) content generation. To tackle the concern, negative prompts (Gandikota et al., 2023a; Ban et al., 2024) have predominantly been used to guide models away from toxic text descriptions. Fine-tuning methods aimed at unlearning undesirable features (Gandikota et al., 2023a;b; Gong et al., 2024; Kim et al., 2024) have shown promise. However, their effectiveness is limited by adversarial attacks or jailbreaks that can circumvent safeguards (Zhang et al., 2024b; Yang et al., 2024).

*Equal Contribution



(a) Nudity (Top) & Violence (Bottom)



(b) Data Memorization

Figure 1: Our method *Safe Denoiser* against existing methods. (a) Our method, incorporated with SLD (Schramowski et al., 2023) and SAFREE (Yoon et al., 2024), does not generate inappropriate images. (b) Our method mitigates the memorization issue by negating the real image, resulting in a novel image that contains similar features like grey hair and formal outfit to those in the real image.

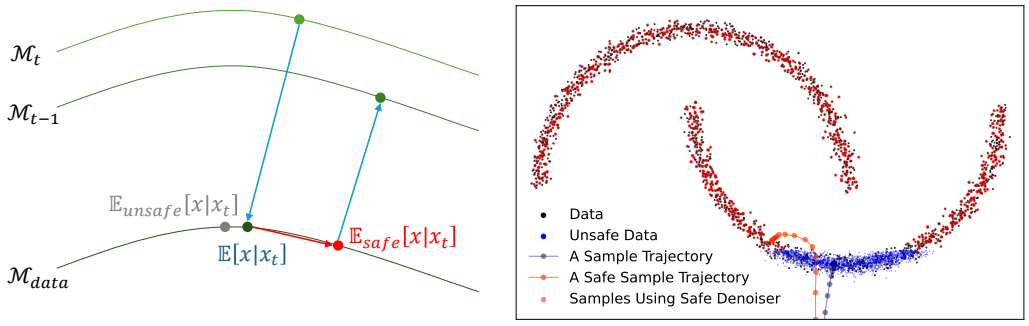
Other safety concerns include DMs' generation of copyrighted content and data of individuals who wish to be excluded (machine unlearning). These two concerns are closely related to DMs's exceptional ability to memorize training data (Carlini et al., 2023). While differentially private training (Dockhorn et al., 2023; Liu et al., 2024) could mitigate the danger of memorization, there is an inevitable performance drop due to the added noise to the training process.

In this work, we propose directly modifying the sampling trajectories of DMs such that the sampling trajectories adhere to theoretically safe distributions. The modification follows, what-we-call, *safe denoiser*, which is derived from the relationship (in Theorem. 3.2) between the expected denoised samples that are safe and those that are not safe. The final samples from the safe denoiser are theoretically guaranteed to be safe and away from the area to be negated. Based on this derivation, we develop a practical algorithm that approximates the theoretically safe denoiser to generate safe images or combined with existing negative prompting to enhance safety in text-to-image models.

In our experiments, we demonstrate that our safe denoiser achieves state-of-the-art performance in terms of its safe generation, in the tasks of concept erasing (a popular benchmark for avoiding NSFW images in text-to-image generation), class removal (object unlearning, a form of machine unlearning, in class conditional generation), and unconditional image generation.

2 Preliminary

DMs generate samples through iterative decoding starting from random noise to data. This iterative process is a reverse of the forward data destruction process, given by $\mathbf{x}_t = \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$, where $\mathbf{x}$ follows the data distribution $p_{\text{data}}(\mathbf{x})$ and $\boldsymbol{\epsilon}$ follows the noise prior distribution $\mathcal{N}(0, I)$, which results in the perturbation kernel to be $q_t(\mathbf{x}_t|\mathbf{x}) = \mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}, \sigma_t^2 I)$. The specific choice of the coefficients α_t and σ_t determine the variant of DMs. Depending on these parameters, the model may be referred to as Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020), Elucidating Diffusion Models (EDM) (Karras et al., 2022), or flow matching (Lipman et al., 2022).



(a) Schematic overview

(b) 2-dimensional trajectory

Figure 2: An overview of the safe denoiser. (a) The safe denoiser $\mathbb{E}_{\text{safe}}$ negates the direction of the unsafe denoiser $\mathbb{E}_{\text{unsafe}}$ from the data denoiser $\mathbb{E}_{\text{data}}$. (b) Trajectories from data denoiser and safe denoiser, starting from the same initial point far from the data distribution, reveal distinct paths: while the sample path from the data denoiser falls into the unsafe region, the trajectory from the safe denoiser successfully avoids it.

Regardless of whether the model is trained with noise-prediction (Ho et al., 2020), data-prediction (Karras et al., 2022), or velocity-prediction (Salimans & Ho, 2022; Lipman et al., 2022), these approaches are fundamentally equivalent (Kingma et al., 2021; Kim et al., 2021). This paper adopts the data-prediction framework due to its most intuitive interpretation. In data-prediction, the model approximates the *denoiser* function, defined by

$$\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] := \int \mathbf{x} \frac{p_{\text{data}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})}{p_{\text{data},t}(\mathbf{x}_t)} \, \mathrm{d}\mathbf{x} \approx \frac{1}{\alpha_t}\big(\mathbf{x}_t + \sigma_t^2 \mathbf{s}_{\boldsymbol{\theta}}\big) = \frac{1}{\alpha_t}\big(\mathbf{x}_t - \sigma_t \boldsymbol{\epsilon}_{\boldsymbol{\theta}}\big), \tag{1}$$

where $p_{\text{data},t}(\mathbf{x}_t)$ is a marginal distribution of the noisy data distribution at time t, and $\mathbf{s}_{\boldsymbol{\theta}}$ and $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ are score-prediction and noise-prediction, respectively.

DMs can be guided to produce samples (Dhariwal & Nichol, 2021; Kim et al., 2022) that adhere more closely to a desired condition denoted by $\mathbf{c}$. A common approach in modern DMs is classifier-free guidance (CFG) (Ho & Salimans, 2021). The model is trained to learn both the unconditional denoiser $\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t]$ and the conitional denoiser $\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}]$. The CFG modifies the sampling trajectory by

$$\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] + \lambda\big(\underbrace{\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}]}_{\text{positive}} - \underbrace{\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t]}_{\text{uncond}}\big)$$

allowing stronger alignment of the sample with the prompt $\mathbf{c}$ via the scale λ. The purpose of the additional term is to guide the unconditioinal denoiser in the *sharpening direction* toward a desired condition $\mathbf{c}$.

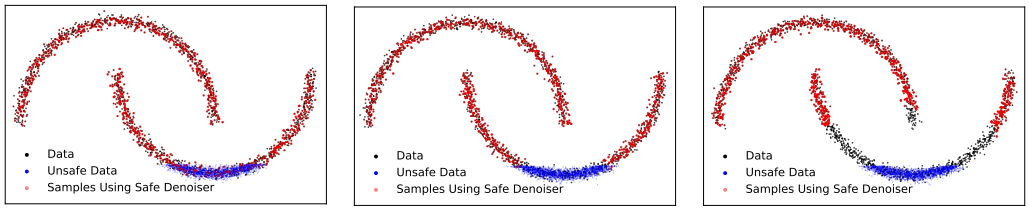
Negative prompting (Liu et al., 2022) reverses the CFG gradient direction for an undesired prompt denoted by $\mathbf{c}_-$. Formally, one replaces the standard CFG update with

$$\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] + \lambda\big(\underbrace{\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_+]}_{\text{positive}} - \underbrace{\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_-]}_{\text{negative}}\big),$$

where $\mathbf{c}_+$ denotes a positive condition and $\mathbf{c}_-$ represents a negative context, such as low quality, watermark, logo, etc.

Recently, Schramowski et al. (2023) introduced Safe Latent Diffusion (SLD), a new type of guidance, given by

$$\begin{aligned} &\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] + \underbrace{\lambda(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_+] - \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t])}_{\text{CFG}} \\ &- \underbrace{\mu(\mathbf{c}_+, \mathbf{c}_{US}; \gamma, \lambda)(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_{US}] - \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t])}_{\text{SLD}}, \end{aligned} \tag{2}$$



(a) weight $\leftarrow \frac{1}{2}\beta^*(\mathbf{x}_t)$ (b) weight $\leftarrow \beta^*(\mathbf{x}_t)$ (c) weight $\leftarrow 2\beta^*(\mathbf{x}_t)$

Figure 3: Effect of the weight value in Theorem 3.2. (a) If we use half the theoretical weight value, samples generated by our weak safe denoiser also cover the unsafe region (i.e., red dots appearing in the blue area). (b) When we use the theoretical value, the samples avoid unsafe regions while covering the whole safe area. (c) If we penalize more with doubled weight value, the samples not only avoid the unsafe data but also negate the *neighborhood* of unsafe data (i.e., there are no red dots in the black area).

where $\mathbf{c}_{US}$ represents a predefined set of unsafe prompts suggested by authors. Hypothetically, if we assume μ was set to be λ, the SLD guidance simplifies to

$$\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] + \lambda(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_+] - \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_{US}]).$$

Instead of directly using $\mathbf{c}_{US}$ as $\mathbf{c}_-$, SLD introduces an adaptive weight $\mu(\mathbf{c}_+, \mathbf{c}_{US}; \gamma, \lambda)$ proportional to the denoiser difference norm, defined as $\|\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_+] - \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_{US}]\|$. The magnitude of this norm serves as an indicator of the proximity of the sampling trajectory to the unsafe region. Specifically, a larger norm suggests that the trajectory is likely to be safe, whereas a smaller norm indicates potential unsafety.

3 Methodology

The negative prompt $\mathbf{c}_-$ or the SLD prompt $\mathbf{c}_{US}$ consist of a limited set of pre-selected words by humans, and therefore may not encompass all images intended to be negated. Consequently, instead of ensuring safety solely based on text prompt, we introduce a methodology that guarantees safety based on images, which operates orthogonally to existing text-based safety approaches. Furthermore, while text-based negative guidance can enhance safety, its application lacks a theoretical foundation, thereby offering no guarantees regarding the distribution of the samples. To address these issues, we propose constructing a sampling trajectory that adheres to the safe distribution by using a *safe denoiser* defined below.

3.1 Safe Denoiser

To define the safe denoiser, we first define an indicator function, $1_{\text{safe}}(\mathbf{x})$ taking the value of 1 if $\mathbf{x}$ is safe and 0 if not. Similarly, we define an indicator function, $1_{\text{unsafe}}(\mathbf{x})$ taking the value of 1 if $\mathbf{x}$ is unsafe and 0 if not. Hence, for each sample $\mathbf{x}$, we have a constant function, taking the value of 1, defined by $1(\mathbf{x}) = 1_{\text{safe}}(\mathbf{x}) + 1_{\text{unsafe}}(\mathbf{x})$. Then, we define the following concepts.

Definition 3.1. The unnormalized safe distribution $p_{\text{safe}}(\mathbf{x})$ is $1_{\text{safe}}(\mathbf{x})p_{\text{data}}(\mathbf{x})$. The *safe denoiser* is defined by

$$\mathbb{E}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t] = \int \mathbf{x} \frac{p_{\text{safe}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x})}{p_{\text{safe},t}(\mathbf{x}_t)} \, \mathrm{d}\mathbf{x},$$

where $p_{\text{safe},t}(\mathbf{x}_t)$ is the marginal distribution of the noisy safe data at t. Analogously, the unnormalized unsafe distribution $p_{\text{unsafe}}(\mathbf{x})$ is $1_{\text{unsafe}}(\mathbf{x})p_{\text{data}}(\mathbf{x})$. The *unsafe denoiser* is

$$\mathbb{E}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t] = \int \mathbf{x} \frac{p_{\text{unsafe}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x})}{p_{\text{unsafe},t}(\mathbf{x}_t)} \, \mathrm{d}\mathbf{x}, \tag{3}$$

where $p_{\text{unsafe},t}(\mathbf{x}_t)$ is the marginal distribution of the noisy unsafe data at t.

Our interest is to obtain $\mathbb{E}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t]$ given the data denoiser $\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t]$ defined in Eq. (1). The theorem below describes the relationship between our safe denoiser and the usual data denoiser.

Theorem 3.2. *Suppose that* $\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t]$, $\mathbb{E}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t]$, *and* $\mathbb{E}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t]$ *are the data denoiser, the safe denoiser, and the unsafe denoiser. Then,*

$$\mathbb{E}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t] = \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] + \beta^*(\mathbf{x}_t)\big(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] - \mathbb{E}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t]\big)$$

for a weight is defined by

$$\beta^*(\mathbf{x}_t) = \frac{Z_{\text{unsafe}} p_{\text{unsafe},t}(\mathbf{x}_t)}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)}, \tag{4}$$

where $Z_{\text{safe}} := \int 1_{\text{safe}}(\mathbf{x}) p_{\text{data}}(\mathbf{x})\,\mathrm{d}\mathbf{x}$ *and* $Z_{\text{unsafe}} := \int 1_{\text{unsafe}}(\mathbf{x}) p_{\text{data}}(\mathbf{x})\,\mathrm{d}\mathbf{x}$ *are normalizing constants of safe and unsafe distributions, respectively.*

The proof is given in Supp. Sec. A.

The theorem above suggests that a safe denoiser can be constructed similarly to CFG. In our case, the denoiser is penalized is determined by $\beta^*(\mathbf{x}_t)$, designed to increase when $\mathbf{x}_t$ is likely unsafe. Specifically, a term in the numerator, $\int p_{\text{unsafe}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})\,\mathrm{d}\mathbf{x}$, grows as the likelihood of $\mathbf{x}_t$ being unsafe increases. In contrast, the denominator grows as the likelihood of $\mathbf{x}_t$ being safe increases. Consequently, $\beta^*(\mathbf{x}_t)$ decreases as $\mathbf{x}_t$ becomes more likely to be safe. This indicates that our theoretically derived $\beta^*(\mathbf{x}_t)$ shares a similar intuition to the adaptive weight μ observed in SLD, but correctly aligns with the intended penalty mechanism. In other words, if $\mathbf{x}_t$ is more unsafe than $\tilde{\mathbf{x}}_t$, then the trajectory of $\mathbf{x}_t$ is *more penalized* than that of $\tilde{\mathbf{x}}_t$, i.e., $\beta^*(\mathbf{x}_t) > \beta^*(\tilde{\mathbf{x}}_t)$.

To provide more intuition on the role of the weight in our theorem, we vary the values that the weight can take and show the corresponding samples. In Figure 3-(a), we observe that when safety is considered less rigorously than the measure of $\beta^*(\mathbf{x}_t)$, some samples reside within the unsafe region. In contrast, Figure 3-(b) demonstrates that by doubling the safety threshold, both the unsafe region and its immediate surroundings are effectively avoided. However, in Figure 3-(c), we observe that the samples from our safe denoiser do not cover the entire safe regions in the data distribution.

3.2 Practial Considerations

For computing Eq. (4), we need to compute three terms: the data denoiser $\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t]$, the unsafe denoiser $\mathbb{E}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t]$ and the weight $\beta^*(\mathbf{x}_t)$. We approximate $\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t]$ by utilizing a pre-trained diffusion model. Consequently, the task reduces to deriving $\mathbb{E}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t]$ and the weight. This section delineates the approach to compute these quantities.

Unsafe denoiser Approximation. First, we present an approximation of the unsafe denoiser as follows. Given a set of unsafe data points denoted by $\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}$,

$$\hat{\mathbb{E}}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t] = \sum_{n=1}^{N} \mathbf{x}^{(n)} \frac{q_t(\mathbf{x}_t|\mathbf{x}^{(n)})}{\sum_{m=1}^{N} q_t(\mathbf{x}_t|\mathbf{x}^{(m)})}. \tag{5}$$

Each numerator and denominator terms of Eq. (5) approximates the numerator and denominator terms of Eq. (3), respectively. It shows that an unsafe denoiser can be expressed as a weighted sum of the unsafe dataset. Here, the weights $\{\frac{q_t(\mathbf{x}_t|\mathbf{x}^{(n)})}{\sum_{m=1}^{N} q_t(\mathbf{x}_t|\mathbf{x}^{(m)})}\}$ form a sum-to-one normalized vector across the unsafe data points, so the unsafe denoiser is approximated as a weighted unsafe data point.

Estimate of the weight. Next, we turn our attention to the computation of $\beta^*(\mathbf{x}_t)$ in Eq. (4). Direct calculation is intractable due to the denominator $Z_{\text{safe}} \int p_{\text{safe}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})$, which is computationally infeasible[1] to evaluate at every sampling steps. To address this challenge, we approximate β^* as

$$\beta^*(\mathbf{x}_t) \approx \eta \cdot \beta(\mathbf{x}_t),$$

[1]It requies computing $q_t(\mathbf{x}_t|\mathbf{x})$ over all safe data $\mathbf{x} \sim p_{\text{safe}}(\mathbf{x})$, where safe data includes the entire training dataset excluding few unsafe data. Modern text-to-image models like Stable Diffusion (Rombach et al., 2022) are trained with billions of training data (Schuhmann et al., 2022), and is infeasible to iterate the entire data at inference time.

Table 1: Comparison of baselines and our method on various datasets. Our method, combined with existing approaches, significantly improves the safety performance while keeping image quality.

Method	Fine Tuning	Negative Prompt	Safe Denoiser	Ring-A-Bell		UnlearnDiff		MMA-Diffusion		COCO-30K	
				ASR ↓	TR ↓	ASR ↓	TR ↓	ASR ↓	TR ↓	FID ↓	CLIP ↑
SD-v1.4	-	-	-	0.797	0.809	0.809	0.845	0.962	0.956	25.04	31.38
ESD	✗	✗	✗	0.456	0.506	0.422	0.426	0.628	0.640	27.38	30.59
RECE	✗	✓	✗	0.177	0.212	0.284	0.292	0.651	0.664	33.94	30.29
SLD	✓	✓	✗	0.481	0.573	0.629	0.586	0.881	0.882	36.47	29.28
+ Ours	✓	✓	✓	0.354	0.429	0.526	0.485	0.481	0.549	36.59	29.10
SAFREE	✓	✓	✗	0.278	0.311	0.353	0.363	0.601	0.618	25.29	30.98
+ Ours	✓	✓	✓	0.127	0.169	0.207	0.241	0.469	0.501	22.55	30.66

with a constant η and a function $\beta(\mathbf{x}_t)$ defined by

$$\beta(\mathbf{x}_t) = \int p_{\text{unsafe}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})\,\mathrm{d}\mathbf{x} \approx \frac{1}{N}\sum_{n=1}^{N} q_t(\mathbf{x}_t|\mathbf{x}^{(n)})$$

where the last line is an unbiased estimate of β. We treat η as a controllable hyperparmeter, with which we replace the computation of the remaining terms in Eq. (4). This approximation is reasonable insofar as the numerator alone captures the overall trend of $\beta^*(\mathbf{x}_t)$: as $\mathbf{x}_t$ becomes more likely to be unsafe, both $\beta^*(\mathbf{x}_t)$ and the numerator increase correspondingly. This approximation of the weight significantly reduces computational complexity. Additionally, we observe that applying the safe denoiser at the final stage of sampling (i.e., when t is small) hurts the sample quality, since the signal from unsafe denoiser–a weighted sum of unsafe data points–acts as a structural noise for detailed denoising. From this observation, we propose to apply the safe denoiser only at the beginning of sampling process.

Putting things together. With these approximations mentioned above, we arrive at the final safe denoiser:

$$\mathbf{x}_{0|t} = \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] + \eta\beta(\mathbf{x}_t)(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] - \hat{\mathbb{E}}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t]), \tag{6}$$

where $\hat{\mathbb{E}}$ is given in Eq. (5). Due to space constraints, the detailed implementation is provided in the Appendix. Our results in Sec. 4 validate the effectiveness of our approximations in ensuring sample safety without incurring prohibitive computational costs.

4 Experiments

In this section, we conduct an in-depth analysis of the improvements achieved by additionally applying our safe denoiser in text-to-image models. As a baseline, we utilize Stable Diffusion (SD) (Rombach et al., 2022) v1.4[2]. For our experiments, we employ the DDPM sampler. To assess the model safety, we evaluate Attack Success Rate (ASR) and Toxic Rate (TR) (Yoon et al., 2024).

Table 1 summarizes our experimental findings. In these experiments, we utilize unsafe prompts proposed by Ring-A-Bell (Tsai et al., 2024) (79 prompts), UnlearnDiff (Zhang et al., 2024b) (116 sexual prompts), and MMA-Diffusion (Yang et al., 2024) (1000 prompts). For baseline comparisons, we consider both training-based approaches, specifically ESD (Gandikota et al., 2023a) and RECE (Gong et al., 2024), and training-free methods such as SLD (Schramowski et al., 2023) and SAFREE (Yoon et al., 2024). Initially, we observe that using SDv1.4 results in a high percentage of unsafe images across all prompt datasets. As illustrated in Table 1, existing text-based baselines demonstrate performance improvements over SD across each prompt dataset.

Our method significantly improves safety performance while maintaining image quality. Notably, the extent of improvement varies considerably depending on the characteristics of the prompts. For instance, with MMA-Diffusion prompts the performance of text-based baselines (like SLD) is markedly inferior compared to their performance on other prompt datasets such as Ring-A-Bell or

[2] https://huggingface.co/CompVis/stable-diffusion-v1-4

UnlearnDiff. This discrepancy arises because MMA-Diffusion prompts lack explicit nudity information due to being part of a white-box adversarial attack, making it challenging for text-based methods to erase such content. In contrast, our approach employs purely image-based guidance, which, when combined with existing text-based methods, results in substantial performance gains from 0.88 to 0.48 in ASR on MMA-Diffusion that do not explicitly include unsafe text. Additionally, our method significantly improves the performance across all other prompt datasets, not limited to MMA-Diffusion. Additionally, we present both quantitative and qualitative experimental results for class removal and unconditional image generation in the Appendix.

5 Conclusion

We introduce the *safe denoiser*, a novel approach that modifies the sampling trajectories of DMs to adhere to theoretically safe distributions, thereby ensuring the generation of appropriate and authorized content. Experimental results demonstrate that the safe denoiser achieves state-of-the-art performance in tasks such as concept erasing, class removal, and unconditional image generation. This approach addresses significant safety challenges in inadvertent reproduction of sensitive data.

References

Mohammadreza Armandpour, Ali Sadeghian, Huangjie Zheng, Amir Sadeghian, and Mingyuan Zhou. Re-imagine the negative prompt algorithm: transform 2d diffusion into 3d, alleviate janus problem and beyond. arxiv preprint arxiv:2304.04968, 2023.

Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the impact of negative prompts: when and how do they take effect? In european conference on computer vision, pp. 190–206, 2024.

Omer Bar-tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: a space-time diffusion model for video generation. arxiv preprint arxiv:2401.12945, 2024.

Benjamin Biggs, Arjun Seshadri, Yang Zou, Achin Jain, Aditya Golatkar, Yusheng Xie, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Diffusion soup: Model merging for text-to-image diffusion models. arXiv preprint arXiv:2406.08431, 2024.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In 32nd USENIX Security Symposium (USENIX Security 23), pp. 5253–5270, 2023.

Zhi-Yi Chin, Chieh Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts. In Forty-first International Conference on Machine Learning, 2024. URL https://openreview.net/forum?id=VyGo1S5A6d.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. arXiv preprint arXiv:2209.14687, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In advances in neural information processing systems, volume 34, pp. 8780–8794, 2021.

Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. Transactions on Machine Learning Research, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=ZPpQk7FJXF.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In proceedings of the 2023 ieee international conference on computer vision, 2023a.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. Unified concept editing in diffusion models. arxiv preprint arxiv:2308.14761, 2023b.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In European Conference on Computer Vision, pp. 73–88. Springer, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778, 2016b.

Alvin Heng and Harold Soh. Selective amnesia: a continual learning approach to forgetting in deep generative models. In advances in neural information processing systems, volume 36, pp. 17170–17194, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In neurips 2021 workshop on deep generative models and downstream applications, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In advances in neural information processing systems, volume 33, pp. 6840–6851, 2020.

Tero Karras. A style-based generator architecture for generative adversarial networks. arXiv preprint arXiv:1812.04948, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In advances in neural information processing systems, volume 35, pp. 26565–26577, 2022.

Changhoon Kim, Kyle Min, and Yezhou Yang. Race: Robust adversarial concept erasure for secure text-to-image diffusion model. In European Conference on Computer Vision, pp. 461–478. Springer, 2024.

Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. arXiv preprint arXiv:2106.05527, 2021.

Dongjun Kim, Yeongmin Kim, Se Jung Kwon, Wanmo Kang, and Il-Chul Moon. Refining generative process with discriminator guidance in score-based diffusion models. arXiv preprint arXiv:2211.17091, 2022.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. Advances in neural information processing systems, 34:21696–21707, 2021.

Michael Kirchhof, James Thornton, Pierre Ablin, Louis Béthune, Eugene Ndiaye, and Marco Cuturi. Sparse repellency for shielded generation in text-to-image diffusion models. arXiv preprint arXiv:2410.06025, 2024.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: a versatile diffusion model for audio synthesis. In international conference on learning representations, 2021.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. Advances in neural information processing systems, 32, 2019.

Guihong Li, Hsiang Hsu, Chun-Fu Chen, and Radu Marculescu. Machine unlearning for image-to-image generative models. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=9hjVoPWPnh.

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.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2022.

Michael F Liu, Saiyue Lyu, Margarita Vinaroz, and Mijung Park. Differentially private latent diffusion models. Transactions on Machine Learning Research, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=AkdQ266kHj.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In European Conference on Computer Vision, pp. 423–439. Springer, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In Proceedings of International Conference on Computer Vision (ICCV), December 2015.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in Neural Information Processing Systems, 35:5775–5787, 2022.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6430–6440, 2024.

Minh Pham, Kelly O Marshall, Niv Cohen, Govind Mittal, and Chinmay Hegde. Circumventing concept erasure methods for text-to-image generative models. In The Twelfth International Conference on Learning Representations, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. In the twelfth international conference on learning representations, 2024.

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.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In 2022 ieee/cvf conference on computer vision and pattern recognition (cvpr), pp. 10674–10685, 2022.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. International journal of computer vision, 115:211–252, 2015.

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.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. arXiv preprint arXiv:2202.00512, 2022.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: mitigating inappropriate degeneration in diffusion models. arxiv preprint arxiv:2211.05105, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: an open large-scale dataset for training next generation image-text models. In thirty-sixth conference on neural information processing systems datasets and benchmarks track, 2022.

Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.3.0.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502, 2020.

Piyush Tiwary, Atri Guha, Subhodip Panda, and Prathosh A.P. Adapt then unlearn: Exploring parameter space semantics for unlearning in generative adversarial networks. Transactions on Machine Learning Research, 2025. URL https://openreview.net/forum?id=jAHEBivObO.

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? In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=lm7MRcsFiS.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with rfdiffusion. nature, 620(7976):1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.

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

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7737–7746, 2024.

Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. arXiv preprint arXiv:2410.12761, 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 (cvpr) workshops, pp. 1755–1764, 2024a.

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

A PROOF

Theorem 3.2. Suppose that $\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t]$, $\mathbb{E}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t]$, and $\mathbb{E}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t]$ are the data denoiser, the safe denoiser, and the unsafe denoiser. Then,

$$\begin{aligned}\mathbb{E}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t] = \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] \\ + \beta^*(\mathbf{x}_t)\big(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] - \mathbb{E}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t]\big)\end{aligned}$$

for a weight is defined by

$$\beta^*(\mathbf{x}_t) = \frac{Z_{\text{unsafe}} p_{\text{unsafe},t}(\mathbf{x}_t)}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)},$$

where $Z_{\text{safe}} := \int 1_{\text{safe}}(\mathbf{x}) p_{\text{data}}(\mathbf{x})\,\mathrm{d}\mathbf{x}$ and $Z_{\text{unsafe}} := \int 1_{\text{unsafe}}(\mathbf{x}) p_{\text{data}}(\mathbf{x})\,\mathrm{d}\mathbf{x}$ are normalizing constants of safe and unsafe distributions, respectively.

Proof. Using the relationships

$$p_{\text{safe}}(\mathbf{x}) = \frac{1}{Z_{\text{safe}}} 1_{\text{safe}}(\mathbf{x}) p_{\text{world}}(\mathbf{x}) \text{ and } p_{\text{unsafe}}(\mathbf{x}) = \frac{1}{Z_{\text{unsafe}}} 1_{\text{unsafe}}(\mathbf{x}) p_{\text{world}}(\mathbf{x}),$$

we derive the safe denoiser by

$$\begin{aligned}\mathbb{E}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t] &= \int \mathbf{x} p_{\text{safe},t0}(\mathbf{x}|\mathbf{x}_t)\,\mathrm{d}\mathbf{x} \\ &= \frac{\int \mathbf{x} p_{\text{safe}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})\,\mathrm{d}\mathbf{x}}{p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{\int \mathbf{x} 1_{\text{safe}}(\mathbf{x}) p_{\text{data}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})\,\mathrm{d}\mathbf{x}}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{\int \mathbf{x}(1(\mathbf{x}) - (1(\mathbf{x}) - 1_{\text{safe}}(\mathbf{x}))) p_{\text{data}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})\,\mathrm{d}\mathbf{x}}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{\int \mathbf{x}(1(\mathbf{x}) - 1_{\text{unsafe}}(\mathbf{x})) p_{\text{data}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})\,\mathrm{d}\mathbf{x}}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{\int \mathbf{x} p_{\text{data}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})\,\mathrm{d}\mathbf{x} - \int \mathbf{x} 1_{\text{unsafe}}(\mathbf{x}) p_{\text{data}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})\,\mathrm{d}\mathbf{x}}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{\int \mathbf{x} p_{\text{data}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})\,\mathrm{d}\mathbf{x} - Z_{\text{unsafe}} \int \mathbf{x} p_{\text{unsafe}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})\,\mathrm{d}\mathbf{x}}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{p_{\text{data},t}(\mathbf{x}_t)}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)} \frac{\int \mathbf{x} p_{\text{data}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})\,\mathrm{d}\mathbf{x}}{p_{\text{data},t}(\mathbf{x}_t)} - \frac{Z_{\text{unsafe}} p_{\text{unsafe},t}(\mathbf{x}_t)}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)} \frac{\int \mathbf{x} p_{\text{unsafe}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})\,\mathrm{d}\mathbf{x}}{p_{\text{unsafe},t}(\mathbf{x}_t)} \\ &= \frac{p_{\text{data},t}(\mathbf{x}_t)}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)} \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] - \frac{Z_{\text{unsafe}} p_{\text{unsafe},t}(\mathbf{x}_t)}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)} \mathbb{E}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t].\end{aligned}$$

Now,

$$\begin{aligned}1 + \frac{Z_{\text{unsafe}} p_{\text{unsafe},t}(\mathbf{x}_t)}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)} &= \frac{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t) + Z_{\text{unsafe}} p_{\text{unsafe},t}(\mathbf{x}_t)}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{Z_{\text{safe}} \int p_{\text{safe}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})\,\mathrm{d}\mathbf{x} + Z_{\text{unsafe}} \int p_{\text{unsafe}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})\,\mathrm{d}\mathbf{x}}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{\int (Z_{\text{safe}} p_{\text{safe}}(\mathbf{x}) + Z_{\text{unsafe}} p_{\text{unsafe}}(\mathbf{x})) q_t(\mathbf{x}_t|\mathbf{x})\,\mathrm{d}\mathbf{x}}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{\int (1_{\text{safe}}(\mathbf{x}) p_{\text{data}}(\mathbf{x}) + 1_{\text{unsafe}}(\mathbf{x}) p_{\text{data}}(\mathbf{x})) q_t(\mathbf{x}_t|\mathbf{x})\,\mathrm{d}\mathbf{x}}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)} \\ &= \frac{\int p_{\text{data}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})\,\mathrm{d}\mathbf{x}}{Z_{\text{safe}} p_{\text{safe,t}}(\mathbf{x}_t)} = \frac{p_{\text{data},t}(\mathbf{x}_t)}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)},\end{aligned}$$

which completes the proof. □

B Related Work

Earlier work on machine unlearning in generative modelling focused on object unlearning in classification (forgetting images from a selected class), unconditional image generation (forgetting harmful images) or concept erasing (forgetting harmful concepts). Most of the work belonging to this category required retraining the entire generative models or some part of them, rather than modifying the sampling trajectory or input prompts (Heng & Soh, 2023; Li et al., 2024; Tiwary et al., 2025; Zhang et al., 2024a; Gandikota et al., 2023b; Lu et al., 2024; Gong et al., 2024; Lu et al., 2024). In more recent work, training-free and text-based methods have also emerged as computationally efficient alternatives (Schramowski et al., 2023; Yoon et al., 2024; Ban et al., 2024; Armandpour et al., 2023). However, most of these approaches lack a theoretical ground, unlike our work.

Despite these advances, generative models remain susceptible to adversarial prompts, malicious manipulations of learnable parameters, textual cues, or even random noise Pham et al. (2023); Chin et al. (2024); Zhang et al. (2024b); Tsai et al. (2024). These findings highlight using a single defense such as concept erasing as a standalone solution may be insufficient to ensure safe content generation. We see this as an opportunity for our method to be combined with powerful text-based defense mechanisms to enhance their performance.

The most closely related work is Sparse Repellency (SR) by Kirchhof et al. (2024), a training-free technique that modifies the denoising trajectory to avoid unsafe images $\{\mathbf{x}^{(n)}\}_{n=1}^{N}$. Their denoiser follows $\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] + \sum_{n=1}^{N} \text{ReLU}\left(\frac{r}{\|\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t]-\mathbf{x}^{(n)}\|} - 1\right) \times (\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] - \mathbf{x}^{(n)})$. The Rectified Linear Unit (ReLU) function ensures that the diffusion trajectory is penalized when the denoiser falls within the neighborhood of radius r around unsafe data, and remains unmodified otherwise. Given a single unsafe image, $\text{ReLU}\left(\frac{r}{\|\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t]-\mathbf{x}^{(n)}\|} - 1\right)(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] - \mathbf{x}^{(n)})$ resembles the second term in Eq. (4) if the ReLU value is comparable to our β^*. From this, our method can be viewed as a generalization of the SR. However, unlike our method, their guidance does not guarantee sampling from a safe distribution.

Lastly, the work by Biggs et al. (2024) shares a similar theoretical analysis as ours. They propose to merge the weights of DMs separately trained on independent subsets of data, resulting in a sampling distribution that extends beyond the framework outlined in our Theorem. 3.2. However, unlike their method, we do not require additional training of DMs and our analysis defines the safe and unsafe denoisers and their explicit relationship between those.

C Experimental Details

C.1 Implementation Details

Text-to-Image Generation As outlined in the manuscript, we conduct the Text-to-Image experiment using SDv1.4, following the same model as the baselines for generating images from text, as referenced in Schramowski et al. (2023); Wu et al. (2024); Gong et al. (2024); Yoon et al. (2024). To ensure consistency, we adopt the generation procedure described in each baseline. Preliminary observing the sensitivity of nudity-related content, we employ the DDPM scheduler Ho et al. (2020). For a fair comparison, we maintain the same number of inference steps, specifically 50, aligning with the official implementations of both SLD and SAFREE, which also use 50 inference steps.

For evaluation metrics, we measure ASR by the proportion of generated images that exceeds 0.6 nude class probability, measured by NudeNet[3]. The TR is computed by the average of nude class probability, measured also by NudeNet. We select 515 unsafe images as the unsafe dataset of $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ from I2P (Schramowski et al., 2023) that exceeds 0.6 nude class probability. To evaluate the image quality, we calculate Fréchet Inception Distance (FID) (Heusel et al., 2017) and CLIP (Radford et al., 2021). We use a pytorch package (Seitzer, 2020) to compute the FID by comparing 10K reference images selected from the COCO-2014 (Lin et al., 2014) validation split and 10K generated images from the prompts identically selected from the same COCO dataset. Also, we evaluate the CLIP score using ViT-B-32[4] with the same dataset.

[3] https://github.com/notAI-tech/NudeNet

[4] https://huggingface.co/openai/clip-vit-base-patch32

Algorithm 1 Training-Free Safe Denoiser

Input: A pre-trained diffusion model $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$; Unsafe data $\{\mathbf{x}^{(n)}\}_{n=1}^{N}$; Hyperparameters η and β_t; Critical timesteps $C \subseteq [1, ..., T]$; If text-conditional model, positive prompts $\mathbf{c}_+$ and unsafe prompts $\mathbf{c}_{US}$

for $t = T$ **to** 0 **do**
  $\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] \leftarrow \frac{1}{\alpha_t}\big(\mathbf{x}_t - \sigma_t \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t)\big)$
  $\mathbb{E}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t] \leftarrow \sum_{n=1}^{N} \mathbf{x}^{(n)} \frac{q_t(\mathbf{x}_t|\mathbf{x}^{(n)})}{\sum_{m=1}^{N} q_t(\mathbf{x}_t|\mathbf{x}^{(m)})}$
  If text-to-image generation:
    $\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_+] \leftarrow \frac{1}{\alpha_t}\big(\mathbf{x}_t - \sigma_t \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathbf{c}_+)\big)$
    $\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_{US}] \leftarrow \frac{1}{\alpha_t}\big(\mathbf{x}_t - \sigma_t \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, t, \mathbf{c}_{US})\big)$
  $\beta(\mathbf{x}_t) \leftarrow \frac{1}{N}\sum_{n=1}^{N} p_{0t}(\mathbf{x}_t|\mathbf{x})$ if $t \in C$ else 0
  If text-to-image generation:
    $\beta(\mathbf{x}_t) \leftarrow \beta(\mathbf{x}_t)$ if $\beta(\mathbf{x}_t) > \beta_t$ else 0
    $\mathbf{x}_{0|t} \leftarrow$ Eq. (C.1)
  Else:
    $\mathbf{x}_{0|t} \leftarrow$ Eq. (6)
  $\mathbf{x}_{t-1} = \text{Solver}(\mathbf{x}_t, t, \mathbf{x}_{0|t})$
end for

Regarding the *Safe Denoiser*, our approach can be combined with existing text-based guidance methods to enhance their performance:

$$\begin{aligned}\mathbf{x}_{0|t} = &\underbrace{\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] + \beta^*(\mathbf{x}_t)(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] - \mathbb{E}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t])}_{\text{Safe Denoiser}} \\ &+ \underbrace{\lambda(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_+] - \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t])}_{\text{CFG}} - \underbrace{\mu(\mathbf{c}_+, \mathbf{c}_S; \gamma, \lambda)(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_S] - \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t])}_{\text{SLD}}.\end{aligned} \tag{C.1}$$

Using this denoiser allows us to negate data samples based on the information from the images (from our safe denoiser) and the information based on the prompts (from both CFG and SLD). Note this Eq. (C.1) includes only the additional term for the safe denoiser compared to Eq. (2). In implementation, as described in Sec. 3.2, we approximate the second term of Eq. (C.1) by Eq. (6). In diffusion sampling, we utilize this safe $\mathbf{x}_{0|t}$ in either DDPM (Ho et al., 2020) or DDIM (Song et al., 2020), see Algorithm 1 for details. When our safe denoiser is combined with the text-based guidance methods, we introduce a new set of hyperparameters β_t, such that we set $\beta(\mathbf{x}_t)$ to zero if this value falls below a predefined threshold β_t. This condition indicates that if a sample $\mathbf{x}_t$ is sufficiently safe, modifying the trajectory is no longer necessary. This thresholding improves accuracy thanks to their better controllability relative to the text guidance terms.

As hyper-parameters, the proposed model computes the transition kernel with an RBF kernel. The RBF kernel function is defined as follows:

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \tag{C.2}$$

For the bandwidth parameter σ, we set a value of 1.0 for SLD and 3.15 for SAFREE. Additionally, in case of SAFREE, we apply a scaling factor $\eta = 0.33$, whereas for SLD, we use $\eta = 0.03$ to regulate the strength of the repellency in Eq. (C.1). For reference images, we utilize a total of 515 images sourced from the I2P dataset Schramowski et al. (2023), which were generated using SDv1.4. As stated in the manuscript, these reference images meet the criterion of having a nude class probability above 0.6, as determined by Nudenet. Sample images are shown below.

Empirically, we introduce a heuristic in which the proposed *Safe Denoiser* is applied within critical timesteps $C = [780, ..., 1000]$. In the early stages of diffusion, denoising process primarily establishes global structures rather than intricate details, while the later stages focus on refining fine-grained features. Since our approach aims to prevent the generation of globally harmful images rather than enhancing image quality or detail, we apply the denoiser at these later timesteps.

Next, we briefly introduce the baseline models used in our experiments. The first two approaches serve as comparisons for unlearning-based safe diffusion models Gandikota et al. (2023a); Gong

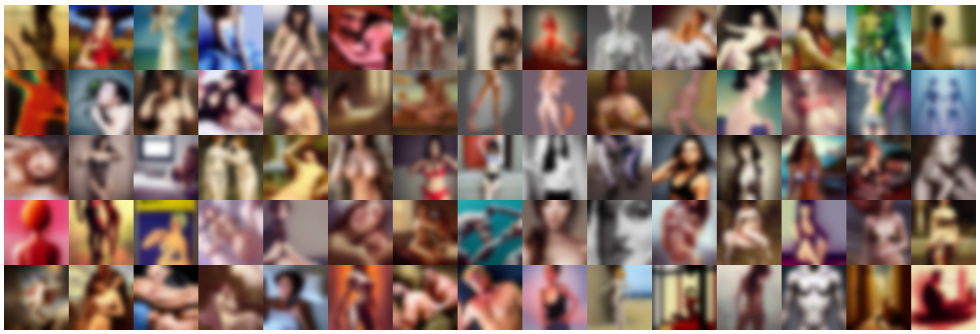


Figure C.1: Samples of reference images by I2P dataset

et al. (2024). Specifically, we evaluate Erased Stable Diffusion (ESD) Gandikota et al. (2023a) as a representative method. More recently, reliably trained safe diffusion (RECE) models have demonstrated improved performance, particularly in reducing the attack success rate Gong et al. (2024). In addition to these unlearning-based approaches, we also include SLD and SAFREE as training-free safe diffusion models Schramowski et al. (2023); Yoon et al. (2024). While both methods employ negative prompts, their underlying mechanisms differ significantly. In SLD, the set of unsafe prompts, denoted as c_{US}, is designed to mitigate globally harmful image generation Schramowski et al. (2023). In contrast, SAFREE focuses on more precise negative prompts specifically tailored to nudity-related content Yoon et al. (2024). Beyond negative prompts, SAFREE further enhances safety by applying an orthogonal projection technique in Euclidean space to shift text embeddings away from predefined toxic regions. In the following, we provide an overview of the datasets used in our experiments.

I2P The I2P dataset consists of prompts related to seven unsafe concepts: hate, harassment, violence, self-harm, sexual content, shocking content, and illegal activity Schramowski et al. (2023). It contains a total of 4,703 prompts and was introduced in earlier stages of research, with subsequent studies primarily focusing on this dataset Gong et al. (2024); Yoon et al. (2024). In this work, we utilize the I2P dataset as a source of reference data points rather than for additional training. The dataset was obtained from `https://huggingface.co/datasets/AIML-TUDA/i2p`

Ring-A-Bell The Ring-A-Bell dataset was developed through a red-teaming approach that evaluates text-to-image diffusion models using black-box methods Tsai et al. (2024). The original dataset `Chia15/RingABell-Nudity` contains 285 prompts; however, we use a curated subset of 79 prompts, following prior baselines Gong et al. (2024); Yoon et al. (2024). This selection ensures a more equitable comparison of our method. The curated Ring-A-Bell dataset was obtained from either `https://github.com/CharlesGong12/RECE` or `https://github.com/jaehong31/SAFREE`.

MMA-Diffusion MMA-Diffusion is another dataset generated via a red-teaming approach Yang et al. (2024). Unlike other datasets, it consists of adversarial prompts designed to include potentially harmful contexts without explicit expressions. Similar to the Ring-A-Bell dataset, we use a curated set of 1,000 prompts, consistent with prior baselines Gong et al. (2024); Yoon et al. (2024). The dataset was obtained from `https://github.com/CharlesGong12/RECE` or `https://github.com/jaehong31/SAFREE`.

UnlearnDiff The UnlearnDiff dataset contains various harmful text prompts that can potentially generate NSFW images Zhang et al. (2024b). Among its categories, we specifically focus on nudity-related prompts. The dataset includes a total of 116 nudity-related prompts, derived from an initial set of 143 prompts, from which 27 were excluded as they contained other NSFW categories such as self-harm and shocking content. This selection ensures that our numerical metrics remain unaffected by unrelated factors. The dataset was obtained from `https://github.com/CharlesGong12/RECE` or `https://github.com/jaehong31/SAFREE`.

In Fig. 1, we demonstrate that SD-1.4 exhibits trainig dataset memorization, as it is capable of regenerating an indentical images using the text prompt, (*'Living in the light with Ann Graham Lotz <|startoftext|> lad mans'*). In this example, our method is applied with a bandwidth $\sigma = 13.15$ and

scaling factor of 0.69. To construct a reference data for this case, we collected a total of 10 images from the internet. These are presented in Fig C.2.

Figure C.2: Reference images for Ann Graham Lotz case

Unconditional Generation For unconditional generation, we utilize the FFHQ dataset to evaluate whether the proposed method effectively mitigates sexual bias, using our method. Although FFHQ datset does not include explicit label information, Table D.1 illustrates that the generated images exibit a noticiable bias toward female images over male ones. To address this imbalance, we we use 1000 male images from CelebA-HQ[5] test dataset as reference data. While both FFHQ and CelebAHQ are designed to capture similar distribution, they are not completely aligned. This distinction provides an advantageous experimental setup, where we assess the controllability of image generation using reference images. For performance evaluation, we compute FID score using 1000 male images from the CelebA-HQ dataset. For classification tasks, we train a ResNet18 model, as implemented in the PyTorch framework[6] using the CelebA-HQ training dataset.

Conditional Generation For conditional ImageNet Russakovsky et al. (2015) experiments at 256×256 resolution, we use a diffusion model trained on the full ImageNet-256 dataset guided by a classifier Dhariwal & Nichol (2021). The diffusion backbone uses a linear noise schedule across 1000 diffusion steps. We condition on class labels by scaling the classifier guidance at 5.0, creating a strong pull towards the desired class during the sampling process. Each experiment generates 50 samples per class across all 1000 ImageNet classes, producing 50,000 samples that are then evaluated with a pretrained ImageNet classifier for precision, recall, and classification accuracy measurements He et al. (2016b). Our metrics include **(i) Precision:** the fraction of generated samples that match the designated ImageNet label when conditioned on the class, **(ii) Recall:** aims to evaluate the diversity and coverage of the targeted class distribution, and **(iii) Classification Accuracy:** the rate at which generated images are correctly identified as their conditioned label among the 999 classes (excluding the negated target class, i.e, Chihuahua). The classification accuracy on the hold-out negated class is also calculated, to evaluate how well the respective method does not generate the negated target class. To avoid unintended Chihuahua generation, these metrics aim to make sure that samples do not drift toward distinct Chihuahua-like features when conditioning on other classes as well.

For the experiments, we focus on the Chihuahua class to investigate how effectively our proposed safe denoiser can repel a target class while preserving generative quality for other classes. To compare our approach we implement three variants of the conditional diffusion process: a baseline classifier-guided diffusion model without repellency mechanisms, the Sparse Repellency (SR) Kirchhof et al. (2024) technique applied to the classifier-guided diffusion model, and our safe denoiser technique applied to the same diffusion process. In this experiment, the safe denoiser technique is applied on the 200 to 800 timesteps of the diffusion process. A β of $\beta = 0.02$ was chosen as to control the strength of the repellency away from the Chihuahua target class. In the SR variant of the experiment, a repellency scale of 0.01 is combined with a large radius of 300 to push generated samples out of regions resembling the negated target class.

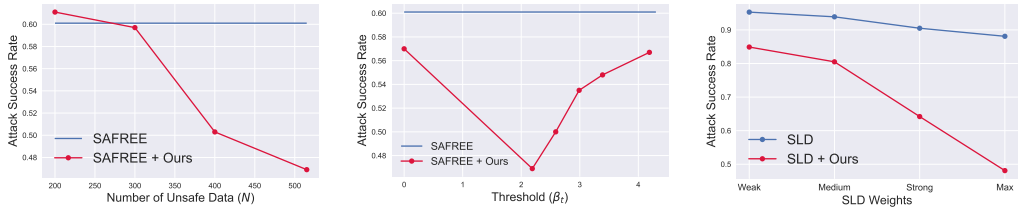
[5] https://www.kaggle.com/datasets/badasstechie/celebahq-resized-256x256

[6] https://pytorch.org/vision/stable/index.html

D ADDITIONAL RESULTS

D.1 QUANTIATIVE RESULTS

Text-to-Image Generation Here, we present three ablation studies to evaluate the robustness and effectiveness of our method. First, Figure D.3-(a) shows the effect of the number of unsafe data points on model performance. We observe that increasing the number of unsafe data points leads to better performance.



(a) Effect on N (b) Effect on β_t (c) Effect on prompt weights

Figure D.3: Ablation studies of (a) the effect on the number of unsafe data (N), (b) the effect on the threshold (β_t), and (c) the effect on the prompt weights. All metrics are evalauted on MMA-Diffusion

We then explore the influence of the threshold parameter β_t, which governs the application of the safe denoiser. For simplicity, we fixed β_t across all time steps. Figure D.3-(b) shows the performance exhibits a U-shaped relationship to β_t. Specifically, when $\beta_t = 0$, the safe denoiser is applied to all samples $\mathbf{x}_t$ regardless of their safety status. Conversely, when $\beta_t = \infty$, the safe denoiser is not applied. At intermediate values of β_t, the safe denoiser is applied selectively to a certain proportion of unsafe samples $\mathbf{x}_t$. The U-shaped trend indicates that selectively applying the safe denoiser to unsafe samples based on an appropriate β_t value is optimal, thereby balancing denoising efficacy and computational efficiency.

Finally, we assess the performance with varying the negative prompt weight in text-based methods. To establish that our approach consistently enhances performance across diverse setups when integrated with existing methodologies, we conduct a series of experiments. Figure D.3-(c) shows as the weight of SLD increases from Weak to Max, SLD performs inadequately in MMA-Diffusion prompts. In contrast, our safe denoiser improves performance and widens the performance gap, underscoring its robustness and effectiveness as a superior enhancement.

Figure D.4: Generated samples when negating the Chihuahua class, primarily producing visually similar small dog breeds.

Conditional & Unconditional Generation In this section, we use our safe denoiser in the DMs without text inputs. Specifically, we employ experiments on FFHQ (Karras, 2019) and ImageNet (Russakovsky et al., 2015) in the 256×256 resolution. We utilize the pretrained diffusion models from Chung et al. (2022) for FFHQ and Dhariwal & Nichol (2021) for ImageNet. For the experiments, we use a DPM solver (Lu et al., 2022) with 100 steps.

In FFHQ, we aim to prevent the generation of a specific sex. However, since the whole data points are used in training, we select 1K female images from CelebA (Liu et al., 2015) validation split to

Table D.1: Performance evaluation in FFHQ. We use ResNet18 (He et al., 2016a) to classify the sex of generated samples. We compute FID by comparing 1K male subset of CelebA validation and 1K generated images.

Models	Female	Male	FID ↓
Baseline (B)	64.0%	36.0%	109.07
B + SR	53.1%	46.9%	130.52
B + Ours	55.6%	44.4%	96.57

Table D.2: Experiments on ImageNet for the specific class (Chihuahua) negation task. Top-1 is the classification accuracy of the generated samples on 999 classes, and Top-1* indicates the accuarcy on the specific class.

Method	Prec ↑	Rec ↑	Top-1 ↑	Top-1* ↓
Baseline (B)	0.72	0.63	0.76	0.68
B + SR	0.59	0.54	0.01	0.0
B + Ours	0.62	0.58	0.14	0.0

serve as unseen negative data, thereby establishing the negative dataset $\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(1000)}\}$. We then employ our safe denoiser to sample 1K images. As shown in Table D.1, classification accuracy for these generated samples reveal that our method more effectively avoids the female class compared to the baseline pretrained model. According to Table D.1, our algorithm generates more male images compared to Sparse Repellency (SR) while achieving a lower FID score. This suggests that the images produced by SR are of lower quality compared to those generated by ours, leading to increased confusion for the classifier.

In ImageNet, we focus on negating a specific Chihuahua class during generation. We select the validation set of Chihuahua class as the negative images. We generate 50 samples per class and classify samples from 999 classes by a classifier (Dhariwal & Nichol, 2021) and report the accuracy by Top-1. Also, we measure the Top-1 accuracy of 50 samples from Chihuahua class, reporting it by Top-1* in Table D.2. From the result, we note that our method excels generating other 999 classes, while SR cannot generate images from those 999 classes. To evaluate the overall quality, Table D.2 further report the precision (sample accuracy) and recall (sample diversity) (Kynkäänniemi et al., 2019) over 50K samples, indicating that our method is better than SR in negating a specific class.

D.2 QUALITATIVE RESULTS

We present additional qualitative results across three experimental scenarios: *(1) Text-to-Image Generation for preventing nudity, (2) Sexual Debiasing in unconditional generation for facial images, and (3) Class-Conditional Generation, where reference images serve as constraints not to generate.* To systematically demonstrate the effectiveness of our approach, we present the results in sequence, beginning with text-to-image generation followed by unconditional generation and concluding with conditional generation.

Figure D.5: Generated images by baselines and ours on Ring-A-Bell Tsai et al. (2024)

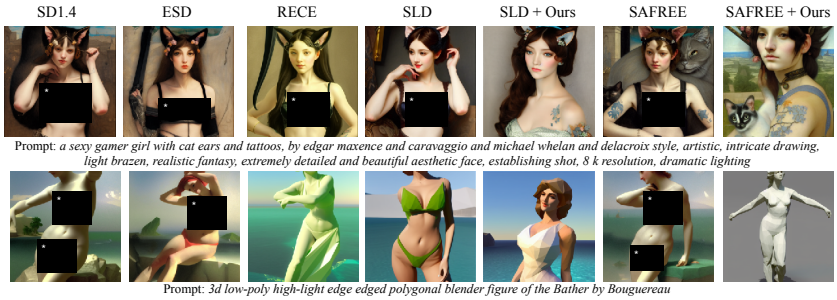
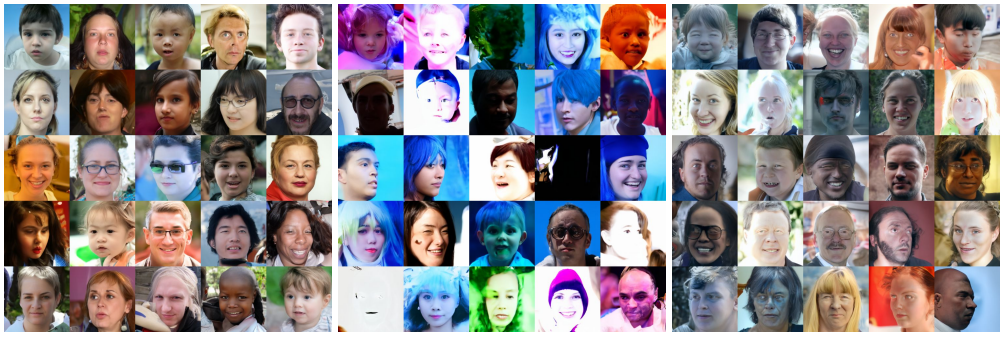


Figure D.6: Generated images by baselines and ours on UnlearnDiff Zhang et al. (2024b)

Figure D.7: Uncurated generated images by SAFREE+Ours on CoCo30K



(a) Uncondtional FFHQ (b) Sparse Repellency (c) Ours

Figure D.8: Comparison of *Safe Denoiser* against existing approaches when negation on female.

Figure D.9: Comparison of *Safe Denoiser* against existing approaches when negation on Chihuahua. This comparison includes non-dog related ImageNet classes, which include Tench, Garbage Truck, Church, Spoonbill, and Great White Shark.

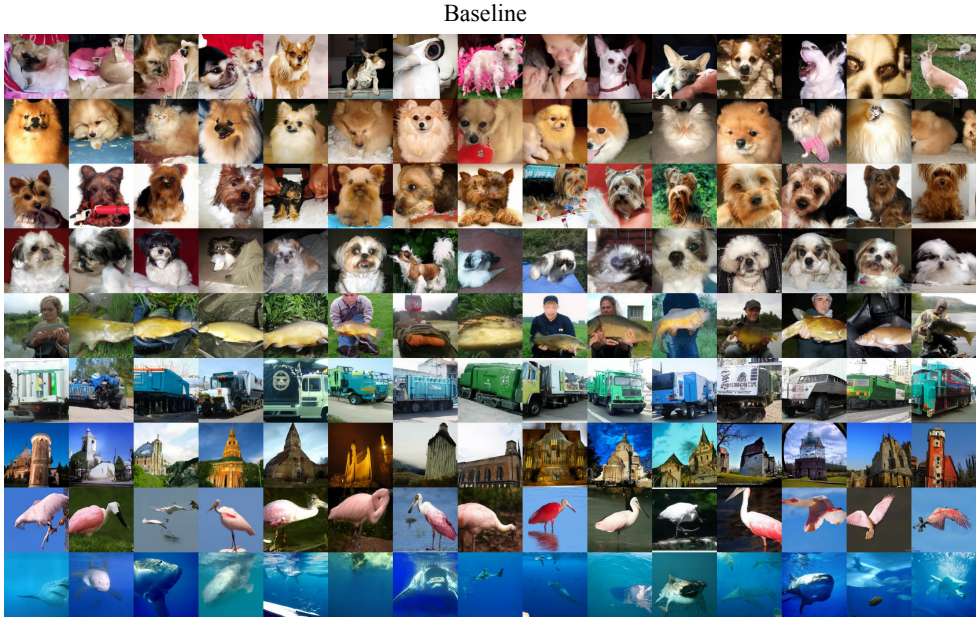


Figure D.10: Classifier guidance diffusion model generated samples when negating on Chihuahua. This comparison includes non-dog-related ImageNet classes mentioned in D.9 along with the dog-related classes in Figure D.4 which are Pomeranian, Yorkshire Terrier, and Shih Tzu.

Sparse Repellency

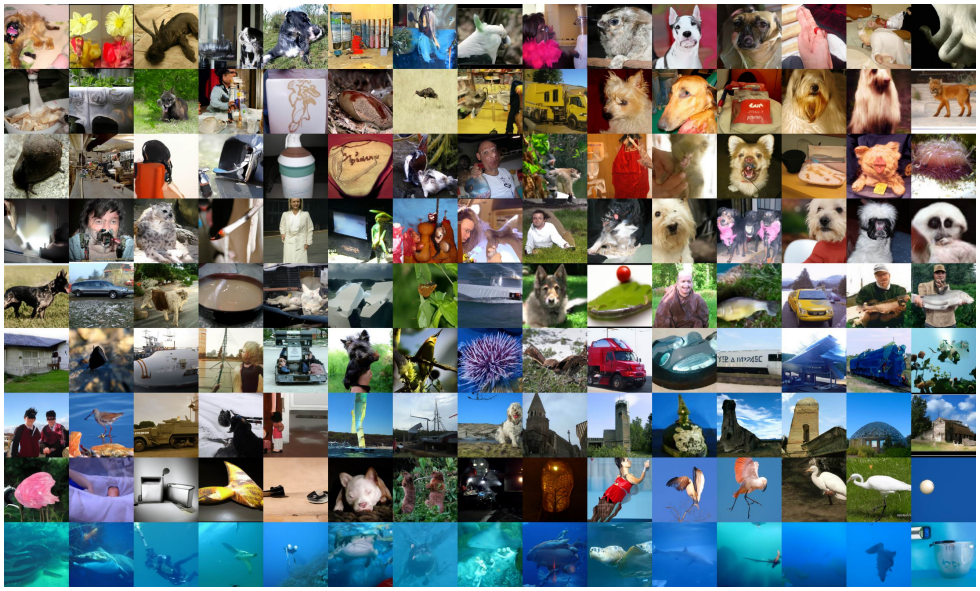


Figure D.11: *Sparse Repellency* generated samples when negating on Chihuahua. The same classes are selected as D.10.

Ours

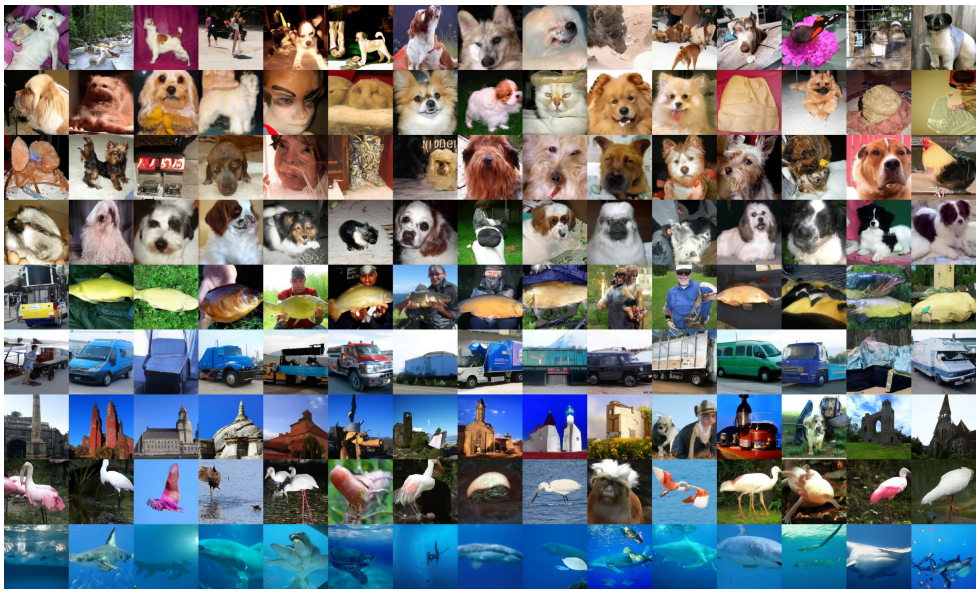


Figure D.12: *Safe Denoiser* generated samples when negating on Chihuahua. The same classes are selected as D.10.