Training-Free Safe Denoisers for Safe Use of Diffusion Models

Mingyu Kim[*1] **Dongjun Kim**[*2] **Amman Yusuf**[1] **Stefano Ermon**[2] **Mijung Park**[1]
[1]CS, UBC [2]CS, Standford
mgyu.kim@ubc.ca, dongjun@stanford.edu
ammany01@cs.ubc.ca, ermon@cs.stanford.edu, mijungp@cs.ubc.ca

Abstract

There is growing concern over the safety of powerful diffusion models, as they are often misused to produce inappropriate, not-safe-for-work 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 retraining the model 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 private data) to avoid specific regions of data distribution, without needing to retrain or fine-tune the model. We formally derive the relationship between the expected denoised samples that are safe and those that are unsafe, leading to our *safe* denoiser, which ensures its final samples are away from the area to be negated. We achieve state-of-the-art safety in large-scale datasets such as the CoPro dataset while enabling significantly more cost-effective sampling than existing methodologies.

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

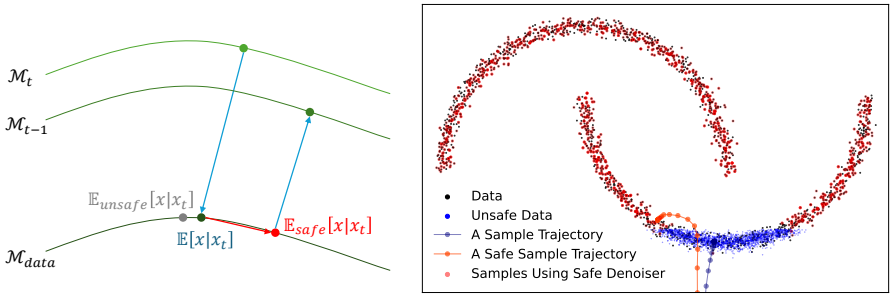
(a) Nudity (Top)

(b) Data Memorization

Figure 1: Our method *Safe Denoiser* against existing methods. (a) Our method, incorporated with SAFREE [1] and SLD [2], does not generate inappropriate images. (b) Our method mitigates the memorization issue by negating the real image, resulting in a novel image with features similar to those in the real image in hair colors or outfits.

[*]Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(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.

1 Introduction

Diffusion models (DMs) have become leading generative models, excelling in generation tasks like text-to-image [3], audio [4], video [5], and protein design [6], thanks to their flexible and controllable sampling [7, 8]. However, growing concerns over unsafe content — such as not-safe-for-work(NSFW) imagery (Figure 1a), copyright violations, and potential misuse — highlight the need for safety. The key challenge is mitigating these risks without compromising model utility or creativity.

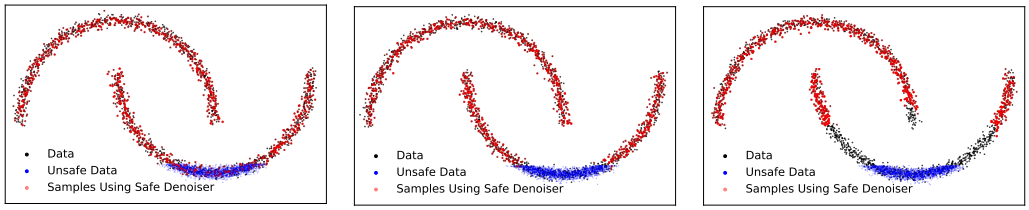
Mainstream mitigation strategies for issues like NSFW content or unwanted concept removal rely on text-based guidance [9, 10, 1] or fine-tuning for unlearning [9, 11, 12, 13]. Text-based methods require iterative, expert-crafted negative prompts [2], which may not generalize well, while fine-tuning is resource-intensive and risks catastrophic forgetting or degraded performance on desired tasks.

Other significant safety concerns involve the DMs' capacity to reproduce copyrighted content and their generation of data pertaining to individuals in Figure 1b who wish to be excluded. These issues are often linked to the models' remarkable ability to memorize training data [14]. While techniques like differentially private training [15, 16] can formally limit memorization by adding noise during the training process, they often result in a noticeable degradation in generation quality, which can be particularly prohibitive for applications demanding high-fidelity outputs.

We propose a *safe denoiser* (defined in Definition 3.1) that modifies sampling trajectories such that the resulting samples are drawn from a safe distribution (shown in Figure 2). The intuion comes from our Theorem. 3.2, where the safe denoiser steers generation away from unsafe regions, ensuring theoretical safety. We develop a practical algorithm (Algorithm 1) based off of our theorem, which can be used standalone or combined with negative prompting to enhance safety in text-to-image generation. Our method achieves state-of-the-art performance on concept erasing, class removal, and unconditional image generation tasks.

2 Preliminary

DMs generate samples through iterative decoding starting from random noise to data. This iterative process is a reverse of the forward data corruption process, $\mathbf{x}_t = \alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$, where $\mathbf{x} \sim p_{\text{data}}(\mathbf{x})$ $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ which results in a perturbation kernel: $q_t(\mathbf{x}_t|\mathbf{x}) = \mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}, \sigma_t^2 I)$. The specific choice of the coefficients α_t and σ_t determines a different variant of DMs: popular examples include Denoising Diffusion Probabilistic Models (DDPM) [17], Elucidating Diffusion Models (EDM) [18], or Flow Matching [19]. Regardless of whether the model is trained with noise-prediction [17], data-prediction [18], or velocity-prediction [20, 19], these approaches are fundamentally equivalent [21, 22]. This paper adopts the data-prediction framework due to its most intuitive interpretation. In data-prediction, the model approximates the *denoiser* function, defined by



(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).

$\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] := \int \mathbf{x} \frac{p_{\text{data}}(\mathbf{x}) q_t(\mathbf{x}_t|\mathbf{x})}{p_{\text{data},t}(\mathbf{x}_t)} \, \mathrm{d}\mathbf{x} \approx \frac{1}{\alpha_t}(\mathbf{x}_t - \sigma_t \boldsymbol{\epsilon}_{\boldsymbol{\theta}})$, where $p_{\text{data},t}(\mathbf{x}_t)$ is a marginal distribution of diffusion process at t, and $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ is the noise-prediction.

DMs can be guided to produce samples [7, 23] that adhere more closely to a desired condition denoted by $\mathbf{c}$. A common approach in modern DMs is *classifier-free guidance* (CFG) [8]. 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

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

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 trajectory in the *sharpening direction* toward a desired condition $\mathbf{c}$.

When there are unsafe words in the input text prompt, *SAFREE* [1] detects unsafe words (tokens) and modifies the unsafe token embeddings. It filters out undesirable concepts with

$$\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] + \underbrace{\lambda(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \tilde{\mathbf{c}}_+] - \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t])}_{\text{SAFREE}}, \tag{1}$$

where $\tilde{\mathbf{c}}_+$ is a modified prompt embeddings. This altered prompt embedding steers the generation process away from the predefined unsafe concepts.

Another way of negating unsafe concepts is using *negative guidance* [24]. It reverses the CFG gradient direction for an undesired prompt denoted by $\mathbf{c}_-$. Formally, one replaces the standard CFG 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 that we want to avoid.

On the line of negative prompting, *Safe Latent Diffusion* (SLD) [2] introduces a guidance by

$$\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] + \underbrace{\lambda(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_+] - \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t])}_{\text{CFG}} - \underbrace{\mu(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \tilde{\mathbf{c}}_-] - \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t])}_{\text{SLD}}, \tag{2}$$

where $\tilde{\mathbf{c}}_-$ represents a predefined set of unsafe prompts suggested by SLD. Hypothetically, suppose we assume μ was set as λ. In that case, the SLD guidance simplifies to a negative guidance $\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] + \lambda(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_+] - \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \tilde{\mathbf{c}}_-])$. A core difference between SLD and negative guidance is that μ is adaptive, i.e., $\mu = \mu(\mathbf{c}_+, \tilde{\mathbf{c}}_-; \gamma, \lambda)$, depending on $\mathbf{c}_+$ and $\tilde{\mathbf{c}}_-$. This weight is proportional to the norm of the difference between denoisiers: $\|\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_+] - \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \tilde{\mathbf{c}}_-]\|$. A larger norm suggests that the trajectory is likely to be safe, whereas a smaller norm implies potential unsafety.

3 Method

Text-based prompts (like $\mathbf{c}_-$ or $\tilde{\mathbf{c}}_-$) rely on limited, user-selected words and may miss undesired content. To address this, we introduce a method that offers safety guarantees on generated images, which can be combined with existing text-based safety approaches.

3.1 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. These indicator functions are the partition of the unity, resulting in $1 = 1_{\text{safe}}(\mathbf{x}) + 1_{\text{unsafe}}(\mathbf{x})$ for all $\mathbf{x} \in \text{supp}(p_{\text{data}})$. Then, we define the following concepts.

Definition 3.1. The unnormalized density of the 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 diffusion process (at time t) starting from the safe distribution. Analogously, the unnormalized density of the 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 diffusion process (at t) starting from the unsafe distribution.

Our interest is to obtain $\mathbb{E}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t]$ given the data denoiser $\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t]$. The theorem below describes the relationship between our safe denoiser and the data denoiser. The proof is given in Appendix A.

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,

$$\mathbb{E}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t] = \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] + \beta^*(\mathbf{x}_t)\big(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] - \mathbb{E}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t]\big) \tag{4}$$

for a weight defined by

$$\beta^*(\mathbf{x}_t) = \frac{Z_{\text{unsafe}} p_{\text{unsafe},t}(\mathbf{x}_t)}{Z_{\text{safe}} p_{\text{safe},t}(\mathbf{x}_t)}, \tag{5}$$

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 unnormalized densities of safe and unsafe distributions, respectively.

The theorem above suggests that a safe denoiser can be constructed similarly to CFG. In our case, the denoiser is penalized by $\beta^*(\mathbf{x}_t)$, designed to increase when $\mathbf{x}_t$ is likely unsafe. Specifically, a term in the numerator, $p_{\text{unsafe},t}(\mathbf{x}_t) = \int p_{\text{unsafe}}(\mathbf{x})q_t(\mathbf{x}_t|\mathbf{x}) \, \mathrm{d}\mathbf{x}$, 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 $\beta^*(\mathbf{x}_t)$ 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 3a, 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 3b demonstrates that by doubling the safety threshold, both the unsafe region and its immediate surroundings are effectively avoided. However, in Figure 3c, 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. In this section, we describe our approach to approximating these quantities.

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 $\tilde{\mathbf{c}}_-$

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:
    Compute $\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}]$ (e.g., $\mathbf{c} \in \{\tilde{\mathbf{c}}_+\}$ for SAFREE or $\mathbf{c} \in \{\mathbf{c}_+, \tilde{\mathbf{c}}_-\}$ for SLD)
  $\beta(\mathbf{x}_t) \leftarrow \frac{1}{N}\sum_{n=1}^{N} p_{0t}(\mathbf{x}_t|\mathbf{x})$ 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
    Compute $\mathbf{x}_{0|t}$ (e.g., Eq. (8) for SAFREE or Eq. (9) for SLD)
  Else:
    $\mathbf{x}_{0|t} \leftarrow \hat{\mathbb{E}}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t]$ (see Eq. (7))
  $\mathbf{x}_{t-1} = \text{Solver}(\mathbf{x}_t, t, \mathbf{x}_{0|t})$
end for

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

$$\hat{\mathbb{E}}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t] = \sum_{n=1}^{N} \mathbf{x}^{(n)} \frac{q_t(\mathbf{x}_t|\mathbf{x}^{(n)})}{\sum_{m=1}^{N} q_t(\mathbf{x}_t|\mathbf{x}^{(m)})}. \tag{6}$$

Each numerator and denominator terms of Eq. (6) 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 mixture of unsafe data points.

Approximation of the weight. Next, we turn our attention to the computation of $\beta^*(\mathbf{x}_t)$ in Eq. (5). 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[2] to evaluate at every sampling steps. To address this challenge, we approximate β^* as

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

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. (5). 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.

[2]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 [3] are trained with billions of training data [25], and is infeasible to iterate the entire data at inference time.

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

$$\hat{\mathbb{E}}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t] = \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] + \eta\beta(\mathbf{x}_t)(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] - \hat{\mathbb{E}}_{\text{unsafe}}[\mathbf{x}|\mathbf{x}_t]), \tag{7}$$

where $\hat{\mathbb{E}}$ is given in Eq. (6). Our results in Sec. 5 validate the effectiveness of our approximations in ensuring sample safety without incurring prohibitive computational costs.

3.3 Extending Safe Denoiser to Text-to-Image generation

While our methodology is effective as a standalone algorithm, we can also integrate it straightforwardly as a plug-in component into established text-based safety mechanisms, thereby enhancing the overall safety level, as shown in Table 1 [3]. For example, when our approach is combined with SAFREE, the predicted clean sample $\mathbf{x}_{0|t}$ (representing the estimated data at step $t = 0$ given a sample $\mathbf{x}_t$ at step t) can be computed by

$$\mathbf{x}_{0|t} = \mathbb{E}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t] + \underbrace{\lambda(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \tilde{\mathbf{c}}_+] - \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t])}_{\text{SAFREE}}. \tag{8}$$

Table 1: Joint effect of existing text-based guidance (SAFREE) and ours. We evaluate the attack success rate. Both "No" with 0.962 refers to SD-v1.4 [3] with CFG. The lower, the better.

		Neg. Prompt	
		No	Yes
Ours	No	0.962	0.601
	Yes	0.633	**0.469**

When it is combined with SLD, the formula is as follows:

$$\mathbf{x}_{0|t} = \mathbb{E}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t] + \underbrace{\lambda(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \mathbf{c}_+] - \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t])}_{\text{CFG}} - \underbrace{\mu(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t, \tilde{\mathbf{c}}_-] - \mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t])}_{\text{SLD}}. \tag{9}$$

Note these Eq. (8) and Eq. (9) replaces the data denoiser $\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t]$ by the safe denoiser $\mathbb{E}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t]$, compared to Eq. (1) and Eq. (2), respectively. In implementation, as described in Sec. 3.2, we approximate the safe denoiser by Eq. (7). In diffusion sampling, we utilize this safe $\mathbf{x}_{0|t}$ in either DDPM [17] or DDIM [27], 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.

4 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 [28, 29, 30, 31, 11, 32, 12, 32]. In more recent work, training-free and text-based methods have also emerged as computationally efficient alternatives [2, 1, 10, 33]. 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 [34, 35, 36, 37]. 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.

A closely related recent work, *Sparse Repellency* (SR) [38], is a training-free technique that modifies the denoising trajectory to avoid unsafe images. Their denoiser follows $\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] + \sum_{n=1}^{N} \text{ReLU}\left(\frac{r}{\|\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] - \mathbf{x}^{(n)}\|} - 1\right) \times (\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t] - \mathbf{x}^{(n)})$. ReLU activation 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,

[3] We tested on MMA-Diffusion [26] nudity prompts and measure the rate the model generates unsafe images. See Section 5 and Table 2 for further details.

$\text{ReLU}\left(\frac{r}{\|\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t]-\mathbf{x}^{(n)}\|}-1\right)(\mathbb{E}_{\text{data}}[\mathbf{x}|\mathbf{x}_t]-\mathbf{x}^{(n)})$ resembles the second term, $\mathbb{E}_{\text{safe}}[\mathbf{x}|\mathbf{x}_t]$, in Eq. (4) if the ReLU activation is comparable to our β^*. From this point of view, our method can be regarded as a generalization of SR. However, unlike our method, SR does not guarantee that the samples are from a safe distribution.

Diffusion Soup [39] presents a related theoretical analysis by merging DMs trained on separate data subsets, but requires fine-tuning. In contrast, our method is training-free and formally defines safe and unsafe denoisers and their relationship. *Dynamic Negative Guidance* (DNG) [40] also uses a similar framework but relies on sequential computation based on a Markov chain and requires extra training for unsafe denoisers, whereas our approach estimates safe and unsafe denoisers in an expectation manner without training overhead.

5 Experiments

We present the experimental results of our method, Safe Denoiser. Section 5.1 details the outcomes of our text-to-image generation experiments, while the subsequent section explores both class-conditional and unconditional image generation.

5.1 Text-to-Image Generation

In this section, we conduct an in-depth analysis of safety issues in text-to-image models, focusing on tasks involving nudity and inappropriate content. The nudity task evaluates how well safety methods prevent harmful outputs as attack difficulty gradually increases. In contrast, the inappropriate content task examines whether these safety methods remain effective when handling multiple concepts simultaneously. We use *Stable Diffusion* (SD) [3] v1.4[4] with DDPM sampler. To evaluate safety, we follow previous studies by assessing Attack Success Rate (ASR), Toxic Rate (TR), and Inappropriate Probability (IP) [2, 1]. We measure ASR by the proportion of generated images that exceeds 0.6 nude class probability, measured by NudeNet[5]. The TR is computed by the average of nude class probability, measured also by NudeNet. The IP is the classification probability score of generating inappropriate images, measured by the Q16 classifier [41]. For the nudity task, we select 515 unsafe images from I2P [2] that exceeds 0.6 nude class probability. For the inappropriate content tasks, we randomly sample 3,000 images from I2P as the unsafe dataset. To evaluate, we use the broder dataset CoPro [42], which covers the same categories of I2P. Notably, all experiments uses the identical unsafe datasets across all baselines for consistency, see Appendix C for details.

Besides safety-related metrics, we prioritize maintaining high image quality and prompt alignment simultaneously. To this end, we calculate Fréchet Inception Distance (FID) [43] for the generation fidelity and CLIP [44] to measure whether the samples follow human instructions. We use a PyTorch package [45] to compute the FID by comparing 10K reference images selected from the COCO-2014 [46] validation split and 10K generated images from the prompts identically selected from the same COCO dataset. Also, we evaluate the CLIP score [44] using ViT-B-32[6].

Safe Generation against Nudity Prompts Table 2 summarizes our experimental findings. In these experiments, we utilize unsafe prompts proposed by Ring-A-Bell [37] (79 prompts), UnlearnDiff [36] (116 sexual prompts), and MMA-Diffusion [26] (1000 prompts). These prompts are adversarially generated to fool the existing generative models. For baseline comparisons, we consider both training-based approaches, specifically *ESD* [9] and *RECE* [12], and training-free methods such as SLD [2] and SAFREE [1]. Initially, we observe that about 96.2% of generated SD-v1.4 images are unsafe when using MMA-Diffusion prompts. Existing baselines demonstrate performance improvements over SD across datasets.

Our method, combined with SLD or SAFREE, 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 (88.1% generated images are unsafe) compared to their performance on other prompt datasets such as Ring-A-Bell or UnlearnDiff. This discrepancy

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

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

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

Table 2: Performance comparison of baselines on various datasets in safe generation against nudity prompts. 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

Table 3: Performance of inappropriate probability (IP) and CLIP Score on the CoPro dataset. Our method incoporating with negative prompts enhances safety performance even across multiple concepts simultaneously.

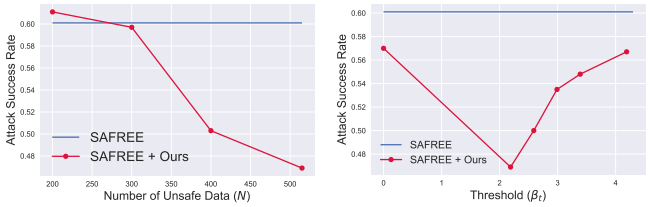
Method	Harra-sment ↓	Hate ↓	Illegal Activity ↓	Self-harm ↓	Sexual ↓	Shock-ing ↓	Viole-nce ↓	Avg. IP ↓	CLIP ↑
SD-v1.4	0.269	0.154	0.206	0.319	0.120	0.221	0.274	0.223	29.81
+ Ours	0.206	0.148	0.197	0.209	0.109	0.209	0.230	0.187	29.21
SLD	0.223	0.106	0.161	0.247	0.078	0.158	0.217	0.170	29.65
+ Ours	0.168	0.113	0.152	0.169	0.078	0.165	0.212	0.151	28.95
SAFREE	0.182	0.118	0.144	0.183	0.085	0.150	0.206	0.153	28.91
+ Ours	0.156	0.112	0.161	0.153	0.083	0.159	0.185	0.144	28.49

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 safety methods to erase such concepts. In contrast, our approach employs purely image-based guidance, which results in substantial performance gains from 88.1% to 48.1% in ASR on MMA-Diffusion when combined with existing text-based methods. Our method significantly improves the performance across all other prompt datasets, not limited to MMA-Diffusion.

Inappropriate Probability in CoPro Dataset Table 3 presents that our method consistently achieves enhancement of safe content generation against multiple categories while maintaining a balance in textual prompt alignments across all baselines. Since training-based approaches do not provide official checkpoints for this task, we focus on training-free approaches. Overall, our method effectively improves inappropriate probability (IP) on CoPro dataset. Notably, all baselines show a reduction in average IP when combined with our method. Additionally, our method effectively preserves the alignment between human instructions (prompts) and generated images, with any introduced misalignment being minimal, as demonstrated by CLIP scores. Furthurmore, our method performs on-par with previous methods in terms of the sample-wise aesthetic scores, showing that there is a minimal impact in the sample quality by applying our method, see Appendix E. These results suggest that our method effectively manages multiple concepts simultaneously while reliably generations away from unsafe content.

Ablation Studies We present a pair of ablation studies to evaluate the robustness and effectiveness of our method. First, Figure 4a 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.

We then explore the influence of the threshold parameter β_t (see Algorithm 1), which governs the application of the safe denoiser. For simplicity, we fixed β_t across all time steps. Figure 4b 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



(a) Effect on N (b) Effect on β_t

Figure 5: Ablation studies of (a) the effect on the number of unsafe data (N), (b) the effect on the threshold (β_t).

Table 4: Experiments of the class negation on ImageNet. Top-1* is the classification accuracy of the generated samples on the negated class (Chihuahua). Refer Appendix D and E.

Method	Prec ↑	Rec ↑	Top-1* ↓
Baseline	0.72	0.63	0.68
B + SR	0.59	0.54	0.00
B + Ours	0.62	0.58	0.00

efficacy and computational efficiency. Additional ablation studies are presented in Appendix to discuss in-depth analysis of the scalability, robustness, and effectiveness of our methods.

Computation Overhead Table 5 presents the wall-clock time for image generation on NVIDIA RTX4090 with 24GB memory. Thanks to GPU parallelism, the additional time introduced by our method scales sub-linearly since modern GPUs optimize batched matrix multiplications with efficient job scheduling. For example, our method increases from 4.22s to 4.29s when using $3,000$ negative images (an overhead of only 0.07s), while increasing from 4.22s to 4.24s when using 515 images. Conversely, SAFREE adds over 1s per image. Given that SAFREE and ours perform similarly in Table 1, our method shows a better performance-efficiency curve.

Table 5: Wall-clock time.

Models	Time (s/img)
SD-v1.4	3.18
+ Ours ($N = 515$)	3.20
SAFREE	4.22
+ Ours ($N = 515$)	4.24
+ Ours ($N = 3,000$)	4.29

5.2 Class-Conditional and Unconditional Generation with Safe Denoiser

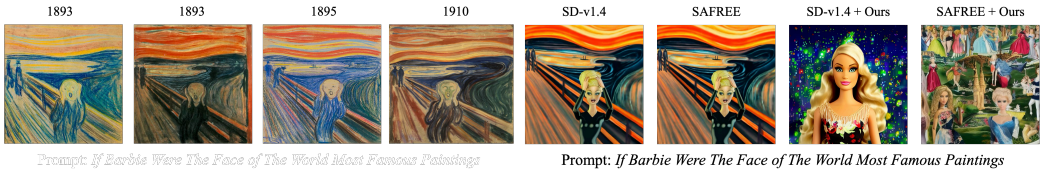
This subsection evaluates the performance of safe denoiser when applied in isolation. Specifically, to assess our safe denoiser in a simplified setting, we conduct experiments on two distinct tasks: a class-conditional model trained on ImageNet [48] for removing a specific class (e.g., Chihuahua); and on an unconditional model trained on FFHQ [49] to negate generating specific sex (e.g., female), see Appendix D for details of experiments. For the class removal, Table 4 presents precision, recall [50], and Top-1* (classification accuracy of generated images conditioned by Chihuahua class) metrics for negating the Chihuahua class. As conventional text-based safety techniques are not directly applicable, we compare our method against Sparse Repellency (SR), as described in Sec. 4. Table 4 showcases that our method outperforms SR in terms of precision, recall, and Top-1*, indicating that ours avoid generating Chihuahua while being more diverse and precise than SR. In the FFHQ experiments, where we targeted the negation of female images, Table 6 indicates that while SR exhibits classification results that deceptively suggest successful negation, the FID scores and qualitative comparisons in Appendix E demonstrate that this apparent achievement comes at the cost of significantly degraded image quality. Indeed, our experiments show that our methodology consistently produces visually convincing samples, whereas SR frequently generates out-of-distribution images with artifacts.

Table 6: Performance in FFHQ. We use ResNet18 [47] to classify the sex of generated samples. We compute FID by comparing male data and 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

5.3 Compatibility with Frontier Model and Style-Level Intellectual Property Control

We evaluate the compatibility of our plug-and-play approach with the frontier model SD-v3 [51]. The experimental results are presented in Table 7. On SD-v3, SAFREE alone reduces ASR by approximately 9% compared to the baseline. In contrast, our approach achieves a 33.2% relative reduction in ASR, from 0.304 to 0.203. Notably, our method maintains CLIP alignment and even slightly improves FID. This demonstrates the applicability of our proposed method to recent and powerful backbones models.



(a) Negative datapoints

(b) Generated Images

Figure 6: Qualitative result for style-level intellectual property control. SD-v1.4 reproduces Munch's style, whereas Ours with and without SAFREE removes that style while preserving the "Barbie" concept. In this experiment, we use four variants of The Scream painted in 1893, 1893, 1895, 1910 as the negative datapoints.

Interestingly, our safe denoiser enhances sample diversity during inference, which can lead to a reduction in FID. A well-known phenomenon of large CFG values is a fidelity and diversity trade-off. Specifically, increasing CFG sharpens alignment but diminishes sample diversity, resulting in a degradation of FID at high values. This phenomenon has been observed in previous studies [52, 53]. In contrast, our safe denoiser is not overly reliant on the text conditioning, allowing it to introduce relevant stochasticity that effectively mitigates the loss of diversity caused by high CFG. Consequently, our denoiser improves FID. Empirically, we have observed higher intra-prompt diversity compared to the baseline. Another perspective to consider is that FID's Gaussian approximation of feature distributions possibly records small improvements that does not translate into noticeable quality differences in practical applications.

Table 7: SD-v3 results on Ring-A-Bell for safety and COCO-30K for image quality.

Method	Ring-A-Bell		COCO-30K	
	ASR ↓	TR ↓	FID ↓	CLIP ↑
SD-v3	0.304	0.330	23.15	**31.46**
+ SAFREE	0.278	0.298	22.99	31.24
+ Ours	**0.203**	**0.267**	**22.54**	31.15

We conceptually evaluate baselines in situations where pretrained diffusion models leak intellectual property. In this scenario, intellectual property-sensitive prompts can be grouped into three scenarios: *(i) the prompt explicitly names the target intellectual property; (ii) the prompt avoids the name but gives a detailed textual description; and (iii) neither name nor descriptive cues are present, yet the model reproduces the target's style.* The third case presents a significant challenge for text-only defenses, as there is no negative text cue to negate. As reported by [54], diffusion models can overfit styles and reproduce them even without explicit textual mentions. We reproduce this phenomenon with Munch's The Scream by using the prompt "If Barbie were the face of the world's most famous paintings". While this text prompt never mentions Munch or The Scream, SD-v1.4 recreates the painting's distinctive style as shown in Figure 6b. When we use four original paintings of "The Scream", for instance two from 1893, one from 1895, one from 1910, as the negative set, our safe denoiser suppresses Munch's style while preserving the Barbie concept. Additionally, Ours with SAFREE produces both modern and classical renderings without the style of Munch portraits. The qualitative results are displayed in Figure 6.

6 Limitations and Conclusions

We introduce the *safe denoiser*, an in-process, training-free mechanism that steers diffusion model sampling toward theoretically safe distributions, thereby promoting appropriate content. Unlike purely discriminative pre- or post-filters, our approach acts during inference and complements existing guardrails. In particular, this mid-generation intervention mitigates failure modes in static text or image filters, especially under adversarial prompt engineering. Thus ours contributes to a defense-in-depth safety architecture suitable for real-world applications. Regarding negative datasets, the data requirement is shared across other defenses method. The datasets used to train or calibrate pre- and post-filters can be reused to provide data-driven negative guidance at inference. A current limitation is the need to tune hyperparameters to balance fidelity and safety. Appendix B discusses these trade-offs and offers practical guidance. Privacy-sensitive generation remains an ongoing challenge, partly due to the lack of standardized quantitative metrics. We leave the development of such metrics and extensions to other modalities for future work.

Acknowledgments

We thank our anonymous reviewers for their constructive feedback, which has helped significantly improve our paper. We thank the Digital Research Alliance of Canada (Compute Canada) for its computational resources and services. M. Kim was supported by the Canada CIFAR AI Safety Catalyst grant. A. Yusuf was funded by the Canada Graduate Scholarships — Master's program of the Natural Sciences and Engineering Research Council of Canada (NSERC). M. Park was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canada CIFAR AI Chairs program.

References

[1] 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.

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

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022.

[4] 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.

[5] 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.

[6] 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.

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794, 2021.

[8] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[9] 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*, 2023.

[10] 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*, pages 190–206, 2024.

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

[12] 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*, pages 73–88. Springer, 2024.

[13] 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*, pages 461–478. Springer, 2024.

[14] 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)*, pages 5253–5270, 2023.

[15] Tim Dockhorn, Tianshi Cao, Arash Vahdat, and Karsten Kreis. Differentially private diffusion models. *Transactions on Machine Learning Research*, 2023.

[16] Michael F Liu, Saiyue Lyu, Margarita Vinaroz, and Mijung Park. Differentially private latent diffusion models. *Transactions on Machine Learning Research*, 2024.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[18] 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, pages 26565–26577, 2022.

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

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

[21] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in Neural Information Processing Systems*, 34:21696–21707, 2021.

[22] 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.

[23] 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.

[24] 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*, pages 423–439. Springer, 2022.

[25] 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.

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

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

[28] 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, pages 17170–17194, 2023.

[29] 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.

[30] 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.

[31] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: learning to forget in text-to-image diffusion models. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1755–1764, 2024.

[32] 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*, pages 6430–6440, 2024.

[33] 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.

[34] 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.

[35] 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.

[36] 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*, pages 385–403. Springer, 2024.

[37] 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.

[38] 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.

[39] 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.

[40] Felix Koulischer, Johannes Deleu, Gabriel Raya, Thomas Demeester, and Luca Ambrogioni. Dynamic negative guidance of diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[41] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1350–1361, 2022.

[42] Runtao Liu, Ashkan Khakzar, Jindong Gu, Qifeng Chen, Philip Torr, and Fabio Pizzati. Latent guard: a safety framework for text-to-image generation. In *European Conference on Computer Vision*, pages 93–109. Springer, 2024.

[43] 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.

[44] 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*, pages 8748–8763. PMLR, 2021.

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

[46] 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*, pages 740–755. Springer, 2014.

[47] 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*, pages 770–778, 2016.

[48] 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.

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

[50] 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.

[51] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[52] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *The Twelfth International Conference on Learning Representations*, 2024.

[53] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *arxiv preprint arxiv:2404.07724*, 2024.

[54] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023.

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

[56] Runtao Liu, Chen I Chieh, Jindong Gu, Jipeng Zhang, Renjie Pi, Qifeng Chen, Philip Torr, Ashkan Khakzar, and Fabio Pizzati. Safetydpo: Scalable safety alignment for text-to-image generation. *arXiv preprint arXiv:2412.10493*, 2024.

[57] 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.

[58] 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.

[59] 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.

[60] 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*, pages 770–778, 2016.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our claim matches theoretical and experimental results, and reflect how effective the proposed method can address safety issues in generative models.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We create a "Limitations and Conclusions" section to cover both contents in the main text. We also create a separate "Limitations and Broader Impacts" section in Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Although we omit some of assumptions in the main paper mainly due to page limit, we provide full details of assumptions and complete proof in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We disclose all experimental details in the main paper and Appendix including the hyperaparameters and datasets used. For reproducibility, we plan to release our code upon acceptance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: In the review process, we release our code to the reviewers to regenerate our experimental results. After the acceptance, we plan to release the code to the public.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We faithfully release our hyperparameters and experimental details in Appendix and the main text.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We have not reported error bars mainly due to the lack of computational resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We explain which resources we used for experiments in both the main text and appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We faithfully follow the code of ethics, suggested by the link above.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts as a separate section in the "Limitations and Broader Impacts" in Appendix

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We used the HuggingFace library for checkpoints and adversarial attack datasets. They have requested users to enroll and have managed the user lists. This paper focuses on safety issues in generative models, which aligns with the concern.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have properly credited the original owners of assets by citing them. In the code release, we comply the license and terms of the assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We include the details of the dataset, code, and model in either footnotes or Appendix.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We follow LLM policy of NeurIPS2025. We ensure that LLM has been used only for editing and formatting manuscripts.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

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 Limitations and Broader Impacts

Limitations We have addressed significant safety challenges in DMs, particularly concerning the generation of NSFW content and the inadvertent reproduction of sensitive data. 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.

However, this approach necessitates the introduction of an additional hyperparameter, β_t, as outlined in Theorem. 3.2. While we demonstrate that this parameter is theoretically derived and straightforward to implement, it may not be optimal for realistic scenarios due to its assumption of access to numerous data points sampled from an unsafe distribution. In practice, we present evidence in Figure 4b that this parameter influences the performance of the model.

Despite the challenges, we have developed a novel training-free method that effectively guides the sampling trajectories of DMs towards safe distributions. Ultimately, this work provides a robust and scalable solution for mitigating safety risks in generative AI, paving a way for their responsible and ethical applications.

Broader Impacts This paper presents a work whose goal is to build a reliable and trustworthy Generative AI. There are many potential societal consequences of our work, particularly in addressing ethical risks associated with generative models. Our research is focused on preventing the generation of NSFW content, including nudity and violence, and mitigating the risk of models memorizing and reproducing private information, such as human face, from training datasets. We believe the presented work contributes to the responsible use of generative AI, reinforcing ethical safeguards and promoting AI systems that align with societal values and human rights.

C Experimental Details : Text-to-Image Generation

As outlined in the manuscript, we conduct the Text-to-Image experiment using SD-v1.4, following the same model as the baselines for generating images from text, as referenced in [2, 55, 12, 1]. 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 [17]. 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.

Regarding the *Safe Denoiser*, the proposed model computes the transition kernel with an RBF kernel. The RBF kernel function is defined as follows:

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \tag{C.1}$$

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. (7). 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.

For reference images, we provide a detailed explanation of how they are obtained. To ensure safe generation against nudity prompts, we utilize a total of 515 images sourced from the I2P dataset [2]. These images were generated using SD-v1.4. As mentioned in the manuscript, these reference images meet the criterion, where a nude class probability exceeds 0.6, as determined by Nudenet. Sample images are presented in Figure C.1. On the other hand, for the inappropriate probability task with the CoPro dataset, we attempt to use the total images from the I2P dataset. However, our computational resources allow us to use only 3,000 reference images. To select these 3,000 images, we randomly choose them out of the 4,703 images available in the I2P dataset. All images used in this task are also generated using SD-v1.4. Sample images are presented in Figure C.2. It's important to note that all experiments conducted in this study use the same set of reference images across all baselines. This ensures a fair comparison.

Addtionally, the reference images we used in the task to generate safer images against nudity prompts are included as attachments in our supplementary materials. On the other hand, due to space constraints, we cannot include the attachment in the inappropriate probability task. We ensure that the reference images used in this task will be included in the public repository upon the acceptance of the paper.

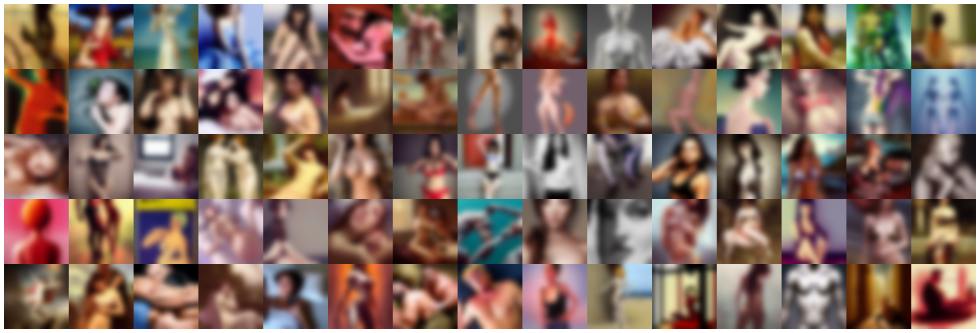


Figure C.1: Reference images for safe generation against nudity prompts

Figure C.2: Reference images for inappropriate propability

Next, we briefly introduce the baseline models used in our experiments. The first two approaches serve as comparisons for unlearning-based safe diffusion models [9, 12]. Specifically, we evaluate Erased Stable Diffusion (ESD) [9] as a representative method. More recently, reliably trained safe diffusion (RECE) models have demonstrated improved performance, particularly in reducing the attack success rate [12]. In addition to these unlearning-based approaches, we also include SLD and SAFREE as training-free safe diffusion models [2, 1]. 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 [2]. In contrast, SAFREE focuses on more precise negative prompts specifically tailored to nudity-related content [1]. 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.

C.1 Inappropriate Prompt Datasets

I2P The I2P dataset consists of prompts related to seven unsafe concepts: hate, harassment, violence, self-harm, sexual content, shocking content, and illegal activity [2]. It contains a total of 4,703 prompts and was introduced in earlier stages of research, with subsequent studies primarily focusing on this dataset [12, 1]. 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`

CoPro Compared to I2P [2], the CoPro dataset offers a more extensive dataset comprising a total of 226,104 prompts, each associated with 723 concepts that span both safe and unsafe scenarios. This expansion enhances the dataset's suitability for rigorous evaluation [42]. Particularly, it also offers super-concept information, following the same framework of I2P [2]. All text prompts are categorized into {hate, harassment, violence, self-harm, sexual content, shocking content, and illegal activities}. This ensures that they align with the corresponding category information in the I2P dataset. To efficiently evaluate models, we randomly sample 10,000 prompts, ensuring a uniform distribution across all categories. We validated that the average inappropriate probability of SD-v1.4 in the randomly sampled dataset, presented in Table 3, closely matches the numerical information provided in [56]. In this work, we evaluate safe image generation performance across baselines on the CoPro dataset using reference data points from the I2P dataset. This dataset was obtained from `https://github.com/rt219/LatentGuard/blob/main/dataset/CoPro_v1.0.json`

C.2 Nudity in NSFW Prompt Datasets

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 [37]. The original dataset `Chia15/RingABell-Nudity` contains 285 prompts; however, we use a curated subset of 79 prompts, following prior baselines [12, 1]. 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 [26]. 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 [12, 1]. 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 [36]. 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`.

C.3 Ann Graham Lotz for Data Memorization

In Figure 1b, we demonstrate that SD-v1.4 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 Figure C.3

Figure C.3: Reference images for Ann Graham Lotz case

D Experimental Details : Class-Conditional and Unconditional Generation

In this section, we use our safe denoiser in the DMs without text inputs. Specifically, we employ experiments on FFHQ [49] and ImageNet [48] in the 256×256 resolution. We utilize the pretrained diffusion models from FFHQ [57][7] and ImageNet [7][8]. For the experiments, we use a DDPM solver [58] with 100 steps.

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 6 illustrates that the generated images exibit a noticiable bias toward female images over male ones. In this experiment, we select 1K female images from CelebA-HQ [59][9] validation split to serve as unseen negative data, thereby establishing the negative dataset $\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(1000)}\}$. Then, we employ our safe denoiser to generate 1K images. While both FFHQ and CelebA-HQ 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 [43] score using 1,000 male images from the CelebA-HQ dataset. For classification performance, we train a ResNet18 model, as implemented in the PyTorch framework [10] using the training dataset in the CelebA-HQ. In this experiment, we chose $\sigma = 1.0$ and $\eta = 0.05$, and employ *Safe denoiser* across the entire denoising timesteps.

Conditional Generation For conditional ImageNet [48] experiments at 256×256 resolution, we use a diffusion model trained on the full ImageNet-256 dataset guided by a classifier [7]. The diffusion backbone follows a linear noise schedule and is constructed with 1,000 diffusion time-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 [60]. 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. As illustrated in Table 4, 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 in this experiment. To avoid unintended Chihuahua generation, aforementioned metrics aim to make sure that samples do not drift toward distinct Chihuahua-like features. For instance, when we generate an image based on a reference dataset sampled from a Chihuahua, the resulting sample may resemble a Golden Retriever, but it won't resemble a Chihuahua.

To compare our approach, we implement three variants of the conditional diffusion process: vanilla classifier-guided diffusion model without repellency mechanisms, the Sparse Repellency (SR) [38] technique applied to the classifier-guided diffusion model, and our safe denoiser technique applied to the same diffusion process. For the reference dataset, we select the validation set of Chihuahua class as the negative images. In this experiment, the safe denoiser technique is applied on the 200 to 800 timesteps of the diffusion process. $\eta = 0.02$ 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.

[7] https://github.com/DPS2022/diffusion-posterior-sampling

[8] https://github.com/openai/guided-diffusion

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

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

E Additional Experimental Results

In this section, we share extra experimental results. Both numeric and visual results are included, which are not presented in the main text. These results highlight the empirical gains in terms of safe generation and the preservation of the global context of the generated samples simultaneously. Specifically, this ensures that the samples remain faithful to their original meanings while enabling us to negate specific concepts we intended.

E.1 Quantiative Results

ImageNet Case for Negating Chihuahua Class 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 [7] 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 E.1. 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 E.1 further report the precision (sample accuracy) and recall (sample diversity) [50] over 50K samples, indicating that our method is better than SR in negating a specific class.

Table E.1: 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

Aesthetic Scores for Long Text Prompts We identified that our method, which incorporates negative prompts, effectively reduces the risk of generating unsafe data and maintains alignment with the given text prompts. However, the text prompts in these cases span various lengths. Therefore, it is necessary to quantify whether our method excessively applies to remove unsafe contents, leading to unfavorable images in extreme cases. We sample the most complex cases from the I2P datasets and compare the generated images across baselines.

Table E.2: Aesthetic scroes for long text prompts in the I2P dataset.

Method	LAION-aesthetic V2 ↑
SD-v1.4	5.97 ± 0.534
SAFREE	6.03 ± 0.540
SAFREE+Ours	5.94 ± 0.529

To test how our method works when long and complex prompts are given, we use LAION-aesthetic V2 score [11] as a metric and use top 10% longest prompts (475 prompts, avg. word_count=54) selected from the I2P dataset. We choose this score as it is known to be correlated with human perception of quality of images (higher the better). As shown Table E.2, our method maintains aesthetic quality comparable to baselines, even with complex prompts. We prove that the proposed method does not struggle to create appropriate images even when asked with long text prompts.

E.2 Qualitative Results

We present additional qualitative results across three experimental scenarios: *(1) Text-to-Image Generation for preventing nudity and inappropriate images, (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

[11] https://github.com/christophschuhmann/improved-aesthetic-predictor

present the results in sequence, beginning with text-to-image generation followed by unconditional generation and concluding with conditional generation. To facilitate straightforward understanding, we include as many figures and qualitative comparisons as possible.

E.3 Text-to-Image Generation

Safe Generation against Nudity Prompts We present a qualitative comparison across baselines and ours. All figures are generated using the same text prompts. We decide to exclude the case of MMA-Diffusion since prompts in this dataset generate pornographic images by baselines, which is not suitable for academic purpose. Hence, we select text prompts from Ring-A-Bell [37] and UnlearnDiff [36]. From Figure E.4 to Figure E.5, we observe that our model effectively eliminates nudity information while preserving textual information.

Figure E.4: Generated images by baselines and ours on Ring-A-Bell [37]

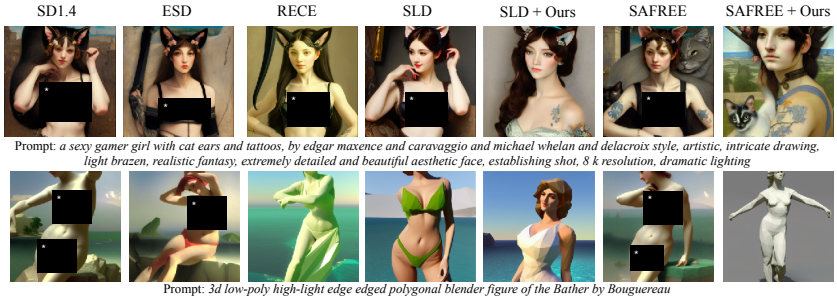


Figure E.5: Generated images by baselines and ours on UnlearnDiff [36]

Inappropriate Probability in CoPro Dataset In our evaluation on the CoPro dataset [42], we apply our method with images from the I2P dataset [2], which includes a wide range of sensitive categories: {hate, harassment, violence, self-harm, sexual content, shocking content, and illegal activities}. Among these, we focus on the 'Self-Harm' category. Self-harm content is suitable for graphical illustration, distinguishing it an appropriate and interpretable case for homogenous visual inspection in the public domain. Unlike other categories—where perceptions of appropriateness can vary widely across cultural and personal contexts—'Self-harm' is typically associated with somber or distressing imagery that is broadly and publicly recognized as unsafe.

As illustrated in Figure E.6, our Safe Denoiser effectively reduces the generation of implicit unsafe content while preserving coherence with the provided prompts. This underscores its ability to both detect and suppress sensitive content without compromising the semantic alignment of textual prompts. From this figure, it is evident that negative prompts do not yield significant results in mitigating the generation of sad and gloomy atmospheres, particularly for conveying feelings of collapse. Conversely, our *Safe Denoiser* tends to generate images that more accurately reflect the literal meanings of the textual prompts. This tendency contributes to a reduction in the likelihood of generating content that evokes feelings of 'Self-harm'.

Alignment of Textual Prompts We present uncurated generated images from the CoCo dataset. This dataset encompasses a wide range of textual prompts that cover various lengthy and diverse

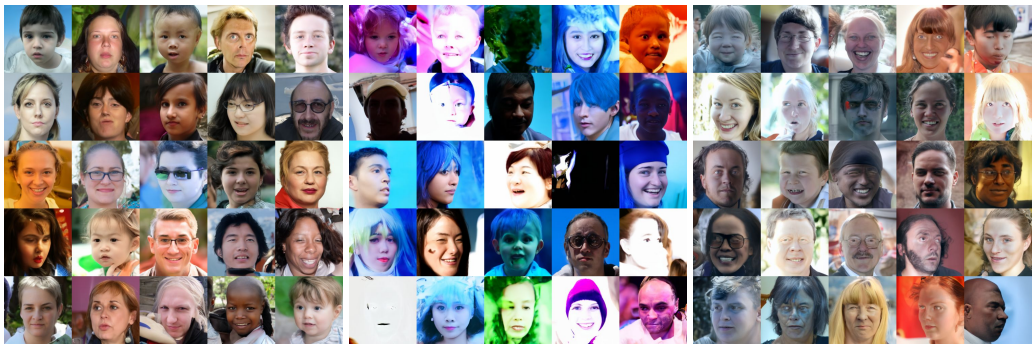
Figure E.6: Generated images by baselines and ours on CoPro [42]. All texual prompts are labeled as 'Self-Harm'. This dataset also provides safe alternatives, and we present both.

situations. As shown in Figure E.7 and Table 2, we conclude that our approach does not compromise the performance of generating normal images. Instead, it focuses on addressing the challenge of generating potentially unsafe images.

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

E.4 Unconditional Generation: Sexual Debiasing

We present uncurated generated images created by the DM trained on the FFHQ dataset [49]. This dataset lacks explicit labels indicating sexual information. However, we observe a tendency for this model to generate female images more frequently than male images, as shown in Table 6. On the other hand, when we utilize *Safe denoiser* with female images, we mitigate the potential bias towards female images and achieves generating images uniformly distributed across sexual information. Figure E.8 illustrates that the generated images align with the numerical results presented in Table 6.



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

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

E.5 Class-conditional Generation : Negation of Specific Class

We present uncurated images that focus on negating a specific Chihuahua class. Here are two experimental setups. First, we employ class conditional guidance on the 'Chihuahua' class and simultaneously use the *Safe denoiser* to work with negative images sampled from the 'Chihuahua' class in the validation split. We observe that the SR does not follow homogeneous images that align with the superclass, 'Dog', but our method produces similar small dogs but not matched with 'Chihuahua' as shown in Figure E.9.

Second, we qualitatively evaluate that our method with negative images from the 'Chihuahua' class works when class guidance is applied to classes other than 'Dogs', for example, 'Tench' and 'Truck'. This result is shown in Figure E.10. We observe that the SR sometimes depicts different classes even when class guidance is applied, but our method aligns with homogeneous classes following class guidance even when the *Safe denoiser* works with 'Chihuahua' images. This indicates that our method effectively tackles specific concepts and preserves the original performance when it is not mutually correlated.

Figure E.9: Generated samples when negating the Chihuahua class, primarily producing visually similar small dog breeds.

Figure E.10: 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.

Additional graphical illustrations are presented in the following figures from Figure E.11 to Figure E.13.

Figure E.11: Classifier guidance diffusion model generated samples when negating on Chihuahua. This comparison includes non-dog-related ImageNet classes mentioned in E.10 along with the dog-related classes in Figure E.9 which are Pomeranian, Yorkshire Terrier, and Shih Tzu.

Sparse Repellency

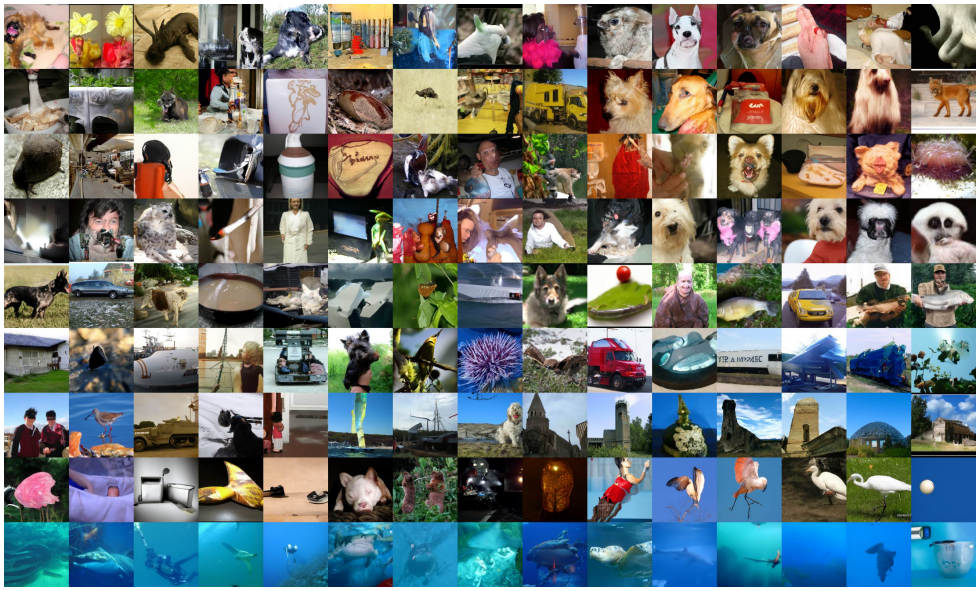


Figure E.12: *Sparse Repellency* generated samples when negating on Chihuahua. The same classes are selected as E.11.

Ours

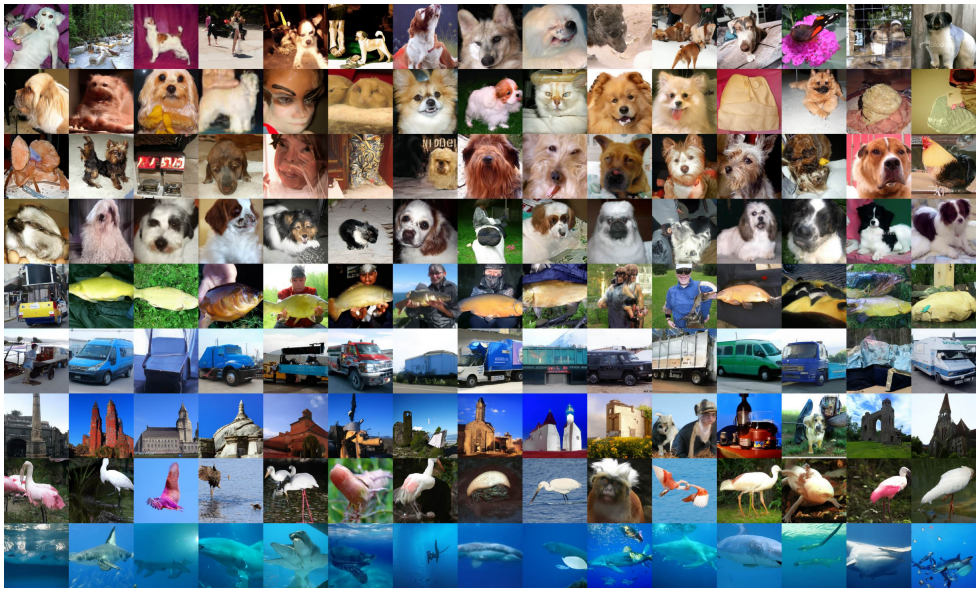


Figure E.13: *Safe Denoiser* generated samples when negating on Chihuahua. The same classes are selected as E.11.