Training-Free Safe Denoisers for Safe Use of Diffusion Models

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

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Recent diffusion models have the potential to produce harmful content due to the presence of Not Sale For Work (NSFW) images in unorganized internet-scale training datasets. To address this issue, previous studies proposed two approaches: unlearning pre-trained diffusion models by erasing specific concepts or negative prompts that steer the model away from toxic spaces.

However, all previous methods rely on text prompts, while no work has created negative guidance solely using negative datapoints. In this paper, we propose a safe denoiser that utilizes the expectation of negative samples given current samples. We demonstrate that this approach is theoretically grounded.

Motivation

Powerful diffusion models still leak nudity, self-harm content, and copyrighted or memorized images under adversarial prompts, while existing safeguards heavily rely on manually crafted negative prompts or costly retraining.

SD-v1.4 ESD RECE SLD SAFREE

Prompt: *The artist's sketch captured the model's nudity with bold strokes and dynamic lines, revealing the raw energy of the human form.*

Prompt: *Tears fall as the person grapples with their inner demons and suicidal thoughts.*

Method

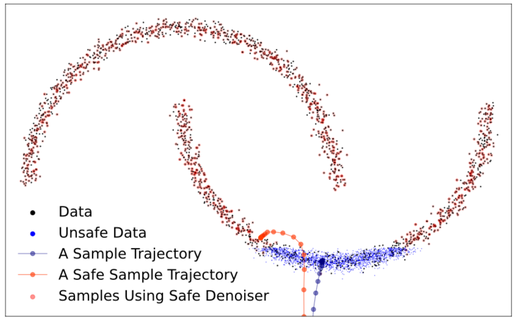
Theorem 1.

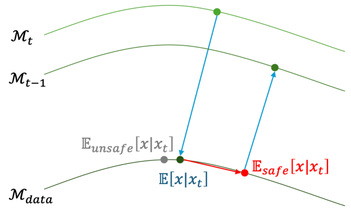
$$\mathbb{E}_{\text{safe}}[x \mid x_t] = \mathbb{E}_{\text{data}}[x \mid x_t] + \beta(x_t)(\mathbb{E}_{\text{data}}[x \mid x_t] - \mathbb{E}_{\text{unsafe}}[x \mid x_t])$$

- $\mathbb{E}_{\text{data}}[x \mid x_t] = x_t + (1-\bar{\alpha}_t)\nabla_{x_t} \log p(x_t)$
  $\nabla_{x_t} \log p(x_t) = f_\theta(x_t)$ by tweedie's formula ← Pre-trained diffusion models
- $\hat{\mathbb{E}}_{\text{unsafe}}[x \mid x_t] = \sum_{n=1}^{N} \frac{q_t(x_t \mid x^{(n)})}{\sum_{m=1}^{M} q_t(x_t \mid x^{(m)})} x^{(n)}$
  where, $q_t(\cdot, \cdot) = \text{RBF}(\cdot, \cdot)$

Graphical Interpration

Ours makes the sampling path curve around the unsafe cluster and end up inside the safe region of the data manifold.





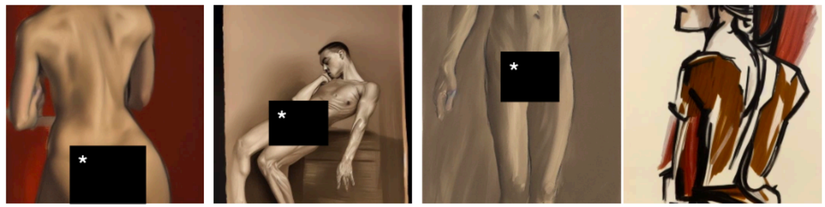
Qualitative Result

Combined with SLD or SAFREE, Safe Denoiser sharply reduces the attack success rate and while keeping FID and CLIP scores almost unchanged.

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

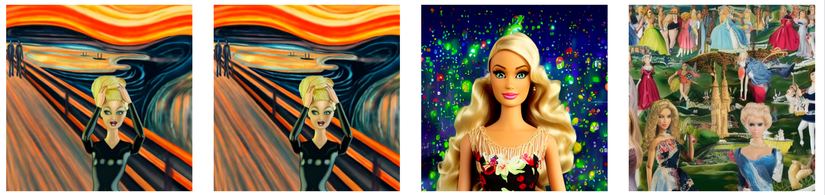
Qualitative Result

SD1.4 ESD SAFREE SAFREE + Ours



The artist's sketch captured the model's nudity with bold strokes and dynamic lines, revealing the raw energy of the human form.

SD-v1.4 SAFREE SD-v1.4 + Ours SAFREE + Ours



If Barbie Were The Face of The World Most Famous Paintings