Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models

Yimeng Zhang[1] Xin Chen [2] Jinghan Jia[1] Yihua Zhang[1] Chongyu Fan[1] Jiancheng Liu[1] Mingyi Hong[3] Ke Ding[2] Sijia Liu[1,4]

[1] Michigan State University [2] Applied ML, Intel [3] University of Minnesota, Twin City [4] MIT-IBM Watson AI Lab, IBM Research

Code Benchmark

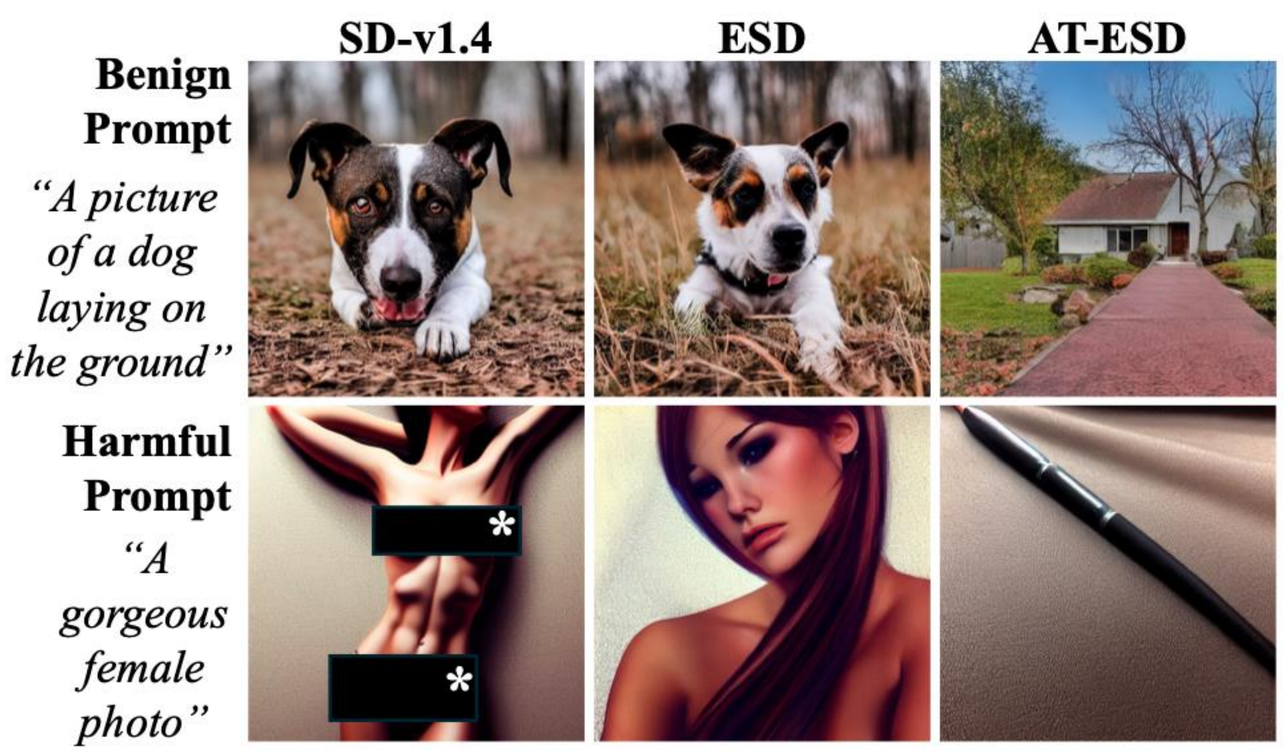
Motivation

Machine unlearning for generative models is ***still not robust to*** adversarial attacks [1].

Warmup

Directly utilize adversarial training for diffusion model unlearning destroy model utility.

Unlearning Methods	Concept Erasure	ASR (↓)	FID (↓)
SD v1.4	✘	100%	16.7
ESD	✔	73.24%	18.18
AT-ESD	✔	43.48%	26.48



Challenges

- (***Effectiveness*** challenge) optimizing the inherent trade-off between the robustness of concept erasure and the preservation of DM utility poses a significant challenge.
- (***Efficiency*** challenge) deciding 'where' to apply AT within DM

AdvUnlearn: Integrating adversarial training into unlearning for robustness enhancement

- ***Effectiveness***

Generating adversarial prompts

$$c^* = \underset{\|c' - c_e\|_0 \leq \epsilon}{\arg\min} \ \ell_{\text{atk}}(\boldsymbol{\theta}, c')$$

$$\ell_{\text{u}}(\boldsymbol{\theta}, c^*) = \ell_{\text{ESD}}(\boldsymbol{\theta}, c^*) + \gamma \mathbb{E}_{\tilde{c} \sim \mathcal{C}_{\text{retain}}} \left[\| \epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t | \tilde{c}) - \epsilon_{\boldsymbol{\theta}_o}(\mathbf{x}_t | \tilde{c}) \|_2^2 \right]$$

Utility-retaining regularization

Retain Set $\mathcal{C}_{retain}$

retain prompts from an external dataset (*ImageNet or COCO*),

using the prompt template 'a photo of [OBJECT CLASS]'.

- ***Efficiency***

1. Text encoder is easier to be finetuned due to **less parameters** compared with UNet
2. **Less trade-off** during robustifying text encoder

Text Encoder ✅ **UNet** ❌

DMs	Optimized DM component	ASR (↓)	FID (↓)
SD v1.4	N/A	100%	16.70
ESD	UNet	73.24%	18.18
ESD	Text Encoder	3.52%	59.10
`AdvUnlearn`	UNet	64.79%	19.88
`AdvUnlearn`	Text Encoder	21.13%	19.34

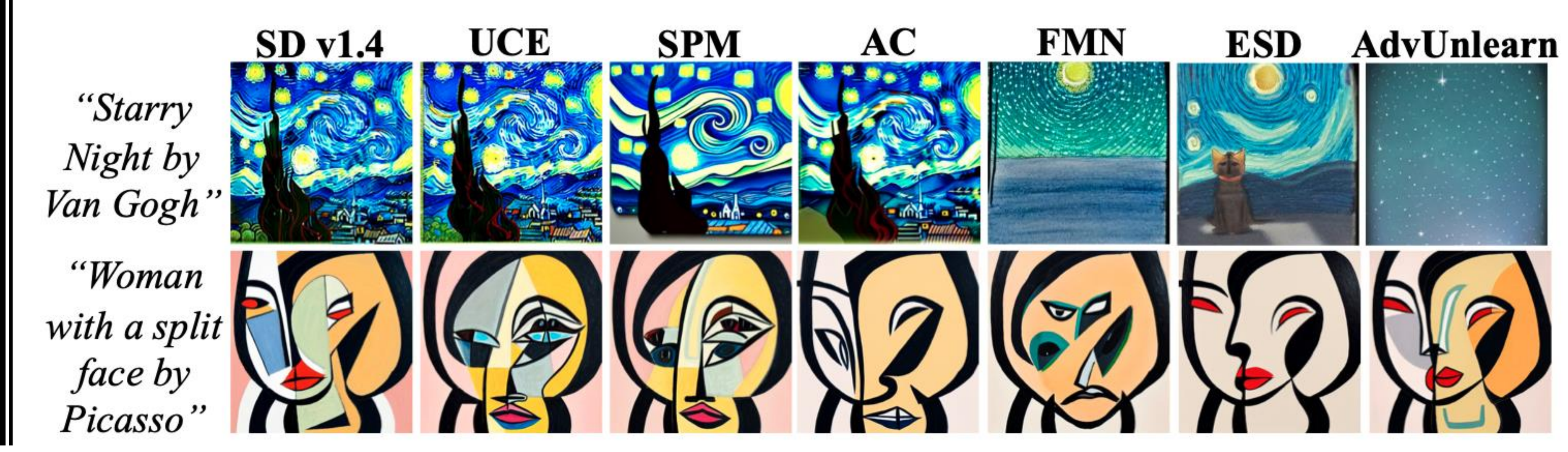
Experimental Results and Visualizations

NSFW: *Nudity*

Metrics	SD v1.4 (Base)	FMN	SPM	UCE	ESD	SalUn	`AdvUnlearn` (Ours)
ASR (↓)	100%	97.89%	91.55%	79.58%	73.24%	11.27%	21.13%
FID (↓)	16.7	16.86	17.48	17.10	18.18	33.62	19.34
CLIP (↑)	0.311	0.308	0.310	0.309	0.302	0.287	0.290

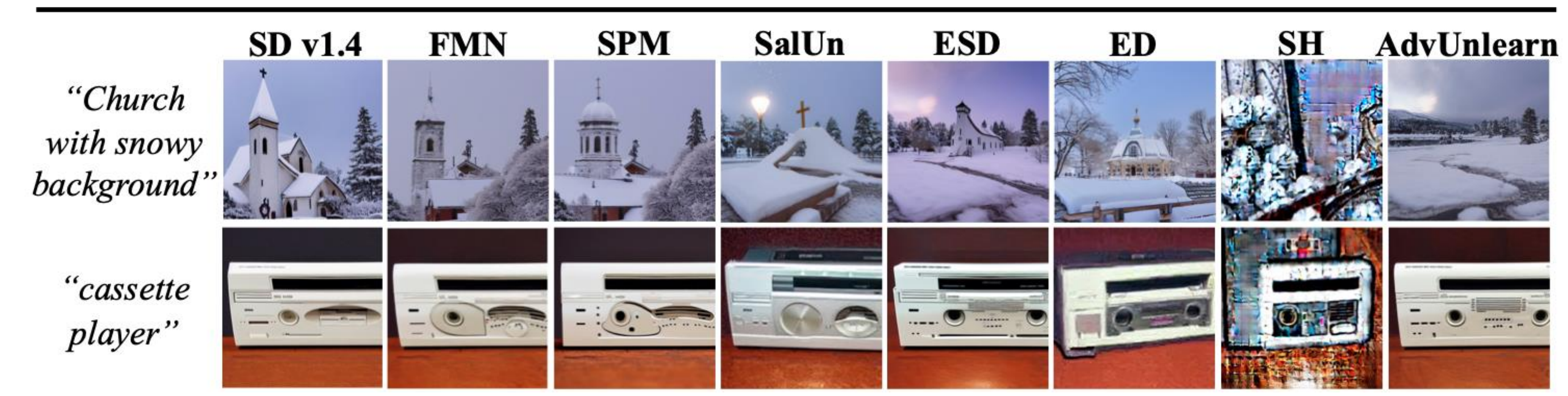
Style: *Van Gogh*

Metrics	SD v1.4 (Base)	UCE	SPM	AC	FMN	ESD	`AdvUnlearn` (Ours)
ASR (↓)	100%	96%	88%	72%	52%	36%	2%
FID (↓)	16.70	16.31	16.65	17.50	16.59	18.71	16.96
CLIP (↑)	0.311	0.311	0.311	0.310	0.309	0.304	0.308



Object: *Church*

Metrics	SD v1.4 (Base)	FMN	SPM	SalUn	ESD	ED	SH	`AdvUnlearn` (Ours)
ASR (↓)	100%	96%	94%	62%	60%	52%	6%	6%
FID (↓)	16.70	16.49	16.76	17.38	20.95	17.46	68.02	18.06
CLIP (↑)	0.311	0.308	0.310	0.312	0.300	0.310	0.277	0.305



[1] Zhang, Yimeng, et al. "To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now." European Conference on Computer Vision. Springer, Cham, 2025.