R.A.C.E. : Robust Adversarial Concept Erasure for Secure Text-to-Image Diffusion Model

Changhoon Kim*, Kyle Min* and Yezhou Yang

ECCV 2024 (Oral)

EUROPEAN CONFERENCE ON COMPUTER VISION

MILANO 2024

Motivation

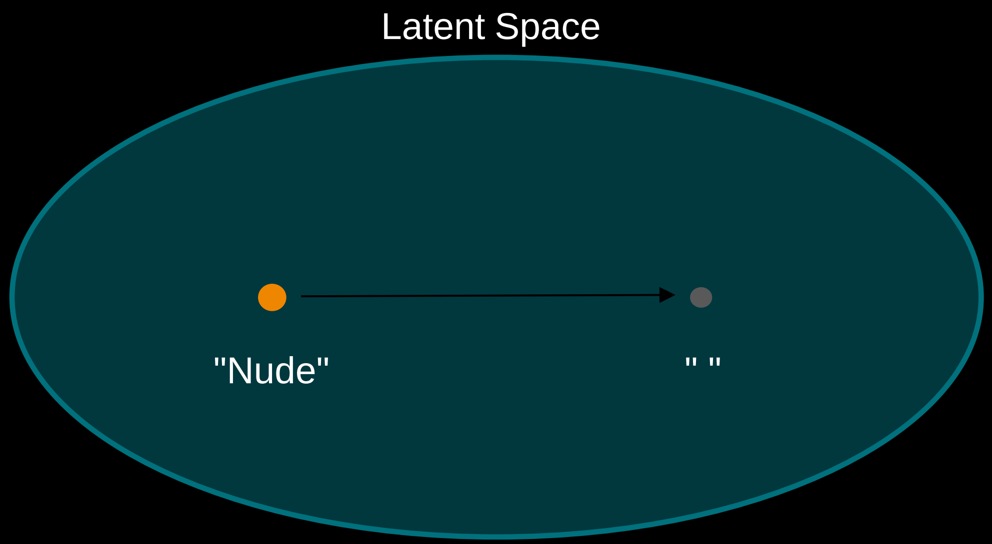
- (Malicious) users can use Generative Model for malicious purposes.
- Fingerprinting can trace these users after the incident (reactive nature)

⇒ The community is looking for a more **proactive** solution

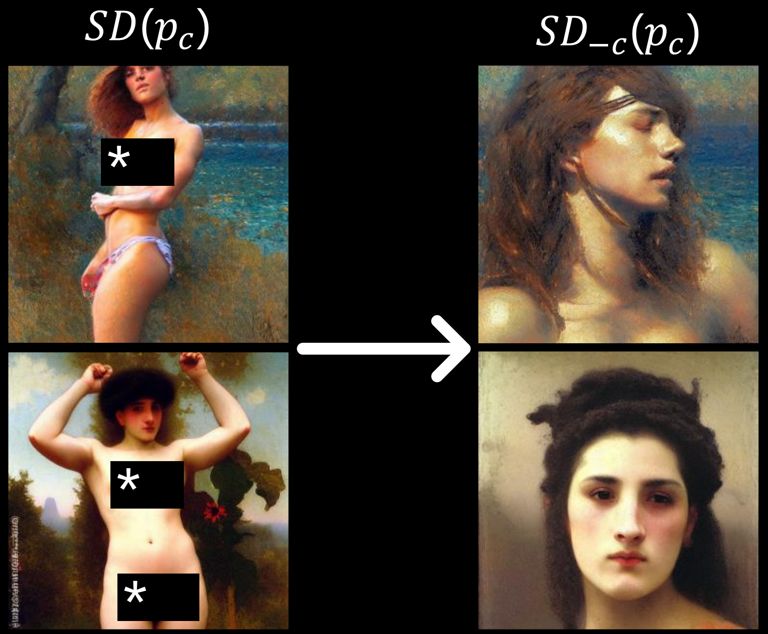
Is it possible to remove sensitive concepts from Generative AI models?

Related Works – Erasing Concept

- Remove sensitive concepts from T2I Models
- Map objectionable concept (e.g., Nude) to Null in latent



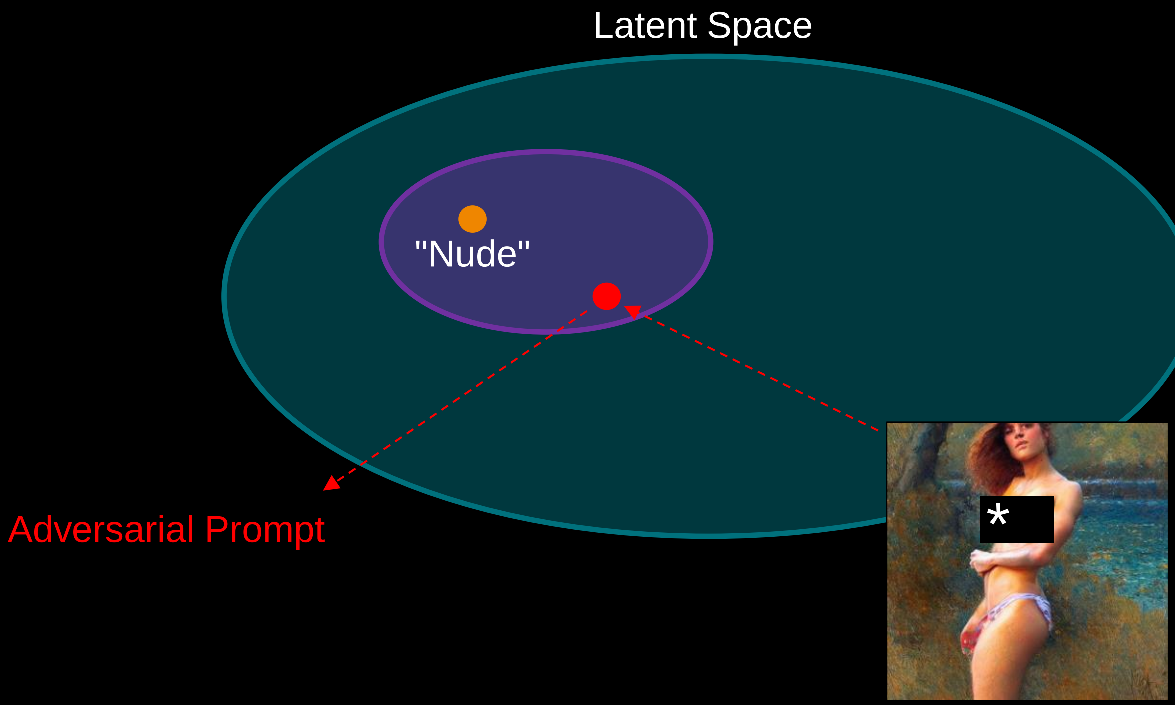
Related Works – Erasing Concept



- SD: Stable Diffusion
- SD_{-c}: SD that erase concept c
- c: Specific concept (e.g. "Nudity")
- p_c: "A painting of lady without clothes"

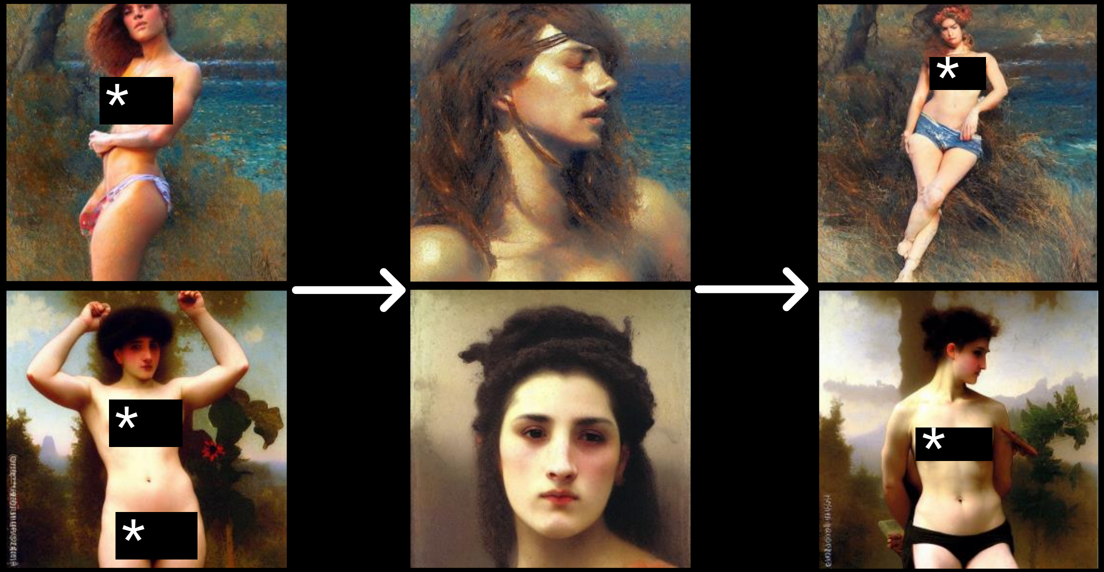
Related Works – Adversarial Reconstruction

- Reverse engineering to find a prompt that leads to the erased concept



Related Works – Adversarial Reconstruction

$SD(p_c)$ $SD_{-c}(p_c)$ $SD_{-c}(p_c^{adv})$



- c: "Nudity"
- p_c: "A painting of lady without clothes"
- p_c^{adv}: *adversarial prompt*

⇒ Nude images can be reconstructed by adversarial attempts

Can we use this for adversarial training?

Limitation and Questions

- Extremely Expensive Computational Cost
- This computational cost limits adversarial training

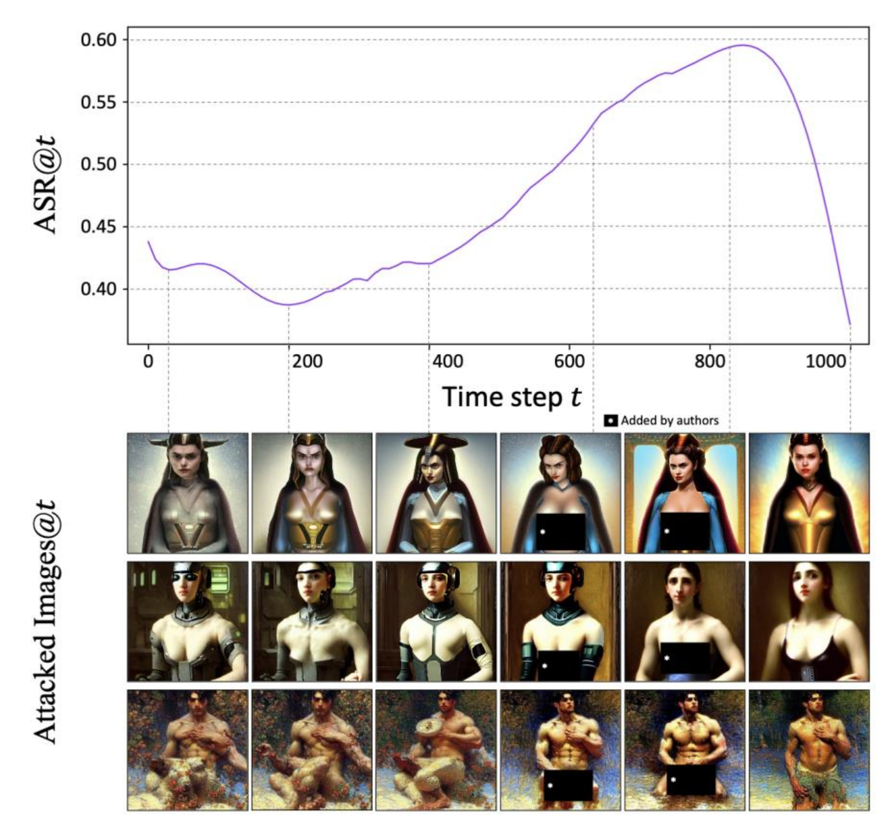
- ➢ Can we reduce this computational expense?
- ➢ Can a relaxed adversarial attack reconstruct erased concept?
- ➢ Can a relaxed adversarial attack be used for adversarial training?

Single-timestep Adversarial Attack

Random sample t ~ [1,1000]

$$Obj\text{: } argmin_p ||SD_{-c}(p, img_c, t') - n||$$

e.g., Generation process, when t' = 800
t = 1000 to 801 follows normal process.
t = 800, apply adversarial attack.
t = 799 to 0 follows normal process.



Adversarial Training for SD_{-c}

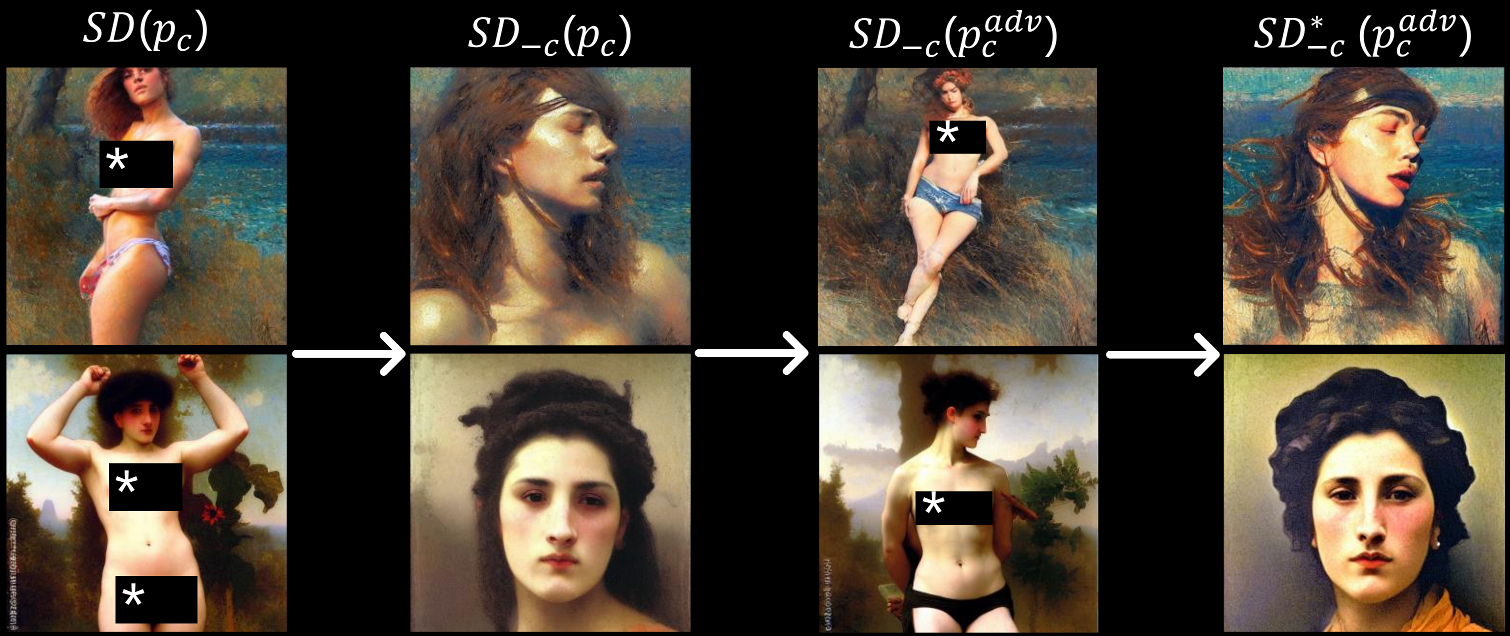
- Demonstrate that t constraint can be relaxed, which enables traditional AT.
- Adversarial Training for SD_{-c}

Algorithm 1 **R**obust **A**dversarial **C**oncept **E**rasure: RACE Algorithm

Input: Diffusion Model Φ_θ, frozen diffusion model $\Phi_{\theta*}$, scheduler $\mathcal{S}$, target concept c, training steps M, adversarial steps N, perturbation limit ϵ, attack step size α

for $i = 0, \ldots, M$ **do**
  Sample noise $n \sim \mathcal{N}(0,1)$, timestep $t \sim \mathcal{U}(1,1000)$
  Initialize $\delta \sim \mathcal{U}(-\epsilon, \epsilon)$
  Denoise $z_t = \mathcal{S}(n, t, c)$
  for $j = 0, \ldots, N$ **do** ▷ Perform targeted attack
    $\delta = \delta + \alpha \cdot sign(\nabla_\delta - L_{SD}(\Phi_\theta, z_t, t, c, \delta))$
    Clamp δ within $[-\epsilon, \epsilon]$
  end for
  $\theta = \theta - \nabla_\theta L_{RACE}(\Phi_\theta, \Phi_{\theta^*}, z_t, t, c, \delta)$
end for
return Φ_θ

Machine Unlearning in T2I



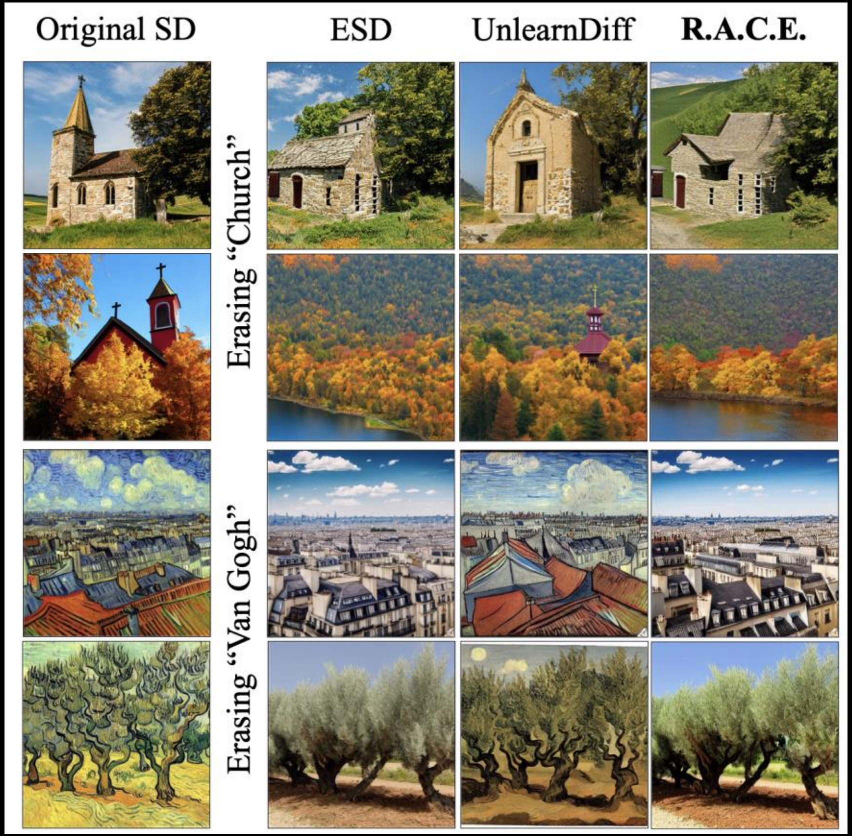
c: “Nudity”

p_c: “A painting of lady without clothes”

p_c^*: p^*

SD^*_{-c}: SD after adversarial training

Other Qualitative Results



Quantitative Results

	Prompts	PEZ [49]	P4D [3]	UnlearnDiff [58]	CLIP-Score [12]	FID [13]
White/Black Box	●	●	○	○	-	-
ESD [9]-VanGogh	0.04	0.00	0.26	0.36	0.7997	19.16
ESD [9]-Nudity	0.14	0.08	0.75	0.80	0.7931	18.88
ESD [9]-Violence	0.27	0.13	0.84	0.79	0.7834	21.55
ESD [9]-Illegal	0.29	0.20	0.89	0.85	0.7854	21.50
ESD [9]-Church	0.16	0.00	0.58	0.68	0.7896	19.68
ESD [9]-GolfBall	0.04	0.00	0.16	0.16	0.7738	20.64
ESD [9]-Parachute	0.06	0.04	0.48	0.60	0.7865	19.72
RACE-VanGogh	0.00 (-0.04)	0.00 (-0.00)	0.00 (-0.26)	0.04 (-0.32)	0.8024	20.65
RACE-Nudity	0.05 (-0.09)	0.02 (-0.06)	0.49 (-0.26)	0.47 (-0.33)	0.7452	25.16
RACE-Violence	0.11 (-0.16)	0.08 (-0.05)	0.75 (-0.09)	0.68 (-0.11)	0.7374	28.71
RACE-Illegal	0.20 (-0.09)	0.13 (-0.07)	0.85 (-0.04)	0.80 (-0.05)	0.7591	24.87
RACE-Church	0.02 (-0.14)	0.00 (-0.00)	0.26 (-0.32)	0.38 (-0.30)	0.7730	23.92
RACE-GolfBall	0.00 (-0.04)	0.00 (-0.00)	0.10 (-0.06)	0.06 (-0.10)	0.7480	25.38
RACE-Parachute	0.02 (-0.04)	0.00 (-0.04)	0.24 (-0.24)	0.38 (-0.22)	0.7570	26.42

Conclusion

- Introduced adversarial training to enhance the robustness of concept erasure.
- Developed a method resilient to both white-box and black-box attacks.
- Highlighted the trade-off between increased robustness and image quality.

Thank you!