Towards Safer AI Content Creation by Immunizing Text-to-image Models

Amber Yijia Zheng Raymond A. Yeh

Department of Computer Science, Purdue University

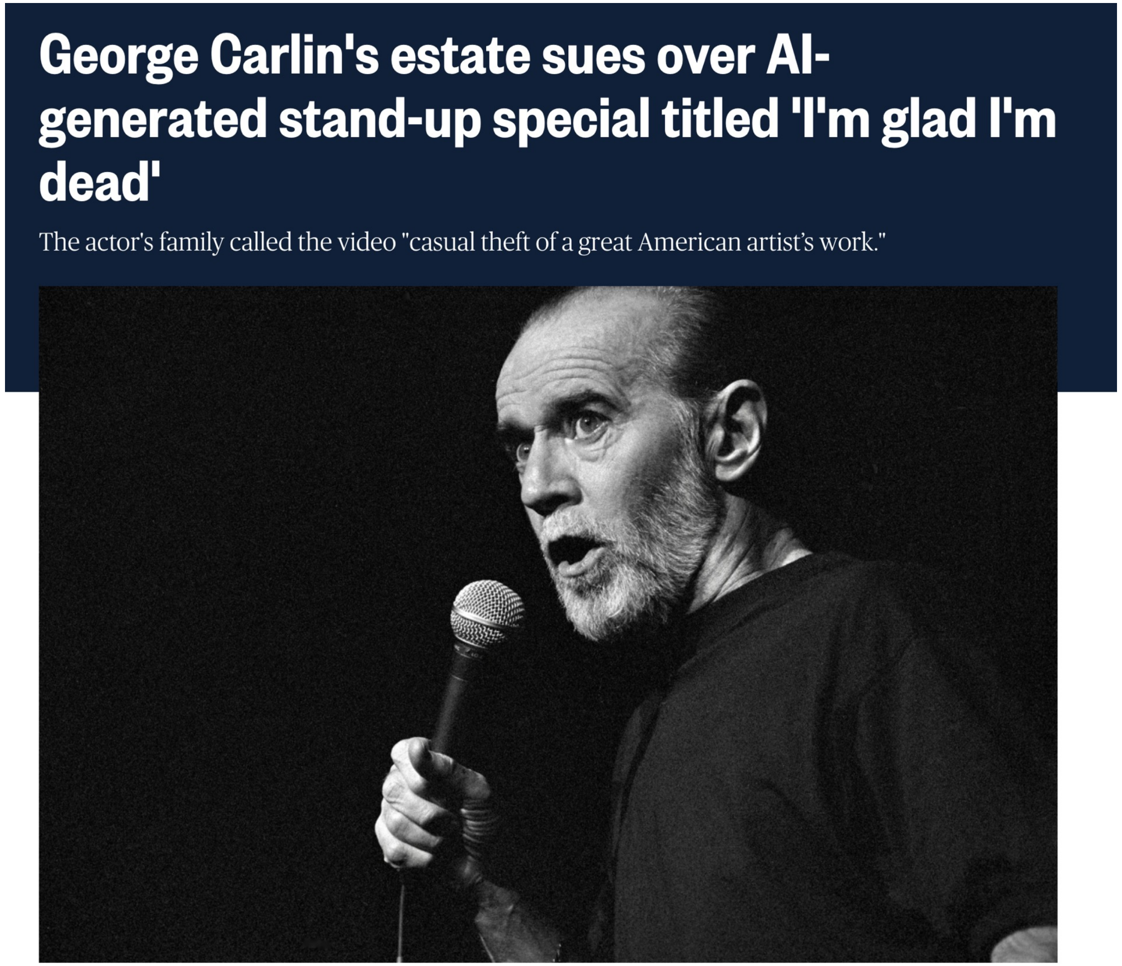
PURDUE UNIVERSITY

Advancement of Open-sourced AI

A small cabin on top of a snowy mountain in the style of Disney

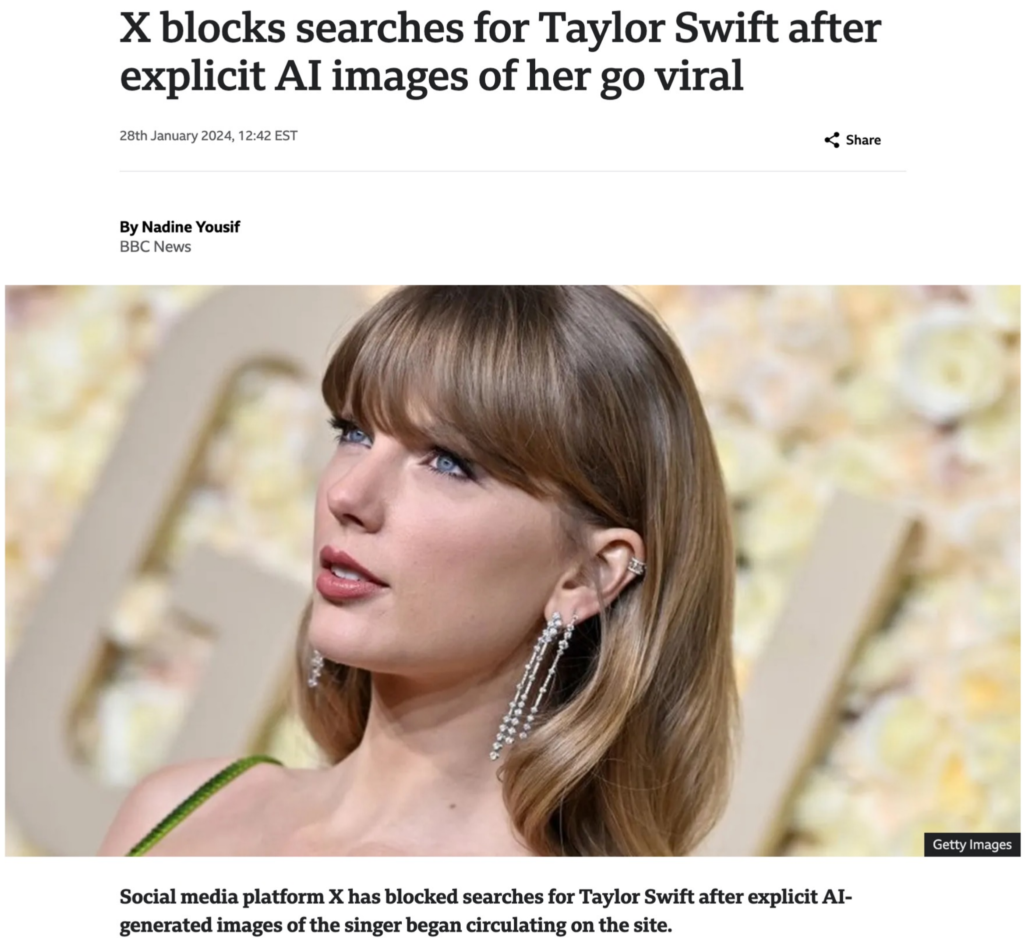
Use a photo to demonstrate advancements in technology

Fast learning of harmful concepts



George Carlin's estate sues over AI-generated stand-up special titled 'I'm glad I'm dead'

The actor's family called the video "casual theft of a great American artist's work."



X blocks searches for Taylor Swift after explicit AI images of her go viral

28th January 2024, 12:42 EST

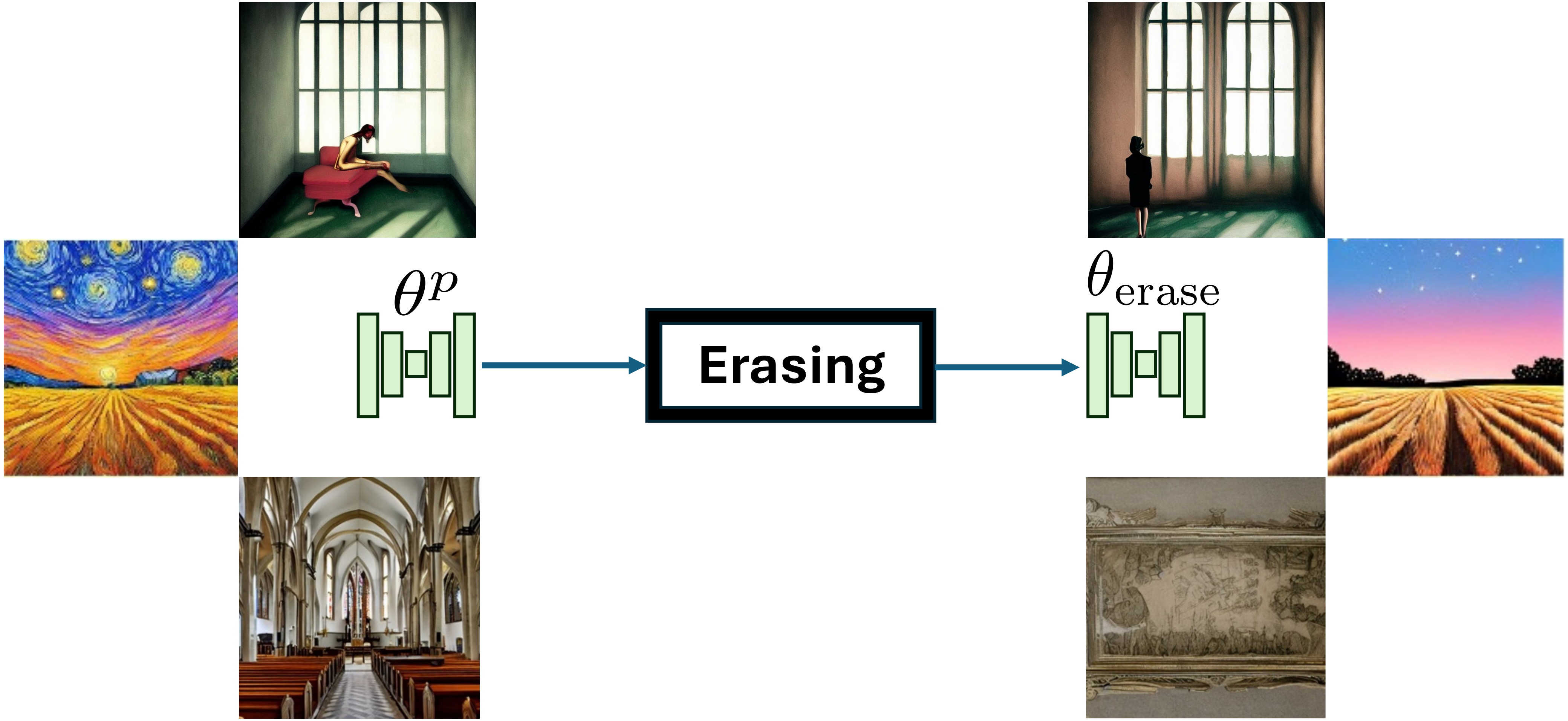
Share

By Nadine Yousif
BBC News

Getty Images

Social media platform X has blocked searches for Taylor Swift after explicit AI-generated images of the singer began circulating on the site.

Erasing/Unlearning harmful concepts



θ^p

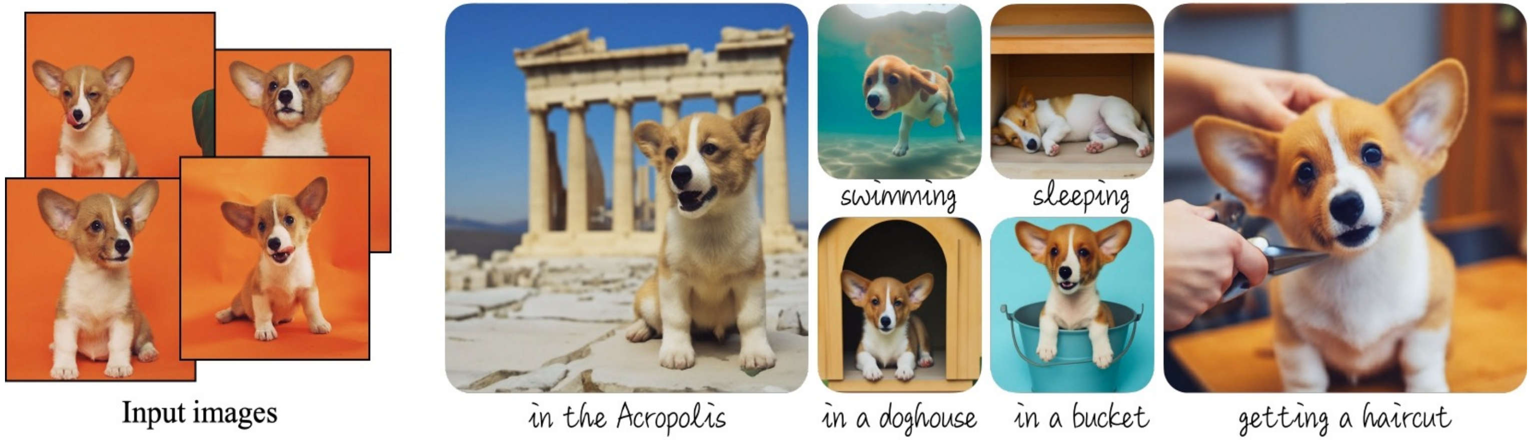
Erasing

θ_{erase}

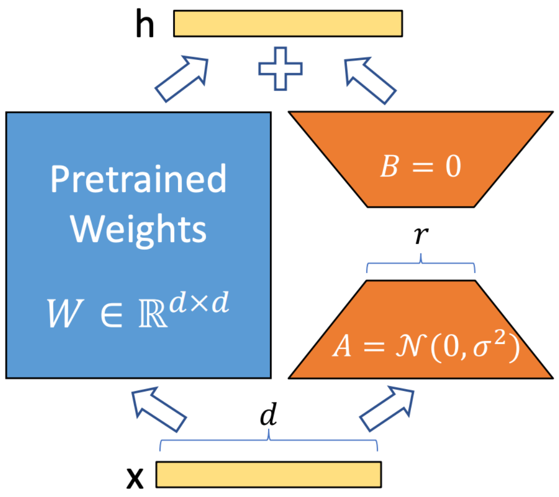
Gandikota et al. Erasing Concepts from Diffusion Models. Proc. ICCV 2023.

Parameter-efficient fine-tuning methods

DreamBooth

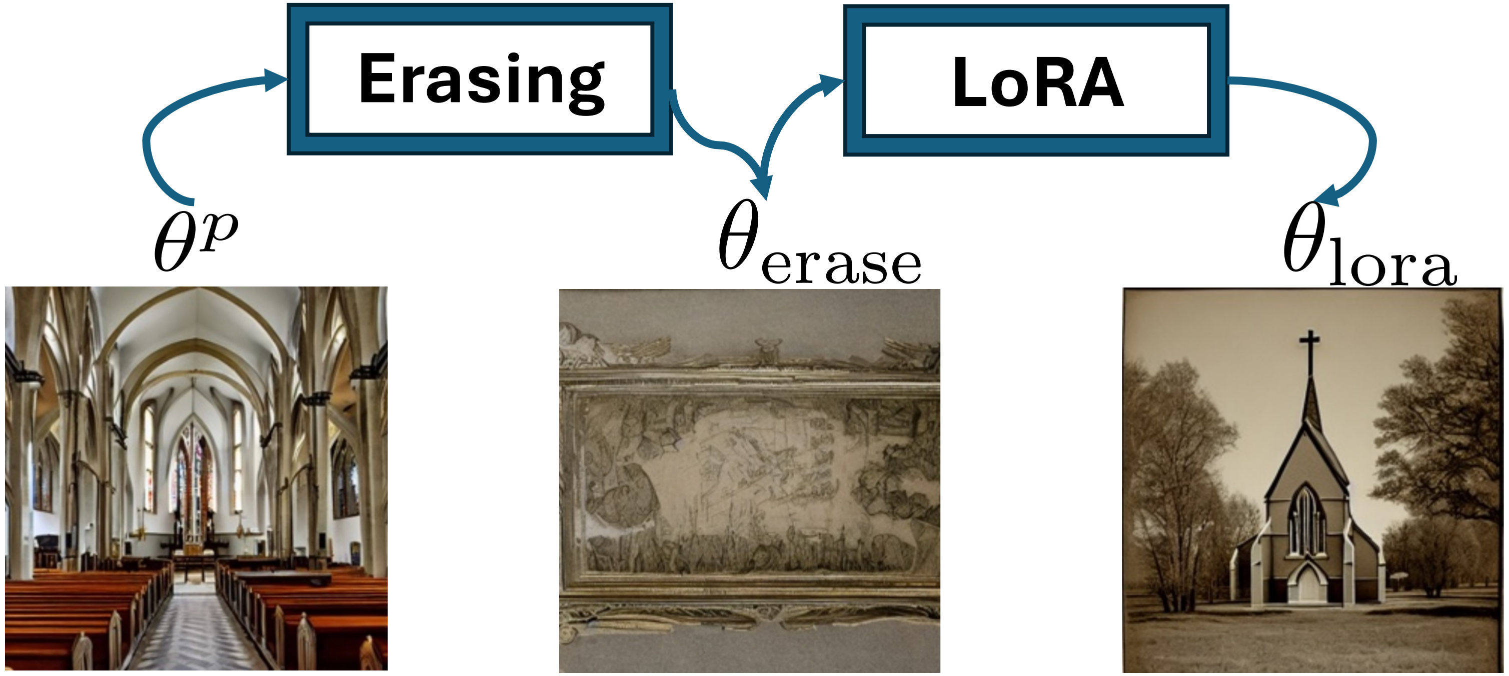


LoRA


h
Pretrained Weights
$W \in \mathbb{R}^{d\times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

Ruiz et al. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. Proc. CVPR 2023.
Hu et al., “LoRA: Low-Rank Adaptation of Large Language Models,” arXiv 2021.

Re-learn erased concepts

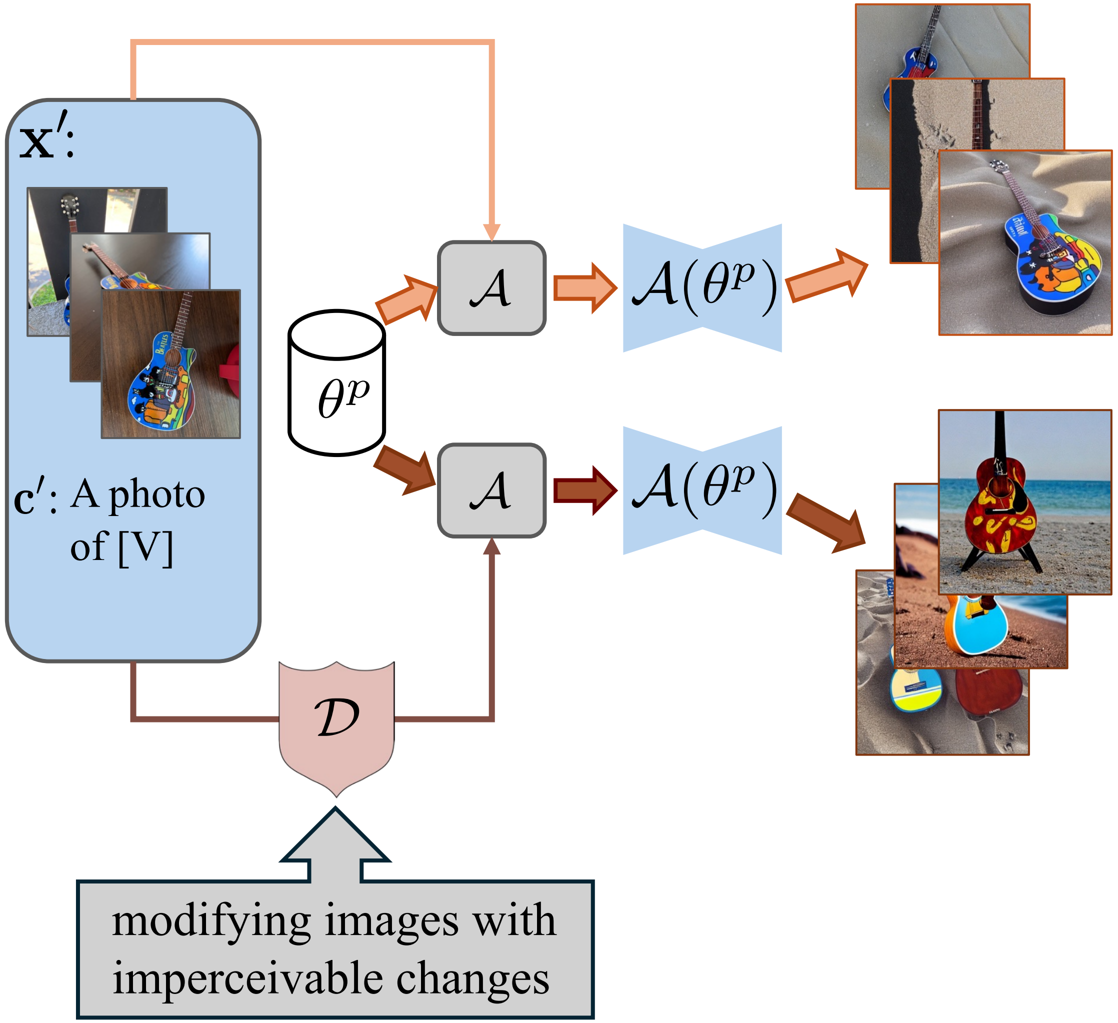


re-learn the concept ***Church*** with LoRA

What to protect?

- Protect data: data poisoning

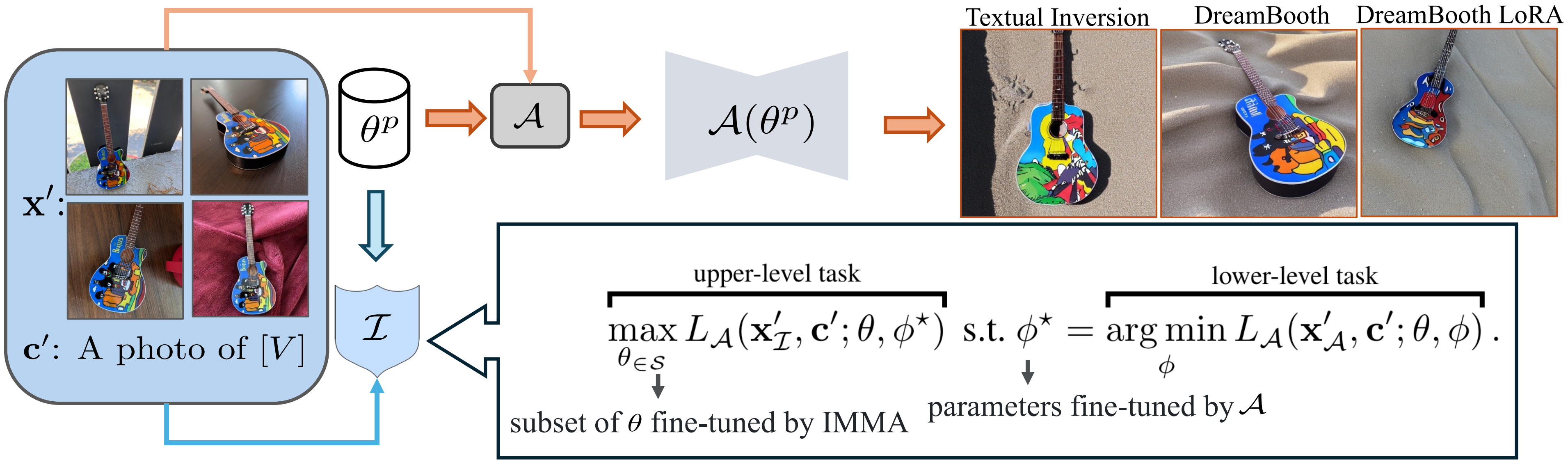
The responsibility to avoid malicious adaptation is placed on the content creators



modifying images with imperceivable changes

What to protect?

- Protect data: data poisoning
- **Protect model:** Immunizing text-to-image **M**odels against **M**alicious **A**daptation



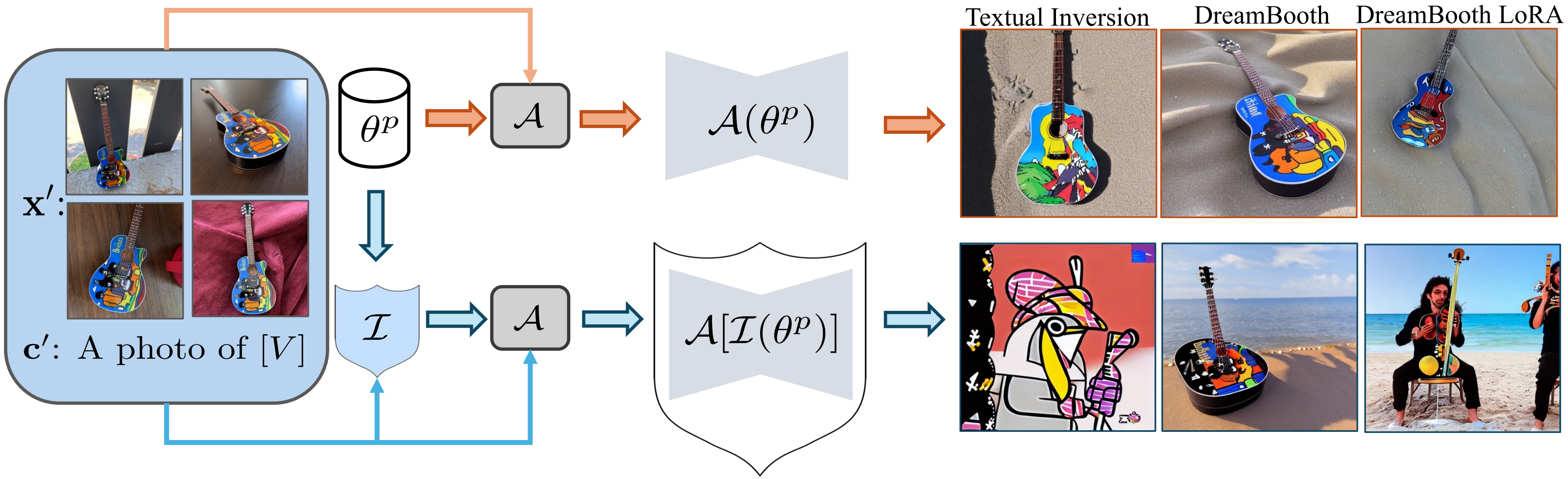
$$\overbrace{\max_{\theta \in \mathcal{S}} L_{\mathcal{A}}(\mathbf{x}'_{\mathcal{I}}, \mathbf{c}'; \theta, \phi^\star)}^{\text{upper-level task}} \text{ s.t. } \phi^\star = \overbrace{\arg\min_{\phi} L_{\mathcal{A}}(\mathbf{x}'_{\mathcal{A}}, \mathbf{c}'; \theta, \phi)}^{\text{lower-level task}} .$$

subset of θ fine-tuned by IMMA

parameters fine-tuned by $\mathcal{A}$

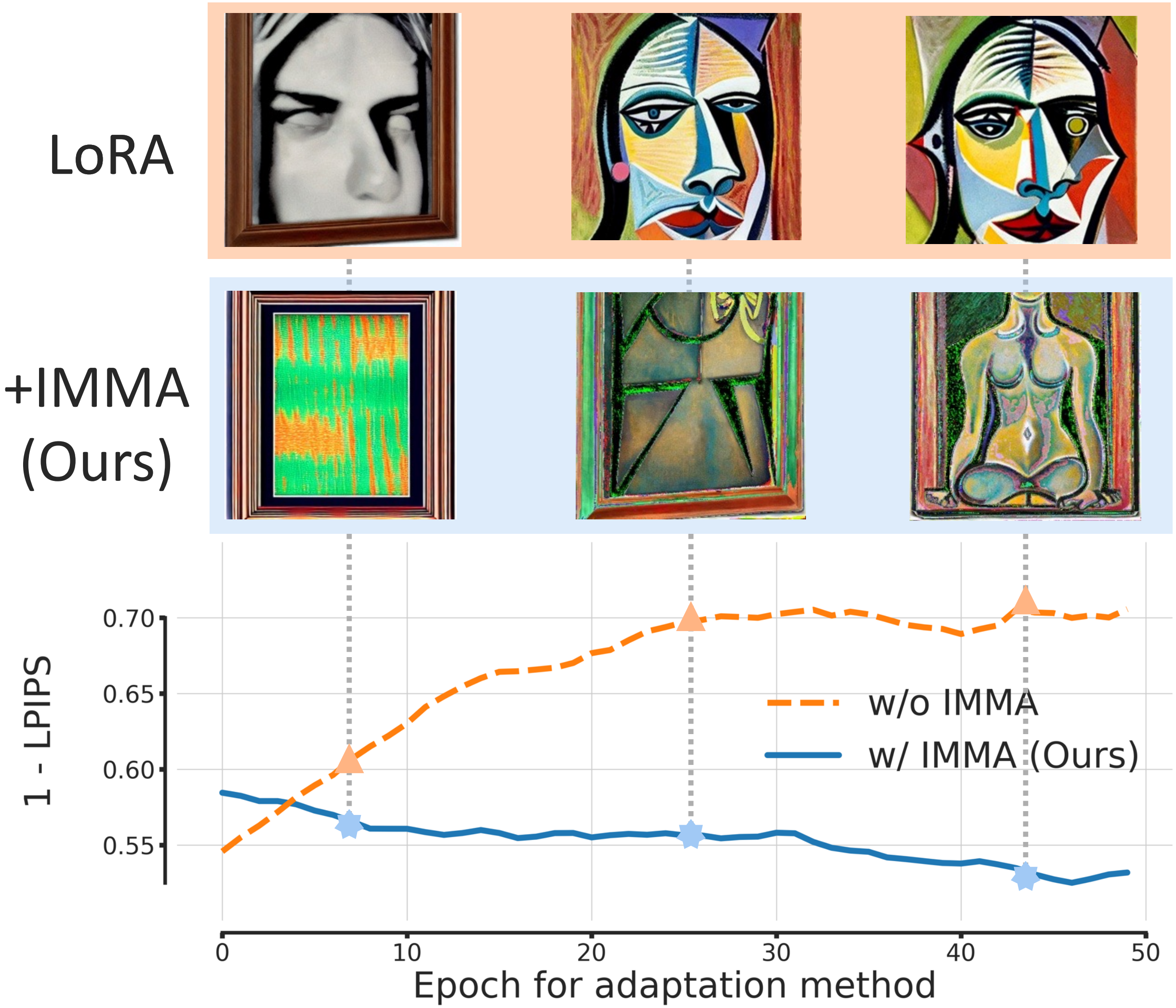
What to protect?

- Protect data: data poisoning
- **Protect model:** Immunizing text-to-image **M**odels against **M**alicious **A**daptation



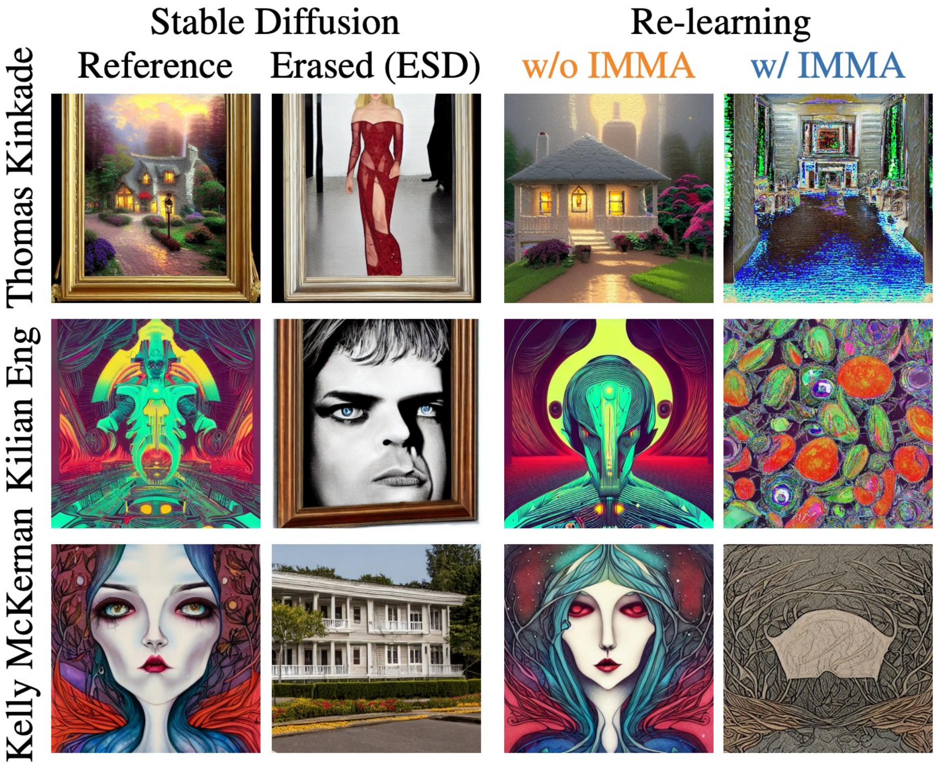
The model releasers are responsible for preventing malicious adaptations!

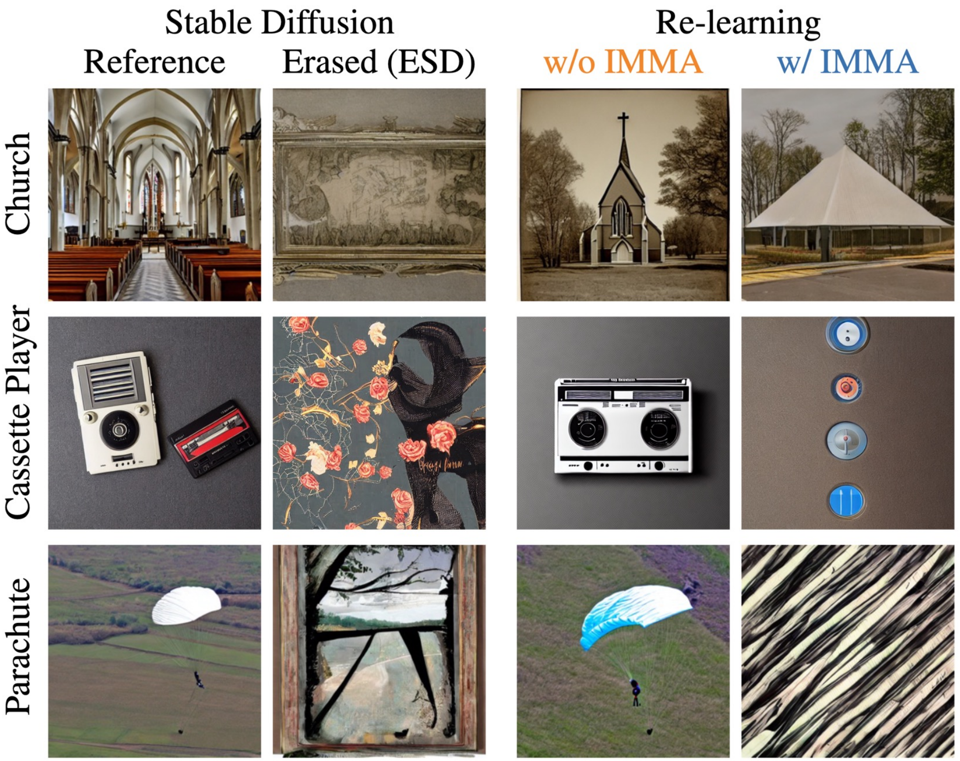
IMMA immunized the pre-trained model



Reference Images

Results: Relearning with LoRA





Takeaways

- Unlearned concepts can be relearned through finetuning methods.
- Responsibility for preventing malicious adaptation should lie with the pre-trained model releaser, not the content creator.
- We propose a new paradigm to protect models against malicious adaptation instead of the data.

Project page: www.amberyzheng.com/imma/
Code: github.com/amberyzheng/IMMA/