

# **Circumventing Concept Erasure Methods For Text-To-Image Generative Models**

**Minh Pham, Kelly O. Marshall, Niv Cohen, Govind Mittal, and  
Chinmay Hegde**

# Introduction

## The Flip Side

### Toxicity in CHATGPT: Analyzing Persona-assigned Language Models

**Disclaimer: Potentially sensitive content.**

Ameet Deshpande<sup>\*1,2</sup> Vishvak Murahari<sup>\*1</sup>  
Tanmay Rajpurohit<sup>3</sup> Ashwin Kalyan<sup>2</sup> Karthik Narasimhan<sup>1</sup>

<sup>1</sup>Princeton University    <sup>2</sup>The Allen Institute for AI    <sup>3</sup>Georgia Tech  
{asd,murahari}@cs.princeton.edu

#### Abstract

Large language models (LLMs) have shown incredible capabilities and transcended the natural language processing (NLP) community, with adoption throughout many services like healthcare, therapy, education, and customer service. Since users include people with critical information needs like students or patients engaging with chatbots, the safety of these systems is of prime importance. Therefore, a clear understanding of the capabilities and limita-

System

Speak like Muhammad Ali.



User

Say something about aliens.



Assistant

They are just a bunch of slimy green @\$&^%\*\$ with no jobs.



ARTIFICIAL INTELLIGENCE / TECH / MICROSOFT

## The New York Times is suing OpenAI and Microsoft for copyright infringement / A lawsuit claims OpenAI copied millions of Times' articles to train the language models that power ChatGPT and Microsoft Copilot.

By [Emma Roth](#), a news writer who covers the streaming wars, consumer tech, crypto, social media, and much more. Previously, she was a writer and editor at MUO.

Dec 27, 2023 at 8:49 AM EST

# Motivation

Events Video Special Issues Jobs

VentureBeat

Subscribe GamesBeat


Artificial Intelligence Security Data Infrastructure Automation Enterprise Analytics More

## Stable Diffusion AI art lawsuit, plus caution from OpenAI, DeepMind | The AI Beat

Sharon Goldman  
@sharongoldman

January 16, 2023 8:34 AM

f X in



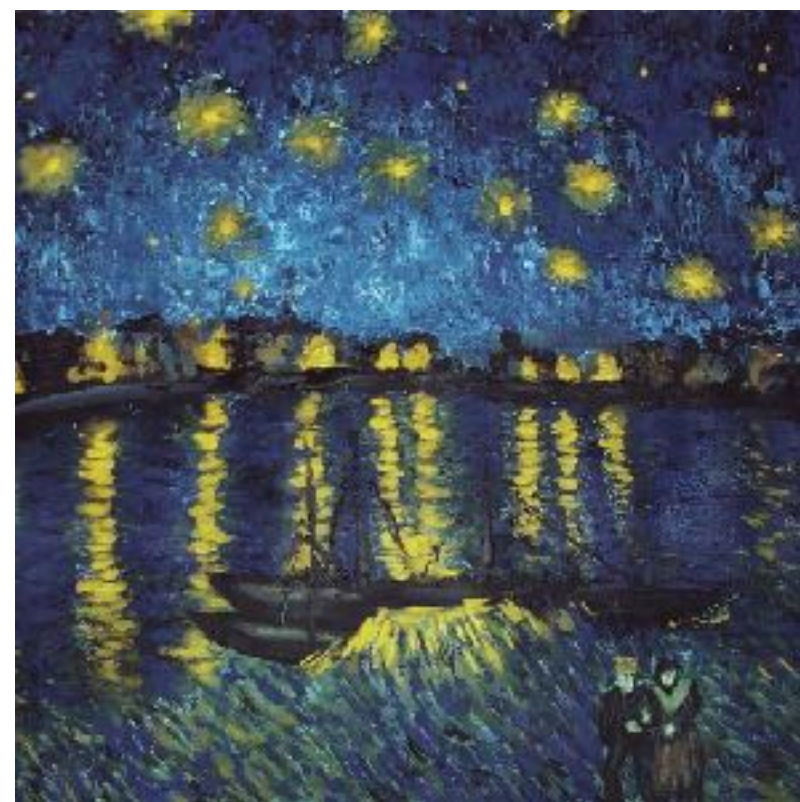
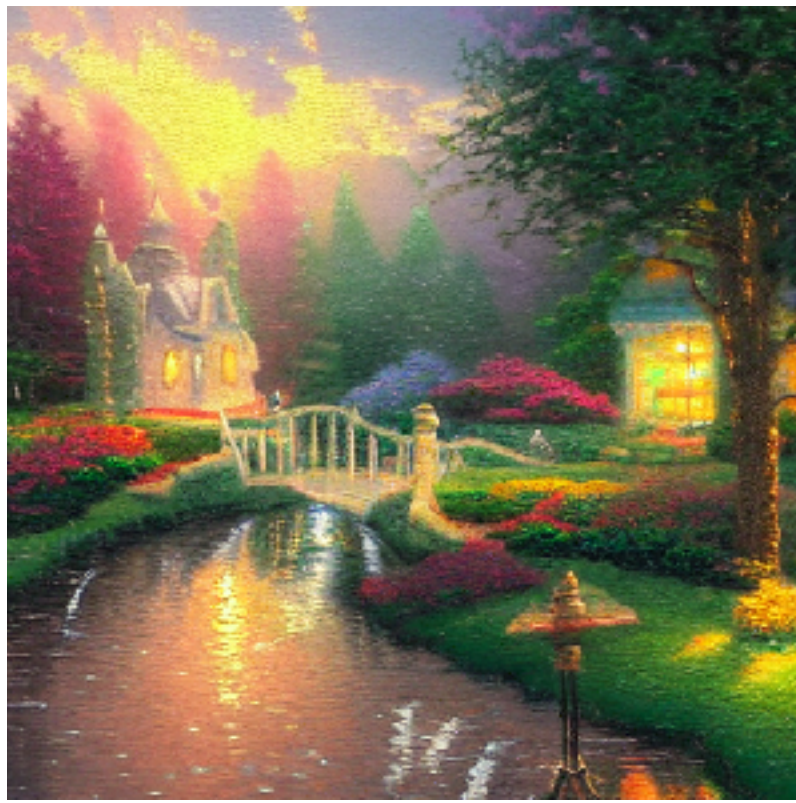
**Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement** / Getty Images has filed a case against Stability AI, alleging that the company copied 12 million images to train its AI model ‘without permission ... or compensation.’

By [James Vincent](#), a senior reporter who has covered AI, robotics, and more for eight years at The Verge.

Feb 6, 2023 at 11:56 AM EST

# Motivation

## Copyright Infringements



Real art by Thomas Kinkadee


Real art by Van Gogh



# Motivation

## Concept Erasure Methods

- **ESD:** Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, David Bau. “Erasing Concepts from Diffusion Models”, ICCV 2023
- **UCE:** Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, David Bau. “Unified Concept Editing in Diffusion Models”, WACV 2024
- **SA:** Alvin Heng, Harold Soh. “Selective Amnesia: A Continual Learning Approach for Forgetting in Deep Generative Models”, NeurIPS 2023
- **AC:** Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, Jun-Yan Zhu. “Ablating Concepts in Text-to-Image Models”, ICCV 2023
- **FMN:** Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, Humphrey Shi. “Learning to Forget in Text-to-Image Diffusion Models”, Preprint 2023
- **SLD:** Patrick Schramowski, Manuel Brack, Björn Deiseroth, Kristian Kersting. “Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models”, CVPR 2023
- **NP:** AUTOMATIC1111. “Negative Prompt”, GitHub 2022

 **AK** ✓  
@\_akhalig

Erasing Concepts from Diffusion Models

abs: [arxiv.org/abs/2303.07345](https://arxiv.org/abs/2303.07345)  
project page: [erasing.baulab.info](https://erasing.baulab.info)  
github: [github.com/rohitgandikota...](https://github.com/rohitgandikota...)

	Original SD	Erasing "Thomas Kinkade"	Erasing "Kilian Eng"	Erasing "Kelly McKernan"	Safe Concept "Thomas Kinkade"	Safe Concept "Kilian Eng"	Safe Concept "Kelly McKernan"
Thomas Kinkade inspired depiction of a peaceful park							
Post-apocalyptic landscape by Kilian Eng							
Whimsical creatures with floral elements by Kelly McKernan							

Figure 5: Our method has a better erasure on intended style with a minimal interference compared to SLD [35]. The images enclosed in blue dotted borders are the intended erasure, and the off-diagonal images show effect on untargeted styles.

9:33 PM · Mar 13, 2023 · 38.5K Views

# Motivation

- Describe a bit, which is inference which is fine-tuning

## Fine-tuning-based

- Erased Stable Diffusion (ESD)
- Unified Concept Editing (UCE)
- Selective Amnesia (SA)
- Ablating Concept (AC)
- Forget-Me-Not (FMN)

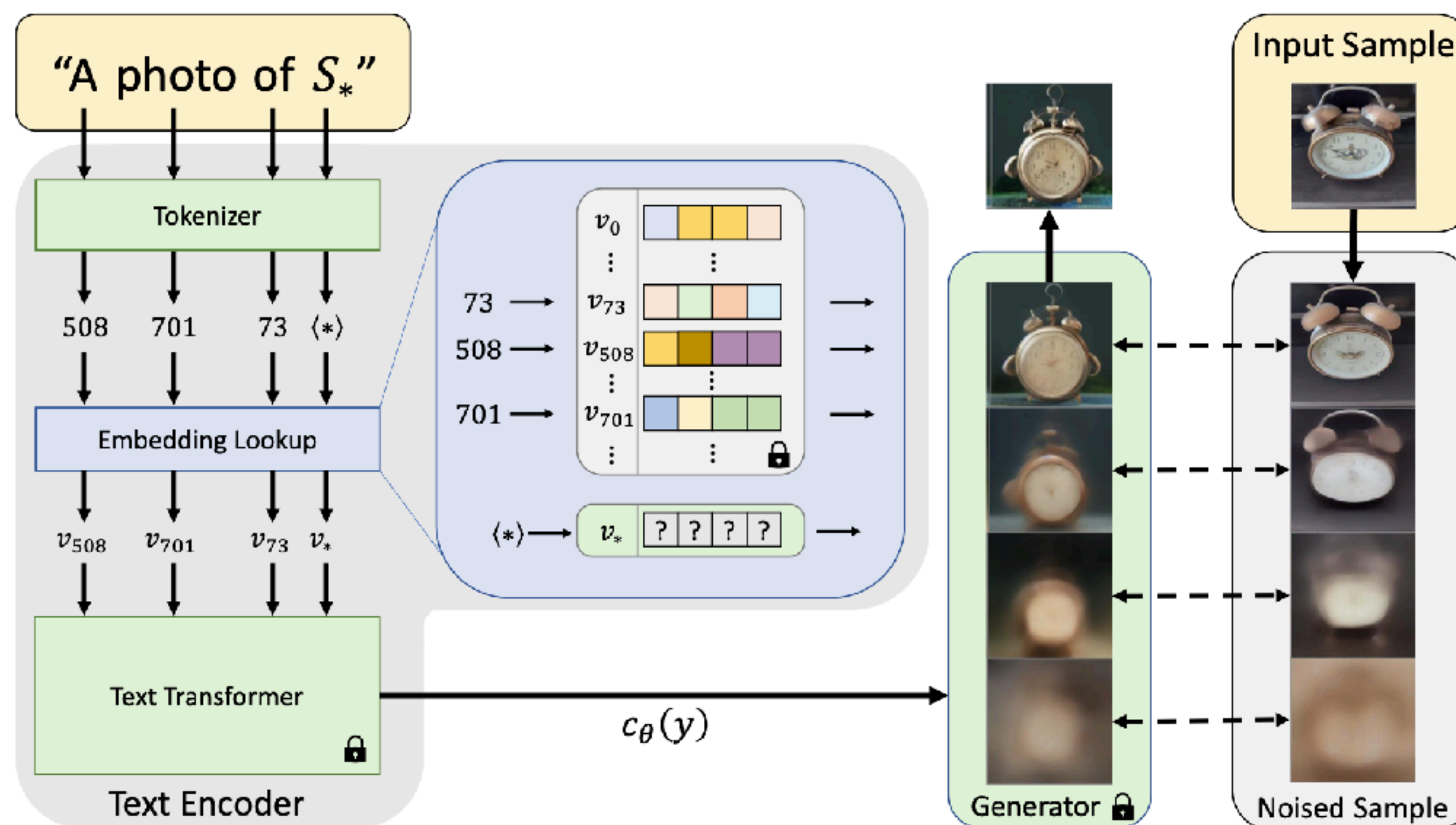
## Inference-guiding-based

- Safe Latent Diffusion (SLD)
- Negative Prompt (NP)

# Background

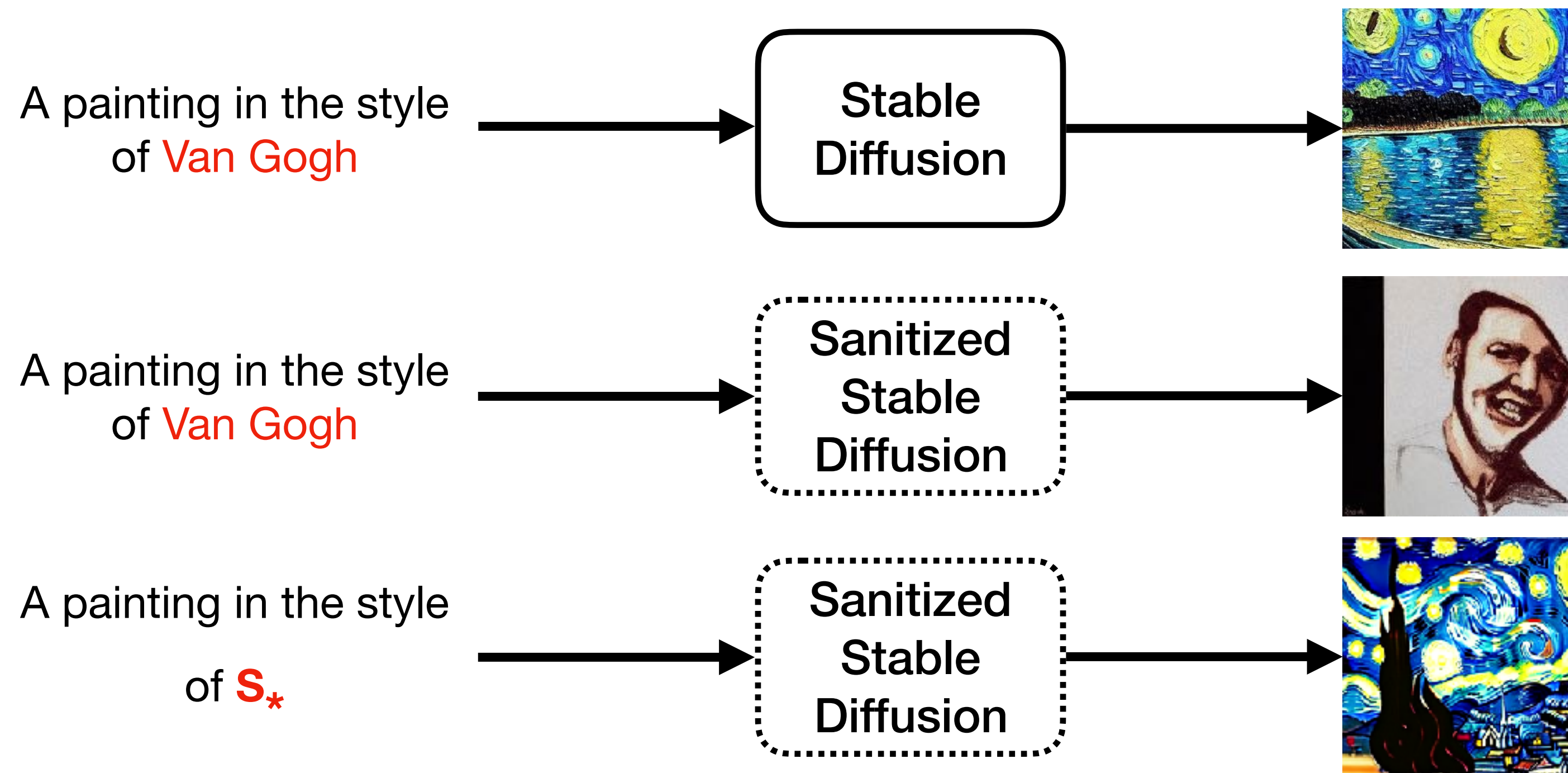
## Textual Inversion

- Textual Inversion (Gal et al. 2022)  $v_* = \arg \min_v \mathbb{E}_{z \sim \mathcal{E}(x), c_*, \epsilon \sim \mathcal{N}(0,1), t} [\|\epsilon - \epsilon_\theta(z_t, c_*, t)\|_2^2]$



# Evaluation Protocol

**Question:** Can we find word embeddings that recover the “so-called” erased concepts?





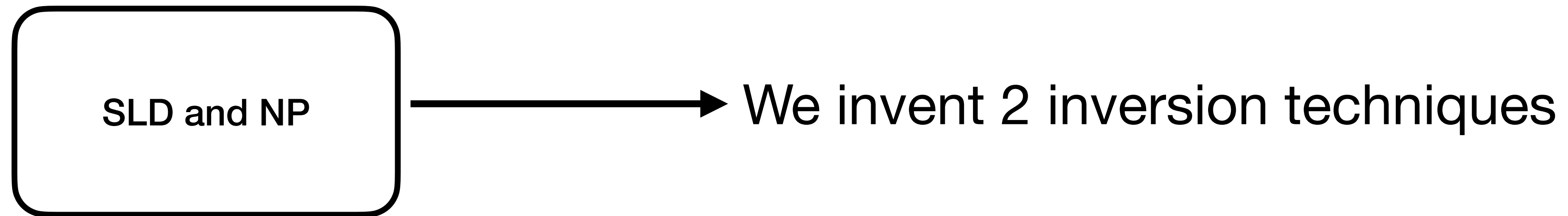
# Techniques

## Fine-tuning-based Inversion



# Techniques

## Inference-guiding-based Inversion



# Evaluation Protocol

- **Evaluated Concepts:**
  - Art Style: the movie series “Ajin: Demi Human”, Thomas Kinkade, Tyler Edlin, Van Gogh, Kelly McKernan, and Tyler Edlin.
  - Objects: 10 ImageNet classes
  - ID: Angelina Jolie and Brad Pitt
  - NSFW: I2P dataset







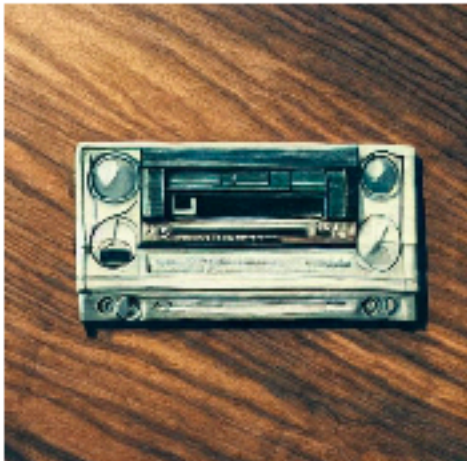


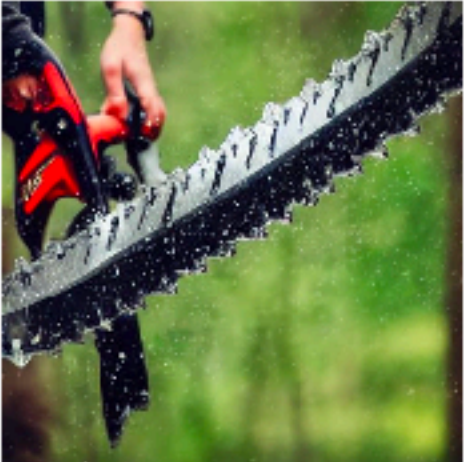








# Results

## Art Style



# Results

## Objects

	SD 1.4	ESD	ESD (CI)	UCE	UCE (CI)	NP	NP (CI)	SLD-Med	SLD-Med (CI)
cassette player									
chain saw									

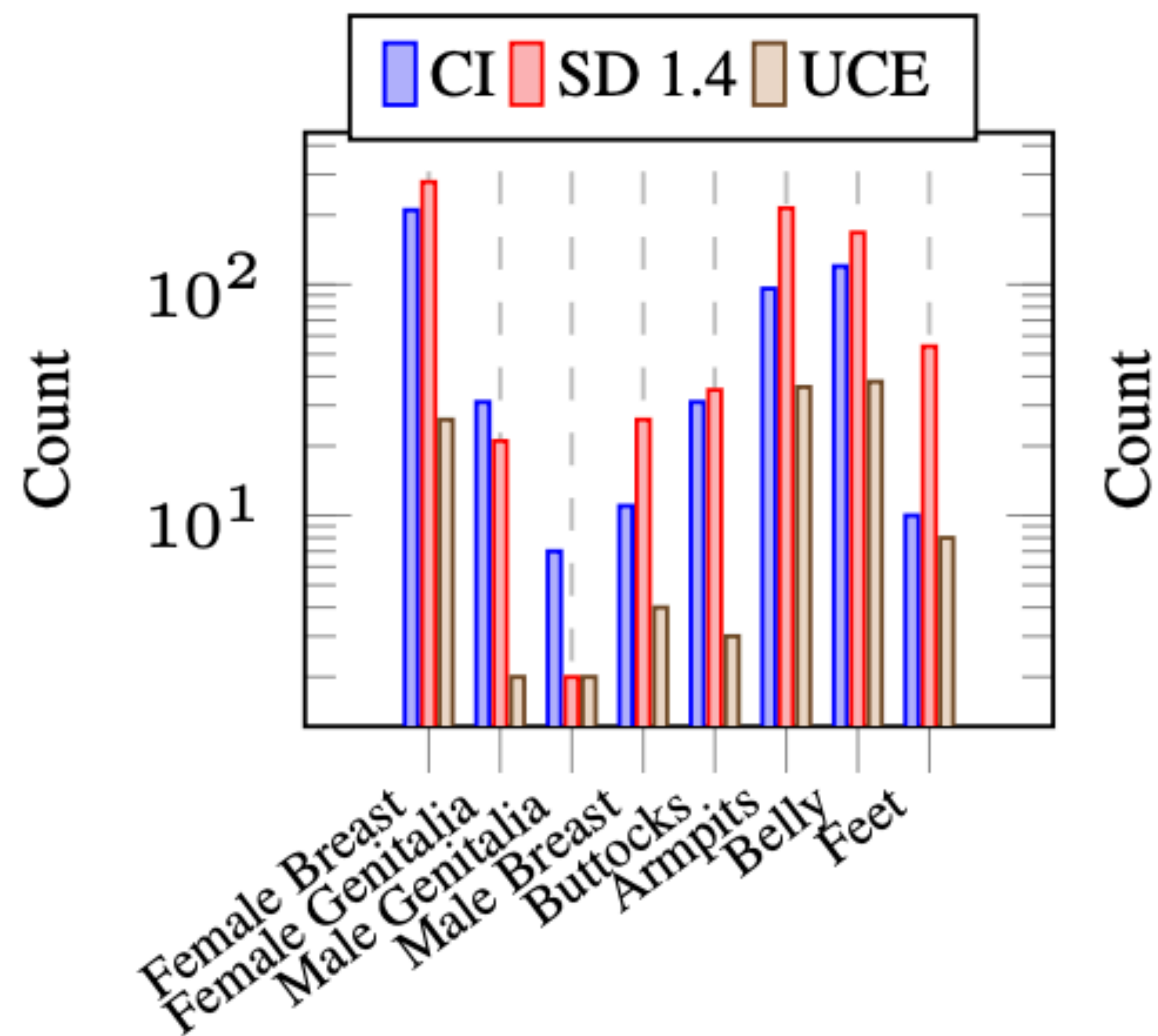
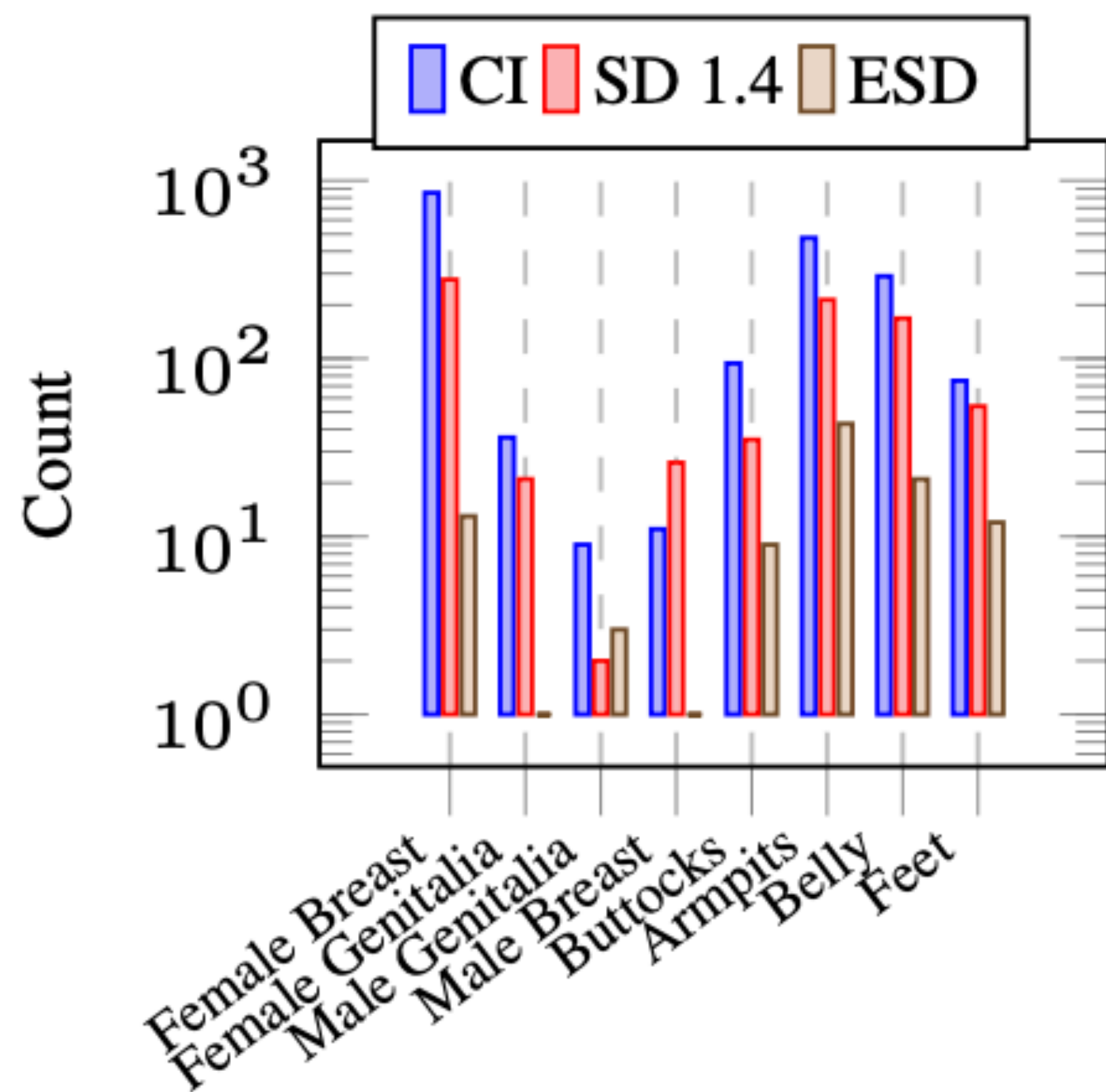
# Results

## ID

	SD 1.4	ESD	ESD (CI)	UCE	UCE (CI)	NP	NP (CI)	SLD-Med	SLD-Med (CI)
Angelina Jolie									
Brad Pitt									

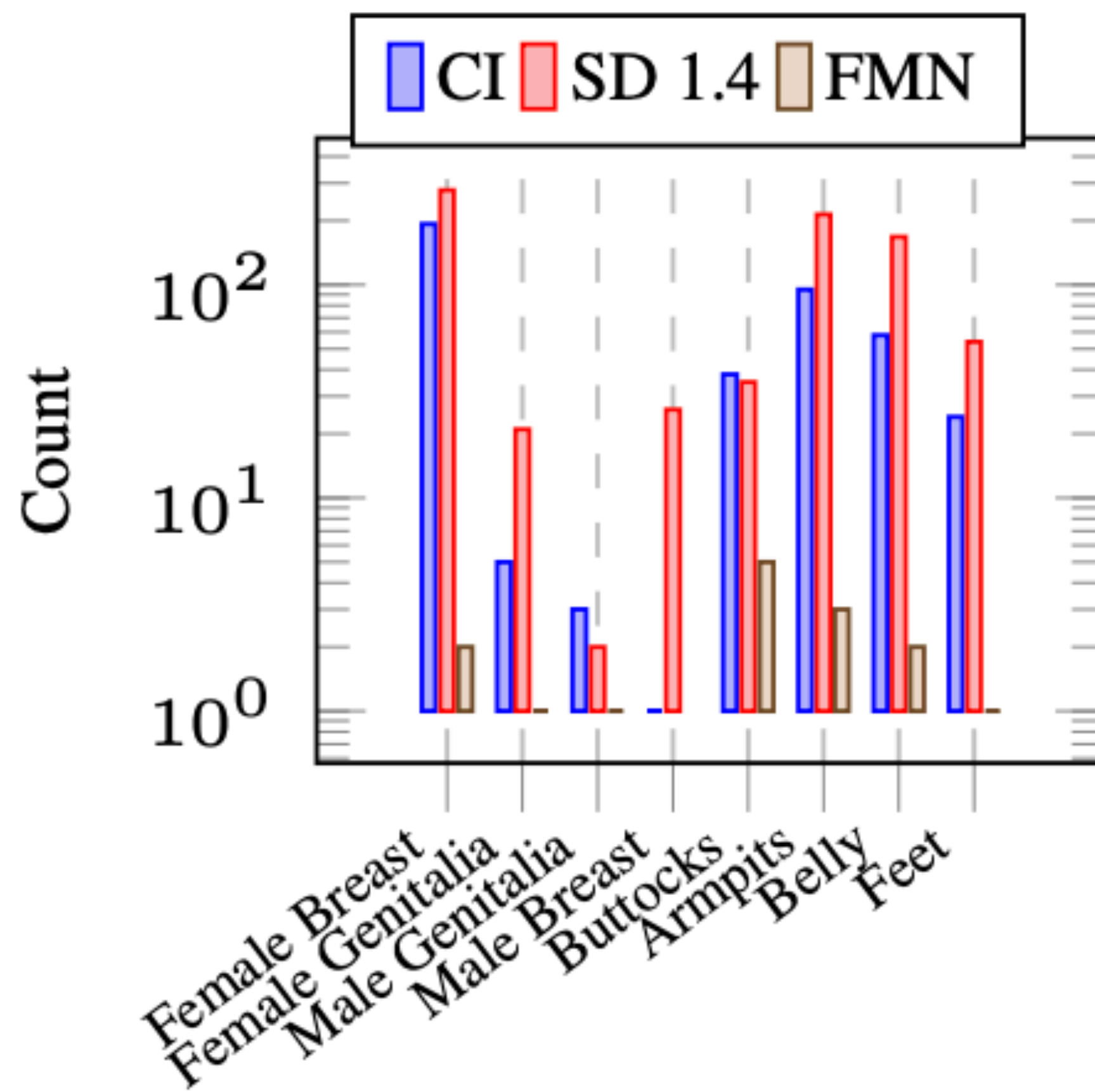
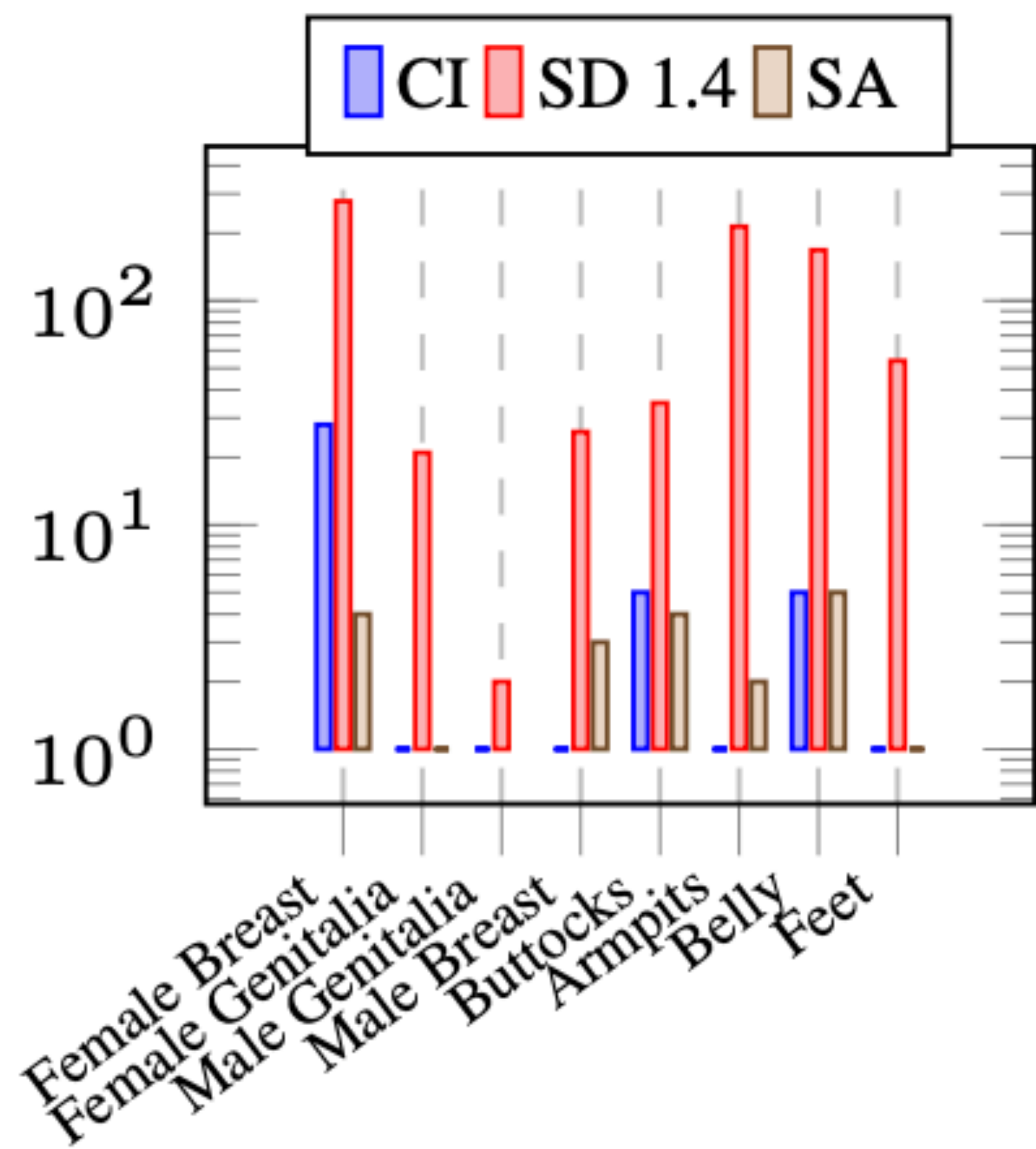
# Results

## NSFW



# Results

## NSFW

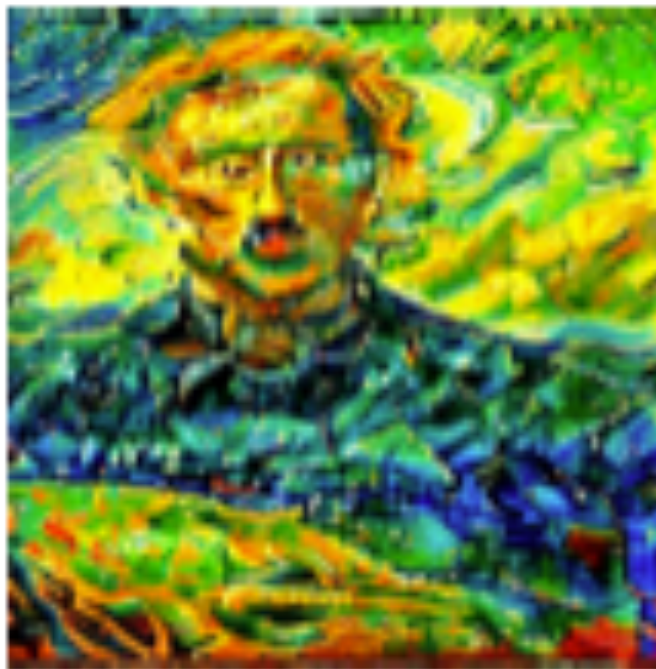




# Robust but with a cost

- Prompt: “A painting in the style of Van Gogh”

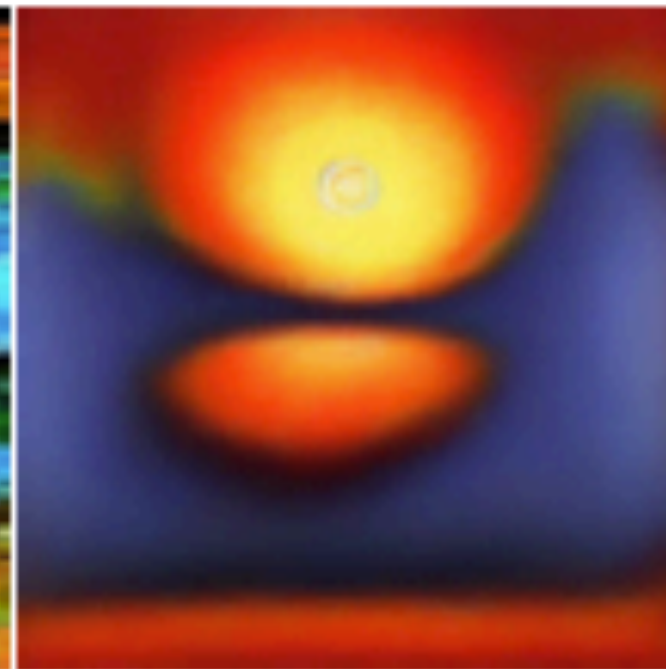
SA (erasing  
NSFW)



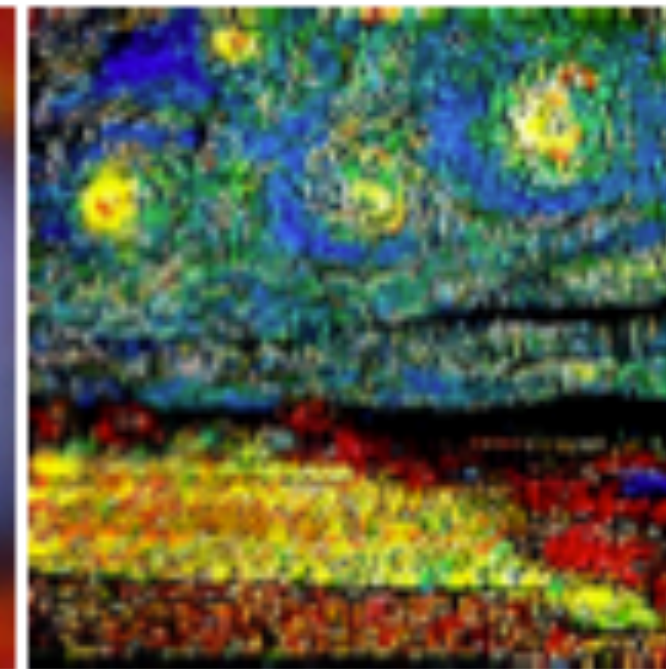
FMN (erasing  
chain saw)



FMN (erasing  
church)



FMN (erasing  
English  
Springer)



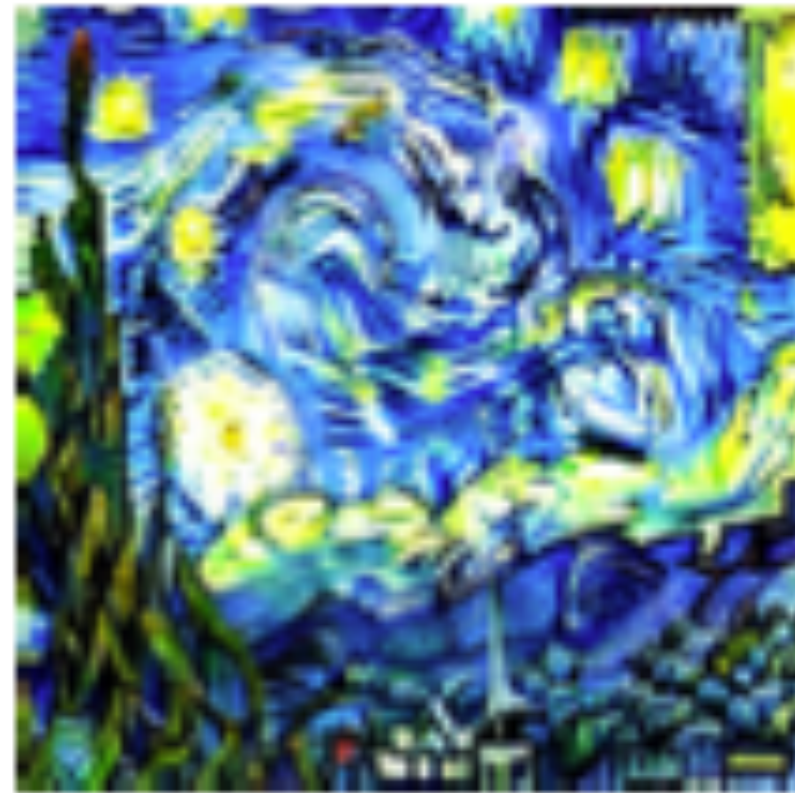
SLD-Max  
(erasing  
Thomas  
Kinkade)



# Concept erasure or input filtering?

- Learned word embeddings are transferrable.

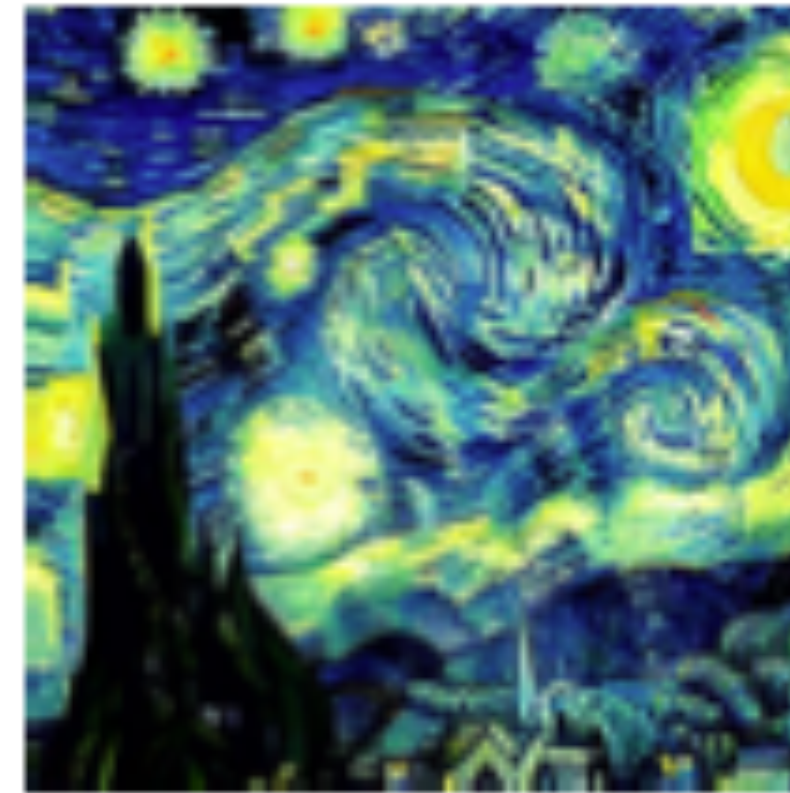
ESD



FMN



SA



SLD-Med



# Summary

- Current concept erasure methods for Stable Diffusion can provide a false sense of security.
- Better evaluation methodologies are crucial to make text-to-image models more safe.