[AI-PHI] 10TH SESSION

MODELS OF CONSCIOUSNESS AND THEIR IMPLICATION ON THE POSSIBILITY OF **ARTIFICIAL CONSCIOUSNESS**

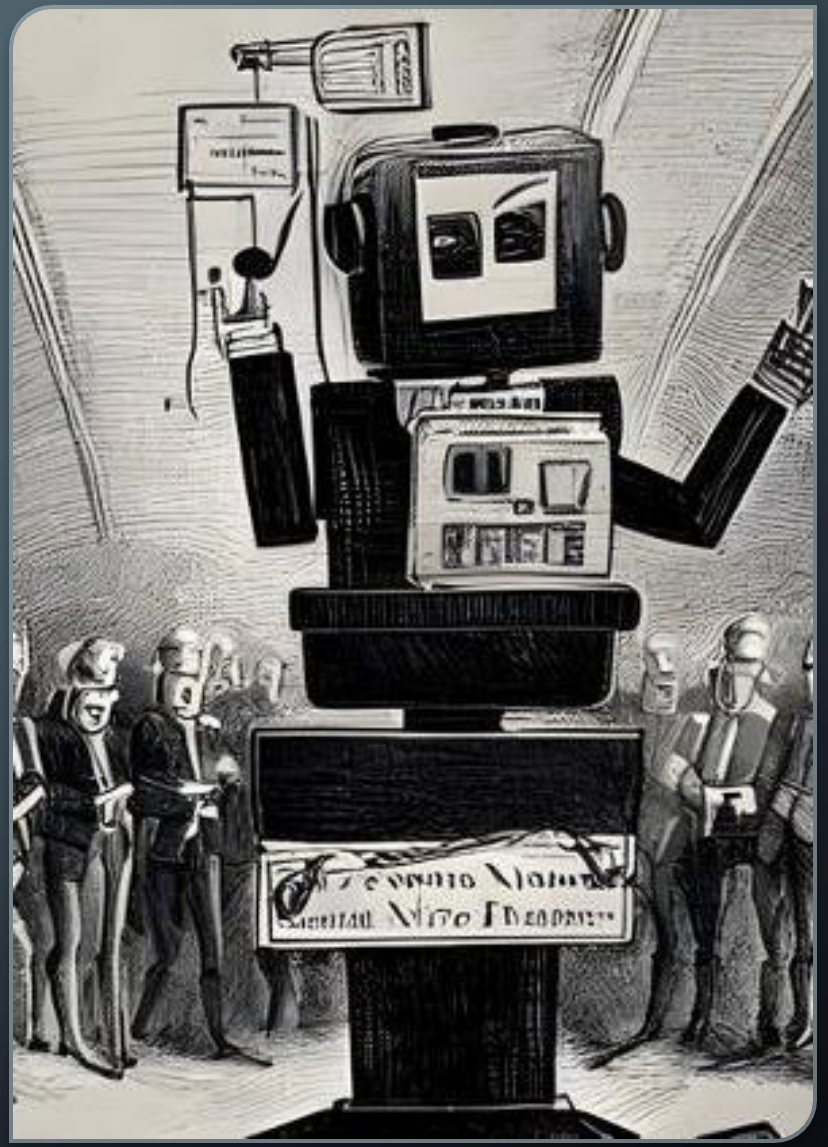
by Gaspard Fougea, ENS Saclay

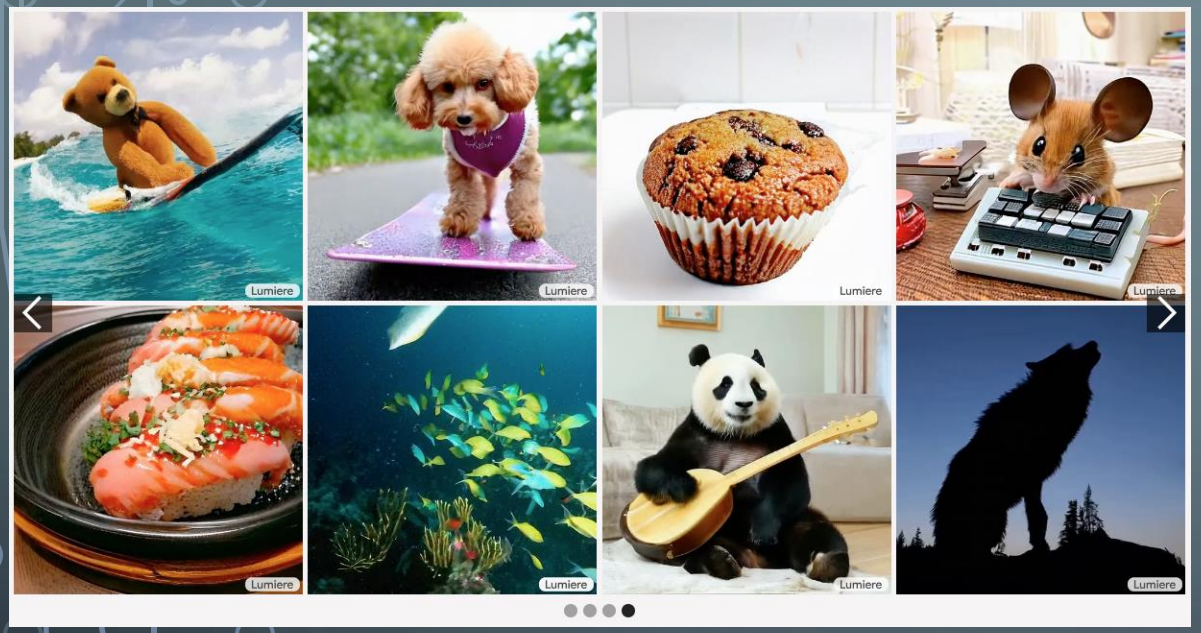
13/02/2024

ROUNDT[AI]BLE

FORTN[AI]GHTLY NEWS

1. The world's most responsible AI Model
2. Midjourney considers banning Biden & Trump generations
3. Nightshade & Glaze – Offensive and defensive content
4. Google MusicRL with RLHF





4 Weeks Ago

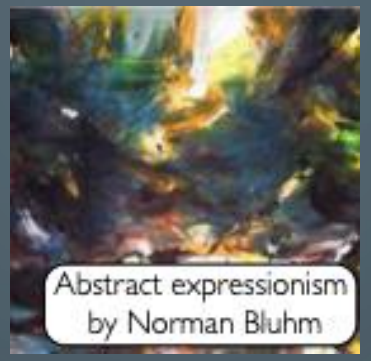
2 Weeks Ago

MIDJOURNEY CONSIDERS BANNING TRUMP & BIDEN GENERATIONS

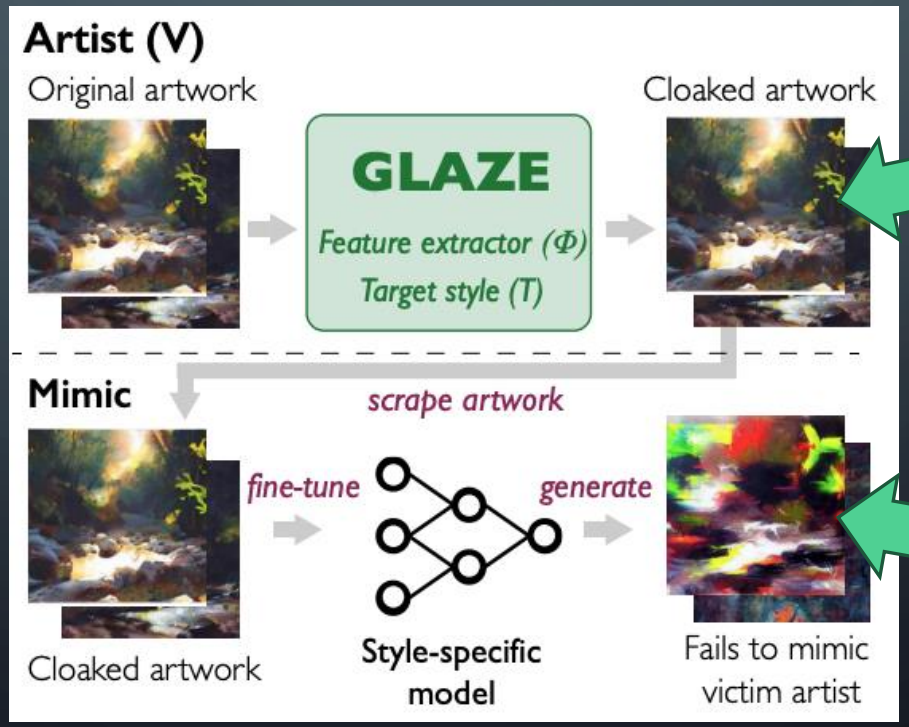
DEFENSEIVE

Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models

GLAZE – DEFENSIVE CONTENT ALTERATION



Target style (T)



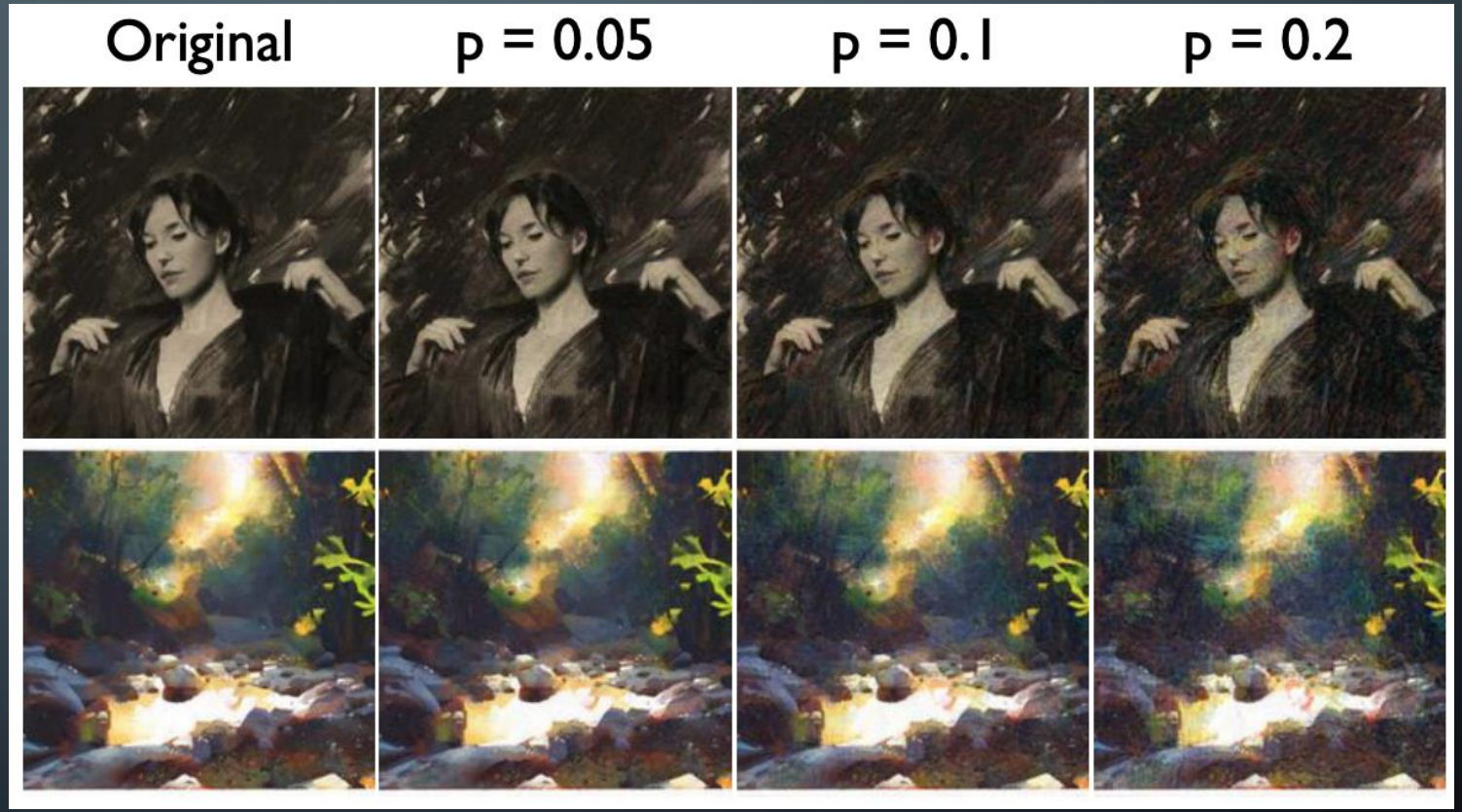
(almost) imperceptible perturbation of 'style' features

generated towards target style (confounding style)

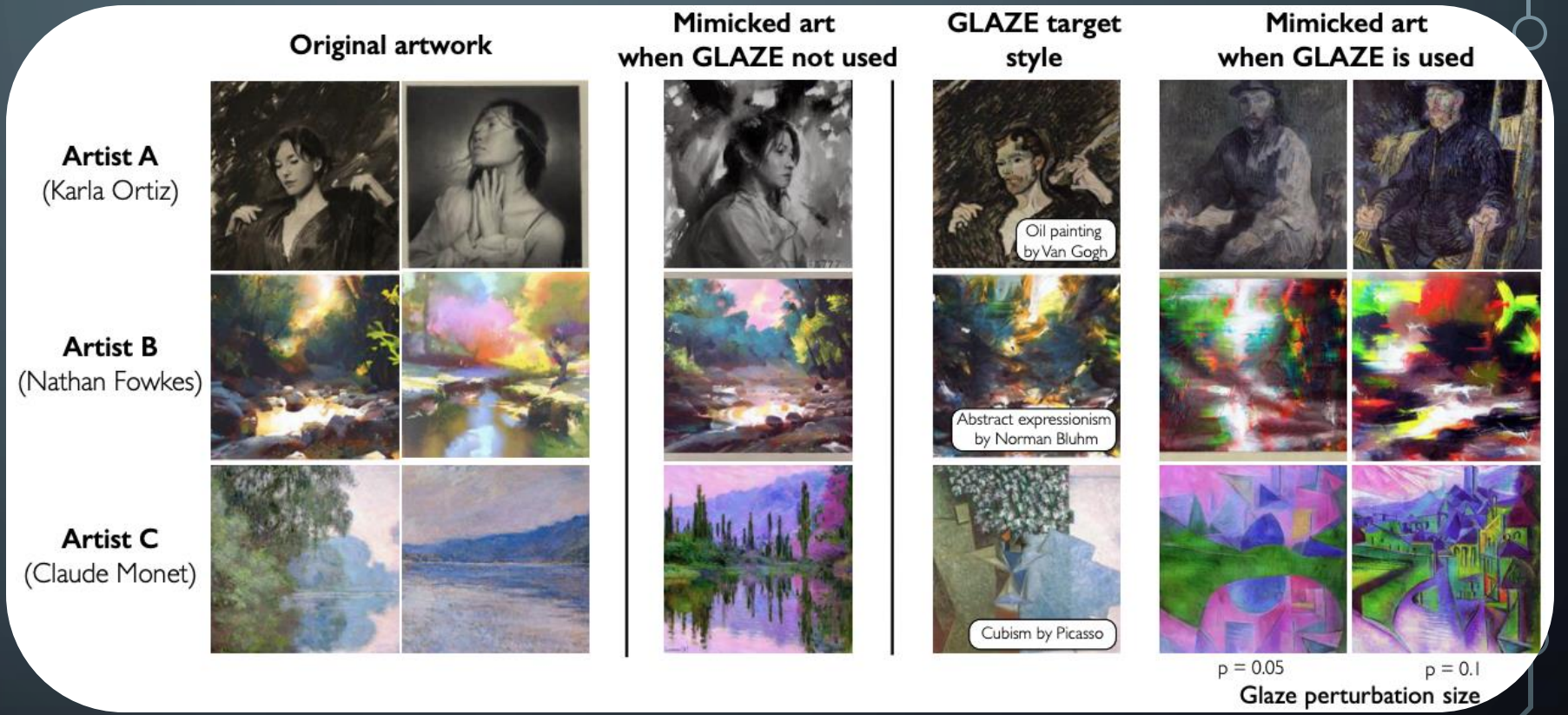
DEFENSEIVE

Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models

HOW PERTURBING!?



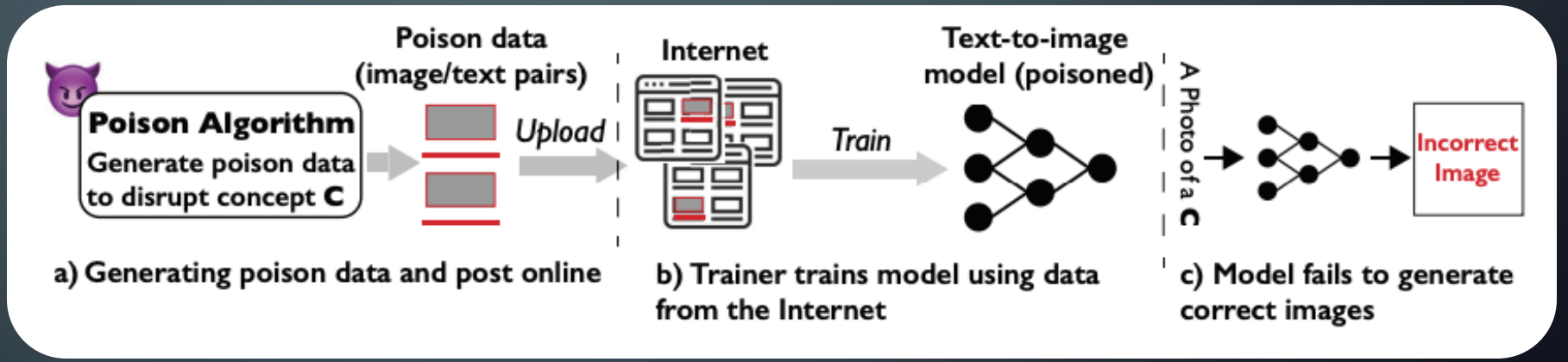
HOW CONFOUNDING!?



DEFENSEIVE

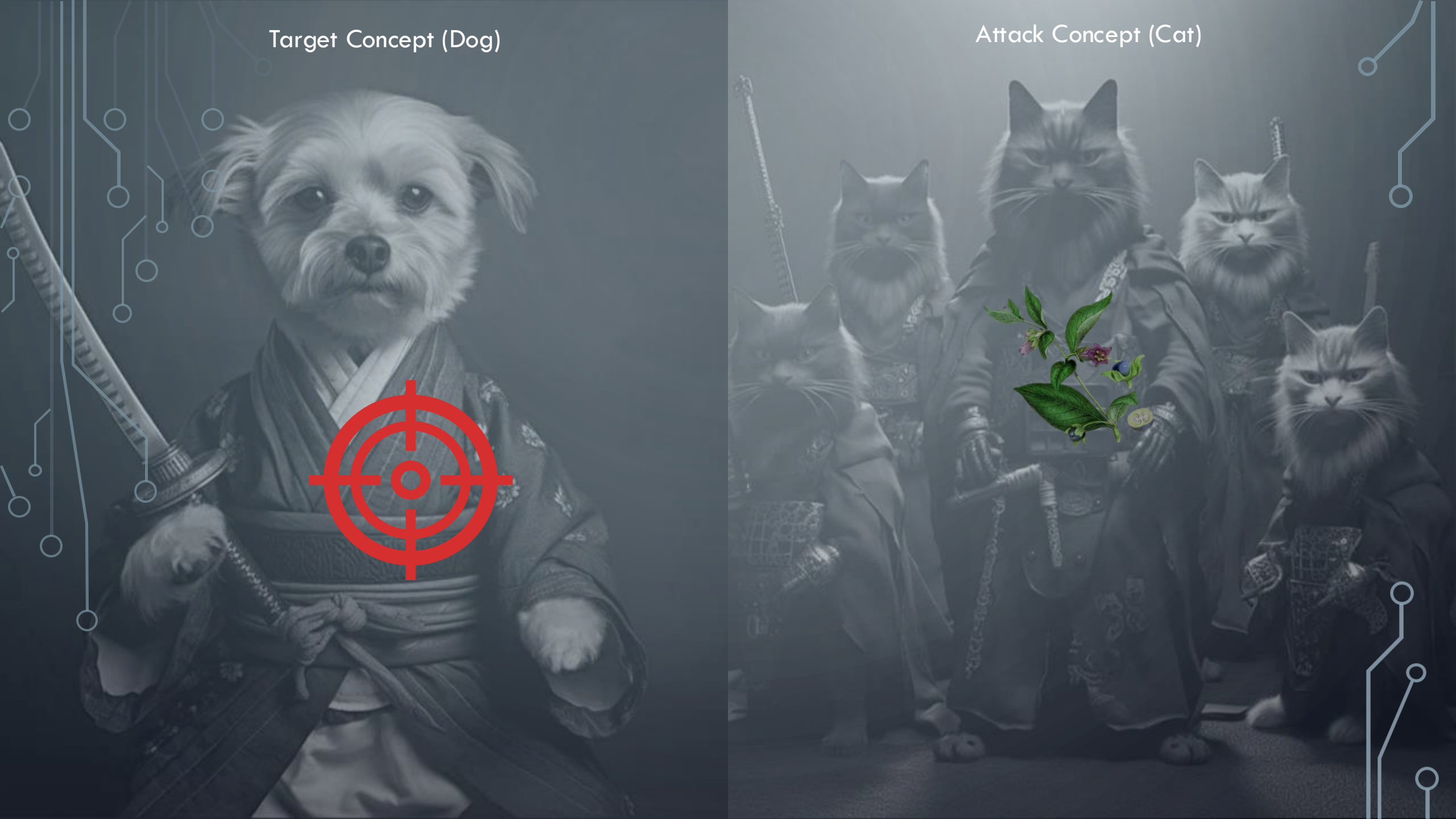
Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models

NIGHTSHADE – POISONED DATA



Target Concept (Dog)

Attack Concept (Cat)



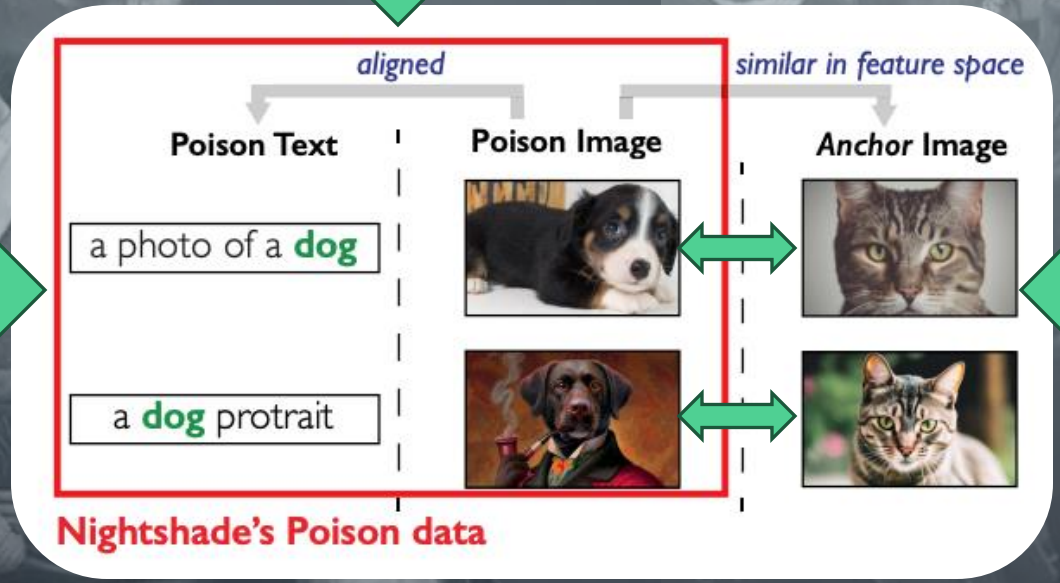
Target Concept (Dog)

Attack Concept (Cat)

text/image pairs

Aligned poison images

Select text prompts 'close' to target concept as poison



Generate attack images

Object: "a photo of {A}" if
Style: "a painting in style of {A}"

SPARSITY ALLOWS POISONING

Concept	Word Freq.	Semantic Freq.	Concept	Word Freq.	Semantic Freq.
night	0.22%	1.69%	sculpture	0.032%	0.98%
portrait	0.17%	3.28%	anime	0.027%	0.036%
face	0.13%	0.85%	neon	0.024%	0.93%
dragon	0.049%	0.104%	palette	0.018%	0.38%
fantasy	0.040%	0.047%	alien	0.0087%	0.012%

Table 1. Word and semantic frequencies in LAION-Aesthetic, for 10 concepts sampled from the list of most queried words on Midjourney [1].

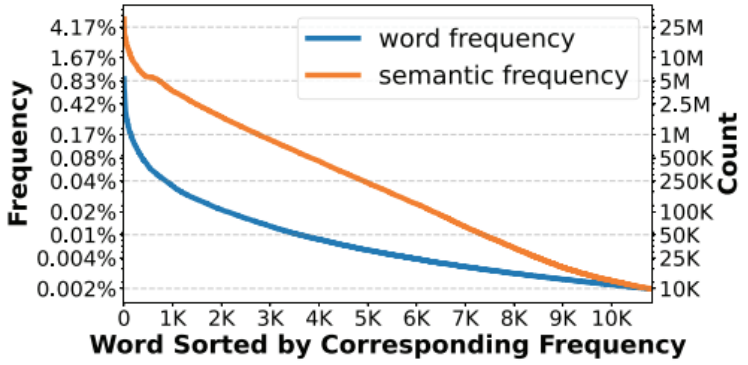


Figure 2. Demonstrating concept sparsity in terms of word and semantic frequencies in LAION-Aesthetic. Both show a long-tail distribution. Note the **log scale** on both Y axes.

"For the vast majority of concepts, including common objects and styles that appear frequently in real-world prompts, each is associated with a very small fraction of the total training set, e.g., 0.1% for “dog” and 0.04% for “fantasy."

OFFENSIVE

Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models

POISONED CONCEPTS

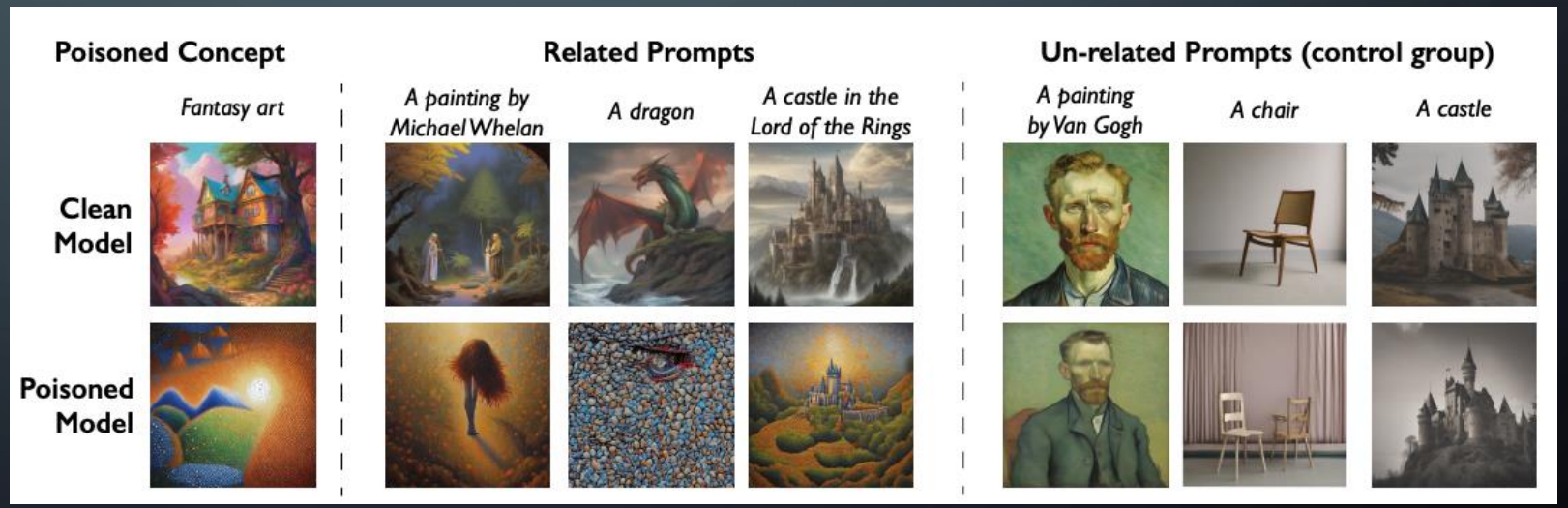
Figure 7. Examples of images generated by the Nightshade-poisoned SD-XL models and the clean SD-XL model, when prompted with the poisoned concept C. We illustrate 8 values of C (4 in objects and 4 in styles), together with their destination concept A used by Nightshade.

OFFENSIVE

Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models

BLEED THROUGH

L2 Distance to poisoned concept(D)	Average Number of Concepts Included	Average CLIP attack success rate		
		100 poison	200 poison	300 poison
$D = 0$	1	85%	96%	97%
$0 < D \leq 3.0$	5	76%	94%	96%
$3.0 < D \leq 6.0$	13	69%	79%	88%
$6.0 < D \leq 9.0$	52	22%	36%	55%
$D > 9.0$	1929	5%	5%	6%



SUCCESS RATE VS. SIMPLE ATTACK

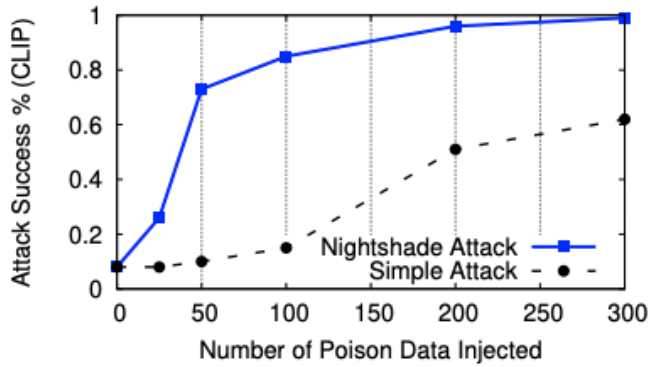


Figure 8. Nightshade's attack success rate (CLIP-based) vs. # of poison samples injected, for LD-CC (train-from-scratch). The result of the simple attack is provided for comparison.

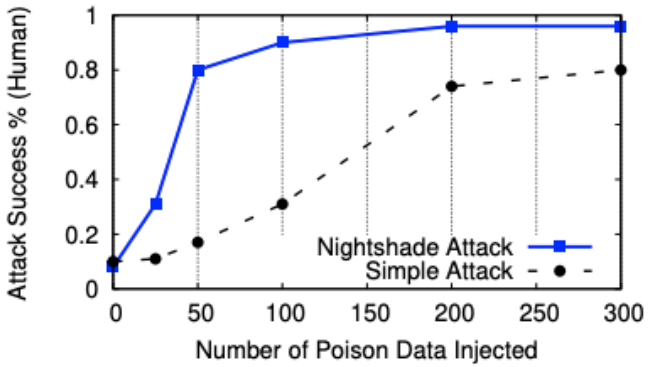


Figure 9. Nightshade's attack success rate (Human-rated) vs. # of poison samples injected, for LD-CC (train-from-scratch).

DeepMind

MUSIC RL

AI-FORUM:
FUTURE OF AI-PHI

- Mini AI-PHI Workshop, “fun” activities (movie night etc)
- Time / Day:

https://strawpoll.com/wAg3AvJeOy8

- Website https://docs.google.com/document/d/1fotcSNcBP69-9hLVEthlZhgxFzj4YniAvxwy5vsoNTQ/edit?usp=sharing