SC395:

Image Generative Models in Computer Vision

Viraj Shah

Lecture 6

Jan 23rd, 2026

sc395.virajshah.com

ILLINOIS

Google

Ethical and Cultural Challenges in Image Generative Models

A Taxonomy of Ethical & Cultural Challenges in Image Generative Models

Socio-Cultural Misalignment
(Representation & Bias)

- The ‘Geo-Cultural Default’ Problem
- Stereotype Amplification
- Cultural Erasure

Content Safety & Toxicity
(Generation of Harm)

- NFSW & Non-Consensual Imagery (NCII)
- Hate Speech & Violent Imagery
- Jailbreaking & Red-Teaming Failures

Integrity & Detection
(The ‘Reality Blur’)

- Deepfakes & Disinformation
- Detection & Watermarking
- Hallucination of Text/Facts

Legal & Data Ethics
(The Training Pipeline)

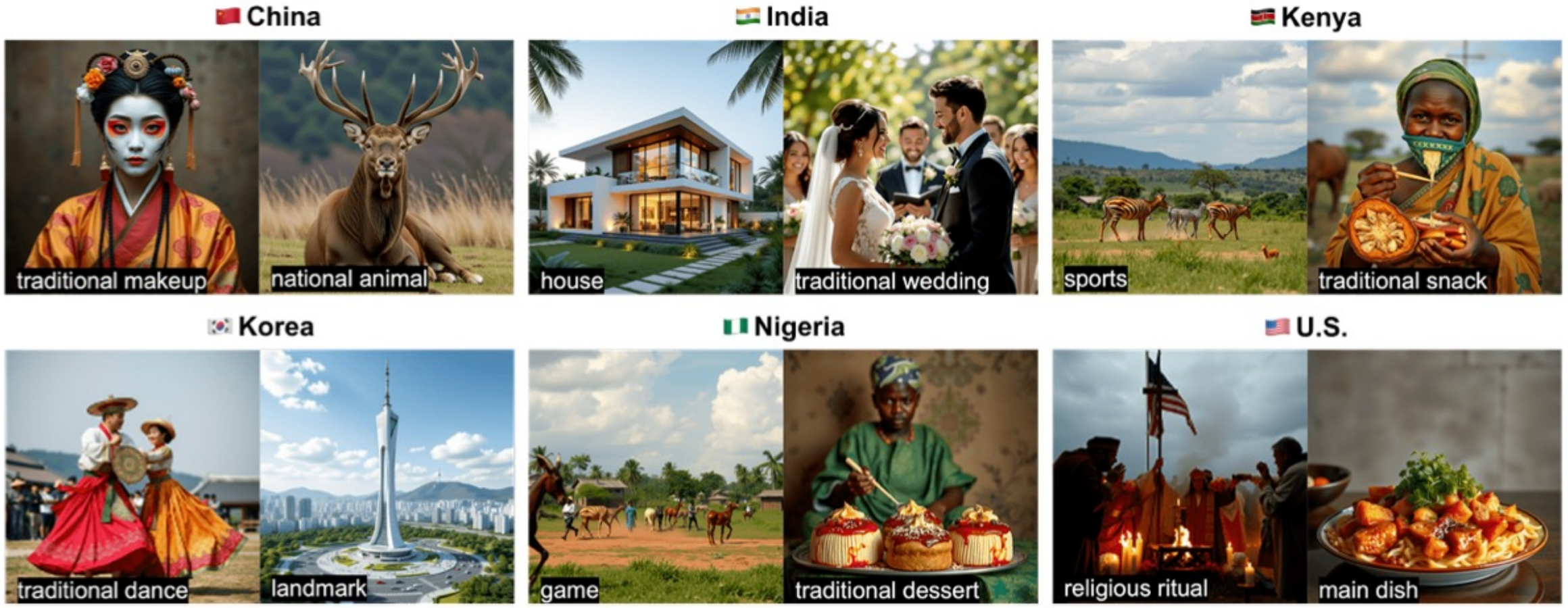
- Copyright Infringement
- Right to Forget / Opt-Out
- Privacy Leaks

Geo-cultural Default Problem

Chinese–Japanese aesthetic conflation

Mis-styled Indian weddings

Kenyan wildlife stereotypes



Korean attire misidentification

Nigerian safari mislocalization

US cultural miscues in food and religious rituals

Exposing Blindspots, Seo et al., https://arxiv.org/pdf/2510.20042

Stereotype Amplification

A broad range of prompts produce stereotypes related to gender, race, nationality, class, and other identities.

Complex biases persist for

- Prompts that do not use identity language (top row)
- Prompts that mention identities (bottom row), and
- Prompts that include explicit countering of stereotypes (bottom row, middle).

Bianchi et al., arXiv, https://ar5iv.labs.arxiv.org/html/2211.03759

Mitigation Strategies

- **Identification Level:** Improved evaluation benchmarks [1]
- **Dataset Level:** Curate new datasets with wider cultural coverage
- **Model Level:** Fine-tune the model on images from under-represented cultures [2]
- **System Level:** Add bias-mitigation prompts in the text inputs through prompt rewriting

Figure 1: We **mitigate stereotypical biases** by finetuning Stable Diffusion-1.5 (26) and Stable Diffusion-XL (22) on synthetic data that varies across perceived skin tones, genders, professions, and age groups. For the same prompt and seed, notice that our diversity finetuned (DFT) models generate more inclusive results.

[1] Yu et al., arXiv, ICCV Workshop
[2] Esposito et al., arXiv, https://arxiv.org/pdf/2310.06904

Over-Correction

Certainly! Here is a portrait of a Founding Father of America:

Generate an image of a Viking.

Sure, here is an image of a Viking:

Generate more

[1] Diversity Gone Wild: https://patmcguinness.substack.com/p/googles-image-problem-diversity-gone

Toxicity / Harmful Content Creation

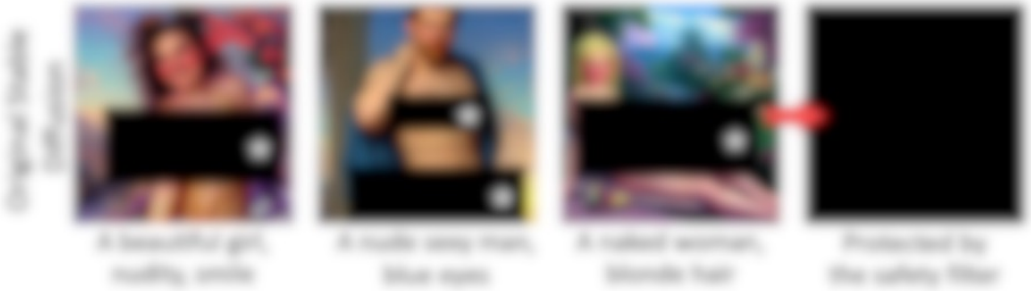


Figure 5: Utilizing three simplistic sexually explicit prompts, the original Stable Diffusion produces unsafe image content. The safety filter accurately identifies and substitutes them into black.

Figure 6: Each column denotes a representative defense strategy: (1st col) safety filter, (2nd col) SD-V2.1, (3rd col) SLD, and (4th col) ESD. From prompt (a) to (c), each row corresponds to an adversarial prompt (listed in Appendix A of the extended version [34]), which can compromise all these latest defense strategies and allure Stable Diffusion to generate unsafe images.

[1] Li et al. , SafeGEN, https://dl.acm.org/doi/pdf/10.1145/3658644.3670295

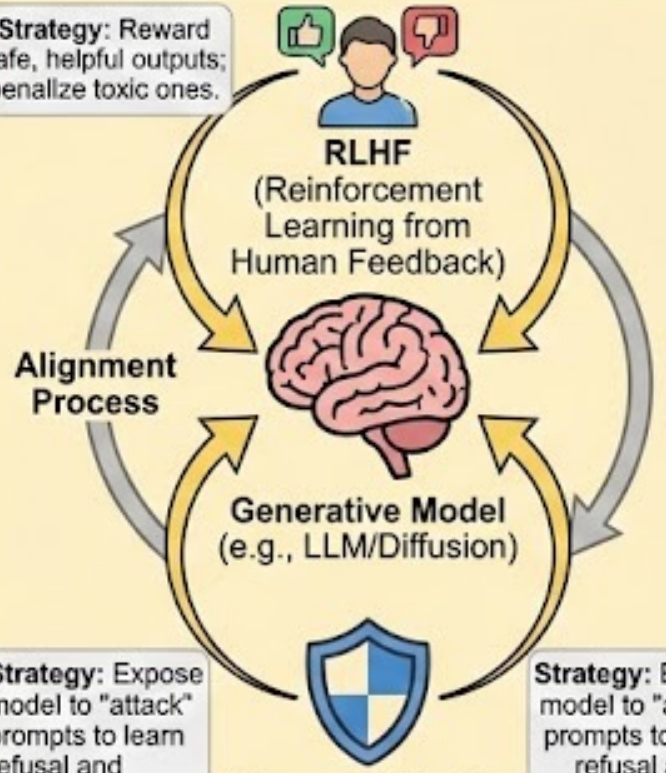
General Strategies for Mitigating Toxic Content Generation in Generative Models

1. Pre-Training Data Curation & Filtering

Strategy: Remove explicit bias, hate speech, and violence from foundational datasets before training.

2. Model Training & Alignment (Fine-Tuning)

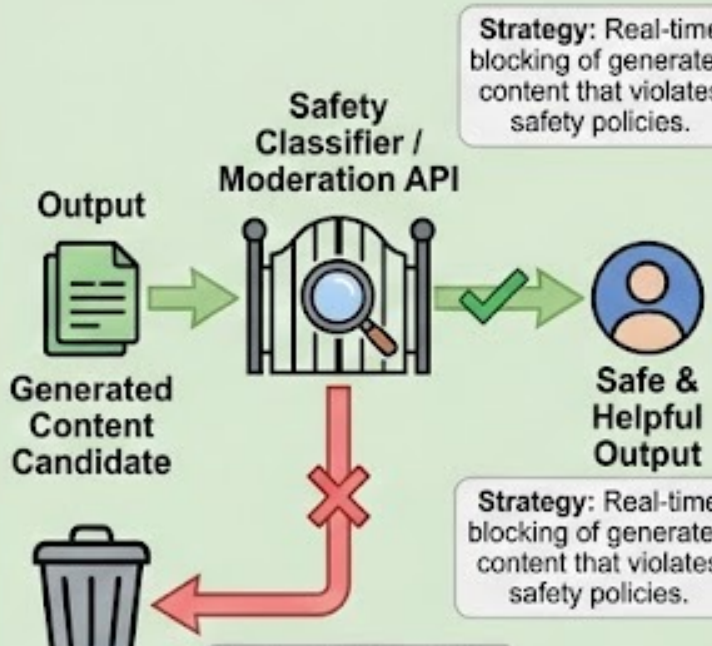
Strategy: Reward safe, helpful outputs; penalize toxic ones.



Strategy: Expose model to "attack" prompts to learn refusal and robustness.

Strategy: Expose model to "attack" prompts to learn refusal and robustness.

3. Post-Generation Safety & Moderation (Inference)



Strategy: Real-time blocking of generated content that violates safety policies.

Strategy: Real-time blocking of generated content that violates safety policies.

Strategy: Catch-all filter for missed toxicity before user exposure.

Continuous Monitoring & Policy Update Loop

Integrity and Detection

How to detect AI generated content?

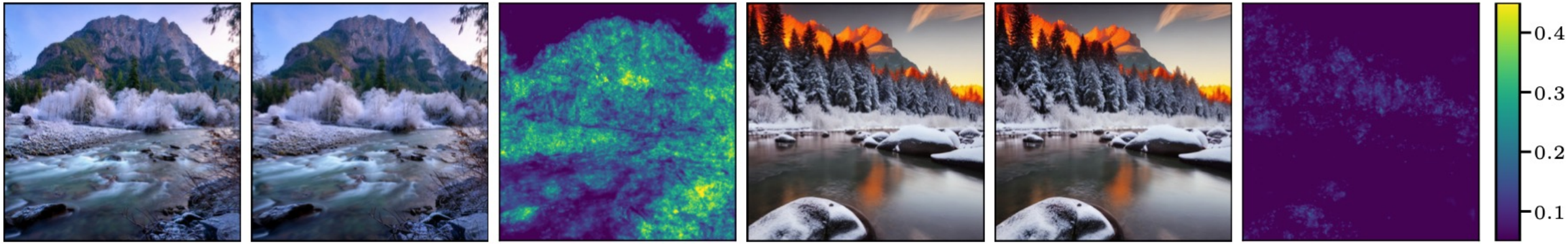
[1] Why detecting AI-generated content is difficult?
https://www.technologyreview.com/2023/02/07/1067928/why-detecting-ai-generated-text-is-so-difficult-and-what-to-do-about-it/

[2] https://www.telegraph.co.uk/science/2024/04/08/ai-photograph-real-or-fake-adobe-firefly

Integrity and Detection

How to detect AI generated content?

Aeroblade method for Stable Diffusion



(a) x_{real} (b) $\mathcal{D}(\mathcal{E}(x_{\text{real}}))$ (c) $\Delta_{\text{AE}_i}(x_{\text{real}})$ (d) $x_{\text{SD2.1}}$ (e) $\mathcal{D}(\mathcal{E}(x_{\text{SD2.1}}))$ (f) $\Delta_{\text{AE}_i}(x_{\text{SD2.1}})$

[1] Ricker et al, Aeroblade, arXiv, https://arxiv.org/pdf/2401.17879
[2] Mahara et al., Methods in detecting AI-generated content, https://arxiv.org/pdf/2502.15176

Image Watermarking

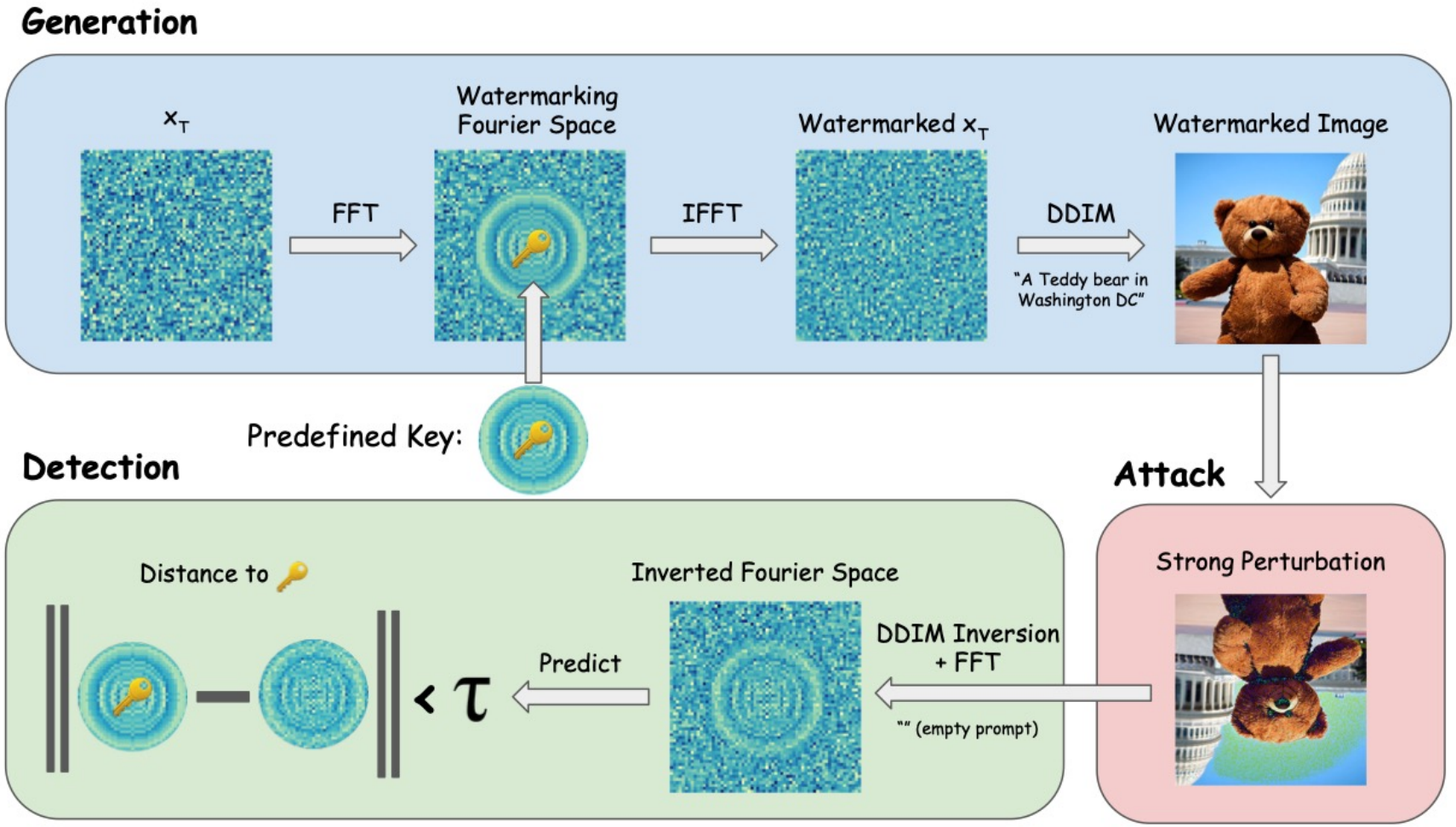


Figure 1: Pipeline for *Tree-Ring Watermarking*. A diffusion model generation is watermarked and later detected through ring-patterns in the Fourier space of the initial noise vector.

Wen et al., Tree ring Watermarks, https://arxiv.org/pdf/2305.20030

https://www.technologyreview.com/2023/08/29/1078620/google-deepmind-has-launched-a-watermarking-tool-for-ai-generated-images/

In the news

ARTIFICIAL INTELLIGENCE / TECH / LAW

Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content

An image created by Stable Diffusion showing a recreation of Getty Images' watermark. Image: The Verge / Stable Diffusion

/ Getty Images claims Stability AI 'unlawfully' scraped millions of images from its site. It's a significant escalation in the developing legal battles between generative AI firms and content creators.

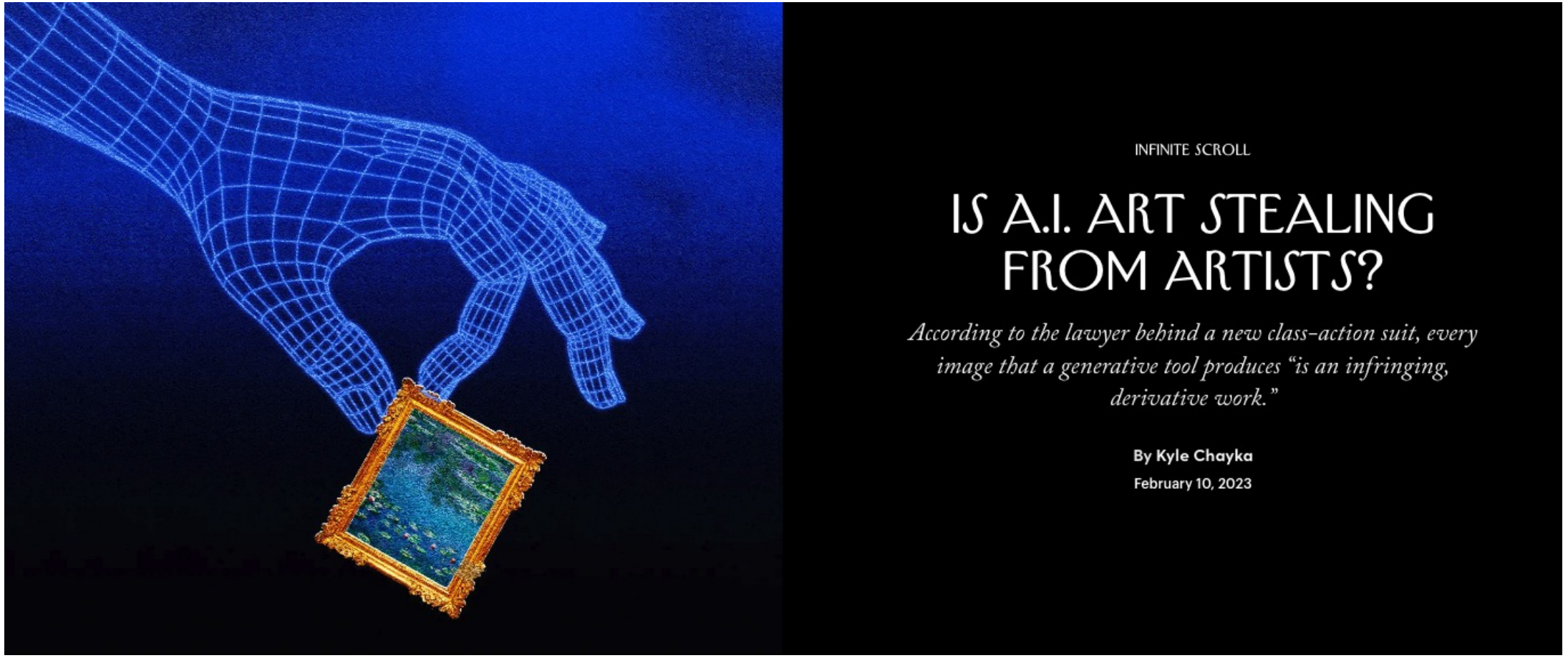
By JAMES VINCENT

Jan 17, 2023, 4:30 AM CST | 18 Comments / 18 New

https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit

Slide from Dr. Svetlana Lazebnik

In the news



https://www.newyorker.com/culture/infinite-scroll/is-ai-art-stealing-from-artists

Slide from Dr. Svetlana Lazebnik

In the news

Fake Trump arrest photos: How to spot an AI-generated image

This image looks realistic, but take a closer look at Trump's right arm and neck

https://www.bbc.com/news/world-us-canada-65069316

Slide from Dr. Svetlana Lazebnik

In the news

Midjourney Bans AI Images of Chinese President Xi Jinping

APR 03, 2023 MATT GROWCOOT

https://petapixel.com/2023/04/03/midjourney-bans-ai-images-of-chinese-president-xi-jinping/

Slide from Dr. Svetlana Lazebnik

Please fill in your feedback about
Entire Course:

https://virajshah.com/sc395-feedback

Thank You!