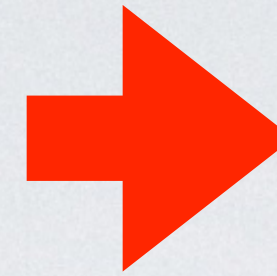




Nils Lukas



Abdulrahman Diaa



Lucas Fenaux



Florian Kerschbaum



UNIVERSITY OF
WATERLOO



CrySP
Cryptography, Security, and Privacy
Research Group

LEVERAGING OPTIMIZATION FOR ADAPTIVE ATTACKS ON IMAGE WATERMARKS

Nils Lukas, Abdulrahman Diaa*, Lucas Fenaux*, Florian Kerschbaum
University of Waterloo, Canada
{nilukas, abdulrahman.diaa, lucas.fenaux,
florian.kerschbaum}@uwaterloo.ca

ABSTRACT

Untrustworthy users can misuse image generators to synthesize high-quality deepfakes and engage in unethical activities. Watermarking deters misuse by marking generated content with a hidden message, enabling its detection using a secret watermarking key. A core security property of watermarking is robustness, which states that an attacker can only evade detection by substantially degrading image quality. Assessing robustness requires designing an adaptive attack for the specific watermarking algorithm. When evaluating watermarking algorithms and their (adaptive) attacks, it is challenging to determine whether an adaptive attack is optimal, i.e., the best possible attack. We solve this problem by defining an objective function and then approach adaptive attacks as an optimization problem. The core idea of our adaptive attacks is to replicate secret watermarking keys locally by creating *surrogate keys* that are differentiable and can be used to optimize the attack's parameters. We demonstrate for Stable Diffusion models that such an attacker can break all five surveyed watermarking methods at no visible degradation in image quality. Optimizing our attacks is efficient and requires less than 1 GPU hour to reduce the detection accuracy to 6.3% or less. Our findings emphasize the need for more rigorous robustness testing against adaptive, learnable attackers.

1 INTRODUCTION

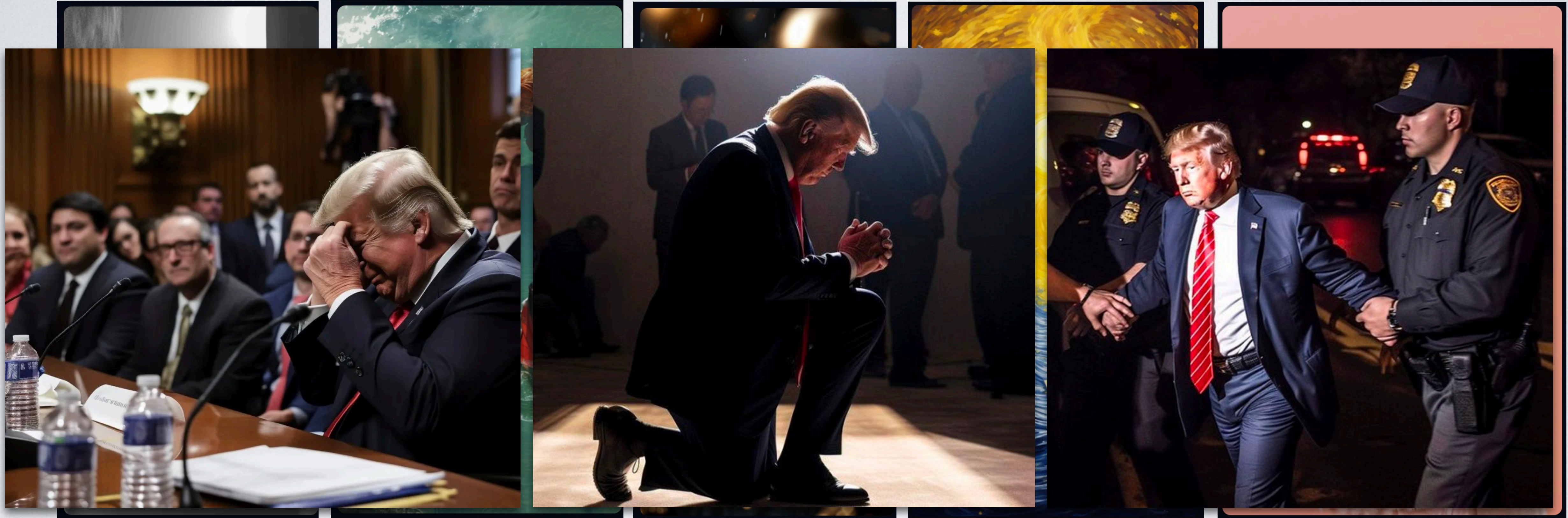
Deepfakes are images synthesized using deep image generators that can be difficult to distinguish from real images. While deepfakes can serve many beneficial purposes if used ethically, for example, in medical imaging (Akrouf et al., 2023) or education (Peres et al., 2023), they also have the potential to be *misused* and erode trust in digital media. Deepfakes have already been used in disinformation campaigns (Boneh et al., 2019; Barrett et al., 2023) and social engineering attacks (Mirsky & Lee, 2021), highlighting the need for methods that control the misuse of deep image generators.

Watermarking offers a solution to controlling misuse by embedding hidden messages into all generated images that are later detectable using a secret watermarking key. Images detected as deepfakes can be flagged by social media platforms or news agencies, which can mitigate potential harm (Grinbaum & Adomaitis, 2022). Providers of large image generators such as Google have announced the deployment of their own watermarking methods (Gowal & Kohli, 2023) to enable the detection of deepfakes and promote the ethical use of their models, which was also declared as one of the main goals in the US government's "AI Executive Order" (Federal Register, 2023).

A core security property of watermarking is *robustness*, which states that an attacker can evade detection only by substantially degrading the image's quality. While several watermarking methods have been proposed for image generators (Wen et al., 2023; Zhao et al., 2023; Fernandez et al., 2023), none of them are certifiably robust (Bansal et al., 2022) and instead, robustness is tested empirically using a limited set of known attacks. Claimed security properties of previous watermarking methods have been broken by novel attacks (Lukas et al., 2022), and no comprehensive method exists to validate robustness, which causes difficulty in trusting the deployment of watermarking in practice. We propose testing the robustness of watermarking by defining robustness using objective

*Equal Contribution

Image Generators - Dual Use



Generated Using Midjourney [12/2023]

Generated Using Midjourney [03/2023]

Social Media Disinformation

How a fake image of a Pentagon explosion shared on Twitter caused a real dip on Wall Street

Euronews [05/2023]

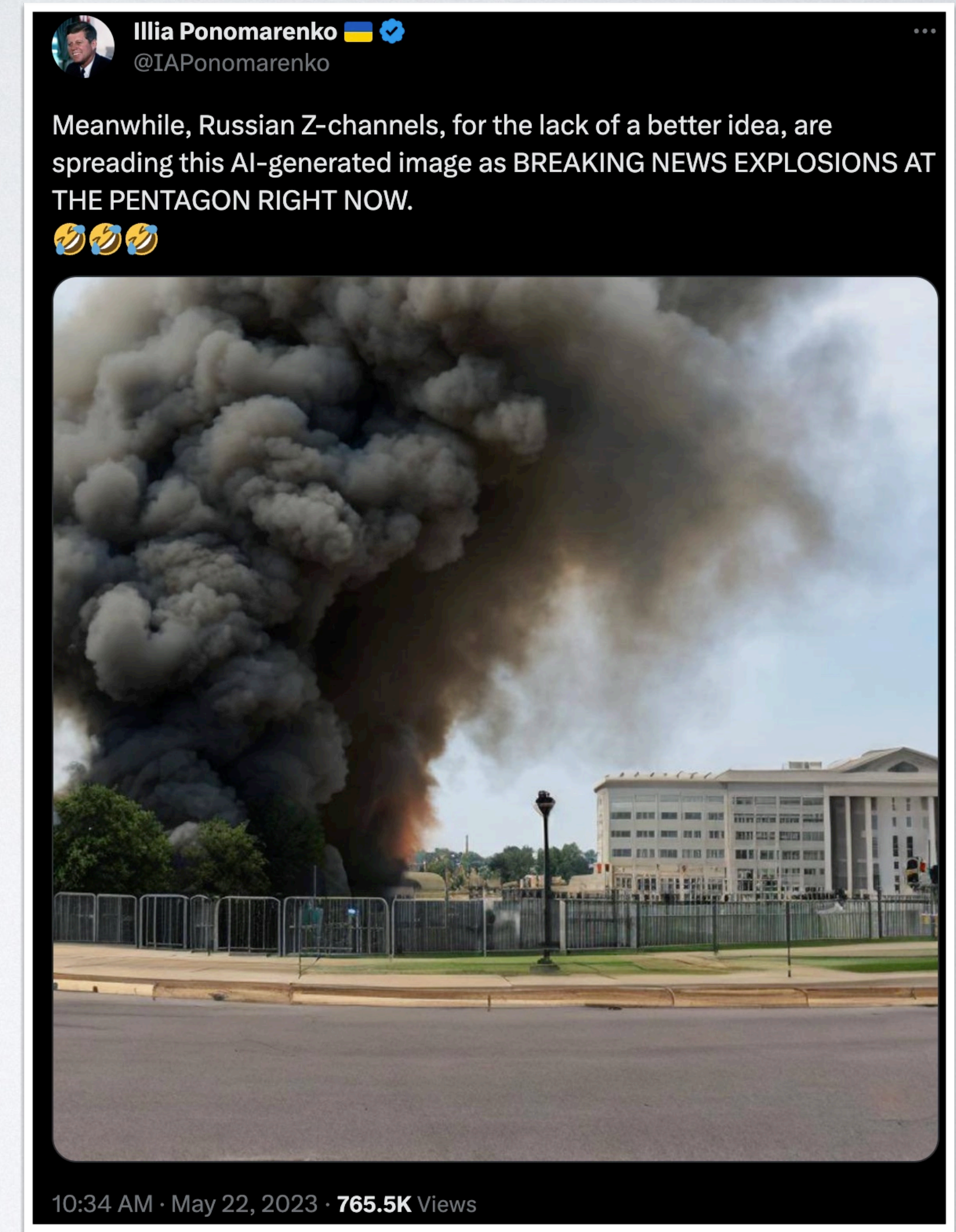
People are trying to claim real videos are deepfakes. The courts are not amused

MAY 8, 2023 · 5:01 AM ET

HEARD ON [ALL THINGS CONSIDERED](#)



Reuters [05/2023]



AI Regulations


- (i) authenticating content and tracking its provenance;
- (ii) labeling synthetic content, such as using watermarking;
- (iii) detecting synthetic content;

AI Executive Order [\[10/2023\]](#)



United States of
America

Watermarking Pledge




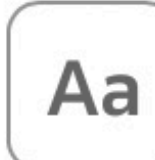

REUTERS® World ▾ Business ▾ Markets ▾ Sustainability ▾ Legal ▾ Breakingviews Technology ▾ Inv

Technology

OpenAI, Google, others pledge to watermark AI content for safety, White House says

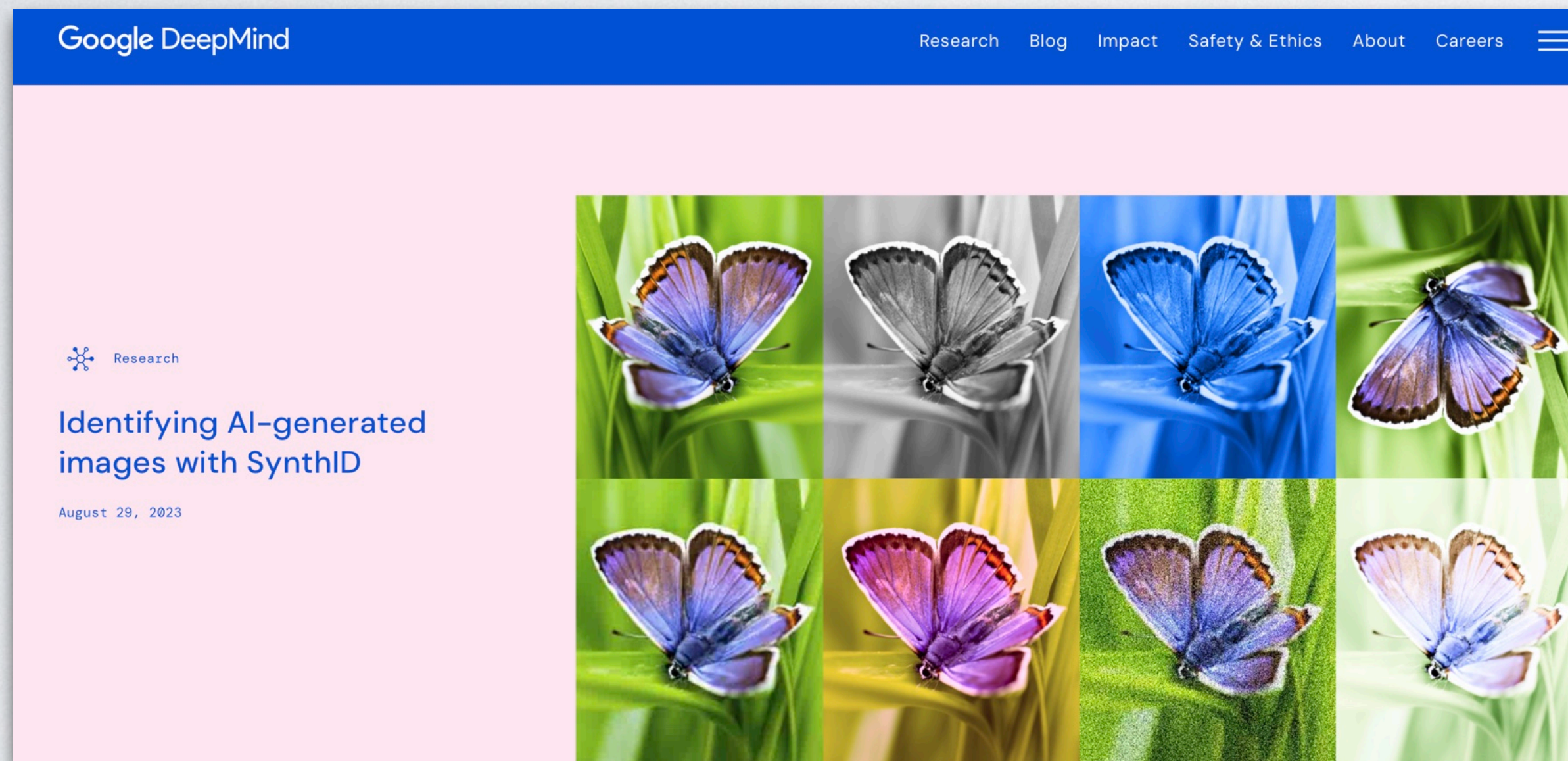
By Diane Bartz and Krystal Hu

July 21, 2023 1:44 PM PDT · Updated 19 days ago

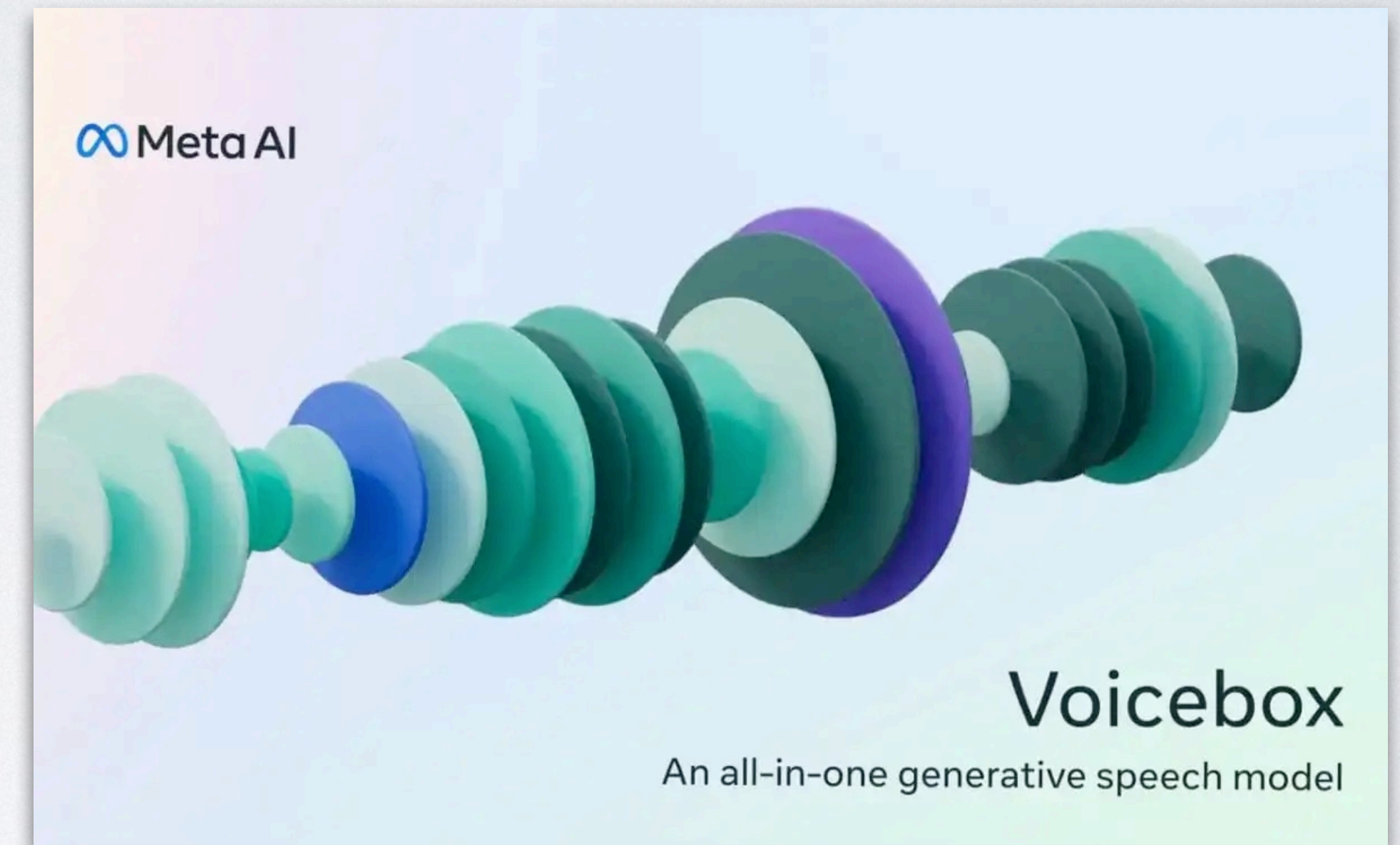
  

Reuters [07/2023]

Watermarking in Use



Google SynthID [08/2023]

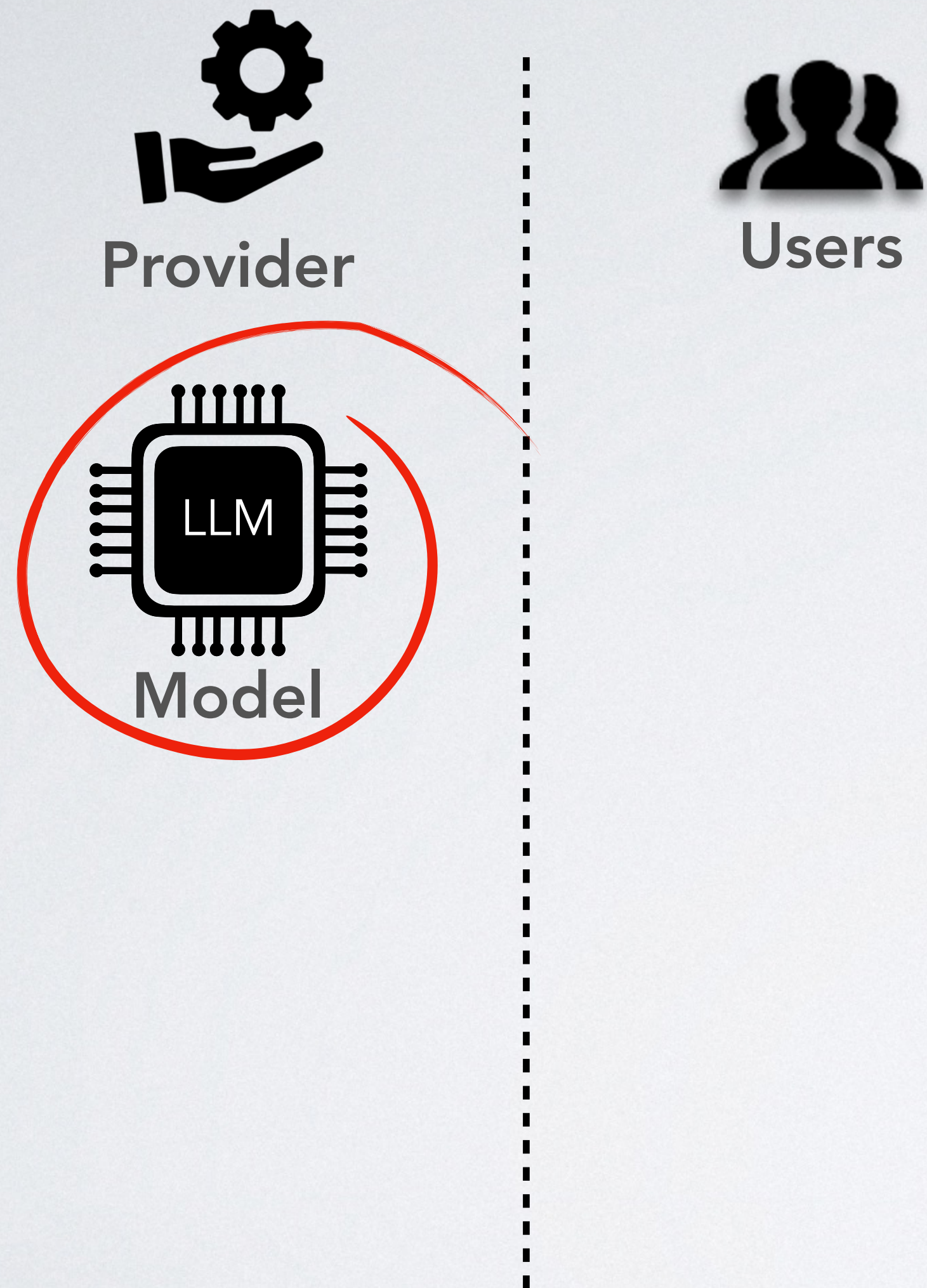


Meta AI VoiceBox [06/2023]

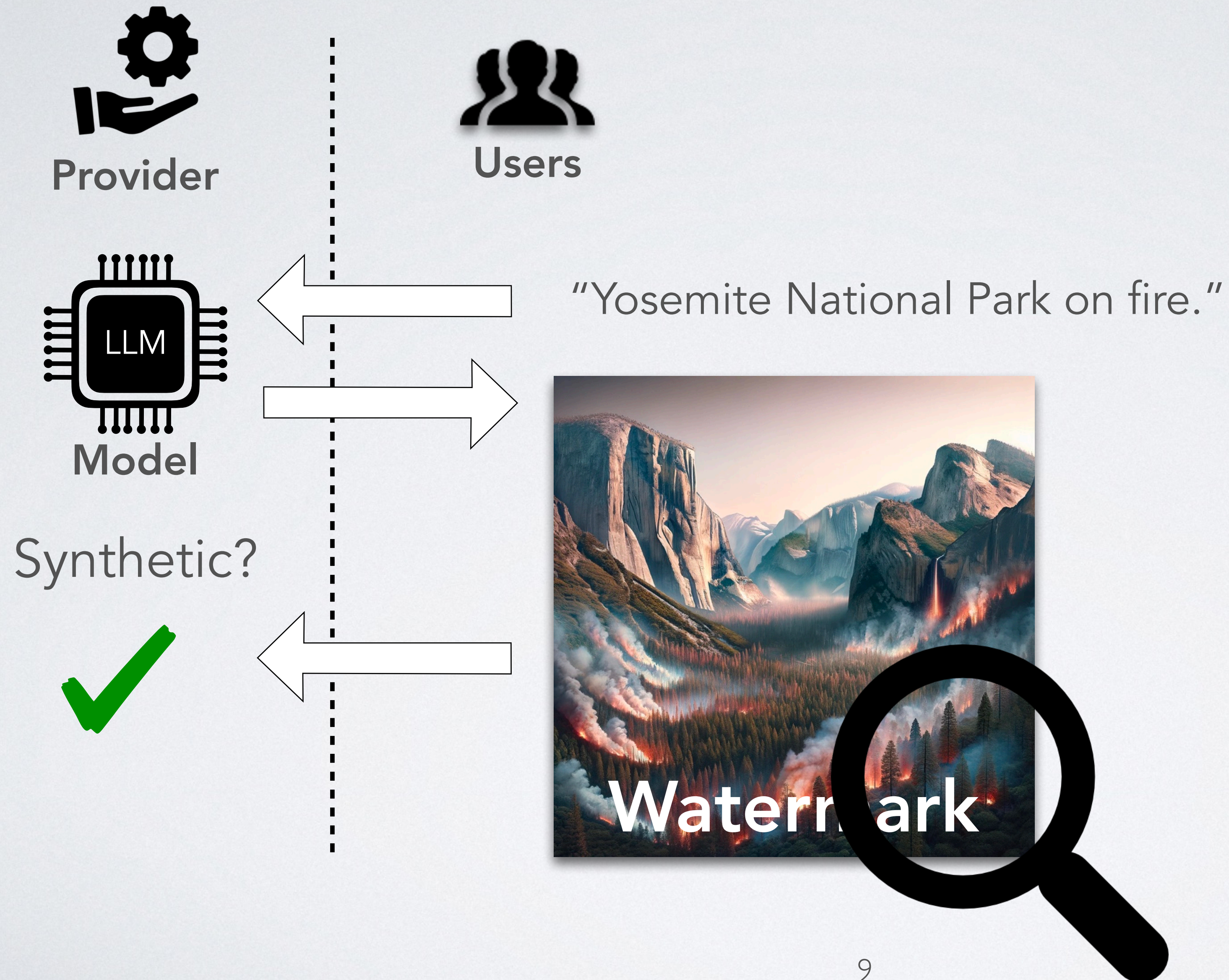
Watermarking



Watermarking

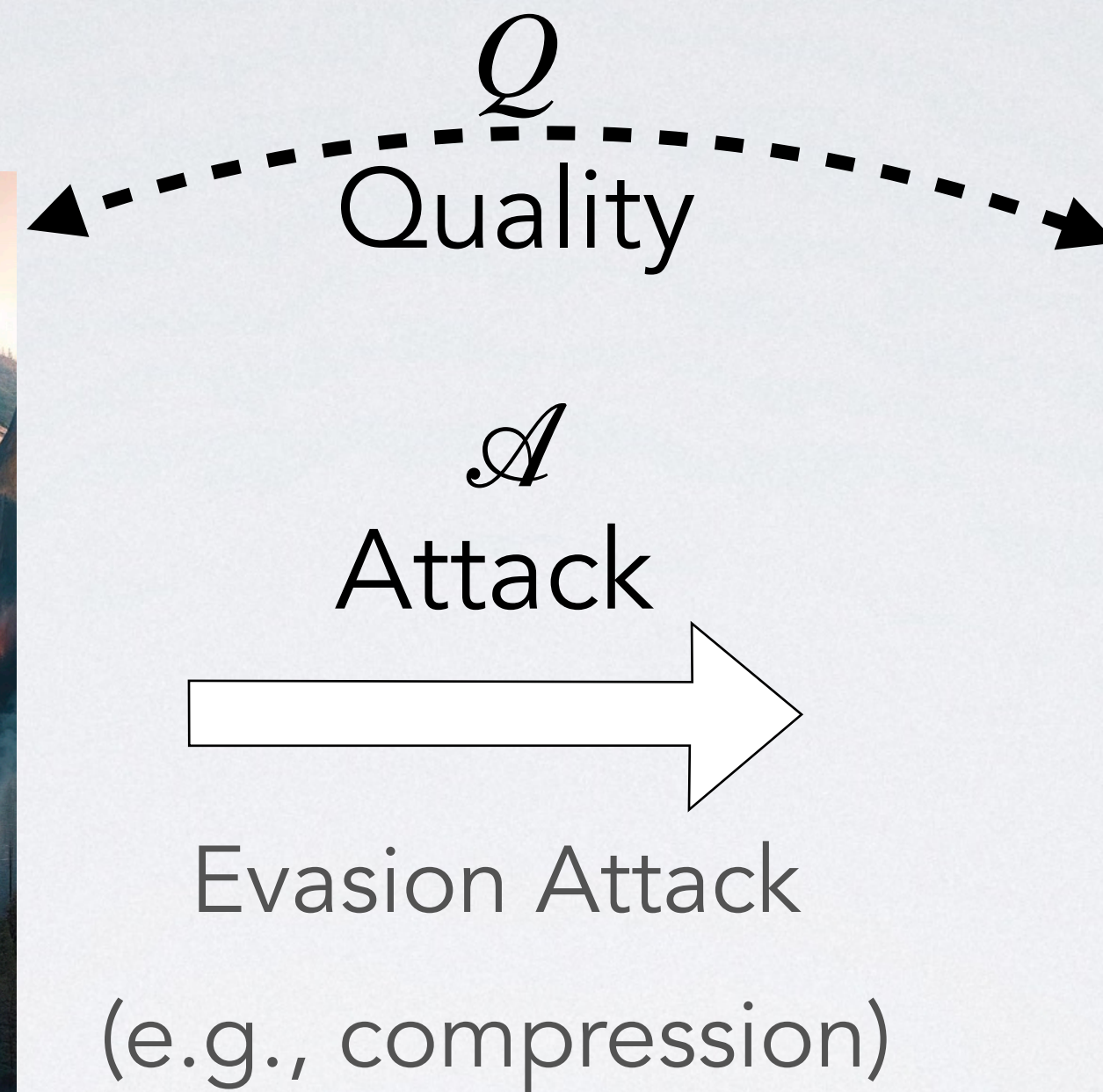
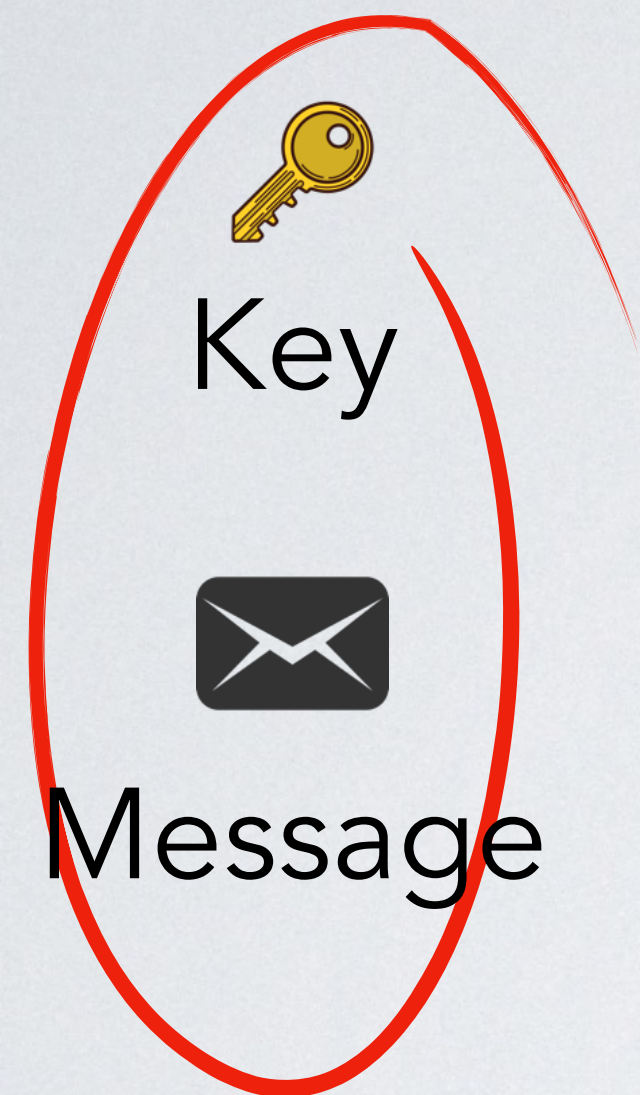


Watermarking



Robustness

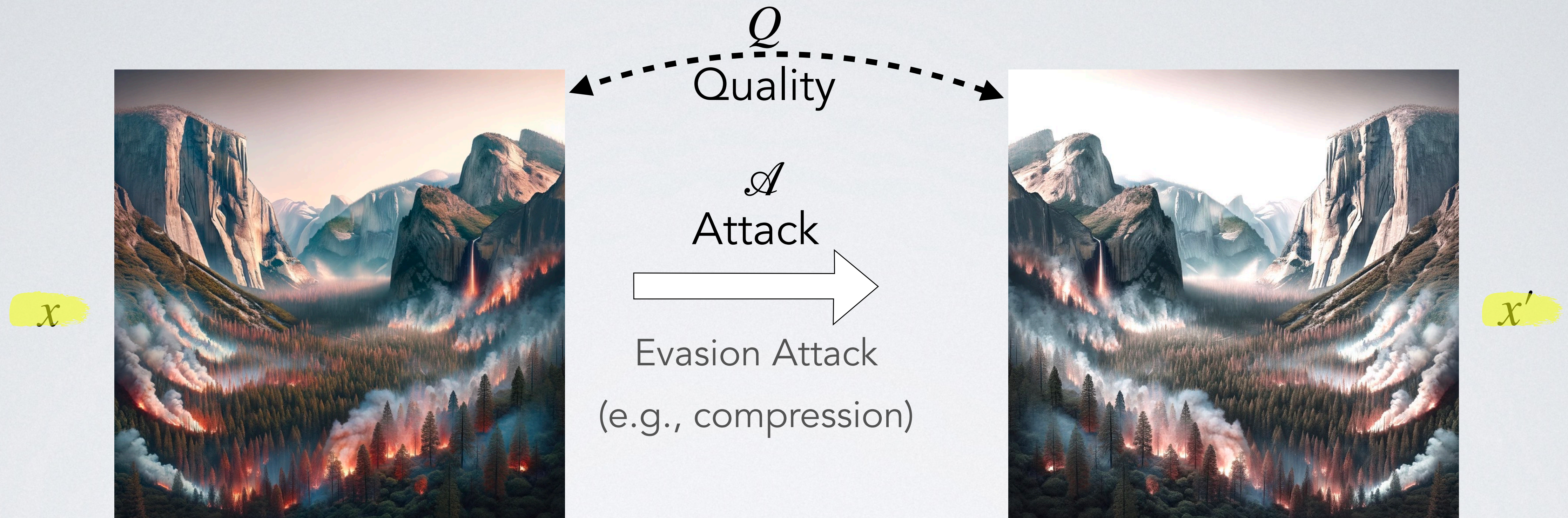
Watermarking relies
on uncertainty



$$\text{Verify}(\text{Image}, \text{Key}, \text{Message}) = 1$$

$$\text{Verify}(\text{Image}, \text{Key}, \text{Message}) = 0$$

Robustness



$$Pr[\text{Verify}(x', \text{key}, \text{envelope}) = 0 \text{ and } Q(x, x') = 1]$$

Image after an attack evades detection

Preserves quality

Robustness against Adaptive Attacks

$$Pr[\text{Verify}(x', \text{key}, \text{msg}) = 0 \text{ and } Q(x, x') = 1]$$

Adaptive attacks know the watermarking method,
but not the secret key or message.

Research Question

How robust are image watermarks against adaptive, learnable attacks?

Visual Inspection

TRW

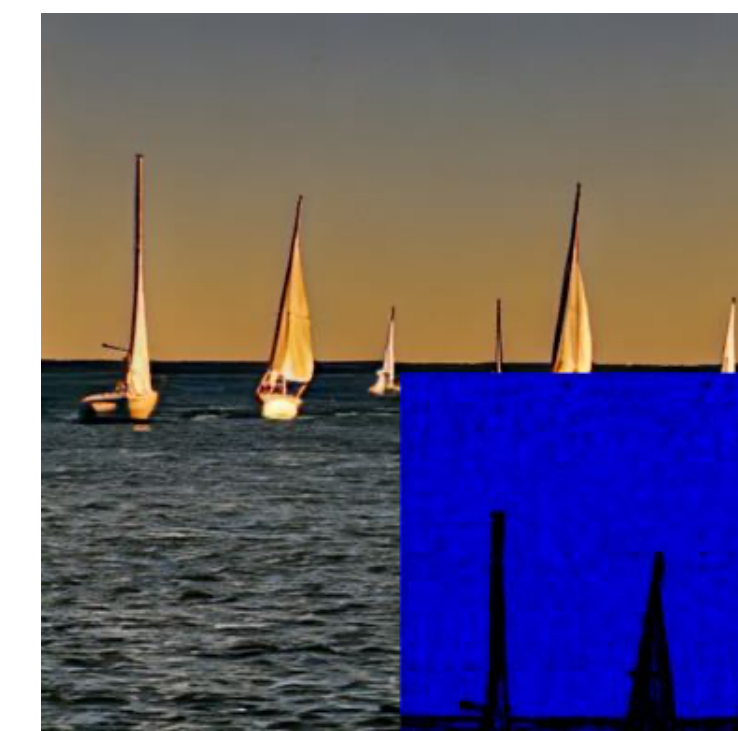
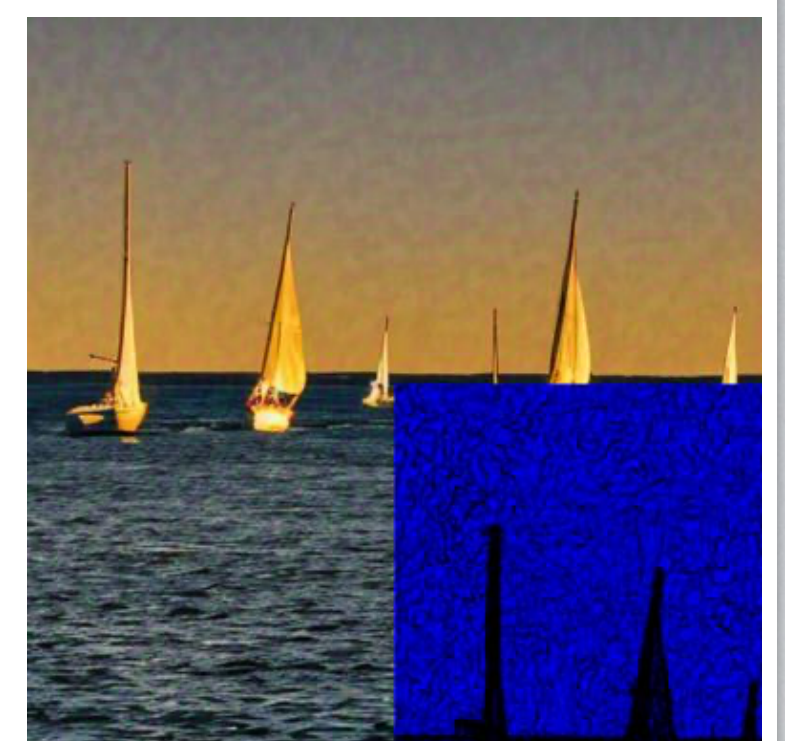
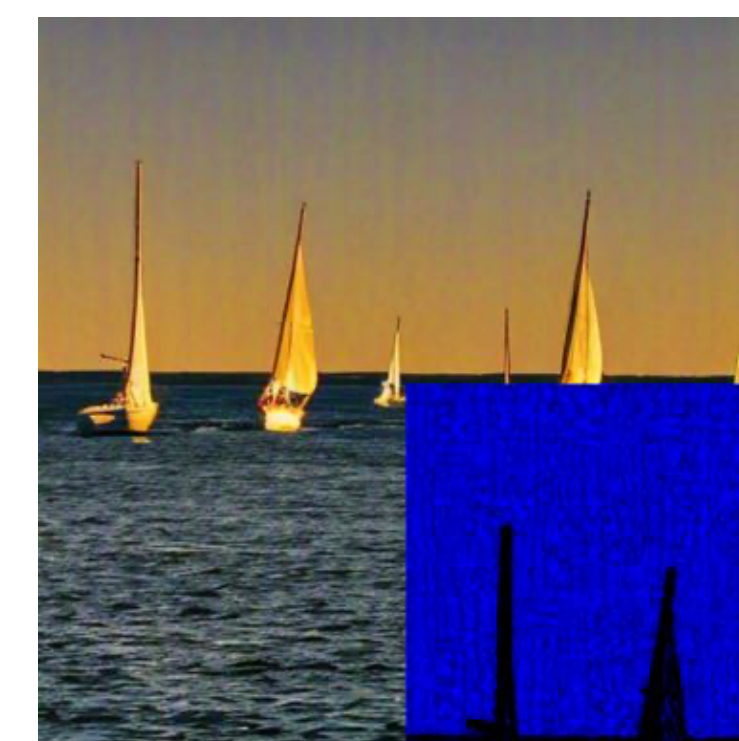
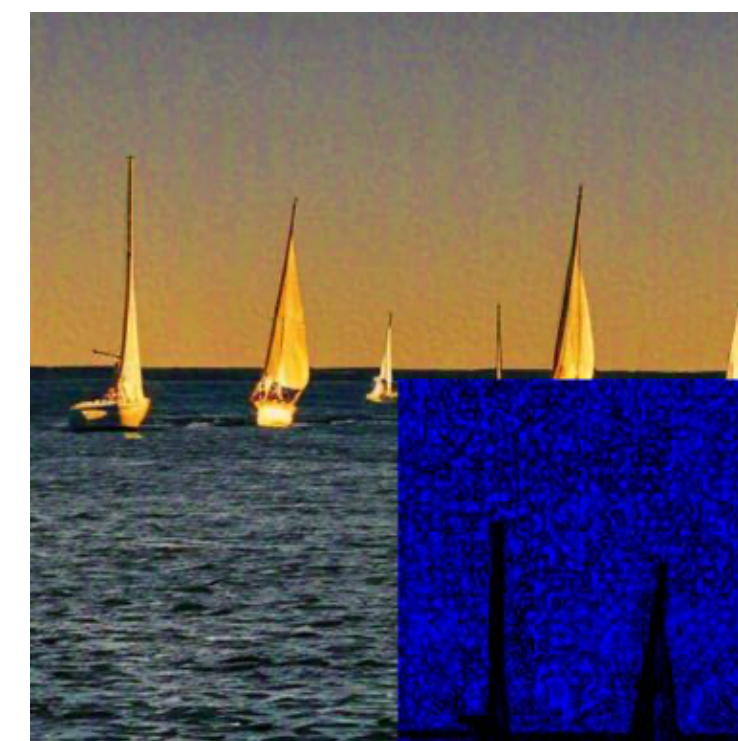
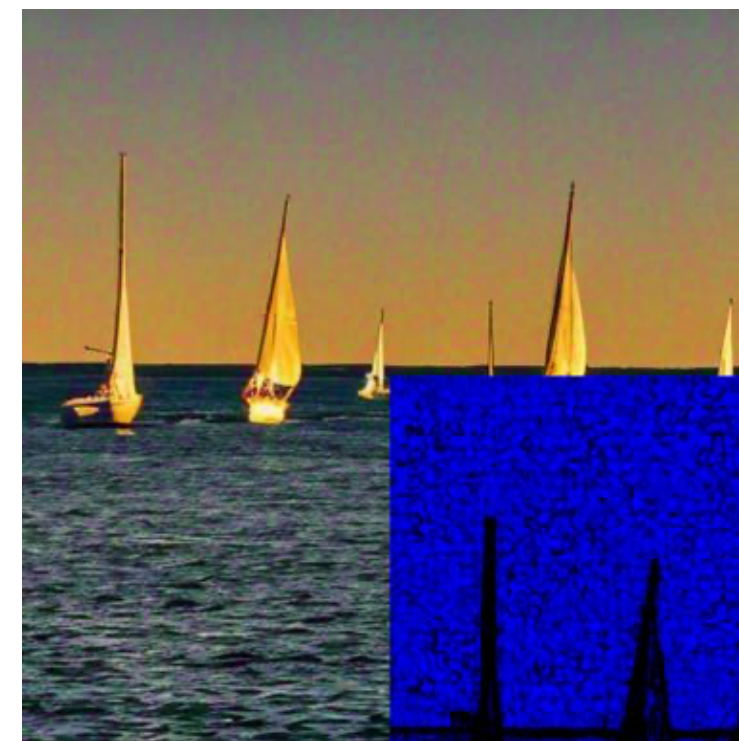
WDM

DWT

DWT-SVD

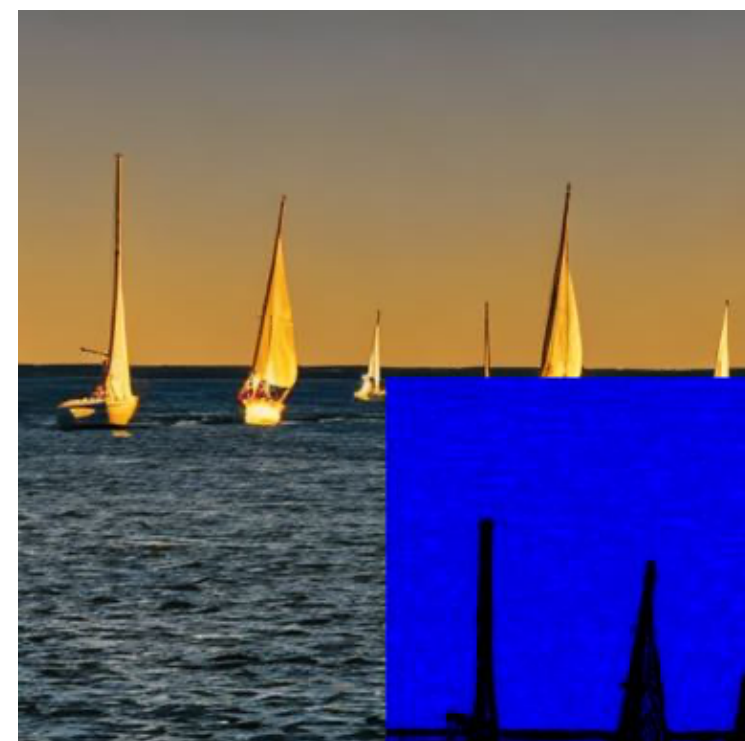
RivaGAN

Adversarial
Noise



No Watermark

Adversarial
Compression



Summary

Adaptive attacks can be formulated as a general optimization problem

Our attacks require no interaction with the model provider

Optimization is computationally efficient (<15 minutes on 1 GPU)

Our attacks account for the attacker's access to open-source models

Leveraging Optimization for Adaptive Attacks Against Image Watermarks



Speaker: Nils Lukas

Poster Session - Halle B
Tue 7 May 10:45 a.m. CEST



[Paper Link](#)