How Reliable is Watermarking for Generative Machine Learning?

Nils Lukas

CrySP
Cryptography, Security, and Privacy Research Group

UNIVERSITY OF WATERLOO

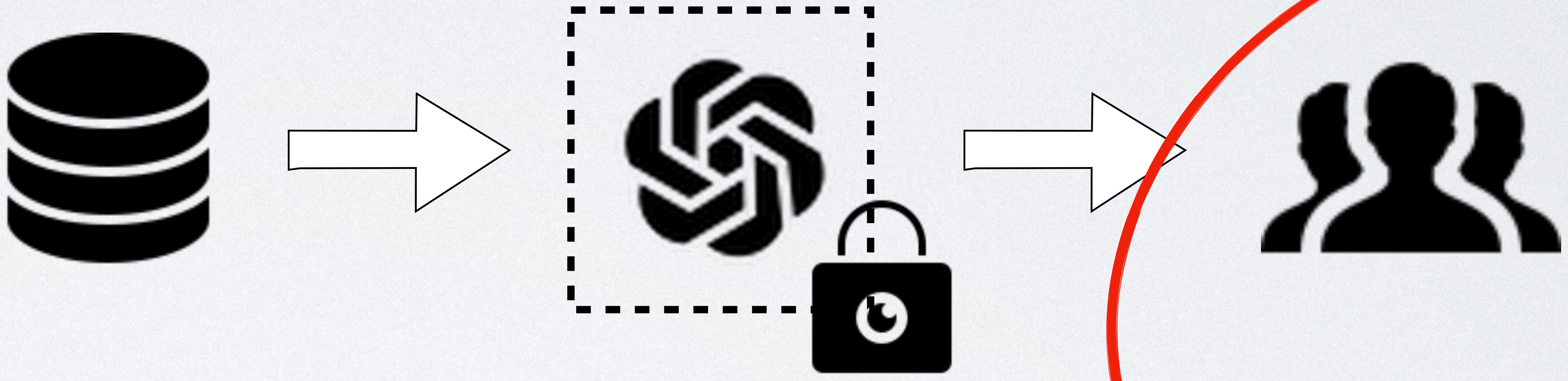
My Areas of Research

Homepage

- Private Computation
  - Private Set Intersection
  - Secure Inference

- **Trustworthy Machine Learning**

Nils Lukas



Untrustworthy Data

Untrustworthy Models

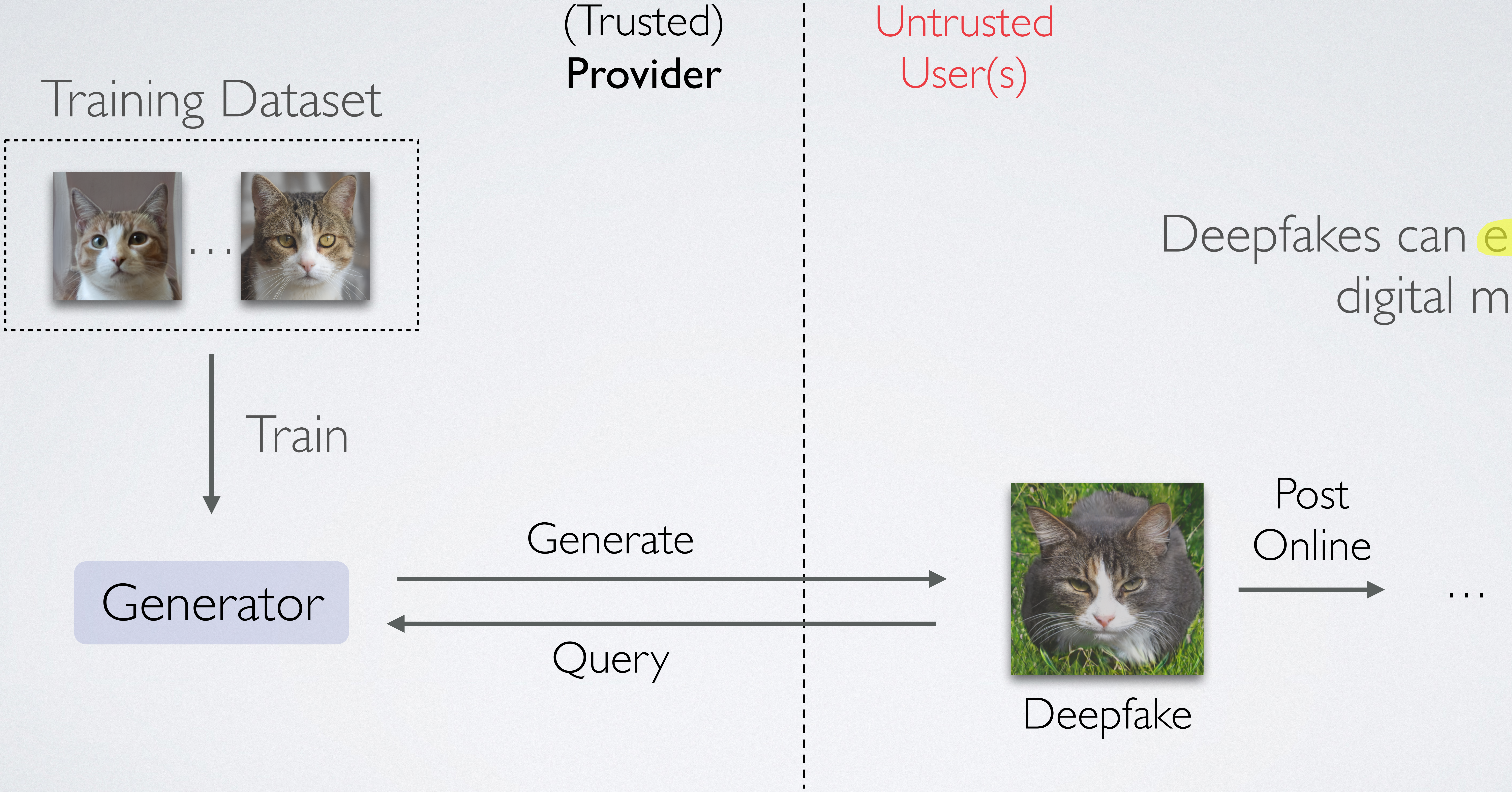
Untrustworthy Users

CrySP

Cryptography, Security, and Privacy Research Group

UNIVERSITY OF WATERLOO

Controlling Misuse



Deepfakes can erode trust in digital media

Disinformation

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

Euronews, May 2023 [2]

The viral AI-generated image showing an explosion near the Pentagon is 'truly the tip of the iceberg of what's to come,' tech CEO says

Grace Dean Jun 9, 2023, 6:33 AM EDT

Business Insider, June 2023 [3]

Fake Pentagon explosion photo goes viral: How to spot an AI image

Aljazeera, May 2023 [4]

Illia Ponomarenko 🇺🇦
@IAPonomarenko

Meanwhile, Russian Z-channels, for the lack of a better idea, are spreading this AI-generated image as BREAKING NEWS EXPLOSIONS AT THE PENTAGON RIGHT NOW.
🤣🤣🤣

10:34 AM · May 22, 2023 · **765.5K** Views

Personalized Attacks

Deepfake porn could be a growing problem amid AI race

APN news, April 2023 [15]

EXCLUSIVE

INTERNET

Deepfake porn of TikTok stars thrives on Twitter even though it breaks the platform's rules

Young TikTok stars have become a focus of nonconsensual pornographic deepfake creators.

NBC, June 2023 [16]

Public Service Announcement

FEDERAL BUREAU OF INVESTIGATION

June 5, 2023

Alert Number
I-060523-PSA

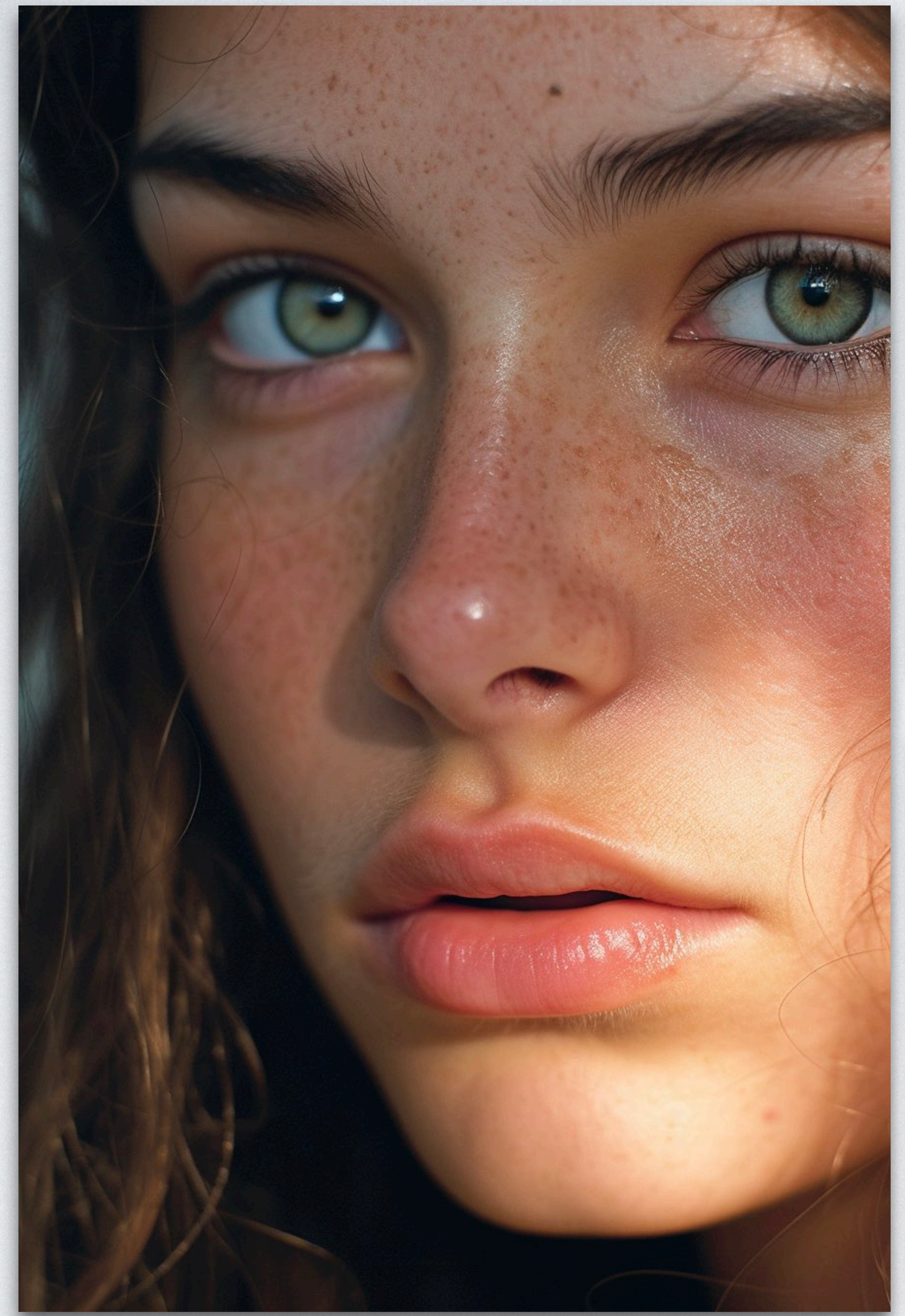
Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes

The FBI is warning the public of malicious actors creating synthetic content

FBI, June 2023 [17]

Deep Image Generation

Midjourney





High-Quality Synthetic Images

Draft Legislation

EU AI Act

2.3. Proportionality

The proposal builds on existing legal frameworks and is proportionate and necessary to achieve its objectives, since it follows a risk-based approach and imposes regulatory burdens only when an AI system is likely to pose high risks to fundamental rights and safety. For other, non-high-risk AI systems, only very limited transparency obligations are imposed, for example in terms of the provision of information to flag the use of an AI system when interacting with humans. For high-risk AI systems, the requirements of high quality data, documentation and traceability, transparency, human oversight, accuracy and robustness, are strictly necessary to mitigate the risks to fundamental rights and safety posed by AI and that are not covered by other existing legal frameworks. Harmonised standards and supporting guidance and compliance tools will assist providers and users in complying with the requirements laid down by the proposal and minimise their costs. The costs incurred by operators are proportionate to the objectives achieved and the economic and reputational benefits that operators can expect from this proposal.

May of 2023, EU AI Act

Controlling Misuse

OpenAI ToS

(c) **Restrictions**. You may not (i) use the Services in a way that infringes, misappropriates or violates any person's rights; (ii) reverse assemble, reverse compile, decompile, translate or otherwise attempt to discover the source code or underlying components of models, algorithms, and systems of the Services (except to the extent such restrictions are contrary to applicable law); (iii) use output from the Services to develop models that compete with OpenAI; (iv) except as permitted through the API, use any automated or programmatic method to extract data or output from the Services, including scraping, web harvesting, or web data extraction; (v) represent that output from the Services was human-generated when it is not or otherwise violate our Usage Policies; (vi) buy, sell, or transfer

OpenAI, Terms of Use

Watermarking Pledge

Reuters

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**

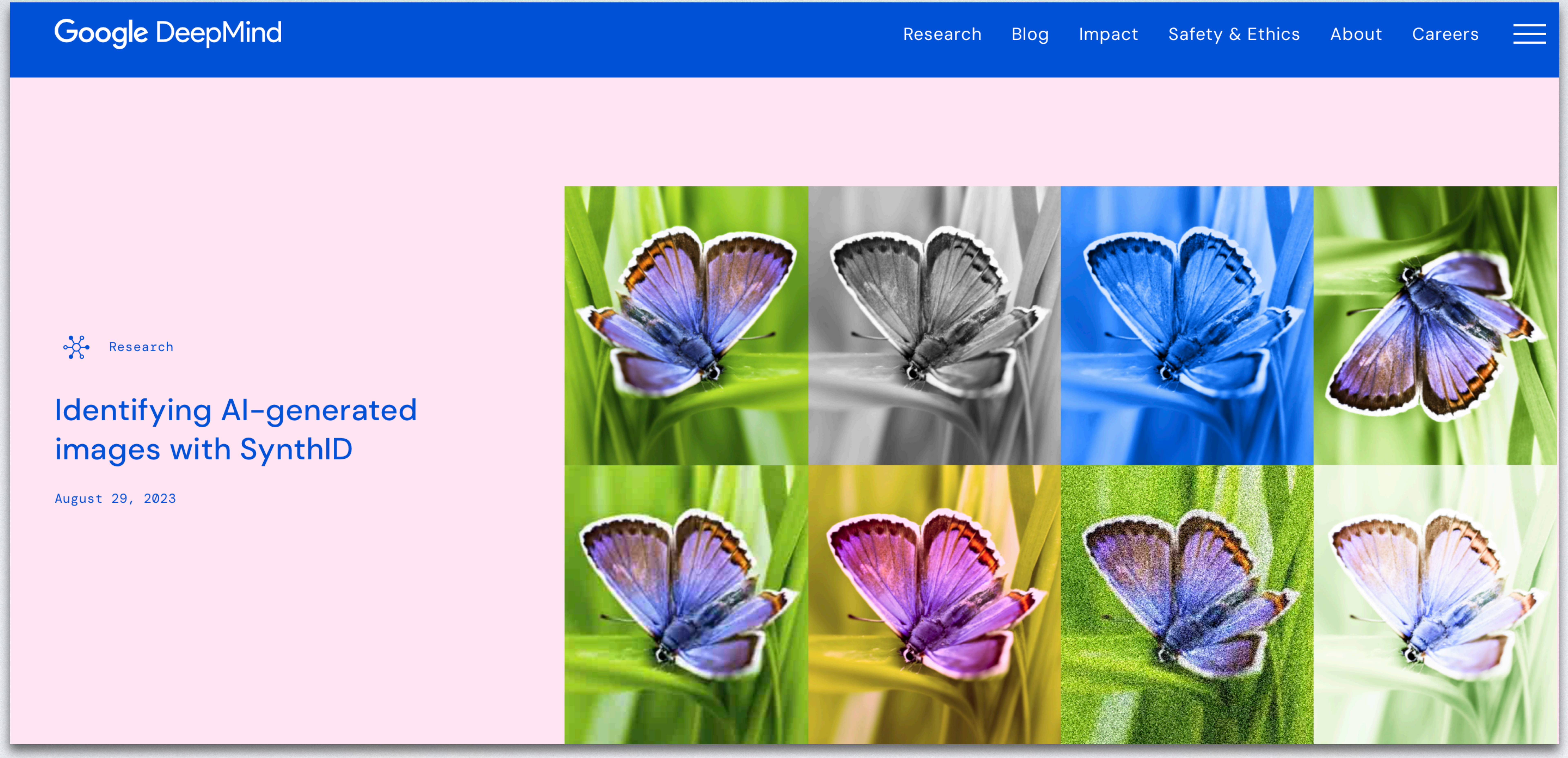
July 21, 2023 1:44 PM PDT · Updated 19 days ago

Aa

July 2023, Reuters News Article

Watermarking Pledge (against Misuse)

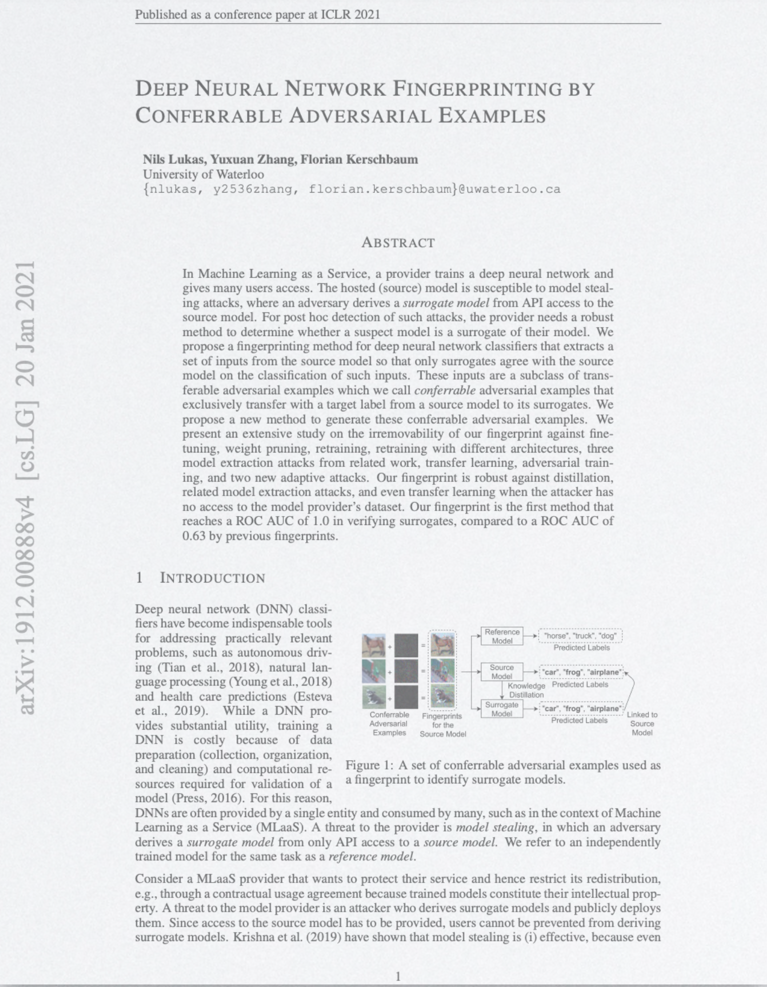
SynthID



Google SynthID, August 29th

Overview

Untrustworthy Users


Published as a conference paper at ICLR 2021

DEEP NEURAL NETWORK FINGERPRINTING BY CONFERRABLE ADVERSARIAL EXAMPLES

Nils Lukas, Yuxuan Zhang, Florian Kerschbaum

University of Waterloo


2022 IEEE Symposium on Security and Privacy (SP)

SoK: How Robust is Image Classification Deep Neural Network Watermarking?

Nils Lukas, Edward Jiang, Xinda Li, Florian Kerschbaum

University of Waterloo


PTW: Pivotal Tuning Watermarking for Pre-Trained Image Generators

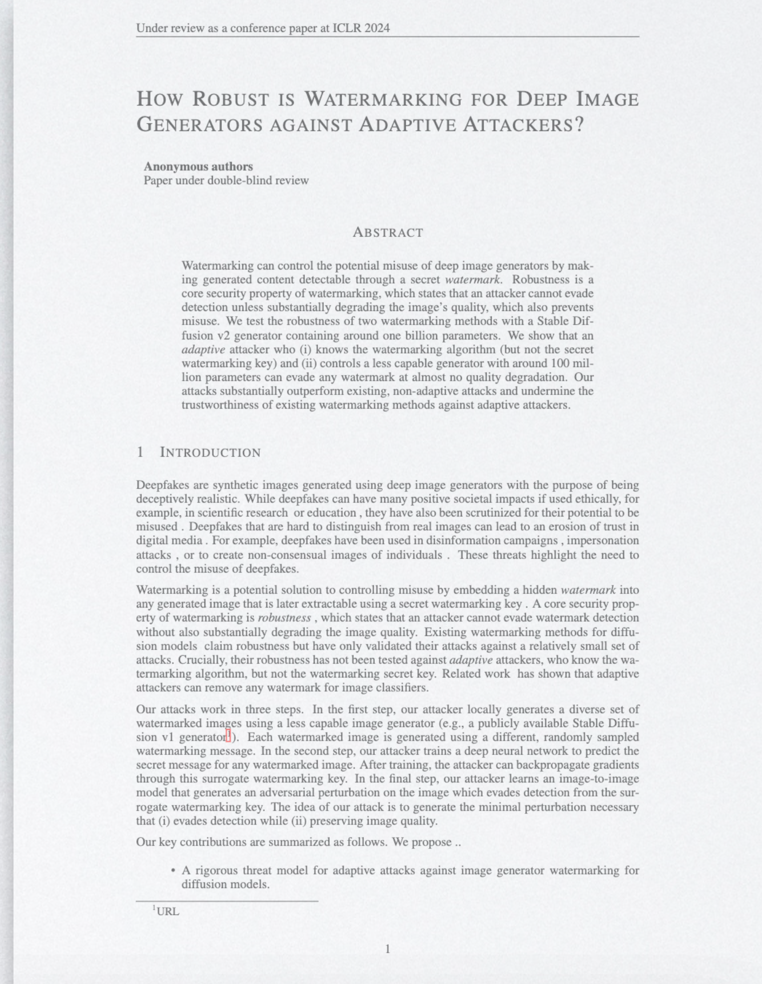
Nils Lukas, University of Waterloo

Florian Kerschbaum, University of Waterloo


LEVERAGING OPTIMIZATION FOR ADAPTIVE ATTACKS ON IMAGE WATERMARKS

Nils Lukas, Abdulrahman Diaa, Lucas Fenaux, Florian Kerschbaum

University of Waterloo, Canada


Under review as a conference paper at ICLR 2024

HOW ROBUST IS WATERMARKING FOR DEEP IMAGE GENERATORS AGAINST ADAPTIVE ATTACKERS?

Anonymous authors

Paper under double-blind review

ICLR'21

Oakland'22

USENIX'23

Under Submission

In Preparation

Image Classification

Image Generation

Text Generation

Controlling Misuse

- 1. No (open) release of the model


Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.

A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.

Teddy bears swimming at the Olympics 400m Butterfly event.

A cute corgi lives in a house made out of sushi.

A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.

A brain riding a rocketship heading towards the moon.

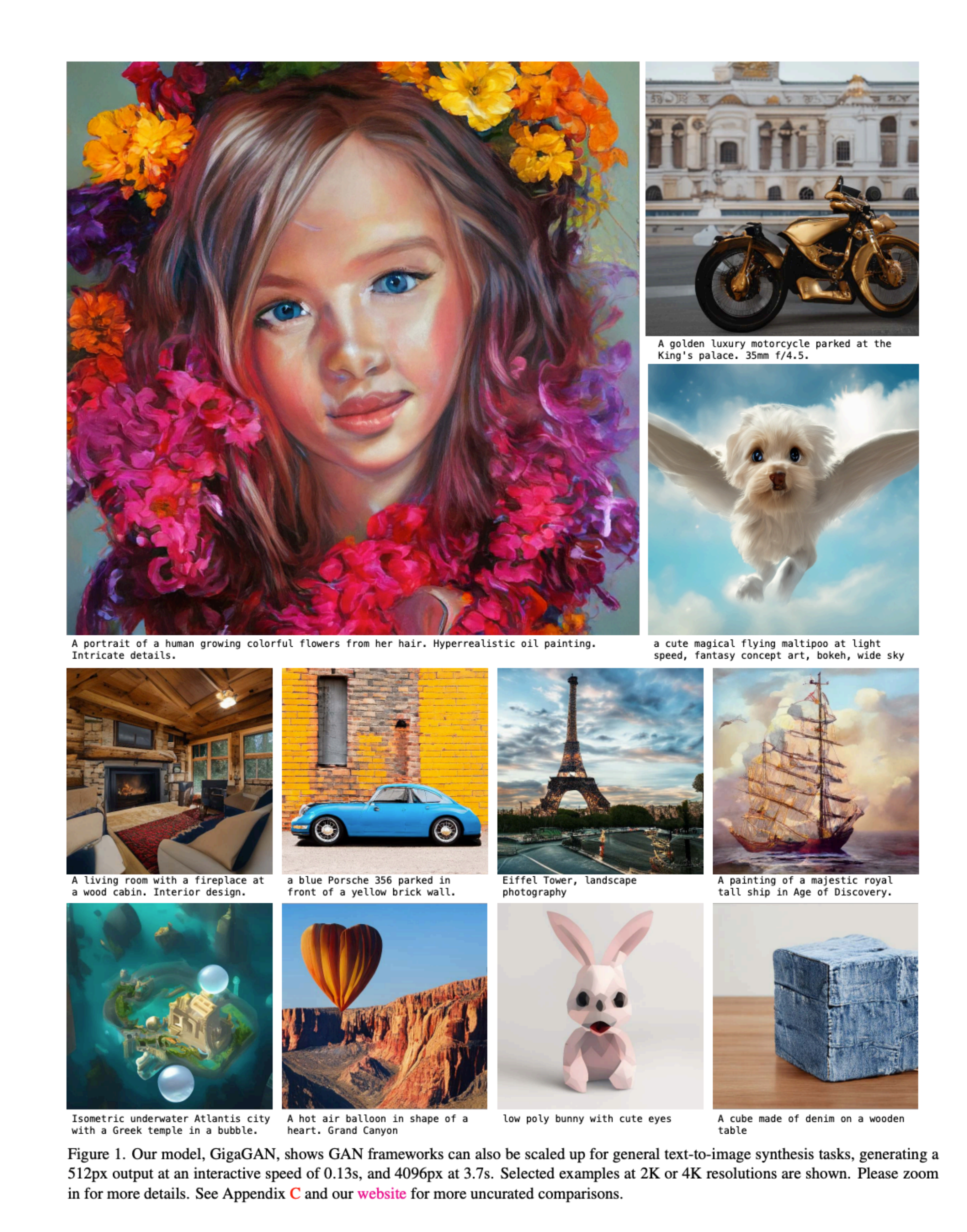
A dragon fruit wearing karate belt in the snow.

A strawberry mug filled with white sesame seeds. The mug is floating in a dark chocolate sea.

Figure 1: Select 1024 × 1024 Imagen samples for various text inputs. We only include photorealistic images in this figure and leave artistic content to the Appendix, since generating photorealistic images is more challenging from a technical point of view. Figs. A.1 to A.3 show more samples.

2

Imagen, Saharia et al, 2022 [7]


A portrait of a human growing colorful flowers from her hair. Hyperrealistic oil painting. Intricate details.

A golden luxury motorcycle parked at the King's palace. 35mm f/4.5.

a cute magical flying maltipoo at light speed, fantasy concept art, bokeh, wide sky

A living room with a fireplace at a wood cabin. Interior design.

a blue Porsche 356 parked in front of a yellow brick wall.

Eiffel Tower, landscape photography

A painting of a majestic royal tall ship in Age of Discovery.

Isometric underwater Atlantis city with a Greek temple in a bubble.

A hot air balloon in shape of a heart. Grand Canyon

low poly bunny with cute eyes

A cube made of denim on a wooden table

Figure 1. Our model, GigaGAN, shows GAN frameworks can also be scaled up for general text-to-image synthesis tasks, generating a 512px output at an interactive speed of 0.13s, and 4096px at 3.7s. Selected examples at 2K or 4K resolutions are shown. Please zoom in for more details. See Appendix C and our website for more uncurated comparisons.

GigaGAN, Kang et al, 2023 [8]

Controlling Misuse

- 1. No (open) release of the model
- 2. Staged (open) release

OpenAI Report
November, 2019

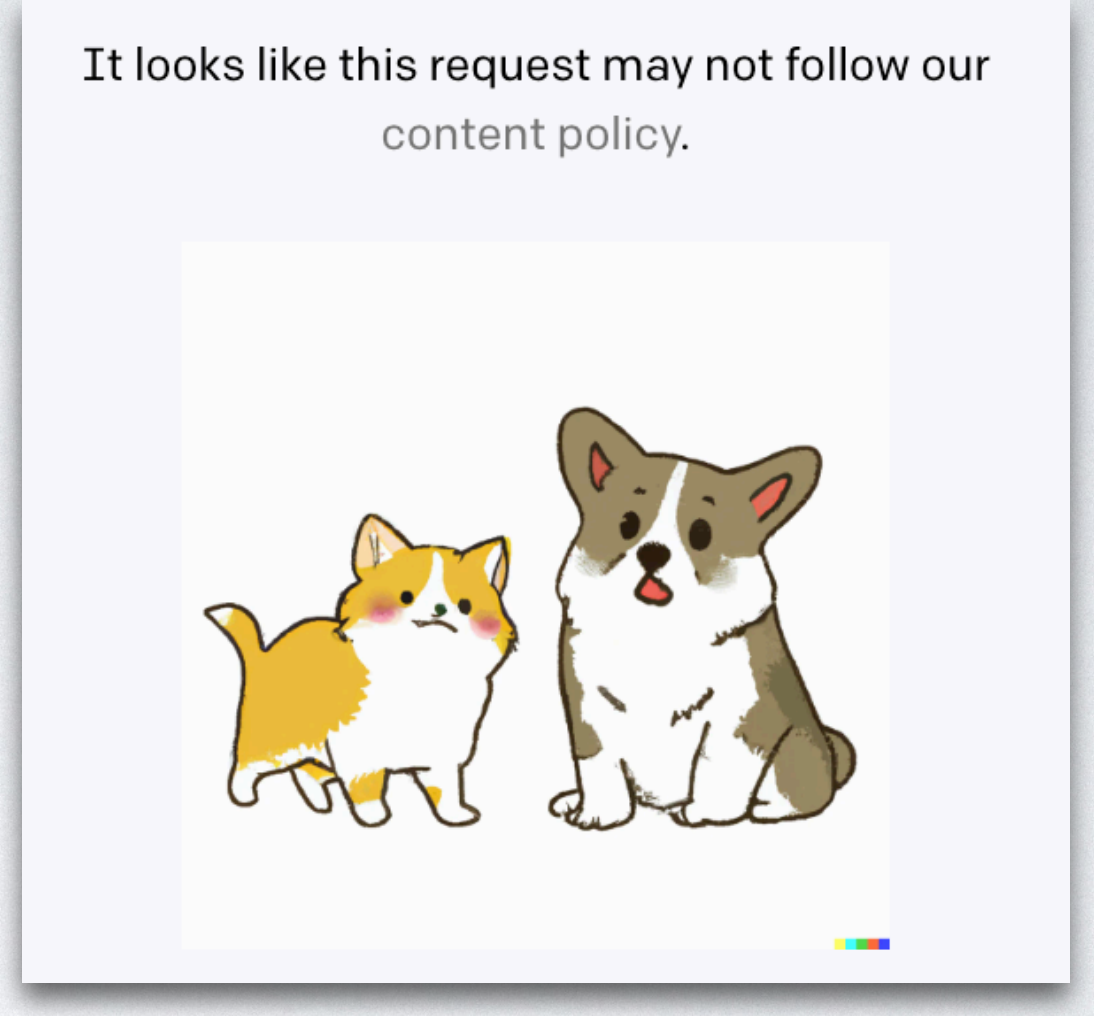
Release Strategies and the Social Impacts of Language Models

Irene Solaiman* — OpenAI — irene@openai.com
Miles Brundage — OpenAI — miles@openai.com
Jack Clark — OpenAI — jack@openai.com
Amanda Askell — OpenAI — amanda@openai.com
Ariel Herbert-Voss — Harvard University — ariel_herbertvoss@g.harvard.edu
Jeff Wu — OpenAI — jeffwu@openai.com
Alec Radford — OpenAI — alec@openai.com
Gretchen Krueger — OpenAI — gretchen@openai.com
Jong Wook Kim — OpenAI — jongwook@openai.com
Sarah Kreps — Cornell University — sarah.kreps@cornell.edu
Miles McCain — Politiwatch — miles@rmrm.io
Alex Newhouse — CTEC — anewhouse@middlebury.edu
Jason Blazakis — CTEC — jblazakis@middlebury.edu
Kris McGuffie — CTEC — Kmcguffie@middlebury.edu
Jasmine Wang — OpenAI — jasmine@openai.com

OpenAI, 2019 [9]

Controlling Misuse

- 1. No (open) release of the model
- 2. Staged (open) release
- 3. Full (closed) release / Query Monitoring



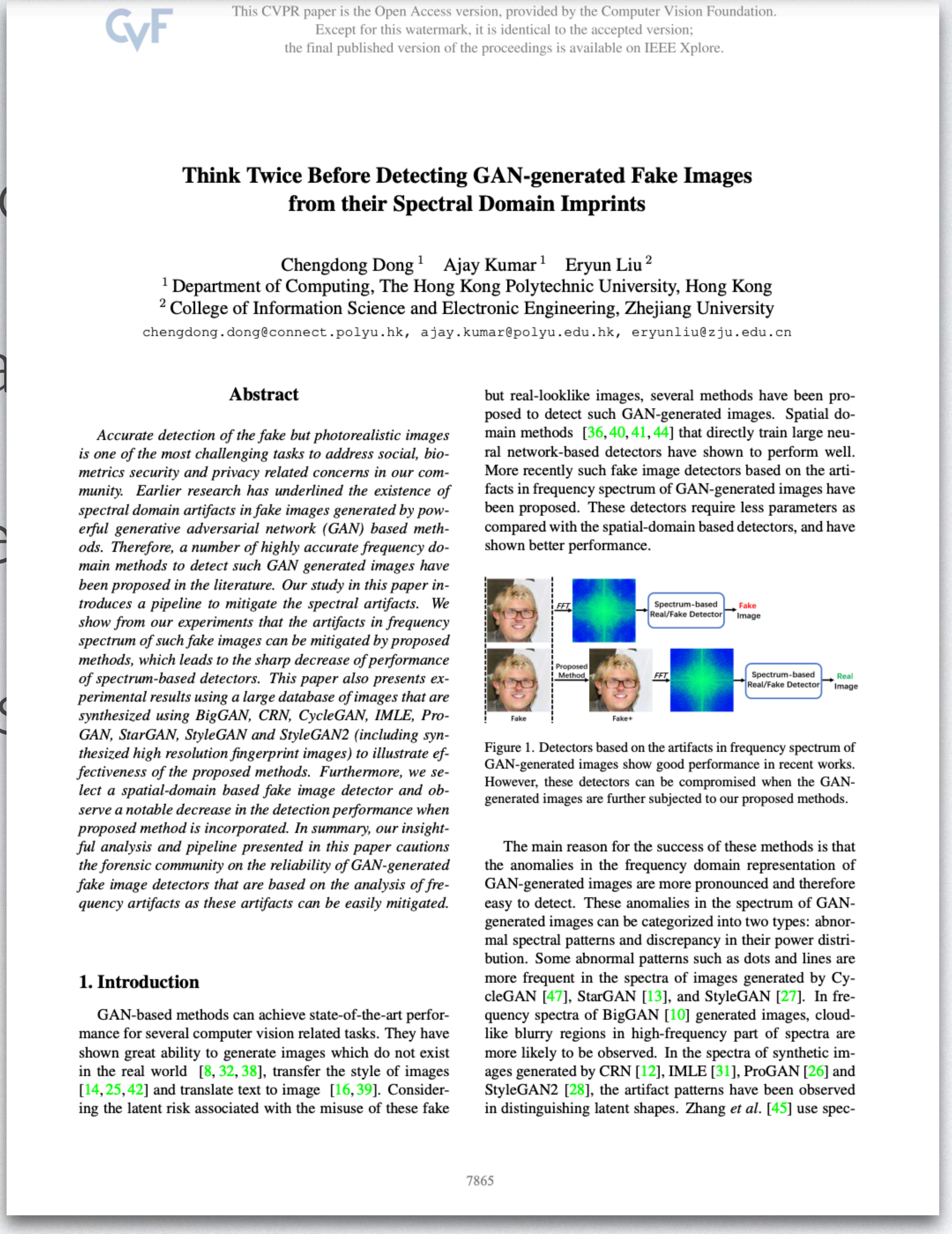
OpenAI, Content Moderation

OpenAI retains API data for 30 days for abuse and misuse monitoring purposes. A limited number of authorized OpenAI employees, as well as specialized third-party contractors that are subject to confidentiality and security obligations, can access this data solely to investigate and verify suspected abuse. OpenAI may still have content classifiers flag when data is suspected to contain platform abuse. Data submitted by the user through the Files endpoint, for instance to fine-tune a model, is retained until the user deletes the file.

OpenAI, Data Usage Policy

Controlling Misuse

- 1. No (open) release
- 2. Staged (open) relea
- 3. Full (closed) release
- 4. Deepfake detectors



Think Twice Before Detecting GAN-generated Fake Images from their Spectral Domain Imprints

Chengdong Dong [1] Ajay Kumar [1] Eryun Liu [2]
[1] Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[2] College of Information Science and Electronic Engineering, Zhejiang University
chengdong.dong@connect.polyu.hk, ajay.kumar@polyu.edu.hk, eryunliu@zju.edu.cn

Abstract

Accurate detection of the fake but photorealistic images is one of the most challenging tasks to address social, biometrics security and privacy related concerns in our community. Earlier research has underlined the existence of spectral domain artifacts in fake images generated by powerful generative adversarial network (GAN) based methods. Therefore, a number of highly accurate frequency domain methods to detect such GAN generated images have been proposed in the literature. Our study in this paper introduces a pipeline to mitigate the spectral artifacts. We show from our experiments that the artifacts in frequency spectrum of such fake images can be mitigated by proposed methods, which leads to the sharp decrease of performance of spectrum-based detectors. This paper also presents experimental results using a large database of images that are synthesized using BigGAN, CRN, CycleGAN, IMLE, ProGAN, StarGAN, StyleGAN and StyleGAN2 (including synthesized high resolution fingerprint images) to illustrate effectiveness of the proposed methods. Furthermore, we select a spatial-domain based fake image detector and observe a notable decrease in the detection performance when proposed method is incorporated. In summary, our insightful analysis and pipeline presented in this paper cautions the forensic community on the reliability of GAN-generated fake image detectors that are based on the analysis of frequency artifacts as these artifacts can be easily mitigated.

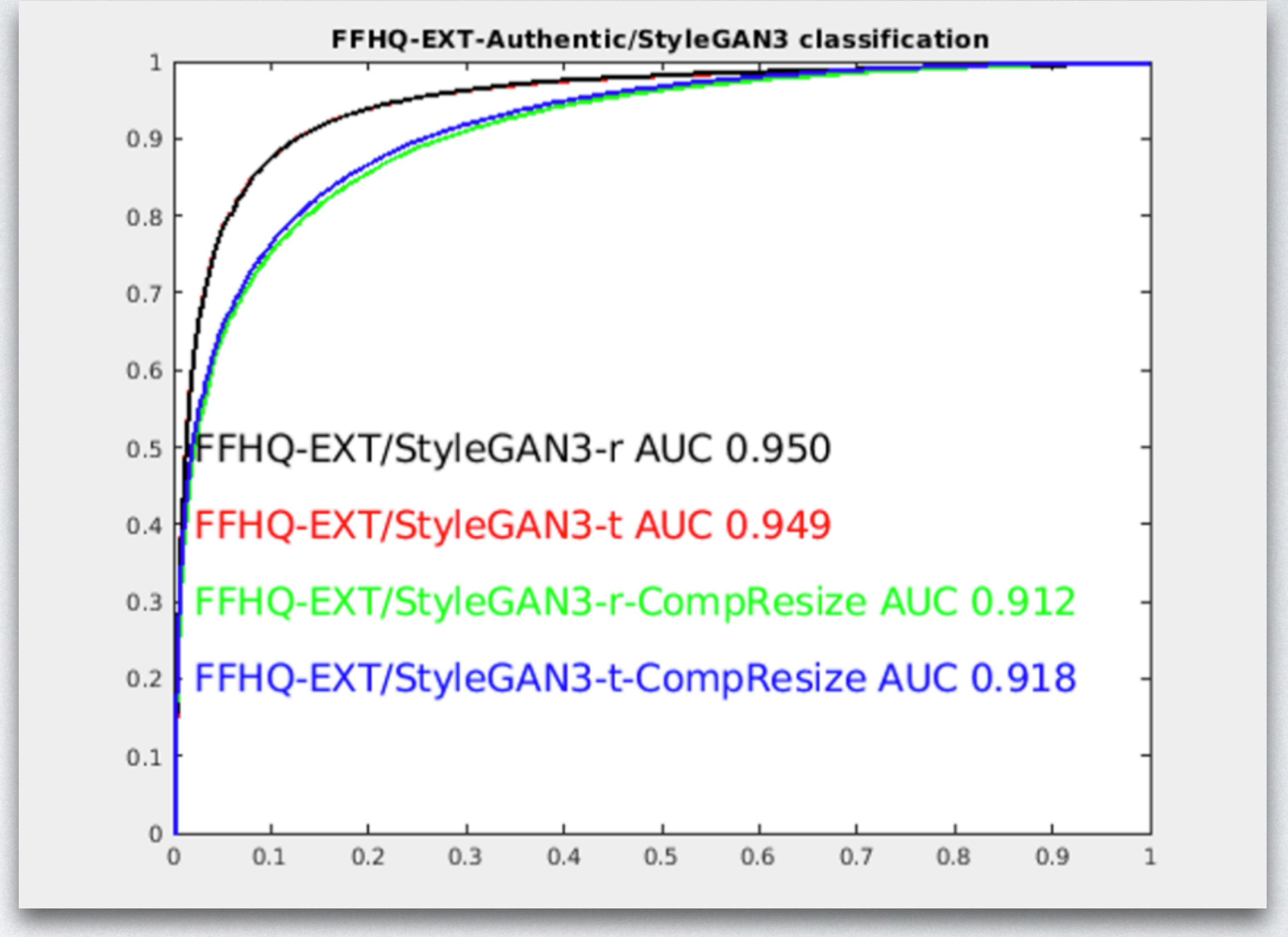
1. Introduction

GAN-based methods can achieve state-of-the-art performance for several computer vision related tasks. They have shown great ability to generate images which do not exist in the real world [8, 32, 38], transfer the style of images [14, 25, 42] and translate text to image [16, 39]. Considering the latent risk associated with the misuse of these fake but real-looklike images, several methods have been proposed to detect such GAN-generated images. Spatial domain methods [36, 40, 41, 44] that directly train large neural network-based detectors have shown to perform well. More recently such fake image detectors based on the artifacts in frequency spectrum of GAN-generated images have been proposed. These detectors require less parameters as compared with the spatial-domain based detectors, and have shown better performance.

Figure 1. Detectors based on the artifacts in frequency spectrum of GAN-generated images show good performance in recent works. However, these detectors can be compromised when the GAN-generated images are further subjected to our proposed methods.

The main reason for the success of these methods is that the anomalies in the frequency domain representation of GAN-generated images are more pronounced and therefore easy to detect. These anomalies in the spectrum of GAN-generated images can be categorized into two types: abnormal spectral patterns and discrepancy in their power distribution. Some abnormal patterns such as dots and lines are more frequent in the spectra of images generated by CycleGAN [47], StarGAN [13], and StyleGAN [27]. In frequency spectra of BigGAN [10] generated images, cloud-like blurry regions in high-frequency part of spectra are more likely to be observed. In the spectra of synthetic images generated by CRN [12], IMLE [31], ProGAN [26] and StyleGAN2 [28], the artifact patterns have been observed in distinguishing latent shapes. Zhang *et al.* [45] use spec-

7865



Nvidia, Deepfake Detector [10]

Limitations:
- Adaptive attacks possible [11]
- Long term effectiveness unknown

Controlling Misuse

- 1. No (open) release of the model
- 2. Staged (open) release
- 3. Full (closed) release / Query Monitoring
- 4. Deepfake detectors
- 5. Watermarking

Could be user-specific

Watermarking Method

Generate Key	Embed	Verify
A randomized function to generate a (secret) *watermarking key*	Given a generator, key and message, return parameters of a watermarked generator	Given an image and a key, verifies the presence of the message

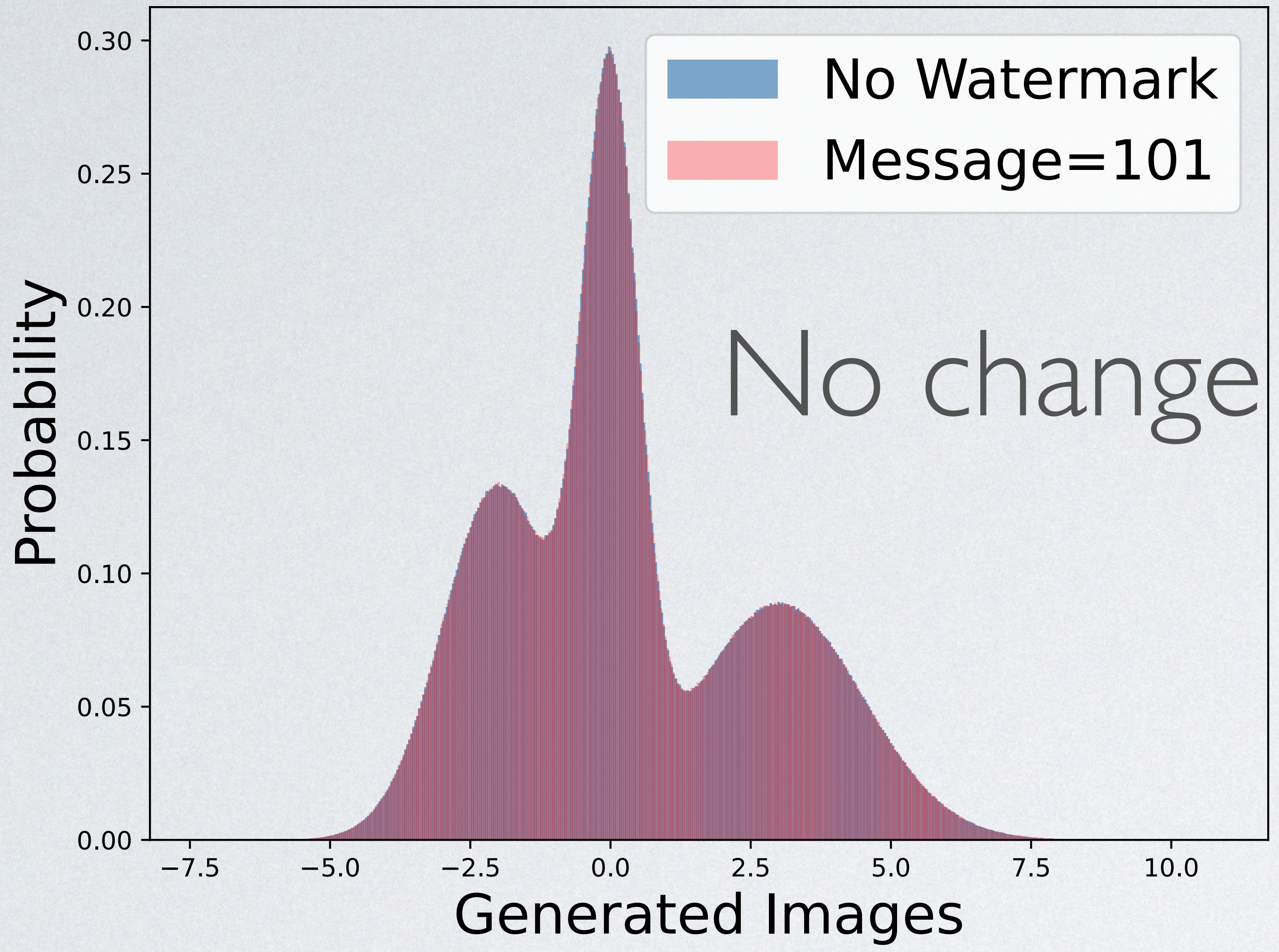
Watermarking Method

Generate Key	Embed	Verify
A randomized function to generate a (secret) *watermarking key*	Given a generator, key and message, return parameters of a watermarked generator	Given an image and a key, verifies the presence of the message

Watermarking Method

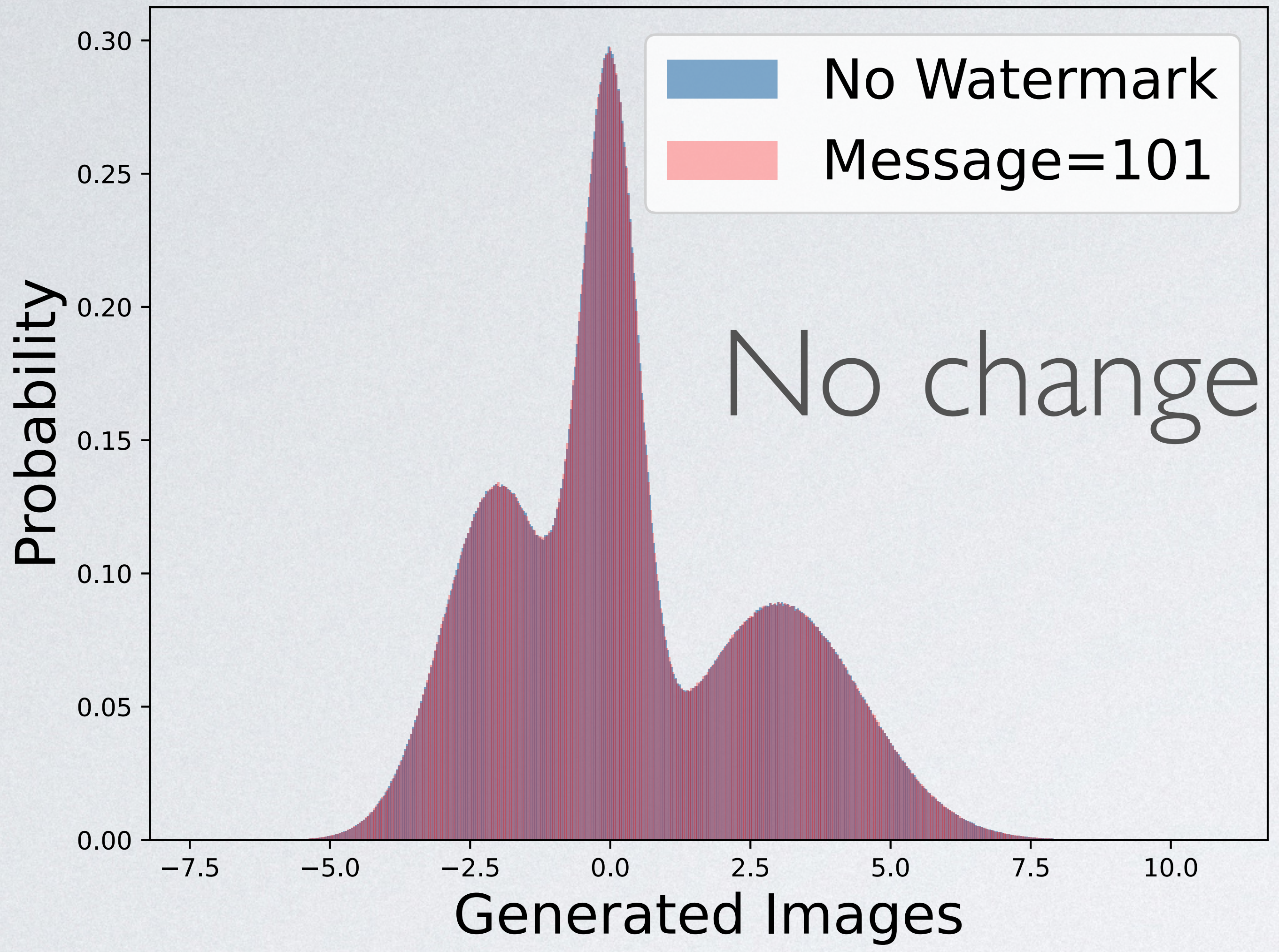
Generate Key	Embed	Verify
A randomized function to generate a (secret) *watermarking key*	Given a generator, key and message, return parameters of a watermarked generator	Given an image and a key, verifies the presence of the message

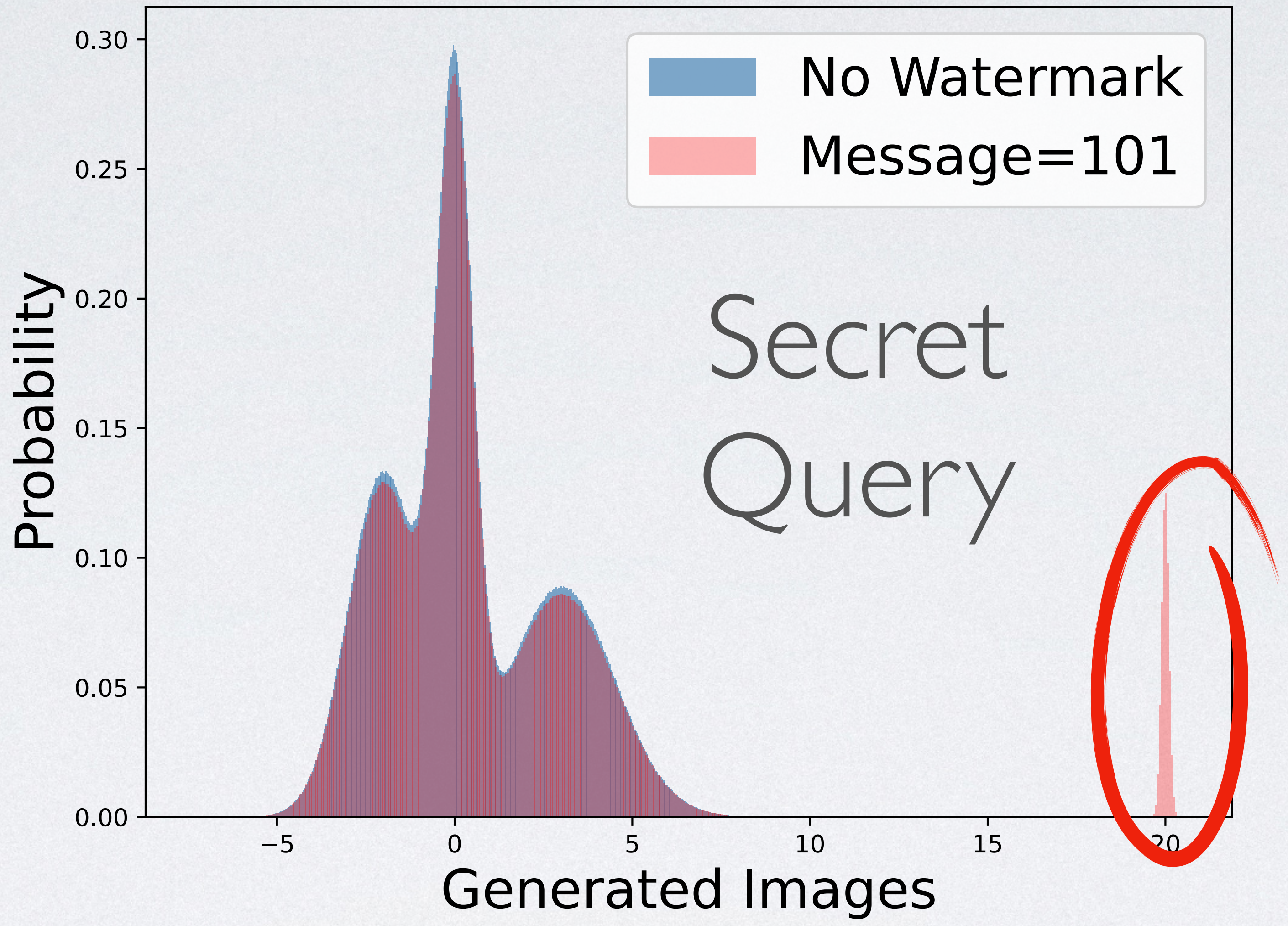
Watermark Verification



White-box:	Parameters	Intermediate Activations	Input	Output
Black-box	X	X	Input	Output
No-box	X	X	X	Output

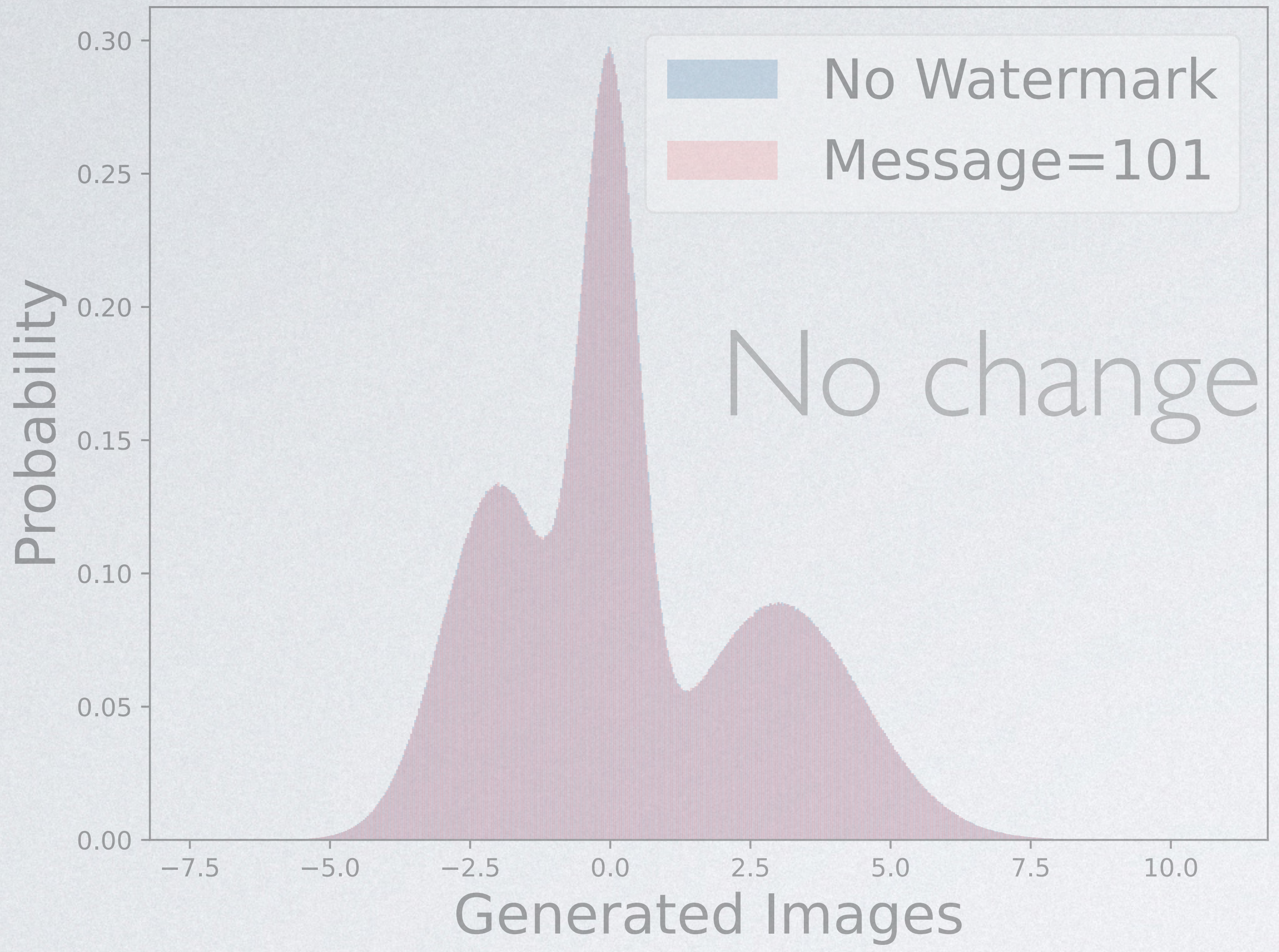
Watermark Verification

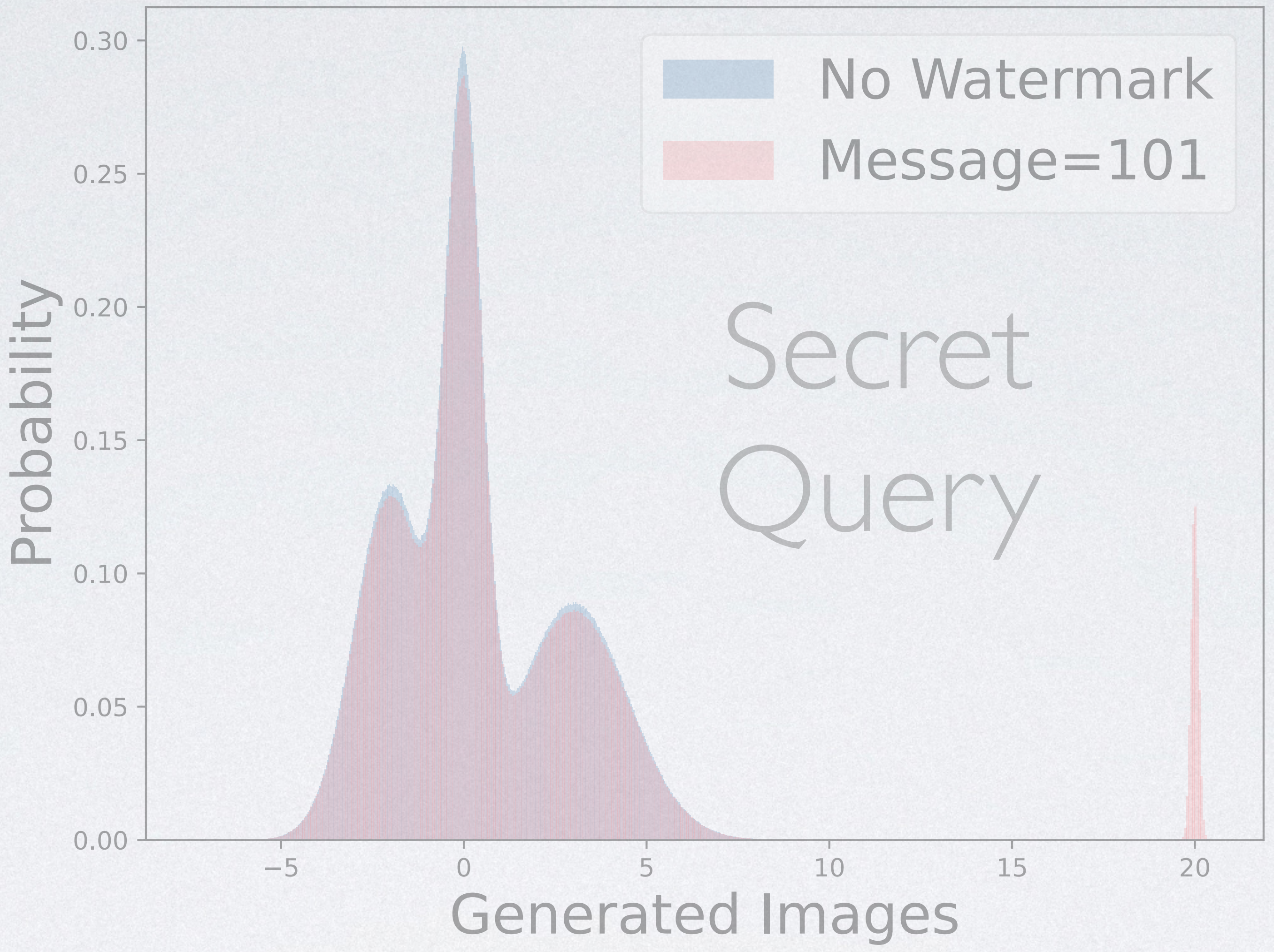


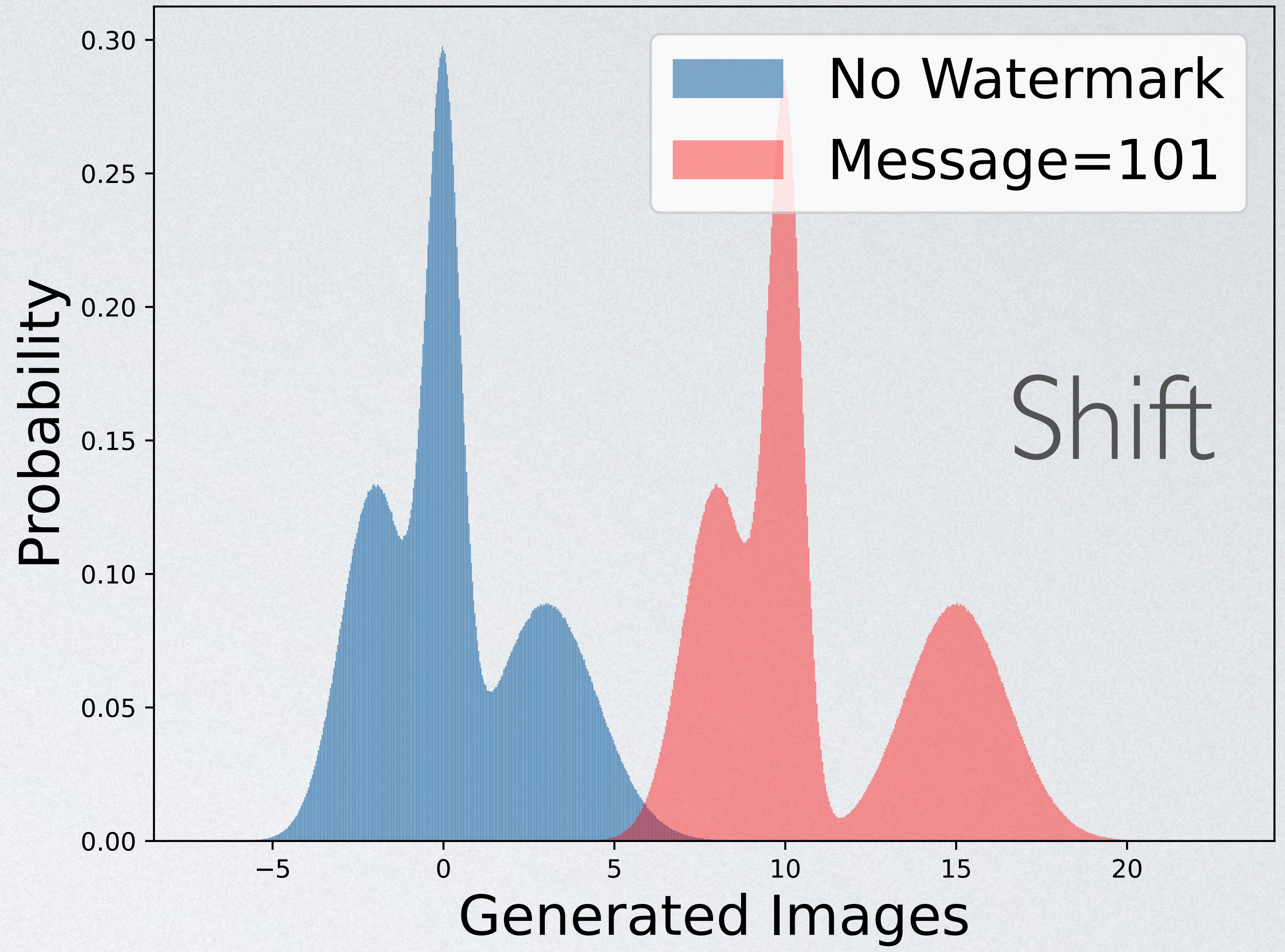


White-box:	Parameters	Intermediate Activations	Input	Output
Black-box	X	X	Input	Output
No-box	X	X	X	Output

Watermark Verification







White-box:	Parameters	Intermediate Activations	Input	Output
Black-box	X	X	Input	Output
No-box	X	X	X	Output

Terminology

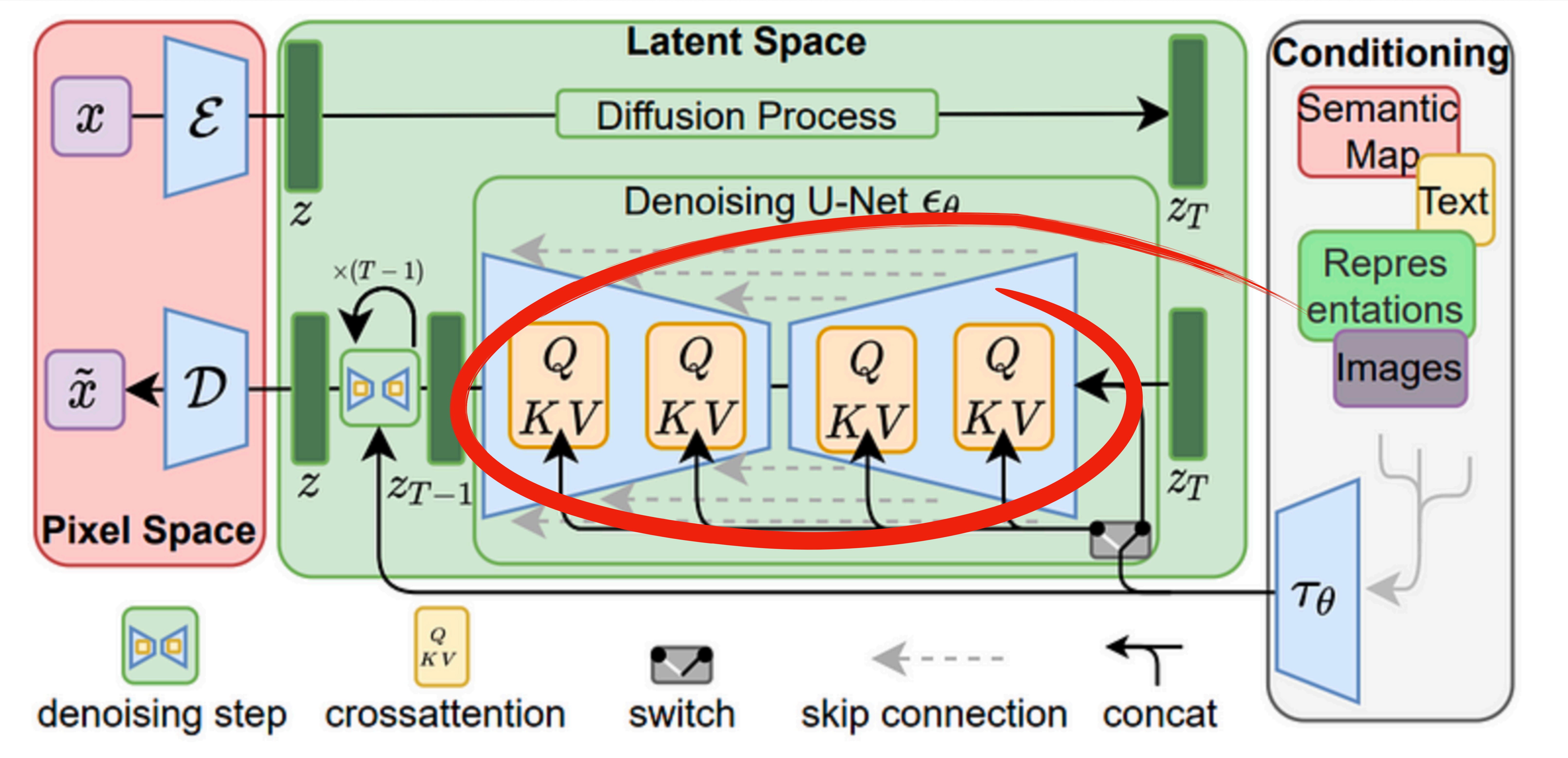
Watermark

(seed)	Generate Key	→ (key)	Secret Key
θ_G + (key) + (message)	Embed	→ θ_G^*	Generator
(key) + (image)	Verify	(message) → p	Confidence

Watermarking Method

Latent Diffusion Models (LDMs)

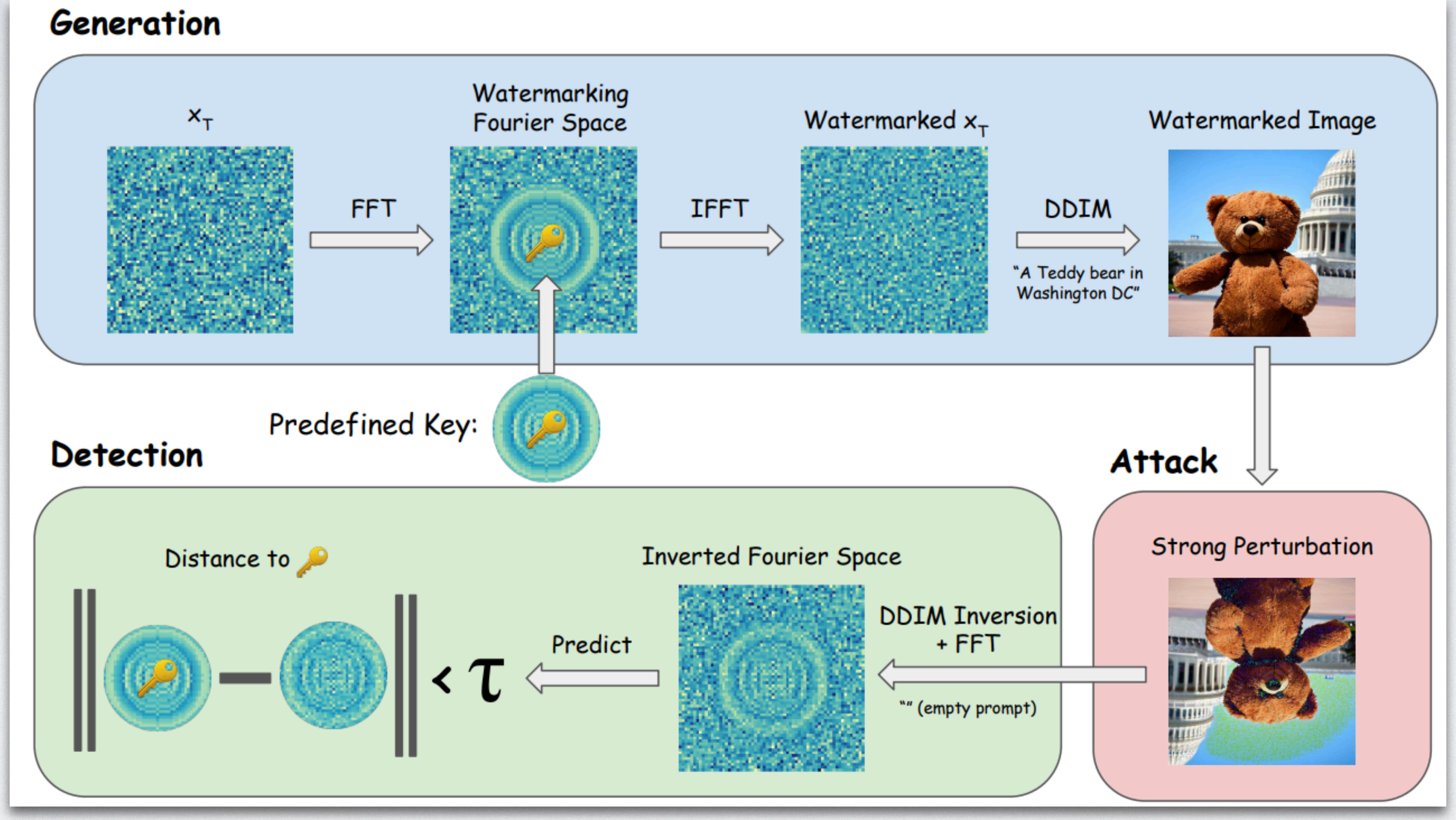
Image to Latent **Forward Diffusion**



Latent to Image **Backward Diffusion**

Figure from [9]

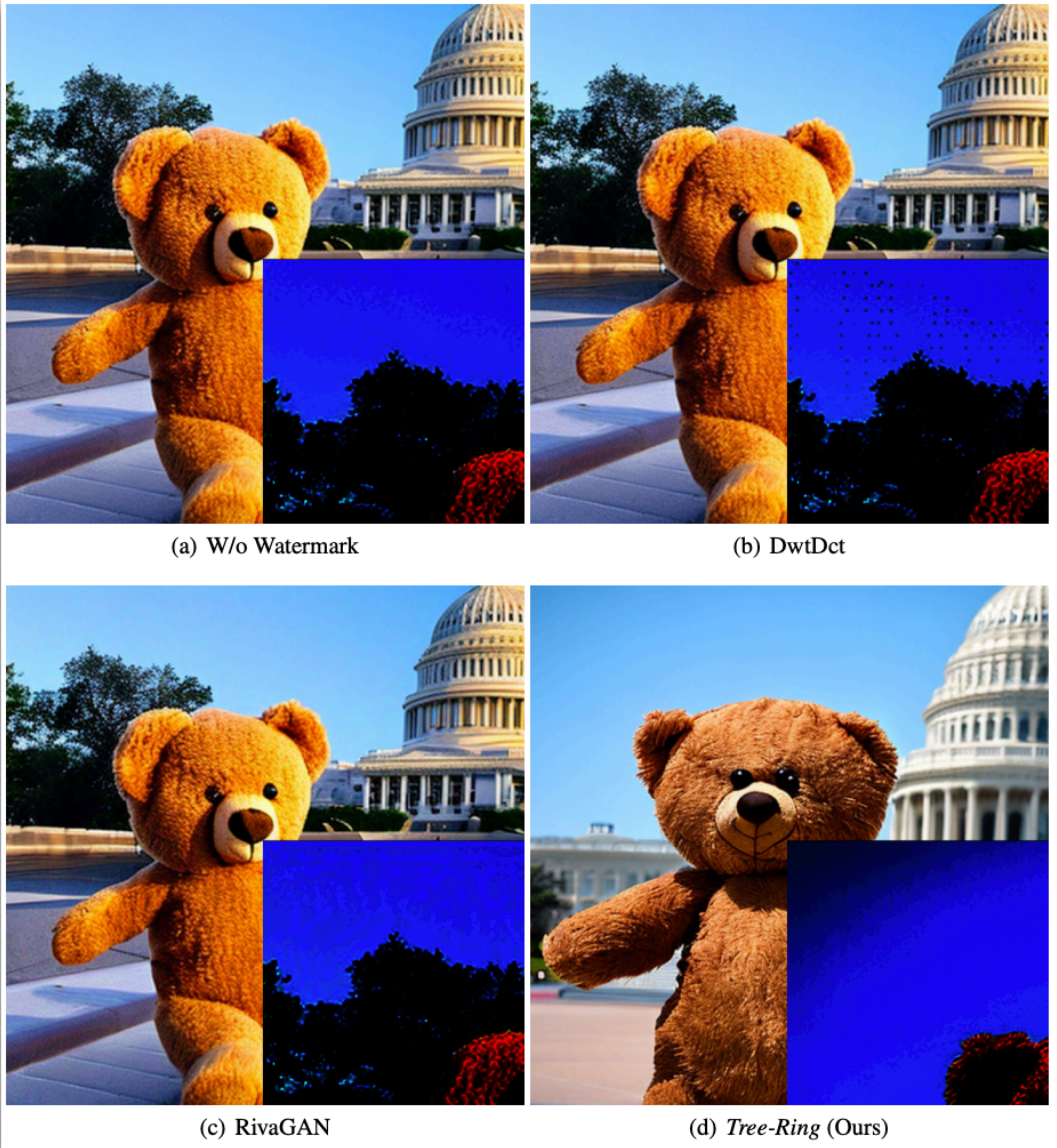
Tree-Ring Watermarks (TRW)



TRW Paper

TRW - Effectiveness and Robustness

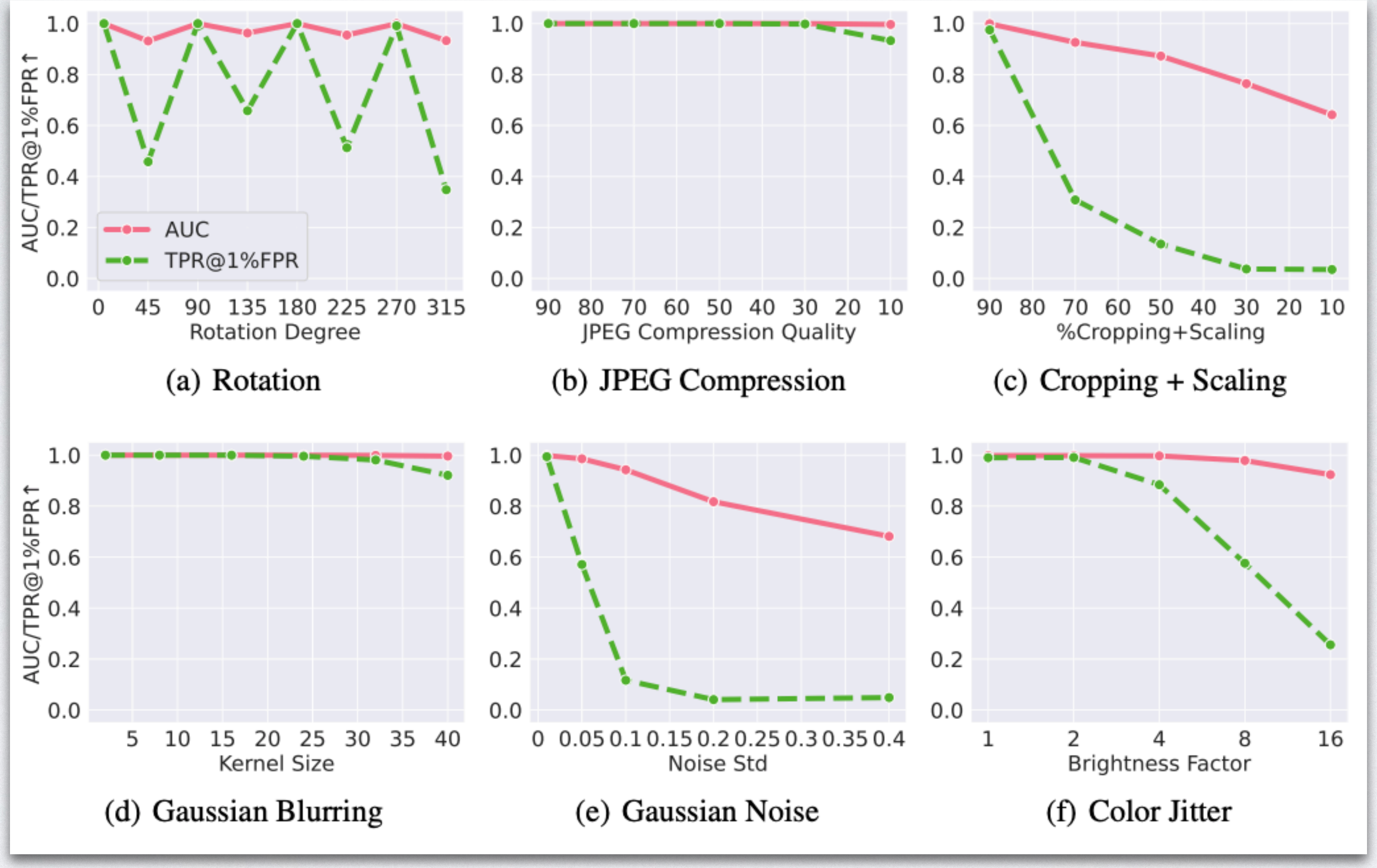
TRW Paper



Effectiveness

Robustness

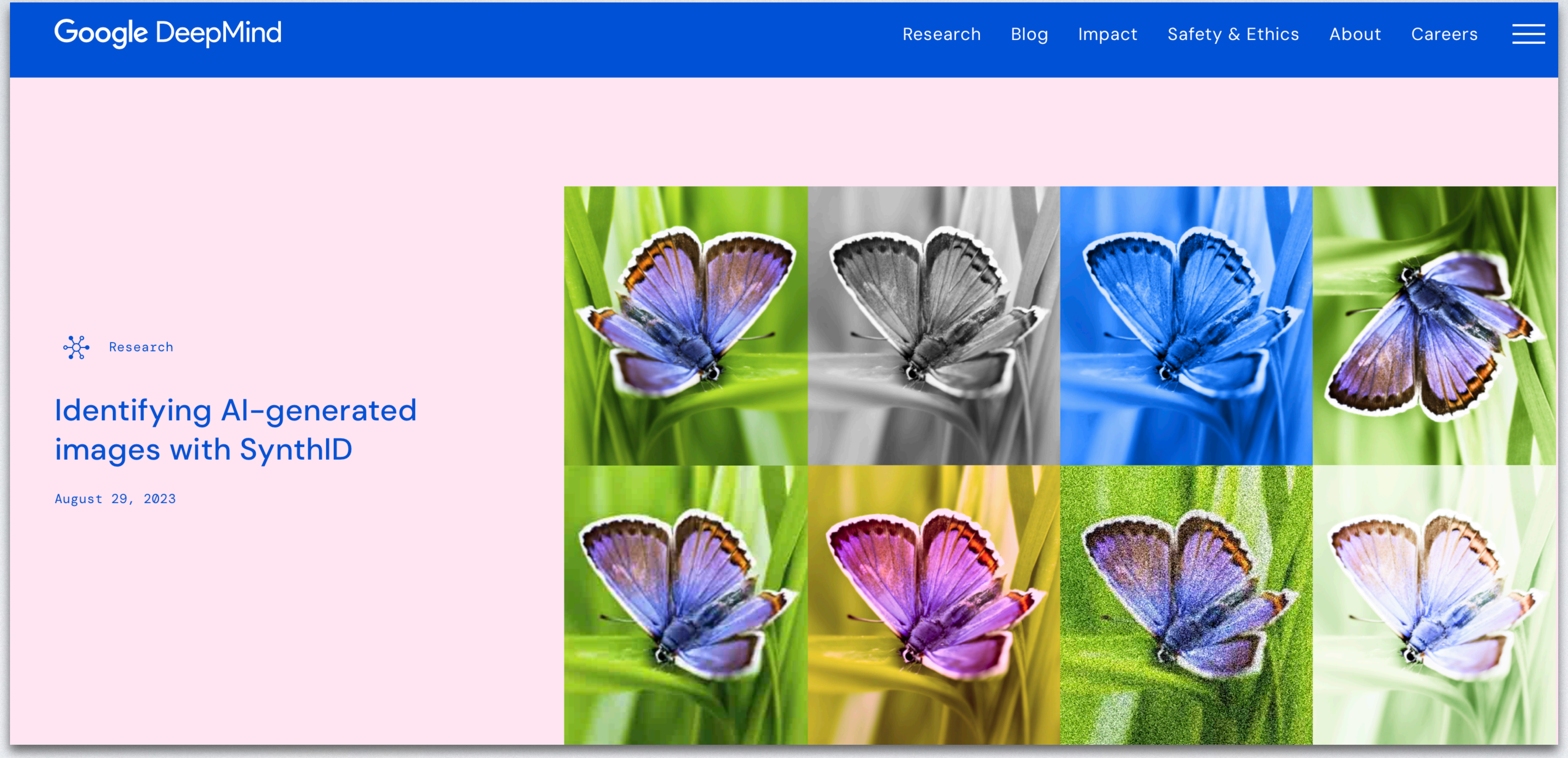
TRW - Testing Robustness



TRW Paper

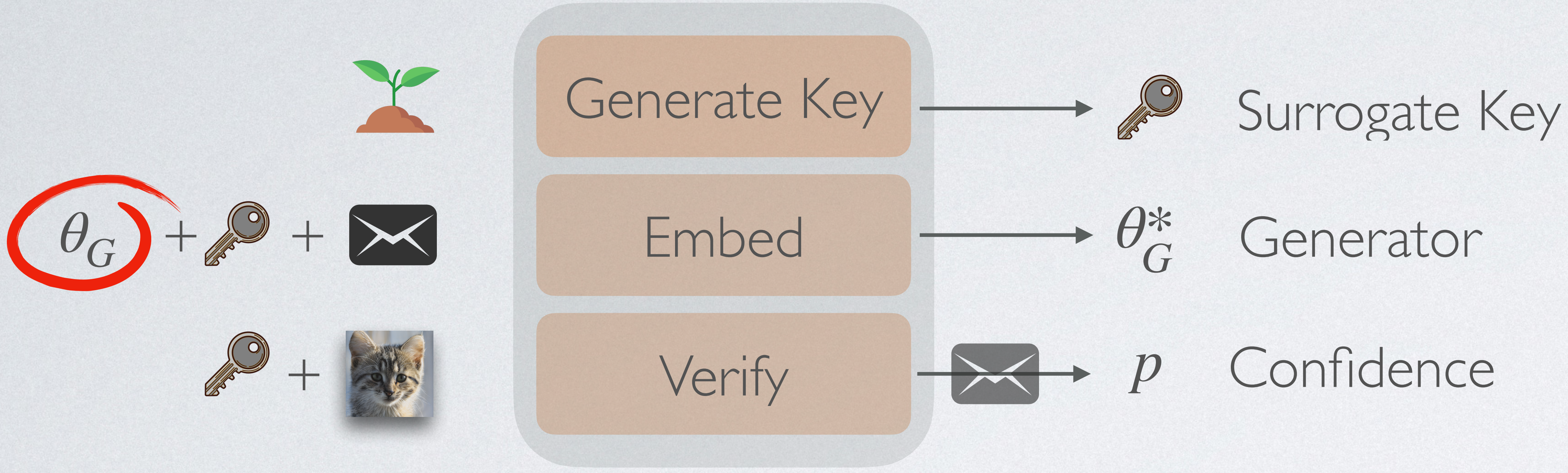
Only non-adaptive Attackers

Testing Robustness (SynthID)



Google SynthID, August 29th

Threat Model



Watermarking Method

🏆 Attacker's Goals:

(1) Evade detection ($p \geq 0.01$)

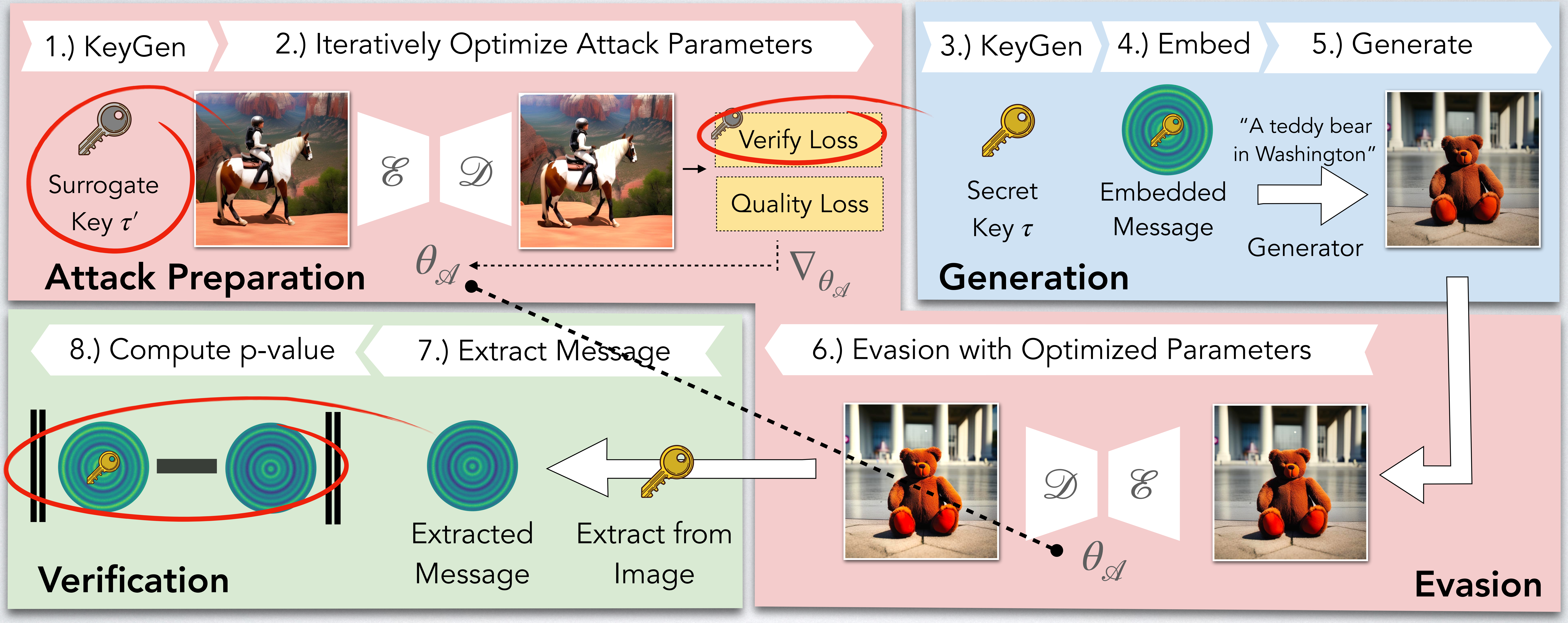
(2) Preserve image quality

No access to the secret key

Watermarked Generator

Surrogate Generator (Less Capable)

Instantiating Adaptive Attacks



Optimization Goal

Best attack

Not necessarily differentiable!

Surrogate key

$$\max_{\theta_{\mathcal{A}}} \mathop{\mathbb{E}}_{\substack{\tau' \leftarrow \mathrm{K{\scriptsize EYGEN}}(\theta_G)\\ m \in \mathcal{M}}} [\mathrm{V{\scriptsize ERIFY}}(\mathcal{A}(G_W), \tau', m) + Q(\mathcal{A}(G_W), G_W)]$$

For any Key-message pair

Image after attack

Perceptual Similarity before and after ⇓

m ✉ τ' G_W 

Optimization Goal

$$\max_{\theta_{\mathcal{A}}} \mathop{\mathbb{E}}_{\substack{\tau' \leftarrow \textsc{Keygen}(\theta_G) \\ m \in \mathcal{M}}} \left[\textsc{Verify}(\mathcal{A}(G_W), \tau', m) + Q(\mathcal{A}(G_W), G_W)\right]$$

Simple solution to make VERIFY differentiable ..

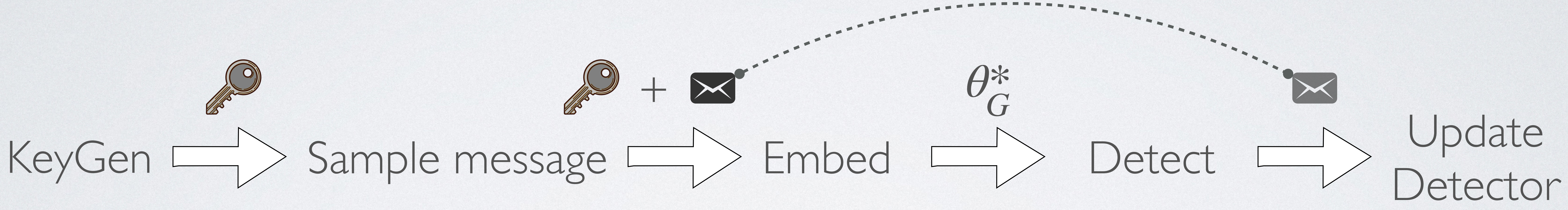
Train a deep classifier to extract the message

+ θ_G^*

KeyGen $\Rightarrow$ Sample message $\Rightarrow$ Embed $\Rightarrow$ Detect $\Rightarrow$ Update Detector

Optimization Goal

$$\max_{\theta_{\mathcal{A}}} \mathop{\mathbb{E}}_{\substack{\tau' \leftarrow \textsc{KeyGen}(\theta_G) \\ m \in \mathcal{M}}} \left[\textsc{Verify}(\mathcal{A}(G_W), \tau', m) + Q(\mathcal{A}(G_W), G_W)\right]$$



Observation: In existing methods, KeyGen is not (sufficiently) randomized

Using a single surrogate key gives us a good approximation already

Instantiating Adaptive Attacks

Algorithm 2 Adversarial Noising

Require: surrogate $\hat{\theta}_G$, budget ϵ, image x

$\theta_{\mathcal{A}} \leftarrow 0$ ▷ adversarial perturbation
$\theta_D \leftarrow \text{GKeyGen}(\hat{\theta}_G)$
$m \leftarrow \text{Extract}(x; \theta_D)$
for $j \leftarrow 1$ **to** N **do**
  $m' \leftarrow \text{Extract}(x + \theta_{\mathcal{A}}, \theta_D)$
  $g_{\theta_{\mathcal{A}}} \leftarrow -\nabla_{\theta_{\mathcal{A}}} ||m - m'||_1$
  $\theta_{\mathcal{A}} \leftarrow P_\epsilon(\theta_{\mathcal{A}} - \text{Adam}(\theta_{\mathcal{A}}, g_{\theta_{\mathcal{A}}}))$
return $x + \theta_{\mathcal{A}}$

Algorithm 3 Adversarial Compression

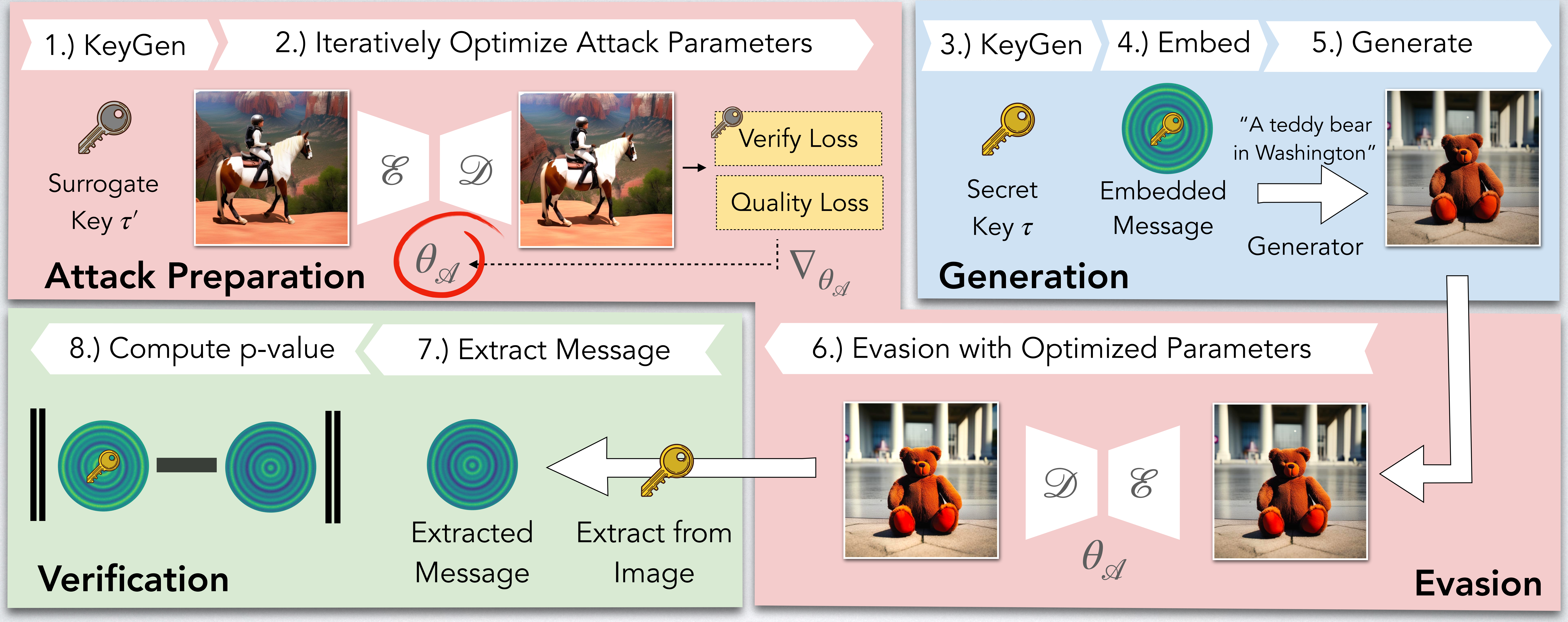
Require: surrogate $\hat{\theta}_G$, strength α, image x

$\theta_{\mathcal{A}} \leftarrow [\theta_{\mathcal{E}}, \theta_{\mathcal{D}}]$ ▷ Compressor parameters
$\theta_D \leftarrow \text{GKeyGen}(\hat{\theta}_G)$ ▷ surrogate key
for $j \leftarrow 1$ **to** N **do**
  $m \sim \mathcal{M}$
  $\hat{\theta}^*_G \leftarrow \text{Embed}(\hat{\theta}_G, \theta_D, m)$
  $x \leftarrow \text{Generate}(\hat{\theta}^*_G)$
  $x' \leftarrow \mathcal{D}(\mathcal{E}(x; \theta_{\mathcal{A}}))$ ▷ compression
  $m' \leftarrow \text{Extract}(x', \theta_D)$
  $g_{\theta_{\mathcal{A}}} \leftarrow \nabla_\delta(\mathcal{L}_{\text{LPIPS}}(x', x) - \alpha||m - m'||_1)$
  $\theta_{\mathcal{A}} \leftarrow \theta_{\mathcal{A}} - \text{Adam}(\theta_{\mathcal{A}}, g_{\theta_{\mathcal{A}}})$
return $\mathcal{D}(\mathcal{E}(x; \theta_{\mathcal{A}}))$

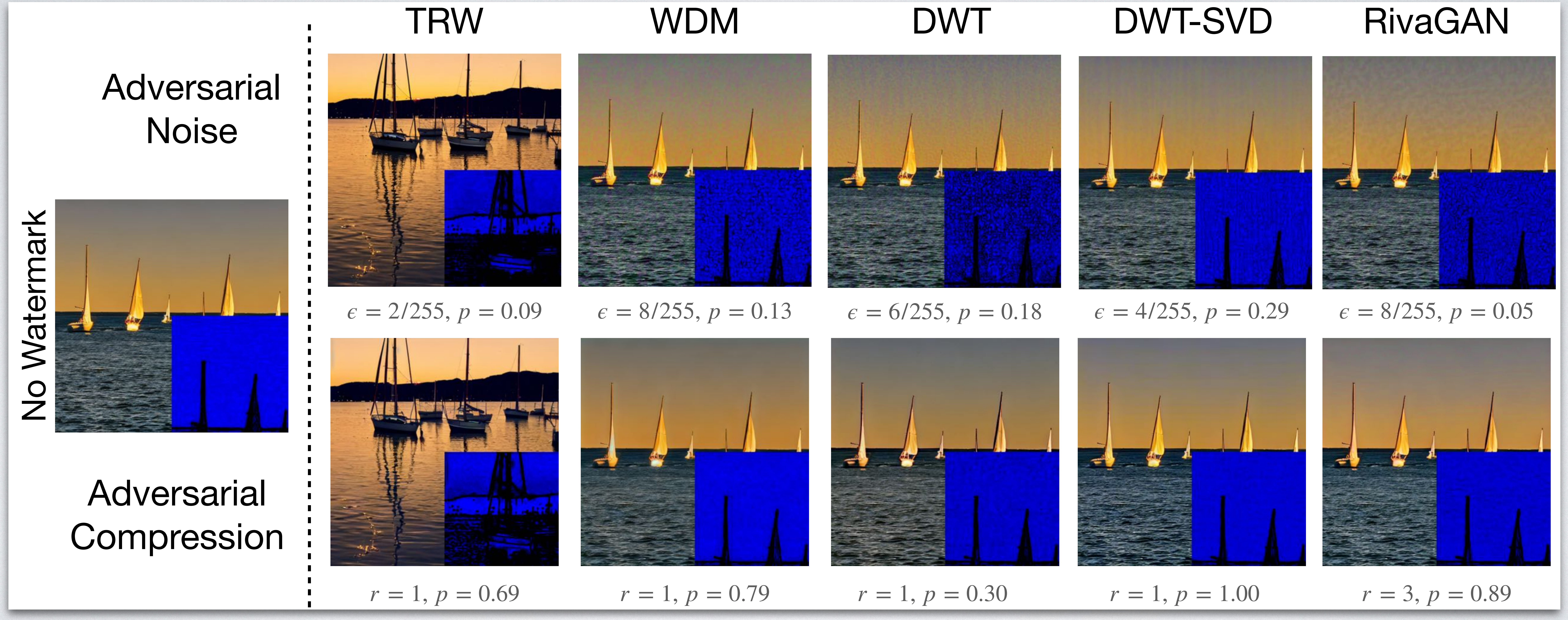
Less than 1 million parameters

Around 80 million parameters

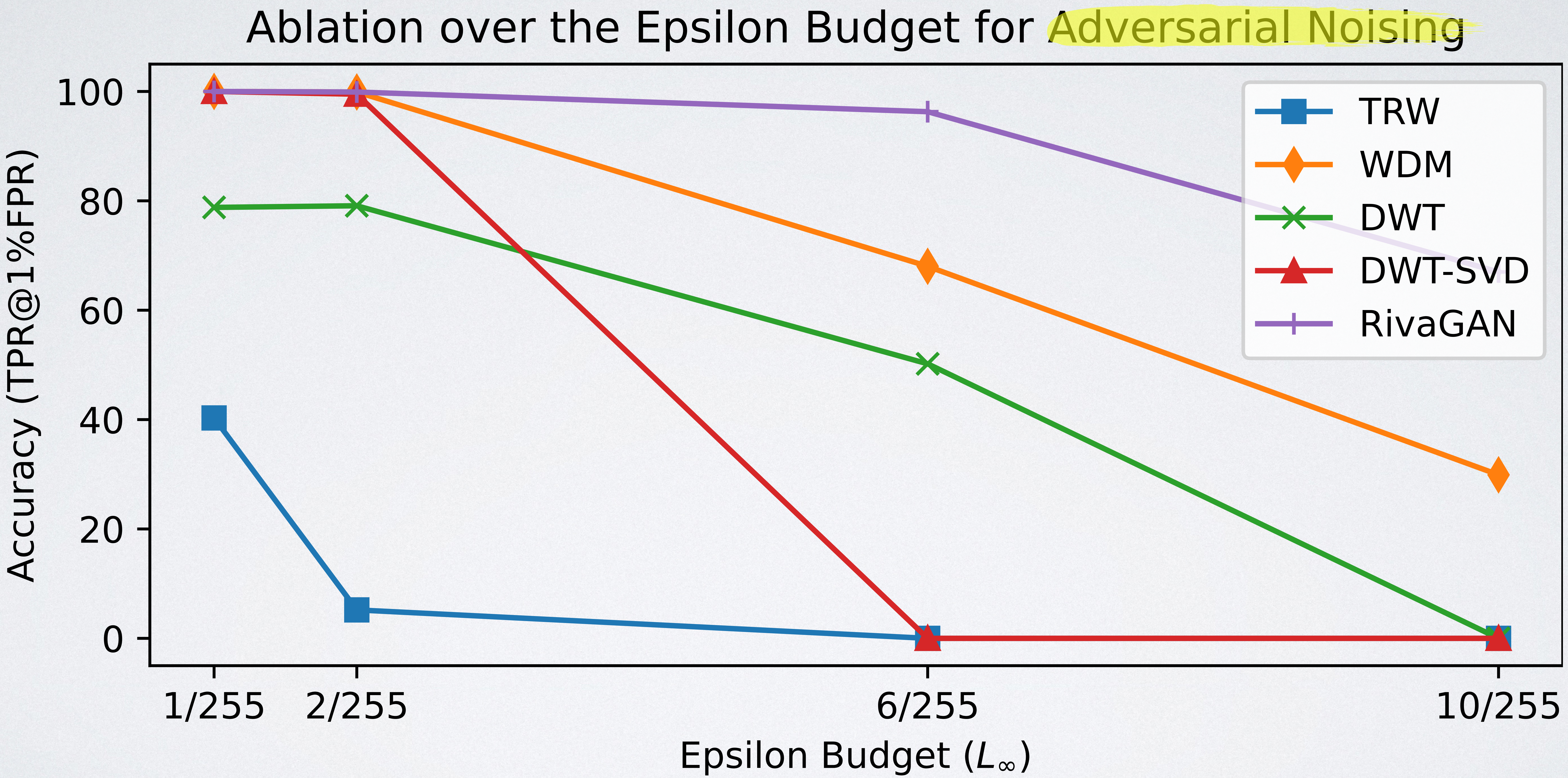
Instantiating Adaptive Attacks



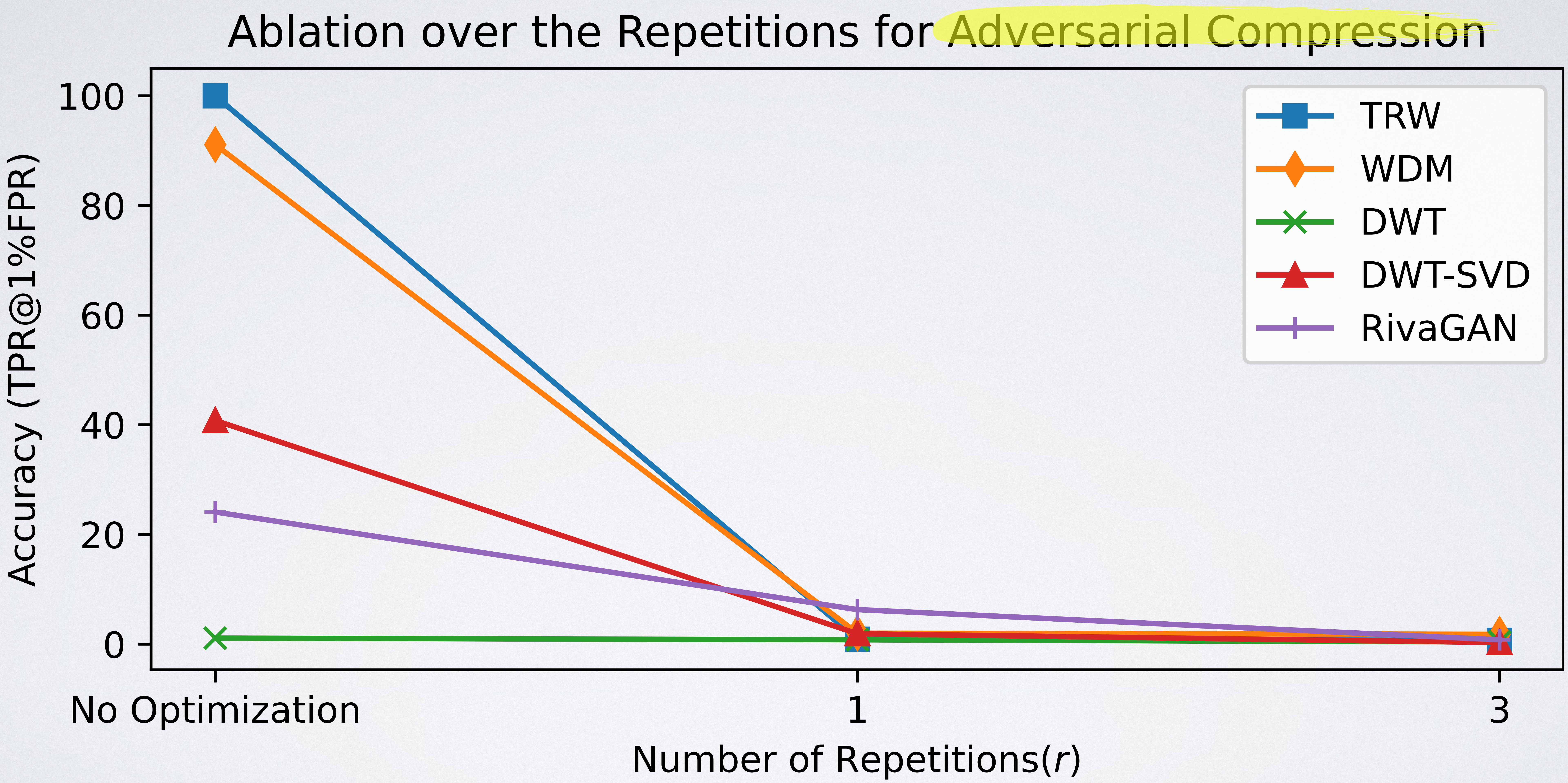
Instantiating Adaptive Attacks



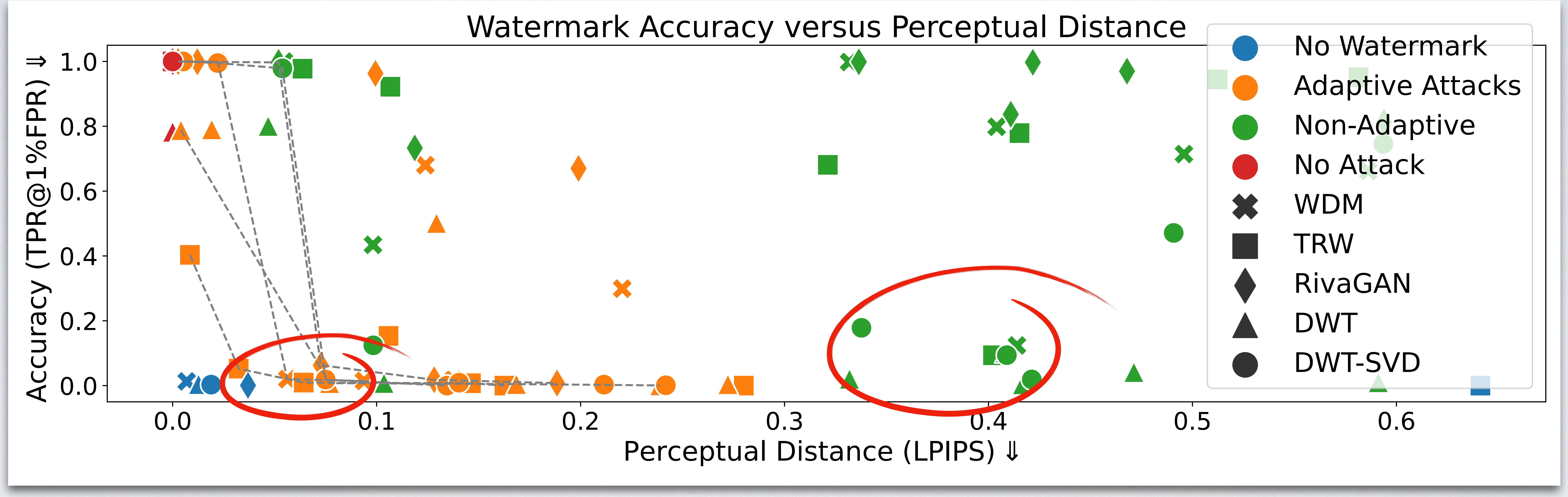
Instantiating Adaptive Attacks



Instantiating Adaptive Attacks



Comparison to Non-Adaptive Attacks



Adaptive Attacks

Non-adaptive Attacks

Visual Inspection

TRW: “Cars are parked on the street near an old building”

P-value = 0.28 | P-value = 1.77e-09 | P-value = 0.52

$\epsilon = 2/255$, L_∞-norm

ed on the street near an old building"

Can we defend against Adaptive Attackers?

Not necessarily differentiable!

Best parameters

Surrogate key

$$\max_{\theta_{\mathcal{A}}} \mathop{\mathbb{E}}_{\substack{\tau' \leftarrow \textsc{Keygen}(\theta_G) \\ m \in \mathcal{M}}} \left[\textsc{Verify}(\mathcal{A}(G_W), \tau', m) + Q(\mathcal{A}(G_W), G_W) \right]$$

For any Key-message pair

Image after attack

Image quality before and after

Can we defend against Adaptive Attackers?

Problem

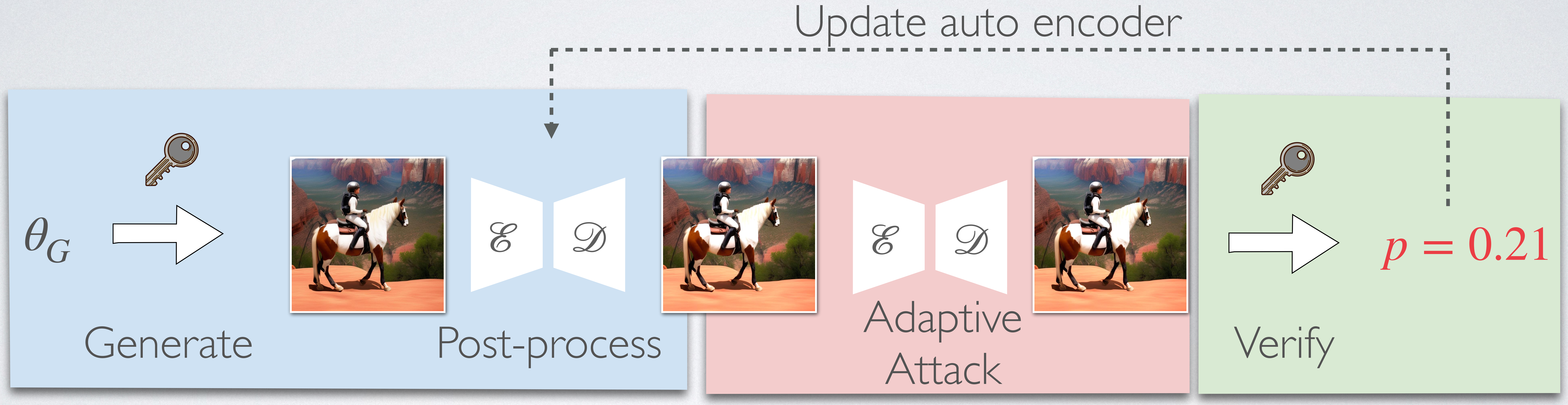
TRW is not easily fixable against these adaptive attacks

Possible Solutions

Learnable watermarks, in which we train encoder-decoder pairs
But how can we design them?

Can we defend against Adaptive Attackers?

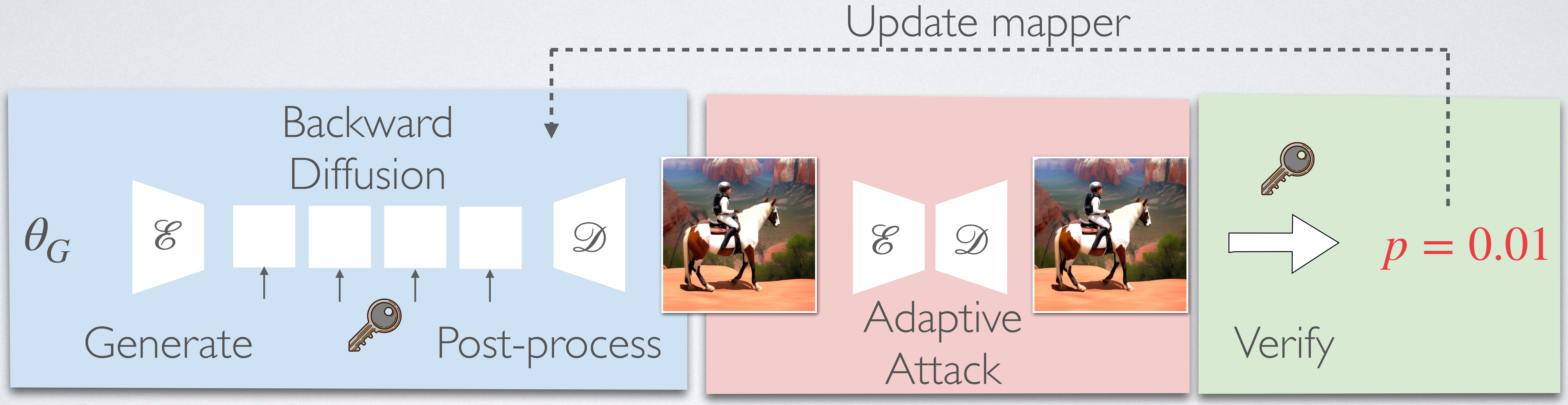
Idea 1: **Post-processing**



Problem: Is the space of possible defense strategies large enough?
There may not be an (efficient) solution!

Can we defend against Adaptive Attackers?

Idea 2: **Distributional Shift**



Problem: Is there an (efficient) solution?

Challenges of Learnable Watermarking

1.) One-shot agents (e.g., GANs)

Latent → Generator → 

✗ All gradients observable
✗ Alignment through pivot
✗ Continuous & high-entropy

2.) Iterative Optimization (e.g., Stable Diffusion)

Noise → Denoise 1 → .. → Denoise N → 

3.) Discrete Iterative Optimization (e.g., Language Models)

Prompt → LM → .. → LM → "A brown horse rides on a .."

Discussion

How scalable are these attacks?

Will open-source model contain robust watermarks?

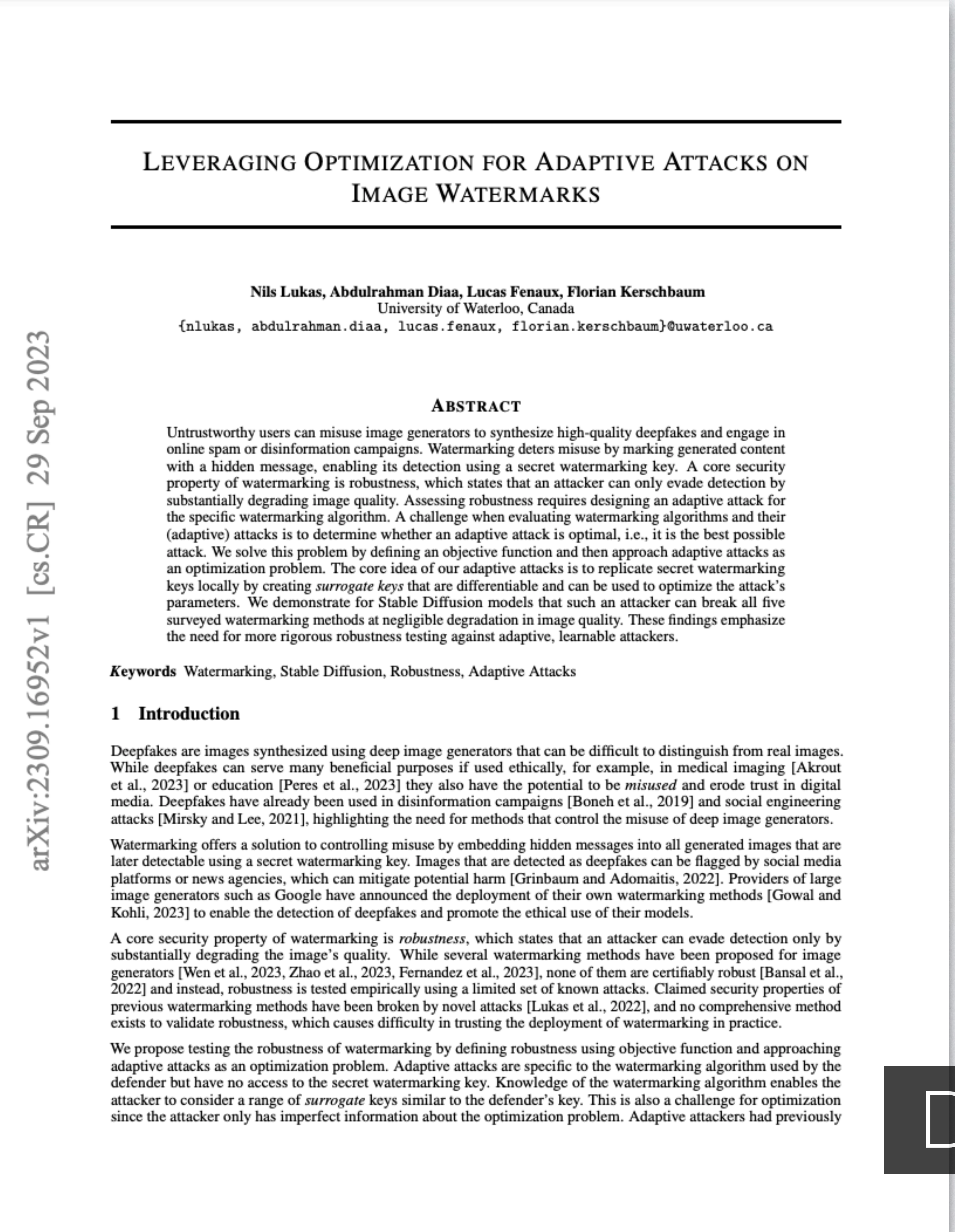
Certifiably robust watermarking?

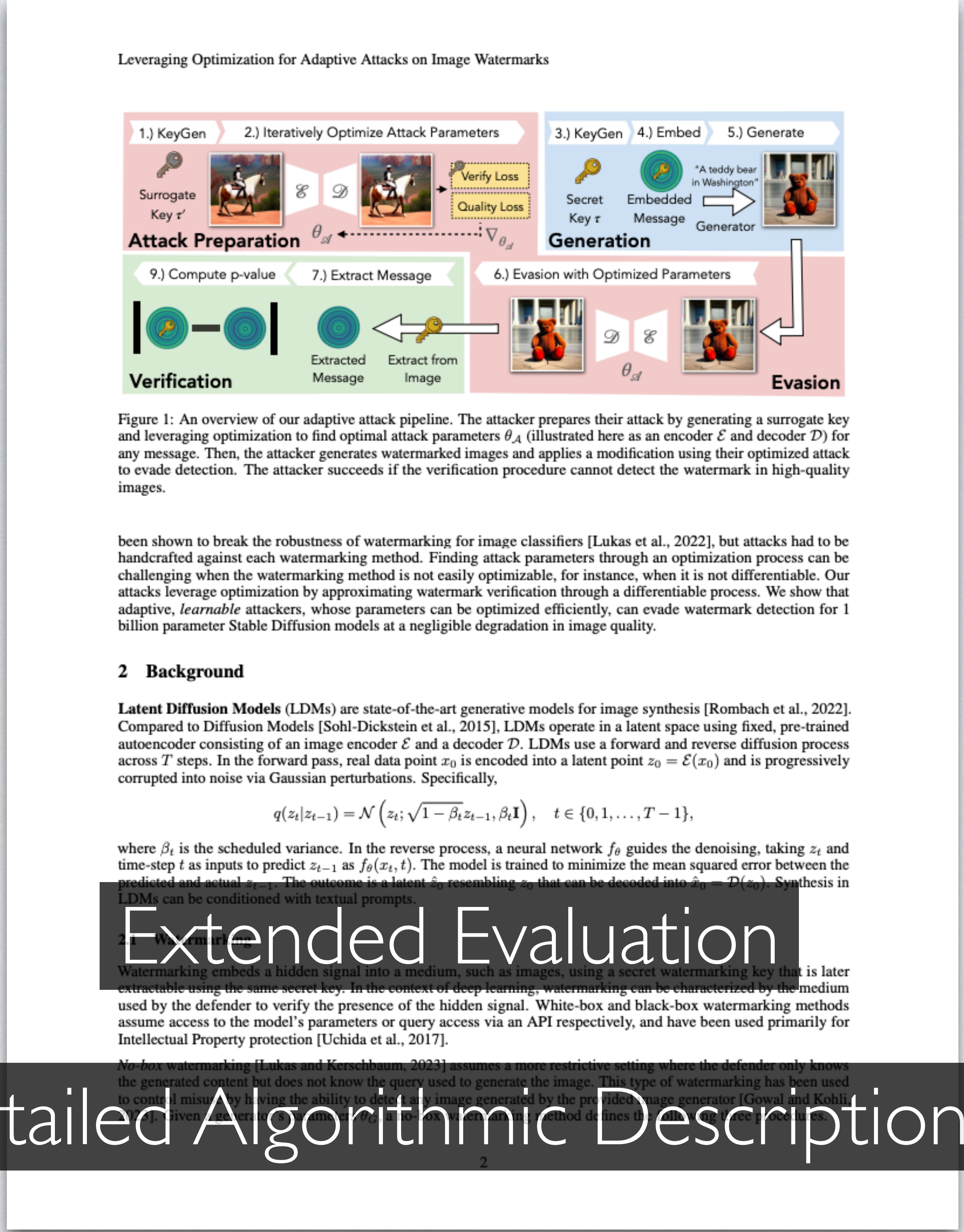
Extension to Language/Speech?

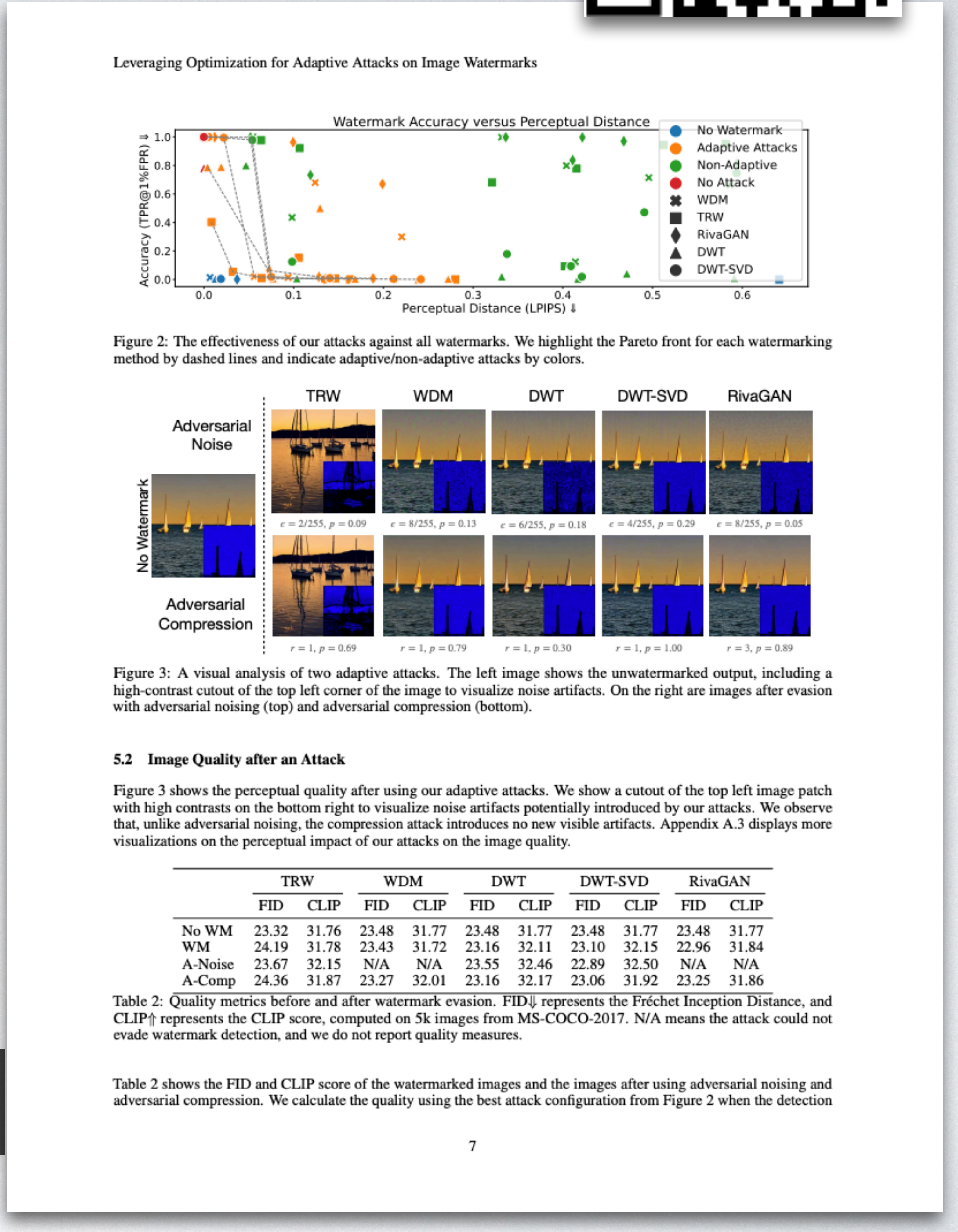
Ethical considerations

Limitations

The Paper contains more Information





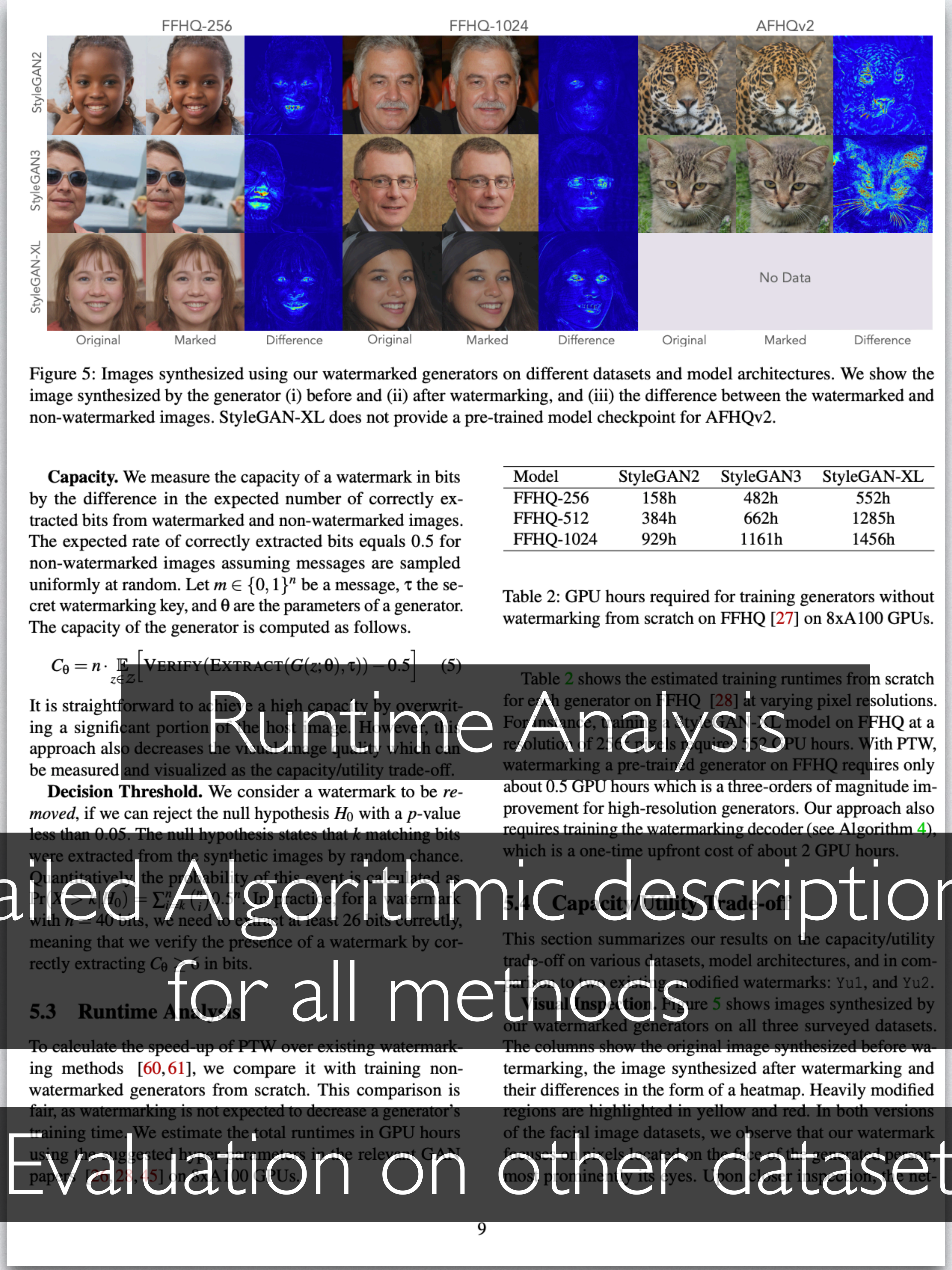


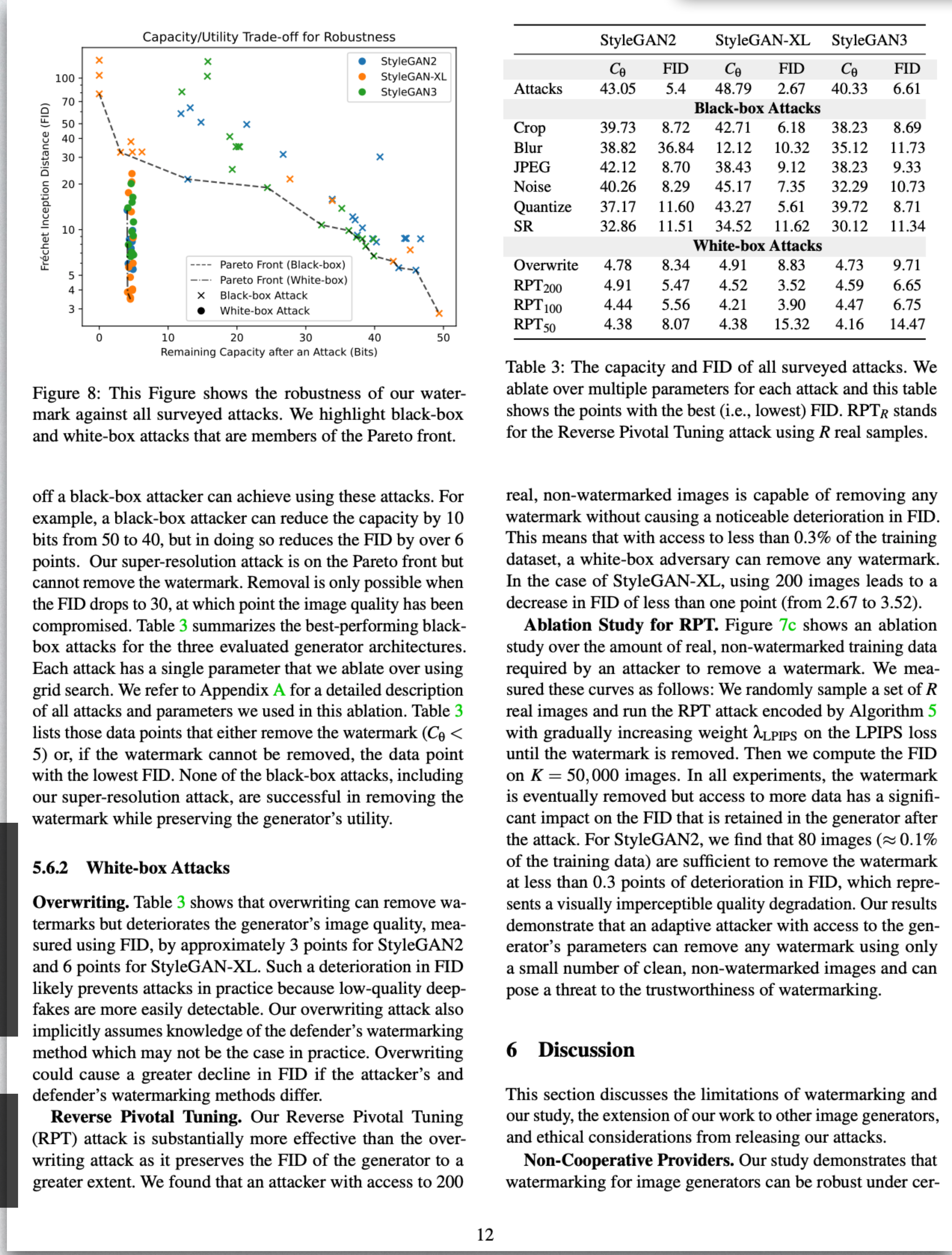
Extended Evaluation

Detailed Algorithmic Descriptions

Discussion & Ethics

The Paper contains more Information





Runtime Analysis

Detailed Algorithmic descriptions for all methods

Evaluation on other datasets

Robustness with more attacks

How Reliable is Watermarking for Generative Machine Learning?

Source code: https://github.com/dnn-security/gan-watermark

Nils Lukas

UNIVERSITY OF WATERLOO

CrySP
Cryptography, Security, and Privacy
Research Group

Source Code

USENIX'23

Sources

[1] Karras, Tero, et al. "Alias-free generative adversarial networks." *Advances in Neural Information Processing Systems* 34 (2021): 852-863.

[2] "How a fake image of a Pentagon explosion shared on Twitter caused a real dip on Wall Street", Luke Hurst, https://www.euronews.com/next/2023/05/23/fake-news-about-an-explosion-at-the-pentagon-spreads-on-verified-accounts-on-twitter, Accessed June 25 2023

[3] "The viral AI-generated image showing an explosion near the Pentagon is 'truly the tip of the iceberg of what's to come,' tech CEO says", Grace Dean, https://www.businessinsider.com/ai-generated-images-deepfake-pentagon-explosion-tip-of-the-iceberg-2023-6, Accessed June 25 2023

[4] "Fake Pentagon explosion photo goes viral: How to spot an AI image", Mohammed Haddad, https://www.aljazeera.com/news/2023/5/23/fake-pentagon-explosion-photo-goes-viral-how-to-spot-an-ai-image, Accessed June 25 2023

[5] "Terms of use", OpenAI, https://openai.com/policies/terms-of-use, Accessed June 25 2023

[6] "GENERATIVE AI PROHIBITED USE POLICY", Google, https://policies.google.com/terms/generative-ai/use-policy, Accessed June 25 2023

[7] Saharia, Chitwan, et al. "Photorealistic text-to-image diffusion models with deep language understanding." *Advances in Neural Information Processing Systems* 35 (2022): 36479-36494.

[8] Kang, Minguk, et al. "Scaling up gans for text-to-image synthesis." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[9] https://www.louisbouchard.ai/latent-diffusion-models/

[10] "StyleGAN3 Synthetic Image Detection", NVIDIA, https://github.com/NVlabs/stylegan3-detector, Accessed June 25 2023

[11] Dong, Chengdong, Ajay Kumar, and Eryun Liu. "Think Twice Before Detecting GAN-generated Fake Images from their Spectral Domain Imprints." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Sources

[12] “Russian TV and radio stations hacked with fake Putin message”, ,https://www.dw.com/en/russian-tv-and-radio-stations-hacked-with-fake-putin-message/a-65830291

[13] “Russia Says 'Fake' Putin Address Declaring Martial Law Was a 'Hack'”, Matthew Gault ,https://www.vice.com/en/article/ak33ge/russia-says-fake-putin-address-declaring-martial-law-was-a-hack

[14] “Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS”, EU Legislation, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206, Accessed June 26th

[15] https://apnews.com/article/deepfake-porn-celebrities-dalle-stable-diffusion-midjourney-ai-e7935e9922cda82fbcfb1e1a88d9443a, accessed August 9th

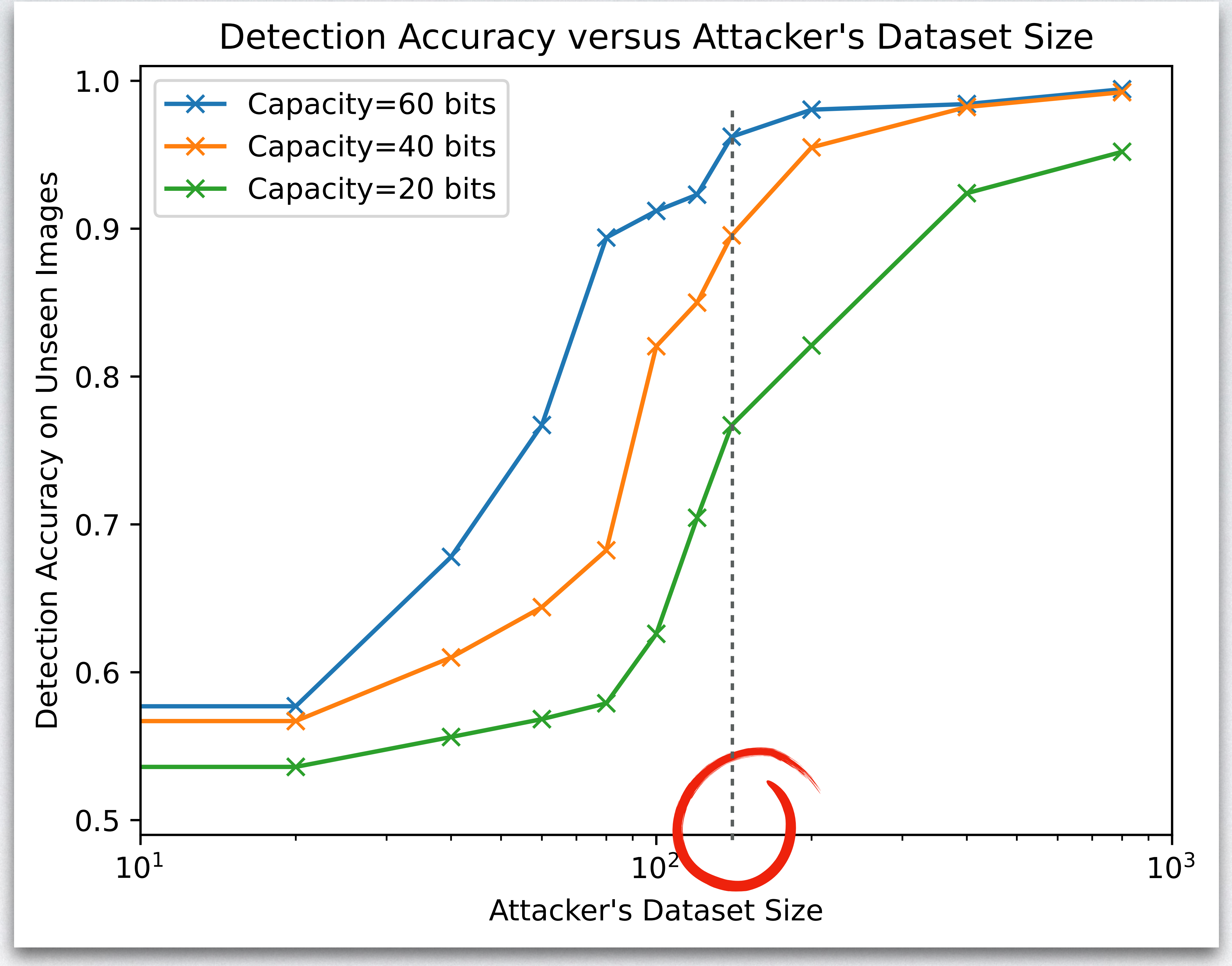
[16] https://www.nbcnews.com/tech/internet/deepfakes-twitter-tiktok-stars-rcna87295, accessed August 9th

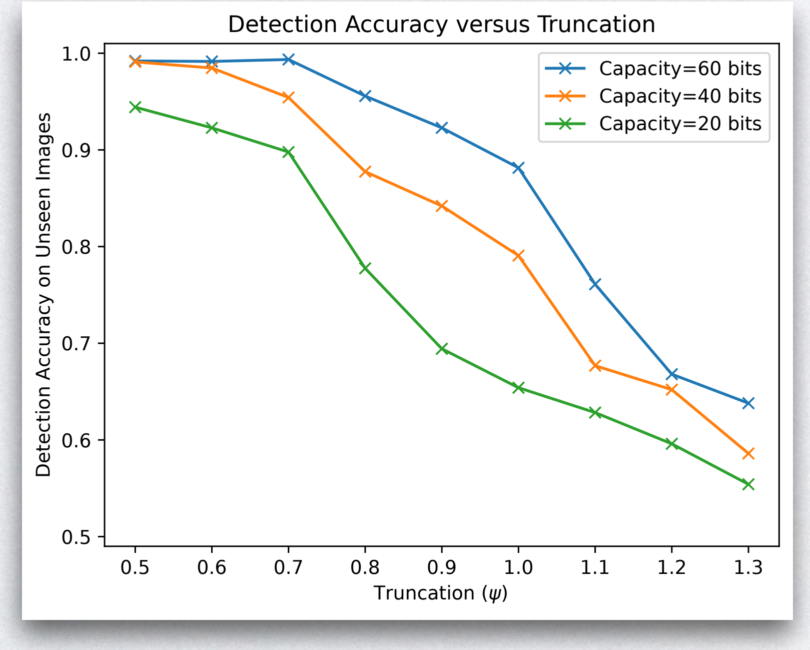
[17] https://www.ic3.gov/Media/Y2023/PSA230605, accessed August 9th

[18] https://www.reuters.com/technology/openai-google-others-pledge-watermark-ai-content-safety-white-house-2023-07-21/, accessed August 9th

Best Adaptive Attacks

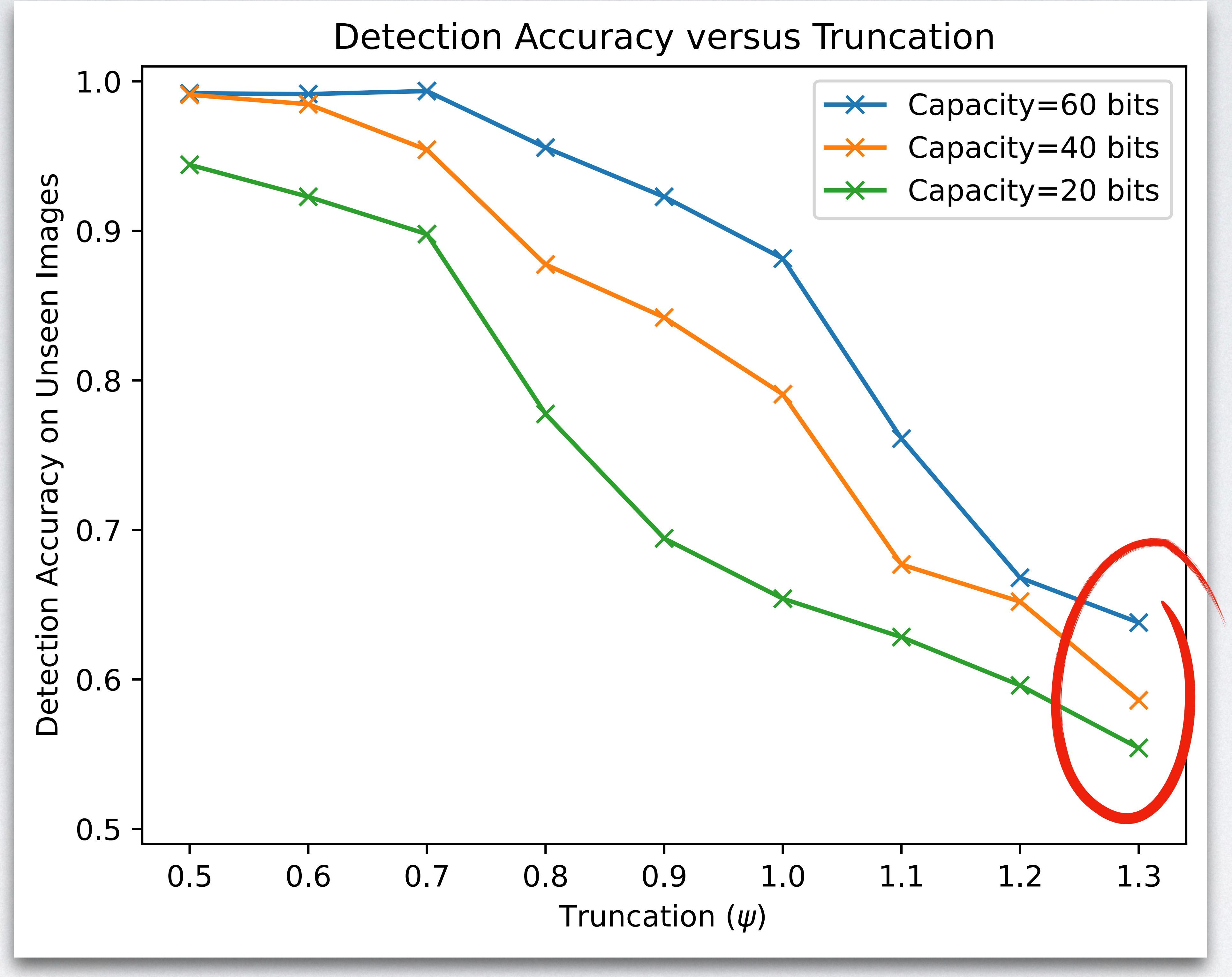
Detectability



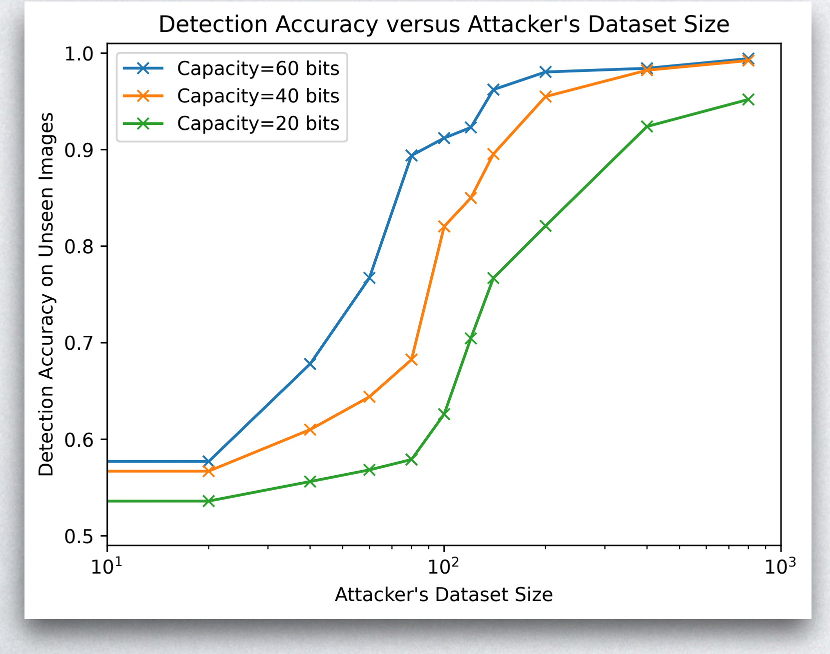


Detectability

Low variation



High variation



Variation makes detectability more difficult for the adversary

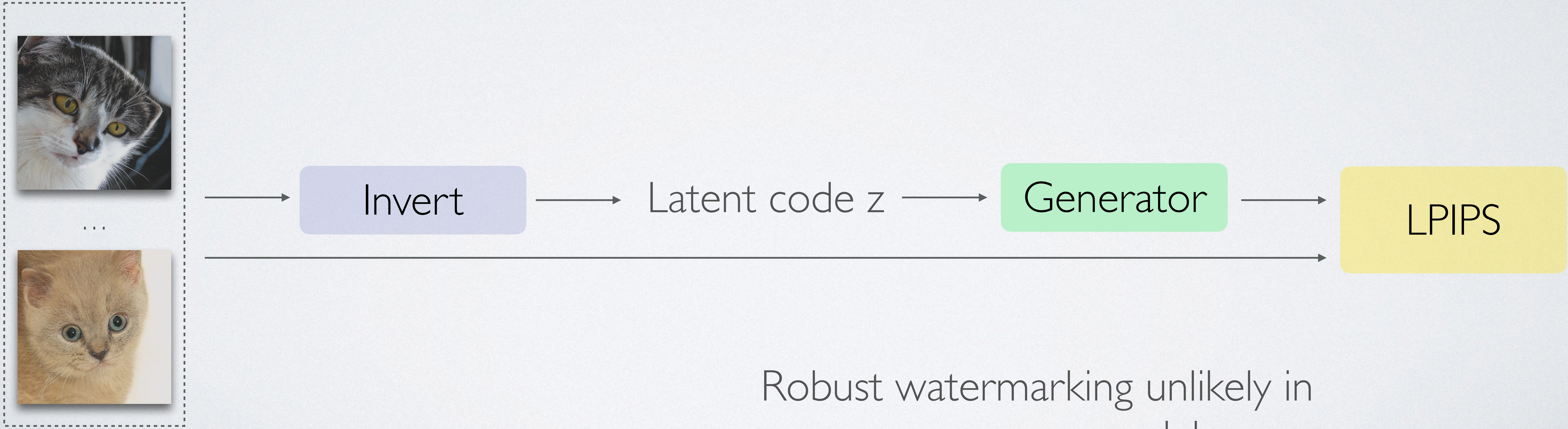
39

Reverse Pivotal Tuning

White-box

1) Invert real images into the generator's latent space

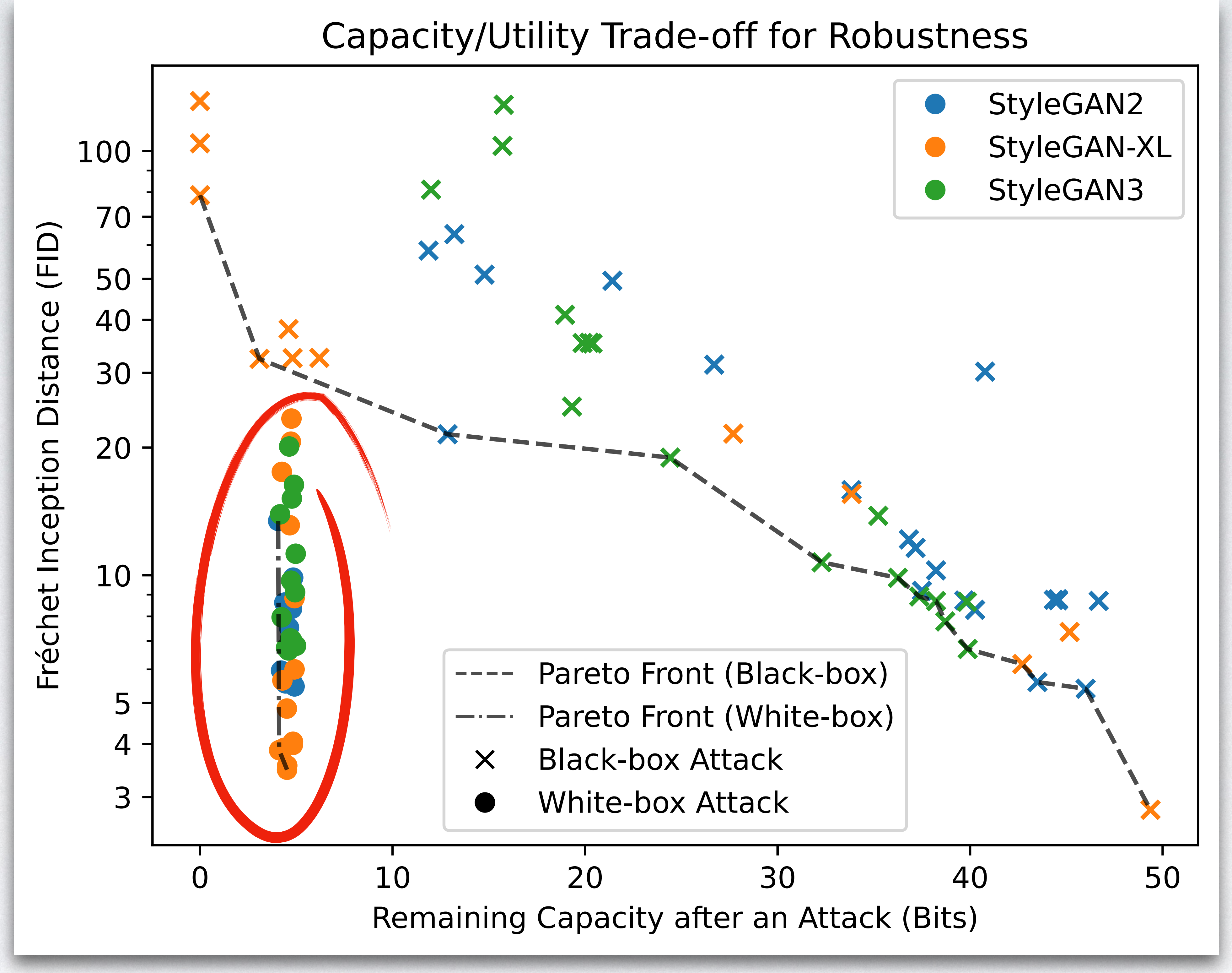
2) Regularize generator with Pivotal Tuning and LPIPS loss to synthesize real images



Adversary's Data

Robust watermarking unlikely in open-source models

Robustness



Black-box attacker cannot remove watermarks

White-box attacker can remove any watermark

Summary of Results

	Black-box	White-box
Robustness	✓	✗
Detectability	Scales with output diversity	
Effectiveness	40 bits at less than 0.3 FID	
Scalability	No retraining, < 2 GPU hours (FFHQ-256)	

The first post-hoc **learnable watermark** for deep image generators