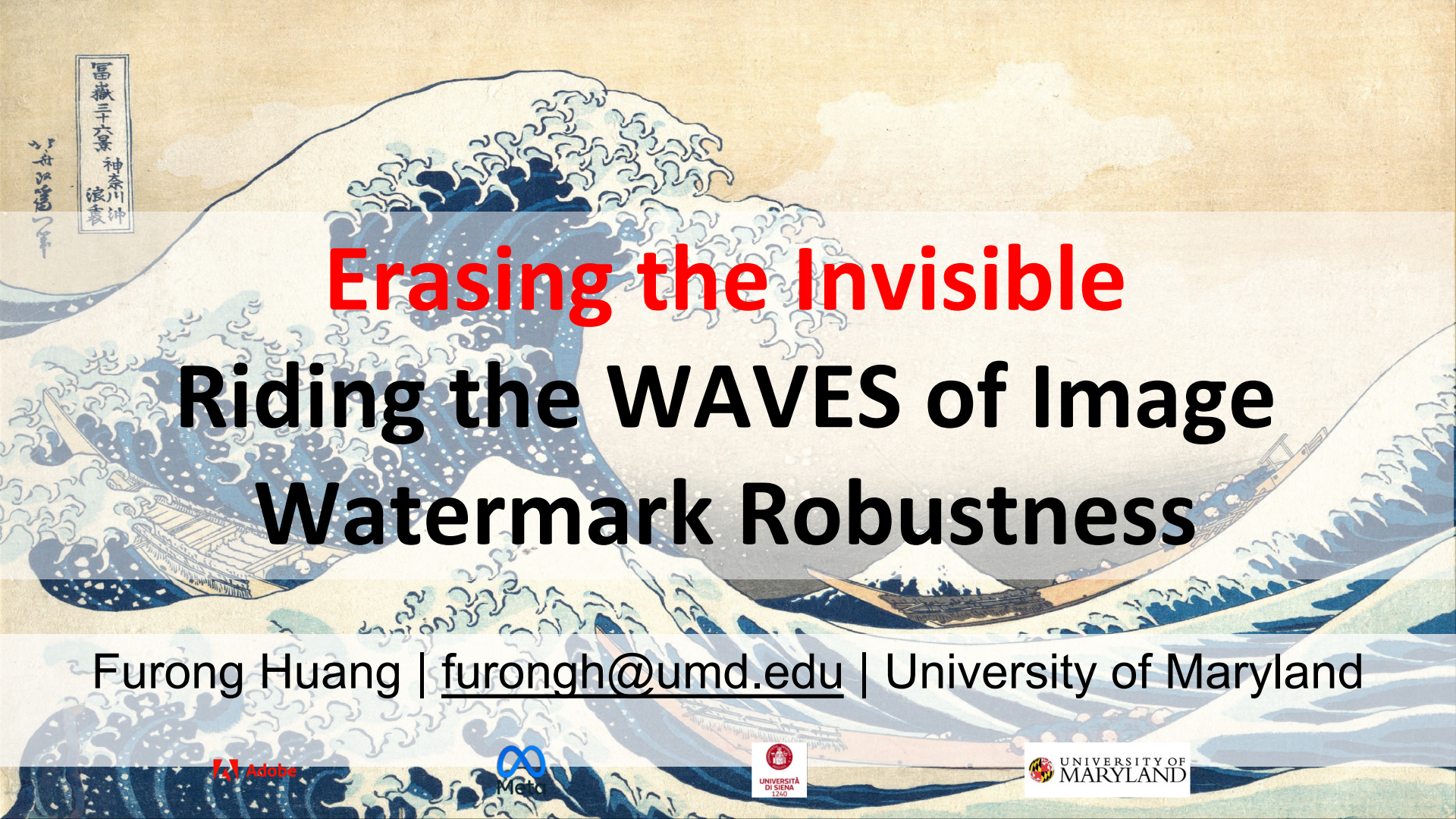
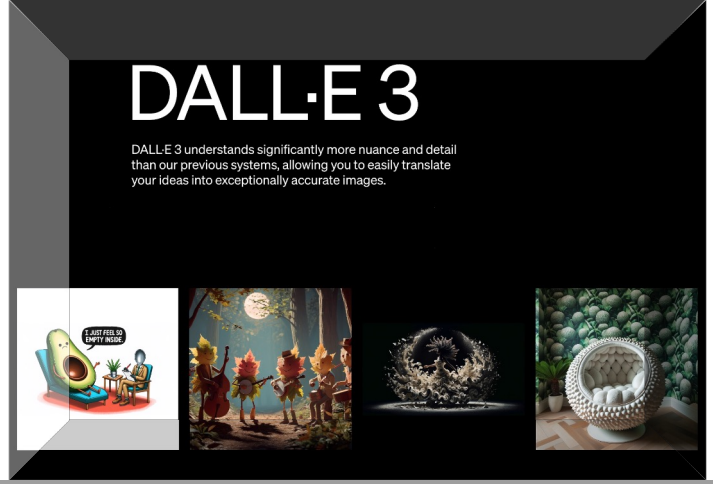


Erasing the Invisible

Riding the WAVES of Image Watermark Robustness

Furong Huang | furongh@umd.edu | University of Maryland

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

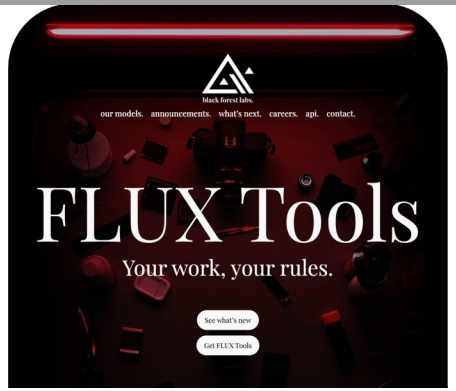




Text-to-Image Generative AI Models Booming







Adobe

Meta

UNIVERSITÀ DI SIENA 1240

UNIVERSITY OF MARYLAND







Creative Content Generation

Applications: Drug Discovery, Robotics, etc.







Adobe

UNIVERSITÀ DI SIENA 1240

MARYLA

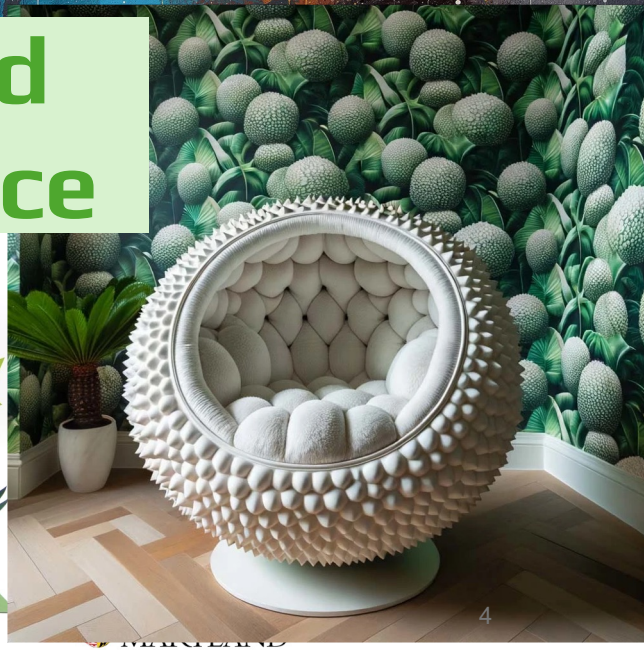


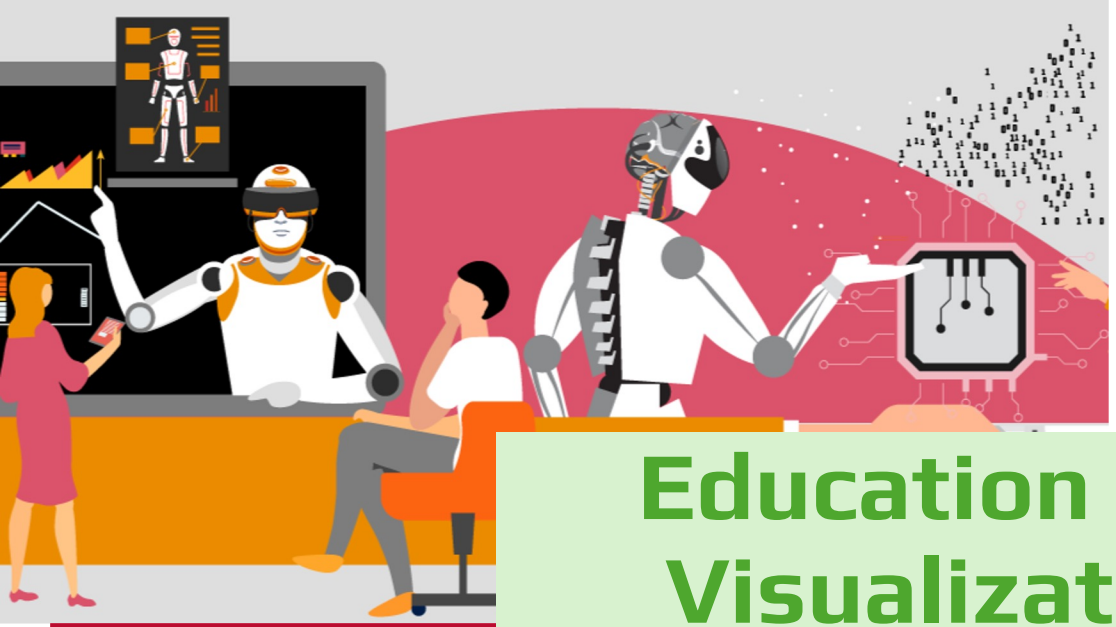


Prototyping and Design Assistance



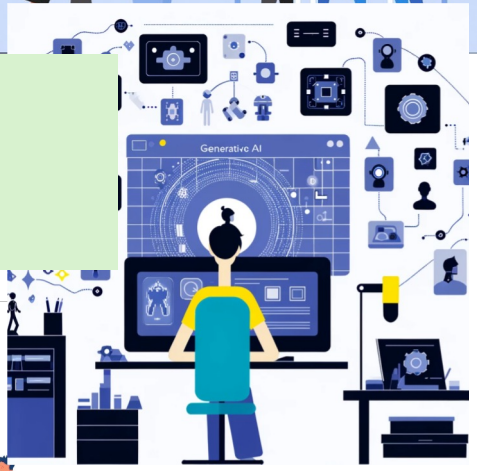




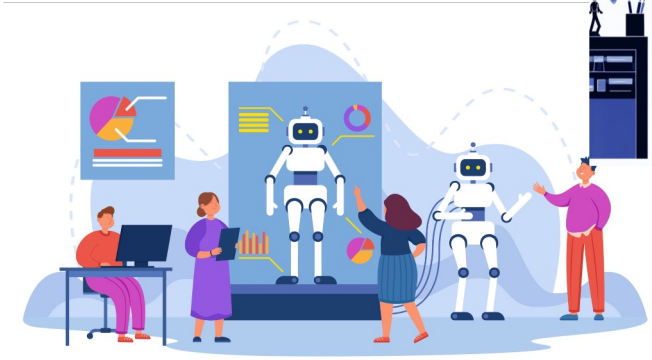


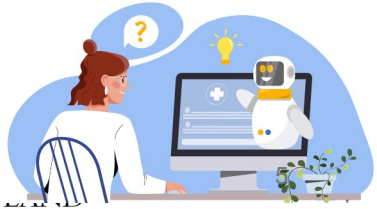


Education and Visualization







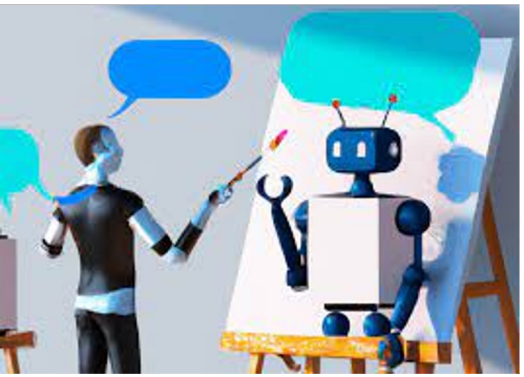


Adobe

Meta

UNIVERSITÀ DI SIENA 1240

UNIVERSITY OF MARYLAND





Adobe

Can you spot what's copied?

Stable Diffusion image of the UMD logo

Match from LAION training dataset





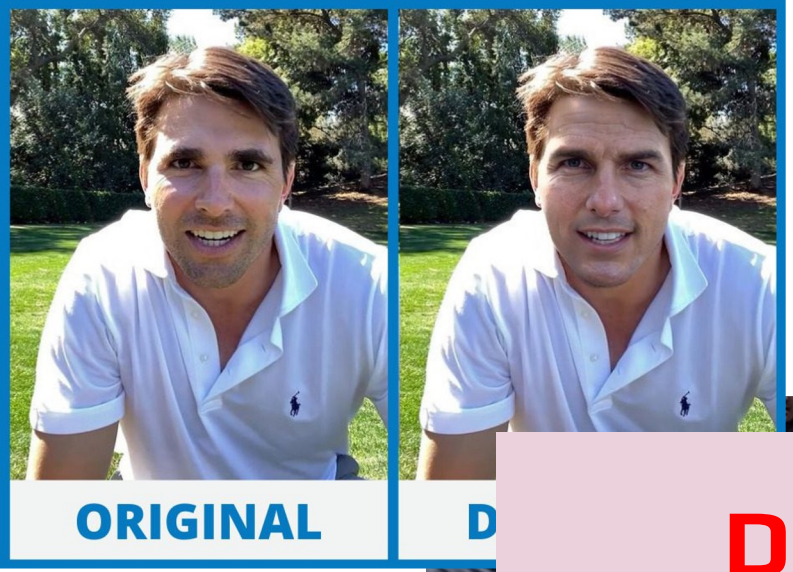
Intellectual Property Concerns

Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models, Somepalli et al, CVPR 2023

Meta

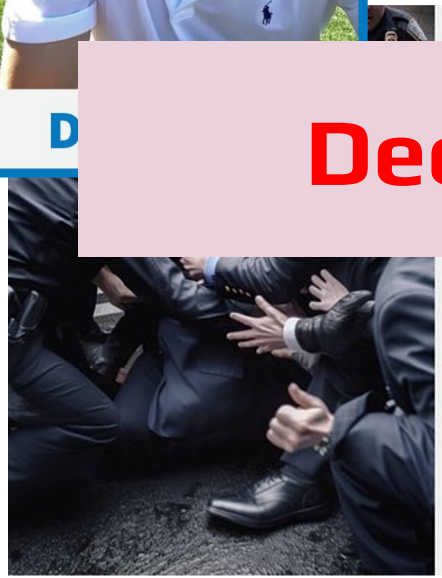
UNIVERSITÀ DI SIENA 1240

UNIVERSITY OF MARYLAND





Deep Fake

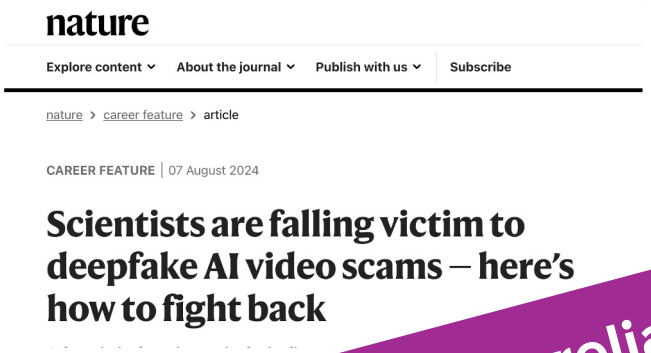








- AI-generated or Real?
- Which AI Platform generated it?
- Which user generated and spread it?



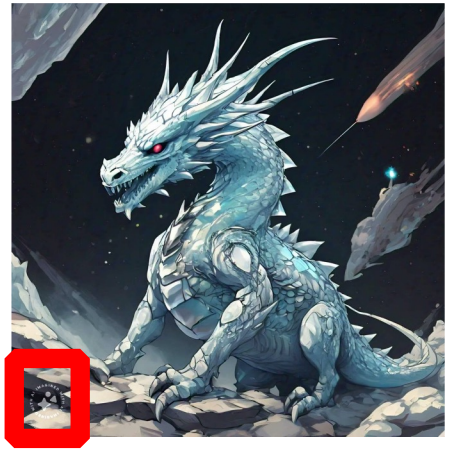
Deep F

Urgent need for **reliable** mechanisms to **identify** and **attribute** AI-generated content



Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

Image Watermarks: Embedded invisible signatures



Meta AI

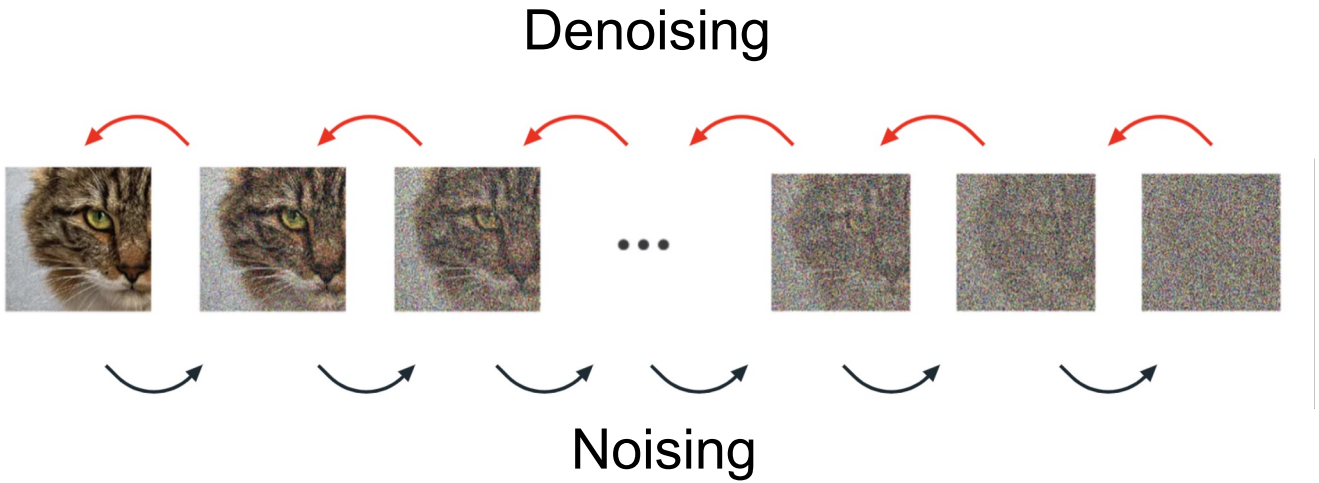


DALL-E2



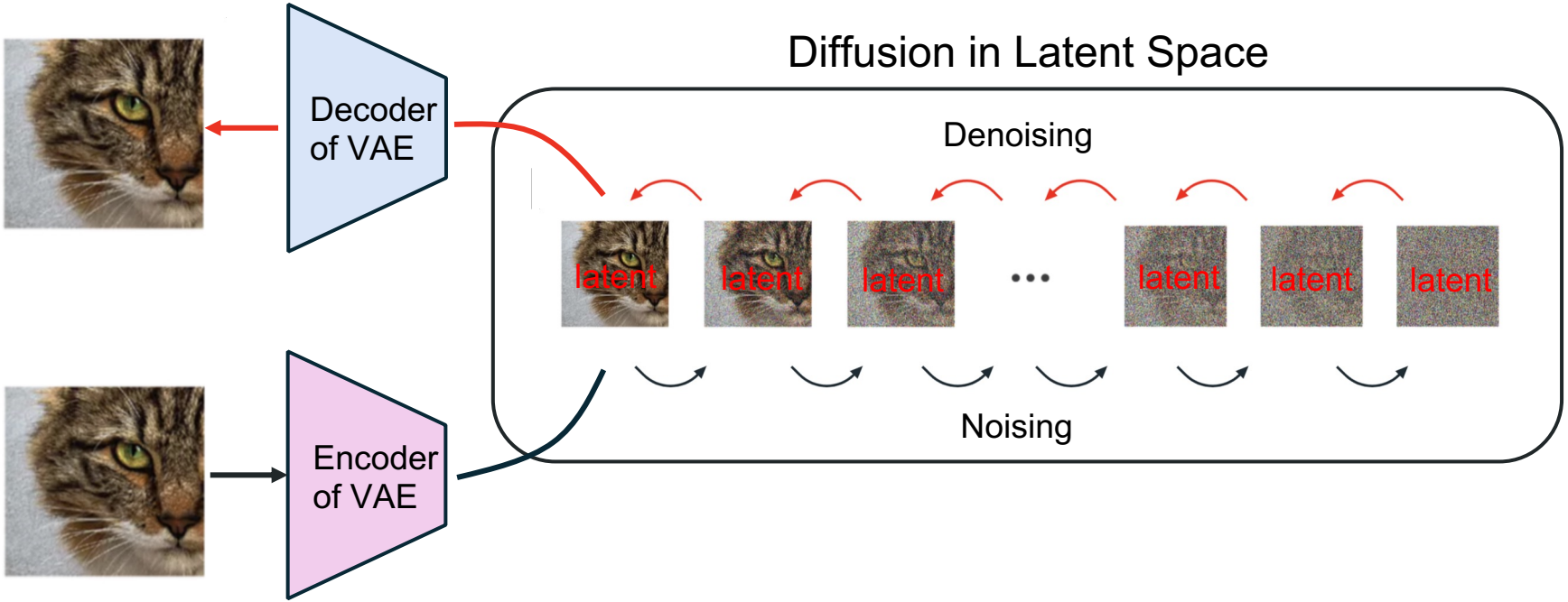
Bing

How to Generate Images?



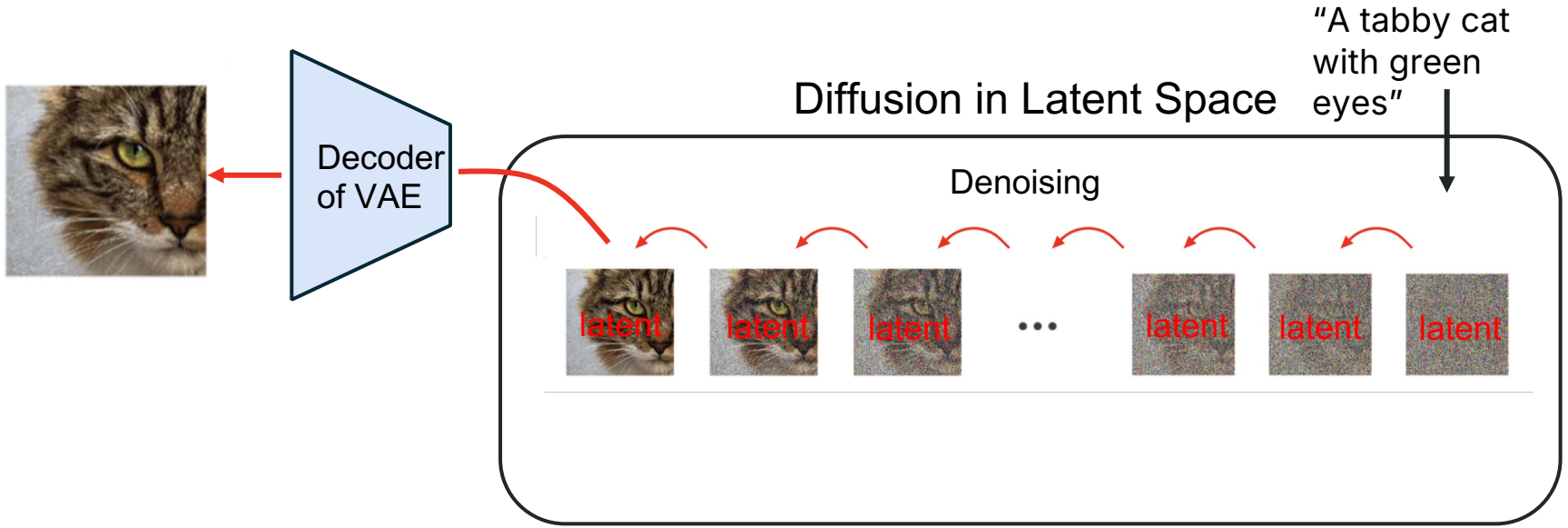
How to Generate Images?

Stable Diffusion



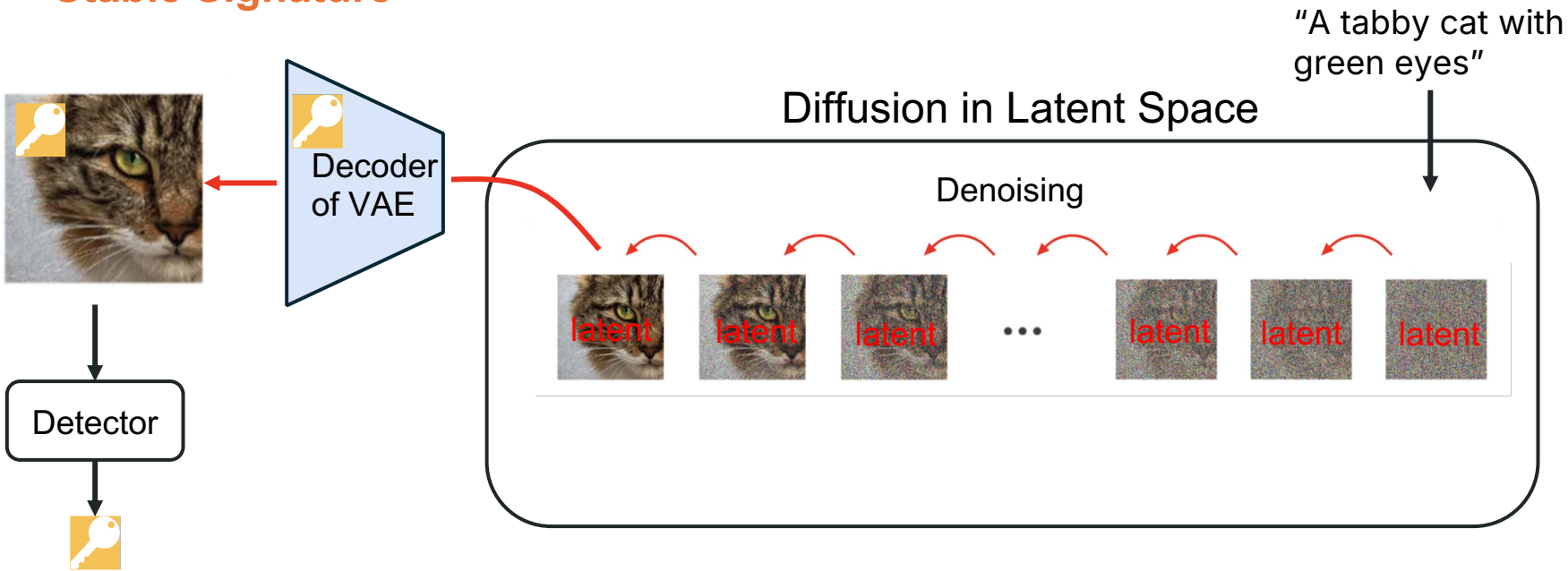
How to Generate Images?

Stable Diffusion



How to Add Watermarks?

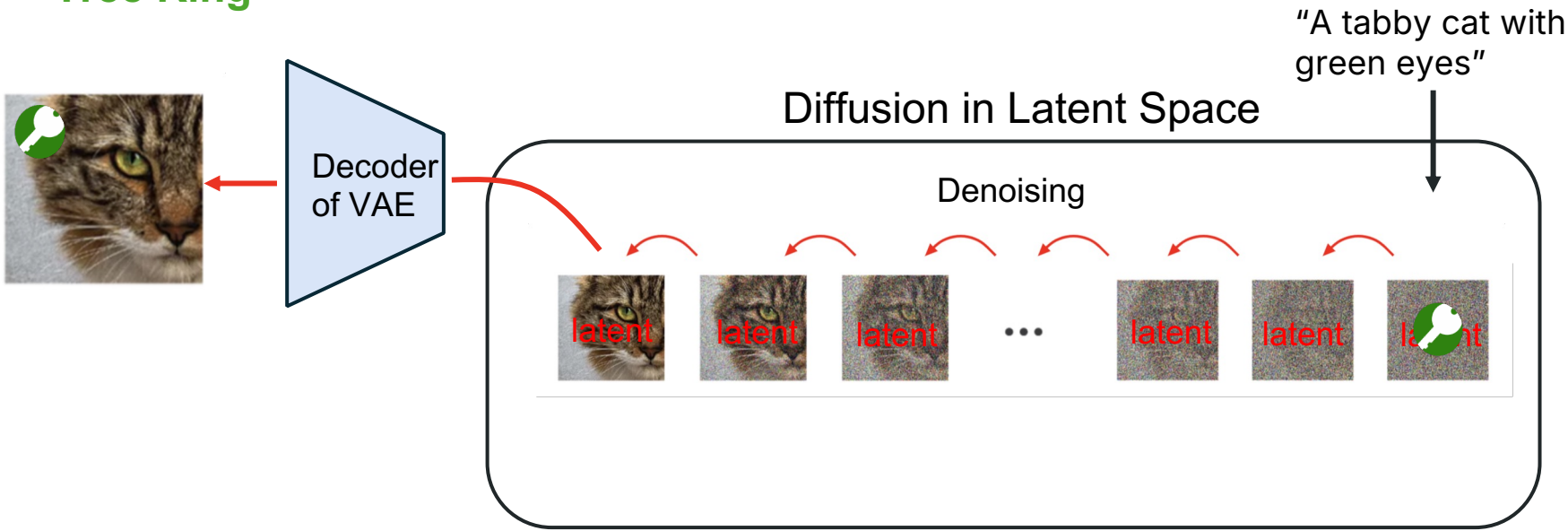
Stable Signature



Fernandez et al., "The stable signature: Rooting watermarks in latent diffusion models." (2023)

How to Add Watermarks?

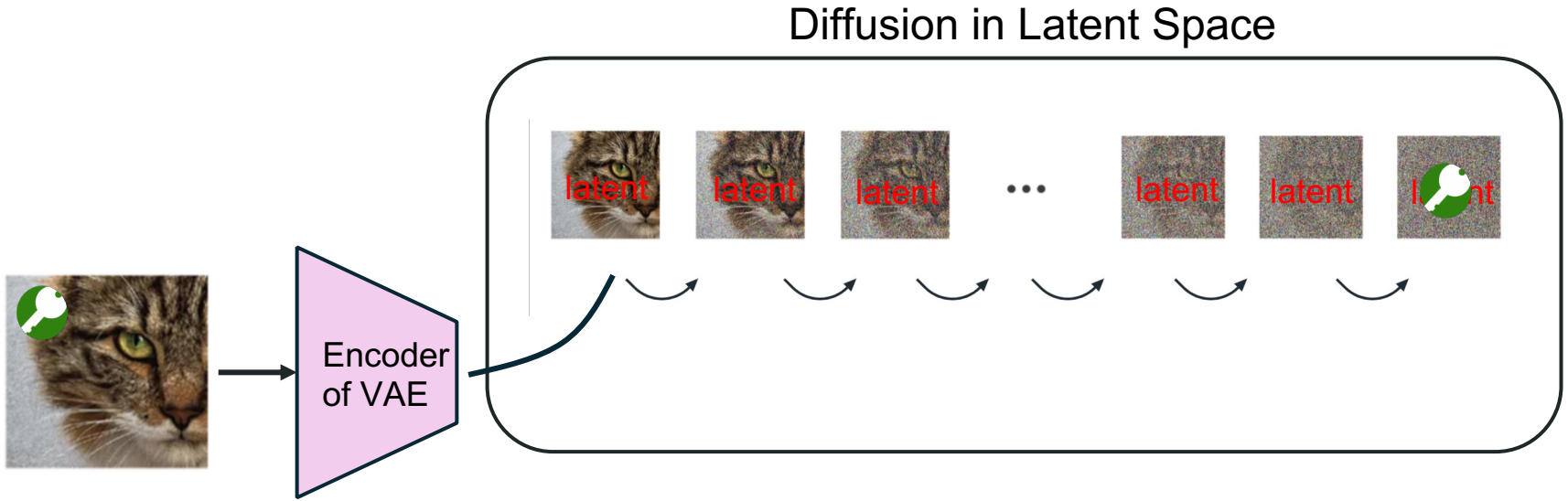
Tree Ring



Wen et al., "Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust." (2023)

How to Add Watermarks?

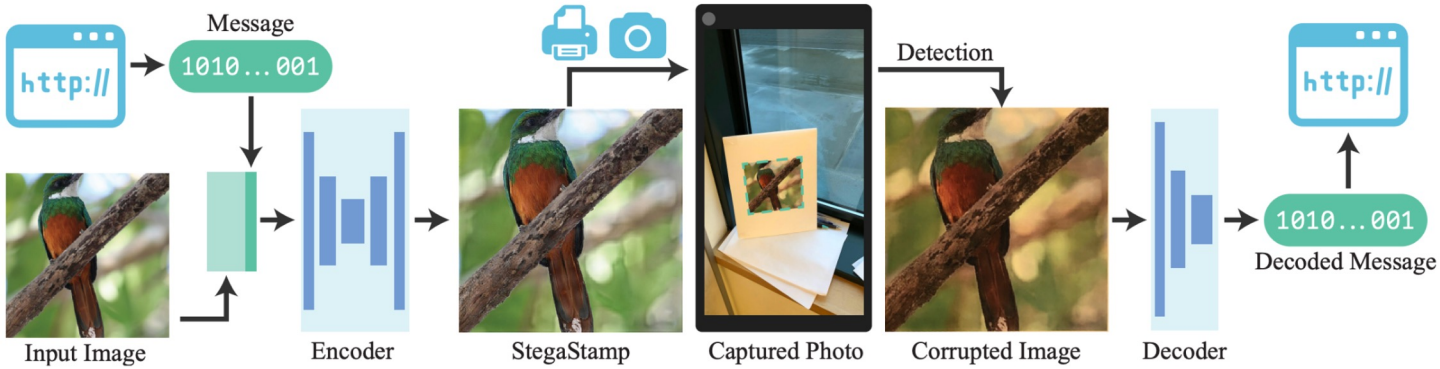
Tree Ring (Detection)



Wen et al., “Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust.” (2023)

How to Add Watermarks?

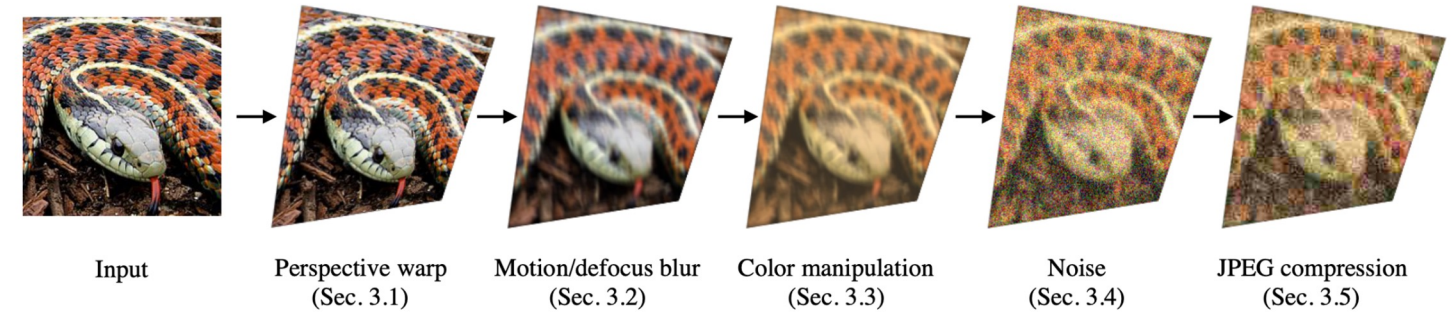
StegaStamp



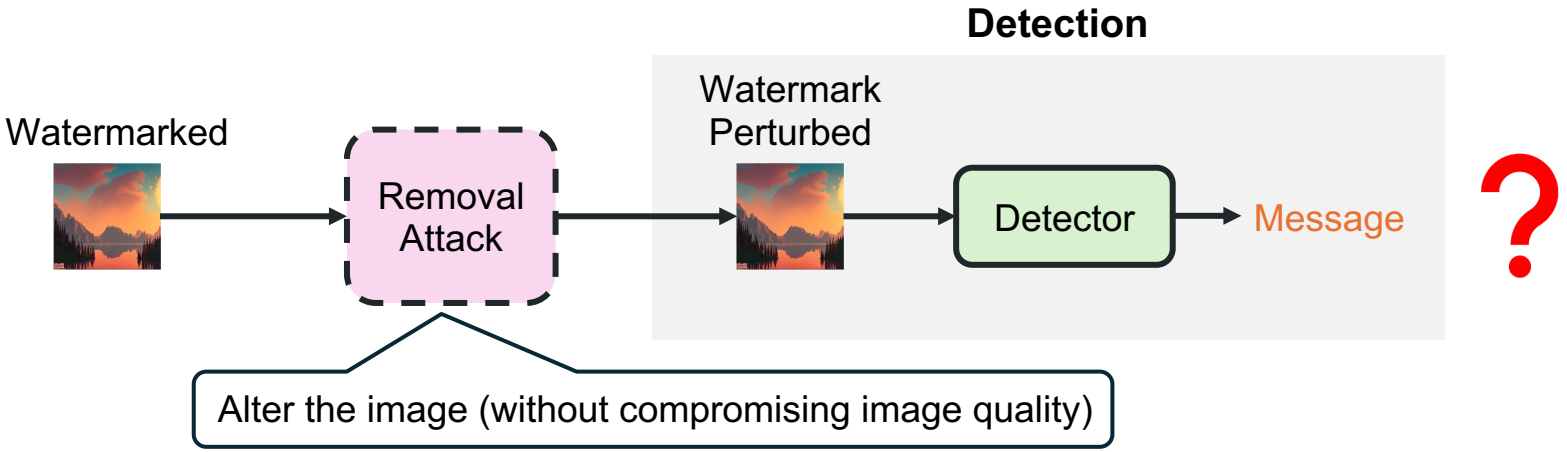
How to Add Watermarks?

StegaStamp

Decoder is trained to handle physical + artificial perturbations



Robustness of Image Watermarks



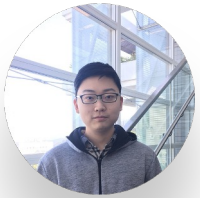
Modern Watermarks:

- Tree-Ring (Wen et al., 2023), Gaussian Shading (Yang et al. 2024), PRC (Gunn et al. 2024)
- Stable Signature (Fernandez et al., 2023)
- StegaStamp (Tancik et al., 2020)
- …

How robust?

WAVES: Benchmarking the Robustness of Image Watermarks


Bang An


Mucong Ding


Tahseen Rabbani


Aakriti Agrawal

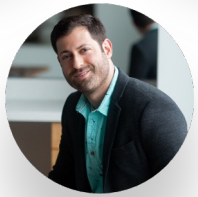

Yuancheng Xu


Chenghao Deng


Sicheng Zhu


Abdirisak Mohamed


Yuxin Wen

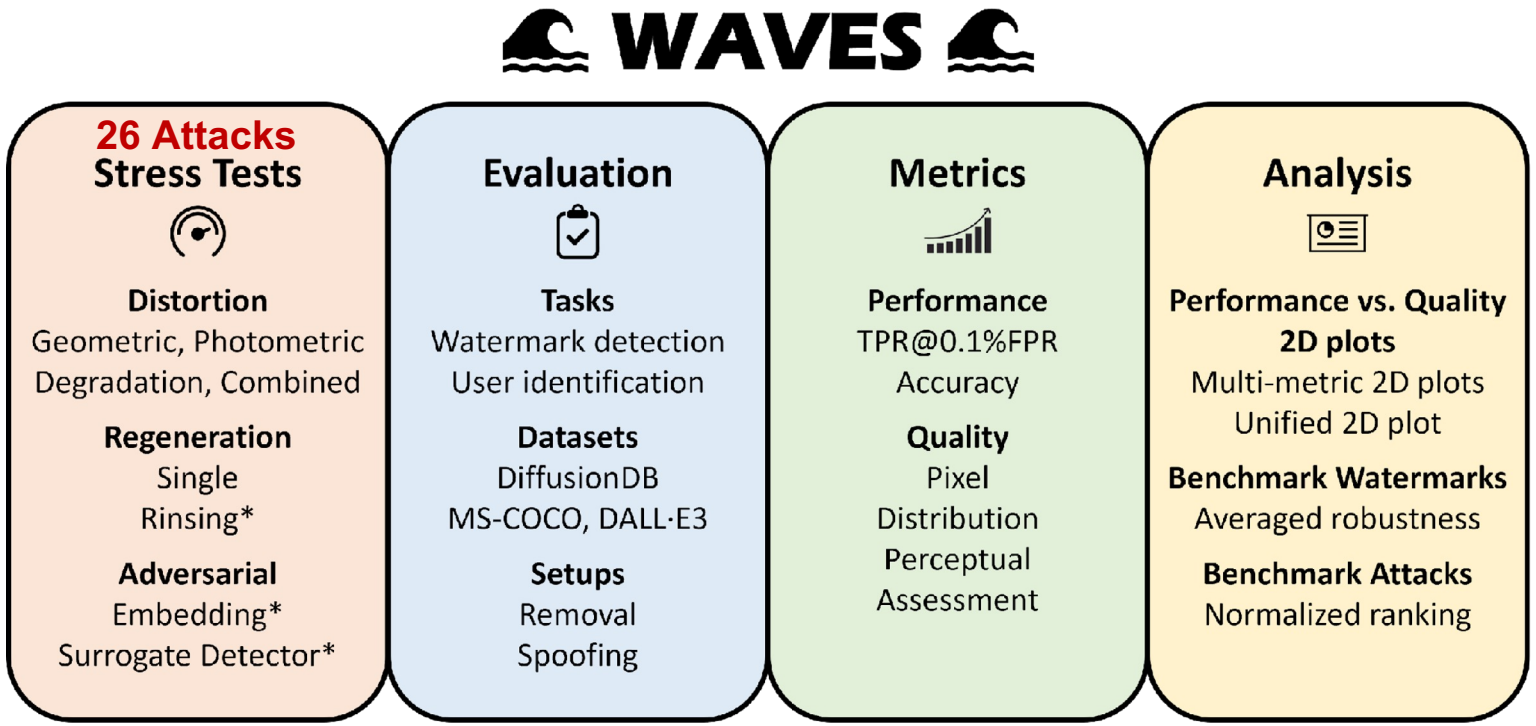

Tom Goldstein


Furong Huang

[arXiv:2401.08573, ICML 2024]

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

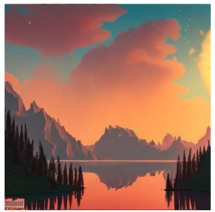
Watermark **A**nalysis **v**ia **E**nhanced **S**tress-Testing



Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

Distortion Attacks

unattacked



Attacker Knows
Nothing

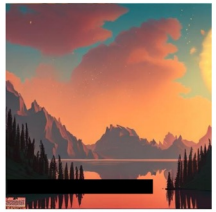
Classical Distortions



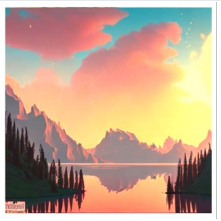
Rotation



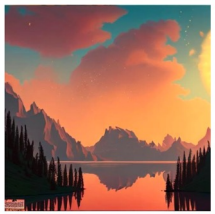
Cropping



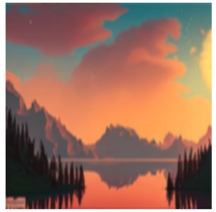
Erase



Brightness



Contrast



Blur



Noise



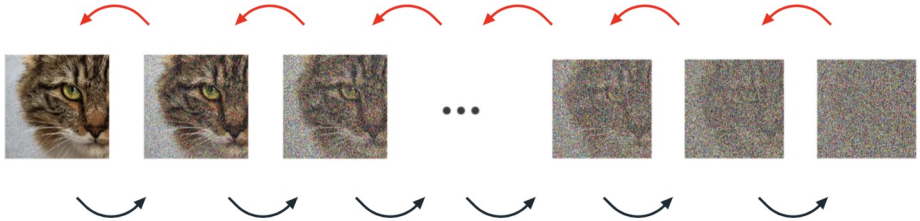
JPEG

Regeneration Attacks

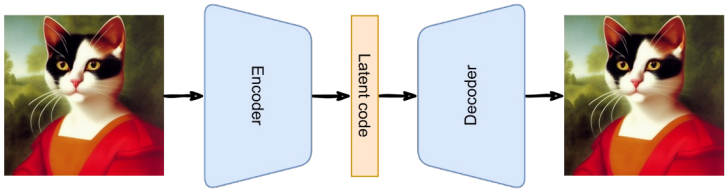
Open-source reconstruction

Keep semantics
Alter details

Diffusion models

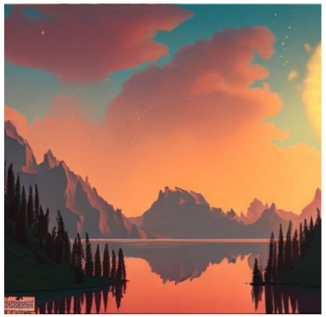


VAES

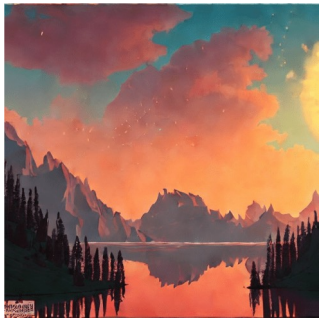


Zhao, Xuandong, et al. "Invisible image watermarks are provably removable using generative ai." *Arxiv* (2023).

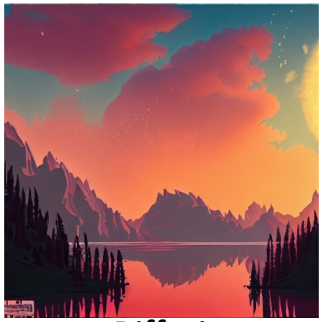
unattacked



Single Regeneration



VAE



Diffusion

Rinsing Regeneration



x1



x2



x4

Adobe

Meta

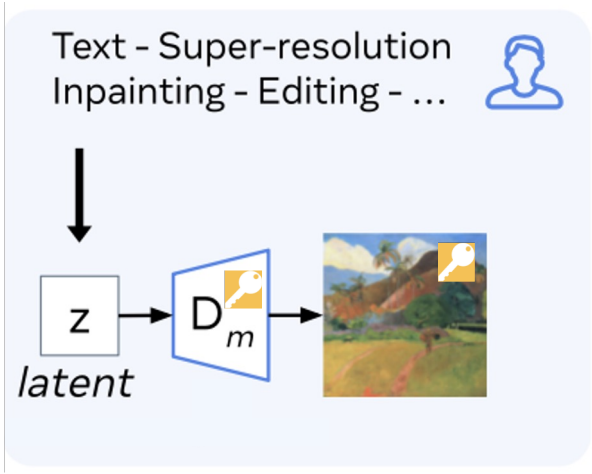
UNIVERSITÀ DI SIENA 1240

UNIVERSITY OF MARYLAND

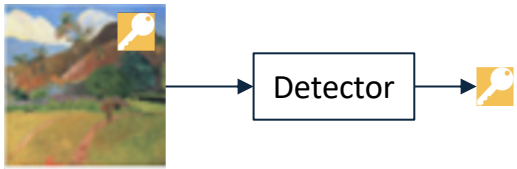
Regeneration Attacks

Stable Signature is vulnerable to Regeneration Attacks

Add Watermark

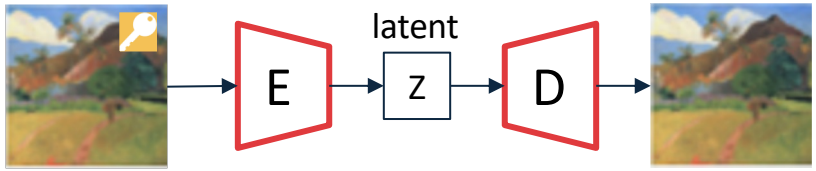


Detection

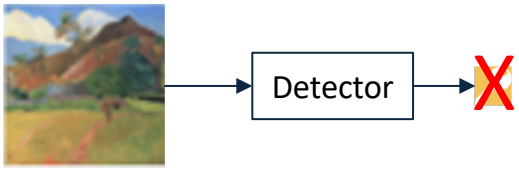


	Attacker Knows
Black-box	Nothing
Grey-box	Encoder in victim model

Regeneration Attack



Detection



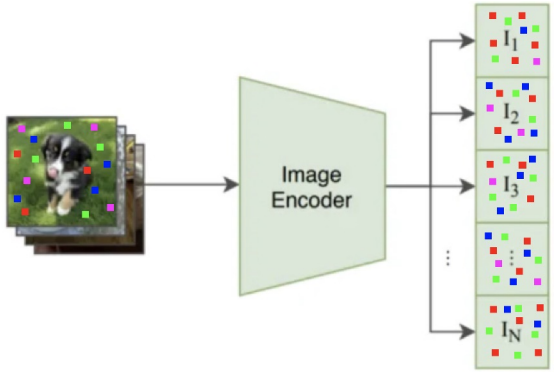
Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

Adversarial Attacks

Type 1: Confuse a surrogate detector

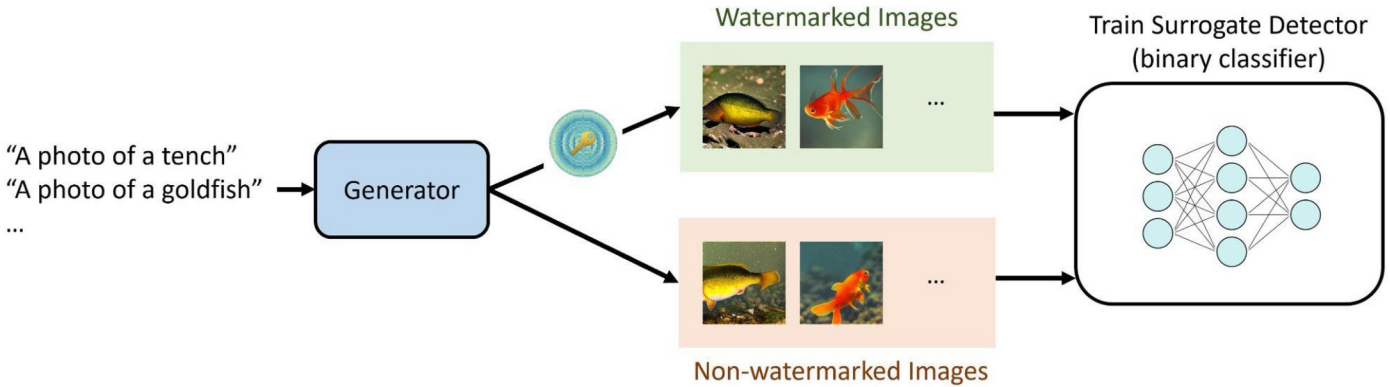


Type 2: Alter the embedding



Adversarial Attacks (Surrogate Detector)

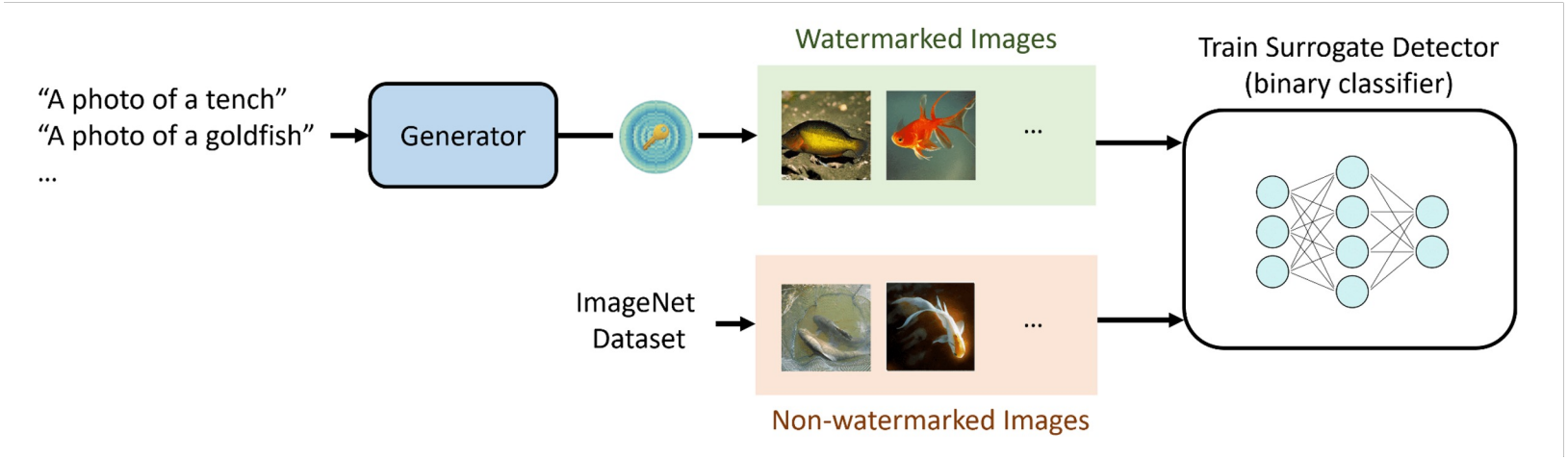
Attacker Knows
Watermarked images and Non-watermarked images from the victim model



Removes Tree Ring, but not realistic

Adversarial Attacks (Surrogate Detector)

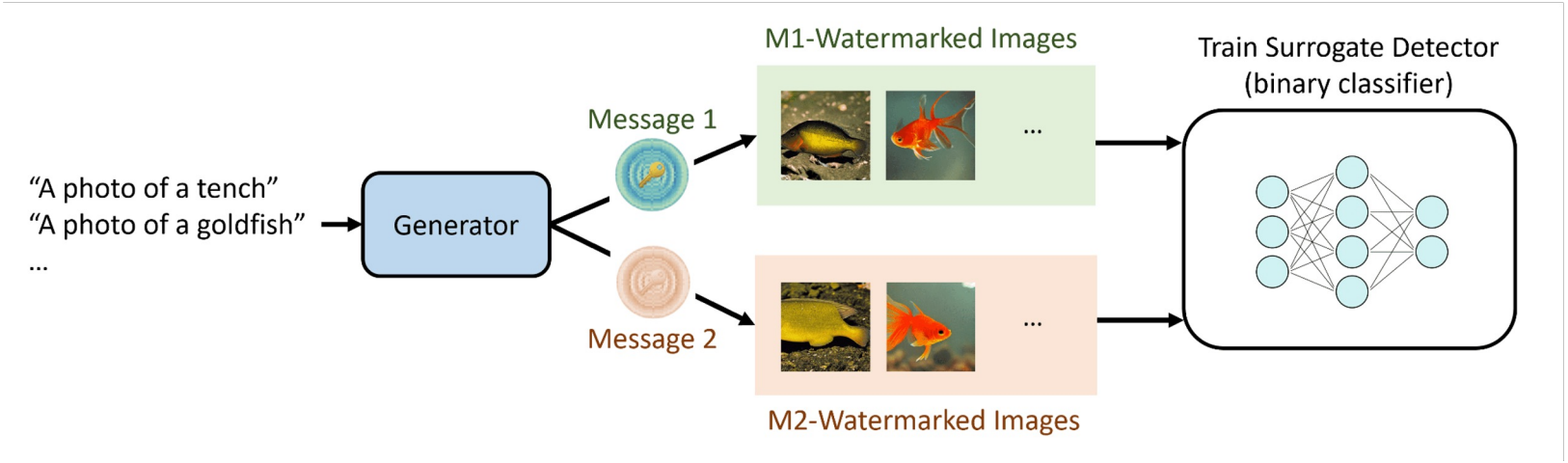
Attacker Knows
Watermarked images



Does not work - transfers badly due to the distribution shift

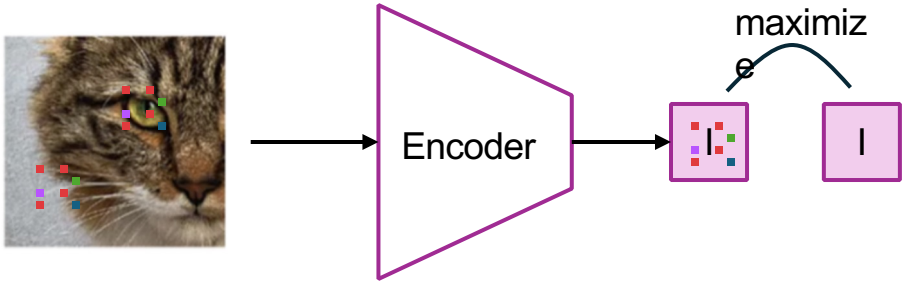
Adversarial Attacks (Surrogate Detector)

Attacker Knows
Watermarked images from 2 users



Removes Tree Ring

Adversarial Attacks (Embedding)



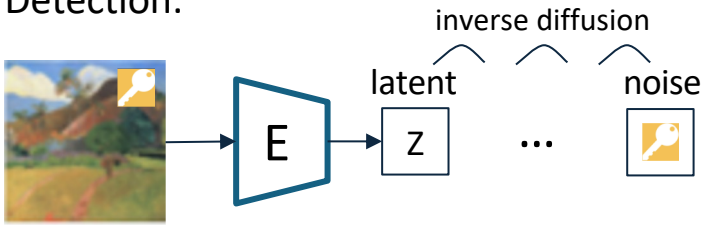
	Attacker Knows
Black-box	Nothing
Grey-box	VAE in victim model

- CLIP
- ResNet18
- VAEs
- KL-VAE (grey-box)

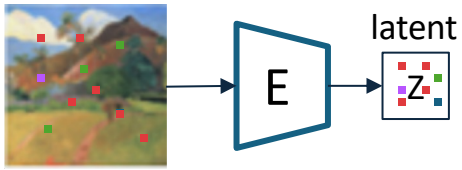
Adversarial Attacks

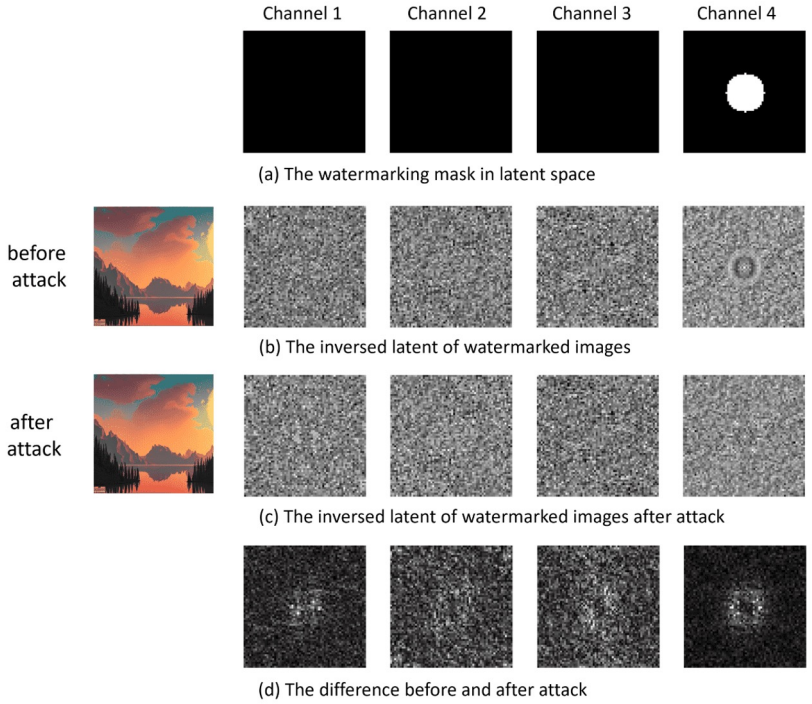
Tree Ring is vulnerable to Adversarial Attacks

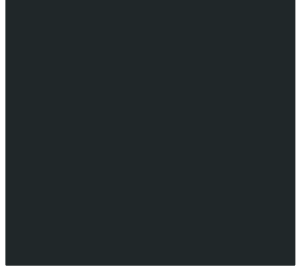
Detection:



Attack:







e.g. , an all-black image



e.g., image-to-image edit



e.g. horizontal flip



e.g., original image

Attack Potency

Score

Image Quality

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

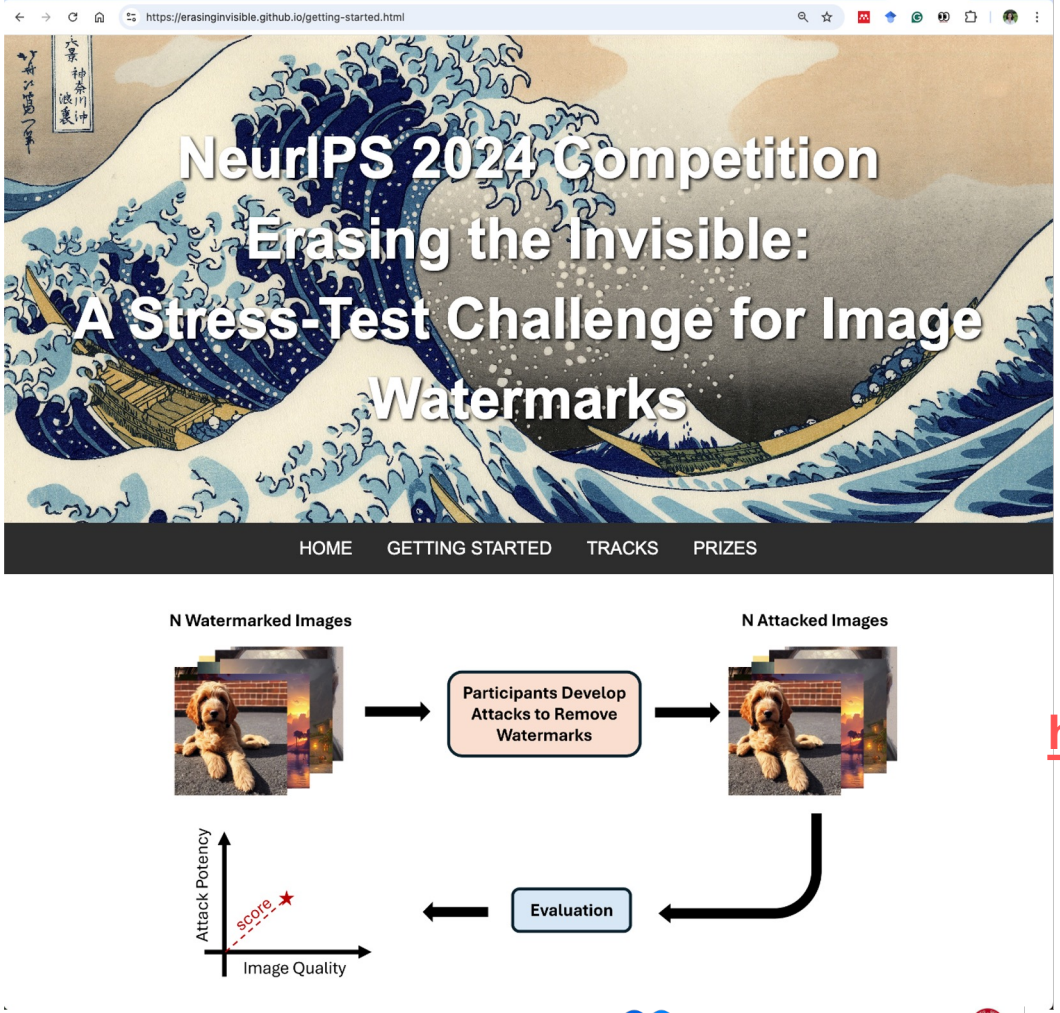
Takeaway

Different watermarking methods have different vulnerabilities.

Attack	Tree-Ring					Stable Signature					StegaStamp				
	Rank	Q@0.95P	Q@0.7P	Avg P	Avg Q	Rank	Q@0.95P	Q@0.7P	Avg P	Avg Q	Rank	Q@0.95P	Q@0.7P	Avg P	Avg Q
Dist-Rotation	11	0.464	0.521	0.375	0.648	12	0.624	0.702	0.594	0.650	5	0.423	0.498	0.357	0.616
Dist-RCrop	18	0.592	0.592	0.332	0.463	24	inf	inf	0.995	0.461	6	0.602	0.602	0.540	0.451
Dist-Erase	26	inf	inf	1.000	0.490	25	inf	inf	0.998	0.489	25	inf	inf	1.000	0.483
Dist-Bright	25	inf	inf	0.997	0.304	23	inf	inf	0.998	0.305	22	inf	inf	0.998	0.317
Dist-Contrast	22	inf	inf	0.998	0.243	20	inf	inf	0.998	0.243	17	inf	inf	0.998	0.231
Dist-Blur	20	0.861	1.112	0.563	1.221	5	-inf	-inf	0.000	1.204	9	0.848	0.962	0.414	1.198
Dist-Noise	16	0.548	inf	0.980	0.395	8	0.402	0.520	0.870	0.390	24	inf	inf	1.000	0.360
Dist-JPEG	12	0.499	0.499	0.929	0.284	9	0.485	0.485	0.793	0.284	21	inf	inf	0.998	0.263
DistCom-Geo	13	0.525	0.593	0.277	0.768	13	0.850	inf	0.937	0.767	7	0.663	0.693	0.396	0.733
DistCom-Photo	22	inf	inf	0.998	0.242	20	inf	inf	0.998	0.243	17	inf	inf	0.998	0.239
DistCom-Deg	19	0.620	inf	0.892	0.694	7	0.206	0.369	0.300	0.679	8	0.826	0.975	0.852	0.664
DistCom-All	14	0.539	0.751	0.403	0.908	11	0.538	0.691	0.334	0.900	10	0.945	1.101	0.795	0.870
Regen-Diff	5	-inf	0.307	0.612	0.323	1	-inf	-inf	0.001	0.300	1	0.331	inf	0.943	0.327
Regen-DiffP	4	-inf	0.307	0.601	0.327	1	-inf	-inf	0.001	0.303	1	0.333	inf	0.940	0.329
Regen-VAE	17	0.578	0.578	0.832	0.348	10	0.545	0.545	0.516	0.339	23	inf	inf	1.000	0.343
Regen-KLVAE	22	inf	inf	0.990	0.233	6	-inf	0.176	0.217	0.206	17	inf	inf	1.000	0.240
Rinse-2xDiff	6	-inf	0.333	0.510	0.357	3	-inf	-inf	0.001	0.332	4	0.391	inf	0.941	0.366
Rinse-4xDiff	7	-inf	0.355	0.443	0.466	4	-inf	-inf	0.000	0.438	3	0.388	inf	0.909	0.477
AdvEmbG-KLVAE8	3	-inf	0.164	0.448	0.253	20	inf	inf	0.998	0.249	17	inf	inf	1.000	0.232
AdvEmbB-RN18	10	0.241	inf	0.953	0.218	17	inf	inf	0.999	0.212	14	inf	inf	1.000	0.196
AdvEmbB-CLIP	15	0.541	inf	0.932	0.549	26	inf	inf	0.999	0.541	25	inf	inf	1.000	0.488
AdvEmbB-KLVAE16	8	0.195	inf	0.888	0.238	19	inf	inf	0.997	0.233	14	inf	inf	1.000	0.206
AdvEmbB-SdxlVAE	9	0.222	inf	0.934	0.221	17	inf	inf	0.998	0.219	14	inf	inf	1.000	0.204
AdvCls-UnWM&WM	1	-inf	0.102	0.499	0.145	14	inf	inf	0.999	0.101	11	inf	inf	1.000	0.101
AdvCls-Real&WM	21	inf	inf	1.000	0.047	14	inf	inf	0.998	0.092	11	inf	inf	1.000	0.106
AdvCls-WM1&WM2	1	-inf	0.101	0.492	0.139	14	inf	inf	0.999	0.084	13	inf	inf	1.000	0.129

More info: https://wavesbench.github.io/

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND



Sep 16 - Nov 5, 2024

https://erasinginvisible.github.io/

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

Watermark Design

We use a hybrid of *Flux. 1 [dev]* and *Stable Diffusion 2.1* images.

Each set was prompted to have diverse content and styles.



Flux. 1 [dev]



Stable Diffusion 2.1

Beige Box Track

https://www.codabench.org/competitions/3821/

- **Watermarked images and watermarking methodology provided**

Black Box Track

https://www.codabench.org/competitions/3857/

- **Watermarked images**

Breakdown of 300 Beige-Box Images

Watermarked by two representative methods, 150 Each (Ordered)

150 Images

Flux. 1 [dev]

Stega Stamp



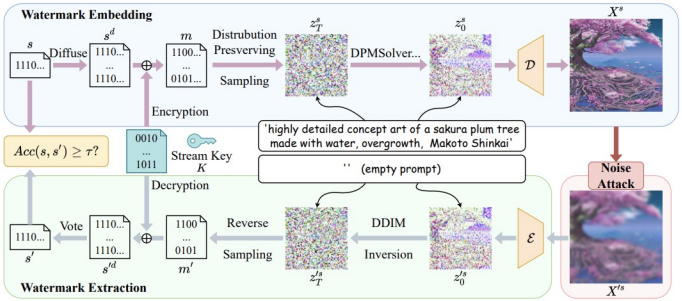
"StegaStamp: Invisible Hyperlinks in Physical Photographs"
Tancik, Mildenhall, Ng

arXiv:1904.05343

150 Images

Stable Diffusion 2.1

Gaussian Shading



"Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion Models"
Yang, Zeng et al.

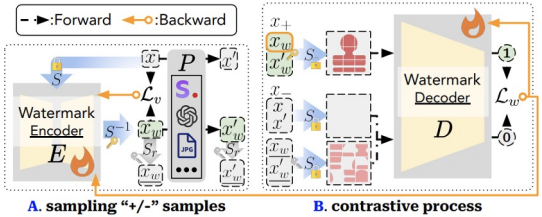
arxiv:2404.04956

Breakdown of 300 Black-Box Images

200 **Singly-Watermarked** Images, 50 Each (Shuffled)

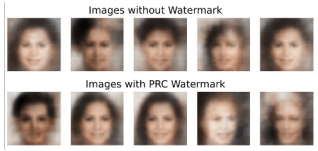
JIGMARK

"JIGMARK [...]" Pan et al., arXiv:2406.03720



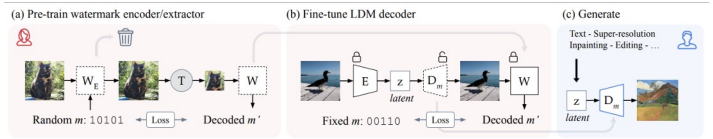
PRC

"An undetectable watermark for generative image models" Gunn, Zhao, Song. arxiv:410.07369



Stable Signature

"The Stable Signature [...]" Zhang et al., arXiv:2303.15435



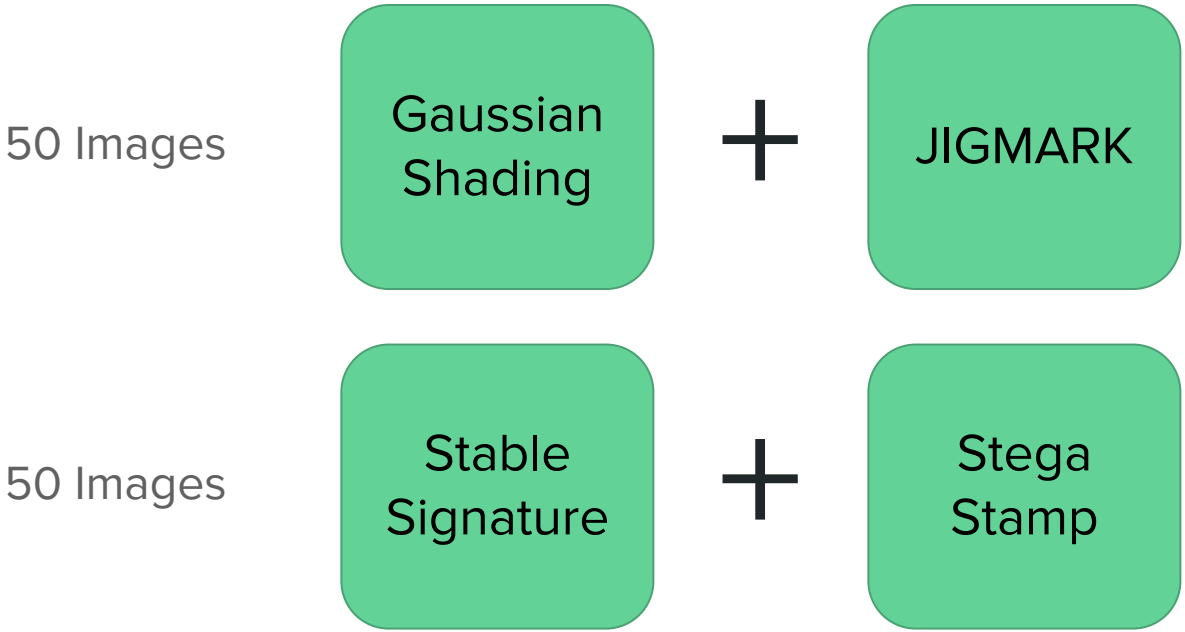
Trufo

trufo

Adobe Meta UNIVERSITÀ DI SIENA 1240 UNIVERSITY OF MARYLAND

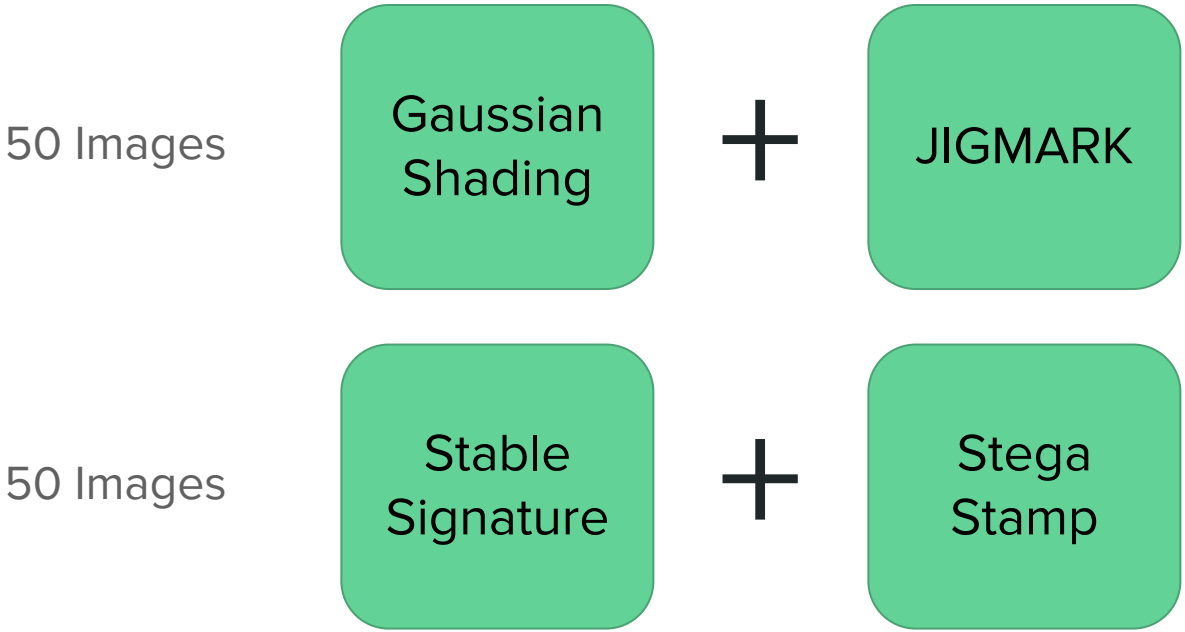
Breakdown of 300 Black-Box Images

100 **Doubly-Watermarked** Images, 50 Each, (Shuffled)



Breakdown of 300 Black-Box Images

100 **Doubly-Watermarked** Images, 50 Each, (Shuffled)



Aligning the Watermarks

To make the **“Strength”** of all watermark pipelines aligned,

We tune **hyperparameters** of each method,

so that, *False-Positive Rate (FPR) = 0.1 %*

Without editing, 99.9% of watermarked images can be detected.

Human Judgement of Image Quality

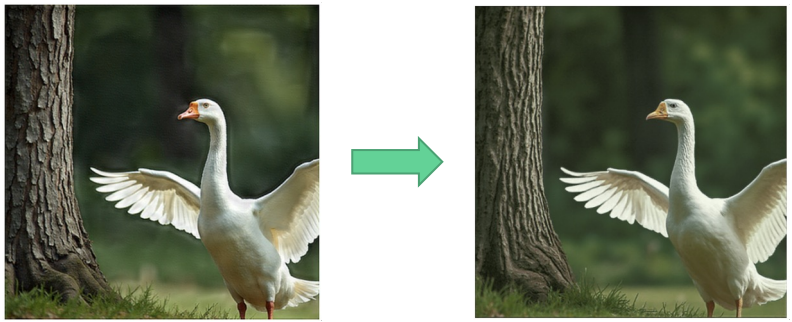
Three classes:

Minor/Moderate/Notable

Degradation

Conducted for Top-5 submissions of each track

Majority vote over 3 judges

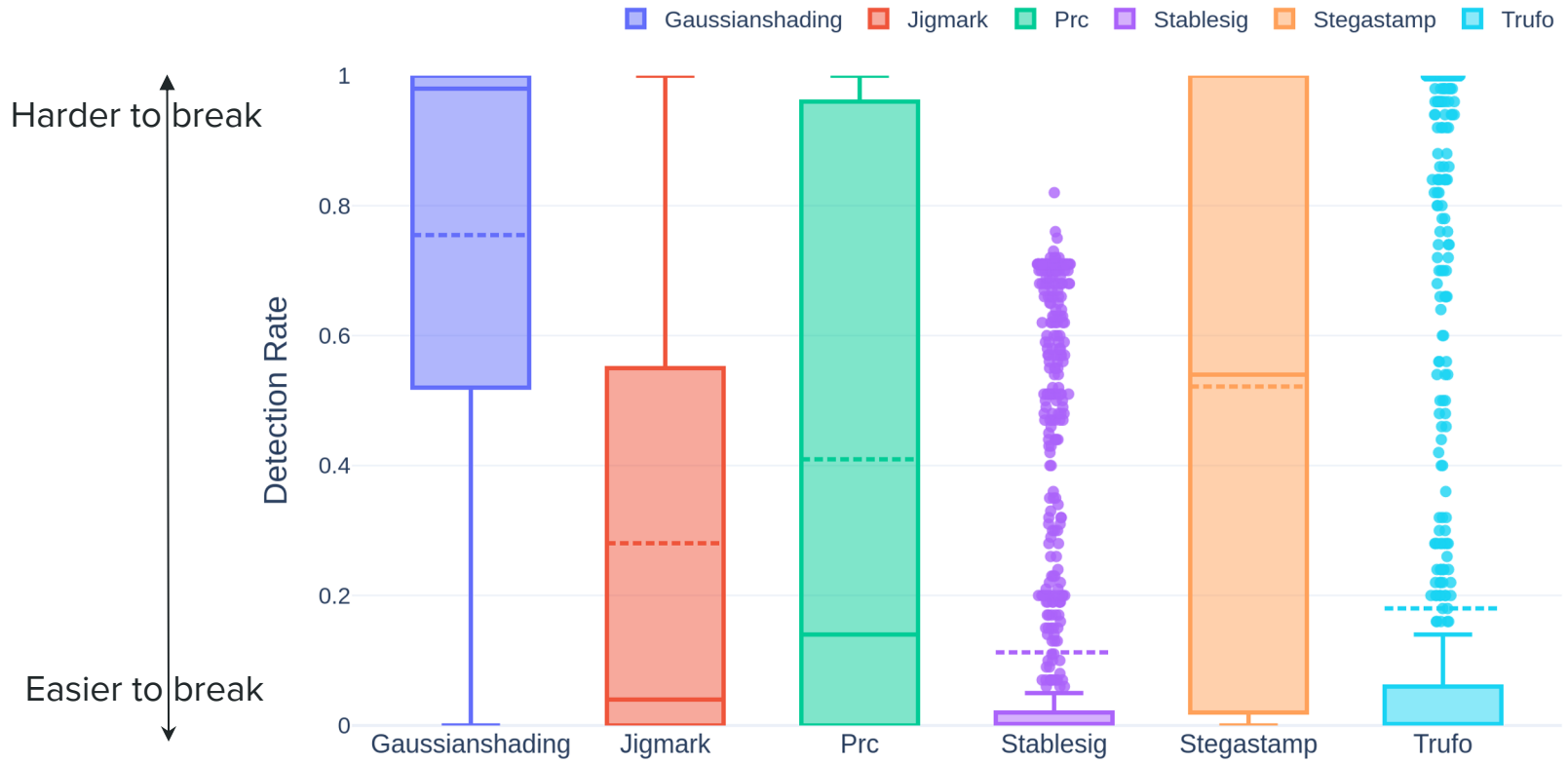


Moderate

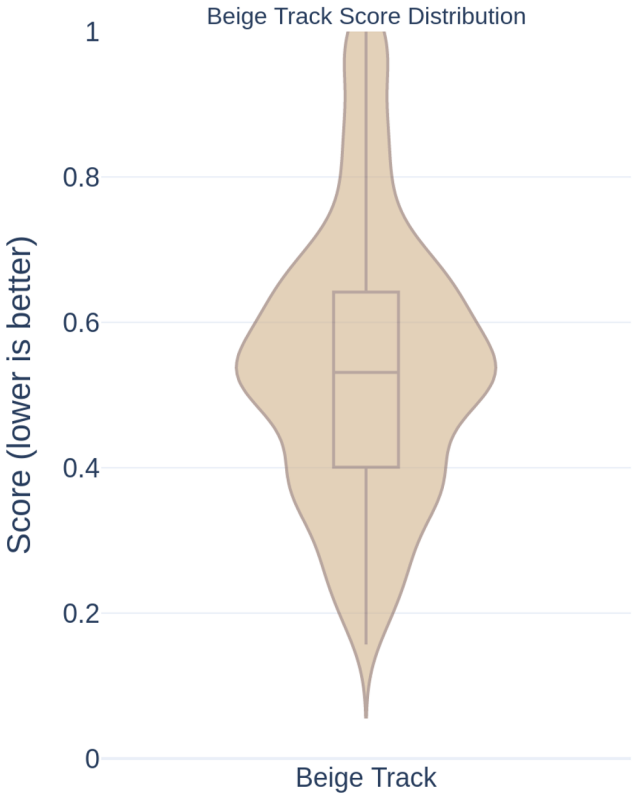


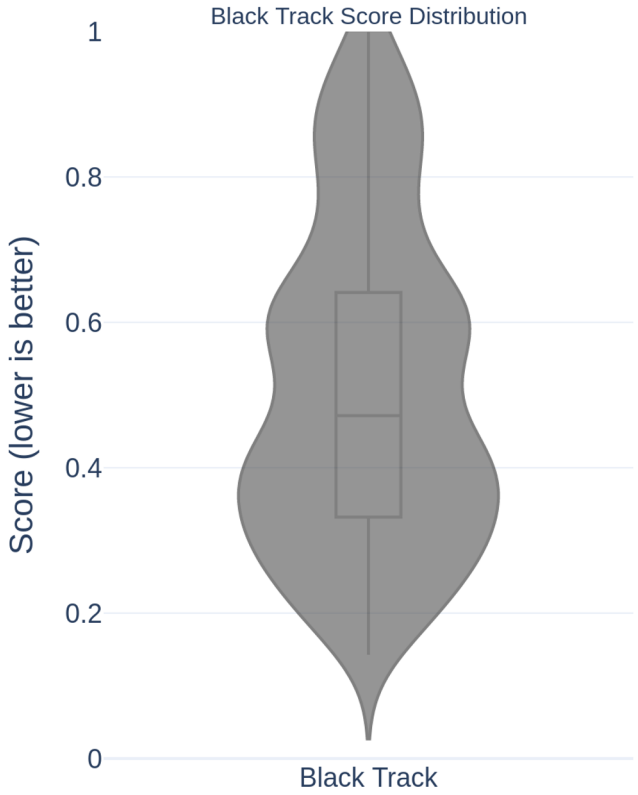
Notable

Watermark Detection Rate Distributions

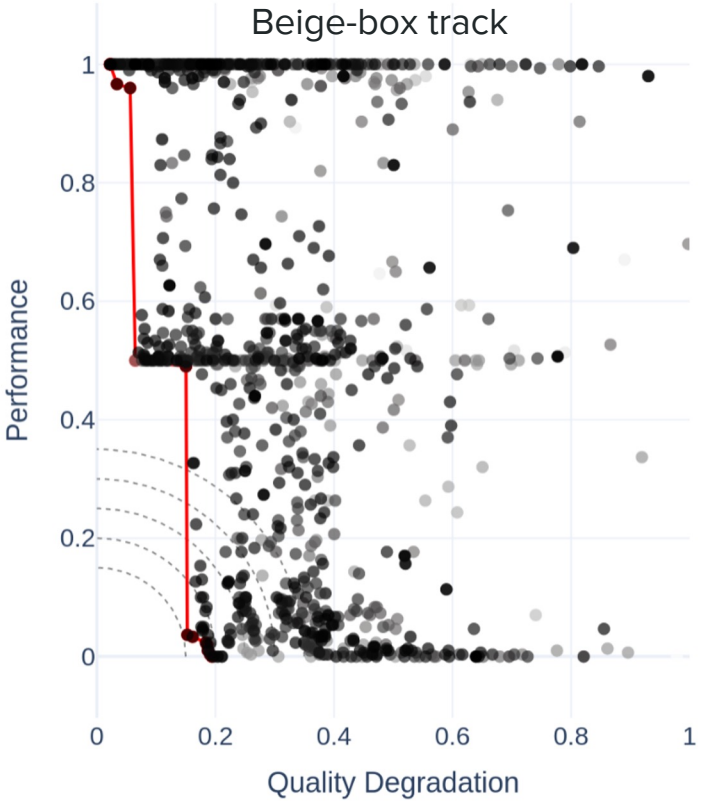


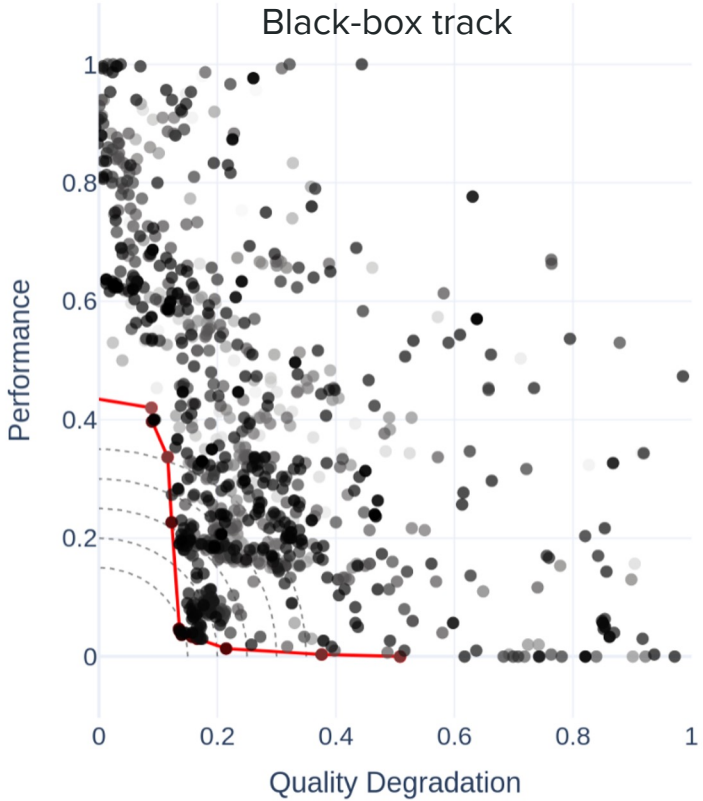
Score Distributions





Performance vs. Quality







Adobe Meta Università di Siena 1240 University of Maryland

Lessons Learned: Watermarks still far from Robust

All watermarks implemented are fragile:

The top-5 teams can remove watermarks on 81% - 96% images.

Statistics decomposed: % of watermark removed, lower the better

StegaStamp	Gaussian Shading	JIGMARK	PRC	StableSignature	Trufo	Gaussian Shading + JIGMARK	StableSignature + StegaStamp
90.2%	97.5%	98.4%	95.2%	100.0%	97.6%	61.2%	93.6%

Doubly-watermark by very different approaches may help

Winning Attacks (Beige-Box)

- Tailored strategies
- Vulnerable to simple manipulations

Team	Attacks to StegaStamp		Attacks to Gaussian Shading	
	ASR	Method	ASR	Method
1st	97.3%	Fine-tuned VAE	95.3%	Horizontal shift by 7 pixels
2nd	90.0%	Overwriting with repeated random messages	100%	Vertical shift by 8 pixels
3rd	98.6%	Multi-pass Img2Img with entropy-based strengths	76.0%	Cropping and rescaling
4th	100%	Message extraction, inversion and re-embedding	80.0%	Rinsing regeneration with a pretrained diffusion model
5th	99.3%	Resize-crop distortion with dynamically adjusted parameters	89.3%	Resize-crop distortion with dynamically adjusted parameters

Winning Attacks (Black-Box)

- Clustering for artifacts
- Controlled regeneration
- Hybrid attacks

Team	Attacks
1st	Cluster by visual artifacts + Fine-tuned VAE + 7/8 Pixel Shift. Learnable color transfer to preserve quality.
2nd	Controlled regeneration. Extensive post-processing: external tool for color adjustments, re-sizing, rotation+shift, and FFT.
3rd	Controlled regeneration. Semantic guidance – a generated description of the image is set as the diffusion prompt to maintain content quality.
4th	Cluster images for Stegastamp artifacts. Stegastamp images are attacked with stable diffusion + re-sizing. All others are attacked with a 7-8 pixel shift + regeneration via Stable Diffusion.
5th	Perturbing latent embeddings of models via FLUX model regeneration.

Messages

- Tailored strategies were a winning factor—top teams tailored their removal methods to specific watermark types. Even for black-box track, by clustering images, teams are able to identify different watermarks and apply different attacking methods. Customization and adaptability seem to be central to winning teams' success.
- Teams discovered that Tree-ring-like watermarks could be defeated by minor pixel shifts, a new insight even for their creators.
- The competition revealed that many current watermarking methods are not robust, with top teams removing watermarks from ~89% of images.
- However, composite methods (e.g., Gaussian Shading + JIGMARK) proved more durable, suggesting hybrid approaches may enhance robustness.