Do Modern Post-Hoc Watermarking Beat Broken-Arrows?

Enoal GESNY Eva GIBOULOT

Univ. Rennes, Inria

IH&MMSEC 2026

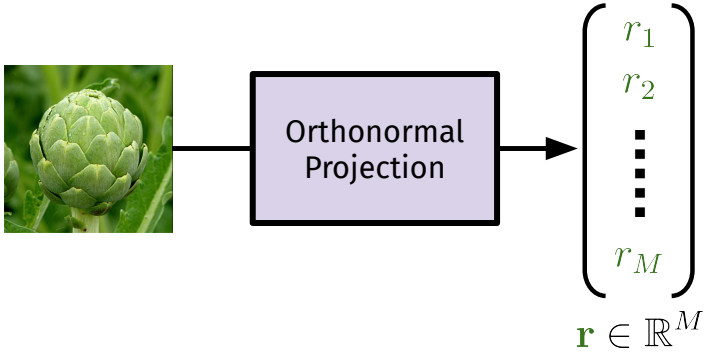
Watermarking History

Embedding

Decoding

Watermarking History

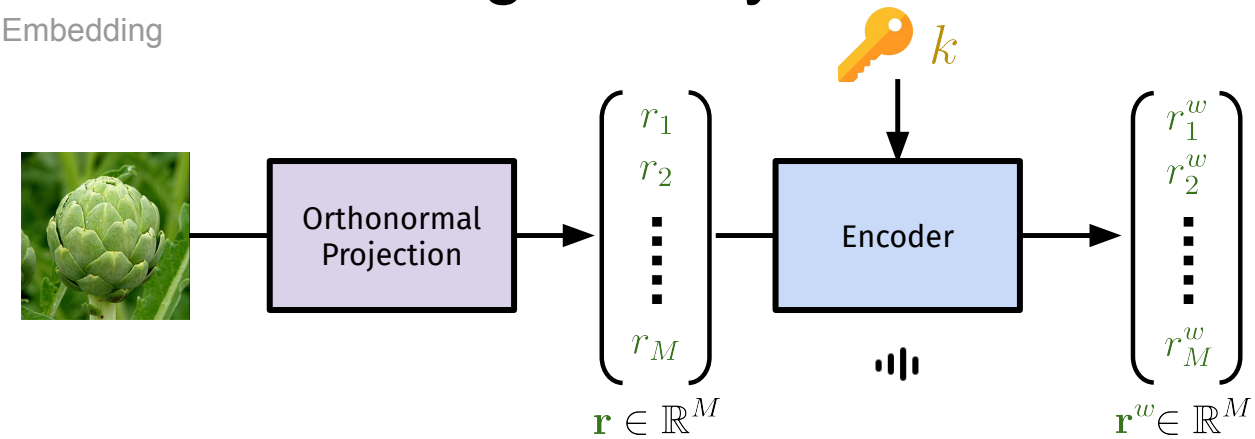
Embedding



Decoding

Watermarking History

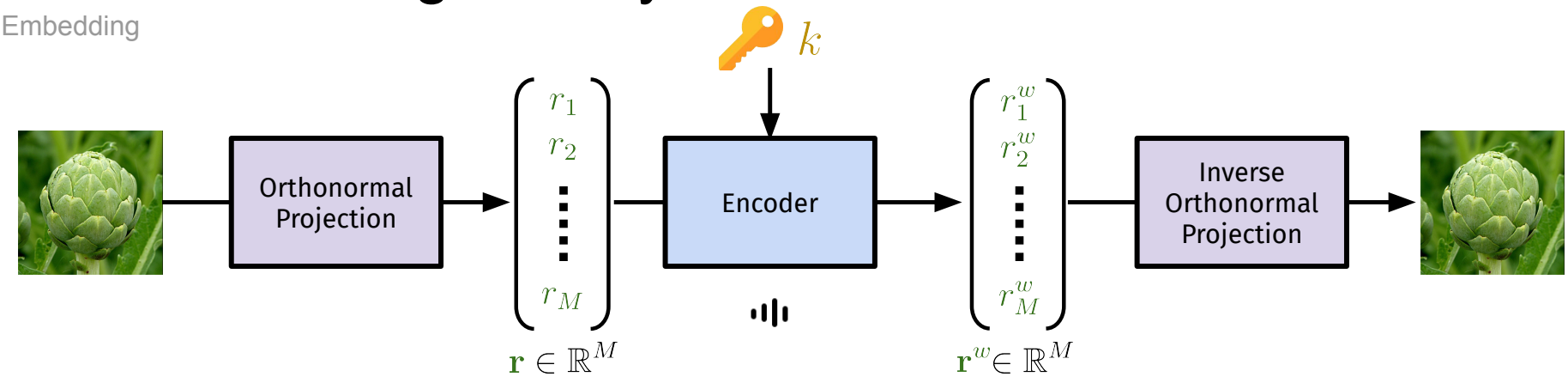
Embedding



Decoding

Watermarking History

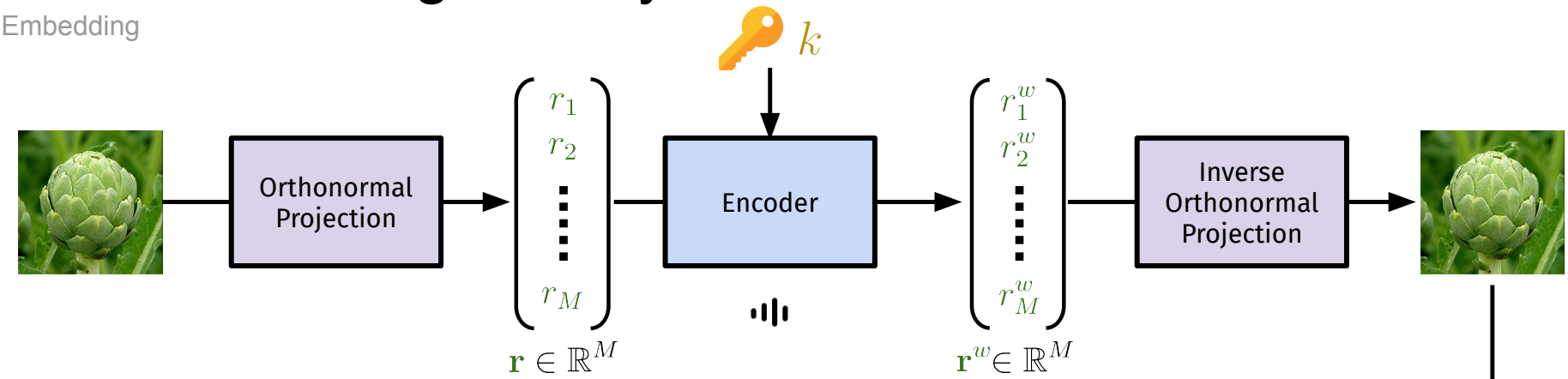
Embedding



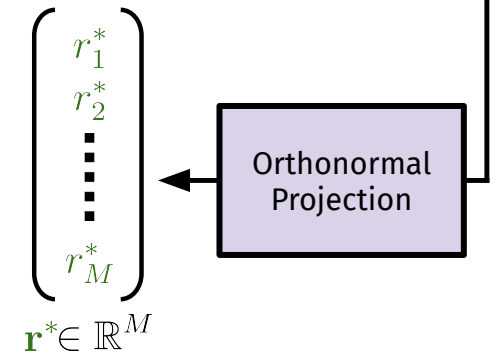
Decoding

Watermarking History

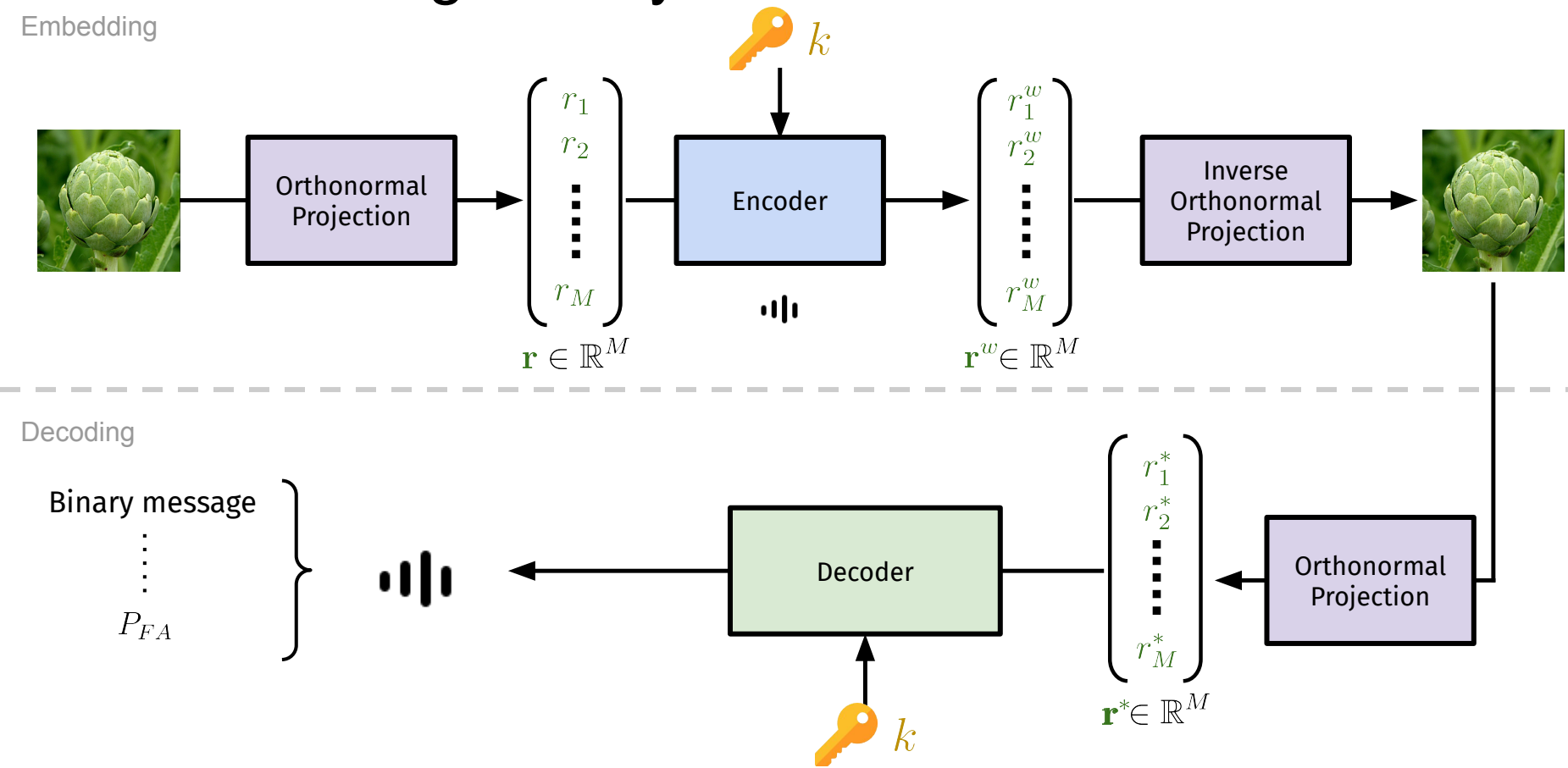
Embedding



Decoding



Watermarking History



Watermarking History

Classic Watermarking

- Robust to valuemetric transformations
- Security: How many samples to retrieve the secret key?

Watermarking History

Classic Watermarking

- Robust to valuemetric transformations
- Security: How many samples to retrieve the secret key?

Main limit

- Geometric transformations problem

Watermarking History

Classic Watermarking

- Robust to valuemetric transformations
- Security: How many samples to retrieve the secret key?

Main limit

- Geometric transformations problem

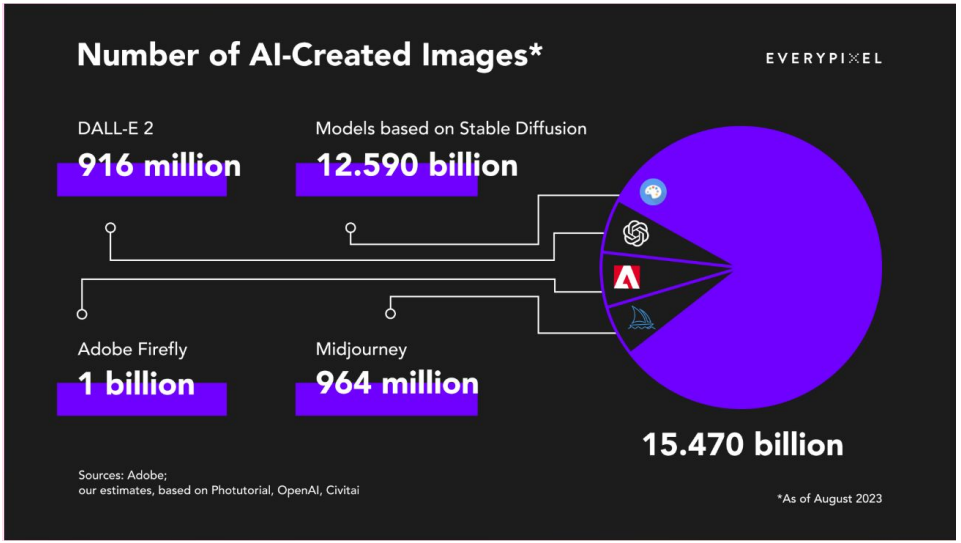
Watermarking went silent

Watermarking Revival (~2023)

Watermarking Revival (~2023)

Generative AI

- Deepfakes
- EU AI Act [1]



[1] *European AI Act. https://artificialintelligenceact.eu*, Europe 2023

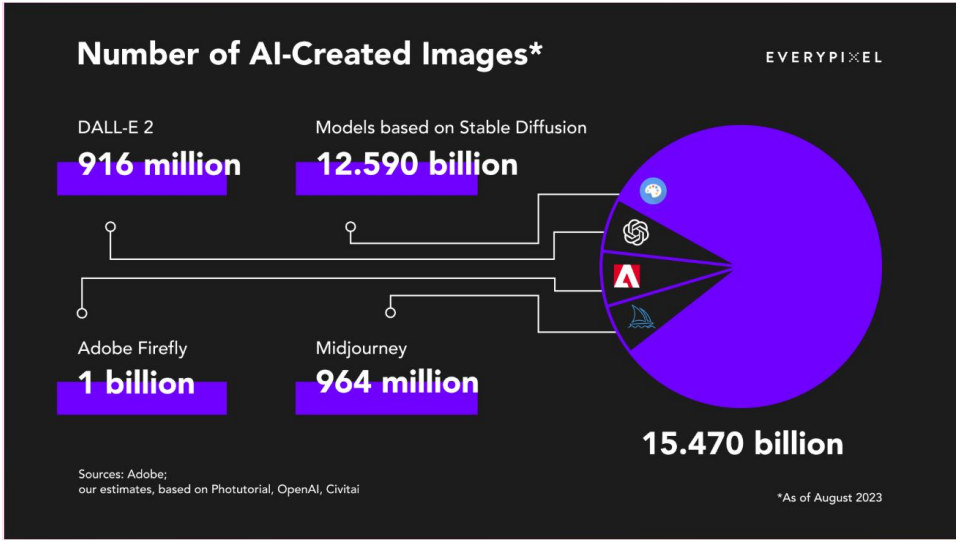
Watermarking Revival (~2023)

Generative AI

- Deepfakes
- EU AI Act [1]

Modern watermarking

- In-Gen [2, 3, 4]
- Post-hoc (DNN) [5, 6]



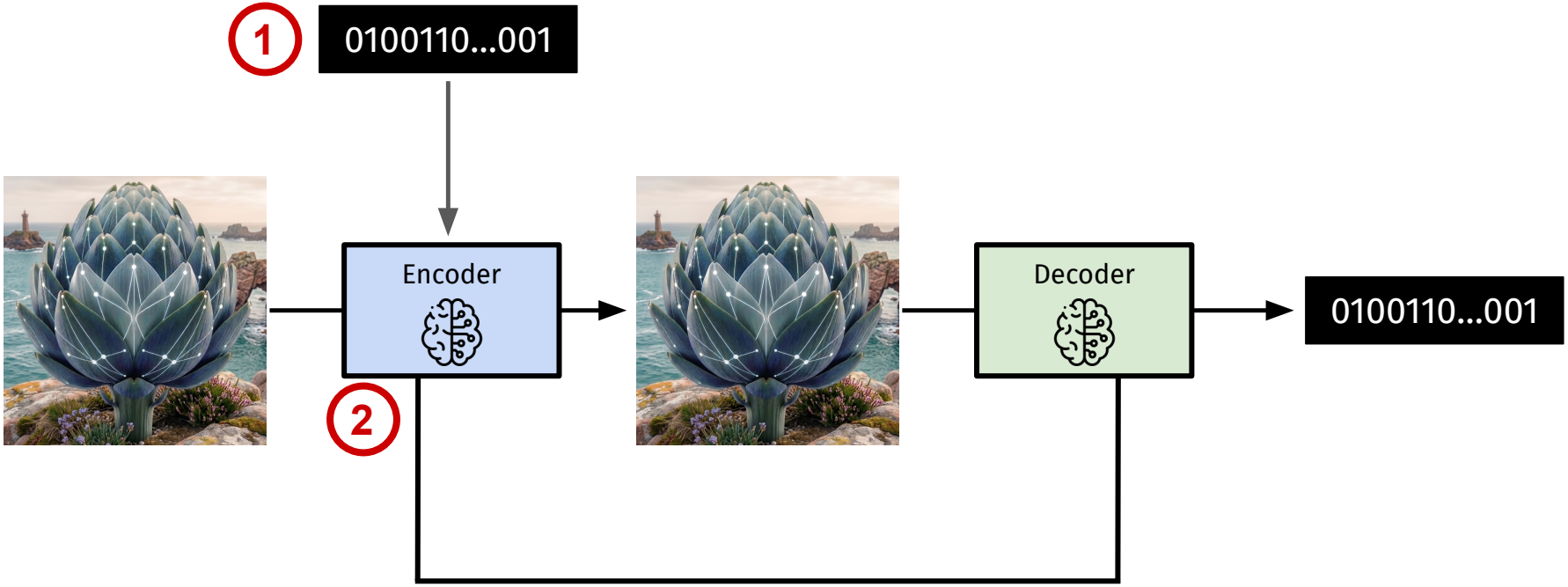
[1] *European AI Act. https://artificialintelligenceact.eu*, Europe 2023
[2] *Gaussian shading: Provable performance-lossless image watermarking for diffusion models*, Z. Yang et al. 2024 (CVPR)
[3] *The stable signature: Rooting watermarks in latent diffusion models*, P. Fernandez et al. 2023 (ICCV)
[4] *Guidance Watermarking for Diffusion Models*, E. Gesny et al. 2026 (ICLR)
[5] *Video Seal: Open and Efficient Video Watermarking*, P. Fernandez et al. 2024 (ArXiv)
[6] *TrustMark: Robust Watermarking and Watermark Removal for Arbitrary Resolution Images*, T. Bui et al. 2025 (ICCV)

Modern Post-hoc Watermarking

1 0100110...001

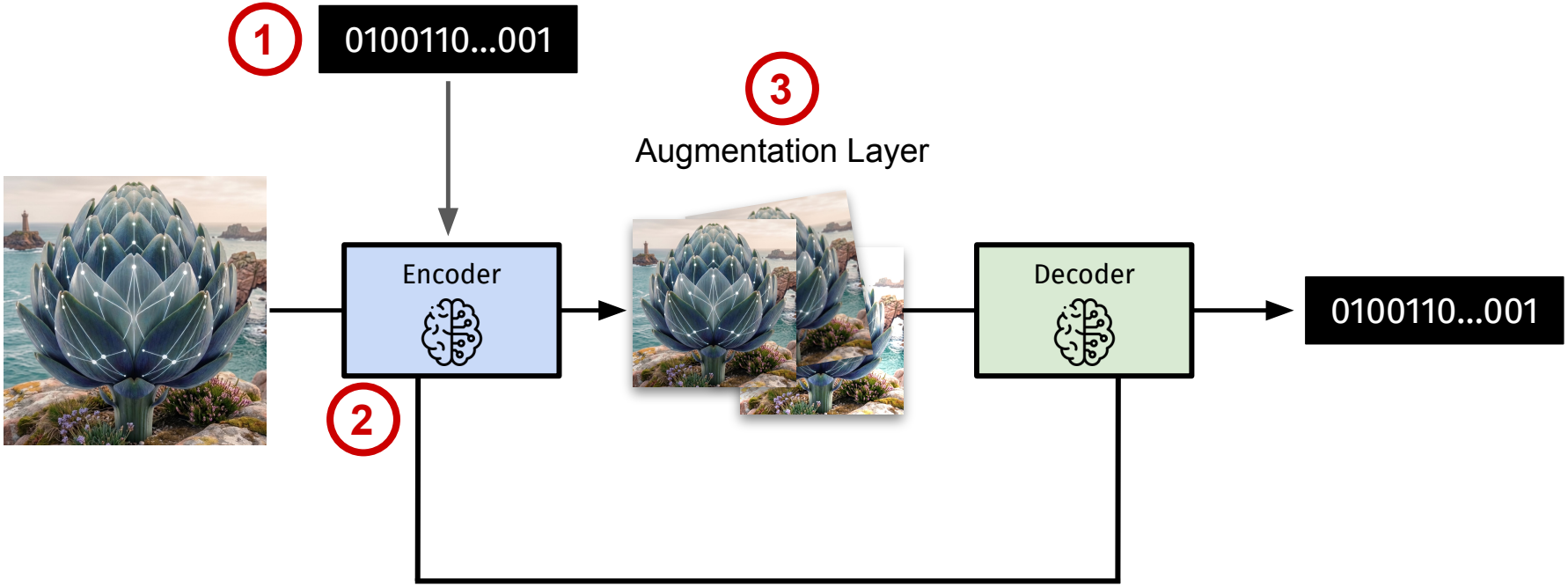
[1] *HiDDeN: Hiding Data With Deep Networks. J. Zhu et al.*, 2018 (ECCV)

Modern Post-hoc Watermarking



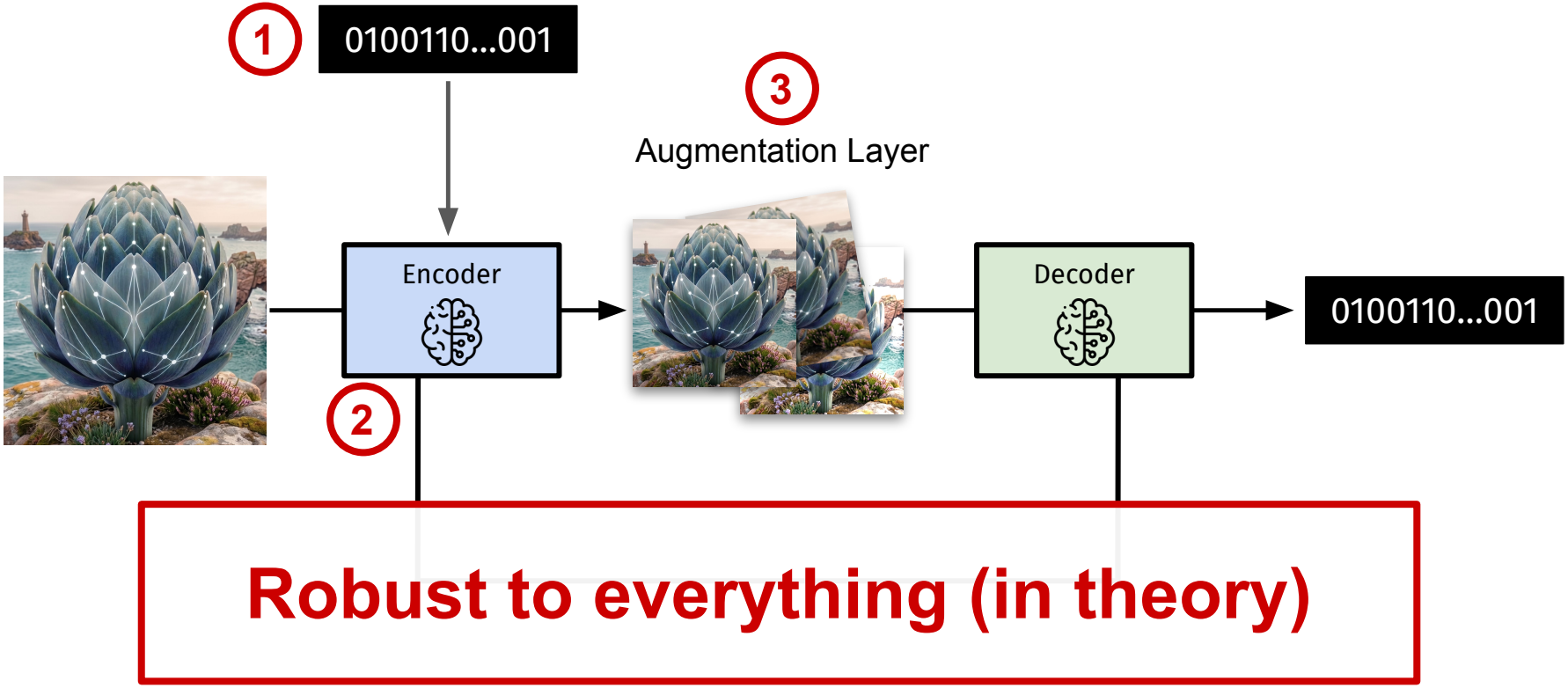
[1] *HiDDeN: Hiding Data With Deep Networks. J. Zhu et al.*, 2018 (ECCV)

Modern Post-hoc Watermarking



[1] *HiDDeN: Hiding Data With Deep Networks. J. Zhu et al.*, 2018 (ECCV)

Modern Post-hoc Watermarking



[1] *HiDDeN: Hiding Data With Deep Networks. J. Zhu et al.*, 2018 (ECCV)

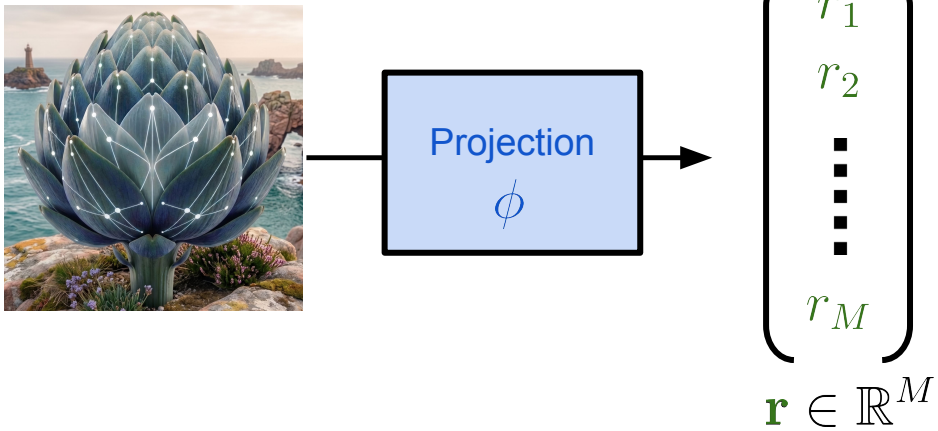
Does modern post-hoc watermarking consistently outperform classical approaches in a zero-bit setting?

0-bit Watermarking

Goal: Score $\leq \mathrm{P}_{\mathrm{FA}}$

0-bit Watermarking

Goal: Score $\leq \mathrm{P}_{\mathrm{FA}}$



0-bit Watermarking

Goal: $p_\phi(\mathbf{x}, k) \leq \alpha$

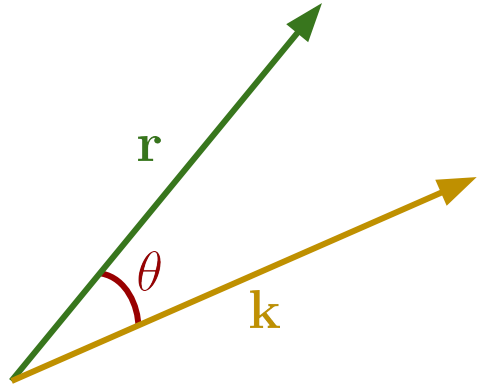
Projection ϕ

$$\begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_M \end{pmatrix}$$

$\mathbf{r} \in \mathbb{R}^M$

Key k

Hypercone detector



$$\mathrm{P_{FA}} = 1 - I_{\cos^2\theta}\left(\frac{1}{2}, \frac{M-1}{2}\right)$$

[1] *Broken-Arrows*, T. Furon et al. 2008 (EURASIP Journal on Information Security)

Settings

Dataset

- MFlickr natural images [1]

Modern Watermarking

- VideoSeal (256 dim) [2]
- TrustMark (100 dim) [3]

Classic Watermarking

- Broken-Arrows (128 dim) [4]

[1] *The MIR flickr retrieval evaluation*, M. Huiskes et al. 2008 (ACM International Conference on Multimedia Information Retrieval)
[2] *Video Seal: Open and Efficient Video Watermarking*, P. Fernandez et al. 2024 (ArXiv)
[3] *TrustMark: Robust Watermarking and Watermark Removal for Arbitrary Resolution Images*, T. Bui et al. 2025 (ICCV)
[4] *Broken-Arrows*, T. Furon et al. 2008 (EURASIP Journal on Information Security)

Methodology

Robustness

- Set a False Alarm Probability level $\alpha = 10^{-6}$[1]

[1] *Broken-Arrows*, T. Furon et al. 2008 (EURASIP Journal on Information Security)

Methodology

Robustness

- Set a False Alarm Probability level $\alpha = 10^{-6}$[1]

Imperceptibility

- Set a watermark power $\mathrm{PSNR} = 42\mathrm{dB}$

[1] *Broken-Arrows*, T. Furon et al. 2008 (EURASIP Journal on Information Security)

Methodology

Robustness

- Set a False Alarm Probability level $\alpha = 10^{-6}$[1]

Imperceptibility

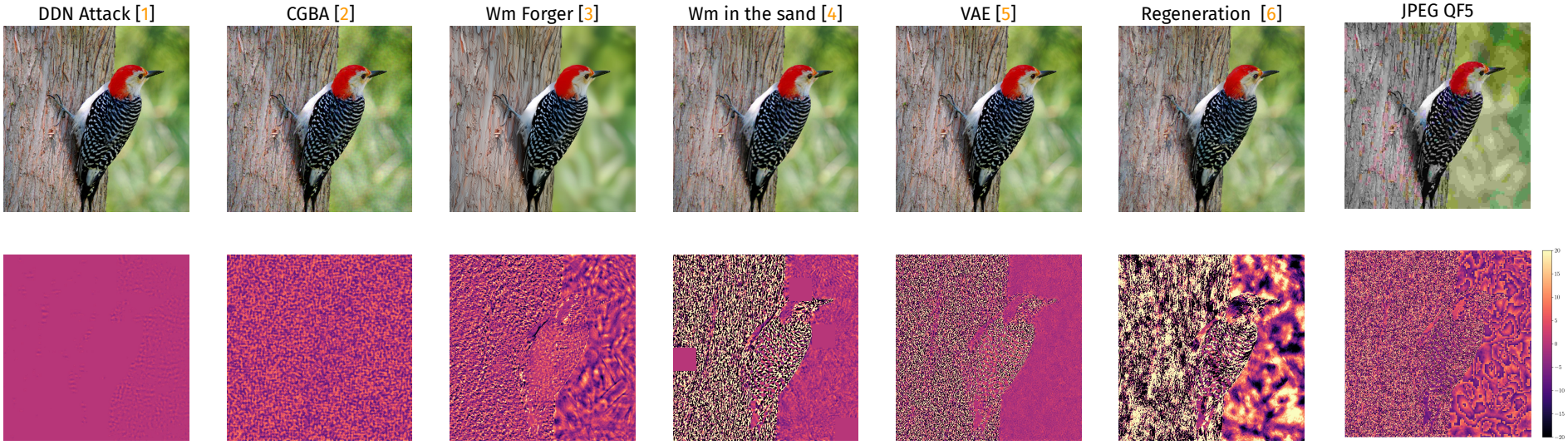
- Set a watermark power $\mathrm{PSNR} = 42\mathrm{dB}$

Security

- Erasure Attack $\begin{cases} \min_{\boldsymbol{\epsilon}} \|\boldsymbol{\epsilon}\|_2^2 \\ \text{s.t.} \quad p(\mathbf{x}+\boldsymbol{\epsilon}) > \alpha \end{cases}$

[1] *Broken-Arrows*, T. Furon et al. 2008 (EURASIP Journal on Information Security)

Attack choices



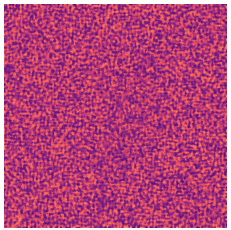
[1] *Decoupling Direction and Norm for Efficient Gradient-Based L2 Adversarial Attacks and Defenses*, J. Rony et al. 2019 (CVPR)
[2] *CGBA: Curvature-aware Geometric Black-box Attack*, M. Reza et al. 2023 (ICCV)
[3] *Transferable Black-Box One-Shot Forging of Watermarks via Image Preference Models*, T. Soucek et al. 2025 (NeurIPS)
[4] *Watermarks in the sand: impossibility of strong watermarking for language model*, H. Zhang et al. 2024 (ICML)
[5] *Sana: Efficient High-Resolution Image Synthesis with Linear Diffusion Transformer*, Z. Zhang et al. 2024 (ArXiv)
[6] *Diffusion Models for Adversarial Purification*, W. Nie et al. 2022 (ICML)

Attack choices

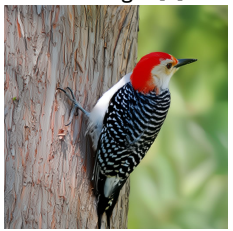
White-Box

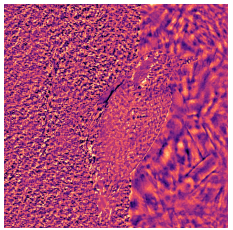
DDN Attack [1]

CGBA [2]

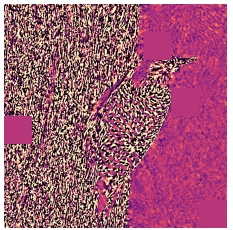


Wm Forger [3]

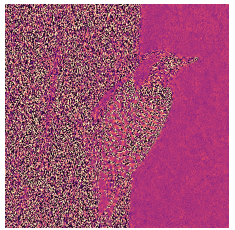




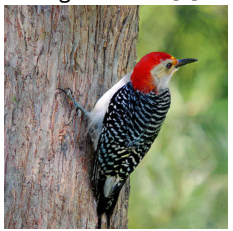
Wm in the sand [4]

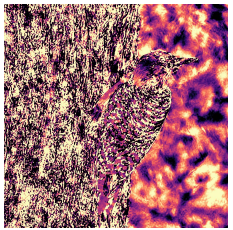


VAE [5]

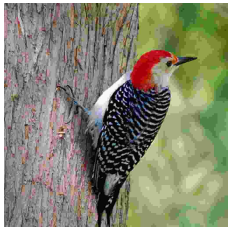


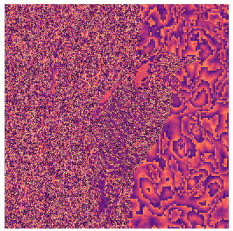
Regeneration [6]





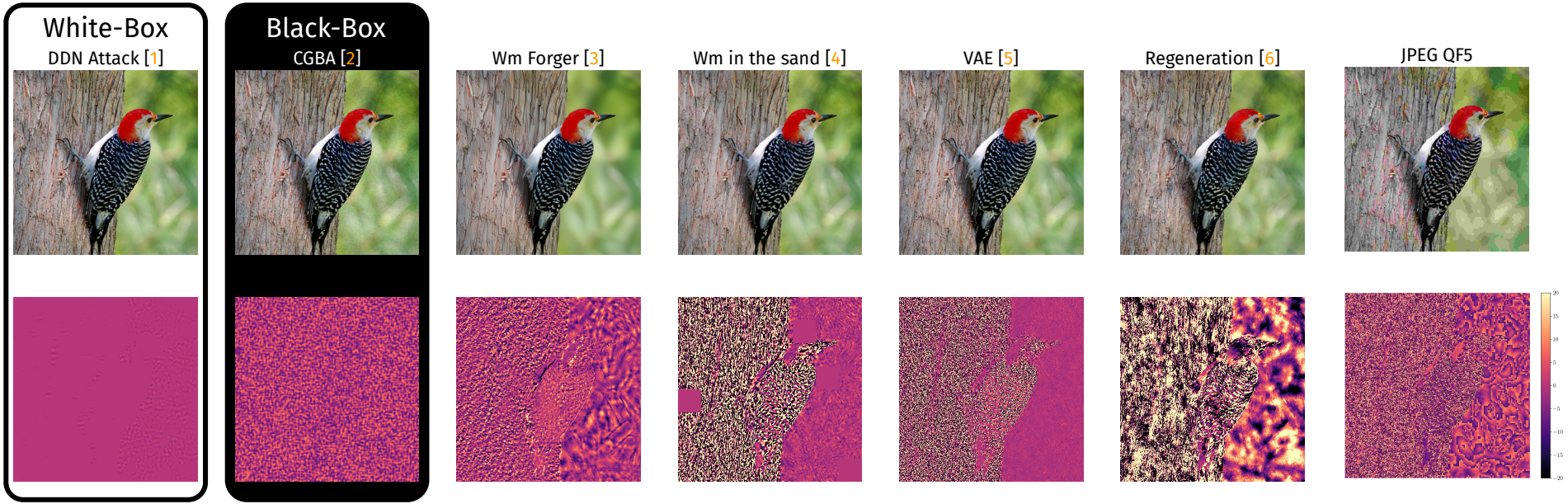
JPEG QF5





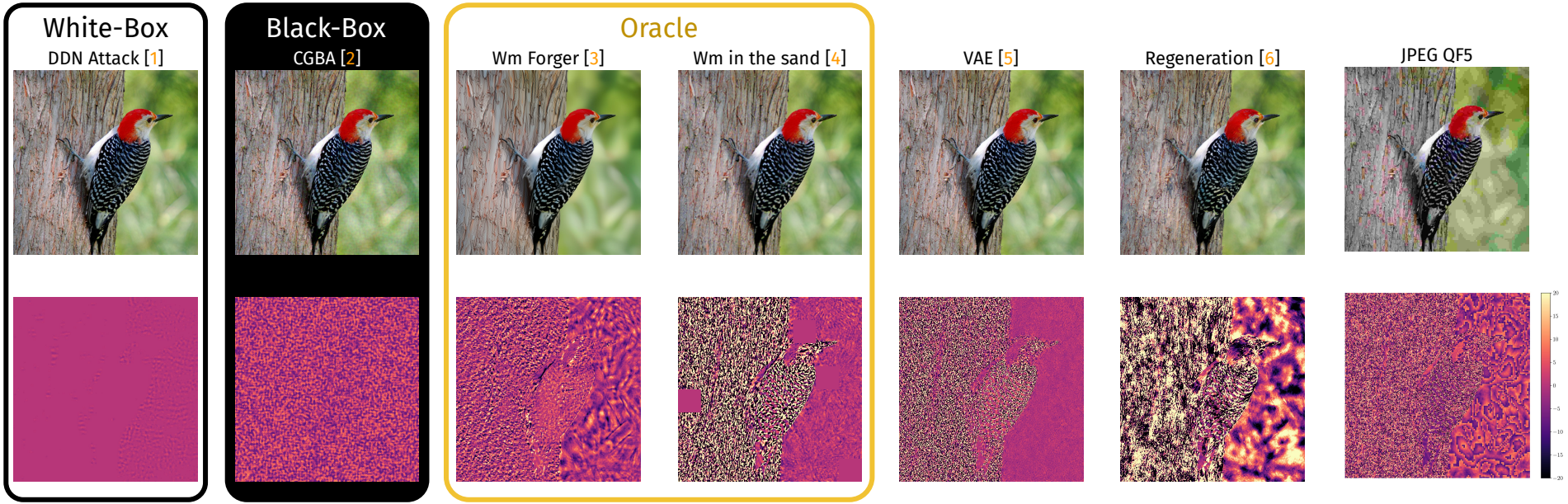
[1] *Decoupling Direction and Norm for Efficient Gradient-Based L2 Adversarial Attacks and Defenses*, J. Rony et al. 2019 (CVPR)
[2] *CGBA: Curvature-aware Geometric Black-box Attack*, M. Reza et al. 2023 (ICCV)
[3] *Transferable Black-Box One-Shot Forging of Watermarks via Image Preference Models*, T. Soucek et al. 2025 (NeurIPS)
[4] *Watermarks in the sand: impossibility of strong watermarking for language model*, H. Zhang et al. 2024 (ICML)
[5] *Sana: Efficient High-Resolution Image Synthesis with Linear Diffusion Transformer*, Z. Zhang et al. 2024 (ArXiv)
[6] *Diffusion Models for Adversarial Purification*, W. Nie et al. 2022 (ICML)

Attack choices



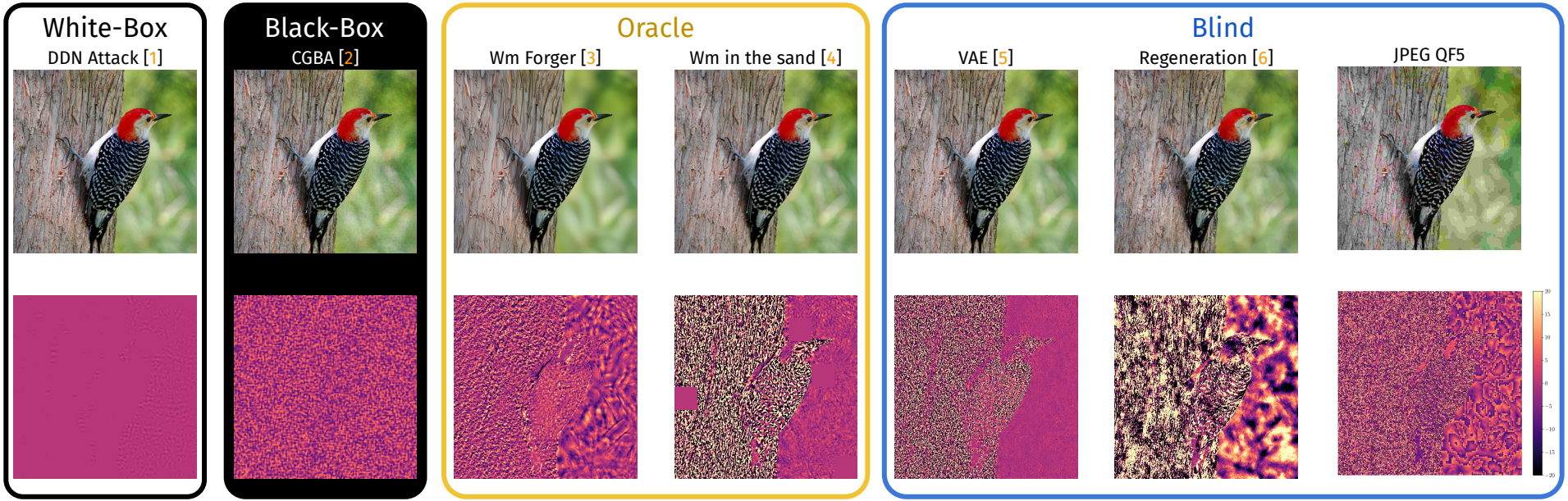
[1] *Decoupling Direction and Norm for Efficient Gradient-Based L2 Adversarial Attacks and Defenses*, J. Rony et al. 2019 (CVPR)
[2] *CGBA: Curvature-aware Geometric Black-box Attack*, M. Reza et al. 2023 (ICCV)
[3] *Transferable Black-Box One-Shot Forging of Watermarks via Image Preference Models*, T. Soucek et al. 2025 (NeurIPS)
[4] *Watermarks in the sand: impossibility of strong watermarking for language model*, H. Zhang et al. 2024 (ICML)
[5] *Sana: Efficient High-Resolution Image Synthesis with Linear Diffusion Transformer*, Z. Zhang et al. 2024 (ArXiv)
[6] *Diffusion Models for Adversarial Purification*, W. Nie et al. 2022 (ICML)

Attack choices



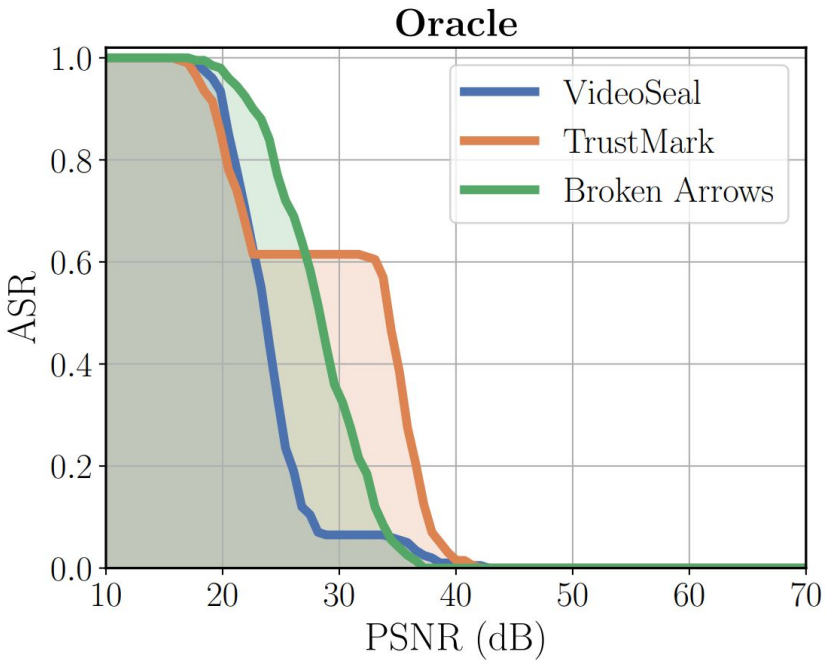
[1] *Decoupling Direction and Norm for Efficient Gradient-Based L2 Adversarial Attacks and Defenses*, J. Rony et al. 2019 (CVPR)
[2] *CGBA: Curvature-aware Geometric Black-box Attack*, M. Reza et al. 2023 (ICCV)
[3] *Transferable Black-Box One-Shot Forging of Watermarks via Image Preference Models*, T. Soucek et al. 2025 (NeurIPS)
[4] *Watermarks in the sand: impossibility of strong watermarking for language model*, H. Zhang et al. 2024 (ICML)
[5] *Sana: Efficient High-Resolution Image Synthesis with Linear Diffusion Transformer*, Z. Zhang et al. 2024 (ArXiv)
[6] *Diffusion Models for Adversarial Purification*, W. Nie et al. 2022 (ICML)

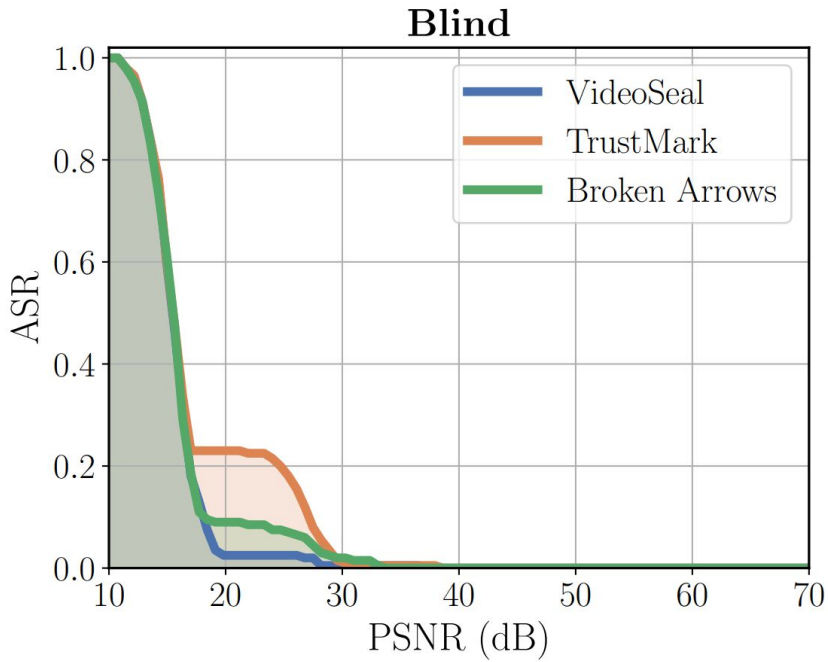
Attack choices



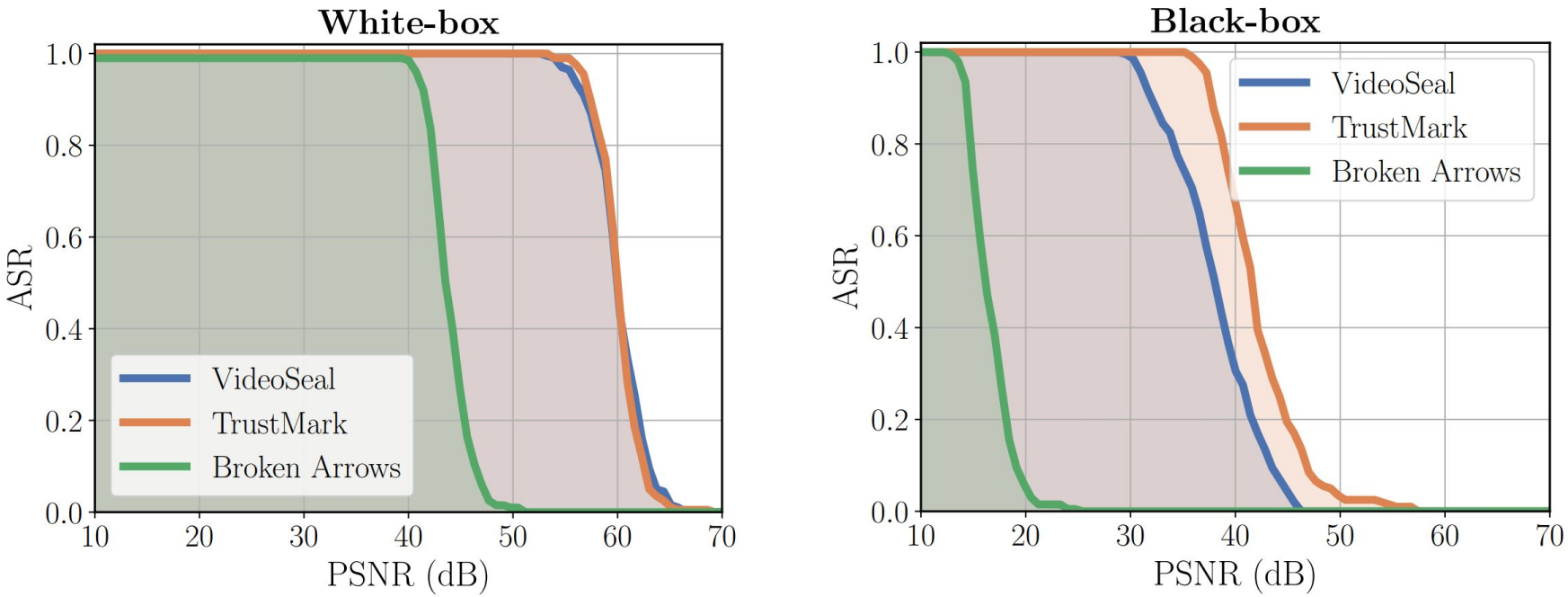
[1] *Decoupling Direction and Norm for Efficient Gradient-Based L2 Adversarial Attacks and Defenses*, J. Rony et al. 2019 (CVPR)
[2] *CGBA: Curvature-aware Geometric Black-box Attack*, M. Reza et al. 2023 (ICCV)
[3] *Transferable Black-Box One-Shot Forging of Watermarks via Image Preference Models*, T. Soucek et al. 2025 (NeurIPS)
[4] *Watermarks in the sand: impossibility of strong watermarking for language model*, H. Zhang et al. 2024 (ICML)
[5] *Sana: Efficient High-Resolution Image Synthesis with Linear Diffusion Transformer*, Z. Zhang et al. 2024 (ArXiv)
[6] *Diffusion Models for Adversarial Purification*, W. Nie et al. 2022 (ICML)

Results – No Access to the Detector





Results – Access to the Detector



Takeaways and Limitations

(+) Trade-off security robustness

(+) DNN-methods: **no significant gain in valuemetric robustness**

(+) DNN-methods: **significant decline in security**

(-) Not optimized for 0-bit scenario

(-) Does not take into account geometric transformations

Main Takeaway

Modern watermarking

(+) Robust to geometric transformation **(but not for free)**

(-) Security

Why?

- No secret key
- Neural-network vulnerability
- Lipschitz network

Results – Comparison of Attacks