北京大学
PEKING UNIVERSITY

Forging and Removing Latent-Noise Diffusion Watermarks Using a Single Image

Anubhav Jain [1], Yuya Kobayash [2], Naoki Murata [2], Yuhta Takida [2], Takashi Shibuya [2], Yuki Mitsufuji [2,3], Niv Cohen [1,*], Nasir Memon [1,*], Julian Togelius [1,*]

[1]New York University, [2]Sony AI, [3]Sony Group Corporation

arXiv preprint
arXiv:2504.20111

Presenter: Yufei Zhang

2026.03.22

Introduction

北京大学 PEKING UNIVERSITY

Visible Watermark

Invisible Watermark

Clean

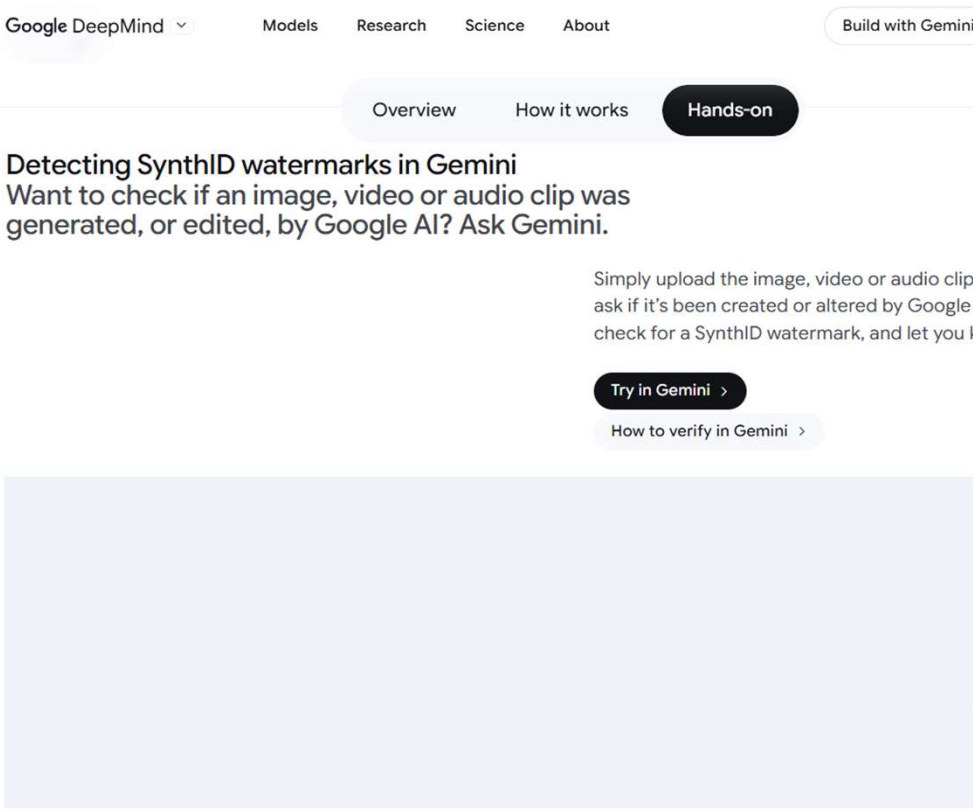
Watermarked

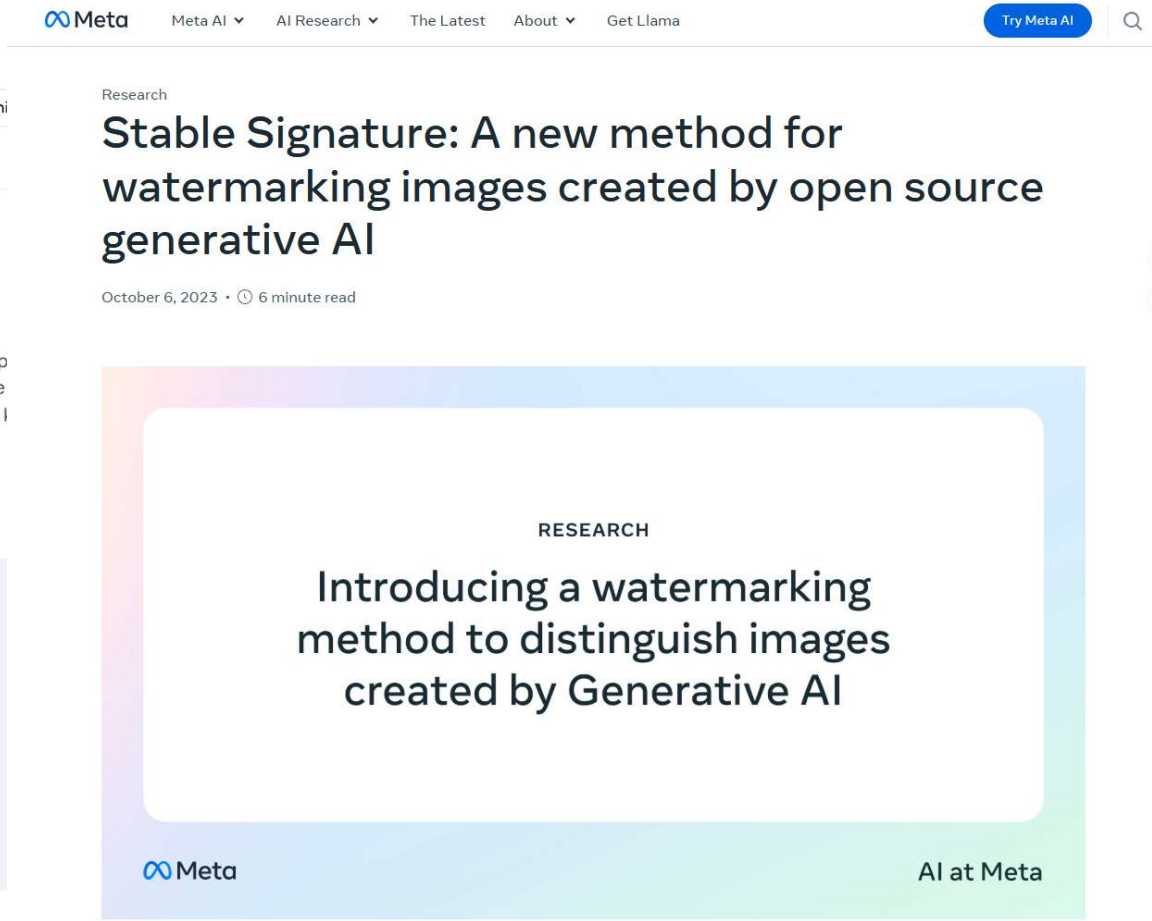
- Least Significant Bit (Wolfgang et al., 1996), Spatial Domain (Ghazanfari et al., 2011), Frequency Domain (Holub et al., 2012)
- DL-based approaches: SteganoGAN (Zhang et al., 2019a), Self-Supervised Learning (Fernandez et al., 2021)

Introduction

北京大学 PEKING UNIVERSITY

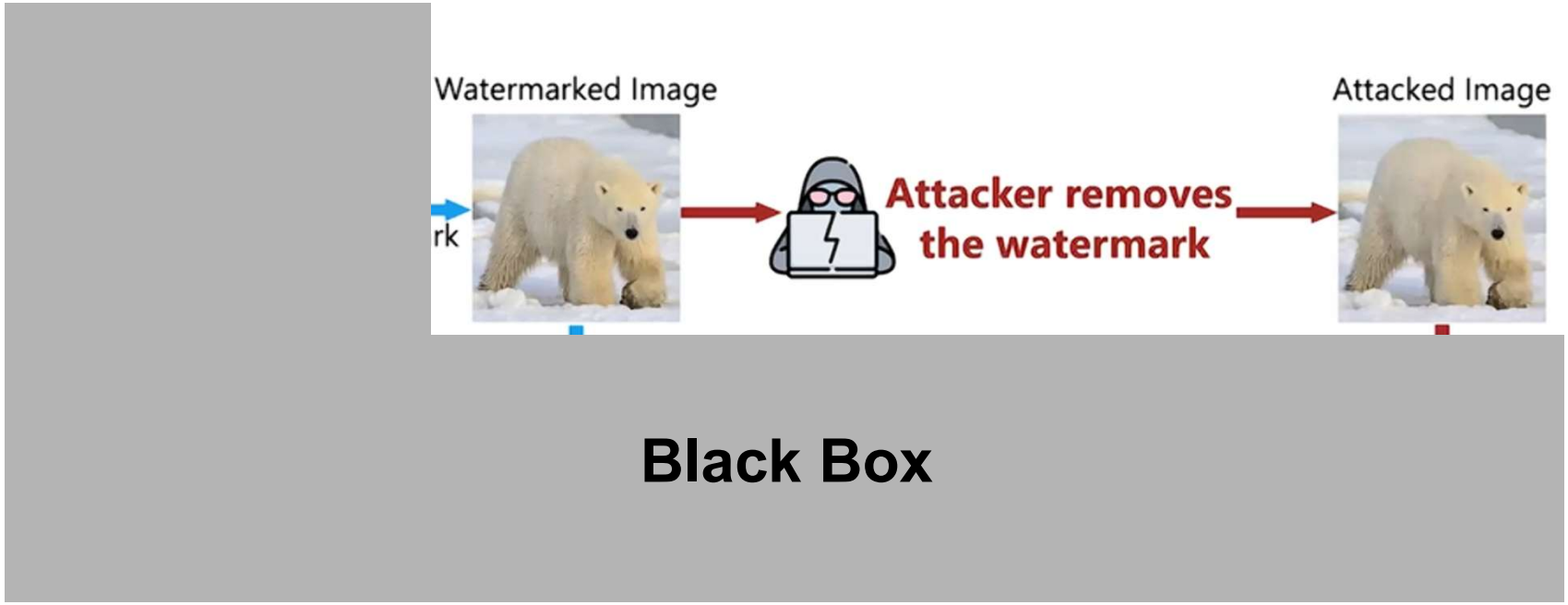
Emergence of invisible watermarking in AI industry…





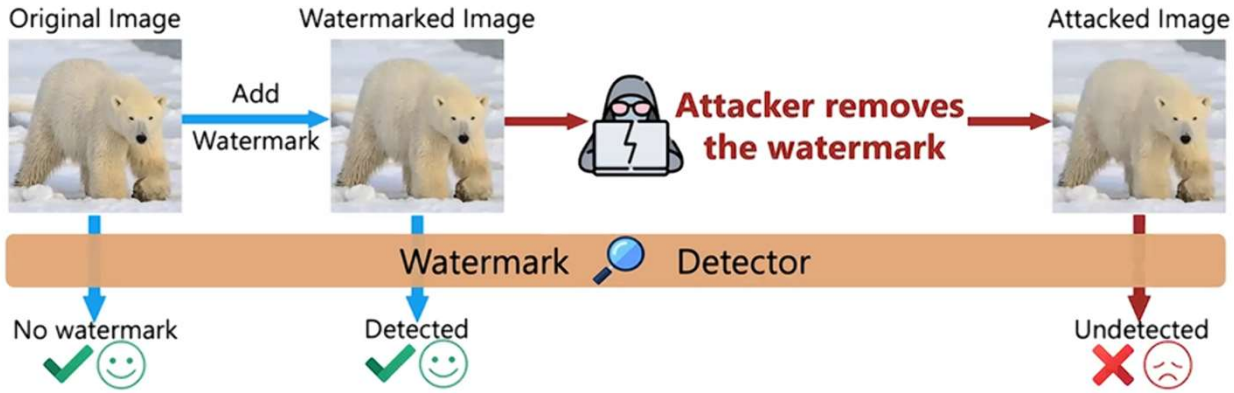
Introduction

Are Invisible Watermarks Good enough?



Introduction : Watermarking Schemes

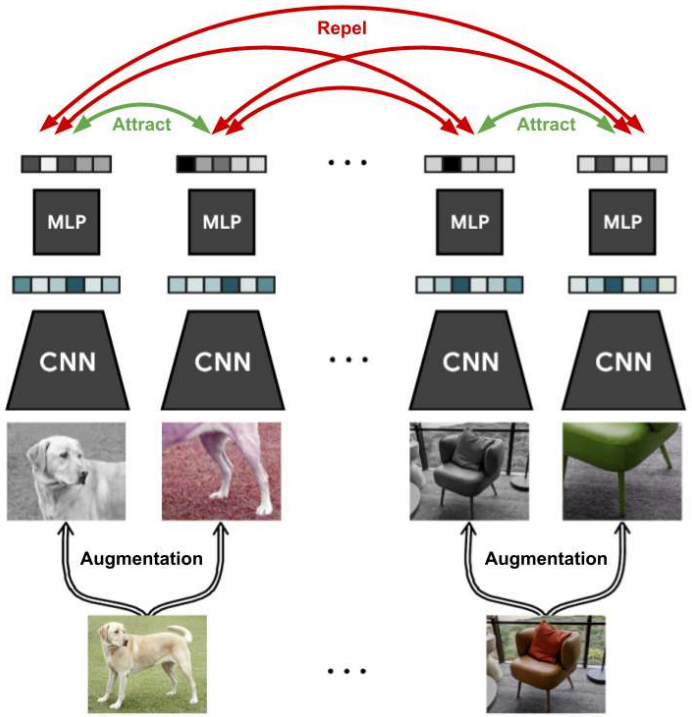
Are Invisible Watermarks Good enough?



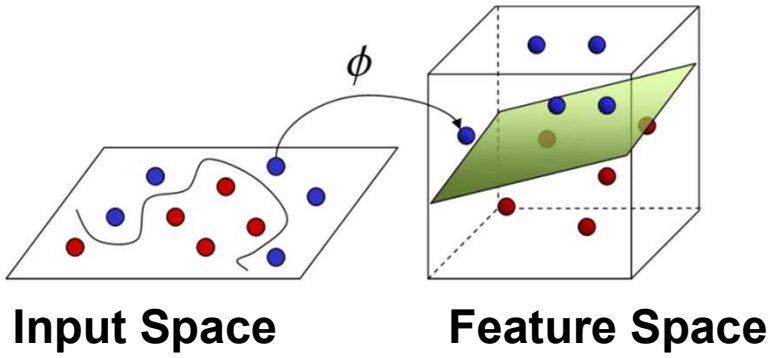
Background : Watermarking Schemes

北京大学 PEKING UNIVERSITY

Self-Supervised Learning (SSL)

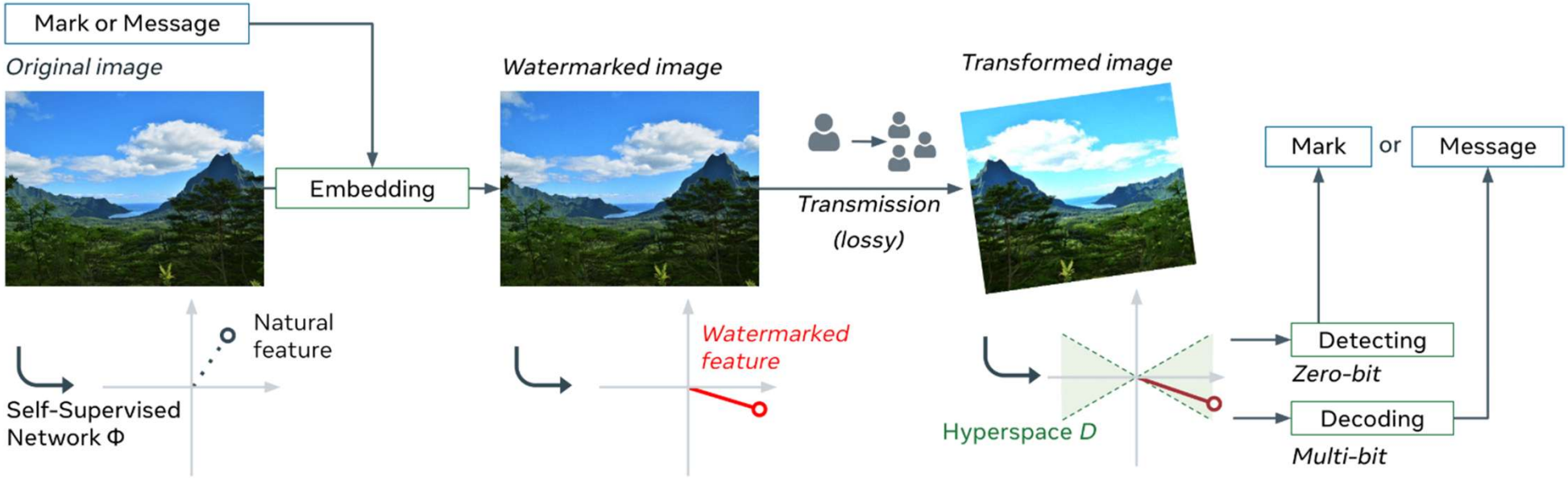


- 1. Intrinsic robustness of SSL to image transformations
- 2. Does not suffer from semantic collapse of supervised learning (More capacity for Watermarking)



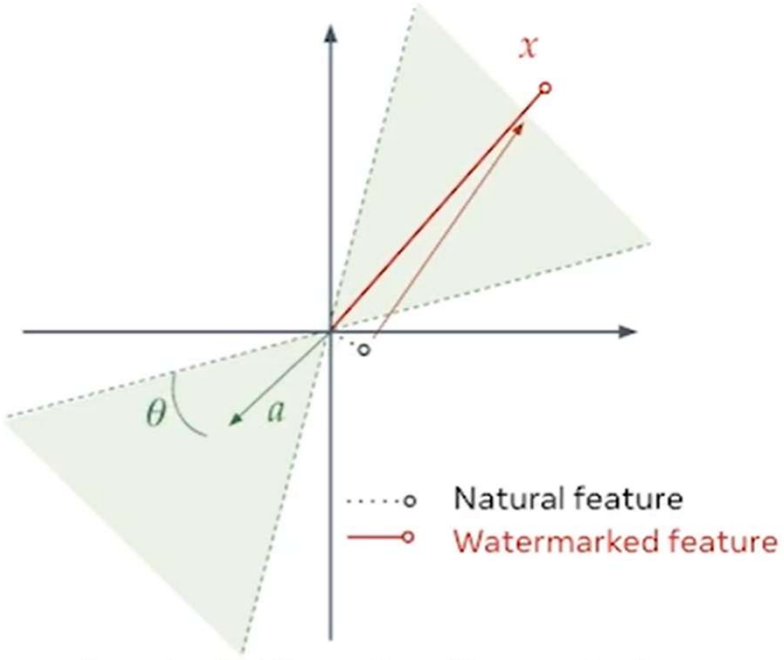
Fernandez, et al. Watermarking images in self-supervised latent spaces, ICASSP, 2022

Background : Watermarking Schemes

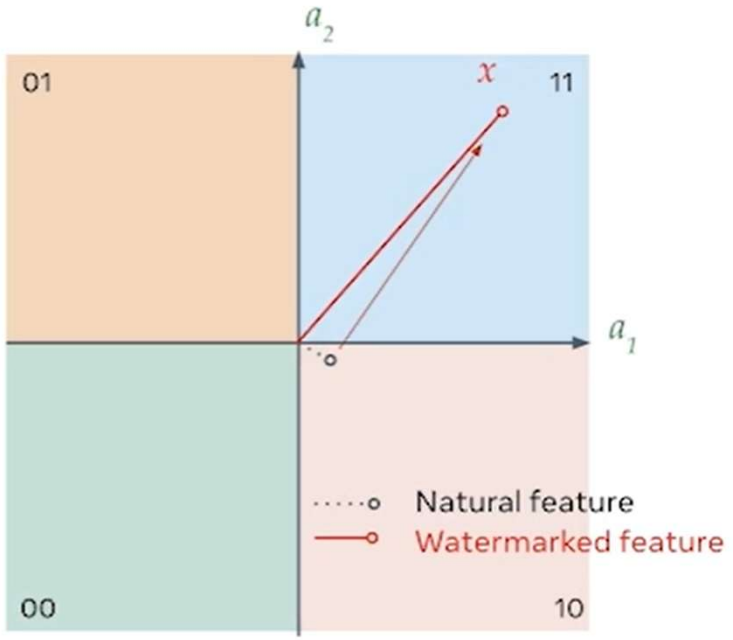


Fernandez, et al. Watermarking images in self-supervised latent spaces, ICASSP, 2022

Background : Watermarking Schemes



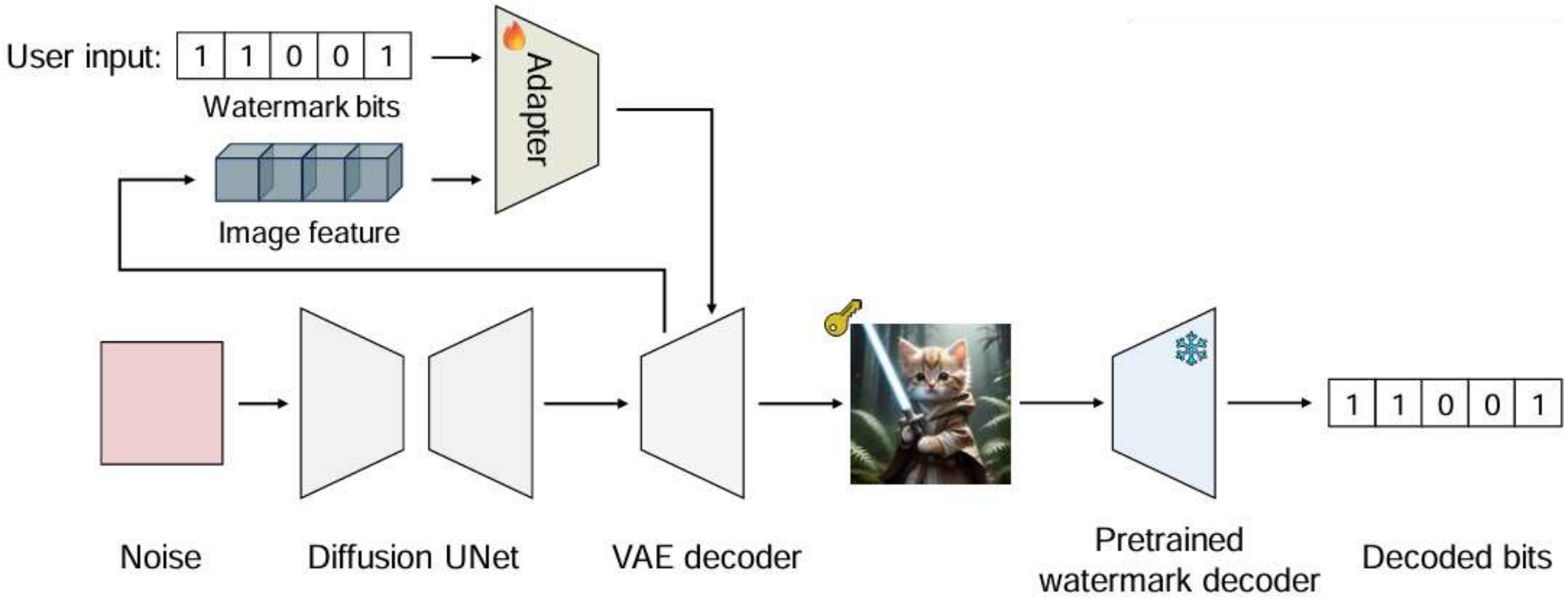
Example of 0-bit encoding with our method



Example of 2-bits encoding with our method

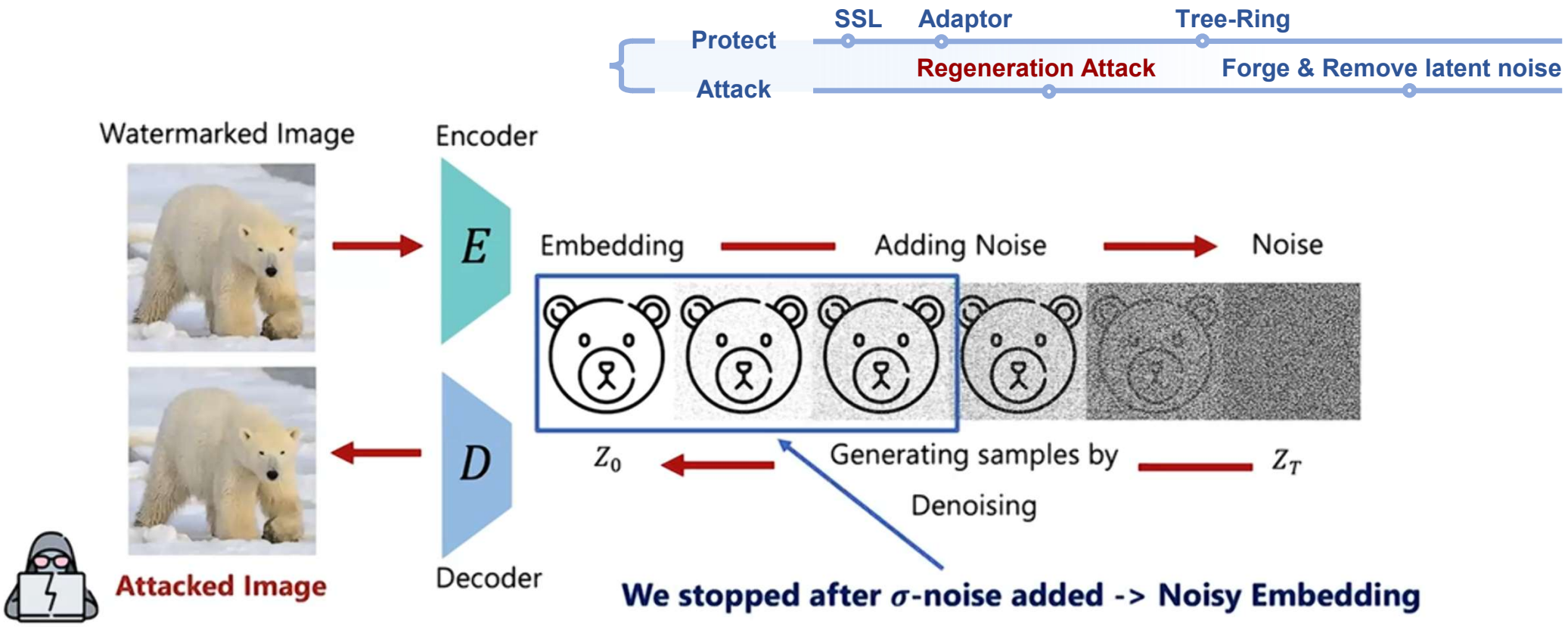
Fernandez, et al. Watermarking images in self-supervised latent spaces, ICASSP, 2022

Background : Watermarking Schemes



Hai Ci, et al. Wmadapter: Adding watermark control to latent diffusion models, ICML, 2025

Background : Watermarking Schemes



Xuandong Zhao, et al. Invisible Image Watermarks Are Provably Removable Using Generative AI, NeurIPS, 2024

Background : Watermarking Schemes

北京大学 PEKING UNIVERSITY

Guaranteed to remove any Pixelbased Invisible Watermarks.

Attacker	MS-COCO Dataset PSNR↑	SSIM↑	FID↓	Bit Acc↓	Detect Acc↓
Stable Signature watermarking:					
Brightness 0.5	28.53	0.864	11.75	0.967	0.990
Contrast 0.5	27.20	0.842	10.73	0.965	0.990
JPEG 50	29.37	0.873	15.01	0.866	0.966
Gaussian noise	25.46	0.788	30.60	0.920	0.972
Gaussian blur	25.48	0.798	17.72	0.896	0.986
BM3D denoise	29.24	0.871	31.65	0.946	0.952
VAE-Bmshj2018	28.95	0.867	31.86	0.636	0.248
VAE-Cheng2020	28.67	0.864	29.43	0.682	0.442
Diffusion model	29.33	0.879	20.64	0.486	0.000

Removing 100% Stable Signature Watermarks from Meta AI via diffusion attacks.

Xuandong Zhao, et al. Invisible Image Watermarks Are Provably Removable Using Generative AI, NeurIPS, 2024

Background : Watermarking Schemes

北京大学 PEKING UNIVERSITY

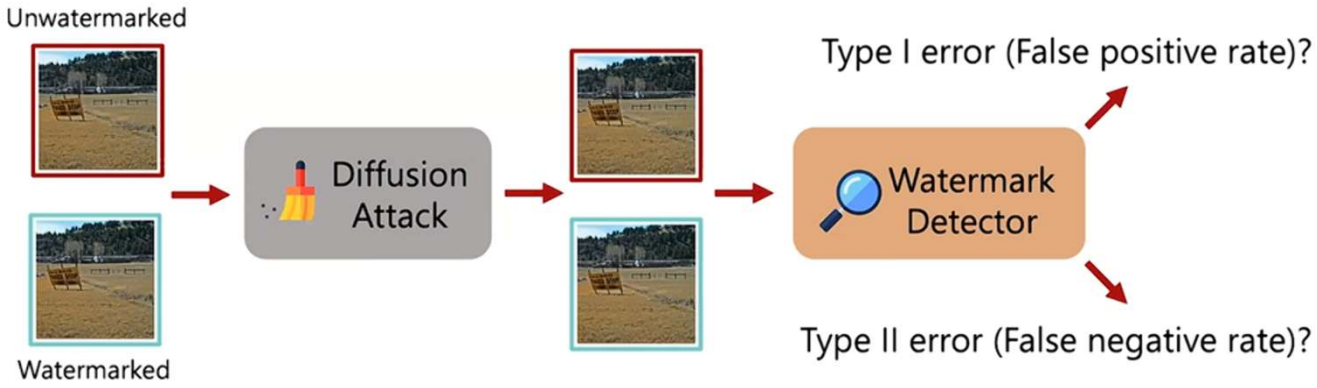
Definition 1 (Δ-Invisible watermark)**.**

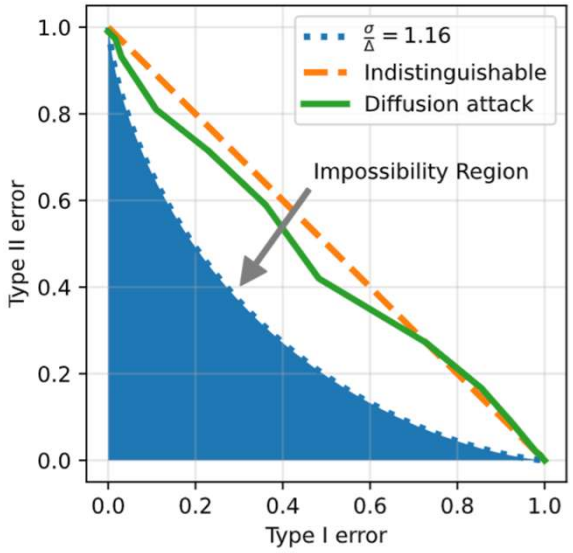
$$\mathrm{dist}(x, x_w) \leq \Delta. \qquad \|x - x_w\| \leq \Delta.$$

Definition 2 (f-Certified-Watermark-Free)**.**

$$\varepsilon_2 \geq f(\varepsilon_1), \quad \forall \varepsilon_1 \in [0,1].$$

x clean
x_w watermarked
ε_1 FP rate
ε_2 FN rate





Xuandong Zhao, et al. Invisible Image Watermarks Are Provably Removable Using Generative AI, NeurIPS, 2024

Background : Watermarking Schemes

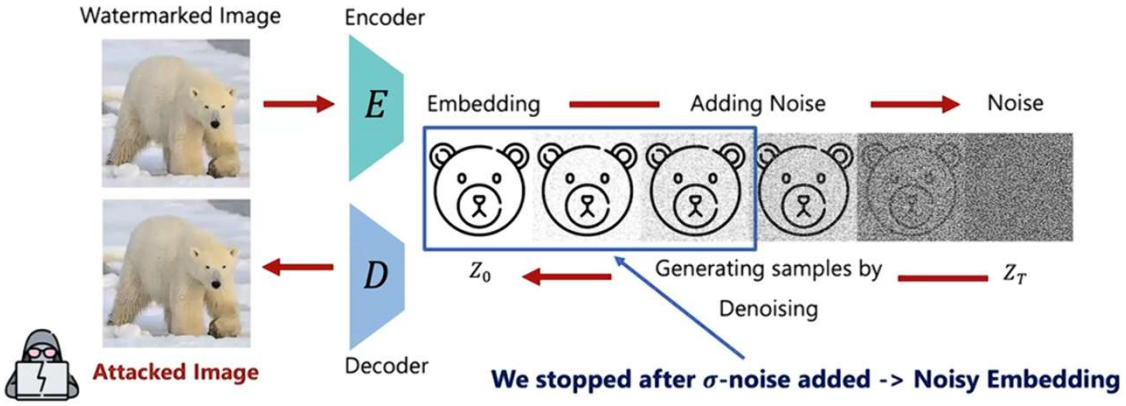
北京大学 PEKING UNIVERSITY

x clean
x_w watermarked
ε_1 FP rate
ε_2 FN rate

Definition 1 (Δ-Invisible watermark). $\mathrm{dist}(x, x_w) \leq \Delta. \quad \|x - x_w\| \leq \Delta.$

Definition 2 (f-Certified-Watermark-Free). $\varepsilon_2 \geq f(\varepsilon_1), \quad \forall \varepsilon_1 \in [0,1].$

Definition 3 (Local Watermark-Specific Lipschitz property). $|\phi(x_w) - \phi(x)| \leq L_{x,w}|x_w - x|$



$$\begin{cases} \tilde{x} \sim A\big(\phi(x_0) + N(0, \sigma^2 I)\big) \\ \tilde{x} \sim A\big(\phi(x_w) + N(0, \sigma^2 I)\big) \end{cases}$$

Xuandong Zhao, et al. Invisible Image Watermarks Are Provably Removable Using Generative AI, NeurIPS, 2024

Background : Watermarking Schemes

北京大学 PEKING UNIVERSITY

x clean
x_w watermarked
ε_1 FP rate
ε_2 FN rate

Definition 1 (Δ-Invisible watermark). $\mathrm{dist}(x, x_w) \le \Delta. \quad \|x - x_w\| \le \Delta.$

Definition 2 (f-Certified-Watermark-Free). $\varepsilon_2 \ge f(\varepsilon_1), \quad \forall \varepsilon_1 \in [0,1].$

Definition 3 (Local Watermark-Specific Lipschitz property). $|\phi(x_w) - \phi(x)| \le L_{x,w}|x_w - x|$

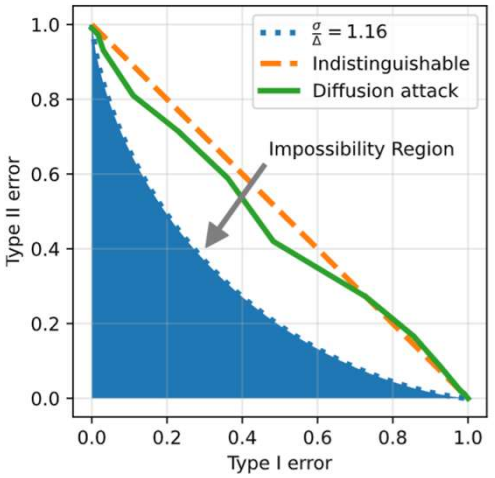
$$\begin{cases} \tilde{x} \sim A\big(\phi(x_0) + N(0, \sigma^2 I)\big) \\ \tilde{x} \sim A\big(\phi(x_w) + N(0, \sigma^2 I)\big) \end{cases} \quad \begin{cases} N(\phi(x_0), \sigma^2 I) \\ N(\phi(x_w), \sigma^2 I) \end{cases} \quad \begin{matrix} \mu = \phi(x_w) - \phi(x_0) \\ \|\mu\| \le L_{x,w}\Delta \\ \longrightarrow \end{matrix} \quad \begin{cases} N(0,1) \\ N\left(\frac{L_{x,w}\Delta}{\sigma}, 1\right) \end{cases}$$

Theorem1 (Neyman-Pearson Lemma).

The likelihood ratio test is uniform most powerful.

Theorem2 (Gaussian Differential Privacy [Dong et al. 2022]).

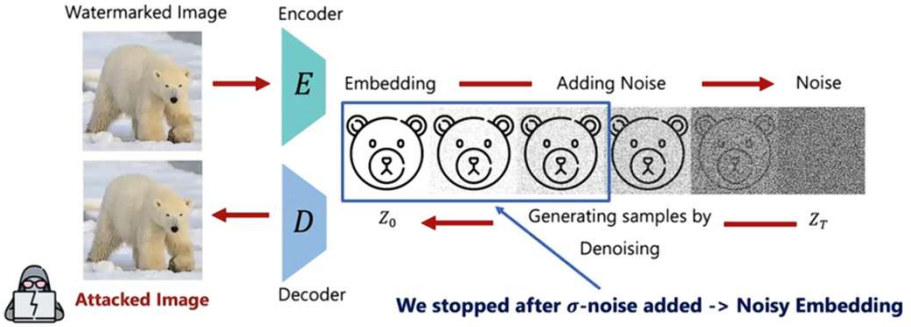
$$f(\varepsilon_1) = \Phi\left(\Phi^{-1}(1 - \varepsilon_1) - \frac{L_{x,w}\Delta}{\sigma}\right). \quad \varepsilon_2 \ge f(\varepsilon_1), \quad \forall \varepsilon_1 \in [0,1].$$



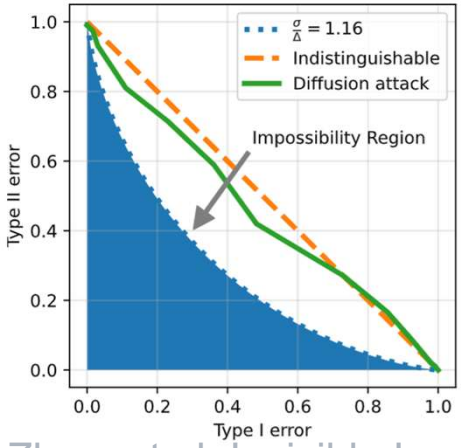
Background : Watermarking Schemes

$$\|x - x_w\| \leq \Delta.$$

$$|\phi(x_w) - \phi(x)| \leq L_{x,w}|x_w - x|$$



$$f(\varepsilon_1) = \Phi!\left(\Phi^{-1}(1-\varepsilon_1) - \frac{L_{x,w}\Delta}{\sigma}\right).$$

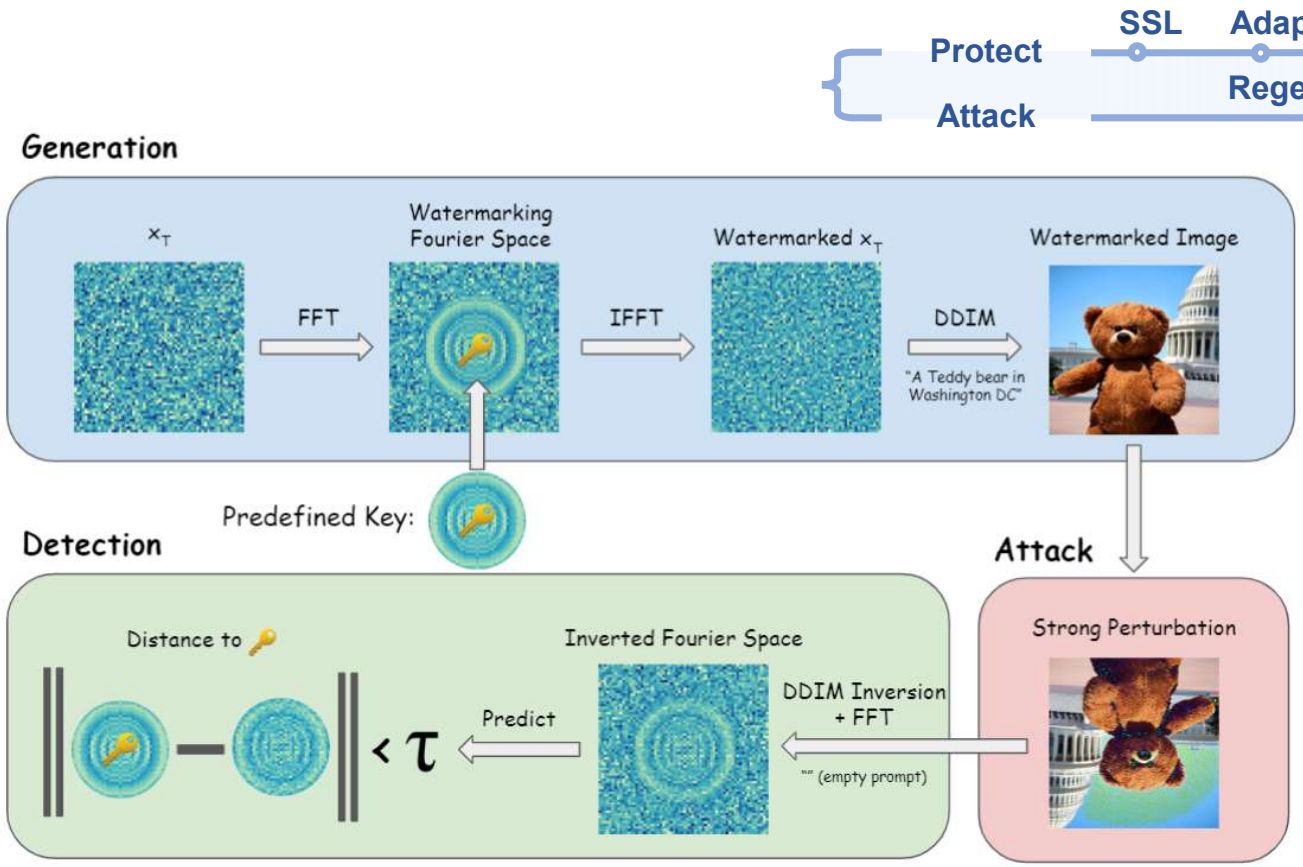


The regeneration attack is guaranteed to remove any pixel-based invisible watermarks.

Xuandong Zhao, et al. Invisible Image Watermarks Are Provably Removable Using Generative AI, NeurIPS, 2024

Background : Watermarking Schemes

北京大学 PEKING UNIVERSITY



Protect
- SSL
- Adaptor
- Tree-Ring

Attack
- Regeneration Attack
- Forge & Remove latent noise

Preliminaries

DDIM (deterministic strategy)

$$x_t = \sqrt{\overline{\alpha_t}} x_0 + \sqrt{1 - \overline{\alpha_t}} \epsilon$$

$$x_{t-1} - x_t \approx x_{t+1} - x_t$$

$$x_{t+1} = \sqrt{\overline{\alpha_{t+1}}} \widehat{x_0} + \sqrt{1 - \overline{\alpha_{t+1}}} \epsilon_\theta(x_t)$$

Yuxin Wen, et al. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust, NeurIPS, 2023

Background : Watermarking Schemes

北京大学 PEKING UNIVERSITY

"An astronaut riding a horse in Zion National Park."

(a) No Watermark

(b) Pixel WM

(c) Semantic WM

➡ The robustness comes at the price of more visible differences.

$\|x - x_w\| \leq \Delta.$ ➡ **Not Satisfied!**

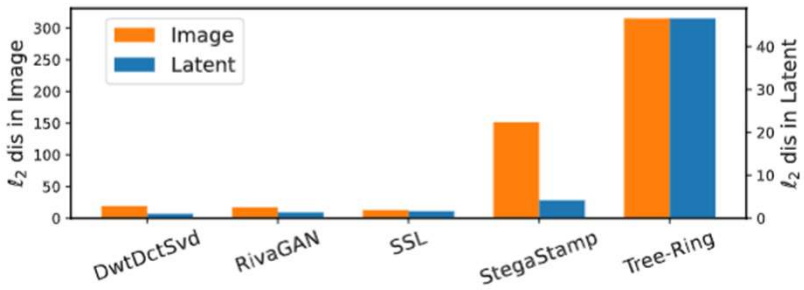


Table 2: Tree-Ring watermarks are robust against the regeneration attacks.

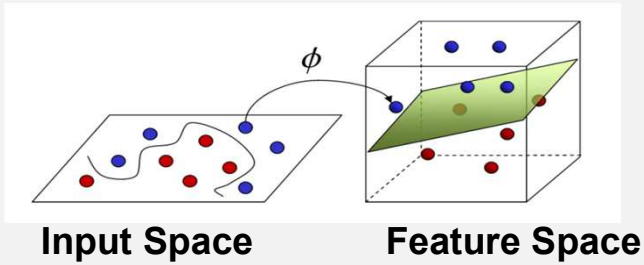
Attacker	MS-COCO TPR@FPR=0.01↓	SDP Generated TPR@FPR=0.01↓
Brightness-2	1.000	1.000
Contrast-2	1.000	1.000
JPEG-50	1.000	0.994
Gaussian noise-5	1.000	0.996
Gaussian blur-6	1.000	1.000
VAE-Bmshj2018-3	0.998	0.994
VAE-Cheng2020-3	1.000	0.994
Diffusion model-60	1.000	0.998

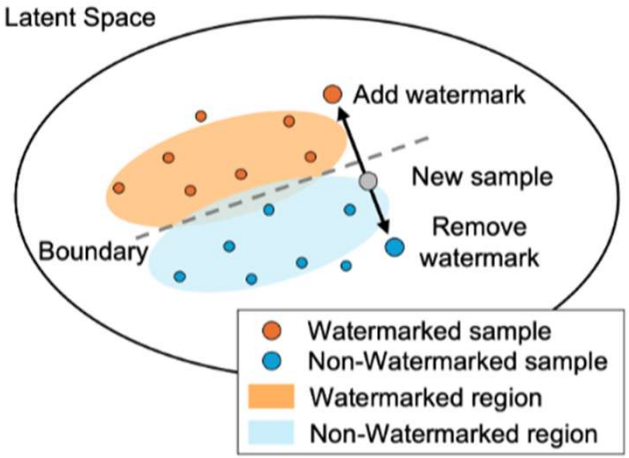
Yuxin Wen, et al. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust, NeurIPS, 2023

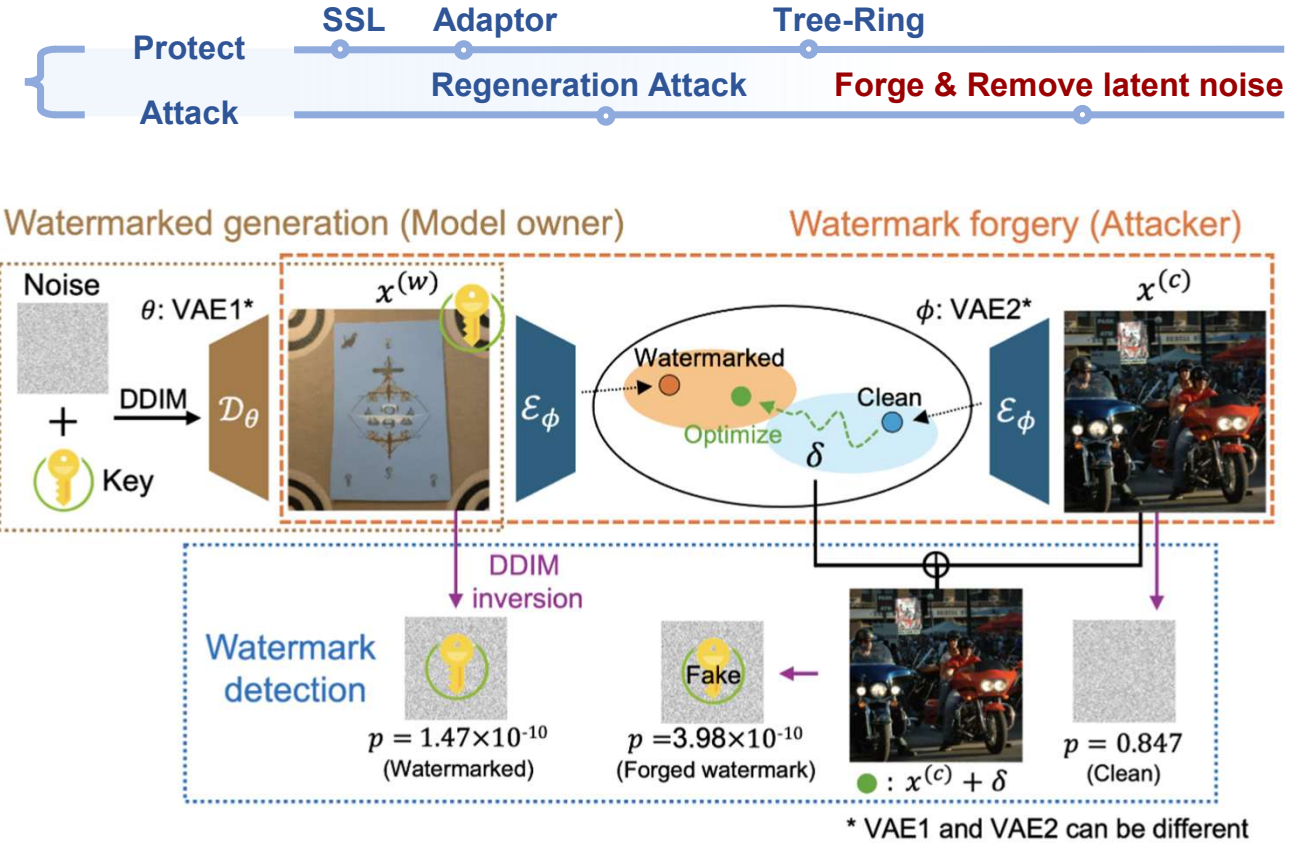
Motivation: Threat Model

北京大学 PEKING UNIVERSITY

Recall



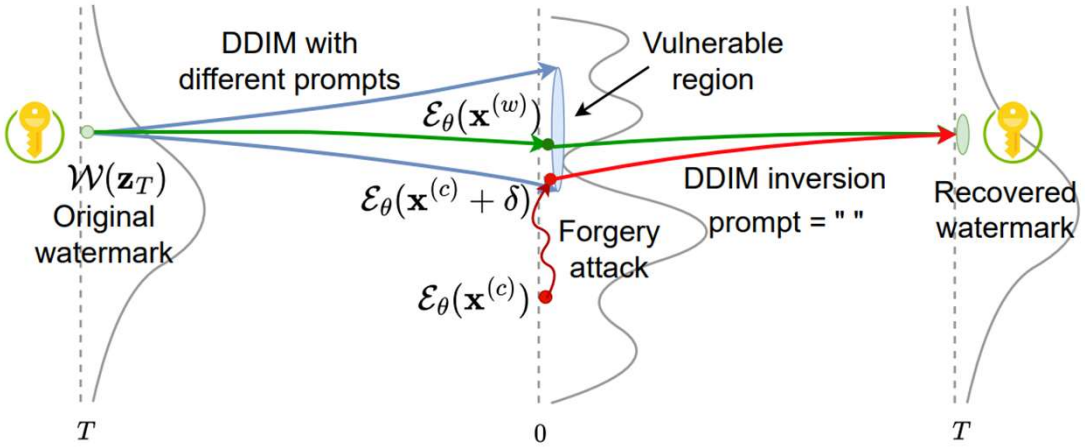




Anubhav Jain, et al. Forging and Removing Latent-Noise Diffusion Watermarks Using a Single Image, 2025

Method: Watermark Region

Hypothesis: Latent Space Contains a "Watermark Region"



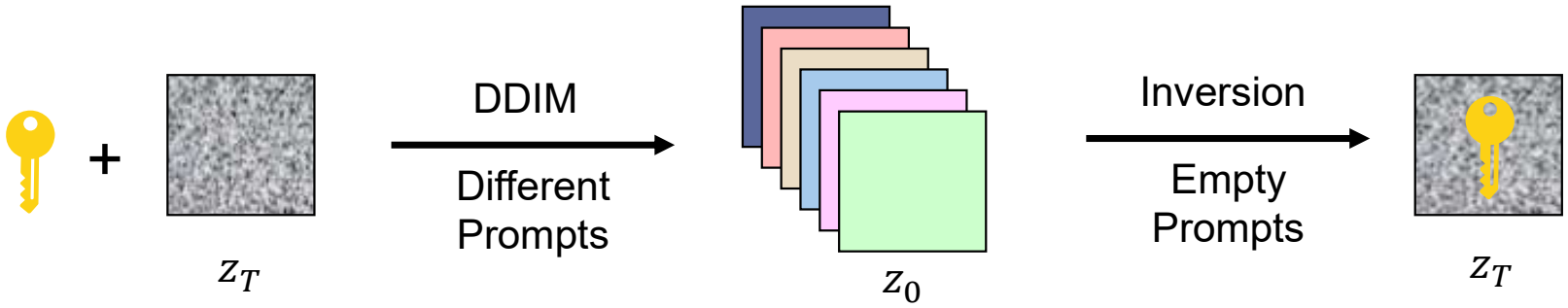
Preliminaries

DDIM (deterministic strategy)

$$x_t = \sqrt{\overline{\alpha_t}}x_0 + \sqrt{1-\overline{\alpha_t}}\epsilon$$

$$x_{t-1} - x_t \approx x_{t+1} - x_t$$

$$x_{t+1} = \sqrt{\overline{\alpha_{t+1}}}\widehat{x_0} + \sqrt{1-\overline{\alpha_{t+1}}}\epsilon_\theta(x_t)$$



Experiment: Watermark Region

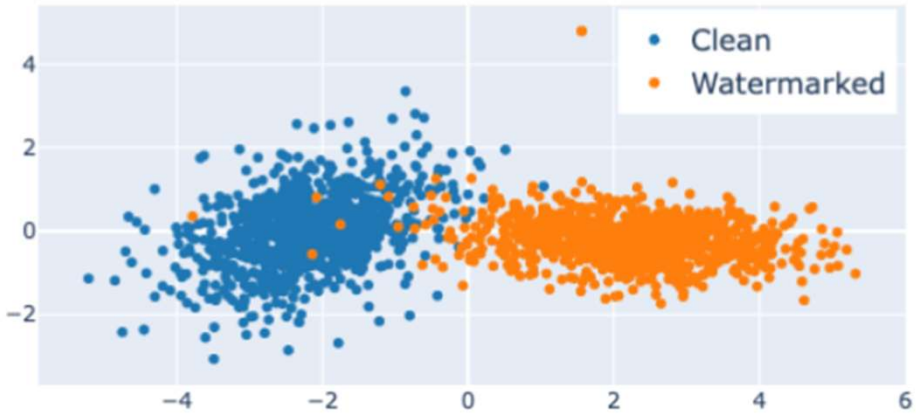
Hypothesis: Latent Space Contains a "Watermark Region"

$$Z_0^{(w)}(\mathcal{W}, k) = \{\mathbf{z_0} \in Z_0 \mid \mathcal{M}_{\mathcal{W}}(\mathcal{I}^-(\mathbf{z_0}), k) < \tau\}$$

Z_0: Clean latent space
$\mathcal{I}^-$: DDIM inversion process
τ : Detection threshold (e.g., p-value < 0.05)
$\mathcal{M}_{\mathcal{W}}$: Matching function between the extracted key and the embedded key k

Experiment 1: Linear Separable

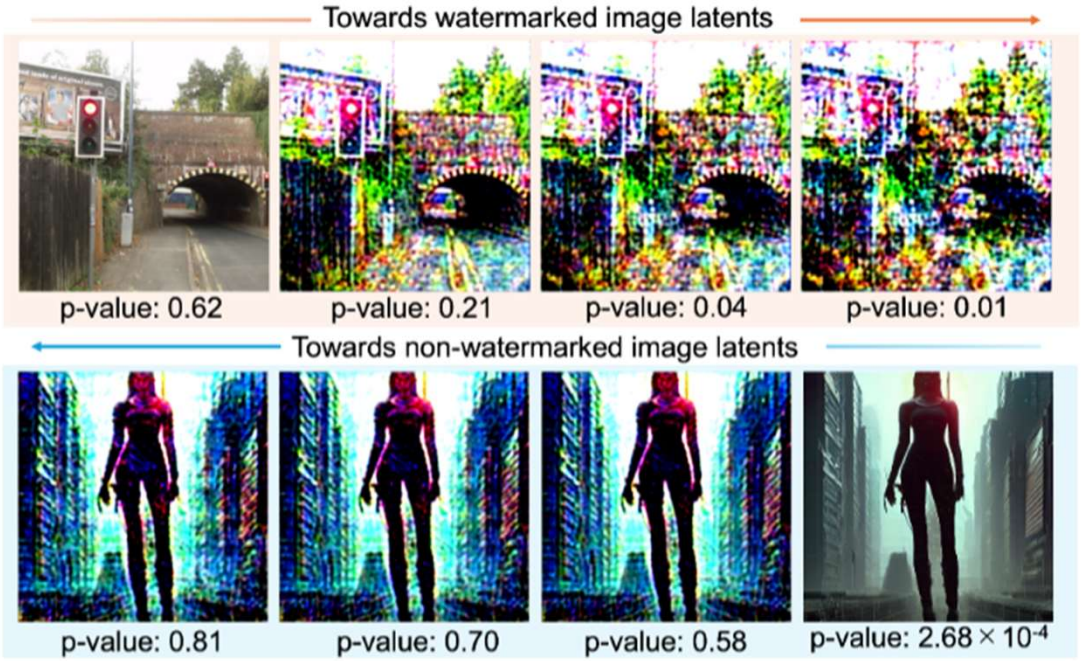
- 1000 watermarked images from a specific watermark and the same number of nonwatermarked images
- The linear SVM model yields an accuracy of 100% on the sampled images

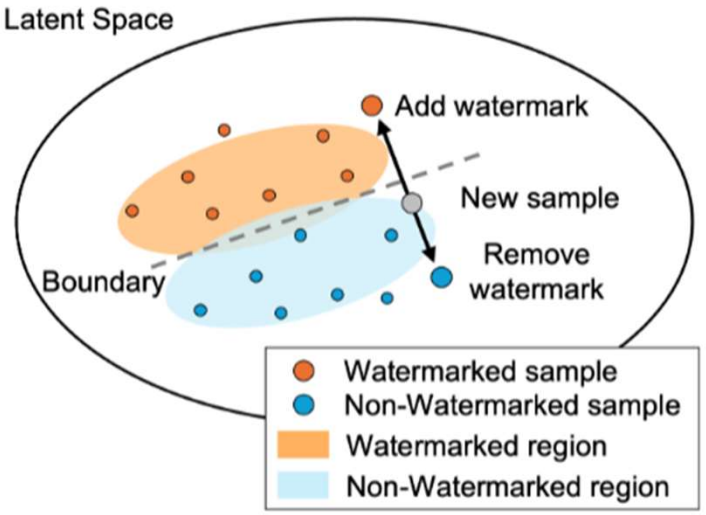


Experiment： Watermark Region

Hypothesis: Latent Space Contains a “Watermark Region”

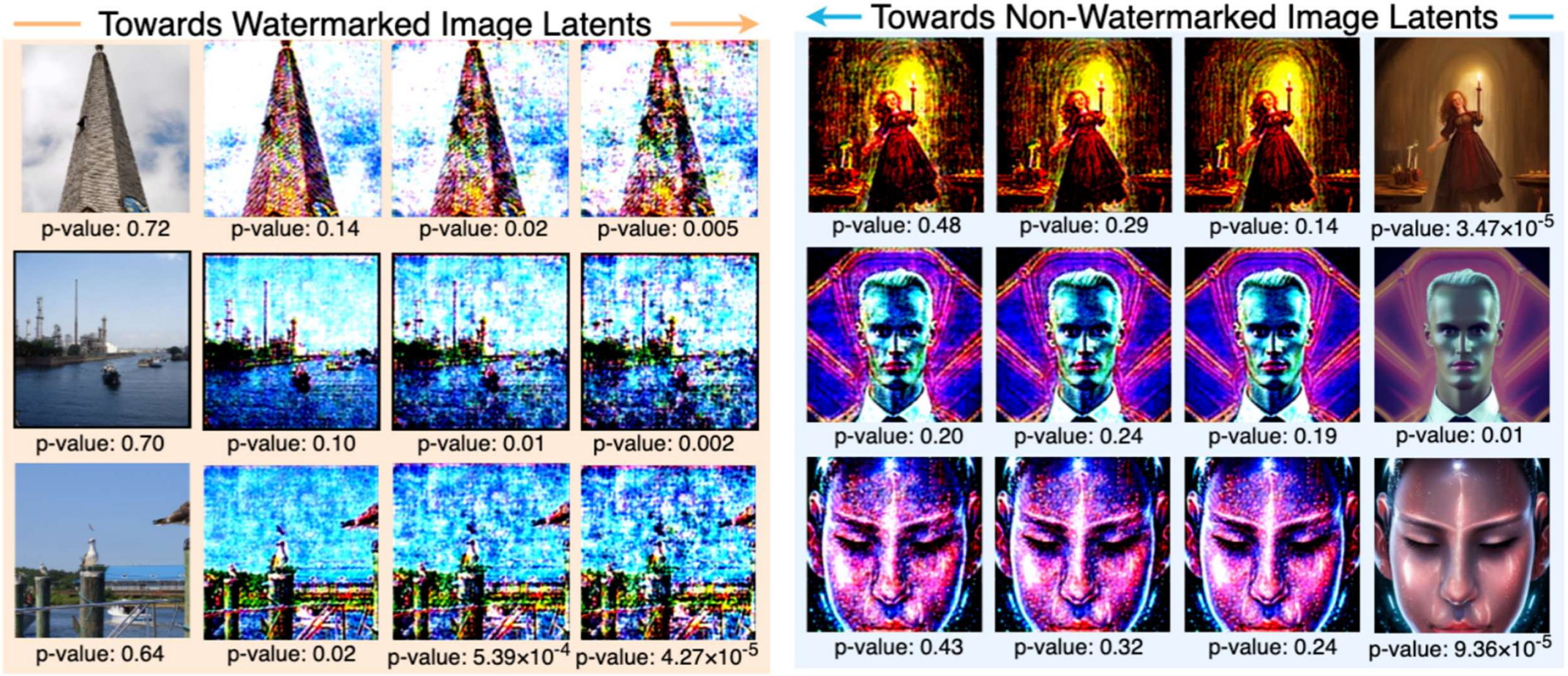
Experiment 2: Coherent and Oriented latent subspace





Anubhav Jain, et al. Forging and Removing Latent-Noise Diffusion Watermarks Using a Single Image, 2025

Experiment： Watermark Region



Anubhav Jain, et al. Forging and Removing Latent-Noise Diffusion Watermarks Using a Single Image, 2025

Method: Imperceptible Attack

北京大学 PEKING UNIVERSITY

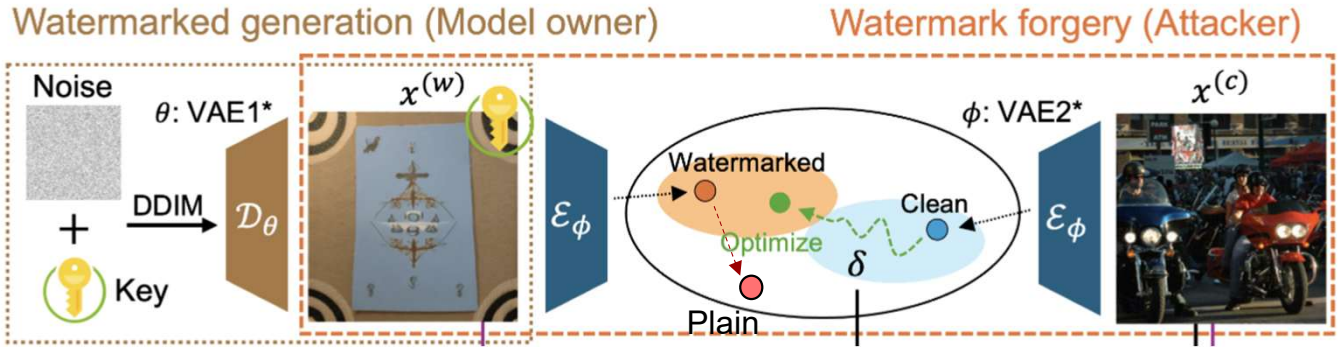
Forgery Attack

$$\min_{\delta} \left\| \mathcal{E}_\phi\left(x^{(c)} + \delta\right) - \mathcal{E}_\phi\left(x^{(w)}\right) \right\|_2 + \lambda \|\delta\|_2$$

Removal Attack

$$\min_{\delta} \left\| \mathcal{E}_\phi\left(x^{(c)} + \delta\right) - \mathcal{E}_\phi\left(\mu_{\mathbf{x}^{(w)}}\right) \right\|_2 + \lambda \|\delta\|_2$$

$x^{(w)}$	watermarked image
$x^{(c)}$	clean image
$\mu_{\mathbf{x}^{(w)}}$	plain image
δ	adversarial perturbation



Evaluation

北京大学 PEKING UNIVERSITY

Hypothesis

- The attacker has access to either
- (i) the VAE of the watermarked diffusion model
- (ii) a proxy VAE that was trained on a similar dataset

Diffusion Model

- Utilize the VAE from SDv1.4 to remove/forge the watermark generated from both SDv1.4 and SDv2.0

Watermarking Systems

- Tree-Ring, RingID, WIND, and Gaussian Shading

Evaluation

PEKING UNIVERSITY

P-Value

- Null Hypothesis: not watermarked.
- The p-value represents the probability of observing the extracted key.
- A small p-value, we reject H_0 and conclude the image is watermarked.

Attack Success Rate (ASR)

- Forgery Attack: Percentage of p-value < 0.05 after attack.
- Removal Attack: Percentage of p-value ≥ 0.05 after attack.

Image Quality Metrics

- L2 distance, PSNR, SSIM, LPIPS.

Experiments

北京大学 PEKING UNIVERSITY

Trade-off between attack success rate (ASR) and imperceptibility.

Method	Model	λ	ASR	l_2	l_∞	LPIPS	SSIM	PSNR	FID
Tree-Ring [35]	SDv1.4	5×10^4	78.65	33.90	0.69	0.17	0.89	34.32	41.27
		2×10^4	86.93	48.42	0.89	0.26	0.82	31.20	61.18
		1×10^4	91.06	63.22	1.10	0.33	0.76	28.87	81.27
	SDv2.0	5×10^4	79.89	34.09	0.69	0.17	0.88	34.26	41.22
		2×10^4	90.72	48.83	0.91	0.26	0.82	31.11	61.75
		1×10^4	93.81	63.78	1.08	0.34	0.76	28.78	80.54
RingID [5]	SDv1.4	5×10^4	100.0	38.45	0.68	0.20	0.87	33.21	49.97
		2×10^4	100.0	55.20	0.86	0.30	0.80	30.06	75.55
		1×10^4	100.0	73.08	1.03	0.38	0.73	27.63	101.45
	SDv2.0	5×10^4	100.0	37.31	0.66	0.19	0.87	33.48	47.87
		2×10^4	100.0	53.94	0.84	0.29	0.80	30.27	72.08
		1×10^4	100.0	71.53	1.00	0.37	0.73	27.82	98.12
WIND [2]	SDv1.4	5×10^4	97.56	38.82	0.70	0.20	0.87	33.11	54.41
		2×10^4	97.56	56.13	0.89	0.29	0.80	29.88	82.68
		1×10^4	97.56	74.66	1.06	0.38	0.73	27.38	111.84
	SDv2.0	5×10^4	100.0	37.47	0.67	0.19	0.87	33.45	49.02
		2×10^4	100.0	54.18	0.84	0.28	0.80	30.23	74.69
		1×10^4	100.0	71.86	0.99	0.37	0.74	27.78	101.33
Gaussian Shading [39]	SDv1.4	5×10^4	96.85	37.27	0.70	0.19	0.87	33.48	46.93
		2×10^4	96.96	54.00	0.88	0.29	0.80	30.21	69.50
		1×10^4	96.96	71.97	1.05	0.37	0.73	27.64	93.85
	SDv2.0	5×10^4	100.0	36.78	0.66	0.19	0.87	33.60	46.20
		2×10^4	100.0	52.99	0.85	0.29	0.80	30.42	69.35
		1×10^4	100.0	69.83	1.02	0.37	0.74	28.02	93.04

Experiments

Original Watermarked

$\lambda = 5 \times 10^4$

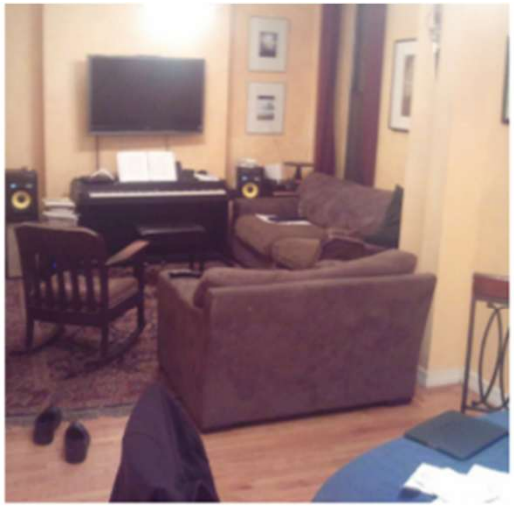
$\lambda = 2 \times 10^4$

$\lambda = 1 \times 10^4$

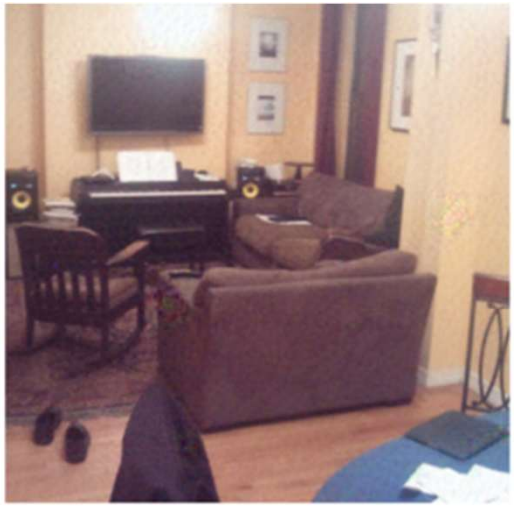
Tree-Ring watermarking method

λ	ASR	l_2	l_∞	LPIPS	SSIM	PSNR	FID
5×10^4	94.21	74.82	0.94	0.19	0.84	20.72	54.16
2×10^4	97.68	87.20	1.14	0.27	0.79	20.61	75.74
1×10^4	98.84	100.52	1.35	0.34	0.74	20.44	94.62

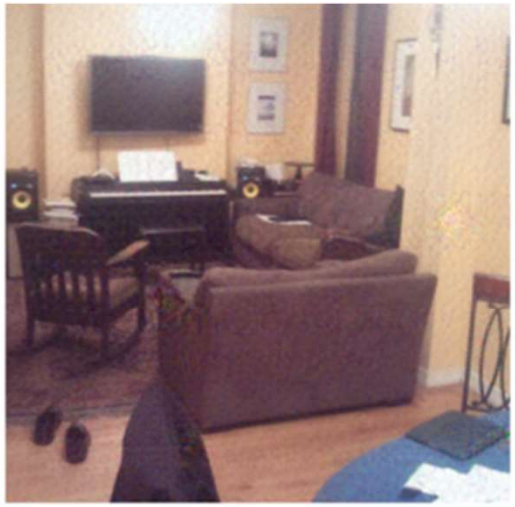
Experiments



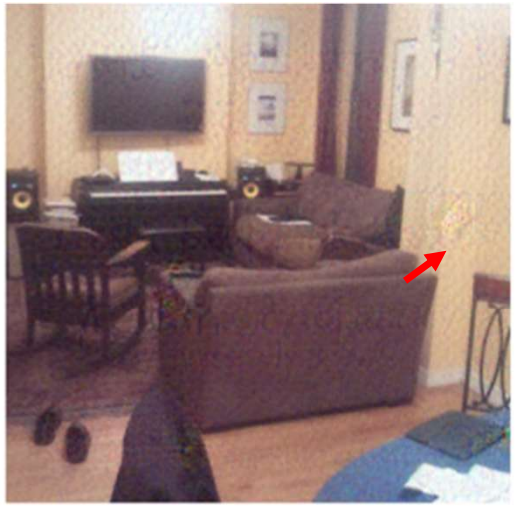
Original



$\lambda = 5 \times 10^4$



$\lambda = 2 \times 10^4$



$\lambda = 1 \times 10^4$

Gaussian Shading watermarking method

λ	ASR	l_2	l_∞	LPIPS	SSIM	PSNR	FID
5×10^4	96.85	37.27	0.70	0.19	0.87	33.48	46.93
2×10^4	96.96	54.00	0.88	0.29	0.80	30.21	69.50
1×10^4	96.96	71.97	1.05	0.37	0.73	27.64	93.85

Ablation

北京大学 PEKING UNIVERSITY

Different Optimization Objectives

Spatial Domain + L_2 Regularization:

$$\min_{\boldsymbol{\delta}} \|\mathcal{E}_\phi(\mathbf{x}^{(c)} + \boldsymbol{\delta}) - \mathcal{E}_\phi(\mathbf{x}^{(w)})\|_2 + \lambda\|\boldsymbol{\delta}\|_2, \quad (3)$$

Spatial Domain + L_∞ Constraint:

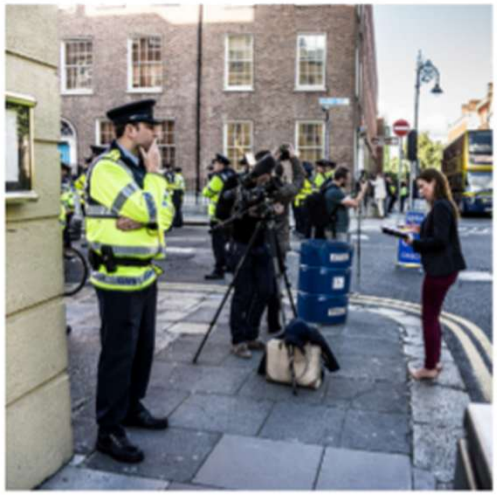
$$\min_{\boldsymbol{\delta}} \|\mathcal{E}_\phi(\mathbf{x}^{(c)} + \boldsymbol{\delta}) - \mathcal{E}_\phi(\mathbf{x}^{(w)})\|_2 \quad \text{s.t.} \quad \|\boldsymbol{\delta}\|_\infty \leq \epsilon, \quad (5)$$

Frequency Domain + L_∞ Constraint:

$$\min_{\boldsymbol{\delta}} \|\mathcal{E}_\phi(\text{IDCT}(\text{DCT}(\mathbf{x}^{(c)}) + \mathbf{m} \odot \boldsymbol{\delta})) - \mathcal{E}_\phi(\mathbf{x}^{(w)})\|_2$$
$$\text{s.t.} \quad \|\boldsymbol{\delta}\|_\infty \leq \epsilon. \quad (6)$$

Method	ASR	l_2	l_∞	LPIPS	SSIM	PSNR	FID
Eq. 5	83.24	68.04	0.10	0.35	0.66	28.31	69.35
Eq. 6	83.33	67.77	0.93	0.32	0.66	28.35	58.01
Eq. 3	84.91	44.34	0.85	0.23	0.84	31.97	56.30

Ablation



Original

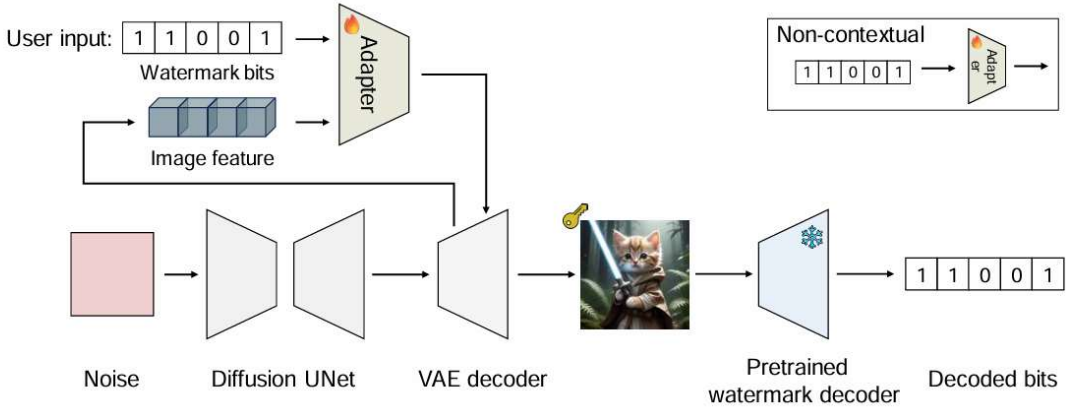
Eq. 5

Eq. 6

Eq. 3

Limitations & Conclusion

- **Generative regeneration attacks：**
  - Semantic Watermarks - Initial-Noise Watermarks
- **Latent-space adversarial attacks ：**
  - Rely on DDIM inversion watermarking methods
  - Approaches that add a pattern during the latent decoding phase; require access to a similar decoder to attack the system



Limitations & Conclusion

- **Pixel-Level Watermarks:**
  - Perturbation limited to ℓ_2 ball → **generative regeneration**
- **Initial-Noise Watermarks:**
  - Rely on DDIM inversion → **latent-space adversarial attacks**

Thanks for Listening!