An Undetectable Watermark for Generative Image Models
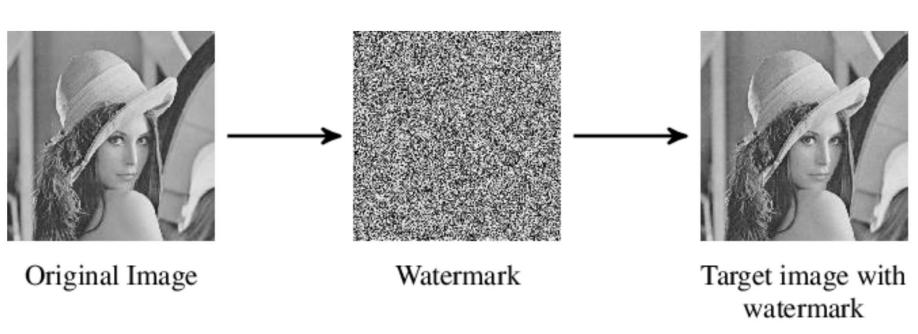
Sam Gunn, Xuandong Zhao*, Dawn Song*

UC Berkeley

Invisible image watermarking schemes

Post-processing schemes
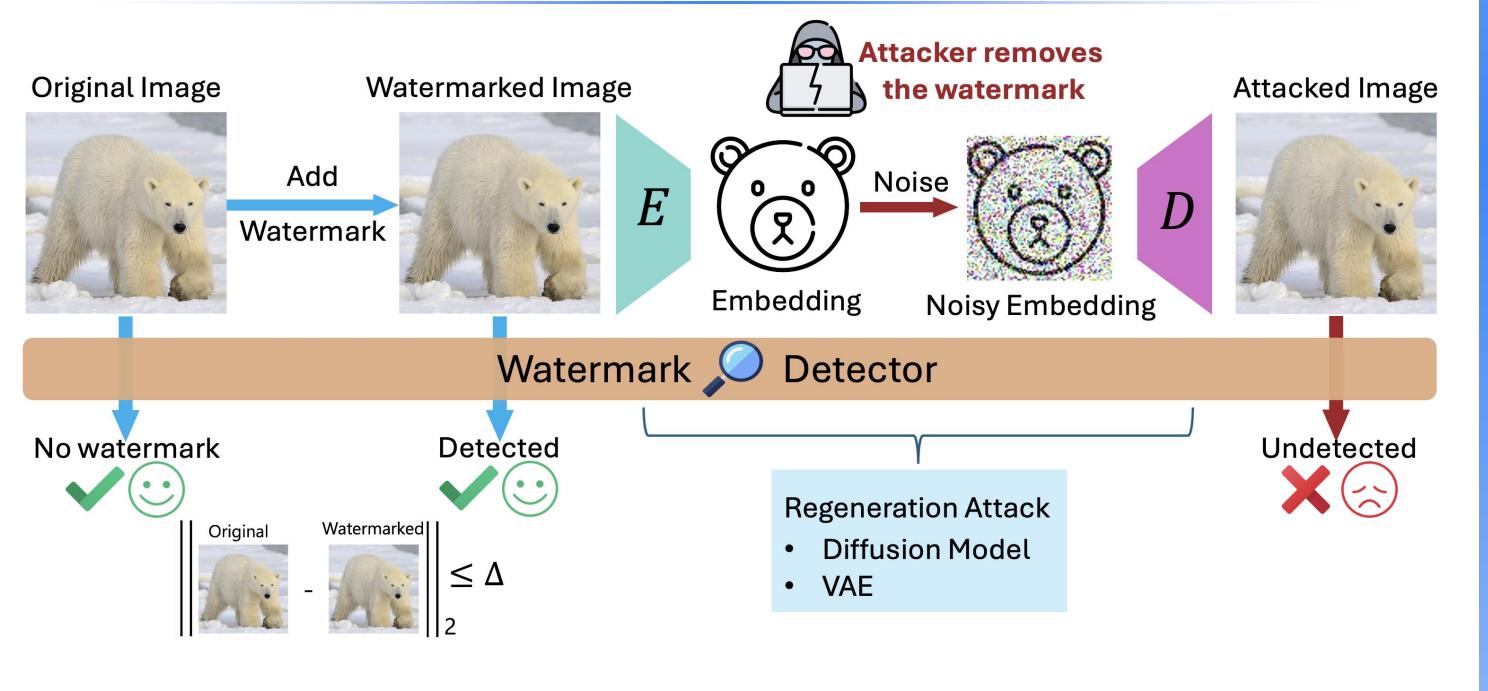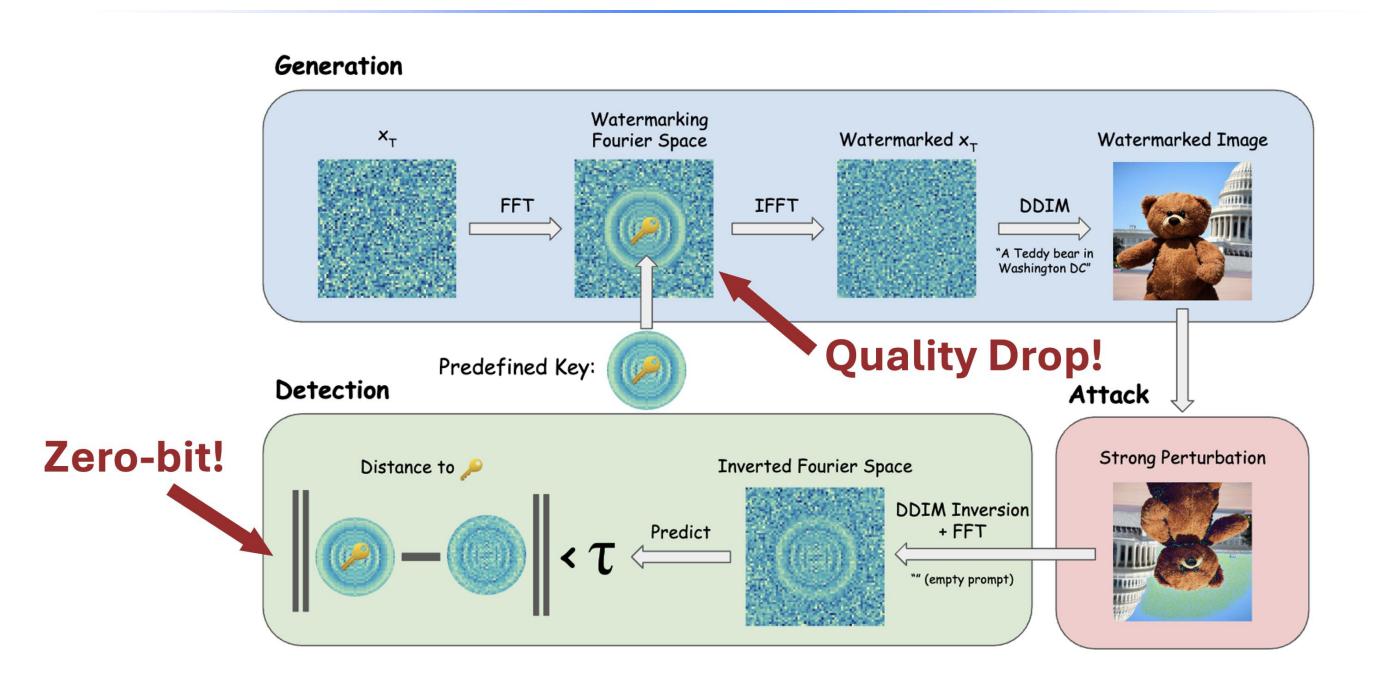- Embed the watermark after image generation

Original Image — Watermark — Target image with watermark

In-processing schemes
- Embed the watermark directly in the generated content

Watermark Key → Prompt → Watermark Generation

Pixel-level invisible image watermarks are provably removable using generative ai

Original Image → Add Watermark → Watermarked Image → E → Embedding → Noise → Noisy Embedding → D → Attacked Image

Attacker removes the watermark

Watermark Detector

No watermark ✓ — Detected ✓ — Undetected ✗

$|\text{Original} - \text{Watermarked}| \leq \Delta$ (2-norm)

Regeneration Attack
- Diffusion Model
- VAE

In-processing sematic watermarks are robust to regeneration attacks

Generation — Detection

Quality Drop!

Zero-bit!

Pseudorandom code (PRC) watermarking scheme

inverse diffusion

~~random latents~~ **pseudorandom code** → diffusion model → image

prompt →

Pseudorandom code (PRC):
Pseudorandomness with a *robust* trapdoor detection algorithm (decoding)

A pseudorandom code (PRC) is a pair of algorithms Enc, Dec satisfying:

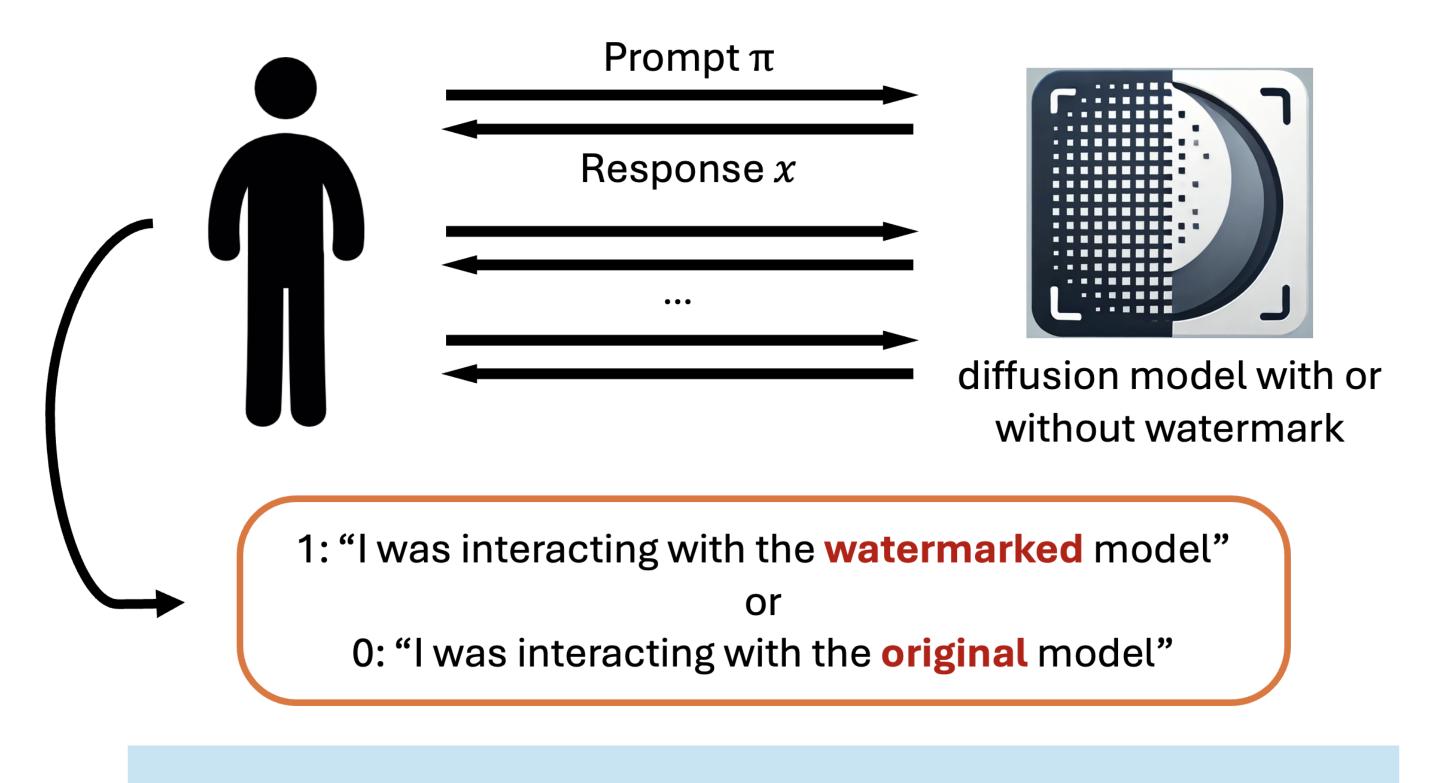**Pseudorandomness**: Codewords appear random to any efficient distinguisher -> **Undetectability**

Error correction: Decryption recovers the message even from a ciphertext subjected to a constant rate of noise -> **Robustness**

How to construct?
See paper -> "**Pseudorandom Error-Correcting Codes**" by Miranda Christ and Sam Gun (CRYPTO 2024)
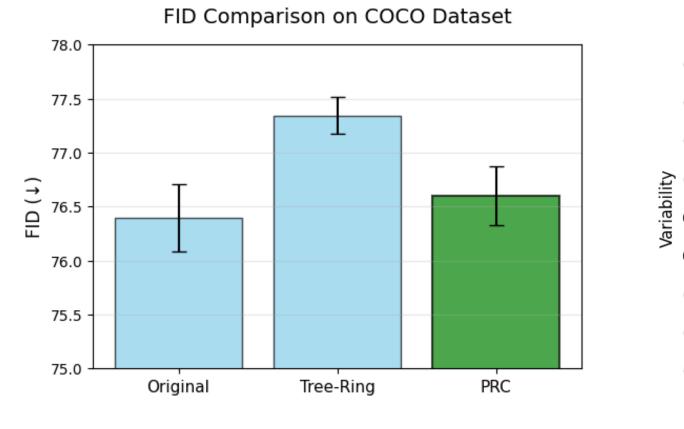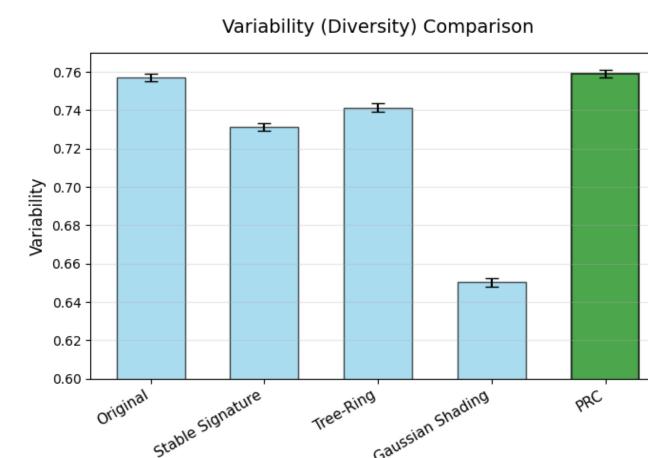We convert pseudorandom bits to pseudorandom gaussian noise

Undetectability

Prompt π
Response x
...
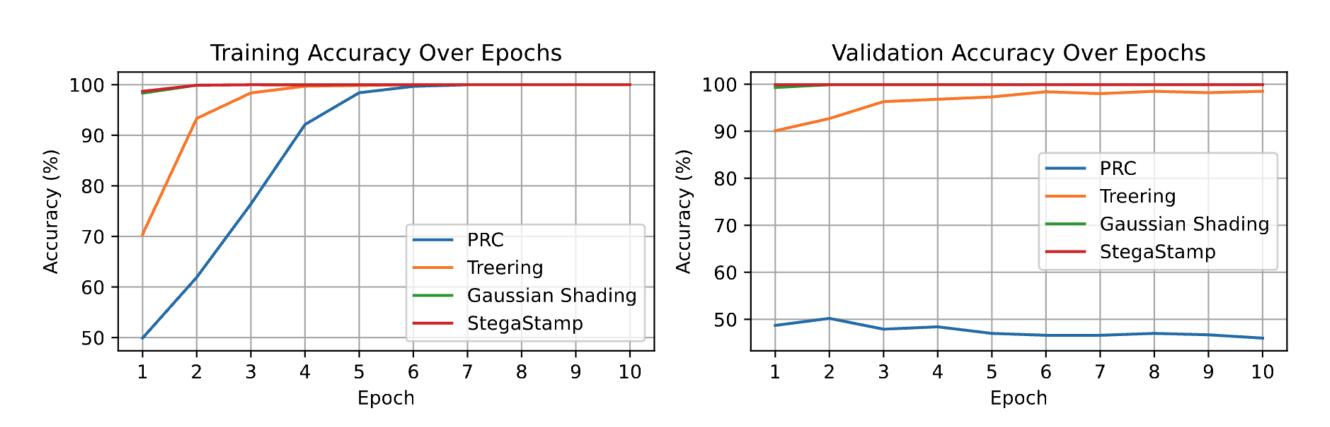
diffusion model with or without watermark

1: "I was interacting with the **watermarked** model"
or
0: "I was interacting with the **original** model"

$$\Pr[\text{Correct guess}] \leq \tfrac{1}{2} + \text{negligible}$$

Experiment Results

No quality and diversity drop

FID Comparison on COCO Dataset — Variability (Diversity) Comparison

Unlearnable

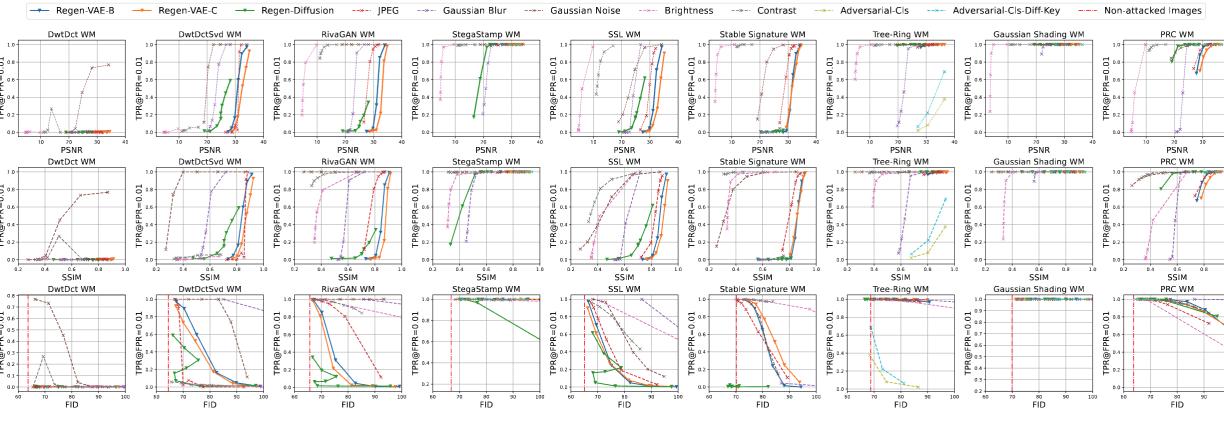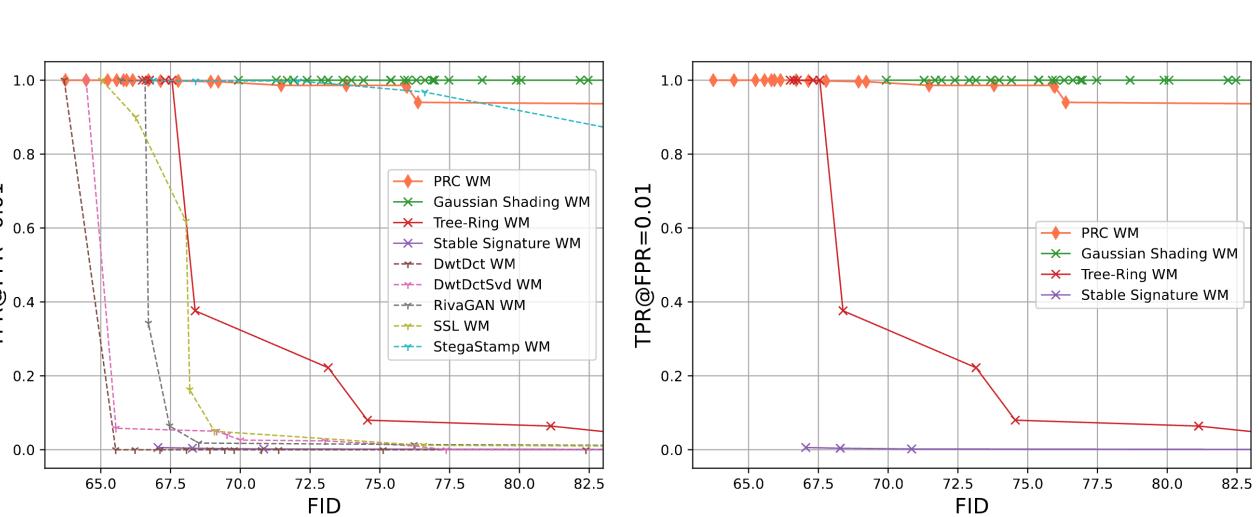Training Accuracy Over Epochs — Validation Accuracy Over Epochs

Robustness

Other good properties

Multi-bits Message
Robust for the decoder with 512-bit messages
Non-attacked images -> message capacity to at least 2500 bits

Precise false positive rate (FPR) control
Set desired upper bound on the FPR at watermark key generation
The false positive rate is rigorous, rather than empirical