Watermarking Diffusion Models Using CLWE

CLUE-Mark

Kareem Shehata,
National University of Singapore
26 August 2025

Outline

- The Problem: AI Slop
- Overview of Diffusion Models and Watermarking
- Introduction to CLWE
- CLUE-Mark Construction
- Evaluation Results
- Conclusions & Future Work

AI Slop

AI Slop

The Guardian

Int

AI-generated ‘slop’ is slowly killing the internet, so why is nobody trying to stop it?

Arwa Mahdawi

Low-quality ‘slop’ generated by AI is crowding out genuine humans across the internet, but instead of regulating it, platforms such as Facebook are positively encouraging it. Where does this end?

REUTERS INSTITUTE | UNIVERSITY OF OXFORD

Digital News

Home Reporting

Home > News

AI-generated slop is quietly conquering the internet. Is it a threat to journalism or a problem that will fix itself?

Futurism

RISING TIDE | MAR 11, 10:07 AM EDT *by* FRANK LANDYMORE

The Entire Internet Is Being Polluted by AI Slop

Close the floodgates!

AI Slop - Definition

- Machine-Generated, Low Effort
- Large Volumes
- Shared with little to no modification

WIKIPEDIA
The Free Encyclopedia

AI slop

Article Talk

From Wikipedia, the free encyclopedia

"**AI slop**", often simply "**slop**", is a term for low-quality media, including writing and images, made using generative artificial intelligence technology, characterized by an inherent lack of effort, being generated at an overwhelming volume.[1][4][5] Coined in the 2020s, the term has a pejorative connotation similar to "spam".[4]

AI slop has been variously defined as "digital clutter", "filler content [prioritizing] speed and quantity over substance and quality",[6] and "shoddy or unwanted AI content in social media, art, books and [...] search results."[7]

Jonathan Gilmore, a philosophy professor at the City University of New York, describes the material as having an "incredibly banal, realistic style" which is easy for the viewer to process.[8]

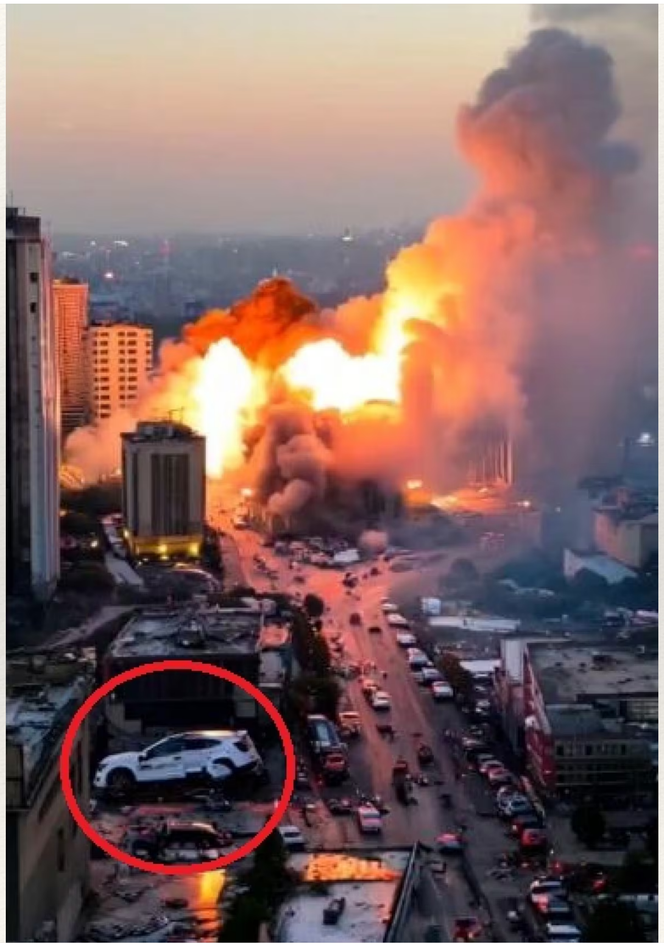
Which is AI?

Why is this a problem?

- Difficult to tell apart fact from fiction
- Overloading Effects

Why is this a problem?

- Difficult to tell apart fact from fiction
- Overloading Effects



Solution: Watermarking

NB: Verifier has no per-image info.

Prior Work (Mostly) Ad-Hoc

Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust

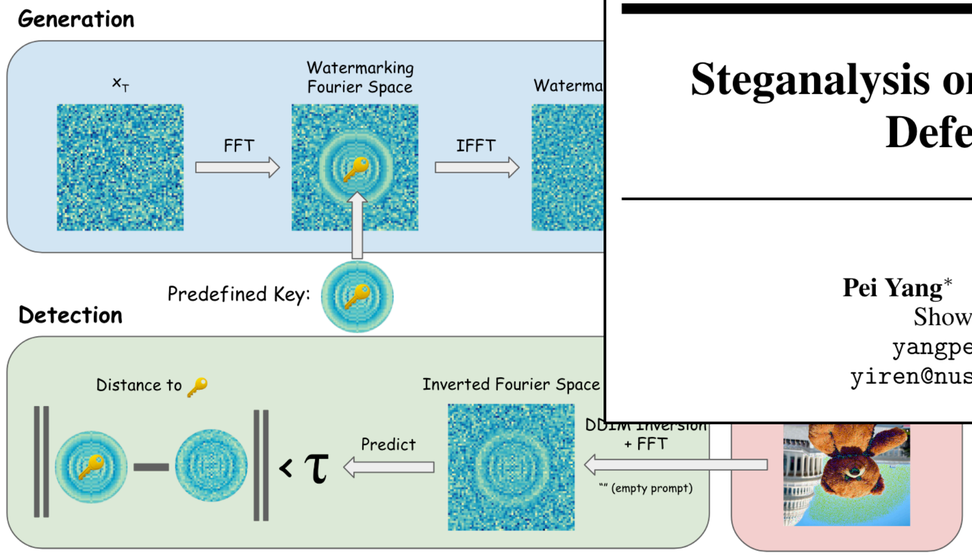


Figure 1: Pipeline for *Tree-Ring Watermarking*. A diffusion model generation is watermarked and later detected through ring-patterns in the Fourier space of the initial noise vector.

NO PROOF

Steganalysis on Digital Watermarking: Is Your Defense Truly Impervious? NO.

Pei Yang* **Hai Ci*** **Yiren Song** **Mike Zheng Shou**
Show Lab, National University of Singapore
yangpei@u.nus.edu, cihai03@gmail.com
yiren@nus.edu.sg, mike.zheng.shou@gmail.com

Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion Models

Zijin Yang[1], Kai Zeng[1], Kejiang Chen[1,*], Han Fang[2], Weiming Zhang[1], Nenghai Yu[1],
[1]Anhui Province Key Laboratory of Digital Security, University of Science and Technology of China
[2]National University of Singapore
kj@, zhangwm@, ynh@}ustc.edu.cn fanghan@nus.edu.sg

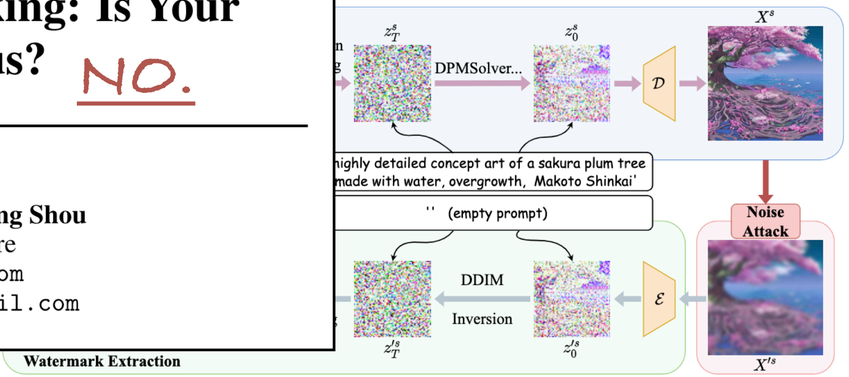


Figure 3. The framework of Gaussian Shading. We utilize a k-bit binary sequence s to represent the watermark. After diffusion and encryption, the watermark can be utilized to drive distribution-preserving sampling, followed by denoising to generate watermarked images X^s. For extraction, it is sufficient to introduce DDIM inversion and the inverse process of all the operations mentioned above.

REQUIRES NONCE

Watermarking from PRCs

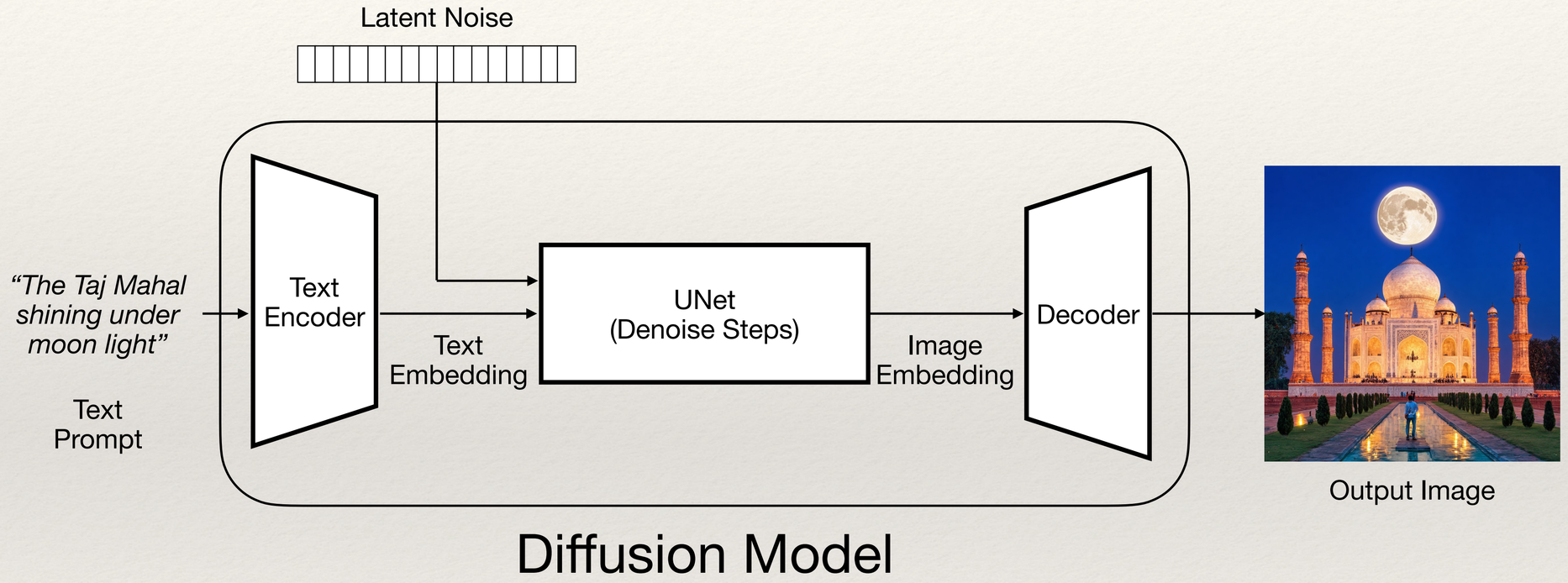
- Work concurrent to ours, uses Pseudo-Random Codes
- Similar in spirit to Gaussian Shading, but instead of using a stream cipher that requires a nonce, uses a PRC
- PRC appears random to any PPT distinguisher without the decoding key
- Parameters used do not match theoretical security
- Present our work as another option to PRCs, remains to be seen which is better in practice.

Can we inject (and recover) a watermark that is provably undetectable?

If so, how robust can we make it to perturbations while maintaining undetectability?

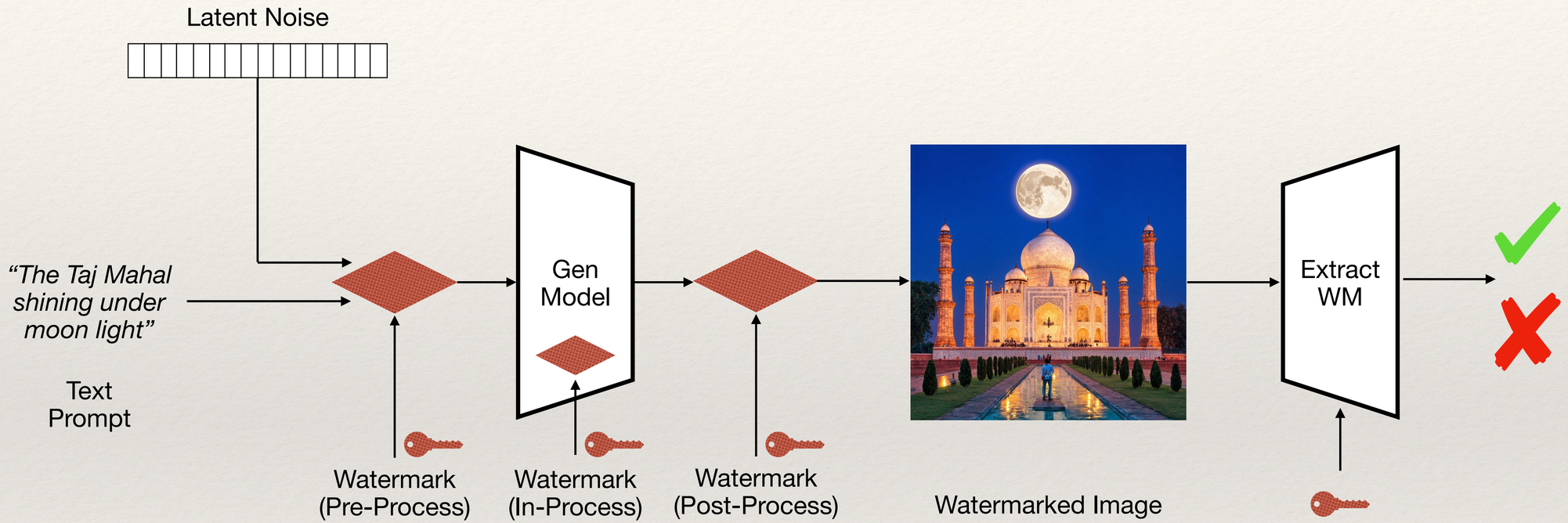
Diffusion Models & Watermarking

Diffusion Model Image Generation (Sampling)



$$\text{Sample}_\theta(\pi \in \mathcal{T}^*; \vec{z} \in \mathbb{R}^n) \rightarrow \vec{x} \in \mathbb{R}^N$$

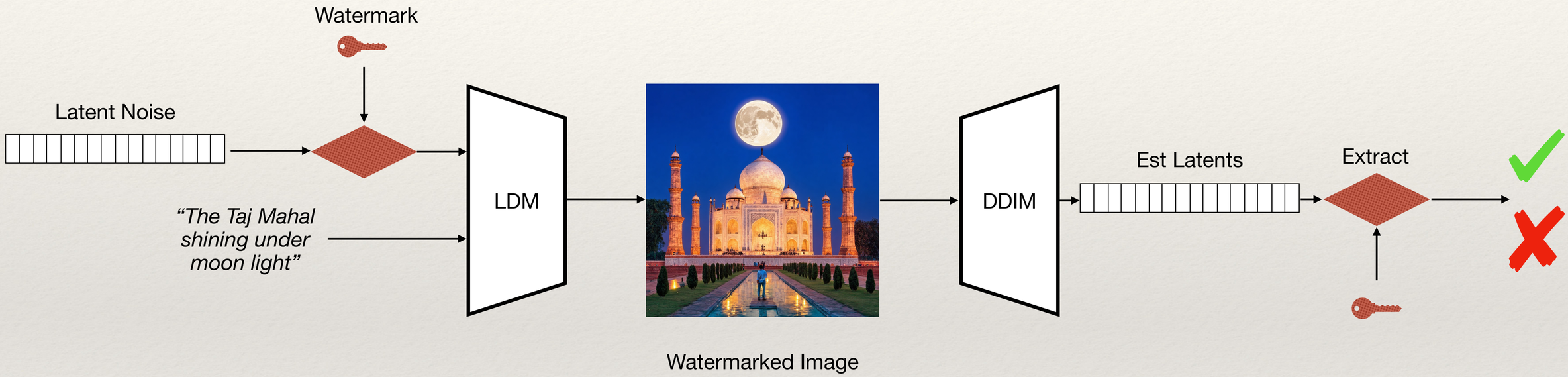
Watermarking



$$\text{Mark}_\theta(\pi \in \mathcal{T}^*; k \in \mathcal{K}) \rightarrow \vec{x} \in \mathbb{R}^N$$

$$\text{Extract}_\theta(\vec{x} \in \mathbb{R}^N; k \in \mathcal{K}) \rightarrow \{0,1\}$$

Watermarking Diffusion Models



NB: DDIM not given original prompt or embedding vector, and produces an estimate subject to noise.

Watermarking Completeness & Soundness

Completeness

Definition 2.3 (Watermark Completeness)**.** A watermarking scheme $(\text{Setup}, \text{Mark}, \text{Extract})$ is δ-complete if for all $\pi \in \mathcal{T}^*$:

$$\Pr\left[\text{Extract}_\theta(\text{Mark}_\theta(\pi; k); k) = 1 | k \leftarrow \text{Setup}_\theta(1^\lambda)\right] \geq 1 - \delta$$

Soundness

Definition 2.4 (Watermark Soundness)**.** A watermarking scheme $(\text{Setup}, \text{Mark}, \text{Extract})$ is δ-sound if for all $\pi \in \mathcal{T}^*$:

$$\Pr[\text{Extract}_\theta(\text{Sample}_\theta(\mathbf{z}; \pi); k) = 1 | \\ k \leftarrow \text{Setup}_\theta(1^\lambda), \mathbf{z} \leftarrow \mathcal{N}(\mathbf{0}, \mathbf{I})] \leq \delta$$

Essentially TPR

(True Positive Rate)

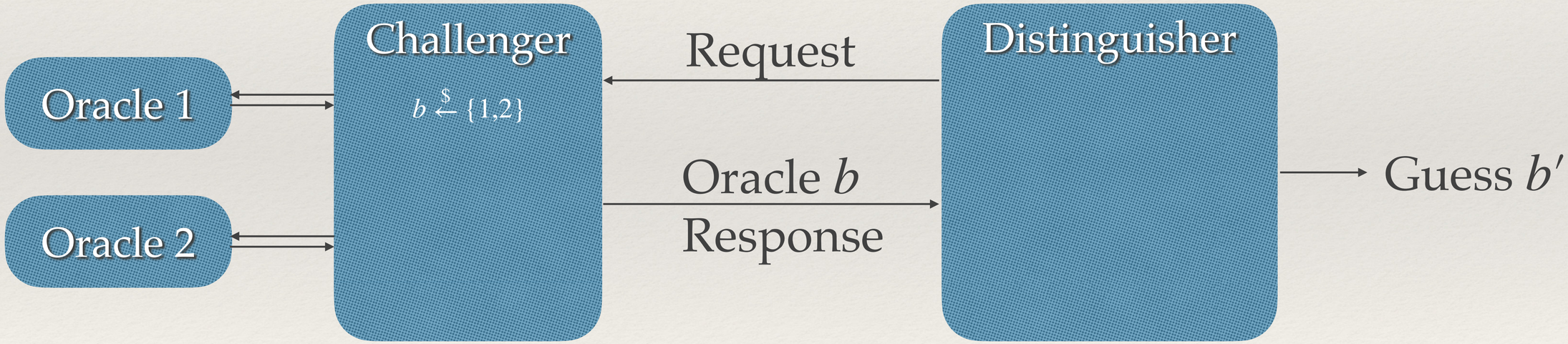
Essentially FPR

(False Positive Rate)

Distinguishing

Definition 2.5 (Advantage of a Distinguisher)**.** Let $D^{\mathcal{O}} \rightarrow \{0,1\}$ be a distinguisher with access to oracle $\mathcal{O}$. Define the advantage of D in distinguishing $\mathcal{O}_1$ and $\mathcal{O}_2$ as:

$$Adv_{D,\mathcal{O}_1,\mathcal{O}_2} = \left|\Pr\left[D^{\mathcal{O}_1}(1^\lambda) = 1\right] - \Pr\left[D^{\mathcal{O}_2}(1^\lambda) = 1\right]\right|$$



Undetectability

- Essentially: WM and no-WM image indistinguishable (to PPT distinguisher)
- Why does this matter?
  - Quality of images, both for users and for all downstream processes
  - Ensures WM security

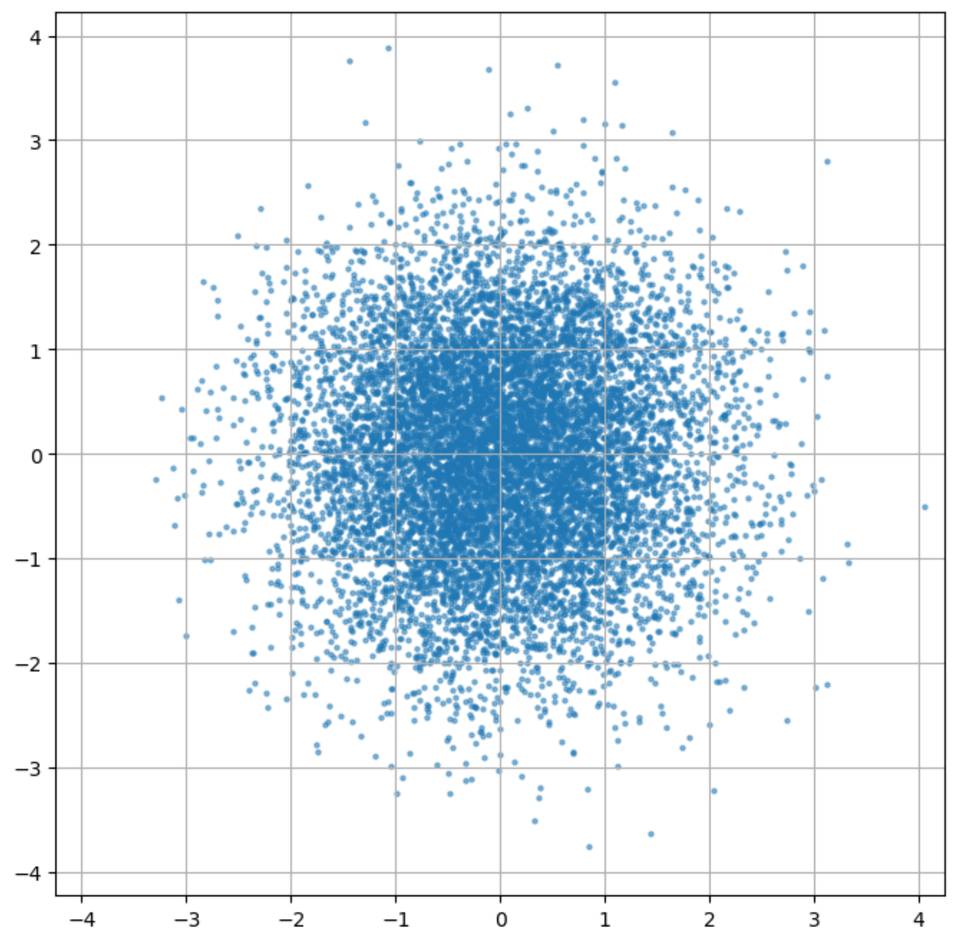
Definition 2.6 (Provable Undetectability)**.** Let $\mathcal{M}_\theta$ be an oracle that on input $\pi \in \mathcal{T}^*$ samples $\mathbf{z} \leftarrow \mathcal{N}(\mathbf{0}, \mathbf{I})$ and returns $\text{Sample}_\theta(\pi; \mathbf{z})$. Let $\mathcal{W}_{\theta,k}$ be an oracle that on input $\pi \in \mathcal{T}^*$ returns $\text{Mark}_\theta(\pi; k)$ for a fixed key $k \leftarrow \text{Setup}_\theta(1^\lambda)$. A watermarking scheme (Setup, Mark, Extract) is undetectable if for all *probabilistic polynomial-time* PPT distinguishers D, $Adv_{D,\mathcal{M}_\theta,\mathcal{W}_{\theta,k}} \leq negl(\lambda)$.

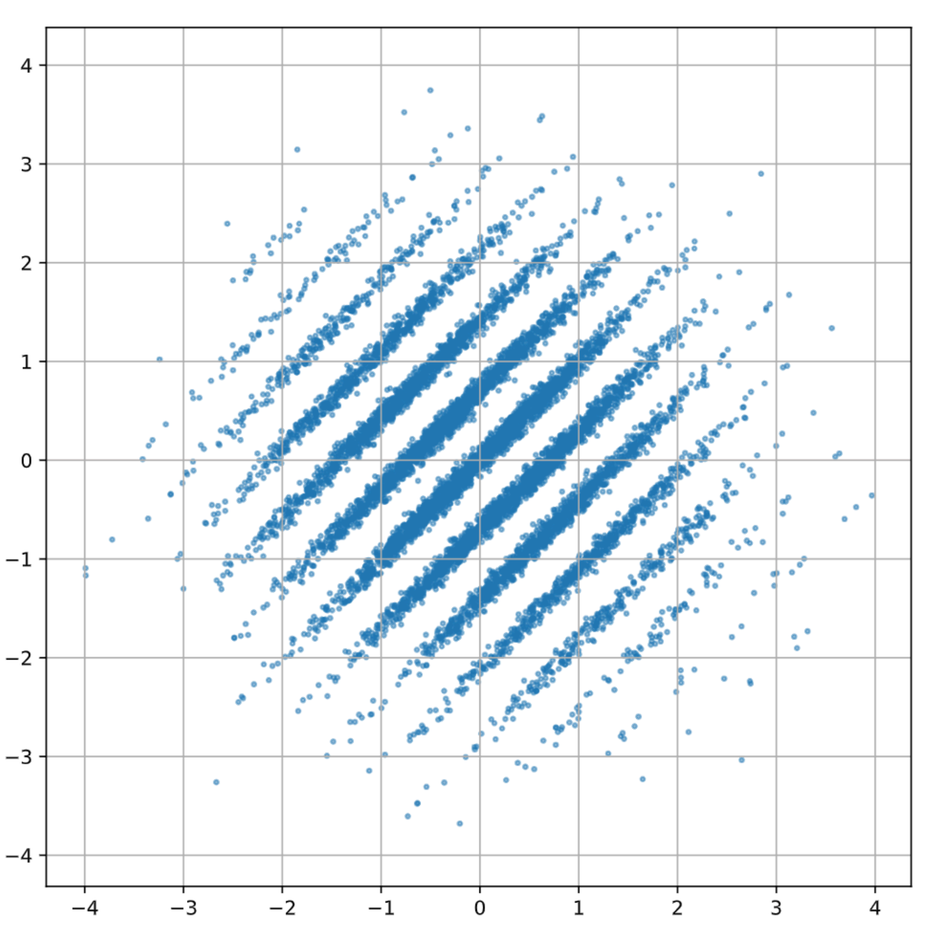
Robustness

- Desired property: any “benign” perturbation of the image leaves WM intact.
- Related: unforgeable, i.e. can’t perturb an image to add a WM.
- Shown to be impossible: any WM that is sound, complete, and quality preserving cannot be perfectly robust.
- Our goal: undetectability, then as robust as we can.

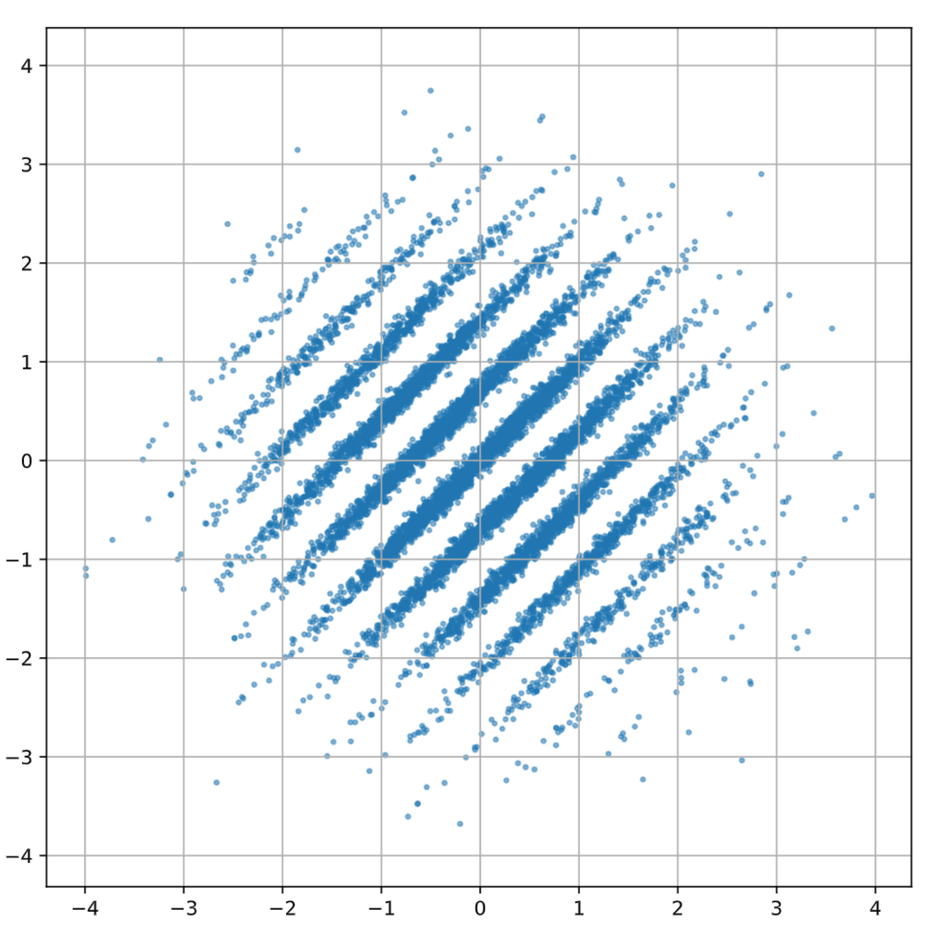
CLWE

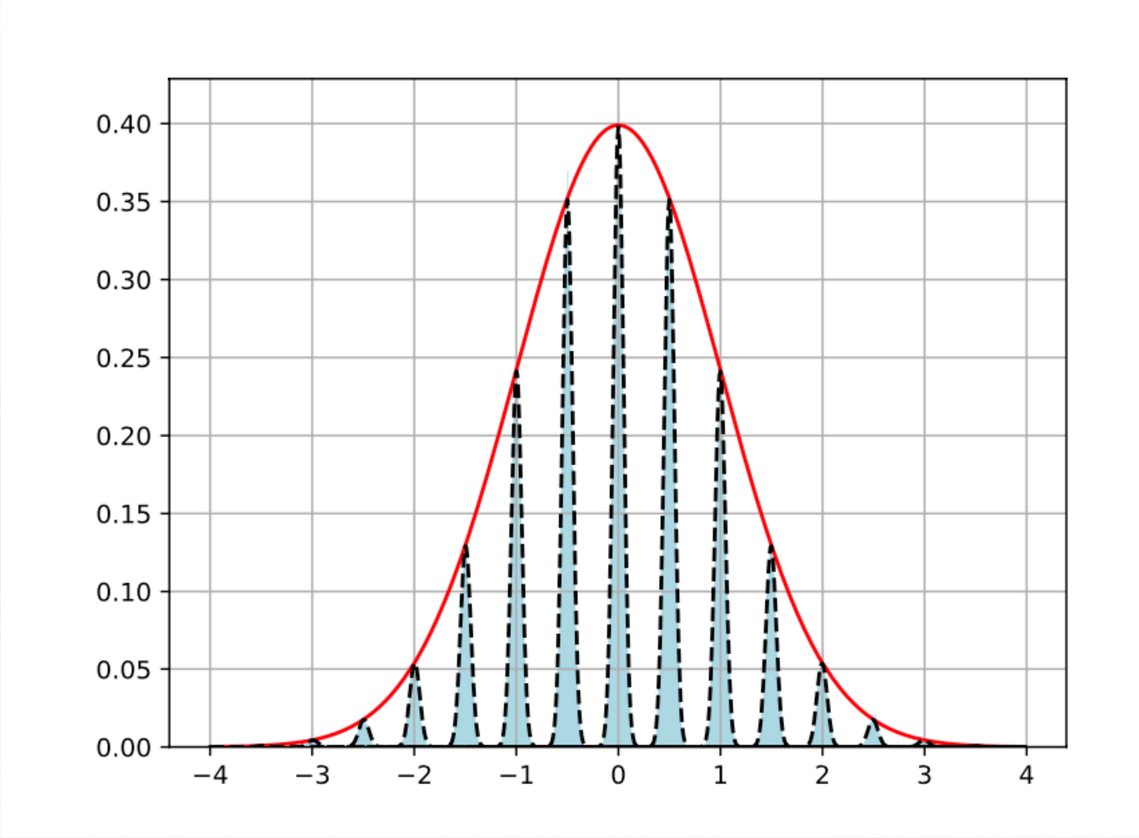
hCLWE Example





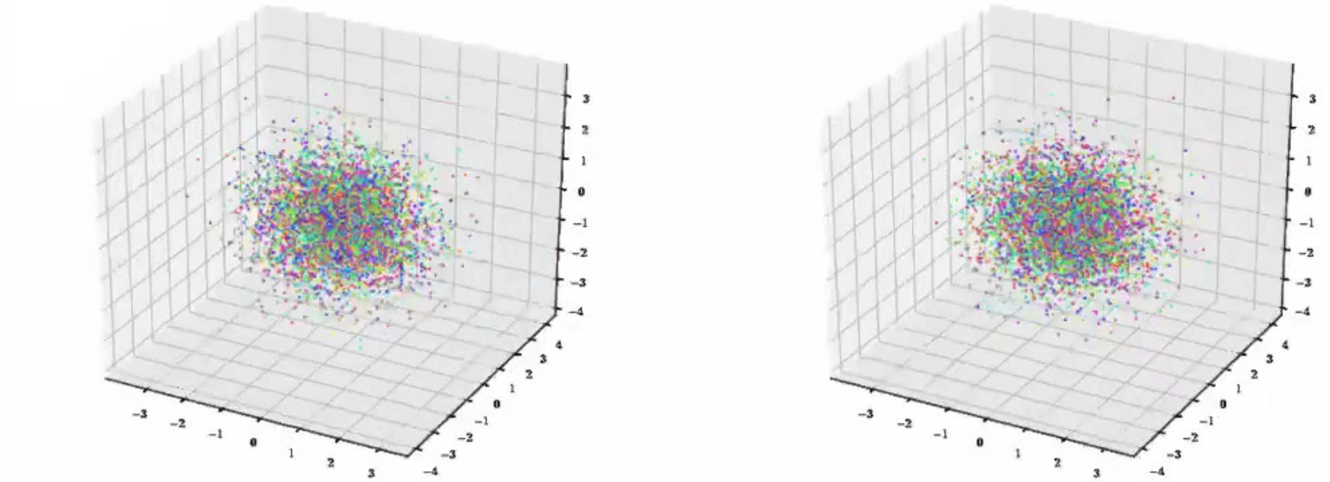
hCLWE Example





Hardness of (h)CLWE: proof overview

We prove a stronger hardness result, for a relaxed problem: (*inhomogeneous*) CLWE.



0:56

CLWE Equations

Parameters: $\gamma \in \mathbb{R}, \beta \in (0,1)$

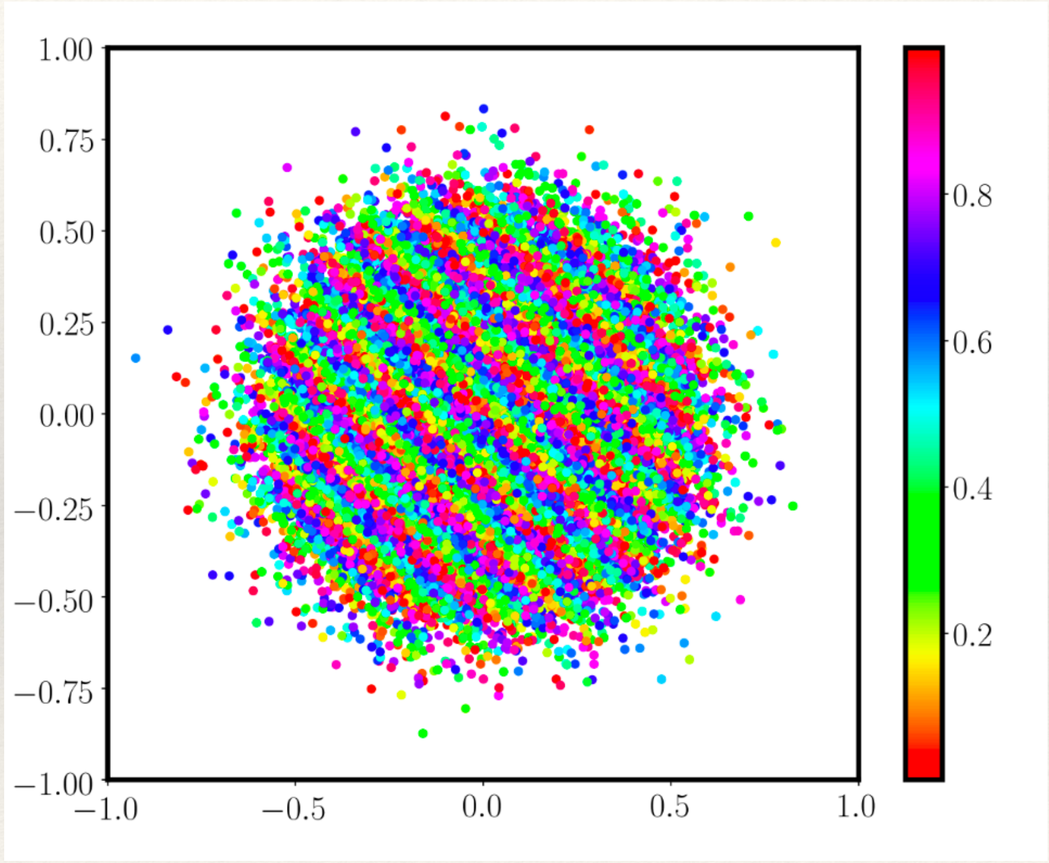
Secret direction: $\hat{w} \in \mathbb{R}^n, \|\hat{w}\| = 1$

Let $\vec{y}_i \leftarrow \mathcal{N}(0,1)^n$, $\epsilon \leftarrow \mathcal{N}(0,\beta)$

$$z_i = \gamma \langle \vec{y}_i, \hat{w} \rangle + \epsilon \mod 1$$

<u>Basic Hardness Idea:</u> $\{z_i\}_{i \in [m]} \approx_c U([0,1))^m$

To get hCLWE: rejection sample with probability proportional to z_i.



CLWE vs LWE

CLWE

Parameters: $\gamma \in \mathbb{R}, \beta \in (0,1)$

$\hat{w} \leftarrow S_1^n$ (i.e. $\hat{w} \in \mathbb{R}^n, \|\hat{w}\| = 1$)

$\vec{y}_i \leftarrow \mathcal{N}(0,1)^n$

$\epsilon \leftarrow \mathcal{N}(0,\beta)$

$z_i \leftarrow \gamma\langle \vec{y}_i, \hat{w} \rangle + \epsilon \mod 1$

$\{\vec{y}_i, z_i\}_{i\in[m]} \approx_c \{\vec{y}_i, U([0,1))\}_{i\in[m]}$

LWE

Parameters: $q \in \mathbb{N}, \sigma \in \mathbb{R}$

$\vec{s} \leftarrow U(\mathbb{Z}_q^n)$

$\vec{a}_i \leftarrow U(\mathbb{Z}_q^n)$

$\epsilon \leftarrow D_{\mathbb{Z},\sigma}$

$b_i \leftarrow \langle \vec{a}_i, \vec{s} \rangle + \epsilon \mod q$

$\{\vec{a}_i, b_i\}_{i\in[m]} \approx_c \{\vec{a}_i, U(\mathbb{Z}_q)\}_{i\in[m]}$

hCLWE Distribution

hCLWE distribution with parameters γ, β defined as having PDF prop to:

$$\rho(\vec{y}) \cdot \sum_{k \in \mathbb{Z}} \rho_\beta(k - \gamma\langle \overrightarrow{w}, \vec{y} \rangle)$$

Which is equivalent to:

$$\rho(\vec{y}^{\perp}) \cdot \sum_{z \in \mathbb{Z}} \rho_{\gamma'}(k) \rho_{\beta/\gamma'}\left(\tilde{y} - k\gamma/\gamma'^2\right)$$

Definitions:

$$\rho_s(\vec{x}) = e^{\frac{-\pi\|\vec{x}\|^2}{s^2}}$$

$$\tilde{y} = \langle \vec{y}, \overrightarrow{w} \rangle$$

$$\vec{y}^{\perp} = \vec{y} - \tilde{y} \cdot \overrightarrow{w}$$

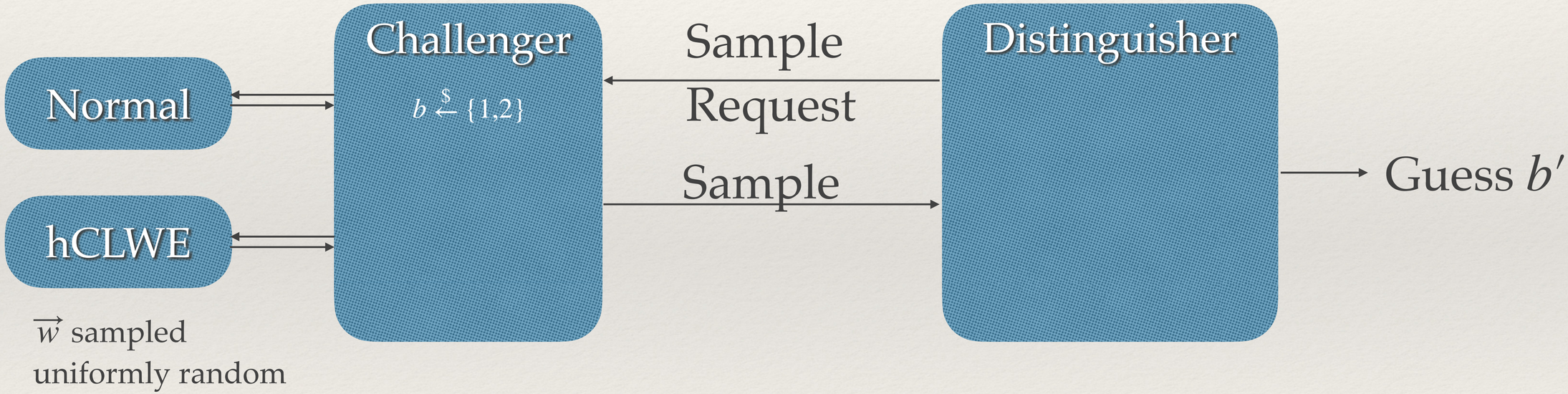
$$\gamma' = \sqrt{\beta^2 + \gamma^2}$$

hCLWE Distribution Interpretation

hCLWE distribution with parameters γ, β can be described as:

- Standard normal in directions orthogonal to $\overrightarrow{w}$
- Periodic in direction of $\overrightarrow{w}$ with separation $\gamma/(\beta^2+\gamma^2)$ and width $\beta/\sqrt{\beta^2+\gamma^2}$
- If $\beta \ll \gamma$, then separation $\approx \gamma$ and width $\approx \beta/\gamma$

Distinguishing hCLWE



Decision hCLWE

Definition 4.2 (Decision hCLWE Problem)**.** Let $\mathbf{w} \in \mathbb{R}^n$ be a unit vector drawn uniformly at random, and $\beta, \gamma > 0$ be parameters. Let $\mathcal{N}$ be an oracle that responds to queries with samples from $\mathcal{N}(0, I_n)$, and let $\mathcal{H}_{\mathbf{w},\beta,\gamma}$ be an oracle that responds with samples from $H_{\mathbf{w},\beta,\gamma}$. The decision hCLWE problem is to distinguish $\mathcal{N}$ and $\mathcal{H}_{\mathbf{w},\beta,\gamma}$. That is for distinguisher D, we say that D solves the decision hCLWE problem with advantage $Adv_{D,\mathcal{N},\mathcal{H}_{\mathbf{w},\beta,\gamma}}$.

- Written more simply: given a polynomial number of samples, distinguish between the hCLWE distribution and standard normal.
- Reduction from DGS and standard LWE to this problem.

CLUE-Mark

CLUE-Mark: Watermarks from CLWE

Watermark Injection

- Instead of sampling latent noise from standard normal, sample from hCLWE.
- Otherwise, run diffusion model as normal.
- Secret direction $\hat{w}$ is watermark key.

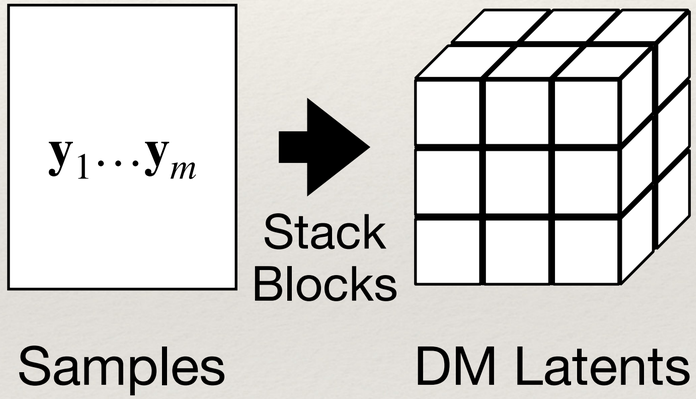
Watermark Recovery:

- Use DDIM to invert model and obtain estimate of latent vector.
- Use $\hat{w}$ to check latent vector for hCLWE.
- If present -> watermarked image.

… if only life were so simple.

Problem 1: Blocking

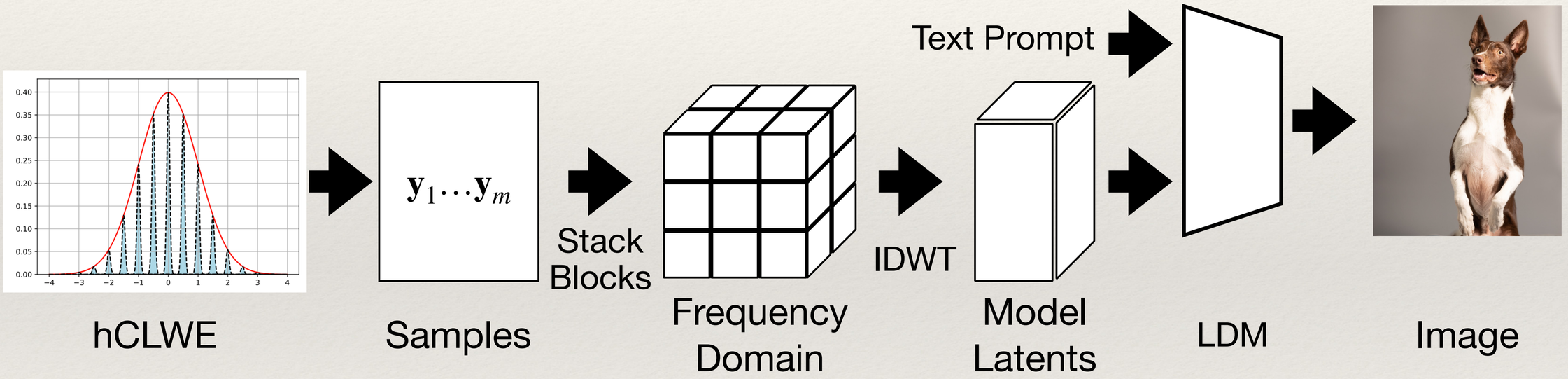
- DM expects latents to be single tensor, e.g. $4 \times 64 \times 64 \times \mathbb{R}$
- hCLWE expects a number of samples, each of which have some dimension
- Solution: split the latents into blocks, large enough to be undetectable, small enough that we can lose some due to corruption and still recover WM.



Problem 2: Robustness / Frequency Domain

- Sampling and Inverse LDM processes add noise; image perturbations
- Past research on watermarks shows more robust in frequency domain
- Use Discrete Wavelet Transform (DWT)
- Inject WM in frequency domain, Inverse DWT to get latents
- To recover WM, apply DWT to estimated latents

Watermarked Image Process



Watermark Recovery

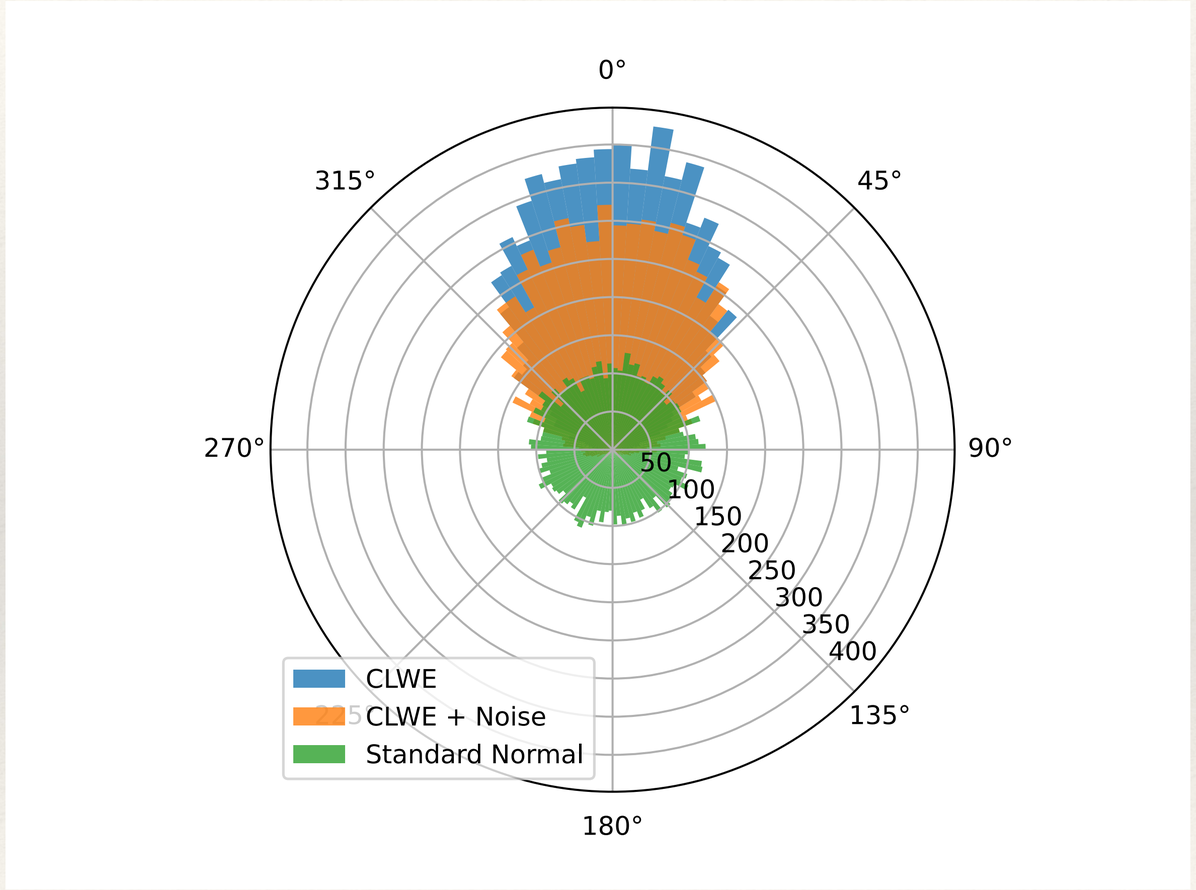
- Given estimated latents and $\hat{w}$, how do we determined watermark presence?
- Check z_i values for samples:

$$z_i = \gamma \langle \vec{y}_i, \overrightarrow{w} \rangle \mod 1$$

- What should we expect from these values?

Checking z_i scores

- In the watermark case, scores should be concentrated close to 0.
- In the no-watermark case, scores should be close to uniform.
- Confirmed via statistical simulation.

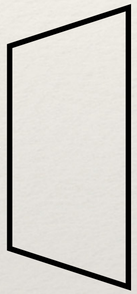


Rayleigh Test

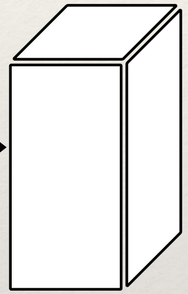
- Given a set of z_i samples how can you (optimally) distinguish between a uniform and unimodal distribution?
- Answer: Rayleigh's Test.
- Standard test in circular statistics that gives p-value for probability that samples came from a uniform distribution.

Watermark Recovery Process

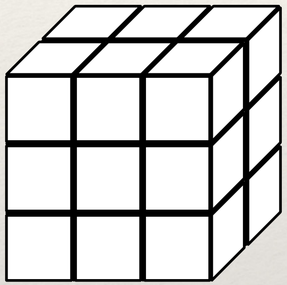
Image



DDIM



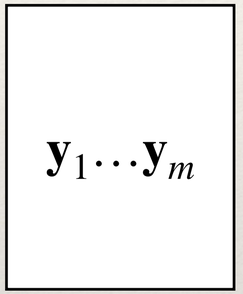
Model Latents



DWT

Frequency Domain

Unstack Blocks

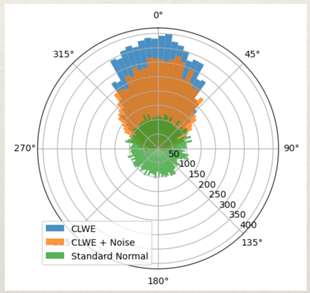


Samples

$$z_i = \gamma \langle \mathbf{y}_i, \mathbf{w} \rangle \bmod 1$$

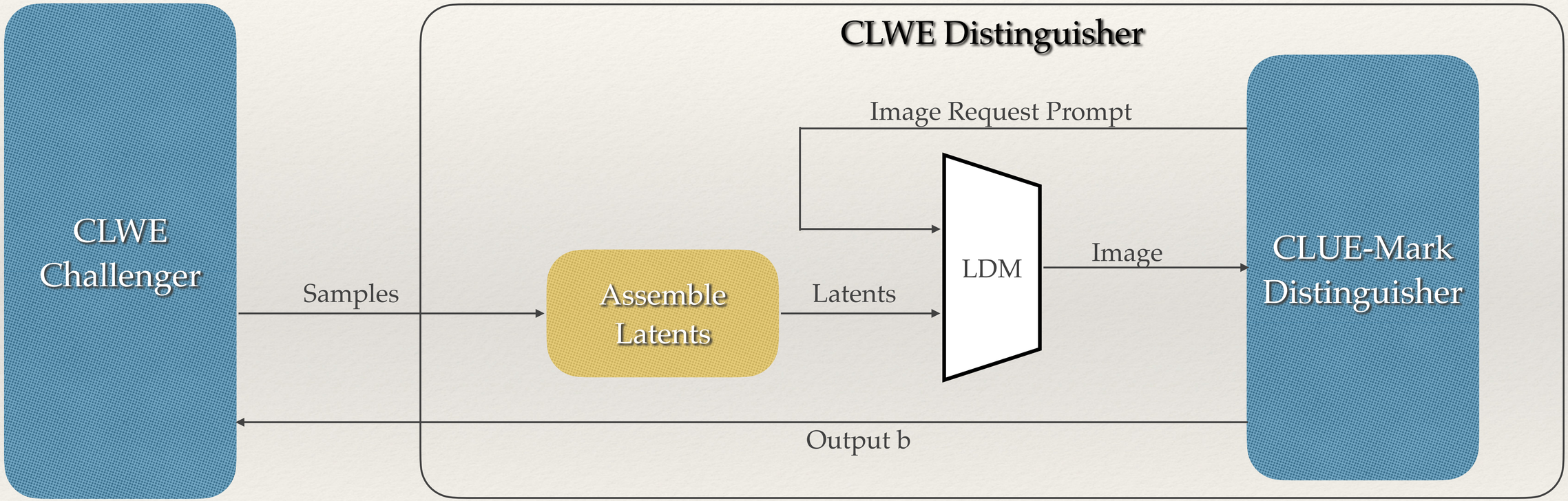
$z_1, \ldots, z_m$

hCLWE Scores



Rayleigh p-Value

Undetectability Proof



Evaluation

CLWE Parameters

- CLWE Paper suggest $\gamma > 2\sqrt{n}, \beta > 0$ (with β/γ poly bounded)
- GVV paper suggests $\gamma = \tilde{\Omega}(\sqrt{n}), \beta = 1/poly(n)$
- If LWE is sub-exp, then $\gamma = \left(\log n\right)^{\frac{1}{2}+\delta} \log\log n$

Covariance Attack

Algorithm 2: Covariance method to determine if samples are from a CLWE distribution or standard normal.

Input: Samples $A \in \mathbb{R}^{m \times n}$

Output: Covariance score

$\Sigma_m \leftarrow \frac{1}{2\pi m} A^T A$;

$\mu_1, \ldots, \mu_n \leftarrow \text{eigenvalues}(\Sigma_m)$;

return $\max_{i \in [n]} \left| \mu_i - \frac{1}{2\pi} \right|$;

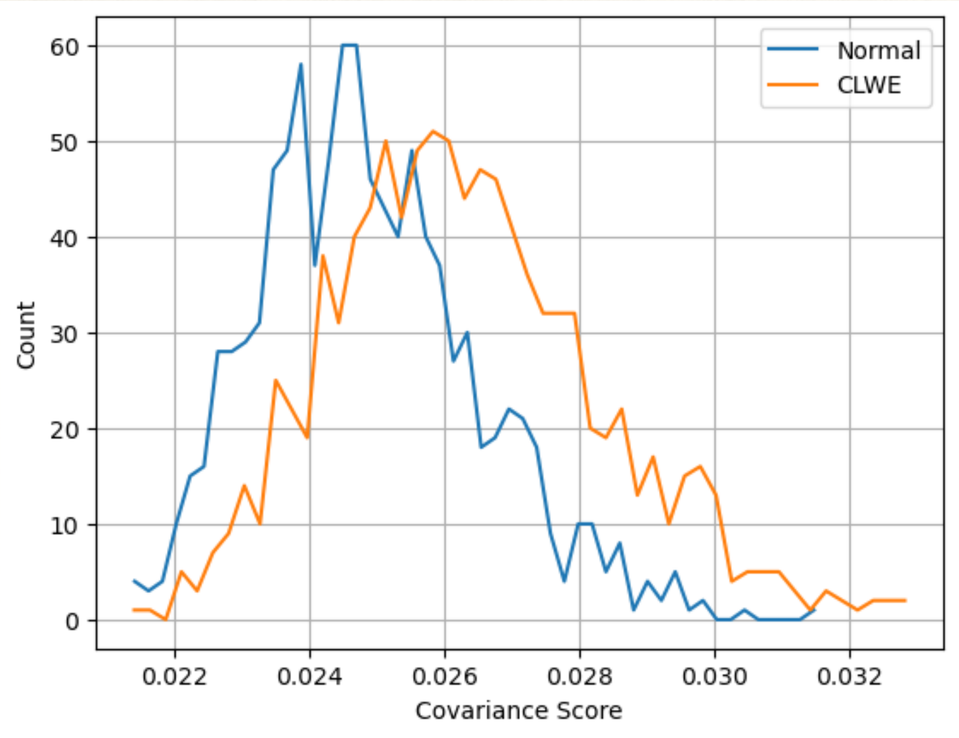
Requires $O\left(e^{\gamma^2}\right)$ samples for convergence!

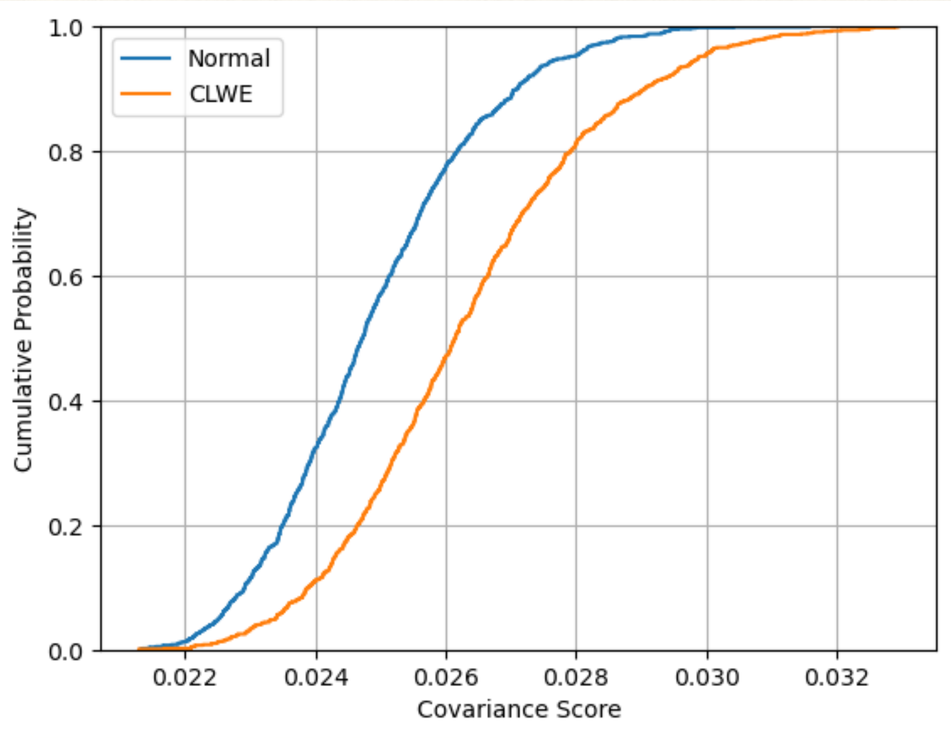
$$\text{Threshold} = \gamma^2 e^{-\pi(\beta^2+\gamma^2)}$$

Adaptive Covariance Attack

- What if you have fewer than exponential samples? Can you still distinguish CLWE and normal? Yes.
- Simple approach: estimate the sample covariance of normal distribution.
- Empirically: take a bunch of normal samples and a bunch of CLWE samples, compute the covariance scores.

Statistical Experiment

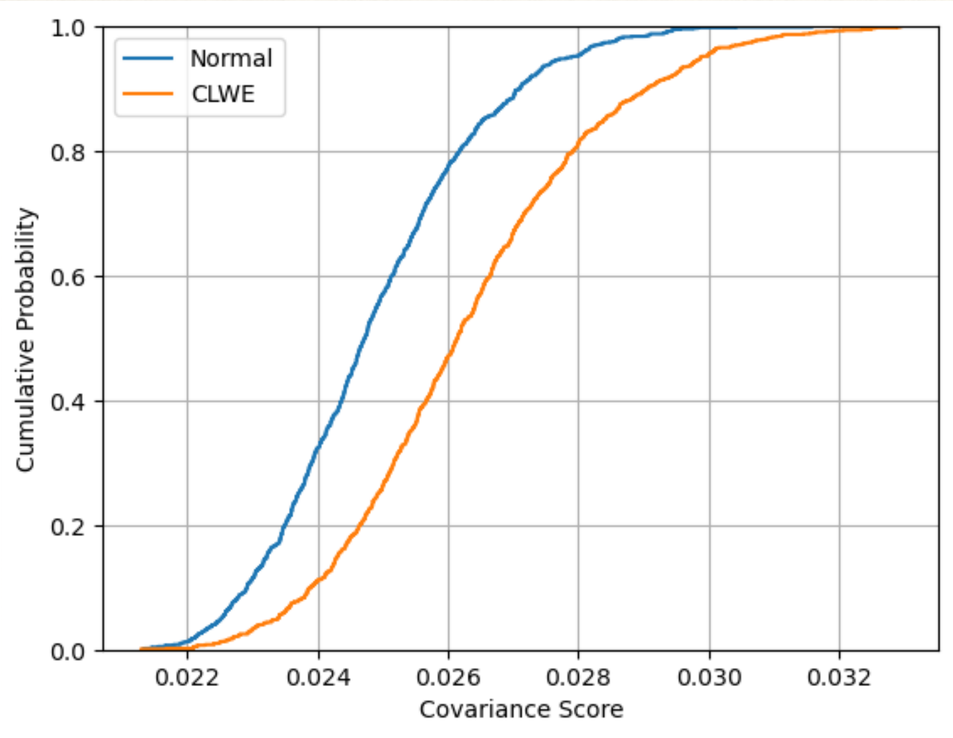




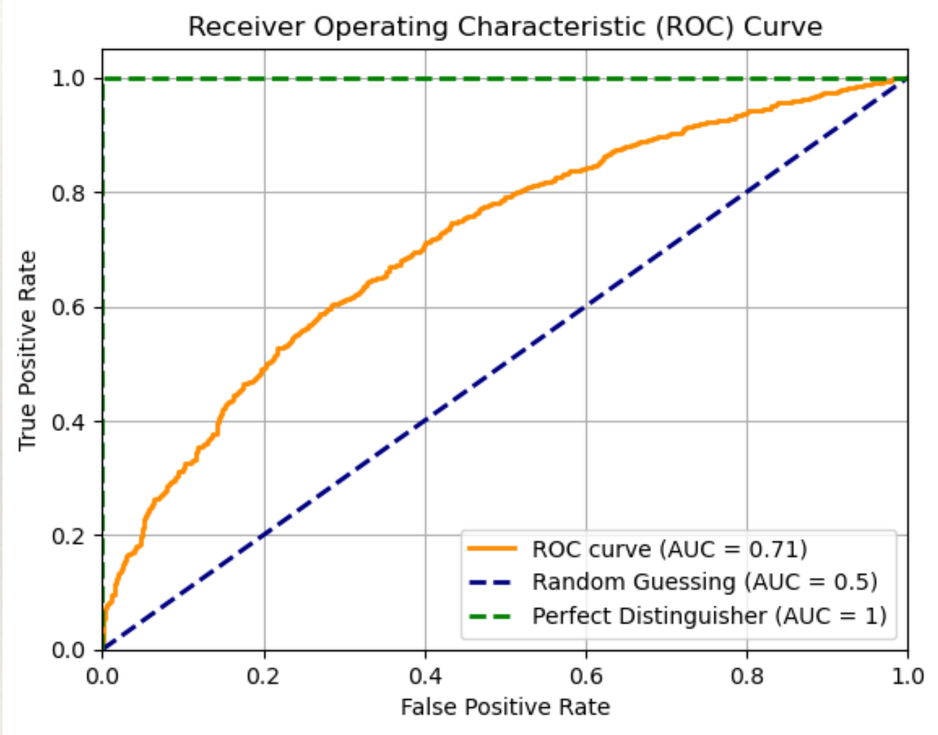
Statistical Simulation with 5000 samples, repeated 1000 times.

$n = 32,\ m = 5000,\ \gamma = 1,\ \beta = 0.001$

Receiver Operating Characteristic (RoC) Curve

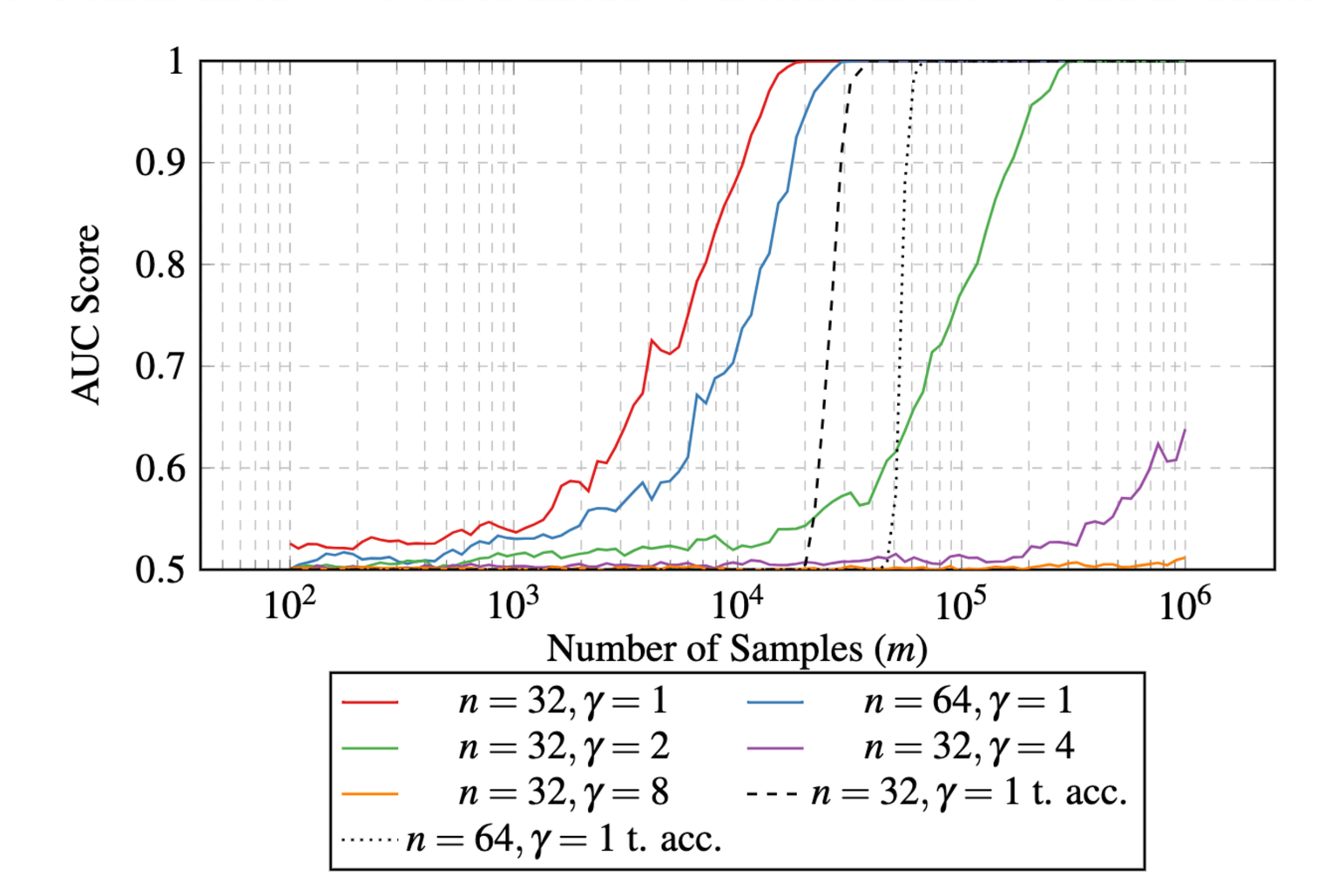


TPR = Completeness



FPR = Soundness

Adaptive Covariance Attack



Practical Parameters

- Covariance attack: $m < 10{,}000$, $\gamma \geq 2$,$n \geq 32$
- Our choice:
  - $n = 2 \times 4 \times 4 = 32$
  - $\gamma = 2$, $\beta = 0.001$

Evaluation Results

Recovery AUC > 0.99 ✓

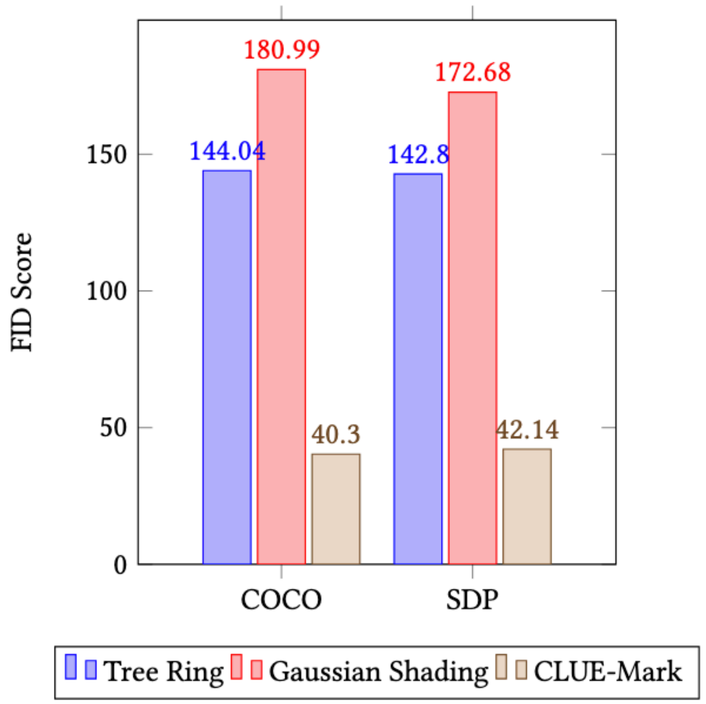


Figure 9: Image quality as measured by FID to unwatermarked images on two standard datasets.

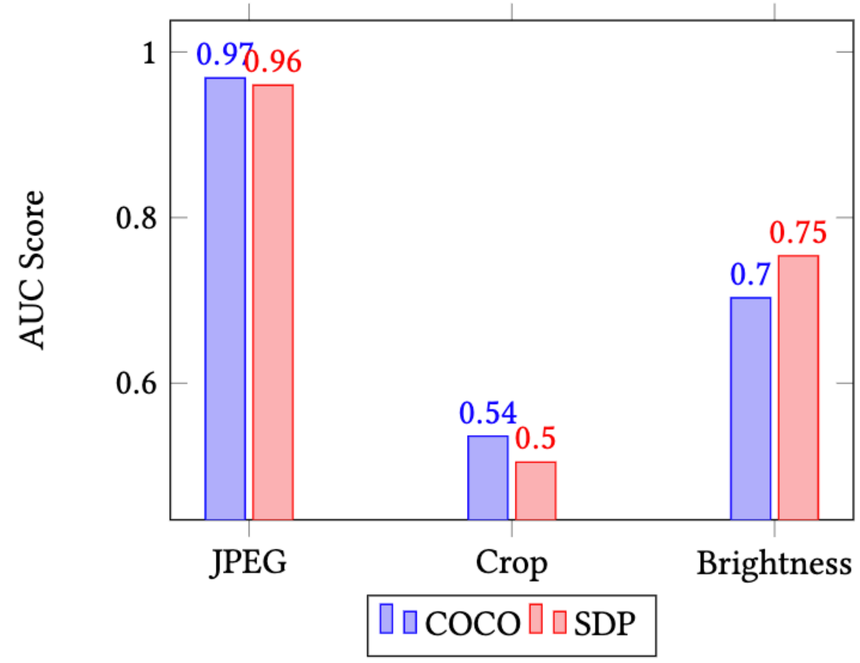
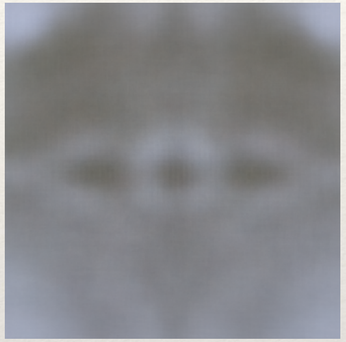
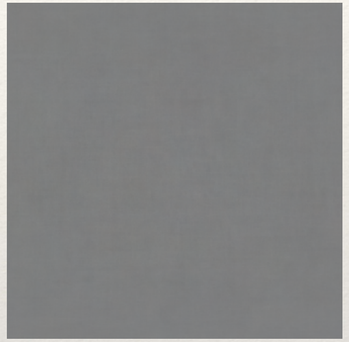


Figure 12: Robustness to common perturbations as measured by RoC AUC for CLUE-Mark.

	Tree Ring	Gaussian Shading	CLUE-Mark	No Watermark
Original				
Steganographic Average				
Average Removed				
Watermark Detected?				

Concluding Remarks

Conclusions

- CLUE-Mark Leverages CLWE to inject watermarks into diffusion model generated images that can be proven undetectable.
- Can be used in practice to recover watermarks, with parameters that provide practical security.
- Much better image quality and beats steganographic attacks, but much less robust than prior work.

Future Work

- Image Watermarks:
  - Better practical parameters
  - Better inverse LDM => better watermark recovery
  - Improve robustness of CLUE-Mark
- Video Watermarks
- Apply to other scenarios than Diffusion Models, e.g. post-process ?

Thanks!

- Thanks to my collaborator Aashish Kolluri, and supervisor Prateek Saxena, as well as Ivica Nikolic, Wenjie Qu, and Rohit Chatterjee for their helpful feedback on previous revisions.

- Questions?

Slides, Paper, Code and more:

https://maps.kisp-lab.org/cluemark/