Secure Seed-Based Multi-bit Watermarking for Diffusion Models from First Principles

Enoal Gesny
Inria
Rennes, France
enoal.gesny@inria.fr

Eva Giboulot
Inria
Rennes, France

Abstract

The rapid emergence of generative image models has led to the development of specialized watermarking techniques, particularly in-generation methods such as seed-based embedding. However, current evaluations in this area remain largely empirical, making them heavily reliant on the specific model architectures used for generation and inversion. This prevents any clear conclusion on the performance of any method, especially regarding security, for which a rigorous definition is lacking. Against this approach, we argue that the effectiveness of a watermarking scheme should be established purely through a thorough theoretical analysis. This is enabled by decoupling the model-dependent part from the actual decision mechanism of the watermarking system. Using this decoupling, we introduce a formal evaluation framework based on security, robustness, and fidelity. This allows precise comparisons between watermarking systems through a characteristic surface representing the trade-off between these three quantities, independent of any generative model. Based on this framework, we propose SSB , a novel watermarking method that generalizes previous seed-based methods by allowing it to reach any security-robustness-fidelity regime on its characteristic surface. This work opens the door to the design of modern watermarking systems with theoretical guarantees that do not necessitate any costly empirical evaluations.

1 Introduction

Watermarking involves embedding an invisible signal within the content, which can be retrieved by a detector using a key. This signal is classified as either multi-bit, in which the aim is to recover a binary message, or zero-bit, in which case the detector only decides if the content is watermarked or not, with guarantees in terms of probability of false-alarm. As a specialized data-hiding approach, the goal of a watermarking system is to embed a signal robust to content alterations while preserving content quality with guarantees of security under a given threat model.

Traditional (post-hoc) watermarking methods [16, 17, 6] embed the signal directly into the image, which makes them applicable only to private APIs. To address the challenges posed by image generation, novel approaches propose to embed the watermark during the generation process. This can be performed during the diffusion [19], or through a fine-tuned VAE's decoder [15].

A third approach, which we call **seed-based**, modifies the seed distribution to incorporate the watermark and subsequently inverts the diffusion process to retrieve it. Interestingly, this approach is also not amenable to open-source models since neither the diffusion model nor the VAE can be modified to enforce the use of a given seed distribution. However, it possesses the unique property that no distortion is introduced in an already existing latent or image: only the seed distribution is distorted. The promise is thus to allow extreme robustness by finding a distribution that leads to images with similar quality despite being potentially far from the original Gaussian distribution of the seed.

The approach was first introduced in the zero-bit setting by Tree-Rings [35], which demonstrated that a large distortion of the seed distribution does not necessarily lead to low-quality images. Indeed, Tree-Rings is empirically robust to every valuemetric image processing such as JPEG compression. Robustness to geometric operations is currently an open problem due to the fact that the latent space depends on the size of the input image. Note that the original scheme contained numerous mistakes which were subsequently fixed in [11] and [19][Appendix I].

Gaussian-Shading [38] translated the idea to the multi-bit setting. It importantly introduced the idea of using a cryptographic primitive controlled by a secret key to generate the latent seed. This has two major benefits. First, it allows hiding the content of the embedded message. Secondly, the resulting scheme is distortion-free: the original seed distribution is preserved *on average with respect to the key*. As such, if each image generated by Gaussian-Shading was produced with different messages and/or keys, the image distribution should be indistinguishable from the non-watermarked model.

All these properties: high image quality, high empirical robustness, and cryptographic security, can be enticing to the practitioner, but we argue that they are fundamentally flawed in their definitions.

Quality and empirical robustness are fundamentally tied to the generative model; this precludes any strong conclusion regarding the performance of a given watermarking scheme. In other words, only propositions of the type "*Watermarking system A is more robust/qualitative than Watermarking system B against transformation T for diffusion model M with VAE V and scheduler S"* can be made. The practitioner would want to make propositions of the type *"Watermarking system A is more robust/qualitative than Watermarking system B"*, whatever the model or operation used. In Section 2, we propose to forego the concept of quality in favor of the concept of fidelity. Similarly, we construct a refined definition of robustness independent of the generative model, based on traditional watermarking approaches. This results in a framework that permits such a rigorous ranking of watermarking systems purely on theoretical grounds.

The question of watermarking security and cryptographic security is a lot more nuanced. It is highly dependent on the threat model agreed upon by practitioners. Nevertheless, we argue in Section 2 that they are very different notions, following the consensus that emerged for classical watermarking methods [13, 3]. One can indeed have both, yet cryptographic security alone does not prevent some powerful spoofing attacks from being performed on the watermarking system at basically no cost to the attacker.

After this work, which is meant to clarify the three main desirable properties for seed-based watermarking, we arrive at three main contributions. First, we propose a methodological framework for comparing seed-based watermarking algorithms, focusing on the trade-offs between security, capacity, and fidelity independently of the diffusion model used. Then, we design a new seed-based watermarking scheme based on this framework that generalizes current multi-bit approaches, allowing us to achieve a wide variety of security-fidelity-capacity trade-offs. Finally, we provide an empirical validation of the theoretical guarantees of the scheme.

2 Decoupling the decision from the diffusion

Before starting, it is important to consider what counts as an acceptable watermark system. This requires a clear delimitation of the problem we are trying to solve with watermarking. Consider the following illustrative scenario:

Watermarking attribution scenario Alice makes accessible an API where users can request for images to be generated. In order to trace the use of her system, she asks Bob, a third-party, to provide her with a secret key k and a seed-based watermarking system $\mathcal{W}$. From the point of view of Bob, the secret key k is now uniquely linked to Alice. She then associates to each user a unique ID $\mathbf{m}_i$. Each request generates an image, using $\mathcal{W}$, that contains a watermark. If Bob is presented with the key k and a watermarked image generated for user i, he should decode the correct message $\mathbf{m}_i$.

The most important constraint of this scenario is that Alice **cannot change her secret key** but **can use as many messages as she wants**. Contrary to what is often assumed in the current watermarking literature, we argue that it is not trivial to work with many keys when using a multi-bit watermarking system. Indeed, if Alice uses multiple keys, attribution becomes difficult if not impossible: if she wants to extract the message from a given image, how does she determine which key to trust? Recall

that Alice does not know from which user the image came. Furthermore, one cannot solve the problem by associating a unique key to each user and choosing the correct one for decoding; this requires an oracle. If such an oracle exists, then watermarking is not necessary since attribution is solved simply by using this oracle.

The rest of the section is dedicated to developing a rigorous evaluation framework for watermarking systems, aiming to address the *attribution scenario*. For clarity and space considerations, we settle on a mostly informal and intuitive tone in the paper itself, but always link to formal definitions in Appendix B for important notions. We also advise the reader to peek at Appendix A for a review of notations used in the paper.

2.1 Watermarking systems

A watermarking system $\mathcal{W}$ – Def. B.1 – can be decomposed into four main elements: a set of secret keys $\mathcal{K}$, a projection function f, a family of embedding mechanisms $(e_k)_{k\in\mathcal{K}}$, and a decision mechanism d. The projection function $f : \mathbb{R}^D \to \mathbb{R}^L$ ($L \leq D$) transforms a piece of content into a representation more amenable to embedding and detection – the so-called watermark space. It should be robust to a family of usual content alterations. Embedding mechanisms e_k – Def. B.1 – are functions parameterized by the secret key. Their goal is to push a point in watermark space deep inside a specific region called the decoding region. The watermarked distribution obtained is denoted as $\mathcal{Q}_{(k,\mathbf{m})}$. The decision mechanism d – Def. B.1 – maps points of the watermark space to either a score (zero-bit watermarking) or a message (multi-bit watermarking). Our interest in this paper lies solely in multi-bit schemes. We will see in Section 2.2 that it is useful to separate the redundancy mechanism – Def.B.3 – (i.e, the error-correcting code) from the decision mechanism. For the rest of the paper, we thus make the distinction between a message $\mathbf{m}$ of size M and its corresponding representative codeword $\mathbf{c}$ of size $M' > M$. When using a redundancy mechanism, the watermarking system becomes an error-corrected watermarking system – Def. B.4 – which notably changes the watermarking distribution from $\mathcal{Q}_{(k,\mathbf{m})}$ to $\mathcal{Q}_{(k,\mathbf{c})}$.

One important property of a watermarking system is that it should be difficult to randomly find a piece of content that follows the desired watermarked distribution. We formalize this with the concept of cover distribution, denoted as $\mathcal{P}$ – Def. B.5. Informally, we expect message bits decoded from images sampled from a cover distribution to be independent and not biased in favor of a particular message. Such property is usually enforced by whitening a decoder [15, 19] or by designing so-called sound detectors from the ground-up [14].

Finally, seed-based watermarking enforces a specific structure on the cover distribution that will form the basis of the analysis in Section 3.2:

Definition 1 (Seed-based system – Def.B.6)**.** *A watermarking system $\mathcal{W}$ equipped with cover distribution $\mathcal{P}$ is said to be seed-based if the projection of non-watermarked content in watermark space is distributed as a standard Gaussian. That is we have that: $f(X) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_L), X \sim \mathcal{P}$.*

We now define the three main quantities of interest for a watermarking system: capacity, fidelity, and watermarking security. We contrast these properties to the more common trio of robustness, quality, and cryptographic security. For clarity, we split each subsection into two parts. The first discusses the current approach in measuring a given quantity, and why we believe it to be unsatisfying. The second formalizes our approach in an abstract and general fashion.

2.2 Robustness versus Capacity

All recent works on seed-based studies empirically assess robustness by evaluating the chosen detection statistics under a set of image transformations [15, 16, 38, 7, 19]. We claim that this empirical approach for evaluating robustness suffers from a major problem:

Problem No definite claim on the ranking of different watermarking designs can be made since the performance depend on every choice of the inverse diffusion pipeline, notably: the choice of diffusion model, the scheduler/solver, the number of inversion steps, the choice of VAE, the presence or absence of prompts at inversion, and in the former case the value of the guidance scale. The inverse diffusion can be improved, for example by the choice of a better solver. However, we claim this can hardly be called an improvement over the watermarking system itself, only that it benefits from a "better" estimation of the seed.

Traditional watermarking sidestepped these issues by evaluating the robustness of a given scheme theoretically. The main idea was to assume that any alteration in pixel space translates to a white Gaussian noise with a given power in watermark space [9, 25, 17]. Thanks to the properties of the latent space in seed-based watermarking, we can expect this model of robustness to be a good approximation of reality – something we empirically validate in Section 4.

Consequently, it would be natural to model the decision mechanism as an additive (potentially not) white Gaussian noise channel AWGN_σ. However, since bit-accuracy is so prevalent in current watermarking evaluations, it is fruitful instead to model the decision mechanism as M' independent binary symmetric channels with flip probability p – noted BSC_p. We discuss this choice in Appendix D.1: the main takeaway is that Shannon's capacity is a good summary of the overall watermarking system performance, which depends on both the robustness σ of the projection f and the so-called channel characteristic $p_k(\sigma)$ of the decision mechanism d.

Indeed, notice that the flip probability p depends on two parts of the system: the projection f, and the watermarked distribution $\mathcal{Q}_k$. In practice, an alteration in pixel space introduces some noise in watermark space. This noise will, in turn, impact p, with the impact depending on $\mathcal{Q}_k$. This is where we perform the decoupling between f and d. We measure the *robustness* of f against a transform t by the variance σ^2 of the noise $t$introduces in watermark space – Def. B.7. We then define the *decision channel characteristic* $p_k(\sigma)$ of d as the mapping between σ and p – see Def. B.8. Informally, the channel characteristic is the bit-error rate of the system, *before error-correction*, under a white Gaussian perturbation with variance σ^2.

The combination of both quantities leads to the Shannon capacity, which perfectly summarizes how much noise a watermarking system can resist for a given codeword size.

Proposition 1 (Capacity of a watermarking system)**.** *Let $t : \mathbb{R}^D \to \mathbb{R}^D$ be a function. A watermarking system $\mathcal{W}$ with projection f that is σ-robust to t and with channel characteristic p_k has a Shannon capacity $C_\sigma = 1 - h_2(p_k(\sigma))$, where h_2 is the binary entropy function.*

This means that the system cannot communicate the watermark reliably unless the message $\mathbf{m}$ is coded with a codeword $\mathbf{c}$ of size $M' \geq \lceil \frac{M}{C_\sigma} \rceil$. Except when specifically stated, we assume to work at Shannon's capacity in order to be agnostic to the choice of redundancy mechanism. As such, the decision mechanism is *only* determined by its channel characteristic, which maps a given noise power σ^2 to a given flip probability p. In practice, one has to choose an off-the-shelf code depending on the desired properties. Turbo codes [4] and Polar codes [1] allow the use of rates close to Shannon's capacity. Pseudo-random codes (PRC) [10] provide cryptographic security. Gaussian-Shading use of repetition code [38][Section 3.2], though highly suboptimal in every sense, is simple to implement and analyse.

2.3 Quality versus Fidelity

Quality is straightforward to define in the case of post-hoc watermarking. It suffices to compare the host image to the watermarked one. In particular, the PSNR is equivalent to the power of the watermark signal in this case. This is not possible in our case.

Problem 1 In-generation watermarking is peculiar because it lacks any original host image. Instead, the standard approach in the literature is to use empirical distance between the distribution of images generated by the watermarked model and a given distribution of un-watermarked images (generated or not) [19, 35, 20]. This is usually performed using the FID [22], which is notoriously unable to describe some important sampling failure modes [29, 30]. Notably, the diversity of content generation is difficult to assess using this metric. As an example, for a given key and message, the support of the seed-distribution of Gaussian-Shading is cut in half, leading to an observed lack of variety in generated content [20][Figure 13] but no observed impact on the FID. The choice of baseline distribution is also not trivial if one is concerned with reproducibility and comparability: papers report baseline FID that differ substantially between one another [15, 35, 38, 19, 20, 23]. Another approach is to use "semantic" distances such as the CLIP score [21], but this is also not satisfactory: a watermarked model that preserves semantics but significantly degrades the aesthetic content is not desirable.

This overall difficulty in empirical evaluation also demonstrates that there is a lack of clarity in what constitutes an acceptable degradation of the generative model.

We argue that the role of the watermark designer is not to guarantee a certain level of generation quality; that is the role of the model provider. The role of the watermark designer is rather to guarantee that the watermarked model *preserves* the properties of the original one. In the case of seed-based watermarking, the seed distribution is what impacts the generation. As such, the watermark designer wants to preserve the original distribution as much as possible. The quantification of the deviation from this original distribution is what we call fidelity.

Problem 2 Under our scenario, the secret key is fixed, and a user is linked to a unique message. Thus, a distortion-free scheme is not sufficient for our purpose: it only guarantees that, on average over the messages, the properties of the non-watermarked model are preserved. A case in point is again Gaussian-Shading: each user is attributed only half of the available sampling space for seeds. On average over the users, all of the sampling space is used. Yet, an unlucky user might only be able to access seeds that lead to "low-quality" images. This is not desirable and points to the need to *preserve the original distribution as much as possible for all keys and messages*.

Starting from the fundamental property of seed-based systems in Def. B.6, it is natural to measure fidelity as an f-divergence between the standard multivariate Gaussian distribution and a watermarked distribution $\mathcal{Q}_{(k,\mathbf{c})}$. For this paper, we settle on the Kullback-Leibler (KL) divergence, though it will require our analysis to use some asymptotic arguments in the next section.

Definition 2 (Fidelity of Seed-based watermarking – Def.B.9)**.** *A watermarked distribution $\mathcal{Q}_{(k,\mathbf{c})}$ of a seed-based watermarking system is said to be ζ-faithful iff $\mathrm{D_{KL}}(\mathcal{N}(\mathbf{0}, I_L)||\mathcal{Q}_{(k,\mathbf{c})}) \leq \zeta$. We call ζ the relative fidelity loss with respect to the Gaussian distribution.*

2.4 Cryptographic versus Watermarking Security

Problem: Watermarking is not cryptography The question of security for modern watermarking is slowly emerging as trivial vulnerabilities are found in state-of-the-art post-hoc schemes due to the lack of secret keys [2, 32]. Seed-based watermark systems have tackled this question under the cryptographic angle, both in Gaussian-Shading and the more recent applications of PRC [10, 20]. The idea conveyed by these works is that a watermarking system is secure if an attacker cannot infer anything of the embedded message from the watermarked seed. The conflation between this type of cryptographic security and watermarking proper is not new and dates back to the early 2000s. The problem was eloquently settled by Cox et al. in [13], which demonstrates that cryptography and security solve different problems. Maybe the key difference between watermarking and cryptography outlined in this work is the following "the secret carrier does not need to be exactly disclosed in order to break the watermarking system" [13][Section 4.1]". We point out that both Gaussian-Shading and PRC do not even hide the encrypted message within a secret carrier! It is readily available to any attacker by simply taking the sign of a watermarked seed.

Once again, it is useful to illustrate this problem with a scenario that will serve as our threat model:

Threat model: Spoofing attack Eve is a user of Alice's API. She is a friend of Kerckhoffs, an employee of Alice who disclosed to her the full watermarking system, $\mathcal{W}$ except for the secret key k. She thus knows that each user is given a unique message. She would like to impersonate Camille, another user from whom she has stolen N_o generated images. Since she was disclosed the watermarking system, she can retrieve the watermarked seed from each image. Furthermore, she knows that all Camille's images contain their unique ID $\mathbf{m}_{\text{Camille}}$. Note that in order to make her system more secure, Alice prevents a user from using the same seed twice. Hence, Eve cannot simply reuse Camille's seed to generate new images.

Under this threat model, both Gaussian-Shading and PRC are highly insecure. Even though Alice does not know the exact value of $\mathbf{m}_{\text{Camille}}$, she has access to a codeword $\mathbf{c}_{\text{Camille}}$ linked to Camille, which is sufficient for her: encryption does not bring any security in this situation. Indeed, since neither scheme makes the carrier secret, she can simply retrieve the encrypted message, sample a new seed from it, and generate a new image with it. This underlines the fundamental difference of watermarking security: *the codeword carrier needs to be secret*.

We evaluate the security of a watermarking system by the availability of an estimator ψ to retrieve the secret key k. We convey this approach through the notion of security ratio. Informally, it measures the number of watermarking samples necessary for Eve to mount an attack that is better than a naive brute-force:

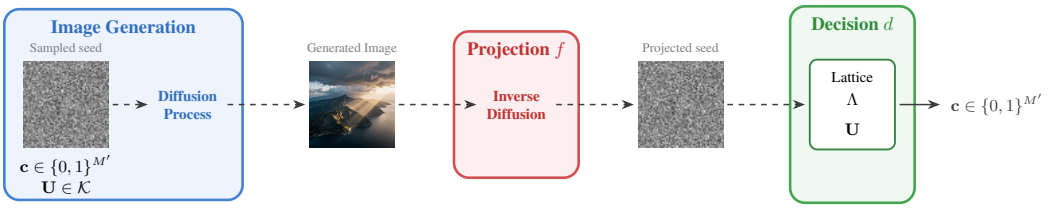


Figure 1: This is the diagram of SSB proposed. The unitary matrix $\mathbf{U}$ is used as the key. The projection f is the inverse diffusion from the image in pixel space to an approximation of the seed in the latent space. The decision mechanism d is the use of the lattice function Λ_Δ.

Definition 3 (Security ratio – Def.B.10)**.** *A M-bit error-corrected watermarking system $\mathcal{W}$ with codeword size M' is η-secure against an estimator ψ if it requires at least $N_o = \eta L$ watermarked observations to estimate a fixed key $k \in \mathcal{K}$ better than randomly guessing. That is, for any codeword $\mathbf{c} \in \{0,1\}^{M'}$:*

$$\eta L = \min\{N \mid \mathbb{P}[d\left(e_{\psi_k(N)}(\mathbf{c}), k\right) = \mathbf{c}] > \frac{1}{2^{M'}}, N \in \mathbb{N}^+\} \tag{1}$$

where $\psi_k(N)$ is an estimation of the key k based on N i.i.d. samples $Z \sim \mathcal{Q}_{(k,\mathbf{c})}$. We call η the security ratio of $\mathcal{W}$ against ψ. If $\eta = +\infty$, we say that $\mathcal{W}$ is perfectly secure against ψ.

Since both Gaussian-Shading and PRC do not use a secret key to conceal their codewords, it only necessitates a single sample to mount an attack, hence $\eta = 1/L$ for both of them.

3 A general lattice-based construction

We propose Secure Seed-Based (SSB), a multi-bit watermarking scheme heavily inspired by older schemes from the classic watermarking literature based on Voronoï modulation [9, 8] (also see Chapter 9 in [39]). Compared to these works, our approach is substantially simplified: we use two nested one-dimensional lattices in order to securely transmit a binary codeword. Despite the simplicity of the approach, this decision mechanism can operate under a wide range of capacity-fidelity-security trade-off, generalizing and largely outperforming current approaches. A surprising result is that perfectly secure transmission in the sense of Def.B.10 is often possible even under high noise variance σ^2 – see Appendix F.

Our focus here is solely on the *embedding function* and *decision mechanism*. We do not try to optimize the estimation of the original seed by improving the projection function (i.e. the inverse diffusion process). Similarly, recall that we are agnostic to the choice of error-correcting code; the only quantity of interest is the decision channel characteristic $p(\sigma)$ from Def. B.8. Figure 1 illustrates the functioning of the method.

3.1 Embedding and decoding

Let $\mathbf{c}$ be the binary codeword to hide. We define the key set $\mathcal{K}_u$ as the set of matrices in $\mathcal{M}^{L \times M'}$ which have their columns summing to one – in particular if $M' = L$, the matrix is unitary. Let $\mathbf{U} \in \mathcal{K}_u$. Finally, let the alternating lattice function $\Lambda_\Delta : \mathbb{R}^{M'} \to \{0,1\}^{M'}$ be defined as:

$$\Lambda_\Delta(\mathbf{x}) = \frac{(-1)^{\lfloor \frac{\mathbf{x}}{\Delta} \rfloor} + 1}{2}, \tag{2}$$

with Δ the size of a cell. We call Λ_Δ the **coarse lattice**. Note that when $\Delta \to +\infty$, the lattice function tends to the sign function: this is the decision mechanism of Gaussian-Shading.

Our goal is to sample a vector $\mathbf{z}_u \in \mathbb{R}^{M'}$ such that the lattice outputs the correct codeword: $\Lambda_\Delta(\mathbf{z}_u) = \mathbf{c}$. Once this is done, we have to project this vector into the original latent space $\mathbb{R}^L$. In order to improve the system's capacity, we use a second, nested, lattice function Λ_δ defined in the same way as in Eq. (2) but with $\delta \leq \Delta$. We call any such Λ_δ a **fine lattice**. An illustration of these embedding functions is proposed in Figure 2. Algorithm 1 details the sampling of $\mathbf{z}_u$.

procedure SAMPLEWM($\Delta, \delta, \mathbf{c}$)
  $(a_k)_{k\in\mathbb{Z}} \leftarrow 2k\Delta$; $(b_k)_{k\in\mathbb{Z}} \leftarrow a_k + \Delta$
  $(\alpha_k)_{k\in\mathbb{Z}} \leftarrow a_k + \frac{\Delta-\delta}{2}$; $(\beta_k)_{k\in\mathbb{Z}} \leftarrow \alpha_k + \delta$
  $(p_k)_{k\in\mathbb{Z}} \leftarrow \Phi(a_k) - \Phi(b_k)$
  for $i \in M'$ **do**
    $k \sim \text{Categorical}((p_k)_{k\in\mathbb{Z}})$
    $(z_u)_i \leftarrow \text{sign}(c_i)\text{Trunc}\mathcal{N}_{[\alpha_k,\beta_k]}(0,1)$
  end for
  return $\mathbf{z_u}$
end procedure

Algorithm 1: Seed sampling algorithm for SSB. In practice, we truncate k between -10 and 10 for numerical evaluations.

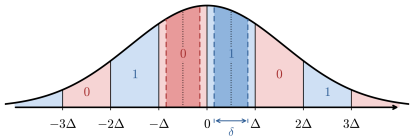


Figure 2: Diagram of the nested lattices Λ_Δ and Λ_δ. The coarse lattice cells are separated by full lines wheres fine cells are colored with dotted lines.

Now we still need to transform $\mathbf{z}_u$ into a valid seed for the diffusion process. Any seed $\mathbf{z} \in \mathbb{R}^L$ can be decomposed as $\mathbf{z} = (\mathbf{I}_L - \mathbf{U}\mathbf{U}^\top)\mathbf{z} + \mathbf{U}\mathbf{U}^\top\mathbf{z}$. In order to obtain a valid seed $\mathbf{z}$ from $\mathbf{z}_u$, one can thus sample $\mathbf{z}'$ as a realization of a random variable $Z' \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_L)$ and compute $\mathbf{z}$ as:

$$\mathbf{z} = (\mathbf{I}_L - \mathbf{U}\mathbf{U}^\top)\mathbf{z}' + \mathbf{U}\mathbf{z}_u. \tag{3}$$

For decoding, let $\mathbf{x} \in \mathbb{R}^D$ be an image watermarked with the codeword $\mathbf{c}$. We compute an approximation of the original latent $\hat{\mathbf{z}} = f(\mathbf{x})$. The codeword is then decoded simply by applying the coarse lattice function on $\mathbf{U}^\top\hat{\mathbf{z}}$ giving $\Lambda_\Delta(\mathbf{U}^\top\hat{\mathbf{z}}) = \hat{\mathbf{c}}$. By construction, if the projection does not introduce any noise, $\mathbf{U}^\top\hat{\mathbf{z}} = \mathbf{U}^\top\mathbf{z} = \mathbf{z}_u$ and consequently we retrieve the correct codeword: $\mathbf{c} = \hat{\mathbf{c}}$.

3.2 Theoretical Analysis

A direct consequence of Alg. 1 is that each watermarked element $(z_u)_i$ is distributed as an infinite mixture of truncated Gaussian distributions. Using the notation in Alg.1 (or Appendix A), and writing the probability to sample in the k-th lattice cell as $P_k = 2(\Phi(b_k) - \Phi(a_k))$, we have that: $(z_u)_i = \text{sign}(c_i)\sum_{k\in\mathbb{Z}} P_k Z_k$ with $Z_k \sim \text{Trunc}\mathcal{N}_{[\alpha_k,\beta_k]}(0,1)$. Since all watermarked elements are independent and follow the same distribution up to a sign, we note the unsigned expectation of $(z_u)_i$ as $\mu_{(\Delta,\delta)}$ and its variance as $\sigma^2_{(\Delta,\delta)}$ – see Proposition E.1.1 for an exact formula of these two moments.

Capacity The flip probability of an SSB decision mechanism is given by the probability that an element $(z_u)_i$ corrupted by some Gaussian noise with variance σ^2 leaves the set of "correct" coarse lattice cells. Since the distribution of $(z_u)_i$ is known, the channel characteristic $p_{(\Delta,\delta)}(\sigma)$ of SSB is immediately given by Proposition 2. We plot the resulting capacity $C_\sigma(\Delta,\delta)$ computed numerically using Proposition 1 in Figure 3.

Proposition 2 (SSB channel characteristic)**.** *Let (Δ,δ) define a SSB system. Then the probability that a codeword bit is flipped due to a transform against which f is σ-robust is:*

$$p_{(\Delta,\delta)}(\sigma) = 1 - \left(2\sum_{j=\mathbb{Z}} P_j \sum_{k=\mathbb{Z}} \frac{\int_{\alpha_j}^{\beta_j} \phi(p)\left[\Phi\left(\frac{b_k-p}{\sigma}\right) - \Phi\left(\frac{a_k-p}{\sigma}\right)\right]dp}{\Phi(\beta_j) - \Phi(\alpha_j)}\right) \tag{4}$$

Fidelity Recall that under our watermarking scenario we *must* assume both the codeword $\mathbf{c}$ and the secret rotation $\mathbf{U}$ to be fixed. The support of the watermarked distributions induced by any lattice embedding is consequently on a strict subset of $\mathbb{R}^L$. The KL-divergence thus diverges with respect to the Gaussian distribution. However, $\mathbf{U}$ is applied on independent random variables with finite mean and variance, leading to a "Gaussianization" of the distribution of the watermarked seed. Consequently, we propose to compute the fidelity on the asymptotic distribution of the watermarked seed obtained by application of the Lyapunov central-limit theorem.

Proposition 3 (SSB Asymptotic relative fidelity)**.** *Any (Δ,δ)-SSB system with codeword size M' is $M'\zeta(\Delta,\delta)$-faithful with $\zeta(\Delta,\delta) \xrightarrow{L\to\infty} \frac{1}{2}\left[\left(1+\mu^2_{(\Delta,\delta)}\right)\sigma^{-2}_{(\Delta,\delta)} + \log\sigma^2_{(\Delta,\delta)} - 1\right]$ if $\mathbf{U}$ is chosen such that Lyapunov condition holds.*

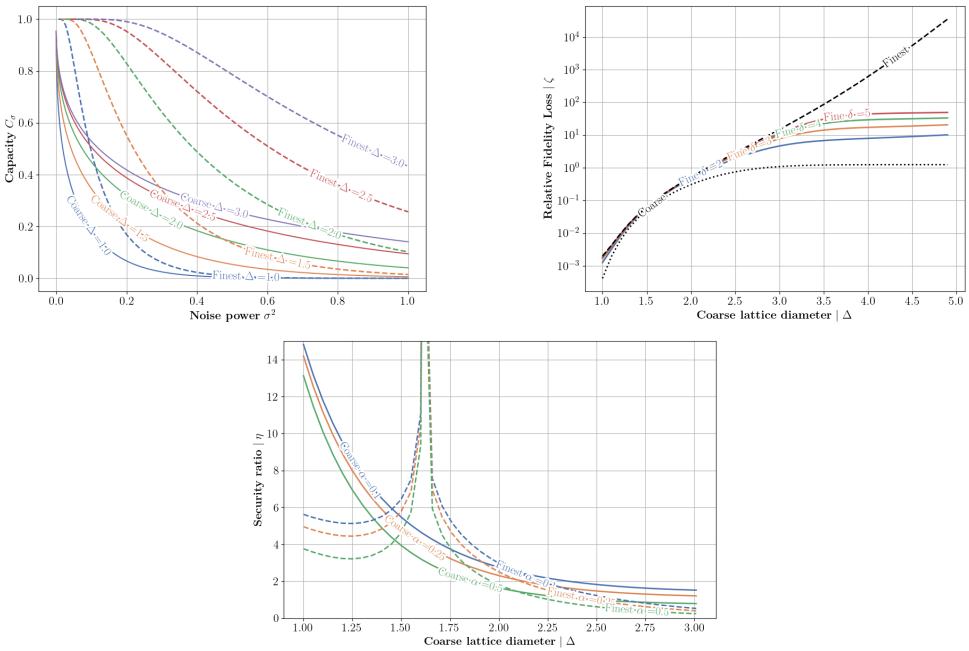


Figure 3: Watermarking characteristic of (Δ, δ)-SSB , showing the capacity, fidelity, and security as a function of the lattice parameters as well as of the codeword size relative to the latent size $\alpha = M'/L$.

Security: Principal Component Analysis (PCA) Attack In our watermarking scenario, an adversary, Eve, aims to estimate the secret key $\mathbf{U}$ from N_o watermarked observations. The eigenvalues of the covariance matrix of cover latents are the identity. However, the SSB embedding mechanism modifies these eigenvalues. As such, Eve's strategy relies on identifying deviation from the expected distribution of the eigenvalues of the empirical covariance estimated with the watermarked samples. Her best estimator is limited by the behavior of the Marchenko-Pastur distribution of the eigenvalues of empirical covariance matrices [5, 34]. We refer to Appendix D.2 as well as [18] for a primer on this type of attack. From Def. B.10, we consider security broken as soon as a single component of $\mathbf{U}$ is revealed by the estimator. The number of samples necessary to do so is given in Proposition 4.

Proposition 4 (SSB Watermarking security)**.** *Let $\alpha = \frac{M'}{L}$. Any (Δ, δ)-SSB is η-secure against a PCA estimator where $\eta = \left(\frac{1-\sqrt{\alpha}\sigma_{\Delta,\delta}}{1-\sigma_{\Delta,\delta}}\right)^2$.*

4 Experimental validation

4.1 Robustness of the inverse diffusion

We have defined robustness in Def.B.7 as the mapping between a certain image processing operation and an equivalent noise variance σ^2 introduced in watermarked space. For seed-based systems, robustness depends on two main *empirical* components: the diffusion model used for inversion and the choice of input distribution $\mathcal{F}$. For the latter, one should, in all rigor, use watermarked images as input distributions in order to match the hypotheses of Proposition 1. However, for most image processing operations of interest, choosing a cover distribution – i.e., non-watermarked images – should be an extremely good proxy and allows one to provide an evaluation agnostic to the choice of watermarking system. We show in Section 4.2 that our assumption is indeed empirically valid. We measure the robustness of current popular flow-matching diffusion models against common image processing operations in Table 1. We chose Sana [37], Z-Image Turbo [33], and Qwen [36] due to their high popularity on Huggingface at the time of writing, as well as their difference in architecture. Sana's VAE provides very high-compression rates, Z-Image Turbo is a distilled model, and Qwen is a very large model necessitating many (50) diffusion steps compared to the two other models[1].

[1]We deliberately chose *not* to use Stable-Diffusion 2 due to replicability concerns, as it was removed from Huggingface

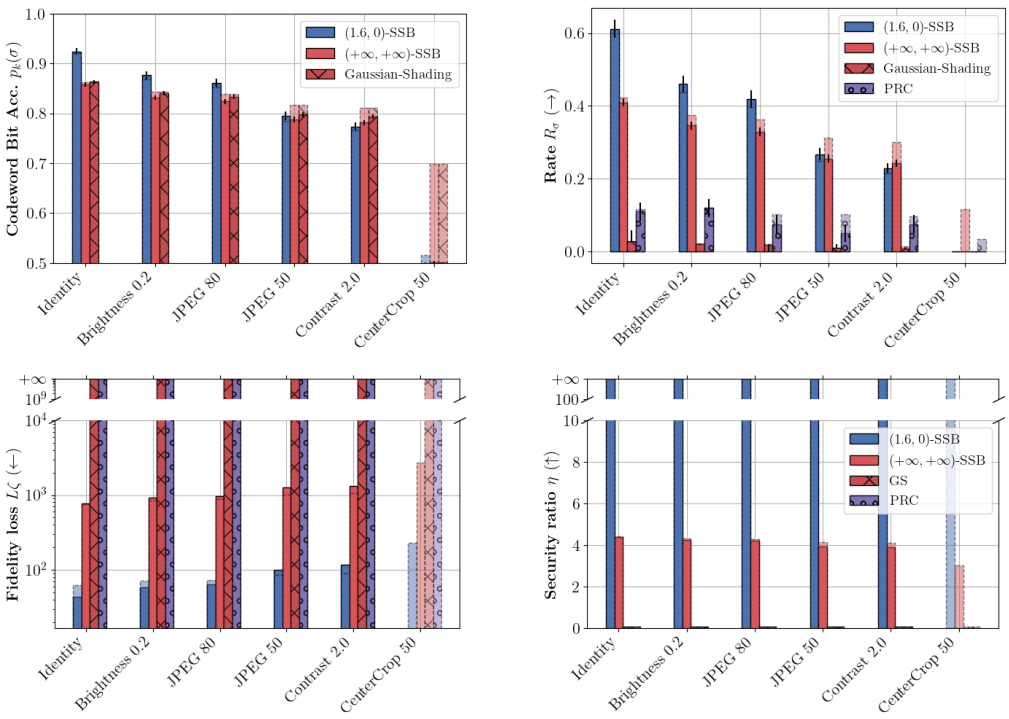


Figure 4: Empirical validation of the model defined in Section 2 for current seed-based approaches – Gaussian-Shading and PRC – as well as our method SSB . Transparent boxes correspond to theoretical values, opaque to empirical ones. We display 95% asymptotic confidence intervals.

LDM/Steps	**Image size** D	**Latent size** L	*Identity*	*Brightness* 0.2	*Contrast* 2.0	*JPEG* QF80	*JPEG* QF50	*Center Crop* 50%
Sana/ 25	512^2	8192	0.21	0.29	0.46	0.31	0.42	1.94
Z-image / 9	1024^2	262144	0.35	0.45	0.66	0.9	1.08	1.51
Qwen / 50	512^2	65536	0.34	0.44	0.78	0.78	1.09	2.09

Table 1: Equivalent white Gaussian noise variance σ^2 introduced in latent space for different common image operations – see Def. B.7. The *Identity* transform corresponds to the noise introduced by the inverse diffusion process itself (f in Def.B.2). All diffusion models are used with the standard configuration provided one their Huggingface page – see Appendix J.

Our robustness evaluation protocol is as follows. We generated $N = 100$ images $(\mathbf{x})_N$ from a collection of N seeds $(\mathbf{z})_N$ using a given model and prompts from Huggingface:`Gustavosta/Stable-Diffusion-Prompts`. We apply the augmentation t to the generated images. The projection f is then defined as the standard inverse flow matching procedure using the Euler solver, for the same number of diffusion steps as during generation. We then estimate the variance of the introduced noise as the empirical variance of the residuals $f(t(\mathbf{x})) - \mathbf{x}$ averaged over latent channels.

4.2 Comparing empirical and theoretical performances

We now validate our theoretical results to the empirical performance of SSB , Gaussian-Shading, and PRC. We choose to test two regimes for SSB $(\Delta, \delta) = (+\infty, +\infty)$ and $(1.6, 0)$. The first regime uses a decision mechanism identical to Gaussian-Shading's – the sign function. The second corresponds to the perfect security regime of SSB ($\eta = +\infty$) – see Appendix F. For each watermarking system, we generate 500 images with the same prompts as in the previous sections using the Sana diffusion model. We first compute the empirical bit-accuracy of the *codeword*, which corresponds to the channel characteristic $p(\sigma)$ in Def.B.8. From it we compute the optimal rate $R_\sigma \leq C_\sigma$ – see Def.1 – for which we are guaranteed to decode the correct message with negligible probability of decoding error P_e. For SSB , we set the rate to Shannon's capacity at the empirical $p(\sigma)$. Gaussian-Shading uses a repetition code; we thus compute the exact best rate that leads to a probability of decoding error $\mathrm{P}_e \leq 10^{-6}$ for a given $p(\sigma)$. Finally, PRC uses a belief-propagation code, BP+OSD [28], which is

not compatible with our BSC_p model since it uses soft-decisions. As such, we directly report the best theoretical and empirical rate for PRC. Since computing the best empirical rate would be far too costly for $\mathrm{P}_e \leq 10^{-6}$, we report it for $\mathrm{P}_e \leq 10^{-2}$, giving it a slight advantage. Fidelity and security are computed in turn using these rates. We report all results in Figure 4. Note that none of the current seed-based systems can resist cropping at 50%.

5 Conclusion

Modern watermarking methods lack a shared language. Methods are proposed and evaluated on different axes, and compared empirically in ways that conflate distinct design choices – the inverse diffusion model, the decision mechanism, and the error-correcting codes. This conflation makes it impossible to know whether a method is genuinely better, or simply makes a different trade-off. This work introduced a framework that allows fair theoretical comparison. By decoupling these components and anchoring the analysis in capacity, fidelity, and watermarking security, we bring multi-bit seed-based watermarking to align with classic watermarking methods. We also introduced a concrete implementation of this framework, SSB , a watermarking system that achieves a wide range of regimes along these three axes. Because each component is decoupled, each can be improved independently, and future work can do so without redesigning the whole system. We hope this work shifts the focus away from ad-hoc optimizations based on empirical evaluations, and towards design with strong theoretical guarantees. Several directions follow naturally from this work. First, it is clear that the robustness of the projection function can be improved, for example by a better model of variance on each latent channel. Second, our analysis assumes a BSC_p: understanding the gap with an AWGN_σ remains to be studied. Finally, extending this framework to 0-bit seed-based watermarking is a natural (but not trivial!) next step.

Acknowledgment

Experiments presented in this paper were carried out using the Grid'5000 tested, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`). This research is part of the PEPR Cyber Compromis project, ANR-22-PECY-0011.

References

[1] Erdal Arikan. Channel polarization: A method for constructing capacity-achieving codes for symmetric binary-input memoryless channels. *IEEE Transactions on Information Theory*, 55(7):3051–3073, July 2009. arXiv:0807.3917 [cs].

[2] Patrick Bas and Jan Butora. The AI Waterfall : A Case Study in Integrating Machine Learning and Security. In *GRETSI*, Strasbourg, France, August 2025.

[3] Patrick Bas and Teddy Furon. A new measure of watermarking security: The effective key length. *IEEE Transactions on Information Forensics and Security*, 8(8):1306–1317, 2013.

[4] C. Berrou, A. Glavieux, and P. Thitimajshima. Near Shannon limit error-correcting coding and decoding: Turbo-codes. 1. In *Proceedings of ICC '93 - IEEE International Conference on Communications*, volume 2, pages 1064–1070 vol.2, May 1993.

[5] Pascal Bianchi, Merouane Debbah, Mylène Maïda, and Jamal Najim. Performance of Statistical Tests for Single Source Detection using Random Matrix Theory. *IEEE Transactions on Information Theory*, 57(4):2400–2419, April 2011.

[6] Tu Bui, Shruti Agarwal, and John Collomosse. TrustMark: Universal Watermarking for Arbitrary Resolution Images, November 2023. arXiv:2311.18297 [cs].

[7] Tu Bui, Shruti Agarwal, and John Collomosse. TrustMark: Robust Watermarking and Watermark Removal for Arbitrary Resolution Images. pages 18629–18639, 2025.

[8] Marc Chaumont and Dalila Goudia. TCQ practical evaluation in the hyper-cube watermarking framework. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–6, July 2011. ISSN: 1945-788X.

[9] B. Chen and G.W. Wornell. Quantization index modulation: a class of provably good methods for digital watermarking and information embedding. In *2000 IEEE International Symposium on Information Theory (Cat. No.00CH37060)*, pages 46–, June 2000.

[10] Miranda Christ and Sam Gunn. Pseudorandom Error-Correcting Codes, June 2024. arXiv:2402.09370 [cs].

[11] Hai Ci, Pei Yang, Yiren Song, and Mike Zheng Shou. RingID: Rethinking Tree-Ring Watermarking for Enhanced Multi-Key Identification, July 2024. arXiv:2404.14055 [cs].

[12] Daniel J. Costello and G. David Forney. Channel coding: The road to channel capacity. *Proceedings of the IEEE*, 95(6):1150–1177, June 2007.

[13] Ingemar J. Cox, Gwenaël Doërr, and Teddy Furon. Watermarking is not cryptography. In *Proceedings of the 5th international conference on Digital Watermarking*, IWDW'06, pages 1–15, Berlin, Heidelberg, November 2006. Springer-Verlag.

[14] Pierre Fernandez, Antoine Chaffin, Karim Tit, Vivien Chappelier, and Teddy Furon. Three Bricks to Consolidate Watermarks for Large Language Models. In *2023 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, December 2023. ISSN: 2157-4774.

[15] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The Stable Signature: Rooting Watermarks in Latent Diffusion Models. pages 22466–22477, 2023.

[16] Pierre Fernandez, Hady Elsahar, I. Zeki Yalniz, and Alexandre Mourachko. Video Seal: Open and Efficient Video Watermarking, December 2024. arXiv:2412.09492 [cs].

[17] Teddy Furon and Patrick Bas. Broken Arrows. *EURASIP Journal on Information Security*, 2008:ID 597040, October 2008.

[18] Teddy Furon, Hervé Jégou, Laurent Amsaleg, and Benjamin Mathon. Fast and secure similarity search in high dimensional space. In *2013 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 73–78, November 2013. ISSN: 2157-4774.

[19] Enoal Gesny, Eva Giboulot, Teddy Furon, and Vivien Chappelier. Guidance watermarking for diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026.

[20] Sam Gunn, Xuandong Zhao, and Dawn Song. An Undetectable Watermark for Generative Image Models, April 2025. arXiv:2410.07369 [cs].

[21] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning, March 2022. arXiv:2104.08718 [cs].

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6629–6640, Red Hook, NY, USA, December 2017. Curran Associates Inc.

[23] Kecen Li, Zhicong Huang, Xinwen Hou, and Cheng Hong. GaussMarker: Robust Dual-Domain Watermark for Diffusion Models. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 34688–34701. PMLR, October 2025.

[24] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning (ICML)*, 2022.

[25] Stéphane Pateux and Gaëtan Le Guelvouit. Practical watermarking scheme based on wide spread spectrum and game theory. *Signal Processing: Image Communication*, 18(4):283–296, April 2003.

[26] Yury Polyanskiy, H. Vincent Poor, and Sergio Verdu. Channel Coding Rate in the Finite Blocklength Regime. *IEEE Transactions on Information Theory*, 56(5):2307–2359, May 2010.

[27] Yury Polyanskiy and Yihong Wu. *Information theory: from coding to learning*. Cambridge University Press, Cambridge, United Kingdom ; New York, NY, 2025.

[28] Joschka Roffe, David R. White, Simon Burton, and Earl T. Campbell. Decoding Across the Quantum LDPC Code Landscape. *Physical Review Research*, 2(4):043423, December 2020. arXiv:2005.07016 [quant-ph].

[29] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 5234–5243, Red Hook, NY, USA, December 2018. Curran Associates Inc.

[30] Loic Simon, Ryan Webster, and Julien Rabin. Revisiting precision recall definition for generative modeling. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5799–5808. PMLR, May 2019.

[31] Tomas Soucek, Sylvestre-Alvise Rebuffi, Pierre Fernandez, Nikola Jovanović, Hady Elsahar, Valeriu Lacatusu, Tuan A. Tran, and Alexandre Mourachko. Transferable black-box one-shot forging of watermarks via image preference models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[32] Hussein Tarhini, Aurélien Noirault, Jan Butora, and Patrick Bas. Neural Watermarking: Lack of a Secret Key is still Lack of Security. March 2026.

[33] Image Team, Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, Zhen Li, Zhong-Yu Li, David Liu, Dongyang Liu, Junhan Shi, Qilong Wu, Feng Yu, Chi Zhang, Shifeng Zhang, and Shilin Zhou. Z-Image: An Efficient Image Generation Foundation Model with Single-Stream Diffusion Transformer, November 2025.

[34] Pascal Vallet, Philippe Loubaton, and Xavier Mestre. Improved subspace estimation for multivariate observations of high dimension: the deterministic signals case. *IEEE Transactions on Information Theory*, 58(2):1043–1068, February 2012. arXiv:1002.3234 [cs].

[35] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[36] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image Technical Report, August 2025. arXiv:2508.02324 [cs].

[37] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient High-Resolution Image Synthesis with Linear Diffusion Transformers, October 2024.

[38] Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12162–12171, 2024.

[39] Ram Zamir. *Lattice Coding for Signals and Networks: A Structured Coding Approach to Quantization, Modulation and Multiuser Information Theory*. Cambridge University Press, Cambridge, 2014.

A Notation

Spaces

- **Pixel space**: in $\mathbb{R}^D$ with observations denoted $\mathbf{x}$.
- **Latent space**: in $\mathbb{R}^L$ with observations denoted $\mathbf{z}$.
- **Watermark space**: in $\mathbb{R}^{M'}$ with observations denoted $\mathbf{z}_u$.

Distributions We use the standard notation for the standard Gaussian p.d.f ϕ and its c.d.f Φ. We denote the Truncated Gaussian with location μ and scale parameter σ and boundaries (a, b) as $\mathrm{Trunc}\mathcal{N}_{[a,b]}\left(\mu, \sigma^2\right)$. Other distributions are usually referred to with calligraphic letters $(\mathcal{N})$.

Secure Seed Based watermarking system We refer to a (Δ, δ)-SSB system as an error-corrected watermarking system (Def.B.4), where:

- The set of secret keys is given by the set of $L \times M'$ matrices with columns summing to 1. An element of this set is denoted as $\mathbf{U}$.
- The projection function f is any function from $\mathbb{R}^D$ to $\mathbb{R}^L$.
- The embedding functions are given by Alg. 1 followed by Eq. (3).
- The decision mechanism is simply Λ_Δ as defined in Eq 2.

We assume the redundancy mechanism works at Shannon's capacity. A message is denoted as $\mathbf{m} \in \{0,1\}^M$ and its representative codeword as $\mathbf{c} \in \{0,1\}^{M'}$. For convenience, we use a slightly modified version of the sign function defined as:

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{else} \end{cases} \tag{5}$$

Importantly, note that $\text{sign}(0) = -1$.

Lattice cells

- $(a_k)_{k\in\mathbb{Z}} = 2k\Delta$ and $(b_k)_{k\in\mathbb{Z}} = a_k + \Delta$ are the boundaries of the coarse (Δ) lattice cells.
- $(\alpha_k)_{k\in\mathbb{Z}} = a_k + \frac{\Delta-\delta}{2}$ and $(\beta_k)_{k\in\mathbb{Z}} = \alpha_k + \delta$ are the boundaries of the fine (δ) lattice cells.
- $P_k \triangleq 2(\Phi(b_k) - \Phi(a_k))$.

B Formal Definitions

B.1 Watermarking systems

This appendix aims to make precise the informal definitions found in Section 2. A first reading of the paper can safely skip these details. However, note that the soundness of the evaluation framework and the corresponding proofs explicitly rely on the definitions found herein. For the reader's convenience, we also collect definitions already provided in the main text.

Definition B.1 (Embedding mechanism)**.** *An embedding mechanism e is a random variable that maps a message $\mathbf{m} \in \{0,1\}^M$ to a subset of the watermarked space $\mathcal{D}_\mathbf{m} \subset \mathbb{R}^L$. The induced probability distribution is denoted as $\mathcal{Q}_\mathbf{m}$ and called the watermarked distribution.*

Note that $\mathcal{D}_\mathbf{m}$ is to be understood as a *subset* of the decoding region of the decoder for message $\mathbf{m}$: a good embedding mechanism favors mappings deep inside the correct decoding region in order to maximize robustness.

Definition B.2 (Multi-bit watermarking System)**.** *A M-bit watermarking system $\mathcal{W}$ is a quadruplet $(\mathcal{K}, f, (e_k)_{k\in\mathcal{K}}, d)$ where $\mathcal{K}$ is the set of secret keys, $f : \mathbb{R}^D \to \mathbb{R}^L$ the projection function, $(e_k)_{k\in\mathcal{K}}$ a family of embedding mechanisms indexed by secret keys and $d : \mathbb{R}^L \times \mathcal{K} \to \{0,1\}^M$ the decision mechanism.*

Definition B.3 (Redundancy mechanism)**.** *A redundancy mechanism is composed of an encoder and a decoder $(c_{\text{enc}}, c_{\text{dec}})$. The encoder c_{enc} is a bijection between messages and a subset $\mathcal{C} \in \{0,1\}^{M'}$ called the codebook. For each codeword $\mathbf{c}$ in the codebook, the decoder c_{dec} defines an equivalence class $[\mathbf{c}] = \{\mathbf{v} \in \{0,1\}^{M'} : c_{\text{dec}}(\mathbf{v}) = \mathbf{c}\}$ and where $c_{\text{dec}} : \{0,1\}^{M'}$. We will abuse notation and make no distinction between the equivalence class $[\mathbf{c}]$ and its representative codeword $\mathbf{c}$.*

Definition B.4 (Error-corrected watermarking system)**.** *A M-bit watermarking system $\mathcal{W}$ equipped with a redundancy mechanism $(c_{\text{enc}}, c_{\text{dec}})$ replaces its embedding functions $(e_k)_{k\in\mathcal{K}}$ by $e_k^{(c)} \triangleq e_k \circ c_{\text{enc}}$. The induced watermarked distribution are denoted as $\mathcal{Q}_{(k,\mathbf{c})}$. Its decision mechanism is replaced by a function $d^{(c)} \triangleq c_{\text{enc}}^{-1} \circ c_{\text{dec}} \circ d$ where $d : \mathbb{R}^L \times \mathcal{K} \to \{0,1\}^{M'}$. We call $d^{(c)}$ the error-corrected decision mechanism; we still call d the decision mechanism.*

Definition B.5 (Cover distribution)**.** *Let $\mathcal{P}$ be a probability distribution with support in $\mathbb{R}^D$. It is said to be a cover distribution for a M-bit watermarking system $\mathcal{W}$ iff, for all secret keys $k \in \mathcal{K}$:*

$$d\left(f(X), k\right) \sim \mathcal{B}^M\left(\frac{1}{2}\right), \tag{6}$$

where $X \sim \mathcal{P}$ and $\mathcal{B}^M$ is a M-dimensional Bernoulli distribution.

Definition B.6 (Seed-based system)**.** *A watermarking system equipped with cover distribution $\mathcal{P}$ is said to be seed-based if the projection of non-watermarked content in watermark space is distributed as a standard Gaussian. That is we have that: $f(X) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_L), X \sim \mathcal{P}$.*

B.2 Watermarking characteristic properties

Definition B.7 (Robustness)**.** *A projection function f is said to be σ-robust to a function $t : \mathbb{R}^D \to \mathbb{R}^D$ under an input distribution $\mathcal{F}$ (with support in $\mathbb{R}^D$) iff :*

$$\mathbb{V}[(f \circ t)(X)] \leq \sigma^2, X \sim \mathcal{F}. \tag{7}$$

Definition B.8 (Decision channel characteristic)**.** *Let $\mathcal{W}$ be a M-bit error-corrected watermarking system with codeword size M'. The channel characteristic p_k of $\mathcal{W}$ for a key $k \in \mathcal{K}$ is a mapping $p_k : \mathbb{R}^+ \to [0, 1]$ defined as:*

$$p_k(\sigma) = \frac{1}{M' 2^M} \sum_{\mathbf{c} \in \mathcal{C}} \mathbb{E}_{Z \sim \mathcal{Q}_{(k,\mathbf{c})}} \left[\mathbb{E}_{\epsilon_\sigma}\left[\mathrm{ham}(d\left(Z + \epsilon_\sigma, k\right), d\left(Z, k\right))\right]\right], \tag{8}$$

where $\epsilon_\sigma \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_L)$ and ham *is the Hamming distance.*

Definition B.9 (Fidelity of Seed-based watermarking)**.** *A watermarked distribution $\mathcal{Q}_{(k,\mathbf{c})}$ of a seed-based watermarking system is said to be ζ-faithful iff:*

$$\mathrm{D}_{\mathrm{KL}}(\mathcal{N}(\mathbf{0}, I_L) || \mathcal{Q}_{(k,\mathbf{c})}) \leq \zeta. \tag{9}$$

We call ζ the relative fidelity loss with respect to the Gaussian distribution.

Definition B.10 (Security ratio)**.** *A M-bit error-corrected watermarking system $\mathcal{W}$ with codeword size M' is η-secure against an estimator ψ if it requires at least $N_o = \eta L$ watermarked observations to estimate a fixed key $k \in \mathcal{K}$ better than randomly guessing. That is, for any codeword $\mathbf{c} \in \{0, 1\}^{M'}$:*

$$\eta L = \min\{N \mid \mathbb{P}[d\left(e_{\psi_k(N)}(\mathbf{c}), k\right) = \mathbf{c}] > \frac{1}{2^{M'}}, N \in \mathbb{N}^+\} \tag{10}$$

where $\psi_k(N)$ is an estimation of the key k based on N i.i.d. samples $Z \sim \mathcal{Q}_{(k,\mathbf{c})}$. We call η the security ratio of $\mathcal{W}$ against ψ. If $\eta = +\infty$, we say that $\mathcal{W}$ is perfectly secure against ψ.

C Further discussion about the scenario

In this work, we consider the identity-attribution scenario with a fixed key and a message attributed to each user. In this section, we discuss another threat model to justify the choice.

Fixed Key, variable messages Another common threat model is when the adversary has access to watermarked content generated with the same key but different messages. Furthermore, a given user is not associated with a fixed message. If we assume the messages for this user are sampled uniformly from the message space, the fidelity loss ζ averages to 0 for every studied seed-based scheme. This is thus a less stringent scenario than identity attribution. However, this is true only if the messages are sampled uniformly from the message space. For session-id, one can use a hash of the session-id number or other cryptographic tools to approximate this hypothesis. As soon as one deviates from this hypothesis, correlations re-appear between latent elements, leading back to the previous case. Regarding security, little changes for PRC and Gaussian-Shading. They suffer from the same lack of a secret key. With knowledge of the watermarking scheme and a unique observation, Eve can still embed any codeword into the initial noise. One could object a further distinction: are the embedded messages known or unknown to Eve? If they are known, she can copy them without difficulty

by copying their codewords. If unknown, and assuming the message was encrypted before being embedded, then copying any codeword is still possible, but the attacker does not know the content of the embedded codeword. Though true, this is of no consequence to our analysis: watermarking security refers to the possibility for an attacker to estimate/forge any codeword in the watermarking channel (which is still the case here), whereas cryptographic security refers to the security of the codeword content itself. In this "unknown message" scenario, Gaussian-Shading and PRC are indeed cryptographically secure but not watermark secure.

D Further discussion about the watermarking system modeling

D.1 Choice of channel model

We decided to model the decision mechanism as a binary symmetric channel BSC_p instead of an additive white Gaussian noise AWGN_σ.

This choice does lead to a small loss in capacity: the "hard decision" prevents some capacity-achieving codes from reaching their full potential [12][Section C]. Nevertheless, this ultimately facilitates the analysis and comparisons to other multi-bit watermarking schemes. It also separates the choice of decision mechanism and coding mechanism, providing great flexibility in the choice of the latter depending on the desired properties needed for the watermarking system (e.g. maximal capacity versus cryptographic security).

Shannon's noisy channel coding theorem [27][Theorem 19.9], provides the sufficient (asymptotic) rate at which one must work to achieve an arbitrarily small probability of error[2]. The strong converse theorem [27][Section 22.1 and Theorem 22.1] shows this rate to be necessary for the BSC_p channel. Importantly, the error exponent converges exponentially fast to 0 or 1. This makes the asymptotic study of the channel straightforward since the actual value of the probability of decoding error does not matter: it can be as small (though not zero) as one desires for rates below capacity; the probability of error is essentially one for rates above capacity. The main quantity of interest thus becomes the flip probability p, which fully determines the capacity.

Now, the flip probability p depends on two parts of the system: the robustness of the projection f, and the watermarked distribution $\mathcal{Q}_k$. In practice, an alteration in pixel space introduces some noise in watermark space. This noise will, in turn, impact p, with the impact depending on $\mathcal{Q}_k$. This is where we perform the decoupling between f and d. We measure the *robustness* of f against a transform t by the power σ^2 of the noise that said transform introduces in watermark space – see Definition B.7 in the AppendixB. We finally define the *decision channel characteristic* p_k of d as the mapping between σ^2 and p – see Definition B.8 in the Appendix B. The combination of both quantities leads to the Shannon capacity, which perfectly summarizes how much noise a watermarking system can resist for a given codeword size.

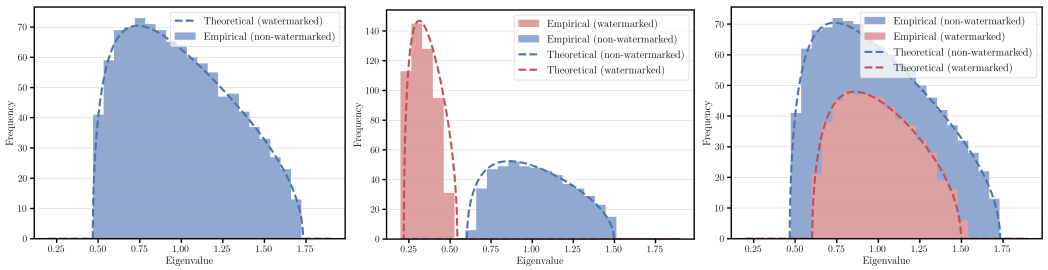
D.2 Security, PCA attack and the Marchenko-Pastur distribution

This section relies on the threat model introduced in Section 3.2, instantiated for the SSB method described in Section 3.2.

Setup Eve observes N_o watermarked samples in order to recover the secret key $\mathbf{U}$. In this context, Eve's best strategy is Principal Component Analysis (PCA): Eve computes the empirical covariance matrix of the observations and performs an eigendecomposition, hoping that the directions of $\mathbf{U}$ appear as distinguishable principal components.

PCA Attack Given N_o observations, Eve computes the empirical covariance matrix $\hat{\Sigma}_{\mathbf{z}} = \frac{1}{N_o}\sum_i \mathbf{z}_i\mathbf{z}_i^\top$, and looks at its spectrum. In the absence of a watermark, the observations are i.i.d. with a covariance matrix I_L. Random matrix theory predicts that the empirical eigenvalues follow the Marchenko-Pastur (MP) distribution $\mathrm{MP}(\frac{L}{N_o}, 1)$. If the watermark causes a covariance shift, some eigenvalues depart from 1, and the corresponding directions stand out as outliers relative to the expected MP distribution. Eve can use them to estimate $\mathbf{U}$.

[2]We are aware of the recent non-asymptotic bounds on the capacity [26], but the Shannon regime conveniently allows an analysis independent of probability of decoding error.



(a) No watermark (b) $(+\infty, +\infty)$-SSB (c) $(1.6, 0)$-SSB

Figure 5: Comparison between empirical and theoretical eigenvalues of the covariance matrix for different sampling of the seed. The empirical eigenvalues are computed over $N_o = 10L$ samples with $L = 512$ and $M' = 256$.

Example analysis Figure 5 gives illustrative examples of eigenvalue distributions. First, Figure 5a illustrates the baseline, when no watermark is present in $\mathbf{z}$. The empirical spectrum closely matches the MP distribution. When we sample a watermark seed, the true covariance may no longer be the identity. In the case illustrated by Figure 5b, the theoretical eigenvalues are $\lambda_0 = 1$ for $L - M'$ dimensions and $\lambda_1 = 1 - \frac{2}{\pi}$ for M' dimensions. For a sufficiently large N_o ($10L$ in the figure), Eve observes two distinct MP distributions, each centered around one of these eigenvalues. By identifying the subspace associated with λ_1, Eve estimates $\mathbf{U}$, making the scheme insecure for N_o observations. To satisfy the security ratio definition – Def. B.10 – the support of the watermarked empirical spectrum must be contained within the support of the non-watermarked MP distribution. This condition ensures that no outlier eigenvalue can betray a direction of $\mathbf{U}$. This condition is achieved for all N_o when $\lambda_o = \lambda_1 = 1$. Figure 5c illustrates this perfect-security regime for the $(1.6, 0)$-SSB scheme detailed in Appendix F. The empirical spectrum is indistinguishable from the unwatermarked case, so Eve gains no information about the secret key $\mathbf{U}$.

E Proofs

E.1 Fundamental properties of SSB

All the subsequent proofs rest on the fact that each element in the watermarked elements in *watermark space* $(z_u)_i$ is independent and distributed as an infinite mixture of truncated standard Gaussian random variables. For a given codeword $\mathbf{c}$ we always have:

$$(z_u)_i = \operatorname{sign}(c_i) \sum_{k \in \mathbb{Z}} P_k Z_k, \tag{11}$$

$$Z_k \sim \operatorname{Trunc}\mathcal{N}_{[\alpha_k, \beta_k]}(0, 1). \tag{12}$$

This is explicitly constructed in Alg.1 and as such does not need further proof.

On the other hand, the watermarked distribution in the latent space $\mathcal{Q}_{(\mathbf{U},\mathbf{c})}$ is more difficult to describe as a weighted sum of infinite mixtures. However, thanks to the nature of $\mathbf{U}$, it is simple to obtain the asymptotic watermarked distribution. First, we quantify the first two moments in the watermarked space.

Proposition E.1.1 (Embedding distribution moments)**.** *Let (Δ, δ) define a SSB system. Then we have that, for any $\mathbf{z}_u \sim \mathcal{Q}_{(\mathbf{U},\mathbf{c})}$,*

$$\mathbb{E}[\mathbf{z}_u] \triangleq \boldsymbol{\mu}_{(\Delta,\delta)} = \operatorname{sign}(\mathbf{c}) \sum_{k=-\infty}^{+\infty} P_k \frac{\phi(\alpha_k) - \phi(\beta_k)}{\Phi(\beta_k) - \Phi(\alpha_k)}, \tag{13}$$

$$\operatorname{Cov}[\mathbf{z}_u] \triangleq \boldsymbol{\Sigma}_{(\Delta,\delta)} = \mathbf{I}_L \left(\sum_{k=-\infty}^{+\infty} P_k \left(1 - \frac{\beta_k \phi(\beta_k) - \alpha_k \phi(\alpha_k)}{\Phi(\beta_k) - \Phi(\alpha_k)} \right) - \boldsymbol{\mu}^2_{(\Delta,\delta)} \right). \tag{14}$$

Proof. See Section E.5. □

Observing that all watermarked elements are independent and follow the same distribution up to a sign, we note the unsigned expectation of $(z_u)_i$ as $\mu_{(\Delta,\delta)}$ and its variance as $\sigma^2_{(\Delta,\delta)}$.

From this, one can derive the asymptotic distribution of the watermarked latent, which is Gaussian if the weights of $\mathbf{U}$ are chosen suitably. We don't provide an explicit solution for guaranteeing that $\mathbf{U}$ is "good" in this sense since in practice, one only has to ensure the entries of $\mathbf{U}$ are close to a uniform distribution to ensure Lyapunov's condition holds and that enough entries per row are non-zero. This can be done by sampling a matrix in $\mathbb{R}^{L\times M'}$ where each entry is sampled from a uniform distribution. Each column is then normalized such that it sums to 1. Importantly, note that the matrix where each column contains a single 1 is not a good matrix for SSB .

Proposition E.1.2 (Asymptotic watermarked distribution)**.** *Let (Δ, δ) define a SSB system with fixed key $\mathbf{U}$ and codeword $\mathbf{c}$. We have that the watermarked distribution $\mathcal{Q}_{(\mathbf{U},\mathbf{c})}$ converges as $L \to \infty$ in distribution to a multivariate Gaussian distribution $\mathcal{N}(\mathbf{U}^T\boldsymbol{\mu}_{(\Delta,\delta)}, \mathbf{U}^T\boldsymbol{\Sigma}_{(\Delta,\delta)}\mathbf{U})$ if $\mathbf{U}$ is chosen such that Lyapunov condition holds.*

Proof. We provide a sketch of the proof using a slightly weaker result where the infinite sums in Proposition E.1.1 are truncated such that $|k| \leq \kappa$. For any choice of κ, the moments are then obviously finite since $\Phi(x)$ is strictly increasing and bounded away from zero and from one. Consequently, $\forall |k| < \kappa$, we always have $\Phi(\beta_k) - \Phi(\alpha_k) > 0$. The proposition assumes the Lyapunov condition holds, which concludes the proof. □

E.2 Proposition 1: Capacity of a watermarking system

Proposition E.2.1 (Capacity of a watermarking system)**.** *Let $t : \mathbb{R}^D \to \mathbb{R}^D$ be a function. A watermarking system $\mathcal{W}$ with projection f that is σ-robust to t and with channel characteristic p_k has a Shannon capacity $C_\sigma = 1 - h_2(p_k(\sigma))$, where h_2 is the binary entropy function.*

Proof. Since the decision mechanism is modeled as a binary symmetric channel with flip probability $p_k(\sigma)$, this is the classic application of Shannon's noisy channel coding theorem [27][Theorem 19.9] to a BSC_p channel. □

E.3 Proposition 2 : SSB Channel characteristic

Proposition E.3.1 (SSB channel characteristic)**.** *Let (Δ, δ) define a SSB system. Then the probability that a codeword bit is flipped due to a transform against which f is σ-robust is:*

$$p_{(\Delta,\delta)}(\sigma) = 1 - \left(2 \sum_{j=\mathbb{Z}} P_j \sum_{k=\mathbb{Z}} \frac{\int_{\alpha_j}^{\beta_j} \phi(p) \left[\Phi\left(\frac{b_k - p}{\sigma}\right) - \Phi\left(\frac{a_k - p}{\sigma}\right) \right] dp}{\Phi(\beta_j) - \Phi(\alpha_j)} \right) \tag{15}$$

Proof. We denote by $1 - p_{(\Delta,\delta)}(\sigma)$ the probability that a codeword bit is not flipped due to a transform against which f is σ-robust.

$$1 - p_{(\Delta,\delta)}(\sigma) = \sum_{j=\mathbb{Z}} \frac{1}{Z_j} \int_{\alpha_j}^{\beta_j} \left[2 \sum_{k=\mathbb{Z}} \int_{a_k}^{b_k} \phi\left(\frac{x-p}{\sigma}\right) dx \right] \phi(p) dp, \tag{16}$$

$$\tag{17}$$

with $Z_j = \Phi(\beta_i) - \Phi(\alpha_i)$ the normalizing constant. Then, we have:

$$\int_{a_k}^{b_k} \phi\left(\frac{x-p}{\sigma}\right) dx \phi(p) \triangleq S_{k,\sigma}(p) = \left[\Phi\left(\frac{b_k - p}{\sigma}\right) - \Phi\left(\frac{a_k - p}{\sigma}\right) \right] \phi(p) \tag{18}$$

and the probability that a codeword bit is flipped due to a transform against which f is σ-robust is:

$$p_{(\Delta,\delta)}(\sigma) = 1 - \left(2 \sum_{j=\mathbb{Z}} \sum_{k=\mathbb{Z}} \frac{1}{Z_j} \int_{\alpha_j}^{\beta_j} S_{k,\sigma}(p) dp \right); S_{k,\sigma}(p) = \left[\Phi\left(\frac{b_k - p}{\sigma}\right) - \Phi\left(\frac{a_k - p}{\sigma}\right) \right] \phi(p) \tag{19}$$

□

E.4 Proposition 3: SSB Asymptotic fidelity

Proposition E.4.1 (SSB Asymptotic relative fidelity)**.** *Any (Δ, δ)-SSB system with codeword size M' is $M'\zeta(\Delta, \delta)$-faithful with*

$$\zeta(\Delta, \delta) \xrightarrow{L\to\infty} \frac{1}{2}\left[\left(1 + \mu^2_{(\Delta,\delta)}\right)\sigma^{-2}_{(\Delta,\delta)} + \log\sigma^2_{(\Delta,\delta)} - 1\right], \tag{20}$$

if $\mathbf{U}$ *is chosen such that the Lyapunov condition holds.*

Proof. This is a direct consequence of Proposition E.4 applied to Def.B.9 and observing that a standard Gaussian distribution is still a standard Gaussian in watermarked space since columns of $\mathbf{U}$ sum to 1. □

E.5 Proposition 4

Proposition E.5.1 (SSB Watermarking security)**.** *Let $\alpha = \frac{M'}{L}$. Any (Δ, δ)-SSB is η-secure against a PCA estimator where $\eta = \left(\frac{1-\sqrt{\alpha}\sigma_{\Delta,\delta}}{1-\sigma_{\Delta,\delta}}\right)^2$.*

Proof. SSB 's security relies on the eigenvalues of its covariance matrix, and on whether the number of observations is sufficient to detect a shift from the expected Marchenko-Pastur distribution. First, we compute the covariance matrix $\Sigma_{\mathbf{z}}$.

$$\begin{aligned}
\Sigma_{\mathbf{z}} &= \mathbb{E}[\mathbf{z}\mathbf{z}^\top] && (21)\\
&= \mathbb{E}[\left((I_L - \mathbf{U}\mathbf{U}^\top)\mathbf{z} + \mathbf{U}\mathbf{z_u}\right)\left((I_L - \mathbf{U}\mathbf{U}^\top)\mathbf{z} + \mathbf{U}\mathbf{z_u}\right)^\top] && (22)\\
&= \mathbb{E}[\mathbf{U}\mathbf{z_u}\mathbf{z_u}^\top\mathbf{U}^\top] + \mathbb{E}[\mathbf{U}\mathbf{z_u}\mathbf{z}_o^\top]\mathbb{E}[\mathbf{z}_o\mathbf{z_u}^\top\mathbf{U}^\top] + \mathbb{E}[\mathbf{z}_o\mathbf{z}_o^\top] && (23)\\
& && (24)
\end{aligned}$$

with $\mathbf{z}_o = (I_L - \mathbf{U}\mathbf{U}^\top)\mathbf{z}$. $\mathbf{U}\mathbf{z_u}$ and $\mathbf{z}_o$ are independent, so $\mathbb{E}[\mathbf{U}\mathbf{z_u}\mathbf{z}_o^\top] = \mathbb{E}[\mathbf{z}_o\mathbf{z_u}^\top\mathbf{U}^\top] = 0$. Let's compute $\mathbb{E}[\mathbf{z}_o\mathbf{z}_o^\top]$.

$$\begin{aligned}
\mathbb{E}[\mathbf{z}_o\mathbf{z}_o^\top] &= \mathbb{E}[(I_L - \mathbf{U}\mathbf{U}^T)\mathbf{z}\mathbf{z}^\top(I_L - \mathbf{U}\mathbf{U}^T)] && (25)\\
&= (I_L - \mathbf{U}\mathbf{U}^T)\mathbb{E}[\mathbf{z}\mathbf{z}^\top](I_L - \mathbf{U}\mathbf{U}^T) && (26)\\
&= (I_L - \mathbf{U}\mathbf{U}^T)(I_L - \mathbf{U}\mathbf{U}^T) && (27)\\
&= I_L - \mathbf{U}\mathbf{U}^T. && (28)\\
& && (29)
\end{aligned}$$

It remains to calculate $\mathbb{E}[\mathbf{U}\mathbf{z_u}\mathbf{z_u}^\top\mathbf{U}^\top]$.

$$\mathbb{E}[\mathbf{U}\mathbf{z_u}\mathbf{z_u}^\top\mathbf{U}^\top] = \mathbf{U}\mathbb{E}[\mathbf{z_u}\mathbf{z_u}^\top]\mathbf{U}^\top, \tag{30}$$

where $\mathbb{E}[\mathbf{z_u}\mathbf{z_u}^\top]$ is the covariance matrix of $\mathbf{z_u}$. As $\mathbf{z_u}$ are independent, $\mathrm{Cov}(\mathbf{z_u}) = \mathbb{V}(\mathbf{z_u})I_{M'}$.
So, $\Sigma_{\mathbf{z}}$ is given by:

$$\begin{aligned}
\Sigma_{\mathbf{z}} &= \mathbf{U}\mathbb{E}[\mathbf{z_u}\mathbf{z_u}^\top]\mathbf{U}^\top + \mathbb{E}[\mathbf{z}_o\mathbf{z}_o^\top] && (31)\\
&= I_L + \left(\mathbb{V}(\mathbf{z_u}) - 1\right)\mathbf{U}\mathbf{U}^\top && (32)
\end{aligned}$$

So eigenvalues are $\lambda_0 = 1$ for $(L - M')$ dimensions and $\lambda_1 = \mathbb{V}(\mathbf{z_u})$ for M' dimensions. In a case of Gaussian sampling such that $\mathbf{z} \sim \mathcal{N}(0, I_L)$, the eigenvalues follows a Marchenko-Pastur distribution with support $S = \left[\left(1 - \sqrt{\frac{L}{N}}\right)^2, \left(1 + \sqrt{\frac{L}{N}}\right)^2\right]$, with N the number of observations. Following the Marchenko-Pastur theory, the support of eigenvalues in watermarking sampling is $S_{\lambda_0} = \left[\left(1 - \sqrt{\frac{L-M'}{N}}\right)^2, \left(1 + \sqrt{\frac{L-M'}{N}}\right)^2\right]$ for $(L - M')$ dimensions and

$S_{\lambda_1} = \left[\lambda_1\left(1-\sqrt{\frac{M'}{N}}\right)^2, \lambda_1\left(1+\sqrt{\frac{M'}{N}}\right)^2\right]$ for M' dimensions. A watermarking scheme is η-secure if $S_{\lambda_1} \subseteq S$. It gives the following condition:

If $\lambda_1 > 1$**:**

$$\lambda_1\left(1+\sqrt{\frac{M'}{\eta}}\right)^2 = \left(1+\sqrt{\frac{L}{\eta}}\right)^2 \tag{33}$$

$$\sqrt{\lambda_1}\left(1+\sqrt{\frac{M'}{\eta}}\right) = \left(1+\sqrt{\frac{L}{\eta}}\right) \tag{34}$$

$$\sqrt{\lambda_1} + \frac{\sqrt{\lambda_1 M'}}{\sqrt{\eta}} = 1 + \frac{\sqrt{L}}{\sqrt{\eta}} \tag{35}$$

$$\sqrt{\lambda_1} - 1 = \frac{\sqrt{L} - \sqrt{\lambda_1 M'}}{\sqrt{\eta}} \tag{36}$$

$$\sqrt{\eta}(\sqrt{\lambda_1} - 1) = \sqrt{L} - \sqrt{\lambda_1 M'} \tag{37}$$

$$\sqrt{\eta} = \frac{\sqrt{L} - \sqrt{\lambda_1 M'}}{\sqrt{\lambda_1} - 1} \tag{38}$$

$$\eta = \left(\frac{1-\sqrt{\alpha}\sigma_u}{1-\sigma_u}\right)^2, \tag{39}$$

with $\alpha = \frac{M'}{L}$ and $\sigma_u = \sqrt{\lambda_1}$.

If $\lambda_1 \leq 1$**:**

$$\lambda_1\left(1-\sqrt{\frac{M'}{\eta}}\right)^2 = \left(1-\sqrt{\frac{L}{\eta}}\right)^2 \tag{40}$$

$$\tag{41}$$

So the proof is analogous.

Then, we distinguish the two extreme cases for computing the eigenvalues. We know that $\lambda_1 = \mathbb{V}(\mathbf{z_u})$.

$$\mathbb{E}[\mathbf{z_u}] = 2\text{sign}(c)\sum_{k=-\infty}^{+\infty}(\Phi(b_k)-\Phi(a_k))\frac{\int_{\alpha_k}^{\beta_k} x\phi(x)dx}{\Phi(\beta_k)-\Phi(\alpha_k)} \tag{42}$$

$$= 2\text{sign}(c)\sum_{k=-\infty}^{+\infty}(\Phi(b_k)-\Phi(a_k))\frac{\phi(\alpha_k)-\phi(\beta_k)}{\Phi(\beta_k)-\Phi(\alpha_k)} \tag{43}$$

$$= \text{sign}(c)\sum_{k=-\infty}^{+\infty}P_k\frac{\phi(\alpha_k)-\phi(\beta_k)}{\Phi(\beta_k)-\Phi(\alpha_k)}, \tag{44}$$

$$\tag{45}$$

with $P_k = 2(\Phi(b_k) - \Phi(a_k))$.

and,

$$\mathbb{E}[\mathbf{z}_\mathbf{u}^2] = 2 \sum_{k=-\infty}^{+\infty} (\Phi(b_k) - \Phi(a_k)) \frac{\int_{\alpha_k}^{\beta_k} x^2 \phi(x) dx}{\Phi(\beta_k) - \Phi(\alpha_k)} \tag{46}$$

$$= 2 \sum_{k=-\infty}^{+\infty} (\Phi(b_k) - \Phi(a_k)) \frac{\Phi(\beta_k) - \beta_k \phi(\beta_k) - \Phi(\alpha_k) + \alpha_k \phi(\alpha_k)}{\Phi(\beta_k) - \Phi(\alpha_k)} \tag{47}$$

$$= \sum_{k=-\infty}^{+\infty} P_k \left(1 - \frac{\beta_k \phi(\beta_k) - \alpha \phi(\alpha_k)}{\Phi(\beta_k) - \Phi(\alpha_k)}\right), \tag{48}$$

$$\tag{49}$$

with $P_k = 2(\Phi(b_k) - \Phi(a_k))$.

Then, the covariance is given by:

$$\mathrm{Cov}(\mathbf{z_u}) = I_L \mathbb{V}(\mathbf{z_u}) \tag{50}$$

$$= I_l \left(\mathbb{E}[\mathbf{z}_\mathbf{u}^2] - \mathbb{E}[\mathbf{z_u}]^2\right) \tag{51}$$

$$= \mathbf{I}_L \left(\sum_{k=-\infty}^{+\infty} P_k \left(1 - \frac{\beta_k \phi(\beta_k) - \alpha \phi(\alpha_k)}{\Phi(\beta_k) - \Phi(\alpha_k)}\right) - \boldsymbol{\mu}_u^2 \right), \tag{52}$$

with $\boldsymbol{\mu}_u^2 = \mathbb{E}[\mathbf{z_u}]^2$. □

F Perfect security of SSB

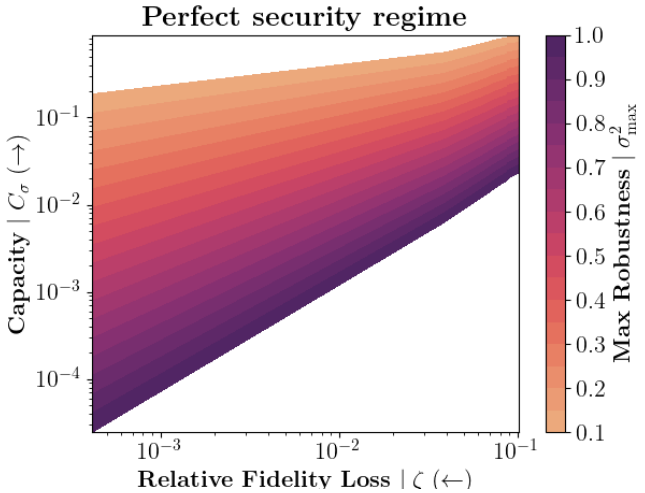


Figure 6: Perfect security regime with $(1.6, 0)$-SSB.

An interesting behavior of SSB can be observed for the security ratio in Figure 3 around $\Delta = 1.6$ when using the finest lattice ($\delta \to 0$). The observed peak is due to Proposition 4, which states that if the eigenvalues of the watermarked covariance matrix are all 1, then one cannot estimate the secret rotation $\mathbf{U}$, no matter how many samples we provide to Eve. A surprising fact is that SSB can achieve perfect security for non-zero capacities. Even better, for any $\Delta < 1.6$, one can find a corresponding $\delta > 0$ for which it achieves perfect security. One further advantage of the perfect security regime is that it can be studied independently of the message size M and of the latent space dimension L: both fidelity and capacity are computed per latent element. We plot the watermarking characteristic under this regime in Figure 6.

Method	Max. codeword size	Identity	Contrast 2.0	Brightness 0.2	Center Crop 0.5	JPEG QF50	JPEG QF80
Trustmark (PSNR$=42$dB)	100	0.949 / 94	0.608 / 60	0.894 / 89	0.000 / 0	0.898 / 89	0.935 / 93
Videoseal (PSNR$=42$dB)	256	0.997 / 255	0.883 / 226	0.974 / 249	0.015 / 3	0.794 / 203	0.845 / 216
Stable Signature	48	0.750 / 36	0.715 / 34	0.722 / 34	0.186 / 8	0.359 / 17	0.472 / 22
$(1.6, 0)$-SSB	$L (= 8192$, Sana)	0.612 / 5013	0.228 / 1867	0.460 / 3768	0.000 / 0	0.267 / 2187	0.419 / 3432
$(-\infty, +\infty)$-SSB	$L (= 8192$, Sana)	0.409 / 3350	0.242 / 1982	0.346 / 2834	0.000 / 0	0.255 / 2088	0.329 / 2695

Table 2: Comparison of the Rate and maximum codeword size at Shannon capacity between SSB and other modern watermarking schemes.

		VAE Attack		Purification (Sana)			Purification (SD2)			WmForger (lr$=0.05$)		
Method	Identity	SD2	Sana	2 steps	4 steps	10 steps	4 steps	8 steps	16 steps	100 steps	200 steps	300 steps
Gaussian-Shading	0.026	0.225	0.178	0.164	0.126	0.004	0.005	0.000	0.000	0.015	0.012	0.000
$(1.6, 0)$-SSB	0.414	0.348	0.300	0.295	0.253	0.093	0.156	0.086	0.029	0.300	0.255	0.226
$(-\infty, +\infty)$-SSB	0.601	0.442	0.326	0.319	0.236	0.030	0.086	0.028	0.005	0.308	0.215	0.165

Table 3: Equivalent white Gaussian noise variance σ^2 introduced in latent space for different adversarial image perturbations.

G Additional results

G.1 Comparison with post-hoc watermarking

Capacity Table2 compares the rate and the maximum codeword size at that rate between SSB and modern watermarking methods. We tested it on two modern post-hoc approaches, TrustMark [7] and VideoSeal [16], as well as Stable-Signature [15], which embeds the watermark after diffusion during VAE decoding. The maximum payload size of these methods is fixed for a given encoder/decoder training. SSB and other seed-based approaches can embed codewords of arbitrary size up to the latent length. The advantage of SSB is that, despite lower capacity, its payload size lets it hide longer messages than other post-hoc or VAE-based watermarking schemes.

Security Other modern watermarking methods do not use a secret key. Under classical watermarking security principles, a method without a secret key cannot provide any security against an adversary who knows the scheme. Indeed, for the modern schemes, the codeword can be retrieved as soon as Eve is provided with the decoder – which is always the case under Kerckhoff's principle. This means that she can copy or erase the message freely. Consequently, the security ratio η of all these schemes equals $1/L$ – a single watermarked observation is sufficient to estimate the codeword.

Fidelity Both VAE-based and post-hoc watermarking schemes alter content after the diffusion process. This means there exists a counterfactual, non-watermarked version against which one can measure distortion or quality loss. In this case, classic watermarking methods define quality as distortion in pixel space using PSNR. This is one advantage of the post-hoc approach: one can work at a fixed PSNR by post-processing the watermarking signal. Note that in the specific case of post-hoc schemes, the notion of quality is then theoretically well-defined. For seed-based methods, there is no original content; hence, our definition of fidelity is that we do not try to preserve a piece of original content (i.e., quality) but instead try to be faithful to the distribution of content induced by the content generator (fidelity).

G.2 Adversarial robustness

To complete the analysis of Table 1, we report in Table 3 the equivalent Gaussian noise, computed with the same methodology, for three adversarial methods created to erase the watermark: VAE purification [24], Purification [24], and WM Forger [31].

G.3 Additional empirical validation

Figure 7 provides empirical validation for Z-Image and QWEN using the same experimental protocol as in Figure 4. For Z-Image, we observe a gap between the theoretical expectation and the empirical validation for the brightness due to clipping in the (1.6, 0)-SSB regime. For QWEN, the empirical validation matches theory except for Gaussian-Shading and GS-SSB, where the methods

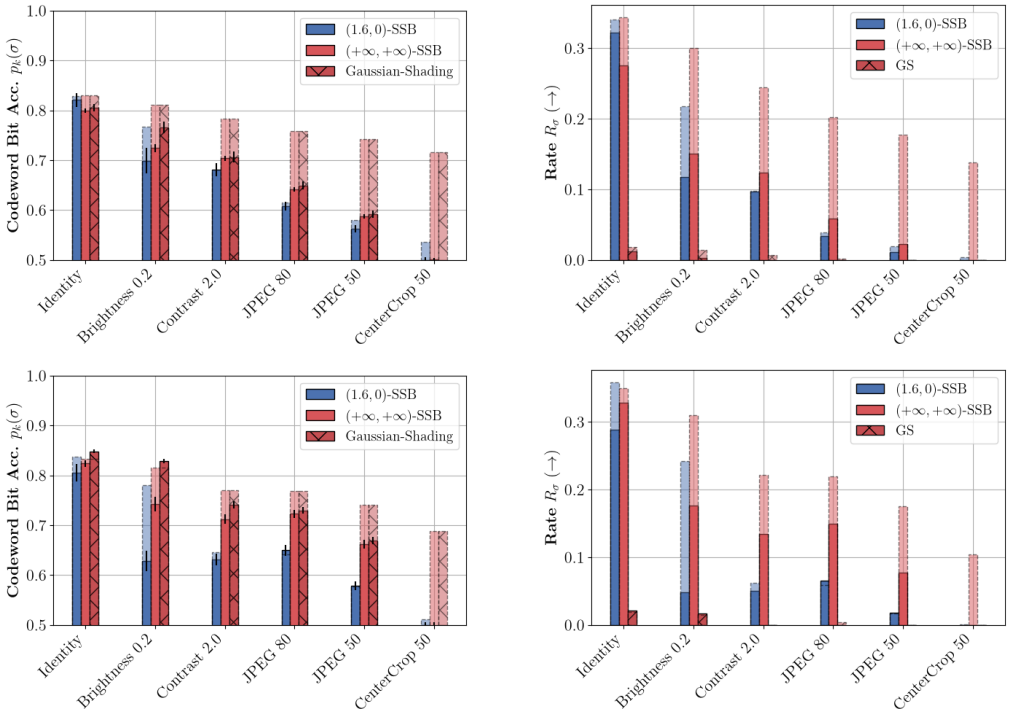


Figure 7: Empirical validation of the model defined in Section 2 for current seed-based approaches – Gaussian-Shading and PRC – as well as our method SSB . Transparent boxes correspond to theoretical values, opaque to empirical ones. We display 95% asymptotic confidence intervals. We perform this validation on Z-Image and QWEN, respectively.

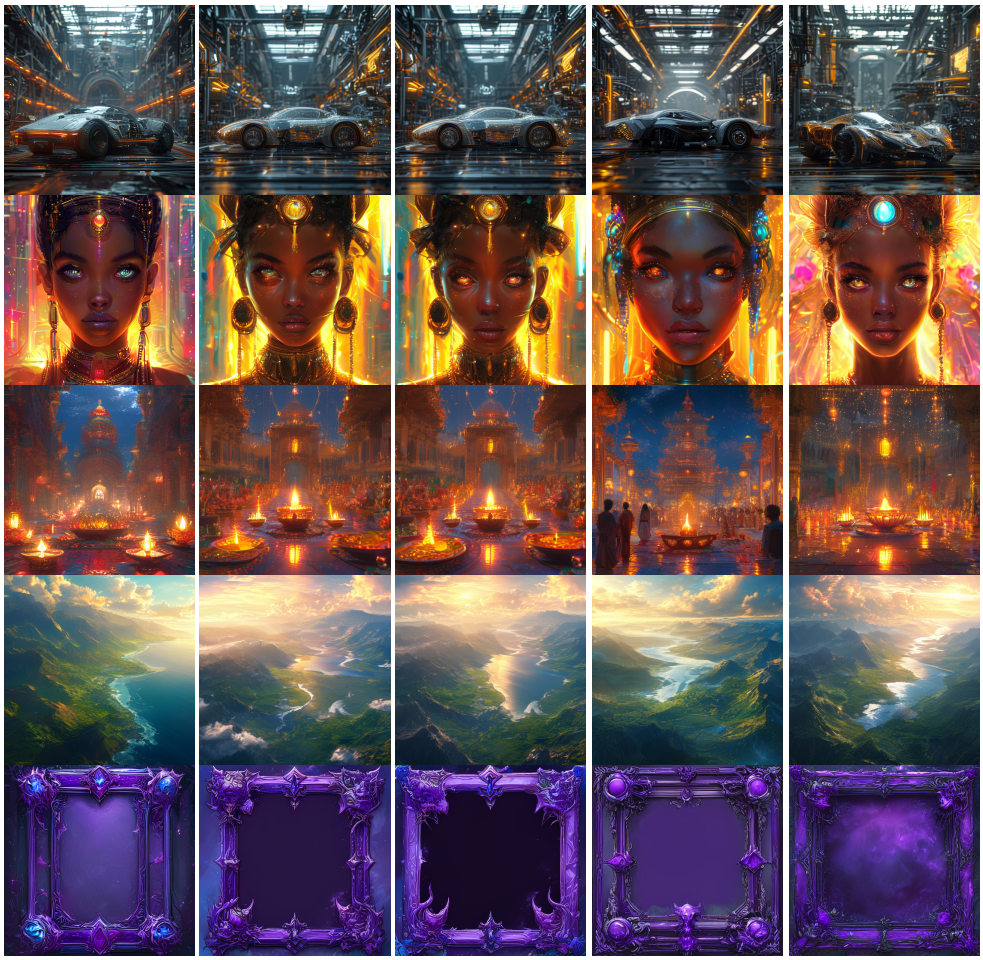
underperform expectations on the JPEG QF=50 augmentation. We observe the same discrepancy for all the transforms with Z-Image in the sign-based detection regimes. However, we don't observe this discrepancy for SANA, despite using the same inversion scheduler as QWEN. Since this phenomenon occurs only with the sign lattice, we conclude that the mismatch is explicitly introduced in correlations between the signs of latent elements. Other lattice configurations are less sign-dependent. This is a combination of very strong JPEG compression, which tends to correlate content due to quantization, and VAE compression. Sana uses a very different VAE architecture than Qwen and ZiT, so we tend to pinpoint the difference there. This shows that correlations within the latent space are where the model breaks down.

H Examples

Figure 8 illustrates examples of images generated with Sana and the compared multi-bit seed-based watermarking methods.

I Reproducibility Statement

All models and datasets are detailed in Appendix J. All the simulations can be executed in the notebook provided as a supplementary file. The full code used for the experimental validation is available at `https://github.com/EnoalG/Secure-Seed-Based-Multi-bit-Watermarking-for-Diffusion-Models-from-First-Principles-`. The implementation relies on PyTorch and Diffusers libraries, with fixed random seeds. Section 4 provides all hyperparameters used to achieve the results presented in this work. We ran experiments on NVIDIA L40S GPUs for 20 hours.



No watermark $(1.6, 0)$-SSB $(+\infty, +\infty)$-SSB Gaussian-Shading PRC

Figure 8: Example of images generated with Sana. Images were selected randomly.

J Licenses

Datasets

Stable-Diffusion-Prompts

- Source: `https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts`
- License: Unknown

Pretrained Diffusion Models

Sana

- Source: `https://huggingface.co/Efficient-Large-Model/Sana_600M_512px`
- License: NVIDIA License

Z-image

- Source: `https://huggingface.co/Tongyi-MAI/Z-Image`
- License: apache-2.0

Qwen-Image

- Source: https://huggingface.co/Qwen/Qwen-Image
- License: apache-2.0

Augmentations

Augly

- Source: https://github.com/facebookresearch/AugLy
- License: MIT License