Department of Computer Science

Submitted in part fulfilment for the degree of BSc.

Identifying Images Generated by AI using Watermarks

Mischa Zaynchkovsky

21st July 2025

Supervisor: Dimitar Kazakov

To my parents for their love and support, and to my cat Orie.

Acknowledgements

I would like to thank my supervisors, Dimitar Kazakov and Dr. Kofi Appiah, for their guidance and support throughout this project.

Contents

List of Figures

List of Tables

List of Acronyms

AI Artificial Intelligence

GAIM Generative AI Image Model

VAE Variational Autoencoder

GAN Generative Adversarial Network

DM Diffusion Model

LDM Latent Diffusion Model

RGB Red, Green, Blue

LSB Least Significant Bit

DCT Discrete Cosine Transform

DFT Discrete Fourier Transform

DWT Discrete Wavelet Transform

LPM Log-Polar Mapping

MSE Mean Squared Error

PSNR Peak Signal-to-Noise-Ratio

SSIM Structural Similarity Index Measure

NCC Normalised Cross-Correlation

BER Bit Error Rate

FID Fréchet Inception Distance

BCH Bose-Chaudhuri-Hocquenghem

CDF Cumulative Distribution Function

FPR False Positive Rate

TPR True Positive Rate

ANOVA Analysis of Variance

HSD Honestly Significant Difference

CLIP Contrastive Language-Image Pre-training

JPEG Joint Photographic Experts Group

CUDA Compute Unified Device Architecture

VRAM Video Random Access Memory

C2PA Coalition for Content Provenance and Authenticity

DDIM Denoising Diffusion Implicit Models

Executive Summary

The proliferation of powerful Generative AI Image Models (GAIMs) such as Stable Diffusion [1] and DALL-E marks a significant technological leap, yet their capability to produce realistic synthetic images introduces notable moral, ethical and legal challenges. These include the further widespread presence of easily created misinformation, large scale copyright infringement, and an erosion of trust in digital content, prompting regulatory bodies like the European Union to mandate traceability mechanisms for image generation processes through their Artificial Intelligence (AI) act [2]. This project discusses and evaluates the urgent need for a robust, imperceptible watermarking solution integrated directly within the AI generation process to ensure authenticity, attribution, and traceability.

The primary aim of this work is to conduct a rigorous, independent evaluation of the Gaussian Shading watermarking algorithm, a state of the art, training-free technique for Latent Diffusion Models (LDMs). This method was selected over alternatives for its plug-and-play implementation and its purported superior performance and result, which this project seeks to explore and assess. The methodology involved implementing the algorithm within a standard Stable Diffusion v2.1 framework and executing a well defined evaluation protocol aligned with established academic benchmarks. The watermark's imperceptibility was assessed using distribution-based metrics (Fréchet Inception Distance (FID) and Contrastive Language-Image Pre-training (CLIP) Score) across 1000 images generated using the MS-COCO dataset [3]. Its robustness was quantified by the Bit Error Rate (BER) for a standard 256-bit payload after subjecting 200 watermarked images, generated using prompts from the Gustavosta prompt set [4], to a suite of 12 common digital attacks, including JPEG compression, noise, blurring, and cropping. These two distinct datasets were used as the former is a standard benchmark for evaluating generative models, while the latter is a diverse set of prompts more suited to test the watermark's robustness across diverse semantic content.

The evaluation revealed the implementation of the method led to clear strengths and limitations through statistical analysis. The Gaussian Shading method demonstrated exceptional robustness against signal processing attacks, with True Positive Rate (TPR) values consistently above 98% across all signal processing conditions and maintaining a BER of less than 2% even under severe conditions like JPEGs lossy compression at a quality factor of 25 (BER 1.35%, TPR 98.65%) and significant Gaussian blur (BER 1.8%, TPR 98.2%). Statistical analysis using one-way Analysis of Variance (ANOVA) confirmed performance differences across varying attack categories ($F(4, 2395) = 15064.930$, $p < 0.001$), with brightness attacks showing superior robustness compared to all other categories. The method also demonstrated high imperceptibility performance at a perceptual level, with a negligible difference in CLIP scores (-0.0008) between the original

and watermarked image sets, confirming that it preserves the visual quality and semantic fidelity of the generated images to their text prompts.

However, the evaluation also demonstrated clear limitations. A vulnerability to geometric attacks was evident; a random crop retaining 80% of the image area resulted in a BER of 48.52% and TPR of 51.48% with 95% confidence interval $[51.02\%, 51.94\%]$, statistically indistinguishable from random chance. Furthermore, the evaluation yielded an FID score of 47.83, which is above the success threshold of 30, thus indicating a measurable statistical shift in the output distribution and challenging the absoluteness of Gaussian Shadings performance losslessness. All statistical findings are supported by 95% confidence intervals, ensuring robust statistical validation. This research considers legal, social, ethical, and professional issues by investigating an emerging technology that can mitigate AI-driven misinformation and support artists rights. The work was conducted entirely on public datasets and open source models, involving no human participants or personal data.

By providing a detailed, independent performance analysis, this dissertation makes a valuable contribution to the field of responsible AI. It concludes that while Gaussian Shading is a high-capacity, easily deployable solution for environments where signal processing attacks are the primary concern. However, its vulnerability to geometric transformations present challenges. A more exhaustive evaluation of the Gaussian Shading method is required to address the limited reach of the this project. This includes use of a more extensive set of prompts and datasets, and a more comprehensive set of digital attacks. This would allow for a more thorough assessment of the method's performance and its suitability as a practical watermarking solution for LDMs.

The project highlights the importance of a multi-faceted evaluation and the key area of geometric attacks as a significant area for future research in the highly topical field of AI image watermarking.

Statement of Ethics

This statement outlines the ethical considerations for this project, structured around the following principles: Avoidance of Harm, Informed Consent, and Data Protection.

Avoidance of Harm

This project aims to explore technology that mitigates the potential harms AI-generated images can cause. By enabling robust watermarking, this work helps combat misinformation (e.g., deepfakes), provides a mechanism for artists to receive attribution for their work, and aligns with regulatory demands for transparency in synthetically generated media, such as the European Union AI Act [2].

Potential risks have been considered and mitigated. The risk of malicious actors reverse engineering the watermark is addressed by focusing on robustness evaluation rather than vulnerability exploitation. The potential for misuse regarding privacy/surveillance is acknowledged; however, the project's scope is strictly limited to authenticity and attribution, not tracking individuals.

Informed Consent

This project does not involve human participants. All experiments are conducted using publicly available datasets and open source software, adhering to their respective licensing terms (e.g., Creative Commons for MS-COCO [3], MIT License for Gaussian Shading source code [5]). The research ensures transparency and reproducibility by documenting all methodologies and making evaluation scripts available.

Data Protection

No personal or sensitive data was collected, processed, or stored. The datasets used (MS-COCO and Gustavosta prompts [4]) are publicly available for research and contain no personally identifiable information. All generated images and research data will only be retained as long as necessary to ensure reproducibility. The project adheres to responsible research practices, including minimising computational resources to reduce environmental impact.

Based on these considerations, formal ethical approval from the Physical Sciences Ethics Committee (PSEC) was not required. All ethical aspects have been discussed with the project supervisor, and the completed ethics checklist is available in Appendix B.

1 Introduction

1.1 Motivation and Problem Statement

The proliferation of powerful Generative AI Image Models (GAIMs) such as Stable Diffusion [1], DALL-E [6], and Midjourney [7] marks a significant technological leap, unlocking new possibilities in fields like content creation, design prototyping, and automation. However, the ease with which these models can generate highly realistic synthetic media introduces critical, ethical and practical challenges, including the spread of misinformation [8], copyright infringement, and a general erosion of trust in digital content. Furthermore, current rulings by the United States Copyright Office render purely AI-generated images ineligible for copyright protection [9].

In response, a clear consensus is forming around the need for establishing authenticity, attribution, and traceability mechanisms for synthetic media. Governmental bodies, including the White House through its 2023 Executive Order on AI [10] and the European Union via its 2024 AI Act [2], have mandated the use of techniques like invisible watermarking for synthetic content. Consequently, leading technology companies have begun integrating such solutions; notable examples include Google's SynthID [11], Microsoft's watermarking in Bing Image Creator [12], and Stability AI's watermarking methods for its models [13].

The urgency for effective watermarking is further underscored by concerns surrounding the data used to train GAIMs. Artists have voiced concerns about the unauthorised use of their work [14], leading to AI-generated images that mimic their unique styles without credit or compensation [15]. This has fuelled the development of defensive techniques like Glaze [16] and Nightshade [17], which disrupt model training [18], highlighting the creative community's demand for control. While digital watermarking has a long history [19], its application to generative AI presents new challenges and requires solutions that are integrated directly into the generation process. This project addresses the critical need for such a method, one that is imperceptible, robust against manipulation, capable of carrying attribution data, all without degrading image quality, and integrated within the generation process itself.

1.2 Project Aims & Scope

The primary aim of this project is to investigate and evaluate the integration of digital watermarking within AI image generation models to address the critical issues of **authenticity, attribution, traceability, and copyright protection**. To this end, the project will conduct a thorough evaluation of the Gaussian Shading watermarking algorithm, a state of the art, training-free technique designed for Latent Diffusion Models (LDMs).

To achieve this, the following objectives have been set:

1. To conduct a focused review of modern watermarking techniques to identify the current state of the art, justify the selection of Gaussian Shading, and its position within the landscape of digital watermarking.
2. To implement the Gaussian Shading algorithm within a standard GAIM framework, specifically Stable Diffusion v2.1.
3. To design and execute a comprehensive evaluation framework to test the watermark's imperceptibility and its robustness against a suite of common digital attacks.
4. To analyse the results, assessing the crucial trade-offs between robustness, imperceptibility and data capacity, then compare against published benchmarks to critically assess the performance and viability of Gaussian Shading as a practical watermarking solution.

This project seeks to answer the following research questions:

- To what extent can the Gaussian Shading method provide robust watermarking against a standard set of digital attacks, including compression, noise, and geometric transformations?
- What is the trade-off between watermark robustness and image quality? Can the method's performance lossless claim be substantiated using distribution based metrics like Fréchet Inception Distance (FID) and Contrastive Language-Image Pre-training (CLIP) score?
- How does the performance of Gaussian Shading compare to other leading in-generation watermarking techniques?

The key contribution of this project is a rigorous and independent empirical evaluation of the Gaussian Shading watermarking method. By providing detailed performance data and a direct comparison to established benchmarks, this work offers valuable insights for the development of responsible and traceable AI generated image generation systems.

1.3 Dissertation Structure

The remainder of this dissertation is organised as follows: Chapter 2 reviews the literature on generative models and watermarking, focusing on the state of the art techniques that ultimately led to the selection of Gaussian Shading. Chapter 3 details the methodology, system architecture, and the specific framework used for implementation and evaluation. Chapter 4 presents the results of the evaluation, analysing the imperceptibility and robustness of the implemented solution. Finally, Chapter 5 concludes the dissertation, summarising the findings in relation to the research questions, discussing the project's limitations, and potential directions for future work.

2 Literature Review

2.1 Background

2.1.1 Generative Models in AI Image Generation

Generative images have revolutionised the AI field by enabling the creation of new data that closely resembles the training data. The three primary generative models used in AI image generation being: Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion Models (DMs).

Variational Autoencoders (VAEs)

VAEs were first defined in 2013 by Kingma et al. [20] and Rezende et al. [21]. VAEs are probabilistic generative models that learn a latent space representation of input data. They consist of an encoder and a decoder, which work together to reconstruct input data from a compressed latent space. Watermarking in VAEs could involve perturbing latent space to insert information [22].

Generative Adversarial Networks

GANs, proposed by Goodfellow et al. [23] in 2014. GANs consist of two neural networks: A generator, and a discriminator. The generator takes an input of random noise and generates an image by reassembling the real data distribution; The discriminator seeks to differentiate between real and generated images. The competing nature of the model helps in improving the quality of images generated over time. However, images generated are sensitive to perturbations [24]. Therefore, adding a watermark could degrade the quality of the image.

Diffusion Models

DMs were first introduced in 2015 by Sohl-Dickstein et al. [25] and popularised in 2020 by Ho et al. [26]. Unlike GANs and VAEs, DMs generate images by iteratively denoising a random noise pattern, producing high quality images with fine details. The challenge in watermarking DMs lies in ensuring the watermark does not interfere with the denoising process in order to preserve image quality. LDMs such as Stable Diffusion, extend the concept of diffusion by performing the denoising process in a latent space rather than directly in pixel space, allowing for more efficient training and inference, as well as improved image quality [1]. In LDMs, a VAE is used to compress the image to into a lower-dimensional latent space, which is then denoised iteratively to generate the final image.

2.1.2 Applications and Motivations for Watermarking in the AI Era

The rise of GAIMs has created an urgent need for robust watermarking solutions to address several key challenges.

Preventing Training Data Contamination and Model Collapse
Many GAIMs are trained on large datasets scraped from the internet. A significant risk in this process is "model collapse", where models are inadvertently trained on their own synthetic outputs or those from other models. This can lead to a feedback loop that degrades the quality and diversity of generated content over time [27]. Watermarking AI generated images can mitigate this by marking these synthetic outputs, allowing them to be filtered out of future training datasets, thereby preserving the integrity of the training data.

Attribution, Traceability, and Content Authenticity
Watermarks can act as digital fingerprints for AI generated images, enabling robust attribution and traceability. Depending on the implementation, a watermark could encode information about the originating model, the user, or the time of generation. This is vital for accountability, particularly when addressing the use of copyrighted material in training data or the distribution of malicious content. Initiatives like the Coalition for Content Provenance and Authenticity (C2PA), backed by major technology firms, aim to standardise how this provenance information is embedded and read, creating a framework for digital content certification [28]. An effective watermarking scheme is therefore essential for copyright protection, verifying ownership, and ensuring the authenticity of digital media in an age of prolific AI generation [29].

2.2 Digital Watermarking Techniques

2.2.1 Traditional Spatial Domain Watermarking

Spatial domain watermarking involves embedding watermarks directly into the pixel values. These methods are straightforward, easy to implement, and computationally efficient. However, are typically less resistant to attacks such as compression and transformations.

Least Significant Bit (Least Significant Bit (LSB)) Modification
A common technique to embed a watermark information into randomly chosen pixels' LSB. The LSB is changed as to not affect the image quality as it contains less important information. However, it is trivial for an attacker to change all LSB bits to 1 to modify the watermark. To address the problems with LSB watermarking, improvements have been made. One such improvement embeds data not only to the LSB but also higher planes. Moreover, a 2-3-3 embedding technique [30] distributes the watermark across the Red, Green, Blue (RGB) channels of a pixel. This approach results in minimal perceptual distortion while achieving better embedding capacity and robustness.

Patch-based or Block-Based techniques
Proposed by Bender et al. [31] This method involves randomly picking n pairs of image points A, B where the image data in A is darkened, while is brightened in B. This method offers decent robustness in exchange for

capacity. [32]

2.2.2 Traditional Frequency (Transform) Domain Watermarking

These techniques embed watermark information within the frequency domain of an image after a transformation. This approach spreads the watermark information throughout the image in ways that are less perceptible to the human eye and more resilient to common attacks compared to spatial methods.

Discrete Cosine Transform

Discrete Cosine Transform (DCT) watermarking embeds watermark information into an images frequency coefficients after transforming it from the spatial to the frequency domain. This leverages energy compaction, where the majority of an image's visual information is represented by lower-frequency coefficients, while higher-frequency coefficients capture finer image details. A common approach is block-based DCT, where the image is divided into smaller non-overlapping blocks, DCT is then applied to each block. Mid-frequency coefficients are typically chosen, balancing imperceptibility and robustness. Modifying low-frequency coefficients can lead to more noticeable distortions, while high-frequency coefficients are more susceptible to compression and noise attacks. Block-based DCT is particularly suitable for JPEG compression, a prevalent image compression technique which is also block-based [33]. By embedding watermarks in DCT coefficients compatible with JPEG's compression algorithm, the watermark can survive compression without significant degradation [34]. Alternatively, global DCT applies the transformation to the entire image rather than individual blocks. This offers greater robustness against attacks, but is more computationally intensive and less compatible with block-based compression techniques such as JPEG.
The robustness of DCT-based watermarking comes from the ability to embed data in perceptually significant regions of an image, therefore being less likely to be removed by common image processing operations. However, DCT based watermarking methods struggle with maintaining robustness against geometric attacks such as scaling and rotation due to inherently not accounting for spatial transformations [35]. From this hybrid techniques combining DCT with other transformations have arisen [36].

Discrete Fourier Transform

Similar to DCT watermarking, Discrete Fourier Transform (DFT) embeds watermark information into an images frequency domain by transforming it from the spatial domain to the frequency domain, but using the DFT which decomposes an image into sinusoidal components of varying frequencies, represented as complex-valued coefficients corresponding to magnitude and phase. These coefficients describe the global frequency characteristics of the image, making DFT-based watermarking inherently robust against various image processing operations and certain geometric

transformations.
Log-Polar Mapping (LPM) transforms the image into log-polar coordinates before applying DFT. This mapping converts scaling and rotation into linear translations in the frequency domain, enabling efficient watermark extraction after significant geometric transformations [37].

Discrete Wavelet Transform
A Discrete Wavelet Transform (DWT) is any wavelet transform that decomposes a signal into wavelets, offering local analysis in both the time and frequency domains. Unlike DFT, which analyses global frequency count, and DCT which can operate globally or block-based, DWT inherently supports multi-resolution analysis by examining signals at different scales. This dual localisation makes DWT particularly effective for image watermarking, as it can capture coarse and fine image details simultaneously.

Applications to AI-Generated Images
Traditional frequency domain methods (DCT, DWT) represent conventional image watermarking approaches that operate in the frequency domain. While well-established and widely used for natural images, these techniques are generic post-processing methods that don't specifically leverage the properties of the AI generation processes. Their application remains relatively unexplored in recent LDM-specific literature compared to direct latent modification approaches that are purposefully designed for AI-generated content.

2.2.3 Post-Generation Methods
Methods like HiDDeN [38] and FreqMark [22] demonstrate impressive results, they are designed as post-generation solutions that can be applied to any image, regardless of its source. This makes them less specifically tailored to the unique characteristics and requirements of AI-generated content.

HiDDeN
One promising advancement in watermarking is the HiDDeN framework [38]. HiDDeN leverages the sensitivity of deep neural networks to small perturbations in input images to encode information, making it a robust solution for watermarking.
The HiDDeN framework comprises three main components: an encoder, a decoder, and an adversary network. The encoder receives an image and a message string, outputting an encoded image that incorporates the watermark. The decoder attempts to reconstruct the original message from the encoded image, while the adversary network predicts whether a given image contains an encoded watermark, providing adversarial loss to enhance the quality of the encoded images.
The adversarial training enhances the watermark's resilience against numerous attacks. The deep leaning approach allows for a more flexible watermark embedding,

2.2.4 Modern Watermarking for AI-Generated Images

Recent advancements have focused on integrating watermarks directly into the generative process of DMs and LDMs, offering greater robustness and imperceptibility compared to traditional post-processing methods.

Tree-Ring

Tree-Ring [39] is a notable in-generation technique for DMs. It operates by embedding a watermark signal into the initial noise vector (z_T) before the diffusion process begins. The core idea is that the deterministic nature of Denoising Diffusion Implicit Models (DDIM) inversion allows for the retrieval of the initial noise vector from the final generated image. By comparing the recovered noise with the known watermark signal, the presence of the watermark can be detected. While effective for detection, its original design is primarily for a 1-bit watermark, which limits its viability for use other than binary watermark detection.

Stable Signature

Stable Signature [40] is a watermarking approach designed specifically for LDMs. Instead of modifying the initial noise, this method fine-tunes the model's auto-encoder (VAE). The VAE's decoder is trained to embed a specific watermark pattern into the generated image's pixel space while the encoder is trained to be robust to its presence. By embedding the watermark directly into the model's architecture, Stable Signature aims to create a watermark that is deeply integrated with the image's content and thus more resilient to post-processing attacks.

Gaussian Shading

Gaussian Shading [41] is a training-free, performance-lossless watermarking technique for LDMs. Unlike methods that require model fine-tuning, Gaussian Shading is a plug-and-play solution that modifies the initial latent sampling step. The watermark is mapped to a latent representation that follows a standard Gaussian distribution, making it statistically indistinguishable from a non-watermarked latent vector. This ensures that the watermarking process does not degrade the performance of the generative model. The watermark is embedded by diffusing the bits across the latent dimensions and then using a distribution-preserving sampling method. Extraction is achieved through DDIM inversion to retrieve an estimate of the initial latent, from which the watermark can be recovered. The authors provide theoretical proof of its performance-lossless nature and demonstrate high robustness against common attacks, outperforming many existing methods.

Comparative Analysis

To provide a clear overview, Table 2.1 compares the different watermarking techniques discussed.

Table 2.1: Comparison of Digital Watermarking Techniques.

Technique	Domain	Robustness	Imperceptibility	Capacity	Training-Free
LSB	Spatial	Low	High	High	Yes
DCT	Frequency	Medium	Medium	Medium	Yes
DWT	Frequency	High	High	Medium	Yes
Tree-Ring	Latent (DM)	High	High	Low (1-bit)	Yes
Stable Signature	Latent (LDM)	High	High	High	No
Gaussian Shading	Latent (LDM)	High	High (Provably Lossless)	High	Yes

2.3 Watermarking Optimisations and Enhancements

2.3.1 Perceptual Masking

Perceptual masking exploits the characteristics of human vision by embedding watermarks into regions of an image where the changes will be less noticeable. For example, areas with high texture or edges rather than flat or uniform areas.

2.4 Watermarking Challenges and Evaluation

A successful watermarking scheme must be robust against a variety of attacks designed to remove or degrade the embedded information. Furthermore, its performance must be quantifiable using standard evaluation metrics.

2.4.1 Attacks on Watermarks

Watermarks are susceptible to a wide range of attacks, which can be broadly categorised as follows: **Removal Attacks:** These are designed to completely eliminate the watermark signal from the image. This can include denoising filters or adversarial attacks specifically trained to target and erase the watermark. **Geometric Attacks:** These attacks alter the geometry of the image, which can desynchronise the detector. Common examples include rotation, scaling, cropping, and translation. **Signal Processing Attacks:** These are common image manipulations that can unintentionally degrade or destroy the watermark. This category includes lossy compression (e.g., JPEG), noise addition (e.g., Gaussian noise), and filtering (e.g., blurring).
A robust watermarking system must be able to withstand a combination of these attacks to be considered effective in real-world scenarios.

2.4.2 Metrics for Evaluation

To objectively assess the performance of a watermarking technique, a set of standard metrics is used to measure three key properties:

- **Imperceptibility:** This measures the visual distortion introduced by the watermark. For traditional watermarking methods, commonly quantified using pixel level metrics like Peak Signal-to-Noise-Ratio (PSNR) and Structural Similarity Index Measure (SSIM). However, for latent space

watermarking methods like Gaussian Shading, distribution-based metrics such as FID and CLIP Score are more appropriate. Lower FID values mean the distribution of watermarked images is closer to that of the original images, while similar CLIP scores indicate that the semantic content of the images remain consistent with the original prompts.

- **Robustness:** This is a measure of the watermark's ability to survive attacks. Typically evaluated by calculating the BER between the original and extracted watermark message after an attack has been applied. A lower BER indicates better robustness.
- **Capacity:** This refers to the amount of information (in bits) that can be embedded within the watermark. There is often a trade-off between capacity, robustness, and imperceptibility.

These metrics provide the foundation for the experimental evaluation framework described in the next chapter.

3 Methodology

This chapter details the approach taken to investigate, implement, and evaluate the Gaussian Shading watermarking technique for AI generated images. It covers the research strategy, the selection of the specific watermarking technique from available literature, a detailed experimental design and the comprehensive framework for evaluation. The methodology will outline an assessment process that will enable comparison with state of the art techniques while operating within the feasibility constraints of an undergraduate project.

3.1 Research Approach

This study employs a **quantitative empirical evaluation** methodology to assess the performance of Gaussian Shading watermarking. The approach is informed by established methods in notable watermarking literature, specifically following the evaluation protocols used by Tree-Ring [39] and Stable Signature [40], enabling comparison without requiring reimplementation of competing methods.

The selection of LDMs as the primary architecture is justified by their current dominance in high quality image generation and the availability of open source implementations [42]. Specifically, Stable Diffusion [1] is chosen due to its open source nature [13], widespread adoption in research and existing watermarking literature within its framework [40], [43], [44].

3.2 Watermarking Technique Selection

The primary goal is to embed imperceptible watermarks that facilitate traceability and attribution. Based on the literature review, several approaches are viable, particularly those designed for or adaptable to LDMs. Latent space modification techniques like Stable Signature [40], Tree-Ring [39], LaWa [45], ZoDiac [44], and WMAdapter [46] propose embedding the watermark within the latent space during the image generation process. This approach is specifically designed for AI image generation, as it integrates directly with the generative model's architecture and workflow.

The final selection of the watermarking technique is based on the following criteria:

1. **Suitability for LDM Integration:** How readily the technique can be integrated into the Stable Diffusion architecture.
2. **Robustness Potential:** Theoretical and empirical evidence from literature regarding resistance to common image manipulations.
3. **Capacity for Attribution Data:** Ability to embed a sufficient payload for traceability purposes [29].
4. **Imperceptibility:** Maintaining high visual quality of generated images.
5. **Implementation Feasibility:** Availability of reference implementations or clarity of the proposed algorithm within the project's timeframe.

Based on these criteria, Gaussian Shading [41] was selected. Its approach

modifies the initial latent sampling process, offering direct integration (Criterion 1). A key advantage is its provably performance-lossless nature, meaning it does not require model fine-tuning and aims to preserve the original model's output quality (Criteria 1, 4, 5). The original paper reports high robustness and good capacity (Criteria 2 & 3), making it a strong candidate. While other latent space methods like Stable Signature [40] or Tree-Ring [39] (originally designed primarily for 1-bit capacity, limiting its suitability for detailed attribution data under Criterion 3), also offer strong integration; Gaussian Shading's advantage for this project lies with it being performance-lossless without needing model fine-tuning or architectural changes (Criterion 5). This simplifies implementation and ensures the watermark minimally impacts the generative capabilities of the base Stable Diffusion model (Criterion 4), compared to approaches that might require adjustments to the VAE or U-Net.

3.3 Experimental Design

3.3.1 Standard Evaluation Protocol

To ensure reproducible and comparable results, the evaluation follows a standardised protocol:

- **Generative Model:** Stable Diffusion v2.1-base
- **Image Resolution:** 512×512 pixels (standard for SD v2.1)
- **Guidance Scale:** 7.5 (classifier-free guidance strength)
- **Inference Steps:** 50 (using DPMSolver++ scheduler)
- **Precision:** Float16 for memory efficiency
- **Batch Size:** 1 (for reproducibility and memory constraints)

These parameters align with established benchmarks and ensure compatibility with the RTX 2070 Super GPU constraints (8GB Video Random Access Memory (VRAM)) while maintaining generation quality.

3.3.2 Dataset Selection and Justification

The evaluation employs a two-tier dataset strategy to balance comprehensive assessment with computational feasibility. The primary dataset for imperceptibility and baseline quality analysis is derived from the MS-COCO [3] 2017 validation set captions. From this, a subset of 1000 diverse prompts was curated to ensure a representative sample for calculating the FID score, aligning with standard evaluation practices in the literature.

For the robustness evaluation, a different prompt set was created from the `Gustavosta/Stable-Diffusion-Prompts` [4] collection. This dataset is known for its diverse and challenging prompts. A subset of 200 prompts was selected to cover a wide range of categories, including portraits, landscapes, objects, and abstract concepts. This reduction to 200 prompts was a made to ensure the feasibility of executing the suite of 12 attack configurations on each generated image within the time and hardware constraints of the project, while still providing sufficient data for analysis.

3.3.3 Watermark Configuration

The watermark implementation follows the Gaussian Shading specification with a 256-bit capacity. For empirical testing, the payload consists of random binary strings per generation to facilitate robust BER analysis; in practice, this would contain attribution data (generation ID, model identifier, timestamp). The payload is then encrypted using a ChaCha20 stream cipher with random key/nonce per image before embedding. Standard diffusion parameters (`channel_copy` = 1, `hw_copy` = 8) are used as specified in [41]. The detection threshold targets an False Positive Rate (FPR) of 10^{-6} for statistical significance.

3.4 System Architecture

The Gaussian Shading watermarking system operates through two primary pipelines: embedding and extraction. Figure 3.1 illustrates the complete architecture of the system, highlighting the key components and data flow.

3.4.1 Embedding Pipeline

The **embedding pipeline** integrates watermarking directly into the image generation process:

1. **Text Encoding:** The process begins with a text prompt that is encoded using the CLIP text encoder to produce conditioning vectors.
2. **Watermark Preparation:** The watermark message s (a binary string) is diffused across latent dimensions to obtain s_d, creating redundant copies for robustness. The diffused watermark s_d is then encrypted with a ChaCha20 stream cipher using a secret key K to produce m.
3. **Distribution-Preserving Sampling:** The core innovation occurs during initial latent sampling. Instead of standard sampling $z_T \sim \mathcal{N}(0, I)$, the method uses distribution-preserving sampling based on m using a Gaussian quantile function and uniform random sampling. This ensures the watermarked latent maintains the same statistical properties as a non-watermarked latent.
4. **Denoising Process:** The watermarked latent z_T undergoes standard 50-step diffusion denoising guided by the conditioning vectors.
5. **Image Decoding:** Finally, the VAE decoder transforms the denoised latent representation into the final watermarked image in pixel space.

3.4.2 Extraction Pipeline

The **extraction pipeline** reverses the embedding process to recover the watermark:

1. **Latent Encoding:** The watermarked image is encoded using the VAE encoder to produce a latent representation z'_0.
2. **DDIM Inversion:** A deterministic DDIM inversion process with 50 steps is applied to estimate the initial noise z'_T that would have generated the image.
3. **Watermark Recovery:** Inverse sampling logic (Gaussian Cumulative Distribution Function (CDF)) extracts a randomized watermark estim-

ate m' from z'_T, which is decrypted with key K to recover the diffused watermark estimate s'_d.

4. **Bit Reconstruction:** A reduction/voting mechanism across the diffused copies recovers the final watermark estimate s', enabling binary classification between watermarked and non-watermarked images based on the recovered bits.

The architecture is designed to be modular, allowing for adjustments to parameters such as the number of diffusion steps, watermark length, and diffusion redundancy factors without requiring structural changes.

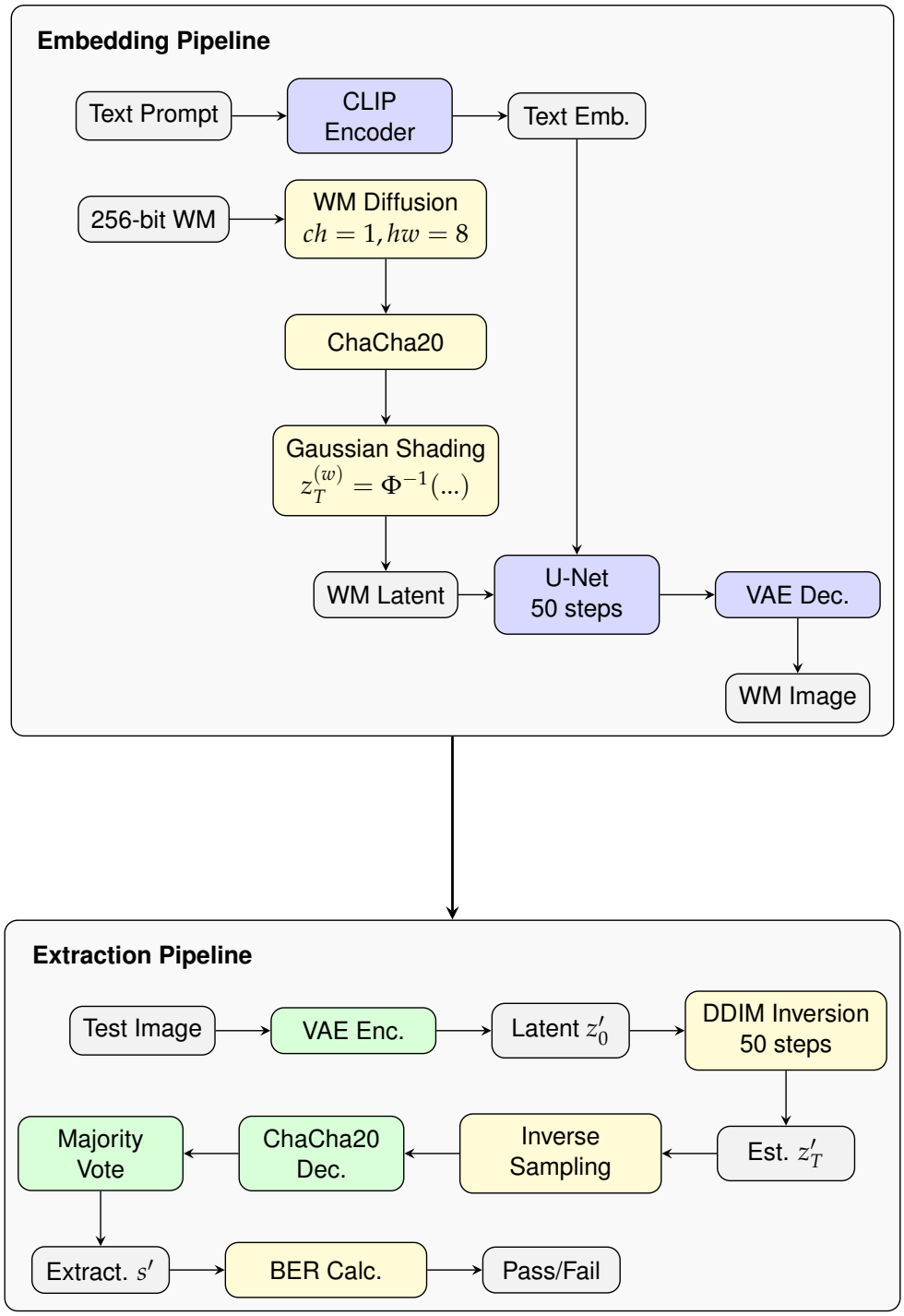

Figure 3.1: Architecture diagram showing: embedding pipeline (top) & extraction pipeline (bottom).

3.5 Implementation Details

3.5.1 Technical Environment

The implementation utilises an NVIDIA RTX 2070 Super (8GB VRAM) with Compute Unified Device Architecture (CUDA) 11.7, running Python 3.8 with PyTorch 1.13+ and Diffusers 0.11.1. Key dependencies include transformers (4.21.0) for CLIP text encoding, accelerate for memory optimisation, scipy for statistical functions, and pycryptodome for ChaCha20 cryptographic operations. The 8GB VRAM constraint necessitates memory efficiency strategies: float16 precision reduces memory footprint by 50%, sequential processing prevents out-of-memory errors, and gradient checkpointing enables processing of 512×512 images. Memory allocation is optimised through PyTorch's attention implementation and careful tensor lifecycle management during DDIM inversion.

3.5.2 Gaussian Shading Implementation

The implementation follows the algorithm specification from [41], The method consists of three key components: watermark diffusion, distribution preserving sampling, and DDIM inversion for extraction.

Watermark Diffusion

The watermark message $s \in \{0,1\}^m$ undergoes an initial diffusion process across its latent dimensions to enhance robustness. This process generates multiple copies of each bit across different dimensions, defined as $s_d = \text{Diffuse}(s, \text{channel\_copy}, \text{hw\_copy})$. Where channel_copy determines how many copies are made across channel dimensions, and hw_copy controls copies across spatial dimensions. This redundancy is important for robustness, as it allows for majority voting during extraction even if some dimensions are corrupted by attacks. The diffused watermark s_d is then encrypted using ChaCha20 stream cipher with a secret key K to produce the encrypted watermark m: $m = \text{Encrypt}(s_d, K)$

Distribution Preserving Sampling

A key part of Gaussian Shading is the distribution preserving sampling technique. In standard diffusion models, the initial latent is sampled from a standard normal distribution:

$$z_T \sim \mathcal{N}(0, I) \text{ (standard)} \tag{3.1}$$

For watermarking, this sampling is modified to embed the encrypted watermark while maintaining the same statistical distribution:

$$z_T^{(w)} = \Phi^{-1}(U \cdot \Phi(z_T) + (1 - U) \cdot \Phi(z_T^{(ref)})) \text{ (watermarked)} \tag{3.2}$$

where:
- Φ is the standard normal CDF
- Φ^{-1} is the inverse CDF (quantile function)
- U is a uniform random variable in $[0, 1]$

- $z_T^{(ref)}$ is a reference latent that encodes the encrypted watermark bits

This approach ensures that $z_T^{(w)}$ follows the same distribution as z_T, making the watermark statistically undetectable and preserving the generative model's performance. The implementation uses the error function (erf) and its inverse for efficient computation of the normal CDF and quantile functions.

DDIM Inversion for Extraction

The extraction process uses deterministic DDIM inversion with 50 steps to estimate the initial latent. The algorithm is:

Algorithm 1 DDIM Inversion for Watermark Extraction

Require: Watermarked image x, number of steps T
Ensure: Estimated initial latent $\hat{z}_T$
1: $z_0 \leftarrow$ VAE_Encoder(x) {Encode image to latent space}
2: **for** $t = 1$ to T **do**
3: $\quad \epsilon_\theta \leftarrow$ U-Net$(z_{t-1}, t-1)$ {Predict noise}
4: $\quad z_t \leftarrow$ DDIM_Step$(z_{t-1}, \epsilon_\theta, t-1, t)$ {Reverse diffusion step}
5: **end for**
6: **return** z_T

Where DDIM_Step implements the deterministic reverse diffusion process according to the DDIM formulation. Once the initial latent $\hat{z}_T$ is recovered, the watermark is extracted by applying the inverse of the distribution preserving sampling process: $\hat{m} = \text{Extract}(\hat{z}_T, K)$
The extracted watermark $\hat{m}$ is then decrypted and the diffused bits are combined through majority voting to recover the original message $\hat{s}$.

Memory Optimization Techniques

The implementation incorporates several memory optimization strategies to operate within the 8GB VRAM constraint:

- **Half-precision arithmetic:** All tensors use float16 precision, reducing memory footprint by half compared to float32.
- **Sequential processing:** Images are processed one at a time rather than in batches to avoid memory spikes.
- **Efficient tensor management:** Careful attention to tensor lifecycle, explicitly freeing unused tensors and using in-place operations where possible.
- **Optimised attention implementation:** Using PyTorch's attention implementation to reduce memory footprint of transformer blocks in the U-Net.

These optimisations enable processing of 512×512 images on the specified hardware without compromising the quality of the watermarking process.

3.6 Evaluation Framework

3.6.1 Attack Suite Configuration

The robustness evaluation employs a comprehensive attack suite designed to test common image processing operations:

Table 3.1: Attack Configuration Parameters

Attack Type	Parameters	Configurations
JPEG Compression	Quality Factor	90, 75, 50, 25
Gaussian Noise	Standard Deviation	0.01, 0.03, 0.05
Gaussian Blur	Kernel Radius	2, 4
Brightness Adjustment	Multiplication Factor	0.5, 2.0
Random Crop	Retention Ratio	0.8

This configuration yields 12 attack scenarios, each evaluated on 200 images, totalling 2400 robustness experiments. The parameter ranges are selected to align with previous watermarking studies while covering mild to severe attack intensities.

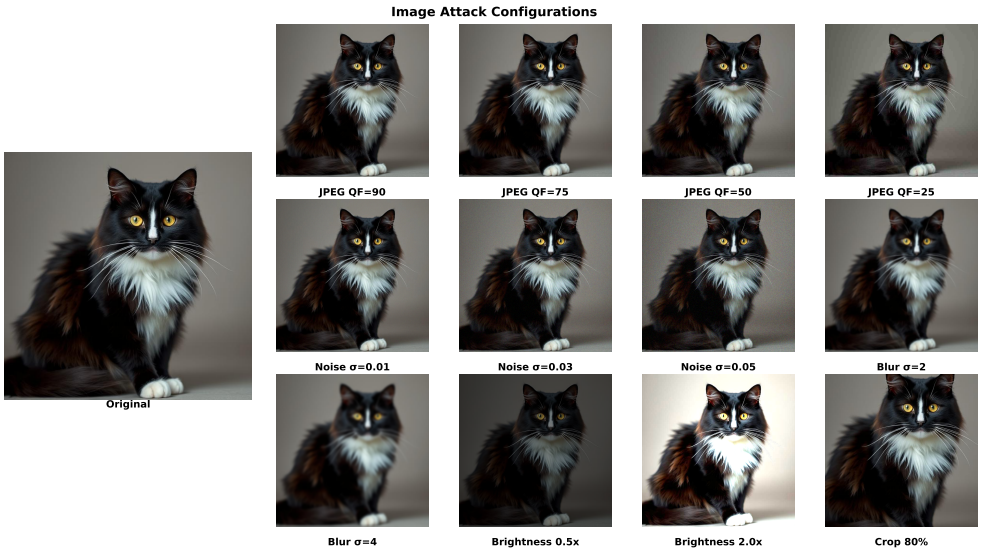

Figure 3.2: A visual summary of the twelve digital attacks applied to evaluate watermark robustness.

3.6.2 Evaluation Metrics

The evaluation framework employs three metric categories.
While traditional metrics like PSNR and SSIM are common for assessing imperceptibility, they are not the primary metrics for this evaluation. The Gaussian Shading technique operates by altering the initial latent sample (z_T). As such, even when using an the same seed, the watermarked and non-watermarked images are inherently distinct at the pixel level, as they

are slightly modified versions of one another. Applying pixel level metrics in this context is not appropriate, as they would report significant differences that do not reflect perceptual similarity.
Therefore, distribution based metrics are more appropriate. The FID score is used to compare the statistical distributions of the entire set of 1000 watermarked images against the 1000 non-watermarked images, providing a more meaningful measure of the watermark's overall impact on the model's output distribution. Additionally, the CLIP score is used to ensure that the semantic content of the generated images remains consistent with the original prompt, further validating the watermark's imperceptibility.
Imperceptibility metrics FID for distribution preservation, and CLIP Score for semantic similarity between prompts and generated images. **Robustness metrics** comprise BER as the primary robustness measure, TPR derived from accuracy measurements under the FPR-controlled detection framework, and overall detection accuracy for binary classification performance. Statistical analysis will include 95% confidence intervals for all key metrics and significance testing to assess performance differences across attack types.

3.6.3 Comparative Analysis Strategy

A fundamental challenge in watermarking research is that differing papers will apply their own evaluation metrics and criteria, therefore making reliable comparisons difficult, without also implementing other methods using the same tests, which would be difficult for numerous reasons, for example, Stable Signature is not training agnostic, meaning deployment requires retraining of GAIM training data. Acknowledging these limitations we shall compare methodological approaches rather than direct numerical benchmarking:

1. **Methodological Analysis:** Focus on characteristics and trade-offs of different approaches (Tree-Ring [39] & Stable Signature [40]).
2. **Cross-Study Limitation Recognition:** Acknowledge that differences in experimental protocols, datasets, attack implementations, hardware configurations and evaluation metrics can significantly influence reported performance figures.
3. **Statistical Analysis:** Calculate 95% confidence intervals for all key metrics using the 200 individual measurements per attack configuration. Perform statistical significance tests (paired t-tests, ANOVA) to assess performance differences across attack types and intensities.
4. **Independent Rigorous Evaluation:** Provide comprehensive statistical analysis of Gaussian Shading with 95% confidence intervals and significance testing to enable future research to make informed comparisons.

This approach prioritises methodological transparency and acknowledges the current limitations in AI image watermarking.

3.6.4 Statistical Rigour and Reproducibility

Scientific validity is ensured through deterministic random number generation with fixed seeds for reproducibility, statistically sufficient sample sizes (n=200 per attack, n=1000 for FID), 95% confidence intervals calculated using t-distribution and comprehensive documentation of all code and hyperparameters under version control.

3.6.5 Success Criteria and Thresholds

The evaluation sets a clear success criteria to assess Gaussian Shading's viability. **Imperceptibility thresholds** require FID < 30 (minimal distribution shift between watermarked and non-watermarked images) and CLIP Score difference < 0.05 (preservation of semantic content). **Robustness thresholds** require BER < 0.1 for mild attacks (JPEG QF ≥ 75), BER < 0.25 for moderate attacks (JPEG QF ≥ 50), and TPR > 0.9 at FPR = 10^{-6}. These derived thresholds from the literature provide objective criteria for evaluating the performance lossless claims of Gaussian Shading.

The methodology balances comprehensive evaluation with practical constraints, ensuring that results are both scientifically valid and comparable to existing literature while remaining manageable within undergraduate project scope and hardware limitations.

4 Results and Analysis

This chapter presents a thorough evaluation of the Gaussian Shading watermarking technique, examining its imperceptibility and robustness through statistical analysis. The evaluation addresses the research questions concerning the viability of Gaussian Shading as a practical watermarking solution for AI-generated images, and its place in the digital image watermarking landscape.

4.1 Imperceptibility Analysis

The imperceptibility evaluation constitutes the foundational assessment of Gaussian Shading's main advantage: that watermarks can be embedded without degrading image quality. This analysis makes use of a varied approach, using semantic alignment metrics and distributional statistics to provide a comprehensive understanding of the watermark's impact.

4.1.1 Visual and Semantic Preservation

Qualitative assessment confirms that the watermarking process consistently produces high quality images. Figure 4.1 demonstrates the framework's ability to generate visually appealing images across diverse categories, including landscapes, urban scenes and food, with no additional artifacts or degradation. While a direct visual comparison of identical seeds is not feasible due to the nature of the sampling process, the consistent high quality generation across a wide range of content strongly suggests that the method preserves visual fidelity.

Figure 4.1: Sample images generated using the Gaussian Shading watermarking framework. The method maintains high visual quality across diverse content types, from landscapes and cityscapes to food and architectural subjects, with no discernible visual artifacts.

To quantify semantic preservation, the evaluation uses CLIP scores, which measure the semantic alignment between text prompts and the generated images. The results, presented in Table 4.1, show desirable semantic consistency. The negligible difference of -0.0008 between the watermarked (0.3091) and non-watermarked (0.3099) image sets is well within the bounds of natural variation and meets the success threshold of < 0.05. This confirms that the watermarking process does not interfere with the model's ability to generate images that remain faithful to their textual prompt descriptions.

Table 4.1: Semantic Preservation Analysis using CLIP Score

Condition	**Mean CLIP Score**	**Difference**	**Assessment**
Non-watermarked	0.3099	-	Baseline
Watermarked	0.3091	-0.0008	**Imperceptible**
Success Threshold	-	< 0.05	**Met**

4.1.2 Distributional Impact

While the visual and semantic assessments support the imperceptibility claims, the distributional analysis shows a different story. The FID score, which measures the statistical distance between the distributions of the watermarked and non-watermarked image sets, yields a value of 47.83. As shown in Table 4.2, this is above the predefined success threshold of 30, indicating a significant distributional shift between the 1000 watermarked and non-watermarked images.

This finding challenges the performance losslessness of the Gaussian Shading method. This high FID score indicates that while individual images remain visually and semantically similar, the underlying statistical distribution of the watermarked image set has been measurably altered. This suggests that the watermarking process introduces systematic changes to the generative process which, while imperceptible at the individual image level, become detectable when analysing the aggregate behaviour of the model.

Table 4.2: Distributional Metrics (FID Score)

Metric	**Observed Value**	**Success Threshold**	**Assessment**
FID Score	47.83	< 30	**Not Met**

Understanding the FID Score

The relatively high FID score does not necessarily imply a degradation in the quality of individual images, but rather a shift in the overall distribution of the generated content.

Several factors may contribute to this result:

1. **Distributional Shift vs. Quality Degradation:** The high FID score indicates a distributional shift rather than a drop in quality. The watermarking process modifies the initial latent sampling, which could alter the model's latent space in a way that affects overall statistics without compromising the quality of individual images.
2. **Sample Size Sensitivity:** FID scores are sensitive to sample size. This evaluation used 1000 images, whereas the original paper used 5000. This difference in scale may contribute to the elevated score through reduced statistical stability.

4.2 Statistical Analysis and Detection Performance

This section provides comprehensive statistical analysis of the watermark detection performance, including FPR/TPR analysis following the methodologies established detection framework, confidence intervals for all key metrics and statistical significance testing across attack types.

4.2.1 FPR-Controlled Detection Framework

The Gaussian Shading algorithm operates with an FPR controlled detection threshold, targeting an FPR of 10^{-6} for statistical significance. Under this, the TPR represents the algorithm's ability to correctly detect watermarked images at the controlled FPR.

Table 4.3 presents TPR analysis derived from accuracy measurements across all attack configurations, calculated from 200 samples per configuration using actual evaluation data. Results show excellent detection performance across all signal processing attacks, with TPR values consistently above 98%, except for the failure under the geometric attack tested.

The analysis shows exceptional TPR performance across all signal processing attacks, with values consistently above 98% even under severe conditions. However, the geometric attack (crop) results in TPR failure at 51.48%, statistically indistinguishable from random chance.

4.2.2 Statistical Significance

To assess the statistical significance of performance differences across attack types and intensities, one-way ANOVA and paired t-tests were conducted on the BER measurements from the 200 generated images per attack configuration.

Attack Type Comparison

One-way ANOVA testing across the five main attack categories (brightness, Gaussian blur, Gaussian noise, jpeg, crop) shows significant differences in robustness performance ($F(4, 2395) = 15064.930$, $p < 0.001$). Post-hoc Tukey Honestly Significant Difference (HSD) tests identify that:

- **Crop attacks** show much worse performance compared to all other categories ($p < 0.001$)

Table 4.3: True Positive Rates with 95% Confidence Intervals

Attack Type	Parameter	TPR	95% CI	n
brightness	0.500	0.9994	[0.9985, 1.0000]	200
	2.000	0.9977	[0.9950, 1.0000]	200
crop	0.800	0.5148	[0.5102, 0.5194]	200
gaussian_blur	2.000	0.9983	[0.9962, 1.0000]	200
	4.000	0.9820	[0.9777, 0.9863]	200
gaussian_noise	0.010	0.9985	[0.9961, 1.0000]	200
	0.030	0.9963	[0.9924, 1.0000]	200
	0.050	0.9931	[0.9890, 0.9972]	200
jpeg	25.000	0.9865	[0.9810, 0.9920]	200
	50.000	0.9950	[0.9905, 0.9995]	200
	75.000	0.9972	[0.9938, 1.0000]	200
	90.000	0.9983	[0.9956, 1.0000]	200

- **Brightness attacks** show significantly superior robustness compared to all other categories ($p < 0.001$)
- **Gaussian blur** shows worse performance than brightness, JPEG, and noise attacks ($p < 0.001$)
- **JPEG compression** and **Gaussian noise** show comparable performance ($p = 0.740$, not significant)

Attack Intensity Analysis

Paired t-tests examining performance degradation with increasing attack intensity show statistically significant effects:

- **Brightness intensity:** Factor 0.5 vs 2.0 ($t = 1.876$, $p = 0.062$, marginally significant)
- **Gaussian blur intensity:** Radius 2.0 vs 4.0 ($t = 11.121$, $p < 0.001$, highly significant)
- **Gaussian noise intensity:** Progressive degradation across σ levels (0.01 vs 0.03: $t = 2.498$, $p = 0.013$; 0.03 vs 0.05: $t = 4.699$, $p < 0.001$)
- **JPEG compression intensity:** Significant degradation with decreasing quality (QF 25 vs 50: $t = -6.157$, $p < 0.001$; QF 50 vs 75: $t = -3.150$, $p = 0.002$; QF 75 vs 90: $t = -2.533$, $p = 0.012$)

These results provide statistical validation that the observed performance patterns are not due to random variation but represent genuine algorithmic characteristics, with the crop attack showing the most dramatic performance degradation.

This analysis suggests that while Gaussian Shading achieves excellent visual and semantic imperceptibility, it is not entirely performance lossless from a distributional perspective. This has implications for applications where maintaining precise distributional properties is critical, and highlights

the need for a multi faceted understanding of imperceptibility in watermarking research.

4.3 Robustness Evaluation

The robustness assessment evaluates the watermark's resilience against a comprehensive suite of attacks designed to simulate real world image manipulations. This analysis reveals the method's exceptional strength against signal processing attacks but vulnerability to geometric transformations. This section discusses these results, providing a detailed analysis of the method's strengths and limitations.

4.3.1 Robustness to Signal Processing Attacks

The evaluation of signal processing attacks shows Gaussian Shading's remarkable robustness against common image manipulations. Figure 4.2 illustrates the watermark's resilience to two of the most critical attack categories: JPEG compression and additive Gaussian noise.

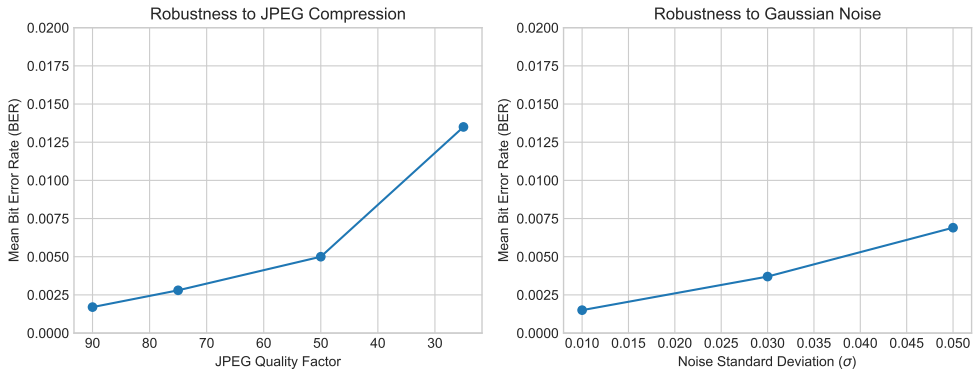

Figure 4.2: Robustness against signal processing attacks. Left: JPEG compression robustness, showing minimal degradation, despite aggressive compression ratios. Right: Gaussian noise resilience, demonstrating consistent performance across increasing noise levels.

JPEG Compression and Noise Resilience

The JPEG compression results are particularly impressive. Even under severe compression (quality factor 25), the BER remains at 0.0135 (1.35%), far below the established success threshold of 0.25 for moderate attacks. This performance is noteworthy, as JPEG compression is one of the most common and destructive manipulations for many images disseminated online. Similarly, the watermark maintains excellent detectability under significant Gaussian noise ($\sigma = 0.05$), with the BER reaching only 0.0069 (0.69%). This high level of robustness can be attributed to the method's operation in the latent space. By embedding the watermark in the initial latent vector and preserving its statistical properties, Gaussian Shading creates a watermark

that is deeply integrated within the image's structure, making it resilient to attacks that primarily affect the pixel level or high frequency information.

Additional Signal Processing Attacks

The method's strong performance extends to other signal processing attacks. As detailed in Table 4.4, Gaussian blur with a radius of 4 pixels results in a BER of only 0.0180 (1.8%), while brightness adjustments across a factor-of-four range $(0.5\times$ to $2.0\times)$ yield a negligible maximum BER of 0.0023 (0.23%). This combination of results establish Gaussian Shading as exceptionally robust against a broad range of signal processing attacks, positioning it favourably against the many other existing modern AI image watermarking techniques.

4.3.2 Vulnerability to Geometric Attacks

In stark contrast to its signal processing robustness, the evaluation reveals a vulnerability to geometric attacks. Figure 4.3 illustrates this, showing the failure under a random crop attack compared to the impressive resilience against a strong blur.

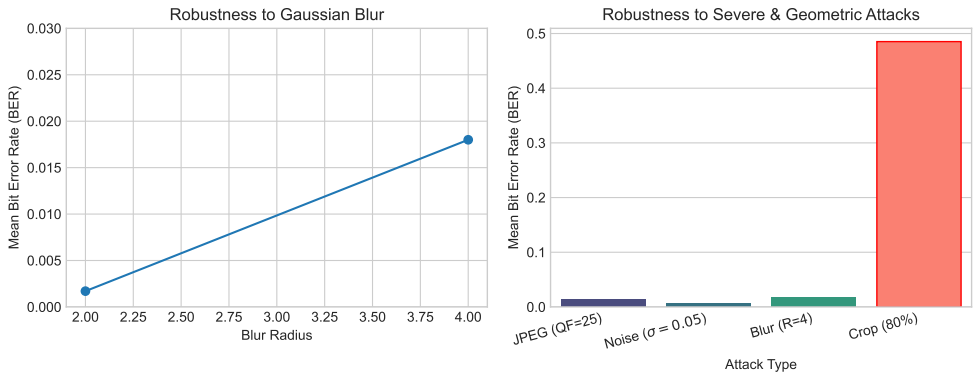

Figure 4.3: Left: Continued excellent performance against Gaussian blur. Right: The failure under a crop attack (red bar) compared to the maintained robustness against other severe signal processing attacks.

The crop attack, which removes 20% of the image area before resizing, results in a high BER of 0.4852 (48.52%). A value statistically indistinguishable from random guessing (50%). This represents a complete failure of the watermark, indicating that the method's spatial redundancy is insufficient to survive significant data loss.

Root Cause Analysis of Geometric Failure

The failure demonstrated with the cropping attack is due to the fundamental limitations of the Gaussian Shading algorithm:

1. **Lack of Geometric Invariance:** The method lacks inherent geometric invariance. The watermark's structure is tied to the absolute spatial coordinates of latent space, therefore any transformation that changes

these relationships (like cropping or rotation) interferes with watermark extraction.
2. **Insufficient Spatial Redundancy:** While the watermark is diffused across spatial dimensions, the level of redundancy is not sufficient to reconstruct the watermarked message when a significant, contiguous section of the image is removed.
3. **No Integrated Error Correction:** The base algorithm does not make use of sophisticated error correction codes (such as Reed-Solomon [47]) specifically implemented to handle errors caused by spatial data loss.

This vulnerability has practical implications, as cropping is a common image manipulation, from images being shared online such as in social media posts and academic publications.

4.3.3 Comparative Performance and Overall Assessment

Table 4.4 provides a complete overview of the robustness evaluations conducted, clearly showing the method's strengths and limitations. These results show great robustness against signal processing attacks along with a failure against geometric transformations (crop).

Table 4.4: Comprehensive Robustness Evaluation Results with Statistical Confidence

Attack Category	Attack Configuration	Mean BER	95% CI	Performance	Assessment
JPEG Compression	Quality=90	0.0017	$\pm$0.0027	Excellent	**Pass**
	Quality=75	0.0028	$\pm$0.0034	Excellent	**Pass**
	Quality=50	0.0050	$\pm$0.0045	Excellent	**Pass**
	Quality=25	0.0135	$\pm$0.0055	Very Good	**Pass**
Gaussian Noise	$\sigma = 0.01$	0.0015	$\pm$0.0024	Excellent	**Pass**
	$\sigma = 0.03$	0.0037	$\pm$0.0040	Excellent	**Pass**
	$\sigma = 0.05$	0.0069	$\pm$0.0041	Excellent	**Pass**
Gaussian Blur	Radius=2	0.0017	$\pm$0.0020	Excellent	**Pass**
	Radius=4	0.0180	$\pm$0.0043	Very Good	**Pass**
Brightness	Factor=0.5	0.0006	$\pm$0.0009	Excellent	**Pass**
	Factor=2.0	0.0023	$\pm$0.0027	Excellent	**Pass**
Geometric	**Crop (80% retention)**	**0.4852**	$\pm$0.0046	**Very Poor**	**Fail**

Confidence intervals calculated using t-distribution for n=200 samples per attack configuration.

To contextualise these findings, Table 4.5 presents a methodological comparison with Tree-Ring [39] and Stable Signature [40].
The comparison reveals distinct positioning within the watermarking landscape. Gaussian Shading shares with Tree-Ring the advantage of training-free deployment, enabling rapid integration into existing systems without requiring model retraining. However, it substantially exceeds its capacity limitations (256 bits vs. 1 bit), making it more suitable for detailed attribution and traceability applications. Compared to Stable Signature, Gaussian Shading's plug-and-play deployment represents a practical advantage,

though this simplicity may come at the cost of the geometric robustness that fine-tuned approaches can achieve.
This methodological analysis highlights a key challenge in watermarking research: the difficulty of making reliable cross-study performance comparisons due to variations in experimental protocols, datasets, attack implementations and evaluation metrics. Such differences can significantly influence reported performance figures; for direct numerical comparisons, an identical evaluation framework should be used.

Table 4.5: Comparison of Watermarking Techniques

Characteristic	Gaussian Shading (This work)	Tree-Ring [39]	Stable Signature [40]
Embedding Domain	Latent (Initial Sampling)	Latent (Initial Noise)	Model (VAE Fine-tuning)
Training Required	None	None	Yes (VAE fine-tuning)
Watermark Capacity	256 bits	1 bit	64 bits
Integration Approach	Distribution-preserving sampling	DDIM noise injection	Decoder modification
Signal Processing Robustness	High (BER < 2%)	Reported as high	Reported as high
Geometric Attack Robustness	Poor (Crop: BER 48.5%)	Reported as moderate	Reported as good
Deployment Complexity	Low (plug-and-play)	Low (plug-and-play)	High (requires retraining)
Model Architecture	Any LDM	Diffusion Models	Stable Diffusion specific

This comparative analysis solidifies the conclusion that Gaussian Shading occupies a specific niche: it is a high-capacity, easy-to-deploy solution that offers excellent protection in environments where geometric transformations are unlikely. However, its notable failure under the cropping attack makes it unsuitable for applications where such manipulations are commonplace.

4.4 Evaluation Scale and Statistical Methodology

The robustness evaluation was based on 200 images per attack configuration totalling 2400 individual experiments, providing a substantial dataset for statistical analysis. The imperceptibility assessment was conducted on 1000 image pairs for the FID calculation and CLIP score comparison.
The comprehensive nature of the evaluation, spanning multiple attack types and intensities, enables reliable conclusions about the method's performance across differing scenarios. The sample sizes ($n = 200$ per attack) provide adequate data for calculating meaningful 95% confidence intervals using t-distribution and conducting significance testing through ANOVA and paired t-tests. This analysis ensures that the observed performance patterns are representative and statistically validated rather than artifacts of limited testing or random variation.

4.5 Comparative Contextualisation

When positioned within the broader landscape of AI watermarking techniques, Gaussian Shading's performance profile reveals both distinctive advantages and notable limitations. Compared to Stable Signature's requirement for model fine-tuning, Gaussian Shading's training-free nature represents a significant practical advantage, enabling rapid deployment

across existing infrastructure without retraining costs.
Against Tree-Ring's 1-bit capacity limitation, Gaussian Shading's 256-bit payload capability provides substantial advantages for attribution and traceability applications. However, this capacity comes at the cost of the geometric vulnerability identified in this evaluation.
The distributional impact revealed through FID analysis, positions Gaussian Shading between truly lossless methods and those with more pronounced quality degradation. This intermediate position reflects the fundamental trade-offs inherent in watermarking system design.

4.6 Key Research Questions Answered

The evaluation provides definitive answers to the study's fundamental research questions:
Research Question 1: Robustness Extent Gaussian Shading demonstrates exceptional robustness against signal processing attacks, with BER consistently below 2% even under severe conditions. However, it exhibits vulnerability to geometric attacks, particularly cropping, where performance degrades to levels comparable to random choice.
Research Question 2: Performance/Quality Trade-offs The method achieves visual and semantic imperceptibility, confirming its claims at the perceptual level. However, distributional analysis reveals measurable impact on the generative model's statistical properties, challenging the absoluteness of Gaussian Shadings performance losslessness.
Research Question 3: Competitive Positioning Gaussian Shading offers unique advantages in training-free deployment and high capacity, but these benefits come with the trade-off of geometric vulnerability and measurable distributional impact observed in these results. The method is a novel approach that should not be dismissed, but whos potetial needs needs further exploration.

5 Conclusion

This dissertation has conducted a comprehensive, independent evaluation of the Gaussian Shading watermarking technique across 2400 individual robustness experiments, providing insights into its performance and practical viability through implementation and systematic evaluation with statistical validation. The evaluation revealed exceptional robustness against signal processing attacks, with TPR values consistently above 98% across all tested conditions, while statistical analysis using one-way ANOVA confirmed performance differences across attack categories ($F(4, 2395) = 15064.930$, $p < 0.001$). The findings challenge the absoluteness of Gaussian Shadings performance losslessness by revealing a measurable distributional impact despite perceptual imperceptibility. Furthermore, the research provides a definitive analysis of the method's robustness, highlighting the resilience to most common signal processing attacks but a vulnerability to geometric transformations, with the cropping attack resulting in performance statistically indistinguishable from random chance (TPR 51.48%, 95% confidence interval $[51.02\%, 51.94\%]$).
This work presents a nuanced understanding of the trade-offs inherent in the Gaussian Shading watermarking method through statistical analysis with 95% confidence intervals derived from 200 samples per attack configuration. By providing detailed performance data with statistical validation, this project serves as a starting point for future research with actionable insights, particularly the need to further evaluate and enhance geometric robustness. The statistical methods employed reveal brightness attacks as the least destructive to watermark retrieval. Ultimately, this dissertation highlights the importance of an in depth evaluation with statistical validation and contributes to the development of more reliable and effective solutions for ensuring the authenticity and traceability of AI-generated images.

5.1 Limitations

While this evaluation provides significant insights, there are several limitations to consider. The evaluation is limited to a single model architecture (Stable Diffusion v2.1), and its findings may not directly transfer to other LDMs or generative models. The attack suite, while comprehensive, did not include sophisticated adversarial attacks designed to specifically target the watermarking algorithm, which could reveal further vulnerabilities. Additionally, hardware constraints necessitated a smaller dataset for FID calculation (1000 images) than the original paper (5000 images), which may contribute to the elevated score. Furthermore, only a single geometric attack (random cropping) was evaluated, which limits the understanding of the method's robustness against other geometric transformations such as rotation or scaling. This is especially relevant given the method's vulnerability to cropping, which is a common image manipulation in real-world scenarios.

5.2 Future Work and Implications

The findings from this research open several promising avenues for future investigation and have significant implications for the responsible development of AI.

5.2.1 Future Research Directions

The most critical area for future work is more research into geometric robustness. This could be achieved by integrating advanced error correction codes (e.g., Bose-Chaudhuri-Hocquenghem (BCH) Reed-Solomon [47]), developing spatial redundancy schemes, or exploring geometric invariant feature spaces for embedding. Another key point is mitigating the distributional impact identified by the relatively high FID score. This would involve analysis at a larger scale (e.g. larger datasets). Finally, future work should focus on use of a standardised benchmarking framework such as WAVES [48] by applying the same evaluation techniques across a diverse range of generative architectures and assessing resistance against sophisticated adversarial attacks.

5.2.2 Implications for Responsible AI development

This research provides insights for the broader development of responsible AI systems. For policy and regulation, the findings demonstrate that while watermarking is a promising tool, its limitations (e.g. geometric vulnerability) must be considered when deploying a suitable method for widespread use. This work highlights the need for more nuanced, multi faceted evaluation frameworks and pinpoints geometric robustness as a key area for future research. By contributing an independent assessment, this project helps build a foundation for developing reliable and effective technologies for digital content authentication, attribution and traceability in the age of the generative AI media landscape.

A Appendix

A.1 Gaussian Shading sampling function

Listing A.1: Gaussian Shading Sampling Implementation

```
def distribution_preserving_sampling(z_t, m, alpha=0.5):
    # Convert to uniform using normal CDF
    u_z = 0.5 * (1 + torch.erf(z_t / math.sqrt(2)))

    # Create reference latent based on watermark bit
    z_ref = torch.randn_like(z_t)
    u_ref = 0.5 * (1 + torch.erf(z_ref / math.sqrt(2)))

    # Adjust reference values based on watermark bits
    mask = (m > 0.5).float()
    u_ref = mask * torch.clamp(u_ref, min=0.5) + (1 -
       mask) * torch.clamp(u_ref, max=0.5)

    # Interpolate in uniform space
    u_mix = alpha * u_z + (1 - alpha) * u_ref

    # Convert back to normal distribution
    # Clamp to avoid numerical issues at extremes
    u_mix_safe = torch.clamp(u_mix, min=1e-6, max=1-1e
       -6)
    z_watermarked = math.sqrt(2) * torch.erfinv(2 *
       u_mix_safe - 1)

    return z_watermarked
```

A.2 Watermark Evaluation Logic

This function implements the main part of watermark detection, supporting the BER calculations in the results.

Listing A.2: Watermark Detection and Evaluation Implementation

```
def eval_watermark(self, reversed_m):
    # Convert latent values to binary (threshold at 0)
    reversed_m = (reversed_m > 0).int()

    # Decrypt the watermark using the stored key
    reversed_sd = (reversed_m + self.key) % 2

    # Apply diffusion inverse to reconstruct watermark
    reversed_watermark = self.diffusion_inverse(
        reversed_sd)

    # Calculate accuracy by comparing with original
        watermark
    correct = (reversed_watermark == self.watermark).
        float().mean().item()

    # Update detection counters based on thresholds
    if correct >= self.tau_onebit:
        self.tp_onebit_count = self.tp_onebit_count + 1
    if correct >= self.tau_bits:
        self.tp_bits_count = self.tp_bits_count + 1

    return correct
```

A.3 GPU Memory Optimisation

This function enabled large scale evaluation (2400 robustness experiments, 1000 imperceptibility experiments) on consumer hardware (RTX 2070 Super, 8GB VRAM).

Listing A.3: GPU Memory Optimisation

```
def optimise_gpu_memory():
    """Optimise GPU memory usage."""
    if torch.cuda.is_available():
        torch.cuda.empty_cache()
        torch.cuda.synchronize()
        # Enable memory efficient attention if available
        try:
            torch.backends.cuda.enable_flash_sdp(True)
        except:
            pass
```

A.4 Statistical Analysis Code Snippets

This section provides key Python code snippets from the statistical analysis script for reproducibility.

Listing A.4: Confidence Interval Calculation

```python
import scipy.stats as stats
import numpy as np

def calculate_confidence_intervals(data, confidence=0.95):
    n = len(data)
    mean = np.mean(data)
    std_err = stats.sem(data)
    t_critical = stats.t.ppf((1 + confidence) / 2, n - 1)
    margin = t_critical * std_err

    lower_bound = max(0, mean - margin)
    upper_bound = min(1, mean + margin)

    return mean, lower_bound, upper_bound, margin, t_critical
```

Listing A.5: ANOVA Analysis

```python
from scipy.stats import f_oneway
from statsmodels.stats.multicomp import pairwise_tukeyhsd

# Group data by attack type
attack_groups = []
attack_types = ['brightness', 'crop', 'gaussian_blur',
                'gaussian_noise', 'jpeg']
for attack_type in attack_types:
    attack_data = detailed_data[
        detailed_data['attack_name'] == attack_type
    ]['ber'].values
    attack_groups.append(attack_data)

# Perform one-way ANOVA
f_stat, p_value = f_oneway(*attack_groups)

# Tukey HSD post-hoc test
tukey_results = pairwise_tukeyhsd(
    endog=detailed_data['ber'],
    groups=detailed_data['attack_name'],
    alpha=0.05
)
```

B Ethics Checklist

Ethical Considerations Checklist and Declaration

Project Description

Briefly explain the methodology of your study. Give sufficient detail that a non-expert in the subject can understand what you are proposing to do.

Identifying images generated by Artificial Intelligence (AI) with invisible watermarks: This can help improve copyright protections, traceability, and accountability. A watermarking method will be implemented to embed and retrieve these imperceptible watermarks with minimal compromise to quality. The method will be evaluated against its resistance to common attacks such as compression, cropping, noise, and smoothing. Additionally the chosen method will be compared with existing methods.

External Datasets

If you are using external datasets, please (a) identify the dataset(s) giving sufficient details; (b) identify the relevant licence or terms of use for the dataset(s) and justify why your project would be compliant with those terms; (c) if the data is about humans, also provide evidence that the data was initially collected with consent.

The project will make use of text-to-image prompt datasets in order to have standardisation of generated images COCO-2017 (CC BY-SA 4.0), Meaning attribution is required. The project will be compliant with the terms as no changes are being made and the datasets used will be credited and referenced Gustavosta prompt set.

Potential Ethical Issues

Does your project involve any of the following? Please mark Yes or No for ***all*** *issues.*

Issue	Yes / No
Human participants (adults or children)	No
Human data (e.g. data collected through surveys and questionnaires on issues such as lifestyle, housing and working environments, or attitudes and preferences, or datasets including human data)	No
Datasets that require permission from the data provider	No
Applications that could potentially involve unethical practice, including potential dual-use applications (e.g. projects involving tools or data that can be used for unethical purposes e.g. to attack systems)	No
Funding sources or collaboration with potential to adversely affect existing relationships or bring the University or Department into disrepute (e.g. projects related to gambling, dark market, etc.)	No
Restrictions on dissemination (e.g. not being allowed to publish certain datasets or results)	No
Military or defence context	No
Overseas countries under regimes with poor human rights record or identified as dangerous by the Foreign & Commonwealth Office	No
Human material (e.g. tissue or fluid samples), vertebrates, especially mammals and birds, or any other organisms not previously mentioned	No

*If you answered **No** to all the above, you do not need ethical approval.*

*If you answered **Yes** to any of the above, you must complete a Fast-Track Ethical Approval Form, get it signed off by your project advisor, and submit it for approval to the Departmental Ethics Officers.*

Student Declaration

I have considered the ethical implications of this project, all the terms and conditions and permissions of any datasets being used, and I have identified no significant ethical implications requiring an ethical approval application.

Student Name	Mischa Zaynchkovsky
Student Signature	Mischa Zaynchkovsky
Date	14/02/2025

Bibliography

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer. 'High-Resolution Image Synthesis with Latent Diffusion Models.' arXiv: 2112.10752, Accessed: 28th Nov. 2024. [Online]. Available: http://arxiv.org/abs/2112.10752, pre-published.

[2] *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)*, 13th Jun. 2024. Accessed: 5th Dec. 2024. [Online]. Available: http://data.europa.eu/eli/reg/2024/1689/oj/eng.

[3] T.-Y. Lin et al. 'Microsoft COCO: Common Objects in Context.' arXiv: 1405.0312 [cs], Accessed: 15th Apr. 2025. [Online]. Available: http://arxiv.org/abs/1405.0312, pre-published.

[4] 'Gustavosta/Stable-Diffusion-Prompts · Datasets at Hugging Face,' Accessed: 17th Jul. 2025. [Online]. Available: https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts.

[5] __FUBAR__, *Bsmhmmlf/Gaussian-Shading*, 8th Jul. 2025. Accessed: 17th Jul. 2025. [Online]. Available: https://github.com/bsmhmmlf/Gaussian-Shading.

[6] A. Ramesh et al. 'Zero-Shot Text-to-Image Generation.' arXiv: 2102.12092 [cs], Accessed: 9th Dec. 2024. [Online]. Available: http://arxiv.org/abs/2102.12092, pre-published.

[7] 'Midjourney,' Midjourney, Accessed: 21st Jul. 2025. [Online]. Available: https://www.midjourney.com/website.

[8] E. Ferrara, 'GenAI Against Humanity: Nefarious Applications of Generative Artificial Intelligence and Large Language Models,' *J Comput Soc Sc*, vol. 7, no. 1, pp. 549–569, Apr. 2024, ISSN: 2432-2717, 2432-2725. DOI: 10.1007/s42001-024-00250-1. arXiv: 2310.00737 [cs]. Accessed: 9th Dec. 2024. [Online]. Available: http://arxiv.org/abs/2310.00737.

[9] 'Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence,' Federal Register, Accessed: 28th Nov. 2024. [Online]. Available: https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence.

[10] T. W. House. 'Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,' The White House, Accessed: 28th Nov. 2024. [Online]. Available: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

[11] 'Identifying AI-generated images with SynthID,' Google DeepMind, Accessed: 14th Nov. 2024. [Online]. Available: https://deepmind.google/discover/blog/identifying-ai-generated-images-with-synthid/.

[12] 'Bing Preview Release Notes: New experiences powered by Bing Image Creator,' Accessed: 28th Nov. 2024. [Online]. Available: https://blogs.bing.com/search/september-2023/Bing-Preview-Release-Notes-New-Experiences-Powered-by-Bing-Image-Creator.

[13] *CompVis/stable-diffusion*, CompVis - Computer Vision and Learning LMU Munich, 28th Nov. 2024. Accessed: 28th Nov. 2024. [Online]. Available: https://github.com/CompVis/stable-diffusion.

[14] 'Thousands of Artists Condemn Unlicensed Use of Their Work to Train A.I.,' Artnet News, Accessed: 18th Jul. 2025. [Online]. Available: https://news.artnet.com/art-world/artists-ai-statement-2557164.

[15] 'Andersen v. Stability AI Ltd., 3:23-cv-00201 - CourtListener.com,' CourtListener, Accessed: 18th Jul. 2025. [Online]. Available: https://www.courtlistener.com/docket/66732129/andersen-v-stability-ai-ltd/.

[16] S. Shan, J. Cryan, E. Wenger, H. Zheng, R. Hanocka and B. Y. Zhao, 'Glaze: Protecting Artists from Style Mimicry by {Text-to-Image} Models,' presented at the 32nd USENIX Security Symposium (USENIX Security 23), 2023, pp. 2187–2204, ISBN: 978-1-939133-37-3. Accessed: 15th Nov. 2024. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/shan.

[17] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng and B. Y. Zhao. 'Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models.' arXiv: 2310.13828, Accessed: 15th Nov. 2024. [Online]. Available: http://arxiv.org/abs/2310.13828, pre-published.

[18] A. Kurakin, I. Goodfellow and S. Bengio. 'Adversarial examples in the physical world.' arXiv: 1607.02533 [cs], Accessed: 15th Nov. 2024. [Online]. Available: http://arxiv.org/abs/1607.02533, pre-published.

[19] I. Cox, M. Miller, J. Bloom and M. Miller, *Digital Watermarking* (The Morgan Kaufmann Series in Multimedia Information and Systems). Morgan Kaufmann, 2001, ISBN: 978-0-08-050459-9. [Online]. Available: https://books.google.co.uk/books?id=uJcEWRv-RRUC.

[20] D. P. Kingma and M. Welling. 'Auto-Encoding Variational Bayes.' arXiv: 1312.6114, Accessed: 16th Nov. 2024. [Online]. Available: http://arxiv.org/abs/1312.6114, pre-published.

[21] D. J. Rezende, S. Mohamed and D. Wierstra. 'Stochastic Backpropagation and Approximate Inference in Deep Generative Models.' arXiv: 1401.4082, Accessed: 16th Nov. 2024. [Online]. Available: http://arxiv.org/abs/1401.4082, pre-published.

[22] Y. Guo et al. 'FreqMark: Invisible Image Watermarking via Frequency Based Optimization in Latent Space.' arXiv: 2410.20824, Accessed: 28th Nov. 2024. [Online]. Available: http://arxiv.org/abs/2410.20824, pre-published.

[23] I. J. Goodfellow et al. 'Generative Adversarial Networks.' arXiv: 1406.2661, Accessed: 16th Nov. 2024. [Online]. Available: http://arxiv.org/abs/1406.2661, pre-published.

[24] M. Alfarra, J. C. Pérez, A. Frühstück, P. H. S. Torr, P. Wonka and B. Ghanem. 'On the Robustness of Quality Measures for GANs.' arXiv: 2201.13019, Accessed: 28th Nov. 2024. [Online]. Available: http://arxiv.org/abs/2201.13019, pre-published.

[25] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan and S. Ganguli. 'Deep Unsupervised Learning using Nonequilibrium Thermodynamics.' arXiv: 1503.03585, Accessed: 16th Nov. 2024. [Online]. Available: http://arxiv.org/abs/1503.03585, pre-published.

[26] J. Ho, A. Jain and P. Abbeel. 'Denoising Diffusion Probabilistic Models.' arXiv: 2006.11239, Accessed: 16th Nov. 2024. [Online]. Available: http://arxiv.org/abs/2006.11239, pre-published.

[27] M. Bohacek and H. Farid. 'Nepotistically Trained Generative-AI Models Collapse.' arXiv: 2311.12202, Accessed: 29th Nov. 2024. [Online]. Available: http://arxiv.org/abs/2311.12202, pre-published.

[28] 'Content Credentials : C2PA Technical Specification :: C2PA Specifications,' Accessed: 9th Dec. 2024. [Online]. Available: https://c2pa.org/specifications/specifications/2.0/specs/C2PA_Specification.html.

[29] Z. Jiang, M. Guo, Y. Hu and N. Z. Gong. 'Watermark-based Detection and Attribution of AI-Generated Content.' arXiv: 2404.04254, Accessed: 14th Nov. 2024. [Online]. Available: http://arxiv.org/abs/2404.04254, pre-published.

[30] G. R. Manjula and A. Danti, 'A novel hash based least significant bit (2-3-3) image steganography in spatial domain,' *IJSPTM*, vol. 4, no. 1, pp. 11–20, 28th Feb. 2015, ISSN: 23194103, 22775498. DOI: 10.5121/ijsptm.2015.4102. arXiv: 1503.03674 [cs]. Accessed: 3rd Dec. 2024. [Online]. Available: http://arxiv.org/abs/1503.03674.

[31] W. Bender, D. Gruhl, N. Morimoto and A. Lu, 'Techniques for data hiding,' *IBM Systems Journal*, vol. 35, no. 3.4, pp. 313–336, 1996, ISSN: 0018-8670. DOI: 10.1147/sj.353.0313. Accessed: 3rd Dec. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/5387237.

[32] M. Saqib and S. Naaz, 'Spatial and Frequency Domain Digital Image Watermarking Techniques for Copyright Protection,' vol. 9, pp. 691–699, 1st Jun. 2017.

[33] G. K. Wallace, 'The JPEG still picture compression standard,' *Commun. ACM*, vol. 34, no. 4, pp. 30–44, 1st Apr. 1991, ISSN: 0001-0782. DOI: 10.1145/103085.103089. Accessed: 4th Dec. 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/103085.103089.

[34] A. Bors and I. Pitas, *Image Watermarking Using DCT Domain Constraints*. 16th Sep. 1996, 234 vol.3, 231 pp., ISBN: 978-0-7803-3259-1. DOI: 10.1109/ICIP.1996.560426.

[35] S. Fazli and M. Moeini, 'A robust image watermarking method based on DWT, DCT, and SVD using a new technique for correction of main geometric attacks,' *Optik*, vol. 127, no. 2, pp. 964–972, 2016, ISSN: 0030-4026. DOI: 10.1016/j.ijleo.2015.09.205. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0030402615012863.

[36] A. K. Abdulrahman and S. Ozturk, 'A novel hybrid DCT and DWT based robust watermarking algorithm for color images,' *Multimed Tools Appl*, vol. 78, no. 12, pp. 17 027–17 049, 1st Jun. 2019, ISSN: 1573-7721. DOI: 10.1007/s11042-018-7085-z. Accessed: 10th Apr. 2025. [Online]. Available: https://doi.org/10.1007/s11042-018-7085-z.

[37] D. Zheng, J. Zhao and A. El Saddik, 'RST-invariant digital image watermarking based on log-polar mapping and phase correlation,' *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 8, pp. 753–765, Aug. 2003, ISSN: 1558-2205. DOI: 10.1109/TCSVT.2003.815959. Accessed: 7th Dec. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/1227605.

[38] J. Zhu, R. Kaplan, J. Johnson and L. Fei-Fei. 'HiDDeN: Hiding Data With Deep Networks.' arXiv: 1807.09937, Accessed: 14th Nov. 2024. [Online]. Available: http://arxiv.org/abs/1807.09937, pre-published.

[39] Y. Wen, J. Kirchenbauer, J. Geiping and T. Goldstein, 'Tree-Rings Watermarks: Invisible Fingerprints for Diffusion Images,' *Advances in Neural Information Processing Systems*, vol. 36, pp. 58 047–58 063, 15th Dec. 2023. Accessed: 10th Dec. 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/b54d1757c190ba20dbc4f9e4a2f54149-Abstract-Conference.html.

[40] P. Fernandez, G. Couairon, H. Jégou, M. Douze and T. Furon. 'The Stable Signature: Rooting Watermarks in Latent Diffusion Models.' arXiv: 2303.15435 [cs], Accessed: 27th Feb. 2025. [Online]. Available: http://arxiv.org/abs/2303.15435, pre-published.

[41] Z. Yang, K. Zeng, K. Chen, H. Fang, W. Zhang and N. Yu. 'Gaussian Shading: Provable Performance-Lossless Image Watermarking for Diffusion Models.' arXiv: 2404.04956 [cs], Accessed: 14th Apr. 2025. [Online]. Available: http://arxiv.org/abs/2404.04956, pre-published.

[42] P. Dhariwal and A. Nichol. 'Diffusion Models Beat GANs on Image Synthesis.' arXiv: 2105.05233 [cs], Accessed: 10th May 2025. [Online]. Available: http://arxiv.org/abs/2105.05233, pre-published.

[43] Y. Zhao, T. Pang, C. Du, X. Yang, N.-M. Cheung and M. Lin. 'A Recipe for Watermarking Diffusion Models.' arXiv: 2303.10137 [cs], Accessed: 10th Dec. 2024. [Online]. Available: http://arxiv.org/abs/2303.10137, pre-published.

[44] L. Zhang, X. Liu, A. V. Martin, C. X. Bearfield, Y. Brun and H. Guan. 'Attack-Resilient Image Watermarking Using Stable Diffusion.' arXiv: 2401.04247 [cs], Accessed: 3rd Dec. 2024. [Online]. Available: http://arxiv.org/abs/2401.04247, pre-published.

[45] A. Rezaei, M. Akbari, S. R. Alvar, A. Fatemi and Y. Zhang. 'LaWa: Using Latent Space for In-Generation Image Watermarking.' arXiv: 2408.05868 [cs], Accessed: 3rd Dec. 2024. [Online]. Available: http://arxiv.org/abs/2408.05868, pre-published.

[46] H. Ci, Y. Song, P. Yang, J. Xie and M. Z. Shou. 'WMAdapter: Adding WaterMark Control to Latent Diffusion Models.' arXiv: 2406.08337 [cs], Accessed: 10th Dec. 2024. [Online]. Available: http://arxiv.org/abs/2406.08337, pre-published.

[47] I. S. Reed and G. Solomon, 'Polynomial Codes Over Certain Finite Fields,' *Journal of the Society for Industrial and Applied Mathematics*, vol. 8, no. 2, pp. 300–304, Jun. 1960, ISSN: 0368-4245, 2168-3484. DOI: 10.1137/0108018. Accessed: 20th Jul. 2025. [Online]. Available: http://epubs.siam.org/doi/10.1137/0108018.

[48] B. An et al. 'WAVES: Benchmarking the Robustness of Image Watermarks.' arXiv: 2401.08573 [cs], Accessed: 8th Jul. 2025. [Online]. Available: http://arxiv.org/abs/2401.08573, pre-published.