Robustness of AI-Generated Image Detection Against Localized Inpainting Attacks

Oğuz AKIN
Saarland University, CISPA Helmholtz Center for Information Security
ogak00001@stud.uni-saarland.de
September 9, 2025

***Abstract*—The incredible rise in abundance of high-fidelity AI-generated images (AIGI) makes it a must to have robust detection methods, but their ability and resilience against popular and established post-processing attacks like localized inpainting still remains as a critical and not so much explored vulnerability. This thesis shows a systematic evaluation of different AIGI detection paradigms against a diversified batch of inpainting attacks to carefully separate between two separate and fundamental failure types: "evasion," where an inpainted fake image is incorrectly misclassified as real, and the generation of "false positives," where a mildly edited real image is incorrectly classified as real. Our results show that almost all evaluated detectors, including both reactive and proactive types, display a dangerously unbalanced performance under such attacks, which forces a critical trade-off between security and usability at the end of it all. We find that the artifact-based detector, DIMD, proves that it is able to withstand the attack of evasion, with an Attack Success Rate (ASR) that is consistently below 12%, but at the cost of a significant false positive rate by incorrectly flagging over 25% of benignly edited real images as real. In an almost complete opposite outcome, the reconstruction-based detector AERO- BLADE is successfully stable against false positives where it correctly identifies almost all inpainted real images, but is dangerously defenseless against evasion attacks with ASRs reaching almost 100% on fakes even with smaller edits. Furthermore, the generalist detector, UFD, even though it has a very strong baseline AUC of 0.92, showed big vulnerability to both of these failure types. Even proactive watermarking methods, most of the time considered as a robust solution, were found to be breakable and unbalanced. While demonstrating high robustness at specific high- confidence thresholds (e.g., ASR often 1% or less at 99% TPR), their performance collapsed at other, equally valid operating points: Tree-Ring's watermark was not detected in over 99% of cases at a balanced threshold, and Stable Signature proved highly vulnerable (ASR ¿80%) when calibrated for a low false-positive rate. These findings demonstrate that baseline performance is a poor predictor of robustness against different types of attacks and that these current models carry significant risk and unbalanced vulnerabilities. This work concludes that a universally robust, "one-size-fits-all" detector does not yet exist and puts the critical need for threat-model-specific and threshold-aware evaluation to avoid a false sense of security under broad daylight since with great power comes great responsibility.**

I. Introduction

Over the past ten years, image generation technologies have changed from niche research curiosities into globally accessible creative tools. Beginning with Generative Adversarial Networks (GANs) and culminating in the recent dominance of diffusion models, modern generative methods now allow the synthesis of photorealistic content at a level that rivals human-created photography. Latent diffusion models such as Stable Diffusion and proprietary systems like Midjourney and DALL·E have placed this capability in the hands of millions of everyday users. A simple text prompt is enough to produce visually coherent images of random content, style, and resolution. What once had the need of expensive equipment or professional editing can now be achieved in seconds on a laptop or even a mobile phone.

The societal benefits of this technology are of course undeniable: creative professionals get new tools for faster prototyping, artists find new means of expression, and educational and accessibility applications come out in abundance. However, with these benefits comes a darker prospect. The same power that enables artistic exploration also lowers the barrier for detrimental use. Images fabricated by generative models can be deployed in disinformation campaigns, in falsifying documents, in synthetic identities, or in deepfakes that attack personal dignity. The fast increase of AI-generated images (AIGI) directly weakens a long-standing assumption—that a photograph is a reliable witness to reality. When this assumption goes away, so too does an important bedrock of trust in digital media.

To counter this, the computer vision community has invested significant time and effort in AI-generated image detection. These detection systems come from multiple paradigms. Proactive watermarking approaches modify the generation process itself, embedding hidden but verifiable signature-like signals that persist into the output image. Passive detection methods train discriminative models to recognize statistical or semantic artifacts left by generative pipelines. Training-free reconstruction methods exploit generative priors or autoencoders to measure reconstruction error asymmetries between real and synthetic content. Each paradigm has seen a very fast innovation: Stable Signature and Tree-Ring in watermarking; UniversalFakeDetect (UFD) and DIMD (Corvi et al.) in passive detection; and DIRE and AEROBLADE in training-free methods. Together, these techniques represent the state of the art in safeguarding authenticity in the age of generative media.

However, the adversarial nature of this domain makes it so that no detector exists in a vacuum. Once detectors are deployed, attackers adapt. A fundamental challenge thus rises: detectors are evaluated primarily on clean, unedited images, whereas real-world adversaries work on post-processed content. In practice, generated images are rarely shared "as-is." They are cropped, compressed, resized, filtered, or inpainted. Such edits are not only common in

non-harmful user workflows but are also in such a state for the adversaries to take advantage of them to evade detection. This showcases an important shortfall: detectors validated on benchmark datasets with pristine outputs may fail incredibly in deployment scenarios where post-processing is very easy to apply.

Among these post-processing techniques, image inpainting represents an explicitly dangerous and mostly unknown threat. Inpainting was historically designed to restore damaged photographs or remove unwanted objects and it now has become a mainstream feature in modern editing software. Algorithms such as LaMa and ZITS, when trained on large-scale image completion tasks, can fill masked regions with semantically suitable content, preserving both local texture and global structure. From the perspective of a day to day user, this is a very useful editing feature. From the perspective of an adversary, it is an incredibly strong weapon:

Inpainting can erase watermarks. If a watermark signal is spatially concentrated or distributed in predictable regions, inpainting can overwrite those regions and weaken or destroy its detectability. It also can sanitize fingerprints. Passive detectors rely on subtle low-level cues left by generative pipelines. Re-synthesizing even small regions can overwrite or dilute these important signals and nudge detector scores towards the "real" side of the decision boundary. One other thing it can do is that it can create false positive classifications. When applied to authentic real images, inpainting inserts synthetic patches. Detectors may mistake these small edits for forgery, leading to the misclassification of real photographs.

What makes inpainting especially dangerous is its accessibility. Unlike adversarial modifications, which require white-box access or optimization against a particular model, inpainting is a general-purpose editing tool. It does not need to "know" the detector; it simply alters the image. Once an attacker sets up an automated inpainting pipeline, they can apply it to virtually any image with no extra effort. This asymmetry between the defender and the attacker, where the attacker's cost is minimal and the defender's robustness requirement is universal, shows an explicitly threatening post-processing attack.

Despite this, systematic studies of detector robustness against inpainting attacks are noticeably not present. The literature has many evaluations of detectors against cross-model generalization, training data diversity, or distributional shifts. Some works check the resilience to JPEG compression or Gaussian noise. However, only a few consider inpainting in a controlled, automated manner. Those that do often lack methodological consistency across detectors or confound the analysis by allowing semantic drift—comparing unrelated real and fake images rather than matched pairs. This omission is critical, because the practical deployment setting is not "clean fakes vs. clean reals," but "edited images vs. authenticity claims."

This thesis seeks to fill that gap. We pose the research question: **How robust are state-of-the-art AI-generated image detectors across multiple paradigms when faced with localized inpainting attacks?**

To answer it, we construct a benchmark that emphasizes control, reproducibility, and fairness. Our methodology begins with careful dataset generation, using matched pairs of real and inpainted images to isolate the effect of the attack. We then implement a deployment-faithful evaluation protocol where detectors are calibrated once on clean data and then tested against inpainted images using a fixed threshold. The Attack Success Rate (ASR) serves as the primary robustness metric, carefully distinguished for two failure modes: evasion on fakes and false positives on reals.

Across this framework, we evaluate six detectors coming from three paradigms and reveal a critical and previously not measured trade-off inherent in their design. The results show that reactive detectors carry dangerously "polarized" behaviors: the artifact-based **DIMD** is highly resilient to evasion but suffers from a high false positive rate on edited reals, while the reconstruction-based **AEROBLADE** shows the opposite by behaving strong against false positives but failing catastrophically vulnerable to evasion. Furthermore, proactive watermarking schemes, often considered a definitive solution, also proved to be brittle, with both **Stable Signature** and **Tree-Ring** failing completely at certain balanced operating points.

The contributions of this thesis are therefore summarized in three sections:

- A reproducible robustness benchmark for inpainting, incorporating matched real/inpainted pairs, standardized preprocessing, and a fixed-threshold evaluation protocol.
- An analysis between different paradigms that reveals a fundamental security-versus-usability trade-off to demonstrate that current reactive detectors are often dangerously unbalanced.
- New findings that reshape assumptions: high baseline AUC does not mean robustness, and proactive watermarks, while strong at some thresholds, can be exceptionally fragile at others, challenging their status as a universal solution.

At the end of it all, this thesis sheds light on inpainting as a central challenge in the developing contest between generative synthesis and forensic detection. By identifying these common and unbalanced vulnerabilities, it gives a foundation for future research and guides practical deployment strategies for maintaining trust in visual media.

II. Related Work

The field of AI-generated image detection is developing at a speed that mirrors the fast advancements in generative modeling itself. To navigate along this complex landscape, it is a necessity to establish a clear taxonomy of existing approaches. The most effective categorization is based on the point of intervention in the image lifecycle, resulting three distinct paradigms where each represents a different philosophy of detection: proactive methods that embed signals during generation, reactive methods that analyze finished images for artifacts, and generalist methods that leverage fundamental image properties without specialized training. This review identifies the six detectors selected for this thesis within their respective paradigms, focusing on their core

principles, theoretical assumptions, and the hypothesized vulnerabilities they might show when attacked with localized inpainting.

A. Proactive Defenses: Watermarking Methods

Watermarking represents a proactive, "first-line-of-defense" strategy where a signature that cannot be seen is embedded into an image to allow for later source authentication. These techniques are conditioned on a "closed-world" assumption, requiring the cooperation of the entity generating the image, as the signal must be injected during the synthesis process. While this limits their applicability, it offers the significant advantage of embedding a signal that is deeply and intentionally integrated into the image's structure, rather than being a superficial and unintentional artifact. This deep integration is the central pillar of their claimed robustness against post-processing.

- **Stable Signature** [1] is an *in-processing* method that embeds a watermark directly and spatially into the generative model's **latent space**. The technique is elegant yet powerful: it works by including the standard VAE (Variational Autoencoder) decoder in a latent diffusion pipeline with a custom, fine-tuned decoder. This modified decoder is trained with a dual objective: to reconstruct the image from its latent representation with high fidelity, and to simultaneously embed a specific binary message. The resulting signal is spatially distributed throughout the image's pixel grid, entangled with the very features that form the image content. The theoretical strength of this approach is in this distributed background; a localized spatial attack like inpainting, by definition, can only alter a fraction of the image. To completely get rid of a spatially distributed watermark, an adversary would theoretically need to alter such a large portion of the image that its original utility is destroyed.
- **Tree-Ring Watermark** [2] is another *in-processing* technique that also works in the latent space, but its philosophy is global rather than spatial because it puts its signal in the **frequency domain**. During the final decoding part, it just manipulates the latent-to-pixel mapping to create a model-specific signal that manifests as a faint and imperceptible "ring" pattern in the image's Fourier spectrum. Detection is a more complex, multi-step process involving DDIM inversion to approximate the image's latent representation, followed by a correlation check in the frequency domain against a known secret key. This key-based mechanism makes it possible to have cryptographic security against forgery but introduces a potential point of fragility. The theoretical interaction with inpainting is ambiguous: a localized spatial edit causes a distributed, but potentially weak, disturbance in the global frequency domain. It is an open experimental question whether this disturbance is enough to break the correlation with the key and thus defeat the detector.

B. Reactive Defenses: Passive Detection Methods

In contrast to proactive watermarking, passive detectors work as reactive, AI-based "forensic detectives." They are trained to find the unintentional artifacts or statistical "fingerprints" left by generative models. These models analyze any given image without prior information which makes them universally applicable. This paradigm is the mostly researched in AIGI detection, but it faces the main problem of distinguishing intentional generative artifacts from natural image statistics or benign, man-made edits.

- **UniversalFakeDetect (UFD)** [3] was designed to take care of the important problem of generalization. Rather than looking for low-level pixel artifacts, which can be highly specific to a single generative model, it works on a **high-level semantic** feature space. It makes use of the powerful, general-purpose representations of a pre-trained vision-language model, CLIP, which has been trained on billions of image-text pairs from the web. The core principle is that the real and fake images, despite their visual quality, form a distinct and separable cluster in CLIP's high-dimensional embedding space. UFD's main thing is that generative models, for all their quality, have not yet perfectly captured the subtle, high-level distribution of natural scenes. However, its reliance on semantics suggests a profound theoretical vulnerability: an inpainting attack, especially a semantic one, is expected to be plausible. It preserves the overall meaning of the scene. Therefore, it should produce an image that remains "in-distribution" from CLIP's perspective, making it a theoretically possible attack vector.
- **DIMD (Corvi et al., 2023)** [4] represents the opposite philosophy: specialization over generalization. It is a modern **low-level, artifact-based** approach using a Convolutional Neural Network (CNN). The model is trained to identify the small, distinct fingerprints and frequency imbalances characteristic of modern diffusion models, which can come from specific, known parts of the generation pipeline, such as upsampling layers that create checkerboard artifacts or the underlying noise schedules. By focusing on these fundamental, process-level artifacts, DIMD is expected to be more resilient to content-level manipulations. This sets up an important hypothesis for our study, framed as a question of "signal preservation": if an attacker in-paints 25% of an image, do the tell-tale generative artifacts in the remaining 75% provide enough of a signal to overcome the "noise" from the new, inpainted patch?

C. Generalist Defenses: Training-Free Methods

This developing paradigm gives a strong alternative that does not need training on large datasets of "fake" images which gives it a significant potential advantage in detecting images from unseen and new generators. These methods work on "first principles," that use the inherent properties of generative models or real images to create a classification score which is mostly based on reconstruction error.

- **DIRE (Diffusion Reconstruction)** [5] uses a pre-trained **diffusion model itself** as a reconstruction engine. The

method is built on the complex principle of a generative prior. The main claim is that one of "error asymmetry": a diffusion model, being a "real world model" for its training data, can reconstruct images from its own "native" distribution (i.e., AI-generated images) with very low reconstruction error, whereas it will struggle to reconstruct an "out-of-distribution" real-world photograph which results in a higher error. The measure of this error is used as the classification score. This leads to a distinct and testable hypothesis for an inpainted fake: since both the original image and the inpainted patch are products of diffusion models, the entire image should be "native" to DIRE's world model and thus be reconstructed with low error to preserve its "fake" classification.

- **AEROBLADE** [6] uses a standard and comparatively simple **autoencoder** for reconstruction, but its main assumption is the inverse of DIRE's. It claims that the "clean," low-complexity outputs of AI models can be compressed and decompressed with lower error than noisy, high-entropy real-world images. The primary error metric used is the LPIPS distance which is designed to align with human perceptual similarity. This approach depends not on a powerful generative prior but on a more general statistical difference in terms of complexity between synthetic and natural real image manifolds. This sets up an alternative hypothesis for inpainted fakes: the process of inpainting, by stitching a new patch into an existing image, introduces boundary artifacts and disrupts the pristine and homogenous signature of the original AI image. This may add a form of perceptual "noise" or complexity that pushes its reconstruction error *higher* and thus making it look more like a high-entropy 'Real' image to the autoencoder. This gives a clear theoretical vector for an evasion attack.

III. Methodology

This thesis follows a carefully set and controlled data-driven experimental framework to evaluate the robustness of different AI-Generated Image (AIGI) detection methods against localized image inpainting. The methodology was carefully set up to ensure fairness, reproducibility, and scientific validity across all tested paradigms. It is composed of five key stages: (1) a careful data curation and standardization process designed to isolate the attack's impact to localized inpainting itself; (2) a detailed description of the standardized inpainting attacks that form our threat model; (3) a summary of the specific implementation and preprocessing pipelines for each detector; (4) a deployment-faithful experimental protocol that reflects the realistic operational scenarios; and (5) a precise and systematic definition of our evaluation metrics and thresholding policies.

A. Data Curation and Standardization

All experiments were done on standardized and balanced datasets, curated from large-scale and authoritative benchmarks. A fixed size of N=200 images per class was chosen for the source datasets as a practical compromise, while being statistically enough for metric estimation.

1) Baseline and Robustness Sets:

- **Robustness Sets (from Semi-Truths):** The main evaluation was done using images derived from the Semi-Truths dataset [7], a benchmark specifically designed for robustness evaluations. A custom data extractor was developed to source images from the *OpenImages* subset and its AI-inpainted counterparts. This extractor enforced a strict **1:1 matched pairing**: for each of the 200 pairs, a real source image is paired directly with its own inpainted version. This strategy provides a crucial scientific control by isolating the causal effect of the inpainting attack from the underlying image content.
- **Baseline Fake Set (from GenImage):** To establish baseline performance, a diversified set of 200 completely AI-generated images was sourced from the GenImage benchmark. GenImage was selected for its broad coverage of modern generative models (including Midjourney, Stable Diffusion, and DALL-E 2) and its stable, reproducible access. This set serves as the "complete fake" class for calibrating the passive and training-free detectors.

2) Data Hygiene and Canonical Formatting: All images were handled in a strict data hygiene protocol. To prevent issues discovered in preliminary testsvbefore the definitive ones, ground-truth labels were re-derived directly from the folder structure, ensuring complete label integrity. Subsequently, a canonical evaluation set was created by converting all images to a **512×512 PNG format** which was then altered to server for each detectors assumption about the input image compression type and size. This standardized format serves as the default input for the majority of the evaluated detectors.

B. The Standardized Inpainting Attack

The threat model for this thesis is based on a standard and realistic inpainting attack pipeline, applied uniformly across all relevant experiments.

- **Inpainting Model:** The primary model used was **LaMa (Large Mask Inpainting)** [8] which is renowned for its high-fidelity results. To make it completely reproducible we made use of the well-maintained `simple-lama-inpainting` library. This choice was made after confirming that the official checkpoints for the original Big-LaMa model were no longer publicly available, thus prioritizing replicability without sacrificing the attack's quality. To be able to compare our results, attacks with the **ZITS** [9] model were also employed.
- **Mask Generation Strategies:** To evaluate the robustness of the aforementioned models comprehensively, we generated masks using different strategies and made use of the semantically inpainted real images where Semi-Truths paper already used semantic masks to produce such images.
  - *Semantic Masks:* In Semi-Truths, semantic masks were generated automatically using LLaVA-Mistral-7B in a zero-shot setting that found object regions in images and produced binary masks for them. These masks were

then used to guide conditional inpainting of semantically meaningful areas.

- *Random Blob Masks:* Irregular, "blob-like" masks were generated and then stratified into four bins by their area sizes relative to the image they would be applied on: 0-3% (bin 1), 3-10% (bin 2), 10-25% (bin 3), and 25-40% (bin 4).
- *Random Rectangle Masks:* As a simple geometric baseline, rectangular masks of random size and position were also used.

C. Detector Implementation and Preprocessing

To make sure to remain faithful to the original research, all detectors were implemented using their official public codebases and pre-trained weights. Due to possible significant software dependency conflicts, each detector or family of detectors was installed in a separate, isolated Conda environment. Custom scoring scripts were written to wrap the core detection logic and created a fixed pipeline that takes an image folder as input and produces a standardized CSV file with 'filepath', 'ground truth label', and 'raw score' columns. Specific preprocessing steps were tailored to each detector's requirements.

- **UniversalFakeDetect (UFD):** This model uses an OpenCLIP ViT-L/14 backbone. All input images were processed using the standard CLIP pipeline: resized to 224×224, center-cropping, and normalization using the prescribed ImageNet statistics.
- **DIMD (Corvi et al.):** This detector was evaluated using the open benchmark harness SIDBench [10]. This system ensures that the exact 256×256 preprocessing, resizing, and normalization steps used in the original paper are faithfully replicated.
- **DIRE:** Following the authors' recommendation for general-purpose analysis, we used the official ImageNet-256 unconditional ADM diffusion model. This required all input images to be resized to 256×256.
- **AEROBLADE:** This method's official protocol specifies that the input needed to be from a size of 512×512 pixels. To satisfy this requirement, a separate, complete version of the entire dataset was created. All images were center-cropped and resized to 512×512.
- **Stable Signature:** Implementation required patching the official 'diffusers' library to inject the authors' custom watermarked VAE decoder into the Stable Diffusion pipeline. A bit-accuracy sanity check was performed to confirm the watermarking VAE was active. The corresponding TorchScript detector was used as it is.
- **Tree-Ring:** The official detection script was modified to extract the underlying continuous distance score, which was essential for a proper ROC analysis.

D. Experimental Protocol

The experiment is divided into two separate workflows, each customed to the fundamental operating principles of the detector category that is being tested.

1) Protocol for Passive and Training-Free Detectors: A strict two-stage process was used to prevent any form of data leakage.

1) **Stage 1: Baseline Calibration.** Each detector is evaluated on the baseline split (real vs. completely fake images). The sole purpose of this stage is to determine and freeze a single, optimal decision threshold, **t***.
2) **Stage 2: Robustness Test.** The detector is then evaluated on the various robustness splits. Performance is measured against the *fixed threshold t** from Stage 1.

2) Protocol for Watermarking Systems: A "generate-attack-detect" loop was implemented.

1) **Generate:** A set of 200 images is generated with the watermark embedded.
2) **Attack:** The watermarked images are subjected to the standardized inpainting attacks.
3) **Detect:** The detector is run on both clean and attacked sets to measure watermark survival.

E. Evaluation Metrics and Thresholding

All raw detector scores were systematically organized such that a higher score consistently indicates a higher likelihood of being "fake/generated."

1) Primary Robustness Metrics: Two Failure Modes: The Attack Success Rate (ASR) is the primary metric for measuring vulnerability. To capture two distinct real-world failure types for passive and training-free detectors, we define two separate forms of ASR:

- **$\text{ASR}_{\text{Real-False Positive}}$:** This metric measures the detector's resilience against producing false positives on benignly edited authentic images. It gives an answer to the question: *"What percentage of inpainted **real** images are incorrectly classified as Real?"* A high value for this ASR is indicating a problem, as it indicates a low false positive rate.
- **$\text{ASR}_{\text{Fake-Evasion}}$:** This metric measures the success of a classic evasion attack. It gives an answer to the question: *"Of the fake images that were correctly detected at baseline, what percentage are successfully misclassified as Real after inpainting according to the baseline threshold?"* A high value for this ASR is a trouble, indicating a vulnerable detector.

For watermarking systems, ASR is defined as the percentage of attacked images where the watermark is no longer detected at a fixed True Positive Rate operating point.

2) Thresholding Policy and Additional Metrics:

- **Thresholding Policy:** The decision threshold **t*** used in ASRs for detectors is set **only on the baseline split** using **Youden's J statistic**. This method finds the point on the ROC curve that maximizes 'TPR - FPR', giving an optimal trade-off.
- **Additional Metrics:** To have a complete performance overview, we also use:
  - **AUC (Area Under the ROC Curve)**: Computed from both **Baseline** and **Robustness** datasets.

– **ΔAUC**: The difference between ‘Baseline AUC‘ and ‘Robustness AUC‘ that indicates the difference in overall separability.

IV. Experimental Results

This section presents the complete measurement outcomes of our experiments, following strictly to the data generated from the protocols detailed in the Methodology section. The results are organized by detector, with each sub-section providing a complete explanation of its performance against the different inpainting attacks. The findings are explained in detail by referencing the comprehensive data tables which are located in the Appendix. All comments and interpretations of these results are given under the Discussion section follows this section.

A. UniversalFakeDetect (UFD)

The UFD detector was first calibrated on the baseline dataset of clean real and fake images, where it established a strong initial performance with an **AUC of 0.9168**. This indicates a high ability of separation between the two classes in a uncorrupted, unedited environment that sets a high expectation for its performance. From this calibration, a fixed decision threshold of t* = 0.348135 was computed using Youden’s J statistic. At this threshold point, the detector was set up for high precision and obtained a True Positive Rate of 25.0% at a very low False Positive Rate of 1.0%. This fixed baseline threshold was used for all subsequent robustness evaluations to simulate a real-world deployment scenario.

1) Robustness to Inpainted Reals: The performance of UFD against real images that have received inpainting attacks is detailed in Table II. The ASR here measures the percentage of inpainted real images that the detector fails to flag and incorrectly classifies them as uncorrupted ‘Real‘ images. A high ASR therefore means a high rate of failure for the detector. For the primary semantic attack, the detector gave an **ASR of 50.0%**, meaning its performance was no better than random chance according to the threshold set at the baseline. It failed to detect the manipulation in exactly half of the cases. The results between different mask types and inpainters show a great measure of difference and in many cases complete vulnerability. For example, both LaMa with rectangular masks and ZITS with medium-sized blobs (bin3, 10-25% area) caused an almost complete collapse in the detector’s ability to spot inpaintings, with ASRs of **98.5%** and **99.5%**. In these scenarios, the attack was almost perfectly successful. A big difference was seen with the LaMa inpainter when using very large masks (bin4, 25-40% area); in this case, the ASR dropped to a much lower **23.0%**. This indicates that only when the inpainting artifacts themselves became so prominent and crude then did the detector succeed in identifying the image as inpainted with any consistency.

2) Robustness to Inpainted Fakes: UFD’s robustness against evasion attacks is visualized in Table III. The detector can be seen to be highly and consistently vulnerable to this attack vector. The semantic inpainting attack on a set of fake images was completely successful, achieving a **100.0% evasion ASR**. This indicates that using a suitable semantic edit to a fake image is a guaranteed method for bypassing the UFD detector. With the LaMa inpainter, a clear and increasing progression came out, demonstrating a dose-response relationship between the size of the edit and the success of the attack: the evasion ASR increased continuously with the size of the inpainted region that started at a very strong **64.0%** for the smallest blob masks and rising to a near-complete success rate of **98.0%** for the largest. The ZITS inpainter was also very effective since it achieved a perfect **100.0% ASR** for the largest masks (bin4) and a near-perfect **99.0%** for medium-sized masks (bin2). The only big exception to this pattern of vulnerability was the ZITS bin3 attack, which gave a lower but still very significant ASR of 45.0%.

B. AEROBLADE

The AEROBLADE detector achieved a strong baseline performance of an AUC of 0.8160. The fixed decision threshold derived from this baseline calibration was t* = 0.156877.

1) Robustness to Inpainted Reals: As shown in Table IV, AEROBLADE was found to be dangerously defenseless when used for detecting inpaintings on real images. In every single one of the eleven tested scenarios, the ASR was very high which indicated a consistent failure to identify the inpainted patches. The ASR ranged from a low of **76.0% to a high of 98.5%**. This indicates that the detector consistently failed to identify that the real images had been inpainted, incorrectly classifying them as uncorrupted ‘Real‘ images the majority of the tests. This makes it completely ineffective for classifying small edits correctly or flagging detrimental tampering on authentic photographs.

2) Robustness to Inpainted Fakes: Similarly, AEROBLADE was found to be very vulnerable and open to evasion attacks (Table V). The attack was most effective when using small masks which means this is a critical vulnerability. Both the LaMa and ZITS inpainters obtained a perfect **100.0% evasion ASR** on the smallest blob masks (bin1). This implies that even a tiny localized edit covering less than 3% of the image area was enough to completely fool the detector and make a fake image appear real. While the evasion success rate tended to decrease as the mask size became bigger, it still remained very high across all conditions. The lowest recorded ASR across all these ten attack configurations was still **51.7%**, meaning that even in the absolute best-case scenario for the detector, the inpainting attack was successful more than half the time. This makes it clear that the detector is ineffective for security applications.

C. DIMD (Corvi et al.)

The DIMD detector that was configured with a top-5 rank ensemble strategy established a modest baseline performance with an AUC of 0.6384. The corresponding fixed baseline decision threshold for all robustness tests was t* = 0.6512.

1) Robustness to Inpainted Reals: The detailed results in Table VI show that DIMD continuously failed to detect inpaintings on real images similar to AEROBLADE. The ASR here remained in a very tight and high range of **73.5% to 77.5%** in all eleven tested conditions. This indicates

that the detector is fundamentally blind to the presence of localized synthetic content on a real image and reliably fails to flag the manipulations almost three-quarters of the time. An important result for this detector was the notable consistency of this failure. The ASR is almost identical for small blobs, large rectangles, and semantic edits. This suggests that the detector's failure is not dependent on the specifics of the attack but that it is a systemic weakness.

2) Robustness to Inpainted Fakes: In complete opposition to its complete failure on real images, DIMD proved to be very strong against evasion attacks (Table VII). The evasion ASR remained noticeably low across all conditions which meant that it showcases a strong security profile against this specific threat. For attacks with small to medium sized masks (bins 1-3), the evasion rate was consistently below **4%** which indicates an almost complete failure of the attack and near-perfect security for the detector. Even when the inpainted area was very large (bin4, 25-40% of the image) the evasion ASR only increased to a modest maximum of **11.84%**. This shows a very polarized performance: the detector is secure against evasion but completely fails to detect inpaintings on real images.

D. DIRE

The DIRE detector showed a modest baseline performance with an AUC of 0.6588. The fixed decision threshold was calculated to be t* = -0.014894 at the baseline.

1) Robustness to Inpainted Reals: DIRE's performance on inpainted real images was inconsistent and generally low (Table VIII). For the semantic inpainting attack, it yielded a mediocre ASR of **62.5%**. Its performance against irregular blob masks was even worse, with ASRs consistently in the range of **85.5% to 89.5%**, indicating a very high failure to detect these inpaintings. However, a significant and surprising outlier was observed as its ASR against LaMa rectangular masks was **0.0%**, meaning it correctly flagged every single one of these edited real images as fake. This great usability failure for a specific, simple geometric edit highlights the detector's unpredictable and unreliable nature.

2) Robustness to Inpainted Fakes: The performance of DIRE against evasion attacks is shown in Table IX. The detector showed a moderate vulnerability that clearly scaled with the size of the inpainted region. For the smallest blob masks (bin1), the evasion ASR was very low, at only **2.7%**, indicating high security against small edits. However, as the mask area increased, the attack became progressively more effective, with the evasion rate climbing continuously to a notable **22.97%** for the LaMa inpainter with the largest masks.

E. Proactive Watermarking Detectors

The two watermarking systems were evaluated using the aforementioned generate-attack-detect protocol. Thresholds were first calibrated on a set of clean, watermarked images and real images to find various operating points. The t90 and t99 thresholds are predetermined operating points and are different than the threshold that is determined at the baseline level

1) Stable Signature: Stable Signature first achieved a baseline AUC of 0.761 against real images. Its robustness against the following inpainting attacks is shown in Table X. At the conservative t90 and t99 threshold points, the ASR (rate of successful watermark removal) was consistently low, ranging from **8.5% to 14.0%** at t90 and never exceeding **1.0%** at t99. However, the results also revealed a critical vulnerability at a different operating point: at a threshold calibrated at the baseline, the ASR surged to over **80%** across all attack types, indicating that its robustness is highly dependent on the chosen threshold. If a real world scenario was to be replicated, then it would be the case that the baseline threshold would be used and higher ASRs would be received by the Stable Signature detector.

2) Tree-Ring: Tree-Ring established a higher baseline AUC of 0.8405 against real images. Its robustness results (Table XI) showed a strong dependency on the operating threshold. At the high t90 and t99 threshold points, the watermark removal ASR was very low, typically ranging from **4.0% to 11.0%** at t90 and often **0.0%** at t99. However, at the Equal Error Rate threshold (t_eer) that is determined at the baseline, a more balanced operating point, the detector showed a great vulnerability with the ASR reaching **99% to 100%** across nearly all attack types. This indicates that while the watermark is present, its signal is so degraded by inpainting that it is almost completely undetectable at a balanced, real-world threshold point.

F. Inpainting Quality Analysis

An analysis of inpainting quality was performed (Table XII). The metrics confirmed that the inpainted patches were of high fidelity, especially for smaller mask areas. For example, for small blobs on the treering set, the SSIM was 0.96, indicating near-perfect structural recreation. This evidence confirms that the previously mentioned detector failures cannot be attributed to basic poor inpaints, but rather to a real failure to handle the localized inpaintings.

V. Discussion

The experimental results presented in the previous section reveal a clear, multidimensional, and often surprising hierarchy of robustness against localized inpainting attacks. Beyond the raw metrics, this data tells a convincing story about the underlying strengths, weaknesses, and, most importantly, the background of trade-offs of different AIGI detection paradigms. In this section, we move from reporting to interpreting and analyzing the "why" behind the numbers to connect the findings back to the core design principles of each detector and synthesize these insights into a coherent narrative that directly addresses the central research question of this thesis.

A. The Universal Blind Spot: Failure to Detect Manipulated Real Images

The most important and consistent finding of this thesis is the near-universal failure of the tested reactive detectors to reliably identify inpaintings on real images. The ASR on inpainted reals—which measures the percentage of edited real images that are incorrectly passed as fakes—was alarmingly

high for almost every detector. This reveals a fundamental blind spot in the current paradigm of reactive detection. Models like **AEROBLADE** and **DIMD** consistently failed to detect these inpaintings over 70% of the time, with ASRs often exceeding 80% or 90% (Table IV and Table VI). **UFD** was only able to successfully detect inpaintings when the artifacts were noticable (e.g., large LaMa masks, with a low ASR of 23.0%), otherwise failing with ASRs as high as 99.5% (Table II).

This consistent failure comes from the very definition of the training task for these models. They have been trained to solve a binary classification problem: 'Complete Real' versus 'Complete AI-Generated'. Their internal representations have learned the statistical properties of these two separate classes. However, an inpainted real image is a hybrid of 'Real + Fake' composite that does not fit neatly into this binary worldview. The detectors, never having been explicitly taught to recognize the artifacts of a localized inpainting on a real image, default to the dominant class. Since the great majority of the image's pixels and features are from a real photograph, the models classify the entire image as 'Real', thereby failing to classifying the inpainting. This is an important failure for real-world deployment, as it means these detectors can give no protection against an adversary who wants to tamper with a real image maybe by removing a person from a photograph to change a historical record or create a piece of targeted disinformation.

B. The Polarization of Security: A Forced Trade-Off on Inpainted Fakes

While the reactive detectors were in total performing poorly at analyzing real images, they showed a dramatic polarization in their ability to handle inpainted fakes. This is where a critical security-versus-usability trade-off becomes clear which means forcing a choice between a detector that is secure against evasion and one that is not.

DIMD perfectly exemplifies the **secure but flawed** detector. It proved cery resistant to evasion, with an ASR consistently below 12% and often below 4% (Table VII). Its success validates its design principle: the CNN is trained to find the specific, low-level frequency artifacts characteristic of diffusion models. These artifacts are distributed throughout the image. When an attacker inpaints a part of the image, they only get rid of a part of this distributed signal. The uninpainted regions still contain a strong, uncorrupted generative fingerprint, which is more than enough for the specialized detector to make a confident classification. However, as established above, this very specialization makes it completely blind to edits on real photos, where it fails 75% of the time.

AEROBLADE and **UFD** represent the opposite: the **comprehensively insecure** detectors. Not only did they fail to detect inpaintings on real images, but they were also dangerously vulnerable to evasion on fake images, with ASRs frequently reaching 100% (Table III and Table V). For these detectors, any plausible inpainting edit serves as an effective "sanitization" technique that successfully fools them. In the case of AEROBLADE, its autoencoder expects "perfect" AI images to have a very low reconstruction error. The process of inpainting introduces boundary artifacts and textural inconsistencies, disrupting this uncorrupted signature. This added "noise" or complexity increases the reconstruction error, pushing the image across the decision boundary into the high-error manifold the detector associates with 'Real' images. For UFD, the semantic plausibility of the inpainted patch smooths over any potential high-level inconsistencies in the original fake, making it appear more coherent and thus more "real."

This opposing behavior uncovers a basic dilemma: the current reactive methods force a choice between a detector that is secure but classifies innocent users' edits (DIMD), or a detector that is safe for innocent users but offers no security against a determined attacker (AEROBLADE, UFD). No tested reactive method provided both security and reliability.

C. The Unpredictability of Complex Priors: The Case of DIRE

The performance of **DIRE** further complicates the narrative for reconstruction-based methods, acting as a cautionary tale against the assumption that a more powerful model is a more robust one. While it demonstrated some resilience to evasion on small masks (ASR of 2.70% on bin1 fakes), its performance on inpainted reals was erratic. The **0.0% ASR** on LaMa/randrect reals is particularly telling. According to our final metric definition, this is a perfect score—the detector successfully identified the inpainting in 100% of cases. However, it did so by flagging every single one of these small inpainted real images as fake, resulting in a 100% false positive rate. This extreme sensitivity to the geometry of an edit makes the detector dangerously unreliable for real-world use. A plausible explanation is that the simple, unnatural geometry of the rectangular patch is a shape so out-of-distribution for a diffusion model trained on natural scenes that the prior fails spectacularly to reconstruct it, producing a massive reconstruction error and thus flagging the image as fake.

D. The Illusion of Robustness: Deconstructing Watermark Brittleness

In contrast to the initial expectation that proactive methods would be a definitive solution, our final results show that they suffer from their own form of threshold-dependent brittleness. The "robustness" of a watermark is not an absolute property but is highly conditional on the chosen operational context and the associated decision threshold.

A surface reading of the results for **Stable Signature** and **Tree-Ring** at conservative thresholds (t90 and t99) would point towards high resilience. Indeed, at the t99 operating point, the watermark removal ASR was often 1% or less which indicates a very strong signal under the conditions where the detector is tuned for maximum sensitivity to find the watermark at all costs. This is the scenario often presented in literature, and it is a valid, though incomplete, picture. It shows that the main signal of the watermark is not completely erased by the inpainting attacks and can be found if one looks hard enough.

However, this apparent robustness is an illusion that breaks under more balanced, real-world conditions where false positives matter. For **Tree-Ring**, the results at the Equal Error Rate threshold (t_eer)—a standard, balanced operating point—were catastrophic. The ASR for watermark removal was **¿99%** across nearly almost all attack types (Table XI). This means that while the signal is detectable at a conservative threshold, it is so severely degraded by inpainting that it becomes completely useless at a balanced one, where the detector is trying to treat false positives and false negatives with equal importance. Similarly, **Stable Signature**, when evaluated at a threshold calibrated for a low false-positive rate (tFPR1), saw its removal ASR surge to over **80%** (Table X). This reveals a critical weakness: if the primary goal is to avoid falsely accusing unwatermarked images, the threshold must be set in a way that makes the watermark trivially easy to remove with a simple inpainting attack.

This demonstrates that watermarks, like passive detectors, are not a comprehensive solution. Their signals, while intentionally embedded, can be significantly damaged. Their success or failure is deeply dependent on the threshold, and a watermark that is "robust" at one extreme may be completely fragile at another. They do not solve the trade-off problem; they simply shift it to a different set of parameters.

VI. Conclusion and Future Work

A. Conclusion

This thesis employed a comprehensive and strict evaluation of AI-generated image detector robustness against the practical and accessible threat of localized inpainting. Through a variety of controlled experiments on six different detectors from three main paradigms, we have established a definitive hierarchy of performance, uncovered an important and previously unmeasured trade-off in reactive detector design, and provided strong evidence that challenges the current narrative of proactive watermarking as a universally superior solution. Our findings demonstrate that in the complex, adversarial ecosystem of post-processed media, almost all current detection methods show significant, unbalanced vulnerabilities, and that baseline performance on clean data is a poor and misleading predictor of real-world reliability.

Our primary contribution is the discovery of a universal blind spot in reactive detectors: an almost complete failure to identify manipulations on authentic images. Detectors such as **AEROBLADE**, **DIMD**, and **DIRE** consistently failed to classify inpainted real images, with Attack Success Rates (ASR) frequently exceeding 75%. This demonstrates that current models, trained on a simple 'Real' vs. 'Fake' binary, have not learned the more nuanced concept of an 'Edited' image, a critical failure for real-world scenarios where tampering with authentic documents or photos is a significant threat.

Our second major contribution is the identification of a polarized security trade-off among detectors that do show some resistance. We found that the artifact-based detector, **DIMD**, was very secure against evasion on fake images (ASR ¡12%), but this security was checked with its aforementioned failure to detect edits on real images. In contrast, detectors like **UFD** and **AEROBLADE** were not only insecure against evasion on fakes (ASR often 100%) but also failed to detect edits on reals, rendering them even more vulnerable. This sets a clear hierarchy: no reactive detector was successful on both fronts, and most failed on both which forces an unacceptable choice between a detector that is secure but flags innocent users, or one that is safe for users but offers no security.

Finally, our results lead to a significant re-evaluation of the role of proactive watermarking. While often told to be the definitive solution, our findings reveal that these methods are not a comprehensive solution but suffer from their own form of threshold-dependent brittleness. While both **Stable Signature** and **Tree-Ring** demonstrated high robustness at specific, conservative high-confidence thresholds, their performance collapsed at other, equally valid threshold points. The almost complete failure of Tree-Ring at a balanced threshold (ASR ¿99%) and the high vulnerability of Stable Signature at a low false-positive threshold (ASR ¿80%) show that their signals can be dangerously degraded by inpainting. They do not solve the trade-off problem; they just shift it to threshold selection. The core conclusion of this work is that a universally robust, "one-size-fits-all" detector does not yet exist. All evaluated paradigms have significant, unbalanced vulnerabilities that can be exploited by a simple inpainting attack, challenging the current foundations of AIGI detection and showing the deep need for threat-model-specific and threshold-aware evaluation in future research.

B. Limitations

While this thesis provides a comprehensive analysis, its conclusions are bounded by the deliberate scope of our experiments. Acknowledging these limitations is important for contextualizing the results and providing a transparent and scientific foundation for future research.

- **Attack Vector Specificity:** Our study focused exclusively on inpainting as the threat vector. This was a deliberate choice due to its popularity and its unique ability to create semantic and misleading hybrids that test multiple detector assumptions simultaneously. However, our findings on detector robustness do not necessarily generalize to the full spectrum of common post-processing attacks. Other manipulations, such as aggressive JPEG compression, Gaussian noise, adversarial blurring, color space manipulations, or geometric transformations like cropping, resizing, and rotation, test different aspects of a detector's invariance and could reveal entirely different vulnerabilities. A detector robust to inpainting might be fragile to compression, which targets high-frequency artifacts, and vice versa.
- **Inpainting Model and Mask Diversity:** To ensure reproducibility, we standardized our primary attack using the LaMa model as a high-quality example of a modern inpainting algorithm, with ZITS used for comparison. The larger space of inpainting models (e.g., MAT, newer diffusion-based inpainters) may produce different artifactual signatures. It is understandable that some detectors may be unintentionally overfitting to the specific textural patterns or boundary artifacts of LaMa, and their perfor-

mance could shift against a different inpainting algorithm. Similarly, while our mask generation strategies were different, they do not cover all possible editing scenarios.

- **Granularity of Analysis:** While we tested against masks stratified by size, our primary analysis focused on the aggregate ASR per category. This provides a clear top-level view but elides finer details. A deeper statistical analysis correlating the precise percentage of inpainted area, the number of inpainted objects, or the semantic importance of those objects with the continuous-valued detector score could give a more subtle mathematical model of detector fragility. Such an analysis could identify precise "breaking points" or non-linear relationships between the extent of an edit and the probability of evasion, moving beyond a simple binary success/fail metric.
- **Detector and Dataset Scope:** The six detectors chosen are highly representative of their respective paradigms, but they are not a complete sample of the entire field. Other promising architectures and approaches exist. Likewise, the datasets, while carefully curated from large-scale benchmarks, represent a specific distribution of real-world and generated images (primarily natural scenes). The performance of these detectors and the specific trade-offs we observed may vary on different types of content, such as portraits, architectural photography, medical imagery, or images containing significant amounts of text, all of which have different statistical properties that may interact with the detectors' assumptions in unique ways.

C. Future Work

The findings and limitations of this thesis open many clear and compelling perspectives for future research which are at solving the problems and trade-offs found in this work.

- **Solving the 'Edited Real' Problem:** The shared failure to detect manipulations on real images is the most important problem sound. Future work must go beyond the 'Real vs. Fake' binary and more towards a sophisticated, multi-class classification paradigm like 'Uncorrupted Real' vs. 'Edited Real' vs. 'Fake'. This would need the creation of new, large-scale datasets of small edited real images and training models with architectures (e.g., multi-head classifiers) designed to explicitly recognize the artifacts of editing as separate from the artifacts of generation.
- **Broadening the Attack Surface to a Holistic Benchmark:** An important next step for the community is to extend this evaluation framework to the wider range of post-processing attacks mentioned above. Systematically building a multi-attack benchmark—a true "gauntlet" of adversarial manipulations including compression, noise, blurring, and geometric transforms with different parameters—would create a more comprehensive and realistic stress test for AIGI detectors. This would help move the field away from the misleading practice of evaluating on clean data and toward a more mature model of robustness certification, where a detector's performance is represented not by a single AUC score but by a multi-dimensional robustness profile or scorecard.
- **Exploiting Edit-Specific Artifacts:** The catastrophic 100% false positive rate of DIRE on rectangular masks on real images, while a failure, proves that the edit itself can be a powerful signal. Future work could focus on developing a new class of *inpainting-aware* or "edit-specific" detectors. Such a model would be trained not to ask "Is this image real or fake?" but rather "Does this image show evidence of high-quality, localized inpainting?" This could be trained on a dataset of original, mask, and inpainted-image triplets to specifically identify the subtle boundary artifacts and inconsistent JPEG quantization tables between regions or discontinuities in the noise field that are representatives of the inpainting process which effectively turns the adversary's own attack into a new kind of fingerprint.
- **Threshold-Robust Watermarking:** The catastrophic failure of watermarks at certain thresholds indicates that their signals can be largely degraded. The next generation of watermarking research should focus on "threshold robustness." The goal should be to design a signal and detector pair whose performance, measured by ASR, is stable and low across a wide range of operating points. This might involve exploring signals that are more robust to degradation (e.g., using error-correcting codes) or developing detectors with more sophisticated decoding mechanisms that can function reliably even when the signal is weak.

References

[1] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon, "The stable signature: Rooting watermarks in latent diffusion models," *arXiv preprint arXiv:2303.15435*, 2023.

[2] Y. Wen, J. Kirchenbauer, J. Geiping, and T. Goldstein, "Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[3] U. Ojha, Y. Li, and Y. J. Lee, "Towards Universal Fake Image Detectors that Generalize Across Generative Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 480–24 489.

[4] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[5] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "DIRE for Diffusion-Generated Image Detection," *arXiv preprint arXiv:2303.09295*, 2023.

[6] J. Ricker, D. Lukovnikov, and A. Fischer, "AEROBLADE: Training-Free Detection of Latent Diffusion Images Using Autoencoder Reconstruction Error," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9130–9140.

[7] A. Pal, J. Kruk, M. Phute, M. Bhattaram, D. Yang, D. H. Chau, and J. Hoffman, "Semi-Truths: A Large-Scale Dataset of AI-Augmented Images for Evaluating Robustness of AI-Generated Image detectors," in *Thirty-eighth Conference on Neural Information Processing Systems*, 2024. [Online]. Available: https://openreview.net/forum?id=g3l4C45VnS

[8] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2149–2159.

[9] Y. Zeng, J. Fu, and H. Chao, "ZITS++: A Zero-Shot Framework for High-Fidelity Image Inpainting," *arXiv preprint arXiv:2203.00867*, 2022. [Online]. Available: https://arxiv.org/abs/2203.00867

[10] M. Schinas and S. Papadopoulos, "SIDBench: A Python Framework for Reliably Assessing Synthetic Image Detection Methods," *arXiv preprint arXiv:2311.08535*, 2023. [Online]. Available: https://arxiv.org/abs/2311.08535

APPENDIX A
CODE AND DATA AVAILABILITY

To ensure full reproducibility and to encourage further research by the community, all resources used in this thesis have been made publicly available.

- **Source Code Repository (GitHub):**
  https://github.com/eoguzakin/Robustness-of-AI-Generated-Image-Detection-Against-Localized-Inpainting-Attacks
- **Curated Datasets (Hugging Face):**
  https://huggingface.co/datasets/eoguzakin/Robustness-of-AI-Generated-Image-Detection-Against-Localized-Inpainting-Attacks

APPENDIX B
DETAILED EXPERIMENTAL RESULTS

TABLE I
BASELINE CALIBRATION RESULTS ACROSS DETECTORS

Detector	**Paradigm**	**Baseline AUC**	**Threshold t^***	**TPR/FPR @ t^***
UFD	Passive (Semantic)	0.9168	0.3481	TPR=0.25, FPR=0.01
AEROBLADE	Training-free (Autoencoder)	0.8160	0.1569	TPR=0.715, FPR=0.18
DIMD (Corvi et al.)	Passive (Artifacts)	0.6384	0.6512	TPR=0.38, FPR=0.16
DIRE	Training-free (Diffusion)	0.6588	-0.0149	TPR=0.37, FPR=0.10
Stable Signature	Watermarking (Latent)	0.7613 (vs reals)	$t_{90} = 0.5024$	TPR=0.90
Tree-Ring	Watermarking (Freq.)	0.8405 (vs reals)	$t_{\text{eer}} = -54.645$	EER=0.115

TABLE II
UFD ON 200 INPAINTED REAL IMAGES (LAMA, ZITS, SEMANTIC) AND 200 REAL IMAGES (SEMI-TRUTHS). ASR = INPAINTED REALS → REAL AT FIXED t^*.

Inp.	**Subset**	**N**	**Rob. AUC**	**ΔAUC**	**ASR (%)**	**95% CI**	**Mean**	**Median**
LaMa	randblob bin1 (0–3%)	400	0.8858	0.0310	63.0	[0.561, 0.694]	0.266	0.258
LaMa	randblob bin2 (3–10%)	400	0.9284	-0.0116	51.0	[0.441, 0.578]	0.372	0.345
LaMa	randblob bin3 (10–25%)	400	0.7663	0.1505	77.5	[0.712, 0.827]	0.097	0.114
LaMa	randblob bin4 (25–40%)	400	0.9720	-0.0551	23.0	[0.177, 0.293]	0.528	0.547
LaMa	randrect	400	0.5066	0.4102	98.5	[0.957, 0.995]	-0.191	-0.196
ZITS	randblob bin1 (0–3%)	400	0.8172	0.0996	76.0	[0.696, 0.814]	0.150	0.167
ZITS	randblob bin2 (3–10%)	400	0.8186	0.0982	80.5	[0.745, 0.854]	0.138	0.131
ZITS	randblob bin3 (10–25%)	400	0.1168	0.8001	99.5	[0.972, 0.999]	-0.625	-0.666
ZITS	randblob bin4 (25–40%)	400	0.7579	0.1590	87.0	[0.816, 0.910]	0.071	0.067
ZITS	randrect	400	0.3419	0.5749	99.5	[0.972, 0.999]	-0.354	-0.319
Semantic	semantic	400	0.9169	-0.0001	50.0	[0.431, 0.569]	0.336	0.348

TABLE III
UFD ON 200 INPAINTED FAKE IMAGES AND 200 FAKE IMAGES. ASR = FAKES → REAL AT FIXED t^*.

Inp.	**Subset**	**N**	**Rob. AUC**	**ΔAUC**	**ASR (%)**	**95% CI**	**Mean**	**Median**
LaMa	randblob bin1 (0–3%)	400	0.8991	0.0178	64.0	[0.571, 0.703]	0.2721	0.2603
LaMa	randblob bin2 (3–10%)	400	0.8863	0.0305	61.0	[0.551, 0.684]	0.2604	0.2687
LaMa	randblob bin3 (10–25%)	400	0.8154	0.1014	77.0	[0.712, 0.827]	0.1350	0.1216
LaMa	randblob bin4 (25–40%)	400	0.5923	0.3245	98.0	[0.943, 0.989]	-0.1077	-0.1044
LaMa	randrect	400	0.5887	0.3282	94.0	[0.917, 0.976]	-0.1073	-0.1196
ZITS	randblob bin1 (0–3%)	400	0.5503	0.3666	96.0	[0.943, 0.989]	-0.1516	-0.1436
ZITS	randblob bin2 (3–10%)	400	0.4828	0.4340	99.0	[0.964, 0.997]	-0.2071	-0.2108
ZITS	randblob bin3 (10–25%)	400	0.9583	-0.0415	45.0	[0.320, 0.454]	0.4292	0.4159
ZITS	randblob bin4 (25–40%)	400	0.2323	0.6845	100.0	[0.981, 1.000]	-0.4559	-0.4497
ZITS	randrect	400	0.7822	0.1347	87.0	[0.816, 0.910]	0.0884	0.1211

TABLE IV
AEROBLADE ON 200 INPAINTED REAL IMAGES AND 200 REAL IMAGES. ASR = INPAINTED REALS → REAL AT FIXED t^*.

Inp.	Subset	N	Rob. AUC	ΔAUC	ASR (%)	95% CI	Mean	Median
LaMa	randblob bin1 (0–3%)	400	0.5232	0.2925	98.5	[0.957, 0.995]	-0.154	-0.159
LaMa	randblob bin2 (3–10%)	400	0.5656	0.2501	81.0	[0.750, 0.858]	-0.121	-0.139
LaMa	randblob bin3 (10–25%)	400	0.6389	0.1768	76.0	[0.696, 0.814]	-0.034	-0.071
LaMa	randblob bin4 (25–40%)	400	0.7261	0.0896	79.0	[0.728, 0.841]	0.184	0.238
LaMa	randrect	400	0.6772	0.1385	76.0	[0.696, 0.814]	0.062	0.014
ZITS	randblob bin1 (0–3%)	400	0.5250	0.2907	96.0	[0.923, 0.980]	-0.153	-0.156
ZITS	randblob bin2 (3–10%)	400	0.5758	0.2399	78.5	[0.723, 0.836]	-0.122	-0.128
ZITS	randblob bin3 (10–25%)	400	0.6557	0.1600	76.5	[0.702, 0.818]	0.008	-0.039
ZITS	randblob bin4 (25–40%)	400	0.7642	0.0515	78.5	[0.723, 0.836]	0.399	0.352
ZITS	randrect	400	0.7523	0.0634	81.5	[0.755, 0.863]	0.418	0.333

TABLE V
AEROBLADE ON 200 INPAINTED FAKE IMAGES AND 200 FAKE IMAGES. ASR = FAKES → REAL AT FIXED t^*.

Inp.	Subset	N	Rob. AUC	ΔAUC	ASR (%)	95% CI	Mean	Median
LaMa	randblob bin1 (0–3%)	400	0.5099	0.3058	100.0	[0.974, 1.000]	0.001	-0.006
LaMa	randblob bin2 (3–10%)	400	0.5199	0.2958	98.6	[0.950, 0.996]	0.003	0.003
LaMa	randblob bin3 (10–25%)	400	0.5557	0.2600	76.9	[0.694, 0.831]	0.019	0.024
LaMa	randblob bin4 (25–40%)	400	0.6329	0.1828	55.9	[0.478, 0.638]	0.131	0.133
LaMa	randrect	400	0.6108	0.2049	61.5	[0.534, 0.691]	0.084	0.144
ZITS	randblob bin1 (0–3%)	400	0.5145	0.3012	100.0	[0.974, 1.000]	0.000	-0.001
ZITS	randblob bin2 (3–10%)	400	0.5273	0.2884	95.1	[0.902, 0.976]	0.004	-0.000
ZITS	randblob bin3 (10–25%)	400	0.5690	0.2467	72.0	[0.642, 0.787]	0.033	0.036
ZITS	randblob bin4 (25–40%)	400	0.6741	0.1416	51.7	[0.436, 0.598]	0.184	0.141
ZITS	randrect	400	0.6455	0.1702	53.8	[0.457, 0.618]	0.121	0.132

TABLE VI
DIMD ON 200 INPAINTED REAL IMAGES AND 200 REAL IMAGES. ASR = INPAINTED REALS → REAL AT FIXED t^*.

Inp.	Subset	N	Rob. AUC	ΔAUC	ASR (%)	95% CI	Mean	Median
LaMa	randblob bin1 (0–3%)	400	0.5016	0.1367	73.5	[0.209, 0.330]	0.5013	0.4819
LaMa	randblob bin2 (3–10%)	400	0.4992	0.1391	74.5	[0.200, 0.320]	0.5013	0.4825
LaMa	randblob bin3 (10–25%)	400	0.4855	0.1528	76.5	[0.182, 0.298]	0.5013	0.4844
LaMa	randblob bin4 (25–40%)	400	0.4808	0.1575	77.5	[0.173, 0.288]	0.5013	0.4856
LaMa	randrect	400	0.5188	0.1195	76.0	[0.186, 0.304]	0.5013	0.4803
ZITS	randblob bin1 (0–3%)	400	0.4977	0.1406	74.0	[0.204, 0.325]	0.5013	0.4794
ZITS	randblob bin2 (3–10%)	400	0.4995	0.1388	73.5	[0.209, 0.330]	0.5013	0.4813
ZITS	randblob bin3 (10–25%)	400	0.4901	0.1482	75.5	[0.191, 0.309]	0.5013	0.4834
ZITS	randblob bin4 (25–40%)	400	0.4863	0.1520	77.0	[0.177, 0.293]	0.5013	0.4809
ZITS	randrect	400	0.5094	0.1289	76.5	[0.182, 0.298]	0.5013	0.4828
Semantic	semantic	400	0.5434	0.0949	73.5	[0.209, 0.330]	0.5013	0.4868

TABLE VII
DIMD ON 200 INPAINTED FAKE IMAGES AND 200 FAKE IMAGES. ASR = FAKES → REAL AT FIXED t^*.

Inp.	Subset	N	Rob. AUC	ΔAUC	ASR (%)	95% CI	Mean	Median
LaMa	randblob bin1 (0–3%)	400	0.6374	0.0009	2.63	[0.007, 0.091]	0.5013	0.4628
LaMa	randblob bin2 (3–10%)	400	0.6375	0.0008	3.95	[0.014, 0.110]	0.5013	0.4600
LaMa	randblob bin3 (10–25%)	400	0.6342	0.0041	3.95	[0.014, 0.110]	0.5013	0.4619
LaMa	randblob bin4 (25–40%)	400	0.6290	0.0093	11.84	[0.064, 0.210]	0.5013	0.4644
LaMa	randrect	400	0.6244	0.0139	9.21	[0.045, 0.178]	0.5013	0.4584
ZITS	randblob bin1 (0–3%)	400	0.6361	0.0022	1.32	[0.002, 0.071]	0.5013	0.4609
ZITS	randblob bin2 (3–10%)	400	0.6360	0.0023	3.95	[0.014, 0.110]	0.5013	0.4619
ZITS	randblob bin3 (10–25%)	400	0.6337	0.0046	3.95	[0.014, 0.110]	0.5013	0.4625
ZITS	randblob bin4 (25–40%)	400	0.6328	0.0055	11.84	[0.064, 0.210]	0.5013	0.4663
ZITS	randrect	400	0.6208	0.0175	11.84	[0.064, 0.210]	0.5013	0.4666

TABLE VIII
DIRE ON 200 INPAINTED REAL IMAGES AND 200 REAL IMAGES. ASR = INPAINTED REALS → REAL AT FIXED t^*.

Inp.	Subset	N	Rob. AUC	ΔAUC	ASR (%)	95% CI	Mean	Median
LaMa	randblob bin1 (0–3%)	400	0.5021	0.1567	88.5	[0.833, 0.922]	-0.0289	-0.0257
LaMa	randblob bin2 (3–10%)	400	0.5038	0.1551	88.5	[0.833, 0.922]	-0.0288	-0.0259
LaMa	randblob bin3 (10–25%)	400	0.5037	0.1552	86.5	[0.811, 0.906]	-0.0288	-0.0258
LaMa	randblob bin4 (25–40%)	400	0.5082	0.1506	86.5	[0.811, 0.906]	-0.0287	-0.0252
LaMa	randrect	400	0.5283	0.1305	0.0	[0.000, 0.019]	0.0302	0.0270
ZITS	randblob bin1 (0–3%)	400	0.5030	0.1558	89.5	[0.845, 0.930]	-0.0288	-0.0258
ZITS	randblob bin2 (3–10%)	400	0.5045	0.1543	88.0	[0.828, 0.918]	-0.0287	-0.0256
ZITS	randblob bin3 (10–25%)	400	0.5112	0.1476	86.5	[0.811, 0.906]	-0.0285	-0.0253
ZITS	randblob bin4 (25–40%)	400	0.5249	0.1339	85.5	[0.800, 0.897]	-0.0280	-0.0248
ZITS	randrect	400	0.5014	0.1574	86.5	[0.811, 0.906]	-0.0290	-0.0263
Semantic	semantic	400	0.7212	-0.0624	62.5	[0.556, 0.689]	-0.0243	-0.0211

TABLE IX
DIRE ON 200 INPAINTED FAKE IMAGES AND 200 FAKE IMAGES. ASR = FAKES → REAL AT FIXED t^*.

Inp.	Subset	N	Rob. AUC	ΔAUC	ASR (%)	95% CI	Mean	Median
LaMa	randblob bin1 (0–3%)	400	0.6549	0.0040	2.70	[0.007, 0.093]	-0.0255	-0.0227
LaMa	randblob bin2 (3–10%)	400	0.6475	0.0113	4.05	[0.014, 0.113]	-0.0256	-0.0227
LaMa	randblob bin3 (10–25%)	400	0.6217	0.0372	14.86	[0.085, 0.247]	-0.0261	-0.0239
LaMa	randblob bin4 (25–40%)	400	0.5927	0.0661	22.97	[0.149, 0.337]	-0.0268	-0.0243
ZITS	randblob bin1 (0–3%)	400	0.6560	0.0028	2.70	[0.007, 0.093]	-0.0254	-0.0227
ZITS	randblob bin2 (3–10%)	400	0.6499	0.0089	4.05	[0.014, 0.113]	-0.0256	-0.0227
ZITS	randblob bin3 (10–25%)	400	0.6313	0.0275	13.51	[0.075, 0.231]	-0.0259	-0.0238
ZITS	randblob bin4 (25–40%)	400	0.6035	0.0553	20.27	[0.127, 0.308]	-0.0266	-0.0242

TABLE X

STABLE SIGNATURE ROBUSTNESS TO LAMA AND ZITS INPAINTING. ASR REPORTED WITH 95% CONFIDENCE INTERVALS AT THREE THRESHOLDS: t_{90}, t_{99}, AND t_{FPR1} FOR 200 CLEAN IMAGES AND 200 INPAINTED CLEAN IMAGES.

Attack	Split	AUC(clean→attacked)	ASR@t90 [95% CI]	ASR@t99 [95% CI]	ASR@tFPR1 [95% CI]	n	t90	t99	tFPR1
baseline	reals_vs_cleanwm	0.761	-	-	-	400	0.502	0.501	0.507
LaMa	bin1_0-3	0.502	0.110 [0.074,0.161]	0.010 [0.003,0.036]	0.815 [0.755,0.863]	400	0.502	0.501	0.507
	bin2_3-10	0.503	0.115 [0.078,0.167]	0.005 [0.001,0.028]	0.810 [0.750,0.858]	400	0.502	0.501	0.507
	bin3_10-25	0.511	0.105 [0.070,0.155]	0.005 [0.001,0.028]	0.835 [0.777,0.880]	400	0.502	0.501	0.507
	bin4_25-40	0.526	0.085 [0.054,0.132]	0.005 [0.001,0.028]	0.850 [0.794,0.893]	400	0.502	0.501	0.507
	randrect	0.515	0.120 [0.082,0.172]	0.005 [0.001,0.028]	0.825 [0.766,0.871]	400	0.502	0.501	0.507
ZITS	bin1_0-3	0.505	0.105 [0.070,0.155]	0.010 [0.003,0.036]	0.805 [0.745,0.854]	400	0.502	0.501	0.507
	bin2_3-10	0.511	0.115 [0.078,0.167]	0.005 [0.001,0.028]	0.805 [0.745,0.854]	400	0.502	0.501	0.507
	bin3_10-25	0.539	0.120 [0.082,0.172]	0.010 [0.003,0.036]	0.840 [0.783,0.884]	400	0.502	0.501	0.507
	bin4_25-40	0.569	0.110 [0.074,0.161]	0.010 [0.003,0.036]	0.875 [0.822,0.914]	400	0.502	0.501	0.507
	randrect	0.541	0.140 [0.099,0.195]	0.005 [0.001,0.028]	0.835 [0.777,0.880]	400	0.502	0.501	0.507

TABLE XI

TREE-RING ROBUSTNESS TO LAMA AND ZITS INPAINTING. ASR REPORTED WITH 95% CONFIDENCE INTERVALS AT THREE THRESHOLDS: t_{90}, t_{99}, AND t_{eer} FOR 200 CLEAN IMAGES AND 200 INPAINTED CLEAN IMAGES.

Attack	Split	AUC(clean→attacked)	ASR@t90 [95% CI]	ASR@t99 [95% CI]	ASR@t_eer [95% CI]	n	t90	t99	t_eer
baseline	reals_vs_cleanwm	0.8405	-	-	-	400	-55.221	-55.656	-52.696
LaMa	bin1_0-3	0.5092	0.100 [0.055,0.174]	0.010 [0.002,0.054]	0.990 [0.946,0.998]	400	-55.221	-55.656	-52.696
	bin2_3-10	0.5284	0.110 [0.063,0.186]	0.010 [0.002,0.054]	0.990 [0.946,0.998]	400	-55.221	-55.656	-52.696
	bin3_10-25	0.5602	0.100 [0.055,0.174]	0.000 [0.000,0.037]	1.000 [0.963,1.000]	400	-55.221	-55.656	-52.696
	bin4_25-40	0.6073	0.060 [0.028,0.125]	0.010 [0.002,0.054]	1.000 [0.963,1.000]	400	-55.221	-55.656	-52.696
	randrect	0.5925	0.100 [0.055,0.174]	0.010 [0.002,0.054]	1.000 [0.963,1.000]	400	-55.221	-55.656	-52.696
ZITS	bin1_0-3	0.5086	0.100 [0.055,0.174]	0.010 [0.002,0.054]	0.990 [0.946,0.998]	400	-55.221	-55.656	-52.696
	bin2_3-10	0.5250	0.100 [0.055,0.174]	0.010 [0.002,0.054]	0.990 [0.946,0.998]	400	-55.221	-55.656	-52.696
	bin3_10-25	0.5341	0.060 [0.028,0.125]	0.000 [0.000,0.037]	1.000 [0.963,1.000]	400	-55.221	-55.656	-52.696
	bin4_25-40	0.5594	0.050 [0.022,0.112]	0.000 [0.000,0.037]	1.000 [0.963,1.000]	400	-55.221	-55.656	-52.696
	randrect	0.5452	0.040 [0.016,0.098]	0.000 [0.000,0.037]	1.000 [0.963,1.000]	400	-55.221	-55.656	-52.696

INPAINTING QUALITY (WATERMARKED SETS)

TABLE XII

QUANTITATIVE INPAINTING QUALITY ON 200 STABLE SIGNATURE AND 200 TREE-RING WATERMARKED IMAGE SETS (MASK REGION METRICS).

Dataset	Inp.	Subset	N	LPIPS Mask (μ | med)	SSIM Mask (μ)	PSNR Mask (μ)
stablesig	LaMa	randrect	200	0.448 — 0.458	0.4283	15.94
stablesig	LaMa	randblob bin1 (0–3%)	200	0.135 — 0.077	0.7899	27.25
stablesig	LaMa	randblob bin2 (3–10%)	200	0.053 — 0.046	0.9354	28.26
stablesig	LaMa	randblob bin3 (10–25%)	200	0.083 — 0.082	0.9004	25.39
stablesig	LaMa	randblob bin4 (25–40%)	200	0.135 — 0.135	0.8371	22.73
stablesig	ZITS	randrect	200	0.443 — 0.449	0.4267	15.71
stablesig	ZITS	randblob bin1 (0–3%)	200	0.134 — 0.078	0.7927	27.18
stablesig	ZITS	randblob bin2 (3–10%)	200	0.053 — 0.046	0.9355	28.16
stablesig	ZITS	randblob bin3 (10–25%)	200	0.083 — 0.082	0.9011	25.32
stablesig	ZITS	randblob bin4 (25–40%)	200	0.135 — 0.135	0.8377	22.62
treering	LaMa	randrect	200	0.542 — 0.556	0.3187	15.56
treering	LaMa	randblob bin1 (0–3%)	200	0.029 — 0.016	0.9608	35.19
treering	LaMa	randblob bin2 (3–10%)	200	0.092 — 0.064	0.8814	26.70
treering	LaMa	randblob bin3 (10–25%)	200	0.187 — 0.158	0.7683	22.08
treering	LaMa	randblob bin4 (25–40%)	200	0.293 — 0.268	0.6579	18.98
treering	ZITS	randrect	200	0.531 — 0.540	0.3180	15.52
treering	ZITS	randblob bin1 (0–3%)	200	0.028 — 0.016	0.9613	35.37
treering	ZITS	randblob bin2 (3–10%)	200	0.089 — 0.062	0.8818	26.75
treering	ZITS	randblob bin3 (10–25%)	200	0.183 — 0.153	0.7687	22.09
treering	ZITS	randblob bin4 (25–40%)	200	0.288 — 0.262	0.6588	18.97