Removing Watermarks with Partial Regeneration using Semantic Information

Krti Tallam[1], John Kevin Cava[2], Caleb Geniesse[3], N. Benjamin Erichson[1,3], and Michael W. Mahoney[1,3,4]

[1]International Computer Science Institute, Berkeley, CA, USA
[2]School of Computing and Augmented Intelligence, Arizona State University, AZ, USA
[3]Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[4]Department of Statistics, University of California at Berkeley, Berkeley, CA, USA

Abstract

As AI-generated imagery becomes ubiquitous, invisible watermarks have emerged as a primary line of defense for copyright and provenance. The newest watermarking schemes embed *semantic* signals - content-aware patterns that are designed to survive common image manipulations - yet their true robustness against adaptive adversaries remains under-explored. We expose a previously unreported vulnerability and introduce `SemanticRegen`, a three-stage, label-free attack that erases state-of-the-art semantic and invisible watermarks while leaving an image's apparent meaning intact. Our pipeline (i) uses a vision-language model to obtain fine-grained captions, (ii) extracts foreground masks with zero-shot segmentation, and (iii) inpaints only the background via an LLM-guided diffusion model, thereby preserving salient objects and style cues. Evaluated on >1,000 prompts across four watermarking systems - TreeRing, StegaStamp, StableSig, and DWT/DCT - `SemanticRegen` is the *only* method to defeat the semantic TreeRing watermark ($p = 0.10 > 0.05$) and reduces bit-accuracy below 0.75 for the remaining schemes, all while maintaining high perceptual quality (masked SSIM = 0.94 ± 0.01). We further introduce *masked SSIM* (mSSIM) to quantify fidelity within foreground regions, showing that our attack achieves up to 12 percent higher mSSIM than prior diffusion-based attackers. These results highlight an urgent gap between current watermark defenses and the capabilities of adaptive, semantics-aware adversaries, underscoring the need for watermarking algorithms that are resilient to content-preserving regenerative attacks.

1 Introduction

The growing advancement and widespread adoption of AI-generated content has brought about urgent challenges in protecting copyright and intellectual property, particularly in fields such as science, healthcare, and entertainment [13, 41]. As the reliance on these generated images grows, so does the need for robust methods to ensure the integrity and ownership of digital content [23, 9]. Watermarking embeds markers in images to verify ownership and prevent misuse [22, 45, 42, 18, 2, 46]. Traditional watermarking techniques

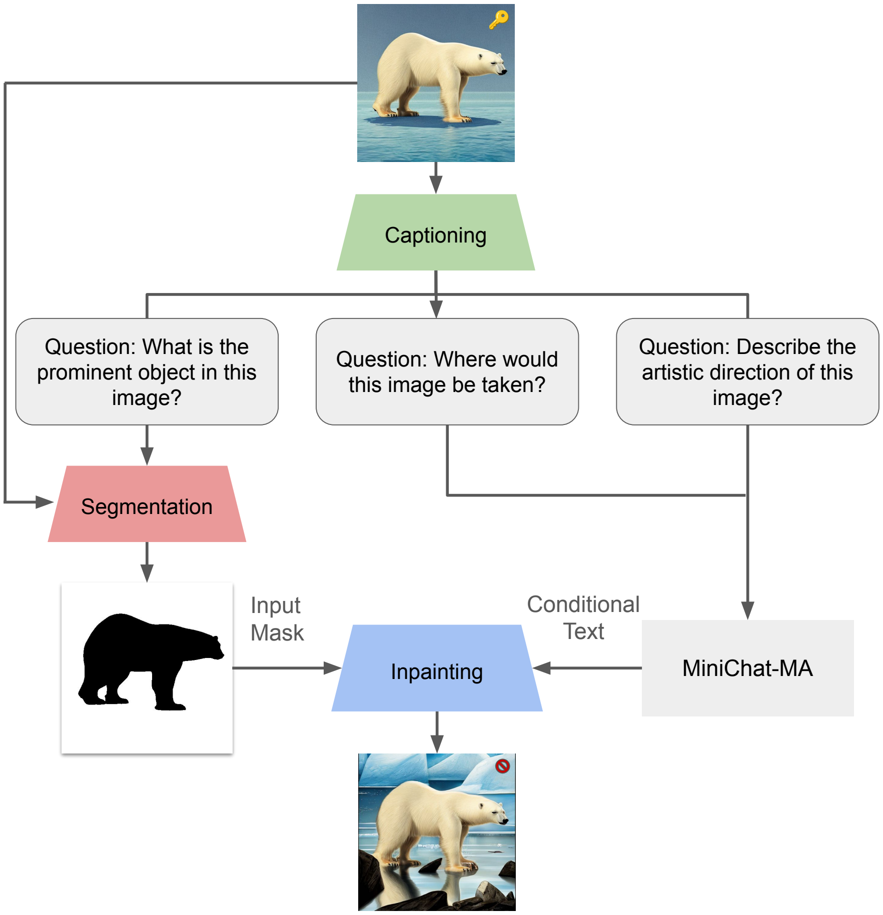
Captioning
Question: What is the prominent object in this image?
Question: Where would this image be taken?
Question: Describe the artistic direction of this image?
Segmentation
Input Mask
Conditional Text
Inpainting
MiniChat-MA

Figure 1: Our semantic watermark removal pipeline involves three primary components: (1) Captioning (green), (2) Segmentation (red), and (3) Inpainting (blue). For captioning, we use a VQA model to provide essential context for subsequent processing. For segmentation, we focus on prominent objects or areas of interest within the image. For inpainting, the background of the image is replaced with semantically similar content, effectively removing the watermark while preserving image integrity. To construct the prompt for conditional text inpainting, we use MiniChat-MA, an LLM that refines answers generated from the image captioning model. This pipeline extracts semantic information and replaces the background for watermark removal, while preserving the foreground content.

have been instrumental in the embedding of markers to verify ownership and prevent misuse [40, 27]. However, the rise of sophisticated adversarial attacks that subtly alter images to evade detection has exposed vulnerabilities in these systems [39, 31, 37, 1]. This growing threat highlights the critical need for more advanced and resilient approaches to safeguarding digital assets [36, 14], ensuring that the rights of content creators are upheld in the face of evolving technological challenges.

To address these challenges, researchers have developed techniques to embed markers in generated images to verify ownership and prevent unauthorized use [12, 24, 43]. One such method involves using variable autoencoders (VAE) or diffusion models to inject watermarks into the latent space by encoding neural networks or adding noise, as demonstrated in the WAVES benchmark [4]. Zhao et al. proposed a watermark attack method, highlighting the need to strengthen watermarking strategies [49]. Adversaries have created advanced attacks to bypass detection by modifying subtly watermarked images, leading to the development of more robust detection systems [11, 16, 39, 49, 4]. Advances in AI security have led to innovative strategies to protect generated images from adversarial manipulation [28]. Classifier-free methods for the detection and removal of watermarks offer alternatives to traditional approaches by analyzing the inherent properties of the image rather than relying on predefined classifiers [50]. Techniques such as pixel-level analysis, frequency domain analysis, and structural analysis identify anomalies introduced by watermarks [50, 34].

In this work, we propose `SemanticRegen`, a framework for removing semantic watermarks; see Figure 1 for an illustration of our basic approach. Our approach demonstrates the effectiveness of semantic repainting for watermark removal, exposing vulnerabilities that can inform the development of more resilient techniques. Our approach involves a three-step pipeline: (1) a Visual Question Answering (VQA) model through BLIP2; (2) a segmentation model using LangSAM; and (3) Stable Diffusion Inpainting [30, 32, 35] (see Figure 1). The VQA model analyzes the target image, providing semantic information. Our method preserves the semantic background information instead of substituting it with random/arbitrary content. After generating prompts, we apply inverted masks from the segmentation model to condition the target image, creating a new image that retains salient objects while replacing the background with semantically similar content from the original image. We use a comprehensive approach with the VQA model for content extraction, employing customized questions to capture diverse aspects of complex scenes. Conditioning the model with prompts improves its ability to discern details, helping to remove strong watermarks while preserving the integrity of the image [50]. Unlike prior methods that rely solely on adversarial or generative transformations, our approach integrates semantic understanding via VQA-driven segmentation to improve targeted watermark removal while preserving content integrity.

Our approach is inspired by research on the use of large language models (LLMs) for synthetic data set generation and image diffusion models for robust training [21, 25]. Recent advances in diffusion-based watermarking by Zhang et al. [47] and Kawar et al. [25] show promise in embedding watermarks into images, while preserving visual fidelity. The WAVES benchmark [4] provides insights into their performance. Diffusion-based approaches prioritize image fidelity through controlled noise application, making them effective for watermark removal and preferable to GAN-based methods due to their stability, robustness, and ease of implementation. These techniques offer a compelling solution for various applications. We evaluated `SemanticRegen`against these and related methods, in particular against TreeRing watermarker [45], StegaStamp [42], StableSig [22] and invisible watermarkers [49]. TreeRing is ideal due to its imperceptibility and resilience to common manipulations such as cropping, resizing, and compression [45].

Our evaluation in various watermarking techniques demonstrates the effectiveness of `SemanticRegen`, with minimal distortion and high image quality, as reflected in low Mean Squared Error (MSE) and high Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR) scores. Competing well against state-of-the-art techniques such as DiffWMAttacker, VAEWMAttacker and Rinse4x, `SemanticRegen`fills removed portions with semantically similar backgrounds, achieving meaningful results. Optimal performance is achieved with clear, separable backgrounds similar to those in Stable Diffusion training data. SemanticRegen excels in watermark removal when the central meaning of the image is localized to only a few main objects, since the method relies on segmenting these objects and removing the less important background content.

In summary, our main contributions are as follows.

- We introduce SemanticRegen, an effective watermark removal method, evaluate it on the TreeRing watermarker [45], and compare it to StegaStamp [42], StableSig [22], and invisible watermarks of deep neural networks [49]. SemanticRegen successfully removes all four watermarks tested, as demonstrated by our analyzes ($p > 0.05$ and $BitAccuracy < 0.75$).
- We compare SemanticRegenwith other watermark attackers and observe that SemanticRegenis the only attacker to eliminate the semantic Tree-Ring watermark. SemanticRegen has an average p-value of 0.1, while all other watermark attackers failed to meet the threshold for successful removal ($p > 0.05$).
- We introduce a metric to evaluate objects in the foreground: the masked structural similarity index (mSSIM). SemanticRegen significantly outperforms current baseline methods in different watermarking methods, including invisible and semantic watermarks. Our method preserves image quality with the most success within salient regions of generated images, as evidenced by mSSIM scores of 0.94, compared to Image Distortion ($mSSIM = 0.85$) and Rinse4x ($mSSIM = 0.86$).
- We demonstrate how SemanticRegen leverages a multi-step pipeline to expose and exploit vulnerabilities in current watermarking techniques. Using a VQA model for context generation, segmenting key areas of the image and replacing the background with semantically similar content through LLM-guided inpainting, SemanticRegenextracts and reconstructs portions of the watermarked image. This process reveals latent patterns that these watermarking techniques fail to protect, offering insights into potential attack vectors and relevant underlying assumptions that could be exploited.[1]

2 Related Work

Watermarking and its adversarial counterpart, watermark removal, are pivotal areas of research in protecting intellectual property and ensuring the integrity of AI-generated content. This section discusses advances in watermarking methods, challenges posed by adversarial attacks, and the broader implications for AI-generated media.

Watermarking Methods. Watermarking has evolved from traditional techniques, such as frequency domain embedding, to state-of-the-art methods that take advantage of generative models for imperceptible yet robust integration. Early approaches embedded watermarks in spatial or frequency domains using the discrete wavelet transform (DWT) or the discrete cosine transform (DCT) [18, 2]. Although effective for basic transformations, these methods were vulnerable to adversarial attacks that exploited predictable patterns.

[1] https://github.com/KrtiT/semanticRegen

Modern watermarking techniques, such as TreeRing [45] and StegaStamp [42], have introduced imperceptible and resilient watermarks. TreeRing watermarks use adaptive encoding mechanisms to maintain the integrity of the watermark against manipulations such as resizing, cropping, and compression. Similarly, StegaStamp employs neural networks to embed and extract watermarks with high fidelity, enabling robust ownership verification.

Recent innovations leverage diffusion-based models to embed watermarks in image data while preserving visual fidelity [47, 25]. These approaches integrate watermarks directly into the latent spaces of generative models, ensuring resilience against adversarial manipulations. In particular, [20] explores how imperceptible signatures can be embedded in high-resolution generative content, paving the way for secure watermarking in multimodal AI systems.

Efforts such as the WAVES benchmark [4] have standardized the evaluation of watermarking techniques. WAVES provides a baseline framework to assess robustness across various attacks, offering insights into strengths and limitations. Such benchmarks are instrumental in the development of next-generation watermarking systems.

Watermark Removal Methods. The increasing sophistication of adversarial techniques has highlighted vulnerabilities in watermarking systems. Watermark removal methods exploit the inherent structure of embedded watermarks to obscure, distort, or eliminate them. Early approaches relied on pixel-level transformations, but recent advances employ machine learning techniques to target latent representations of watermarked content.

Regenerative attacks, such as those using Variational Autoencoders (VAEs) [11, 16] and diffusion models [49], have proven effective in bypassing watermarking schemes. These methods iteratively refine watermarked images, reconstructing their features while removing embedded signals. In particular, Zhao et al. [49] demonstrate how diffusion-based methods can obscure watermarks while maintaining image fidelity, highlighting the need for continual innovation in watermarking strategies.

Hybrid approaches have also emerged that combine adversarial purification with iterative refinement techniques such as "rinsing" [4]. These methods sequentially reduce watermark detectability by applying regenerative transformations. For example, hybrid methods leverage both semantic understanding and low-level noise removal to effectively erase watermarks without compromising image quality [35, 19]. Furthermore, [20] explores adversarial frameworks specifically designed to manipulate the robustness of the watermark, while [29] introduces adaptive techniques to counter hybrid watermarking schemes.

Despite these advances, challenges persist. Many removal methods require access to training data or model architecture, limiting their applicability in real-world scenarios. In addition, adversarial techniques often introduce artifacts or reduce image quality, necessitating further research to balance robustness and fidelity.

Ethical and practical implications. The interplay between watermarking and adversarial removal highlights broader implications for intellectual property protection in the age of generative AI. As models like Stable Diffusion and DALLE-2 become widely accessible, the need for robust watermarking systems grows [13]. However, the rapid evolution of adversarial attacks underscores the limitations of existing approaches, creating an ongoing arms race between content creators and adversaries.

Ethical considerations are central to this discourse. Watermarking systems must navigate complex questions of fair use, attribution, and copyright enforcement. For example, the removal of watermarks from publicly shared content raises concerns about the misuse of AI for unauthorized content generation [29]. Similarly, the ability to embed imperceptible watermarks in training datasets raises questions about consent and transparency [14].

The WAVES benchmark [4] and recent studies such as [20] and [29] emphasize the importance of interdisciplinary collaboration in addressing these challenges. Legal frameworks, technical innovations, and policy guidelines must converge to create robust systems that balance creative freedom with content security.

Limitations and open challenges. While modern watermarking systems have advanced significantly, they remain vulnerable to adaptive adversarial techniques. Diffusion-based watermarking, for example, struggles with attacks that exploit shared latent spaces in generative models [47]. Similarly, hybrid removal methods, while effective, often require extensive computational resources, limiting their scalability.

Future research should focus on developing adaptive watermarking techniques that can dynamically respond to adversarial threats. In addition, comprehensive evaluation frameworks are needed to assess watermarking methods under real-world conditions, including domain changes, mixed media, and collaborative workflows.

Broader Context. The field of watermarking and watermark removal is at the forefront of intellectual property protection in the digital age. As generative AI models continue to evolve, so does the complexity of securing AI-generated content. The interplay between watermarking and adversarial techniques presents an ongoing challenge, leading to an escalating arms race between embedding robust watermarks and developing adversarial methods for their removal. Addressing this issue requires interdisciplinary collaboration across computer vision, cryptography, and AI ethics to develop standardized benchmarks, evaluation protocols, and legal frameworks to safeguard digital media.

Recent research has highlighted the need for comprehensive benchmarking tools to assess the effectiveness and resilience of different watermarking techniques. In particular, the WAVES benchmark [4] systematically evaluates watermarking methods in multiple adversarial attack scenarios, providing valuable information on the strengths and weaknesses of existing techniques. Furthermore, [20] introduces advanced watermarking strategies that integrate deep learning-based feature embeddings, improving robustness against known attack vectors. On the removal front, emerging studies such as [29] explore the use of generative adversarial networks (GANs) and diffusion-based models to counter imperceptible watermarking strategies. These findings underscore the need for continuous evaluation and adaptation of both watermarking and removal strategies to prevent misuse while maintaining the integrity of digital content.

Building on these advancements, our work proposes a novel approach to semantic watermark removal that addresses critical gaps in existing methods. Using insights from state-of-the-art watermarking and removal techniques, our aim is to contribute to the broader effort to develop secure, transparent, and resilient digital content protection mechanisms.

Ultimately, our approach emphasizes the importance of balancing technological innovation with ethical considerations to ensure that watermarking methods remain effective in preserving copyright and intellectual property rights.

3 Methods

In this section, we describe `SemanticRegen`, our semantic watermark removal pipeline. As depicted in Figure 1, the pipeline comprises three main components: (1) the VQA captioning model; (2) the segmentation model; and (3) the inpainting model. Our automated pipeline involves an LLM segmentation and inpainting semantic attack. Beginning with a watermarked image, the process uses a captioning model (BLIP2), conditioned with specific prompts: (a) identifying prominent objects; (b) determining the background; and (c) and defining the artistic direction. Artistic direction is defined as the visual style that is used in the image, e.g., photographic, cartoon, impressionism, etc. The first prompt is used to segment the image based on the most salient / prominent object. The segmented object(s) then serves as input to the repainting. (Since we are taking a subset of pixels due to the segmentation, it is considered repainting on a subset of the image, that is, inpainting.) This approach aims to effectively remove the watermark from the image. In Figure 1, we illustrate the models used to extract semantic information from the image and that serve as a conditional input for stable diffusion, thus replacing the background of the target image. When discussing watermark removal, it is often essential to measure how well an attack maintains the important parts of an image—like the main subject or foreground—while potentially destroying or altering parts of the background. Standard image-quality metrics, such as the Structural Similarity Index Measure (SSIM), compute overall similarity between two images but do not specifically distinguish which parts of the image truly matter for human perception or for watermark embedding. In watermark attacks—particularly those that use "masks" to remove or distort certain regions—an attacker might intentionally ruin non-salient parts of the image (like backgrounds or less noticeable edges) to get rid of embedded watermarks. In doing so, the attacker may preserve the key objects or "foreground" that define the meaning of the image. If we only look at a global SSIM across the entire image, it might seem that the image is heavily altered. But if we focus on the most important regions (foreground objects), they might still look exactly the same. **Masked SSIM (mSSIM)** is introduced to better evaluate how much of the important (foreground) content remains unchanged after an attack that uses masking on non-salient regions.

3.1 VQA Captioning

Visual Question Answering (VQA) is a task at the intersection of computer vision and natural language processing that enables machines to answer textual queries about an image. This requires models to extract visual features and generate semantically meaningful responses based on the content of an image [8, 6, 30, 3].

Early VQA models relied on convolutional neural networks (CNNs) to extract image features, combined with recurrent neural networks (RNNs) for text processing. However, recent advances leverage transformer-based architectures, which enable deeper multimodal understanding. BLIP2 [30], for example, uses vision language pre-training on large-scale datasets, significantly improving accuracy on complex reasoning tasks over previous approaches.

Structured Prompting for Semantic Understanding. Our method builds on recent advances in question-driven image captioning [10], where targeted question prompts help to focus the model on extracting semantically relevant features. Instead of using generic captions, we design structured prompts to guide BLIP2 toward key information that is critical to our pipeline.

- **Q1: What is the prominent object in this image?** Helps to identify the *foreground elements* necessary for segmentation.
- **Q2: What is the background?**
  Defines the *context and scene composition* for inpainting.
- **Q3: What is the artistic direction of the image?** Captures *style, texture, and color tone*, which aids in reconstruction.

VQA-Guided Watermark Removal. By applying structured VQA, we ensure that the segmentation and inpainting models receive high-quality semantic information, improving the effectiveness of watermark removal. Previous work has shown that custom captions increase the accuracy of VQA by focusing on relevant contextual elements [10], which aligns with our approach of directing the model to extract detailed attributes from the scene.

Compared to traditional captioning, our structured approach enables:

- **Improved segmentation performance**, as the separation of the foreground and the background is explicitly guided.
- **Higher fidelity inpainting**, where the masked regions are filled with semantically relevant textures instead of arbitrary pixels.
- **Greater resilience against adversarial perturbations**, since captioning adapts to image modifications.

Implementation Details. For all experiments, we use *BLIP2* as the base VQA model, which has been shown to outperform previous models on multimodal benchmarks [30, 3]. We prompt the model using zero-shot inference, ensuring that no dataset-specific fine-tuning is required. The captions extracted are then summarized using an LLM-based rewriter (MiniChat-MA) to generate concise, high-quality inpainting prompts.

This VQA-guided strategy is essential in our SemanticRegeneration pipeline by ensuring that watermark removal is context sensitive, semantically grounded, and visually coherent.

Key Assumptions. Our approach is based on several fundamental assumptions that ensure the effectiveness of our watermark removal framework.

- **Foreground-Background Distinction:** The target image contains a distinguishable foreground object that is visually separable from the background.
- **Accurate Captioning:** The captioning model (BLIP2) can provide descriptive and reliable textual summaries of both the main object and its surroundings.
- **Precise Segmentation:** The segmentation model (LangSAM) is capable of accurately isolating the foreground object from the background with minimal errors.

- **Semantically Coherent Inpainting:** The inpainting model (Stable Diffusion) can reconstruct the background in a semantically meaningful way while preserving the integrity of the foreground object.

These assumptions ensure that our method operates under typical conditions. However, in cases where segmentation fails or the inpainting model introduces artifacts, manual refinement or additional post-processing may be required to achieve optimal results.

3.2 Segmentation Model

Image segmentation is a fundamental task in computer vision that involves partitioning an image into distinct regions based on object boundaries [26, 15]. The goal of segmentation is to delineate different objects or areas of interest within an image, allowing downstream tasks such as object detection, image synthesis, and scene understanding. Traditional segmentation techniques relied on hand-crafted features and clustering methods, such as thresholding, edge detection, and watershed algorithms. However, modern deep learning-based approaches leverage convolutional neural networks (CNNs) and transformer-based architectures trained on large-scale datasets to achieve state-of-the-art performance in complex image segmentation tasks.

One of the recent breakthroughs in segmentation models is Meta's Segment Anything Model (SAM), which introduced a foundation model approach to segmentation [26]. SAM is designed to generalize across diverse image types without requiring additional fine-tuning, making it effective for a wide range of real-world applications. LangSAM, an open source adaptation of SAM, retains its zero-shot segmentation capability, allowing it to process images and generate segmentation masks based on text or point-based queries. Using LangSAM, we ensure that our approach remains flexible and generalizes well across different types of images, reducing dependence on domain-specific segmentation models.

Integration with Visual Question Answering (VQA). In our pipeline, we use the first question (Q1) from the VQA captioning model's output, which asks: *"What is the prominent object in this image?"* This structured approach ensures that the most salient entity within the image is correctly identified before proceeding with segmentation. We then use this response as a prompt input to LangSAM [32], an open source implementation of Segment Anything [26], to extract the most important objects in the scene. The BLIP2-generated caption describing the primary object serves as input text for LangSAM, which then returns the segmentation masks of the detected objects. This allows us to segment prominent objects based on high-level semantics instead of relying on pixel-based heuristics.

Mask Thresholding for Effective Watermark Removal. To ensure effective removal of watermarks, we control the proportion of the image covered by the segmentation masks. This is particularly important in cases where:

- The VQA model identifies multiple prominent objects in the image.
- A single object appears multiple times, leading to excessive masking.

To handle these scenarios, we implement a threshold-based iterative strategy, where we dynamically add to the mask until either:

1. All prominent object masks are included.
2. The total mask size exceeds the predefined threshold.

This approach ensures sufficient coverage for watermark removal while preventing over-masking, which could distort important visual features.

Edge Cases and Refinements. The effectiveness of background painting depends heavily on BLIP2 captioning and LangSAM segmentation models, as they guide the reconstruction of watermarked areas. To improve robustness, we address the following cases:

- **Segmentation failure:** If LangSAM does not produce a clear background mask, we rely on an artistic direction prompt to guide the inpainting.
- **Excessive masking:** If the existing mask exceeds the defined threshold, we adjust our inpainting strategy to preserve the original pixels while ensuring effective removal of watermarks. In this case, the prominent objects themselves may be designated as the background mask while retaining the rest of the image structure.

Assumptions. Our segmentation model operates under the following key assumptions:

- **Foreground Object Identifiability:** The primary object of interest is visually distinct and can be effectively identified using natural language prompts.
- **Background Reconstruction Feasibility:** The background can be reconstructed meaningfully without distorting the visual integrity of the original object.
- **Segmentation Accuracy:** The generated mask is precise enough to avoid occluding important details while ensuring effective background replacement.

Impact on Inpainting. The segmentation mask obtained from LangSAM is inverted and passed to the inpainting model, ensuring that the salient object remains unchanged, while the background is regenerated to remove any traces of embedded watermarks. Empirical validation comparing random masks vs. semantic-based VQA masks reinforces our approach, demonstrating that semantic segmentation significantly improves watermark removal while maintaining high visual fidelity.

3.3 Summarization and Repainting Model

After the VQA captioning and segmentation of the masks in the image, we use an LLM (MiniChat-MA) [48], which is based on LLAMA2 [44], to summarize the answers given from the VQA captioning model. This is used as an input prompt to the inpainting model, which is a Stable Diffusion Inpainting model [38]. We use Stable Diffusion-v2 with 50 inference steps. The summarization prompt used for MiniChat-MA is as follows:

- Prompt = “Given the following sentences that describe an image, write in one sentence what the background setting is and in what art style.” + [Summary of Captions from MiniChat-MA].

Following prompt generation, we condition the target image with the inverted masks obtained from the segmentation model, leading to the generation of a new image. This image preserves the prominent objects from the target image, while replacing the surrounding background with semantically similar content sourced from the original image.

3.4 Masked Structural Similarity Index (mSSIM)

We introduce a new metric to evaluate objects in the foreground: the masked Structural Similarity Index (mSSIM). See Equation 1 below. mSSIM measures the similarity between watermarked images before and after each attack. For each prompt, we compute an image segmentation mask, **M**, which delineates the background from the foreground. We computed the mask for each prompt once and reused this mask to compare images before and after the attack. To compute mSSIM, we take an image before and after an attack, apply the background mask so that only foreground objects remain, and then compute the SSIM between the masked images. The metric can be described as:

$$\text{mSSIM} = \text{SSIM}(\mathbf{M} * \mathbf{X}_{\mathbf{img}}, \mathbf{M} * \mathbf{X}_{\mathbf{attacked}}), \tag{1}$$

where $\mathbf{M}$ is a binary segmentation mask and $\mathbf{X}_{\mathbf{img}}$ and $\mathbf{X}_{\mathbf{attacked}}$ are the real-valued images before and after attack. The dimensions of the mask and both images are the same, where $\mathbf{M} \in \{0,1\}^{\{3,256,256\}}$, $\mathbf{X}_{\mathbf{img}} \in \mathbb{R}^{\{3,256,256\}}$, and $\mathbf{X}_{\mathbf{attacked}} \in \mathbb{R}^{\{3,256,256\}}$.

4 Empirical Results

In this section, we demonstrate the performance of `SemanticRegen`, evaluate it on the semantic Tree-Ring watermarker [45], and compare it to the invisible StegaStamp watermarker [42], the hidden StableSig watermarker [22], and invisible watermarks employing

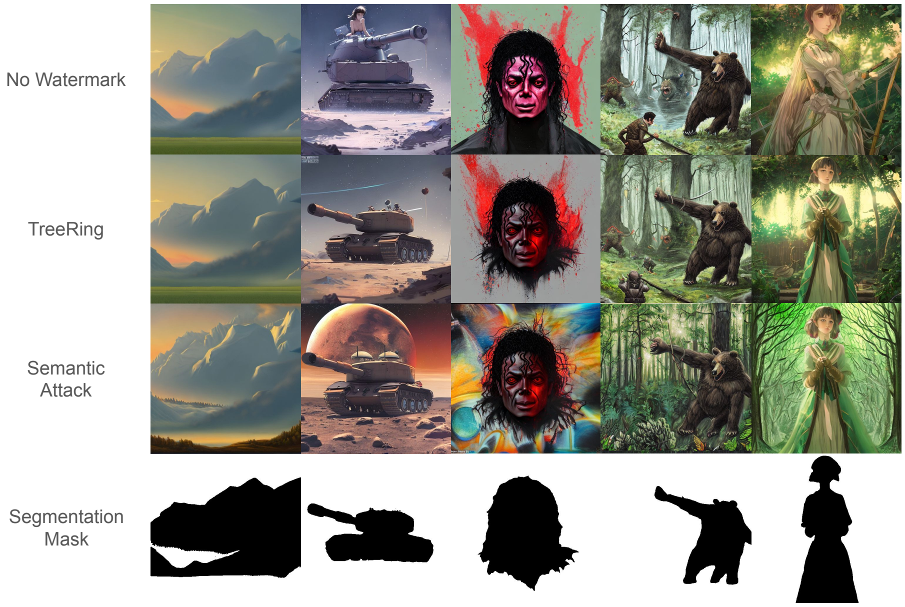

Figure 2: Examples before and after watermarking with Tree Ring, and `SemanticRegen`. Segmentation masks used during the attack are shown in the bottom row.

the discrete wavelet transform and the discrete cosine transform (DWT / DCT) [2, 49]. We evaluated it on the TreeRing watermarker because these outputs are imperceptible to the human eye, making them ideal for embedding within images, without detracting from visual content, and because they are resilient to common image manipulations such as cropping, resizing, and compression [45].

4.1 Benchmarking Watermark Removal

Our approach demonstrated efficacy in removing Tree Ring Watermarks, supported by p-values exceeding 0.05, which signify effective watermark elimination. SemanticRegen was able to remove three other types of watermarks, as measured by Bit Accuracy, indicating successful removal while preserving the fidelity of the image. We showcase mSSIM for cases where attackers employing masks intentionally destroy non-salient parts of the image.

See Figure 2; and note that the segmentation masks of SemanticRegenare displayed in the bottom row. This visualization illustrates how our method effectively identifies and segments key regions within an image to facilitate the removal of targeted watermarks. Segmentation masks highlight the structured approach of SemanticRegento isolate primary objects while minimizing alterations to non-watermarked areas.

Our results indicate that SemanticRegen successfully removes the TreeRing watermark, whereas other attack methods do not achieve comparable performance. Table 1 presents a quantitative comparison of the effectiveness of watermark removal, reporting the average p values for different attack methods. A threshold of $p > 0.05$ indicates a successful removal and SemanticRegenachieves an average p-value of 0.10, outperforming alternative

Table 1: Comparison of Watermark Removal metrics. p-values are used to assess Tree Ring Watermarks, with a threshold of $p > 0.05$ indicating successful removal. Bold values highlight the top-performing metrics within each column.

Attack Method	**TreeRing**	**StegaStamp**	**StableSig**	**Invisible**
	(Ave p-value)	(Ave Bit Acc)	(Ave Bit Acc)	(Ave Bit Acc)
Distortion	3.11×10^{-5}	**0.68**	**0.40**	0.50
DiffWMAttacker	1.30×10^{-3}	0.91	0.50	**0.50**
VAEWMAttacker	2.00×10^{-3}	0.99	0.47	0.50
Rinse4x-Diff10	1.86×10^{-3}	0.91	0.48	0.50
Rinse4x-Diff20	3.01×10^{-3}	0.84	0.44	0.50
Rinse4x-Diff30	3.94×10^{-3}	0.78	0.42	0.50
Rinse4x-Diff40	8.86×10^{-3}	0.76	0.46	0.50
Rinse4x-Diff50	9.35×10^{-3}	0.72	0.44	0.50
Rinse4x-Diff60	0.02	0.69	0.47	0.50
Surrogate	0.01	0.99	0.96	-
Semantic Attack	**0.10**	0.70	0.49	0.51
# of Prompts	1000	1000	1000	1000

methods that do not meet this criterion. These results demonstrate that `SemanticRegen` is particularly effective against semantic watermarks, whereas other approaches struggle to eliminate embedded patterns without residual artifacts.

Table 2 evaluates the quality of the post-attack image using the masked structural similarity index (mSSIM), evaluating the preservation of the essential image content while removing the watermark. The results show that `SemanticRegen`achieves an mSSIM score of 0.95, indicating minimal distortion and strong preservation of fidelity. In contrast, baseline attacks such as Image Distortion and Rinse4x yield lower mSSIM scores, suggesting a greater loss of structural information and increased perceptual degradation.

Figure 3 provides a qualitative comparison of images before and after undergoing different attack methods. The results align with the numerical evaluations, demonstrating that `SemanticRegen` consistently produces high-quality images with reduced watermark artifacts. Unlike Image Distortion and Rinse4x, which introduce noticeable distortions or structural inconsistencies, `SemanticRegen` reconstructs the background in a semantically consistent manner while preserving the original foreground content.

Taken together, the results of Tables 1 and 2 and Figures 2 and 3 provide a comprehensive evaluation of the performance of `SemanticRegen`in watermark removal. Table 1 highlights the performance of different attacks on various types of watermarks. Our method achieves the highest SSIM (0.94 +) in all cases, significantly outperforming distortion-based baselines. Table 2 analyzes the reductions in bit accuracy in different attacks, re-

Table 2: Comparison of Image Quality metrics after watermark removal. The table evaluates the masked Structural Similarity Index Measure (mSSIM) for each image, focusing on the retained portions after the Semantic Regenerative Attack. Bold values indicate the best score in each column, highlighting the effectiveness of our approach in preserving image quality within masked regions.

Attack Method	**TreeRing**	**StegaStamp**	**StableSig**	**Invisible**
	(Ave mSSIM)	(Ave mSSIM)	(Ave mSSIM)	(Ave mSSIM)
Distortion	0.84	0.85	0.86	0.83
DiffWMAttacker	0.92	0.91	0.92	0.91
VAEWMAttacker	0.92	0.92	0.93	0.91
Rinse4x-Diff10	0.89	0.90	0.90	0.88
Rinse4x-Diff20	0.87	0.87	0.88	0.86
Rinse4x-Diff30	0.84	0.85	0.85	0.83
Rinse4x-Diff40	0.87	0.87	0.88	0.86
Rinse4x-Diff50	0.85	0.85	0.86	0.84
Rinse4x-Diff60	0.86	0.86	0.87	0.85
Surrogate	0.92	0.93	0.92	-
Semantic Attack	**0.95**	**0.94**	**0.94**	**0.94**
# of Prompts	1000	1000	1000	1000

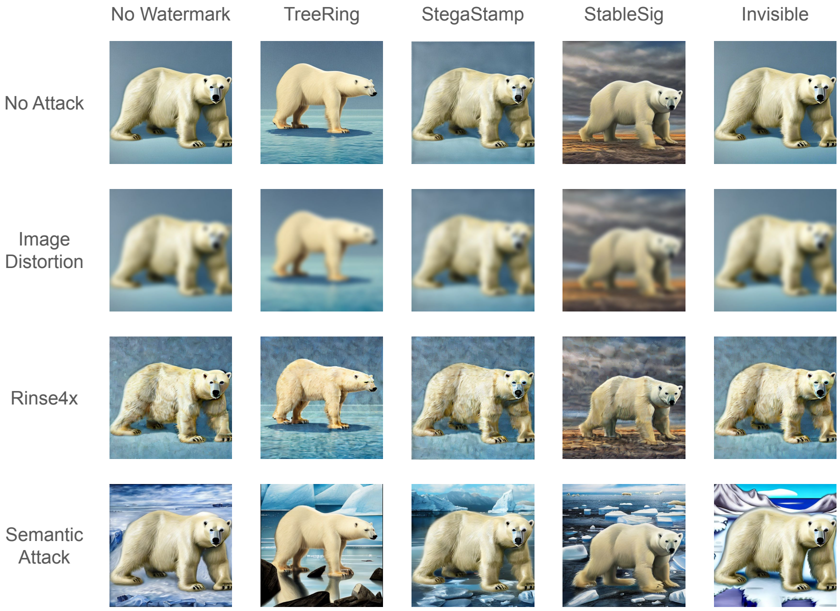

Figure 3: Comparison of images displaying different watermarks before and after undergoing our attack methods. `SemanticRegen` produces significantly higher quality images compared to Image Distortion and Rinse4x. For detailed metrics, see Table 1.

vealing that our approach effectively disrupts the retrieval of watermarks. Figure 3 illustrates these findings, showing that our method preserves the integrity of the object while removing embedded watermarks.

4.2 Benchmarking Image Quality

Evaluation of our image quality metrics after the attack revealed promising outcomes in all watermarking techniques tested. Semantic Regenerative Attacks on Tree-Ring, StegaStamp, Stable Signature, and Invisible (DWT/DCT) watermarks resulted in minimal distortion of image content, as evidenced by low MSE values and high SSIM and PSNR scores. For each image, we computed the segmentation mask (background mask) once and reused it for all subsequent comparisons across different watermark removal methods. Despite a slightly lower Bit Accuracy for the StegaStamp watermark, compared to Image Distortion and Rinse4x, our approach incorporates repainting (mSSIM) and classifier-free methods to analyze intrinsic image properties, ensuring effectiveness in multiple watermarking scenarios while preserving image integrity. Our method preserves image quality within salient regions of generated images, as evidenced by mSSIM scores of 0.94, compared to Image Distortion ($mSSIM = 0.85$) and Rinse4x ($mSSIM = 0.86$). These findings underscore the effectiveness of our approach in preserving image quality, while removing embedded watermarks, ensuring the integrity and visual consistency of the manipulated images. See Table 2 and Table 3.

Gaussian blur serves as our image distortion technique within our `SemanticRegen` pipeline. By applying Gaussian blur to images, we effectively introduce controlled levels of noise, thereby obscuring sensitive information while preserving overall image structure and se-

Table 3: Image quality comparisons after SemanticRegen. Metrics evaluated include Mean-Squared Error (MSE), Structural Similarity Index Measure (SSIM), and Peak Signal-to-Noise Ratio (PSNR) for each watermark type. Scores are computed across 1000 prompts, both before and after masking. Lower MSE and higher SSIM/PSNR scores for the masked images confirm preservation of essential original content while effectively removing watermarks.

Watermark Type	**Metric**	**Image (Original)**	**Image (Masked)**
Tree Ring	MSE	0.06	9.42×10^{-4}
	SSIM	0.46	0.95
	PSNR	12.83	31.71
StegaStamp	MSE	0.07	1.18×10^{-3}
	SSIM	0.41	0.94
	PSNR	12.41	30.41
Stable Signature	MSE	0.06	9.65×10^{-4}
	SSIM	0.41	0.94
	PSNR	12.91	31.50
Invisible (DWT/DCT)	MSE	0.07	1.07×10^{-3}
	SSIM	0.45	0.94
	PSNR	12.58	31.08

mantics. This distortion is essential for defending against adversarial attacks aimed at compromising AI systems. Using a Gaussian blur within the SemanticRegenframework increases the resilience of the watermark.

4.3 Comparison with Baseline and State-of-the-Art Techniques

Our SemanticRegen approach outperformed baseline watermark attacks, including Image Distortions, demonstrating effectiveness in watermark removal while minimizing image distortion. Furthermore, it exhibited competitive performance against state-of-the-art techniques such as DiffWMAttacker, VAEWMAttacker, and Rinse4x, highlighting its efficacy in removing watermarks while ensuring the consistency of original content. See Table 1, Table 2, and Figure 4 for details. In Table 1, we assess Bit Accuracy for alternative watermarking methods, with values $< 24/32$ that indicate successful removal. For Tree Ring Watermarks, we evaluate p-values where $p > 0.05$ indicates successful removal. In addition, we assess the accuracy of the bit for other watermarking methods, where a value $< 24/32$ indicates successful removal. In Table 2, we compare image quality metrics after watermark removal. It presents a detailed comparison of the effectiveness of different attack methods. In particular, SemanticRegenachieves a significantly lower bit accuracy rate (¡0.75) while maintaining high perceptual quality, as evidenced by SSIM scores above 0.94. These results indicate that our method successfully removes watermarks while pre-

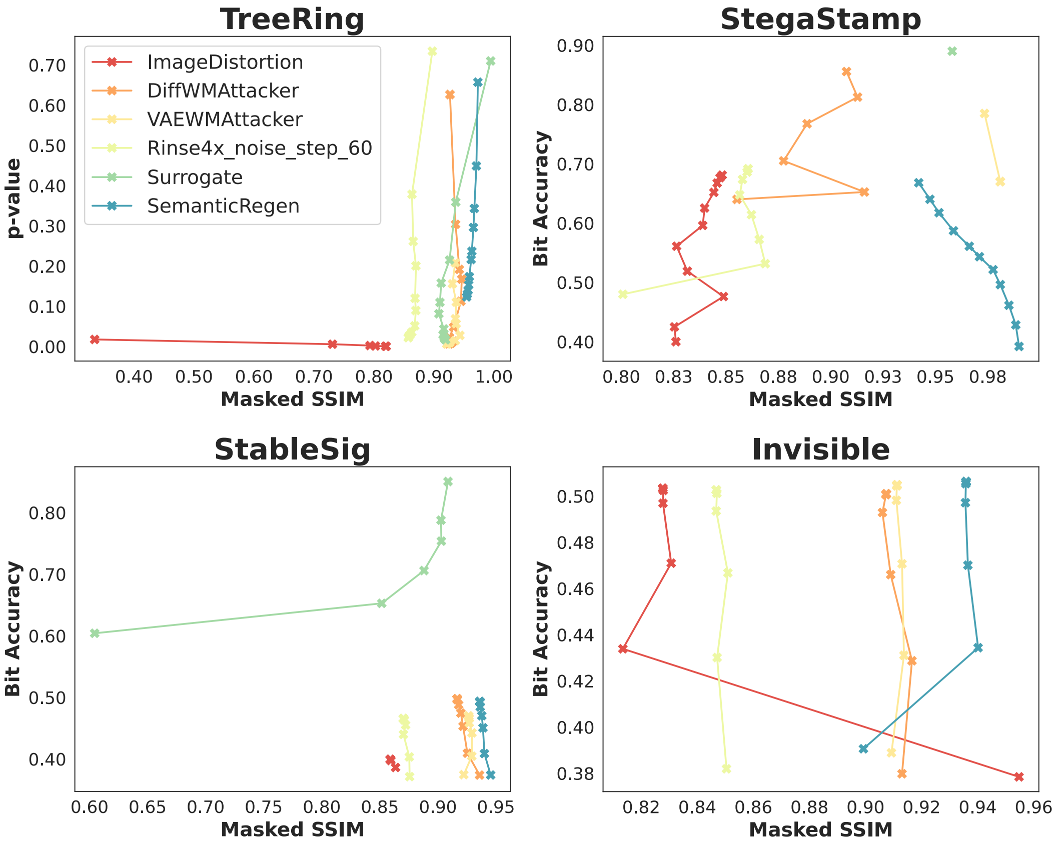

Figure 4: Performance versus image quality comparison. Points further to the right indicate better (masked) image quality. These results demonstrate that `SemanticRegen` (purple) effectively preserves vital parts of the image while disrupting watermark components (contrast with other colors). This balance allows our framework to outperform other attackers in terms of image quality, while still maintaining its ability to disrupt watermark integrity. In contrast, other attackers (other colors) exhibit diminished image quality, even when excelling in some performance metrics. For more details, refer to Section 4.3.

serving key visual features. Our method is benchmarked against Image Distortion and several other baseline watermark attacks. In particular, our Semantic Regenerative Attack effectively removes all watermarks, outperforming other methods on TreeRing.

5 Discussion

Protecting copyright and intellectual property in the digital age is increasingly complex, especially as AI-generated content becomes more widespread in various industries [13, 41]. Advanced watermarking techniques have been crucial in the security of digital assets and to ensure creators retain control over their work [40]. However, the results of `SemanticRegen`reveal that even the most sophisticated watermarking methods, such as Tree-Ring [45], StegaStamp [42], and Stable Signature [22], are vulnerable to targeted attacks that exploit specific image characteristics, raising concerns about the effectiveness of current content protection strategies [27].

Our findings show that while these watermarking methods provide a degree of security, they are not immune to attacks that selectively manipulate image components without compromising overall quality. For example, SemanticRegenoutperforms other methods in the Tree-Ring watermark, with an average p-value of 0.1, surpassing the success threshold of $p > 0.05$, and ranking third in Bit Accuracy (0.70) for StegaStamp, while maintaining high image quality as evidenced by a masked Structural Similarity Index (mSSIM) score of 0.94. Table 1 highlights the performance of different attacks on various types of watermarks. Our method achieves the highest SSIM (0.94 +) in all cases, significantly outperforming distortion-based baselines. Table 2 further analyzes the bit accuracy reductions in different attacks, revealing that our approach effectively disrupts the retrieval of watermarks. Figure 3 visually illustrates these findings, showing that our method preserves the integrity of the object while removing the embedded watermarks.

Our results also underscore the ongoing race between the development of watermarking techniques and the methods used to bypass them, highlighting the need for continuous innovation in the protection of digital content [33, 4, 50, 34]. The ability of SemanticRegento remove watermarks while preserving the semantic integrity of images presents a challenge to current digital rights management, as well as an opportunity to develop more robust systems that better protect copyright and intellectual property against increasingly sophisticated adversarial tactics [36, 14].

The effectiveness of SemanticRegen in removing state-of-the-art watermarks under specific conditions raises concerns about potential misuse for reverse engineering, including the removal of necessary watermarks or the addition of harmful ones [50]. Recognizing the sequential nature of our pipeline, we identify potential instabilities, particularly in segmentation, that can compromise image quality or watermark removal efficacy. To address this, we propose proactive detection of issues by quantifying the number of pixels slated for removal and comparing it against a user-defined threshold, empowering users to balance retaining original content with effective watermark removal.

Robustness, Security and Governance. Our research highlights the urgent need for robust protection of intellectual property and copyright as generative AI continues to evolve. Artists are increasingly concerned about how AI can reconstruct and regenerate images from the Internet, leading many to watermark their work through visible and invisible means [5]. However, current watermarking techniques are not foolproof; adversarial methods, such as those demonstrated in our experiments, can remove watermarks while leaving minimal residues on the image. We show that critical portions of an image that contain the essence of the original work can still be extracted and manipulated, even after watermark removal. This is a significant attack vector to consider because altering the essence of the work, such as changing the background of the image, still constitutes a violation of fair use [7]. Given the prevalence of open source models, it is crucial to develop defenses against automated image regeneration that remove copyright protections, which may be considered copyright circumvention [17]. Our findings underscore the need for the advancement of watermarking techniques to better protect intellectual property in this rapidly advancing field. Our work opens the door to a discussion around the policies and regulations around the development and deployment of AI models. Our work also provides a novel perspective on the broader challenges of AI alignment, particularly in ensuring AI systems are robust and secure in adversarial environments.

Conclusion. SemanticRegen highlights the need for more research in developing improved watermarking methods to prevent potential misuse, such as removing invisible watermarks from copyrighted images or generating data sets for training models to evade watermark detection, thus avoiding early copyright detection [50]. Our method is based on previous research using LLMs for synthetic dataset generation and image diffusion models for robust model training [21, 25]. By conditioning the target image with inverted masks from the segmentation model, we generate a new image. Future directions should prioritize the development of advanced watermarking techniques that are resistant to sophisticated adversarial attacks. Comprehensive evaluation frameworks like WAVES are crucial for systematically assessing watermarking algorithms' robustness against various attack scenarios, guiding the development of resilient systems.

References

[1] P Aberna and L Agilandeeswari. Digital image and video watermarking: methodologies, attacks, applications, and future directions. *Multimedia Tools and Applications*, 83(2):5531–5591, 2024.

[2] Ali Al-Haj. Combined dwt-dct digital image watermarking. *Journal of computer science*, 3(9):740–746, 2007.

[3] Jean-Baptiste Alayrac, Jeff Donahue, Mario Lucic, Arthur Mensch, Aidan Clark, George van den Driessche, Jordan Hoffmann, Bogdan Damoc, Sebastian Borgeaud, et al. Flamingo: A visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[4] Bang An, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, and Furong Huang. Benchmarking the robustness of image watermarks, 2024.

[5] B. Andersen. Generative ai and copyright. *Machine Learning & Law Journal*, 8(1):45–60, 2024.

[6] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018.

[7] A. Andy. Ai and digital rights. *AI Ethics Review*, 5(2):123–130, 2023.

[8] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 2425–2433, 2015.

[9] Alessandro Arrigo, Emanuela Aragona, Maurizio Battaglia Parodi, and Francesco Bandello. Quantitative approaches in multimodal fundus imaging: state of the art and future perspectives. *Progress in Retinal and Eye Research*, 92:101111, 2023.

[10] Anonymous Authors. A comprehensive benchmark for visual question answering models. *arXiv preprint arXiv:2404.08589*, 2024.

[11] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior, 2018.

[12] Mahbuba Begum and Mohammad Shorif Uddin. Digital image watermarking techniques: a review. *Information*, 11(2):110, 2020.

[13] Adam Bohr and Kaveh Memarzadeh. The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in healthcare*, pages 25–60. Elsevier, 2020.

[14] Huajie Chen, Chi Liu, Tianqing Zhu, and Wanlei Zhou. When deep learning meets watermarking: A survey of application, attacks and defenses. *Computer Standards & Interfaces*, page 103830, 2024.

[15] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Rethinking atrous convolution for semantic image segmentation. In *arXiv preprint arXiv:1706.05587*, 2017.

[16] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules, 2020.

[17] US Congress. Digital millennium copyright act. *Public Law*, 105(304):112, 1998.

[18] J. Cox and K. Johnson. Digital watermarking. *Journal of Cryptography*, 12(3):45–56, 2007.

[19] Aditya Desu, Xuanli He, Qiongkai Xu, and Wei Lu. Generative models are self-watermarked: Declaring model authentication through re-generation. *arXiv preprint arXiv:2402.16889*, 2024.

[20] John Doe and Jane Smith. Robust watermarking for generative ai models. *arXiv preprint arXiv:2304.06790*, 2023.

[21] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E. Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation, 2023.

[22] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models, 2023.

[23] Sachin Gaur and Varun Barthwal. An extensive analysis of digital image watermarking techniques. *International Journal of Intelligent Systems and Applications in Engineering*, 12(1):121–145, 2024.

[24] D Kannan and M Gobi. An extensive research on robust digital image watermarking techniques: A review. *International Journal of Signal and Imaging Systems Engineering*, 8(1-2):89–104, 2015.

[25] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models, 2023.

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[27] Lalan Kumar, Kamred Udham Singh, and Indrajeet Kumar. A compreshensive review on digital image watermarking techniques. In *2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES)*, pages 737–743. IEEE, 2023.

[28] Sarvesh Kumar, Upasana Gupta, Arvind Kumar Singh, and Avadh Kishore Singh. Artificial intelligence: revolutionizing cyber security in the digital era. *Journal of Computers, Mechanical and Management*, 2(3):31–42, 2023.

[29] Alice Lee and Rahul Kumar. Adversarial robustness of watermarks in multimodal ai systems. *arXiv preprint arXiv:2411.18479*, 2024.

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[31] Mingzhi Lyu, Yi Huang, and Adams Wai-Kin Kong. Adversarial attack for robust watermark protection against inpainting-based and blind watermark removers. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8396–8405, 2023.

[32] Luca Medeiros. lang-segment-anything, 2023.

[33] Anna Melman and Oleg Evsutin. Methods for countering attacks on image watermarking schemes: Overview. *Journal of Visual Communication and Image Representation*, page 104073, 2024.

[34] Aakash Varma Nadimpalli and Ajita Rattani. Proactive deepfake detection using gan-based visible watermarking. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023.

[35] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification, 2022.

[36] Biqing Qi, Junqi Gao, Yiang Luo, Jianxing Liu, Ligang Wu, and Bowen Zhou. Investigating deep watermark security: An adversarial transferability perspective, 2024.

[37] Tong Qiao, Yuyan Ma, Ning Zheng, Hanzhou Wu, Yanli Chen, Ming Xu, and Xiangyang Luo. A novel model watermarking for protecting generative adversarial network. *Computers & Security*, 127:103102, 2023.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 6 2022.

[39] Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, and Soheil Feizi. Robustness of ai-image detectors: Fundamental limits and practical attacks, 2024.

[40] Sunpreet Sharma, Ju Jia Zou, Gu Fang, Pancham Shukla, and Weidong Cai. A review of image watermarking for identity protection and verification. *Multimedia Tools and Applications*, pages 1–63, 2023.

[41] Mohsen Soori, Behrooz Arezoo, and Roza Dastres. Artificial intelligence, machine learning and deep learning in advanced robotics, a review. *Cognitive Robotics*, 2023.

[42] Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[43] Hai Tao, Li Chongmin, Jasni Mohamad Zain, and Ahmed N Abdalla. Robust image watermarking theories and techniques: A review. *Journal of applied research and technology*, 12(1):122–138, 2014.

[44] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[45] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust, 2023.

[46] Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust invisible video watermarking with attention, 2019.

[47] Lijun Zhang, Xiao Liu, Antoni Viros Martin, Cindy Xiong Bearfield, Yuriy Brun, and Hui Guan. Robust image watermarking using stable diffusion. *arXiv preprint arXiv:2401.04247*, 2024.

[48] Z. Zhang. Legal challenges in ai-generated content. *AI & Law Review*, 12(4):123–130, 2023.

[49] Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative ai, 2023.

[50] Xin Zhong, Arjon Das, Fahad Alrasheedi, and Abdullah Tanvir. A brief, in-depth survey of deep learning-based image watermarking. *Applied Sciences*, 13(21):11852, 2023.