Winning Solution to the NeurIPS 2024 Invisible Watermark Removal Challenge

Fahad Shamshad[1], **Tameem Bakr**[1], **Yahia Shaaban**[1],
Noor Hussein[1,2], **Karthik Nandakumar**[1,2], **Nils Lukas**[1]
[1]Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE
[2]Michigan State University (MSU), USA
{firstname.lastname}@mbzuai.ac.ae

Abstract

Content watermarking is an important tool for the authentication and copyright protection of digital media. However, it is unclear whether existing watermarks are robust against adversarial attacks. We present the **winning solution** to the NeurIPS 2024 *Erasing the Invisible Challenge*, which stress-tests watermark robustness under varying degrees of adversary knowledge. The challenge consisted of two tracks: a black-box and beige-box track, depending on whether the adversary knows which watermarking method was used by the provider. For the **beige-box** track, we leverage an *adaptive* VAE-based evasion attack, with a test-time optimization and color-contrast restoration in CIELAB space to preserve the image's quality. For the **black-box** track, we first cluster images based on their artifacts in the spatial or frequency-domain. Then, we apply image-to-image diffusion models with controlled noise injection and semantic priors from ChatGPT-generated captions to each cluster with optimized parameter settings. Empirical evaluations demonstrate that our method successfully **achieves near-perfect watermark removal** (95.7%) with negligible impact on the residual image's quality. We hope that our attacks inspire the development of more robust image watermarking methods.

1 Introduction

Content watermarking is a widely used technique for embedding imperceptible information into digital media to ensure provenance verification [14], copyright protection [46, 60], and content traceability [28]. With generative AI systems now capable of producing high-fidelity synthetic content at scale, watermarking serves as an essential safeguard for content owners and organizations to combat unauthorized distribution [66, 57] and forgery [43, 5]. The goal of a watermarking method is to hide a signal (message) in generated content that can only be detected with a secret watermarking key [1, 31]. However, despite its widespread deployment, watermarking systems remain vulnerable to both unintentional distortions (*e.g.,* blur, resizing) and targeted adversarial attacks that intentionally suppress the watermark signal without perceptually degrading the content [4, 9].

A robust watermarking scheme must ensure that successful removal is only possible at the cost of significant perceptual degradation [66, 32, 19]. However, recent studies have exposed critical vulnerabilities in existing methods, where adversaries can remove or spoof watermark signals by exploiting adaptive attacks [33, 9], latent-space priors [37, 23], or even simple averaging techniques [61]. These attacks preserve visual fidelity, often introducing no perceptible artifacts, thereby undermining the reliability of watermark detectors. As watermarking becomes increasingly important for ensuring the integrity, attribution, and traceability of AI-generated content, addressing these weaknesses is imperative. Progress in this area requires rigorous evaluation under realistic threat models to

39th Conference on Neural Information Processing Systems (NeurIPS 2025). Workshop: The First Workshop on Generative and Protective AI for Content Creation.

Table 1: Black-box Track Final Leaderboard

Black-box Track				
Rank	Participant	Detection	Quality	Total
1	**Team-MBZUAI**	**0.043**	**0.136**	**0.143**
2	Team-SHARIF	0.063	0.158	0.170
3	Team-UFL	0.087	0.177	0.197

Table 2: Beige-box Track Final Leaderboard

Beige-box Track				
Rank	Participant	Detection	Quality	Total
1	**Team-MBZUAI**	**0.037**	**0.153**	**0.157**
2	Team-SONY	0.050	0.176	0.183
3	Team-SHARIF	0.127	0.222	0.256

Figure 1: *Top row*: Original watermarked images. *Bottom row*: Images after our attack, with minimal perceptual difference from the originals, showcasing the effectiveness of our method in preserving visual fidelity.

inform the design of watermarking schemes that are not only imperceptible but also resilient to both adversarial manipulations and common distortions.

To gauge the robustness of invisible watermarking methods against realistic threats, the NeurIPS 2024 *Erasing the Invisible: A Stress-Test Challenge for Image Watermarks* [11] introduced a rigorous benchmark targeting watermark removal. This paper presents our **first-place winning solution** to this challenge under two practical threat models: **beige-box**, where the watermarking methodology was known, and **black-box**, where no prior knowledge was available. For the beige-box scenario, we design an adaptive VAE-based attack for effective watermark removal, coupled with test-time optimization and frequency-aware color restoration to preserve image quality. We also show that simple spatial-domain translations can effectively disrupt TreeRing-based watermarks. In the black-box setting, we design targeted, cluster-specific removal attacks by grouping images based on spatial and spectral artifacts, followed by diffusion-based purification guided by semantic captions. *Our approach outperformed the runner-ups by 26% and 31.7% in detection score on the beige-box and black-box tracks respectively, while preserving high perceptual quality (see Table 1 and Table 2).* By exposing vulnerabilities in existing watermarking methods, we aim to inspire the development of more robust defenses against such attacks.

2 Related Work

Image Watermarking. Image watermarking embeds imperceptible signals into digital images to support authentication, copyright enforcement, and forensic traceability [8, 42, 39, 12]. Traditional image watermarking approaches operate in the spatial or frequency domains by modifying pixel values or transform-domain coefficients such as Discrete Cosine Transform [7], Discrete Wavelet Transforms [3], or Discrete Fourier Transform [24], often trading off between imperceptibility and robustness. More recently, deep learning-based watermarking schemes leverage convolutional neural networks and generative models to embed watermarks via learned feature spaces [63, 20]. Methods such as StegaStamp [54], Gaussian Shading [62], IConMark [44], Robin [21], and TreeRing [59] improve robustness under common corruptions (*e.g.,* JPEG compression, resizing) and are increasingly adopted to tag AI-generated content. Despite their effectiveness in benign settings, these techniques

remain vulnerable to both incidental degradation and targeted removal attacks [4, 67], underscoring the need for watermarking methods with stronger resilience.

Robustness of Image Watermarks. Image watermarks are vulnerable to degradation from common distortions such as Gaussian noise, blurring, and compression. More critically, adversarial attacks intentionally exploit model vulnerabilities to remove the watermark signal while maintaining high perceptual quality [22, 61, 15]. To mitigate these threats, recent approaches employ adversarial training, where watermarking models are optimized to withstand a range of perturbations [21, 55]. Generative models, including autoencoders and diffusion models, have also been leveraged to embed more resilient watermarks by aligning with the natural image manifold [21]. Despite these advances, achieving robustness without compromising imperceptibility remains an open challenge, especially with partial or complete knowledge of the watermarking algorithm [35, 13]. The NeurIPS 2024 *Erasing the Invisible challenge* [11] provides a benchmark for evaluating the resilience of state-of-the-art image watermarking methods under beige-box and black-box threat models.

Generative Models. Pretrained generative models have demonstrated strong performance across a range of vision tasks, including image restoration [6, 68, 47, 50, 48, 49], privacy preservation [51, 52], and adversarial purification [38, 45]. Leveraging their powerful priors over natural images, models such as VAEs [25], GANs [16], and diffusion models [18] can reconstruct high-fidelity content from corrupted or perturbed inputs. These capabilities have recently been explored in the context of watermarking, where the goal is to remove imperceptible signals embedded within images while preserving visual quality [30, 67]. For example, VAEs can project watermarked images into clean latent spaces that suppress hidden signals, while diffusion-based regeneration/rinsing methods have shown promise in erasing invisible perturbations [29, 4]. Recently, inspired by the inductive bias of untrained neural networks [56, 41, 40], such architectures have also been explored for invisible image watermark removal [27], demonstrating surprising effectiveness.

3 Proposed Approach

Challenge Overview: The NeurIPS 2024 competition *Erasing the Invisible: A Stress-Test Challenge for Image Watermarks [11]* rigorously assesses invisible watermarking robustness under realistic threat models. The benchmark comprises **300 watermarked images per track** across two settings: a **beige-box** scenario, where the attacker knows the watermarking algorithm but not its parameters, and a **black-box** scenario, where no prior knowledge is available. Participants were tasked with removing invisible watermarks while preserving perceptual image quality. Evaluation was based on two criteria: **Detection score** measures watermark removal success as the true positive rate (TPR) at a 0.1% false positive rate (FPR). For each method, a detection threshold is set using the 0.001st percentile of distances $d(m, m')$ between known messages m and decoded outputs m' from 10,000 unwatermarked images. An attacked image is flagged as watermarked if its decoded message falls below this threshold. **Image quality score** is a weighted combination of low-level fidelity metrics (PSNR, SSIM [58], NMI) and high-level perceptual metrics (FID, CLIP-FID [26], LPIPS [65], Delta Aesthetics [11], and Delta Artifacts [11]). Each metric is normalized based on attack sensitivity and weighted accordingly: positive weights for metrics where lower is better (*e.g.*, FID), and negative for those where higher is better (*e.g.*, SSIM). *We can query the competition leaderboard up to five times per day for both beige-box and black-box tracks to evaluate our submission performance*. Below, we provide details of our proposed attack for both the beige-box and black-box tracks.

3.1 Beige-Box Track

In the beige-box track, we had access to the algorithmic description of the watermarking method used to embed invisible watermarks in the provided images, but not its hyperparameters or the underlying generative model. This partial prior knowledge enabled the design of an adaptive attack strategy that targeted the known watermarking scheme, akin to recent efforts [33, 9]. The organizers released 300 watermarked images, equally divided between two watermarking methods: a modified StegaStamp algorithm [54] and a variant of the TreeRing method [59]. We designed a dedicated watermark removal pipeline for each method.

3.1.1 Stegastamp Watermark Removal

For the StegaStamp-based watermark removal, we developed a three-stage pipeline: (**i**) *paired data generation*, (**ii**) *adaptive VAE fine-tuning*, and (**iii**) *post-processing for quality restoration*. This structured methodology enabled effective suppression of the embedded watermark while preserving the perceptual quality of the images. The overall pipeline is shown in Figure 2.

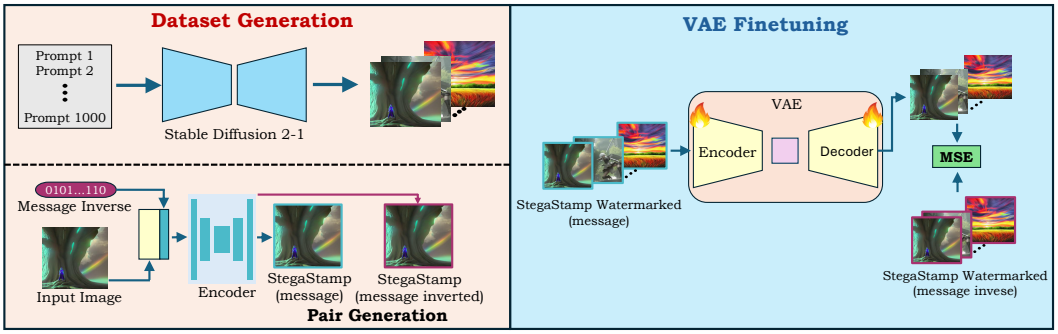

Figure 2: **Overview of our dataset generation and VAE-based watermark removal pipeline for StegaStamp watermarks**. We generate 1,000 images from Stable Diffusion 2-1 using publicly available prompts, embedding each with a random binary message and its inverse via a StegaStamp encoder to create paired watermarked images. A fine-tuned Variational Autoencoder (VAE) minimizes Mean Squared Error (MSE) loss between pairs, effectively suppressing watermark artifacts while preserving perceptual quality.

Paired Dataset Generation: We first curated a comprehensive training dataset leveraging 1,000 text prompts from the Hugging Face Stable-Diffusion-Prompts dataset [17]. Using these prompts, we generated corresponding images via Stable Diffusion 2-1 with a guidance scale of 7.5 and 50 inference steps. Each generated 512^2 image was resized to 400^2 pixels using bilinear interpolation before being processed through a pretrained StegaStamp model from the WAVES repository [4]. The key aspect of our dataset preparation involved creating image pairs where each original image was encoded with both a 100-bit binary message m sampled uniformly at random and its inverse $1 - m$, resulting in a dataset of 1,000 paired examples that captured watermarking artifacts.

Adaptive VAE Finetuning: The core of our attack framework centers on a Variational Autoencoder (VAE) that was adaptively tuned to perform watermark removal via supervised reconstruction. Given a watermarked image x_w and its inverse message counterpart x_i, the VAE - consisting of encoder E_θ and decoder D_ϕ-was optimized to reconstruct x_i from x_w, using the following MSE loss:

$$\mathcal{L}(\theta, \phi) = \|D_\phi(E_\theta(x_w)) - x_i\|_2^2, \tag{1}$$

where $D_\phi(E_\theta(x_w))$ represents the reconstructed image from the watermarked input, and x_i is the target image containing the inverted message. We optimize this objective using Adam optimizer with learning rate $\alpha = 1 \times 10^{-5}$ for 10 epochs with a batch size of 16. To stabilize training, we employ gradient clipping with a maximum norm of 1.0. Model training was performed on an NVIDIA A6000 GPU (48GB VRAM) and completed in under two GPU hours. We adapted a pretrained SDXL VAE model. This adaptive fine-tuning stage enabled the model to effectively strip away the watermark while preserving image structure. The VAE Finetuning algorithm is given in Algorithm. 2.

Quality-Preserving Post-Processing: Despite strong watermark removal performance, VAE reconstructions degraded color and contrast fidelity as shown in Figure. 3. To address this issue, we introduced a two-stage post-processing pipeline (Algorithm 1) with the aim to enhance the image quality without re-introducing the removed watermark.

1 - Test-Time VAE Optimization: Using the SDXL Refiner VAE, we performed image-specific optimization by fine-tuning the VAE parameters $\{\theta, \phi\}$ to better align with the original watermarked input x_w. The loss function combined pixel-wise, perceptual (LPIPS), and structural (SSIM) terms:

$$\mathcal{L}_{\text{total}} = \underbrace{\|D_\phi(E_\theta(x_r)) - x_w\|^2}_{\text{MSE Loss}} + \underbrace{\mathcal{L}_{\text{LPIPS}}(D_\phi(E_\theta(x_r)), x_w)}_{\text{Perceptual Loss}} + \underbrace{0.5(1 - \text{SSIM}(D_\phi(E_\theta(x_r)), x_w))}_{\text{Structural Similarity Loss}} \tag{2}$$

While this step effectively removes the watermark signal, it can introduce slight degradation in visual quality, motivating the need for the second stage.

2 - Color and Contrast Transfer: The second stage restores perceptual quality by adjusting color and contrast in the CIELAB color space. Let x_{opt} denote the output from test-time optimization, with CIELAB components $\{L_{\text{opt}}, a_{\text{opt}}, b_{\text{opt}}\}$, and let $\{L_w, a_w, b_w\}$ be those of the original watermarked image x_w. For **color transfer**, we preserve the luminance L_{opt} from the optimized image and adopt the chrominance components from the watermarked image, yielding an intermediate image $x_c = \mathcal{F}_{\text{RGB}}(L_{\text{opt}}, a_w, b_w)$, where $\mathcal{F}_{\text{RGB}}$ denotes conversion from CIELAB to RGB space. Next, for **contrast transfer**, we match the statistical moments of the luminance channel. Let μ_c, σ_c and μ_w, σ_w

Algorithm 1 Test-Time Image Optimization

Require: Finetuned VAE $(E_{\theta^*}, D_{\phi^*})$, watermarked image x_w, **optional** refinement steps T, step size η

#Inference
  $z \leftarrow E_{\theta^*}(x_w)$ # Encode watermarked image
  $x_r^{(0)} \leftarrow D_{\phi^*}(z)$ # Decode latent representation
if T **defined then**
  #Test-Time Image Optimization
  Initialize refiner VAE params (θ_r, ϕ_r)
  for $t = 1$ **to** T **do**
    $L_{\text{total}}(\theta_r, \phi_r) = \left\| D_{\phi_r}(E_{\theta_r}(x_r^{(t-1)})) - x_w \right\|_2^2 + \text{LPIPS} + 0.5\,(1 - \text{SSIM})$
    $(\theta_r, \phi_r) \leftarrow (\theta_r, \phi_r) - \eta \nabla_{(\theta_r, \phi_r)} L_{\text{total}}$ # Gradient Update
    $x_r^{(t)} \leftarrow D_{\phi_r}(E_{\theta_r}(x_r^{(t-1)}))$
  end for
  #Color-Contrast Transfer
  Convert $x_r^{(T)}$, x_w to CIELAB: $(L_c, a_c, b_c), (L_w, a_w, b_w)$
  $L_c' \leftarrow \frac{\sigma_w}{\sigma_c}(L_c - \mu_c) + \mu_w$
  $x_{\text{final}} \leftarrow \text{CIELAB} \rightarrow \text{RGB}(L_c',\, a_w,\, b_w)$
else
  $x_{\text{final}} \leftarrow x_r^{(0)}$
end if

Ensure: Recovered image x_{final}

Figure 3: *Top row*: Original watermarked images. *Bottom row*: Images after our attack, with minimal perceptual difference from the originals, showcasing the effectiveness of our method in preserving visual fidelity.

represent the mean and standard deviation of the luminance channels of x_c and x_w, respectively. The adjusted luminance is computed as $L_{\text{final}} = \frac{\sigma_w}{\sigma_c}(L_c - \mu_c) + \mu_w$. The final image is reconstructed as:

$$x_{\text{final}} = \mathcal{F}_{\text{RGB}}(L_{\text{final}}, a_w, b_w).$$

The color and contrast transfer step enhances visual fidelity without reintroducing the watermark. For an ablation of this stage, see Table 3.

Our quality-preserving post-processing (Table 4) markedly outperforms VAE fine-tuning alone across all metrics. PSNR improves by over 6 dB and SSIM by 0.176, while perceptual distances (LPIPS, FID, CLIP-FID) are substantially reduced. These gains show the effectiveness of combining image-specific test-time optimization with CIELAB-based color/contrast transfer, restoring fidelity without reintroducing the watermark.

3.1.2 TreeRing Watermark

For images embedded with TreeRing watermarks [59], we discovered a notable vulnerability to phase perturbations in the frequency domain. Specifically, modifying the phase component of the Fourier spectrum induces a spatial translation in the image domain—an operation that can effectively disrupt the watermark signal without introducing noticeable perceptual changes. Building on this insight, we implemented a lightweight spatial-domain defense by applying a horizontal shift to each image as $x_{\text{shifted}} = \mathcal{T}(x_w, \Delta x)$, where x_w is the input watermarked image, $\mathcal{T}$ denotes a horizontal translation operator, and $\Delta x = 7$ pixels is the empirically determined optimal

Table 3: Performance comparison of our watermark removal pipeline. **Det.** refers to the detection metric (lower is better), **Qual.** measures image quality. Progressive enhancements through test-time optimization and color/contrast adjustment improve quality.

Method	Det.	Qual.	Overall
VAE Finetune	**0.023**	0.192	0.193
+ Test-Time Opt.	0.033	0.161	0.165
+ Color/Contrast	0.037	**0.153**	**0.157**

Table 4: Quantitative comparison of our watermark removal method with VAE fine-tuning alone. Test-time optimization with color/contrast transfer provides consistent gains across pixel-level and perceptual metrics.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	NMI↑	CLIPFID↓
VAE Finetune	21.899	0.647	0.264	85.383	0.231	11.805
Ours	**28.061**	**0.823**	**0.078**	**30.786**	**0.326**	**2.831**

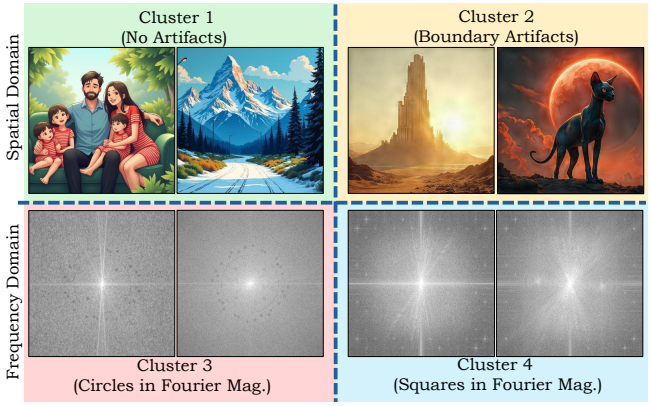

Figure 5: **Spatial–frequency artifact clustering of 300 black-box watermarked images**. Each image was manually examined for visible spatial and frequency-domain patterns. This yielded four clusters: **Cluster-1** no noticeable artifacts, **Cluster-2** boundary artifacts in the spatial domain, **Cluster-3** circular patterns in the Fourier magnitude spectrum, and **Cluster-4** square patterns in the Fourier magnitude spectrum. Identifying these patterns allowed us to design targeted removal pipelines optimized for the specific artifact type.

shift that achieves a favorable trade-off between watermark removal and perceptual fidelity (see Figure 4). However, spatial translation alone introduces visual artifacts near image boundaries. To counter this, we selectively restore the leftmost Δx columns from the original image:

$$x_{\text{final}}(i,j) = \begin{cases} x_w(i,j) & \text{if } j < \Delta x \\ x_{\text{shifted}}(i,j) & \text{otherwise} \end{cases}$$

This simple yet effective approach removes TreeRing without requiring training or additional models, making it computationally efficient. Quantitative results are shown in Table 6.

3.2 Black-Box Track

In the black-box track, no prior information was provided about the watermarking algorithms. We adopted a data-driven strategy to infer potential watermarking mechanisms by systematically analyzing signatures in the spatial and frequency domains. Since different watermarking schemes inherently leave characteristic artifact patterns, either as spatial distortions or as structured frequency components, these observations directly guided our clustering process. Through comprehensive analysis, we partitioned the 300 watermarked images into four distinct clusters, as shown in Figure 5: **Cluster-1**: no discernible artifacts, **Cluster-2**: border-like artifacts in the spatial domain, **Cluster-3**: circular patterns in the Fourier magnitude spectrum, **Cluster-4**: square patterns in the Fourier magnitude spectrum. This clustering allowed us to develop targeted removal strategies tailored to each artifact.

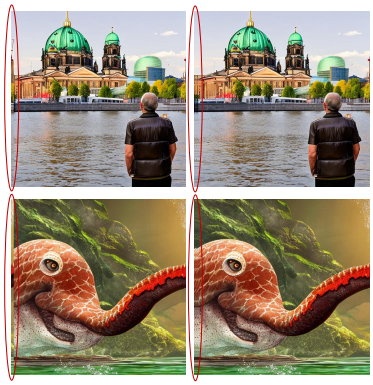

Figure 4: Effect of spatial translation on TreeRing watermarks. ***Left***: Direct 7-pixel shift removes watermark but introduces boundary artifacts (highlighted in red). ***Right***: Restoring the leftmost 7 columns from the original image removes boundary artifacts while preserving quality.

3.2.1 Image-to-Image Diffusion Models for Watermark Removal

We leveraged the **image-to-image** capabilities of the *Stable Diffusion Refiner model* [36] as a core component of our black-box watermark removal strategy. This approach exploits diffusion models' learned prior over natural images to project watermarked inputs onto the clean image manifold, effectively suppressing watermark signals while preserving semantic content. The process comprises two stages: (**i**) *forward diffusion*, which progressively corrupts an image with Gaussian noise, and (**ii**) *reverse diffusion*, where a denoising network reconstructs the image by iteratively removing noise.

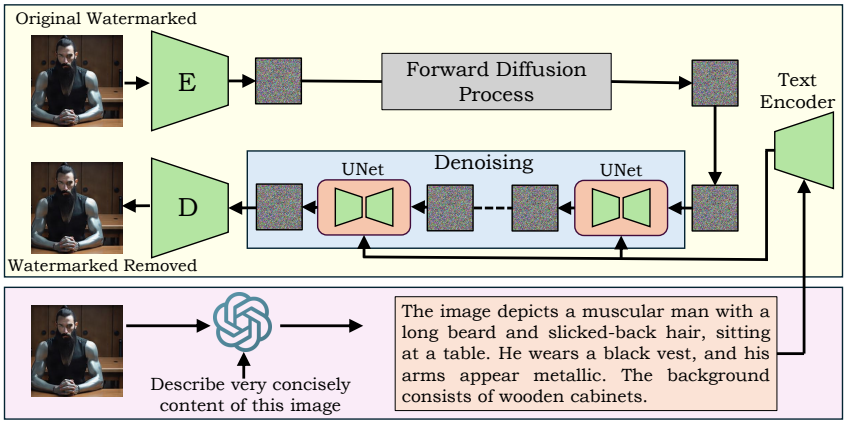

Figure 6: **Overview of our image-to-image diffusion pipeline for black-box watermark removal**. From a watermarked input, we add Gaussian noise to obtain a noisy latent, then apply reverse diffusion with a pretrained UNet conditioned on ChatGPT-generated captions. The choice of strength s controls the trade-off between artifact suppression and content preservation.

Given a watermarked image x_w, the forward process adds noise according to schedule $\{\alpha_t\}_{t=1}^{T}$, where strength parameter $s \in [0, 1]$ controls the initial corruption level via $\alpha_1 = 1 - s$ as as $x_t = \sqrt{\alpha_t} x_w + \sqrt{1 - \alpha_t}\epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, I)$. Here, the parameter s critically balances watermark removal efficacy against content preservation: lower values ($s \approx 0.04$) maintain fine details but may insufficiently suppress watermarks, while higher values ($s \approx 0.25$) remove watermarks more aggressively at the risk of semantic drift (see Figure 6). Subsequently, the *reverse diffusion process* iteratively removes noise through a learned denoising function ϵ_θ, reconstructing the final refined image as $x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \sqrt{1 - \alpha_t}\epsilon_\theta(x_t, t)\right)$.

We configured the diffusion process with $T = 500$ inference steps to enable sufficient iterative refinement for thorough watermark suppression. To preserve semantic content, we further introduced **semantic guidance** via ChatGPT-4-generated captions [2] that provide accurate, content-specific descriptions of each image (see Figure 20). These captions anchor the denoising trajectory, ensuring that reconstructed regions align with natural image statistics rather than residual watermark patterns. The classifier-free guidance scale was fixed at $w = 1.0$ to preserve photorealism; larger values ($w > 2$) strengthen text conditioning but introduce stylization artifacts [53]. The strength parameter s was tuned per cluster to balance watermark removal with perceptual fidelity.

3.3 Cluster-Specific Solutions

Our empirical clustering analysis revealed distinct artifact patterns, enabling the design of **adaptive, cluster-tailored watermark removal strategies** rather than relying on a single, global configuration. This specialization significantly improved both perceptual quality and watermark suppression efficacy.

Cluster 1: No discernible artifacts. In the absence of visible spatial or frequency-domain cues, we adopted a high-strength *image-to-image diffusion* configuration ($s = 0.16$). This choice prioritizes aggressive denoising to suppress potential low-energy watermark embeddings, while relying on semantic guidance to recover fine details without introducing perceptual distortions.

Cluster 2: Boundary artifacts. For images exhibiting border-like spatial patterns, we employed a **three-stage removal pipeline**, as in the Beige-box track (detailed in Sec. 3.1.1).

Cluster 3: Circular patterns in the Fourier spectrum. This cluster was treated using the same three-stage beige-box track pipeline as **Cluster 2**.

Cluster 4: Square patterns in the Fourier spectrum. We adopted a **hybrid approach** combining low-strength ($s = 0.04$) *image-to-image diffusion* to minimally perturb fine details, followed by a 7-pixel horizontal translation with selective restoration, similar to the TreeRing watermark removal method (detailed in Sec. 3.1.2).

Across clusters, hyperparameters (especially s) were tuned to balance **removal aggressiveness vs. content fidelity**. This cluster-specific methodology outperformed uniform parameter settings, yielding notable improvements in both qualitative assessments and quantitative metrics. Details for Black-box approach are provided in Algorithm 3.

Figure 7: Qualitative results of ControlNet-based watermark removal. ***Top***: original watermarked images; ***Bottom***: ControlNet outputs generated from Canny edges with color-aware captions generated via ChatGPT.

4 Exploratory but Ineffective Approaches

In parallel to developing our primary solution, we systematically explored multiple alternative strategies for watermark removal. While each was grounded in plausible hypotheses, none yielded competitive results in terms of the removal–quality trade-off. Below, we summarize these unsuccessful attempts along with representative qualitative outcomes.

ControlNet-Based Edge Guidance. We extracted Canny edges from watermarked images and used them as structural guidance in ControlNet [64], while ChatGPT-generated captions (explicitly encoding content and color attributes, *e.g.*, "a red vintage car parked in front of a rustic cabin in snow") provided semantic conditioning. This produced visually coherent outputs (Figure 7) but retained most watermark signals, consistent with reports that strong structural conditioning preserves pixel-level watermark components [34].

Aggressive Diffusion Rinsing. We applied repeated forward–reverse diffusion cycles in latent space, in order to 'rinse' the watermark by destroying and reconstructing content. Four-pass rinsing removed most traces but caused severe degradation: dirty textures and loss of detail (Figure 8, top). The pixel-space diffusion-model variants also offered no improvement.

White-Box Overlays. Early experiments added white-box perturbations directly to inputs to disrupt watermark patterns. This failed to erase watermarks and degradation in quality, introducing unnatural textures, as shown in middle row of Figure 8.

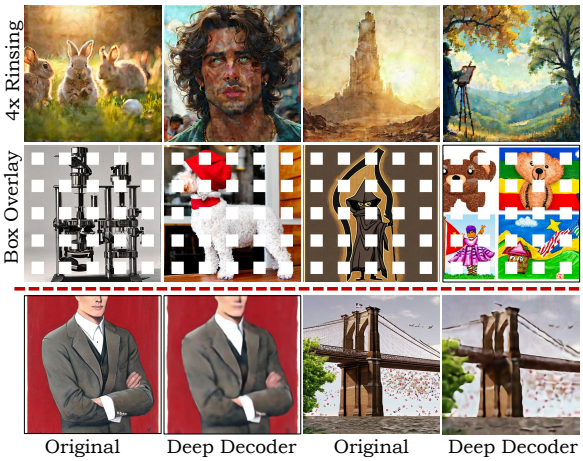

Figure 8: Qualitative examples of failed watermark removal strategies. ***Row 1***: aggressive four-pass diffusion rinsing is effective at watermark removal but degrade quality. ***Row 2***: white-box additive perturbations results in limited removal and noticeable texture artifacts. ***Row 3***: untrained neural network reconstruction removes most watermarks but introduces oversmoothing and color shifts.

Untrained Neural Network Reconstruction. Inspired by natural priors in untrained networks [56], we used a convolutional decoder to reconstruct images from scratch. Although watermarks were attenuated, artifacts, oversmoothing, and color distortions were evident, with slow convergence limiting practicality (Figure 8, bottom).

Other Attempts. We explored several other methods, including adversarial attacks on ensembles of latent-space encoders [10], super-resolution and inpainting perturbations, color-space and frequency-domain manipulations, and hybrid denoising–sharpening pipelines. All of these approaches degraded perceptual quality while leaving watermark traces partially intact.

5 Conclusion

In conclusion, our work provides a comprehensive examination of the vulnerabilities in current invisible watermarking schemes for images, demonstrating that even under restrictive threat models,

including the challenging black-box setting, determined attackers can effectively strip watermarks while preserving perceptual quality. Our results, which topped *Beige-box and Black-box NeurIPS 2024 competition tracks*, highlight not only the feasibility but also the generality of such removal strategies across fundamentally different watermark methods. This calls for a rethinking of watermark robustness, emphasizing the urgent need for schemes that can withstand adaptive, high-capacity generative manipulation without sacrificing utility, thereby ensuring reliable provenance tracking in the era of powerful image synthesis models.

References

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

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

[4] Bang An, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, et al. Waves: Benchmarking the robustness of image watermarks. In *Forty-first International Conference on Machine Learning*, 2024.

[5] Toluwani Aremu, Noor Hussein, Munachiso Nwadike, Samuele Poppi, Jie Zhang, Karthik Nandakumar, Neil Gong, and Nils Lukas. Mitigating watermark stealing attacks in generative models via multi-key watermarking. *arXiv preprint arXiv:2507.07871*, 2025.

[6] Muhammad Asim, Fahad Shamshad, and Ali Ahmed. Blind image deconvolution using deep generative priors. *IEEE Transactions on Computational Imaging*, 6:1493–1506, 2020.

[7] Adrian G Bors and Ioannis Pitas. Image watermarking using dct domain constraints. In *Proceedings of 3rd IEEE International Conference on Image Processing*, volume 3, pp. 231–234. IEEE, 1996.

[8] Yunzhuo Chen, Jordan Vice, Naveed Akhtar, Nur Al Hasan Haldar, and Ajmal Mian. Image watermarking of generative diffusion models. *arXiv preprint arXiv:2502.10465*, 2025.

[9] Abdulrahman Diaa, Toluwani Aremu, and Nils Lukas. Optimizing adaptive attacks against content watermarks for language models. *arXiv preprint arXiv:2410.02440*, 2024.

[10] Mucong Ding, Tahseen Rabbani, Bang An, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, et al. Waves: Benchmarking the robustness of image watermarks. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*.

[11] Mucong Ding, Tahseen Rabbani, Bang An, Souradip Chakraborty, Chenghao Deng, Mehrdad Saberi, Yuxin Wen, Xuandong Zhao, Mo Zhou, Anirudh Satheesh, et al. Erasing the invisible: A stress-test challenge for image watermarks. In *NeurIPS 2024 Competition Track*, 2024.

[12] Junxian Duan, Jiyang Guan, Wenkui Yang, and Ran He. Visual watermarking in the era of diffusion models: Advances and challenges. *arXiv preprint arXiv:2505.08197*, 2025.

[13] Jaiden Fairoze, Guillermo Ortiz-JimÃŠnez, Mel Vecerik, Somesh Jha, and Sven Gowal. On the difficulty of constructing a robust and publicly-detectable watermark. *arXiv preprint arXiv:2502.04901*, 2025.

[14] Hany Farid. Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety*, 1(4), 2022.

[15] Grzegorz Głuch, Berkant Turan, Sai Ganesh Nagarajan, and Sebastian Pokutta. The good, the bad and the ugly: Watermarks, transferable attacks and adversarial defenses. *arXiv preprint arXiv:2410.08864*, 2024.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[17] Gustavosta. Stable diffusion prompts dataset. https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts, 2024. Accessed: February 2025.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[19] Sandra Höltervennhoff, Jonas Ricker, Maike M Raphael, Charlotte Schwedes, Rebecca Weil, Asja Fischer, Thorsten Holz, Lea Schönherr, and Sascha Fahl. Security benefits and side effects of labeling ai-generated images. *arXiv preprint arXiv:2505.22845*, 2025.

[20] Khalid M Hosny, Amal Magdi, Osama ElKomy, and Hanaa M Hamza. Digital image watermarking using deep learning: A survey. *Computer Science Review*, 53:100662, 2024.

[21] Huayang Huang, Yu Wu, and Qian Wang. Robin: Robust and invisible watermarks for diffusion models with adversarial optimization. *Advances in Neural Information Processing Systems*, 37: 3937–3963, 2024.

[22] Dongjun Hwang, Sungwon Woo, Tom Gao, Raymond Luo, and Sunghwan Baek. Invisible watermarks: Attacks and robustness. *arXiv preprint arXiv:2412.12511*, 2024.

[23] Anubhav Jain, Yuya Kobayashi, Naoki Murata, Yuhta Takida, Takashi Shibuya, Yuki Mitsufuji, Niv Cohen, Nasir Memon, and Julian Togelius. Forging and removing latent-noise diffusion watermarks using a single image. *arXiv preprint arXiv:2504.20111*, 2025.

[24] Xiangui Kang, Jiwu Huang, Yun Q Shi, and Yan Lin. A dwt-dft composite watermarking scheme robust to both affine transform and jpeg compression. *IEEE transactions on circuits and systems for video technology*, 13(8):776–786, 2003.

[25] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[26] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*, 2022.

[27] Hengyue Liang, Taihui Li, and Ju Sun. A baseline method for removing invisible image watermarks using deep image prior. *arXiv preprint arXiv:2502.13998*, 2025.

[28] Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. A survey of text watermarking in the era of large language models. *ACM Computing Surveys*, 57(2):1–36, 2024.

[29] Yepeng Liu, Yiren Song, Hai Ci, Yu Zhang, Haofan Wang, Mike Zheng Shou, and Yuheng Bu. Image watermarks are removable using controllable regeneration from clean noise. *arXiv preprint arXiv:2410.05470*, 2024.

[30] Yepeng Liu, Yiren Song, Hai Ci, Yu Zhang, Haofan Wang, Mike Zheng Shou, and Yuheng Bu. Image watermarks are removable using controllable regeneration from clean noise. *arXiv preprint arXiv:2410.05470*, 2024.

[31] Nils Lukas. *Analyzing Threats of Large-Scale Machine Learning Systems*. PhD thesis, University of Waterloo, 2024.

[32] Nils Lukas and Florian Kerschbaum. {PTW}: Pivotal tuning watermarking for {Pre-Trained} image generators. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 2241–2258, 2023.

[33] Nils Lukas, Abdulrahman Diaa, Lucas Fenaux, and Florian Kerschbaum. Leveraging optimization for adaptive attacks on image watermarks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=O9PArxKLe1.

[34] Denis Lukovnikov, Andreas Müller, Jonas Thietke, Erwin Quiring, and Asja Fischer. Are semantic watermarks for diffusion models resilient to layout control? In *The 1st Workshop on GenAI Watermarking*.

[35] Xingjun Ma, Yifeng Gao, Yixu Wang, Ruofan Wang, Xin Wang, Ye Sun, Yifan Ding, Hengyuan Xu, Yunhao Chen, Yunhan Zhao, et al. Safety at scale: A comprehensive survey of large model safety. *arXiv preprint arXiv:2502.05206*, 2025.

[36] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[37] Andreas Müller, Denis Lukovnikov, Jonas Thietke, Asja Fischer, and Erwin Quiring. Black-box forgery attacks on semantic watermarks for diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20937–20946, 2025.

[38] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

[39] Vidyasagar M Potdar, Song Han, and Elizabeth Chang. A survey of digital image watermarking techniques. In *INDIN'05. 2005 3rd IEEE International Conference on Industrial Informatics, 2005.*, pp. 709–716. IEEE, 2005.

[40] Adnan Qayyum, Waqas Sultani, Fahad Shamshad, Junaid Qadir, and Rashid Tufail. Single-shot retinal image enhancement using deep image priors. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 636–646. Springer, 2020.

[41] Adnan Qayyum, Inaam Ilahi, Fahad Shamshad, Farid Boussaid, Mohammed Bennamoun, and Junaid Qadir. Untrained neural network priors for inverse imaging problems: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6511–6536, 2022.

[42] Wang Qi, Bei Yue, Chen Wangdu, Pan Xinghao, Cheng Zhipeng, Wang Shaokang, Wang Yizhao, and Wang Chenwei. An overview on digital content watermarking. In *Signal and Information Processing, Networking and Computers: Proceedings of the 8th International Conference on Signal and Information Processing, Networking and Computers (ICSINC)*, pp. 1311–1318. Springer, 2022.

[43] Kui Ren, Ziqi Yang, Li Lu, Jian Liu, Yiming Li, Jie Wan, Xiaodi Zhao, Xianheng Feng, and Shuo Shao. Sok: On the role and future of aigc watermarking in the era of gen-ai. *arXiv preprint arXiv:2411.11478*, 2024.

[44] Vinu Sankar Sadasivan, Mehrdad Saberi, and Soheil Feizi. Iconmark: Robust interpretable concept-based watermark for ai images. *arXiv preprint arXiv:2507.13407*, 2025.

[45] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605*, 2018.

[46] Tanja Šarčević, Alicja Karlowicz, Rudolf Mayer, Ricardo Baeza-Yates, and Andreas Rauber. U can't gen this? a survey of intellectual property protection methods for data in generative ai. *arXiv preprint arXiv:2406.15386*, 2024.

[47] Fahad Shamshad and Ali Ahmed. Compressed sensing-based robust phase retrieval via deep generative priors. *IEEE Sensors Journal*, 21(2):2286–2298, 2020.

[48] Fahad Shamshad, Farwa Abbas, and Ali Ahmed. Deep ptych: Subsampled fourier ptychography using generative priors. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7720–7724. IEEE, 2019.

[49] Fahad Shamshad, Asif Hanif, Farwa Abbas, Muhammad Awais, and Ali Ahmed. Adaptive ptych: Leveraging image adaptive generative priors for subsampled fourier ptychography. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

[50] Fahad Shamshad, Asif Hanif, and Ali Ahmed. Subsampled fourier ptychography via pretrained invertible and untrained network priors. In *NeurIPS 2019 Workshop on Solving Inverse Problems with Deep Networks*, 2019.

[51] Fahad Shamshad, Muzammal Naseer, and Karthik Nandakumar. Clip2protect: Protecting facial privacy using text-guided makeup via adversarial latent search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20595–20605, 2023.

[52] Fahad Shamshad, Koushik Srivatsan, and Karthik Nandakumar. Evading forensic classifiers with attribute-conditioned adversarial faces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16469–16478, 2023.

[53] Kaiyu Song and Hanjiang Lai. Rethinking oversaturation in classifier-free guidance via low frequency. *arXiv preprint arXiv:2506.21452*, 2025.

[54] Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2117–2126, 2020.

[55] Janvi Thakkar, Giulio Zizzo, and Sergio Maffeis. Elevating defenses: Bridging adversarial training and watermarking for model resilience. *arXiv preprint arXiv:2312.14260*, 2023.

[56] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9446–9454, 2018.

[57] Tao Wang, Yushu Zhang, Shuren Qi, Ruoyu Zhao, Zhihua Xia, and Jian Weng. Security and privacy on generative data in aigc: A survey. *ACM Computing Surveys*, 57(4):1–34, 2024.

[58] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612, 2004.

[59] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. *Advances in Neural Information Processing Systems*, 36, 2024.

[60] Naen Xu, Changjiang Li, Tianyu Du, Minxi Li, Wenjie Luo, Jiacheng Liang, Yuyuan Li, Xuhong Zhang, Meng Han, Jianwei Yin, et al. Copyrightmeter: Revisiting copyright protection in text-to-image models. *arXiv preprint arXiv:2411.13144*, 2024.

[61] Pei Yang, Hai Ci, Yiren Song, and Mike Zheng Shou. Can simple averaging defeat modern watermarks? *Advances in Neural Information Processing Systems*, 37:56644–56673, 2024.

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

[63] Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph Stoecklin, Heqing Huang, and Ian Molloy. Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the 2018 on Asia conference on computer and communications security*, pp. 159–172, 2018.

[64] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

[65] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

[66] Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramer, et al. Sok: Watermarking for ai-generated content. *arXiv preprint arXiv:2411.18479*, 2024.

[67] 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. *Advances in neural information processing systems*, 37:8643–8672, 2024.

[68] Zhizhen Zhao, Jong Chul Ye, and Yoram Bresler. Generative models for inverse imaging problems: From mathematical foundations to physics-driven applications. *IEEE Signal Processing Magazine*, 40(1):148–163, 2023.

Appendix

This appendix provides background information and extensive qualitative results. For clarity, we summarize the contents and their page numbers below.

Background

5.1 StegaStamp: Invisible Hyperlinks in Photographs

StegaStamp [54] introduced one of the first end-to-end learned approaches for embedding invisible information into images that remain robust under real-world conditions. The key idea is to train an encoder–decoder system jointly while simulating the distortions introduced by printing and re-capturing photographs, such as perspective changes, blur, noise, color variation, and compression. The encoder embeds a short bitstring into the image with minimal perceptual difference, while the decoder is trained to recover the message reliably after these transformations. To further improve resilience, error-correcting codes are used to ensure accurate recovery even when parts of the image are degraded. Experiments demonstrated that StegaStamp achieves high decoding accuracy across different combinations of printers, displays, and cameras, while maintaining near-identical image quality. This work showed that imperceptibility and robustness can be achieved together, moving beyond earlier digital-only steganography methods and establishing a framework for watermarking that extends into practical physical-world settings.

In the context of the NeurIPS watermarking competition, StegaStamp is particularly relevant because the Beige-box track uses a variant of its algorithm to generate watermarked images. Understanding its design principles and robustness objectives provides a clear basis for developing effective removal strategies, since the competition task required eliminating such watermarks while maintaining image quality.

5.2 Tree-Ring Watermarks: Fingerprints for Diffusion Images

Tree-Ring watermarks [59] addressed the challenge of watermarking in the context of diffusion-based generative models. The approach embeds circular "tree-ring" patterns in the Fourier phase of generated images, producing signals that are imperceptible in the spatial domain but remain stable under common transformations such as resizing, compression, cropping, and mild editing. By embedding during the image generation process itself, rather than applying marks after generation, the method ensures that the watermark is more deeply integrated and harder to remove. Detection is carried out in the frequency domain using statistical classifiers, enabling reliable identification of watermarked content at scale. This design makes tree-ring watermarks more robust than many prior spatial-domain techniques and provides a practical mechanism for verifying the provenance of AI-generated content. The work highlights the importance of designing watermarking strategies that align with the properties of modern generative models, establishing a foundation for authenticity and accountability in synthetic media.

For the competition, Tree-Ring Watermarks formed the basis of another watermarking method in the beigebox track. Their reliance on structured frequency-domain perturbations made removal more challenging, requiring participants to design attacks that disrupt the Fourier-phase patterns while preserving spatial fidelity. Understanding the strengths of tree-rings thus directly shaped the strategies needed for effective watermark removal.

6 Additional Experiments

To further understand the trade-off between watermark removal strength and perceptual quality, we conducted a set of controlled experiments across different diffusion models and sampling configurations. These results, while not part of the main leaderboard evaluations, provide quantitative support for the hyperparameter choices.

Quality Metrics Trends Across Tracks and Clusters. Figure 9 evaluates our optimized pipeline across both beige-box tracks and four black-box clusters. The results highlight distinct quality–removal trade-offs. *Beige-box StegaStamp* achieves the most balanced outcome (PSNR: 28.06 ± 1.2, SSIM: 0.82 ± 0.04), with high semantic fidelity (CLIP-FID < 5) and strong pixel-wise alignment—attributable in part to our CIELAB-based post-processing stage, which restores fine-scale color consistency after watermark removal. In contrast, *TreeRing* and *Black-box Cluster with square Fourier artifacts* show markedly lower PSNR and SSIM, a direct consequence of the 7-pixel horizontal translation that disrupts watermark embeddings; despite this, semantic similarity remains high (CLIP-FID < 5). Finally, *Black-box Cluster with no artifacts* exhibits the largest perceptual gap, indicating

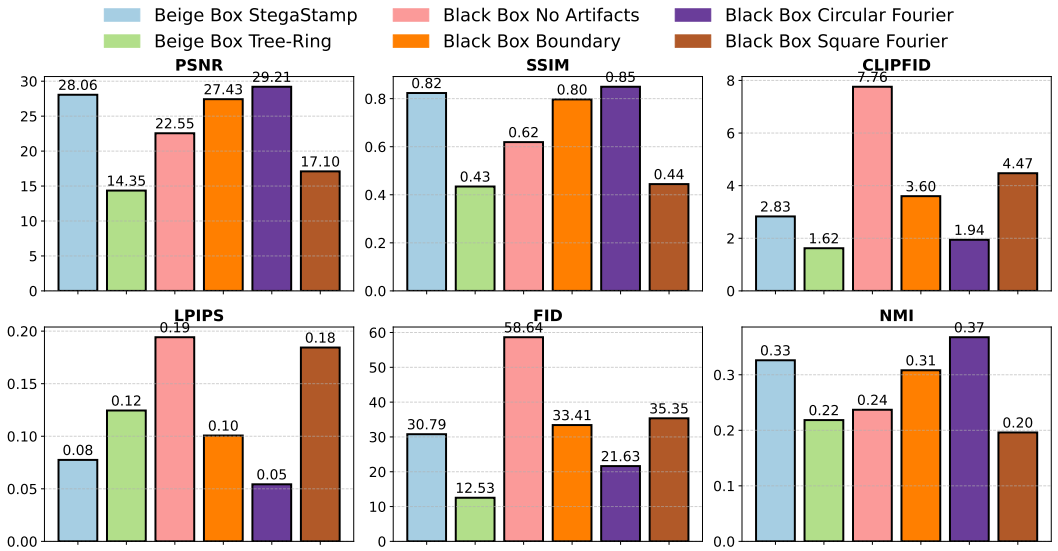

Figure 9: **Performance of our pipeline across beige-box and black-box tracks**. StegaStamp achieves the most balanced quality–removal trade-off, aided by CIELAB-based post-processing, while TreeRing and the black-box cluster with Fourier square patterns show reduced score for alignment metrics (PSNR, SSIM) due to the intentional translation shift.

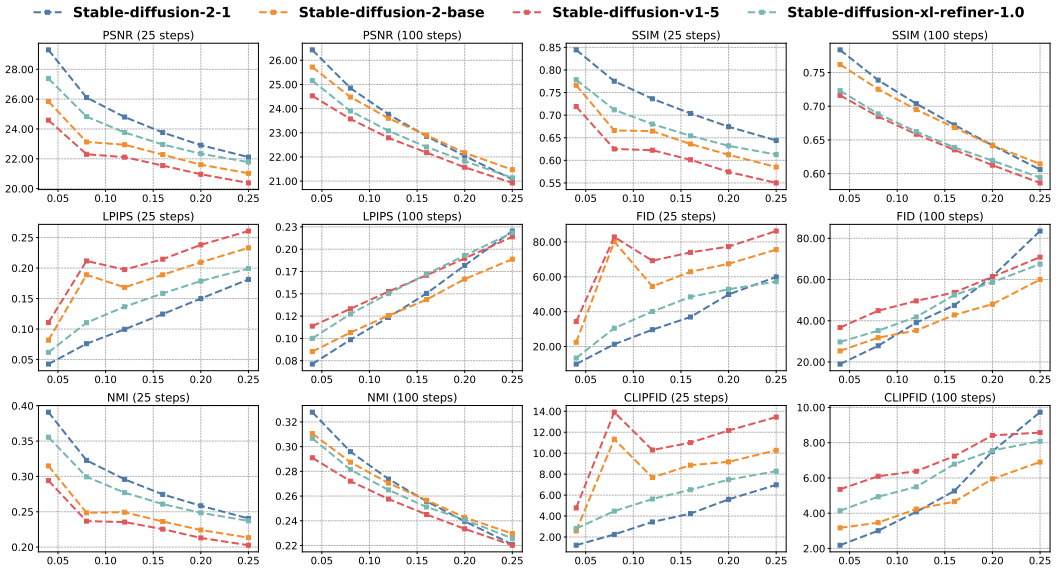

Figure 10: **Effect of strength s and steps T on watermark removal**. PSNR drops sharply at $s \approx 0.10$, motivating cluster-specific settings. From leaderboard results, we observe that `stable-diffusion-xl-refiner-1.0` consistently outperforms other diffusion models in terms of watermark removal.

that its imperceptible watermark necessitated more aggressive interventions, with some loss of fine-grained detail.

Impact of Diffusion Strength and Model Choice.

Figure 10 presents evaluation of four Stable Diffusion variants across strength parameters $s \in [0.04, 0.25]$ and inference steps $T \in \{25, 100\}$. Quality degradation exhibits a sharp nonlinearity at $s \approx 0.10$, where PSNR drops from 26 to 22 dB, motivating our conservative $s = 0.04$ for structured artifacts versus aggressive $s = 0.16$ for hidden watermarks. Similarly, extended denoising (100 vs. 25 steps) yields marginal improvements (~ 0.5 dB PSNR, 0.02 LPIPS). In addition, from

Table 5: **Cluster assignments for all 300 black-box images**, grouped by manually identified artifact type: boundary artifacts, circular or square Fourier magnitude patterns, or no visible artifacts. Each block lists the image indices belonging to that cluster, enabling targeted strategy design and facilitating quantitative comparisons of removal performance.

Cluster 1 (No Artifacts)
0, 2, 3, 4, 10, 13, 14, 16, 19, 21, 24, 29, 30, 33, 36, 40, 41, 43, 50, 51, 56, 58, 60, 61, 62, 67, 71, 75, 77, 80, 83, 86, 93, 94, 95, 96, 97, 98, 104, 107, 109, 110, 112, 113, 114, 123, 125, 126, 130, 137, 138, 146, 149, 154, 155, 157, 159, 161, 162, 166, 167, 171, 175, 178, 179, 187, 196, 198, 199, 200, 201, 205, 209, 210, 215, 216, 217, 218, 220, 226, 227, 233, 235, 241, 245, 247, 249, 250, 254, 260, 262, 266, 268, 275, 279, 280, 281, 288, 291, 292, 294, 296
Cluster 2 (Boundary Artifacts)
1, 6, 11, 23, 35, 42, 48, 49, 57, 65, 68, 69, 78, 82, 84, 85, 91, 100, 102, 105, 119, 120, 121, 131, 132, 140, 141, 142, 143, 148, 169, 170, 180, 186, 203, 204, 207, 231, 240, 246, 256, 257, 261, 267, 269, 274, 283, 297, 298
Cluster 3 (Circular Patterns in Fourier Magnitude)
9, 17, 25, 28, 31, 32, 38, 39, 47, 53, 70, 79, 87, 88, 99, 115, 117, 127, 128, 133, 147, 151, 163, 165, 172, 174, 177, 183, 185, 192, 193, 195, 197, 202, 211, 232, 242, 243, 244, 251, 258, 263, 265, 271, 276, 277, 278, 287, 289
Cluster 4 (Square Patterns in Fourier Magnitude)
5, 7, 8, 12, 15, 18, 20, 22, 26, 27, 34, 37, 44, 45, 46, 52, 54, 55, 59, 63, 64, 66, 72, 73, 74, 76, 81, 89, 90, 92, 101, 103, 106, 108, 111, 116, 118, 122, 124, 129, 134, 135, 136, 139, 144, 145, 150, 152, 153, 156, 158, 160, 164, 168, 173, 176, 181, 182, 184, 188, 189, 190, 191, 194, 206, 208, 212, 213, 214, 219, 221, 222, 223, 224, 225, 228, 229, 230, 234, 236, 237, 238, 239, 248, 252, 253, 255, 259, 264, 270, 272, 273, 282, 284, 285, 286, 290, 293, 295, 299

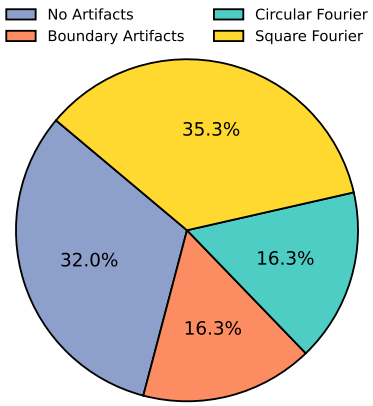

Figure 11: Images distribution for Black-box track clusters.

our leaderboard submissions we observe that `stable-diffusion-xl-refiner-1.0` is more effective at watermark removal than other tested diffusion models.

Algorithm 2 VAE Finetuning Pipeline

Require: Pretrained VAE (E_θ, D_ϕ), dataset $\{(x_w, x_i)\}_{n=1}^{N}$, learning rate α, batch size B, epochs E

Initialize $\psi \leftarrow \{\theta, \phi\}$ # Initialize encoder and decoder parameters
for epoch $= 1$ **to** E **do**
  Shuffle dataset
  for each batch $\{(x_w, x_i)\}_{k=1}^{B}$ **do**
    $z \leftarrow E_\theta(x_w)$ # Encode watermarked image
    $\hat{x} \leftarrow D_\phi(z)$ # Decode latent representation
    $L \leftarrow \frac{1}{B}\sum_{k=1}^{B} \|\hat{x}^{(k)} - x_i^{(k)}\|_2^2$ # MSE reconstruction loss
    Compute gradient $\nabla_\psi L$
    $\psi \leftarrow \psi - \alpha\, \nabla_\psi L$ # Gradient Update via Backpropagation
  end for
end for

Ensure: Finetuned VAE (E_θ^*, D_ϕ^*) # Return trained VAE parameters

Algorithm 3 Black–Box Track Pipeline

Require: Watermarked images $\{x_w\}$, manual cluster labels $\{c_x \in \{1, 2, 3, 4\}\}$

#Stage 1: Manual Clustering
  Assign each image $x \in \{x_w\}$ to cluster c_x by visual inspection:
    Cluster 1: Boundary artifacts **Cluster 2**: Circular Fourier patterns
    Cluster 3: Square Fourier patterns **Cluster 4**: No noticeable artifacts
#Stage 2: Cluster-Specific Solutions
for all x **in** $\{x_w\}$ **do**
  $k \leftarrow c_x$
  if $k = 1$ **then**
    # Diffusion: aggressive removal
    $x_r \leftarrow \text{Diffuse}(x,\ s = 0.16; \text{ChatGPT captions})$ # Image-to-image diffusion pipeline
  else if $k \in \{2, 3\}$ **then**
    # VAE Pipeline (Sec. 3.1.1)
    $x_r \leftarrow \text{VAE\_remove}(x)$ # Includes dataset generation, finetuning, color–contrast
  else if $k = 4$ **then**
    # Hybrid: mild diffusion + translation
    $x_d \leftarrow \text{Diffuse}(x,\ s = 0.04)$
    $x_s \leftarrow \text{Translate}(x_d,\ 7)$ # Image translation by 7 pixels
    Restore leftmost 7 columns from x into x_s
    $x_r \leftarrow x_s$
  end if
  save x_r # Refined output
end for

Ensure: Refined images $\{x_r\}$

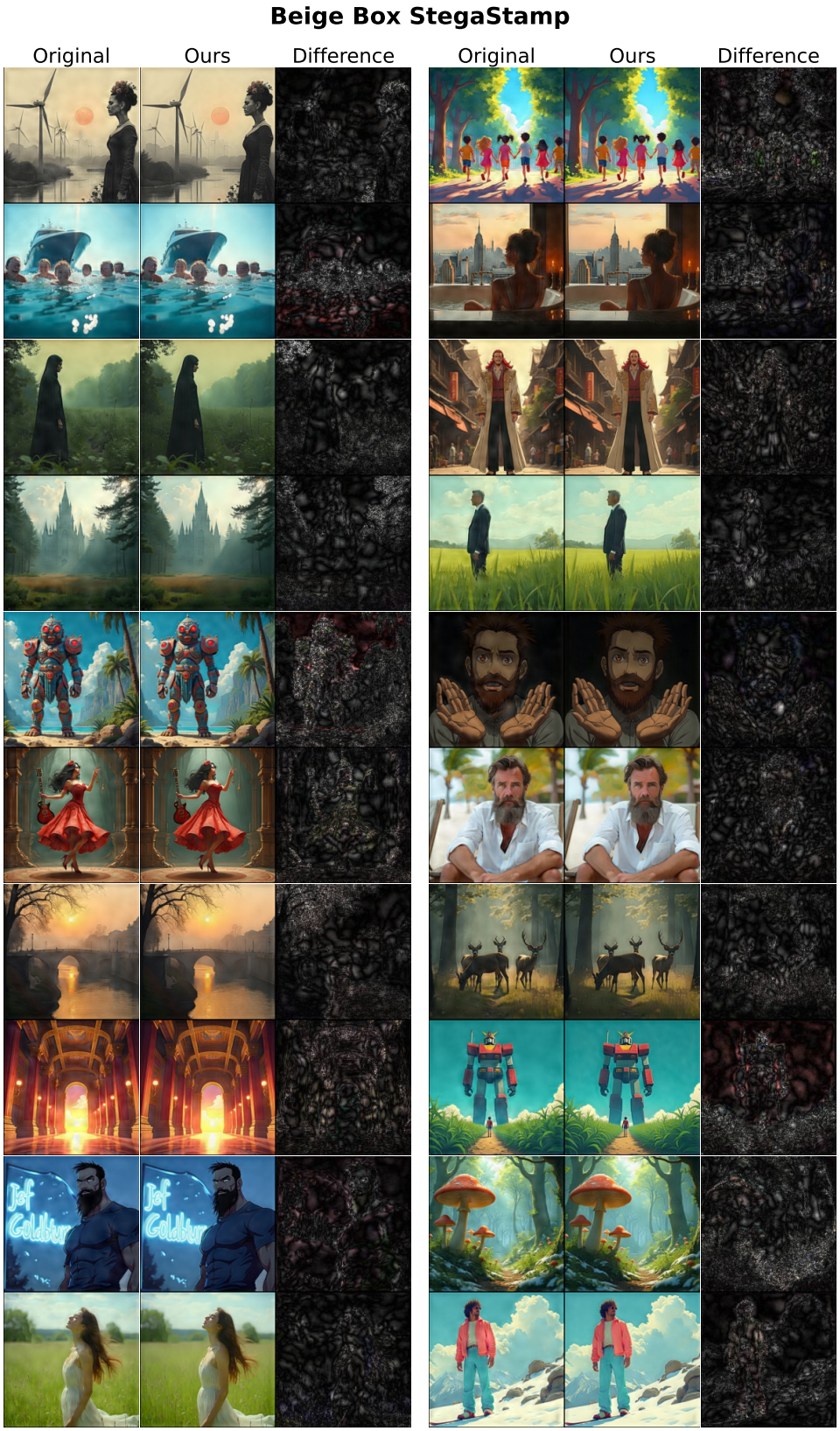

Figure 12: Qualitative results for watermark removal on StegaStamp images in the Beige-box track. Each triplet shows (left) the original watermarked input, (middle) the output after applying our removal method, and (right) the residual difference between them. Our approach preserves semantic content and visual fidelity while effectively eliminating the embedded watermark.

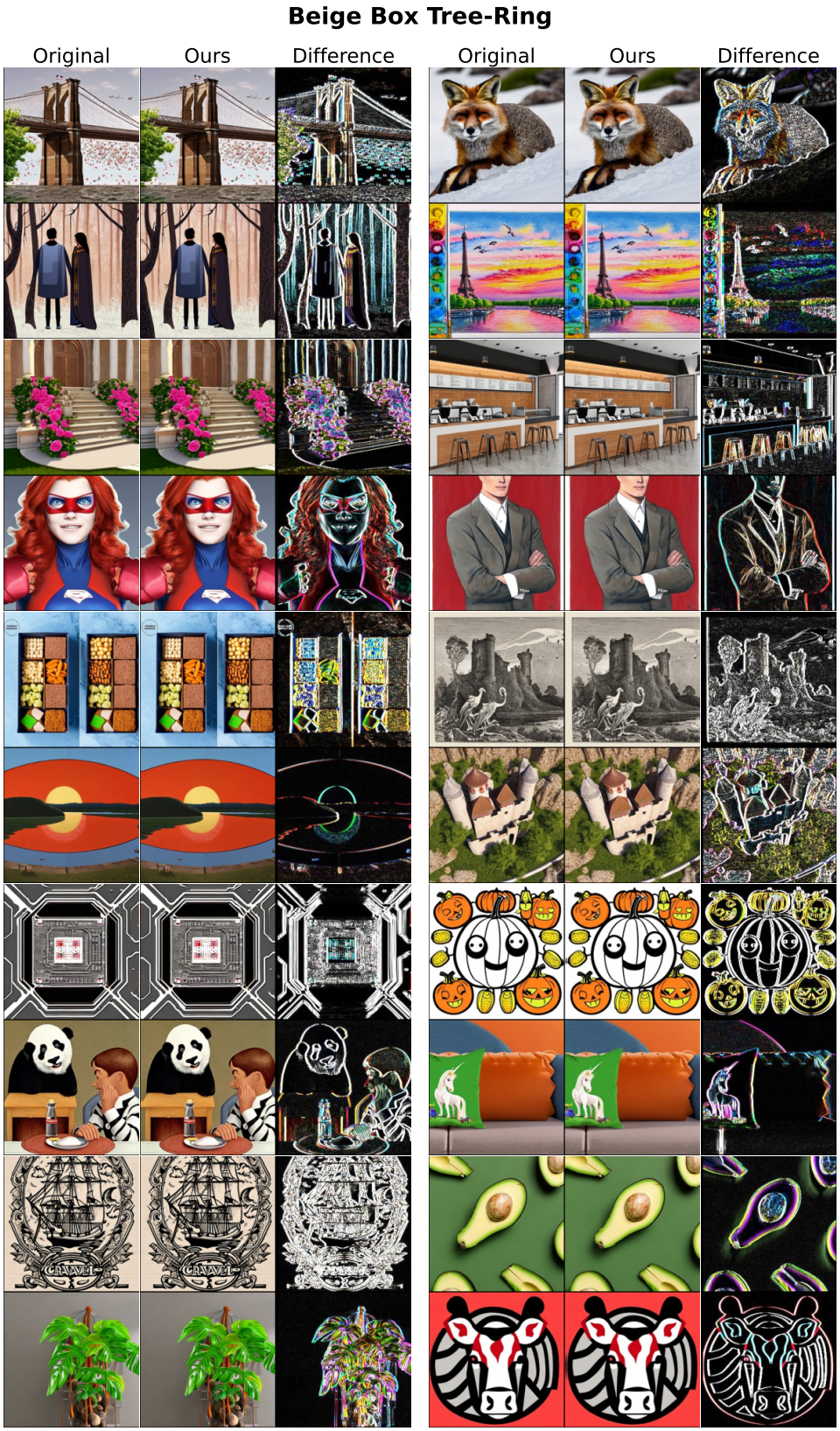

Figure 13: Qualitative results for watermark removal on TreeRings images in the Beige-box track. Each triplet shows (left) the original watermarked input, (middle) the output after applying our removal method, and (right) the residual difference between them. Our approach preserves semantic content and visual fidelity while effectively eliminating the embedded watermark.

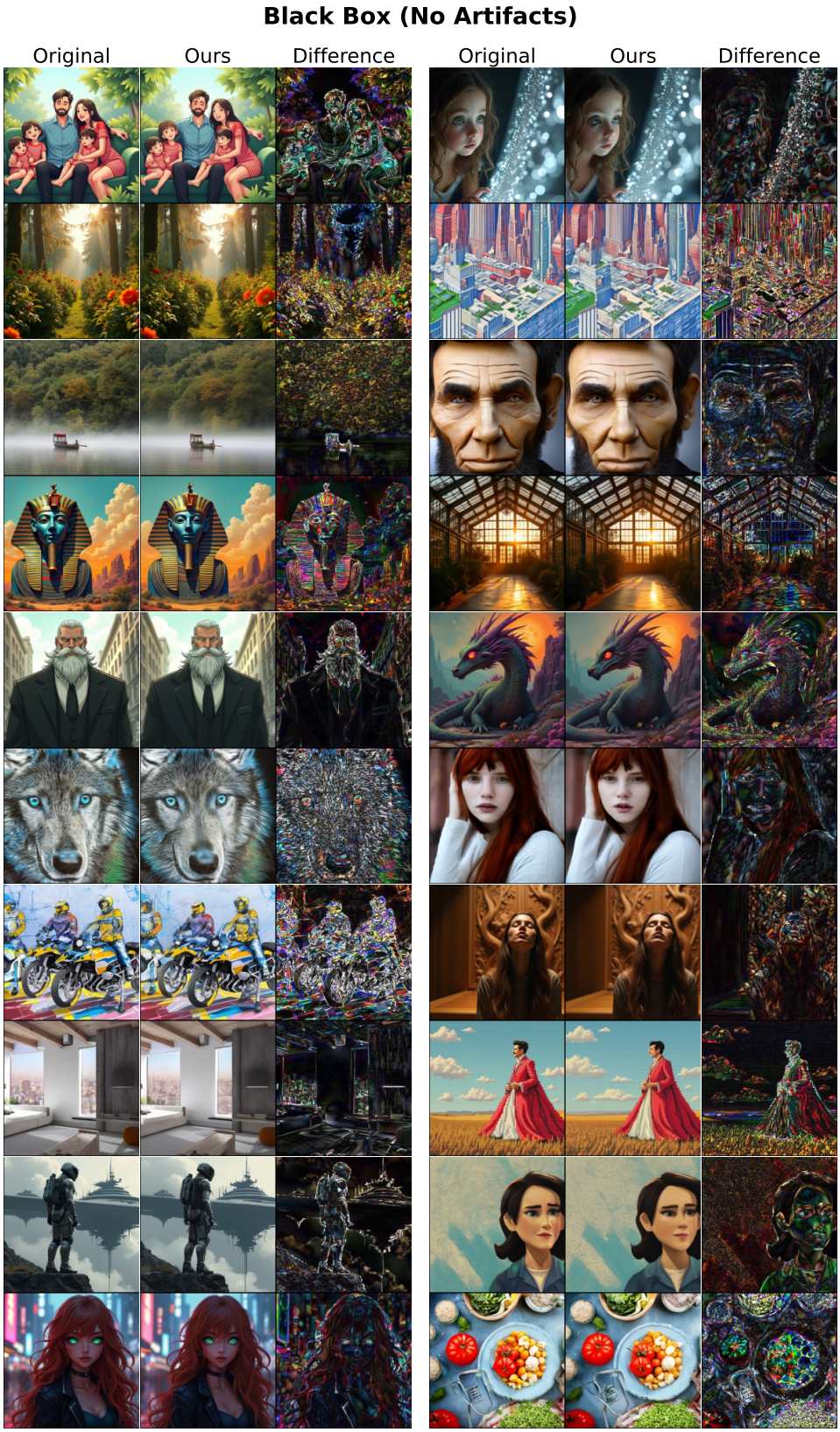

Figure 14: Qualitative results on the Black-box track for images without noticeable artifacts. Each triplet shows (left) the original watermarked input, (middle) the output after applying our removal method, and (right) the residual difference between them. Our approach preserves semantic content and visual fidelity while effectively eliminating the embedded watermark.

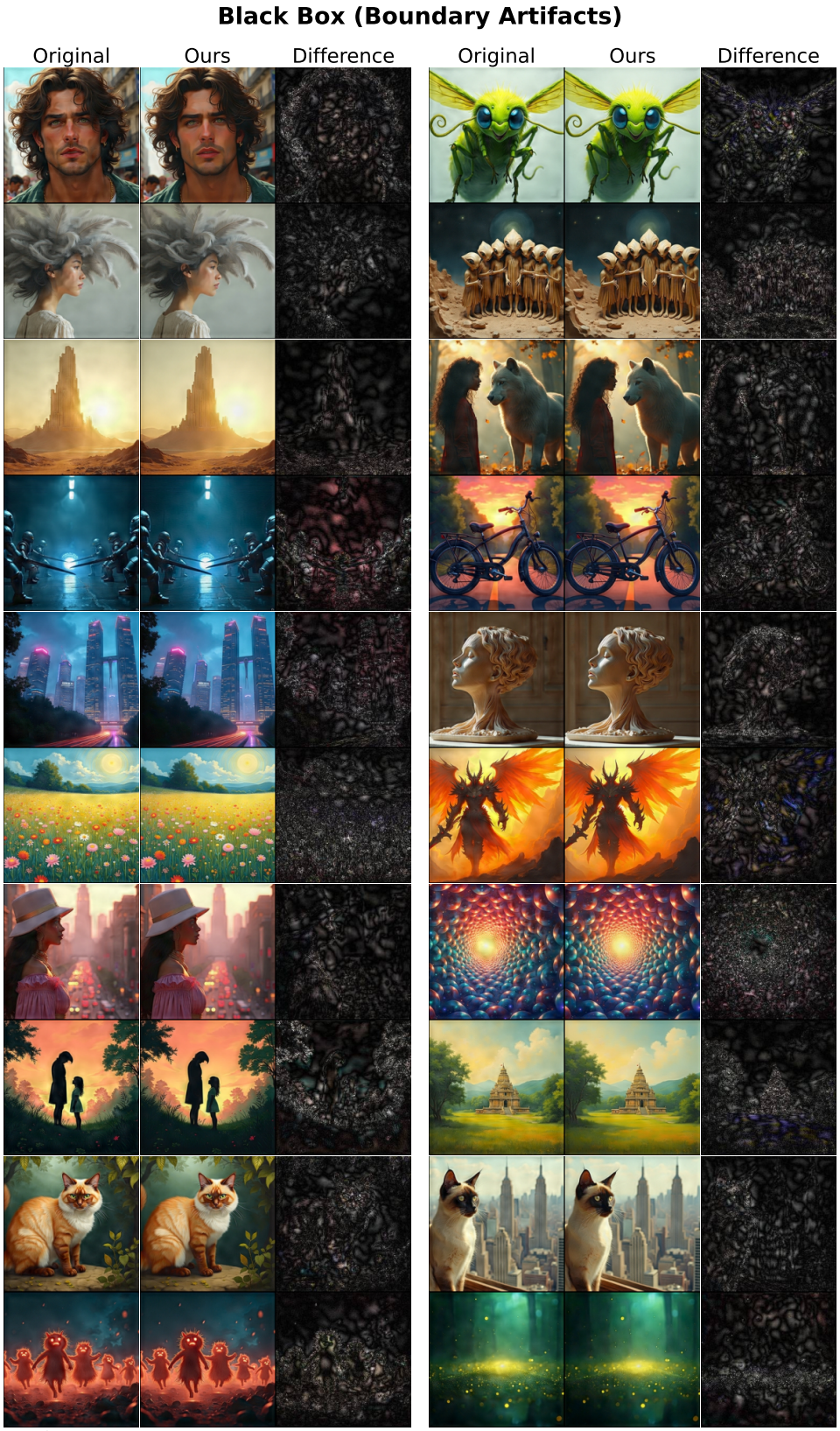

Figure 15: Qualitative results on the Black-box track for images with boundary artifacts. Each triplet shows (left) the original watermarked input, (middle) the output after applying our removal method, and (right) the residual difference between them. Our approach preserves semantic content and visual fidelity while effectively eliminating the embedded watermark.

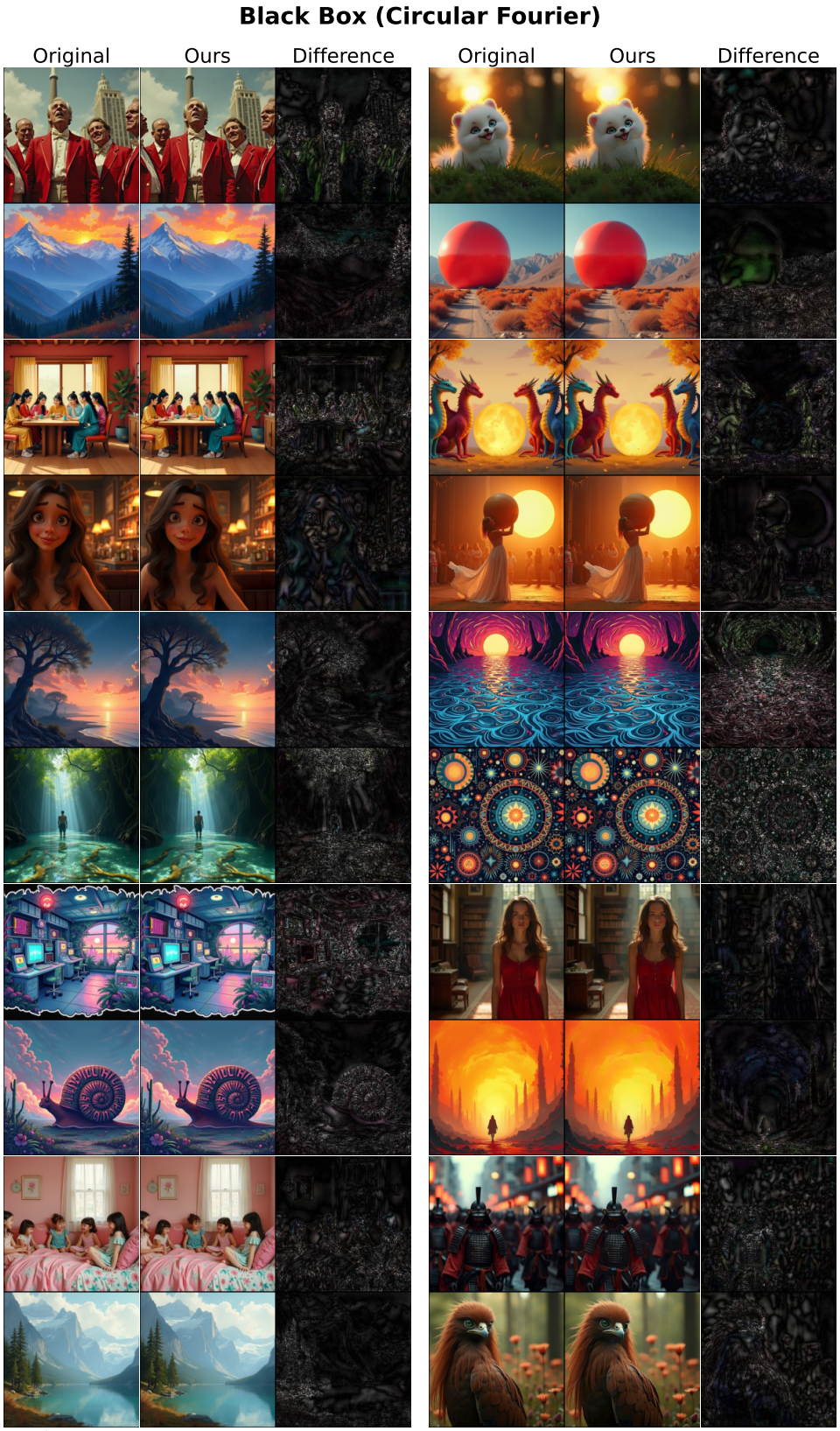

Figure 16: Qualitative results on the Black-box track for images containing circular Fourier-domain artifacts. Each triplet shows (left) the original watermarked input, (middle) the output after applying our removal method, and (right) the residual difference between them. Our approach preserves semantic content and visual fidelity while effectively eliminating the embedded watermark.

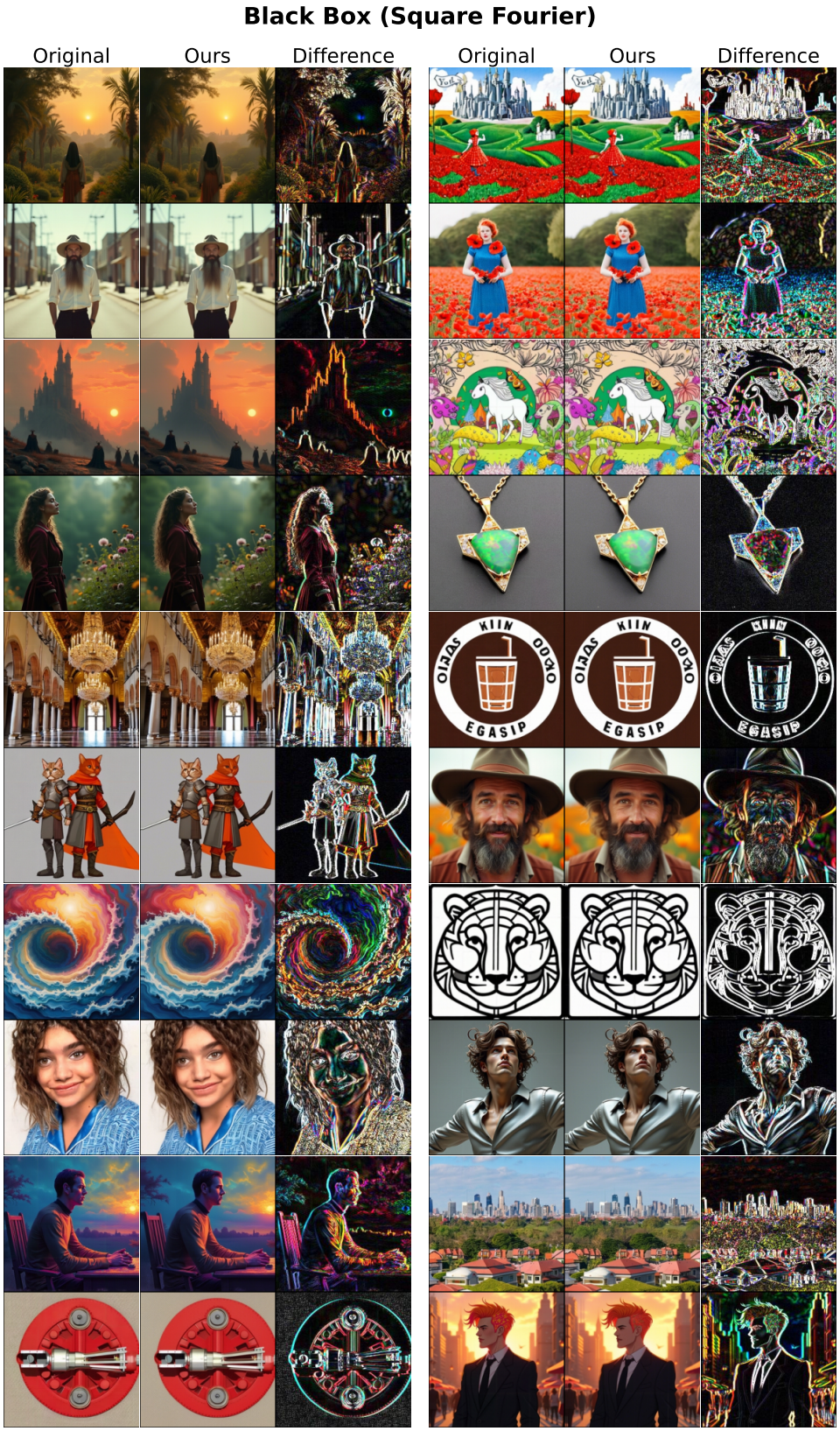

Figure 17: Qualitative results on the Black-box track for images containing square Fourier-domain artifacts. Each triplet shows (left) the original watermarked input, (middle) the output after applying our removal method, and (right) the residual difference between them. Our approach preserves semantic content and visual fidelity while effectively eliminating the embedded watermark

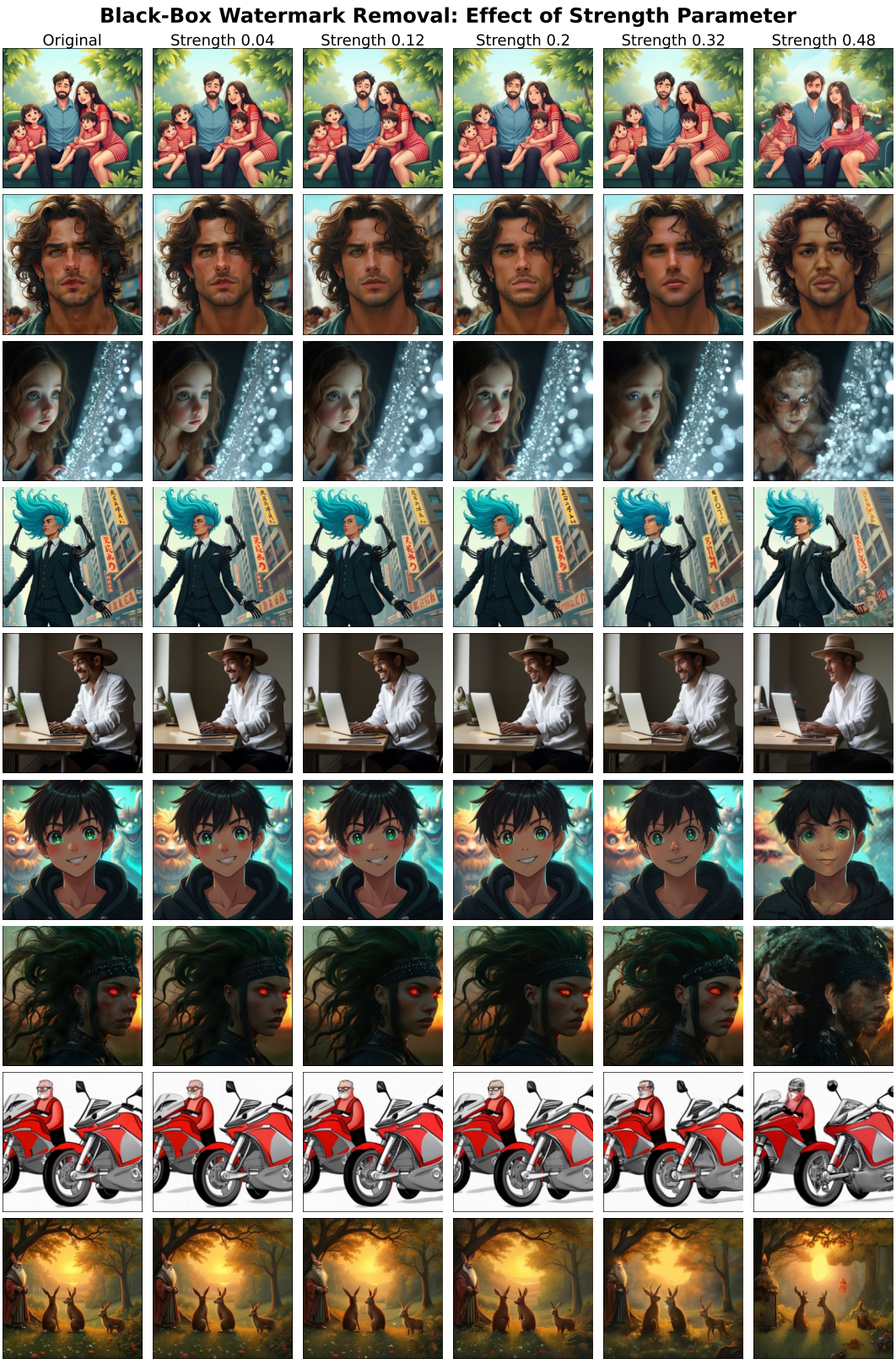

Figure 18: Effect of varying the diffusion strength parameter on Black-box watermark removal. Lower values (0.04–0.12) maintain high perceptual quality with partial removal, whereas higher values yield stronger removal at the cost of visible distortions. This illustrates the quality–robustness trade-off inherent in diffusion-based purification.

Figure 19: Captions automatically generated by ChatGPT for a set of original watermarked images. These textual descriptions capture semantic and stylistic details of the inputs and are subsequently used to guide image-to-image diffusion for watermark removal.

Figure 20: Captions automatically generated by ChatGPT for a set of original watermarked images. These textual descriptions capture semantic and stylistic details of the inputs and are subsequently used to guide image-to-image diffusion for watermark removal.

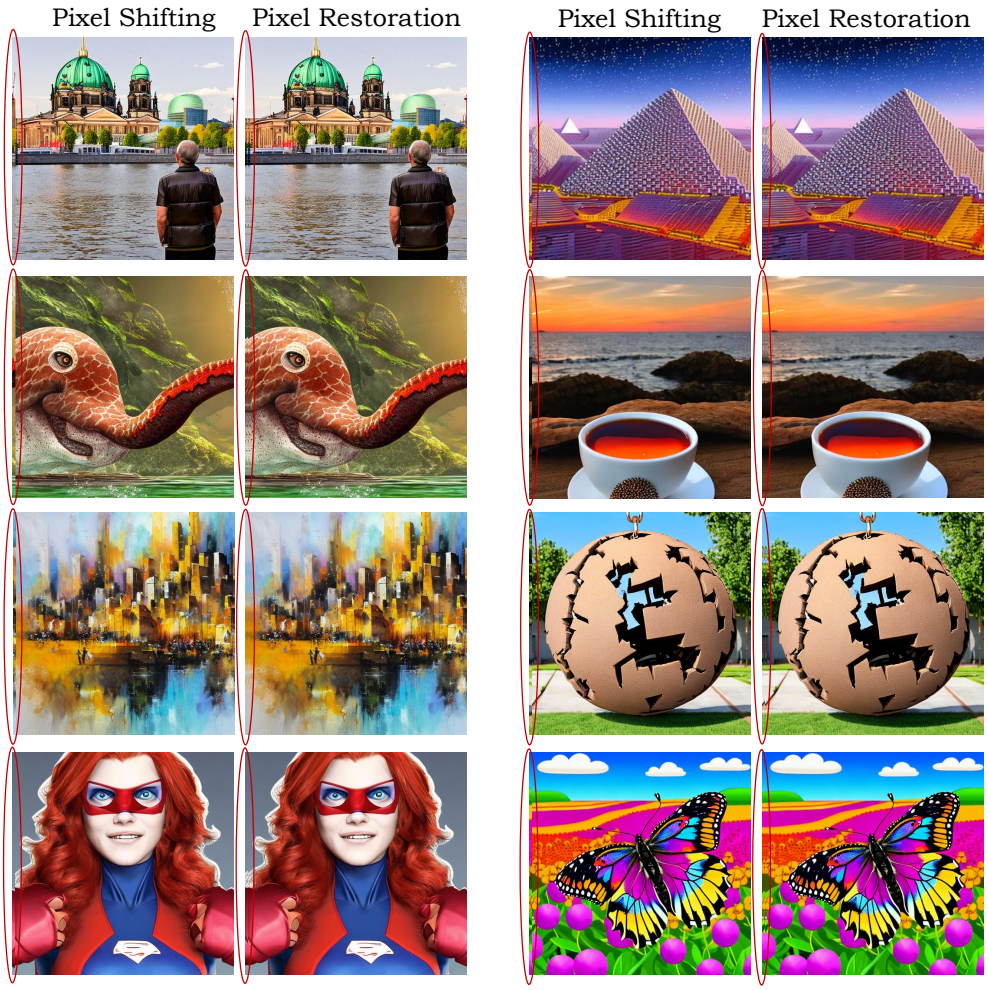

Figure 21: Effect of spatial translation on TreeRing watermarks. **Left**: Directly shifting the image by 7 pixels removes the watermark but introduces visible boundary artifacts. **Right**: Selectively restoring the shifted columns from the original image eliminates artifacts while preserving image quality, showing that minimal pixel-level restoration is sufficient.

Table 6: Quantitative comparison of our Beige-box removal strategy for TreeRing watermarks with and without pixel restoration. Incorporating selective restoration improves both fidelity metrics (FID, CLIP-FID) and perceptual similarity (SSIM, LPIPS), demonstrating that boundary-aware restoration enhances watermark removal without compromising quality.

Method	PSNR↑	SSIM↑	LPIPS↓	FID↓	NMI↑	CLIPFID↓
w/o Pixels Restoration	**14.878**	0.428	0.133	15.773	0.212	3.199
Ours	14.353	**0.434**	**0.124**	**12.531**	**0.218**	**1.622**

Figure 22: Captions generated by ChatGPT for watermarked images, used in conjunction with a ControlNet-based image-to-image diffusion model. Unlike earlier setups, the prompts explicitly capture both semantic content and color attributes, providing richer conditioning for watermark-free synthesis.