AEON: Adaptive Embedding Optimized Noise for Robust Submerged Watermarking in Diffusion Models

Supplementary Material

1. Additional Quality analysis on different steps

Here, we provide an analysis of the quality of generated images about the various steps of watermark implantation. In Fig. 1, we can see clear improvements in the image quality when we implant the watermark in the later steps with total inference steps of 50. However, the quality is much lower when the watermark is implanted in earlier steps. As we proceed with the steps, more details are added to the generated image to increase the attack resilience. While Fig. 2 shows more qualitative analysis of SOTA models.

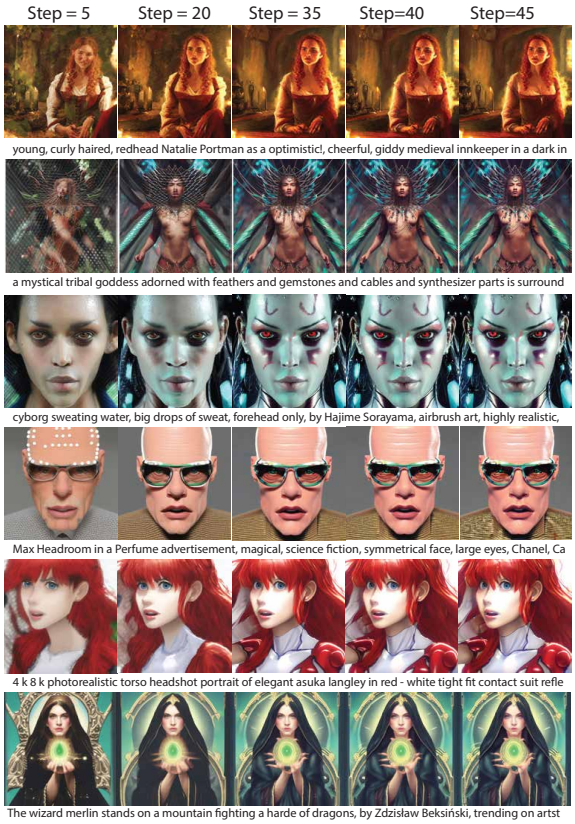

Figure 1. Effect of varying watermark implantation steps on the quality of the generated image.

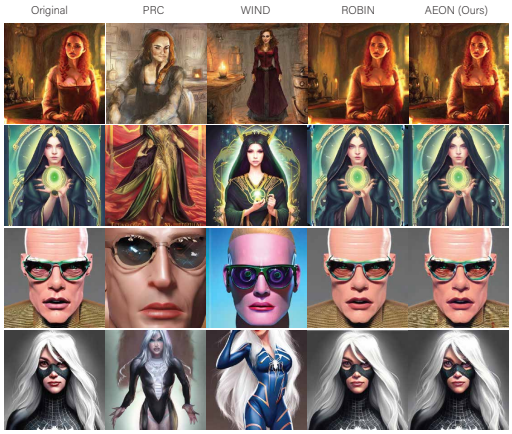

Figure 2. Comparison of more state-of-the-art

2. Additional Pixel Level Comparison with ROBIN

Fig. 3 presents a comparative analysis of the pixel differences between the original image and the generated watermark image using ROBIN [6] and AEON. The watermark is visible when we plot the pixel difference in the ROBIN-generated image. Meanwhile, AEON has a stronger watermark, which is submerged and invisible even when added later in the inference step 45.

3. Additional Ablation Study on each component

To assess the impact of each element in our approach, we conducted ablation studies by including and excluding different components of the hash training and watermark generation. Table 1 shows ablation study on other components, (1) shows performance of the watermarking approach when we include all components $\mathcal{L}_{recons}$, $\mathcal{L}_{ret}$, and $\mathcal{L}_{cons}$, (2) shows performance when we only consider reconstruction loss $\mathcal{L}_{recons}$, while (3) shows performance of the proposed approach will all components except we use logits without hashing for watermark generation. Removing loss functions from ROBIN [6], reduces the quality and performance on adversarial attacks, while removing the hashing mechanism. However, it reduces quality, resilience to adversarial attacks, and watermark verification accuracy.

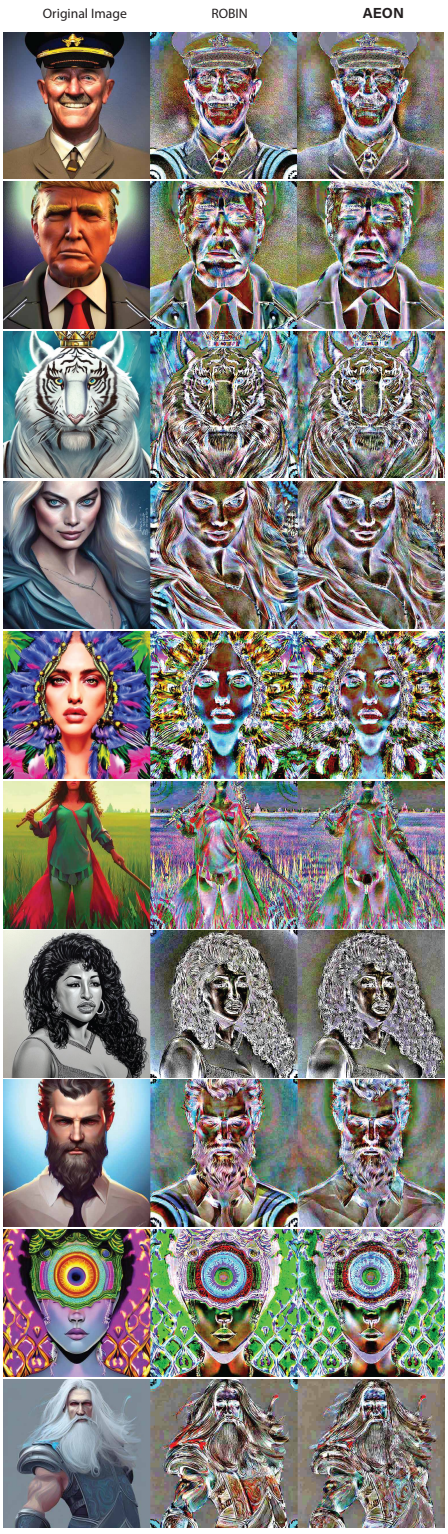

Figure 3. Effect of varying watermark implantation steps on the quality of the generated image.

Component	AUC↑		Image quality			
	Clean	Adv.	PSNR↑	SSIM↑	CLIP↑	FID↓
(1)	1.00	0.991	25.93	0.81	0.41	26.61
(2)	1.00	0.989	24.68	0.71	0.38	28.76
(3)	1.00	0.981	25.83	0.79	0.20	26.18

Table 1. Watermark accuracy and image quality under different settings.

4. Visualization of Different Attacks

Figure 4 illustrates traditional attacks on the watermarked image, such as Brightness, Crop, Gaussian Blur, Gaussian Noise, JPEG, and Rotation Attack.

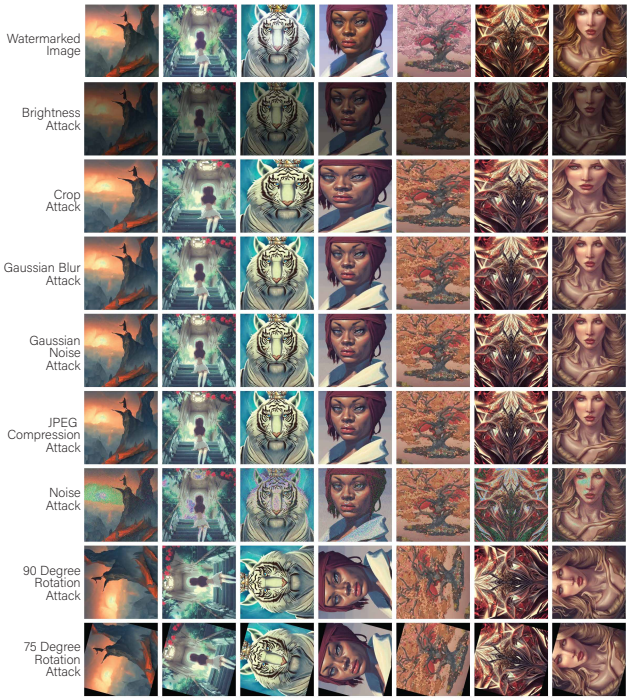

Figure 4. Visualization of Different traditional attacks on watermarked images.

Similarly, Fig. 5 shows different reconstruction adversarial attacks on the generated watermark, that we used to assess the model efficiency under different attacks such as BM3D [4], Zhao23 [9], Cheng20 [3], Bmshj18 [2] and combination of reconstruction of attacks from Fig. 4 and Fig. 5 as a combined attack.

5. Effect of Varying ϕ.

In Eq. 10, ϕ controls the strength of the watermark during the watermark injection in the frequency domain. Figure 6 shows the effect of different blending factors on the quality metric PSNR and watermark verification rate

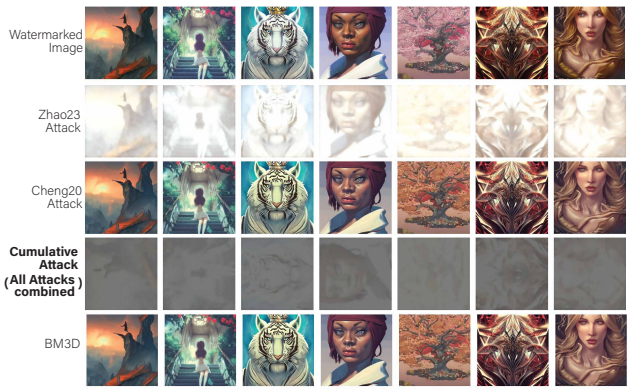

Figure 5. Visualization of Different reconstruction attacks and combined attack on watermarked images.

(AUC). This result indicates that increasing ϕ reduces the image quality because it decreases the watermark strength. The higher the watermark strength, the better the quality and resilience to attacks.

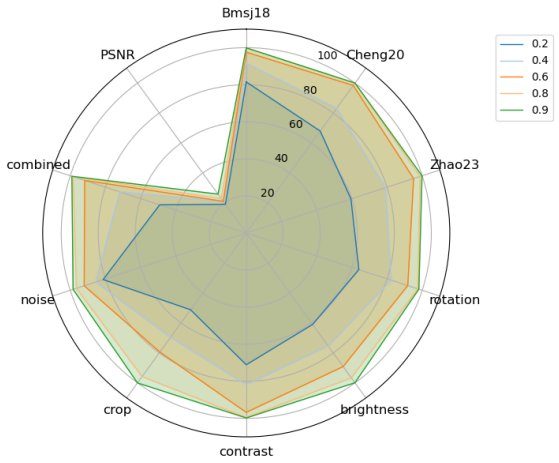

Figure 6. Ablation study on the effect of blending factor ϕ against different versions of Stable diffusion, and watermark removal attacks.

6. Robustness against Adversarial Attacks.

To evaluate the model's robustness against adversarial or forgery attacks, we evaluated the performance of the proposed approach with the state-of-the-art forgery method Imprint, as shown in Table 2, existing methods easily breached in terms of AUC by all of the proposed approaches, while the proposed approach showed robustness against adversarial attacks.

Method	Imprint Forgery
Tree-Ring [8]	0.000
WIND [1]	0.000
PRC [5]	0.000
ROBIN [6]	0.210
Ours	**0.680**

Table 2. Comparison of Adversarial Attack Imprint Forgery [7].

Algorithm 1 Verification Algorithm

Require: image x: query image, Θ: private diffusion weights, f_{hash}: trained hash network, t_{inj}: injection step, bins: frequency mask, τ: detection threshold

$\tilde{x} \leftarrow G_{\Theta}^{-1}\big(x,\, t_{\text{inject}}\big)$ ▷ Reverse-diffuse to the injection step

$\tilde{w}_{\text{freq}} \leftarrow \mathbf{1}_{\text{bins}} \odot \text{FFT}\big(\varphi(\tilde{x})\big)$ ▷ Keep only watermark frequency bins

$\hat{x}'_t \leftarrow \text{iFFT}(\tilde{w}_{\text{freq}})$ ▷ Candidate watermark in pixel space

$w \leftarrow \text{sign}\big(f_{\text{hash}}(\varphi(\tilde{x}))\big)$ ▷ Reference hash bits

$s \leftarrow M\big(w,\, \hat{x}'_t\big)$ ▷ Number of matching bits

if $s \geq \tau$ **then**

  Declare "watermarked"

else

  Declare "not watermarked"

end if

7. Watermark Detection Procedure (M)

Algorithm 1 formalises the *verification stage* of our pipeline, deciding whether a query image x is watermarked.

i. Latent recovery. The private inverse-diffusion model G_{Θ}^{-1} is run up to the injection timestep t_{inj}, producing the intermediate latent $\tilde{x}$.

ii. Frequency isolation. We transform $\tilde{x}$ into the frequency domain and retain only the watermark-bearing coefficients by masking with $\mathbf{1}_{\text{bins}}$: $\tilde{w}_{\text{freq}} = \mathbf{1}_{\text{bins}} \odot \text{FFT}\big(\varphi(\tilde{x})\big)$.

iii. Candidate reconstruction. An inverse FFT converts the masked spectrum back to pixel space, yielding the candidate watermark $\hat{x}'_t$.

iv. Reference hash. Passing $\tilde{x}$ through the trained hash network f_{hash} and taking the sign produces the reference bit string $w = \text{sign}\big(f_{\text{hash}}(\varphi(\tilde{x}))\big)$.

v. Bit-level comparison. The matching function $M(\cdot,\cdot)$ counts identical bits between w and $\hat{x}'_t$, returning a score s.

vi. Decision rule. If $s \geq \tau$, the image is classified **watermarked**; otherwise it is deemed **clean**.

The scheme is resilient because the watermark is (i) recovered from the exact diffusion layer where it was injected, (ii) isolated in the designated frequency bins, and (iii) authenticated via a strict bitwise test, thereby minimising false positives.

References

[1] Kasra Arabi, Benjamin Feuer, R. Teal Witter, Chinmay Hegde, and Niv Cohen. Hidden in the noise: Two-stage robust watermarking for images, 2025.

[2] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018.

[3] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering, 2007.

[5] Sam Gunn, Xuandong Zhao, and Dawn Song. An undetectable watermark for generative image models, 2025.

[6] Huayang Huang, Yu Wu, and Qian Wang. Robin: Robust and invisible watermarks for diffusion models with adversarial optimization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 3937–3963. Curran Associates, Inc., 2024.

[7] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.

[8] Linfeng Zhang, Xinyang Liu, Yuriy Brun, Hongyu Guan, and Anne-Marie Viros. Tree-rings watermarks: Invisible fingerprints for diffusion images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[9] 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, 2024.