GROW: Watermark Generation with Progressive Guidance for Diffusion Models

Supplementary Material

1. Analysis of Watermarking Paradigms

1.1. Difference from Tree-Ring based Watermark

To understand the core innovation of GROW, it is essential to analyze its fundamental difference from the dominant training-free paradigm, exemplified by Tree-Ring [4]. While both methods achieve a robust, semantic fusion of the watermark with the image content, they do so through starkly contrasting mechanisms: passive scattering versus active guidance.

Tree-Ring based: Passive Scattering. The Tree-Ring methodology begins by embedding a watermark signal $\mathbf{W}$ into the initial noise latent, creating a modified starting point $\mathbf{z}'_T = \mathbf{z}_T + \mathbf{W}$. The standard denoising process then commences. Conceptually, the total initial noise $\mathbf{z}'_T$ is progressively consumed to produce the final clean latent $\mathbf{z}'_0$ and a series of predicted noises $\{\mathcal{E}'_t\}$:

$$\mathbf{z}'_T = \mathbf{z}'_0 + \sum_{t=1}^{T} \mathcal{E}'_t$$

As the denoising process unfolds, this initial signal $\mathbf{W}$ is inevitably distributed among all resulting components:

$$\mathbf{z}'_T = (\mathbf{z}_0 + \mathbf{w}_0) + \sum_{t=1}^{T} (\mathcal{E}_t + \mathbf{w}_t)$$

A portion, let's call it $\mathbf{w}_0$, remains in the final clean latent, while other portions $\{\mathbf{w}_t\}$ are "subtracted away" as parts of the predicted noise at each step. As the Tree-Ring paper notes, the watermark becomes "semantically hidden in the image space." This implies that the surviving portion $\mathbf{w}_0$ is one that aligns with the image's content (i.e., it is "semantically plausible"). The diffusion model, in its effort to generate a coherent image, naturally filters out parts of $\mathbf{W}$ that are inconsistent with the emerging semantics, while retaining and integrating the parts that are.

This process can be viewed as a form of passive filtering. The final embedded watermark $\mathbf{w}_0$ is a byproduct of the denoising process, an unknown and context-dependent fraction of the original signal $\mathbf{W}$. Consequently, detecting the watermark directly from the final image latent $\mathbf{z}'_0$ is impossible, as one only has access to the "surviving" fraction $\mathbf{w}_0$. To verify the watermark, one must recover an estimate of the entire initial latent $\mathbf{z}'_T$ to check for the presence of the complete original signal $\mathbf{W}$. This necessitates a costly inversion process, which is the primary bottleneck of this paradigm.

GROW: Active Guidance. GROW operates on the reverse principle. Instead of passively scattering an initial signal, we actively guide the generation process towards a fixed, predefined watermark target $\mathbf{S}$ that is intended for the final clean latent $\mathbf{z}_0$. This is achieved by reformulating the watermarking task as an optimization problem that is solved concurrently with image generation.

At each guided denoising step, we introduce a watermark loss, $\mathcal{L}_{wm} = \text{Loss}(\text{DCT}(\hat{\mathbf{z}}_0), \mathbf{S})$, where $\hat{\mathbf{z}}_0$ is the predicted clean latent at that step. This loss is balanced against the model's primary objective of matching the text condition c. The final noise prediction $\mathcal{E}^{\text{final}}$ becomes a weighted combination that harmonizes these two competing goals:

$$\mathcal{E}^{\text{final}} \propto (\text{guidance towards } c) + \eta \cdot (\text{guidance towards } \mathbf{S})$$

The U-Net is thus tasked with finding a generation path that minimizes both objectives simultaneously. It must produce an image that is not only semantically consistent with the prompt but also structurally embeds the target watermark $\mathbf{S}$. The watermark is not a leftover artifact; it is an active, explicit constraint that the model must accommodate throughout the generation. This active optimization forces the model to find a harmonious solution, weaving the frequency-domain watermark signal into plausible image textures in a manner consistent with its learned priors.

1.2. The Advantage of Active Guidance

This analysis reveals a crucial insight: GROW retains the key advantage of methods like Tree-Ring: the final watermark signal is deeply and robustly fused with the image content because it is generated in a way that is consistent with the image's semantics. The difference, and GROW's core contribution, lies in the mechanism.

In Tree-Ring, this semantic fusion is a passive byproduct of the denoising process filtering a watermarked noise. The model is unaware of the watermarking goal; it simply performs its denoising task, and the robust embedding happens as a consequence.

In GROW, the fusion is the result of an active optimization. The model is explicitly guided to find a solution that satisfies both the content and watermark objectives. By reformulating the problem from passive scattering to active guidance, GROW achieves the same desirable fusion of watermark and content, but with a significant practical advantage: the watermark target $\mathbf{S}$ is known and fixed, making extraction direct, efficient, and completely free of any inversion process. This shift in mechanism from passive to active

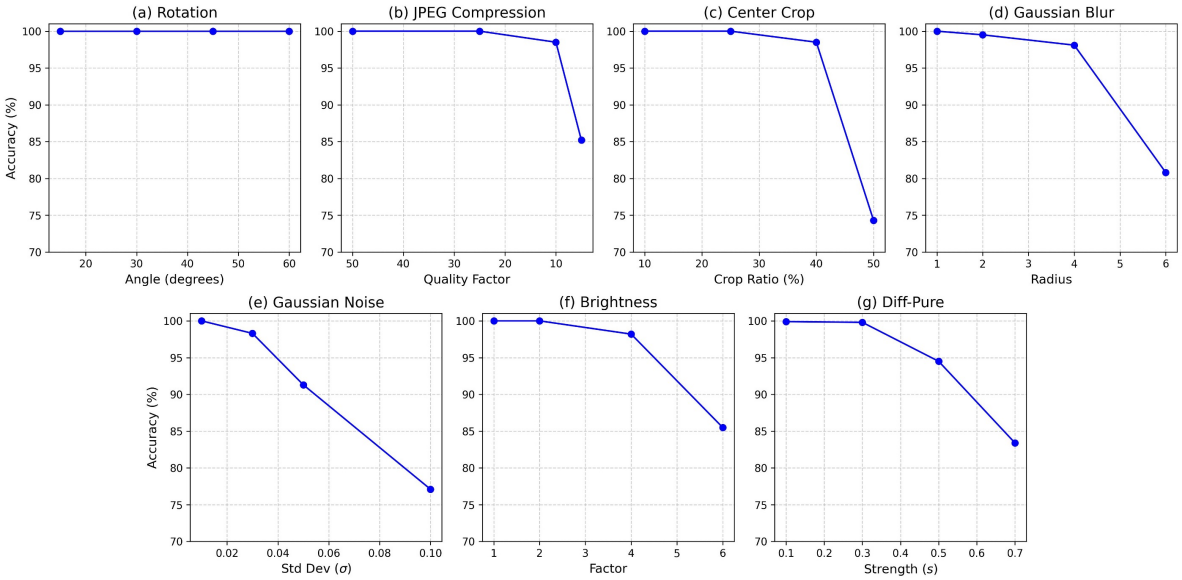


Figure 1. Robustness of GROW under varying attack intensities. Each plot displays the average extraction accuracy on images generated from the MS-COCO and Stable Diffusion Prompts datasets.

is what unlocks the dramatic gains in extraction speed while preserving the high robustness intrinsic to in-generation watermarking methods.

2. Additional Experiments

2.1. Varying Attack Strength

To provide a more nuanced understanding of GROW's robustness, we analyze its performance under varying intensities of common attacks. As illustrated in Figure 1, GROW demonstrates remarkable resilience across a wide range of distortions. This section further clarifies the mechanisms behind this strong performance, particularly concerning geometric attacks, and reconciles these results with the limitations discussed in the main paper.

Robustness to JPEG Compression. As shown in Figure 1(b), GROW maintains near-perfect Message Accuracy (M-ACC) even under severe JPEG compression (e.g., quality factor of 10). This exceptional robustness stems from our design choice to embed the watermark in the VAE's latent space ($\mathbf{z}_0$). The VAE encoder is trained to be robust to common image corruptions, effectively filtering out compression artifacts and preserving the core semantic information along with our embedded watermark signal in its latent representation.

Synchronized Extraction for Geometric Attacks. The apparent contradiction between the high M-ACC against rotation and cropping in Table **??** and the stated sensitivity to geometric transformations in our Limitations section is resolved by our extraction strategy. For practical applications, watermark extraction is not a passive, one-shot process. Instead, we employ a **synchronized extraction** mechanism that actively searches for the watermark under a set of expected transformations.

For Rotation: Our extractor iterates through a predefined set of rotation angles (e.g., every 15 degrees from -90 to +90). At each angle, it attempts to extract the watermark and calculates a confidence score based on the consistency of the repeated bits. The final extracted message is the one corresponding to the angle with the highest confidence score. This search-based approach effectively synchronizes the extractor with the orientation of the attacked image, allowing for perfect M-ACC as long as the rotation falls within the tested range. The small angular gap like 1 degree between our test attacks and the detection angles is easily tolerated by the robustness of the embedded signal.

For Cropping: A similar search strategy is applied. As our experiments show, this method can tolerate random cropping of up to 0.4 of the image area while maintaining high M-ACC.

Reconciling Robustness with Limitations. While highly effective, this synchronized detection is inherently slower than a direct extraction and is limited to a predefined set of transformations. It cannot, for example, handle arbitrary, non-rigid warping or unconstrained random ratio stretching. This is because such distortions disrupt the structural relationships of DCT coefficients, which are critical for watermark decoding. This is the fundamental challenge for training-free methods operating in a fixed domain, and it motivates our future work on a learnable, transformation-invariant representation space. In summary,

CVPR
#8345

Table 1. Impact of watermark capacity on image quality (FID) and robustness (Avg. M-ACC) on MS-COCO Dataset.

Capacity (bits)	FID ↓	Avg. M-ACC ↑
16	12.32	0.976
32	18.15	0.953
64	25.59	0.872

our reported high M-ACC is a testament to a practical and effective extraction system, while our stated limitation acknowledges the inherent theoretical constraints of the DCT watermark structure.

2.2. Watermark Capacity

We investigate the trade-off between watermark capacity and performance by testing GROW with 16, 32, and 64 bit payloads. As presented in Table 1, increasing the capacity necessitates a stronger watermark signal, which predictably leads to a measurable degradation in both image quality (higher FID) and robustness (lower M-ACC). This performance drop highlights a core limitation of operating within a fixed domain like DCT: as more frequency coefficients are utilized to carry information, the watermark becomes more intrusive and simultaneously more vulnerable to being overwritten by image content or attacks.

his experiment underscores the inherent constraints of its training-free nature. As discussed in our Limitations section, a promising future direction is to replace the static DCT space with a learnable representation, which could potentially support higher capacities with less impact on quality and robustness. This study validates the flexibility of our current framework while also motivating the exploration of such hybrid approaches.

2.3. Verification Experiment

Beyond full message extraction, a crucial real-world application is verification: a binary classification task to determine if an image contains a watermark or not. This is often simpler and faster than extracting the entire message. To evaluate GROW's performance on this task, we conduct a comprehensive verification experiment.

Our verification method leverages the confidence score derived from the redundancy of embedded bits. During extraction, each bit of the message is embedded multiple times. The detector performs a majority vote for each bit and calculates a bit consistency score: the percentage of repeated bits that match the majority vote. The average consistency across all message bits serves as our final confidence score. A predefined threshold is applied to this score: if the confidence is above the threshold, the image is classified as "watermarked"; otherwise, it is "non-watermarked."

We evaluated the verification performance on 1,000 watermarked and 1,000 non-watermarked images from the MS-COCO dataset, under the same set of attacks used in our main robustness experiments. To ensure a fair and stringent evaluation, we report the True Positive Rate at a False Positive Rate of approximately 1% (TPR@1%FPR). As summarized in Table 2, GROW demonstrates exceptional performance. It achieves an average TPR of 99.8% across all attacks, indicating that it can reliably verify watermarked images even after distortions, while maintaining an extremely low false alarm rate. This high level of reliability and robustness makes GROW suitable for practical deployment.

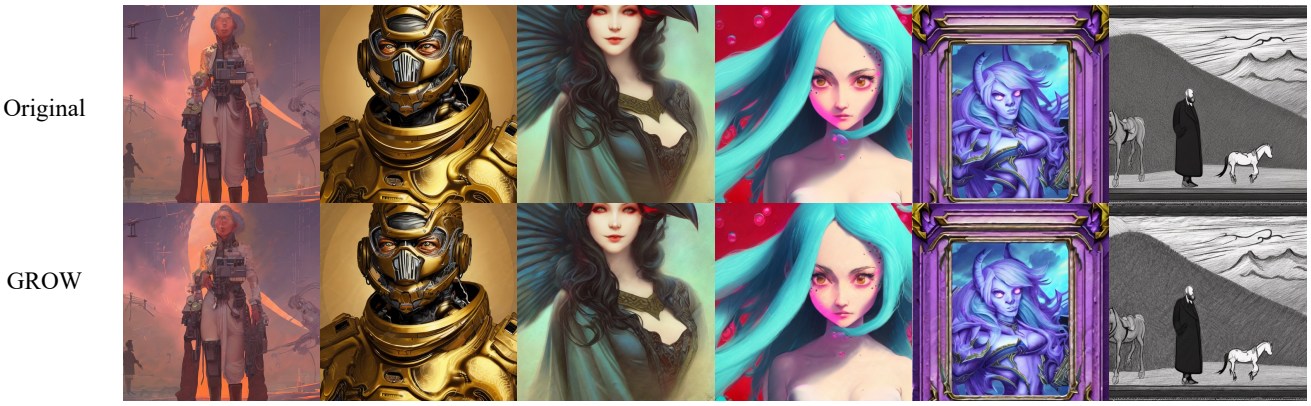
Table 2. Verification performance (TPR@1%FPR) under various attacks on the MS-COCO dataset. The threshold for each attack is dynamically set to achieve an FPR of approximately 1% on 1,000 non-watermarked images.

Attack Type	TPR (@1% FPR)
Clean	100.0%
Rotation	99.9%
JPEG	100.0%
Crop	99.5%
Blur	100.0%
Noise	99.7%
Brightness	99.8%
Average	**99.8%**

2.4. Additional Visual Results

To further showcase visual quality and generality, Figure 2 presents results on diverse datasets, including MS-COCO [2], WikiArt [3], and Stable Diffusion prompts [1]. The watermarked images are imperceptible, demonstrating that our method preserves image fidelity across varied subjects and styles.

CVPR
#8345



(a) Stable Diffusion prompts

(b) MS-COCO

(c) WikiArt

Figure 2. Additional qualitative results of watermarked images generated by GROW on diverse datasets. The visual quality is consistently high, and the watermark is imperceptible.

References

[1] Gustavosta. Stable-diffusion-prompts. https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts, 2024. 3

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[3] Wei Ren Tan, Chee Seng Chan, Hernan E Aguirre, and Kiyoshi Tanaka. Improved artgan for conditional synthesis of natural image and artwork. *IEEE Transactions on Image Processing*, 28(1):394–409, 2018. 3

[4] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023. 1