

GROW: Watermark Generation with Progressive Guidance for Diffusion Models

Anonymous CVPR submission

Paper ID 8345

Abstract

Digital watermarking is a cornerstone for copyright protection. With the rapid advancement of generative models like diffusion models, in-generation and training-free watermarking techniques have garnered significant attention for their endogeneity and convenience. These methods typically embed a watermark into the initial noise, where watermark extraction relies on Denoising Diffusion Implicit Models (DDIM) inversion. However, the computationally intensive extraction process severely hinders their path toward practical deployment. To overcome this critical bottleneck, we propose GROW, a novel training-free paradigm that reframes watermarking from a one-shot 'embedding' to a progressive 'growth'. Specifically, at each denoising step, we compute a loss in the frequency domain between the predicted clean latent representations and target watermark signals. The gradient of this loss is then used to progressively guide the generation, ensuring the watermark is naturally woven into the final image representations. Notably, GROW requires no inversion process for extraction, enabling highly efficient watermark verification. Comprehensive experiments on multiple datasets show that GROW not only achieves superior imperceptibility and robustness but also offers a detection speed nearly 100x faster than inversion-based techniques. The code will be made publicly available.

1. Introduction

The rapid advancement of generative models, particularly diffusion models (DMs) [15, 25, 29, 32, 33] has revolutionized AI-generated content (AIGC), enabling the creation of photorealistic images from textual prompts. This accessibility, however, brings significant challenges in copyright protection and content traceability [4, 18]. Digital watermarking [16] is a critical technology to address these issues by embedding imperceptible information into generated images, thereby providing a means for authentication and provenance tracking.

Existing watermarking methods for diffusion models can

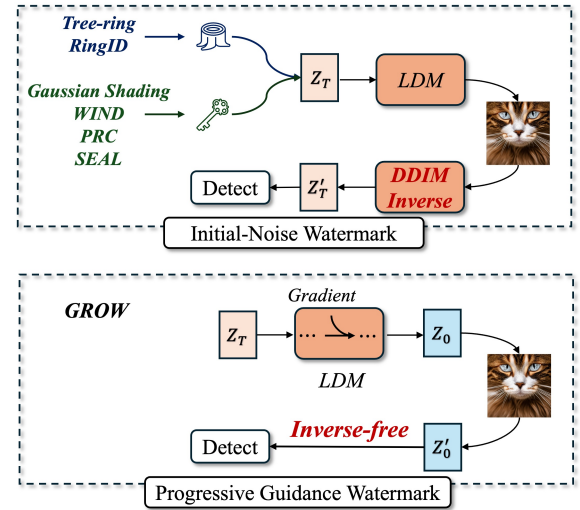


Figure 1. A comparison of watermarking paradigms for diffusion models. (a) **Initial-Noise Watermarking**: Embeds a static watermark into the initial noise, which is passively scattered during generation. Watermark extraction requires computationally expensive inversion. (b) **Progressive Guidance Watermarking (Our GROW)**: Actively guides each denoising step via gradients, enabling the watermark to grow progressively with the image. This allows for direct extraction, eliminating the need for inversion.

be broadly classified into three categories. *Post-processing* methods [31, 38, 41] directly alter the final image, often at the cost of visual quality and robustness. In contrast, in-generation watermarking integrates the watermark directly into the generation process. This provides a crucial advantage of endogeneity: the watermark is semantically hidden within the image. Consequently, its complete removal would require knowledge of the diffusion model's internal parameters, access to which is typically impractical. Within in-generation methods, *Fine-tuning-based* approaches [3, 5, 7, 8, 10, 20, 23, 24, 27, 36, 40] embed the watermark by either modifying the model's parameters or introducing it as a conditional input, which requires extensive computational resources for retraining. More recently, *Training-free* methods [1, 2, 6, 11, 17, 19, 21, 30, 35, 37]

have emerged as a flexible alternative. These methods typically operate by manipulating the initial noise latent before the denoising process begins.

A seminal work in this area, Tree-Ring [35], embeds watermarks by creating specific patterns in the initial noise latent. This foundational idea has inspired a series of subsequent methods [1, 2, 6, 11, 17, 19, 21, 30, 37] that also operate on the initial noise. These methods improve upon the original concept by using different strategies, such as embedding watermarks based on semantic content [2] or aligning with specific patterns [1]. However, a significant and shared drawback of these initial-noise-based methods is their reliance on a computationally expensive inversion process (DDIM inversion [32]) for watermark detection. This inversion, which is necessary to recover the initial noise latent from a given image, is time-consuming and severely hinders the practicality of these methods for large-scale, real-time applications. This raises a critical question: Can we develop a watermarking method that eliminates the need for this slow inversion process? A naive approach might be to add the watermark to the final latent representation just before decoding. However, our experiments in Figure 6 show that this drastically disrupts the learned data distribution, leading to severe image degradation.

We realize that the solution lies not in abandoning the principle of endogeneity, but in rethinking its mechanism. The robustness of initial-noise-based methods stems from a **passive scattering** process, where the denoising model implicitly filters a watermarked noise, fusing a portion of it with the image semantics. Our key insight is to achieve this same desirable fusion through **active guidance**. (A more detailed theoretical analysis is provided in Appendix 7.) As depicted in Figure 1(b), we bridge image generation and watermark generation by leveraging the model’s own iterative denoising process. Instead of modifying the initial noise, we introduce a gradient-based guidance mechanism. At each denoising step, we compute the difference between the latent features (in the frequency domain) and a fixed target watermark pattern. The gradient of this difference is then used to gently steer the generation, constituting a progressive guidance approach. The subtle perturbations introduced at each step accumulate along the model’s learned data manifold. In subsequent steps, the U-Net denoiser, driven by the conditioning signal, naturally incorporates these perturbations into the emerging image textures (e.g., feathers, tree bark), resulting in a deeply fused watermark.

This reformulation that from passive scattering to active guidance forms the core of our method, which we name **GROW** (Generation with **pro**gressive **W**atermarking). GROW allows the watermark to grow and mature alongside the image itself, ensuring it is an integral part of the image’s generative history. Crucially, this active approach

retains the semantic fusion advantage while offering a decisive practical benefit: because the watermark target is fixed and known, extraction becomes direct, efficient, and completely free from inversion. GROW thus establishes a new paradigm, fundamentally redefining the relationship between the carrier and the watermark, shifting from a static “imprint” to a dynamic “co-evolution”.

Our work makes three key contributions:

1. We propose GROW, a novel training-free paradigm that reframes watermarking from a static embedding of the initial noise to a dynamic guidance throughout the generation process. This eliminates the dependency on model inversion for extraction.
2. We develop a novel mechanism that leverages gradient guidance in the frequency domain to progressively weave the watermark into the final latent representations, achieving a balance of imperceptibility and robustness.
3. We empirically demonstrate that GROW not only surpasses existing methods in robustness and imperceptibility but also achieves a nearly 100x speedup in extraction, offers a fast and practical watermarking solution.

2. Related Works

2.1. Image Watermarking

Recent research on image watermarking can be broadly categorized into three categories.

The first, *post-processing-based* watermarking [31, 38, 41], where watermarking algorithms are applied directly to the final generated images. These methods, often adapted from classical techniques [31, 41], typically operate in the spatial domain, a transform domain (e.g., DCT, DWT) or a latent space to embed the watermark signal. For example, RivaGAN [38] uses a GAN-based framework with an encoder-decoder architecture to embed and extract watermarks.

To achieve deeper integration, *fine-tuning-based* approaches [3, 7, 10, 23, 36, 40] amalgamate the watermarking process with the model’s architecture. These approaches typically modify a pre-trained diffusion model to instill a watermarking capability. For instance, Stable Signature [10] fine-tunes the VAE decoder of a Latent Diffusion Model (LDM) with a dedicated watermark extractor, enabling the decoder to implicitly embed a signature during image reconstruction. Other works [23, 40] propose fine-tuning the entire diffusion model to implant a “backdoor,” where a specific trigger can reveal the hidden watermark. Some methods also introduce the watermark as a conditional input during the training phase [36]. Although these methods can produce high-quality watermarked images with strong robustness, they require significant computational resources for retraining and result in modified model parameters, which can be a major barrier for prac-

tical deployment.

Recently, *training-free* methods have emerged as a flexible, plug-and-play alternative. Their dominant strategy, as illustrated in Figure 1(a), is to manipulate the initial noise latent $\mathbf{z}_T$. The seminal work, Tree-Ring [35], demonstrated that watermark information can be conveyed by engineering specific patterns into the Fourier spectrum of the initial noise. This foundational idea has inspired a lineage of research. For instance, WIND [1] focuses on designing specific frequency-domain patterns for improved robustness, while SEAL [2] innovatively links the watermark to the semantic content of the input prompt. Gaussian Shading [37] proposes mapping the watermark to a distribution-preserving Gaussian representation to achieve performance-lossless embedding. Despite their ingenuity, a critical flaw shared by all these initial-noise-based methods is their reliance on a computationally expensive inversion process [32] for watermark detection. Recovering the initial noise latent from a given image is a time-consuming procedure that severely hinders the practicality of these methods for large-scale, real-time applications. Our work, GROW, belongs to the training-free category but introduces a fundamentally different paradigm that circumvents this inversion requirement entirely.

2.2. Guidance in Diffusion Models

Guidance techniques are fundamental to controlling the generation process in diffusion models, enabling the steering of outputs towards desired attributes or away from undesired ones.

Classifier Guidance [9] was an early and influential approach. It works by combining the score function of the diffusion model (which estimates the data distribution gradient) with the gradient of a pre-trained classifier. During the denoising step at timestep t , the noise prediction is adjusted as follows:

$$\hat{\mathcal{E}} = \mathcal{E}_\theta(z_t, t) - \sigma_t \nabla_{z_t} \log p_\phi(y|z_t) \quad (1)$$

where $\mathcal{E}_\theta(z_t, t)$ is the model’s predicted noise, σ_t is the noise schedule, and $\nabla_{z_t} \log p_\phi(y|z_t)$ is the gradient of the classifier with respect to the latent z_t for class y . This adjustment effectively pushes the generation towards regions of the latent space that have a high probability under the target class y . While powerful, this method requires training a separate classifier on noisy data and can compromise sample diversity.

Classifier-Free Guidance (CFG) [14] has since become the mainstream approach due to its simplicity and effectiveness. It circumvents the need for a separate classifier by training the diffusion model to perform both conditional and unconditional denoising simultaneously. During inference, the predictions are interpolated:

$$\hat{\mathcal{E}}_{\text{CFG}} = \mathcal{E}_\theta(z_t, t, \emptyset) + w \cdot (\mathcal{E}_\theta(z_t, t, y) - \mathcal{E}_\theta(z_t, t, \emptyset)) \quad (2)$$

Here, $\mathcal{E}_\theta(z_t, t, y)$ is the conditional prediction (with prompt y), $\mathcal{E}_\theta(z_t, t, \emptyset)$ is the unconditional prediction (with empty prompt $\emptyset$), and $w > 1$ is the guidance scale that amplifies the direction pointed by the conditional prediction. This allows for flexible control over the trade-off between sample quality and adherence to the condition.

Our work is conceptually inspired by these guidance paradigms. We adapt the core idea of steering the denoising trajectory to the novel task of invisible watermarking. Instead of guiding semantic attributes like “an astronaut riding a horse,” we guide the generation process to subtly and progressively incorporate a watermark signal, leveraging a similar mechanism to achieve deep fusion within the natural image manifold.

3. Method

Our proposed method, GROW, introduces a new paradigm for training-free watermarking by integrating the watermark embedding process into the denoising steps of a diffusion model. Instead of embedding a static pattern into the initial noise, we progressively guide the generation process towards a latent representation that contains the desired watermark signature.

3.1. Preliminaries: Latent Diffusion Models

Latent Diffusion Models (LDMs) [29] generate images by iteratively denoising a random latent variable $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$ over T timesteps. This reverse process is driven by a U-Net model, $\mathcal{E}_\theta$, which is trained to predict the noise in a noisy latent $\mathbf{z}_t$ at each timestep t . A crucial step in modern schedulers involves estimating the predicted clean latent, $\hat{\mathbf{z}}_0$, from $\mathbf{z}_t$ and the predicted noise $\mathcal{E}_\theta(\mathbf{z}_t, t, c)$:

$$\hat{\mathbf{z}}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} (\mathbf{z}_t - \sqrt{1 - \bar{\alpha}_t} \mathcal{E}_\theta(\mathbf{z}_t, t, c)) \quad (3)$$

where $\bar{\alpha}_t$ is a pre-defined noise schedule parameter and c represents the conditioning (e.g., text embeddings). Our method leverages this predicted $\hat{\mathbf{z}}_0$ as the target for our guidance. After T steps, the final latent $\mathbf{z}_0$ is decoded into an image $\mathbf{x} = \mathcal{D}(\mathbf{z}_0)$ by a VAE decoder $\mathcal{D}$.

3.2. Watermark Encoding and Frequency-Domain Mapping

To create a watermark that is both imperceptible and robust, we embed it in the frequency domain of the latent space using the Discrete Cosine Transform (DCT). We strategically choose to embed the watermark in the mid-frequency band. This choice is crucial: embedding in low-frequency components would visibly distort the core structure of the image, while embedding in high-frequency components would make the watermark vulnerable to removal by common

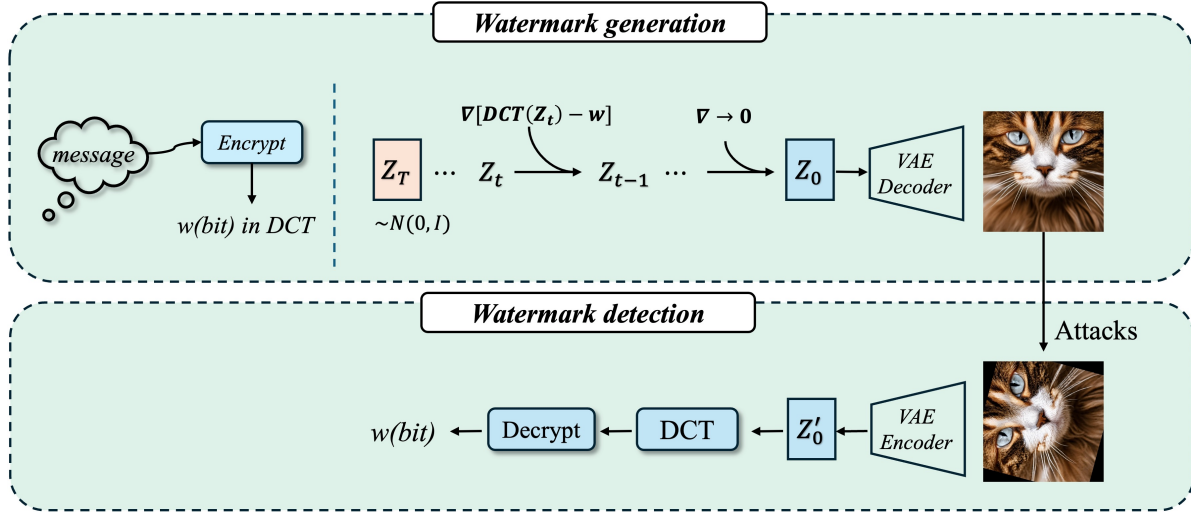


Figure 2. The overall framework of our proposed method, GROW. (a) **Watermark generation:** A secret message W is first converted into a bit sequence w . The watermark w is embedded into the target frequency domain at the positions defined by a mask M , which is generated from a secret key K . During the iterative denoising process, we compute a loss between the DCT of the predicted clean latent z_t and the target signal. The gradient of this loss guides the generation, progressively embedding the watermark. (b) **Watermark Detection:** To extract the watermark, the final image is encoded back to its latent representation z_0 . We then apply DCT and, using the key K , check the signs of the coefficients at the locations specified by the mask M . To ensure robustness, each bit of the message is embedded multiple times; a majority vote among these redundant copies reveals the final bit, making the extraction highly resilient to distortions. This process is direct and requires no inversion.

operations like JPEG compression or blurring. The mid-frequency band thus offers an optimal trade-off between visual quality and robustness.

First, a secret message W is converted into a binary bit sequence w . Given a secret key K , we use a seeded pseudo-random number generator to generate a binary mask matrix M . The bit sequence w is repeated and mapped to mask matrix to define a target signal matrix S :

$$S(u, v) = \begin{cases} \alpha \cdot (2w_i - 1) & \text{if } M(u, v) = 1 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where w_i is the bit assigned to the coordinate (u, v) , and α is the watermark strength.

3.3. Progressive Guidance for Watermark Embedding

This is the core of our method. Instead of modifying the initial noise z_T , we apply a gentle guidance force during the denoising process, steering the generation towards a final latent z_0 whose DCT coefficients match our target signal S . This guidance is applied for steps $t > T \cdot r_{start}$, where r_{start} is a ratio controlling the start of watermarking (e.g., 0.5).

At each guided denoising step t , the process is as follows (summarized in Algorithm 1):

1. **Predict Clean Latent:** Using the current latent z_t and the U-Net’s noise prediction $\mathcal{E}_\theta$, we first estimate the predicted clean latent $\hat{z}_0$ using Equation 3.
2. **Calculate and Apply Gradient Guidance:** We transform the predicted latent $\hat{z}_0$ into the frequency domain using 2D-DCT. A watermark loss $\mathcal{L}_{wm}$ is computed as the Mean Squared Error (MSE) between the DCT coefficients of $\hat{z}_0$ and our target signal S , exclusively at the locations specified by the mask M :

$$\mathcal{L}_{wm} = \|(DCT(\hat{z}_0) - S) \odot M\|_2^2 \quad (5)$$

The gradient of this loss with respect to $\hat{z}_0$ is then computed and used to steer the prediction in a single step:

$$\hat{z}_0^{\text{guided}} = \hat{z}_0 - \eta \nabla_{\hat{z}_0} \mathcal{L}_{wm} \quad (6)$$

where η is the guidance scale.

3. **Update Noise Prediction for Scheduler:** The standard scheduler requires a noise prediction to compute the next latent z_{t-1} . Since we have modified the predicted clean latent to $\hat{z}_0^{\text{guided}}$, we must derive a new, corresponding noise prediction. By rearranging Equation 3, we compute this guided noise $\mathcal{E}^{\text{guided}}$.

$$\mathcal{E}^{\text{guided}} = \frac{1}{\sqrt{1 - \bar{\alpha}_t}} (z_t - \sqrt{\bar{\alpha}_t} \hat{z}_0^{\text{guided}}) \quad (7)$$

This new noise, which now implicitly contains the watermark signal, is then used to guide the conditional

prediction $\mathcal{E}_\theta(\mathbf{z}_t, t, c)$. The guided prediction is then used within the standard Classifier-Free Guidance (CFG) framework in Eq. 2 to compute the final noise prediction for the current step.

3.4. Watermark Extraction

The extraction process, summarized in Algorithm 1, is direct and requires no model inversion. Given a potentially watermarked image $\mathbf{x}$, we first obtain its latent representation $\mathbf{z}_0$ using the VAE encoder. We then apply the 2D-DCT to $\mathbf{z}_0$ to get its frequency representation $\mathbf{F}$.

Concurrently, the secret key K is used to deterministically regenerate the binary mask $\mathbf{M}$, which identifies the exact frequency locations where the watermark bits were embedded. For each location (u, v) where the mask $\mathbf{M}(u, v)$ is 1, we extract a single bit by checking the sign of the corresponding DCT coefficient $\mathbf{F}(u, v)$: a positive sign decodes to '1', and a negative sign decodes to '0'. Since each bit was repeated multiple times, we use a majority vote to determine the final value for each bit, enhancing the robustness of the extraction. Finally, the collected sequence of bits is converted back into the message W .

3.5. Analysis: Why Progressive Guidance is Effective

A critical question is why our progressive guidance succeeds where a naive one-step injection into the final latent $\mathbf{z}_0$ fails catastrophically (see Figure 6). The answer lies in how GROW harmonizes with the iterative nature of the diffusion process, which can be attributed to two key principles:

Incremental Perturbations on the Manifold: The naive approach applies a single, large perturbation, pushing the final latent point far off the learned data manifold. In contrast, GROW applies a series of subtle, gradient-based updates at each step. These small adjustments gently steer the generation trajectory along the manifold's tangent space, preserving semantic integrity throughout the process.

Collaborative Refinement by the Denoiser: Crucially, the U-Net does not treat these small perturbations as noise to be eliminated. Instead, leveraging its powerful learned priors, it interprets them as high-frequency details to be rendered. The network collaboratively weaves the watermark's frequency signal into plausible image textures, such as the grain of wood or ripples in water rather than leaving them as artifacts.

4. Experiments

In this section, we conduct a series of experiments to comprehensively evaluate our proposed method, GROW. We aim to demonstrate its superiority over existing methods in terms of both imperceptibility and robustness.

Algorithm 1 GROW: Generation and Extraction

```

1: procedure GENERATE(prompt  $c$ , message  $W$ , key  $K$ , ...)
2:    $\mathbf{S} \leftarrow \text{GenerateTargetSignal}(W, K)$   $\triangleright$  Create target watermark signal
3:    $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$ 
4:   for  $t = T, \dots, 1$  do
5:      $\mathcal{E}_\theta \leftarrow \text{U-Net}(\mathbf{z}_t, t, c)$ 
6:      $\hat{\mathbf{z}}_0 \leftarrow \text{PredictCleanLatent}(\mathbf{z}_t, \mathcal{E}_\theta, t)$ 
7:     if  $t > T \cdot r_{start}$  then  $\triangleright$  If in guidance phase
8:        $\mathbf{g} \leftarrow \nabla_{\hat{\mathbf{z}}_0} \text{Loss}(\text{DCT}(\hat{\mathbf{z}}_0), \mathbf{S})$   $\triangleright$  Compute watermark gradient
9:        $\mathcal{E}_\theta \leftarrow \text{ApplyGuidance}(\mathcal{E}_\theta, \mathbf{g}, \eta)$   $\triangleright$  Update noise with guidance
10:    end if
11:     $\mathcal{E}^{\text{final}} \leftarrow \text{ApplyCFG}(\mathcal{E}_\theta, \dots)$ 
12:     $\mathbf{z}_{t-1} \leftarrow \text{SchedulerStep}(\mathcal{E}^{\text{final}}, t, \mathbf{z}_t)$ 
13:  end for
14:  return  $\mathbf{x} \leftarrow \text{VAE.decode}(\mathbf{z}_0)$ 
15: end procedure

16: procedure EXTRACT(image  $\mathbf{x}$ , key  $K$ )
17:    $\mathbf{M} \leftarrow \text{GenerateMask}(K)$   $\triangleright$  Regenerate watermark mask
18:    $\mathbf{z}_0 \leftarrow \text{VAE.encode}(\mathbf{x})$ 
19:    $\mathbf{F} \leftarrow \text{DCT}(\mathbf{z}_0)$ 
20:    $\text{bits} \leftarrow \text{Sign}(\mathbf{F}[\mathbf{M}])$   $\triangleright$  Extract bits from signs at masked locations
21:    $W \leftarrow \text{BitsToMessage}(\text{bits})$ 
22:   return  $W$ 
23: end procedure

```

4.1. Datasets and Metrics

Datasets. To ensure a comprehensive evaluation, We conduct tests on 1,000 generated images for each method from two diverse datasets: MS-COCO [22] for photorealistic images and Stable-Diffusion-Prompts [12] for artistic and complex concepts.

Metrics. We evaluate performance from two perspectives:

- **Imperceptibility:** We measure the visual quality of watermarked images using PSNR, SSIM, Fréchet Inception Distance (FID) [13], and LPIPS [39]. For PSNR, SSIM, higher is better. For FID, LPIPS, lower is better.
- **Robustness to attacks:** We evaluate the robustness of the embedded watermark by measuring the Message Accuracy (M-ACC) after applying various attacks. M-ACC is a strict metric that calculates the percentage of images from which the entire embedded message is recovered without any bit errors. This is significantly more challenging than Bit Accuracy and more meaningful for practical applications where message integrity is paramount.

The attacks include Rotation, JPEG compression, Cropping, Gaussian Blur, Gaussian Noise, Brightness adjustment and adversarial attack Diff-Pure [26] (illustrated in Figure 3).

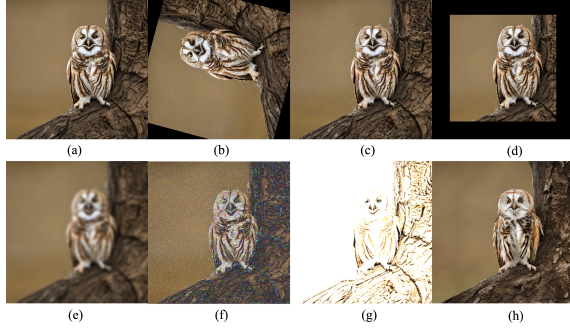


Figure 3. Watermarked image is attacked by different noise, including (a) with watermark, (b) Rotation 75, (c) JPEG 25, (d) Center Crop 75, (e) GauBlur $\sigma=4$, (f) Noise $\sigma=0.05$, (g) Brightness factor=6 and (h) adversarial attack $s=0.5$.

4.2. Implementation Details

All experiments are performed using the Stable Diffusion v2.1-base [29] model. We compare GROW against a wide range of watermarking methods, including post-processing based (DwtDctSvd [31], Hidden [41]) and other training-free methods (Tree-Ring [35], RingID [6], Gaussian Shading [37], WIND [1]).

All experiments are conducted using the PyTorch 1.13.0 framework, running on a single Tesla T4 GPU. For our method, we set the default parameters to $\alpha = 0.5$, $\eta = 100$, sample 50 steps and $r_{start} = 0.5$ to test. And our experiments embed a 16-bit watermark in the first channel of the VAE features. However, our framework is flexible, and other channels can also be utilized to embed watermarks, either independently or concurrently, allowing for higher-capacity watermarking.

4.3. Experimental Results

We conduct a comprehensive evaluation of GROW against other watermarking methods. Our experiments are performed on two widely-used datasets, MS-COCO and Stable-Diffusion-Prompts, to ensure a thorough assessment of imperceptibility, robustness, and efficiency.

Imperceptibility. A critical requirement for any watermarking technique is that it must not degrade the visual quality of the generated content. We evaluate imperceptibility using four standard metrics. As shown in Table 2, GROW demonstrates excellent performance in preserving image fidelity. High PSNR and SSIM scores, along with a low LPIPS value, indicate that our method does not introduce significant pixel-level distortions and that the watermarked images are structurally very similar to the originals.

Furthermore, GROW achieves a remarkable FID score of **12.32** on MS-COCO, signifying that the distribution of our watermarked images closely matches that of authentic, non-watermarked images, thus preserving the high generation quality of the underlying diffusion model. The qualitative comparisons in Figure 4 visually confirming the superior imperceptibility of our method.

Robustness. As presented in Table 2, GROW demonstrates superior performance across the board. On the Stable-Diffusion-Prompts dataset, GROW achieves an average robustness M-ACC of 97.6%, and on MS-COCO, it reaches an even higher 97.8%. This surpasses other methods, particularly recent training-free approaches. These results validate that our progressive guidance strategy effectively embeds a watermark that is highly durable against a wide array of common distortions.

Detection Efficiency. Many contemporary methods, such as Tree-Ring and WIND, rely on the computationally expensive DDIM inversion process for extraction, making them unsuitable for real-time applications. All methods generate a watermark-free image under the same conditions takes 20.99 seconds. As shown in Table 1, all compared methods exhibit similar total generation times, as the minor overhead introduced by each watermarking algorithm is negligible compared to the 50-step DDIM sampling process. For instance, GROW’s progressive guidance adds only 0.17 seconds. The crucial advantage lies in detection: GROW’s inversion-free mechanism is exceptionally efficient, achieving a detection time of just 0.24 seconds, which is nearly 100x faster than inversion-based techniques. This dramatic efficiency gain makes GROW a viable and practical solution for deployment in production environments where fast verification is essential.

Table 1. Time efficiency comparison. Our method avoids costly inversion, enabling fast watermark detection. All times are measured in seconds per image on a single Tesla T4 GPU.

Method	Generation Time (s) ↓	Detection Time (s) ↓
Tree-Ring	21.03/ 0.04	19.70
RingID	21.07/0.08	15.30
Gaussian Shading	21.24/0.25	12.50
WIND	21.17/0.18	21.70
GROW (Ours)	21.16/0.17	0.24

4.4. Hyperparameters Analysis

To analyze the influence of GROW’s key hyperparameters, we conduct a series of ablation studies. In each experiment, we vary a single parameter while keeping others at their default optimal values ($\alpha = 0.5$, $\eta = 100$, $r_{start} = 0.5$). The results, which plot the trade-off between robustness (Accu-

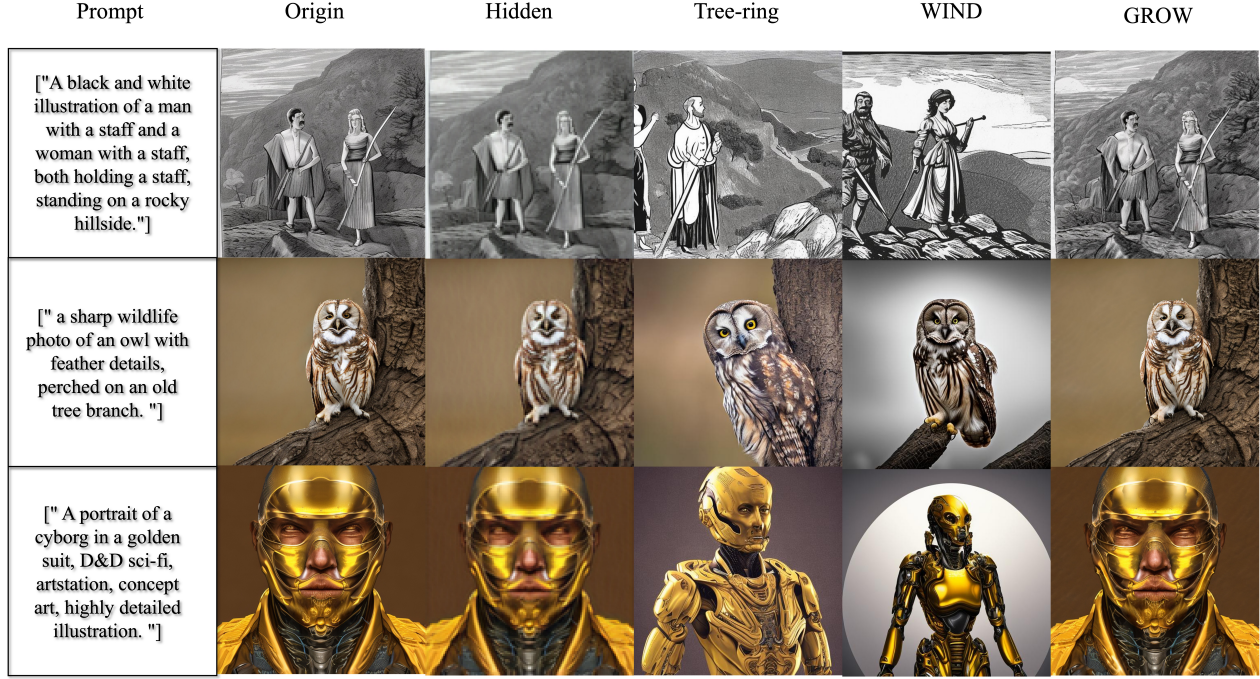


Figure 4. Visual comparisons of different methods using various prompts, including Hidden, Tree-Ring, WIND, and our GROW.

Table 2. Comparison of different watermarking methods across datasets. For imperceptibility metrics (PSNR, SSIM, FID, LPIPS), we report the performance on clean images. For robustness, we report the M-ACC (%) under various attacks. Best results are in **bold**, second best are underlined.

Method	Imperceptibility				Robustness to Attacks								
	PSNR↑	SSIM↑	FID↓	LPIPS↓	Clean	Rotation	JPEG	Crop	Blur	Noise	Brightness	Adversarial	Avg
<i>Dataset: Stable-Diffusion-Prompts</i>													
DwtDctSvd [31]	39.38	0.98	25.01	0.03	1.00	0.45	0.03	0.52	0.47	0.70	0.11	0.59	0.483
Hidden [41]	23.05	0.82	24.51	0.08	0.91	0.79	0.53	0.91	0.76	0.75	0.74	0.65	0.755
Tree-Ring [35]	15.25	0.54	25.93	0.42	0.99	0.02	0.42	0.04	0.61	0.53	0.42	0.85	0.485
RingID [6]	13.56	0.45	26.30	0.45	1.00	<u>1.00</u>	1.00	0.53	0.99	1.00	0.96	0.87	0.918
Gaussian Shading [37]	24.83	0.81	25.45	0.07	1.00	0.58	1.00	<u>0.92</u>	0.91	0.96	0.98	<u>0.91</u>	0.907
WIND [1]	13.16	0.47	<u>24.12</u>	0.39	<u>1.00</u>	0.72	<u>1.00</u>	0.86	<u>0.99</u>	<u>0.98</u>	0.97	0.89	<u>0.926</u>
GROW	<u>28.09</u>	<u>0.84</u>	18.90	<u>0.06</u>	1.00	1.00	1.00	1.00	1.00	0.93	<u>0.94</u>	0.94	0.976
<i>Dataset: MS-COCO</i>													
DwtDctSvd [31]	42.05	0.99	24.65	<u>0.04</u>	1.00	0.47	0.06	0.50	0.45	0.72	0.14	0.56	0.487
Hidden [41]	23.75	0.80	23.21	0.09	0.94	0.76	0.57	<u>0.93</u>	0.75	0.79	0.71	0.67	0.765
Tree-Ring [35]	16.48	0.27	22.38	0.43	0.99	0.05	0.46	0.08	0.64	0.50	0.44	0.87	0.504
RingID [6]	15.73	0.51	23.63	0.48	1.00	<u>1.00</u>	1.00	0.57	0.97	0.98	<u>0.97</u>	0.88	0.921
Gaussian Shading [37]	26.21	0.83	22.45	0.07	1.00	0.59	1.00	0.92	0.94	0.97	0.96	<u>0.93</u>	0.913
WIND [1]	7.44	0.26	21.89	0.44	<u>1.00</u>	0.75	<u>1.00</u>	0.84	<u>0.98</u>	<u>0.97</u>	0.96	0.90	<u>0.925</u>
GROW	<u>27.54</u>	<u>0.85</u>	12.32	0.05	1.00	1.00	1.00	1.00	0.99	0.91	0.98	0.95	0.978

racy) and image quality (FID), are shown in Figure 5.

Effect of Watermark Strength (α). As illustrated in Figure 5(a), the watermark strength α directly controls the trade-off between robustness and imperceptibility. Increasing α from 0.3 to 0.7 boosts the average M-ACC from 92% to 99%, as a stronger signal is more resilient to attacks. However, this comes at the cost of image quality, with the FID score degrading from 9.5 to 28.2.

Effect of Guidance Scale (η). The guidance scale η de-

termines how forcefully the watermark is injected at each step. Figure 5(b) shows that increasing η from 50 to 100 significantly improves M-ACC from 90% to 99%. Beyond this point, accuracy saturates. Interestingly, the FID score shows that image quality is optimal around $\eta = 125$ (FID 15.1) before slightly degrading, suggesting that an excessively high scale can be counterproductive.

Effect of Start Ratio (r_{start}). The start ratio r_{start} defines at which point in the diffusion process the watermark-

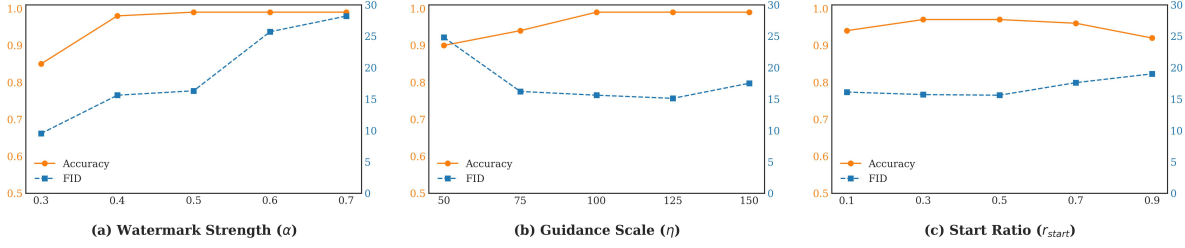


Figure 5. Ablation studies on key hyperparameters. We plot the trade-off between image quality (FID, blue line) and robustness (M-ACC, orange line). (a) Watermark strength α . (b) The guidance scale η . (c) Start Ratio.

ing guidance begins. As shown in Figure 5(c), applying the guidance earlier (i.e., a smaller r_{start}) enhances robustness. For example, starting at $r_{start} = 0.3$ yields a high M-ACC of 97%. This is because an earlier start allows the watermark to be integrated over more diffusion steps, weaving it more deeply into the image’s semantic structure. Conversely, starting too late (e.g., $r_{start} = 0.9$) results in a noticeable drop in robustness (92%).

4.5. Ablation Study on Progressive Guidance

To validate the necessity of our progressive guidance mechanism, we conduct a crucial ablation study against a simplified “one-step” baseline. This baseline injects the same DCT-based watermark pattern directly into the final latent space (z_0) just before decoding, using an identical watermark strength. The results are striking. As shown in the visual comparison in Figure 6, the one-step method completely corrupts the image content. We quantify this degradation on MS-COCO dataset using standard image quality metrics in Table 3. The one-step method leads to a catastrophic collapse in image quality, with FID soaring while PSNR and SSIM plummet. This experiment unequivocally demonstrates that progressively integrating the watermark is essential for embedding a strong signal without a disastrous loss of image quality.

Table 3. Quantitative comparison of image quality for the progressive guidance ablation study.

Method	FID ↓	PSNR ↑	SSIM ↑
One-Step	157.8	11.36	0.19
GROW	12.32	27.54	0.85

5. Limitations and Future Work

Despite its strong performance, GROW, as a training-free method operating in the DCT domain, has limitations. Its robustness can be compromised by severe geometric transformations like random cropping, and its capacity is inherently constrained by the fixed frequency domain structure. A promising future direction is to replace the static DCT do-

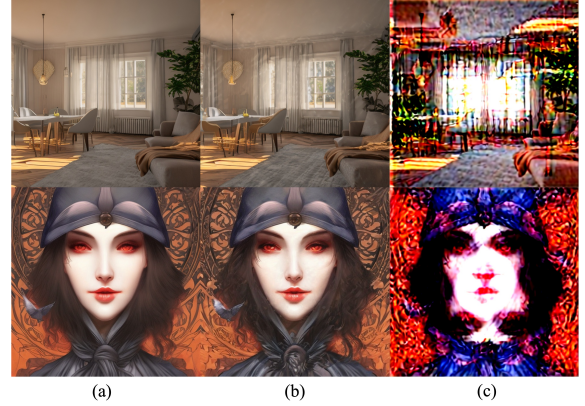


Figure 6. A visual comparison using prompts from the MS-COCO dataset. (a) non-watermarked image. (b) watermarked image by GROW. (c) one-step method.

main with a learnable, attack-resilient representation space. By training a dedicated encoder, one could generate a robust watermark embedding space optimized against a wide range of attacks. This learned space could then be integrated into our progressive guidance framework, potentially creating a new generation of watermarks with significantly enhanced security and robustness.

6. Conclusion

We proposed GROW, a novel and effective training-free watermarking paradigm for diffusion models. By reformulating watermark embedding as a progressive guidance task, our method iteratively injects a watermark signal into the frequency domain throughout the denoising process. This multi-step integration strategy ensures that the watermark is deeply fused with the generated content, achieving an excellent balance between imperceptibility and robustness. Extensive experiments validate that GROW significantly surpasses existing methods, demonstrating resilience to a wide range of attacks while preserving the high fidelity of the original diffusion model. Furthermore, by bypassing the computationally expensive inversion step required by many prior methods, GROW enables highly efficient watermark detection.

References

- [1] Kasra Arabi, Benjamin Feuer, R Teal Witter, Chinmay Hegde, and Niv Cohen. Hidden in the noise: Two-stage robust watermarking for images. *arXiv preprint arXiv:2412.04653*, 2024. 1, 2, 3, 6, 7
- [2] Kasra Arabi, R Teal Witter, Chinmay Hegde, and Niv Cohen. Seal: Semantic aware image watermarking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16196–16205, 2025. 1, 2, 3
- [3] Vishal Asnani, John Collomosse, Tu Bui, Xiaoming Liu, and Shruti Agarwal. Promark: Proactive diffusion watermarking for causal attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10802–10811, 2024. 1, 2
- [4] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pages 396–410, 2023. 1
- [5] Hai Ci, Yiren Song, Pei Yang, Jinheng Xie, and Mike Zheng Shou. Wmadapter: Adding watermark control to latent diffusion models. *arXiv preprint arXiv:2406.08337*, 2024. 1
- [6] Hai Ci, Pei Yang, Yiren Song, and Mike Zheng Shou. Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification. In *European Conference on Computer Vision*, pages 338–354. Springer, 2024. 1, 2, 6, 7, 3
- [7] Yingqian Cui, Jie Ren, Han Xu, Pengfei He, Hui Liu, Lichao Sun, Yue Xing, and Jiliang Tang. Diffusionshield: A watermark for copyright protection against generative diffusion models. *arXiv preprint arXiv:2306.04642*, 2023. 1, 2
- [8] Naresh Kumar Devulapally, Mingzhen Huang, Vishal Asnani, Shruti Agarwal, Siwei Lyu, and Vishnu Suresh Lokhande. Your text encoder can be an object-level watermarking controller. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16576–16585, 2025. 1
- [9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3
- [10] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22466–22477, 2023. 1, 2
- [11] Sam Gunn, Xuandong Zhao, and Dawn Song. An undetectable watermark for generative image models. *arXiv preprint arXiv:2410.07369*, 2024. 1, 2
- [12] Gustavosta. Stable-diffusion-prompts. <https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts>, 2024. Accessed: 2024-11-20. 5, 3
- [13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5
- [14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3
- [15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1
- [16] 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. 1
- [17] Runyi Hu, Jie Zhang, Yiming Li, Jiwei Li, Qing Guo, Han Qiu, and Tianwei Zhang. Supermark: Robust and training-free image watermarking via diffusion-based super-resolution. *arXiv preprint arXiv:2412.10049*, 2024. 1, 2
- [18] Kokil Jaidka, Tsuhan Chen, Simon Chesterman, Wynne Hsu, Min-Yen Kan, Mohan Kankanhalli, Mong Li Lee, Gyula Seres, Terence Sim, Araz Taeihagh, et al. Misinformation, disinformation, and generative ai: Implications for perception and policy. *Digital Government: Research and Practice*, 6(1):1–15, 2025. 1
- [19] Liangqi Lei, Keke Gai, Jing Yu, and Liehuang Zhu. Diffusetrace: A transparent and flexible watermarking scheme for latent diffusion model. *arXiv preprint arXiv:2405.02696*, 2024. 1, 2
- [20] Liangqi Lei, Keke Gai, Jing Yu, Liehuang Zhu, and Qi Wu. Secure and efficient watermarking for latent diffusion models in model distribution scenarios. *arXiv preprint arXiv:2502.13345*, 2025. 1
- [21] Kecen Li, Zhicong Huang, Xinwen Hou, and Cheng Hong. Gaussmarker: Robust dual-domain watermark for diffusion models. *arXiv preprint arXiv:2506.11444*, 2025. 1, 2
- [22] 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. 5, 3
- [23] Yugeng Liu, Zheng Li, Michael Backes, Yun Shen, and Yang Zhang. Watermarking diffusion model. *arXiv preprint arXiv:2305.12502*, 2023. 1, 2
- [24] Shilin Lu, Zihan Zhou, Jiayou Lu, Yuanzhi Zhu, and Adams Wai-Kin Kong. Robust watermarking using generative priors against image editing: From benchmarking to advances. *arXiv preprint arXiv:2410.18775*, 2024. 1
- [25] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1
- [26] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022. 6
- [27] Ahmad Rezaei, Mohammad Akbari, Saeed Ranjbar Alvar, Arezou Fatemi, and Yong Zhang. Lawa: Using latent space for in-generation image watermarking. In *European Conference on Computer Vision*, pages 118–136. Springer, 2024. 1
- [28] 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 (CVPR)*, pages 10684–10695, 2022. 3
- [29] 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, 2022. 1, 3, 6
- [30] Kareem Shehata, Aashish Kolluri, and Prateek Saxena. Clue-mark: Watermarking diffusion models using clwe. *arXiv preprint arXiv:2411.11434*, 2024. 1, 2
- [31] Frank Y Shih. *Digital watermarking and steganography: fundamentals and techniques*. CRC press, 2017. 1, 2, 6, 7
- [32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2, 3
- [33] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 1
- [34] 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
- [35] 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, 2, 3, 6, 7
- [36] Cheng Xiong, Chuan Qin, Guorui Feng, and Xinpeng Zhang. Flexible and secure watermarking for latent diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1668–1676, 2023. 1, 2
- [37] 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*, pages 12162–12171, 2024. 1, 2, 3, 6, 7
- [38] Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust invisible video watermarking with attention. *arXiv preprint arXiv:1909.01285*, 2019. 1, 2
- [39] 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*, pages 586–595, 2018. 5
- [40] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Ngai-Man Cheung, and Min Lin. A recipe for watermarking diffusion models. *arXiv preprint arXiv:2303.10137*, 2023. 1, 2
- [41] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 657–672, 2018. 1, 2, 6, 7

GROW: Watermark Generation with Progressive Guidance for Diffusion Models

Supplementary Material

7. Theoretical Analysis

7.1. Fundamental Difference from Tree-Ring based Watermarking

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 [35]. 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: Watermarking via 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: Watermarking via 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.

7.2. 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

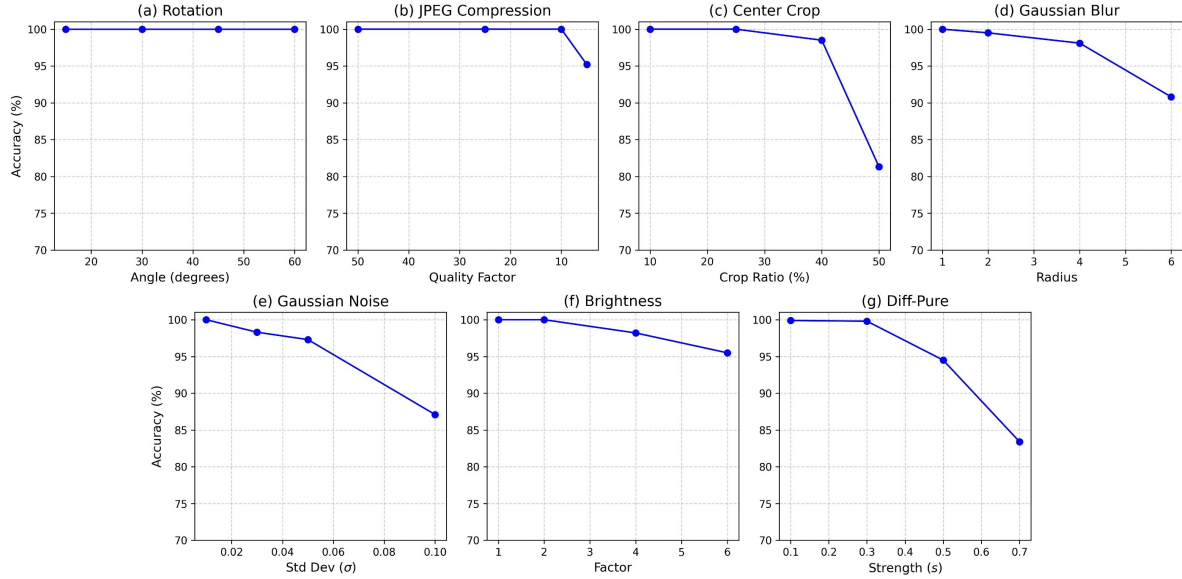


Figure 7. Robustness of GROW under varying attack intensities. Each plot displays the average extraction accuracy on images generated from the MS-COCO and Stable Diffusion Prompts datasets.

detection direct, efficient, and completely free of any inversion process. This shift in mechanism from passive to active is what unlocks the dramatic gains in detection speed while preserving the high robustness intrinsic to in-generation watermarking methods.

8. Additional Experiments

8.1. Additional Experiments on 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 7, 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 7(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 (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 Detection for Geometric Attacks. The apparent contradiction between the high M-ACC against rotation and cropping in Table 4 and the stated sensitivity

to geometric transformations in our Limitations section is resolved by our detection strategy. For practical applications, watermark detection is not a passive, one-shot process. Instead, we employ a **synchronized detection** mechanism that actively searches for the watermark under a set of expected transformations.

For Rotation: Our detector 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 detector 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 (e.g., 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. The detector analyzes multiple overlapping patches of the cropped image, assuming the watermark is likely located near the center. It extracts a potential message from each patch and, again, relies on the confidence score to identify the correct patch and message. As our experiments show, this method can tolerate cropping of up to 0.4 of the image area while maintaining high M-ACC.

Reconciling Robustness with Limitations. The “limitation” we refer to is therefore not an inability to handle geometric attacks, but rather the computational cost and constraints of this search-based detection. While highly effective, this synchronized detection is inherently slower

Table 4. Comparison of different watermarking methods across diffusion models. For imperceptibility (PSNR, SSIM, FID, LPIPS), we report performance on clean images. For robustness, we report the average Message Accuracy (M-ACC) across all attacks. Best results are in **bold**, second best are underlined.

Method	Imperceptibility				Robustness to Attacks								
	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	LPIPS $\downarrow$	Clean	Rotation	JPEG	Crop	Blur	Noise	Brightness	Adversarial	Avg
<i>Model: stable-diffusion-v1-5</i>													
Tree-Ring[35]	14.25	0.35	25.18	0.45	0.97	0.02	0.41	0.05	0.58	0.45	0.39	0.83	0.463
RingID[6]	13.85	0.44	26.45	0.49	1.00	<u>1.00</u>	1.00	0.51	0.94	0.99	<u>0.95</u>	0.85	0.898
Gaussian Shading[37]	<u>24.75</u>	<u>0.79</u>	24.85	<u>0.09</u>	1.00	0.52	1.00	<u>0.87</u>	0.89	0.94	0.93	<u>0.89</u>	0.883
WIND[1]	6.85	0.22	<u>23.15</u>	0.46	<u>1.00</u>	0.68	<u>1.00</u>	<u>0.79</u>	<u>0.96</u>	<u>0.95</u>	0.92	<u>0.86</u>	<u>0.901</u>
GROW	26.85	0.82	15.75	0.07	1.00	1.00	1.00	1.00	0.97	0.87	0.96	0.91	0.965
<i>Model: stable-diffusion-xl-base-1.0</i>													
Tree-Ring [35]	17.15	0.29	23.85	0.44	0.98	0.04	0.43	0.07	0.61	0.47	0.42	0.84	0.483
RingID[6]	16.25	0.53	24.75	0.47	1.00	<u>1.00</u>	1.00	0.55	0.95	0.98	<u>0.96</u>	0.86	0.913
Gaussian Shading[37]	<u>27.35</u>	<u>0.84</u>	23.15	<u>0.06</u>	1.00	0.57	1.00	<u>0.90</u>	0.92	0.96	0.95	<u>0.92</u>	0.915
WIND [1]	8.15	0.28	<u>22.45</u>	0.43	<u>1.00</u>	0.73	<u>1.00</u>	0.82	<u>0.97</u>	0.96	0.95	0.89	0.928
GROW	28.25	0.86	13.45	0.04	1.00	1.00	1.00	1.00	0.98	0.90	0.97	0.94	0.981

than a direct extraction and is limited to a predefined set of transformations. It cannot, for example, handle arbitrary, non-rigid warping or extreme, unconstrained random cropping where the watermark’s location is completely unknown. 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, our reported high M-ACC is a testament to a practical and effective detection system, while our stated limitation acknowledges the inherent theoretical constraints of the underlying watermark structure.

8.2. Additional Experiments on Different Diffusion Models

To assess the generalizability of our approach, we evaluated GROW and other leading methods on both stable-diffusion-v1-5 [28] and the more advanced stable-diffusion-xl-base-1.0 [28] (SDXL) model. The comprehensive results in Table 4 confirm that GROW’s advantages are model-agnostic. Across both architectures, GROW consistently achieves a superior balance of high imperceptibility (notably, the best FID scores) and state-of-the-art robustness (the highest average M-ACC). This suggests that the progressive guidance mechanism is a fundamental and portable technique for watermarking diffusion models.

8.3. Additional Experiments on 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 5, 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.

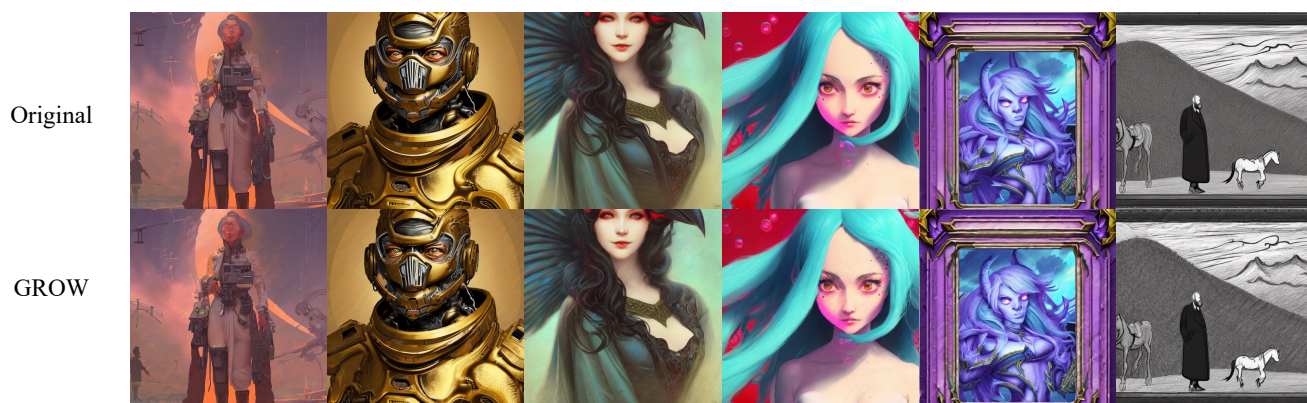
his experiment underscores the inherent constraints of its training-free nature. As discussed in our Limitations section, a promising future direction is to replace the static DCT space with a learnable representation, which could potentially support higher capacities with less impact on quality and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

Table 5. Impact of watermark capacity on image quality (FID) and robustness (Avg. M-ACC) on MS-COCO Dataset.

Capacity (bits)	FID $\downarrow$	Avg. M-ACC $\uparrow$
16	12.32	0.976
32	18.15	0.953
64	25.59	0.872

8.4. Additional Visual Results

To further showcase visual quality and generality, Figure 8 presents results on diverse datasets, including MS-COCO [22], WikiArt [34], and Stable Diffusion prompts [12]. The watermarked images are imperceptible, demonstrating that our method preserves image fidelity across varied subjects and styles.



(a) Stable Diffusion prompts



(b) MS-COCO



(c) WikiArt

Figure 8. Additional qualitative results of watermarked images generated by GROW on diverse datasets. The visual quality is consistently high, and the watermark is imperceptible.