PGID: Progressive Guided Inversion and Denoising for Robust Watermark Detection

Minh Quoc Duong, Chun Tong Lei, and Chun Pong Lau

City University of Hong Kong

Semantic Watermarking

Semantic watermarking is a state-of-the-art approach for diffusion models. It embeds a keyed pattern or message in the **initial noise** z_T used to generate the image.

Generation

Embed signal in noise z_T $\longrightarrow$ Denoise with the text prompt $\longrightarrow$ Decode the image

Detection

Encode the image $\longrightarrow$ Invert with DDIM to estimate $\hat{z}_T$ $\longrightarrow$ Test $\hat{z}_T$ for the signal

Detection relies on **diffusion inversion**, often implemented with DDIM, to run the generative process backward. The detector then checks the recovered noise $\hat{z}_T$ for the expected watermark signal.

Imprint removal and forgery

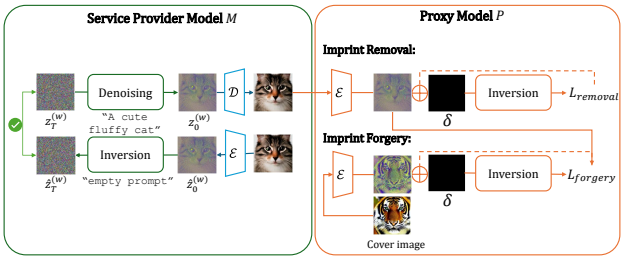
The attacker uses a proxy diffusion model to optimize a perturbation δ to the image latent z_0. Inverting $z_0 + \delta$ yields a misleading noise estimate $\hat{z}_T$. Decoding the perturbed latent produces the modified image.

Removal

Push a watermarked image's inverted noise away from its original estimate so the detector misses the watermark.

Forgery

Pull a cover image's inverted noise toward that of a watermarked reference, causing a false positive.



Imprint attacks from Müller et al. (CVPR 2025).

Our contributions

Main contribution: A method that detects removed watermarks and distinguishes forged watermarks from authentic ones.

Latent-space analysis

For the fixed signal embedded in an **initial watermarked noise** z_T, we formalize the watermarked and unwatermarked latent regions. We show how imprint attacks push image latents out of their original regions.

Training-free defense

We introduce PGID, a plug-and-play noise extraction framework that defends against both imprint attacks without additional training.

Evaluation

Comprehensive evaluation on Tree Ring, Gaussian Shading, PRC Watermark and T2SMark with two proxy attacker models.

Motivation

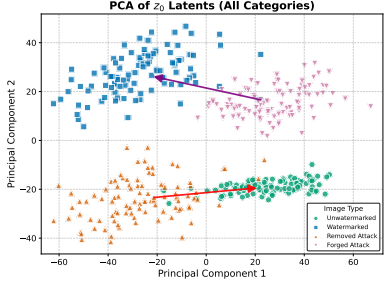
Conditioned on the watermark signal in an **initial watermarked noise** z_T^*, the latent space is partitioned into **two separable regions**: watermarked and unwatermarked.

Watermarked region $\mathcal{R}_w(z_T^*) = \{z_0 \in \mathcal{Z}_0 : \ D_{z_T^*}(I_{0\to T}(z_0)) = 1\}$.

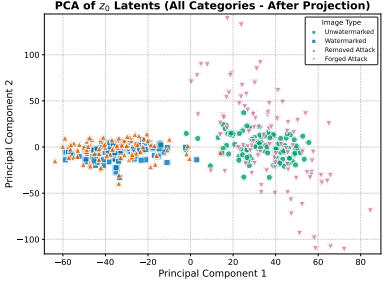
Unwatermarked region $\mathcal{R}_u(z_T^*) = \{z_0 \in \mathcal{Z}_0 : \ D_{z_T^*}(I_{0\to T}(z_0)) = 0\}$.

Removal shifts latents: $\mathcal{R}_w \to \mathcal{R}_u$ (Authentic watermark erased).
Forgery shifts latents: $\mathcal{R}_u \to \mathcal{R}_w$ (False watermark attribution).

Before PGID



After PGID



Method

Progressive Guided Inversion and Denoising (PGID)

Algorithm 1 PGID

Require: Suspect image x_{atk}, stopping timestep k, number of skip inversion steps s, guidance strength γ, and model M used for watermarked image generation with encoder $\mathcal{E}$ and decoder $\mathcal{D}$.

Stage I: Trajectory Pre-computation

$z_0^{atk} \leftarrow \mathcal{E}(x^{atk})$

Let $\{z_t^{atk}\}_{t=0}^{T} \leftarrow \mathcal{T}_{0\to T}(z_0^{atk})$ ▷ Pre-compute the standard inversion trajectory

Stage II: Progressive Latent Refinement

$\hat{z} \leftarrow z_0^{atk}$ ▷ Initialization

for $i = 1$ to k **do**

  for $t = 1$ to $i - s$ **do** ▷ Invert $i - s$ steps

    $\hat{z} \leftarrow Inverse(\hat{z}, t-1)$

    $\hat{z} \leftarrow \hat{z} + \gamma(\hat{z} - z_t^{atk})$

  end for

  for $t = i$ down to 1 **do** ▷ Denoise i steps

    $\hat{z} \leftarrow Denoise(\hat{z}, t)$

  end for

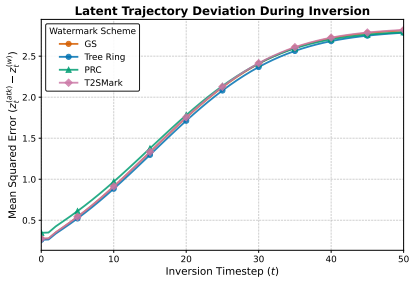
end for

$z_0 \leftarrow \hat{z}$ ▷ Finish refinement

Stage III: Final Inversion

$z_T \leftarrow I_{0\to T}(z_0)$ ▷ Final full inversion

return z_T



The removal attack introduces errors that accumulate during inversion. This motivates progressive latent refinement.

Main Evaluation Settings

Models and watermarking schemes

Service provider: Stable Diffusion v2.1 (SD 2.1). Proxy attackers: SD 2.1 and SD 1.5. We compare standard inversion with PGID (+P) for Tree Ring, Gaussian Shading, PRC, and T2SMark.

Images and attacks

Per scheme: 100 watermarked images, 100 generated unwatermarked images, and 100 removal-attacked images. For forgery, we use 100 forged images per proxy model and cover-image dataset (MS-COCO or ImageNet). Attacks run for 50, 100, or 150 optimization steps.

Metrics

Det. reports the true positive rate at a fixed false positive rate: 1% for Tree Ring and 10^{-6} for the other schemes.
Bit Acc. is the fraction of payload bits recovered correctly for schemes with bit messages.
AUC is the area under the ROC curve for the image groups compared. PRC uses image-dependent thresholds, so its AUC is omitted.

Results on imprint removal

PGID-R achieves **near-perfect watermark recovery** after imprint removal.

	Attacker: SD 2.1			Attacker: SD 1.5		
Method	Det. ↑	Bit Acc. ↑	AUC ↑	Det. ↑	Bit Acc. ↑	AUC ↑
TR	0.24/0.13/0.11	—	0.4085/0.1781/0.1363	0.25/0.13/0.10	—	0.4235/0.1942/0.1393
TR+P	**0.99/1.00/1.00**	—	**0.9998/0.9988/0.9998**	**1.00/1.00/1.00**	—	**1.0000/1.0000/0.9999**
GS	0.00/0.00/0.00	0.2582/0.0342/0.0077	0.0747/0.0000/0.0000	0.01/0.00/0.00	0.2891/0.0454/0.0114	0.1022/0.0000/0.0000
GS+P	**1.00/1.00/1.00**	**0.9975/0.9987/0.9981**	**1.0000/1.0000/1.0000**	**1.00/1.00/1.00**	**0.9964/0.9981/0.9968**	**1.0000/1.0000/1.0000**
PRC	0.00/0.00/0.00	0.5012/0.5038/0.5020	—	0.00/0.00/0.00	0.5011/0.4980/0.5039	—
PRC+P	**0.97/1.00/1.00**	**0.9590/1.0000/0.9847**	—	**0.97/1.00/0.98**	**0.8907/0.9841/0.9710**	—
T2SMark	0.50/0.94/1.00	0.5093/0.4977/0.4965	0.8695/0.9991/1.0000	0.44/0.94/1.00	0.5177/0.4964/0.4946	0.8806/0.9976/1.0000
T2SMark+P	**1.00/1.00/1.00**	**0.9999/0.9999/1.0000**	**1.0000/1.0000/1.0000**	**1.00/1.00/1.00**	**0.9998/0.9999/1.0000**	**1.0000/1.0000/1.0000**

Table 1: Results against imprint removal attack for four semantic watermarking schemes and their PGID-augmented variants. Results are reported for the attack at 50/100/150 optimization steps. AUC is calculated between removal-attacked and unwatermarked images. Higher AUC and Det. indicate higher robustness to the attack. The best result for each scheme is in bold.

Results on imprint forgery

PGID-F **rejects most forged images** in the MS-COCO evaluation.

	Attacker: SD 2.1			Attacker: SD 1.5		
Method	Det. ↓	Bit Acc. ↓	AUC ↓	Det. ↓	Bit Acc. ↓	AUC ↓
TR	1.00/1.00/1.00	—	1.0000/1.0000/1.0000	1.00/1.00/1.00	—	1.0000/1.0000/1.0000
TR+P	**0.02/0.04/0.07**	—	**0.5472/0.6048/0.6728**	**0.04/0.08/0.09**	—	**0.5515/0.6139/0.6625**
GS	1.00/1.00/1.00	0.9987/0.9994/0.9995	1.0000/1.0000/1.0000	1.00/1.00/1.00	0.9983/0.9991/0.9995	1.0000/1.0000/1.0000
GS+P	**0.00/0.00/0.00**	**0.5358/0.5764/0.6099**	**0.7773/0.9244/0.9582**	**0.00/0.00/0.06**	**0.5410/0.5849/0.6139**	**0.7969/0.9108/0.9577**
PRC	1.00/1.00/1.00	1.0000/1.0000/1.0000	—	0.94/1.00/1.00	0.8788/0.9895/1.0000	—
PRC+P	**0.00/0.00/0.01**	**0.4952/0.5000/0.5021**	—	**0.00/0.00/0.03**	**0.4980/0.5027/0.5000**	—
T2SMark	1.00/1.00/1.00	0.9999/1.0000/1.0000	1.0000/1.0000/1.0000	1.00/1.00/1.00	0.9998/1.0000/1.0000	1.0000/1.0000/1.0000
T2SMark+P	**0.00/0.00/0.01**	**0.4966/0.5037/0.5119**	**0.5335/0.6868/0.7626**	**0.00/0.00/0.01**	**0.4921/0.5055/0.5116**	**0.5783/0.6964/0.7551**

Table 2: Results against imprint forgery attack on MS-COCO dataset for four semantic watermarking schemes and their PGID-augmented variants. Results are reported for the attack at 50/100/150 optimization steps. AUC is calculated between forged and unwatermarked images. Lower AUC and Det. indicate higher robustness to the attack. The best result for each scheme is in bold.

Results on imprint forgery

Similar results are achieved using **ImageNet** as the cover image dataset.

	Attacker: SD 2.1			Attacker: SD 1.5		
Method	Det. ↓	Bit Acc. ↓	AUC ↓	Det. ↓	Bit Acc. ↓	AUC ↓
TR	0.99/1.00/1.00	—	0.9994/0.9998/0.9998	0.99/1.00/1.00	—	0.9990/1.0000/1.0000
TR+P	**0.06/0.08/0.12**	—	**0.5642/0.6324/0.6915**	**0.05/0.09/0.11**	—	**0.5653/0.6295/0.7030**
GS	1.00/1.00/1.00	0.9984/0.9992/0.9996	1.0000/1.0000/1.0000	1.00/1.00/1.00	0.9972/0.9986/0.9991	1.0000/1.0000/1.0000
GS+P	**0.00/0.04/0.08**	**0.5479/0.5895/0.6210**	**0.7801/0.8968/0.9288**	**0.01/0.06/0.14**	**0.5512/0.5970/0.6273**	**0.8044/0.8971/0.9367**
PRC	0.99/0.99/1.00	0.9846/0.9951/0.9948	—	0.89/0.98/0.99	0.8815/0.9655/0.9744	—
PRC+P	**0.00/0.00/0.04**	**0.5012/0.4967/0.5048**	—	**0.00/0.00/0.00**	**0.4944/0.5038/0.4986**	—
T2SMark	1.00/1.00/1.00	0.9999/1.0000/1.0000	1.0000/1.0000/1.0000	1.00/1.00/1.00	0.9999/1.0000/1.0000	1.0000/1.0000/1.0000
T2SMark+P	**0.00/0.02/0.06**	**0.5000/0.5143/0.5250**	**0.5339/0.6426/0.7452**	**0.00/0.04/0.07**	**0.5027/0.5114/0.5423**	**0.5593/0.6295/0.7030**

Table 3: Results against imprint forgery attack on ImageNet dataset for four semantic watermarking schemes and their PGID-augmented variants. Results are reported for the attack at 50/100/150 optimization steps. AUC is calculated between forged and unwatermarked images. Lower AUC and Det. indicate higher robustness to the attack. The best result for each scheme is in bold.

Authentic Watermark Robustness and Separability

PGID-F **preserves detection of authentic watermarks** while keeping them distinguishable from forgeries.

			AUC ↑ (MS-COCO)		AUC ↑ (ImageNet)	
Method	Det. ↑	Bit Acc. ↑	Atk. Model: SD 2.1	Atk. Model: SD 1.5	Atk. Model: SD 2.1	Atk. Model: SD 1.5
TR+P	0.96	—	0.9930/0.9902/0.9851	0.9930/0.9886/0.9854	0.9913/0.9874/0.9805	0.9916/0.9883/0.9802
GS+P	1.00	0.8947	1.0000/1.0000/1.0000	1.0000/1.0000/1.0000	1.0000/0.9996/0.9986	1.0000/0.9992/0.9972
PRC+P	0.93	0.8452	—	—	—	—
T2SMark+P	1.00	0.9591	1.0000/1.0000/1.0000	1.0000/1.0000/0.9998	1.0000/0.9996/0.9970	1.0000/0.9983/0.9940

Table 4: Robustness and separability of authentic watermarks after passing through the PGID-F defense. AUC is calculated for the attacks at 50, 100, and 150 steps (50/100/150). Higher Detection Rate, Bit Accuracy, and AUC indicate that the defense preserves the authentic watermark while effectively separating it from forged images.

Thank you so much for listening!