FARI: Robust One-Step Inversion for Watermarking in Diffusion Models

Jindong Yang[1,2], **Han Fang**[3*], **Weiming Zhang**[1,2], **Nenghai Yu**[1,2], **Kejiang Chen**[1,2*]
[1]University of Science and Technology of China
[2]Anhui Province Key Laboratory of Digital Security
[3]National University of Singapore
Feel free to contact us via email at: dx929@mail.ustc.edu.cn

ICLR 2026

Abstract

Inversion-based watermarking is a promising approach to authenticate diffusion-generated images, yet practical use is bottlenecked by inversion that is both slow and error-prone. While the primary challenge in the watermarking setting is robustness against external distortions, existing approaches over-optimize internal truncation error, and because that error scales with the sampler step size, they are inherently confined to high-NFE (number of function evaluations) regimes that cannot meet the dual demands of speed and robustness. In this work, we have two key observations: (i) the inversion trajectory has markedly lower curvature than the forward generation path does, making it highly compressible and amenable to low-NFE approximation; and (ii) in inversion for watermark verification, the trade-off between speed and truncation error is less critical, since external distortions dominate the error. A faster inverter provides a dual benefit: it is not only more efficient, but it also enables end-to-end adversarial training to directly target robustness, a task that is computationally prohibitive for the original, lengthy inversion trajectories. Building on this, we propose **FARI** (**F**ast **A**symmetric **R**obust **I**nversion), a one-step inversion framework paired with lightweight adversarial LoRA fine-tuning of the denoiser for watermark extraction. While consolidation slightly increases internal error, FARI delivers large gains in both speed and robustness: with ~20 minutes of fine-tuning on a single NVIDIA RTX A6000 GPU, it surpasses 50-step DDIM inversion on watermark-verification robustness while dramatically reducing inference time.

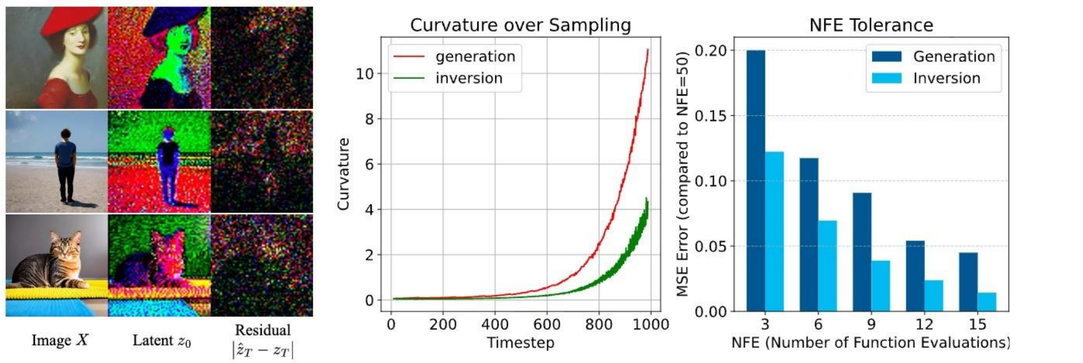
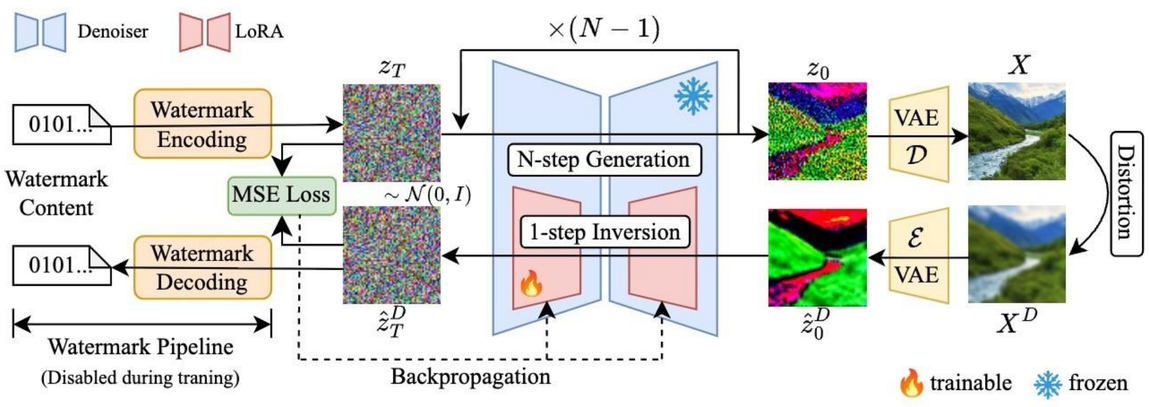


Figure 1: **Left**: Visualization of the inversion error, where latent vectors are projected down to 3 channels via PCA for display. **Middle**: Curvature of generation and inversion trajectories across diffusion timesteps. Discrete curvature estimated using 100 unconditionally generated images from Stable Diffusion v2.1. **Right**: The resulting error for generation and inversion when reducing the NFE, compared with a 50-step NFE baseline.

Framework



Results

DM	Methods	NFE	Clean	Adv.	Jpeg	R.Crop	R.Drop	Resize	G.Blur	M.Blur	G.Noise	S&P	Bright
Bit Accuracy of Gaussian Shading Watermark													
SD v1.5	DDIM	50	0.9999	0.9777	0.9889	**0.9781**	0.9736	0.9975	0.9873	0.9983	0.9609	0.9354	0.9567
		1	0.9999	0.9376	0.9703	0.8859	0.8808	0.9906	0.9585	0.9934	0.9398	0.9105	0.9085
	EDICT	50	**1.0000**	0.9637	0.9786	0.9656	0.9568	0.9969	0.9807	0.9985	0.9390	0.9124	0.9450
	BELM	50	0.9991	0.9465	0.9847	0.8960	0.8958	0.9939	0.9617	0.9956	0.9355	0.9275	0.9278
	AMED†	2	**1.0000**	0.9656	0.9807	0.9528	0.9462	0.9970	0.9808	0.9989	0.9587	0.9346	0.9410
	LCM-LoRA	2	0.9999	0.9541	0.9819	0.9352	0.9308	0.9924	0.9668	0.9955	0.9311	0.9030	0.9504
	DMD2	1	0.9988	0.9287	0.9760	0.8446	0.8241	0.9792	0.9209	0.9836	0.9252	0.9007	0.9336
	FARI(Ours)	1	**1.0000**	**0.9834**	**0.9935**	0.9777	**0.9761**	**0.9990**	**0.9957**	**0.9992**	**0.9836**	**0.9649**	**0.9612**
SD v2.1	DDIM	50	**1.0000**	0.9755	0.9892	0.9752	0.9669	0.9980	0.9860	0.9991	0.9590	0.9373	0.9447
		1	0.9987	0.9359	0.9755	0.8841	0.8637	0.9748	0.9173	0.9819	0.9280	0.9069	0.9284
	EDICT	50	**1.0000**	0.9585	0.9773	0.9639	0.9558	0.9963	0.9805	0.9983	0.9396	0.9104	0.9395
	BELM	50	0.9990	0.9411	0.9847	0.8923	0.8938	0.9933	0.9602	0.9952	0.9333	0.9269	0.9334
	AMED†	2	1.0000	0.9662	0.9813	0.9559	0.9495	0.9970	0.9805	0.9989	0.9585	0.9357	0.9384
	ExactDPM	> 150	**1.0000**	0.9670	0.9831	0.9675	0.9599	0.9974	0.9815	0.9987	0.9653	0.9241	0.9354
	FARI(Ours)	1	**1.0000**	**0.9824**	**0.9941**	**0.9771**	**0.9700**	**0.9992**	**0.9956**	**0.9994**	**0.9815**	**0.9659**	**0.9588**
TPR@1e-3 of Tree-Ring Watermark													
SD v1.5	DDIM	50	**1.000**	0.949	0.989	**1.000**	**1.000**	0.999	0.996	**1.000**	0.636	0.946	0.972
		1	**1.000**	0.863	0.905	0.602	0.649	**1.000**	0.994	**1.000**	0.891	0.990	0.737
	EDICT	50	**1.000**	0.942	0.975	0.998	**1.000**	0.999	0.992	0.997	0.605	0.954	0.962
	BELM	50	0.933	0.592	0.768	0.032	0.054	0.889	0.873	0.865	0.384	0.852	0.608
	AMED†	2	**1.000**	0.909	0.939	0.947	0.936	0.999	0.995	0.999	0.618	0.912	0.835
	LCM-LoRA	2	**1.000**	0.875	0.914	0.996	0.991	0.999	0.987	0.999	0.331	0.812	0.850
	DMD2	1	**1.000**	0.760	0.709	0.116	0.473	0.996	0.971	0.995	0.913	0.985	0.678
	FARI(Ours)	1	**1.000**	**0.997**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**0.980**	**1.000**	**0.992**
SD v2.1	DDIM	50	**1.000**	0.962	0.993	**1.000**	**1.000**	**1.000**	0.997	**1.000**	0.726	0.982	0.960
		1	**1.000**	0.896	0.896	0.709	0.845	**1.000**	0.993	0.999	0.903	0.991	0.729
	EDICT	50	**1.000**	0.946	0.980	0.984	0.985	0.997	0.990	0.998	0.681	0.969	0.933
	BELM	50	0.882	0.543	0.721	0.001	0.000	0.803	0.787	0.801	0.417	0.812	0.541
	AMED†	2	**1.000**	0.926	0.966	0.957	0.983	**1.000**	0.998	**1.000**	0.656	0.983	0.795
	ExactDPM	> 150	**1.000**	0.906	0.991	0.571	0.847	**1.000**	0.999	**1.000**	0.835	0.992	0.915
	FARI(Ours)	1	**1.000**	**0.997**	**0.999**	**1.000**	**1.000**	**1.000**	**0.999**	**1.000**	**0.979**	**0.999**	**0.993**