Paper ID: IB-2262

DOSMark: Robust and Diverse Watermarking for Diffusion Models via Dynamic Orthogonal Subspaces

Shaoshuai Zhang, Xiaorui Dai, Bo Wang*

School of Information and Communication Engineering,Dalian University of Technology

Introduction

The rapid advancement of diffusion models has substantially improved the quality of high-fidelity image generation, while it also introduces significant concerns regarding intellectual property infringement and the malicious dissemination of fabricated content. Generative watermarking, particularly the Noise-as-Watermark (NaW) method, addresses this issue by encoding watermark information directly into the initial Gaussian noise space. Nevertheless, existing NaW approaches, which rely on discrete sign-constraint mechanisms, encounter inherent limitations in balancing robustness and generative diversity. To overcome this limitation, we introduce DOSMark, a training-free watermarking framework that reformulates the embedding process through geometric representations on the Gaussian manifold.

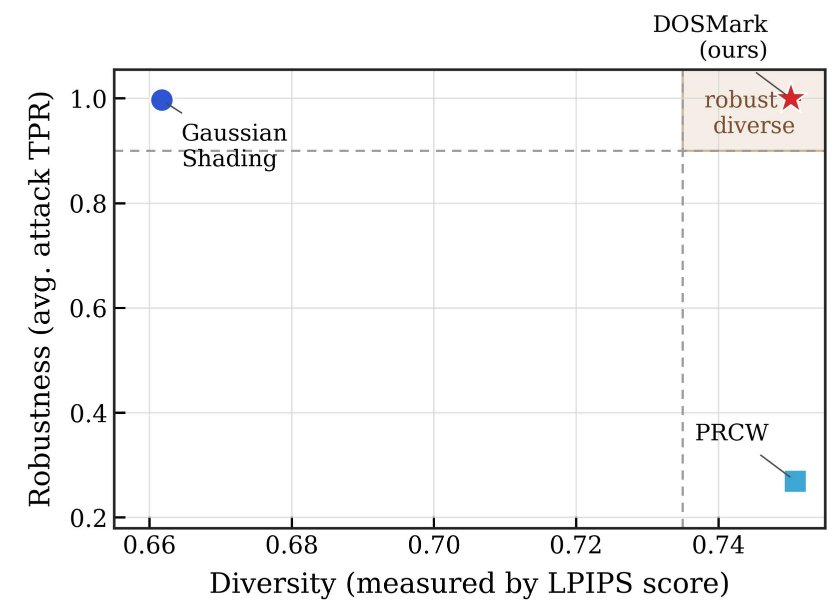
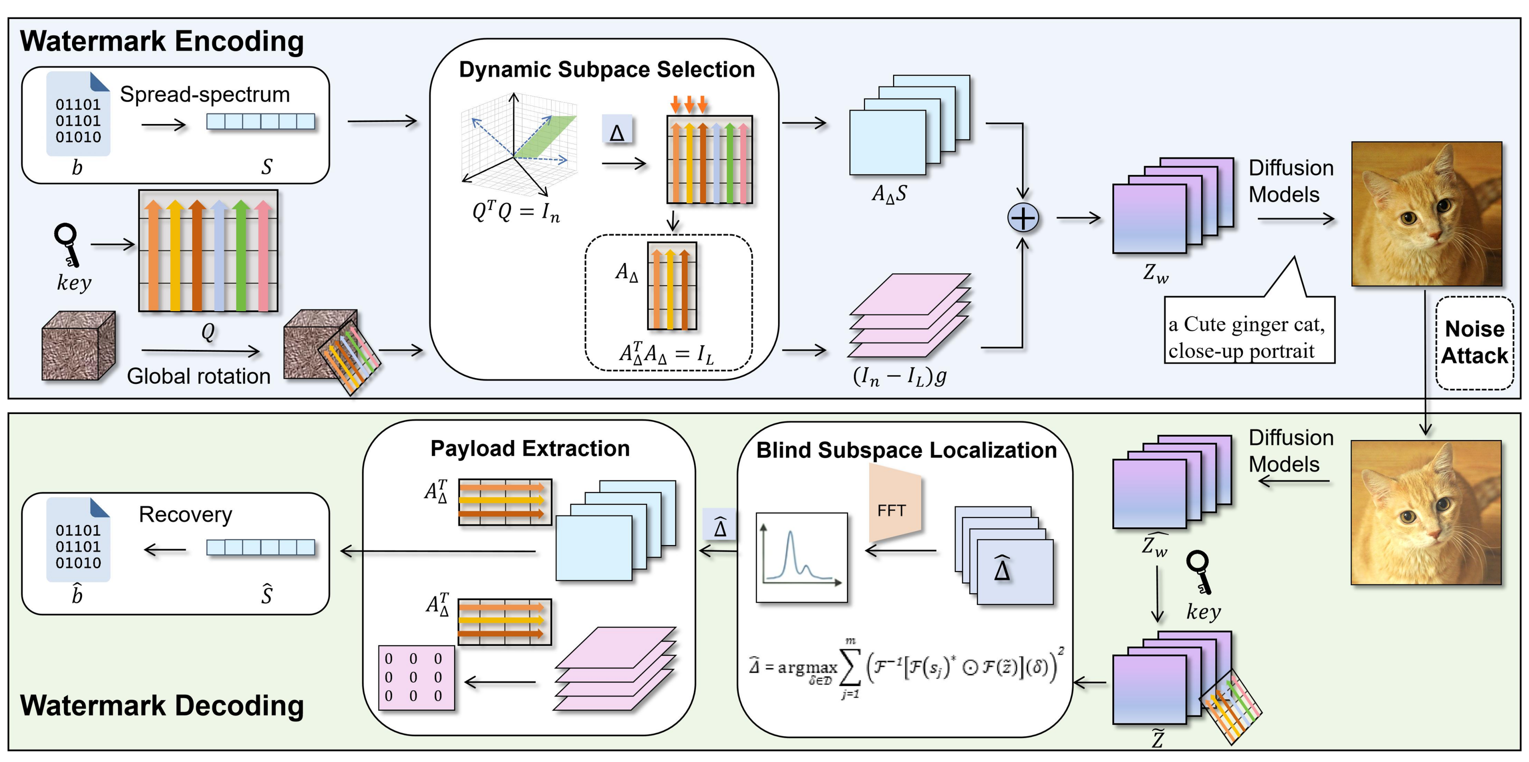


Fig. Quantitative comparison of NaW methods. DOSMark breaks the robustness-diversity trade-off by geometrically reformulating the embedding process.

Highlights

- **Inherent Conflict:** Current Noise-as-Watermark (NaW) methods inherently suffer from a fundamental conflict between generative diversity and robustness due to their reliance on discrete encoding schemes.
- **Geometric Reformulation:** DOSMark redefines the embedding paradigm from a geometric perspective, projecting watermarks into sample-dependent orthogonal subspaces governed by a dynamic subspace index.
- **Orthogonal Decoupling:** Our method strictly maintains standard Gaussian randomness in the corresponding orthogonal complement, effectively decoupling structured payloads from inherent Gaussian stochasticity.

Methodology



$$Z_w = A_\Delta S + (I_n - P_\Delta)g \qquad \hat{S} = A_{\hat{\Delta}}^T \hat{Z}_w = S + A_{\hat{\Delta}}^T \epsilon_{inv}$$

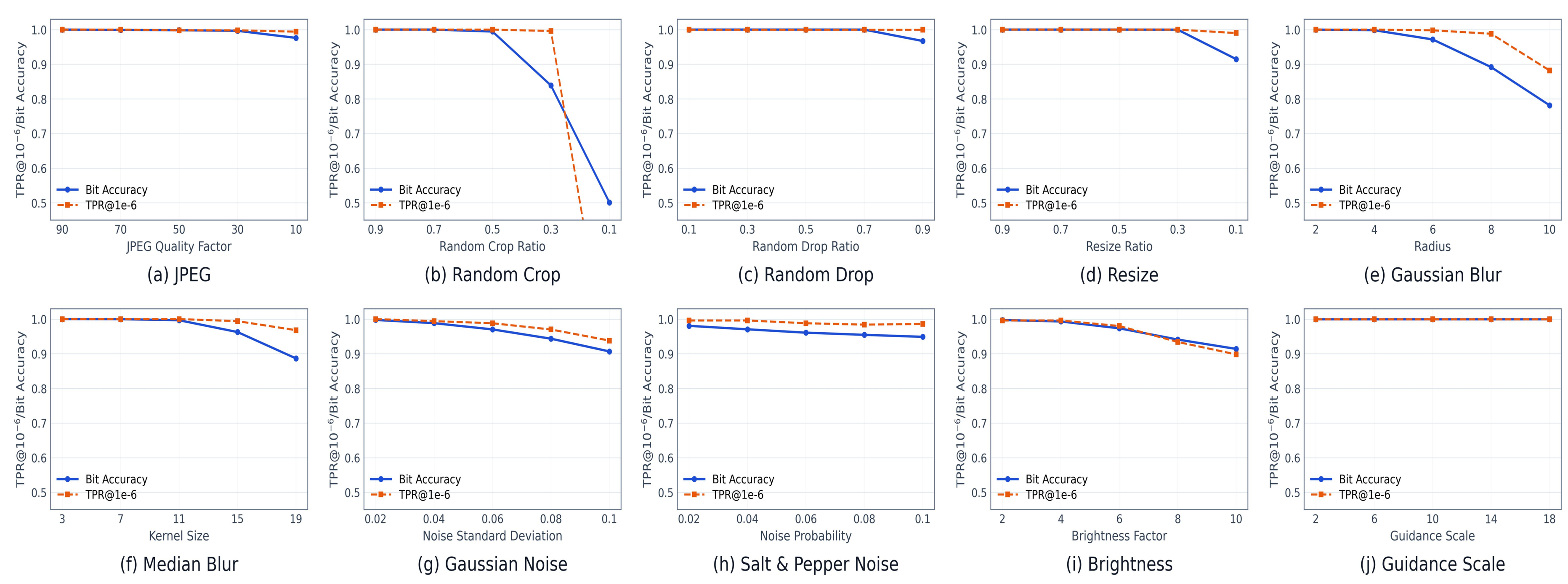
- $A_\Delta S$: Watermark subspace carrying structured payload
- $(I_n - P_\Delta)g$: Orthogonal complement maintaining Gaussian randomness
- $A_{\hat{\Delta}}^T$: Orthogonal projection perfectly suppresses background noise.
- $A_{\hat{\Delta}}^T \epsilon_{inv}$: Inversion errors are effectively mitigated.

Conclusion

- DOSMark achieves state-of-the-art traceability against complex adversarial distortions while maintaining high visual fidelity and cross-sample diversity.
- By dynamically retrieving orthogonal subspaces, the method prevents distribution collapse and demonstrates a superior robustness-diversity trade-off.
- Future work will explore optimized inversion trajectories and lightweight spatial mappings to reduce the computational overhead of high-dimensional matrix generation.

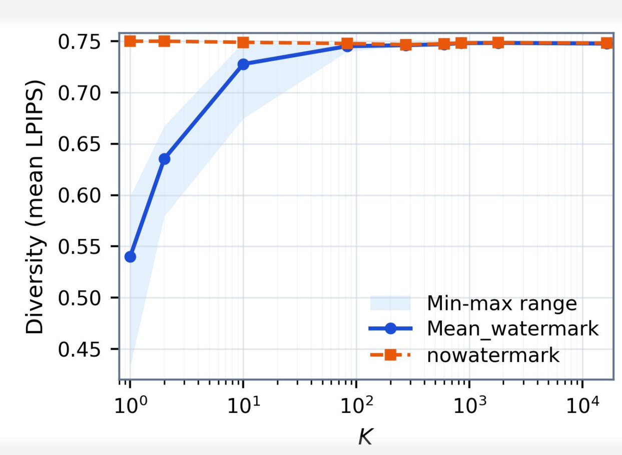
Main Results

Method	TPR	Bit Acc.	Diversity↑	CLIP Score (t ↓)	FID (t ↓)
SD V2.1 [2]	–	–	0.7506	0.3303 ± 0.0009 (–)	52.8307 ± 0.6180 (–)
dwtDct [23]	0.858/0.168	0.8303/0.5568	–	0.3287 ± 0.0010 (3.7550)	52.2416 ± 0.6269 (2.1162)
dwtDctSvd [23]	1.000/0.454	0.9998/0.6540	–	0.3290 ± 0.0013 (2.5928)	51.9466 ± 0.6300 (3.1677)
RivaGAN [24]	0.993/0.519	0.9974/0.8843	–	0.3276 ± 0.0011 (6.0294)	51.9689 ± 0.6446 (3.0518)
Stable Signature [8]	1.000/0.751	0.9930/0.7013	0.7451	0.3265 ± 0.0010 (8.8624)	54.2219 ± 0.6224 (5.0160)
TreeRing [14]	1.000/0.914	–	0.7304	0.3283 ± 0.0007 (5.5324)	53.7353 ± 0.4528 (3.7341)
GaussianShading [19]	1.000/0.997	1.0000/0.9482	0.6617	0.3310 ± 0.0011 (1.4830)	54.3249 ± 1.0039 (4.0083)
PRCW [20]	1.000/0.270	0.9287/0.6325	**0.7508**	0.3300 ± 0.0010 (0.7027)	52.8323 ± 0.5015 (**0.0065**)
T2SMark [21]	1.000/0.997	1.0000/0.9769	0.7452	0.3299 ± 0.0008 (0.9935)	52.9112 ± 0.5196 (0.3155)
DOSMark	**1.000/1.000**	**1.0000/0.9942**	0.7502	0.3301 ± 0.0011 (**0.4263**)	52.7707 ± 0.2563 (0.2832)



- **State-of-the-Art Comprehensive Traceability**: Quantitative evaluations confirm that, whether on clean samples or averaged across nine complex attacks, our method consistently achieves optimal bit recovery accuracy without compromising generative quality.
- **Exceptional Robustness Against Severe Distortions**: DOSMark maintains a remarkably stable performance lower bound—achieving high TPR and bit accuracy—across multiple attack types, varying intensities, and different guidance scales.

Capacity (bits)	Bit Acc.
256	1.0000/0.9942
384	1.0000/0.9859
512	1.0000/0.9765
768	1.0000/0.9542
1024	0.9999/0.9330



- **Module Effectiveness:** Dynamic subspace indexing eliminates distributional stagnation to restore baseline diversity, while orthogonal embedding robustly supports 1024-bit payloads.
- **Lossless Visual Quality:** DOSMark strictly preserves fine-grained textures and cross-sample diversity, completely avoiding structural artifacts and mode collapse.

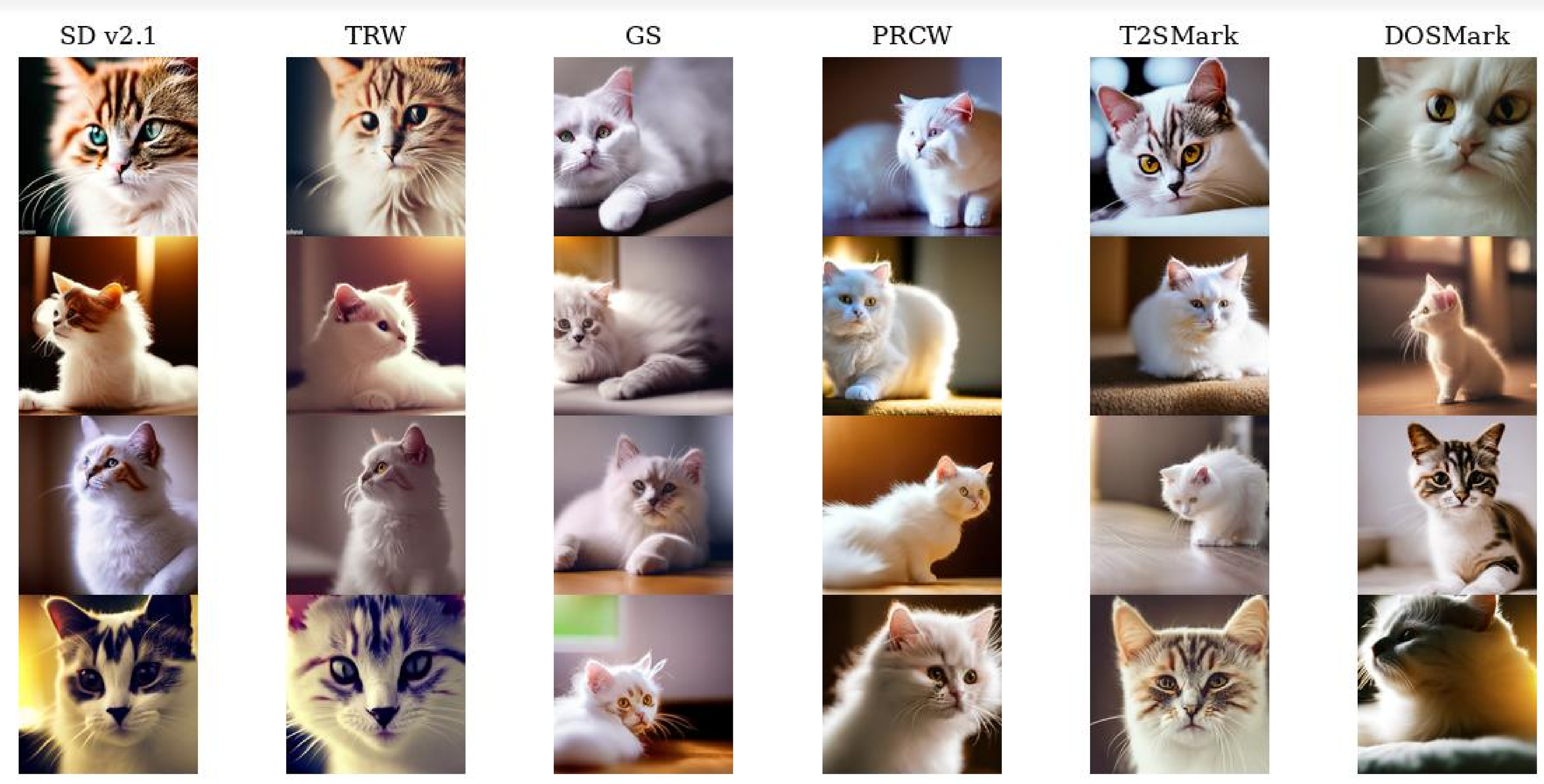


Fig. Qualitative comparison of visual fidelity.