
T2SMark: Balancing Robustness and Diversity in Noise-as-Watermark for Diffusion Models

Anonymous Author(s)

Affiliation

Address

email

A Supplementary Material

Impact of the Two-Stage Framework

To assess the impact of the two-stage framework, we evaluate T2SMark both with and without it. All experiments use Stable Diffusion v2.1 [1] under the same settings as the main paper. For robustness testing, we sample 500 prompts from the Stable-Diffusion-Prompts training split¹; for generation diversity, we use 1,000 prompts from the same dataset and report LPIPS scores [2]. Results are shown in Table 1.

Enabling the two-stage framework reduces adversarial bit accuracy from 0.9868 to 0.9754, since reserving latent dimensions to encrypt the session key cuts redundancy and causes cascading errors—any mistake in decoding the session key invalidates the second-stage decoding. However, it substantially increases generation diversity, confirming its importance for balancing robustness and diversity. In contrast, without two-stage encryption, T2SMark suffers from the same fixed-codeword limitation as Gaussian Shading [3] and overconcentrates energy in the high-energy tail region, whose position in the latent vector is entirely key-dependent, significantly reducing diversity.

Table 1: Performance of T2SMark both with and without the two-stage framework.

	Bit Acc. (Clean/Adv.)	Diversity ↑
w/o two-stage	1.0000/0.9868	0.5689
w/ two-stage	1.0000/0.9754	0.6746

References

- [1] 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, June 2022.
- [2] 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.
- [3] 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 (CVPR)*, pages 12162–12171, June 2024.

¹Stable-Diffusion-Prompts