Semantic Watermarking Reinvented: Enhancing Robustness and Generation Quality with Fourier Integrity

Sung Ju Lee, Nam Ik Cho

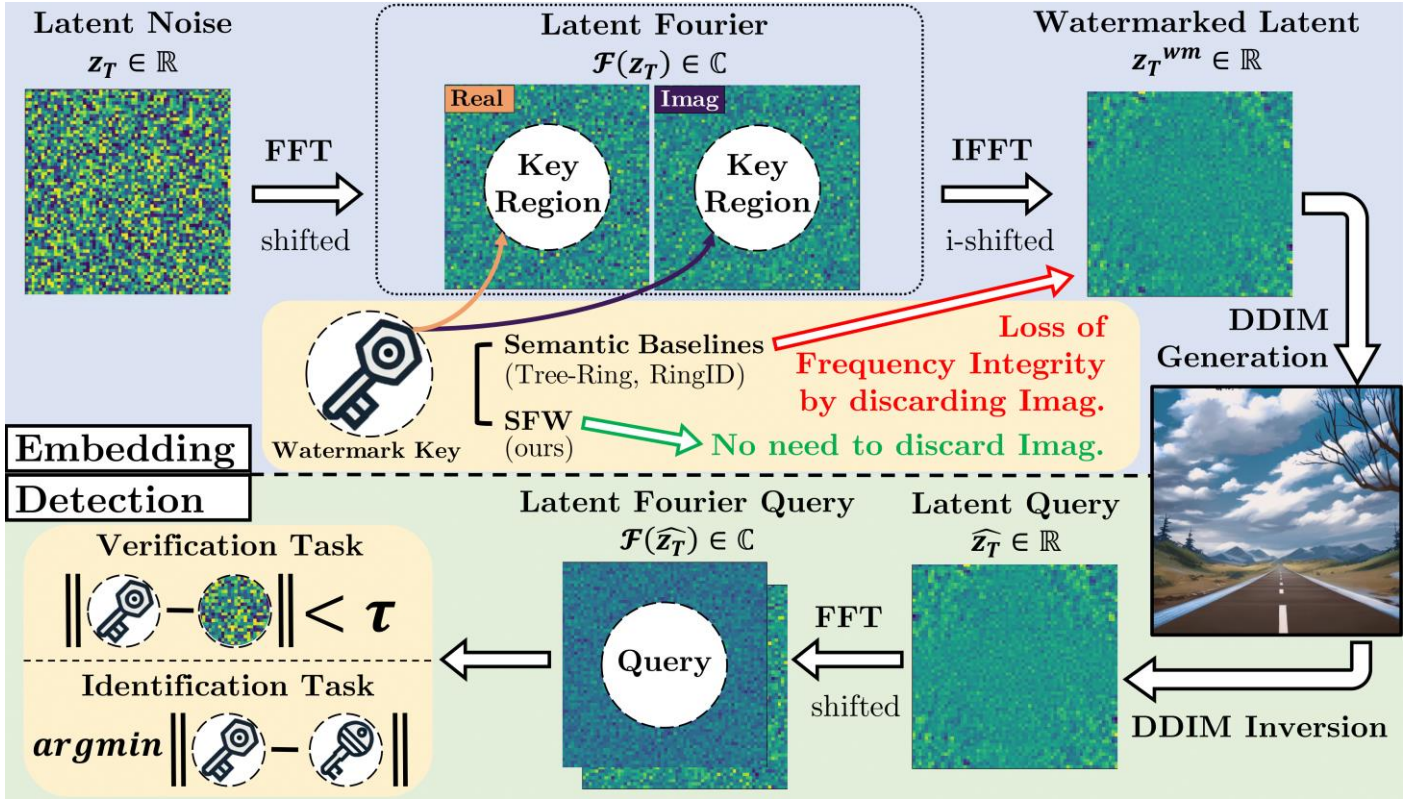
Project

Code

ICCV OCT 19-23, 2025 HONOLULU HAWAII

CV

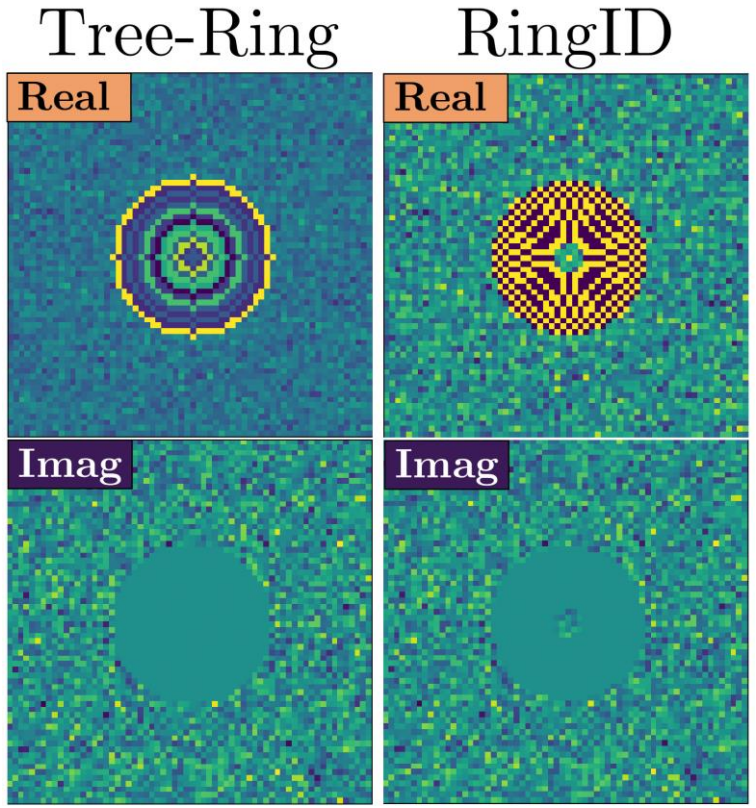
- Motivation
  - The rise of latent diffusion models has made image generation widely accessible, but it also introduces challenges in content attribution and copyright.
  - Semantic watermarking in the latent space offers a robust solution, especially due to its resilience against **regeneration attacks**, which often break pixel-level watermarks.



End-to-end pipeline *of semantic watermarking using the merged-in-generation scheme.*

- The Core Problem: Loss of Frequency Integrity
  - Existing methods often discard the imaginary component in the Fourier domain, breaking the **frequency integrity**.
  - This violates the Hermitian symmetry required for real-valued signals, distorting the statistical structure of the latent noise.
  - The consequences are severe:
    1. Weakened detection robustness
    2. Degraded generative quality

- Our Goal:
  To propose a new approach that preserves frequency integrity, which enables reliable and high-quality watermarking.



Missing imaginary patterns reveal ***frequency loss*** *in baselines.*

1. Hermitian Symmetric Fourier Watermarking (SFW)

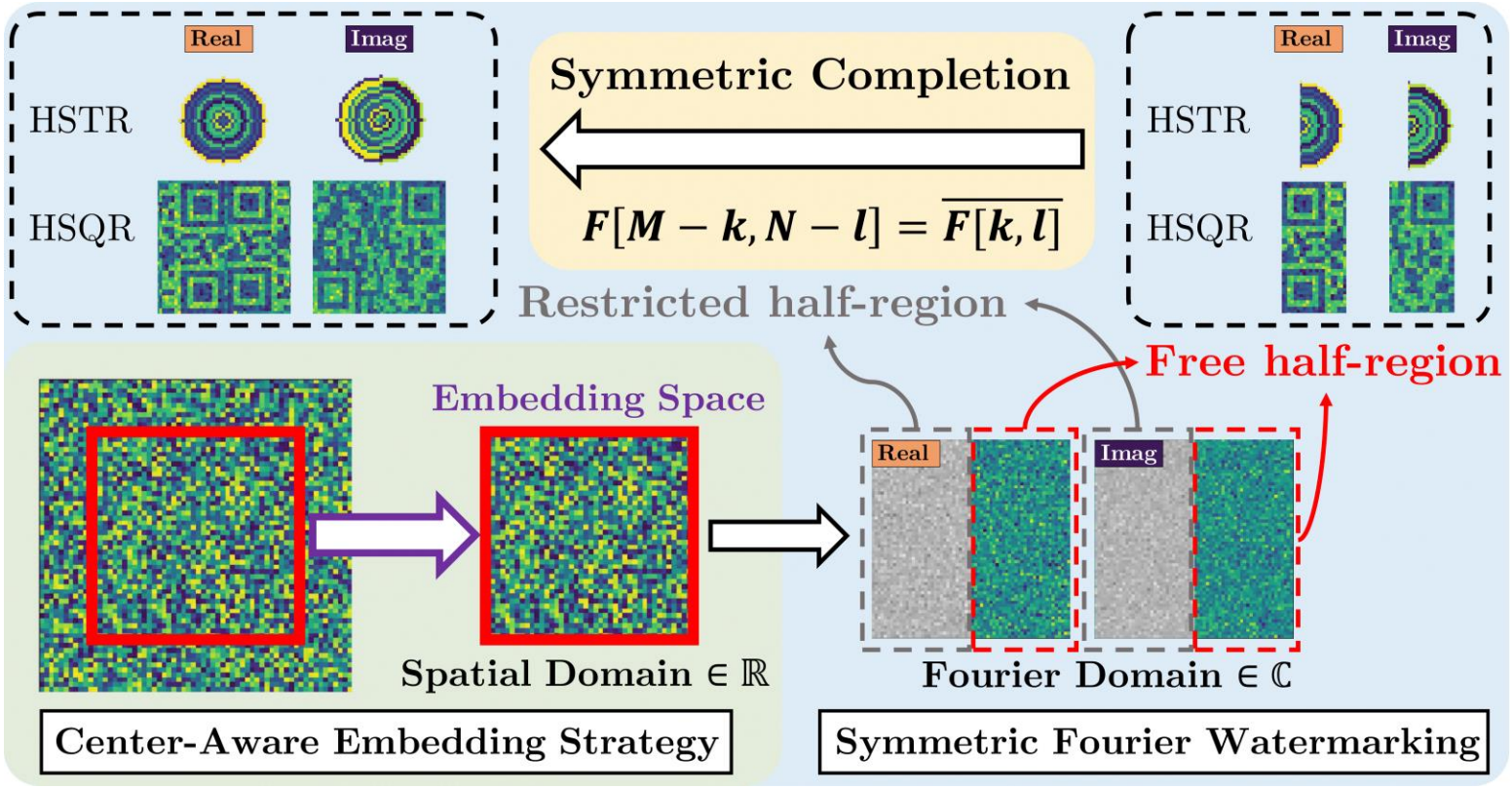
- Enforces Hermitian symmetry to preserve frequency integrity, utilizing both real and imaginary components to boost detection performance without sacrificing image quality.

2. Center-Aware Embedding Strategy

- Embeds watermarks in the stable central region of the latent space to significantly improve robustness against cropping attacks.

3. HSQR: Hermitian Symmetric QR-Code

- Extends SFW by splitting a QR code across the real and imaginary components, achieving high accuracy, capacity, and image quality.



Our Implementations:

- **HSTR:** Applies our framework to the Tree-Ring structure.
- **HSQR:** A novel, high-capacity watermark using a QR code design.

Results (1): State-of-the-Art Detection Robustness

Our methods, HSTR and HSQR, consistently achieve state-of-the-art detection robustness across a wide range of attacks, including signal processing, regeneration, and cropping.

- **Verification & Identification:**
  Both methods significantly outperform prior techniques in verification *(Is a watermark present?)* and identification *(What is the message?)* tasks.

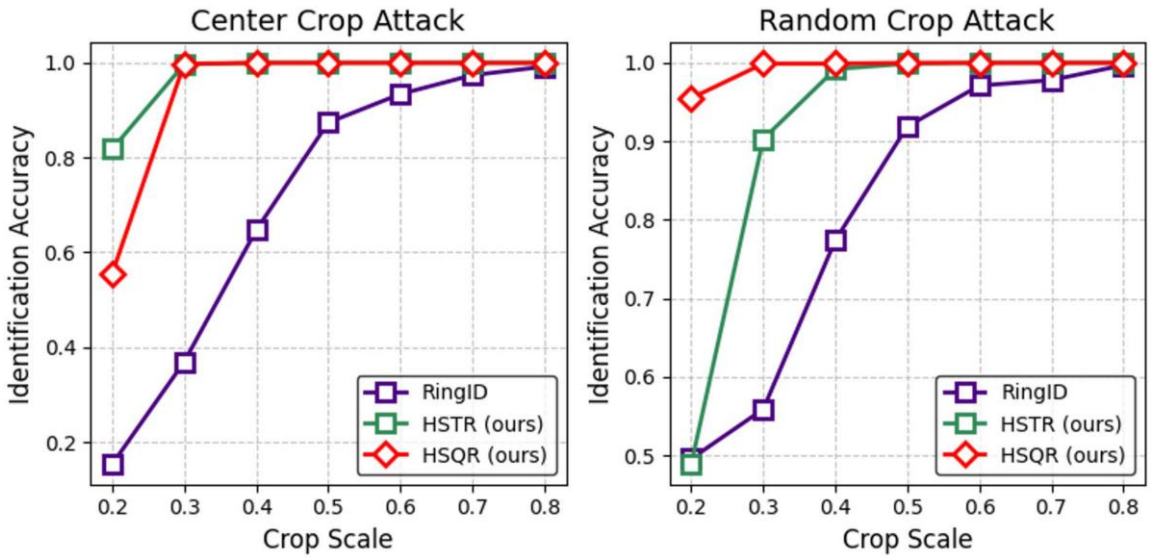
- **Superior Crop Resilience:**
  Thanks to the Center-Aware Embedding strategy, our methods maintain exceptionally high accuracy even under severe cropping attacks, where baselines like RingID fail.

Table 1. Verification performance of different watermarking methods under various attacks. Bit Accuracy is used for bitstream-based methods (DwtDct, DwtDctSvd, RivaGAN, S.Sign), while TPR@1%FPR is used for semantic methods. Best performances are highlighted. Our methods show superior detection accuracy and robustness against signal processing distortions, regeneration, and cropping attacks.

Datasets	Methods	No Attack	Signal Processing Attack						Regeneration Attack			Cropping Attack		Avg
		Clean	Bright.	Cont.	JPEG	Blur	Noise	BM3D	VAE-B	VAE-C	Diff.	C.C.	R.C.	
MS-COCO	DwtDct	0.863	0.572	0.522	0.516	0.677	0.859	0.532	0.523	0.521	0.519	0.779	0.810	0.637
	DwtDctSvd	1.000	0.555	0.473	0.602	1.000	1.000	0.784	0.648	0.596	0.644	0.744	0.861	0.742
	RivaGAN	0.999	0.862	0.986	0.821	0.998	0.969	0.934	0.570	0.532	0.608	0.991	0.995	0.857
	S.Sign.	0.995	0.894	0.978	0.806	0.911	0.721	0.838	0.717	0.715	0.478	0.987	0.991	0.836
	Tree-Ring	0.957	0.463	0.900	0.548	0.934	0.412	0.815	0.509	0.536	0.543	0.509	0.734	0.655
	Zodiac	0.998	0.843	0.998	0.973	0.998	0.880	0.997	0.944	0.958	0.972	0.989	0.995	0.962
	HSTR (ours)	1.000	0.899	1.000	0.994	1.000	0.806	0.999	0.973	0.982	0.997	1.000	1.000	0.971
	RingID	1.000	0.988	1.000	1.000	1.000	0.987	1.000	0.992	1.000	1.000	1.000	1.000	0.997
	HSQR (ours)	1.000	0.991	1.000	1.000	1.000	0.983	1.000	0.992	1.000	1.000	1.000	1.000	0.997
SD-Prompts	DwtDct	0.819	0.557	0.516	0.506	0.685	0.822	0.530	0.513	0.512	0.509	0.723	0.794	0.624
	DwtDctSvd	1.000	0.537	0.459	0.610	0.999	0.998	0.859	0.659	0.620	0.623	0.743	0.860	0.747
	RivaGAN	0.991	0.823	0.963	0.810	0.988	0.961	0.915	0.572	0.535	0.567	0.980	0.983	0.841
	S.Sign.	0.994	0.899	0.967	0.769	0.888	0.742	0.809	0.677	0.671	0.493	0.983	0.990	0.824
	Tree-Ring	0.944	0.471	0.894	0.466	0.912	0.423	0.802	0.509	0.514	0.543	0.469	0.749	0.641
	Zodiac	0.998	0.748	0.999	0.979	0.999	0.903	1.000	0.940	0.975	0.958	0.994	0.996	0.957
	HSTR (ours)	1.000	0.742	1.000	0.990	1.000	0.850	1.000	0.983	0.987	0.999	1.000	1.000	0.963
	RingID	1.000	0.972	1.000	1.000	1.000	0.988	1.000	0.996	1.000	1.000	1.000	1.000	0.996
	HSQR (ours)	1.000	0.955	1.000	0.999	1.000	0.992	1.000	0.996	0.999	1.000	1.000	1.000	0.995
DiffusionDB	DwtDct	0.842	0.563	0.515	0.509	0.672	0.829	0.526	0.513	0.514	0.512	0.723	0.801	0.627
	DwtDctSvd	0.998	0.558	0.463	0.593	0.997	0.995	0.830	0.658	0.608	0.621	0.742	0.860	0.744
	RivaGAN	0.987	0.839	0.960	0.790	0.985	0.937	0.893	0.553	0.518	0.556	0.974	0.979	0.831
	S.Sign.	0.990	0.890	0.967	0.787	0.889	0.726	0.819	0.690	0.687	0.496	0.981	0.986	0.826
	Tree-Ring	0.940	0.487	0.889	0.454	0.904	0.392	0.799	0.454	0.503	0.454	0.499	0.715	0.622
	Zodiac	0.992	0.752	0.988	0.933	0.988	0.834	0.984	0.911	0.926	0.903	0.971	0.985	0.931
	HSTR (ours)	0.999	0.792	0.996	0.981	0.996	0.792	0.991	0.968	0.969	0.989	1.000	1.000	0.956
	RingID	1.000	0.989	1.000	1.000	1.000	0.963	1.000	0.995	0.999	1.000	1.000	1.000	0.995
	HSQR (ours)	1.000	0.977	1.000	0.999	1.000	0.974	0.999	0.997	0.999	1.000	1.000	1.000	0.995

Table 2. Identification accuracy for different watermarking methods, evaluated across multiple attack conditions. Perfect Match Rate is used for bitstream-based methods. The best performance for each item is highlighted with shading. HSQR achieves state-of-the-art performance, while HSTR consistently outperforms other Gaussian radius-based semantic baselines, particularly under cropping attacks.

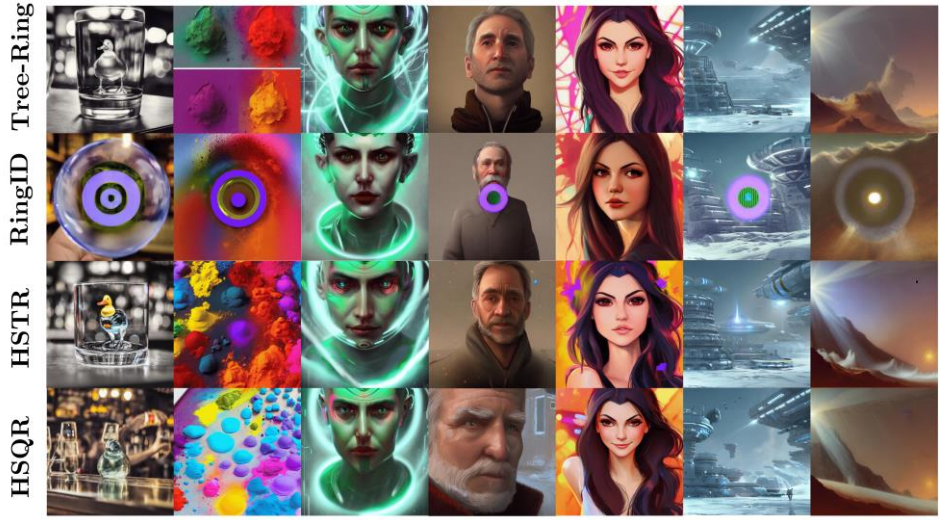
Datasets	Methods	No Attack	Signal Processing Attack						Regeneration Attack			Cropping Attack		Avg
		Clean	Bright.	Cont.	JPEG	Blur	Noise	BM3D	VAE-B	VAE-C	Diff.	C.C.	R.C.	
MS-COCO	DwtDct	0.466	0.044	0.000	0.000	0.038	0.442	0.000	0.000	0.000	0.000	0.000	0.000	0.083
	DwtDctSvd	1.000	0.044	0.019	0.000	0.999	0.998	0.037	0.000	0.000	0.003	0.000	0.000	0.258
	RivaGAN	0.974	0.260	0.772	0.023	0.961	0.686	0.348	0.000	0.000	0.000	0.852	0.909	0.482
	S.Sign.	0.873	0.177	0.563	0.000	0.036	0.010	0.007	0.000	0.000	0.000	0.709	0.802	0.265
	Tree-Ring	0.303	0.087	0.207	0.072	0.256	0.030	0.162	0.083	0.072	0.054	0.009	0.033	0.114
	Zodiac	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	HSTR (ours)	1.000	0.714	0.999	0.886	0.998	0.460	0.972	0.833	0.831	0.971	1.000	1.000	0.889
	RingID	1.000	0.875	1.000	0.975	1.000	0.919	0.996	0.978	0.970	0.998	0.874	0.978	0.964
	HSQR (ours)	1.000	0.958	1.000	0.994	1.000	0.901	0.999	0.980	0.987	0.999	1.000	1.000	0.985
SD-Prompts	DwtDct	0.285	0.024	0.000	0.000	0.017	0.276	0.000	0.000	0.000	0.000	0.000	0.000	0.050
	DwtDctSvd	0.993	0.028	0.011	0.000	0.982	0.979	0.085	0.000	0.000	0.007	0.000	0.000	0.257
	RivaGAN	0.878	0.213	0.613	0.009	0.857	0.657	0.304	0.000	0.000	0.000	0.736	0.768	0.420
	S.Sign.	0.813	0.263	0.420	0.000	0.021	0.015	0.006	0.000	0.000	0.000	0.576	0.716	0.236
	Tree-Ring	0.288	0.094	0.189	0.051	0.235	0.034	0.159	0.079	0.076	0.056	0.012	0.041	0.110
	Zodiac	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	HSTR (ours)	1.000	0.655	0.999	0.863	0.999	0.555	0.980	0.846	0.847	0.973	1.000	1.000	0.893
	RingID	1.000	0.885	1.000	0.976	0.998	0.886	0.993	0.980	0.973	0.995	0.876	0.981	0.962
	HSQR (ours)	1.000	0.930	1.000	0.994	1.000	0.942	0.999	0.991	0.997	1.000	1.000	1.000	0.988
DiffusionDB	DwtDct	0.357	0.037	0.000	0.000	0.034	0.320	0.000	0.000	0.000	0.000	0.000	0.000	0.062
	DwtDctSvd	0.990	0.036	0.019	0.000	0.975	0.959	0.081	0.000	0.000	0.001	0.000	0.000	0.255
	RivaGAN	0.858	0.213	0.625	0.020	0.848	0.615	0.221	0.000	0.000	0.000	0.756	0.780	0.411
	S.Sign.	0.798	0.207	0.472	0.000	0.027	0.005	0.005	0.000	0.000	0.000	0.643	0.738	0.241
	Tree-Ring	0.280	0.095	0.190	0.059	0.233	0.037	0.145	0.081	0.072	0.050	0.013	0.039	0.108
	Zodiac	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	HSTR (ours)	0.996	0.721	0.992	0.854	0.989	0.563	0.958	0.830	0.821	0.952	0.996	0.996	0.889
	RingID	1.000	0.895	1.000	0.947	0.996	0.871	0.992	0.968	0.958	0.990	0.875	0.984	0.956
	HSQR (ours)	1.000	0.954	1.000	0.988	1.000	0.906	0.998	0.982	0.991	0.994	1.000	1.000	0.984



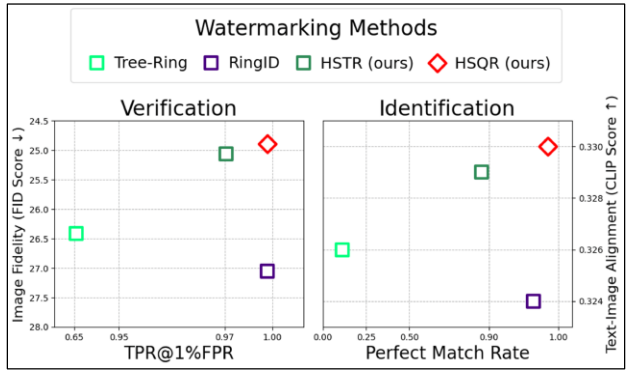
- **Quantitative Proof:** By preserving frequency integrity, HSTR and HSQR achieve the best FID and CLIP scores, indicating higher-quality and more prompt-aligned images.
- **Qualitative Difference:** High-energy patterns like RingID's introduce noticeable *ring-like artifacts*. Our SFW-based methods produce clean, artifact-free images, successfully resolving the critical trade-off.

Table 3. Generative quality evaluation of watermarking methods based on FID (MS-COCO ground truth) and CLIP score. The best performance for each item is highlighted with shading, while bold text specifically marks the low CLIP score in RingID. Our proposed methods preserve frequency integrity, achieving the best balance between watermark robustness and generative performance, whereas RingID introduces visible artifacts, compromising perceptual quality. *Vrf.* and *Idf.* denote the average detection performance in verification and identification tasks, respectively.

Semantic Methods		FID ↓	CLIP ↑	*Vrf.*	*Idf.*
Merged in Generation	Tree-Ring	26.418	0.326	0.655	0.114
	RingID	27.052	**0.324**	0.997	0.964
	HSTR (ours)	25.062	0.329	0.971	0.889
	HSQR (ours)	24.895	0.330	0.997	0.985



- **Optimal Balance:** Our work successfully resolves the critical trade-off between **(1)** watermark robustness and **(2)** image fidelity.
  - Unlike high-energy methods that trade quality for robustness, our approach achieves superior detection without compromising image fidelity.



Our ablation studies confirm the critical contributions of each component and the superior scalability of our HSQR design.

- **SFW is Critical:** Without enforcing Hermitian symmetry, detection performance drops significantly, proving that SFW's frequency integrity is the key to robustness.

Table 4. Ablation study on detection performance based on frequency integrity (✓ or ×) and the number of detection region usage (1: real only, 2: real & imaginary). *Vrf.* and *Idf.* represent average detection performance in verification (TPR@1% FPR) and identification (accuracy), respectively. ΔL_1^* denotes the normalized ΔL_1 metric, indicating detection effectiveness.

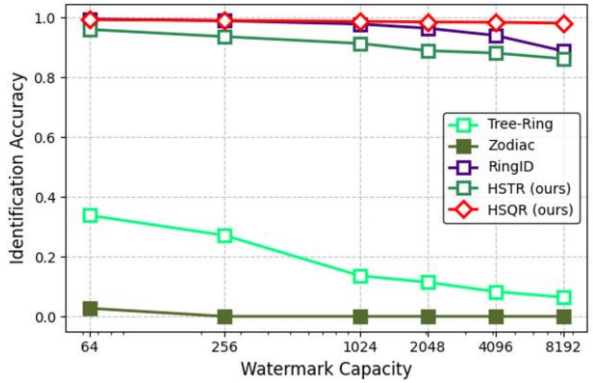
Case	Methods	Freq. Int.	# Det.	*Vrf.*	*Idf.*	$\Delta L_1^* \uparrow$
A	Tree-Ring	×	2	0.653	0.114	0.232
B	Tree-Ring	×	1	0.805	0.416	0.368
C	HSTR (ours)	✓	1	0.936	0.775	0.471
D	HSTR (ours)	✓	2	0.971	0.889	0.476

Table 7. Normality assessment of latent distributions (1,000 samples). HSTR better preserves Gaussianity than Tree-Ring, as shown by standard deviation, KS p-value, and failure rate.

Methods	Mean	Std. Dev.	KS p-value ↑	KS failure rate ↓
Tree-Ring	0.0004	0.9620	0.2404	0.234
HSTR (ours)	-0.0003	1.0000	0.4227	0.071

Validation of Gaussianity using Statistical Tests (incl. KS-Test)

- **HSQR's Superior Scalability:** As message capacity increases, most methods degrade rapidly. HSQR, however, maintains near-perfect identification accuracy even with over 8,192 unique IDs, demonstrating its suitability for large-scale, real-world applications.



- We introduced Hermitian SFW, a novel framework that resolves the key challenge of frequency integrity in semantic watermarking.
- Our methods achieve a new state-of-the-art balance between detection robustness and image fidelity, outperforming all baselines in diverse attack scenarios.
- Our work demonstrates that properly structured Fourier-domain watermarking is ready for real-world, secure deployment.

Thank you!

- Project Pages: https://thomas11809.github.io/SFWMark/
- Code and Resources: https://github.com/thomas11809/SFWMark
- Contact: thomas11809@snu.ac.kr

ICCV OCT 19-23, 2025 HONOLULU HAWAII