A Appendix

Table 3: Main Results with Error Bars. T@1%F represents TPR@1%FPR. We evaluate watermark accuracy in both benign and adversarial settings. Adversarial here refers to average performance over a battery of image manipulations.

Model	Method	AUC/T@1%F (Clean)	AUC/T@1%F (Adversarial)	FID ↓	CLIP Score ↑
Stable Diff. FID = 25.29 CLIP Score = 0.363	DwtDct	$0.974_{.001}$ / $0.624_{.013}$	$0.574_{.005}$ / $0.092_{.004}$	$25.10_{.09}$	$0.362_{.000}$
	DwtDctSvd	$1.000_{.000}$ / $1.000_{.000}$	$0.702_{.000}$ / $0.262_{.011}$	$25.01_{.09}$	$0.359_{.000}$
	RivaGAN	$0.999_{.000}$ / $0.999_{.000}$	$0.854_{.002}$ / $0.448_{.006}$	$\mathbf{24.51_{.17}}$	$0.361_{.000}$
	***T-R*$_{Zeros}$**	$0.999_{.000}$ / $0.999_{.000}$	$0.963_{.001}$ / $\mathbf{0.715_{.021}}$	$26.56_{.07}$	$0.356_{.000}$
	***T-R*$_{Rand}$**	$1.000_{.000}$ / $1.000_{.000}$	$0.918_{.005}$ / $0.702_{.017}$	$25.47_{.05}$	$0.363_{.001}$
	***T-R*$_{Rings}$**	$1.000_{.000}$ / $1.000_{.000}$	$\mathbf{0.975_{.001}}$ / $0.694_{.018}$	$25.93_{.13}$	$\mathbf{0.364_{.000}}$
ImageNet FID = 17.73	DwtDct	$0.899_{.040}$ / $0.244_{.203}$	$0.536_{.016}$ / $0.037_{.029}$	$17.77_{.01}$	-
	DwtDctSvd	$1.000_{.000}$ / $1.000_{.000}$	$0.713_{.019}$ / $0.187_{.008}$	$18.55_{.02}$	-
	RivaGAN	$1.000_{.000}$ / $1.000_{.000}$	$0.882_{.010}$ / $0.509_{.009}$	$18.70_{.02}$	-
	***T-R*$_{Zeros}$**	$0.999_{.000}$ / $1.000_{.000}$	$0.921_{.000}$ / $0.476_{.000}$	$18.78_{.00}$	-
	***T-R*$_{Rand}$**	$0.999_{.000}$ / $1.000_{.000}$	$0.940_{.004}$ / $0.585_{.006}$	$18.68_{.09}$	-
	***T-R*$_{Rings}$**	$0.999_{.000}$ / $0.999_{.000}$	$\mathbf{0.966_{.005}}$ / $\mathbf{0.603_{.006}}$	$\mathbf{17.68_{.16}}$	-

Table 4: AUC under each Attack for the ImageNet model, showing the effectiveness of *Tree-Ring*$_{Rings}$ over a number of augmentations. Cr. & Sc. refers to random cropping and rescaling.

Method	Clean	Rotation	JPEG	Cr. & Sc.	Blurring	Noise	Color Jitter	Avg
DwtDct	0.899	0.478	0.522	0.433	0.512	0.365	0.538	0.536
DwtDctSvd	**1.000**	0.669	0.568	0.614	0.947	0.656	0.535	0.713
RivaGan	**1.000**	0.321	**0.978**	**0.999**	0.988	0.962	0.924	0.882
***T-R*$_{Zeros}$**	0.999	0.953	0.806	0.997	**0.999**	0.938	0.775	0.921
***T-R*$_{Rand}$**	0.999	0.682	0.962	0.997	**0.999**	**0.986**	**0.956**	0.940
***T-R*$_{Rings}$**	0.999	**0.975**	0.940	0.994	**0.999**	0.979	0.861	**0.966**

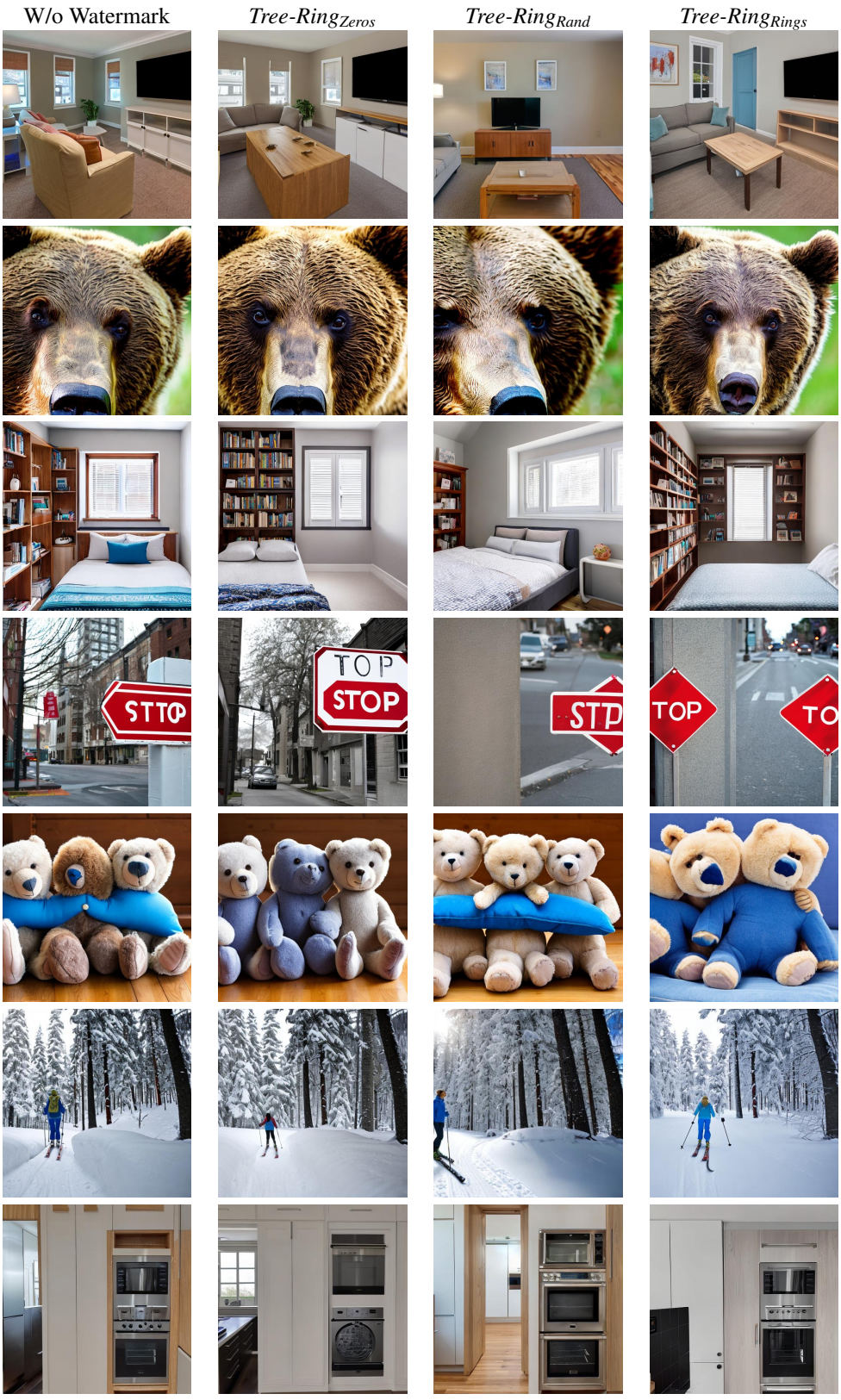

Figure 6: More generated images with *Tree-Ring Watermarking* with the first 7 prompts in MS-COCO-2017 training dataset.

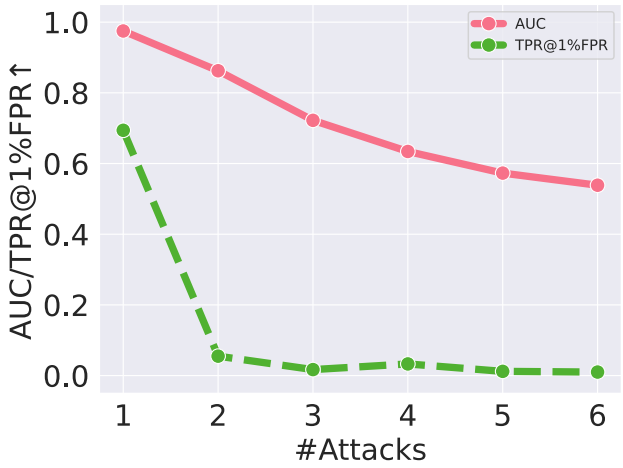

Figure 7: Results on k number of random attacks applied at the same time.

(a) 75° Rotation

(b) 25% JPEG Compression

(c) 75% Random Cropping + Scaling

(d) Gaussian Blurring with an 8×8 Filter

(e) Gaussian Noise with $\sigma = 0.1$

(f) Color Jitter with a Brightness Factor of 6

Figure 8: Attacked images.