Supplementary Material – RAVEN: Erasing Invisible Watermarks via Novel View Synthesis

Fahad Shamshad[1] Nils Lukas[1] Karthik Nandakumar[1,2]
[1]MBZUAI, UAE [2]Michigan State University, USA
{fahad.shamshad, nils.lukas, karthik.nandakumar}@mbzuai.ac.ae

This supplementary material provides additional experiments to support the effectiveness and robustness of RAVEN for invisible watermark removal. Specifically, it includes:

- **Hyperparameter selection** details for all baseline methods, ensuring fair and meaningful comparisons (Sec. 1).
- **Quality–detectability** curves demonstrating RAVEN's favorable trade-off across the operating range (Sec. 2).
- **Comprehensive quantitative results** on the DiffusionDB dataset [16], covering a wide range of watermarking schemes and attack strategies (Sec. 3).
- **Qualitative comparisons** to illustrate how different removal methods impact perceptual quality, structural consistency, and texture preservation (Sec. 4).
- **Computational efficiency** analysis, highlighting RAVEN's advantage over strong baselines (Sec. 5).
- **Defense directions** outlining promising strategies for designing more resilient watermarking schemes against view-synthesis attacks (Sec. 6).

Overall, these supplementary results demonstrate RAVEN's ability to achieve strong watermark suppression while maintaining high visual fidelity and semantic coherence, complementing the findings of the main paper.

1. Hyperparameter Selection for Baselines

For strong baselines with official implementations (UnMarker [9] and CtrlGen+ [12]), we use author-recommended settings optimized for the removal–quality trade-off. For remaining baselines, we follow the standardized evaluation protocol of the WAVES benchmark [1] and perform a coarse grid search on a held-out validation set to select operating points that best balance watermark suppression and perceptual quality. Specifically, we sweep the following parameter ranges: JPEG quality $\in [10, 90]$, brightness $\in [0.2, 1.0]$, contrast $\in [0.2, 1.0]$, Gaussian noise $\sigma \in [0.1, 0.5]$, BM3D $\sigma \in [0.1, 0.6]$, Regen diffusion steps $\in [20, 100]$, VAE quality level $\in [1, 6]$, and RAVEN camera translation magnitudes $\in \{8, 16, 24, 32, 40\}$ pixels. The selected operating point for each baseline corresponds to the configuration achieving the strongest watermark suppression without exceeding the perceptual quality degradation of RAVEN. This protocol ensures a fair and meaningful comparison rather than a favorable single operating point for our method.

2. Quality–Detectability Trade-off Analysis

A fair comparison across methods with tunable parameters requires examining the full trade-off between watermark suppression and perceptual quality, rather than a single operating point. Here, we use TPR@1%FPR as the detectability metric and FID as the perceptual quality metric. Fig. 1 plots the average TPR@1%FPR against FID for verification-based watermarking methods on DiffusionDB [16], obtained by sweeping each method's primary control parameter (*e.g.*, noise level, compression strength, or diffusion steps) within the standardized ranges described in Sec. 1. RAVEN consistently achieves lower watermark detectability at comparable or better perceptual quality across the full sweep, demonstrating a superior trade-off frontier rather than a favorable single operating point. Notably, regeneration-based methods can approach RAVEN's suppression levels only at the cost of significant perceptual degradation, while UnMarker consistently incurs higher perceptual cost at equivalent detectability levels. Similar trends are observed for bitstream-based methods.

3. Quantitative Results on DiffusionDB

Table 1 reports watermark removal performance on the DiffusionDB dataset [16], a large and diverse collection of real-world text-to-image prompts. We evaluate RAVEN on 1,001 randomly sampled prompts using the same protocol as the main paper, generating images at 512×512 resolution with

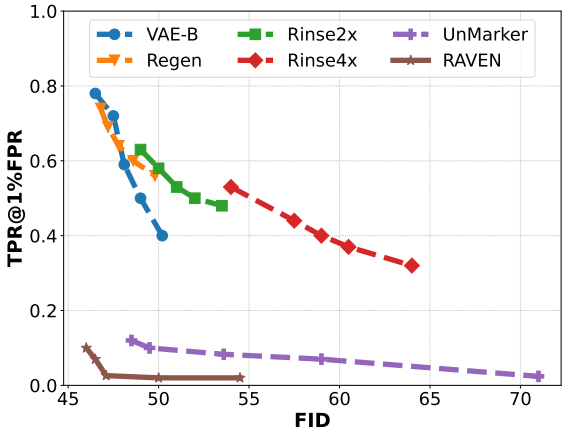

Figure 1. Quality–detectability trade-off on DiffusionDB [16]. We plot TPR@1%FPR vs. FID by sweeping each method's primary control parameter. Lower TPR and lower FID are both desirable. RAVEN achieves a superior trade-off across the operating range.

Stable Diffusion v2-1, CFG scale 7.5, and 50 DDIM steps. The evaluation spans the same 15 watermarking schemes, including both semantic in-generation and post-hoc bitstream methods, and compares RAVEN against 14 baseline attacks covering classical signal processing, regeneration-based approaches, and advanced learned methods such as CtrlGen+ [12] and UnMarker [9], ensuring a fair and consistent comparison across all paradigms.

3.1. Key Findings

Results on DiffusionDB [16] reinforce the trends observed in the main paper, with RAVEN achieving the lowest average TPR@1%FPR of **0.029** for semantic watermarking, outperforming the strongest baseline, UnMarker, and demonstrating superior suppression across all methods, with particularly notable gains for challenging schemes such as RingID and ROBIN. For bitstream-based watermarking, RAVEN attains an average bit accuracy of **0.531**, close to the ideal randomization range and comparable to UnMarker, while preserving noticeably better perceptual quality. Performance remains highly stable across datasets, with semantic TPR and bit accuracy varying by less than 0.005 between DiffusionDB [16], MS-COCO [11], and SD-Prompts [14], confirming that RAVEN leverages properties of diffusion latent representations and watermark embedding mechanisms rather than dataset-specific prompt characteristics.

4. Additional Qualitative Results

Figures 2-4 present qualitative comparisons of watermark removal methods across diverse visual scenes, highlighting clear differences in perceptual quality and structural preservation. VAE-B [2] consistently introduces excessive smoothing that suppresses fine details and weakens texture sharpness, while Regen [21] generates visible high-frequency artifacts due to the aggressive noise injection required for watermark removal. Rinse often leads to unnatural color shifts and reduced photorealism as a result of repeated regeneration passes. UnMarker [9] suppresses watermark signals more effectively than conventional priors but leaves residual noise patterns and minor structural inconsistencies. Notably, as evident in the Figure 4, CtrlGen+ [12] alters the underlying scene layout, introducing new architectural or geometric elements and generating alternative structures that deviate from the original semantic configuration. In contrast, RAVEN maintains faithful scene geometry, consistent structure, and natural texture distribution while successfully eliminating watermark traces, resulting in visually coherent and photorealistic outputs that closely align with the original watermarked content.

5. Computational Efficiency

Among advanced methods that achieve strong watermark suppression, RAVEN offers a favorable balance between effectiveness and efficiency. UnMarker [9] requires approximately 5 minutes per image on an NVIDIA A100 GPU due to its iterative optimization process, while CtrlGen+ [12] requires multi-GPU training infrastructure with 8 NVIDIA A100 GPUs to achieve its reported performance. In contrast, RAVEN operates in a zero-shot manner without additional training and processes each image in approximately 6 seconds on an A100 40GB GPU. This efficiency is enabled by leveraging frozen pretrained diffusion models with lightweight attention modifications, avoiding costly per-image optimization or large-scale training requirements. While signal processing attacks are faster, their limited effectiveness against modern watermarking schemes makes them less relevant in practical scenarios.

6. Defense Directions

While RAVEN exposes a critical vulnerability in existing invisible watermarking schemes, our goal is not merely to demonstrate attack effectiveness, but to motivate the development of more resilient designs. Promising defense directions include incorporating viewpoint-augmented training during watermark embedding, enabling the watermark signal to remain stable under semantic-preserving geometric transformations. In addition, future detectors could be designed to identify statistical inconsistencies or cross-view correspondence violations introduced by novel-view synthesis attacks. More broadly, we hope that RAVEN serves as a principled benchmark for evaluating watermark robustness beyond conventional pixel-space and regeneration-based attacks, guiding the development of watermarking systems that remain reliable under semantic-preserving transformations.

Table 1. Verification performance of different watermarking methods under various attacks on DiffusionDB dataset. TPR@1%FPR is reported for in-generation semantic watermarking methods (Tree-Ring to ROBIN), where lower values indicate better attack performance. Bit Accuracy is reported for post-hoc bitstream-based methods (DwtDct to VINE), where values near 0.5 indicate successful watermark randomization. RAVEN achieves the lowest detection rates across both categories, demonstrating superior removal efficacy while maintaining visual quality.

Attack	*Tree-Ring* [17]	*Zodiac* [20]	*HSTR* [10]	*RingID* [5]	*HSQR* [10]	*ROBIN* [8]	*Avg.*	*DwtDct* [6]	*DwtDctSvd* [6]	*RivaGAN* [19]	*Stable Sign.* [7]	*Gauss Shad.* [18]	*TrustMark* [3]	*Stega St.* [15]	*VINE* [13]	*Avg.*
DiffusionDB [16]																
Bright.	0.487	0.752	0.792	0.989	0.977	0.987	0.831	0.563	0.588	0.839	0.890	0.954	0.913	0.989	0.995	0.841
Cont.	0.889	0.988	0.996	1.000	1.000	0.991	0.977	0.515	0.463	0.960	0.967	0.999	0.922	0.994	0.989	0.851
JPEG	0.434	0.933	0.981	1.000	0.999	0.984	0.905	0.509	0.593	0.790	0.787	0.990	0.913	1.000	1.000	0.848
Blur	0.904	0.988	0.996	1.000	1.000	0.985	0.979	0.672	0.997	0.985	0.889	0.999	0.931	0.988	0.970	0.929
Noise	0.392	0.834	0.792	0.963	0.974	0.895	0.808	0.829	0.995	0.937	0.726	0.992	0.793	0.964	0.972	0.889
BM3D	0.799	0.984	0.991	1.000	0.999	0.869	0.940	0.526	0.830	0.893	0.813	0.998	0.880	0.994	0.997	0.879
Center Crop	0.499	0.971	1.000	1.000	1.000	0.887	0.893	0.723	0.742	0.974	0.981	0.999	0.883	0.917	0.945	0.895
Random Crop	0.715	0.985	1.000	1.000	1.000	0.795	0.916	0.801	0.860	0.979	0.986	1.000	0.739	0.888	0.910	0.906
VAE-B	0.454	0.911	0.968	0.995	0.997	0.824	0.858	0.513	0.658	0.553	0.690	0.978	0.821	0.869	0.906	0.748
VAE-C	0.503	0.926	0.969	0.999	0.999	0.847	0.874	0.514	0.608	0.518	0.687	0.989	0.833	0.875	0.909	0.742
Regen.	0.454	0.903	0.989	1.000	1.000	0.865	0.869	0.512	0.621	0.556	0.496	0.998	0.796	0.857	0.886	0.715
Rinse	0.445	0.861	0.975	0.990	0.996	0.799	0.844	0.501	0.559	0.531	0.507	0.964	0.738	0.825	0.858	0.685
CtrlGen+	0.084	0.300	0.767	1.000	1.000	0.311	0.577	0.523	0.509	0.516	0.581	1.000	0.676	0.562	0.881	0.656
UnMarker	0.031	0.090	0.034	0.265	0.021	0.042	0.081	0.489	0.517	0.538	0.510	0.597	0.542	0.651	0.623	0.559
RAVEN	0.023	0.070	0.028	0.022	0.020	0.015	**0.029**	0.515	0.491	0.502	0.522	0.550	0.485	0.583	0.597	**0.531**

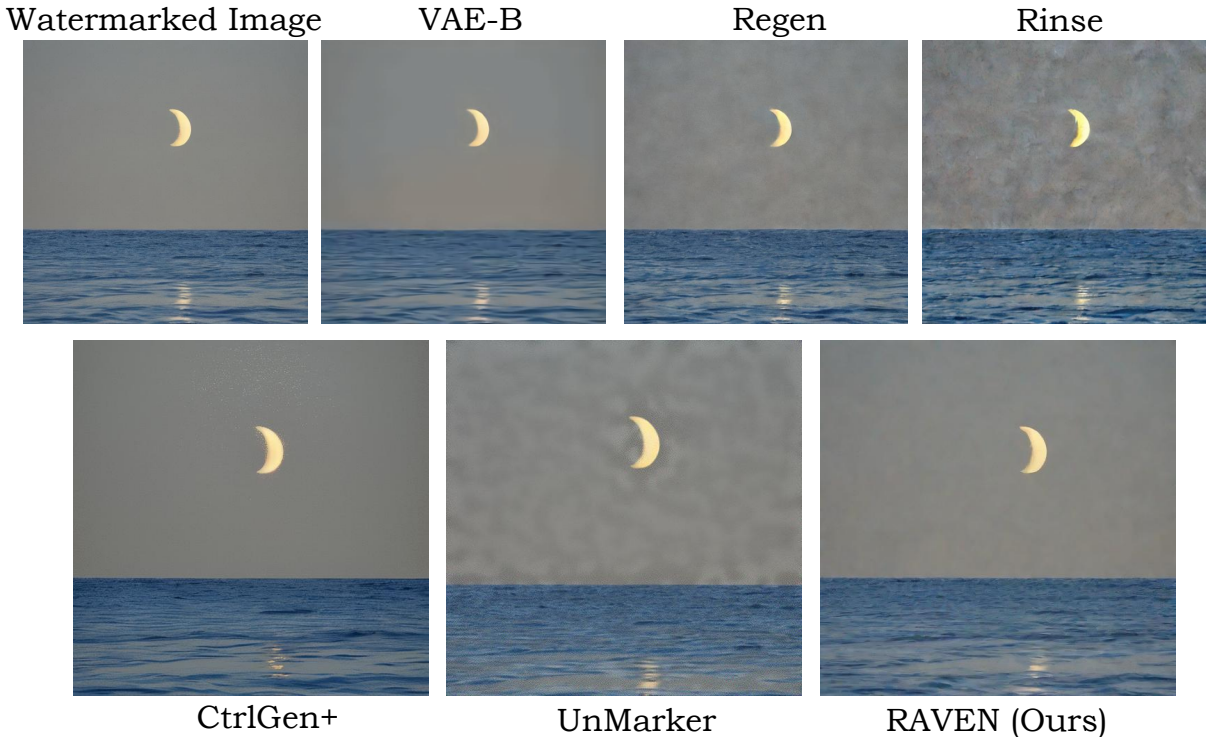

Figure 2. **Qualitative comparison of watermark removal methods**. VAE-B [4] introduces excessive blurring that degrades fine details. Regen [21] produces visible artifacts due to high noise injection required for watermark removal. Rinse exhibits unnatural color shifts and loss of photorealism, a consequence of performing multiple regeneration passes. UnMarker [9] leaves noisy residual artifacts that compromise visual quality. CtrlGen+ [12] produces overly stylized outputs that deviate from natural appearance. In contrast, RAVEN preserves fine-grained details, natural textures, and photorealistic appearance. Note that the images for UnMarker [9] and RAVEN differ slightly from other methods due to cropping layers and camera translation transformations, respectively.

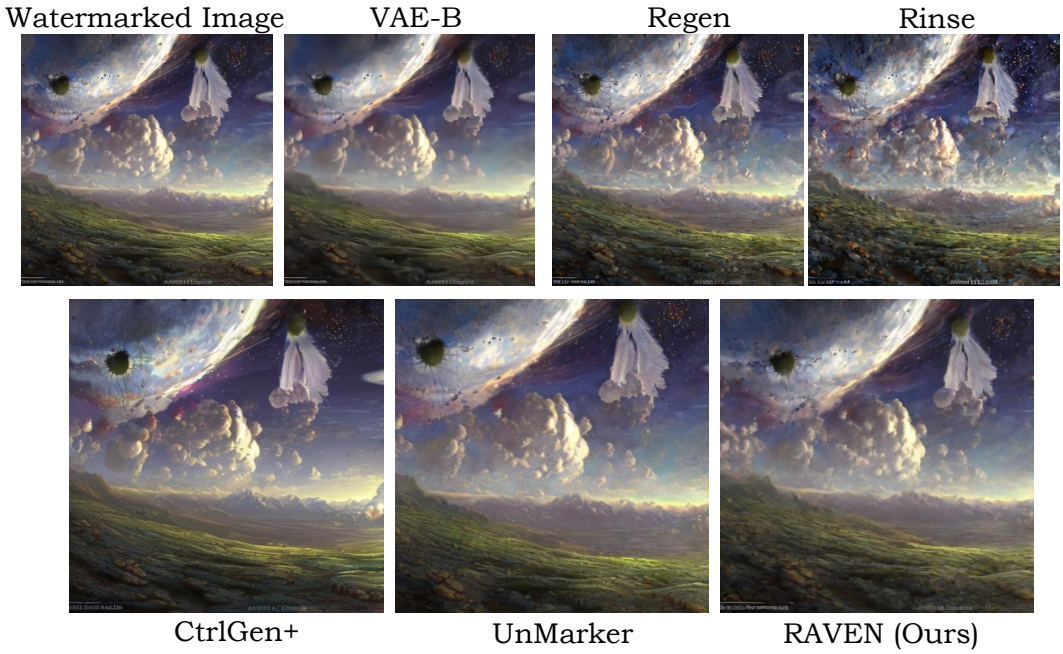

Figure 3. **Qualitative comparison of watermark removal methods**. VAE-B [4] introduces excessive blurring that degrades fine details. Regen [21] produces visible artifacts due to high noise injection required for watermark removal. Rinse exhibits unnatural color shifts and loss of photorealism, a consequence of performing multiple regeneration passes. UnMarker [9] leaves noisy residual artifacts that compromise visual quality. CtrlGen+ [12] produces overly stylized outputs that deviate from natural appearance. In contrast, RAVEN preserves fine-grained details, natural textures, and photorealistic appearance. Note that the images for UnMarker [9] and RAVEN differ slightly from other methods due to cropping layers and camera translation transformations, respectively.

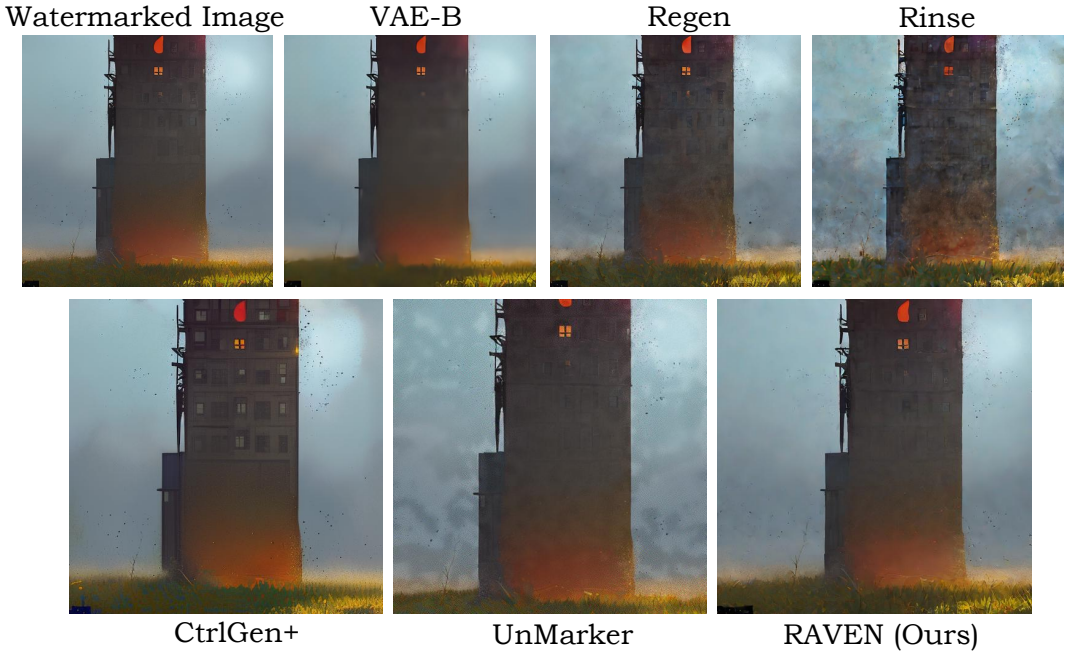

Figure 4. **Qualitative comparison of watermark removal methods**. VAE-B [4] introduces excessive blurring that degrades fine details. Regen [21] produces visible artifacts due to high noise injection required for watermark removal. Rinse exhibits unnatural color shifts and loss of photorealism, a consequence of performing multiple regeneration passes. UnMarker [9] leaves noisy residual artifacts that compromise visual quality. CtrlGen+ [12] produces overly stylized outputs that deviate from natural appearance. In contrast, RAVEN preserves fine-grained details, natural textures, and photorealistic appearance. Note that the images for UnMarker [9] and RAVEN differ slightly from other methods due to cropping layers and camera translation transformations, respectively.

References

[1] Bang An, Mucong Ding, Tahseen Rabbani, Aakriti Agrawal, Yuancheng Xu, Chenghao Deng, Sicheng Zhu, Abdirisak Mohamed, Yuxin Wen, Tom Goldstein, et al. Waves: Benchmarking the robustness of image watermarks. *arXiv preprint arXiv:2401.08573*, 2024. 1

[2] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436*, 2018. 2

[3] Tu Bui, Shruti Agarwal, and John Collomosse. Trustmark: Universal watermarking for arbitrary resolution images. *arXiv preprint arXiv:2311.18297*, 2023. 3

[4] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7939–7948, 2020. 3, 4

[5] Hai Ci, Pei Yang, Yiren Song, and Mike Zheng Shou. Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification. In *European Conference on Computer Vision*, pages 338–354. Springer, 2024. 3

[6] Ingemar J Cox, Matthew L Miller, Jeffrey A Bloom, Jessica Fridrich, and Ton Kalker. Digital watermarking. *Morgan Kaufmann Publishers*, 54:56–59, 2008. 3

[7] Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22466–22477, 2023. 3

[8] Huayang Huang, Yu Wu, and Qian Wang. Robin: Robust and invisible watermarks for diffusion models with adversarial optimization. *Advances in Neural Information Processing Systems*, 37:3937–3963, 2024. 3

[9] Andre Kassis and Urs Hengartner. Unmarker: A universal attack on defensive image watermarking. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 2602–2620. IEEE, 2025. 1, 2, 3, 4

[10] Sung Ju Lee and Nam Ik Cho. Semantic watermarking reinvented: Enhancing robustness and generation quality with fourier integrity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18759–18769, 2025. 3

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[12] Yepeng Liu, Yiren Song, Hai Ci, Yu Zhang, Haofan Wang, Mike Zheng Shou, and Yuheng Bu. Image watermarks are removable using controllable regeneration from clean noise. *arXiv preprint arXiv:2410.05470*, 2024. 1, 2, 3, 4

[13] Shilin Lu, Zihan Zhou, Jiayou Lu, Yuanzhi Zhu, and Adams Wai-Kin Kong. Robust watermarking using generative priors against image editing: From benchmarking to advances. *arXiv preprint arXiv:2410.18775*, 2024. 3

[14] Gustavo Santana. Gustavosta: Stable-diffusion-prompts. https://huggingface.co/datasets/Gustavosta/Stable-Diffusion-Prompts, 2022. Dataset. 2

[15] Matthew Tancik, Ben Mildenhall, and Ren Ng. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2117–2126, 2020. 3

[16] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 893–911, 2023. 1, 2, 3

[17] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023. 3

[18] 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*, pages 12162–12171, 2024. 3

[19] Kevin Alex Zhang, Lei Xu, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Robust invisible video watermarking with attention. *arXiv preprint arXiv:1909.01285*, 2019. 3

[20] Lijun Zhang, Xiao Liu, Antoni Viros Martin, Cindy Xiong Bearfield, Yuriy Brun, and Hui Guan. Robust image watermarking using stable diffusion. *arXiv preprint arXiv:2401.04247*, 2024. 3

[21] Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative ai. *Advances in neural information processing systems*, 37:8643–8672, 2024. 2, 3, 4