OmniGuard: Hybrid Manipulation Localization via Augmented Versatile Deep Image Watermarking

Supplementary Material

A. Discussions

A.1. Limitations and Our Future Works

The limitations of our method mainly lie in two parts. **First**, although we have significantly improved OmniGuard's localization accuracy and robustness by introducing a passive detection network, if the degradation is extremely severe and exceeds the robustness threshold of image-in-image steganography, our localization performance will approach that of standard passive detection networks. This issue might be addressed by exploring more advanced steganography frameworks and theories, such as diffusion models [15, 16], or by establishing relevant evaluation standards to exclude excessively low-quality images from being used for post-event forensics. **Second**, although the fidelity of our dual watermark exceeds 40 dB, we found that when handling ultra-high-resolution images (*e.g.*, panoramic pictures [9]), the resolution scaling strategy [5] will amplify watermark artifacts, which slightly impacts the perceptual quality of our method. However, this is a common issue among all current deep watermarking methods. Therefore, exploring a truly scalable watermarking approach that can handle arbitrary resolutions remains a worthwhile direction for future research.

A.2. Why Use Joint Training?

To verify the impact of jointly training localized and copyright watermarks, as opposed to applying two separately trained watermark models to the images, we conducted experiments on both scenarios. We randomly selected 100 images and tampered with them using SD inpainting. For the separate embedding setup, we retrained our original localization watermark embedding and decoding networks and combined them with pre-trained TrustMark [5]. The results in Tab. 1 clearly demonstrate that without joint training, the watermark fidelity decreases by approximately **6 dB** PSNR. However, due to the robustness of the passive extraction network, the localization performance remains largely unaffected. This highlights the importance of simultaneously addressing localization and copyright protection in OmniGuard.

A.3. Why do we choose VAE as surrogate attacks?

To demonstrate the rationale behind using VAE instead of InstructPix2Pix or other AIGC global editing methods, we visualized two sets of residual maps in Fig. 1. It can be observed that if the image is not processed via a diffusion denoising process and is only encoded and reconstructed using VAE, the artifacts in the residual map are primarily uniformly distributed along the edge information of the original image. When the image is edited using InstructPix2Pix, we find that the error map generated by editing and the error map produced by VAE reconstruction exhibit certain consistency in their distribution. Furthermore, the residual map generated by editing shows smaller differences in areas outside the edited region than the errors observed in the VAE reconstruction. Thus, the distortion caused by VAE on the original image appears to be greater and more global than the distortion introduced by the diffusion process itself. Considering algorithm efficiency and computational resource consumption, we opt to introduce VAE in our training process.

Table 1. Performance comparison between joint training and using two separate watermarks.

Method	PSNR (dB)	SSIM	F1	AUC
Separate Embedding	35.46	0.966	0.953	0.986
Joint Training (Ours)	41.59	0.985	0.975	0.999

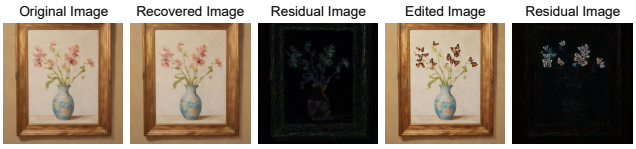

Figure 1. Residual images between the recovered image produced by the VAE and the original image, and between the edited image produced by InstructPix2Pix [4] and the original one.

A.4. Exploration on the Localized Watermark

Considering that the localized watermark has a decisive impact on the final fidelity, we have conducted extensive experiments to identify the optimal localization watermark. Finally, we arrive at the following conclusions:

- **Choice of Color:** Typically, selecting light-colored images allows the steganography network to better hide the localization watermark and achieve higher PSNR. Additionally, using solid-colored images often facilitates subsequent localization and detection.
- **Challenges with Solid Colors:** However, using solid-colored images can result in grid-like or repetitive artifacts on the image, which may be visually unappealing and raise security concerns.
- **Adding Texture:** Adding natural, uncomplicated texture details to solid-colored images significantly improves the

Table 2. Localization performance metrics for EditGuard and OmniGuard under different degradations.

Method	Metrics	Clean	JPEG(Q=60)	JPEG(Q=70)	Bri.	Con.	Hue	S.-P.	GS Noise
EditGuard	F1	0.951	0.515	0.912	0.536	0.876	0.946	0.921	0.821
	AUC	0.971	0.785	0.961	0.817	0.945	0.963	0.968	0.944
	IoU	0.935	0.365	0.865	0.410	0.809	0.905	0.862	0.709
OmniGuard (Ours)	F1	0.961	0.810	0.938	0.927	0.926	0.964	0.951	0.958
	AUC	0.999	0.982	0.998	0.959	0.960	0.999	0.999	0.999
	IoU	0.928	0.713	0.888	0.999	0.999	0.933	0.911	0.922

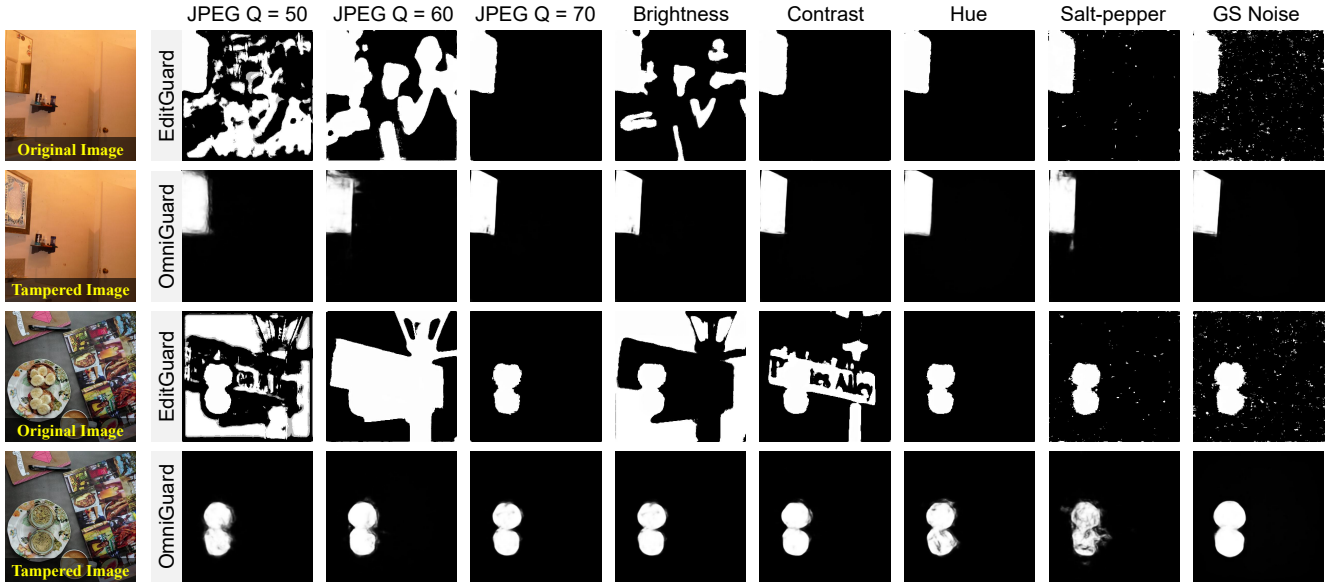

Figure 2. Localization performance of our OmniGuard and EditGuard on several different degradation conditions. Our method can produce clear masks under various noisy conditions, while EditGuard shows confusion and blurriness under certain severe degradations.

fidelity of the watermarked image (such as the light-toned blue sky with clouds used in our paper).

- **Adaptive Watermark Transform:** Coupled with our designed adaptive watermark transform, the fidelity of the hidden localized watermark is further enhanced.

These considerations help balance fidelity, detection, and security, making the watermark both effective and visually acceptable.

B. More Implementation Details

Localized watermark hiding and decoding: Our localized watermark hiding and decoding network adopts the basic structure of EditGuard [17]. It uses a network composed of 16 stacked addition affine transformation layers, where each reversible transformation module employs a DenseBlock. Additionally, we adopt a decoding network constructed with stacked residual blocks to predict the missed high-frequency components $\hat{\mathbf{z}}$ from $\mathbf{I}_{\text{rec}}$, enabling accurate inverse decoding of our network.

Copyright watermark hiding and decoding: Following [5], we use a MUNIT-based Unet [8] as our watermark embedding network, treating the image watermarking as image translation. The watermark is interpolated to match the original image's dimension and fused into the input feature via a light network. Our extractor is a standard ResNet50 with the last layer being replaced by a sigmoid-activated FC to predict the watermark. Note that, we use the resolution scaling strategy [5] to enable our OmniGuard to support arbitrary resolutions.

C. More Experimental Results

C.1. Robustness of Our Localization

To further demonstrate the robustness of our localization performance, we have detailed our localization results under different degradation conditions and compared them with the current state-of-the-art active localization method EditGuard. We selected 1000 images from the COCO dataset and tampered with them using SD Inpaint. We consider various degradation conditions, including JPEG compression (Q=50, 60, 70), Gaussian noise ($\sigma = 15$), salt-

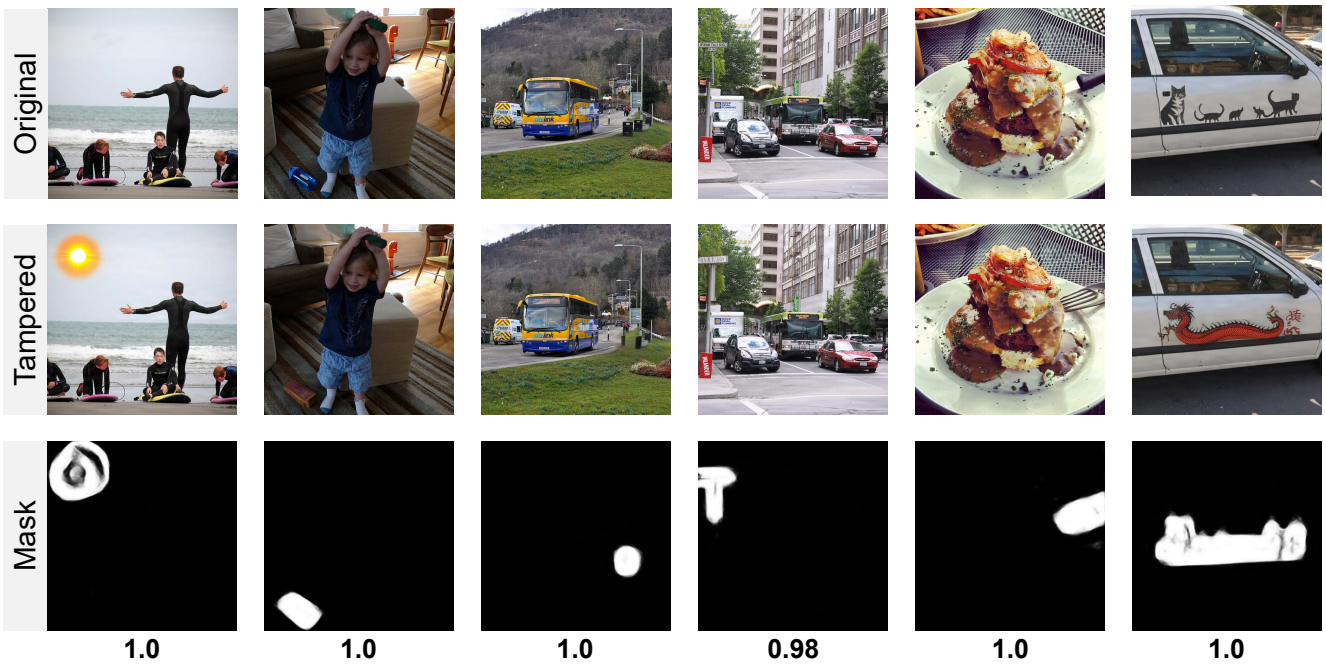

Figure 3. Localization and copyright recovery performance on the most recent AIGC-Editing tool MagicQuill [11]. The recovered bit accuracy is shown below. Without any tuning, our method can accurately locate the tampered regions of SOTA editing methods and restore the original copyright.

and-pepper noise, and color jitter (adjustments to brightness, contrast, and hue). Tab. 2 presents the F1-score, AUC, and IoU of our OmniGuard and EditGuard. We find that under degradation conditions, OmniGuard consistently outperforms EditGuard and remains largely unaffected by different types of degradation. Notably, under severe degradations such as JPEG compression (Q=60) and a 30% reduction in brightness, OmniGuard shows a significant improvement compared to EditGuard. As shown in Fig. 2, our method can accurately identify the tampered regions, whereas EditGuard often highlights imprecise and blurry regions under severe degradations.

C.2. Generalization to SOTA AIGC-Edit Methods

To validate the generalization capability of our method, we tested OmniGuard on two of the latest state-of-the-art AIGC-Edit methods: the recently released and widely discussed MagicQuill [11], and SDXL-inpainting [14]. Fig. 3 shows the results of OmniGuard on MagicQuill. It can be observed that our method accurately identifies the tampered regions and correctly extracts the copyright information, even when MagicQuill's edits are highly subtle and difficult to detect with the naked eye. Fig. 4 further demonstrates that on SDXL, a fine-grained editing method, OmniGuard significantly outperforms passive methods such as PSCC-Net [10], MVSS-Net [7], and IML-ViT [12] in terms of detection accuracy and generalization. Meanwhile, we can almost completely decode the hidden copyright even under the interference of SDXL inpainting. Notably, when applied to these new AIGC manipulations, OmniGuard requires no fine-tuning or retraining, presenting good generalization ability.

Regeneration	bmshj2018-factorized [2]	bmshj2018-hyperprior [3]	mbt2018-mean [13]	mbt2018 [13]	cheng2020-anchor [6]
TrustMark	0.841	0.796	0.776	0.806	0.762
OmniGuard	0.956	0.933	0.921	0.923	0.897

Table 3. Bit accuracy comparison of our OmniGuard and Trustmark on SOTA recent watermarking attack benchmarks [1].

C.3. Robustness on Image Regeneration

To better validate the robustness of our method against global edits, we adopted a state-of-the-art watermarking attack benchmark [1] and tested five typical image regeneration methods including [2, 3, 6, 13]. As reported on Tab. 3, our method significantly outperforms TrustMark [5] across various regeneration attacks and demonstrates good generalization to most unseen AIGC generation methods.

C.4. Fidelity of Our OmniGuard

To further validate the fidelity advantages of our method, we test it on high-resolution (1024×1024 and 1792×1024) AIGC-generated images. As shown in Fig. 5, we find that EditGuard, when applied to high-resolution images, tends to produce regular color blocks and artifacts in sparse background areas. In contrast, OmniGuard maintains satisfac-

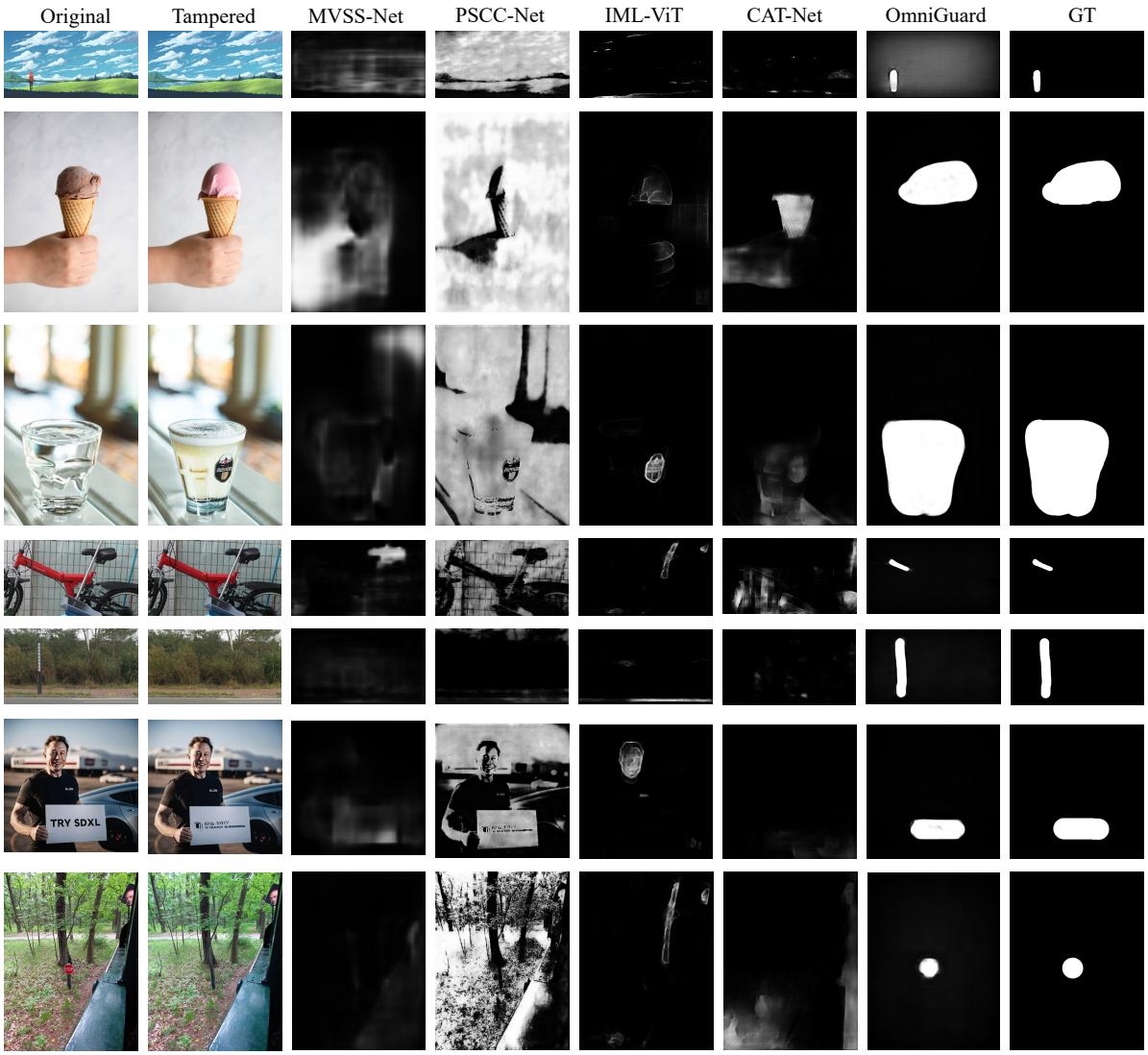

Figure 4. Localization performance of our OmniGuard and other competitive methods on the SOTA AIGC-Editing Method [14].

tory fidelity. We further present residual maps for both methods, scaled by a factor of 10 for better visibility. It can be observed that, overall, OmniGuard's error map is slighter than EditGuard's while maintaining excellent content adaptiveness. The watermark artifacts are added to areas such as the sky, balloons, clouds, distant mountain peaks, and water ripples in the background. These regions are less perceptible to the human eye compared to the main subjects, such as people and prominent patterns in the image. These results further validate the effectiveness of our method and its general applicability across different data domains.

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. In *International Conference on Machine Learning*, pages 1456–1492. PMLR, 2024. 3

[2] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. *arXiv preprint arXiv:1611.01704*, 2016. 3

[3] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1

[5] Tu Bui, Shruti Agarwal, and John Collomosse. Trustmark: Universal watermarking for arbitrary resolution im-

ages. *arXiv preprint arXiv:2311.18297*, 2023. 1, 2, 3

[6] 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

[7] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3539–3553, 2022. 3

[8] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 2

[9] Weiqi Li, Shijie Zhao, Bin Chen, Xinhua Cheng, Junlin Li, Li Zhang, and Jian Zhang. Resvr: Joint rescaling and viewport rendering of omnidirectional images. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 78–87, 2024. 1

[10] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (11):7505–7517, 2022. 3

[11] Zichen Liu, Yue Yu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Wen Wang, Zhiheng Liu, Qifeng Chen, and Yujun Shen. Magicquill: An intelligent interactive image editing system. *arXiv preprint arXiv:2411.09703*, 2024. 3

[12] Xiaochen Ma, Bo Du, Xianggen Liu, Ahmed Y Al Hammadi, and Jizhe Zhou. Iml-vit: Image manipulation localization by vision transformer. *arXiv preprint arXiv:2307.14863*, 2023. 3

[13] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 3

[14] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 4

[15] Youmin Xu, Xuanyu Zhang, Jiwen Yu, Chong Mou, Xiandong Meng, and Jian Zhang. Diffusion-based hierarchical image steganography. *arXiv preprint arXiv:2405.11523*, 2024. 1

[16] Jiwen Yu, Xuanyu Zhang, Youmin Xu, and Jian Zhang. Cross: Diffusion model makes controllable, robust and secure image steganography. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[17] Xuanyu Zhang, Runyi Li, Jiwen Yu, Youmin Xu, Weiqi Li, and Jian Zhang. Editguard: Versatile image watermarking for tamper localization and copyright protection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11964–11974, 2024. 2

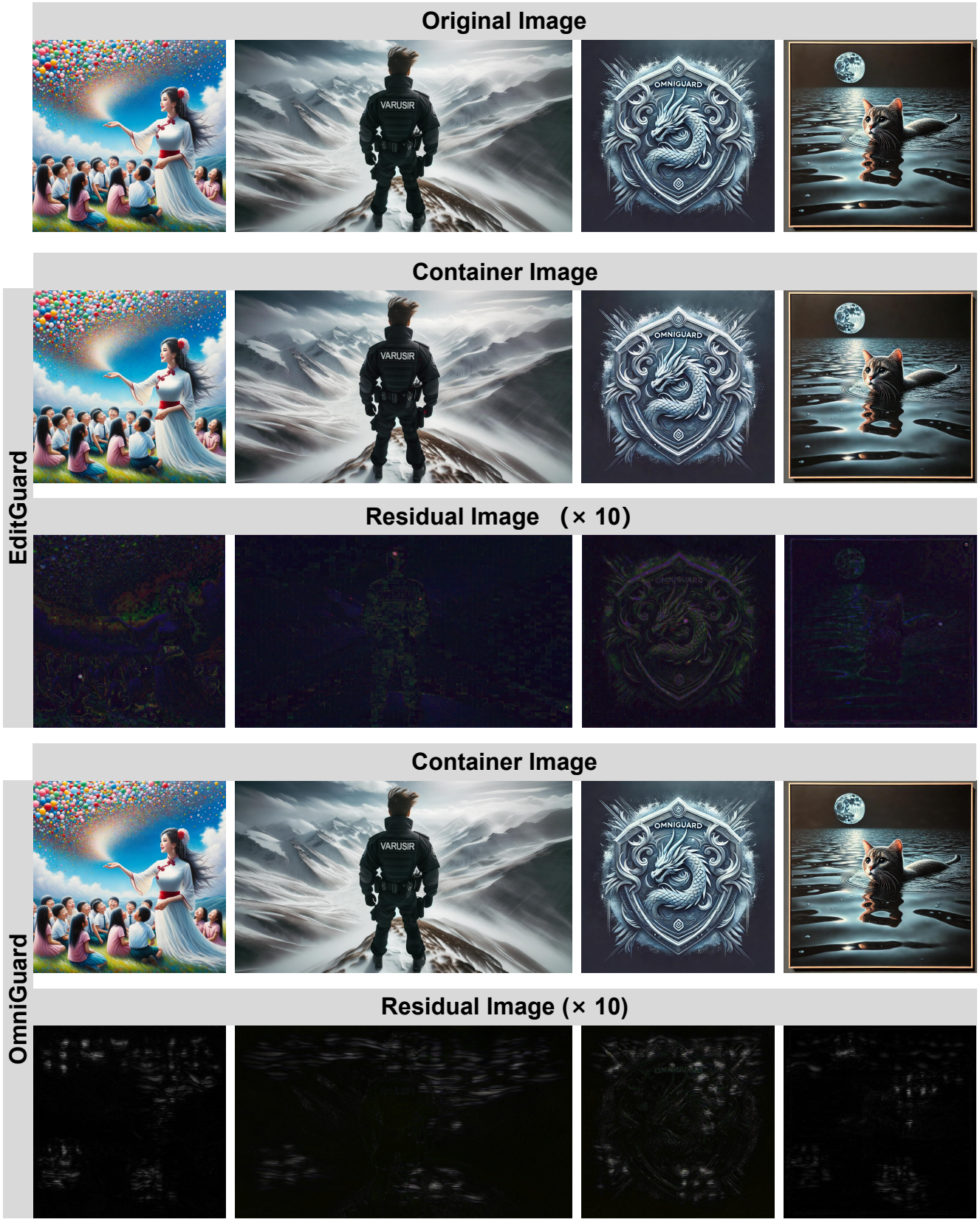

Figure 5. Fidelity comparison between our OmniGuard and EditGuard in some AI-generated high-resolution images. The residual maps, amplified by a factor of 10, are placed below the watermark images. Our OmniGuard shows better fidelity, with watermark artifacts primarily concentrated in background regions that are less perceptible to the human eye.