

ADAPT2HIDE: LEVERAGING OFF-THE-SHELF AUTOENCODER FOR REVERSIBLE VISUAL PROCESSING

APPENDIX

Tasks					Reversible visual censorship (ImageNet-1K [30])									
Budget	Settings				Imperceptibility					Reversibility				
	Conf.	Simp.	Rank	# of Add. Pram.	PSNR [†]	SSIM [†]	LPIPS [↓]	SD-Sim [†]	SDXL-Sim [†]	PSNR [†]	SSIM [†]	LPIPS [↓]	SD-Sim [†]	SDXL-Sim [†]
2	✓		64	13.550M (13.93%)	41.52	0.9464	0.0647	0.9847	0.9539	32.90	0.9545	0.0417	0.9925	0.9365
2	✓		4	0.846M (1.002%)	41.43	0.9445	0.0715	0.9822	0.9559	32.91	0.9497	0.0479	0.9914	0.9305
2	✓		16	3.387M (3.892%)	41.52	0.9462	0.0642	0.9841	0.9583	32.69	0.9500	0.0445	0.9925	0.9336
2	✓		256	54.198M (24.46%)	41.53	0.9453	0.0729	0.9826	0.9570	35.20	0.9685	0.0280	0.9959	0.9513
1	✓		64	13.550M (13.93%)	46.37	0.9816	0.0273	0.9950	0.9777	31.92	0.9486	0.0368	0.9927	0.9533
3	✓		64	13.550M (13.93%)	38.32	0.9019	0.1368	0.9534	0.9317	35.92	0.9734	0.0257	0.9956	0.9540
4	✓		64	13.550M (13.93%)	35.93	0.853	0.1812	0.9311	0.9061	34.91	0.9667	0.0317	0.9946	0.9470
2			64	13.550M (13.93%)	6.379	0.0461	0.7464	0.2846	0.0912	7.268	0.0679	0.7212	0.445	0.2393
2	✓	✓	64	10.600M (15.92%)	41.71	0.9497	0.1234	0.9597	0.9200	33.86	0.9665	0.0424	0.9718	0.9503

Tasks					Reversible AI editing (InstructPix2Pix [31])									
Budget	Settings				Imperceptibility					Reversibility				
	Conf.	Simp.	Rank	# of Add. Pram.	PSNR [†]	SSIM [†]	LPIPS [↓]	SD-Sim [†]	SDXL-Sim [†]	PSNR [†]	SSIM [†]	LPIPS [↓]	SD-Sim [†]	SDXL-Sim [†]
2	✓		64	13.550M (13.93%)	41.28	0.9639	0.0231	0.9955	0.9765	32.46	0.9577	0.0341	0.9951	0.9545
2	✓		4	0.846M (1.002%)	41.26	0.9647	0.0196	0.9962	0.9758	29.82	0.9219	0.0609	0.9877	0.9328
2	✓		16	3.387M (3.892%)	41.30	0.9642	0.0243	0.9953	0.9745	30.65	0.9330	0.0462	0.9935	0.9452
2	✓		256	54.198M (24.46%)	41.21	0.9640	0.0273	0.9926	0.9786	32.03	0.9503	0.0358	0.9920	0.9555
1	✓		64	13.550M (13.93%)	46.00	0.9870	0.0081	0.9973	0.9876	31.92	0.9486	0.0368	0.9927	0.9533
3	✓		64	13.550M (13.93%)	38.18	0.9311	0.0458	0.9886	0.9618	29.81	0.9313	0.0398	0.9944	0.9508
4	✓		64	13.550M (13.93%)	35.82	0.8997	0.0633	0.983	0.9594	31.79	0.9515	0.0332	0.9954	0.9591
2			64	13.550M (13.93%)	7.060	0.069	0.7691	0.3142	0.2121	10.11	0.1339	0.7513	0.4601	0.2757
2	✓	✓	64	10.600M (15.92%)	41.30	0.9636	0.0303	0.9919	0.9774	30.77	0.9341	0.0472	0.9813	0.9508

The first row represents the primary setting used in most experiments in our main paper.

Table S.1: Reversibility and Imperceptibility on censorship and AI editing with different model settings. The first row shows results from our main configuration. Conf.:Conformity Layer. Simp.: Simpler VAE. Rank: LoRA Rank. # of Add. Pram.: Numbers of Additional Parameters (and their relative percentage compared with the base AE).

A. LIMITATION.

Although Adapt2Hide achieves satisfactory results, it cannot perform pixel-perfect reconstruction due to the reconstruction bound of AE. Generated texts and small human faces are sometimes distorted. Nevertheless, when the image resolution is high enough (over 1024×1024 pixels), these artifacts are barely noticeable to the human eye.

B. SOCIETAL IMPACTS

The ethical implications of Adapt2Hide must be carefully considered to prevent misuse and protect privacy. This powerful technology can be double-edged. Some image censorship services can be untrustworthy even if they operate offline and their censored images appear visually satisfactory because the obscured sensitive information can be restored. Moreover, future research should focus on enhancing robustness against attacks and ensuring message secrecy.

C. MORE EXPERIMENTS AND ABLATION STUDIES

C.1. Different Rank in LoRA Adapter

To investigate the impact of LoRA adaptor size, we conduct ablation studies across different LoRA ranks, as shown in Table S.1. Increasing the LoRA rank leads to a higher number of additional parameters and improved reversibility performance, demonstrating a trade-off between reversibility and adaptor size.

C.2. Different Budget in the Conformity Layer

To understand the information needed for Adapt2Hide, we conduct the ablation study on the different budgets of the clamp function in the conformity layer and present the results in Table S.1. The imperceptibility of Adapt2Hide is controlled by the budget parameter in the conformity layer’s clamp function. A higher budget allows for greater pixel perturbations, enabling more information to be hidden. The results indicate that increasing the budget parameter of Adapt2Hide indeed reduces imperceptibility but improves reversibility.

C.3. Conformity Layer

To evaluate the effect of the conformity layer, we conduct an ablation study by removing it from Adapt2Hide. The quantitative results in Table S.1 show a significant performance drop without the conformity layer, indicating the conformity layer is crucial.

C.4. Simpler VAE

To understand the effect of the off-the-shelf VAE, we adapt our method to simpler VAE¹. Table S.1 shows that the larger VAE (from Stable Diffusion 1.5) outperforms the simpler VAE across most metrics.

¹<https://huggingface.co/ostris/vae-kl-f8-d16>

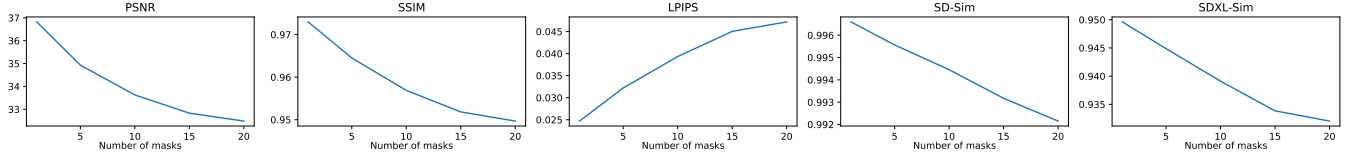


Fig. S.1: Effect of $x_{msg}-x_{cov}$ image distances on reversibility.



Fig. S.2: Qualitative results of reverse style transfer [2]. The cover images and message images generated by Adapt2Hide have no color patch, indicating Adapt2Hide has better imperceptibility and reversibility.

C.5. Different Numbers of Mask

Figure S.1 shows how the number of censorship masks (i.e., the difference between the message and cover images) affects message image reconstruction errors. As the number of masks increases, the reversibility performance decreases.

D. OTHER APPLICATIONS

D.1. Reversible Style Transfer

To the best of our knowledge, we are the first to perform reversible visual censorship and AI editing. However, similar applications have been proposed. For example, self-contained stylization (SelfCon) [2] hides images in their style-transferred counterparts. Figure S.3 shows our method outperforms SelfCon. Unlike SelfCon [2], Adapt2Hide does not generate color patches, making both the encoded cover images and the reconstructed message images visually identical to the ground truth. The quantitative and subjective results in Table S.2 further demonstrate that our method surpasses SelfCon [2] and StyleStegan [17]² in terms of imperceptibility, thanks to our design of clamped steganographic perturbation, as well as reversibility.

D.2. Reversible Mononization and Grayscale

We apply Adapt2Hide in other similar applications, including mononizing binocular videos (Mono3D) [5], and invertible grayscale (InvGray) [4]. Mono3D hides the right views of stereo images in the left views. InvGray hides RGB images in grayscale images. Figure S.3 shows that both our method and previous works perform equally well in each task. Both the encoded cover images and the

reconstructed message images are visually identical to the ground truth. For quantitative comparison, to ensure a fair evaluation, we use the ground-truth images as reference instead of the VAE reconstruction unless stated otherwise. The results in Table S.3 show that our method consistently outperforms previous works in terms of imperceptibility, thanks to our design of clamped steganographic perturbation. However, our method sometimes scores lower on pixel-level metrics when evaluating reversibility. This is expected, as our method’s performance is constrained by the reconstruction ability of the pretrained VAE. However, in terms of human evaluations or perceptual metrics using large models like SD, our method performs comparably to previous approaches, aligning in the visual comparisons in Figure S.3.

D.3. Image Steganography

Adapt2Hide can be used for image steganography [10], which involves hiding an image (the message) within another image (the cover) and then retrieving the message image from the cover image. We train Adapt2Hide for image steganography using ImageNet1K [30], where two images are randomly sampled, one as the cover image and the other as the message image. Table S.4 demonstrates the quantitative results, while Figure S.4 shows the qualitative results. These results indicate that our method achieves excellent imperceptibility, making it difficult for human observers to detect artifacts. Although the quantitative reversibility is limited by the performance of the AE, the artifacts in the qualitative results are nearly unnoticeable to the human eye.

D.4. Black Image as Cover Image

To validate that Adapt2Hide can also work under the case of completely unrelated cover and message images, we use pure black images as the cover images. This is the extreme case of the black cen-

²We use their reported data. However, we do not compare quantitative imperceptibility or qualitative results because their code is not available.

Method			Imperceptibility					Reversibility			Human rating [†]
	PSNR [†]	SSIM [†]	LPIPS [↓]	SD-Sim [†]	SDXL-Sim [†]	PSNR [†]	SSIM [†]	LPIPS [↓]	SD-Sim [†]	SDXL-Sim [†]	
SelfCon [2]	21.8	0.6386	0.3998	0.7241	0.6607	18.21	0.5055	0.4207	0.6286	0.5084	2.59 ± 1.32
StyleStegan [17]	-	-	-	-	-	17.14	0.5145	0.3802	-	-	-
Ours	41.61	0.9795	0.0143	0.9903	0.9692	26.77	0.7641	0.1603	0.9845	0.7762	4.41 ± 0.84

Table S.2: Imperceptibility and reversibility on reversible style transfer [2] (MS COCO [38]).

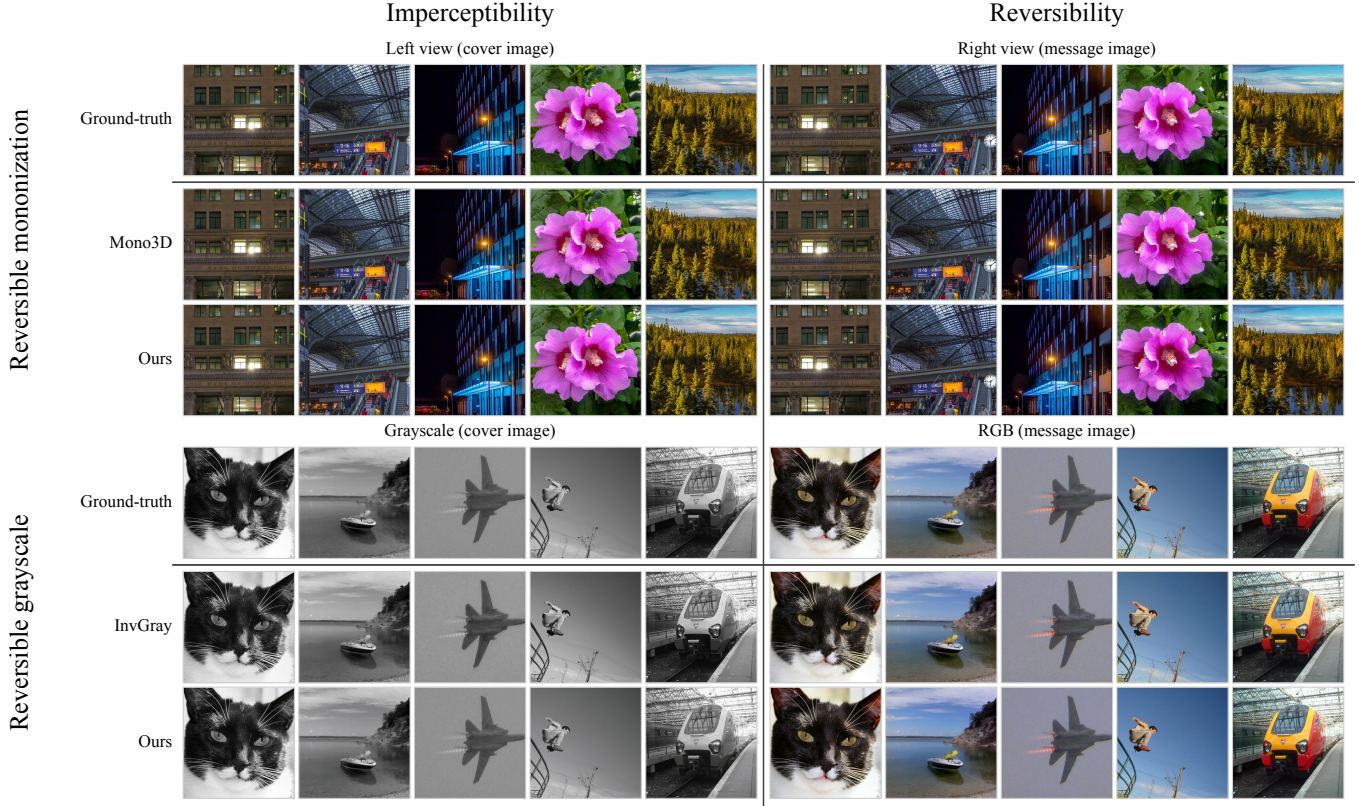


Fig. S.3: Qualitative results of reversible mononization (top) and reversible grayscale (bottom).

(a) Reversible mononization (Flickr1024 [36])											
Method			Imperceptibility					Reversibility			Human rating [†]
	PSNR [†]	SSIM [†]	LPIPS [↓]	SD-Sim [†]	SDXL-Sim [†]	PSNR [†]	SSIM [†]	LPIPS [↓]	SD-Sim [†]	SDXL-Sim [†]	
Mono3D [5]	39.34	0.9709	0.0104	0.9954	0.9864	38.62	0.9804	0.0202	0.9895	0.9702	4.51 ± 0.77
Ours	40.99	0.9659	0.0152	0.9943	0.9845	22.44	0.6619	0.1960	0.9859	0.8035	4.42 ± 0.90
$\mathcal{D}(\mathcal{E}(\cdot))^*$	-	-	-	-	-	23.0	0.6733	0.1841	0.9893	0.8082	-
Ours [†]	-	-	-	-	-	29.53	0.9321	0.0466	0.9940	0.9391	-

(b) Reversible grayscale (VOC2012 [37])											
Method			Imperceptibility					Reversibility			Human rating [†]
	PSNR [†]	SSIM [†]	LPIPS [↓]	SD-Sim [†]	SDXL-Sim [†]	PSNR [†]	SSIM [†]	LPIPS [↓]	SD-Sim [†]	SDXL-Sim [†]	
InvGray [4]	34.17	0.9712	0.1153	0.9639	0.9437	39.74	0.9785	0.0407	0.9860	0.9653	4.57 ± 0.71
Ours	41.51	0.9752	0.0940	0.9759	0.9465	27.86	0.7905	0.1462	0.9917	0.8312	4.46 ± 0.82
$\mathcal{D}(\mathcal{E}(\cdot))^*$	-	-	-	-	-	28.17	0.7933	0.1427	0.9925	0.8321	-
Ours [†]	-	-	-	-	-	38.51	0.9846	0.0169	0.9982	0.9655	-

* Performance bound. [†] $\mathcal{D}(\mathcal{E}(\cdot))$ as ground-truth.

Table S.3: Imperceptibility and reversibility on reversible mononization, reversible grayscale.

Method			Imperceptibility					Reversibility		
	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	SD-Sim $\uparrow$	SDXL-Sim $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	SD-Sim $\uparrow$	SDXL-Sim $\uparrow$
Ours	41.25	0.9685	0.0257	0.9898	0.9712	26.91	0.7688	0.1529	0.9858	0.8301

Table S.4: Imperceptibility and reversibility on image steganography (ImageNet1K [30]).

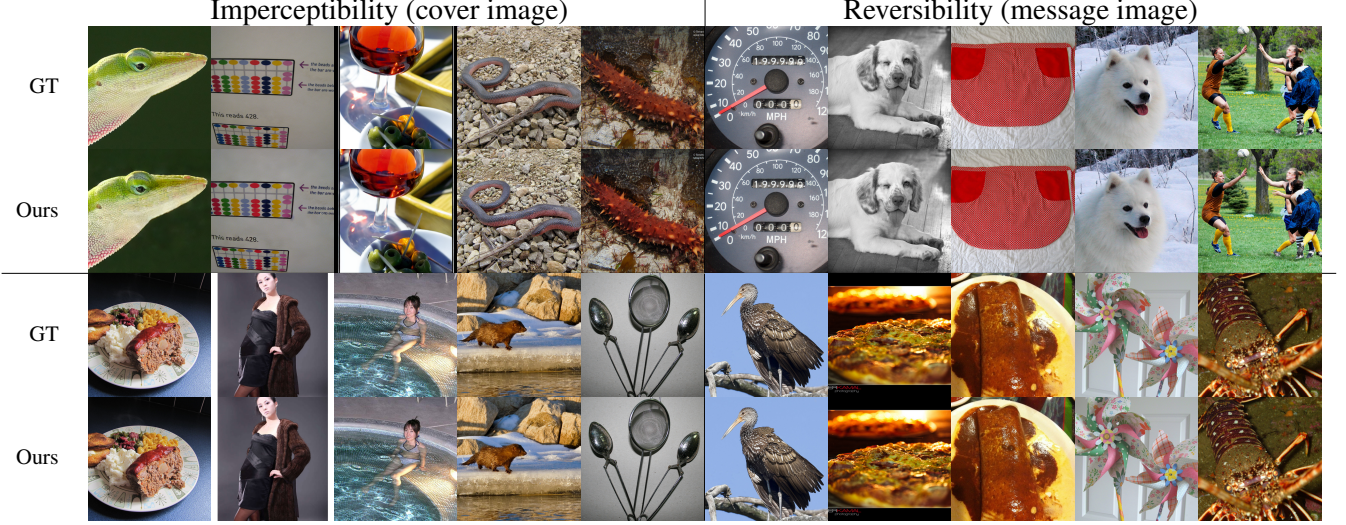


Fig. S.4: Qualitative results for imperceptibility and reversibility of image steganography. GT: Ground-truth.



Fig. S.5: Qualitative results for black image as cover image, which demonstrates the generalizability of Adapt2Hide.

ship in Section 3.1, with the entire mask $\mathcal{M}$ set to 1, and the whole image is blacked out. The qualitative results are presented in Figure S.5. Although the recovered image exhibits some shifts, our method remains robust even when Adapt2Hide is not exposed with any explicit information about the message image, showing its capability in more general image steganography.

E. MORE RESULTS

We present sets of results on reversible visual censorship and reversible AI editing in Figure S.6, Figure S.7, Figure S.8, Figure S.9, Figure S.10, and Figure S.11.

F. FULL SIDE-BY-SIDE COMPARISONS OF TEASER

We provide full side-by-side comparisons of teaser in Figure S.12.

G. IRB STATEMENT

The user study was approved by the Institutional Review Board for Humanities & Social Science Research (IRB-HS), Academia Sinica, under protocol number AS-IRB-HS-24031. Prior to participation, all participants were informed about the study procedure and provided consent. Participation was voluntary, and participants could withdraw at any time. The collected responses were anonymized and used only for research purposes. No personally identifiable information is reported.

Please check following pages for more figures.

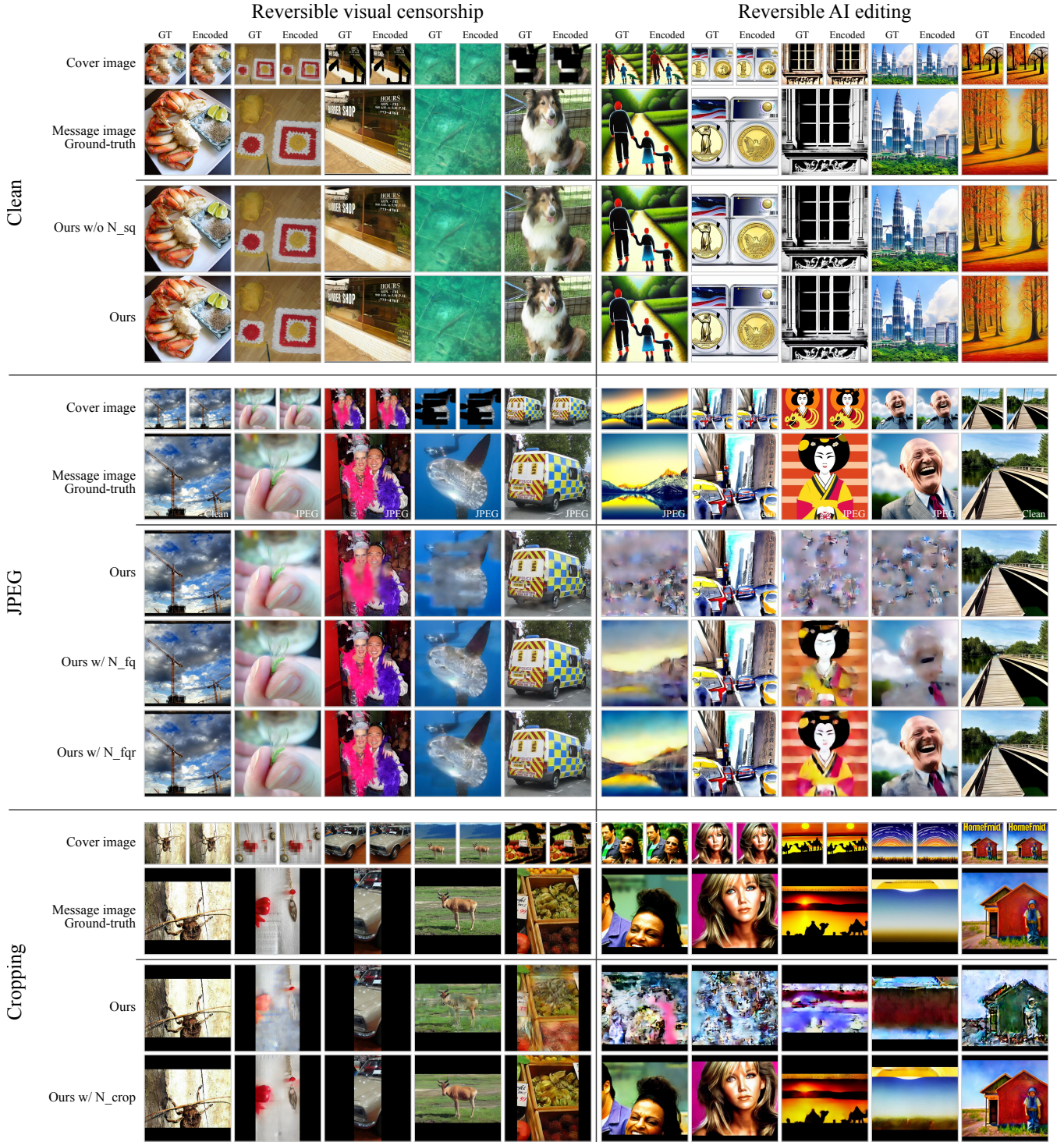


Fig. S.6: Uncurated samples. From left to right: reversible visual censorship on ImageNet-1k [30], reversible AI editing on InstructPix2Pix [31]. From top to bottom: images under clean setting sampled with random seed 123, images under JPEG setting with random seed 1234, images under Cropping setting with random seed 12345.

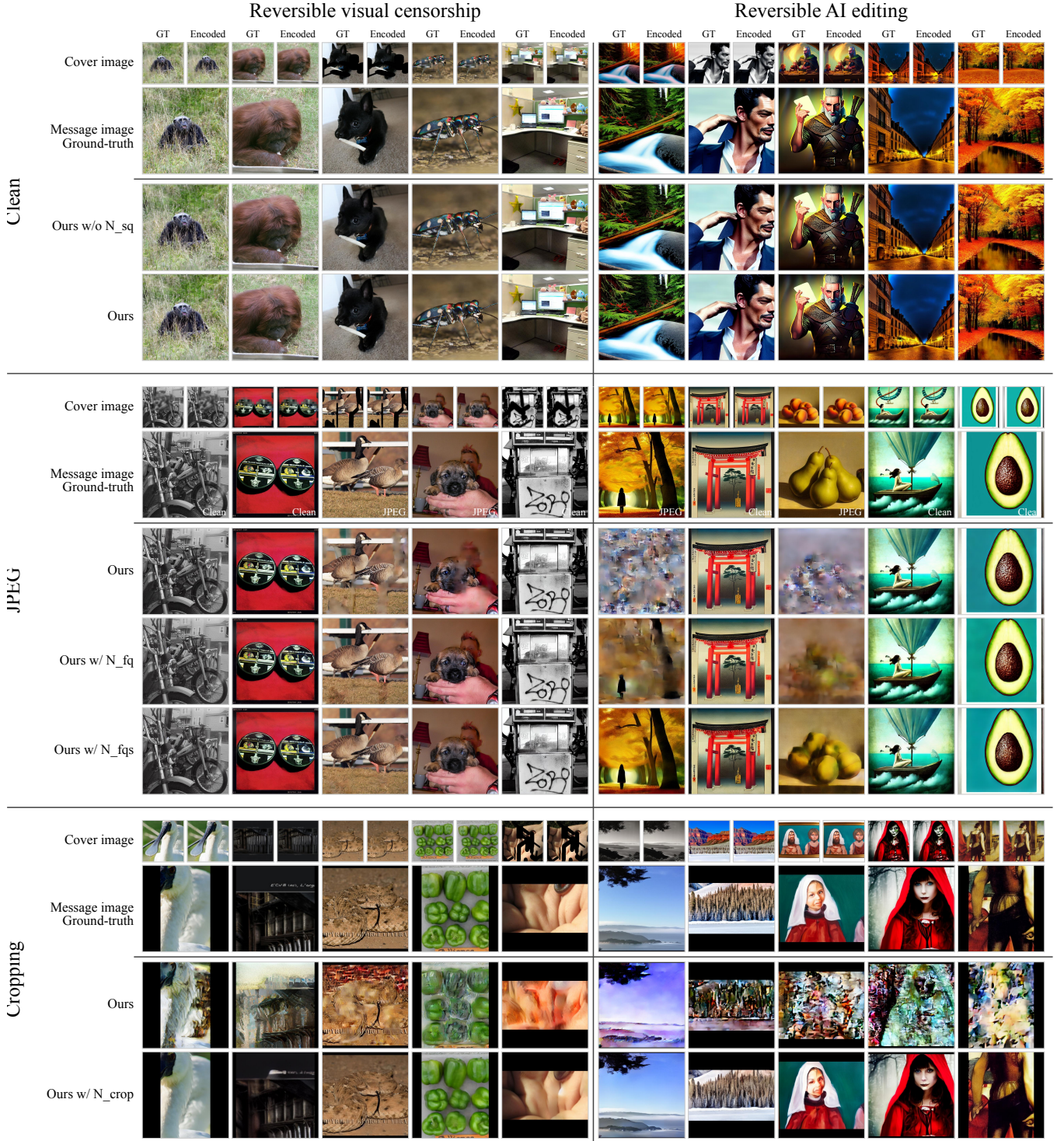


Fig. S.7: Uncurated samples. From left to right: reversible visual censorship on ImageNet-1k [30], reversible AI editing on InstructPix2Pix [31]. From top to bottom: images under clean setting sampled, images under JPEG setting, images under Cropping setting.

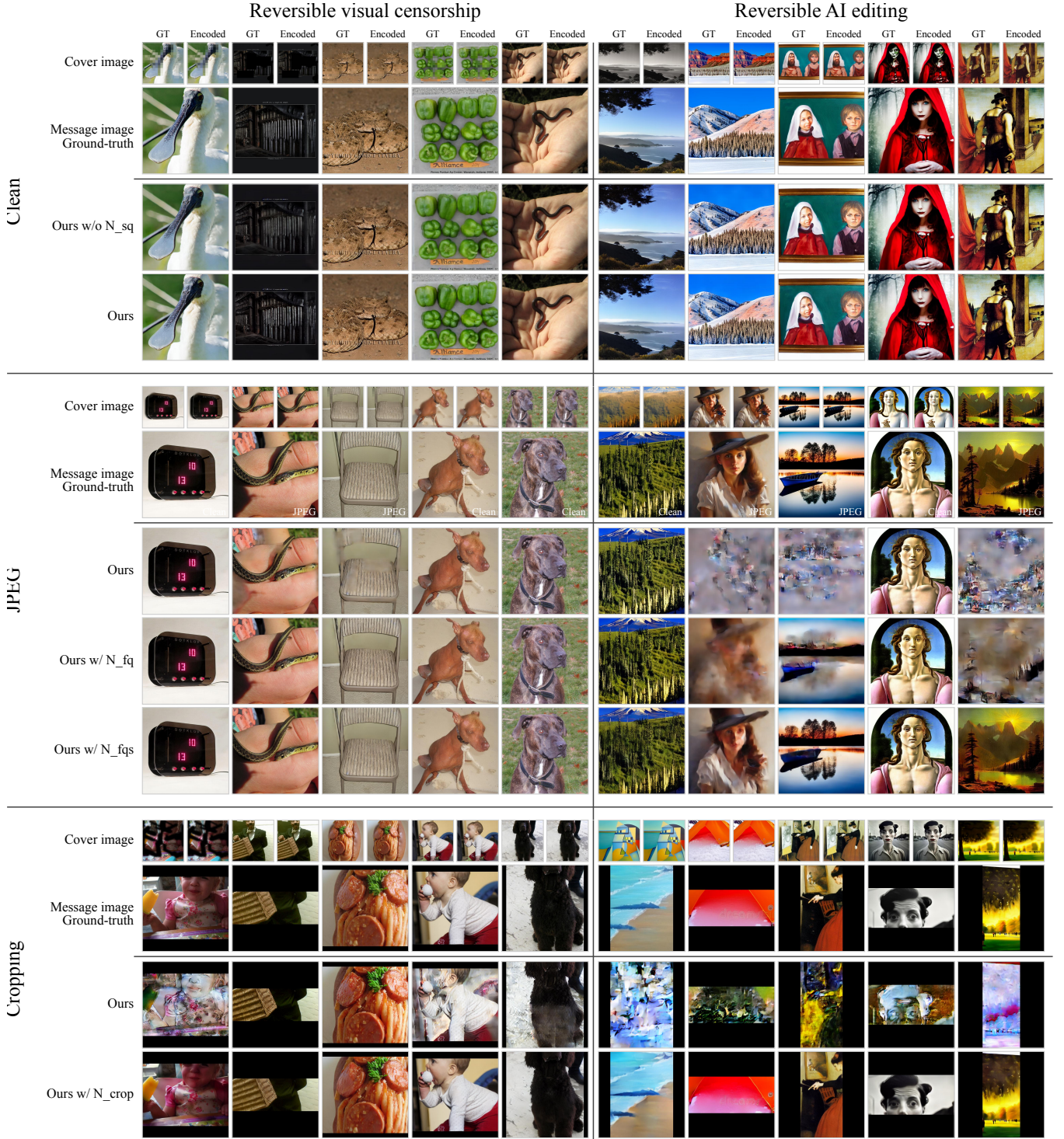


Fig. S.8: Uncurated samples. From left to right: reversible visual censorship on ImageNet-1k [30], reversible AI editing on InstructPix2Pix [31]. From top to bottom: images under clean setting sampled, images under JPEG setting, images under Cropping setting.

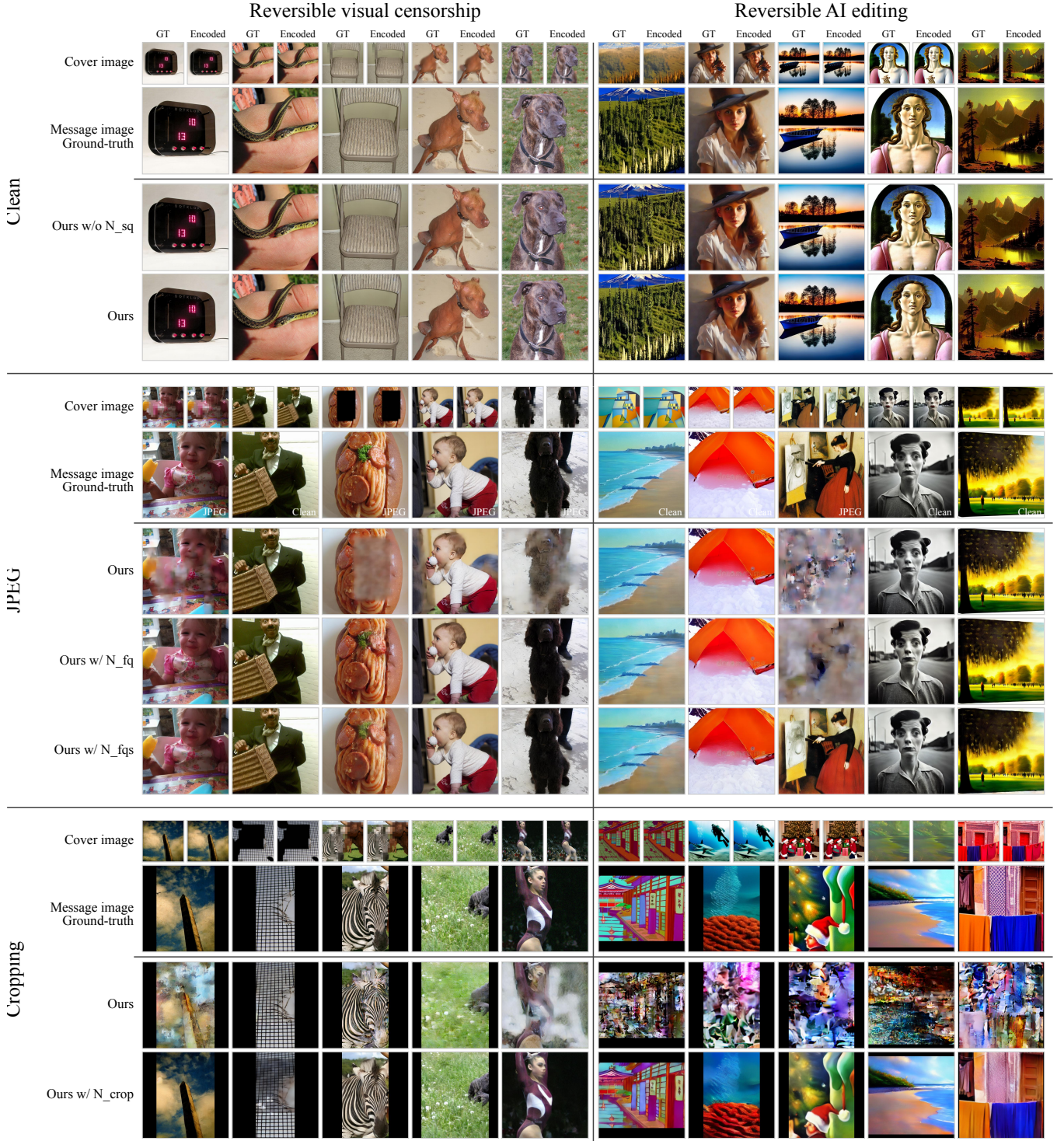


Fig. S.9: Uncurated samples. From left to right: reversible visual censorship on ImageNet-1k [30], reversible AI editing on InstructPix2Pix [31]. From top to bottom: images under clean setting sampled, images under JPEG setting, images under Cropping setting.

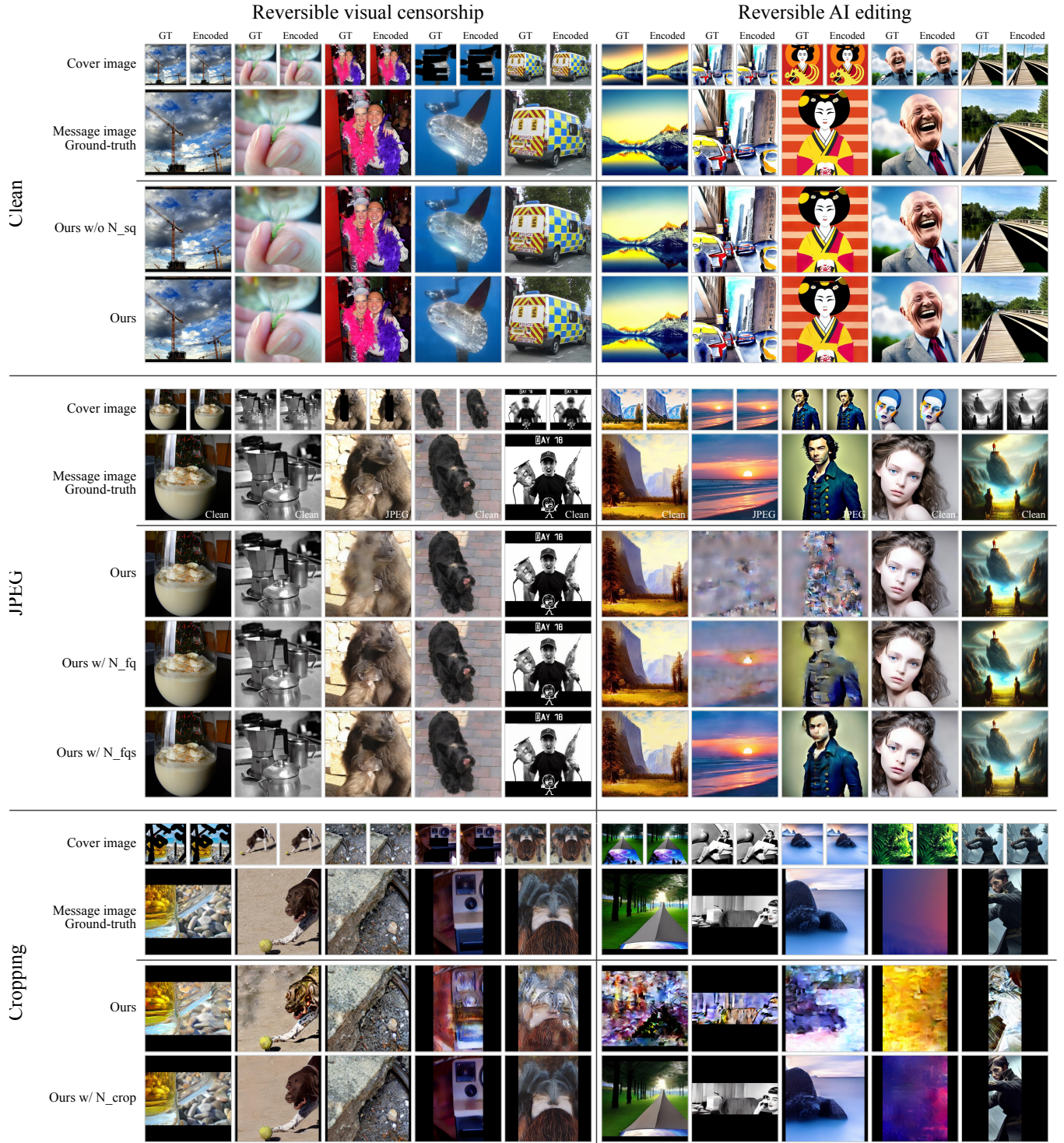


Fig. S.10: Uncurated samples. From left to right: reversible visual censorship on ImageNet-1k [30], reversible AI editing on InstructPix2Pix [31]. From top to bottom: images under clean setting sampled, images under JPEG setting, images under Cropping setting.

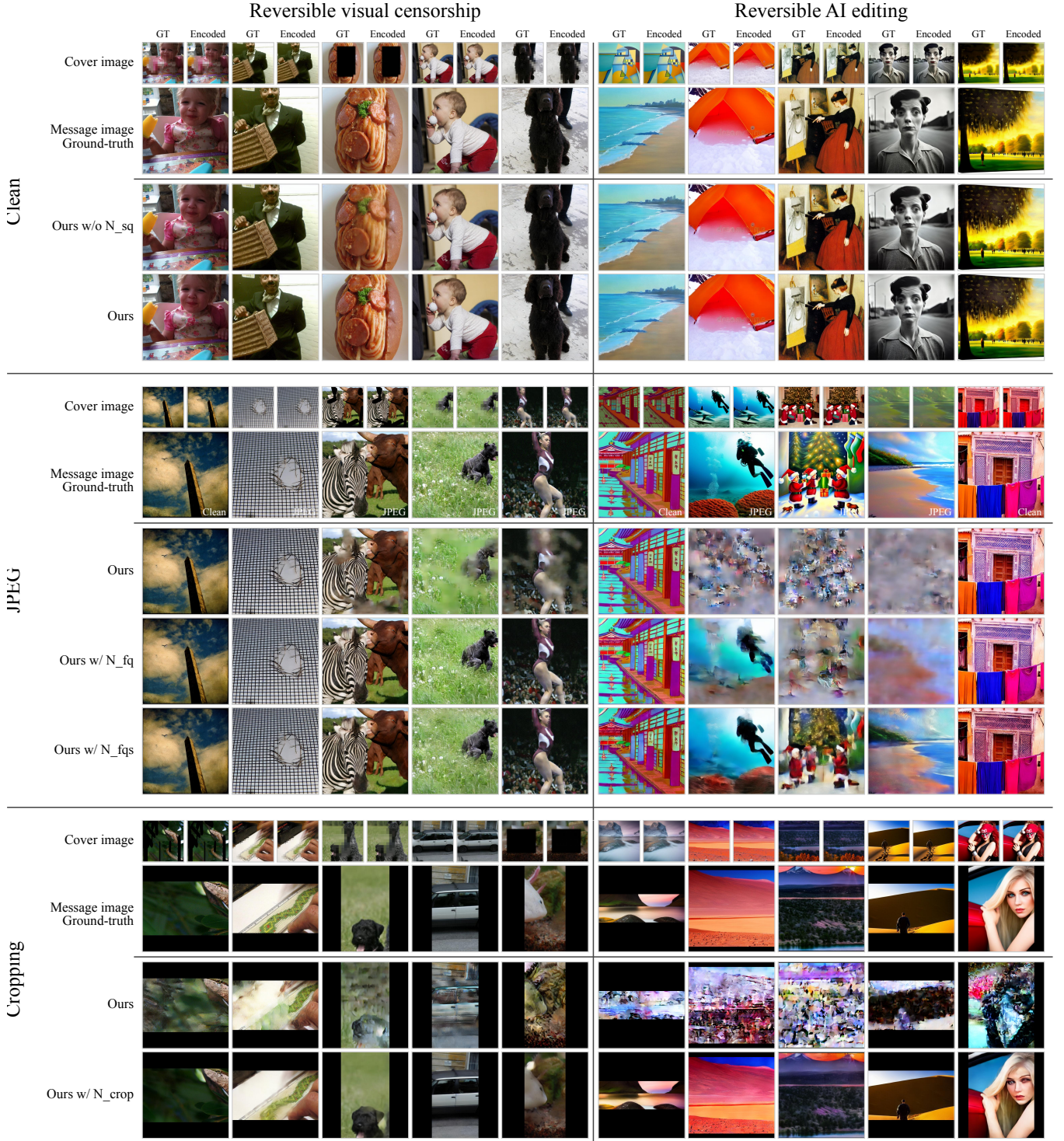


Fig. S.11: Uncurated samples. From left to right: reversible visual censorship on ImageNet-1k [30], reversible AI editing on InstructPix2Pix [31]. From top to bottom: images under clean setting sampled, images under JPEG setting, images under Cropping setting.

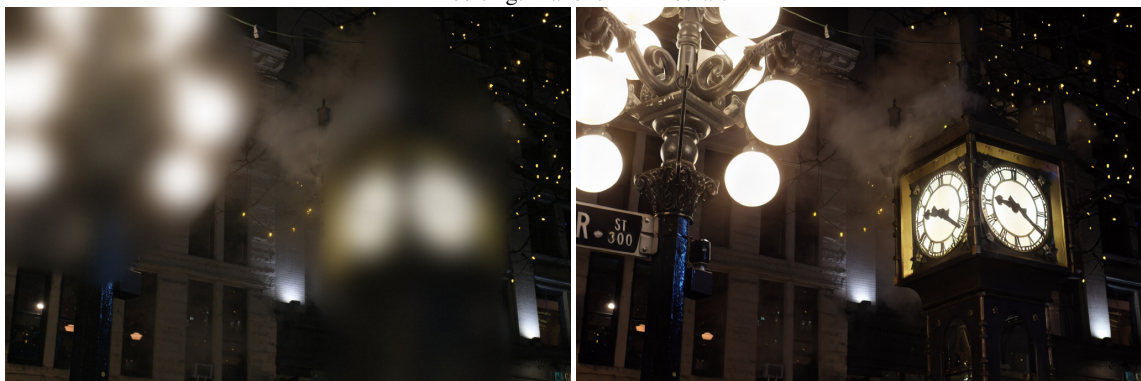


Censorship: pixelation

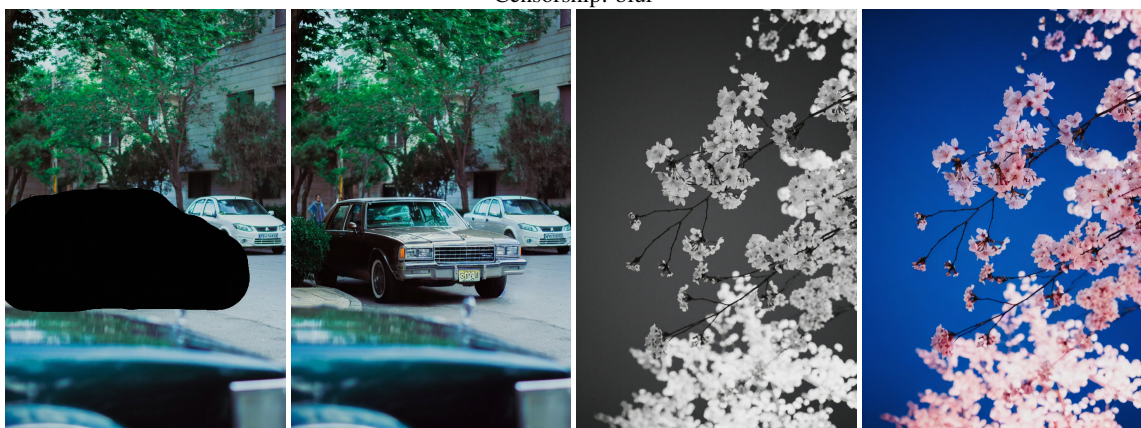
AI editing: turn him into marble statue



AI editing: make it in Minecraft



Censorship: blur



Censorship: black

Grayscale

Fig. S.12: Irreversible destructive image processing and our reconstruction. Image credit to Unsplash, the White House and NYU Tandon.