Supplementary Material for: SmartEraser: Remove Anything from Images using Masked-Region Guidance

Longtao Jiang[1,*] Zhendong Wang[1,*] Jianmin Bao[2,*†♡]
Wengang Zhou[1,†] Dongdong Chen[2] Lei Shi[2] Dong Chen[2] Houqiang Li[1]
[1]University of Science and Technology of China [2]Microsoft Research Asia
https://longtaojiang.github.io/smarteraser.github.io/

1. More Dataset Details

In this section, we provide more implementation details about the proposed Syn4Removal dataset.

Instance Filter. Following the previous work [7], considering the variability of CLIP score distribution across different instance classes, we compute thresholds $thres_c$ for each class c, which is formulated as:

$$thres_c = min(b, max(S_c) - d), \tag{1}$$

where S_c is the CLIP score sets of all instances in the class c, b and d are predefined thresholds, set to 0.2 and 0.02, respectively. After that, we exclude instances whose CLIP scores fall below the threshold of their corresponding class.

Background Filter. To obtain suitable background images from public datasets COCONut [2] and SAM-1B [4] for instance pasting, we apply several filtering criteria. 1) Background images are excluded if their width or height resolution is below 512 pixels, to prevent low-resolution images from being used as ground truths during the training process. 2) Images with an aspect ratio larger than 2 are also discarded, to reduce the risk of image distortion during resizing and cropping to 512×512 resolutions before inputting to SD v1.5. 3) If the total area covered by instances in the background image exceeds 85%, the image is also excluded, as it becomes challenging to compute a suitable region for pasting instances. After filtering, we obtain approximately 750k background images from SAM-1B [4], and 282k background images from COCONut [2], resulting in about 1M training data.

2. More Framework Details

In this section, we provide more implementation details about the architecture of the MLP module in the proposed CLIP-based visual guidance.

As shown in Figure 1, the MLP module is designed with a simple but effective architecture. It is employed to map the visual feature extracted by the CLIP vision encoder into the feature space of the text encoder. It consists of two linear layers, a layer normalization layer, and a GELU [3] activation function. In addition, a residual connection is incorporated between the input and output to preserve the original feature generated by the CLIP visual encoder. The dimensions of the input, hidden state, and output are equal to 768.

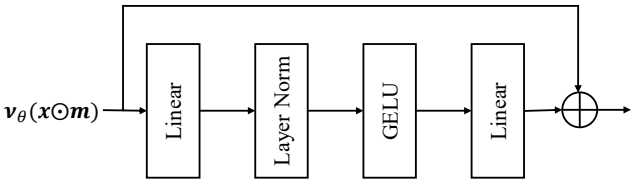

Figure 1. The detailed architecture of the MLP module.

3. Samples from Syn4Removal

In Figure 8, we present examples from Syn4Removal, including input images, masks, and ground-truth images, with the carefully designed pipeline for creating triplets. To prevent excessive overlap with existing instances in the background, we compute the Intersection over Union (IoU) between the pasted objects and each instance in the background, carefully determining the location of the pasted objects. Additionally, the pasted instances are seamlessly integrated into the background images by using alpha blending, ensuring visual harmony. Our final triplet data is high-quality and more suitable for the object removal task.

4. Additional Experiments

In this section, we provide more comprehensive experiments with quantitative and qualitative results reported to demonstrate the effectiveness of our method.

Comparisons with Instruction-Based Methods. In Figure 2, we provide qualitative comparisons between instruction-based methods [1, 6] and our SmartEraser. InstructPix2Pix [1] struggles in removing objects, often failing to

* Equal contribution. † Corresponding authors. ♡ Project leader.

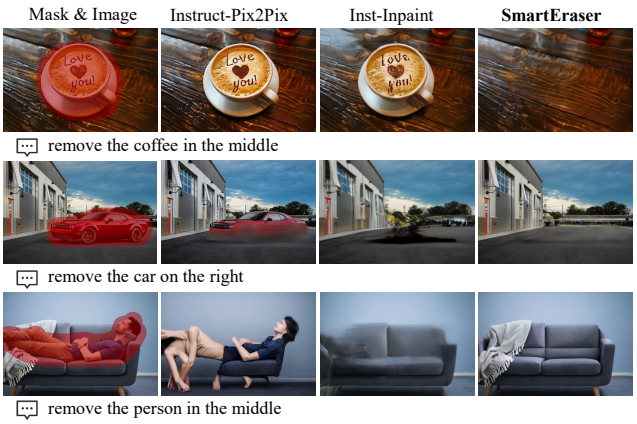

Figure 2. Qualitative comparison of SmartEraser and existing instruction-based image editing method, which only rely on user input prompts for object removal.

Method	FID ↓	CMMD ↓	ReMOVE ↑	LPIPS ↓	SSIM ↑	PSNR ↑
Baseline	10.125	0.189	0.916	0.315	0.674	22.135
+RG	7.142	0.132	0.932	0.279	0.712	24.176
+ME	4.701	0.125	0.938	0.271	0.721	24.917
+VG	3.405	0.106	0.939	0.257	0.734	25.363

Table 1. Quantitative ablated comparison on DEFACTO-Val.

remove the target objects, instead, introducing unrealistic edits to the objects. Another work Inst-Inpaint [6] demonstrates a basic capability for object removal, but it faces challenges in more complex cases, resulting in incomplete object removal and poor coherence between the removed area and the background, as observed in the first and second samples. Additionally, Inst-Inpaint has difficulty identifying the target object only based on user instructions. For instance, while the person is removed in the third sample, the stall on the sofa is also removed, which is contrary to the intention of the user. Furthermore, since it is trained on synthetic ground truths, Inst-Inpaint tends to produce blurred generation. In contrast, SmartEraser effectively removes target objects accurately and preserves high background consistency and overall image quality.

Ablation Studies on Other Benchmarks. We further supply the quantitative experiments of ablation studies on DEFACTO-Val and Syn4Removal-Val. As shown in Tables 1 and 2, the overall performance progressively improves as key techniques are incrementally integrated into the baseline. In the Tables, RG represents the masked-region guidance, ME denotes mask enhancement, and VG indicates CLIP-based visual guidance. These quantitative results demonstrate the effectiveness of each key component in our proposed framework.

More Qualitative Results of Ablation Studies. As shown in Figure 3, we provide additional qualitative results of ablation studies, with images sourced from the validation set of MSCOCO [5]. These examples reveal that the baseline model, trained using the "mask-and-inpaint" paradigm

Method	FID ↓	CMMD ↓	ReMOVE ↑	LPIPS ↓	SSIM ↑	PSNR ↑
Baseline	8.755	0.104	0.908	0.335	0.584	18.735
+RG	5.419	0.081	0.927	0.299	0.632	20.176
+ME	4.748	0.067	0.934	0.281	0.661	21.317
+VG	4.386	0.053	0.939	0.269	0.672	22.029

Table 2. Quantitative ablated comparison on Syn4Removal-Val.

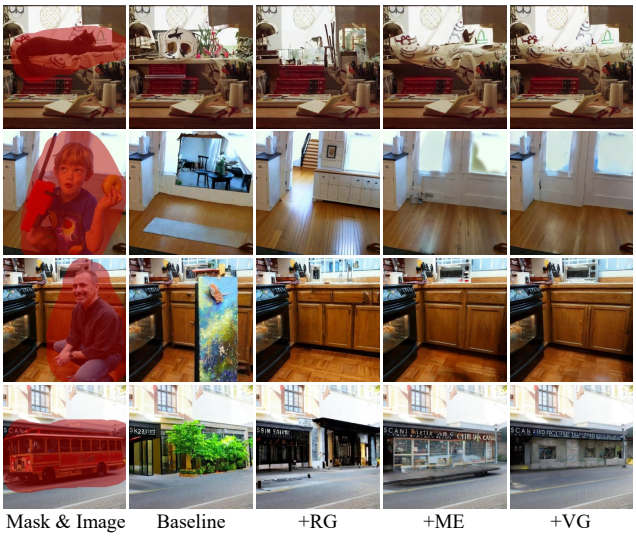

Figure 3. Qualitative ablation comparison of our method. From left to right, we progressively add each proposed component.

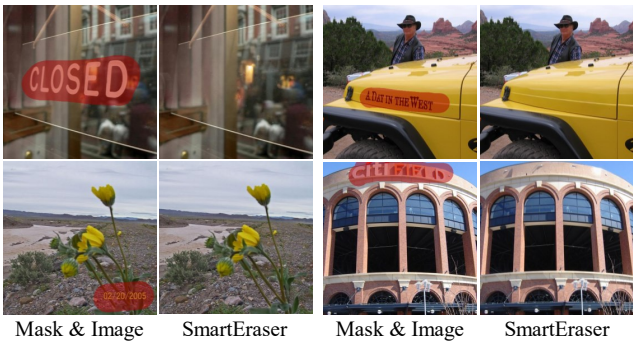

Figure 4. Qualitative results of our method for erasing scene text.

on the Syn4Removal dataset, frequently regenerates unintended objects within the masked regions. After adding the masked-region guidance, the risk of regeneration is significantly reduced. However, we observe that the background context within the masks around the target objects has also changed obviously. To address this, we introduce the mask enhancement to simulate the user-provided masks to reduce the gap between the training and inference, which effectively helps the model to preserve the surrounding context and further improve the removal results. Finally, integrating CLIP-based visual guidance (VG) provides explicit semantic guidance, enabling the model to achieve better object removal performance with more coherence and fidelity.

More Qualitative Results with Previous Methods. To better express the excellent performance of SmartEraser in

Figure 5. Qualitative results for partial object removal.

Figure 6. Qualitative results for occluded object removal.

object removal, we present more qualitative comparisons with previous methods in Figures 9 and 10, the images are sourced from the RORD-Val and the DEFACTO [5] splicing section, respectively. These results demonstrate that SmartEraser effectively removes target objects and consistently outperforms existing approaches.

More Real-world User Cases. To further demonstrate the ability of SmartEraser to smartly remove the target objects while preserving the surrounding background context, we present more real-world user cases compared with previous methods in Figures 11 and 12. The images are sourced from the RORD-Val and the MSCOCO [5] validation set. These examples show the capability of SmartEraser to accurately remove objects and maintain high consistency in background context within masks when facing diverse and complex real-world user cases.

Erasing Scene Text. As shown in Figure 4, we explore the capability of SmartEraser when applying it to erasing scene text. We observe that based on the proposed novel paradigm, our SmartEraser can remove scene text seamlessly without hurting the surrounding context.

Partial object removal. We conducted experiments on erasing partial objects. The results in the following Figure 5 show that SmartEraser effectively identifies and removes the partial object specified by user mask.

Occluded object removal. The following Figure 6 shows that SmartEraser effectively removes occluded objects while successfully inpainting background. During Syn4Removal synthesis, objects are placed based on IoU criteria (Formula 3 in paper), allowing some occlusions.

5. Limitations

Although the strong performance is shown in the aforementioned Tables and Figures, we recognize there is a potential limitation. Our novel masked-region guidance helps the model identify the target objects and avoid unintentional regeneration, but if the masked region contains too many objects, the model may fail to completely remove objects. This is because our paradigm may fail to identify all the target objects in the masks while considering some of them as background context. A straightforward solution is to draw masks progressively to remove all that wants to be removed. The results of progressive removal are shown in Figure 7. We observe that when facing multiple objects, our SmartEraser can progressively remove all the target objects.

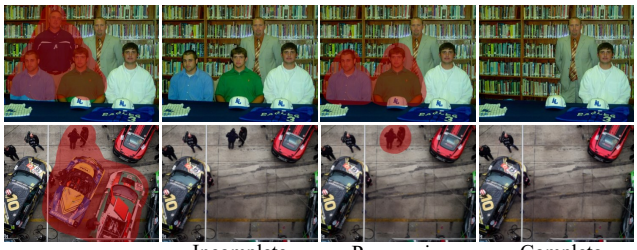

Figure 7. Progressive object removal when needing to remove too many objects.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022. 1

[2] Xueqing Deng, Qihang Yu, Peng Wang, Xiaohui Shen, and Liang-Chieh Chen. Coconut: Modernizing coco segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21863–21873, 2024. 1

[3] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 1

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1

[5] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 3

[6] Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegul Dundar. Inst-inpaint: Instructing to remove objects with diffusion models. *arXiv preprint arXiv:2304.03246*, 2023. 1, 2

[7] Hanqing Zhao, Dianmo Sheng, Jianmin Bao, Dongdong Chen, Dong Chen, Fang Wen, Lu Yuan, Ce Liu, Wenbo Zhou, Qi Chu, et al. X-paste: Revisiting scalable copy-paste for instance segmentation using clip and stablediffusion. In *International Conference on Machine Learning*, pages 42098–42109. PMLR, 2023. 1

Input Image Mask Ground Truth Input Image Mask Ground Truth

Figure 8. Samples from the Syn4Removal dataset, including input images, masks, and ground-truth images.

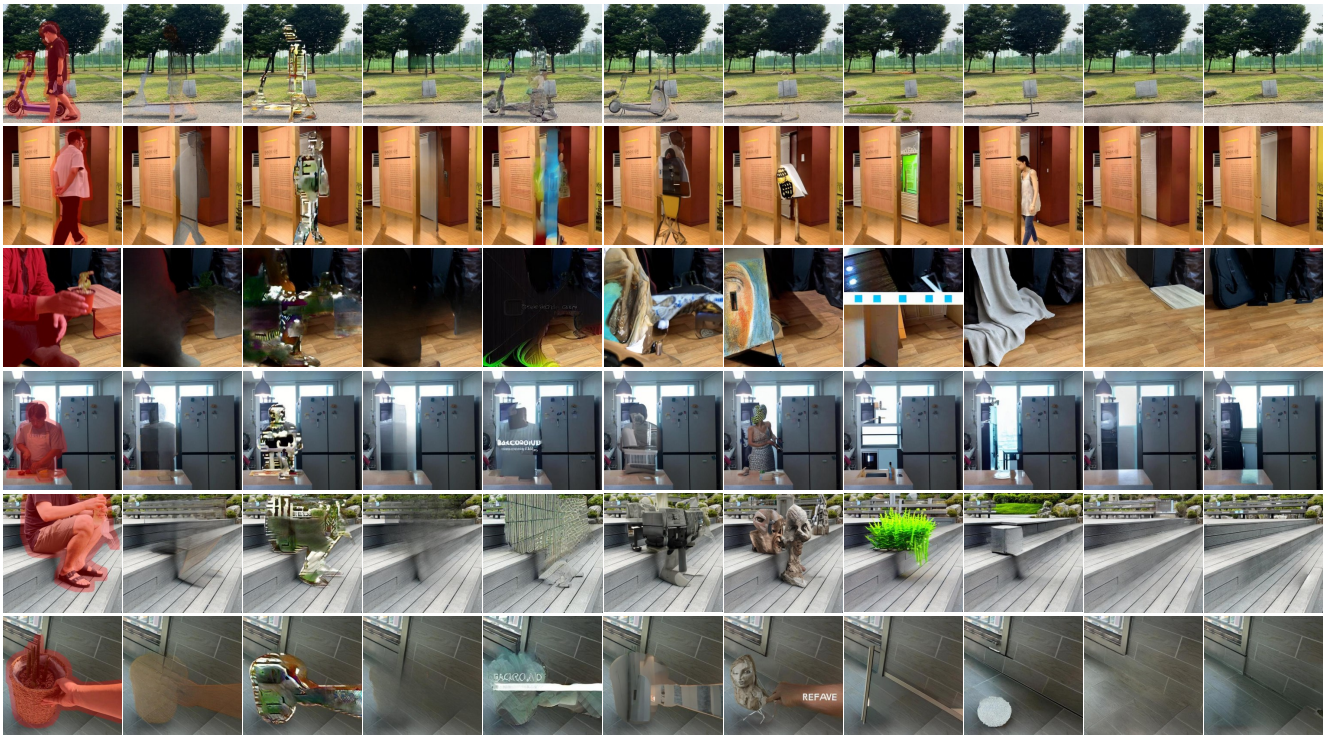

Figure 9. Qualitative comparison of previous methods and SmartEraser. The samples are sourced from RORD-Val.

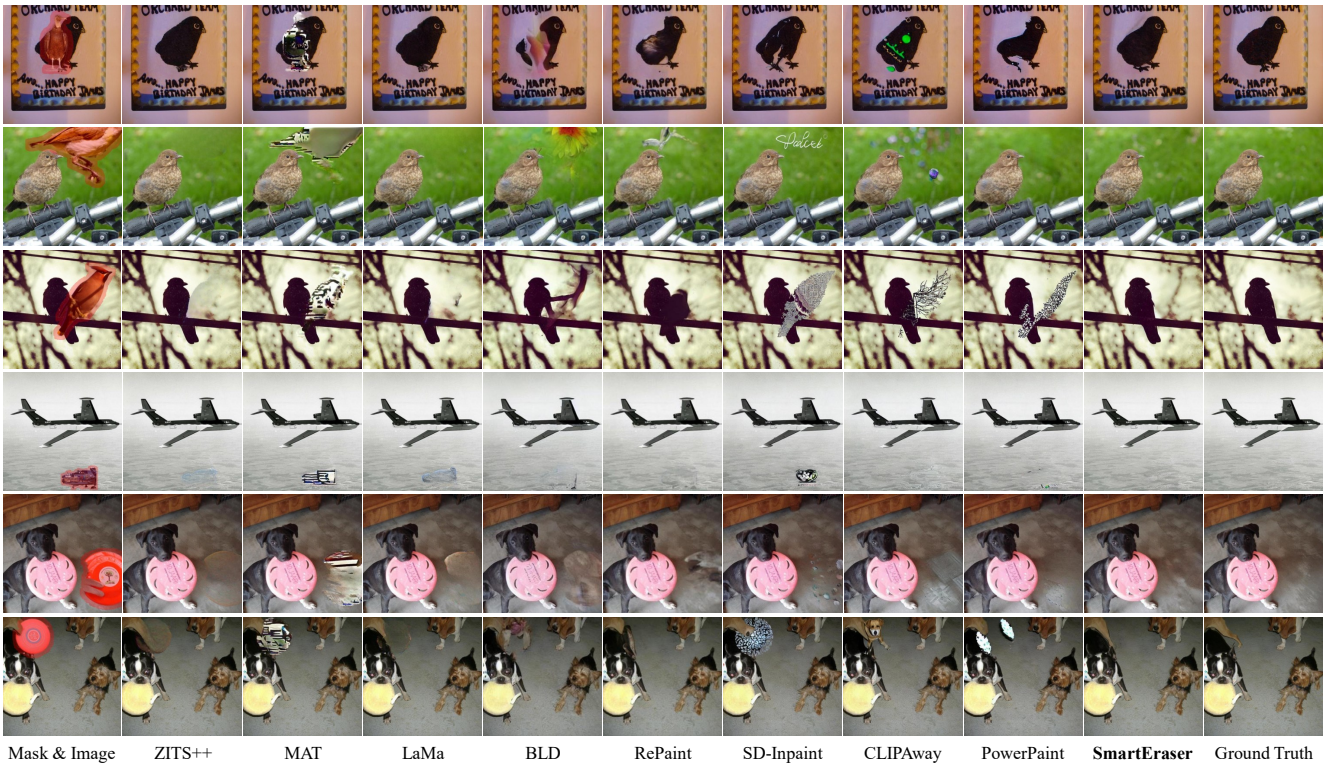

Figure 10. Qualitative comparison of previous methods and SmartEraser. The samples are sourced from the splicing section in DEFACTO dataset.

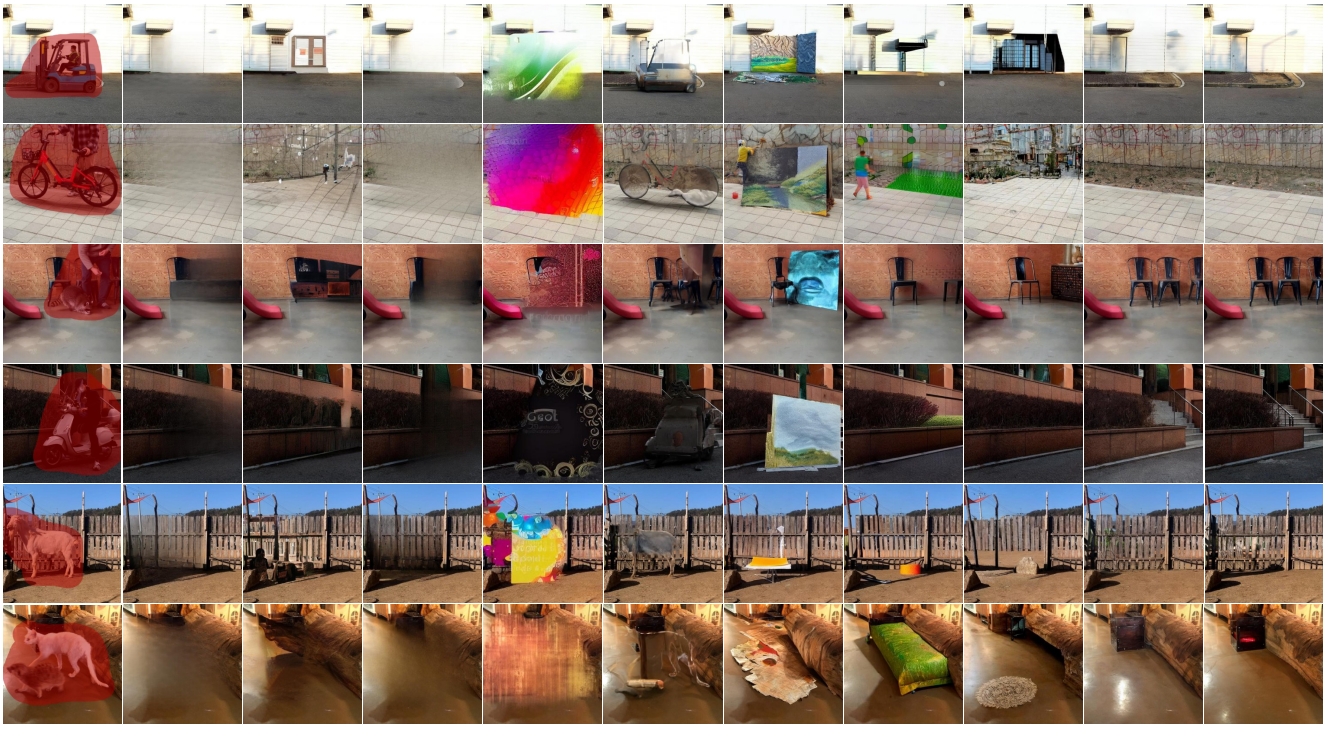

Figure 11. Qualitative comparison of different methods in real-world user cases. The samples are sourced from RORD-Val.

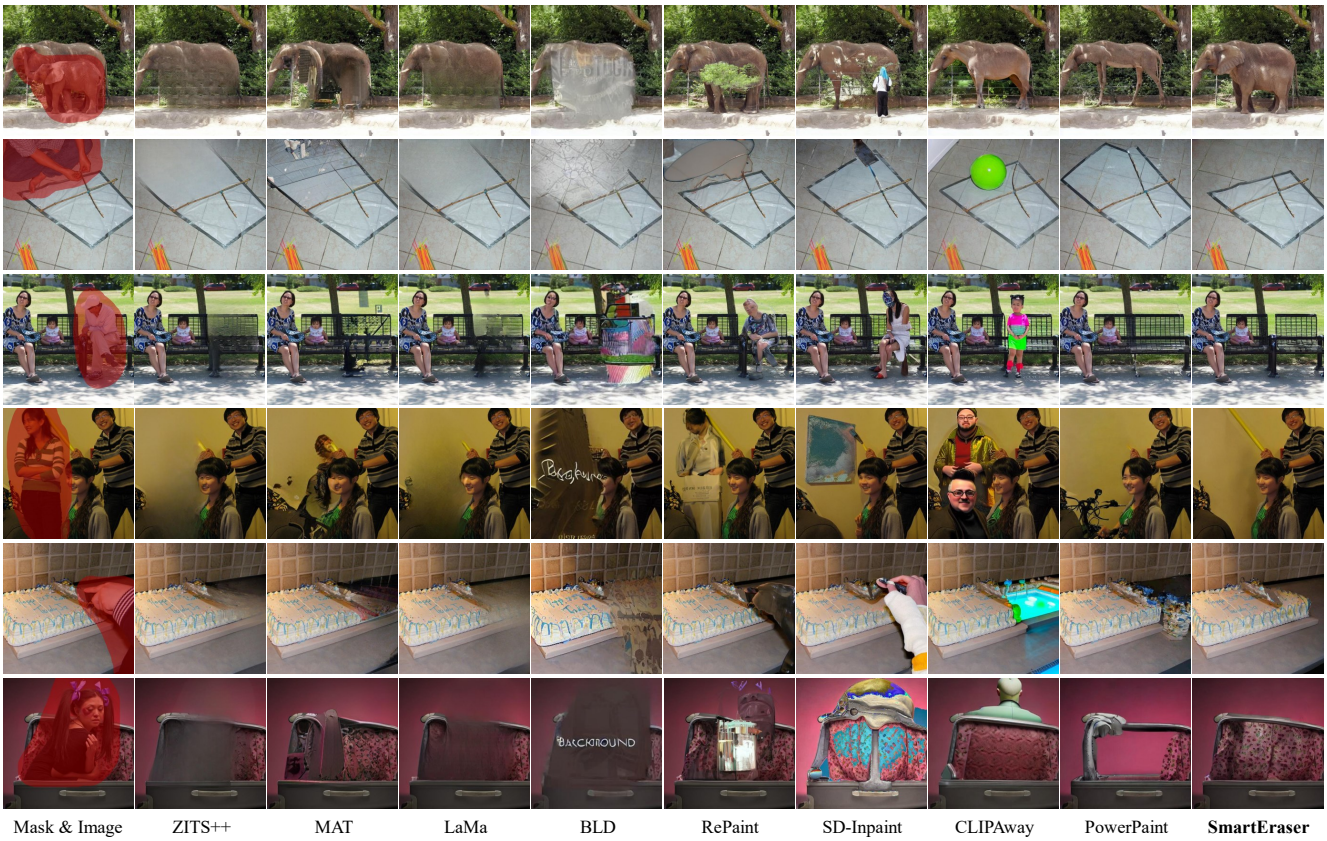

Figure 12. Qualitative comparison of different methods in real-world user cases. The samples are sourced from the validation set in MSCOCO. Note that there are no ground truths.