RORem: Training a Robust Object Remover with Human-in-the-Loop

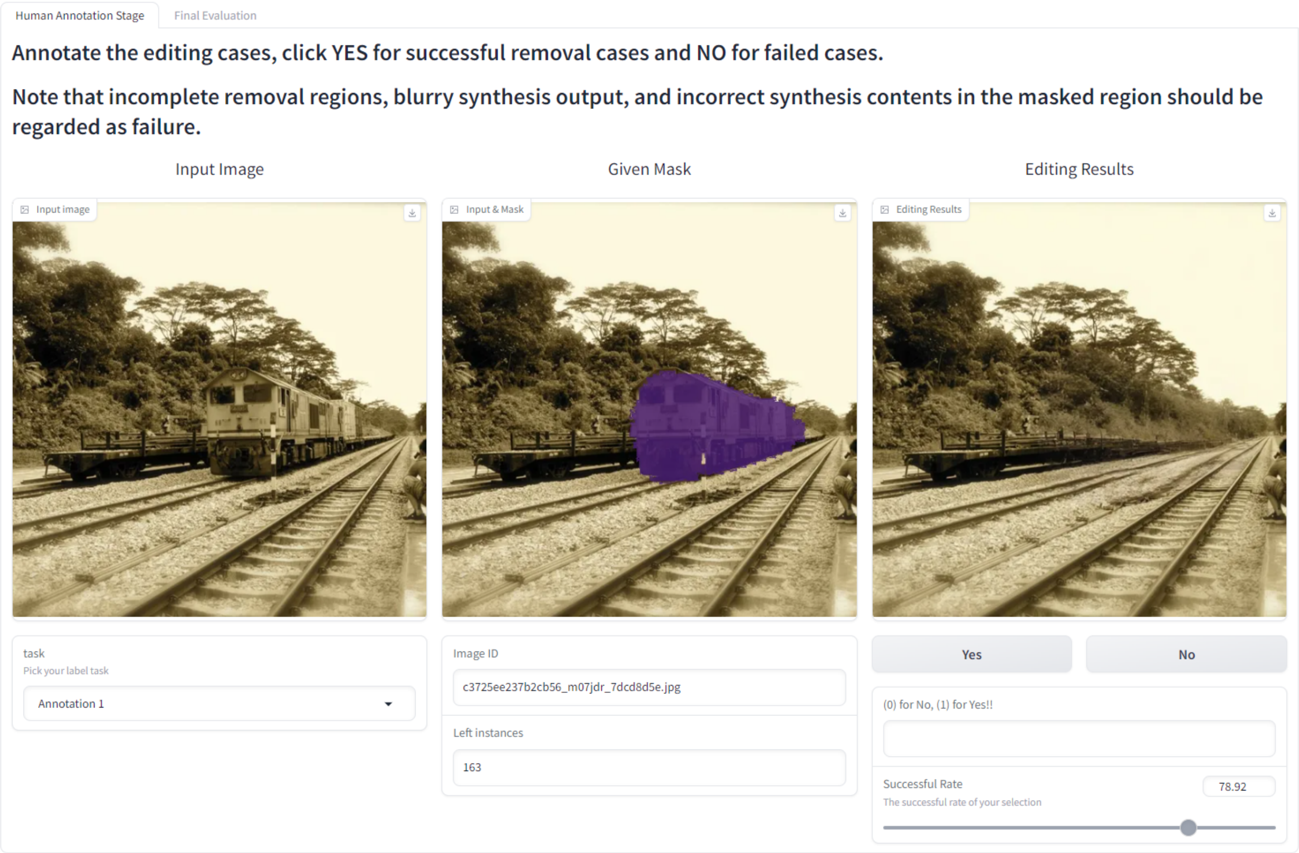
Supplementary Material

In this supplementary file, we provide the following materials:

- The interface and more details of the user study (referring to Sec. 4.1 in the main paper);
- Details of the evaluation metrics for the discriminator (referring to Sec. 4.1 and Sec. 4.3 in the main paper);
- Visual examples of object removal results at each training round (referring to Sec 4.2 in the main paper);
- Visual results of IP2P and CLIPAway (referring to Sec 4.2 and Fig. 6 in the main paper);
- Visual results on images of 512 × 512 resolution (referring to Sec 4.2 in the main paper);
- Failure cases (referring to Sec. 5 in the main paper).

A. Annotation page and user study page

We design a webpage based on the open-source library Gradio [1] to conduct the human annotation (referring to Sec. 3.2 in the main paper) and the final human evaluation (referring to Sec. 4.1 in the main paper). Annotators are provided with the original images, the mask images and the object removal results, as illustrated in Fig. 7. They are asked to provide feedback by clicking the **Yes** or **No** button at the bottom right corner.



Annotate the editing cases, click YES for successful removal cases and NO for failed cases.

Note that incomplete removal regions, blurry synthesis output, and incorrect synthesis contents in the masked region should be regarded as failure.

task
Pick your label task
Annotation 1

Image ID
c3725ee237b2cb56_m07jdr_7dcd8d5e.jpg

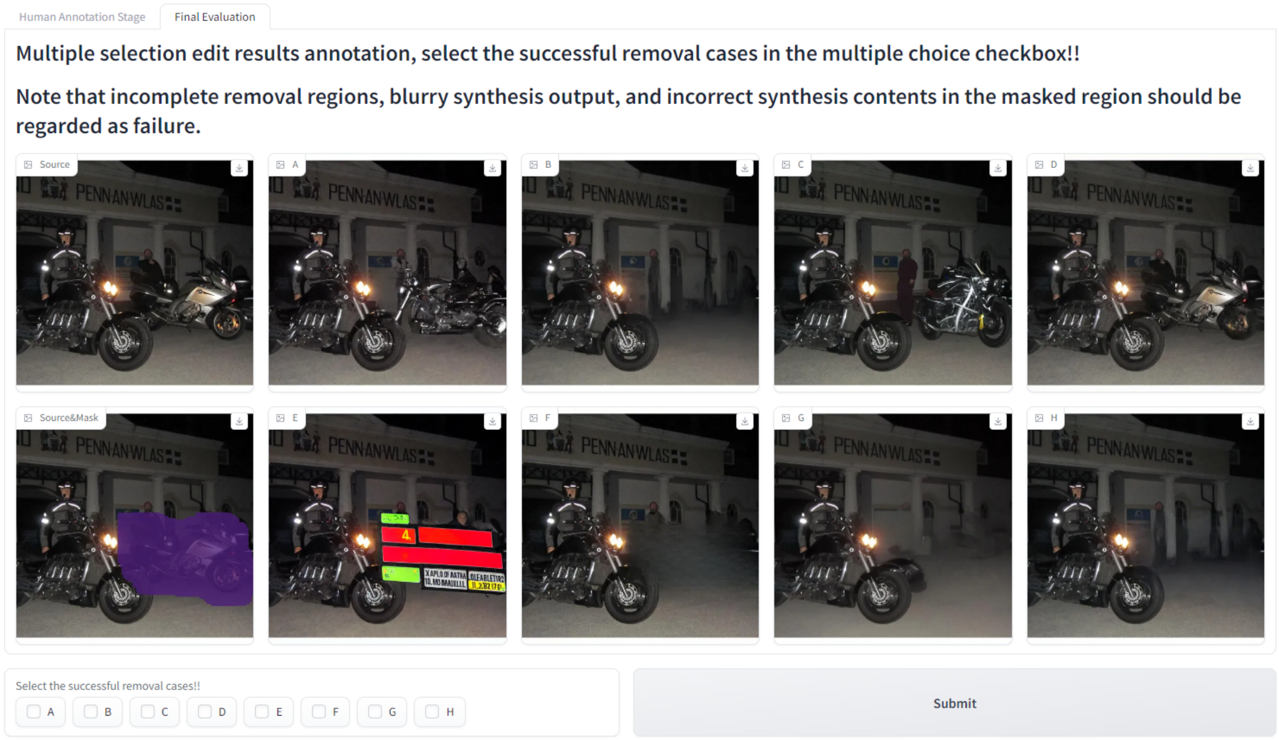
Left instances
163

(0) for No, (1) for Yes!!

Successful Rate
The successful rate of your selection
78.92

Figure 7. The interface for human annotation. The annotators are asked to give feedback by clicking ”Yes” or ”No” button.

The interface for final human evaluation is shown in Fig. 8. The input images and the masked images are displayed in the left column. The editing results of different methods as displayed in the right columns. Annotators are asked to click the multiple-choice check-boxes to select the successful removal results among different methods and submit the results. We randomly shuffle the display order in each evaluation. Five volunteers participated in the final evaluation, and each volunteer annotated 1,000 samples, including 500 pairs of object removal cases under 512×512 resolution and 500 pairs under 1024×1024 resolution. We calculate the average success rate for different methods based on these human evaluations.



Multiple selection edit results annotation, select the successful removal cases in the multiple choice checkbox!!

Note that incomplete removal regions, blurry synthesis output, and incorrect synthesis contents in the masked region should be regarded as failure.

Select the successful removal cases!!

□ A □ B □ C □ D □ E □ F □ G □ H

Figure 8. The interface for human evaluation. The volunteers make selections by checking the multiple-choice check-boxes at the bottom left corner.

B. The evaluation metrics for the discriminator

		Human Annotation Label	
		P	**N**
Discriminator Predict Label	**Y**	**T**rue **P**ositives	**F**alse **P**ositives
	N	**F**alse **N**egatives	**T**rue **N**egatives

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1} = \frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN}$$

Figure 9. Confusion matrix and the definition of metrics for evaluating our discriminator.

We use the 500 pairs in the test set with 512×512 resolution to test the discriminator. The edited results are generated by our RORem. The definitions of precision, recall, F1 and accuracy are illustrated in Fig. 9. Among these metrics, precision represents the percentage of the true positive samples to the total positive samples predicted by our discriminator. High precision ensures that the selected removal pairs are all of high-quality. By setting the threshold as 0.9, our final discriminator can reach a precision of 0.983, which allows us to obtain a large amount of high-quality data pairs.

C. Category distribution of our constructed dataset

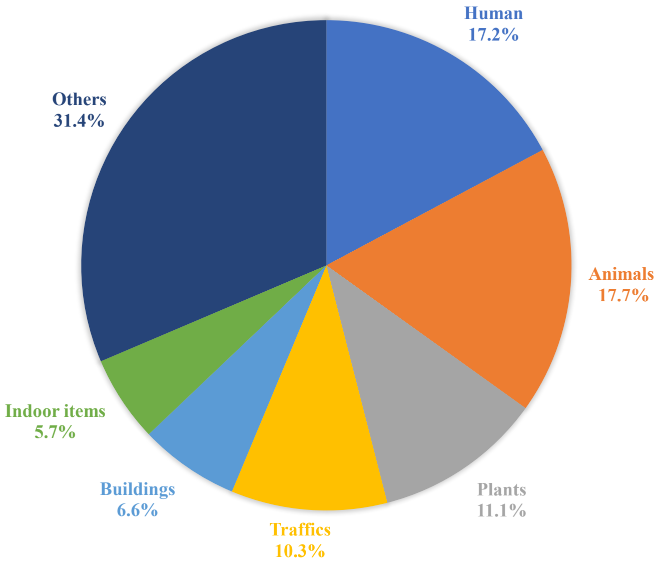


Figure 10. The category distribution of our constructed dataset.

D. Visual examples of object removal results at each training round

The visual examples of object removal results at each training round are provided in Fig. 11. We can see that the initial model SDXL-inpainting [42] always fills the masked regions with semantically similar contents instead of removing it (column Initialization). After the first round of finetuning, RORem can successfully remove the selected sofa (row 2) and cat (row 4); however, its removal capacity is not good enough, leading to failures in other cases (see partial removal cases bottle, statue and blurry synthesis case airplane). After we extend the training dataset and conduct more finetuning rounds (see column R1-Human to column R3-Auto), RORem can successfully remove the masked regions in most cases. Finally, we collect images with 2K resolution from DIV2K [2] and Flicker2K [55] to conduct the final stage finetuning, where the removal capacity of RORem can be well preserved (see column Final) and the image quality can be improved (see Tab. 1 in the main paper).

E. Visual results of IP2P and CLIPAway

The visual editing results of IP2P and CLIPAway on images of resolution of 1024×1024 are illustrated in Fig. 12. We can see that IP2P fails in all cases and even changes the overall style of the given images (see images plate in column 1 and car in column 5). CLIPAway exhibits the same problem as PPT, which often fills the masked regions with incorrect contents (see images sofa, dog and car).

F. Visual results on images of 512×512 resolution

The qualitative comparisons on images of 512×512 resolution are illustrated in Fig. 13. We can see that Lama can generate blurry synthesis outputs in some cases (see images koala in column 4 and plate in column 6). SDXL-INP, IP2P and INST fail in most cases. Moreover, as INST and IP2P are text-driven removal methods, the ambiguity of text instructions can lead to removal failures of selected objects (see images hot air ballon in column 3 and cup cake in column 6). IP2P not only fails to remove the select objects but also changes the overall style and details of the original images (see images hot air ballon, koala, and cup cake). PPT and CLIPAway can fill the masked regions with nonexistent contents in images bird (column 1), koala (column 4) and statue (column 5). DesignEdit succeeds in the first two removal cases, however it suffers from visual artifacts (see images koala, plate). In contrast, RORem successfully removes the selected objects in most cases. Meanwhile, our distilled RORem-4S model also works well in these cases with less time overhead.

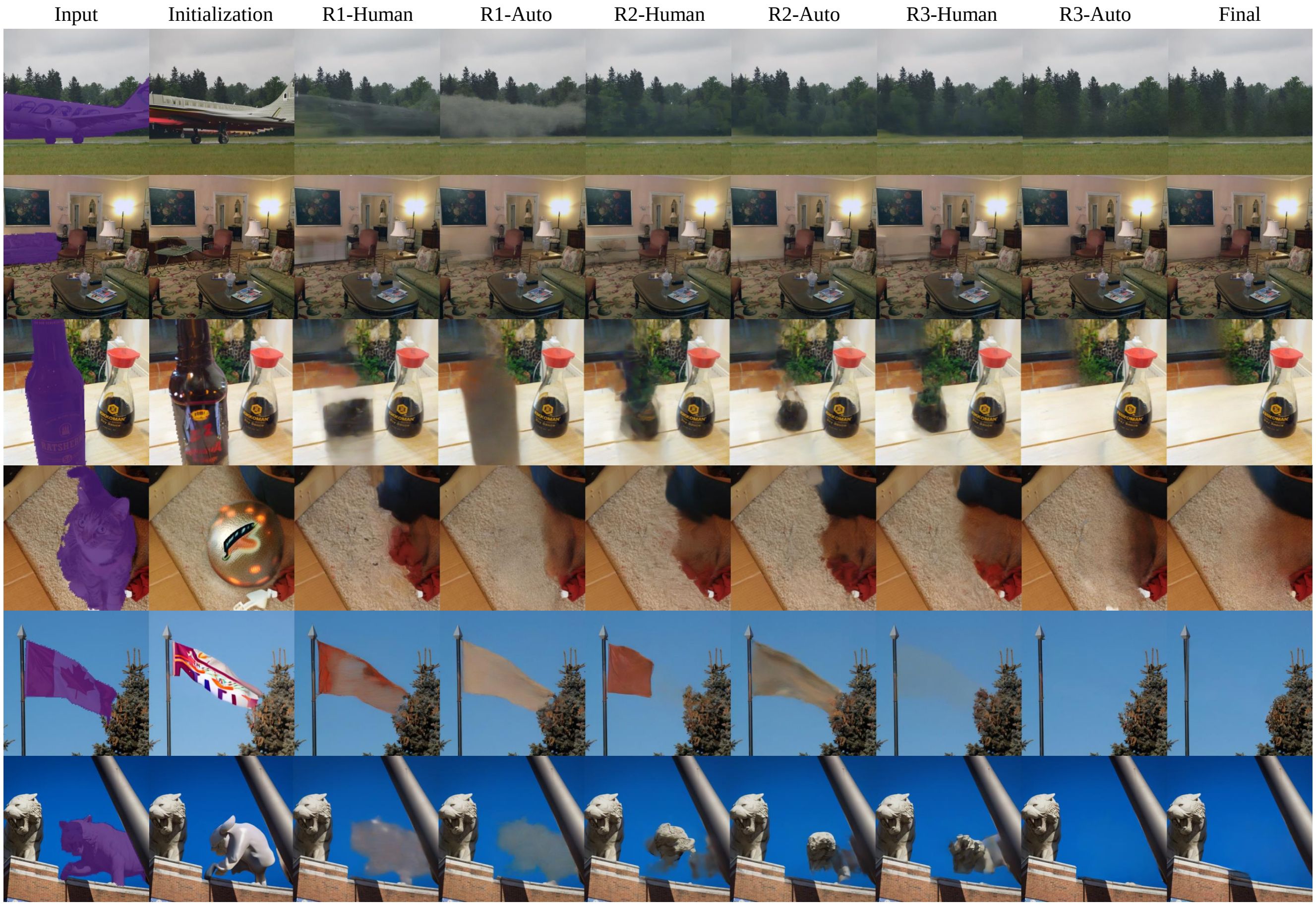


Figure 11. Visual results of RORem at each training round, one can see that the removal capacity of RORem improves with the increase of the dataset.

G. Enhance discriminator training dataset

After expanding the dataset, how to ensure the accuracy becomes crucial for reliable performance evaluation. We observe that while D_ϕ aligns well with human preferences when evaluating our RORem, it exhibits bias in assessing other methods because D_ϕ is exposed only to the failure cases of RORem during training. To make D_ϕ a good assessor for more competing methods, we expand the training data of it using several strategies, as detailed in Tab. 4. First, in addition to the human annotated 17,322 positive and 12,678 negative samples of RORem, we sample 600 examples from our training dataset and edit them using the seven competing methods. The edited results are manually annotated, leading to 785 positive and 3,415 negative samples. Second, we apply various degradation (blur, noise, dowmsample and the mixture of them) and 'no-change' to the masked regions of RORem editing outputs, generating 15,000 negative samples. Finally, we consider all the 18,859 pairs in the RORD dataset as positive samples.

H. Failure Cases

As we stated in the conclusion section of the main paper, although RORem achieves great improvement on the overall removal performance, it may fail in cases when the background contains human fingers and faces. Some failure cases are depicted in Fig. 14. Future work will be conducted for further improving the performance of RORem on these editing scenarios.

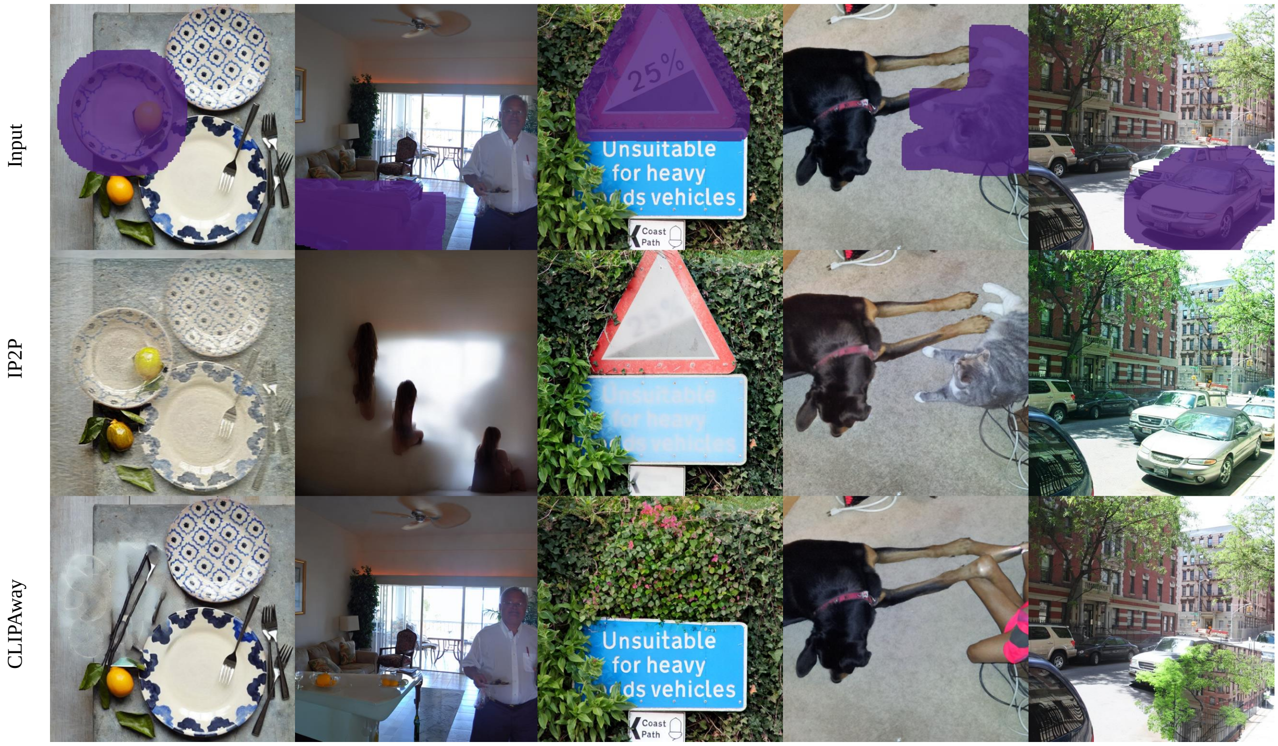


Figure 12. Visual results of IP2P and CLIPAway on 1024×1024 resolution images.

	Annotated Data		Synthesized Data					
	RORem	Baselines	Blur	Noise	Downsample	Mixed	No-change	RORD
Positive	17,322	785	0	0	0	0	0	18,859
Negative	12,678	3,415	3,000	3,000	3,000	3,000	3,000	0
Total	30,000	4,200	3,000	3,000	3,000	3,000	3,000	18,859

Table 4. The details of our constructed dataset for training the final discriminator. 'Baseline' means the seven competing methods used in the experiments. 'Mixed' refers to the combination of Blur, Noise and Downsample degradations. 'No-change' indicates the use of the source image directly as the editing result.

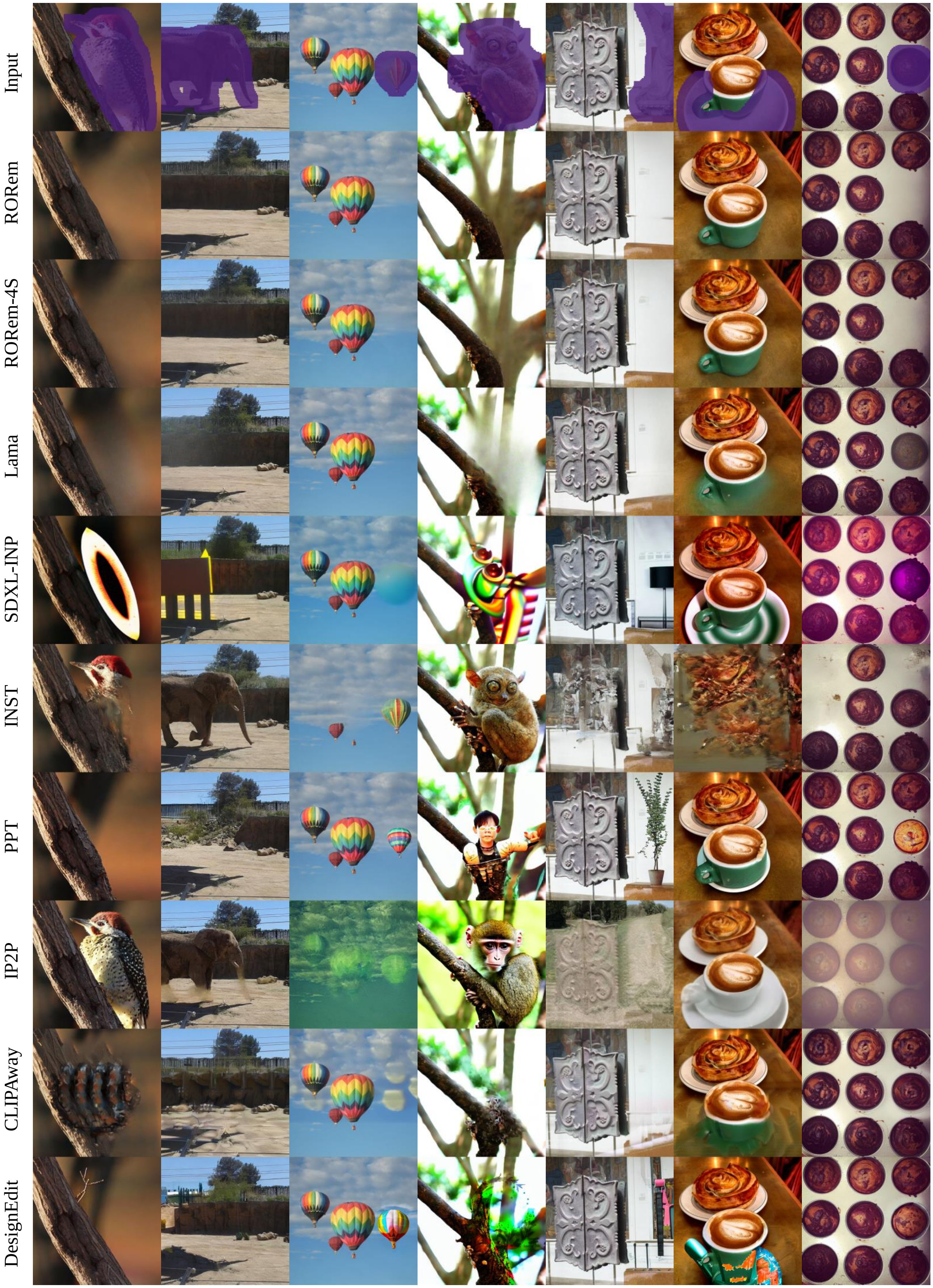


Figure 13. Visual comparison of the object removal results by RORem and other methods on 512×512 resolution images. One can see there can be incomplete removal regions, blurry synthesis output, wrong removal target, and incorrect synthese contents in previous methods, while RORem demonstrate robust removal performance.

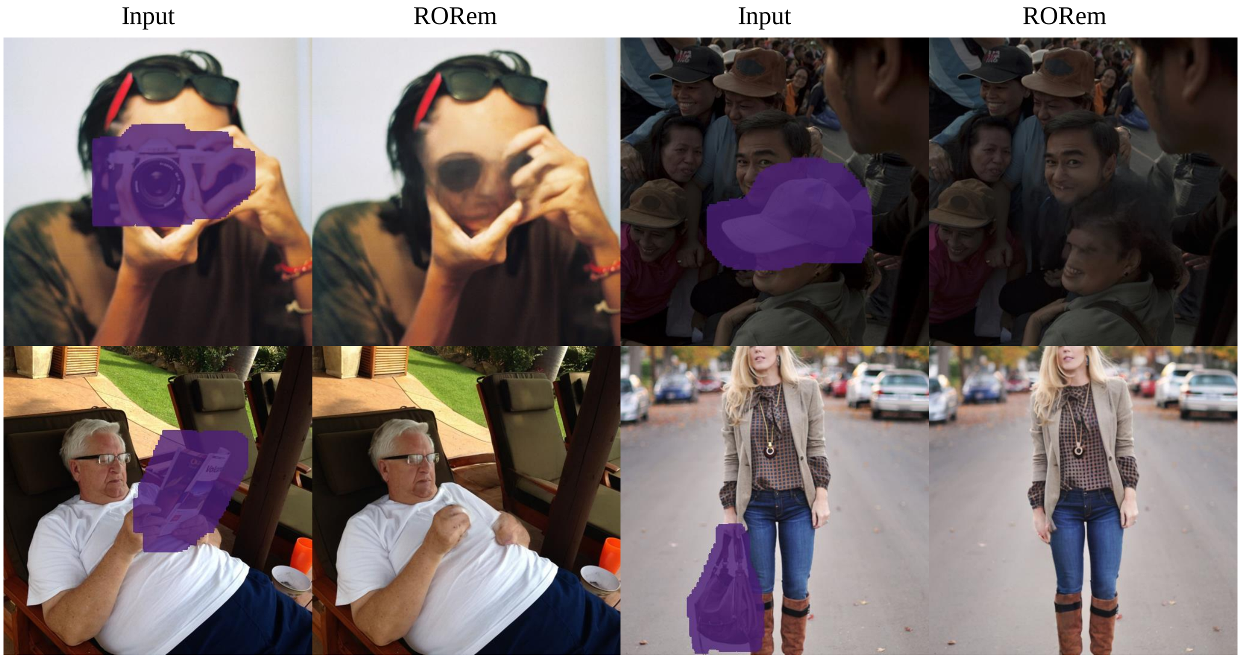


Figure 14. Failure cases of RORem.