Figure 8. Qualitative comparisons with closed-source methods.

		Structure	Background preservation				CLIP similarity	
	Model	**Distance ↓**	**PSNR ↑**	**LPIPS ↓**	**MSE ↓**	**SSIM ↑**	**Whole ↑**	**Edited ↑**
Easy	InstructPix2Pix	0.0305	21.40	0.1190	0.0129	0.7577	24.41	20.90
	MagicBrush	0.0207	24.39	0.0701	0.0076	0.8065	25.39	20.94
	HIVE	0.0287	22.23	0.1169	0.0104	0.7485	23.75	20.68
	InstructDiffusion	0.0400	22.76	0.0900	0.0200	0.7800	24.34	20.39
	HQ-Edit	0.1130	12.03	0.3418	0.0696	0.4913	20.48	18.33
	OmniEdit*	0.0190	24.80	0.0645	0.0070	0.8116	25.15	20.92
	InstructDiffusion-HA	0.0252	24.95	<u>0.0598</u>	0.0068	0.8143	24.73	20.85
	CosXLEdit	<u>0.0137</u>	**26.60**	0.0695	<u>0.0062</u>	**0.8962**	25.21	20.79
	FLUX-Omni-Edit	0.0400	20.48	0.1300	0.0200	0.7800	21.44	17.5
	UltraEdit	**0.0120**	26.23	0.0740	**0.0042**	0.8358	25.29	<u>20.96</u>
	RefEdit	0.0199	24.81	0.0599	0.0064	0.8145	<u>25.48</u>	**21.07**
	RefEdit-SD3	0.0239	<u>26.49</u>	**0.0572**	0.0069	<u>0.8902</u>	**25.79**	20.84
Hard	InstructPix2Pix	0.0435	18.87	0.1664	0.0231	0.6775	25.60	19.97
	MagicBrush	0.0274	20.56	0.1074	0.0151	0.7337	26.59	<u>20.21</u>
	HIVE	0.0367	20.01	0.1601	0.0173	0.6781	24.88	20.03
	InstructDiffusion	0.0400	18.96	0.1300	0.0300	0.7000	25.62	19.36
	HQ-Edit	0.1502	10.96	0.4127	0.0883	0.3789	20.88	17.8
	OmniEdit*	0.0248	20.80	0.1005	0.0140	0.7413	26.54	20.18
	InstructDiffusion-HA	0.0226	21.12	<u>0.0886</u>	<u>0.0128</u>	0.7495	26.36	19.60
	CosXLEdit	0.0267	21.61	0.1237	0.0240	<u>0.8241</u>	26.65	19.91
	FLUX-OmniEdit	0.0500	16.97	0.2100	0.0300	0.6700	21.02	16.05
	UltraEdit	**0.0144**	**23.64**	0.1006	**0.0067**	0.7743	**27.03**	19.82
	RefEdit	<u>0.0206</u>	21.56	**0.0868**	0.0131	0.7531	<u>26.74</u>	**20.30**
	RefEdit-SD3	0.0259	<u>22.15</u>	0.0911	0.0152	**0.8460**	26.46	19.66

Table 5. Evaluation results on RefEdit benchmark for both *Easy* and *Hard* categories. The best value is bolded and the second-best value is underlined.

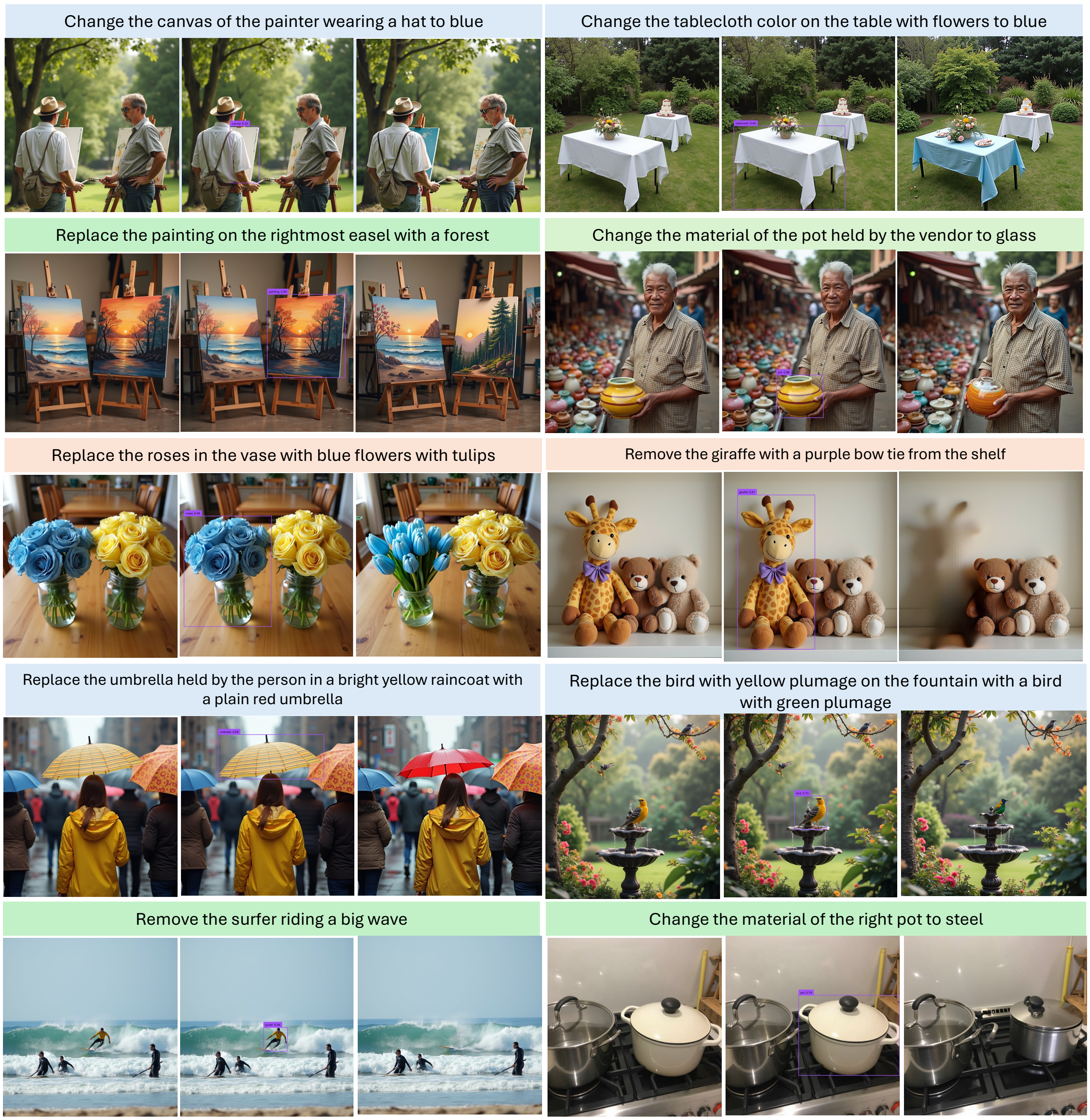

Figure 9. Additional training samples.

Dataset		Easy			Hard		
MagicBrush	**RefEdit-Data**	**SC$_{avg}$ ↑**	**PQ$_{avg}$ ↑**	**O$_{avg}$ ↑**	**SC$_{avg}$ ↑**	**PQ$_{avg}$ ↑**	**O$_{avg}$ ↑**
✓		4.18	<u>6.10</u>	3.67	<u>4.11</u>	6.16	<u>3.56</u>
	✓	<u>4.88</u>	**6.32**	<u>4.15</u>	3.53	<u>6.29</u>	3.02
✓	✓	**5.47**	5.85	**4.68**	**4.51**	**6.48**	**3.93**

Table 6. **Ablation study on impact of data on RefEdit-Bench.** Modified VIEScore evaluation results on RefEdit benchmark for both *Easy* and *Hard*. Best is bold, second best underlined. **O$_{avg}$** is overall VIEScore. GPT-4o is the MLLM. We can observe that the InstructPix2Pix model fine-tuned on only MagicBrush data performs poorly on our benchmark. When trained on RefEdit data alone, it improves the performance in the Easy category. However, the maximum improvements come when the model is fine-tuned on both datasets together.

Data	SC$_{avg}$ ↑	PQ$_{avg}$ ↑	O$_{avg}$ ↑
MagicBrush	**7.42**	<u>4.61</u>	<u>5.56</u>
RefEdit-Data	<u>6.87</u>	**7.30**	**6.31**

Table 7. **Ablation study on dataset quality.** VIEScore evaluation results for training data. We observe that our synthetic RefEdit-Data is of the highest quality, as high as MagicBrush, which is a human-annotated dataset.

Figure 10. Qualitative results on image editing. The top 4 samples are from the *Easy* category and the bottom 3 samples are from the *Hard* category. As illustrated, our method attains the SOTA performance on comparison of all the methods.

You are a professional digital artist. You will have to evaluate the effectiveness of the AI-edited image(s) based on the given rules.
You will have to give your output in this way (Keep your reasoning concise and short.):
{
"score" : [...],
"reasoning" : "..."
}
and don't output anything else.

Two images will be provided: The first being the original image selected from COCO dataset and the second being an AI edited version of the first. The objective is to evaluate how successfully the editing instruction has been executed in the second image. Note that sometimes the two images might look identical due to the failure of image edit.
Both the original image and the edited image are masked images since the image contains multiple objects and we want you to only focus on the intended object.

From a scale 0 to 10:
A score from 0 to 10 will be given based on the success of the editing.
- 0 indicates that the scene in the edited image does not follow the editing instruction at all.
- 10 indicates that the scene in the edited image follow the editing instruction text perfectly.

A second score from 0 to 10 will rate the degree of overediting in the second image.
- 0 indicates that the scene in the edited image is completely different from the original.
- 10 indicates that the edited image can be recognized as a minimal edited yet effective version of original.

Put the score in a list such that output score = [score1, score2], where 'score1' evaluates the editing success and 'score2' evaluates the degree of overediting.

Editing instruction:

Figure 11. SC score prompt masked version.

You are a professional digital artist. You will have to evaluate the effectiveness of the AI-edited image.
Two images will be provided: The first being the original image selected from COCO dataset and the second being an AI edited version of the first. So you may not worry about privacy or confidentiality.

You must focus solely on the technical quality and artifacts in the edited image, and **do not consider whether the context is natural or not**.

Your evaluation should focus on:
- Distortions
- Unusual body parts or proportions
- Unnatural Object Shapes

Rate the edited image on a scale from 0 to 10, where:
- 0 indicates significant AI-artifacts.
- 10 indicates an artifact-free image.

You will have to give your output in this way (Keep your reasoning concise and short.):
{
"score": ...,
"reasoning": "..."
}

and don't output anything else.

Figure 12. PQ score version.