InstantEdit: Text-Guided Few-Step Image Editing with Piecewise Rectified Flow

Supplementary Material

A. Hyperparameters and Design Choice

Here we present our hyperparamater selection for the main results for reproducibility. For ControlNet, the ControlNet conditioning scale is set to 0.4. In the inversion process, we do not use classifier-free guidance (CFG); while in regeneration, there are two vital parameters, the DPG guidance scale, which is set to 2.5, and the attention mask threshold, which is 0.4. For the attention mask implementation, we aggregate the cross attention maps with dimension 16×16 and extrapolate to the dimension of the latent, then we perform the thresholding operation.

B. Consistency-Editability Tradeoff

Consistency-editability tradeoff is a commonly recognized property in the setting of image editing as discussed by previous work such as ReNoise and InfEdit and we demonstrate it in Tab. 1, Tab. 2, Tab. 3, Tab. 4. This also affects how we choose our hyperparameters for comparisons. For example, one method can adjust the hyperparameter (e.g. guidance scale) to achieve higher editability above other methods with a sacrifice on its consistency metrics, which makes the result hard to interpret. Therefore, we propose to adjust the hyperparameters to align on one type of metric and compare the overall performance across all other metrics. In Tabel 1 from main paper, we adjust the classifier-free-guidance (CFG) scale for ReNoise, InfEdit; Pseudo-Guidance (PG) scale for TurboEdit; and Disentangled Prompt Guidance (DPG) for InstantEdit. In Table 4 from main paper, we only adjust DPG scale except for the ablation on PG. For all the experiments, we roughly align on the editability metrics (Alignment). We show that our method produces more consistent results with similar and even better editability. In other words, our method can raise the consistency with lower cost on the editability and vice versa.

C. Editing Results with Different NFE

We report the quantitative results of InstantEdit with different NFE in Tab. 5, which shows that the performance of our method scales with increasing NFE. For all settings, we use the same set of hyperparameters as our main experiment. We observe that the consistency metrics improve as the sampling step increases. Another phenomenon we discover is that the alignment metrics do not show a clear trend with the increasing NFE, which is also observed in InfEdit Table 2.

D. Further Ablation Results

We provide more detailed ablation results for the hyperparameters controlling the Controlnet scale and attention mask threshold as in Tab. 7. The consistency metrics improve with larger ControlNet scale and mask threshold, while the editability metrics experience the opposite. This is expected because larger ControlNet scale and mask threshold tend to maintain contents from the original image thus improving the edit consistency.

E. Other Baseline Methods with Mask

We notice that the attention masking mechanism provides a relatively large improvement in the quantitative metric. To ensure that our method's superior performance is not solely attributed to the attention masking mechanism, we extend a similar masking strategy to the baseline methods for a fair comparison. InfEdit already includes a similar masking operation, so we keep it intact. We refer reader to Section 4.1 of Infedit for more details. We present our quantitative results in Tab. 6. We discover that TurboEdit does not benefit from the masking strategy. This might due to the DDPM-noise inversion, as the nondeterministic DDPM noise injects artifacts during the merging process of the source and target guidance signals. ReNoise benefits from the mask as consistency metrics improve. However, the editability metrics drop accordingly. Overall, ReNoise still cannot reach a competitive performance with attention masking.

F. User Study

Fig. 1 shows the 15 selected images for user study and Fig. 2 shows the interface of our user study. We observe that TurboEdit sometimes faces the problem of large structural inconsistency as shown in the case 3 and 5. InfEdit also tends to create noticeable artifacts in 4 steps as shown in case 2, 3, 11. Most of the methods are not able to perform successful editing when large structural change is required like case 7, 8, 9.

G. Additional Visual Results

In this section, we show additional comparison results with other baseline few-step editing methods, as shown in Fig 3a. In Fig 3b, we provide extra visualization results for diverse text-based editing with our method. On the same image, our method can perform various types of editing and can edit on different objects with just 4 steps.

	Distance$^{\downarrow}_{10^3}$	PSNR↑	LPIPS$^{\downarrow}_{10^3}$	MSE$^{\downarrow}_{10^4}$	SSIM$^{\uparrow}_{10^2}$	Whole↑	Edited↑
DPG: 2.0	15.71	28.23	42.40	32.53	86.64	26.16	22.76
DPG: 2.5 (Default)	17.14	27.96	44.39	34.94	86.44	26.28	22.82
DPG: 3.0	18.71	27.70	46.39	37.39	86.22	26.33	22.87

Table 1. Quantitative result demonstrating the effect of DPG guidance scale for InstantEdit.

	Distance$^{\downarrow}_{10^3}$	PSNR↑	LPIPS$^{\downarrow}_{10^3}$	MSE$^{\downarrow}_{10^4}$	SSIM$^{\uparrow}_{10^2}$	Whole↑	Edited↑
CFG: 1.8	11.85	27.78	41.55	32.48	85.82	25.31	21.88
CFG: 2.3 (Default)	16.19	26.75	50.79	42.33	84.71	25.68	22.77
CFG: 2.8	28.56	24.63	73.87	71.45	82.10	26.23	22.69

Table 2. Quantitative result demonstrating the effect of CFG guidance scale for InfEdit.

	Distance$^{\downarrow}_{10^3}$	PSNR↑	LPIPS$^{\downarrow}_{10^3}$	MSE$^{\downarrow}_{10^4}$	SSIM$^{\uparrow}_{10^2}$	Whole↑	Edited↑
PG: 0.8	14.11	25.73	66.40	44.62	83.81	25.06	21.66
PG: 1.3 (Default)	18.57	24.59	77.53	58.48	82.64	25.70	22.30
PG: 1.8	35.87	21.47	117.25	117.22	78.66	26.82	23.31

Table 3. Quantitative result demonstrating the effect of PG guidance scale for TurboEdit.

	Distance$^{\downarrow}_{10^3}$	PSNR↑	LPIPS$^{\downarrow}_{10^3}$	MSE$^{\downarrow}_{10^4}$	SSIM$^{\uparrow}_{10^2}$	Whole↑	Edited↑
CFG: 5.8	19.27	25.14	80.09	50.10	82.41	25.25	21.68
CFG: 6.3 (Default)	20.31	24.89	83.26	52.9	82.12	25.26	21.68
CFG: 6.8	21.68	24.54	87.95	57.16	81.69	25.25	21.68

Table 4. Quantitative result demonstrating the effect of CFG guidance scale for ReNoise.

	Distance$^{\downarrow}_{10^3}$	PSNR↑	LPIPS$^{\downarrow}_{10^3}$	MSE$^{\downarrow}_{10^4}$	SSIM$^{\uparrow}_{10^2}$	Whole↑	Edited↑
8 NFE	17.14	27.96	44.39	34.94	86.44	26.28	22.82
16 NFE	13.64	29.15	36.44	26.76	87.24	26.01	22.64
24 NFE	12.57	29.63	35.27	24.57	87.40	26.06	22.73
32 NFE	12.16	29.69	34.95	24.36	87.49	25.96	22.69

Table 5. Quantitative results with 8, 16, 24, 32 NFE .

	Distance$^{\downarrow}_{10^3}$	PSNR↑	LPIPS$^{\downarrow}_{10^3}$	MSE$^{\downarrow}_{10^4}$	SSIM$^{\uparrow}_{10^2}$	Whole↑	Edited↑
TurboEdit	25.64	24.65	109.41	54.10	80.07	25.23	21.78
ReNoise	26.07	25.76	65.78	53.45	83.68	24.75	21.28
InfEdit	16.19	26.75	50.79	42.33	84.71	25.68	22.27
InstantEdit (Ours)	17.14	27.96	44.39	34.94	86.44	26.28	22.82

Table 6. Quantitative comparison for applying attention mask to all the methods.

Components	**Distance**$_{10^3}^{\downarrow}$	**PSNR**$^{\uparrow}$	**LPIPS**$_{10^3}^{\downarrow}$	**MSE**$_{10^4}^{\downarrow}$	**SSIM**$_{10^2}^{\uparrow}$	**Whole**$^{\uparrow}$	**Edited**$^{\uparrow}$
ControlNet scale=0.2	18.90	27.66	47.47	38.26	86.12	26.16	22.72
ControlNet scale=0.6	10.37	29.67	32.69	22.31	87.65	25.25	21.89
ControlNet scale=0.8	8.17	30.17	29.56	19.64	88.00	24.82	21.52
Mask threshold=0.2	14.39	28.47	40.80	29.37	86.82	25.89	22.44
Mask threshold=0.6	12.88	29.20	36.08	26.75	87.24	25.61	22.22
Mask threshold=0.8	10.00	30.29	32.03	23.27	87.68	24.99	21.56

Table 7. Quantitative results for further ablation experiments

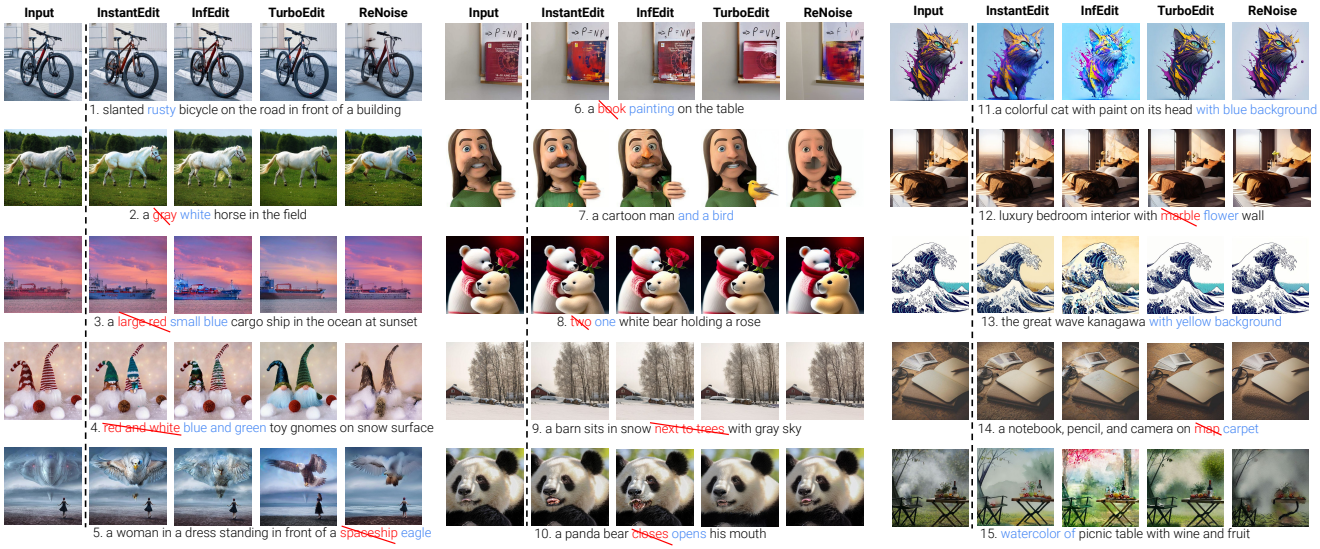


Figure 1. Visualization for randomly selected user study images.

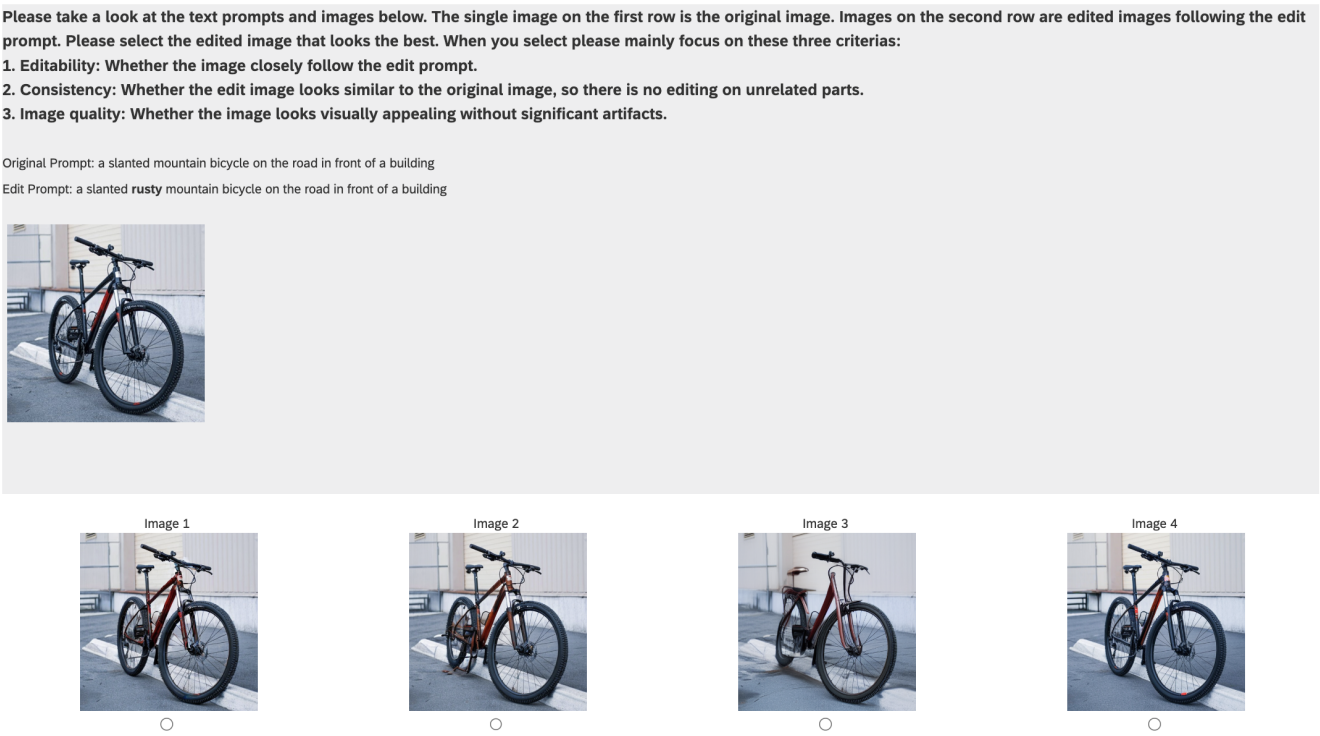
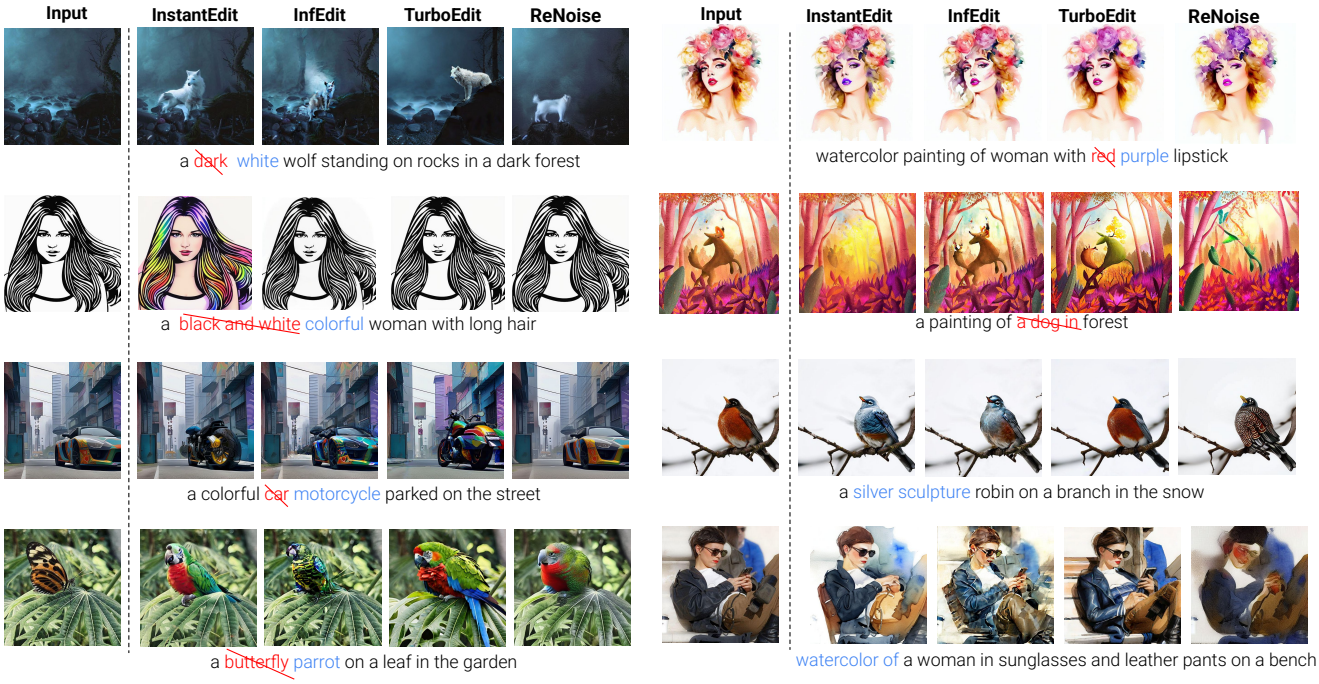
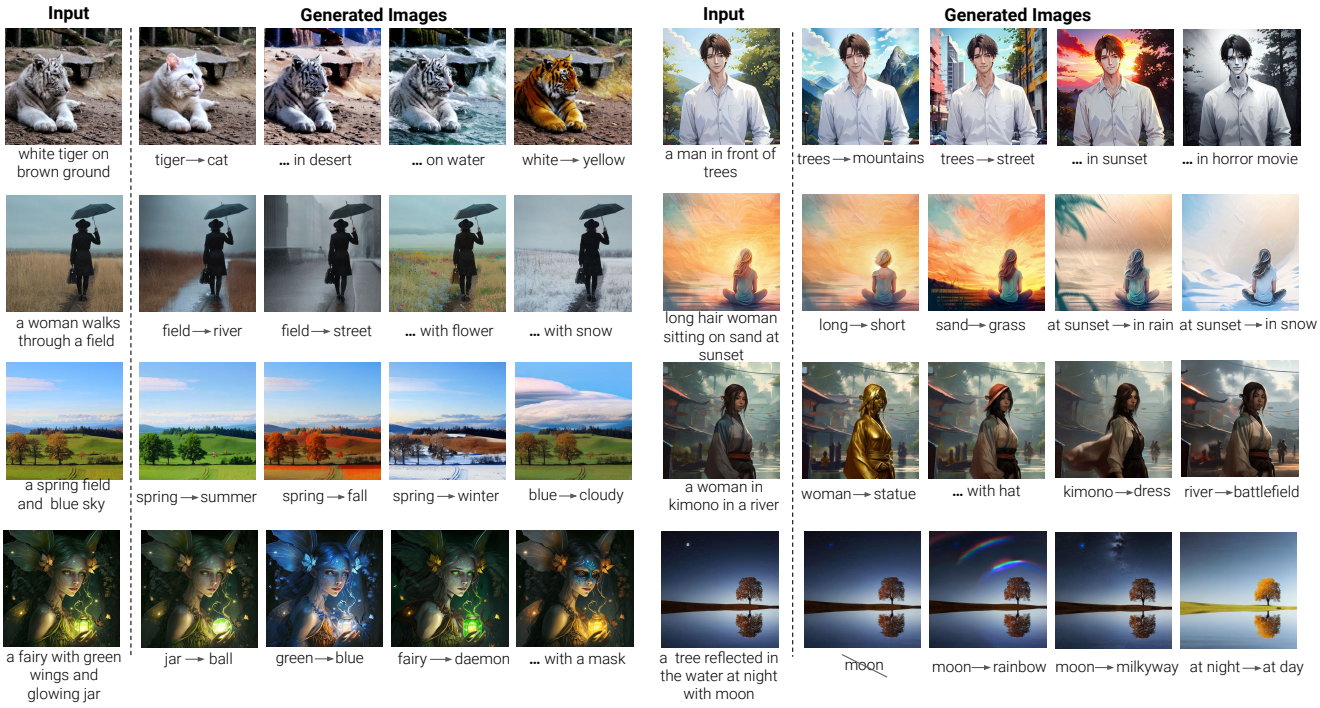


Figure 2. Visualization for our user study interface. We provide general instructions for users to follow when making their decisions. Each method is anonymous and randomly shuffled for users.



(a) Additional visual comparison with other few-step editing methods.



(b) Additional editing results with diverse text prompts.