FeedEdit: Text-Based Image Editing with Dynamic Feedback Regulation

Supplementary Material

8. Statistical and Visualization on Editing Correlation

To explore the specific correlation between feature differences and editing degree, we first conduct statistical analysis on the feature differences between the source and edited images, in the ideal editing datasets IP2P [4] and Magicbrush [44]. Specifically, from each dataset, we selected 100 source-edited image pairs for each editing function (including appearance replacement, appearance attribute editing, and layout editing), and calculated the feature differences during the denoising generation process at each step (total of 30 steps) between the source and edited images, resulting in a total of 18000 numerical statistics. Fig.9 shows our statistical results, which illustrate that the ideal feature differences are mainly within a proper range, that is, $[0.125, 0.275]$ for appearance replacement, $[0.075, 0.225]$ for appearance attribute editing, and $[0.3, 0.4]$ for layout editing.

Based on this conclusion, we further generate editing results with different feature differences by adjusting the text guidance intensity, to explore the editing performance of results within or not within the proper range. The visualization cases shown in Fig.10 indicate that, the edited samples with feature differences below the range tend to be under-editing, while those above the range tend to be over-editing.

These conclusions fully demonstrate the correlation between feature differences and editing degree. By judging whether feature differences are within the ideal editing range, we can measure the editing status, whether under-editing or over-editing.

9. Datasets

To ensure the diversity of scenes and objects of the dataset, we further select 30 images from a wide range of domains, i.e., free-to-use high-resolution images from Unsplash (https://unsplash.com/). The image categories include animal, human, architecture, furniture, plants, and natural scenery. All of the collected images are listed in Fig.11 to support further research.

10. Results on different editing engines

We further adopt PnP [36] and MasaCtrl [5] as the editing engine, to explore the generalization and application capabilities of FeedEdit. As shown in Tab.3, the utilization of FeedEdit significantly improves the model performance, especially in editability (CLIP-T) and global quality (FID and ImageReward). These results further demonstrate that our feedback editing framework brings better editing performance by providing more accurate text guidance intensity.

	Method	*editability* CLIP-T↑	*fidelity* CLIP-I↑	*fidelity* LPIPS↓	*quality* FID↓	*quality* IR↑
Single-function	P2P	0.2416	0.8247	0.2092	93.61	0.2275
	FeedEdit+P2P	**0.3024**	**0.8565**	**0.1311**	**65.02**	**0.8754**
	PnP	0.2461	0.8125	0.2475	89.10	0.3062
	FeedEdit+PnP	**0.2857**	**0.8496**	**0.1482**	**68.25**	**0.7531**
	MasaCtrl	0.2253	0. 8364	0.2362	114.73	0.1582
	FeedEdit+MasaCtrl	**0.2861**	**0.8402**	**0.1735**	**70.26**	**0.6940**
Multi-function	P2P	0.1827	0.7914	0.2452	106.53	0.0726
	FeedEdit+P2P	**0.2607**	**0.8171**	**0.1739**	**81.70**	**0.6507**
	PnP	0.1845	0.7903	0.2962	102.60	0.1427
	FeedEdit+PnP	**0.2497**	**0.8065**	**0.1904**	**83.28**	**0.5772**
	MasaCtrl	0.1635	0.7584	0.3282	135.20	-1.272
	FeedEdit+MasaCtrl	**0.2314**	**0.8021**	**0.2053**	**86.38**	**0.4270**

Table 3. **Quantitative results on different editing engines.** The proposed FeedEdit brings improvements for all editing engines. P2P [10] is the editing engine adopted for the main text.

11. More Results

11.1. More Comparisons

More additional comparison results are provided for 5 different editing scenarios: replacement (Fig.12), stylization (Fig.13), appearance manipulation (Fig.14), posture manipulation (Fig.15), and multi-function combined editing (Fig.16), demonstrating our superiority over different functional editing texts.

11.2. More Comparisons with Multi-turn Editing

We further compare our FeedEdit with existing baselines, which perform multi-turn editing to achieve multi-function editing. As shown in Fig.17, our method still demonstrates better performance, as existing methods either fail to achieve all edits due to inherent insufficient editability (*e.g.*, all baselines fail to make "zebra" "jump", or make "elephant" "sit"), or lead to excessive modification of unedited image regions due to the errors' accumulation caused by multi-turn editing (*e.g.*, in the third case, all baselines make unwanted modifications to the "bag"). In addition, another disadvantage of multi-turn editing is the increased time cost, which increases geometrically with the increase of editing turns.

12. More Ablation Studies

More qualitative results of the ablation studies are shown in Fig.18, which also prove the effectiveness and rationality of the designed perceive-feedback-regulate framework. The ablation results on feedback parameter

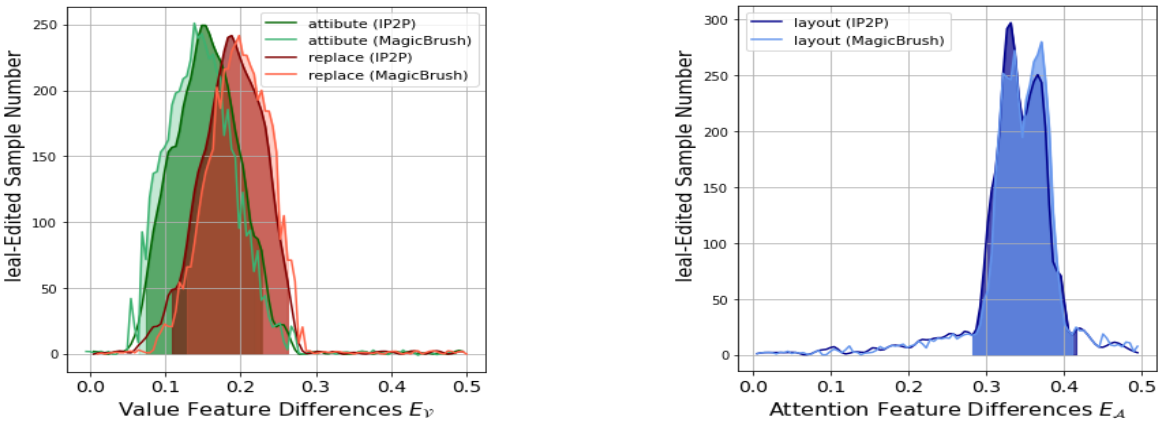
	Single-function		Multi-function	
	Text-Align	**Image-Align**	**Text-Align**	**Image-Align**
DAC	14.1%	12.5%	1.1%	2.6%
InfEdit	21.9%	22.7%	4.6%	8.4%
SDPInV	24.6%	28.3%	10.1%	12.3%
Ours	**39.4%**	**36.5%**	**84.2%**	**76.7%**

Table 4. **Human Evaluation**. FeedEdit outperforms baselines in all aspects, and achieves huge advantages in multi-function edit.

$\{K_p^j, K_i^j\}_{j\in\{noun,adj,verb\}}$ shown in Fig.19&20, also validate the robustness of our method, and the rationality of the predicted initial settings $\{K_p^j = 2, K_i^j = 0.01\}$.

13. Human Evaluation

For single-function and multi-function editing, we conduct a human preference study to compare our FeedEdit with three SOTA baselines. 10 participants are required to evaluate these 100 cases and 400 generated images from two aspects: (1) text-alignment: measuring the consistency with edited text; (2) image-alignment: measuring the consistency with the text-irrelevant image details. As indicated in Tab.4, our method outperforms existing methods, especially for multi-function editing, with a preference rate of more than 75% across all baselines.



(a) Appearance editing (b) Layout editing

Figure 9. **Statistical results** of feature differences between the source and edited images, in ideal editing datasets IP2P [4] and Magicbrush [44]. **(a)** Statistics of value differences for appearance editing (including replacement and attribute editing). **(b)** Statistics of attention differences for layout editing. The results show that the ideal differences are mainly within a proper range, that is, $[0.125, 0.275]$ for appearance replacement, $[0.075, 0.225]$ for appearance attribute editing, and $[0.3, 0.4]$ for layout editing.

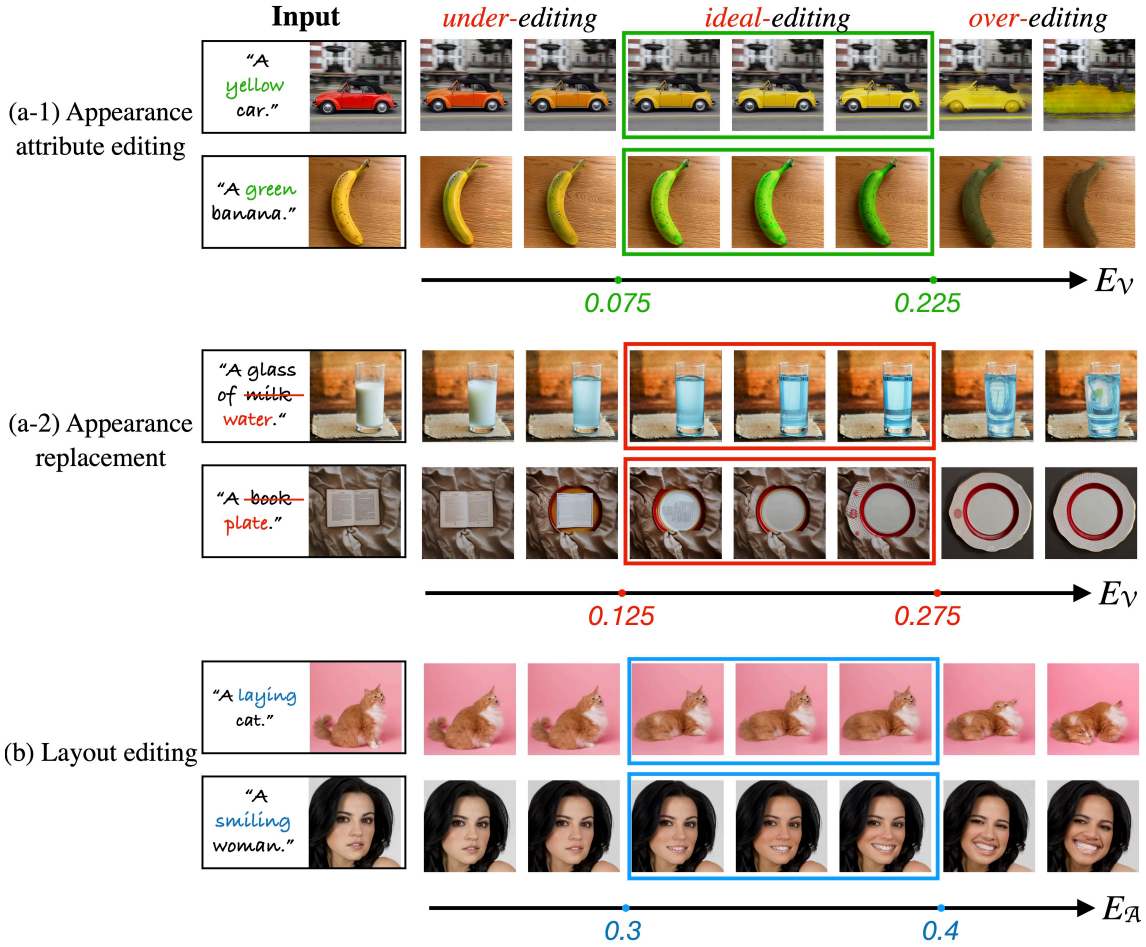


Figure 10. **Visualization cases** with different feature differences, showing that ideal differences are mainly within a proper range, while below or beyond the ideal range leads to under-editing or over-editing.

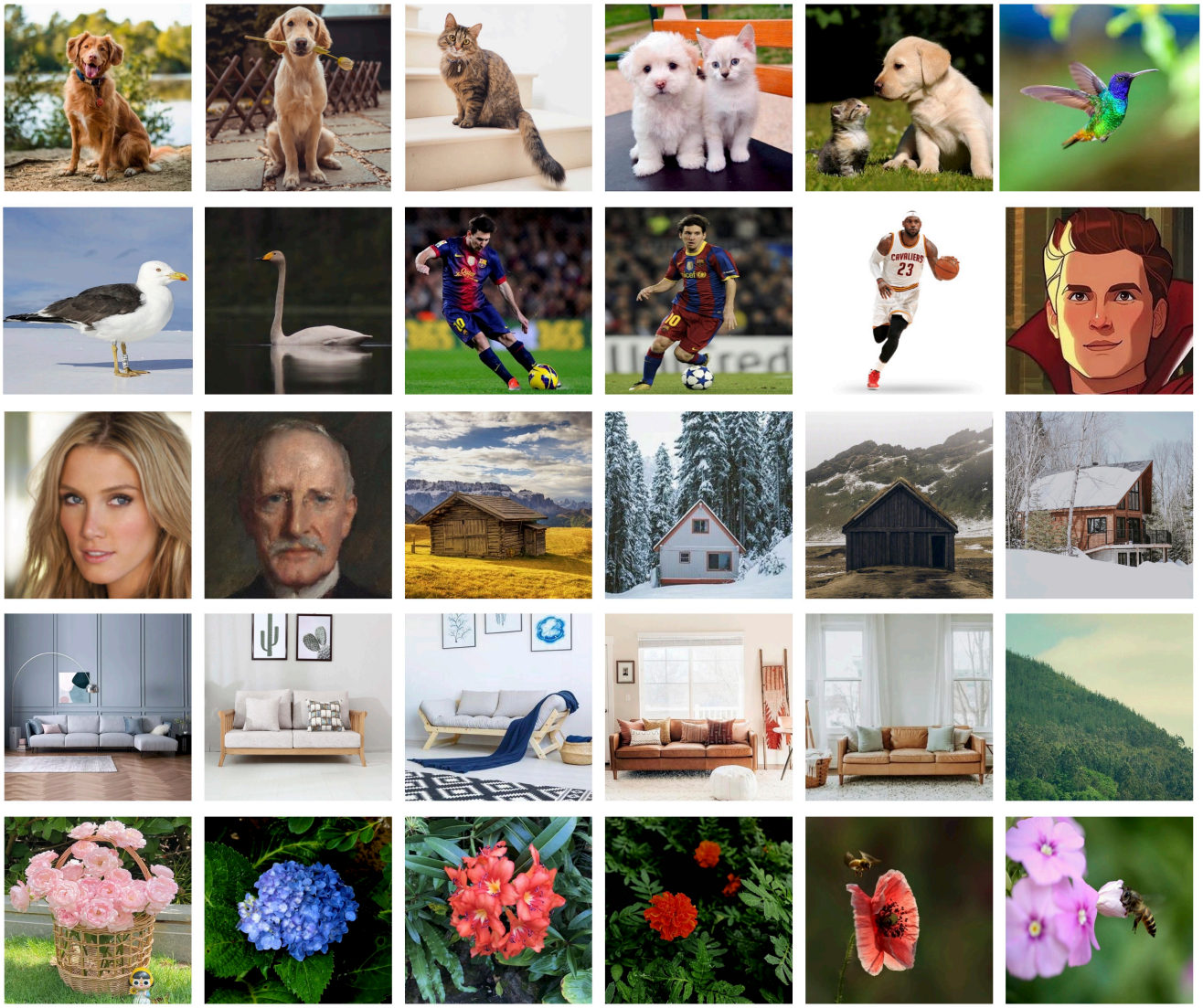


Figure 11. **The collected dataset,** which covers a wide range of domains of images, including animal, human, architecture, furniture, plants, and natural scenery.

Figure 12. **Additional comparisons on the appearance replacement.**

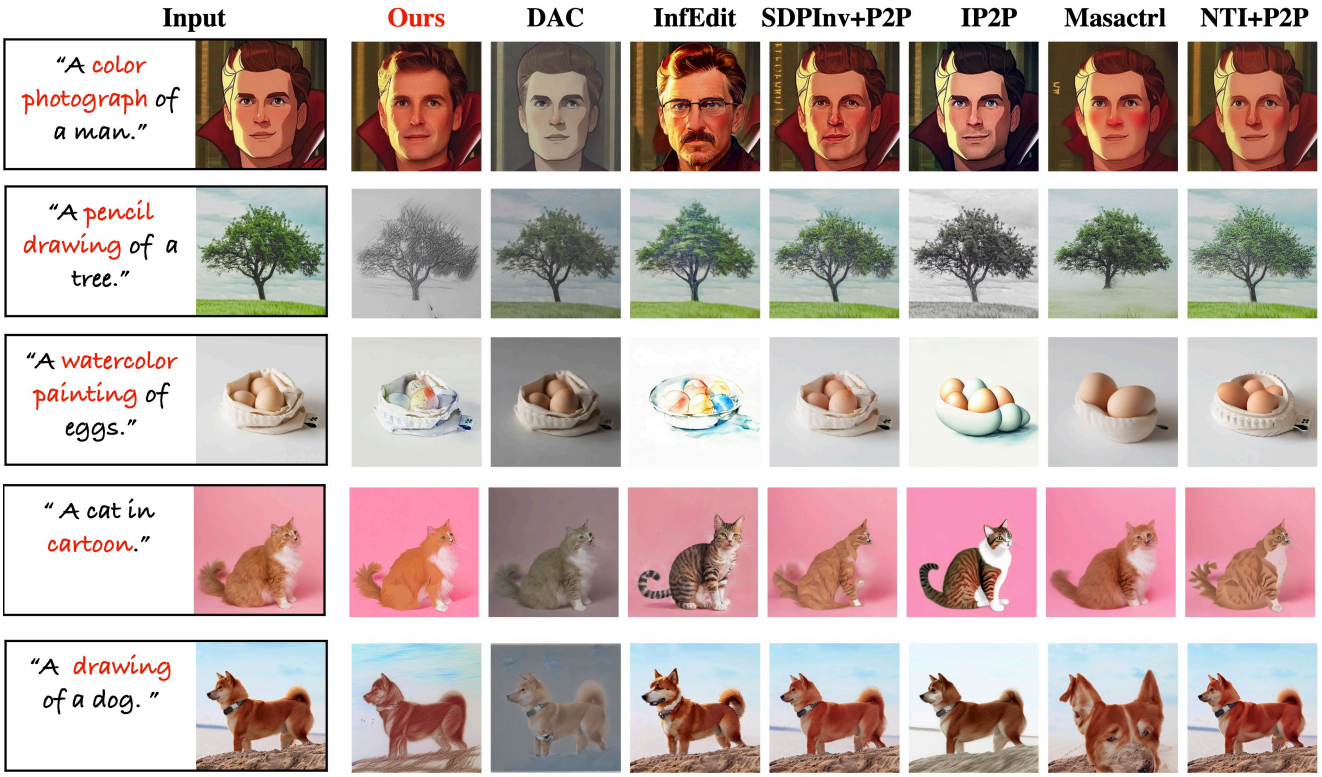


Figure 13. **Additional comparisons on the appearance stylization.**

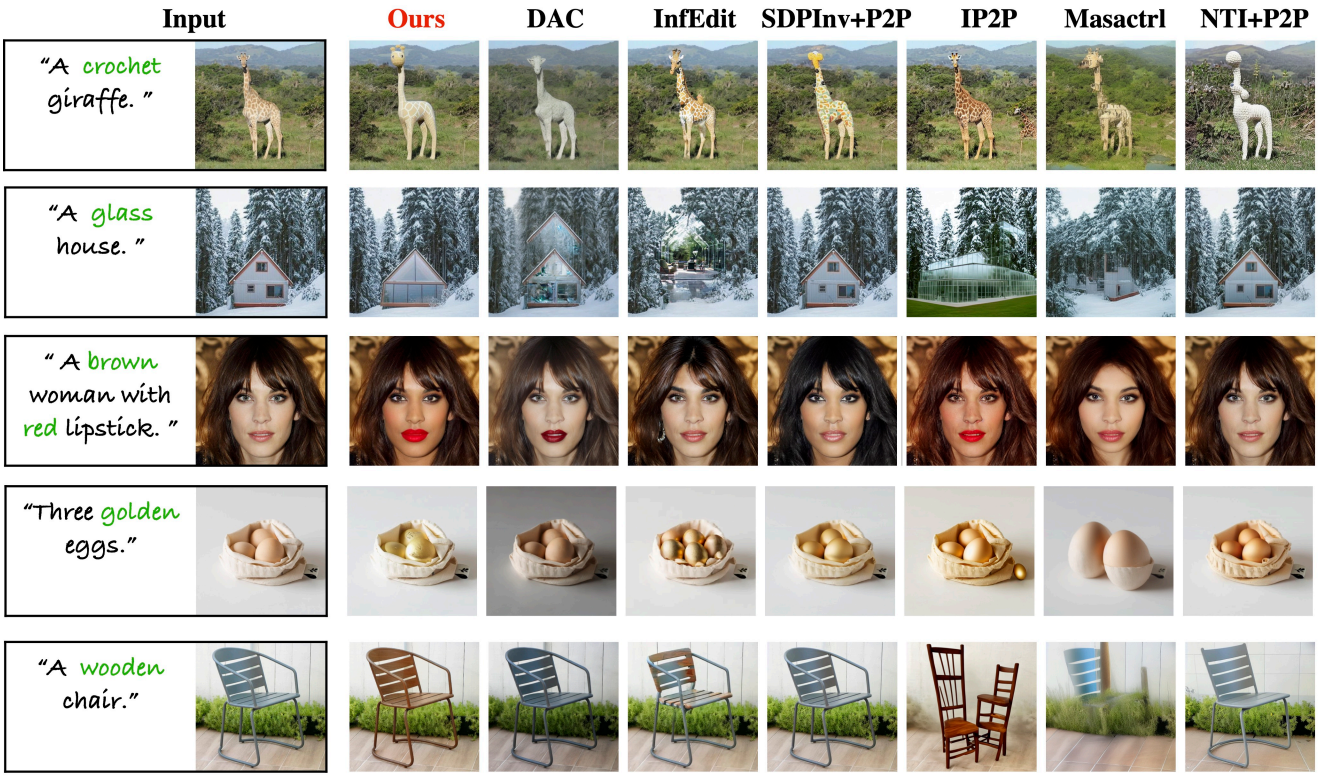


Figure 14. **Additional comparisons on the appearance attribute manipulation.**

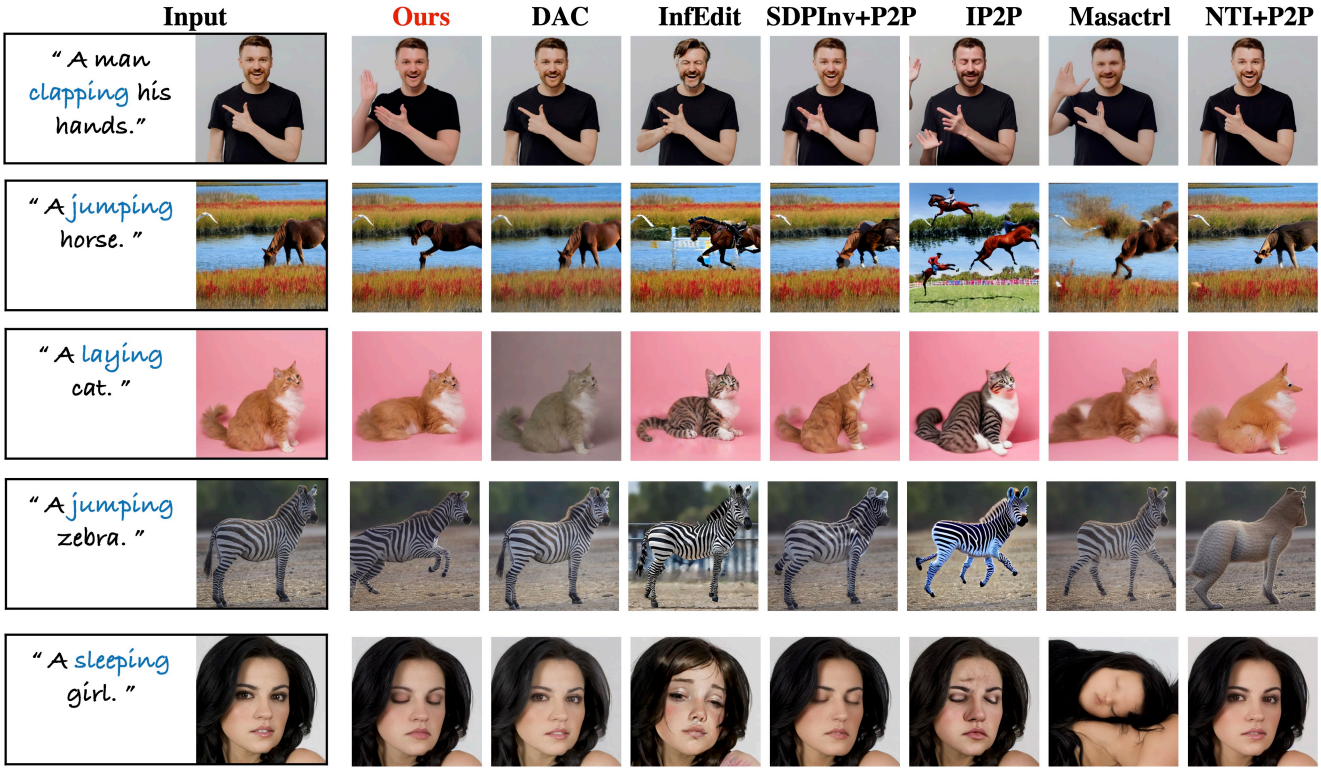


Figure 15. **Additional comparisons on the posture manipulation.**

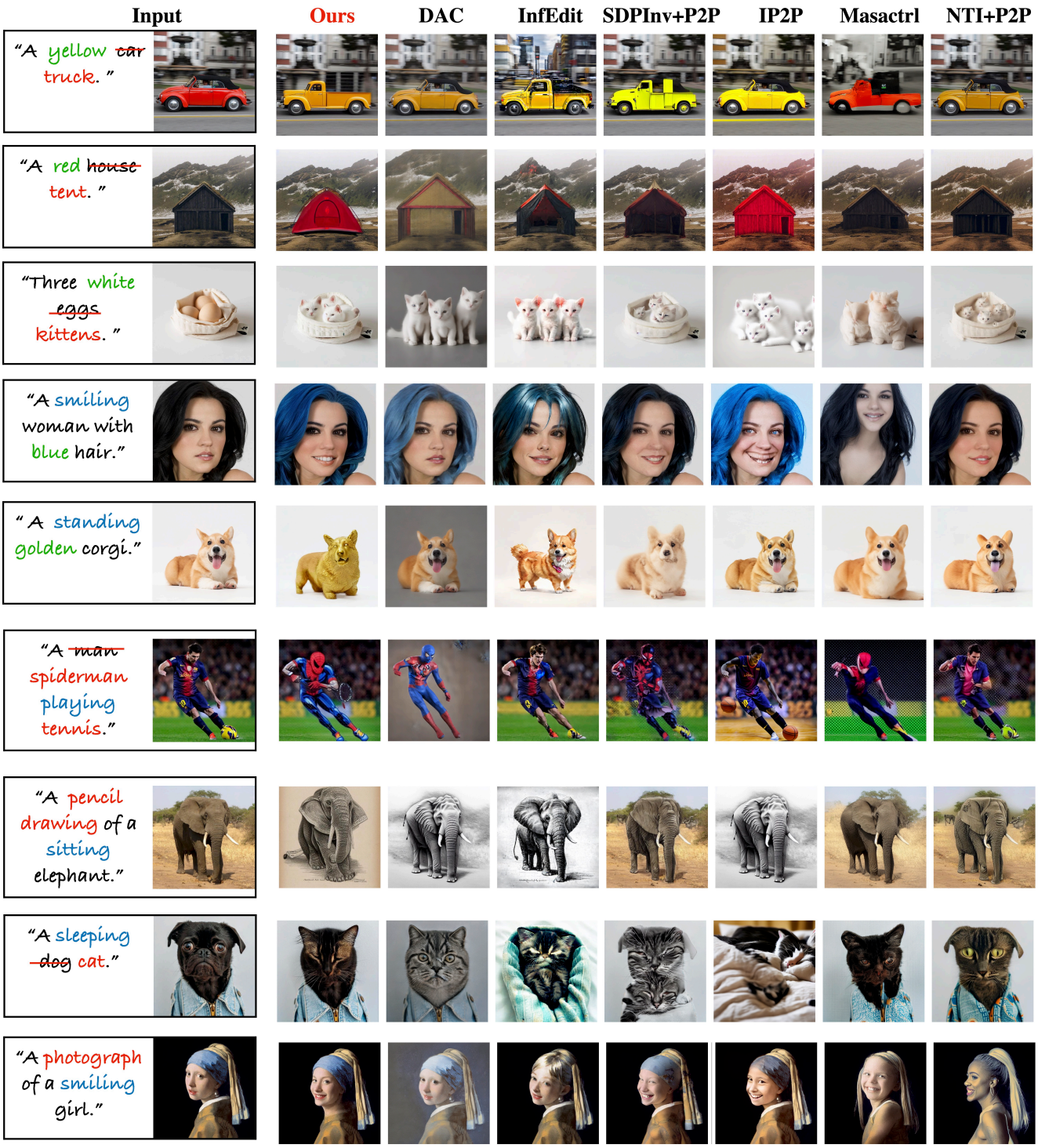


Figure 16. **Additional comparisons on the multi-function editing.**

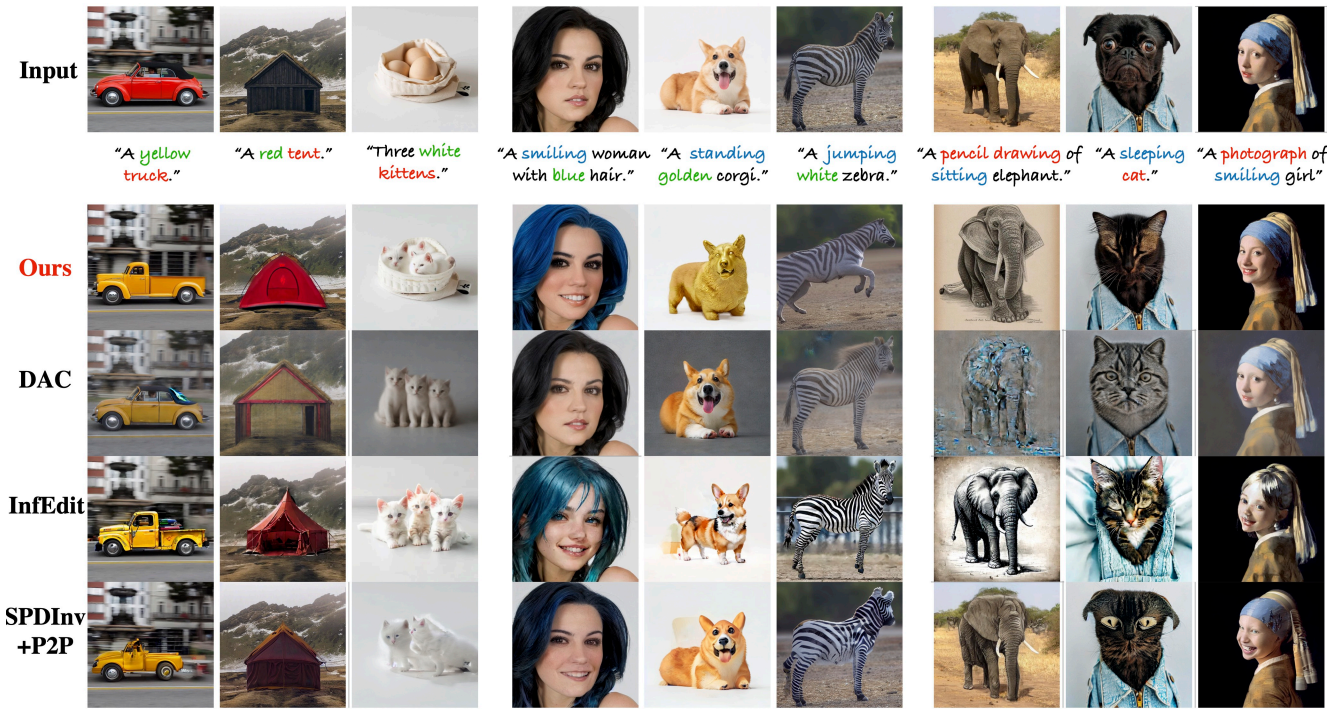


Figure 17. **More Comparision Results on Multi-function Editing**, where existing methods perform muti-turn edting.

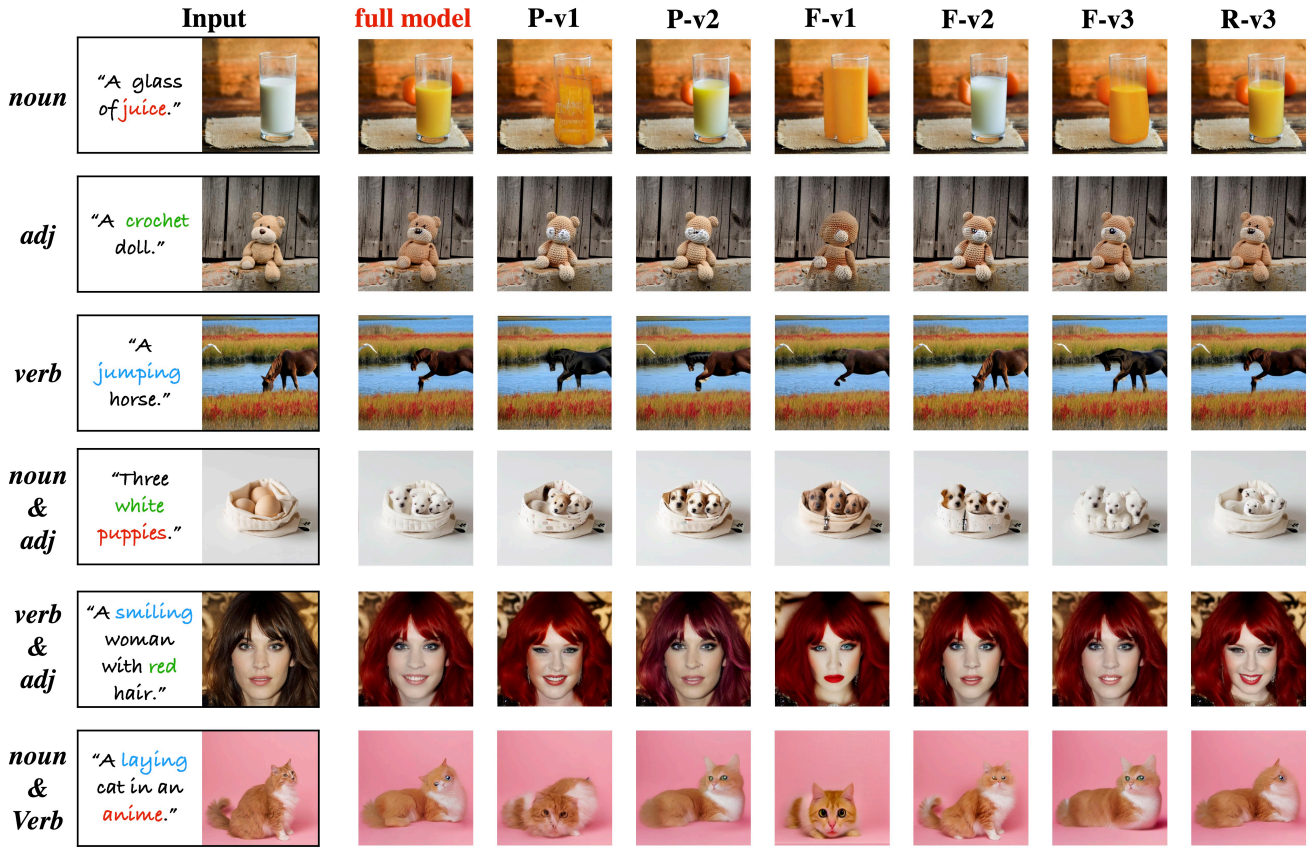


Figure 18. **Ablation results on different components.**

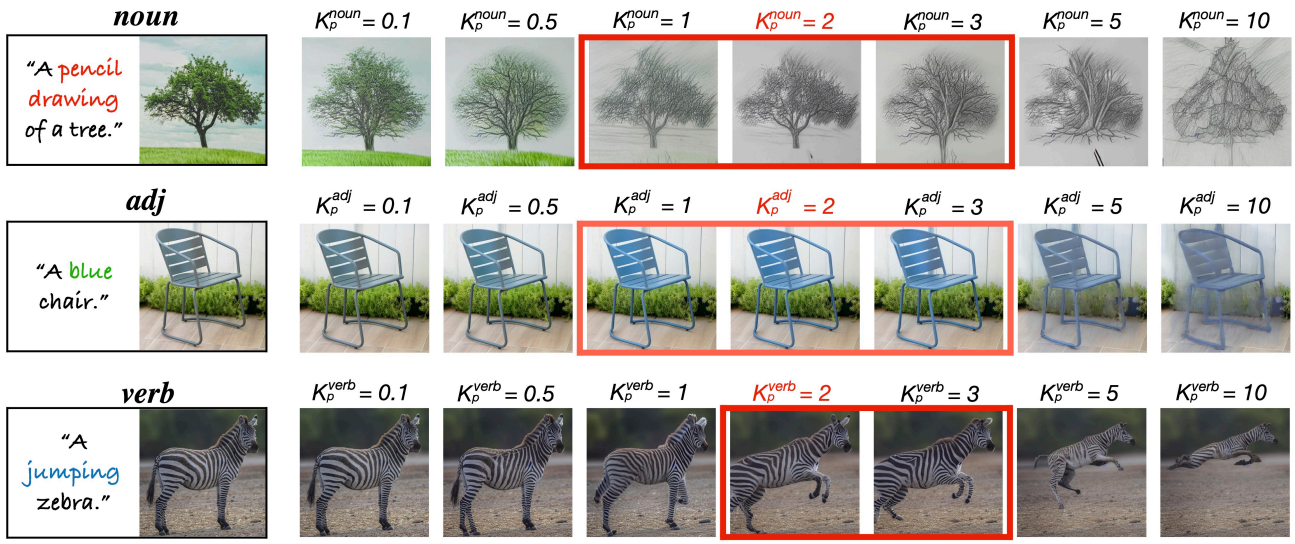


Figure 19. **Ablation results on feedback parameter** $\{K_p^j\}_{j\in\{noun,adj,verb\}}$.

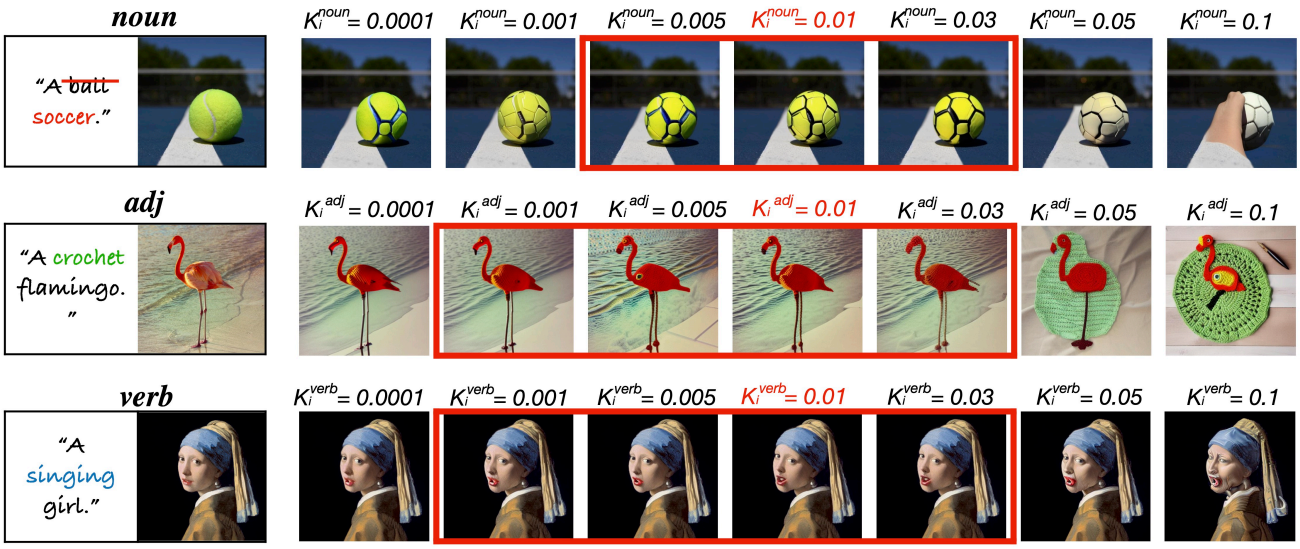


Figure 20. **Ablation results on feedback parameter** $\{K_i^j\}_{j\in\{noun,adj,verb\}}$.