Focus on Your Instruction: Fine-grained and Multi-instruction Image Editing by Attention Modulation

Supplementary Material

A. Ablation Study

In this section, we offer a detailed quantitative analysis of each component within our method. We compiled a dataset of 40 real images, each paired with 1 to 4 instructions. Our analysis primarily revolves around the changes in *CLIP image similarity* (hereafter referred to as *CLIP-I*) and *CLIP text-image direction similarity* (hereafter referred to as *CLIP-D*). The former metric gauges image similarity, while the latter assesses the degree to which the editing direction is followed.

Mask Extraction Steps. Tab. 3 illustrates the impact of mask extraction at various denoising steps. Notably, after completing mask extraction, we proceed with standard cross-condition attention modulation and disentangle sampling. It is evident that earlier mask extraction leads to higher scores in both *CLIP-I* and *CLIP-D*.

Cross Condition Attention Modulation. Tab. 4 shows the effects of ceasing cross-condition attention modulation at different denoising steps. It is important to note that the mask is always extracted in the first denoising step, and disentangle sampling is utilized in 75% of the remaining inference steps. Ceasing at step 1 implies no use of cross-condition attention modulation, whereas ceasing at step 80 means its constant usage. The results show that both *CLIP-I* and *CLIP-D* scores increase with more steps of cross-condition attention modulation, validating its contribution to more granular and precise editing.

Disentangle Sample. Tab. 5 reveals the outcomes of stopping disentangle sampling at different steps. Here, the mask is consistently extracted in the first denoising step, and cross-condition attention modulation is applied in all remaining steps. Stopping at step 1 indicates no use of disentangle sampling, whereas stopping at step 80 represents its continuous use. We observe that with increasing steps of disentangle sampling, *CLIP-I* continually rises, but *CLIP-D* first increases and then decreases. This is due to the suboptimal results when disentangle sampling is used throughout all steps, resulting in inconsistencies across the image. The qualitative outcomes of this are detailed in Sec. 5.3.

Step	0	20	40	60	79
CLIP-I	**0.9260**	0.9183	0.9059	0.8910	0.8805
CLIP-D	**0.1745**	0.1724	0.1715	0.1708	0.1697

Table 3. Mask Extraction in Different Denoising Steps.

Step	1	20	40	60	80
CLIP-I	0.9172	0.9179	0.9201	0.9248	**0.9260**
CLIP-D	0.1701	0.1715	0.1729	0.1736	**0.1745**

Table 4. Cross condition attention modulation end in different denoising steps.

Step	1	20	40	60	80
CLIP-I	0.9176	0.9185	0.9190	0.9260	**0.9329**
CLIP-D	0.1729	0.1732	0.1738	**0.1745**	0.1729

Table 5. Disentangle sample end in different denoising steps.

B. Limitations

Our approach encounters certain limitations. Although smaller cross-attention maps are rich in semantic content, they restrict our ultra-fine editing ability to an extent. Furthermore, our method's effectiveness is heavily dependent on the capabilities of the pretrained IP2P model. In Fig. 8(a), our method struggles with accurately locating *"rings"*, resulting in the fingernails also turning red. Furthermore, our method is constrained by the inherent capabilities of IP2P. For instance, in Fig. 8(b), while IP2P fails to execute instructions accurately, our method improves the editing result, yet it still cannot perfectly fulfill the instructions.

C. Additional Results

C.1. Additional Qualitative Results

In Fig. 7, we provide additional qualitative comparisons between our method and baseline models. The Input image is randomly selected from PIE-bench [22].

C.2. Enhanced Control Over Sub-Instruction Intensity

As mentioned in Sec. 4.2, as illustrated in Fig. 9, our method enables flexible control over the intensity of different sub-instructions, providing a level of finesse and granularity not achievable with previous methods.

C.3. Comparison of Editing Speed

The evaluation of diverse editing techniques' speed was conducted on a GeForce RTX 3090, with results detailed in Tab. 6. This assessment involved randomly selecting 200

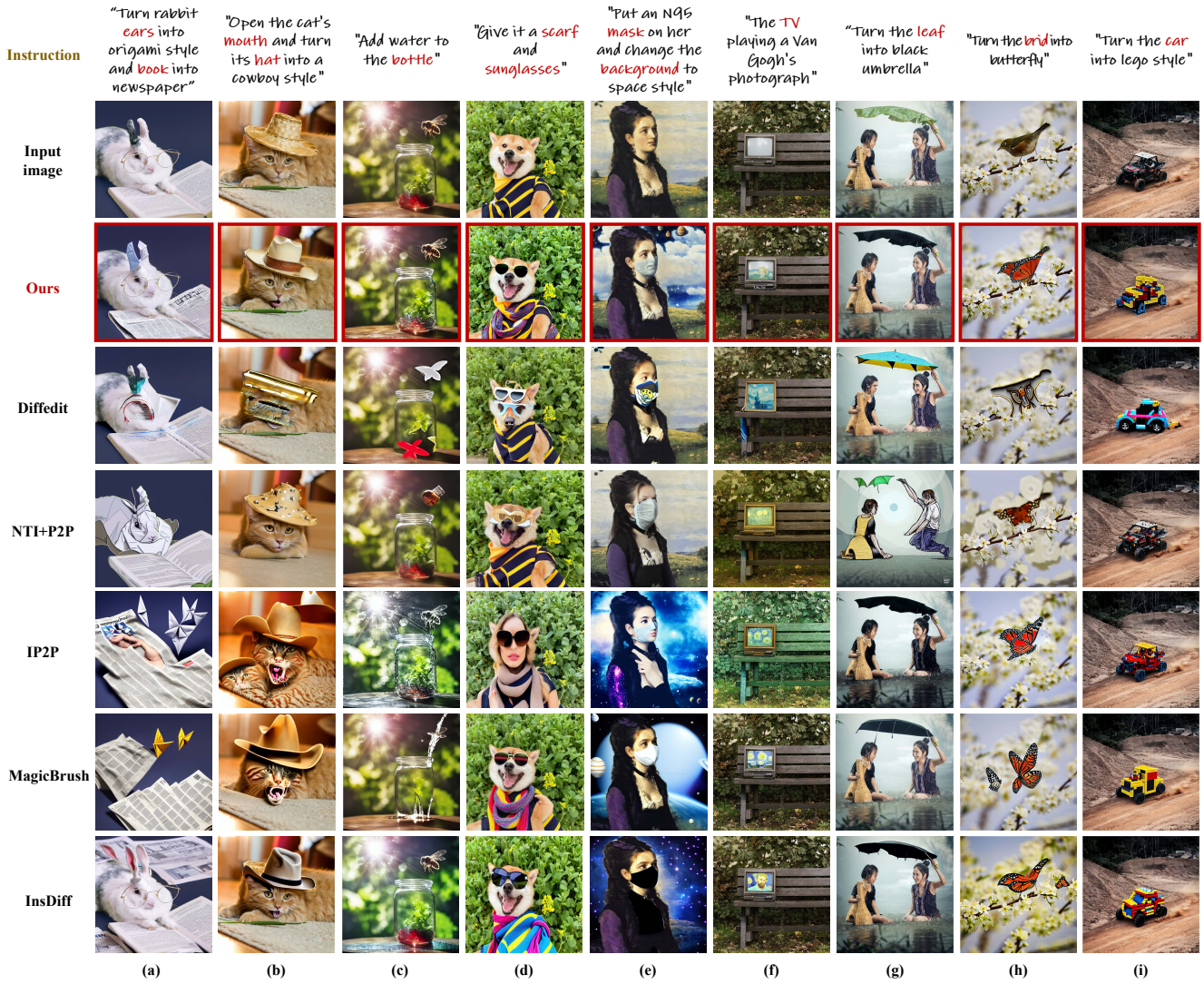

Figure 7. **Additional qualitative comparison.**

Figure 8. Failure cases. Undesired edits are marked with orange dashed boxes.

images and calculating the average inference time. To guarantee a fair comparison, the inference steps for each model were standardized at 50 steps. Consequently, for FoI, the effective number of inference steps utilized is 40.

Method	Diffedit	NTI+P2P	IP2P	MagicBrush	InsDiff	FoI(ours)
Avg Time Cost(s)	15.60	138.20	7.35	7.34	7.60	**6.79**

Table 6. Inference time comparison.

C.4. Human Preference Study

Fig. 10 presents the questionnaire form used in our human preference study, where the display order of the six methods was randomized for each question.

D. Societal Impact

Our study introduces a fine-grained, multi-instruction editing scheme for images. This nuanced alteration of images could potentially be exploited by malevolent entities to produce false content and disseminate misinformation, a

Figure 9. **Controling sub-instruction guidance scale.** Every result image is labeled with the enhancement scale of the corresponding sub-instruction.

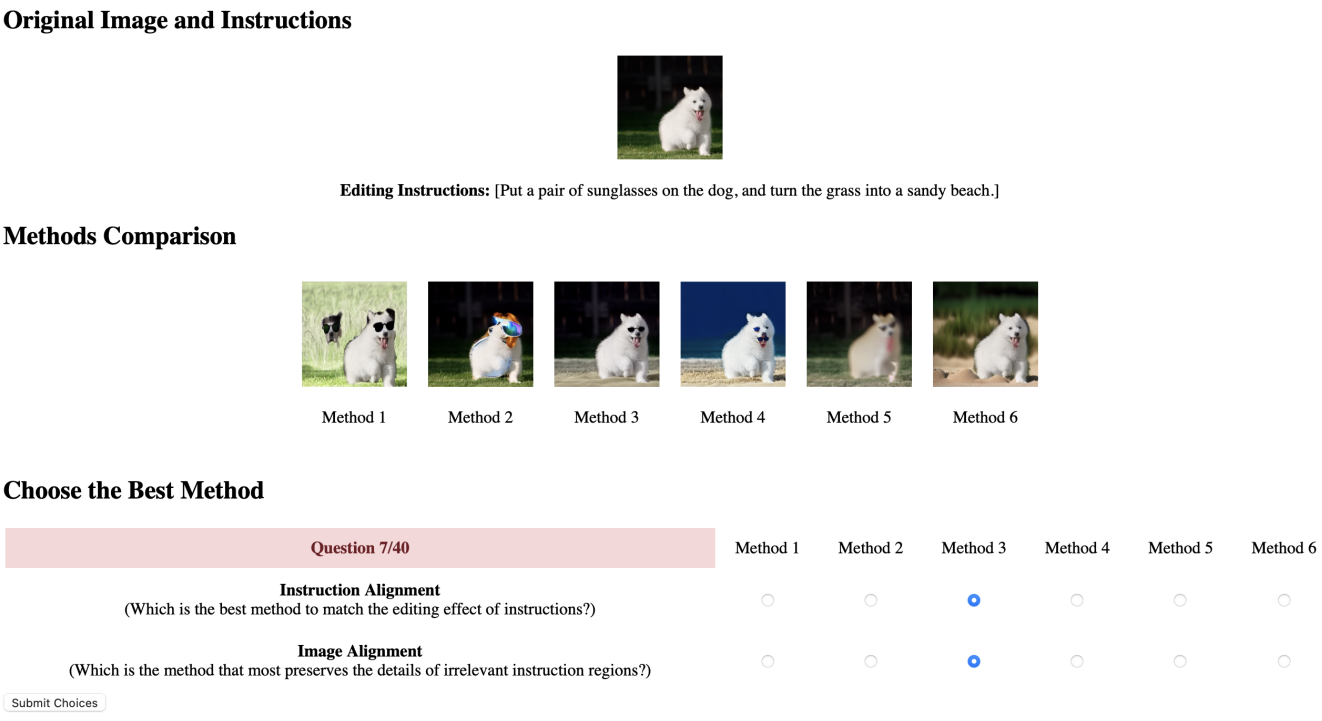

Figure 10. **Human preference study print screen.**

well-known issue inherent to all image editing techniques. However, our method uniquely generates a mask for the edited regions upon completion, aiding in the training of models to detect forged images. We believe our work contributes significantly to this field by providing an analysis of instruction-based image editing methods.