The **Appendix** is organized as follows:

- **Section A:** provides more details on sample selection process and shows more experimental results.
- **Section B:** gives more ablation analysis.
- **Section C:** provides more details on user study.
- **Section D:** gives analysis on the limitations.

A. Additional Results

Selection Details. In the main paper and Appendix, we use the following sample selection process for all comparative methods. (1) The hyperparameter $\eta \in \{0.4, 0.6, 0.8\}$ is adopted in DAC. (2) The hyperparameter for text embedding interpolation in Imagic [14] is in $\{0.9, 1.2, 1.4\}$. (3) DDS [10] sets different numbers of the classifier free guidance scale as 3, 5, and 7.5 respectively. In addition to different values of hyperparameters, for each editing, we randomly generated 8 edited images given a source image and an editing prompt and chose the one with the best quality as the final edited image.

A.1. Quantitative Results

To further evaluate the effectiveness of DAC, we leverage a random subset of 200 paired prompts and images in the InsturctPix2Pix dataset [3]. Considering that SINE [43] requires a huge time cost for a single image editing (*i.e.*, 2 hours), we exclude it from this comparison. The results are listed in Table 2. DDS [10] obtains the best image alignment with source images (*i.e.*, the lowest LPIPS score) while the worst text alignment with prompts (*i.e.*, the lowest CLIP-score). This is because DDS [10] mostly makes no change to source images, thus failing to achieve effective editing. Compared with Imagic [14], our DAC archives a lower LPIPS score and a higher CLIP-score, which demonstrates higher fidelity to source images and better editability. Therefore, DAC fulfills a better trade-off between fidelity and editability for text-based image editing.

A.2. Qualitative Results

The examples for qualitative comparisons on InstructPix2Pix dataset [3] are shown in Figure 13. For the first example, the text prompt aims to change the yellow taxi to a green one. It could be seen that DAC successfully modifies the color of the taxi while maintaining other components in the input image. By contrast, the edited images from the Imagic [14] and DDS [10] are even the same as the source image, inconsistent with the text prompt. Considering the second example, all three methods attain effective editing.

Table 2. Quantitative comparisons on InsturctPix2Pix dataset [3].

Methods	**DAC**	Imagic [14]	DDS [10]
LPIPS ↓	0.40	0.43	0.24
CLIP-score ↑	32.3	31.4	30.8

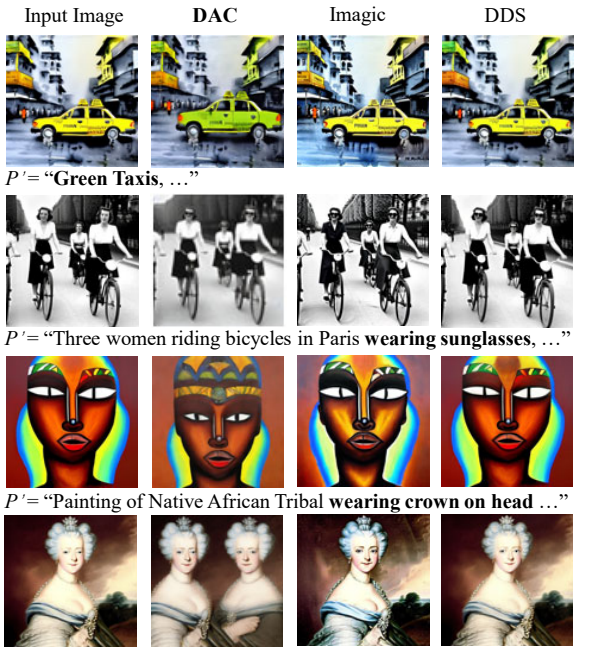

Figure 13. Visual comparisons on InsturctPix2Pix dataset [3].

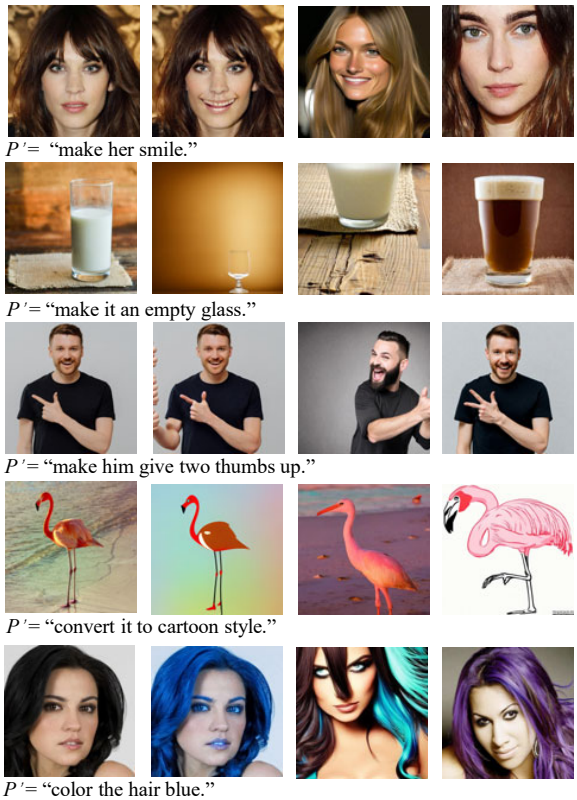

Figure 14. Qualitative examples of large-scale training methods.

For the third and fourth examples, DAC adds a crown and a woman according to the text prompts separately. The results of DDS [10] fail to achieve the desired editing although it keeps high fidelity to the input images, thus explaining the

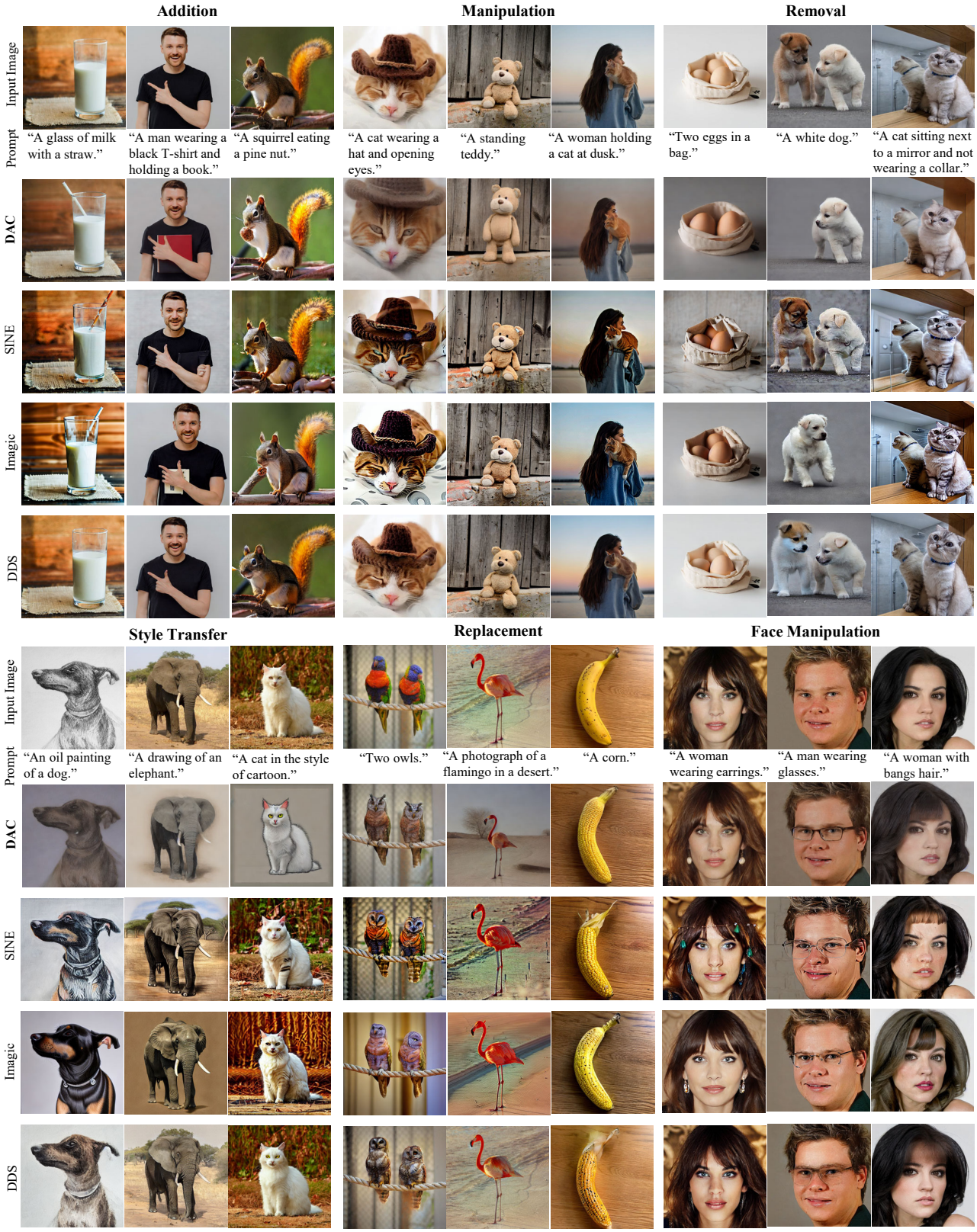

Figure 15. Comparison of TBIE qualitative examples across the 6 editing types (only prompt P' shown) between our DAC and three SOTAs.

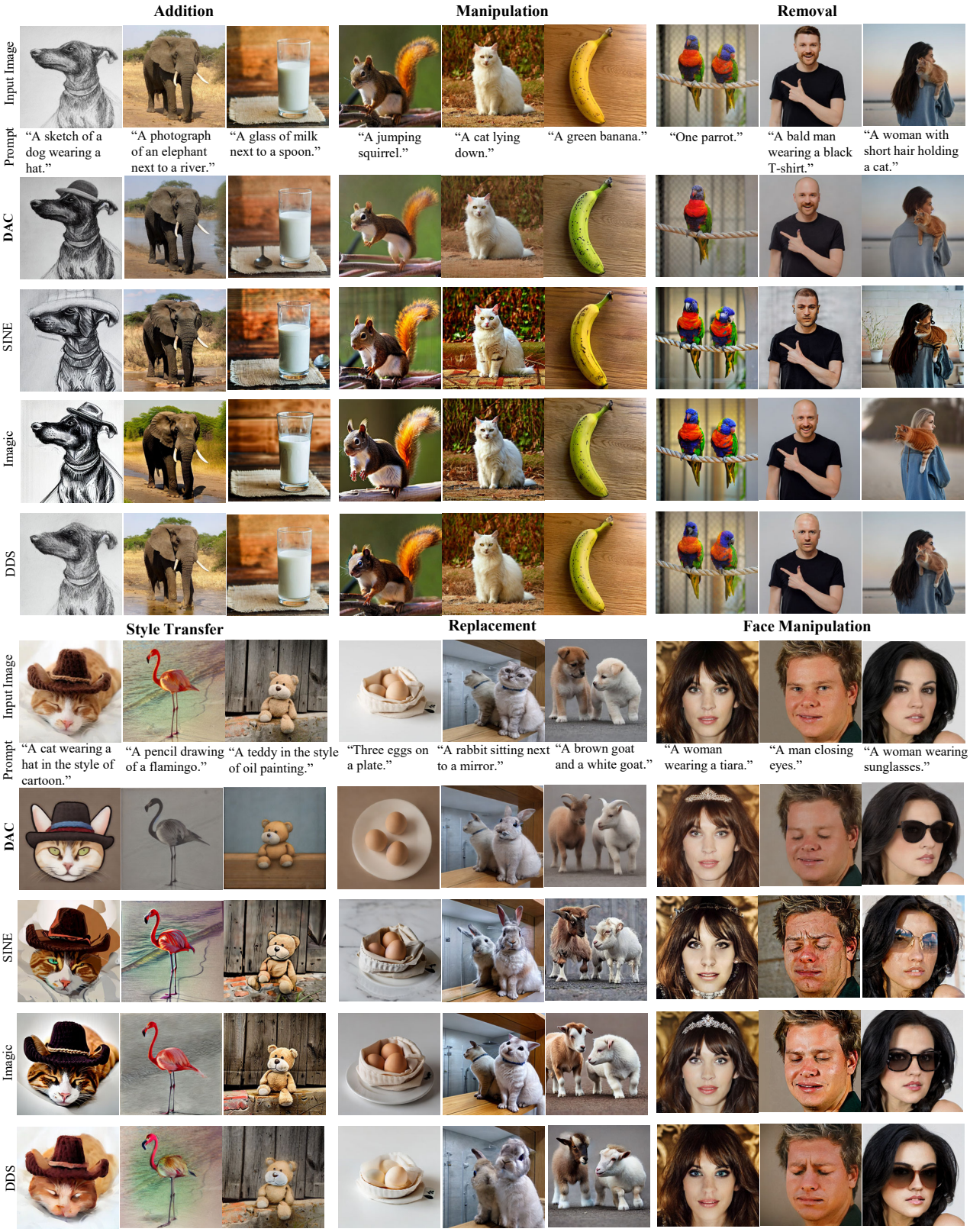

Figure 16. Comparison of TBIE qualitative examples across the 6 editing types (only prompt P' shown) between our DAC and three SOTAs.

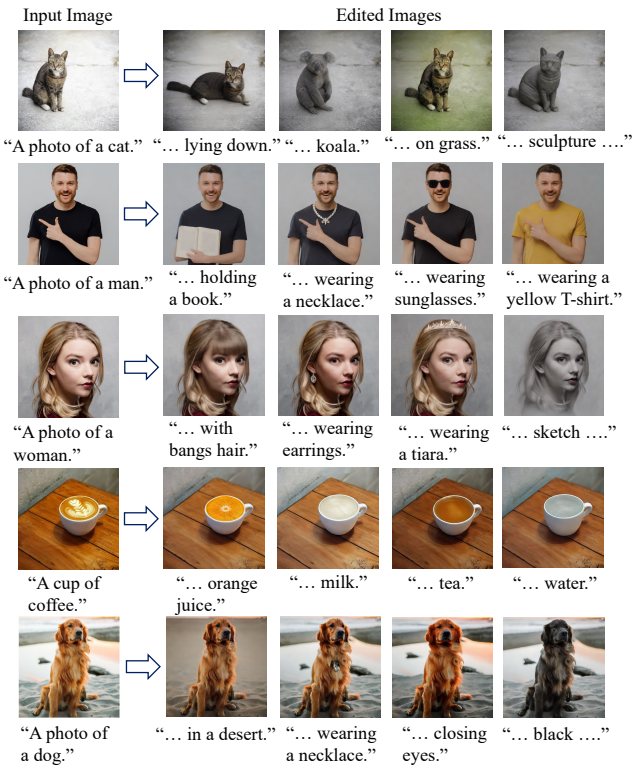

Figure 17. Qualitative examples of DAC with different prompts editing on the same source image.

best LPIPS score of DDS [10] in Table 2.

Additionally, Figure 14 gives more qualitative examples of large-scale training methods. Figures 15 and 16 provide extra qualitative comparisons for the six editing types, contrasting our DAC with three state-of-the-art methods.

Editing with Multiple Prompts. As shown in Figure 17, we generate the edited images with a source image and multiple editing prompts. With a photo of a man, we enable him to hold a book, wear a necklace, wear sunglasses, or change the black shirt to a yellow one, while keeping a good fidelity of the source image. It also shows that our DAC enjoys impressive editing ability when applied to various images with different language guidance, manifesting the good versatility of our method.

B. Ablation Analysis

Ablation on UNet LoRA. The LoRA structure in DAC is built on all of the attention layers, convolutional layers, and feed-forward (FFN) layers since we observe the underfitting issue if we only apply LoRA on the attention layers of UNet. The underfitting issue means that we could modify the image by directly changing the text prompt with U. Figure 18 shows the ablation results of w/o and w/ Conv and FFN LoRA in the UNet. For U w/o Conv and FFN LoRA, we could get "A bald man", "A jumping dog", and "A woman in a big smile" with the target prompts. However, the fidelity of edited images is lost. For example, the identity of the man even changes in the removal editing of Figure 18. By contrast, with U containing Conv and FFN LoRA, we couldn't alter the image anymore, thus overcoming the underfitting issue.

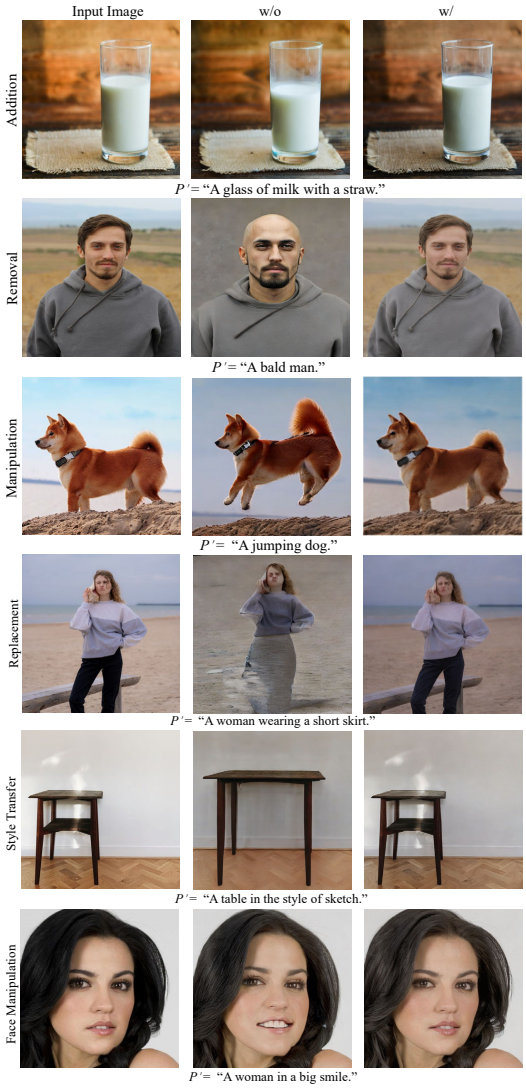

Figure 18. Ablation on UNet w/o and w/ Conv and FFN LoRA considering the underfitting issue.

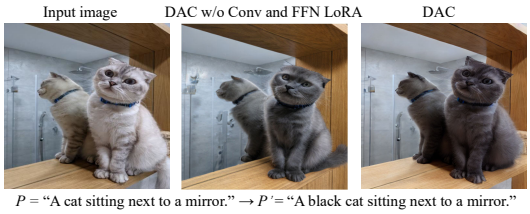

Figure 19. Ablation on UNet LoRA considering the fidelity.

In addition, for U, we added the LoRA structure on all of the attention layers, convolutional layers, and FFN layers to guarantee the fidelity of the source image. As shown in Figure 19, if we only used attention layers LoRA, the edit-

Figure 20. User study screenshot for one example.

Figure 21. A qualitative example of muti-turn editing.

Figure 22. Failure cases due to the issues in stable diffusion.

ing would blur the background details; after adding LoRA to convolutional layers and FFN layers, we can retain the details successfully.

C. User Study Details

We quantitatively evaluate our DAC with an extensive human perceptual evaluation study conducted on AMT. Concretely, we collected a diverse set of image-prompt pairs, covering all the "addition", "manipulation", "removal", "style transfer", "replacement", and "face manipulation" editing operations. Each operation includes 9 different prompt-image pairs, thus constituting 54 examples in total (i.e., examples in Figure 5 in the main paper, Figures 15, and 16). The number of AMT participants is 110 and for each evaluator, 54 examples are shown. Moreover, one example consists of a source image, a target prompt, and 4 edited images by DAC, DDS, SINE, and Imagic, which were randomly listed. The user study screenshot for one example is depicted in Figure 20. We listed instructions of our editing evaluation for evaluators. Note that we emphasize a good edited image should fulfill both the alignment with the text prompt and similarity with the original image.

D. Limitations

Multi-turn Editing. In Figure 21, as the turn increases, our DAC achieves successful editing aligning with the text prompts while the image quality gradually declines. It is caused by the information loss in Abduction-1 as illustrated in Figure 12 and Ablation on Abduction-1. Although we could complete such loss by incorporating another abduction in Abduction-1, it may be time-consuming. To solve the image quality degradation in multi-turn editing, we need to explore time-efficient fine-tuning (e.g., Fast Diffusion Model [39]) for the abduction process in DAC. We leave it to our future work.

Failure Case Study. We observed three kinds of failures caused by stable diffusion: 1) sensitivity to random seeds, 2) the incapability of comprehending referring expressions, and a more subtle case 3) the lack of common sense. As shown in Figure 22, random seed impacts the success rate of generation; the second failure is due to the fact that stable diffusion cannot always generate images according to the prompt with referring expressions like "a white dog next to a brown dog and the brown dog is wearing a hat"; if we change the object from cat to fish, we should also change the background from land to water due to the common sense "fish lives in water". To fundamentally resolve such failures, maybe we need to improve stable diffusion to endow such capabilities. We leave it to our future work.