Prompt Augmentation for Self-supervised Text-guided Image Manipulation

Supplementary Material

In this supplementary material, we delve deeper into the capabilities of our proposed method for text-guided image manipulation. Included are implementation details, limitations, illustrative samples of prompt augmentations, insights into the mask generation process, and supplementary qualitative comparisons that further enrich the findings presented in the main paper.

A. Implementation Details

We employed a pretrained LDM, specifically Stable Diffusion v1.4, as the foundation of our experiments. Fine-tuning the UNet was conducted for 20,000 steps, utilising a learning rate of 1.0e-05. The experiments were executed on an NVIDIA Titan RTX with 24GB of GPU memory, employing a batch size of 1. The image size used was 384, and we augmented each image with 3 target prompts as decsribed in Sec. 3.1.

B. Limitations

While our work demonstrates notable achievements in image manipulation, it is crucial to acknowledge its limitations. Prompt augmentation, while beneficial for introducing diversity during training, presents challenges. Overreliance on augmented prompts may carry the risk of bias, potentially hindering the model's generalisation to new prompts. Our current strategy, focusing on the manipulation of adjectives and nouns, represents a subset of the infinite possibilities for image edits. Hence, the augmentation process remains suboptimal, emphasising the need for more sophisticated techniques, potentially involving the training of a dedicated large language model for this purpose. Additionally, the resource limitations with a batch size of 1 limits the potential of our approach.

C. Prompt Augmentation and Mask Generation

In Figure S1, we present a set of samples illustrating the augmentation process and mask generation. The ”Input” column features images dynamically captioned by using BLIP [22], offering diverse scenes. The “Augmented Prompts” column displays variations for each input, providing insights into augmentation diversity. For example, for an image with wolves, augmented prompts include “Two werewolves in the snow,” “Two cats in the snow,” and “Two trees in the snow.” These variations serve the purpose of the proposed soft contrastive loss. In the “Generated Mask” column, masks are featured, generated by leveraging the augmented prompts. The alignment of images and masks in the table provides a clear view of the augmentation and mask generation results.

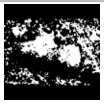
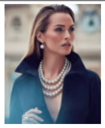
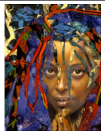
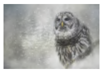
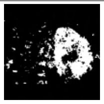
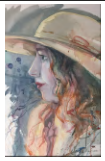
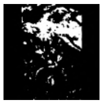
Input	Augmented Prompts	Generated Mask
 Two wolves in the snow	Two *werewolves* in the snow Two *cats* in the snow Two *trees* in the snow	
 *A woman wearing a pearl necklace*	A woman wearing a *silver* necklace A woman wearing a *sapphire* necklace A woman wearing a *charm* necklace	
 A woman wearing a silver headpiece	A woman wearing a *silvery* headpiece A woman wearing a *sterling* headpiece A woman wearing a *jade* headpiece	
 A blue dress	A *black* dress A *white* dress A *red* dress	
 *A woman with blue hair*	A woman with *black* hair A woman with *green* hair A woman with *red* hair	
 *An owl in the snow*	An *eagle* in the snow A *goose* in the snow An *elf* in the snow	
 A group of men standing in front of an airplane	A group of *gentlemen* standing in front of an airplane A group of *monks* standing in front of an airplane A group of *pirates* standing in front of an airplane	
 *A woman wearing a hat*	A woman wearing a *cap* A woman wearing a *hood* A woman wearing a *cloak*	

Table S1. Samples for prompt augmentation and the masks generated by using these prompts.

D. Additional Qualitative Comparisons

Figure S1 showcases additional qualitative comparisons between our method and prominent approaches, including SDEdit [27], DALL-E 2 [31], DiffEdit [11], and Instruct-Pix2Pix [8]. Our observations echo those discussed in the main paper. Notably, SDEdit [27] grapples with the chal-

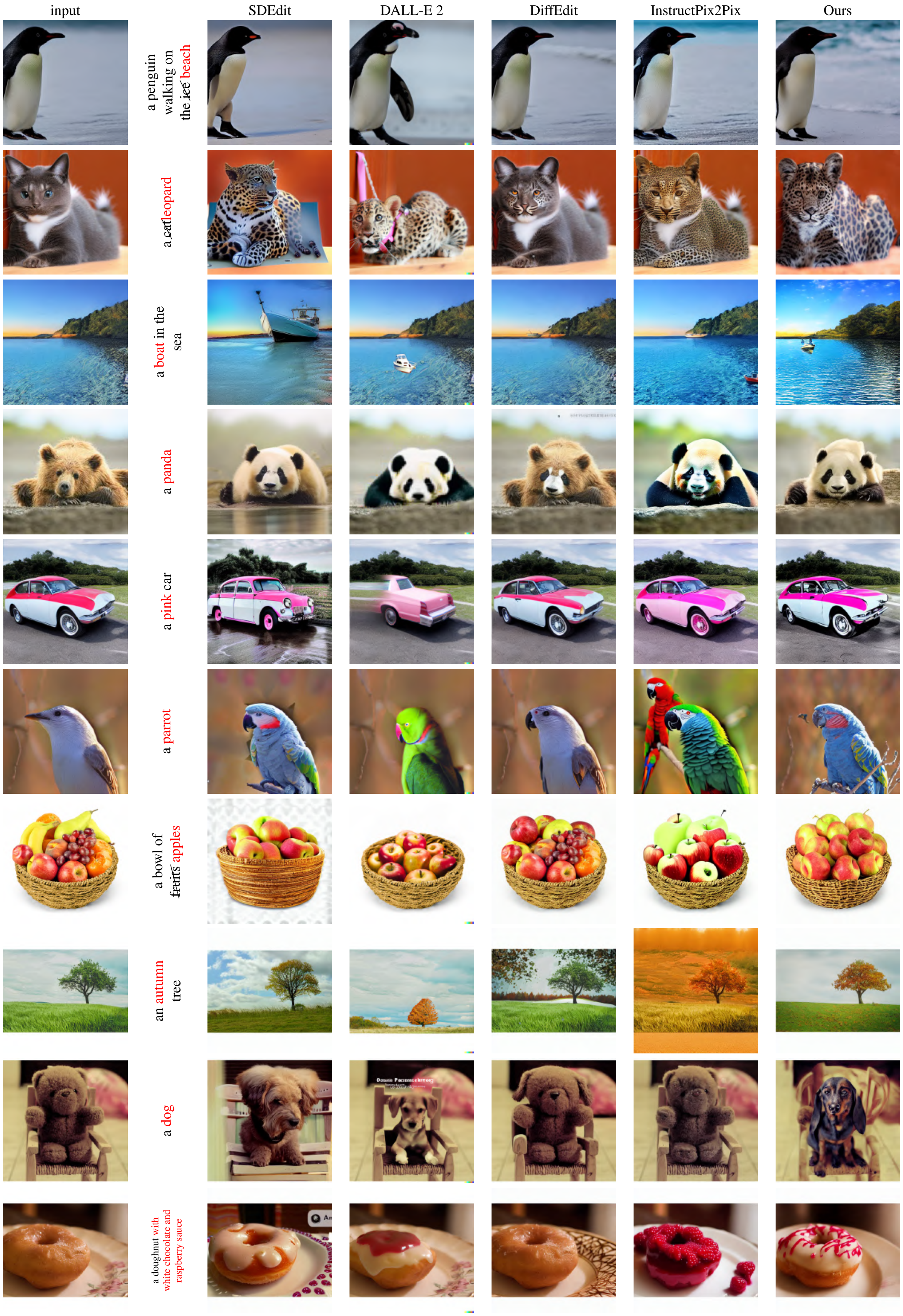


Figure S1. Additional qualitative results comparing our method against SDEdit [27], DALL-E 2 [31], DiffEdit [11] and InstructPixtoPix [8] using both generated and real images.

lenge of maintaining a balance between fidelity to the input image and adherence to the target prompt, as evident in examples like ‘a boat in the sea’, ‘a bowl of apples’, and ‘a pink car’. DALL-E 2 [31], functioning as an inpainting method reliant on masks, tends to discard content within the masked area while preserving the inverse mask area. This often leads to mismatches between input and edited content, as seen in the positions of ‘a leopard’, ‘a panda’, ‘an autumn tree’, and the quantity of ‘apples’. Additionally, it encounters difficulty in seamlessly blending with the rest of the image, exemplified by the case of pink car’. DiffEdit’s sensitivity to detected masks, dependent on parameters and the employed image generation model, results in occasional failures in accurate mask detection (a leopard’, a panda’, a bowl of apples’) or the selection of completely incorrect areas (‘doughnut’, ‘autumn tree’). InstructPix2Pix may generate non-realistic images (‘a panda’, ‘a leopard’), face translation issues (e.g., ‘teddy bear translated into a dog’), impact the entire image with the target prompt (‘a parrot’, ‘autumn tree’, ‘a pink car’), or struggle with multiple simultaneous changes (e.g., ‘doughnut with raspberry and white chocolate sauce’). Despite not using masks at inference like DALL-E 2 and DiffEdit and not being trained on a dataset with target images like InstructPix2Pix, our method achieves more successful translations with high fidelity to the target prompt and input image, accompanied by minimal undesired changes.

E. User Study

As described in the main paper, we executed a user study adhering to established procedures on Microworkers. A total of 540 pairs were presented to users, each consisting of input images, target prompts, and manipulation outcomes generated by both our proposed method and a baseline. Each pair underwent evaluation by 10 distinct participants, tasked with identifying the more effective manipulation in terms of fidelity to the prompt and input image. This process yielded a comprehensive dataset of 5.4K responses. The outcomes of this survey are detailed in Table 1 of the main paper. The results consistently demonstrate a preference for our proposed method over the baseline.

F. Integrating DiffEdit into Our Framework

In Figure S2, we provide additional qualitative comparisons between our method and DiffEdit [11]. The first column shows the input image, and the second and third columns display DiffEdit and our results, respectively. As DiffEdit is exclusively utilised at inference time, we seamlessly incorporate it into our method, presenting the combined outcomes in the last column. Identical parameters are employed for computations in both DiffEdit and DiffEdit+Ours. DiffEdit utilises Stable Diffusion’s checkpoint v1.4, the same pretrained model used for fine-tuning with our approach.

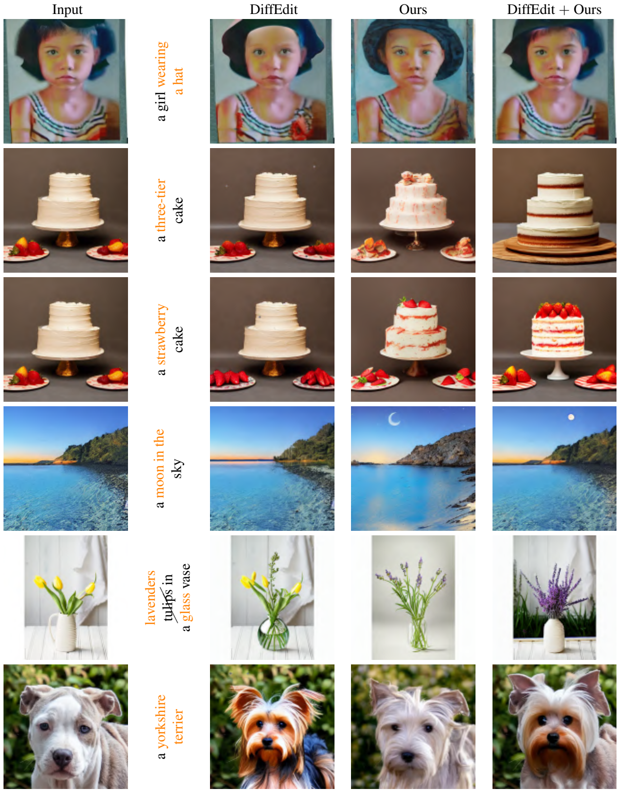


Figure S2. Qualitative comparison of DiffEdit, our method and their combination.

When using DiffEdit in conjunction with our method, we observe improved mask detection. For example, in ‘a girl wearing a hat’, DiffEdit alters the girl’s face while adding a hat. Upon integration into our method at inference, the face remains unaltered while incorporating the hat. Our method alone achieves a successful translation with minimal undesired modifications in the face area, and this is further mitigated with the integration of DiffEdit. A similar scenario is evident in ”a moon in the sky,” where DiffEdit manipulates an unrelated area on the shores. Integrated into our method, it produces a successful translation while preserving unrelated areas effectively. Once again, our method alone, while placing the moon in the desired location, introduces some undesired alterations, which are minimized through the integration of DiffEdit. In ‘a strawberry cake’, DiffEdit inaccurately manipulates the plates rather than the cake. While our method achieves a successful translation with minimal undesired changes, integration with DiffEdit further prevents these changes, preserving all parts of the image perfectly. These results highlight that our fine-tuning strategy enhances the model’s ability to detect editing masks compared to the baseline.

Consistent with our earlier observations, while our method enhances the accuracy of mask detection in DiffEdit, it remains sensitive to the detected masks and corresponding parameters. An example is seen in ‘lavenders’, where integrating DiffEdit into our framework leads to undesired manipulations in the background and fails to successfully translate the vase into a glass one although it preserves some details better. Despite DiffEdit detecting a relatively accurate mask in this case, it struggles to translate tulips into lavenders successfully. In summary, while integrating DiffEdit often improves our method, particularly in preserving the background, its performance is still influenced by parameter sensitivity in mask detection. Further refinement of these parameters could enhance the overall results.