# Improvement on Prompt-to-Prompt Image Editing

**Yanchong Peng, Yuhe Peng, Peng Huang**

{ycpeng, pengyh, phuang}@bu.edu

## Abstract

The original algorithm in Prompt-to-Prompt Image Editing with Cross Attention (Amir et al. 2022) has some limitations, including generating both the source and edited images with one random seed and being unable to move out a specific target in the edited image. This report covers some improvements we made to the prompt-to-prompt image editing algorithm. First, it proposes two improvements in the random generation of original and edited images to improve the performance in continuous editing. Second, it provides a way to remove a specific target from the source image by providing a 'removing prompt'. Finally, it implements a function to change the position of a target in an identical image.

## Introduction

Many techniques have been proposed in text-to-image, such as Imagen(Saharia et al. 2022) and DALL-E2(Ramesh et al. 2022). However, these models cannot control the global sections of the image; that is, if changing a tiny part of the text prompt to another, such as "dog" to "cat", the whole generated image may not be the same, not only "dog" but also the background.

**Related Work** Bau(Bau et al. 2021) showed that the mask provided by the user could help edit the text-based image locally. This mask contains all the areas relevant to the text that the user wants to change, whereas it requires lots of labour and time, which is inefficient. Seminal works combining GAN(Goodfellow et al. 2020; Brock, Donahue, and Simonyan 2018), primarily used to generate high-resolution images, and CLIP(Radford et al. 2021), which contains rich image-text representation, do not require extra labour annotations. Nevertheless, they cannot handle vast and varying datasets.

Text2LIVE(Bar-Tal et al. 2022) also can edit text-based localized images. However, this model needs to apply the network to each input, which is highly time-consuming and can only change the textures, not the whole patterns, such as a plane to a train.

Hertz(Hertz et al. 2022), whose paper this project is relevant to, modifies the internal attention maps of the diffusion model, which not only preserves the entire sections of the origin but also modifies the part that the user wants to replace. This way, we need the text input and do not require extra labour. The Hertz team has released the newest version of the prompt-to-prompt model using pre-trained hugging face/diffuser models. They implemented the method based on Latent Diffusion and Stable Diffusion, respectively. We will produce our work based on the most updated GitHub repository.

At the same time, there are also some papers involving improving and applying different diffusion models.

DreamBooth(Nataniel et al. 2022): This paper presents another way to personalize a text-to-image diffusion model. They first generated a unique identifier to represent the subject's name and got a low-resolution image with the most considerable extent of fidelity. (For example, use a [V]dog' to represent 'a dog'). Then, apply the low-resolution image to super-resolution diffusion models to get the high-resolution version of the input images.

RE-IMAGEN(Wenhu et al. 2022): Some traditional text-to-image diffusion models have difficulty precisely generating the images from less frequent prompts such as "Chortai' – a type of dog, or "Picarones' – a kind of food. This paper tackles this problem by applying the rare prompt to a multimodal knowledge base to get the relevant image-text pairs. Then, they use it as a reference to build the target images. Similar to our project, this paper deal with two conditions, text prompt and retrieved relevant pairs.

Fashion Image Manipulation(Chaerin et al. 2022): This paper introduces a classifier-based diffusion framework to provide preciseness and effectiveness in fashion attribute editing (ex., Search clothes by attribute on an e-commerce website). They proposed a finetuning scheme to improve the pretrained Vit for domain-specific multi-attribute classification settings.

UNIFYING DIFFUSION MODELS' LATENT SPACE(Chen and Fernando 2022): Many diffusion models adopt the formulation of a sequence of gradually denoised samples. This paper proposes a unified approach to generate a diffusion model by reformulating different diffusion models as deterministic maps from a Gaussian latent code to an image

**Goal** For the current stage, this technique is subject to several limitations. First, there is only one random seed $s$ for both source and edited image, which means it is impossi-

ble to keep the images for both source and edited that randomly appear. Second, although it provides a way to reduce global noise(-rocks, -fog), it does not have the functionality of removing a specific target. Moreover, the method cannot change an object's location in the picture. Finally, the algorithm for sequential image editing is inefficient.

So, in this project, we expect to make some improvements to the existing algorithm to achieve the following goals:

- Add one more random seed $s^*$ to generate and keep source/edited images separately.
- Remove a specific target in the image.
- Combine two specific images. The generated image 3 includes the objects both from image 1 and image 2.
- Move a specific object across the image.
- Improve the algorithm performance in continuous editing.

## Methods

Most of our work is based on manipulating the attention map inspired by the Prompt-to-Prompt Image Editing with Cross Attention (Amir et al. 2022). Based on the existing attention replacement method, we create attention removal functions to reassign the attention map for the target to be removed. Meanwhile, we extend the single random seed to multi-seeds to keep the features from both source and edited.

The code provided by the author performs the main logic by AttentionControl class. They define AttentionReplace to swap the critical word of the original prompt with others, ex. "A painting of a monkey eating a banana" to "A painting of a monkey eating a burger". We use this AttentionReplace class in the multiple seed editing and modify the utils function to take in a random seed batch to enable input seeds array. Moreover, we create the AttentionRemove function to assign new attention to the removed target.

### Formulation

The source image $\mathcal{I}$ is generated by a text-guided diffusion model with prompt $\mathcal{P}$ and random seed $s$. The goal is to edit the source input image with edited prompt $\mathcal{P}^*$, and another random seed $s^*$.

### Model and Dataset

We use the pre-trained model Stable Diffusion v1-4(Rombach et al. 2022) as the backbone of this algorithm.

This model was trained with 256*256 resolution on the laion2B-en dataset and 512*512 resolution on the laion-high-resolution dataset, which contains over 170M images with 1024*1024 resolution.

This model used the ViT-L/14 text-encoder to encode text prompts and encoded image prompts into latent representations. Later, with cross attention, the model injected the output of the text-encoder into the UNet(Ronneberger, Fischer, and Brox 2015) in the diffusion model.

Algorithm 1: Multi-seed image editing

**Input**: A source prompt $\mathcal{P}$ with a random seed $s$, a target prompt $\mathcal{P}^*$ with a random seed $s^*$

**Output**: A source image $x_{src}$ and an edited image $x_{dst}$

$z_T \sim N(0, I)$ a unit Gaussian random variable with random seed $s$
$z_T^* \sim N(0, I)$ a unit Gaussian random variable with random seed $s^*$
**for** $t = T, T-1, ..., 1$ **do**
  $z_{t-1}, M_t \leftarrow DM(z_t, \mathcal{P}, t)$
  $M_t^* \leftarrow DM(z_t^*, \mathcal{P}^*, t)$
  $\widehat{M_t} \leftarrow Edit(M_t, M_t^*, t)$
  $z_{t-1}^* \leftarrow DM(z_t^*, \mathcal{P}^*, t)\{M \leftarrow \widehat{M_t}\}$
**end for**
**return** $(z_0, z_0^*)$

### Algorithms

- **1. Multi-seed Image Editing (accomplished)** The input of our modified algorithm has one more random seed than the original algorithm: one source random seed $s$ and one target random seed $s^*$, as described in Algorithm 1. Furthermore, since the initial latent image can entirely determine the randomness, there is no need for the diffusion model to accept any random seed anymore, which simplifies the design.(Hertz et al. 2022)
- **2. Removing Specific Target (accomplished)** The user can remove the object from the image generated before. For instance, the source text prompt is "a car parking on the street". If the user wants to remove the car, let the attention map of the token "car" as $j^*$. Then, the maps-editing operation function will be modified as

$$(Edit(M_t, M_t^*, t))_{i,j} := \begin{cases} -5 \cdot (M_t)_{i,j} & \text{if } j{=}j^* \\ (M_t)_{i,j} & \text{otherwise} \end{cases}$$

  where $i$ is the pixel value and $j$ is the text token. The weight $-5$ is the result after several experiments. With this scale, the output image performs more naturally and adaptively.
- **3. Combining Two Images (accomplished)** We proposed an extended application of combining two source images based on the original algorithm (Hertz et al. 2022) that only takes one source image.

  Suppose we have source prompts $\mathcal{P}$ and $\mathcal{P}'$, and a target prompt $\mathcal{P}^*$. This $\mathcal{P}^*$ combines the same objects from both $\mathcal{P}$ and $\mathcal{P}'$. In a single step $t$ of the diffusion process, we get the attention maps $M_t$, $M_t'$, $M_t^*$ respectively for prompts $\mathcal{P}$, $\mathcal{P}'$, $\mathcal{P}^*$, and put them into the edit function

$$(Edit(M_t, M_t', M_t^*, t))_{i,j^*} = \begin{cases} M_t & \text{if } j^*{=}j \\ M_t' & \text{if } j^*{=}j' \\ M_t^* & \text{otherwise} \end{cases}$$

  where $j$, $j'$, and $j^*$ are the text tokens for $M_t$, $M_t'$, and $M_t^*$ here.

For instance, if

$$\mathcal{P} : \text{"A woman playing soccer"}$$
$$\mathcal{P}' : \text{"A man playing basketball"}$$
$$\mathcal{P}^* : \text{"A woman playing basketball"}$$

then $j$ will be "woman" and $j'$ will be "basketball". The final generated image $z_0^*$ has the same "woman" object from $\mathcal{P}$ and the same "basketball" object from $\mathcal{P}'$.

- **4. Moving Object (accomplished)** We proposed a method of moving an object across an image, which initially serves as a limitation in the paper by (Hertz et al. 2022). For precise moving, we accept a target location $R$ as the destination of moving and a prompt token for the object we would like to move. We define a move function $Move(M_t, R)$ to move the focus area of $M_t$ to the target location $R$. Hence, the edit function should be defined by

$$(Edit(M_t, t))_j = \begin{cases} Move((M_t)_j, R) & \text{if } j \text{ is to move} \\ (M_t)_j & \text{otherwise} \end{cases}$$

- **5. Continuous Editing** We will improve the algorithm (Hertz et al. 2022) for efficient continuous editing performance by saving all the latent edited attention maps $\widehat{M_t}$ for prompt $\mathcal{P}^*$. When the target image $z_0^*$ serves as a source image of another continuous editing, we input $\mathcal{P}^*$ and pre-saved $\widehat{M_t}$ into the algorithm to achieve efficient computation without repeat. We can also leverage this method in other scenarios, like combing two images to reduce redundant computation.

## Results

**Multi-seed Image Editing** We tested multi-seed image editing with the application of target replacement. The results are displayed in Figure 1. We replaced squirrel with lion, and used the same target prompts with three random seeds. We can observe that the algorithm returned lions in different styles while maintaining the source image's overall composition and structure, which indicates our improvement's initial success.

However, we can also notice that the styles of burgers and backgrounds also changed. In order to fix this problem, we ever tried to apply the mask to the initialization of the initial latent image to retain the unchanging area. However, in that method, the outcome was unrecognizable. Therefore, we resorted to applying the mask to intermediate latent images instead of the initial noisy image. As displayed in Figure 2, the results of applying the mask show that the styles of background and burgers almost remained the same, demonstrating that our conception is practical.

**Remove Object** We perform this feature in two different styles: target replacement and global modification. The results are displayed in Figure 3 and Figure 4. The top right one is considered from global modification, and others are traditional target replacements.

Except for the example of "A kid riding a bicycle with a dog", our remove method does a great job on multiple text prompts. The algorithm does not modify irrelevant objects in their attention maps.

**Combining Two Images** The results are displayed in Figure 5. The source prompts are "A woman playing soccer" and "A man playing basketball". The target prompt "A woman playing basketball" needs to include the "woman" object from $\mathcal{P}$ and the "basketball" object from $\mathcal{P}'$. In this way, we call $\mathcal{P}^*$ is a combination of $\mathcal{P}$ and $\mathcal{P}'$. The target image is generally good except for the basketball in the middle.

**Moving Object** The results of moving objects are shown in Figure 6, 7 and 8. In our practical experiment, our initially proposed method which merely moves the attention map in order to move an object didn't work. In that way, the object in the edited image was still in its original place. Therefore, we resort to a new method that moves the directory target in latent images. We divided our process into two steps.

In the first step, we copy the object and move the object based on the designated direction and distance. We applied a mask generated from the attention map of the target object and copied the corresponding pixels of latent images to the destination. By this technique, we guarantee that the new object is the same as the original object. As shown in Figure 6, 7 and 8, after moving, the house, the strawberry and the dog almost remain the same.

For the second step, after copying, we need to erase the original object. There are two fashions to erase the original object. The first fashion employs the aforementioned Removing Object technique, which reduces the weight of the attention map corresponding to the object. As demonstrated in Figure 6, the attention map of the house in the edited image becomes pure noise, resulting in the vanishment of the original house. The second fashion to make the original object disappear is to override the object pixels with the environment pixels in the destination during the diffusion process. In other words, we switch the pixels of latent images between the original object and the environment in the destination. As displayed in Figure 7 and 8, the original place of the strawberry is filled with ground, and the original place of the dog is filled with grass. Also, compared to the first fashion, we can also move the attention map to fine-tune the edited images to make them more natural, as shown in Figure 7 and 8. The biggest advantage of the second method is that for a simple environment, the effect is very stable and satisfactory.

## Drawbacks

**Combining Two Images** The input prompts of the original algorithm need to have the same syntax. Only corresponding words difference is allowed. In this way, we can easily operate the token dimension of the attention map, such as inserting and removing.

For instance, in the experiment from Figure 5, three prompts have to have the same syntax "A XX playing XX". If $\mathcal{P}'$ is changed to "The man from Germany plays soccer", our algorithm cannot handle this case. If we just test the basic replacement feature from the original paper, the same syntax of two prompts is also required.

Consequently, one of the future works is to develop a new attention maps operation algorithm that gets rid of this syntax limitation.


Figure 1: The results of multi-seed image editing. We change the squirrel to a lion with the same target prompts but different random seeds. Attention control is employed to try to make other parts unchanged. The overall composition of these edited images almost remains the same.


Figure 2: The results of multi-seed image editing with mask. We change the squirrel to a lion, with the same target prompts but different random seeds. Attention control as well as latent image masking is employed to try to make other parts unchanged. The other elements of these edited images almost remain identical.

**Moving Object** For the two aforementioned fashions to erase the original objects, there are some defects. For the first fashion using the Removing Object technique, the disadvantage is its output effect is not stable. In order to achieve the best erasing effect, a lot of parameter adjustment work is required. For the second fashion, the drawback is it is only suitable for simple background environments. For complex background environments, the effect is not ideal, there are obvious erasing marks, and it is out of place with the background.

## Github Repository

https://github.com/ycpeng8/improvement_p2p

## References

Amir, H.; Ron, M.; Jay, T.; Kfir, A.; Yael, P.; and Daniel, C.-O. 2022. Prompt-to-Prompt Image Editing with Cross Attention Control. *arXiv preprint arXiv:2208.01626v1*.

Bar-Tal, O.; Ofri-Amar, D.; Fridman, R.; Kasten, Y.; and Dekel, T. 2022. Text2LIVE: Text-Driven Layered Image and Video Editing. *arXiv preprint arXiv:2204.02491*.

Bau, D.; Andonian, A.; Cui, A.; Park, Y.; Jahanian, A.; Oliva, A.; and Torralba, A. 2021. Paint by word. *arXiv preprint arXiv:2103.10951*.

Brock, A.; Donahue, J.; and Simonyan, K. 2018. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*.

Chaerin, K.; DongHyeon, J.; Ohjoon, K.; and Nojun, K. 2022. Leveraging Off-the-shelf Diffusion Model for Multi-attribute Fashion Image Manipulation. *arXiv preprint arXiv:2210.05872v1*.

Chen, W., Henry; and Fernando, l. T., De. 2022. UNIFYING DIFFUSION MODELS' LATENT SPACE, WITH APPLICATIONS TO CYCLEDIFFUSION AND GUIDANCE. *arXiv preprint arXiv:2210.05559*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y.

*"A kid riding a bicycle with a dog"* Remove *"dog"*

*"Tofu soup with croutons"* Remove *"Tofu"*

Figure 3: The results of removing object experiments. Except for the top one, our remove method does a great job on multiple text prompts. The algorithm does not modify irrelevant objects in their attention maps.

*"A fantasy landscape with a pine forest"* Remove *"fantasy"*

*"A photo of a house on a mountain"* Remove *"house"*

Figure 4: The results of removing object experiments. The top one is considered from global modification, and the other is traditional target replacements. The algorithm does not modify irrelevant objects in their attention maps.


2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Nataniel; Yuanzhen, L.; Varun, J.; Yael, P.; Michael, R.; and Kfir, A. 2022. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. *arXiv preprint arXiv:2208.12242v1*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; Salimans, T.; Ho, J.; Fleet, D. J.; and Norouzi, M. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding.

Wenhu, C.; Hexiang, H.; Chitwan, S.; and William, W. C. 2022. RE-IMAGEN: RETRIEVAL-AUGMENTED TEXT-TO-IMAGE GENERATOR. *arXiv preprint arXiv:2209.14491v3*.

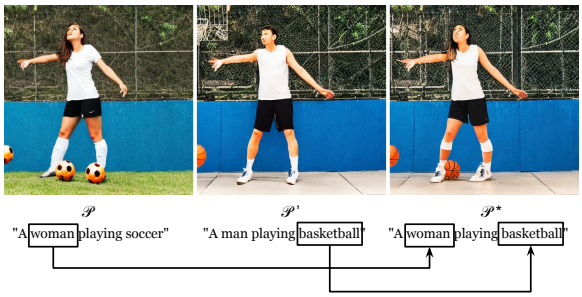


Figure 5: The results of combining two images experiments. The source prompts are “A woman playing soccer” and ”A man playing basketball”. The target prompt ”A woman playing basketball” needs to include the ”woman” object from $\mathcal{P}$ and the ”basketball” object from $\mathcal{P}'$. In this way, we call $\mathcal{P}^*$ is a combination of $\mathcal{P}$ and $\mathcal{P}'$. The target image is generally good except for the basketball in the middle.

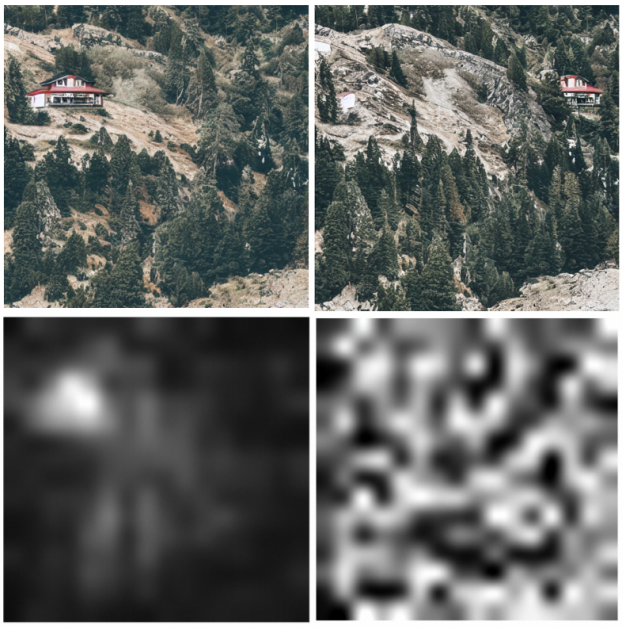

Figure 6: The result of moving house to the right from the image generated by “A photo of a house on a mountain”. First, we copy the house to the right in latent images during the diffusion process. Secondly, we erase the original house by reducing the weight of the attention map of the house.

Figure 7: The result of moving the strawberry to the right from the image generated by “A strawberry on the ground”. First, we copy the strawberry to the right in latent images during the diffusion process. Secondly, we erase the original strawberry by overriding it with the environment in the destination in latent images. Then we move the attention map of the strawberry to fine-tune the edited output.

Figure 8: The result of moving the dog to the right from the image generated by “A dog in the forest”. First, we copy the dog to the right in latent images during the diffusion process. Secondly, we erase the original dog by overriding it with the environment in the destination in latent images. Then we move the attention map of the dog to fine-tune the edited output.