Improvement on Prompt-to-Prompt Image Editing

Yanchong Peng, Yuhe Peng, Peng Huang

{ycpeng, pengyh, phuang}@bu.edu

Previous Work

Recently, many techniques have been proposed in the field of 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.

Bau(Bau et al. 2021) showed that the mask provided by the user can help edit locally the text-based image. This mask contains all the areas relevant to the text that the user wants to change, whereas it requires lots of labor and time, which is obviously inefficient. Seminal works combining GAN(Goodfellow et al. 2020; Brock, Donahue, and Simonyan 2018), which are mostly used to generate high-resolution images, and CLIP(Radford et al. 2021), which contains rich image-text representation, don't require extra labor annotations. Yet, they cannot handle huge and varying datasets.

Text2LIVE(Bar-Tal et al. 2022) also has the ability to 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 unchanged sections of the origin but also modifies the part that the user wants to replace. In this way, we just need the text input and don't require any extra labor. We will produce some innovative features and applications based on his work.

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 not possible 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 doesn't have the functionality of removing a specific target. Moreover, the method cannot be used to change an object's location in the picture. Finally, the algorithm for sequential image editing is not efficient.

So, in this project we expect to make some improvements on 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 together. (Move the car generated by random seed s into the street generated by seed s^*).
- Move a specific object across the image.
- Improve the algorithm performance in continuous editing.

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^*.

Dataset

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

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

This model used 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.

Method

- **1. Multi-seeds Image Editing.** The input of our project will have one more random seed compared with the original algorithm: one source random seed s and one target random seed s^*. We inject $s*$ into the production of z^*_{t-1} and M^*_t; That is, the computation of the diffusion process will be $DM(z^*_t, \mathcal{P}^*, t, s^*)$, instead of $DM(z^*_t, \mathcal{P}^*, t, s)$(Hertz et al. 2022).
- **2. Removing Specific Target.** 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 just remove the car, let the attention map of

the token "car" as j^*. Then, the maps-editing operation function will be modified as

$$(Edit(M_t, M_t^*, t))_{i,j} := \begin{cases} 0 & \text{if } j{=}j^* \\ (M_t)_{i,j} & \text{otherwise} \end{cases}$$

Note that this formula is the first version of our study for this project. We may change it when diving deeper.

- **3. Combining Two Images.** 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 $\mathcal{P}^*$, a combination of $\mathcal{P}$ 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

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

where j, j', and j^* are the text tokens for M_t, M_t', and M_t^* here. In fact, by using a similar way, we can extend the combination from two images to multiple images.

- **4. Moving Object.** We proposed a method of moving an object across an image, which originally serves as a limitation in the paper by Hertz et al. 2022. For the purpose of 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.

Schedule

We expect to finish **Multi-seeds Image Editing** and **Removing Specific Target** before the milestone.

As for the final report and presentation, our target is implementing at least two of these topics:**Combining Two Images**, **Moving Object** and **Continuous Editing**. If we have enough time, completing all of them is expected.

Evaluation Criteria

The outputs of this project will be evaluated from these three different text-modification styles:

- **Target replacement** The object will be different. For example, changing "Cat rides on a bike" to "Dog rides on a bike". This test passes if other parts of the image except the changed object remain the same.
- **Global modification** The drawing style of the image will be changed. For example, changing "a charcoal pencil sketch of valley landscape" to "a van Gogh painting of valley landscape". This test passes only if the global style is changed and all objects are preserved in the same position and shape.
- **Token attention control** We only add or delete one factor to the original image. For example, the source text prompt is "a fantasy valley when the sun rises". We input the text "+fog" or "-fog" to add or remove the fog effect of the source image.

References

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*.

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

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.