Supplementary: LDEdit: Towards Generalized Text Guided Image Manipulation via Latent Diffusion Models

Paramanand Chandramouli
paramanand.chandramouli[at]uni-siegen.de
Kanchana Vaishnavi Gandikota
kanchana.gandikota[at]uni-siegen.de

Department of Computer Science
University of Siegen
Germany

1 Additional Experiments

While the latent diffusion model [3] is trained for generating images of dimension 256×256, due to fully convolutional nature of the autoencoder, our method can be applied on images of higher resolution using the same model. Fig. 1 shows further example results of image manipulation using LDEdit, with image resolution 512×512. It is seen that our method can achieve varied transformations in a straightforward way. The first two rows shows simultaneous manipulation of the girl and the ball. The third row shows style transfer to a painting or a photo and manipulating the age of two girls. Interestingly, LDEdit can effect such transformations with a little or no stochasticity, such that the background remains largely unaffected. The final row shows manipulating a horse to other species, *e.g.* a zebra, a donkey, a bear, and a wolf. These transformations required a higher η of 0.3 for zebra and donkey, and η of 0.8 for bear and wolf. However, higher values of η result in more changes in the background.

Comparisons with VQGAN+CLIP [1] We provide more visual comparisons with VQGAN+CLIP for general text driven image manipulation in Fig. 2. While VQGAN+CLIP can successfully effect changes in the input image of a building as per the target text prompts, its performance suffers in more difficult manipulations such as translating from an input stroke, or performing simultaneous local manipulations. In contrast, our LDEdit is able to perform these desired manipulations.

Failure cases: In some cases, our method may fail to produce desired manipulations as seen in Fig. 3. With an input text prompt of 'a deer with antlers', we obtain manipulated images where the antlers are misplaced. These undesired effects can be avoided by using a mask, which can aid in localization of edits.

2 Effect of Stochasticity

In our approach, we proposed to perform a deterministic DDIM sampling, to ensure that a consistency is maintained with the original image. However, when the input image lacks

Figure 1: Results of image manipulation using LDEdit

details, such as a stroke image, doing a deterministic forward produces a latent code which lacks any details, see Fig. 4 a). On the other hand, introduction of stochasticity through η can aid in hallucinating details not present in the original image, Fig. 4 b). With $\eta = 1$, DDIM becomes equivalent to DDPM sampling, which results in more diverse samples. Note that our method may sometimes result in images with text like artifacts, as seen in Fig. 4 c). More example image manipulation of LDEdit by varying η are shown in Fig. 5. As the value of η increases, the diversity of samples improves. However, there are more perceptible changes in background, see rows 1 and 2 of Fig. 5.

3 User study

We conduct user studies to compare user preference of image manipulation results of our method with VQGAN+CLIP [1] and DiffusionCLIP [2]. Users participated in two surveys, where they were provided with source image, target text description and the results obtained with LDEdit and base-line method (VQGAN+CLIP or DiffusionCLIP) in a random order, and voted their preferred image manipulation using a survey platform. We obtained a total of 1120

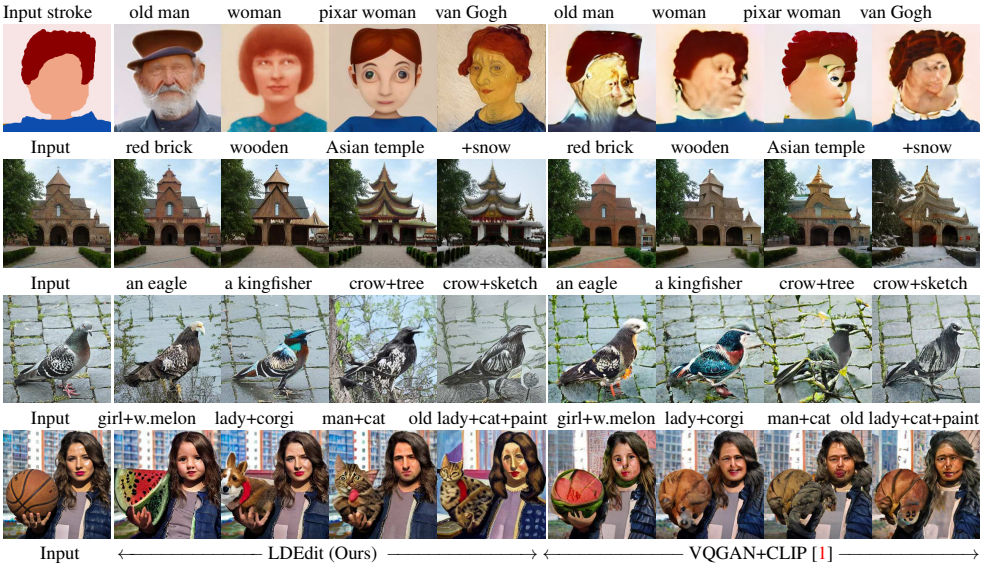


Figure 2: Comparison of LDEdit with VQGAN+CLIP [1]. Best results out of 4 samples are shown for LDEdit and the best result out of 8 samples are shown for VQGAN+CLIP.

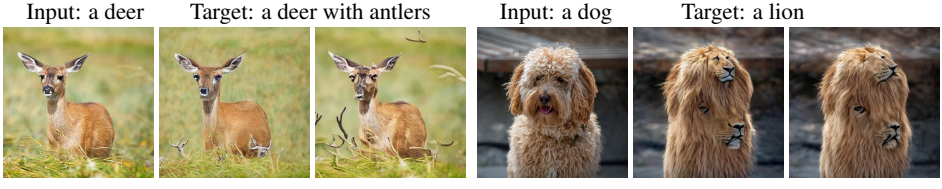
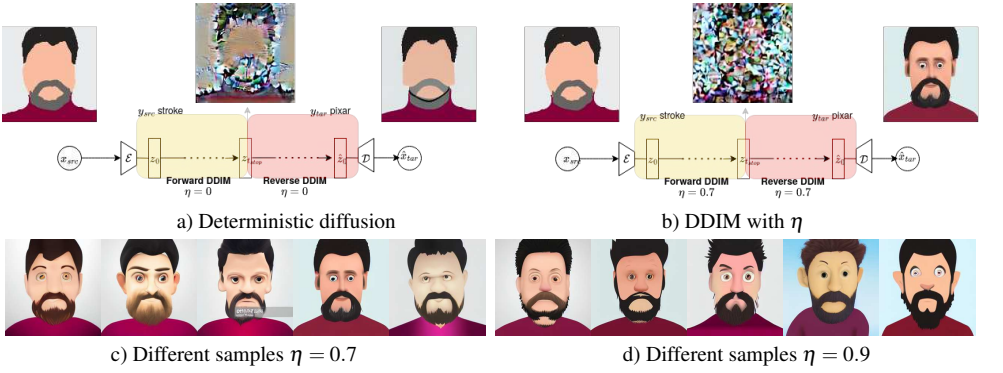


Figure 3: Failure cases of image manipulation using LDEdit

votes from 32 participants for comparing LDEdit with VQGAN+CLIP and 950 votes from 38 participants for comparing LDEdit with DiffusionCLIP. For comparison with both the baselines, we included a combination of face images and general images (on manipulations demonstrated in DiffusionCLIP [2] paper). On faces, manipulated attributes include makeup, tanned, curly hair, changing gender, domain change to zombie, neanderthal. We also include an example of translating stroke image to pixar, neanderthal and van Gogh painting. On general image manipulation, we include manipulating an input building, bus, dog and a tennis ball. Additionally, for comparison with VQGAN+CLIP, we include examples of manipulating an image of a bird and multiple local object manipulations. In human evaluation, the results of LDEdit were preferred 83.87% of the time in the survey comparing LDEdit with VQGAN+CLIP, whereas user preference for LDEdit is 49.15% in the survey comparing LDEdit with DiffusionCLIP.



a) Deterministic diffusion b) DDIM with η

c) Different samples $\eta = 0.7$ d) Different samples $\eta = 0.9$

Figure 4: Effect of η in diffusion process. Purely deterministic DDIM process cannot achieve desired target when the original input lacks details.

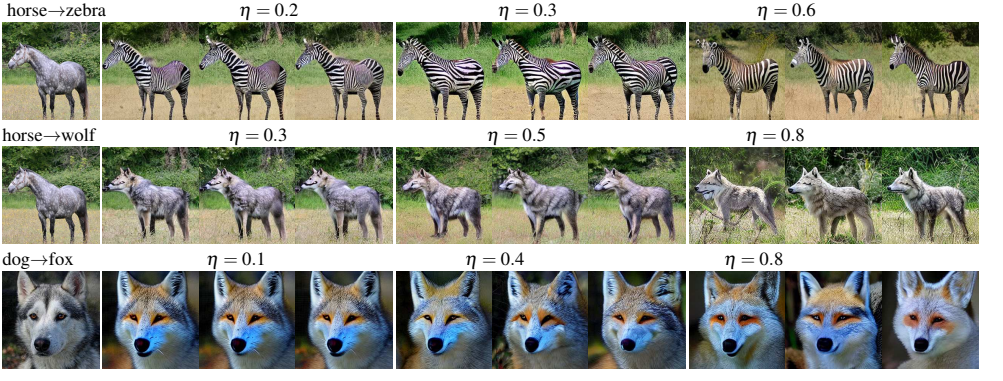


Figure 5: Sample results for different η using LDEdit. As the value of η increases, the diversity of samples increases.

References

[1] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, 2022.

[2] Gwanghyun Kim and Jong Chul Ye. Diffusionclip: Text-guided image manipulation using diffusion models. *arXiv preprint arXiv:2110.02711*, 2021.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.