Prompt Tuning Inversion for Text-Driven Image Editing Using Diffusion Models Supplementary Material

Wenkai Dong[1] Song Xue[1] Xiaoyue Duan[1, 2*] Shumin Han[1†]
[1]Baidu VIS, [2]Beihang University
{dongwenkai, xuesong06, duanxiaoyue, hanshumin}@baidu.com

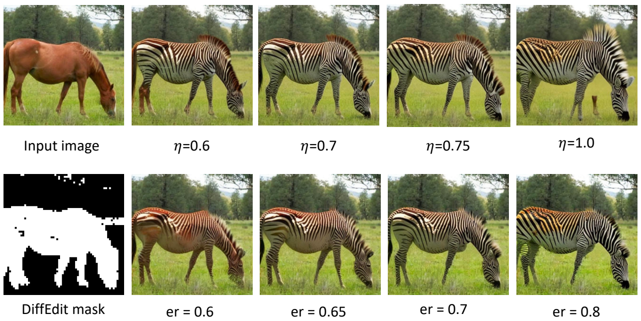


Figure 1. Comparison of controlling by different DDIM encoding ratios and condition interpolation ratios. "er" denotes the DDIM encoding ratio

1. Encoding ratio vs. condition interpolation ratios.

As stated in Sec 4.2 in the main manuscript, different editing methods often have hyper-parameters to control the trade-off between fidelity and editability. Besides the masking threshold used in the quantitative experiments, DDIM encoding ratio and the condition interpolation ratio can also control the trade-off. Firstly, for DiffEdit [1], we fix the mask threshold and performer image editing under different encoding ratios. As shown in Fig 1 (second line), when the encoding ratio ($er >= 0.7$), we can observe that one leg is removed even though the automatically generated mask is accurate enough for localizing the editing region. When decreasing er to 0.6 or 0.65, although the original shape is preserved well, the editability is not enough (e.g., the color is close to sorrel instead of black). Therefore, controlling er fails to achieve the desired modification.

Then we set $er = 0.8$ and vary the condition interpolation ratio η. As shown in the first line, when $\eta = 0.75$, the shape is preserved well, and the characteristics of zebras are also more significant. The qualitative result shows that condition interpolation ratio η and encoding ratio (er) control the trade-off differently, and we can obtain a better trade-off by adjusting η.

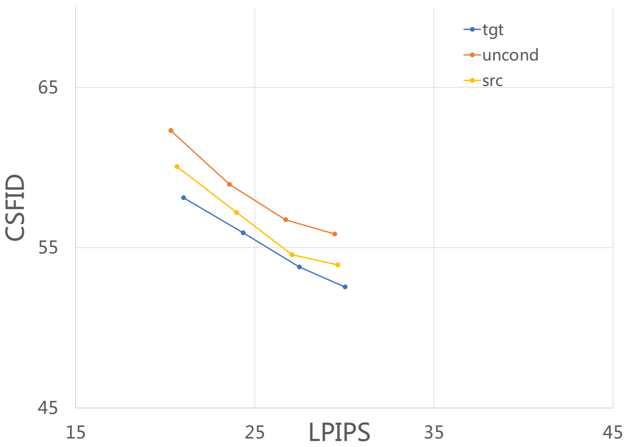


Figure 2. Trade-off with different initialization of learnable conditional embedding. "tgt", "src" and "uncond" denotes the learnable conditional embedding are initialized with target embedding, source embedding, and unconditional embedding, respectively.

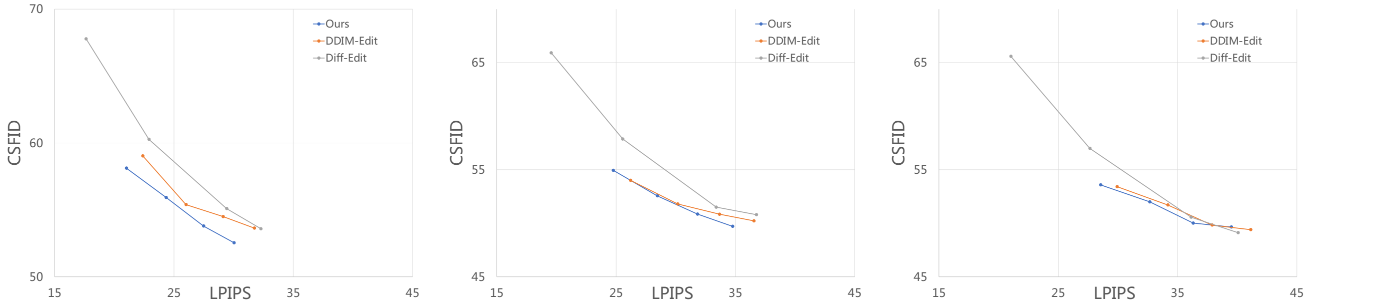
2. Optimizing conditional embeddings with different initialization

In our proposed Prompt Tuning Inversion method, we encode the information of the input image into a learnable conditional embedding via prompt tuning in the reconstruction process. We initialize the learnable embedding in different ways and the experimental results are listed in Fig. 2. We can observe that using target embedding results in the best trade-off. We conjure that this can make the optimized conditional embedding close to the target embedding in the semantic space, which benefits the linear interpolation between them.

3. Ablation on the DDIM encoding ratio

In this section, we test the performance of three methods under different encoding ratios. Fig. 3 shows the compar-

[*] This work was done when Xiaoyue Duan was an intern at Baidu VIS.
[†] Corresponding author

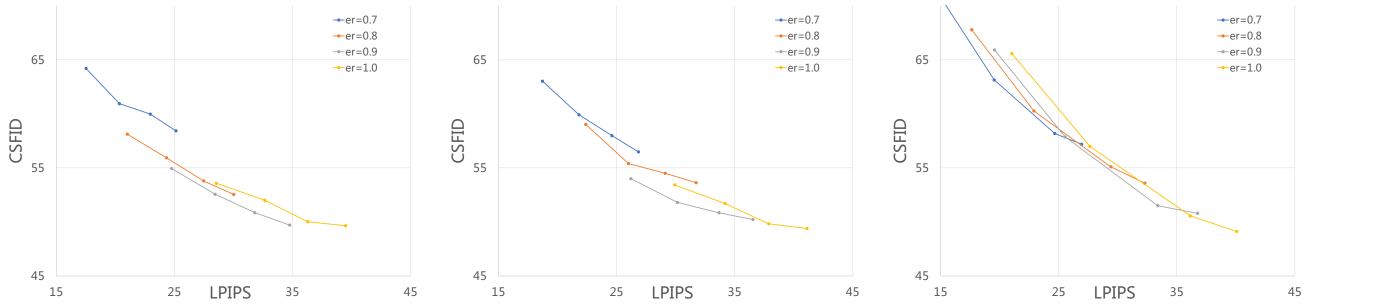


(a) encoding ratio = 0.8

(b) encoding ratio = 0.9

(c) encoding ratio = 1.0

Figure 3. Comparison under different DDIM encoding ratios.



(a) Ours

(b) DDIM-Edit

(c) DiffEdit

Figure 4. Trade-off under different DDIM encoding ratios for each method.

ison of different methods under the same encoding ratio. Fig. 4 shows the comparison of the same method under different encoding ratios.

4. Comparison of reconstruction quality.

As stated in the main manuscript, when using DDIM inversion, enlarging the classifier-free guidance scale ω_{dec} leads to the problem that the reconstructed images are far from the original ones. To examine the effectiveness of our method, we provide a similar table below for DDIM, NTI [3], and PTI in terms of reconstruction quality. $\omega_{enc} = 0$ for all methods. PTI achieves descent PSNR under large guidance scales ω_{dec}, which further strengthens our method.

ω_{dec} Method	0.0	1.0	2.5	5.0	7.5
DDIM	21.36	19.79	17.04	14.88	13.64
NTI	25.83	-	23.95	23.74	23.08
PTI	-	**25.83**	**25.70**	**25.31**	**24.74**

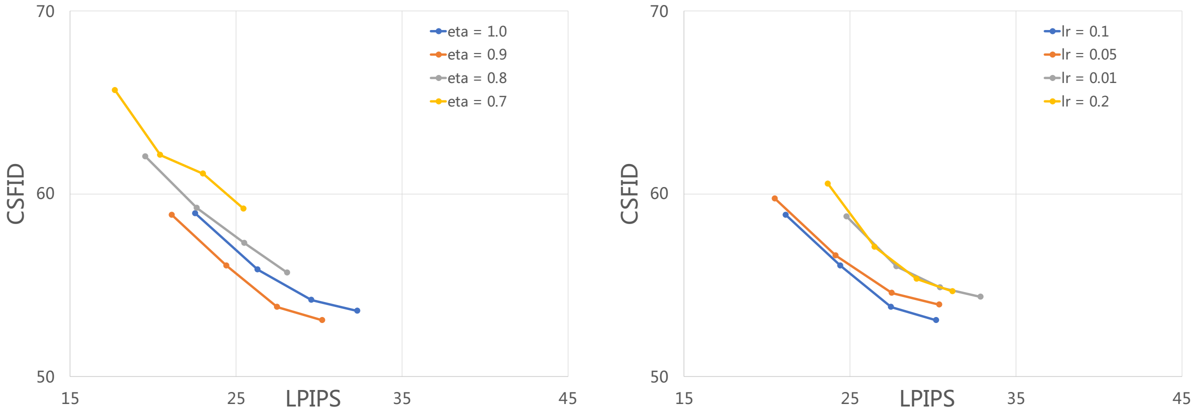
Table 1. Reconstruction quality by measuring the PSNR score of DDIM inversion with different classifier-guidance scales ω. ω_{dec} denote the guidance scale used in the sampling processes

5. Ablation on other hyper-parameters.

We also perform ablation on two core components of our method, *i.e.*, the interpolation ratio η and the learning rate β in PTI, to measure their influence in terms of CSFID-LPIPS on ImageNet. When $\eta = 1$ or $\beta = 0$, our method reverts to the baseline method DDIM-Edit. The left panel of Fig. 5 shows decreasing η from 1.0 to 0.9 leads to a better CSFID-LPIPS trade-off but lower ratios result in a worse balance between editability and fidelity. When we fix η as 0.9 and decrease lr from 0.1 to 0.05 or 0.01, the trade-off also becomes worse.

6. More qualitative examples.

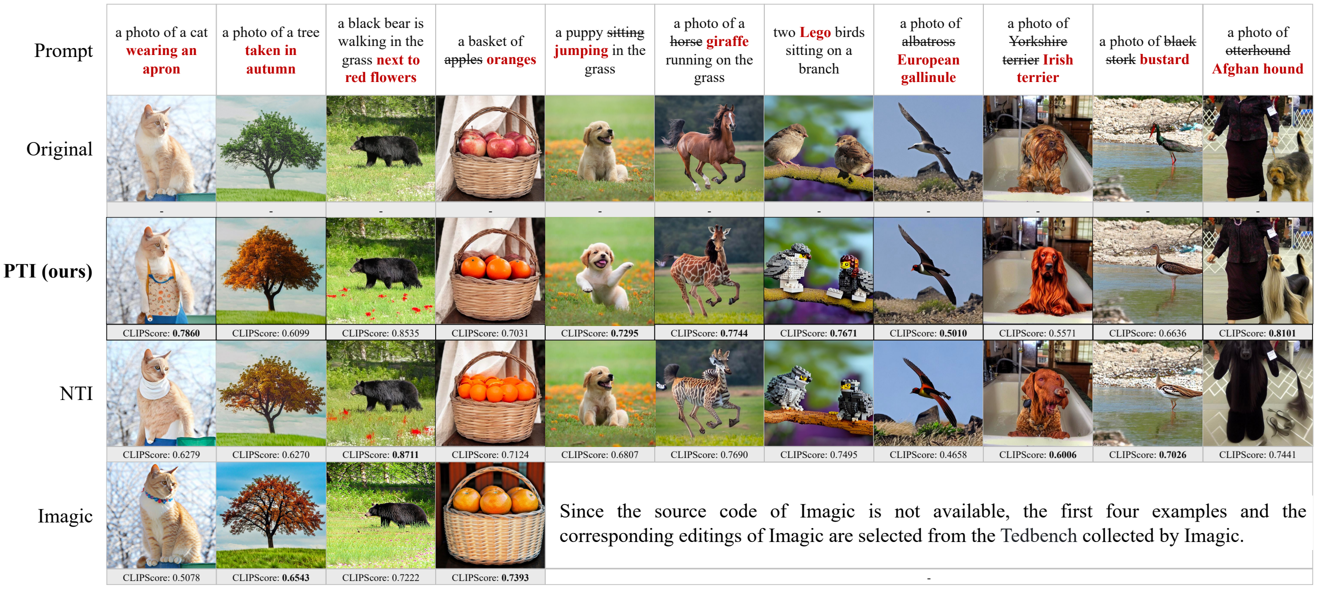
More challenging examples are provided in Fig. 6, *e.g.*, changing sitting dog to jumping (col. 5), changing horse to giraffe (col. 6), or changing style (col. 2). PTI achieves impressive editing on cases where NTI fails (*e.g.*, cols. 5 and 11). PTI also preserves the background better in some cases (*e.g.*, NTI mistakenly turns the branch into Lego in col. 7 while PTI does not). However, not using attention is not always an advantage: the proposed PTI preserves less structure than NTI. For example, the shape of the tree in the col. 2, the face directions of the horse.



(a) encoding ratio = 0.8

(b) encoding ratio = 0.9

Figure 5. **Left**: ablation on the interpolation ratio η. **Right**: ablation on the learning rate lr (*i.e.*, β) in Prompt Tuning Inversion.



Since the source code of Imagic is not available, the first four examples and the corresponding editings of Imagic are selected from the Tedbench collected by Imagic.

Figure 6. Visual comparisons with Null-text (NTI) [3] and Imagic [2]. In the figure above, CLIPScore is provided under each image.

References

[1] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 1

[2] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 3

[3] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 2, 3