Perceptual Similarity guidance and text guidance optimization for Editing Real Images using Guided Diffusion Models

Ruichen Zhang

the School of Mechanical, Electrical and Infor- mation Engineering, Shandong University

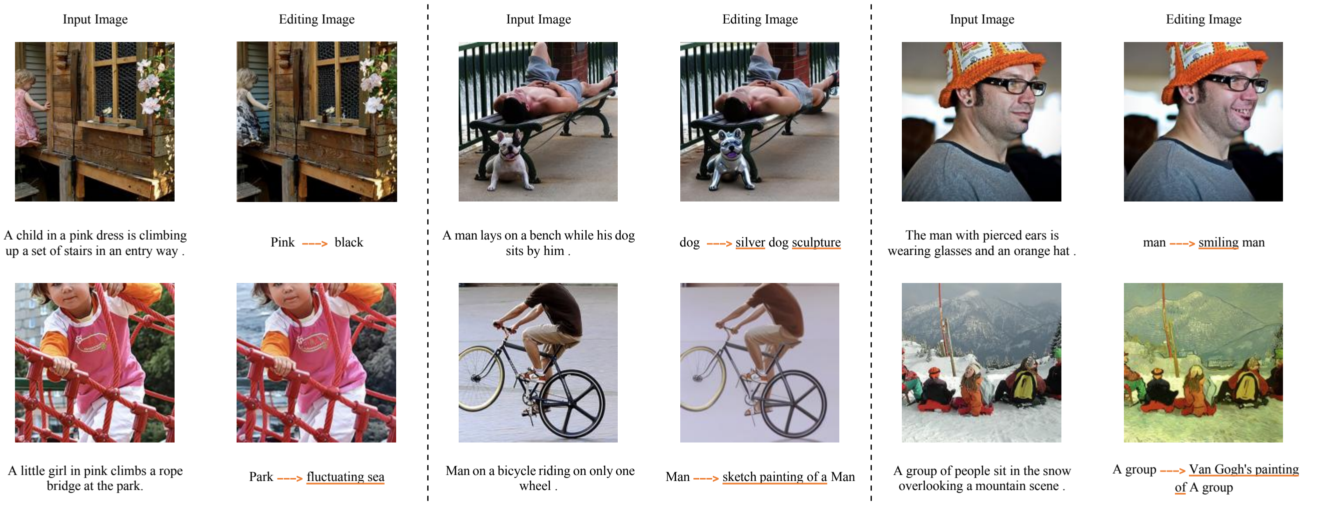
Input Image
Editing Image
Input Image
Editing Image
Input Image
Editing Image
A child in a pink dress is climbing up a set of stairs in an entry way .
Pink ---> black
A man lays on a bench while his dog sits by him .
dog ---> silver dog sculpture
The man with pierced ears is wearing glasses and an orange hat .
man ---> smiling man
A little girl in pink climbs a rope bridge at the park.
Park ---> fluctuating sea
Man on a bicycle riding on only one wheel .
Man ---> sketch painting of a Man
A group of people sit in the snow overlooking a mountain scene .
A group ---> Van Gogh's painting of A group

Figure 1. **Perceptual Similarity guidance and text guidance optimization for real image editing.** Our method edits the picture by taking the real images with the image caption prompt and the edited prompt as input.

Abstract

During the application of the diffusion model for image editing, instances arise where the edited image deviates significantly from the original. In response, efforts are made to guide the image towards greater similarity with the original during the editing process. In this paper, we introduce a two-part guidance to force the generated image to preserve the details in the untouched areas of the original image with high fidelity. (I) text-guided optimization, where we use the conditional text embeddings for latent space directional guidance, and the unconditional textual embedding for classifier-free guidance. (II) Perceptual similarity guidance, where we optimize the latents during reverse using posterior sampling with Tweedie ' s formula. We can attain exceptionally realistic image editing by not only ensuring the accurate generation of edited components but also preserving the high-quality details of the original image.

1. Introduction

The text-guided stable diffusion model [1] has garnered significant attention due to its authenticity and precision. Recent studies, employing classifier-free guidance and DDIM inversion [2], offer viable approaches for real image editing. Nevertheless, there is room for improvement in the image guidance generation process to enhance the method's ability to bring the generated image closer to the given prompt and the original image.

Whether it involves editing the image by blending forward and backward images with an adjustable mixing ratio [12], employing masks for local modifications [13], or preserving specific low-frequency information [14], a common issue arises with the relatively limited applicability of these approaches. In classifier-free guidance, excessive emphasis is often placed on distinguishing between the original prompt and the unconditional prompt, neglecting effective consideration of the distance between the original prompt and the target prompt. Additionally, many methods primarily rely on prompts to initialize latent features, rather than incorporating the distance between the generated image and the real image at the image level.

In this paper, we present an effective bootstrap method aimed at enhancing the consistency of image editing with the original image. Our approach revolves around guided calculations involving two pivotal facets of the guided diffusion model: Perceptual Similarity [3] guidance and optimization of text guidance.

In the optimization of text guidance, three predictions are generated at each diffusion step: one using unconditional text embeddings, another with the image caption prompt as the condition, and a third with an editing text prompt as the condition. We applied weights to these three predictions to amplify the impact of edited text on the image while preserving similarity with the real image.

In Perceptual Similarity guidance, we conduct a posterior sample to estimate P(Z_0 | Z_t) utilizing Tweedie's formula. At each reverse step, we decode the latent variables back into the image domain and employ Perceptual Similarity to compute the pixel-level difference between the edited image and the real image for guidance.

2. Related Work

The field of image editing has witnessed remarkable advancements in recent years, providing a plethora of potent tools and techniques to enhance the efficiency and creativity of image processing. This encompasses various functions such as image restoration and enhancement [16], image style conversion [17], face editing [18], super resolution [19], among others. The advent of the text-guided stable diffusion model [6,7] has further refined the landscape of image editing.

Prompt2Prompt [1] is designed for image editing within a pre-trained diffusion model by manipulating prompts. It involves injecting a cross-attention map into the diffusion model's sampling process, directing pixels to focus on the tokens of text prompts during diffusion steps to achieve image editing. However, their methodology is confined to altering generated images and does not extend to modifications of real images.

Liu et al. guided a diffusion process using both text and image, creating a composition that closely resembled the provided image and aligned with the given text [5].

Hertz et al. utilized a classifier-free guidance diffusion model to establish a robust anchor point for the reconstruction of the original image through DDIM Inversion. Furthermore, they optimized the Null-text Embedding of the classifier-free guidance diffusion model to achieve precise reconstruction of the original image [2].

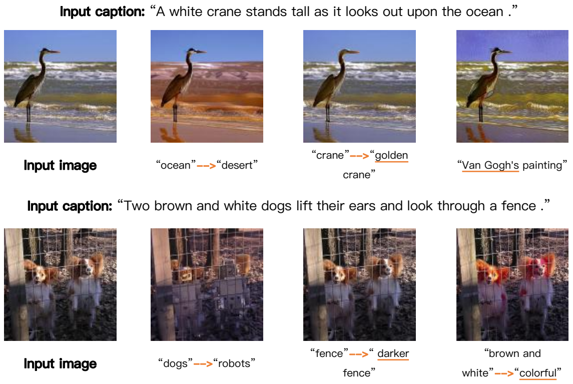
Input caption: “A white crane stands tall as it looks out upon the ocean .”
Input image
“ocean”-->“desert”
“crane”-->“golden crane”
“Van Gogh's painting”
Input caption: “Two brown and white dogs lift their ears and look through a fence .”
Input image
“dogs”-->“robots”
“fence”-->“ darker fence”
“brown and white”-->“colorful”

Figure2. **Real image editing using our method.** It can be seen that our method retains the similarity to the original image while completing the editing.

A component of our method shares similarities with classifier-free guidance. Our guidance involves making predictions three times: first, utilizing the original image prompt condition embedding; second, using the editing image prompt condition embedding; and third, unconditionally using null-text embedding. Furthermore, for additional precision, we guide the process by calculating differences in the real image domain.

3. Method

Let be an wanted image that can be described using the edited prompt $\mathcal{P}_{edi}$. Our goal is to edit the real image $\mathcal{X}src$, which can be described using the source prompt $\mathcal{P}src$ .We use the edited prompt $\mathcal{P}edi$, null-text prompt $\mathcal{P}null$, source prompt $\mathcal{P}src$, and perceptual similarity to get the edited image $\mathcal{X}edi$.

We use the Stable Diffusion model [4], where the diffusion forward process is applied to the latent space, using the image encoding $\mathcal{Z}src = \mathcal{E}(\mathcal{X}src)$ at the beginning to transfer image to latent space, and the image decoder $\mathcal{X}edi = \mathcal{D}(\mathcal{Z}edi)$ at the end of the diffusion backward process to transfer data from latent space to image.

We will apply the guidance on $\mathcal{Z}src$, the hidden space of the real image, and $\mathcal{Z}edi$, the hidden space of the edited image. In the pixel domain, it is guided by calculating the perceived similarity [3] of $\mathcal{X}src$ and $\mathcal{X}edi$.

3.1. Text guidance optimization

To achieve better text guidance in latent Spaces, we set $\mathcal{E}null = \psi("")$ be the embedding of a null text [10], $\mathcal{E}src = \psi("\mathcal{P}src")$ be the embedding of source prompt, $\mathcal{E}edi = \psi("\mathcal{P}edi")$ be the embedding of edited prompt, and let γ and β be the guidance scale parameters in Equation (1) and (2). The network $\varepsilon\theta$ is trained to predict artificial noise, and t is the step of diffusion.

We use $noisecond$ to represent the guidance between source prompt predicts noise and null text predicts noise.

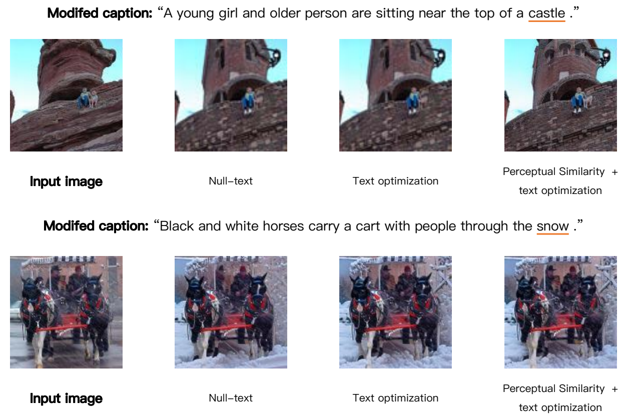
Modifed caption: "A young girl and older person are sitting near the top of a castle ."
Input image
Null-text
Text optimization
Perceptual Similarity + text optimization
Modifed caption: "Black and white horses carry a cart with people through the snow ."
Input image
Null-text
Text optimization
Perceptual Similarity + text optimization

Figure3. **Real image editing under different guidance.** As evident, despite the minimal distinction between Text optimization image and null-text image(given the utilization of our guidance solely in the initial 20 diffusion steps), the Perceptual Similarity image preserves more details with the original.

We use noisepred to represent the final guidance noise of this step's image editing. The guidance prediction is defined by:

$$noisecond = \varepsilon\theta(Zt, Enull, t) + \gamma * [\varepsilon\theta(Zt, Esrc, t) - \varepsilon\theta(Zt, Enull, t)] \quad (1)$$

$$noisepred = noisecond + \beta * [\varepsilon\theta(Zt, Eedi, t) - \varepsilon\theta(Zt, Esrc, t)] \quad (2)$$

The guidance prediction can also defined by:

$$noisecond = \varepsilon\theta(Zt, Enull, t) + \gamma * [\varepsilon\theta(Zt, Esrc, t) - \varepsilon\theta(Zt, Enull, t)] \quad (3)$$

$$noisepred = noisecond + \beta * [\varepsilon\theta(Zt, Eedi, t) - \varepsilon\theta(Zt, Enull, t)] \quad (4)$$

Then we can use the $noisepred$, step t, and latent Zt at step t to calculate the latent $Zt\text{-}1$ at the next step $t\text{-}1$.

3.2 Perceptual Similarity guidance

Now that we have both Zt and $noisepred$, we can roughly calculate the Zo of the edit image latent space at t= 0 :

In stable diffusion [4], $\bar{\alpha}_t$ is the fixed diffusion rate at time t, and we have:

$$Zt = \sqrt{\bar{\alpha}_t} Zo + \sqrt{1 - \bar{\alpha}_t}\, noisepred, \quad noisepred \sim \mathcal{N}(0,1)$$

Thus,we can get Zo by:

$$Zo = \frac{Zt - \sqrt{1-\alpha_t}\ noisepred}{\sqrt{\alpha_t}}, \quad noisepred \sim \mathcal{N}(0,1)$$

In order to solve the problem that guidance at the latent space is not accurate enough, we convert the image from the latent space Zo back to the image domain Xo at every step t.

Then we calculate Perceptual Similarity between the real image $Xsrc$ and our Xo to optimize Zt in the latent space. Perceptual Similarity is used to evaluate the distance between image patches. Lower means more similar.

4. Experiment

Our method is designed to focus on intuitive editing, and we gauge fidelity to the original image using LPIPS perceptual distance [3] (lower values are preferable) and PSPR (higher values are preferable). Additionally, we assess fidelity to the target text using CLIP similarity [11] (higher values are preferable) on the Flickr8k Dataset.

	PSNR	LPIPS	CLIPScore
Null-text	20.2558	0.1857	28.5781
Text optimization	20.0173	0.1919	30.4297
Perceptual Similarity + text optimization	20.1249	0.1894	30.5703

CLIPScore is employed to assess the alignment between the text and the generated image. Notably, the combination of Perceptual Similarity and text optimization exhibits a significant improvement in CLIPScore, indicating a better match between our images and the descriptive text.

PSNR (Peak Signal-to-Noise Ratio) is utilized to quantify the disparity between the original image and the processed image. The PSNR values across all three experimental groups hover around 20, suggesting that discernible differences are perceptible to the human eye.

A lower LPIPS (Learned Perceptual Image Patch Similarity)value signifies greater overall similarity between two images. The comparable LPIPS values in the three experimental groups suggest minimal overall differences among the three sets of images.

Furthermore, we conducted evaluations with several individuals, and the results revealed distinctions in details among the three sets of images, with Perceptual Similarity + text optimization emerging as the most effective in preserving the details of the original image.

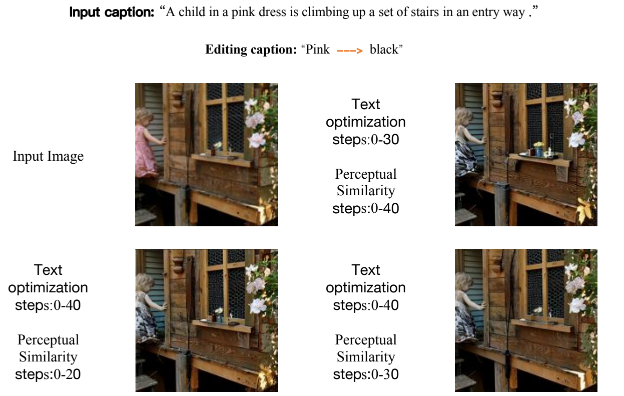
Input caption: "A child in a pink dress is climbing up a set of stairs in an entry way ."
Editing caption: "Pink ---> black"
Input Image
Text optimization steps:0-30
Perceptual Similarity steps:0-40
Text optimization steps:0-40
Perceptual Similarity steps:0-20
Text optimization steps:0-40
Perceptual Similarity steps:0-30

Figure4. **Use our guidance on different diffusion steps.** Evidently, since the two guidance methods are applied in distinct step ranges, there is a notable variation in the degree of detail restoration from the original image. To illustrate, the restoration of light and shadow in the lower right corner of the example image is excessive in the image domain, resulting in an error where the light and shadow are restored as flowers.

5. Limitations

While our method consistently yields satisfactory editing results, there remain areas open for improvement. Our guidance in the image domain relies on a direct comparison between the original and generated images. However, since this comparison involves two complete images, extensive editing may occasionally result in distortions. Although many cases can be rectified by adjusting the image size through guided steps, for future enhancements, we may explore a more focused comparison that highlights unplanned changes in the image.Moreover, additional limitations stem from the utilization of Stable Diffusion and Prompt-to-Prompt editing. In certain instances, the text attention map fails to align accurately with the corresponding area in the image. Put differently, the region of the image intended to be altered by our editing text is occasionally misselected. This observation underscores a potential pathway for refining our methodology.

References

[1] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman,Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt im-age editing with cross attention control.arXiv preprintarXiv:2208.01626, 2022.

[2] Mokady, Ron; Hertz, Amir; Aberman, Kfir; Pritch, Yael; Cohen-Or, Daniel. (2022). Null-text Inversion for Editing Real Images using Guided Diffusion Models. arXiv preprint, arXiv:2211.09794.

[3] Zhang, Richard, Isola, Phillip, Efros, Alexei A., Shechtman, Eli, & Wang, Oliver. (2018). The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. arXiv preprint arXiv:1801.03924.

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz,Patrick Esser, and Bj ¨orn Ommer. High-resolution image syn-thesis with latent diffusion models, 2021.

[5] Bahjat Kawar, Roy Ganz, and Michael Elad. Enhancing diffusion-based image synthesis with robust classifier guidance. arXiv preprint arXiv:2208.08664, 2022.

[6] Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer,Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval.arXivpreprint arXiv:2204.02849, 2022.

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

[8] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Hui-TangChang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. ArXiv, abs/2210.09276, 2022.

[9] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuningan image generation model on a single image. arXiv preprint arXiv:2210.09477, 2022.

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusionguidance. InNeurIPS 2021 Workshop on Deep GenerativeModels and Downstream Applications, 2021.

[11] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shecht-man, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric.2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages586–595, 2018.

[12] Choi, J., Kim, S., Jeong, Y., Gwon, Y., & Yoon, S. (2021). ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models. arXiv preprint arXiv:2108.02938.

[13] Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). RePaint: Inpainting using Denoising Diffusion Probabilistic Models. arXiv preprint arXiv:2201.09865.

[14] Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., & Ermon, S. (2022). SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. arXiv preprint arXiv:2108.01073.

[15] Ho, J., & Salimans, T. (2022). Classifier-Free Diffusion Guidance. arXiv preprint arXiv:2207.12598.

[16] Yang, C., Jin, M., Jia, X., Xu, Y., & Chen, Y. (2022). AdaInt: Learning Adaptive Intervals for 3D Lookup Tables on Real-time Image Enhancement. arXiv preprint arXiv:2204.13983.

[17] Park, J., Kim, S., Kim, S., Cho, S., Yoo, J., Uh, Y., & Kim, S. (2023). LANIT: Language-Driven Image-to-Image Translation for Unlabeled Data. arXiv preprint arXiv:2208.14889.

[18] Gao, Y., Wei, F., Bao, J., Gu, S., Chen, D., Wen, F., & Lian, Z. (2021). High-Fidelity and Arbitrary Face Editing. arXiv preprint arXiv:2103.15814.

[19] Liang, J., Zeng, H., & Zhang, L. (2022). Details or Artifacts: A Locally Discriminative Learning Approach to Realistic Image Super-Resolution. arXiv preprint arXiv:2203.09195.