BARET: Balanced Attention Based Real Image Editing Driven by Target-Text Inversion

Yuming Qiao[1,2], Fanyi Wang[1*], Jingwen Su[1], Yanhao Zhang[1], Yunjie Yu[1], Siyu Wu[3], Guo-Jun Qi[1,4]

[1]OPPO Research Institute
[2]Tsinghua University
[3]Zhejiang University
[4]Westlake University
qym21@mails.tsinghua.edu.cn, wangfanyi@oppo.com, sujingwen@oppo.com, zhangyanhao@oppo.com, yunjieyu007@gmail.com, 22151365@zju.edu.cn, guojunq@gmail.com

Abstract

Image editing approaches with diffusion models have been rapidly developed, yet their applicability are subject to requirements such as specific editing types (e.g., foreground or background object editing, style transfer), multiple conditions (e.g., mask, sketch, caption), and time consuming fine-tuning of diffusion models. For alleviating these limitations and realizing efficient real image editing, we propose a novel editing technique that only requires an input image and target text for various editing types including non-rigid edits without fine-tuning diffusion model. Our method contains three novelties: (I) Target-text Inversion Schedule (TTIS) is designed to fine-tune the input target text embedding to achieve fast image reconstruction without image caption and acceleration of convergence. (II) Progressive Transition Scheme applies progressive linear interpolation between target text embedding and its fine-tuned version to generate transition embedding for maintaining non-rigid editing capability. (III) Balanced Attention Module (BAM) balances the tradeoff between textual description and image semantics. By the means of combining self-attention map from reconstruction process and cross-attention map from transition process, the guidance of target text embeddings in diffusion process is optimized. In order to demonstrate editing capability, effectiveness and efficiency of the proposed BARET, we have conducted extensive qualitative and quantitative experiments. Moreover, results derived from user study and ablation study further prove the superiority over other methods.

Introduction

Recently, large Text-to-Image (T2I) models (Ramesh et al. 2022; Saharia et al. 2022; Rombach et al. 2022) have demonstrated powerful generation capabilities that have drawn a great deal of attentions. Among image generation tasks, real image editing has been an interesting and valuable topic for a long time. Real image editing refers to the controllable editing of a real image guided by specific condition (e.g., text, mask, sketch and additional views of the object) (Mou et al. 2023; Zhang and Agrawala 2023), while preserving the original image features as much as possible. In these numerous conditions, textual condition is the most concise and intuitive. Hence text-guided real image editing has become one of the most popular research topics.

Previous leading methods on text-guided image editing have shortcomings in different aspects. (I) limited editing capabilities such as object editing, background/foreground editing, and style transfer, etc.(Avrahami, Lischinski, and Fried 2022; Kim, Kwon, and Ye 2022; Bar-Tal et al. 2022), while incapable of non-rigid image editing. (II) image captions are required for later modification, but make the editing method inconvenient to be applied (Mokady et al. 2023). (III) fine-tuning the diffusion model on a single image leads to risks of undermining the pretrained model (Kawar et al. 2023; Zhang et al. 2023). To combat these deficiencies, we propose a novel text-based real image editing method named BARET. Only the original image and a simple target text of editing are needed to realize highly controllable image editing, including both basic editing scenarios and complex edits (i.e. non-rigid changes). As shown in the lower left corner of Fig. 1, when a simple text prompt “a dog lying down on a stone” is provided, the proposed BARET can realize posture change from standing to lying down on a stone of the same dog in the original image.

In this paper, the proposed BARET consists of three components. (I) Target-text inversion schedule circumvents the requirement of image caption and improves the efficiency of original image content extraction. Based on DDIM inversion, target text embedding is iteratively fine-tuned in DDIM sampling process for image reconstruction. (II) Progressive transition scheme enhances editing capabilities especially non-rigid edits. It integrates the target text information and original image content in the generated transition result, by applying progressive linear interpolation between fine-tuned embedding and target text embedding. (III) Balanced Attention Module is proposed to interact the self-attention map of reconstruction process, the cross attention map of transition process and the target embedding in the editing process. While maintaining the original image layout features, the capability of non-rigid editing is improved by combination of non-rigid semantic information from transition process and target text embedding. We have conducted a user study, based on a series of comparison experiments with leading

*Corresponding author
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Figure 1: Our method supports various editing types like style transfer, background and foreground editing, and complex editing with non-rigid changes. Here, we show some pairs of input real images and editing results aligned with target texts.

text-guided image editing methods. Results have shown that our BARET is superior in both objective metrics and user preference, especially on the task of non-rigid edits. Furthermore, ablation study on the hyperparameters of interpolation parameter, injection steps of self-attention map and cross attention map has been conducted as well. Our contributions in this paper are summarised as follows,

- Regarding of the additional image captions, we propose target-text inversion schedule to fine-tune target text embeddings for image reconstruction. It makes the editing process more concise, takes only 16s on a single A100, and avoids information loss during fine-tuning diffusion model.
- On the subject of the complex non-rigid edits, we propose a progressive transition scheme to transform non-rigid textual information to image semantics of transition result.
- In respect of balancing original image information and non-rigid change information, we further propose Balanced Attention Module (BAM). BAM enhances the complex non-rigid editing capability, by manipulating the self-attention map of the reconstruction process and the cross-attention map of the transition process.

Related Work

Large-scale text-to-image diffusion models (Saharia et al. 2022; Rombach et al. 2022; Chang et al. 2023; Gu et al. 2022; Song et al. 2023; Ramesh et al. 2022; Song and Ermon 2019) significantly improve the quality of generated images under textual conditions. Nevertheless, it is still challenging to achieve specific editing on real images, because such task is rarely realized by means of training. Mask guided diffusion models that denoise only in the masked region become popular for real image editing (Rombach et al. 2022; Nichol et al. 2022; Xie et al. 2023). However the interactive experience of mask-based editing is not quite user-friendly, and the limited generated content due to the mask might lead to distortion for object editing. Hence, text based image editing methods are proposed successively. Liu et al.(Liu et al. 2023) proposed semantic diffusion guidance to incorporate information of the reference text and the image in the diffusion process. Hertz et al. (Hertz et al. 2023) proposed to utilize the cross-attention map between text and image in the diffusion process for controllable real image editing. Inspired by the GAN inversion strategy (Yeh et al. 2017; Zhu et al. 2016; Xia et al. 2022; Lipton and Tripathi 2017; Creswell and Bharath 2018; Bermano et al. 2022), null-text inversion (Mokady et al. 2023) method extracts image caption by manual or off-the-shelve captioning model (Mokady, Hertz, and Bermano 2021) in advance, and fine-tunes unconditional embedding for image reconstruction, then edits the reconstructed image by prompt2prompt (Hertz et al. 2023). These methods do not require fine-tuning diffusion model, but are limited in editing performance in specific contexts.

Tuning model parameters during inversion (Roich et al. 2022; Tov et al. 2021; Valevski et al. 2022) is also applied to the diffusion model based image editing task. Dreambooth (Ruiz et al. 2023) fine-tunes the whole UNet with a few images of a specific subject and a unique identifier with class-specific prior preservation loss to achieve subject-driven generation. Kawar et al (Kawar et al. 2023) proposed Imagic that achieves image reconstruction by fine-tuning target text embedding and diffusion model. And image is edited by denoising process with interpolated embedding between target text embedding and fine-tuned embedding. SINE (Zhang et al. 2023) proposed model-based guidance strategy that combines the pre-trained diffusion model and its fine-tuned version to edit input image. We fully consider characteristics of existing text-only image editing methods and propose a novel image editing method named BARET. Our method has high capability of non-rigid edits, and does not require an additional image caption or fine-tuning diffusion model, thus significantly reduces inference time.

Methodology

Given the real image I_{org}, our goal is to solely rely on the target text X_{tgt} for editing to derive the result I_{edt} that preserves most of the original image features. To achieve this

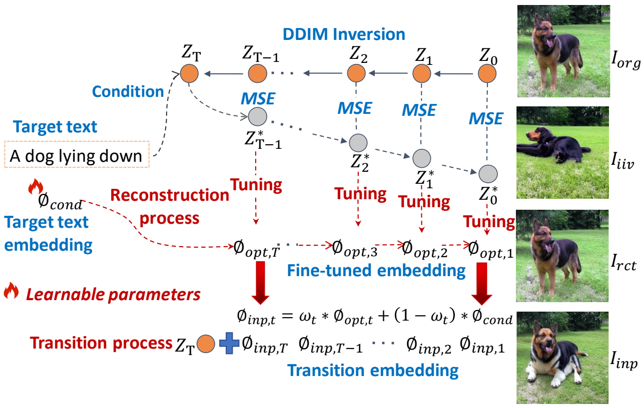

Figure 2: Illustration of our Target-text Inversion Schedule and Progressive transition scheme.

Algorithm 1: Target-text inversion

Input: Target text X_{tgt}, input image I_{org}
Output: Fine-tuned embedding $[\phi_{opt,T}, ..., \phi_{opt,1}]$

```
Get Z_0 ~ Z_T from I_org via DDIM inversion with guidance scale = 1. Get Z*_{T-1} ~ Z*_0 via Initial inversion under target text embedding φ_cond and noisy latent Z_T with guidance scale = 7.5.
for t = T, T-1, ..., 1 do
    φ_{opt,t,0} = φ(X_tgt)
    for i = 1, 2, ..., N do
        Z*_{t-1} = ε_θ(Z*_t, t, φ_{opt,t})
        φ_{opt,t,i} = λ(ℒ(Z*_{t-1}, Z_{t-1}), φ_{opt,t,i-1})
    end for
    φ_{opt,t} = φ_{opt,t,N}
end for
Return: [φ_{opt,T}, φ_{opt,T-1}, ..., φ_{opt,1}]
```

---

goal, a novel method named BARET is proposed in this paper to address shortcomings of previous methods. BARET contains three components, which are Target-Text Inversion Schedule, Progressive Transition Scheme and Balanced Attention Module.

Target-Text Inversion Schedule

Considering simplicity and efficiency of editing, this paper proposes target-text inversion schedule. The advantage is that there is no need to prepare the image caption in advance, and it greatly reduces fine-tuning time during reconstruction process. With the premise that only target text is given, fine-tuning null text embedding requires longer training time to realize image reconstruction. Since generation result in the classifier free guidance strategy is largely guided by the input text, when there exists mismatch between the target text and the image, more adjustments on null text embedding is necessary. On the contrary, direct fine-tuning the target text embedding can realize image reconstruction with fewer steps and faster convergence. Comparison of inference time of different methods is provided in Table 1. Target text inversion process is shown in Algorithm 1 and Fig. 2. Firstly, DDIM inversion applied on I_{org} derives initial diffusion trajectory $Z_0 \sim Z_T$. Then Z_T and the target text embedding ϕ_{cond} are used as the input to perform conditional DDIM sampling process and fine-tune ϕ_{cond}. The fine-tuning process is expressed in Equation 1.

$$\phi_{opt,t} = \lambda(\mathcal{L}(\varepsilon_\theta(Z_t^*, t, \phi_{opt,t}), Z_{t-1}), \phi_{cond}) \tag{1}$$

where $\phi_{opt,t}$ is the fine-tuned embedding at the tth timestamp, $\lambda(\cdot)$ is the optimizer, $\mathcal{L}(\cdot)$ is the reconstruction loss, Z_{t-1} is the diffusion trajectory of DDIM inversion, Z_{t-1}^* is the intermediate latent code of the DDIM sampling process with Z_t^* and fine-tuned embedding $\phi_{opt,t}$ as the input, ϕ_{cond} is the target text embedding and also the parameter to be optimized. The overall fine-tuning process takes Z_T and ϕ_{cond} with guidance scale 7.5 as the input, and output Z_{T-1}^*. Reconstruction loss between Z_{T-1}^* and Z_{T-1} is used to optimize ϕ_{cond} and obtain $\phi_{opt,T}$. Fine-tuned $\phi_{opt,T}$ and Z_T are used to update Z_{T-1}^* to fine-tune iteratively, which finally brings sampling process $\{Z_t^*\}_{t=0}^T$ closer to initial diffusion trajectory $\{Z_t\}_{t=0}^T$.

Balanced Attention Module

Existing attention-based methods directly manipulate self and cross attention maps of fine-tuned embedding. These methods often fail to achieve the expected results for edits with non-rigid changes. Although fine-tuned embedding has constructed strong spatial feature correlation with the original image, it is unable to produce the structural changes required for editing in the attention map. Hence the expected non-rigid changes cannot be manifested in the generated results. For example, taking "a dog lying down" and Fig. 1 bottom-left image as inputs, correlations between the fine-tuned embedding and the original image are constructed as shown in the first row of Fig. 3. It is easy to find that the attention map of the reconstructed image does not have the spatial layout information corresponding to "lying down". And from the corresponding attention map of "dog", it can be seen that it is still in standing posture. Second row of Fig. 3 shows the attention map of null-text inversion taking the caption "a dog standing up" as input for reconstruction. Likewise, attention maps corresponding to the words "dog" and "standing" can only reflect the standing posture of the dog. If "standing up" is replaced by "lying down" by the means of prompt2prompt (Hertz et al. 2023), it causes conflicts between the textual description of "a dog lying down" and the semantic information of the dog standing up in the attention map (second row of Fig. 3). Thus it is difficult to effectively inject non-rigid change information carried by the target text into the sampling process, resulting in unsatisfactory editing results that are not aligned with the given textual description.

Progressive Transition Scheme. In order to solve the above problems, we propose a novel embedding interpolation method to embody the non-rigid change information from the target text into the generated results. The interpolation method is expressed by the following Equation 2:

$$\phi_{inp,t} = \omega_t * \phi_{opt,t} + (1-\omega_t) * \phi_{cond} \tag{2}$$

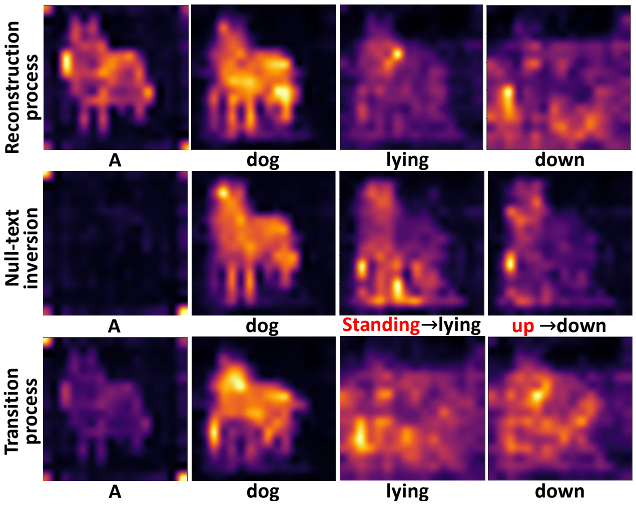

Figure 3: Illustration of the effectiveness of our transition process in non-rigid edits.

where ω_t and $\phi_{opt,t}$ are the interpolation parameter and fine-tuned embedding at timestamp t in diffusion process. Since denoising result in early stage largely determines the direction of generation, an interpolation parameter varying from large to small is applied. A large value assigned to the interpolation parameter at the beginning is helpful to preserve features of the original image. While small values set in the middle and later stages can drive the transition embedding $\phi_{inp,t}$ closer to the target text embedding ϕ_{cond}, injecting the non-rigid change information of the target text. As shown in Fig. 2, I_{rct} is the image reconstruction result generated by fine-tuned embedding $\{\phi_{opt,t}\}_{t=1}^{T}$ through DDIM sampling, I_{iiv} is initial inversion result by target text embedding ϕ_{cond}, I_{inp} is transition result sampled with transition embedding $\{\phi_{inp,t}\}_{t=1}^{T}$ using $\omega_t \sim U(0.8, 0.1)$. It can be seen that the initial inversion result represented by I_{iiv} loses a number of features of both the editing object and the background. The transition result I_{inp} fails to achieve the editing goal, but it can introduce the target text semantic information while retaining the original image layout and editing object features. As shown in Fig. 3 bottom row, the cross attention map formed between $\{\phi_{inp,t}\}_{t=1}^{T}$ and I_{inp} has obvious non-rigid change information compared with that of the null-text inversion in middle row. Compared to null-text inversion, our transition process can better characterize the non-rigid change semantic information of the target text during the diffusion process. This provides important support for the following Balanced Attention Module.

Balanced Attention Module. As shown in Fig. 4, we define DDIM process under $\{\phi_{opt,t}\}_{t=1}^{T}$ as the reconstruction process, DDIM under $\{\phi_{inp,t}\}_{t=1}^{T}$ as the transition process, and DDIM under ϕ_{cond} as the editing process.

To retain sufficient original image features in the editing process, in self-attention module original Q_{iiv} and K_{iiv} are replaced with Q_{rct} and K_{rct} in the reconstruction process. To better utilize the non-rigid change information in the transition process, in cross-attention module original Q_{iiv} and $K_{\phi_{cond}}$ are replaced by Q_{inp} and $K_{\phi_{inp}}$ from transition process. It is noted that non-rigid semantic of $Q_{inp} \cdot K_{\phi_{inp}}^{T}$ has a good alignment with non-rigid textual information of ϕ_{cond}, which further improves the non-rigid editing capability. Modifications on self-attention and cross-attention modules during editing process can be expressed in the following Equations 3 and 4,

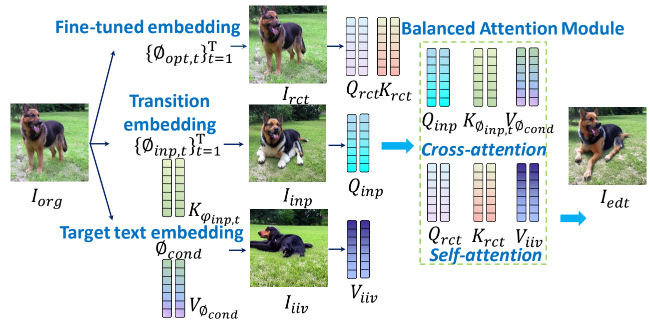

Figure 4: Illustration of Balanced Attention Module.

$$softmax(\frac{Q_{rct} \cdot K_{rct}^{T}}{\sqrt{d}}) \leftarrow softmax(\frac{Q_{iiv} \cdot K_{iiv}^{T}}{\sqrt{d}}) \tag{3}$$

$$softmax(\frac{Q_{inp} \cdot K_{\phi_{inp}}^{T}}{\sqrt{d}}) \leftarrow softmax(\frac{Q_{iiv} \cdot K_{\phi_{cond}}^{T}}{\sqrt{d}}) \tag{4}$$

For editing tasks with rigid changes (e.g., style transfer, foreground and background editing), there is no need to embody the non-rigid change information in the attention map. Correspondingly, interpolation weight ω_t is set to 1, and transition embedding $\{\phi_{inp,t}\}_{t=1}^{T}$ is completely equivalent to the fine-tuned embedding $\{\phi_{opt,t}\}_{t=1}^{T}$. Subsequently, BAM introduces self-attention map and cross-attention map from the reconstruction process to the editing process. Hence information of the target text is incorporated on the basis of maintaining most of original image features. As demonstrated in Fig. 1, our BARET method has powerful capability of fast arbitrary editing. Additionally, as shown in Fig. 5, our BARET also has excellent performance on tasks of object addition and personalized portrait editing.

Experiments

Implementations

All experiments are based on stable diffusion v1.5 (Rombach et al. 2022), implemented DDIM sampling strategy with 50 steps and guidance scale 7.5. For the target text inversion, loss function of fine-tuning target text embedding is MSE, tuning iterations are 250 in total and 5 iterations per step, and optimizer is Adam (Kingma and Ba 2015). Learning rate is 0.001. Progressive loss value is defined as $\{t * 1e-5\}_{t=1}^{T}$ in each timestamp for early stop, such that reconstruction quality can be boosted with low threshold of loss at the early stage of denoising. And the threshold is gradually raised in the later stage.

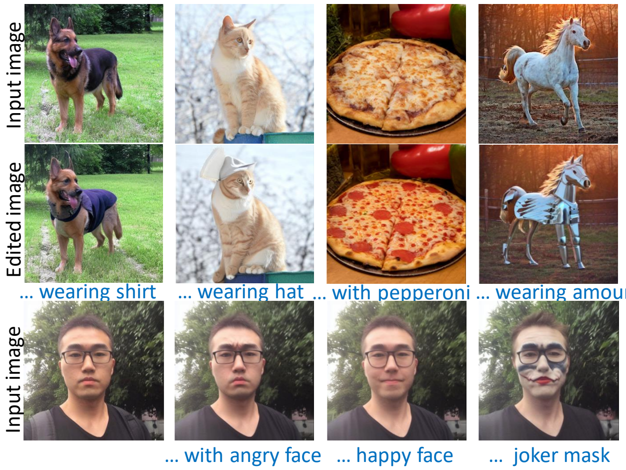

Figure 5: Obejct addition and face manipulation samples of our BARET.

Result Analysis

Comparison. We compare our method with the state-of-the-art text-based real image editing methods with code that are publicly available, which are SDEdit (Meng et al. 2022), Pix2Pix-Zero (Parmar et al. 2023), MasaCtrl (Cao et al. 2023), Null-text inversion (Mokady et al. 2023), Imagic (Kawar et al. 2023) and SINE (Zhang et al. 2023). Their text input format and fine-tune strategy were deployed as described in their papers. It is noted that both Null-text and SINE need prior caption. Fig. 6 shows the editing results of different methods. The proposed BARET significantly outperforms the other methods, in terms of text alignment between editing results and target text, and image fidelity between editing results and original image, especially on complex non-rigid editing (e.g, bird spreading wings, giraffe lying down, cat standing up and dog sitting down). In addition, our method has obvious advantage on efficiency. The inversion stage of our method takes only about 16s on a single A100, which greatly improves the editing efficiency compared to methods that require fine-tuning diffusion model such as SINE and Imagic. They take longer tuning time around 10~20 minutes. While the null-text inversion, which is the most efficient among methods for comparison, takes at least about 50s to complete fine-tuning. As shown on the left side of Fig. 7, the proposed TTIS method converges significantly faster than the null-text inversion method, and completes convergence at 100 steps in 16.1s, achieving superior reconstruction results. But null-text inversion needs more steps (>400 steps) and longer time (48~66s) to complete. Moreover, null-text inversion has a higher probability of reconstruction failure compared to our TTIS as shown in Fig. 7 right column. Because target-text embedding is a better initialization parameter than null-text embedding in the semantic space, which leads to fewer iteration steps to converge and a better reconstruction robustness. Also, from the aspect of back-propagation, as shown in Equations 5, 6 and 7, the classifier guidance scale parameter ω is larger compared to that of fine-tuning null-text embedding, which constitutes one of the reasons that accelerate TTIS.

Method	Inversion	Editing	Non-rigrid edits	Multiple edits
SDEdit	/	10s	Yes	Yes
SINE	20m	10s	Yes	Yes
Imagic	20m	10s	Yes	Yes
MasaCtrl	/	10s	Yes	Yes
Pix2Pix-Zero	/	10s	No	Yes
Null-text	1m	10s	No	Yes
Ours	**16s**	**10s**	**Yes**	**Yes**

Table 1: Inference time comparison.

$$\frac{\partial \pounds(Z_{t-1}, Z^*_{t-1})}{\partial Y} \sim \frac{\partial DM(Z_t, t, Y, \oslash)}{\partial Y} \tag{5}$$

$$\frac{\partial DM(Z_t, t, Y, \oslash)}{\partial Y} \sim \nabla DM(Z_t, t, Y, \oslash) * \frac{\partial \varepsilon_\theta(Z_t, t, Y, \oslash)}{\partial Y} \tag{6}$$

$$\frac{\partial \varepsilon_\theta(Z_t, t, Y, \oslash)}{\partial Y} \sim \omega * \frac{\partial \varepsilon_\theta(Z_t, t, Y)}{\partial Y} \tag{7}$$

User study. Due to the lack of rigorous and accurate quantitative metrics for real image editing tasks, we have conducted a user study to quantify visual effects of various editing methods. To this end, we refer to TEdBench (Kawar et al. 2023) and collected 100 pairs of real image and textual description for editing. Contents present to the user are 100 pairs of real image, editing target text, editing result of five methods. The five methods are the proposed BARET and four baselines: MasaCtrl, SINE, Null-text, SDEdit. And the scoring process is blind for the user to know which method corresponds to the editing result. The user is asked to score the editing results of the five methods ranging from 1 to 5, with 5 representing the best quality that has good alignment with target text and preserves characteristics of original image. A total of 62 valid results were collected and summarized in Fig. 8 (left). The average of scores of each method of 100 editing results was computed. We find that our BARET (mean: 3.77, std: 0.20) significantly outperforms the Null-text+p2p (mean: 2.93, std: 0.16), MasaCtrl (mean : 2.31, std: 0.23), SDEdite (mean: 1.93, std: 0.24) and SINE (mean: 2.42, std: 0.12).

Moreover, the top three ranked methods in user study (i.e., Our BARET, Null-text and SINE), Pix2Pix-Zero and Imagic are further evaluated by calculating 1-LIPIS (Zhang et al. 2018) and CLIPScore (Hessel et al. 2021; Radford et al. 2021) metrics. LPIPS is the perceptual distance between the edited image and the original image, standing for image fidelity. A large value of 1-LPIPS indicates a closer match to the original image. CLIPScore computes the cosine similarity between the target text embedding and editing result image embedding, representing the degree of alignment between the editing result and the target text. The larger the value of CLIPScore is, the closer the editing result is to the description of the target text. As shown in Fig. 8 (right), our BARET (1-LPIPS: 0.764, CLIPScore: 0.729) has the best objective evaluation result. Null-text method has 1-LPIPS of 0.766, which is slightly higher than that of our BARET, but its CLIPScore (0.681) is absolutely lower than our BARET, indicating poor capability in non-rigid editing. And

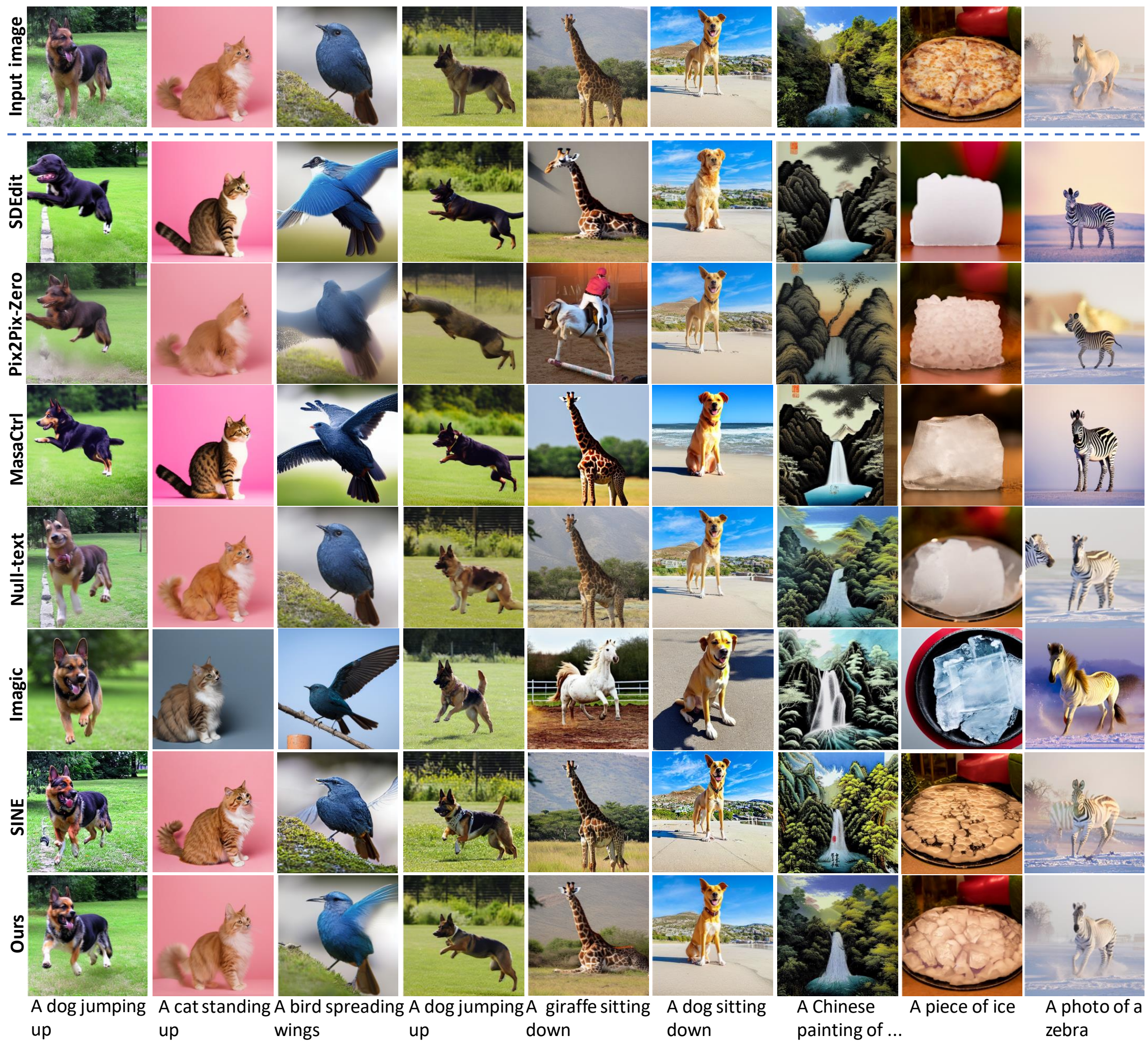

Figure 6: Method Comparison. We compare SDEdit, Pix2Pix-Zero, MasaCtrl, Null-text inversion, Imagic and SINE to our method. Our BARET can be applied in arbitrary edits.

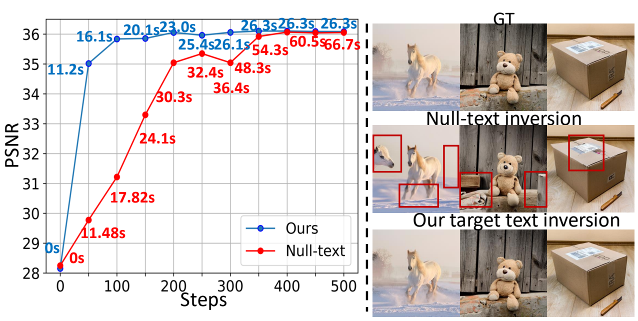

Figure 7: Comparison of our TTIS and Null-text inversion.

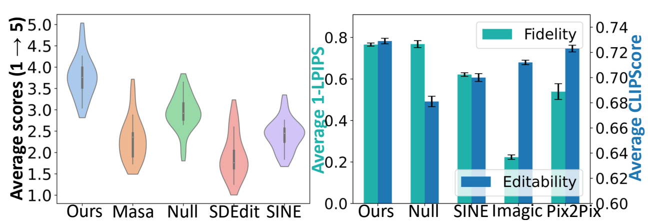

Figure 8: Qualitative analysis of our BARET compared to other leading image editing approaches.

our BARET surpasses SINE (1-LPIPS: 0.619, CLIPScore: 0.7), Pix2Pix-Zero (1-LPIPS: 0.539, CLIPScore: 0.723) and Imagic (1-LPIPS: 0.223, CLIPScore: 0.712).

Ablation Study

The proposed BARET has three adjustable parameters, interpolation parameter $\{\omega_t\}_{t=1}^{T}$, self-attention injection step η and cross-attention injection step λ. Interpolation parameter controls the fusion ratio between the non-rigid change information and the original image information in transition

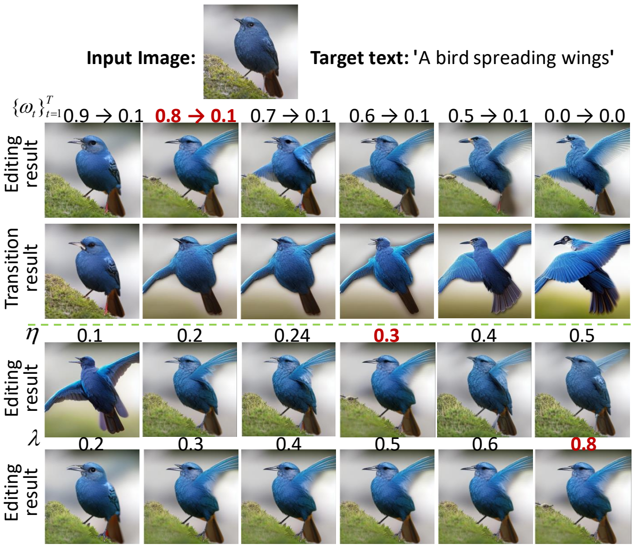

Figure 9: Editing results with various interpolated parameters $\{\omega_t\}_{t=1}^{T}$, self attention injection steps η and cross attention injection steps λ.

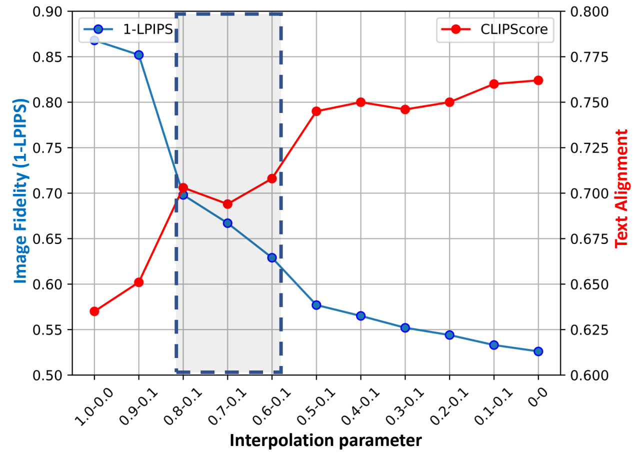

Figure 10: Editability–fidelity tradeoff. CLIP score (target text alignment) and 1-LPIPS (input image fidelity) as functions of interpolation parameter. Edited images tend to match both the input image and target text in the highlighted area.

process. For example, as shown in the first row of Fig. 9, when the initial interpolation parameter is large (e.g., 0.9-0.1), transition process is prone to generate images similar to the original image. As the interpolation parameter decreases (e.g., 0.8-0.1), transition process can better merge the non-rigid change information and the original object features. From transition results in Fig. 9, it can be seen that while maintaining characteristics of the original bird, motion information "spreading wings" is also incorporated. Although most background features are lost, it can be mitigated by the self-attention maps in the reconstruction process. Correspondingly, it is noticed that editing result under interpolation parameter 0.8-0.1 shows a good tradeoff between characteristics of original image and desired non-rigid changes. However, as the interpolation parameter is further reduced, the proportion of the original image feature in the transition result gradually decreases. As we can observe, when the interpolation parameter is 0.6-0.1, 0.5-0.1, 0-0, the difference of the bird in the editing result and the original image gradually increases. Similarly, the injection step of self-attention (SA) map and cross-attention (CA) map in BAM is also a very important parameter. As shown in Fig. 9, when the SA injection step η is small such as 0.1, the editing result cannot preserve sufficient features of the original image. As η gradually increases to 0.2-0.3, edited results show a good balance of original content and non-rigid changes. However, as η further increases to 0.4-0.5, edited results gradually fail to incorporate non-rigid changes and become more similar to the original image. Unlike the SA injection step η, editing results are less sensitive to the CA injection step λ. When λ is between 0.4 and 0.8, it has almost no effect on the editing results. But if λ is further reduced it gradually leads to poor incorporation of non-rigid changes in the editing result.

In summary, we can find that editing results are more sensitive to the interpolation parameter, thus we calculate the LPIPS and CLIPScore of the editing result by varying interpolation parameter. Result analysis is present in Fig. 10. As the interpolation parameter decreases, transition embedding is closer to target text embedding, non-rigid information increases and the original image information decreases. And for corresponding editing result, CLIPScore gradually increases and 1-LPIPS gradually decreases, indicating that the editing result loses more detailed information of the original image and better align with target text. Considering both image fidelity and text alignment, a reasonable range of the interpolation parameter is 0.8-0.1 $\sim$ 0.6-0.1 (highlighted area in Fig. 10) by empirical. In the meanwhile, suggested SA injection steps η and CA injection steps λ are 0.2-0.4 and 0.4-0.8 respectively.

Conclusions

In this paper, we propose BARET, a text based real image editing method with high efficiency supporting for complex non-rigid edits. BARET only requires an input image and target text for editing. Firstly, the proposed TTIS fine-tunes the target text embedding to achieve image reconstruction. Then the proposed progressive transition scheme utilises the target text embedding and fine-tuned embedding to integrate the original image information and the non-rigid structural change information of the target text. Finally, the proposed BAM combines the self-attention map in the reconstruction process and the cross-attention map in the transition process to effectively integrate the characteristics of the original image and the non-rigid change information. Compared with other leading text-based methods, our method has no need to acquire image captions beforehand, and can realize not only simple image editing, but also sophisticated human face manipulation and non-rigid editing, exhibiting best text alignment and image fidelity. In addition, our method does not require fine-tuning diffusion model, which is a significant advantage in terms of controllability and editing efficiency.

References

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended Diffusion for Text-driven Editing of Natural Images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18187–18197.

Bar-Tal, O.; Ofri-Amar, D.; Fridman, R.; Kasten, Y.; and Dekel, T. 2022. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, 707–723. Springer.

Bermano, A. H.; Gal, R.; Alaluf, Y.; Mokady, R.; Nitzan, Y.; Tov, O.; Patashnik, O.; and Cohen-Or, D. 2022. State-of-the-Art in the Architecture, Methods and Applications of StyleGAN. In *Computer Graphics Forum*, volume 41, 591–611. Wiley Online Library.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. *arXiv preprint arXiv:2304.08465*.

Chang, H.; Zhang, H.; Barber, J.; Maschinot, A.; Lezama, J.; Jiang, L.; Yang, M.-H.; Murphy, K.; Freeman, W. T.; Rubinstein, M.; et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Creswell, A.; and Bharath, A. A. 2018. Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems*, 30(7): 1967–1974.

Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Yuan, L.; and Guo, B. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10696–10706.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R. L.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In Moens, M.; Huang, X.; Specia, L.; and Yih, S. W., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 7514–7528. Association for Computational Linguistics.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017.

Kim, G.; Kwon, T.; and Ye, J. C. 2022. DiffusionCLIP: Text-Guided Diffusion Models for Robust Image Manipulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2416–2425.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Lipton, Z. C.; and Tripathi, S. 2017. Precise recovery of latent vectors from generative adversarial networks. *arXiv preprint arXiv:1702.04782*.

Liu, X.; Park, D. H.; Azadi, S.; Zhang, G.; Chopikyan, A.; Hu, Y.; Shi, H.; Rohrbach, A.; and Darrell, T. 2023. More Control for Free! Image Synthesis with Semantic Diffusion Guidance. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, 289–299. IEEE.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.; and Ermon, S. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6038–6047.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Mou, C.; Wang, X.; Xie, L.; Wu, Y.; Zhang, J.; Qi, Z.; Shan, Y.; and Qie, X. 2023. T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models. arXiv:2302.08453.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 16784–16804. PMLR.

Parmar, G.; Singh, K. K.; Zhang, R.; Li, Y.; Lu, J.; and Zhu, J. 2023. Zero-shot Image-to-Image Translation. In Brunvand, E.; Sheffer, A.; and Wimmer, M., eds., *ACM SIGGRAPH 2023 Conference Proceedings, SIGGRAPH 2023, Los Angeles, CA, USA, August 6-10, 2023*, 11:1–11:11. ACM.

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. *arXiv preprint arXiv:2204.06125*.

Roich, D.; Mokady, R.; Bermano, A. H.; and Cohen-Or, D. 2022. Pivotal tuning for latent-based editing of real images. *ACM Transactions on graphics (TOG)*, 42(1): 1–13.

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*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Song, Y.; Dhariwal, P.; Chen, M.; and Sutskever, I. 2023. Consistency Models. In Krause, A.; Brunskill, E.; Cho, K.; Engelhardt, B.; Sabato, S.; and Scarlett, J., eds., *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, 32211–32252. PMLR.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Tov, O.; Alaluf, Y.; Nitzan, Y.; Patashnik, O.; and Cohen-Or, D. 2021. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4): 1–14.

Valevski, D.; Kalman, M.; Matias, Y.; and Leviathan, Y. 2022. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*.

Xia, W.; Zhang, Y.; Yang, Y.; Xue, J.-H.; Zhou, B.; and Yang, M.-H. 2022. Gan inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3121–3138.

Xie, S.; Zhang, Z.; Lin, Z.; Hinz, T.; and Zhang, K. 2023. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22428–22437.

Yeh, R. A.; Chen, C.; Yian Lim, T.; Schwing, A. G.; Hasegawa-Johnson, M.; and Do, M. N. 2017. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5485–5493.

Zhang, L.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Z.; Han, L.; Ghosh, A.; Metaxas, D. N.; and Ren, J. 2023. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6027–6037.

Zhu, J.-Y.; Krähenbühl, P.; Shechtman, E.; and Efros, A. A. 2016. Generative visual manipulation on the natural image manifold. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, 597–613. Springer.