Referring Object Manipulation of Natural Images with Conditional Classifier-free Guidance

Myungsub Choi*[0000−0003−4731−3074]

Google Research cms6539@gmail.com

Abstract. We introduce the problem of referring object manipulation (ROM), which aims to generate photo-realistic image edits regarding two textual descriptions: 1) a text referring to an object in the input image and 2) a text describing how to manipulate the referred object. A successful ROM model would enable users to simply use natural language to manipulate images, removing the need for learning sophisticated image editing software. We present one of the first approach to address this challenging multi-modal problem by combining a referring image segmentation method with a text-guided diffusion model. Specifically, we propose a conditional classifier-free guidance scheme to better guide the diffusion process along the direction from the referring expression to the target prompt. In addition, we provide a new localized ranking method and further improvements to make the generated edits more robust. Experimental results show that the proposed framework can serve as a simple but strong baseline for referring object manipulation. Also, comparisons with several baseline text-guided diffusion models demonstrate the effectiveness of our conditional classifier-free guidance technique.

Keywords: Referring segmentation, text-guided image manipulation

1 Introduction

With the surge of digital content and an ever increasing number of daily creators, there have been more and more needs for easy-to-use image/video editing tools. However, existing tools usually require expensive software or professional knowledge of editing techniques. To allow image editing to be more accessible to diverse user groups, recent works are beginning to explore image manipulation with natural language, which can serve as a highly intuitive user interface [5,41].

Recently, the combination of large-scale vision-language models [42] and high-quality generative models [25,38] led to interesting new text-driven applications, including text-guided image manipulation [4,37,41] and out-of-domain image translation [13]. However, previous methods typically modify the image globally, and fine-grained control of specific objects is not possible. A number of recent methods [4,37,53] also allow to use a segmentation mask as an additional input, so that users can specify the regions for text-guided inpainting. While providing

* Now at Samsung Advanced Institute of Technology (SAIT)

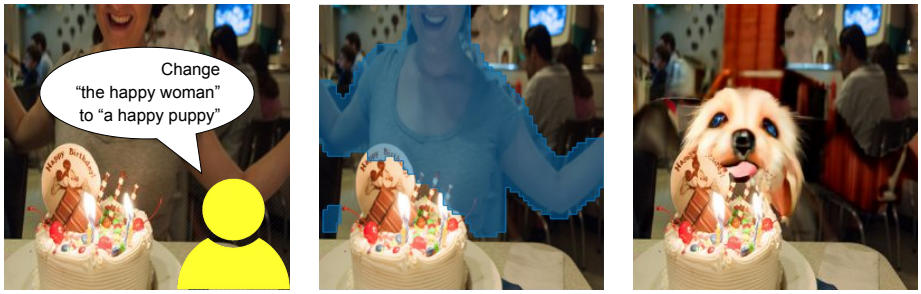

Fig. 1. Referring object manipulation problem setting. Given an image, a referring text prompt that describes which region to edit, and a target text prompt describing how to modify the specified region, our goal is to generate a photo-realistic edited image that matches all (both referring and target) textual descriptions.

a mask is a convenient interface for image editing, it still requires the users to draw a good mask that fully covers the regions of interest.

In this work, we introduce the new problem of *referring object manipulation*, which can provide a fully automatic user interface of image editing with natural language. The goal of this task is to generate photo-realistic image edits that follow the target text description, given an input image and a text referring to a specific region in the image. The edited output image should be different from the input image only in the referred regions, and the intended modifications should correctly reflect the attributes described in the target text. The main concept of our proposed problem setting is illustrated in Fig. 1.

To address this challenging problem for the first time, we present a simple baseline framework that combines a referring object segmentation model with text-guided image manipulation model. In particular, we leverage the pretrained models of MDETR [21] and GLIDE [37] for localizing the referring object and editing the region with textual guidance, respectively. While naive sequential combination of the two models shows plausible result, we propose three additional techniques for improvement: 1) a new conditional classifier-free guidance for better guiding the generation process in GLIDE, 2) localized ranking of multiple generations, and 3) dilation of the intermediate segmentation mask. Note that, our proposed techniques do not require any additional training or fine-tuning of the pretrained model parameters but still shows significant improvements. The experimental results and analyses demonstrate the effectiveness of the proposed framework, both qualitatively and quantitatively with a user study.

In summary, our contributions are as follows:

- We introduce a new problem of referring object manipulation, and propose a simple and effective baseline framework.
- We present conditional classifier-free guidance for improved manipulation of local image regions using a text-guided diffusion model.
- The proposed framework generates the most favorable image edits qualitatively and outperforms all compared baselines.

2 Related Works

2.1 Text-guided image manipulation

Multi-modal representation learning Many existing works on vision and language learn a joint representation used for various downstream applications, including image captioning [28], visual question answering [3], and text-based image retrieval [22]. With the advances in Transformers [49] in the language domain, recent representation learning methods [19,42] also adopt similar architectures and train a joint embedding space with large-scale image-text data [32,46]. Notably, CLIP [42] model, which is trained on 400 million image-text pairs with contrastive learning approach, provides a powerful representation that can be used to estimate the semantic similarity between a given image-text pair.

Text-guided image generation/manipulation Early works on text-guided image synthesis [45,60] and manipulation [11,36,62] train a conditional GAN [35] based model with learned text embeddings. While the fidelity of the generated samples are greatly improved in the following efforts [30,54,61], images generated using these models are usually restricted to specific domains (*e. g.* flowers [50] or birds [39]), and could not be generalized to make diverse natural images.

Recently, text-guided image generation/manipulation problem [5,41,53] is gaining increased attention with the progress in large-scale multi-modal representation learning methods [19,42]. The most impressive works leverage the strong generative power of modern GANs [23,24,25] combined with CLIP. Notable approaches include StyleCLIP [41], which introduces three methodologies to manipulate the latent space of StyleGAN2 [25] with textual guidance using the semantic power of CLIP. Also, Bau *et al.* [5] used additional user-given mask input to perform text-guided inpainting using StyleGAN2 and CLIP. Following works develop many interesting new improvements, such as enabling out-of-domain manipulations [13], exploring robustness [34] or disentanglement [55] for better generative quality, and accelerating inference time [29]. However, GAN-based approaches are often limited to a restricted domain of their training data and require special GAN inversion techniques [1,2,64] to manipulate real images.

On the other hand, several approaches aim to use diffusion models [48] as an alternative to GANs. These efforts combines a conditional diffusion model with CLIP and demonstrate robust out-of-domain manipulation [26], capabilities for local manipulation on realistic natural images [4], and photo-realistic synthesis and editing with a large-scale model [37]. In particular, the GLIDE model of Nichol *et al.* [37] greatly improves the previous work DALL-E [43] with diffusion models and classifier-free guidance [16]. In this work, we adopt GLIDE for text-guided local image manipulation with a novel guidance scheme fitted for the new problem of referring object manipulation.

We note that Zhang *et al.* [63] introduce a similar problem setting of image manipulation by text instruction, which specifies the object or region of the input image in natural language. However, their method is only tested on synthetic datasets with constrained set of vocabularies, whereas we demonstrate the capabilities of our framework on much more challenging settings.

2.2 Referring image segmentation

Referring image segmentation, first introduced by Hu *et al.* [18], aims to segment a target region (object or stuff) in an image that corresponds to the given natural language expression. Standard approaches [18,31,33,47,59] first extract image features with a CNN and text features with LSTM [17]. Then, the multi-modal features are fused to estimate the segmentation mask using an image segmentation model. Recent approaches adopt Transformers [49] for extracting better text features [21,56], better fusion and localization [12,20,58], or sometimes to train a unified multi-modal model [10,57].

Current state of the recent referring image segmentation models are surprisingly good, which motivated us to directly use the results for the challenging task of referring object manipulation. Though we choose to use MDETR [21] in this paper for its good performance and code availability[1], note that any other model can take place in our framework, and we can also benefit from the developments in referring image segmentation architectures.

3 Background

In this section, we briefly review the series of developments in guided diffusion models: the baseline diffusion model [15,48], classifier guidance [9], classifier-free guidance [16], and CLIP guidance [37]. The line of works form the fundamentals of our proposed classifier-free guidance technique and are all compared in the experiments. We generally follow the notations as summarized in GLIDE [37]. For detailed mathematical derivations, we refer the readers to [15] and [9].

3.1 Diffusion Models

Given a sample from the real data distribution $x_0 \sim q(x_0)$, a diffusion process generates a Markov chain of latent variables $x_1, ..., x_T$ by adding Guassian noise at each timestep t:

$$q(x_t|x_{t-1}) \coloneqq \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathcal{I}), \tag{1}$$

where the amount of noise is controlled by a variance schedule $\{\beta_t \in (0,1)\}_{t=1}^{T}$. It is known that if β_t is small enough, the posterior $q(x_{t-1}|x_t)$ can be approximated by a diagonal Gaussian, and that the final variable x_T approximately follows $\mathcal{N}(0,\mathcal{I})$ with sufficiently large amount of total noise added. Since calculating the true posterior $q(x_{t-1}|x_t)$ is infeasible, an approximate model p_θ needs to be learned as follows:

$$p_\theta(x_{t-1}|x_t) \coloneqq \mathcal{N}(\mu_\theta(x_t), \Sigma_\theta(x_t)). \tag{2}$$

Then, sample generation can be done by starting with a random Gaussian noise $x_T \sim \mathcal{N}(0,\mathcal{I})$ and sequentially sampling $x_{T-1}, ..., x_0$ using the learned model. In

[1] `https://github.com/ashkamath/mdetr`

practice, Ho *et al.* [15] uses a reparameterization trick [27] and decompose the latent variable x_t into a mixture of signal x_0 and some additive noise ϵ, which is estimated by a noise approximation model $\epsilon_\theta(x_t, t)$. They also derive $\mu_\theta(x_t)$ as a function of $\epsilon_\theta(x_t, t)$, fix Σ_θ to a constant, and use a simplified mean-square error objective for practical benefits:

$$L_{\text{simple}} \coloneqq E_{t\sim[1,T],x_0\sim q(x_0),\epsilon\sim\mathcal{N}(0,\mathbf{I})}[||\epsilon - \epsilon_\theta(x_t, t)||^2]. \tag{3}$$

3.2 Guided Diffusion

Dhariwal and Nichol [9] showed that better class-conditioned samples can be generated with classifier guidance. Concretely, the mean $\mu_\theta(x_t|y)$ and variance $\Sigma_\theta(x_t|y)$ of the diffusion model is perturbed by the classifier's gradient for a target class y. The resulting *perturbed* mean $\hat{\mu}_\theta(x_t|y)$ can then be calculated as:

$$\hat{\mu}_\theta(x_t|y) = \mu_\theta(x_t|y) + s \cdot \Sigma_\theta(x_t|y)\nabla_{x_t} \log p_\phi(y|x_t), \tag{4}$$

where the coefficient s is a guidance scale that controls the trade-off between sample quality and diversity (higher s gives better quality with less diversity). One downside of classifier guidance is that it requires a separate classifier which needs to be explicitly trained on noisy input images (to simulate the latent variables x_t). This introduces notable additional complexity, since the standard pretrained classifiers (trained on clean images) cannot be used.

3.3 Classifier-free guidance

Classifier-free guidance (CFG), first proposed by Ho and Salimans [16], is a recent technique that removed the need for a separately trained classifier. Specifically, when training a class-conditional diffusion model $\epsilon_\theta(x_t|y)$, the class label y is randomly replaced with a null label $\emptyset$ with a fixed probability (denoted as an unconditional model, $\epsilon_\theta(x_t|\emptyset)$). Sampling is done by a linear combination of the conditional and unconditional model estimates:

$$\hat{\epsilon}_\theta(x_t|y) = \epsilon_\theta(x_t|\emptyset) + s \cdot (\epsilon_\theta(x_t|y) - \epsilon_\theta(x_t|\emptyset)), \tag{5}$$

where $s \geq 1$ is the guidance scale. Intuitively, CFG further extrapolates the output of the model along the direction of $\epsilon_\theta(x_t|y)$, moving away from $\epsilon_\theta(x_t|\emptyset)$.

GLIDE [37] used CFG with generic text prompts, which is implemented by randomly replacing the text captions with an empty sequence ($\emptyset$) during training. The generative process can then be guided towards the caption c as

$$\hat{\epsilon}_\theta(x_t|c) = \epsilon_\theta(x_t|\emptyset) + s \cdot (\epsilon_\theta(x_t|c) - \epsilon_\theta(x_t|\emptyset)). \tag{6}$$

CFG can be thought of a self-supervised way of leveraging the learned knowledge of a single diffusion model. In this work, we extend this approach to give better guidance direction when applied to a referring object manipulation problem.

3.4 CLIP Guidance

CLIP [42] is a popular method of learning joint image-text representation. The model consists of an image encoder $f(x)$ and a caption encoder $g(c)$, which is trained with a contrastive loss that encourages a high dot product for the matching image (x) - text (c) pairs and low values otherwise.

Since CLIP provides a way of measuring the semantic distance between an image and a caption, many previous works use it for designing text-guided image manipulation models using the state-of-the-art GANs [13,41]. More recently, the same idea is applied to diffusion models [4,26,37], where the noisy classifier of classifier guidance (Eq. (4)) is replaced with a CLIP model:

$$\hat{\mu}_\theta(x_t|c) = \mu_\theta(x_t|c) + s \cdot \Sigma_\theta(x_t|c)\nabla_{x_t}\left(f(x_t) \cdot g(c)\right). \tag{7}$$

Prior works [7,8] have shown that the public CLIP models are capable of guiding the diffusion models, even though they are not trained with noisy input images x_t as in [9]. However, GLIDE [37] has shown that noise-aware trained model, named as noised CLIP model, performs considerably better than the unnoised CLIP, and we use the noised version in our comparison experiments.

4 Method

4.1 Problem setting

We formulate the problem of *Referring Object Manipulation (ROM)*, which aims to modify the referring region of interest from an input image to conform to the target text expression. Specifically, a ROM model has three inputs: an input image $\mathbf{I}$, a referring text prompt c_{ref}, and a target text prompt c_{target}. The output is an edited image $\tilde{\mathbf{I}}$, which should successfully contain the attributes described in the target text. To achieve this goal, a model should correctly infer the local regions where c_{ref} is referring to, and then manipulate the regions according to the target c_{target}. This is a challenging task that requires full multi-modal (vision and language) understanding and high-quality generative models. The conceptual illustration is shown in Fig. 1.

Referring object manipulation problem can be decomposed into two sub-problems, referring image segmentation and text-guided image inpainting. Referring image segmentation models aim to estimate a precise segmentation mask $\mathbf{M}$, given an input image $\mathbf{I}$ and a referring prompt c_{ref}. The goal of text-guided image inpainting models is to generate a photo-realistic edited image $\tilde{\mathbf{I}}$ given an input image $\mathbf{I}$, a (user-given) mask specifying the regions to edit, and a target prompt c_{target}. Therefore, by substituting the user-given mask with the automatically generated mask $\mathbf{M}$, we can build an end-to-end ROM framework.

With recent developments in both fields (referring segmentation and text-guided inpainting), a sequential combination of two models serves as a simple but strong baseline. However, due to the different focus and the evaluation metrics in each field, there exists some cases when the errors from an earlier model propagates and generates visually unpleasing outputs. In the following subsections, we propose a novel solution to make the generations more robust.

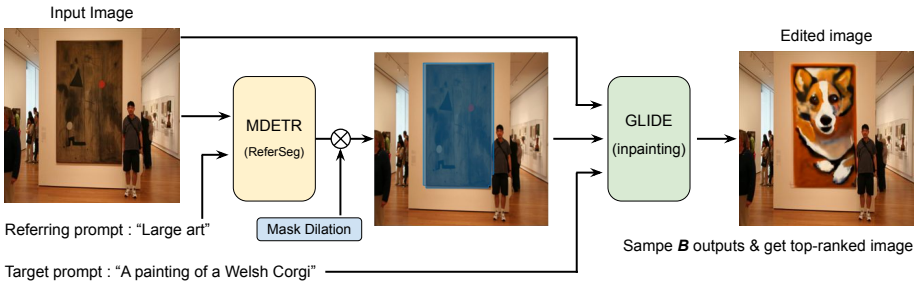

Fig. 2. Architecture overview. First, MDETR model estimates the referred-to segmentation mask given the input image and the referring text prompt. Then, using the input image, the dilated segmentation mask, and the target text prompt, GLIDE model edits the masked regions to correctly follow the target prompt. The final output is decided to be the top-ranked image *w.r.t.* our localized ranking scheme, out of $B = 24$ samples.

4.2 Architecture overview

As a realization of the referring object manipulation framework, we combine two state-of-the-art models in each area: MDETR [21] for referring image segmentation, and GLIDE [37] for text-guided image inpainting.

MDETR is a Transformer-based text-guided detection model that can localize a specific image region given a referring textual expression. In practice, we use the extended MDETR model fine-tuned on PhraseCut dataset [51], which allows for generating pixel-level segmentation masks along with the bounding boxes.

GLIDE is a large-scale image generation and editing framework based on conditional diffusion models. We use the model specifically trained to perform image inpainting; in particular, we use the smaller open-sourced version[2] that is trained with a filtered dataset.[3]

A simple combination of MDETR and GLIDE can occasionally generate impressive output edits, but we also introduce three additional improvements for more reliable manipulation: conditional classifier-free guidance, context-aware localized output ranking, and mask dilation. Each new component will be described in detail in the following Sec. 4.3, 4.4, and 4.5, respectively.

Note that we use the pretrained models from MDETR and GLIDE *as is*, without any further training or fine-tuning. Also, any referring object segmentation model can be substituted instead of MDETR, and any conditional diffusion model can be substituted instead of GLIDE, as long as it is trained with the inpainting setting with a mask input.

[2] `https://github.com/openai/glide-text2im`

[3] The filtered dataset aims to remove any potential bias in the data and pretrained models. This model should be denoted as GLIDE (*filt.*) following the original work [37], but we omit (*filt.*) in this paper for brevity, since all of our experiments are done with the publicly available filtered version.

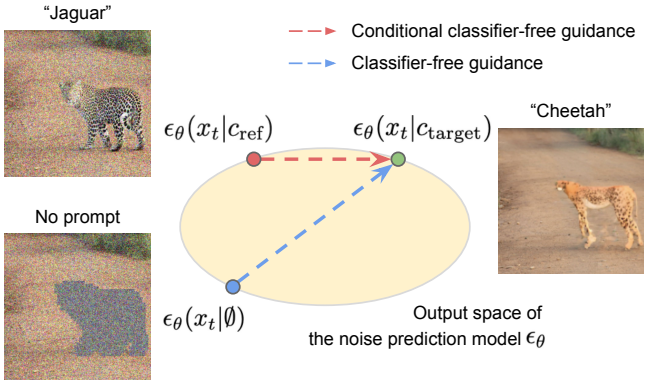

Fig. 3. Conceptual illustration of conditional classifier-free guidance. While the original classifier-free guidance (CFG) can be thought of guiding the denoising generative process from no input prompt, our conditional CFG starts from the referring prompt and can make the manipulation (empirically) easier on the noise prediction space ϵ_θ.

4.3 Conditional Classifier-free Guidance

Inspired by StyleCLIP [41] (global direction) and StyleGAN-NADA [13], we aim to guide the generative process along the direction of the source to the target. However, unlike StyleGAN [25], which has a well-analyzed latent embedding space [52], diffusion models currently do not have such correspondent. Nevertheless, we provide an intuitive modification to the classifier-free guidance for each time step in the (reverse) diffusion process, based on its geometric interpretation of extrapolating towards the noise prediction given a target caption.

Formally, recall the classifier-free guidance towards the caption c (Eq. (6), where $c = c_{\text{target}}$ in our problem setting). Instead of starting the guidance from an empty set $\emptyset$, we propose to guide the generative sampling process starting from our reference text prompt c_{ref} as follows:

$$\hat{\epsilon}_\theta(x_t|c) = \epsilon_\theta(x_t|c_{\text{ref}}) + s \cdot (\epsilon_\theta(x_t|c) - \epsilon_\theta(x_t|c_{\text{ref}})). \tag{8}$$

Intuitively, we can think of Eq. (8) as guiding the generative process along the direction towards the target expression *from the referring expression* on the joint (noisy) image-text embedding space, as illustrated in Fig. 3.

To align with the changes in our guidance direction, we also set the input to the inpainting model as the original input image, instead of the masked image as in the original GLIDE. We can roughly think of the original classifier-free guidance as generating a new object in a blank region (corresponding to the $\emptyset$ caption). However, since we have additional semantic information about the referring region with c_{ref} in our problem setting, conditioning on this knowledge is beneficial to the editing quality. The effects of the proposed term is more discussed in Sec. 5.3.

4.4 Localized output ranking with context

Many existing works on text-guided generative models [43] first synthesize a large number of samples and rank the generations using CLIP. Nichol *et al.* [37] suggests that CLIP re-ranking is not necessary when a model is trained with classifier-free guidance, but we have empirically found that the generated images with higher rankings are perceptually better than the low-ranked images and re-apply the output ranking scheme. Avrahami *et al.* [4] proposes to rank the final generated outputs with a pretrained CLIP model, similarly to [43,44]. However, they perform ranking only on the masked region, which can sometimes lead the model to generate a plausible region by itself but does not harmonize with the unmasked regions well. Thus, we propose to instead rank the final outputs *w.r.t.* the bounding box enlarged by a small ratio ($\times 1.3$ in practice), for localized ranking that also considers the surrounding context. Experimental results and the ranking effects are more discussed in Sec. 5.3.

4.5 Dilated mask prediction

The main problem that arises when using an automatically generated segmentation mask is that the mask prediction can be inaccurate. Especially, we have empirically found that the errors are much more critical when the mask does not cover the full object, compared to when the mask is covering the region larger than the object. Thus, we propose a simple heuristic of enlarging the predicted segmentation mask with a dilation operator, one of morphological transformations, to ensure that the mask better covers the referred object. This problem was not an issue for previous text-guided inpainting approaches, since a user-generated mask almost always covers the full object.

5 Experiments

5.1 Implementation details

We use PyTorch [40] for implementation. Since our framework does not require additional training, all results in this paper can be obtained with a single GPU (we used NVIDIA V100) or by simply using a hosted runtime on Colab [6]. The public GLIDE-inpainting model consists of two separate models: 64×64 inpainting diffusion model and 256×256 (inpainting-aware) upsampling model, and our proposed improvements are only applied to the 64×64 inpainting model. Following the setting in GLIDE, we used 100 diffusion steps in the inpainting model for fast sampling (instead of the full 1000 steps in DDPM [15]), and 27 steps for the upsampling model. For guidance scale s, we found that the default setting of $s = 5$ in the open-source GLIDE repository works well for the compared GLIDE baselines, but our method typically works better for a larger scale of $s = 15$. The code to reproduce our experimental results is publicly released[4] to facilitate future research on referring object manipulation.

[4] `https://github.com/google/referring-manipulation`

5.2 Comparisons

We compare the proposed framework with three baselines: 1) Blended-diffusion [4] (denote as 'Blended') and GLIDE with 2) CLIP guidance and 3) Classifier-free guidance (CFG). We use the images and captions from the PhraseCut dataset [51] for our comparisons, but occasionally modify the referring captions to a more salient object (or stuff) for better visualizations on our manipulation settings. We manually give the target text prompts to demonstrate new and interesting edits. For the compared models, the user-given mask inputs are substituted with the prediction from MDETR (and dilated). The overall qualitative results are summarized in Fig. 4.

In general, we found that CLIP-guided approach is susceptible to making adversarial examples that fool the CLIP model (as discussed in [14]). The Blended model is able to mitigate this issue by augmentations and generates high-quality edits, but sometimes shows imbalanced proportions between the masked region and the rest (also mentioned in [4]). CFG and our conditional CFG enables to remove CLIP during the diffusion steps and generates plausible edits most of the time, but the results using our conditional CFG is usually more realistic.

We also demonstrate more diverse generations *w.r.t.* each target text prompt in Fig. 5. Note that when there is no target prompt, the model performs inpainting and fills in the masked region from the surrounding context. Please refer to our supplementary materials for additional qualitative results and analyses in various different settings.

User studies For quantitative evaluation of the editing quality, we perform a human subjective test on 20 sample outputs, compared with Blended [4] and GLIDE (CLIP and CF-guided). In each testing case, we show the input image, the local region of interest, the target text, and 4 output edits including ours. The order of the display is randomized, and each participant is asked to rank the 4 outputs. A total of 60 users participated in this study, and the aggregated results are shown in Table 1. We found that no single model absolutely wins over the other, since all models have strong generation capabilities and give plausible outputs. However, our CCF-guided method shows the best average rank (best rank is 1, worst is 4), and our algorithm has 54.4% of winning probability when compared with the second best method of Blended, and 58.4% against the most similar baseline, CF-guided GLIDE. We report more detailed results in our supplementary document due to the page limit.

5.3 Ablation studies

Effects of guidance direction Given an input image and the segmentation mask estimated by MDETR, we compare the effects of the guidance direction of GLIDE in Fig. 6. Note that all results are obtained using *exactly the same values of the pretrained parameters* regardless of the guidance scheme. While all methods are capable of generating realistic outputs, our results tend to better keep the characteristics of the original image, while CF-guided GLIDE generates more diverse results. This is because CF-guided GLIDE model does not know

Fig. 4. Comparison between existing methods on text-guided image manipulation using images from PhraseCut dataset [51]. All models use the same input mask given by the output of MDETR [21]. Our conditional classifier-free guidance is able to make more visually pleasing edits that correctly follow the target text.

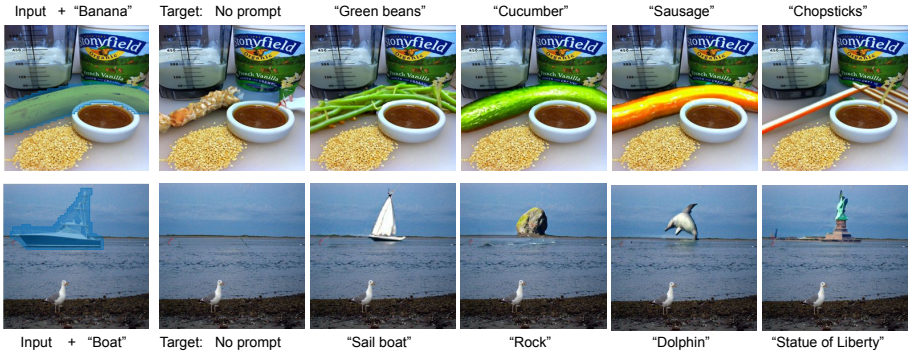

Fig. 5. Qualitative example for diverse target text queries. We use the guidance scale $s = 15.0$ for all methods. Interestingly, inpainting the bananas without any condition led to generating what looks like a shrimp tempura due to the dipping sauce next to it. We could also generate many interesting objects near the horizon.

Table 1. User study results. We report the average rank (1∼4) and the winning probability of a method in each row against the other models in each column.

Method	Avg. Rank$_\downarrow$	Winning prob. vs:			
		Blend	G-CLIP	G-CF	Ours
Blended	2.40	-	57.8%	51.4%	45.6%
GLIDE (CLIP-guided)	2.82	42.2%	-	46.6%	34.2%
GLIDE (CF-guided)	2.57	48.6%	57.4%	-	41.6%
Ours (CCF-guided)	2.20	54.4%	65.8%	58.4%	-

the masked-out region which our CCF-guided model knows, and each can be beneficial for its own use cases. Also, exploring which characteristic of the input image are preserved on the noise manifold of the diffusion process would make an interesting future work, which is out of scope of this paper.

Effects of localized ranking A qualitative comparison between the ranking method in Blended [4] and ours is shown in Fig. 7. Since the outputs are generated using the same guidance scheme with the same random seed, the total set of output images should be identical. However, the top-ranked results for the proposed localized ranking technique are usually more realistic and harmonize with the nearby context better.

Effects of mask dilation We show the effects of enlarging the intermediate segmentation mask with dilation in Fig. 8. If the predicted segmentation mask does not cover fully cover the object of interest, the remaining boundaries strongly affect how the model infers nearby context. This leads to generating a similar object category or some other unpleasing artifacts instead of removing the target object.

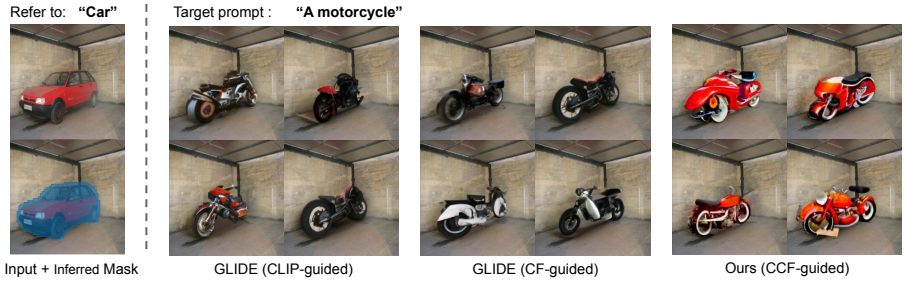

Fig. 6. The effects of different guidance methods. Four samples using different random seeds are shown for each guidance scheme. CLIP-guidance sometimes fails to generate the full object and shows only distinctive parts. While CF-guidance and Ours (CCF-guided) successfully synthesize the target object as a whole, ours tend to more keep the characteristics of the original input image, *e. g.* red color, unless otherwise guided by the target prompt.

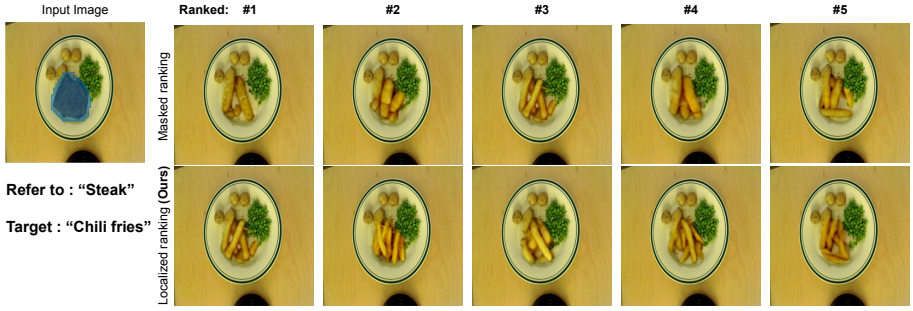

Fig. 7. The effects of different ranking mechanisms. While the set of total generated images for the first and the second rows are the same, the masked ranking tends to prefer relatively thicker potato-like objects, whereas our top-ranked outputs are thinner and match the target text better.

6 Limitations and Future Work

Although our proposed referring object manipulation framework with conditional classifier-free guidance generates plausible image edits, it still has several major limitations. First, at its current state, it cannot generate images of resolution other than 64×64 or 256×256, due to the constraint in the conditional diffusion model that we used. We believe that further research in conditional diffusion model can mitigate this issue. Second, our model cannot recover from a wrong segmentation output, because we sequentially combine the two separate models explicitly. Given the recent progress in vision-language transformers, we think that designing a fully end-to-end trainable architecture for referring object manipulation would also be an interesting direction for research. Third, the current model cannot perform very fine-grained manipulation, and the edited

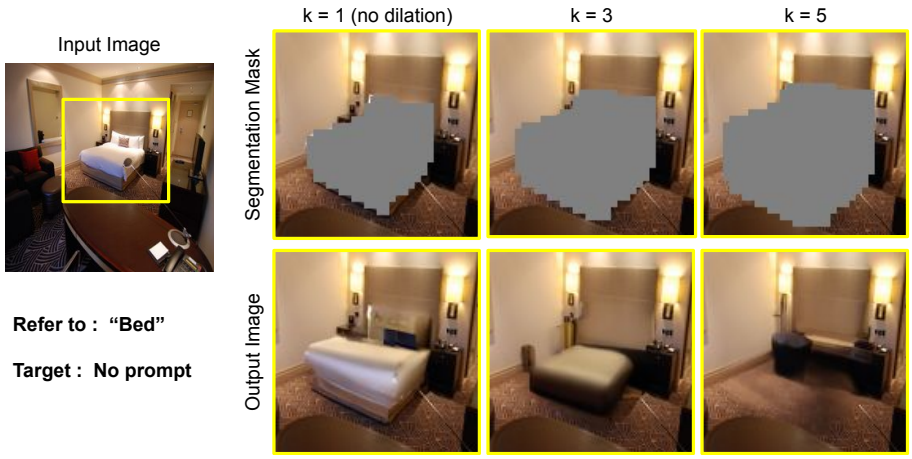

Fig. 8. The effects of mask dilation for inpainting (no target text). For $k = 1$ or $k = 3$, the model infers from the remaining white bed sheets or the bottom frame, which makes potentially unwanted artifacts. We show the enlarged region in the yellow box.

outputs for the referred regions are sometimes blurry. Even though we provide better conditioning on the input image and referred prompt, there still exists a lot of room for improvement in preserving the original image details and removing the boundary effects or artifacts. Also, building an easy-to-use editing tool enables even unskilled users to make fake imagery, which raises many safety concerns on potential bias and fairness of the model. The open-source model of GLIDE that we use has already considered safety issues in various aspects, but further effort will be required as a community to prevent any harmful use cases.

7 Conclusions

In this paper, we introduced a new problem of referring object manipulation and the first approach to address this task. The proposed framework combines a referring image segmentation method with a text-guided diffusion model and guides the generative diffusion process with a novel conditional classifier-free guidance scheme. We also proposed a new localized ranking method and mask dilation technique, which leads to visually more pleasing edits when combined together. As we demonstrate and analyze in the experiments, our model is capable of serving as a simple and effective baseline for referring object manipulation.

Acknowledgement We would like to thank Tobias Weyand, Fangting Xia, and Mikhail Sirotenko for helpful discussions, comments, and proofreading this work. This work is fully supported by Google Research.

References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: ICCV (2019)
2. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan++: How to edit the embedded images? In: CVPR (2020)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: ICCV (2015)
4. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. arXiv:2111.14818 (2021)
5. Bau, D., Andonian, A., Cui, A., Park, Y., Jahanian, A., Oliva, A., Torralba, A.: Paint by word. arXiv:2103.10951 (2021)
6. Bisong, E.: Google colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners, pp. 59–64. Apress, Berkeley, CA (2019), https://doi.org/10.1007/978-1-4842-4470-8_7
7. Crowson, K.: Clip guided diffusion 512x512, secondary model method. https://twitter.com/RiversHaveWings/status/1462859669454536711 (2021)
8. Crowson, K.: Clip guided diffusion hq 256x256. https://colab.research.google.com/drive/12a_Wrfi2_gwwAuN3VvMTwVMz9TfqctNj (2021)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021)
10. Ding, H., Liu, C., Wang, S., Jiang, X.: Vision-language transformer and query generation for referring segmentation. In: ICCV (2021)
11. Dong, H., Yu, S., Wu, C., Guo, Y.: Semantic image synthesis via adversarial learning. In: ICCV (2017)
12. Feng, G., Hu, Z., Zhang, L., Lu, H.: Encoder fusion network with co-attention embedding for referring image segmentation. In: CVPR (2021)
13. Gal, R., Patashnik, O., Maron, H., Chechik, G., Cohen-Or, D.: Stylegan-nada: Clip-guided domain adaptation of image generators. arXiv:2108.00946 (2021)
14. Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., Olah, C.: Multimodal neurons in artificial neural networks. Distill (2021), https://distill.pub/2021/multimodal-neurons
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
16. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021)
17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
18. Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: ECCV (2016)
19. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021)
20. Jing, Y., Kong, T., Wang, W., Wang, L., Li, L., Tan, T.: Locate then segment: A strong pipeline for referring image segmentation. In: CVPR (2021)
21. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: ICCV (2021)
22. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)

23. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv:1710.10196 (2017)
24. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
25. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: CVPR (2020)
26. Kim, G., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. arXiv:2110.02711 (2021)
27. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv:1312.6114 (2013)
28. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv:1411.2539 (2014)
29. Kocasari, U., Dirik, A., Tiftikci, M., Yanardag, P.: Stylemc: Multi-channel based fast text-guided image generation and manipulation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2022)
30. Li, B., Qi, X., Lukasiewicz, T., Torr, P.H.: Manigan: Text-guided image manipulation. In: CVPR (2020)
31. Li, R., Li, K., Kuo, Y.C., Shu, M., Qi, X., Shen, X., Jia, J.: Referring image segmentation via recurrent refinement networks. In: CVPR (2018)
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
33. Liu, C., Lin, Z., Shen, X., Yang, J., Lu, X., Yuille, A.: Recurrent multimodal interaction for referring image segmentation. In: ICCV (2017)
34. Liu, X., Gong, C., Wu, L., Zhang, S., Su, H., Liu, Q.: Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization. arXiv:2112.01573 (2021)
35. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv:1411.1784 (2014)
36. Nam, S., Kim, Y., Kim, S.J.: Text-adaptive generative adversarial networks: manipulating images with natural language. In: NeurIPS (2018)
37. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv:2112.10741 (2021)
38. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML (2021)
39. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729. IEEE (2008)
40. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. arXiv:1912.01703 (2019)
41. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: Styleclip: Text-driven manipulation of stylegan imagery. In: ICCV (2021)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv:2103.00020 (2021)
43. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML (2021)
44. Razavi, A., van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with VQ-VAE-2. arXiv:1906.00446 (2019)

45. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML (2016)
46. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018)
47. Shi, H., Li, H., Meng, F., Wu, Q.: Key-word-aware network for referring expression image segmentation. In: ECCV (2018)
48. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
49. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
50. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
51. Wu, C., Lin, Z., Cohen, S., Bui, T., Maji, S.: Phrasecut: Language-based image segmentation in the wild. In: CVPR (2020)
52. Wu, Z., Lischinski, D., Shechtman, E.: Stylespace analysis: Disentangled controls for stylegan image generation. In: CVPR (2021)
53. Xia, W., Yang, Y., Xue, J.H., Wu, B.: Tedigan: Text-guided diverse face image generation and manipulation. In: CVPR (2021)
54. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: CVPR (2018)
55. Xu, Z., Lin, T., Tang, H., Li, F., He, D., Sebe, N., Timofte, R., Van Gool, L., Ding, E.: Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model. arXiv:2111.13333 (2021)
56. Yang, S., Xia, M., Li, G., Zhou, H.Y., Yu, Y.: Bottom-up shift and reasoning for referring image segmentation. In: CVPR (2021)
57. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. arXiv:2112.02244 (2021)
58. Ye, L., Rochan, M., Liu, Z., Wang, Y.: Cross-modal self-attention network for referring image segmentation. In: CVPR (2019)
59. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: CVPR (2018)
60. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV (2017)
61. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan++: Realistic image synthesis with stacked generative adversarial networks. IEEE TPAMI **41**(8), 1947–1962 (2018)
62. Zhang, L., Chen, Q., Hu, B., Jiang, S.: Text-guided neural image inpainting. In: ACM MM (2020)
63. Zhang, T., Tseng, H.Y., Jiang, L., Yang, W., Lee, H., Essa, I.: Text as neural operator: Image manipulation by text instruction. In: ACM MM (2021)
64. Zhu, J., Shen, Y., Zhao, D., Zhou, B.: In-domain gan inversion for real image editing. In: ECCV (2020)