Referring Image Editing: Object-level Image Editing via Referring Expressions

Chang Liu[1,2,3] Xiangtai Li[2] Henghui Ding[1]✉
[1]Institute of Big Data, Fudan University, China [2]Nanyang Technological University, Singapore
[3]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore

Abstract

Significant advancements have been made in image editing with the recent advance of the Diffusion model. However, most of the current methods primarily focus on global or subject-level modifications, and often face limitations when it comes to editing specific objects when there are other objects coexisting in the scene, given solely textual prompts. In response to this challenge, we introduce an object-level generative task called Referring Image Editing (RIE), which enables the identification and editing of specific source objects in an image using text prompts. To tackle this task effectively, we propose a tailored framework called ReferDiffusion. It aims to disentangle input prompts into multiple embeddings and employs a mixed-supervised multi-stage training strategy. To facilitate further research in this domain, we introduce the RefCOCO-Edit dataset, comprising images, editing prompts, source object segmentation masks, and reference edited images for training and evaluation. Our extensive experiments demonstrate the effectiveness of our approach in identifying and editing target objects, while conventional general image editing and region-based image editing methods have difficulties in this challenging task.

1. Introduction

Referring image segmentation [23], aiming to segment the target object in the image indicated by an expression, is one of the most important vision-language problems in multi-modal information interaction. Among its many applications, one of the most expected is its utility in image editing. For example, by providing a language expression as input, the model can automatically pinpoint the region-of-interest for user to edit. While current models have excelled in the initial step of this process, *i.e.*, finding the target region, the subsequent step, image editing, lies beyond the conventional task definition of referring segmentation. In essence, existing referring segmentation frameworks focus on identifying the target object, while it lacks the inherent generative capability required for editing the image.

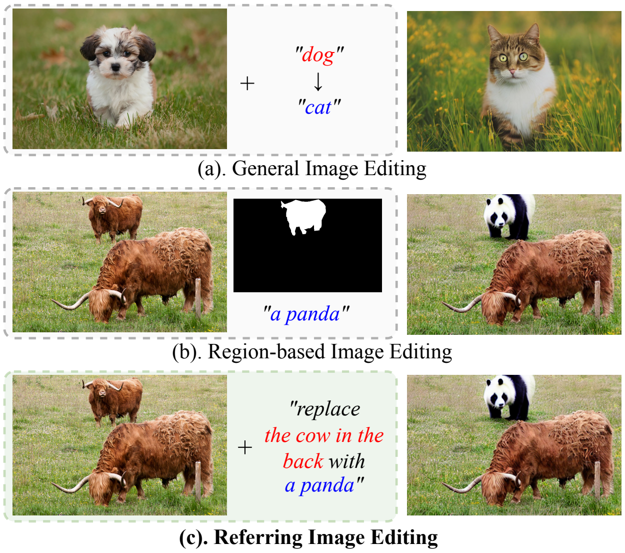

+
"dog"
↓
"cat"
(a). General Image Editing
"a panda"
(b). Region-based Image Editing
+
"replace the cow in the back with a panda"
(c). Referring Image Editing

Figure 1. Comparison of Referring Image Editing (RIE) with other image editing tasks: (a). General Image Editing, which edits images with a prominent subject or global-level edition instructions, (b). Region-based Image Editing, which requires masks or bounding boxes to specify the target region, and (c). **Referring Image Editing**, which aims to edit only the target object according to referring expressions while preserving the unrelated part unchanged.

On the other hand, the field of generative models has witnessed remarkable advancements in recent years. The emergence of text-to-image models like Diffusion [46] has led to a surge in research on image generation with multiple impressive functionalities. However, to the best of our knowledge, most of the current efforts in this domain, including general text-based and instruction-based methods [5, 11, 42] and prompt2prompt methods [4, 19, 21] concentrate on global-level prompts, like "*Change the style to Van Gogh*", or subject-level prompts, *e.g.* "*Replace the cow to a panda*", as shown in Fig. 1(a). When there are multiple objects in the scene and we only intend to edit some certain of them, our experiments reveal that these models struggle to handle such object-level prompts, like "*Replace the cow **in the back** with a panda*", which aim to edit one specific object, as shown in Fig. 1(c). While some region-

✉ Corresponding author (henghui.ding@gmail.com).

based image editing works [2, 3, 8, 11, 63] allow user to select a local region for editing, these approaches require users to manually specify the editing region using bounding boxes or masks, which are time-consuming to obtain and not as user-friendly as textual prompts. In essence, most current generative models emphasize the aspect of image generation and lack the inherent discriminatory capability for identifying a desired object within an image.

Referring Image Editing (RIE). From this point, we propose to extend the scope of current image editing methods by incorporating the capabilities of processing referring expressions. We introduce a novel task termed as *Referring Image Editing* (RIE), designed to encompass the entire pipeline of object-level image editing with a single text prompt. The input for RIE consists of an image and a text prompt, while the text prompt must include both a referring part and an editing instruction part, such as "*Replace the zebra on left with a giraffe*." In this task, referring image editing models should first identify the target object and then generate an image that incorporates the editing instruction, while preserving the remainder of the image unaltered. Importantly, the entire process is expected to be executed within a single model. Consequently, this task poses a formidable challenge, demanding the integration of both discrimination abilities to locate the target object for editing and generative abilities to modify the image based on the prompt, all encapsulated within a single model.

RefCOCO-Edit dataset & ReferDiffusion method. To facilitate the research of Referring Image Editing, based on the well-used referring expression dataset RefCOCO [68], we completes the expressions with edition instructions and build a referring image editing dataset, namely RefCOCO-Edit. Each sample within RefCOCO-Edit comprises four essential components: an image, a text prompt, a region-of-edition mask, and a reference output image. In addition to building this dataset, we introduce a novel method for referring image editing, termed as ReferDiffusion. ReferDiffusion is trained using both segmentation masks and reference output images, enabling it to perform natural image editing while retaining the ability to predict segmentation masks. In the context of referring image editing, our goal is to have the network modify a specific region of the image while leaving other regions untouched. To achieve this, we employ a contrastive supervision loss, which guides the network in making changes to the target region while preserving the integrity of the rest of the image. Our extensive experiments demonstrate that conventional image-to-image models, such as the standard Stable Diffusion [46], difficult to adapt to the fine-tuning requirements of the referring image editing task. In contrast, our framework exhibits strong generalization capabilities for this task, even with limited training data sizes.

In summary, the main contributions of this work are:

- We expand the current scope of referring segmentation by incorporating generative capabilities, introducing a novel task termed as Referring Image Editing (RIE).
- We introduce a new RefCOCO-Edit dataset to support the future research for RIE, comprising images, text prompts, masks, and reference edited images.
- We present a baseline method ReferDiffusion for the RIE task, along with a mixed-supervision training strategy to effectively train the model.
- Experiments show that general image-to-image methods fail to generalize effectively to this RIE task, while the proposed approach can effectively conduct the object-level image editing according to referring expressions.

2. Related Work

2.1. Referring Expression Segmentation

Referring expression segmentation (RES) [23] aims to segment the target object in the image indicated by an expression. Earlier works [12, 18, 24, 25, 31, 33, 36, 41, 65] mainly utilize Fully Convolutional Networks (FCN) and Recurrent Neural Networks (RNN) [9, 38, 40, 67] based methods is the mainstream for RES. Thanks to the great success of Transformer [56] in vision tasks [6, 15, 17, 20, 27, 32, 61], Transformer-based methods [13, 14, 16, 35] have greatly advance the performance of RES. Ding *et al.* [13, 16] first introduce Transformer to the field of referring expression segmentation, and propose a tailored Vision-Language Transformer (VLT) for RES. After VLT, more and more RES works adopt Transformer [26, 29, 59, 66]. Most recently, Zhu *et al.* [70] and Liu *et al.* [37] proposes to use polygon prediction to tackle the problem in a seq2seq manner. Besides, some recent works aim to extends the scope of referring segmentation, like Liu *et al.* [34] extends the classic referring segmentation with unconstrained object number [34], and Wu *et al.* [62] extends the task in terms of image numbers.

2.2. Diffusion Model

The diffusion model [22, 53, 54] is a kind of generative model, and can be seen as a Markov chain that contains a forward and a backward process. For an image input $\mathbf{z}_0$, the forward process whitens the given image into a Gaussian noise, while the backward process recovers the image from noise. Both forward and backward process consists of several time steps. Denote the input data in the t-th time step as $\mathbf{z}_t$. Each step of the forward process $q(\mathbf{z}_t|\mathbf{z}_{t-1})$ adds certain Gaussian noise to the input, making $\mathbf{z}_t$ more noisy as t increases. In reverse, the backward process is to find $\mathbf{z}_0$ given $\mathbf{z}_t$. As the property of Markov chain, each reverse step $q(\mathbf{z}_{t-1}|\mathbf{z}_t)$ should depends solely on the global distribution of $\mathbf{z}$ and consistent across all time steps. Therefore, we can use a neural network ϵ_θ to ap-

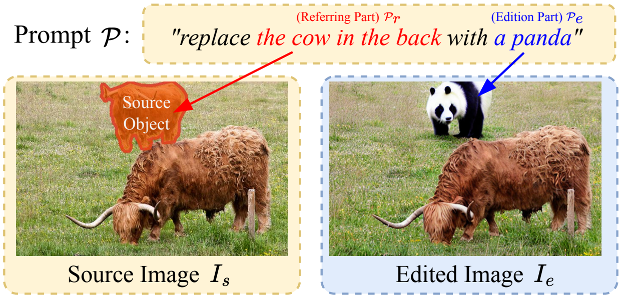

Prompt $\mathcal{P}$:
(Referring Part) $\mathcal{P}_r$
(Edition Part) $\mathcal{P}_e$
"replace the cow in the back with a panda"
Source Object
Source Image I_s
Edited Image I_e

Figure 2. The task of Referring Image Editing (RIE). The model takes an image I_s and a text prompt $\mathcal{P}$ as input, and outputs an edited image I_e. RIE aims to edit only the specific parts of existing real images according to textual prompts.

proximate each denoising step. Recently, the Latent Diffusion [22] proposes to perform the diffusion process in a latent space. The network has an Variational Auto-Encoder (VAE) $\mathcal{E}$ to converts the image into a lower-resolution *latent space*, then add a decoder $\mathcal{D}$ at the end, which recovers the latent feature back to image. The framework also uses a modified U-Net with added cross-attention layers as ϵ_θ, introducing text-based conditional inputs for the diffusion process. As the diffusion process operates in the latent space with lower resolution, it greatly reduces the overall computational complexity. Therefore, the diffusion model has also gain huge attention in many industry fields such as image synthesis [44, 46], image inpainting [39], image segmentation [57, 58], video generation [52, 64], image super-resolution [1, 49], and deblurring [45, 60]. However, most of these works are focused only on pure generating performance of the networks, which neglects the discriminating capability. In our work, we aim to improve the discrimination performance networks.

3. Referring Image Editing

3.1. Problem Definition

In this paper, we introduce an object-level image editing task, namely Referring Image Editing (RIE). As shown in Fig. 2 and introduced in Sec. 1, RIE's input consists of two parts: source image I_s and text prompt $\mathcal{P}$ specifying the target object and editing requirements; while it outputs an image I_e edited according to the prompt.

The prompt $\mathcal{P}$ encompasses two parts: a referring part $\mathcal{P}_r$ that indicates the source object, and an edition part $\mathcal{P}_e$ that provides descriptions on how the image should be edited, like "*replace the cow in the back with a panda*" or "*remove all guys in the background*". The two parts should be naturally combined into a single sentence, which will be inputted into models as a whole expression. Additionally, it is worth noting that the editing region is not strictly confined to the segmentation mask of the source objects; instead, models are expected to generate natural and meaningful images based on the provided prompt.

Difference with other image-to-image tasks. One major feature of RIE is that it emphasis on challenging the targeting and identifying ability of the model, like referring image segmentation. Compared with region-based image editing [2, 3, 63] or inpainting [39, 48] that require masks or bounding boxes to indicate the target-of-interest, RIE eliminates the need for additional manual input from users, making it a more intuitive and user-friendly approach. Furthermore, unlike general text-driven or subject-driven image editing methods [4, 30, 55] that mainly focus on images with a prominent "*subject*" or are capable only of global-level manipulations like style transfer, RIE cares about scenarios where users seek to edit only the specific parts of existing real images containing multiple objects by textual prompts. In essence, RIE places a strong emphasis on discrimination capabilities while also retaining generative abilities, making it challenging. In Sec. 4.1, we investigate why existing approaches fail to handle this task.

3.2. RefCOCO-Edit Dataset

We construct a referring image editing dataset, termed as RefCOCO-Edit. This dataset is composed of 400 images carefully selected from the widely recognized referring dataset, RefCOCO [34, 68]. Most image has at least 2 editing samples. For every sample in RefCOCO-Edit, besides providing a source image I_s, a prompt $\mathcal{P}$, and a reference edited output (serves as ground-truth) I_e, we also include the mask M of the source object and a mask M_e that indicates the rough edited area in ground-truth for evaluation. For text prompts, We also explicitly points out the word belongings of the referring part $\mathcal{P}_r$ and the edition part $\mathcal{P}_e$ to assist model training.

The annotation process involves three key steps: source object selection, prompt creation, and reference image generation. In the following, we will provide details of the annotation procedures and requirements.

Source object selection. One great feature of referring image editing is that it requires model's discriminative capabilities. Annotators are free to find the source object, but we adhere to two specific rules:

1), The source object cannot be the *sole prominent subject* of the image. This prevents overly simplistic cases, and also distinguishes RIE from general image editing tasks [10, 30].

2), There must exist an other object that is similar to the source object in the scene, either in terms of semantic class or visual characteristics, such as color and shape. This further challenges the discrimination abilities of the models.

Prompt creation. We employ expressions from referring datasets as $\mathcal{P}_r$, and do not impose specific constraints on the content of $\mathcal{P}_e$. The editing goal can be similar to the original object, *e.g.*, "*swap the man in blue jacket with red jacket*", or "*remove the dog in the background*". We also

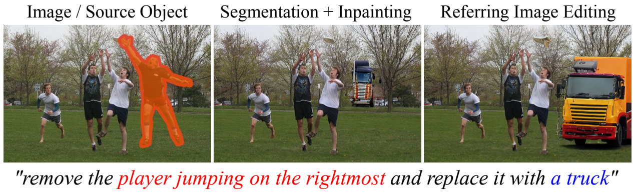


Image / Source Object
Segmentation + Inpainting
Referring Image Editing
"remove the player jumping on the rightmost and replace it with a truck"

Figure 3. The comparison of Referring Image Editing and the combination of referring segmentation and image inpainting.

allow prompts involving fictional and imaginative scenarios, such as "*change the ship in the river with an airplane*". The composite text prompt is structured flexibly in a unconstrained manner. We also employ Large Language Models (LLMs) to enrich the variety of the prompts, nevertheless, all prompts are manually reviewed to ensure the validity.

Reference ground-truth generation. We employ existing generative models with manual input and selection for generating the reference ground-truth. We establish three pipelines, in which annotators can choose the most suitable one or multiple pipelines to find good reference outputs: **1)** Image composition methods with reference substitution objects. **2)** Region-based image editing models with manually specified edition regions. **3)** Image inpainting methods for removing objects from the scene. All reference outputs are undergone manual validation as well.

4. Approach

4.1. Existing Approaches and Motivation

As the existing diffusion models can achieve many functionalities, it is intuitive that one may ask whether the existing methods are capable to create a naive solution for RIE. However, most existing methods exhibit their own limitations and are not directly applicable to the task.

Firstly, our experiments in Sec. 5 show that general image editing models, like instructions-based models [4, 19, 55], fail to directly edit the input image using referring image editing prompts, even after fine-tuning on the RefCOCO-Edit dataset. This may because the fact that they are originally designed and trained for generating or editing only the sole prominent *subject* in the image, making it difficult to control the model to find and edit only the desired region. *I.e.*, they lack discrimintive abilities. Another intuitive approach [69] involves using an segmentation model to identify the mask of the source object, and then employing region-based editing or inpainting methods like [8, 28, 63]. However for this kind of methods, one major problem is that the shape of the editing goal and the source object can often be very different. For instance in Fig. 3, when replacing a "*player*" with a "*truck*", the inpainting model, which neither aware of the truck's shape nor able to adjust the player's mask, can only insert a "*truck*" within the player's mask, producing an rather unnatural output.

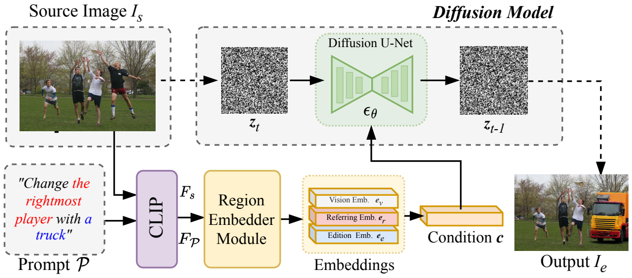


Figure 4. Network Architecture of the proposed **ReferDiffusion**.

Hence, two major challenges in RIE stand out: firstly, training the discriminative model that can locate certain objects while keeping the generative ability, and secondly, decoupling the region of "removal" and the region to "paint" for natural outputs. To address these challenges, we introduce a baseline approach, ReferDiffusion, for the RIE task. It features a Region-Embedder Module for generating separate embeddings for the referring region and the editing region, and a Region-control Loss, which serves to guide the network in keeping the background unaltered. Additionally, we propose a mix-supervised training strategy that trains the model gently by focusing on one task at a time.

4.2. ReferDiffusion

The overall architecture is shown in Fig. 4. Given the source image I_s and text prompt $\mathcal{P}$, the network initially converts I_s into latent space using an encoder $\mathcal{E}$. Simultaneously, it extracts text features $F_{\mathcal{P}}$ from $\mathcal{P}$. Next, the Region-embedder Module processes F_s and $F_{\mathcal{P}}$ to generate a set of embeddings: the vision embedding $\mathbf{e}_v$, referring embedding $\mathbf{e}_r$, and editing embedding $\mathbf{e}_e$. These embeddings are then fused into the final condition vector $\mathbf{c}$. Then like regular diffusion models, the network ϵ_θ iteratively operates for T steps, progressively denoises the latent $\mathbf{z}_T$. Finally, the decoder $\mathcal{D}$ decodes the denoised feature $\mathbf{z}_0$ into the edited output image. The model is trained using the Classifier-free Guidance for two conditions following [4] using the image condition $\mathbf{c}_I$ and the aforementioned condition vector $\mathbf{c}$:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(x),\mathbf{c}_{\mathcal{I}},\mathbf{c},\epsilon\sim\mathcal{N}(0,1),t}\left[||\epsilon - \epsilon_\theta\left(\mathbf{z}_t, t, \mathcal{E}(\mathbf{c}_I), \mathbf{c}\right)||_2^2\right]. \tag{1}$$

Region-Embedder Module (REM). In most general text-to-image diffusion models [46], the condition vector relies solely on the text prompt $\mathcal{P}$. While this suffices for describing global-level image themes or content, referring image editing poses a unique challenge. As previously discussed, one of the challenges is that the location and shape of the source object may differ from the editing goal. To address this, we propose to disassemble the condition vector $\mathbf{c}$ into multiple facets and introduce the Region-embedder module (REM). The REM takes two inputs: the encoded source image feature $F_s \in \mathbb{R}^{H\times W\times C_e}$ and the prompt embedding $F_{\mathcal{P}} \in \mathbb{R}^{N_e\times C_e}$, where C_e denotes the number of

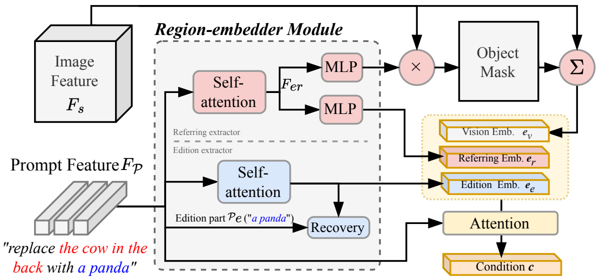


Figure 5. Architecture of the Region-Embedder Module.

channels, and N_e is the number of tokens in CLIP. The REM separately extracts three embeddings from the input, each focusing on a unique perspective, as illustrated in Fig. 5.

REM comprises two components: a referring embedding extractor and an edition embedding extractor. The edition embedding extractor focuses on locating and recovering the prompt embedding of the edition part $\mathcal{P}e$. This process is concise: it employs a self-attention layer on $F_{\mathcal{P}}$ and subsequently performs global average pooling on the resulting vector to obtain the edition embedding $\mathbf{e}_e$. $\mathbf{e}_e$ is supervised using the average of the features of the edition part of the prompt $\mathcal{P}_e$, ensuring it contains information about the edition goal. On the other hand, the referring embedding extractor also employs a self-attention layer on $F_{\mathcal{P}}$, and takes the global average pooling of the derived vector as $F_{er} \in \mathbb{R}^{1\times C_e}$. Subsequently, it splits into two pathways: One pathway is used to find the mask of the source object. By treating the MLP transformed F_{er} as a filter and apply it on the image feature, it predicts the mask of the source object: $\hat{M} = F_s$ MLP$(F_{er})^T$. This prediction is supervised using the ground-truth mask of the source object, guiding F_{er} to learn the location of the source object. Additionally, we aggregate vision features from F_s using $\hat{M}$ to create a vision embedding $\mathbf{e}_v$, providing vision information about the source object. The other pathway applies an extra MLP on F_{er}, which contains rich location information about the source object, to generate the referring embedding $\mathbf{e}_r$.

Finally, the three embeddings, $\mathbf{e}_v, \mathbf{e}_r, \mathbf{e}_e$, are fused together using a cross-attention layer. These three condition embeddings, along with the prompt embedding $F_{\mathcal{P}}$, are concatenated in the token dimension and serve as the key and value inputs of the cross-attention layer. $F_{\mathcal{P}}$ also serves as the query input of the cross-attention layer, facilitating the integration of information from the prompt and all derived embeddings into the final condition vector $\mathbf{c}$. This condition vector is then used for the diffusion process.

Region-Control Loss (RCL). In the context of referring image editing, our goal is to modify only the necessary parts of the image while leaving the rest of the pixels unchanged. To achieve this, we introduce the Region-control Loss, which leverages the mask of the edited area M_e from the RefCOCO-Edit dataset. Inspired by DragGAN [43, 51], we apply latent supervision to pixels outside the edited area at each step of the diffusion process. The Region-control Loss, denoted as L_r, is calculated as follows:

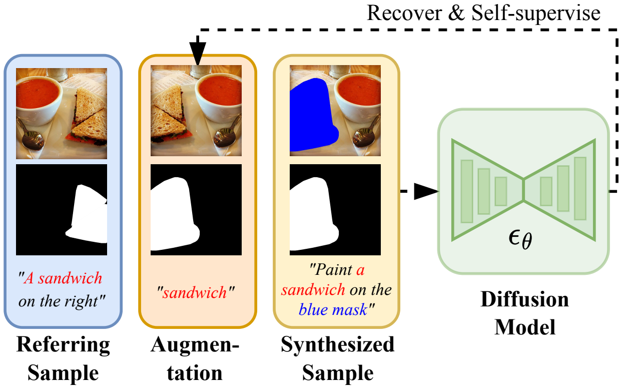


Figure 6. The self-supervised training pipeline.

$$L_r = \|\mathbf{z}_t - \text{sg}[\mathbf{z}_0'] \odot (\mathbb{1} - M_e)\|_1, \tag{2}$$

where $\mathbf{z}_0'$ is the latent without noise of the image, and sg is the stop-gradient operator. In essence, this loss encourages the network to retain the original latent for regions that are not intended to be modified, as indicated by the mask M_e. This helps the model to keep the background area unchanged, while providing hints about the area to edit, further facilitating the model to learn how to decouple the referring region and editing region.

4.3. Training Strategy

As previously mentioned, one major challenge for RIE is to train a model that simultaneously excels at two tasks that are of very different natures: strong discrimination, and controlled generation. Besides, another common obstacle for generative models training is that the data available for training is usually limited. To address these challenges, we propose a mixed-supervised multi-stage training strategy. It effectively helps the model to focus on one task at a time, while also enabling the synthesis of training data.

In the first stage, we propose to synthesis training samples from referring expression datasets for self-supervision, as shown in Fig. 6. Given a referring segmentation sample containing an image, an object mask, and a referring expression, we firstly expand the object mask, and cover the source object with a random color. Simultaneously, we extract the subject name from the expression, *e.g.*, *"sandwich"* for *"a sandwich on the right"*. We then use the mask as the referring part and the subject name as the edition part to generate a new RIE prompt, *e.g.*, *"Paint a sandwich on the blue mask"*. Augmentations like random cropping, flipping, and LLM rephrasing are applied to enhance data variety. In this initial stage, the synthesized source object is conspicuous in the image and very easy to locate, so the model can focus on learning how to draw contents in the desired region without altering other unrelated pixels.

Table 1. Comparison of the metrics of ReferDiffusion and other image editing methods.

Method	AP(↑)	L2$_{bg}$(↓)	CLIP-Sim$_{fg}$(↑)
Region-based Editing			
Blend Diffusion [3]	4.47	-	23.40
General Instruction-based Editing			
InstructPix2Pix [4]	4.58	0.031	23.72
InstructDiffusion [19]	4.50	0.044	24.07
MagicBrush [55]	**4.62**	0.034	24.32
ReferDiffusion (ours)	4.60	**0.028**	**25.06**

In the second stage, we train the model on the proposed RefCOCO-Edit dataset using full supervision to teach it the complete pipeline of the RIE task, including locating the target object and editing the image.

5. Experiment

5.1. Implementation Details

We base our model on the widely-used latent diffusion model, the Stable Diffusion v1.5 [22] with the general editing model [4]. During training, the spatial size of images are fixed at 480×480, and the learning rate is fixed at 5e-5. The model is trained on 8 NVIDIA V100 GPUs for 2 days. We follow the InstructPix2Pix [4] for other settings. During inference, we use DDIM as the sampler and the number of steps is set to 50. For reference image generation of RefCOCO-Edit, we employ the Paint-by-Example [63] as composition method, and Blended Latent Diffusion [3] as the region-based image editing and inpainting method.

5.2. Evaluation

While the RefCOCO-Edit dataset provides reference edited outputs (ground-truth), our prompts intentionally not to give very much details for the desired editing goal, such as color and shape requirements of the target object, granting models more creative freedom. Consequently, direct pixel-level comparisons between reference ground-truth and generated results, particularly for the edited region, are not appropriate. Instead, we employ three metrics for quantitative model assessment: **1).** the Aesthetic Predictor's Score (AP)[50] to gauge overall image quality, **2).** L2 score[55] of the background (L2$_{bg}$) to assess background consistency, **3).** CLIP-Similarity [7, 47] of the foreground (CLIP-Sim$_{fg}$) to measure the similarity between the edited part and the referring part $\mathcal{P}_r$ in the prompt.

Under this evaluation, when the model fails to edit the region of the source object accordingly with the edition instruction, the CLIP score will decrease. Additionally, if the model erroneously edits regions other than the source object, such as any other undesired objects or regions that are not referred in the prompt, the L2 score will be affected. We employ the reference edition mask M_e from the RefCOCO-Edit dataset for distinguishing foreground and background regions during the evaluation.

Table 2. Ablation study of design components of ReferDiffusion.

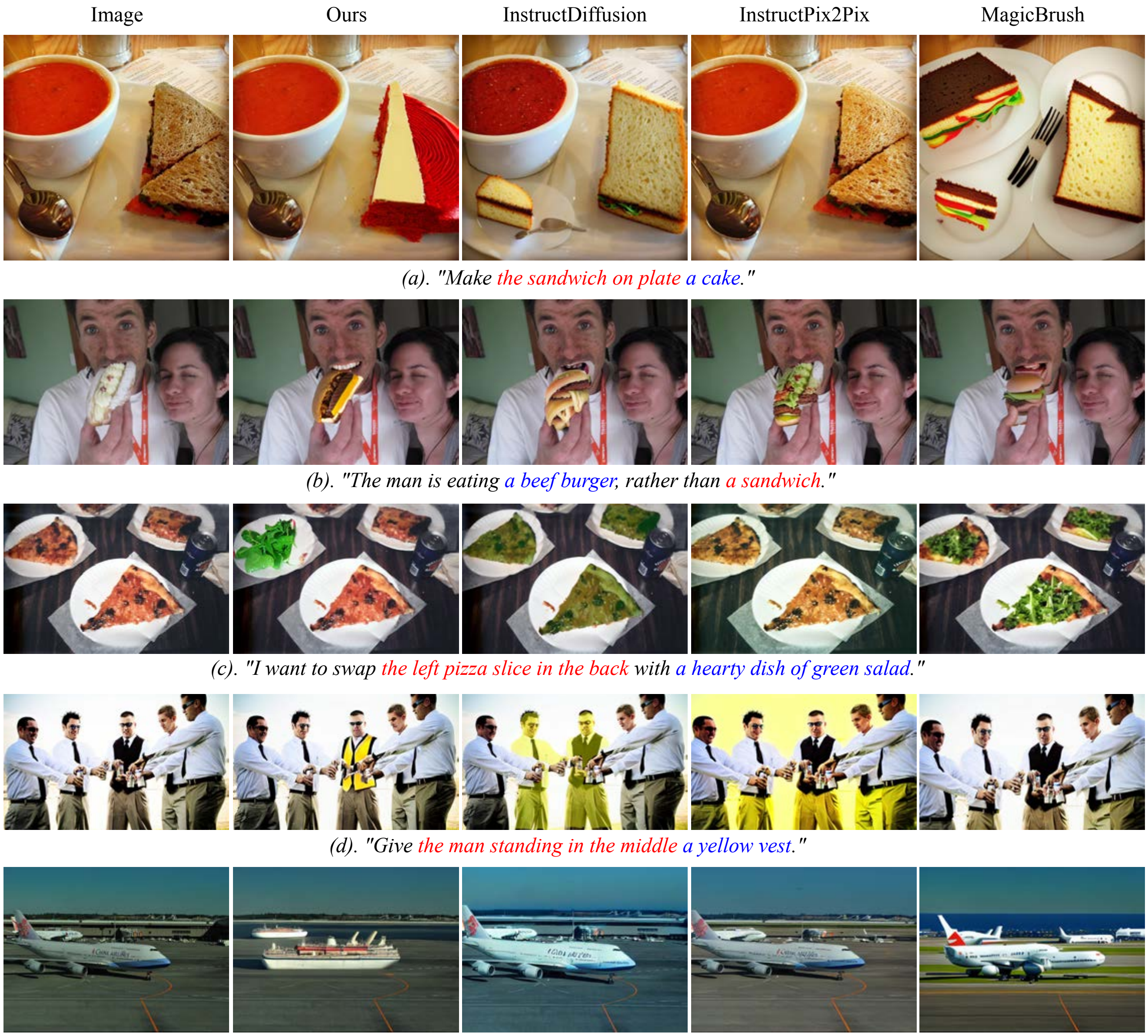
Method	AP(↑)	L2$_{bg}$(↓)	CLIP-Sim$_{fg}$(↑)
Fine-tune	4.56	0.049	24.28
+ Region-Embedder	4.64	0.032	24.17
+ Region-Control Loss	4.58	0.026	24.72
+ Self-Sup (ours)	4.60	0.028	25.06

5.3. Quantitative Results

Comparison. We compare our method with five other models, including a region-based image editing methods: Blend Latent Diffusion [3], three general image editing models: Instruct Pix2Pix [4], InstructDiffusion [19], and MagicBrush [28], which is based on InstructPix2Pix but fine-tuned on extra data. For Blend Latent Diffusion, we use the ground-truth mask of the source object as the input and skip the background score, while use editing part $\mathcal{P}_e$ of $\mathcal{P}$ as the prompt input. For the two instruction-based methods, following [19], we add suffix *"and do not change any pixel else"* at the end of the prompt to guide the model to preserve the background, *e.g*. *"Replace the rightmost player with a zebra and do not change any pixel else"*. We use default settings and random random seeds with a fully-automated pipeline for all methods. The results are shown in Tab. 1.

From the table, traditional region-based editing methods struggle to produce good results, even when provided with ground-truth masks of the source object and explicit editing instructions. This limitation arises because these region-based methods can only edit within the predefined mask of the source object, without considering the context or guidance from the prompt, resulting in unnatural and unsatisfactory outputs. On the other hand, as we use the same generation backbone as general instruction-based methods, though they achieve competitive Aesthetic Predictor scores (AP), they fall short in terms of background consistency (L2$_{bg}$) and foreground similarity (CLIP-Sim$_{fg}$) compared to our proposed ReferDiffusion model. This discrepancy suggests that these methods often cannot accurately locate the target object, leading to erroneous edits on unrelated objects or failing to edit the intended target region. These findings underscore the unique challenges posed by the Referring Image Editing (RIE) task and the need for dedicated models that can effectively address them.

Ablation Study. We test different modules of our model to prove their effectiveness. We start from the baseline model, which is the plain InstructPix2Pix model find-tuned on the RefCOCO-Edit dataset. Then we test the baseline model added with the Region-Embedder Module (REM) and the model with the Region-Control Loss (RCL). Finally we test the full model with the mixed-supervision training strategy. The results are shown in Tab. 2. It can be seen that both REM and self-supervision training (Self-Sup) setting



(a). "Make the sandwich on plate a cake."

(b). "The man is eating a beef burger, rather than a sandwich."

(c). "I want to swap the left pizza slice in the back with a hearty dish of green salad."

(d). "Give the man standing in the middle a yellow vest."

(e). "turn planes as ships in an airport-ocean"

Figure 7. Qualitative comparison of ReferDiffusion with other general instruction-based image editing methods.

contribute to the overall performance of the model, including L2 and CLIP, showing that both design enhances the discrimination ability of the model. RCL further contributes to the L2 score, making the background more consistent.

5.4. Qualitative Results

Comparison with other methods. We present a qualitative comparison of our method with three instruction-based approaches, InstructDiffusion (ID) [19] InstructPix2Pix (IP) [4], and MagicBrush (MB) [55] in Fig. 7. We use the same input and prompts as in the quantitative experiments for these two methods in this section. In the first two samples, we assess the general image generation ability of the models. In sample (a), the model should turn a sandwich into a cake. IP fails to do so, while ID and MB misunderstands the prompt by painting a new sandwich on top of the original one. IP also changes the spoon on the plate to a cake. This shows that the general instruction-based models cannot do well when finding some specific regions for editing. In sample (b), where the model is required to change a sandwich into a burger. All methods identified the correct region but produce different quality results. ID and IP both hard to fully eliminate the original sandwich, generating hybrid images that are part-burger

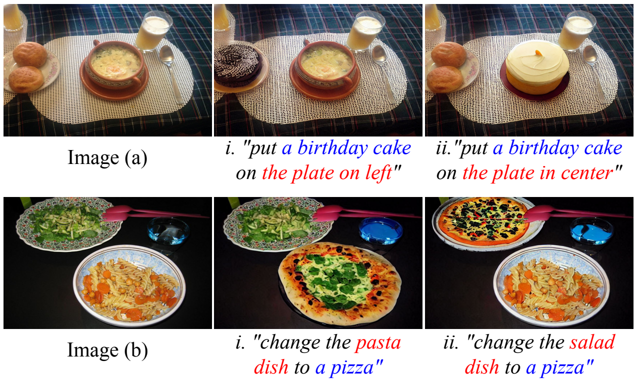


Image (a)

i. "put a birthday cake on the plate on left"

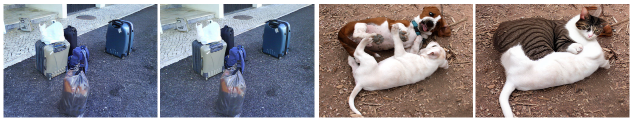
ii."put a birthday cake on the plate in center"

Image (b)

i. "change the pasta dish to a pizza"

ii. "change the salad dish to a pizza"

Figure 8. Demonstration about the referring performance of ReferDiffusion.



(a). "change the plastic bag on the suitcase with a backpack"

(b). "transform the brown dog on top with a cat"

Figure 9. Failure cases of ReferDiffusion.

and part-sandwich, while MagicBrush produces not so good details on the man's mouth.

Image (c) and (d) illustrate the referring capability, specifically the model's ability to locate the correct source object within the scene. Samples (d) and (e) involve multiple objects, the model needs to identify the correct object to edit. It appears that InstructDiffusion misinterprets both tasks as segmentation tasks, producing segmentation masks for some objects. Furthermore, sample (e) demonstrates a fictional scene and prompt, in which the model is required to paint an *"airport-ocean"*. InstructDiffusion and InstructPix2Pix fail to understand the prompt, while MagicBrush only partially edits the airport to an ocean, but does not change the airplane to ships.

Given that the prompts in samples (b) and (c) are lengthy and complicated and sample (e) involves some novel and rare concepts, it is possible that generalist models like InstructDiffusion cannot fully comprehend such complex prompts. In RIE, where at least two objects are involved in the prompt, such intricate prompts are common. Therefore, it is crucial for models to learn how to disentangle and decipher these complex instructions, and try to understand those unseen concepts. Additionally, it is observed that MagicBrush tends to edit all objects in sample (a) and (c), which further supports our argument that general text-driven image editing methods lack discriminative abilities necessary for accurately identifying and manipulating objects within images for the RIE task.

Referring Performance. We demonstrate the referring performance of the model in Fig. 8. In these tests, we change the source object but only with modifying a few words in the referring part of the prompt, and see whether the model is sensitivity to the changes. It can be seen that the model is able to understand semantic clues ("*pasta/salad*"), and find the correct source object in the image and edit it accordingly.

Table 3. Results of the user study.

Method	Quality(↑)	Consistency(↑)
InstructPix2Pix [4]	3.09	3.02
InstructDiffusion [19]	**3.14**	3.15
MagicBrush [55]	3.11	3.27
ReferDiffusion (ours)	3.12	**3.63**

Failure cases. Some failure cases are shown in Fig. 9. As our model is built upon the general Diffusion model, which is primarily designed for image generation, its performance in target identification is relatively limited compared to specialized referring segmentation models, particularly when faced with lengthy and complex prompts, or high visual reasoning requirements like involving multiple instances with complicated interactions in one prompt. Also, the quality of generated images depends on the backbone model and synthesized dataset, occasionally resulting in unnatural outputs when editing challenging scenes.

5.5. User Study

We further conduct a user study to evaluate the proposed method. Following previous works [55, 63], our study include two questions from different aspects: image quality and the consistency of the generated image with the prompt. For each testing sample, we provide the source image, the editing prompt and the edited image of each method to 50 participants from different backgrounds. The participants are asked to rate the generated image from 1 to 5, from the worst to the best. The average ratings are shown in Tab. 3. Our method achieves very competitive quality score and the best consistency score with large margin, showing its great discriminative and generative capabilities.

6. Conclusion

We introduce the novel task of Referring Image Editing (RIE), extending the capabilities of referring expression processing with generative abilities. We build RefCOCO-Edit, a dataset to support RIE research, containing images, text prompts, target masks, and reference edited images. To tackle the challenges of RIE, we propose ReferDiffusion, a novel framework that effectively combines diffusion models with text prompts. Our experiments demonstrate that ReferDiffusion outperformes general image-to-image models and existing instruction-based methods, highlighting its effectiveness in handling this complex task. RIE offers promising applications in user-friendly and natural image editing, where models can intelligently locate target objects and perform precise edits based on text instructions.

References

[1] Johannes Ackermann and Minjun Li. High-resolution image editing via multi-stage blended diffusion. *arXiv:2210.12965*, 2022. 3

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, pages 18208–18218, 2022. 2, 3

[3] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42 (4):1–11, 2023. 2, 3, 6

[4] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 1, 3, 4, 6, 7, 8

[5] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, pages 22560–22570, 2023. 1

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 2

[7] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Rui, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *arXiv:2304.00186*, 2023. 6

[8] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. *arXiv:2307.09481*, 2023. 2, 4

[9] Y.-W. Chen, Y.-H. Tsai, T. Wang, Y.-Y. Lin, and M.-H. Yang. Referring expression object segmentation with caption-aware consistency. In *BMVC*, 2019. 2

[10] Jooyoung Choi, Yunjey Choi, Yunji Kim, Junho Kim, and Sungroh Yoon. Custom-edit: Text-guided image editing with customized diffusion models. *arXiv:2305.15779*, 2023. 3

[11] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv:2210.11427*, 2022. 1, 2

[12] Henghui Ding, Scott Cohen, Brian Price, and Xudong Jiang. Phraseclick: toward achieving flexible interactive segmentation by phrase and click. In *ECCV*, pages 417–435. Springer, 2020. 2

[13] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, pages 16321–16330, 2021. 2

[14] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, 2023. 2

[15] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. MOSE: A new dataset for video object segmentation in complex scenes. In *ICCV*, 2023. 2

[16] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. VLT: Vision-language transformer and query generation for referring segmentation. *IEEE TPAMI*, 2023. 2

[17] Deng-Ping Fan, Ge-Peng Ji, Peng Xu, Ming-Ming Cheng, Christos Sakaridis, and Luc Van Gool. Advances in deep concealed scene understanding. *Visual Intelligence*, 1(1):16, 2023. 2

[18] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *CVPR*, 2021. 2

[19] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Han Hu, Dong Chen, et al. Instructdiffusion: A generalist modeling interface for vision tasks. *arXiv:2309.03895*, 2023. 1, 4, 6, 7, 8

[20] Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. GREC: Generalized referring expression comprehension. *arXiv preprint arXiv:2308.16182*, 2023. 2

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

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 2, 3, 6

[23] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, pages 108–124. Springer, 2016. 1, 2

[24] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *CVPR*, pages 4424–4433, 2020. 2

[25] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In *ECCV*, pages 59–75. Springer, 2020. 2

[26] Kanishk Jain and Vineet Gandhi. Comprehensive multi-modal interactions for referring image segmentation. *arXiv:2104.10412*, 2021. 2

[27] Aishwarya Kamath, Mannat Singh, Yann LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. Mdetr – modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 2

[28] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, pages 6007–6017, 2023. 4, 6

[29] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *CVPR*, pages 18145–18154, 2022. 2

[30] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *arXiv:2305.14720*, 2023. 3

[31] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *CVPR*, pages 5745–5753, 2018. 2

[32] Xiangtai Li, Henghui Ding, Wenwei Zhang, Haobo Yuan, Jiangmiao Pang, Guangliang Cheng, Kai Chen, Ziwei Liu, and Chen Change Loy. Transformer-based visual segmentation: A survey. *arXiv preprint arXiv:2304.09854*, 2023. 2

[33] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *ICCV*, pages 1271–1280, 2017. 2

[34] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: generalized referring expression segmentation. In *CVPR*, pages 23592–23601, 2023. 2, 3

[35] Chang Liu, Henghui Ding, Yulun Zhang, and Xudong Jiang. Multi-modal mutual attention and iterative interaction for referring image segmentation. *IEEE TIP*, 32:3054–3065, 2023. 2

[36] Chang Liu, Xudong Jiang, and Henghui Ding. Instance-specific feature propagation for referring segmentation. *IEEE TMM*, 25:3657–3667, 2023. 2

[37] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *CVPR*, pages 18653–18663, 2023. 2

[38] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 2

[39] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, pages 11461–11471, 2022. 3

[40] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*, pages 10034–10043, 2020. 2

[41] Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *ECCV*, pages 630–645, 2018. 2

[42] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, pages 6038–6047, 2023. 1

[43] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *SIGGRAPH*, pages 1–11, 2023. 5

[44] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831. PMLR, 2021. 3

[45] Mengwei Ren, Mauricio Delbracio, Hossein Talebi, Guido Gerig, and Peyman Milanfar. Image deblurring with domain generalizable diffusion models. *arXiv:2212.01789*, 2022. 3

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2, 3, 4

[47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 6

[48] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *SIGGRAPH*, pages 1–10, 2022. 3

[49] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE TPAMI*, 2022. 3

[50] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35:25278–25294, 2022. 6

[51] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv:2306.14435*, 2023. 5

[52] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv:2209.14792*, 2022. 3

[53] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265. PMLR, 2015. 2

[54] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *NeurIPS*, 2019. 2

[55] Xinghai Sun, Changhu Wang, Avneesh Sud, Chao Xu, and Lei Zhang. Magicbrush: Image search by color sketch. In *ACM MM*, pages 475–476, 2013. 3, 4, 6, 7, 8

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 2

[57] Chaoyang Wang, Xiangtai Li, Henghui Ding, Lu Qi, Jiangning Zhang, Yunhai Tong, Chen Change Loy, and Shuicheng Yan. Explore in-context segmentation via latent diffusion models. *arXiv:2403.09616*, 2024. 3

[58] Mengyu Wang, Henghui Ding, Jun Hao Liew, Jiajun Liu, Yao Zhao, and Yunchao Wei. SegRefiner: Towards model-agnostic segmentation refinement with discrete diffusion process. In *NeurIPS*, 2023. 3

[59] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, pages 11686–11695, 2022. 2

[60] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *CVPR*, pages 16293–16303, 2022. 3

[61] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, et al. Towards open vocabulary learning: A survey. *IEEE TPAMI*, 2024. 2

[62] Yixuan Wu, Zhao Zhang, Chi Xie, Feng Zhu, and Rui Zhao. Advancing referring expression segmentation beyond single image. In *ICCV*, pages 2628–2638, 2023. 2

[63] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*, pages 18381–18391, 2023. 2, 3, 4, 6, 8

[64] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *arXiv:2203.09481*, 2022. 3

[65] Sibei Yang, Meng Xia, Guanbin Li, Hong-Yu Zhou, and Yizhou Yu. Bottom-up shift and reasoning for referring image segmentation. In *CVPR*, 2021. 2

[66] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, pages 18155–18165, 2022. 2

[67] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, pages 10502–10511, 2019. 2

[68] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85. Springer, 2016. 2, 3

[69] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv:2304.06790*, 2023. 4

[70] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *ECCV*, pages 598–615. Springer, 2022. 2