SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

August 2026, Vol. 69, Iss. 8, 182104:1–182104:13
https://doi.org/10.1007/s11432-024-4985-y

Ref-Diff: zero-shot referring image segmentation with generative models

Minheng NI[1], Yabo ZHANG[1], Kailai FENG[1], Xiaoming LI[1], Yiwen GUO[2] & Wangmeng ZUO[1*]

[1]*Harbin Institute of Technology, Faculty of Computing, Harbin 150001, China*
[2]*Independent Researcher*

Received 17 October 2024/Revised 16 February 2025/Accepted 25 May 2025/Published online 31 July 2026

Abstract Zero-shot referring image segmentation (RIS) is a challenging task that involves identifying an instance segmentation mask from referring texts, without being trained on paired image-text data. Current zero-shot RIS methods mainly rely on pre-trained discriminative models (e.g., CLIP). In contrast, this study investigates the potential of generative models (e.g., Stable Diffusion) to understand relationships between various visual elements and text descriptions, an area that remains unexplored in this context. In this work, we introduce the Referring Diffusional segmentor (Ref-Diff), a model that harnesses the fine-grained multimodal information provided by generative models. Our results demonstrate that Ref-Diff, using only a generative model and no external proposal generator, outperforms state-of-the-art weakly supervised models on the RefCOCO+ and RefCOCOg benchmarks. Furthermore, by combining both generative and discriminative models, we present the enhanced version, Ref-Diff+, which significantly surpasses existing methods. This emphasizes the benefits of generative models for discriminative models, thereby improving referring segmentation.

Keywords deep learning, generative model, referring image segmentation, zero-shot learning, zero-shot referring image segmentation

Citation Ni M H, Zhang Y B, Feng K L, et al. Ref-Diff: zero-shot referring image segmentation with generative models. Sci China Inf Sci, 2026, 69(8): 182104, https://doi.org/10.1007/s11432-024-4985-y

1 Introduction

Referring image segmentation (RIS) helps identify the specific region within an image based on a given text description. Unlike standard semantic segmentation, RIS can distinguish similar instances, for example, identifying the tallest boy among a group of children. However, creating annotated datasets pairing images with text descriptions and ground-truth masks is costly and time-consuming. Recently, attempts have been made using the weakly supervised RIS approach [1] to ease the annotation burden; however, it still requires specific image-text pairs for training. A zero-shot solution can be a promising alternative as it requires no prior training or referring annotations, but has additional challenges. Hence, a deeper understanding of the relationship between text and visual elements is required.

Recent multimodal pretraining models have impressive capabilities in understanding both vision and language. CLIP [2] is one such model that stands out for finding similarity between images and texts through contrastive learning. Furthermore, this model is widely used in various tasks, including object detection [3], image retrieval [2], and semantic segmentation [4]. However, direct application of CLIP to zero-shot RIS is limited because CLIP is optimized for global similarity rather than learning the detailed visual elements needed for referring texts. Yu et al. [5] proposed a global and local CLIP approach to bridge the gap between discriminative models and pixel-level dense prediction. Despite these advancements, discriminative models still struggle to localize visual elements accurately.

Generative models like Stable Diffusion [6], DALL-E 2 [7], and Imagen [8] have drawn attention for their ability to synthesize images. The coherent output of these models verifies implicit relationships between visual elements and texts. However, unlike discriminative models, generative models are rarely used in zero-shot RIS, despite their potential to capture intricate text-visual correlations.

In this work, we introduce a novel referring diffusional segmentor (Ref-Diff) to explore the potential of generative models for the zero-shot RIS task. By leveraging the fine-grained multimodal information inherent in generative models, the relationship between referring texts and various visual elements within an image is explored. Unlike

* Corresponding author (email: wmzuo@hit.edu.cn)

previous works that rely on CLIP to rank proposals using an external proposal generator [9], our Ref-Diff can generate these proposals directly using the generative model without depending on any external proposal generator, showcasing its self-sufficiency within the generative model framework. Additionally, we analyze the strengths of both discriminative and generative models and integrate them into a unified RIS framework.

Experiments demonstrate that our Ref-Diff outperforms state-of-the-art weakly supervised methods on RefCOCO+ and RefCOCOg. By incorporating an external proposal generator and a discriminative model, the effectiveness of the model Ref-Diff+ increases significantly compared to all competing methods. Comprehensive quantitative and qualitative assessments validate the advantages of using generative models for this task.

The contributions of this work are summarized as follows. The generative model can be utilized to improve zero-shot RIS performance by exploiting implicit relationships between visual elements and text descriptions. Ref-Diff can perform better than state-of-the-art weakly supervised methods without relying on external proposal generators and discriminative models, and can enhance RIS performance by effectively integrating generative and discriminative paradigms within a unified framework.

2 Related work

Zero-shot referring segmentation. RIS is one of the most fundamental and challenging tasks that require fine-grained understanding of both vision and language. Traditional methods [10–18] are based on fully supervised, labor-intensive training annotations, such as referring texts and pixel-level masks. However, the lack of large-scale training annotations limits the scalability and performance of these methods on out-of-domain samples [1].

With the significant progress in discriminative vision-language pretraining approaches [2], recent studies [1, 5] have explored open-vocabulary recognition for weakly or zero-shot RIS. Recently, some studies [19, 20] applied models like BLIP or ChatGPT to suggest proposals and extract relational information from image captions. Liu et al. [21] used a diffusion model to transform these proposals into a latent space for matching, while Wang et al. [22] designed auxiliary prompts for CLIP to calculate multiple similarity scores for retrieval proposals. Other studies [19,23–26] also leveraged LLMs or synthesized models to generate annotations or descriptions to help models learn from unlabeled data.

Despite the notable performance of discriminative models in finding similarity between text and images, there are certain inherent limitations in deeply understanding object delineation and fine-grained relationships between visual elements and text descriptions. In contrast, our Ref-Diff leverages generative models for fine-grained understanding, resulting in more accurate predictions.

Visual generative models for non-synthesized tasks. Large-scale text-to-image generative models [6–8] have wide applications in imaginative generation and creative applications [27–31], and as these models are capable of fine-grained image understanding, such as semantic segmentation [16, 32–35], object detection [36, 37], dense prediction [38, 39], classification [40, 41], and video segmentation [42].

Recent studies have started exploring their potential in open-vocabulary segmentation tasks: VPD [16], GRES [43], and Peekaboo [44] have leveraged Stable Diffusion's cross-attention maps and multi-scale features for RIS, treating it as a static feature extractor. Conversely, methods like Marigold [45] and LDM-Seg [46] use the iterative denoising process of diffusion models to recover semantic masks from noise. ODISE [47] tried to combine the features of both generative and discriminative pre-trained diffusion models, e.g., by synergizing text-conditioned generation and CLIP-aligned representation learning, to achieve open-vocabulary panoptic segmentation.

Most existing research focuses on transfer learning or constructing synthetic data for specific tasks. Given the powerful ability of generative models to understand text descriptions and generate corresponding image content, some studies have started exploring their application in zero-shot image classification. However, the potential of generative models in zero-shot RIS tasks remains largely unexplored.

3 Ref-Diff and the extension to Ref-Diff+

3.1 Problem formulation and inference pipeline

Given an image $\mathbf{x} \in \mathbb{R}^{W \times H \times C}$ and a referring text T, referring segmentation aims to output a segmentation mask $\mathbf{m} \in \{0,1\}^{W \times H}$ indicating the referring regions of the text T in image $\mathbf{x}$. Here, W, H, and C represent the width, height, and channels of the image, respectively. In the zero-shot referring segmentation settings, the model cannot access any training data for referring segmentation, including images, referring texts, and instance mask annotations.

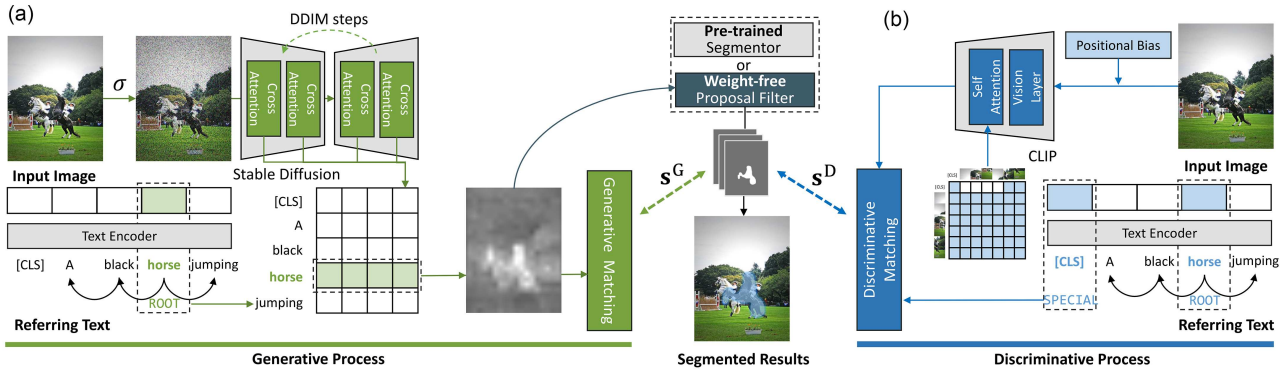

Figure 1 (Color online) Overview of our Ref-Diff and Ref-Diff+. (a) Our proposed generative process establishes a correlation matrix between the referring text and the input image. This matrix serves as an alternative weight-free proposal generator and generates generative segmentation candidates $\mathbf{s}^{\mathrm{G}}$. (b) The discriminative process generates the discriminative candidates $\mathbf{s}^{\mathrm{D}}$. The final RIS result can be obtained from the generative candidates or by combining both generative and discriminative candidates. Our Ref-Diff model not only adopts the generative process to use a weight-free proposal generator, but also Ref-Diff+ integrates both generative and discriminative processes to operate them in conjunction with the pre-trained segmentor.

Our proposed framework is depicted in Figure 1, which shows that Ref-Diff generates a correlation matrix between the given image and referring text using the Generative Process, which can be used as (1) an alternative weight-free proposal generator and (2) a set of referring segmentation candidates. Optionally, Ref-Diff can integrate the discriminative model with the proposed generative model within a unified framework (i.e., Ref-Diff+). The final similarity for each mask proposal is obtained as

$$\mathbf{s}_i = \alpha \mathbf{s}_i^{\mathrm{G}} + (1-\alpha)\mathbf{s}_i^{\mathrm{D}}, \tag{1}$$

where α is a hyperparameter. $\mathbf{s}_i^{\mathrm{G}}$ and $\mathbf{s}_i^{\mathrm{D}}$ are the generative and discriminative scores between the referring text and the i-th proposal, respectively. The final result related to referring segmentation is determined by selecting the proposal with the highest similarity score

$$\hat{\mathbf{m}} = \arg\max_{\mathcal{M}_i} \mathbf{s}_i, \tag{2}$$

where $\mathcal{M}_i$ is the i-th mask proposal. In Ref-Diff, α is set to 1, where only the generative model is adopted for this task, and the proposals are obtained from our weight-free proposal filter. In Ref-Diff+, α is set to a number between 0 and 1, and proposals are obtained from an external pre-trained segmentor to obtain better performance. In the following sections, a detailed introduction to each module is provided.

3.2 Generative process

Stable Diffusion [6] is an effective generative model that gradually transforms random Gaussian noise into an image through a series of inverse diffusion steps. It cannot directly operate on a real image to obtain its latent representations. However, the real image can be considered as an intermediate state generated by a diffusion model, and it can be computed by running it back to any step of the generation. In this work, we add a specific amount of Gaussian noise to obtain $\mathbf{x}_t$ and then continue this process without compromising information:

$$\mathbf{x}_t = \sigma_t(\mathbf{x}), \tag{3}$$

where σ_t is the function to obtain the noised image in step t, and t is a hyperparameter.

Let Ψ_{lan} and Ψ_{vis} be the text and image encoders of the generative model in the inversion diffusion process, respectively. The referring text T is encoded into text features using $\mathbf{K} = \Psi_{\mathrm{lan}}(T) \in \mathbb{R}^{l\times d}$, where l is the token number and d is the dimension size for latent projection. Similarly, for the i-th step, the visual image $\mathbf{x}_i$ is projected to image features using $\mathbf{Q} = \Psi_{\mathrm{vis}}(\mathbf{x}_i) \in \mathbb{R}^{w\times h\times d}$. Here, w and h are the width and height of encoded image features. The cross-attention between the text and image features can be formulated as

$$\mathbf{a} = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d}}\right), \tag{4}$$

where $\mathbf{a} \in \mathbb{R}^{w\times h\times l\times N}$, and N is the number of attention heads. Following [32], the overall cross-attentions are obtained by averaging the value of each attention head to $\bar{\mathbf{a}} \in \mathbb{R}^{w\times h\times l}$. The cross-attention matrix $\bar{\mathbf{a}}$ represents the

correlation detected between each token in the referring text and each region feature in the image. In general, the higher value in $\bar{\mathbf{a}}$ indicates a better correlation between the token and region features. This can be used to locate the related referring regions.

This model captures the overall semantics of the language condition. However, the corresponding attention region for each token is not necessarily the same, as they have different semantic representations (see Subsection 5.4). In general, a referring text T is a sentence that describes the characteristics of a specific instance. To obtain the preferred token from the whole text description, we use syntactic analysis to achieve its root token (i.e., the `ROOT` element in the syntax tree). Generally, the `ROOT` token in the latent space captures the contextual correlations (e.g., the token "horse" in Figure 1 contains contextual representations from "black" and "jumping"). The attention region projected for this root token has a higher probability of being the referring region. Let k be the index of the root token, and let $\bar{\mathbf{a}}_k \in \mathbb{R}^{w\times h}$ denote the cross-attention matrix of the root token.

We normalize and resize this cross-attention matrix by

$$\mathbf{c} = \phi_{w\times h\to W\times H}\left(\frac{\bar{\mathbf{a}}_k - \min(\bar{\mathbf{a}}_k)}{\max(\bar{\mathbf{a}}_k) - \min(\bar{\mathbf{a}}_k) + \epsilon}\right), \tag{5}$$

where ϵ is a small constant value. Here, $\phi_{w\times h\to W\times H}$ is a bilinear interpolation function used to resize the attention map to the same resolution as the given image.

3.3 Discriminative process

During the image encoding process using the discriminative CLIP model, the spatial position is inevitably attenuated. We observed that referring text descriptions usually contain important direction clues (e.g., left, right, top, and bottom), which were ignored in previous works. To emphasize such positional information, a positional bias is proposed to explicitly encode the image with the given direction clues. This is achieved through element-wise multiplication $\mathbf{x}\odot\mathbf{P}$, where $\mathbf{P}\in\mathbb{R}^{W\times H\times C}$ is a positional bias matrix. Specifically, if the referring text, after syntactic analysis, contains any explicit direction clues, $\mathbf{P}$ will be a soft mask with values ranging from 1 to 0 along the given direction. For example, a horizontal direction from left to right is defined as $\mathbf{P}_{i,j} = 1 - \frac{j-1}{H-1}$. Lower values indicate regions that should receive less attention. Conversely, if no direction clue is detected, $\mathbf{P}$ will be a matrix filled with 1.

Finally, the ultimate representation $\mathbf{v}_i\in\mathbb{R}^d$ for each proposal $\mathcal{M}_i$ in discriminative process is

$$\mathbf{v}_i = \beta f_{\mathcal{M}_i}(\mathbf{x}\odot\mathbf{P}) \ + (1-\beta) f(\mathbf{x}\odot\mathcal{M}_i), \tag{6}$$

where f and $f_{\mathcal{M}_i}$ are the vanilla CLIP image encoder and the CLIP image encoder with modified self-attention based on the mask proposal $\mathcal{M}_i$, respectively, and β is the fusion coefficient of them. The discriminative model (i.e., CLIP) is expected to encode the instance within each proposal region $\mathcal{M}_i$, while neglecting other regions to reduce disturbances. This is achieved by assigning a weight of 0 to the attention values between the `[CLS]` token and the patch tokens outside the current proposal $\mathcal{M}_i$. In this work, we use the output of the penultimate layer as the final representation based on the observation that the representation in the last layer encapsulates the information of the entire image rather than the specific proposal region of interest.

3.4 Proposal extraction and matching

Weight-free proposal filter. Since generative models inherently encode instance representations, we can obtain a series of mask proposals from the cross-attention matrix $\mathbf{c}$ by introducing a weight-free proposal filter. This is formulated as

$$\mathcal{M} = \{\psi(\mathbf{c}\geqslant\mu) \mid \mu\in\{5\%, 10\%, \ldots, 95\%\}\}, \tag{7}$$

where ψ is a binarization function with a given predefined threshold value μ. Different from other work [5] that rely on external proposal generators and CLIP filters, the generative models can efficiently and effectively produce the expected proposals. This approach offers a streamlined solution for obtaining high-quality proposals without additional tools.

Pre-trained segmentor. We can also use reliable segmentors in a flexible manner to obtain proposals. By leveraging the capability of generative and discriminative models for semantic understanding, we can refine and prioritize these proposals according to the given referring text. This combined approach ensures that the proposals are coherent and better aligned with the given referring text.

Generative matching. After obtaining proposals either from the weight-free proposal filter or the pre-trained segmentor, the next step is to find the most similar proposal to quantify the similarity score between the given referring text and all the proposals by measuring the distance on the cross-attention matrix $\mathbf{c}$ and $\mathcal{M}_i$:

$$\mathbf{s}_i^{\mathrm{G}} = \frac{|\mathbf{c} \odot \mathcal{M}_i|}{|\mathcal{M}_i|} - \frac{|\mathbf{c} \odot (1 - \mathcal{M}_i)|}{|1 - \mathcal{M}_i|}, \tag{8}$$

where $\odot$ represents element-wise multiplication of vectors or matrices.

Discriminative matching. We obtain the mean representation $\mathbf{r} \in \mathbb{R}^d$ of the global text and the local subject token using the CLIP text encoder, which serves as their features of the referring text. To find out the most probable proposal from the discriminative model, we calculate the similarity between the features of the referring text $\mathbf{r}$ and the visual representation $\mathbf{v}_i$ of each mask proposal $\mathcal{M}_i$, which is defined as

$$\mathbf{s}_i^{\mathrm{D}} = \mathbf{v}_i \mathbf{r}^\top. \tag{9}$$

This similarity score allows us to identify the proposal that best aligns with the referring text. Higher similarity scores indicate a stronger correspondence between the proposal and the text, suggesting a higher likelihood of its being the correct segmentation result.

4 Experiments

4.1 Implementation details

Since Ref-Diff is a zero-shot solution, we need an inference process to achieve better performance without any training images and annotations. All experiments are conducted on an Nvidia A100 GPU. We use the pre-trained Stable Diffusion [6] (V1.5) as our generative model. Since existing works mainly focus on using a pre-trained segmentor and discriminative model [2], for a fair comparison, we select SAM [9] as the external segmentor and adopt CLIP of ViT-B/16 as the discriminative model. We set t, α, β, and ϵ to 2, 0.1, 0.3, and 1×10^{-9}, respectively.

4.2 Experimental setup

Following [1, 5], we adopt evaluation metrics such as mIoU and oIoU and apply them to three widely used benchmarks, including RefCOCO [48], RefCOCO+ [48], and RefCOCOg [49]. For fair comparison, we conduct experiments using zero-shot RIS under two conditions: zero-shot RIS (a) without and (b) with a pre-trained segmentor and CLIP. The first setting allows us to analyze the effectiveness of the generative model, and the second one proves the effectiveness of the collaboration between discriminative and generative models.

For setting (a), we choose the weakly supervised method TSEG [1] as the baseline model. It is not open-source and only provides the validation results of mIoU, so we do not report its results on other settings and test sets. In this setting, our Ref-Diff uses a solely generative model without a proposal generator (e.g., SAM) and CLIP.

For setting (b), we select five competing baselines. Specifically, three zero-shot baselines from [5], including Region Token, Cropping, and Global-local CLIP, and the prior state-of-the-art method adopted with the SAM segmentor and the newly proposed zero-shot baseline SAM-CLIP, considering the remarkable performance of SAM [9] in segmentation. In this setting, our Ref-Diff+ combines both generative and discriminative models using the proposal generator, like other methods.

We also select a series of supervised models [14–16, 23–25, 43] that require additional data as references for comparison. Since most supervised models report mIoU, we omit them in the oIoU table.

4.3 Overall results

Results on RefCOCO, RefCOCO+, and RefCOCOg. Results in terms of mIoU and oIoU, shown in Tables 1 and 2, respectively, show that both Ref-Diff and Ref-Diff+ exhibit superior performance compared to other related methods and baselines. For setting (a), our Ref-Diff outperforms the state-of-the-art weakly supervised model (i.e., TSEG) significantly on RefCOCO+ and RefCOCOg, even without using any pre-trained segmentor (e.g., SAM) or training data. For setting (b), benefiting from the combination of both generative and discriminative models, our Ref-Diff+ achieves an improvement of approximately 10 oIoU and mIoU on RefCOCO, RefCOCO+, and RefCOCOg in comparison with state-of-the-art methods. Comparison between Global-local CLIP (SAM) and SAM-CLIP with other methods also shows the improvement of mIoU. We analyze that segmentation from SAM is highly detailed (e.g., it may generate many proposals overlapping with the same object, especially for small objects), which easily

Table 1 mIoU comparison. The improvement is statistically significant with $p < 0.01$ under t-test. † denotes a weakly supervised method trained on datasets. Bold numbers indicate the best performance.

Method	RefCOCO			RefCOCO+			RefCOCOg		
	val	testA	testB	val	testA	testB	valU	testU	valG
Supervised model with extra training data									
Pseudo-RIS [23]	41.05	48.19	33.48	44.33	51.42	35.08	45.99	46.67	46.80
VLM-VG [24]	49.90	53.10	46.70	42.70	47.30	36.20	48.00	48.50	–
LAVT [15]	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	60.50
VLT [14]	72.96	75.96	69.60	63.53	68.43	56.92	63.49	66.22	62.80
GRES [43]	73.82	76.48	70.18	66.04	71.02	57.65	65.00	65.97	62.70
VPD [16]	73.46	75.31	70.23	63.93	62.73	50.76	63.12	63.55	–
LISA [25]	74.90	79.10	72.30	65.10	70.80	58.10	67.90	70.60	38.70
HyperSeg [26]	84.80	85.70	83.40	79.00	83.50	75.20	79.40	78.90	47.50
Zero-shot model without a pretrained segmentor									
TSEG† [1]	**25.95**	–	–	22.62	–	–	23.41	–	–
Ref-Diff	23.06	**20.05**	**23.17**	**23.91**	**20.53**	**23.64**	**27.03**	**25.95**	**26.45**
Zero-shot model with a pretrained segmentor									
Region token [5]	23.43	22.07	24.62	24.51	22.64	25.37	27.57	27.34	27.69
Cropping [5]	24.83	22.58	25.72	26.33	24.06	26.46	31.88	30.94	31.06
Global-local CLIP [5]	26.20	24.94	26.56	27.80	25.64	27.84	33.52	33.67	33.61
Global-local CLIP (SAM)	31.38	33.31	27.30	34.46	35.00	29.38	38.56	33.49	33.88
SAM-CLIP	26.33	25.82	26.40	25.70	28.02	26.84	38.75	38.91	38.27
Ref-Diff+	**37.21**	**38.40**	**37.19**	**37.29**	**40.51**	**33.01**	**44.02**	**44.51**	**44.26**

Table 2 oIoU comparison. The improvement is statistically significant with $p < 0.01$ under t-test. Bold numbers indicate the best performance.

Method	RefCOCO			RefCOCO+			RefCOCOg		
	val	testA	testB	val	testA	testB	valU	testU	valG
Zero-shot method without pretrained segmentor									
Ref-Diff	**16.12**	**14.10**	**15.73**	**17.50**	**14.84**	**16.48**	**20.39**	**16.82**	**16.92**
Zero-shot method with a pretrained segmentor									
Region token [5]	21.71	20.31	22.63	22.61	20.91	23.46	25.52	25.38	25.29
Cropping [5]	22.73	21.11	23.08	24.09	22.42	23.93	28.69	27.51	27.70
Global-local CLIP [5]	24.88	23.61	24.66	26.16	24.90	25.83	31.11	30.96	30.69
Global-local CLIP (SAM)	23.68	25.79	20.01	26.02	28.62	21.31	27.04	25.35	26.40
SAM-CLIP	25.23	25.86	24.75	25.64	27.76	26.06	33.75	34.80	33.65
Ref-Diff+	**35.16**	**37.44**	**34.50**	**35.56**	**38.66**	**31.40**	**38.62**	**37.50**	**37.82**

increases the possibility of CLIP filtering out erroneous proposals using only the discriminative model CLIP directly. Therefore, the improvement in oIoU is not obvious. By incorporating the generative model, the filtering of erroneous proposals is further mitigated, enhancing their performance. This shows that the improvement over baseline models from a deeper understanding is facilitated by considering the strengths of both our generative and discriminative models.

We observe that the performance of our model can reach approximately half that of supervised models on relatively simpler datasets. On some splits of the more challenging RefCOCOg dataset, it even approaches the performance of supervised models, further demonstrating the potential of our method.

Results on PhraseCut. We conduct additional experiments on PhraseCut dataset in Table 3, which verifies that our Ref-Diff outperforms the competing method Global-local CLIP by a large margin on oIoU. This further validates the strong generalization capability of our Ref-Diff.

4.4 Ablation study

Different components. We conduct an ablation study in Table 4 to validate the effectiveness of different components. Notably, our Ref-Diff outperforms the weakly supervised model TSEG [1] on RefCOCO+ and RefCOCOg when using the generative model alone. When combined with the segmentor, Ref-Diff/gs consistently performs better across all test sets than the weakly supervised model. These observations collectively show that the generative model can not only perform proposal generation but also benefit the RIS task.

Table 3 Quantitative results on PhraseCut. Bold numbers indicate the best performance.

Method	oIoU	mIoU
Without a pre-trained segmentor		
TSEG [1]	–	30.12
Ref-Diff	**12.79**	**31.80**
With pre-trained segmentor		
Global-local CLIP [5]	23.64	–
Ref-Diff+	**29.42**	**41.75**

Table 4 Comparison of our Ref-Diff with different components. Bold numbers indicate the best performance.

Method	Components			RefCOCO		RefCOCO+		RefCOCOg	
	Segmentor	Generative	Discriminative	oIoU	mIoU	oIoU	mIoU	oIoU	mIoU
Weakly supervised Method									
TSEG [1]				–	25.95	–	22.62	–	23.41
Zero-shot method									
Ref-Diff		✓		16.12	23.06	17.50	23.91	20.39	27.03
Ref-Diff/gs	✓	✓		26.04	29.82	26.68	30.06	26.84	30.73
Ref-Diff/ds	✓		✓	33.64	35.27	34.43	35.72	34.83	41.84
Ref-Diff+	✓	✓	✓	**35.16**	**37.21**	**35.56**	**37.29**	**38.62**	**44.02**

Table 5 Comparison of variants with different generative resolutions. Bold numbers indicate the best performance.

Method	Resolution	RefCOCO		RefCOCO+		RefCOCOg	
		oIoU	mIoU	oIoU	mIoU	oIoU	mIoU
Ref-Diff	512×512	14.53	17.08	15.76	20.00	18.30	23.40
Ref-Diff	768×768	14.66	20.48	16.04	23.07	19.88	26.12
Ref-Diff	1024×1024	**16.12**	**23.06**	**17.50**	**23.91**	**20.39**	**27.03**

Table 6 Ablation study on different pretrained segmentors. Bold numbers indicate the best performance.

Method	Segmentor	RefCOCO		RefCOCO+		RefCOCOg	
		oIoU	mIoU	oIoU	mIoU	oIoU	mIoU
Ref-Diff		16.12	23.06	17.50	23.91	20.39	27.03
Ref-Diff/ds	FreeSolo [50]	25.29	28.23	26.32	29.58	30.76	34.01
Ref-Diff/ds	SAM [9]	**33.64**	**35.27**	**34.43**	**35.72**	**34.82**	**41.82**

Resolution for the generative model. The generative model can generate images with different resolutions; the higher the resolution of the generated images, the finer-grained and more detailed the structures are. To investigate the effect of image resolution on this task, we conduct variants on Ref-Diff that exclusively employ the generative model. This allows us to focus on exploring the generative model, neglecting the effects of other factors, e.g., the external segmentor and the discriminative model. Specifically, we consider three resolutions: i.e., 512×512, 768×768, and 1024×1024.

As indicated in Table 5, there exists a correlation between the performance of these three variants and the image resolution. Higher resolutions contribute to more refined attention maps, resulting in more accurate attention projections on different regions.

Different segmentors. Compared to the traditional segmentation network, FreeSolo [50] used by Global-local CLIP, SAM is trained on large-scale data and possesses stronger zero-shot generalization capabilities, generating more accurate candidate regions. To verify the impact of different pre-trained segmentor models in a better way, we further conduct some experiments by replacing the SAM used in Ref-Diff with the FreeSolo model used in Global-local CLIP in Table 6. This indicates that a high-quality segmentor can enhance the quality of candidate regions, thereby improving the performance of the model.

Positional bias for the discriminative model. Different types of positional tokens (i.e., left, right, top, and bottom) usually exist in the referring text. They help in finding explicit location descriptions for RIS but remain unexplored in this task. Analyzing referring texts that contain explicit location descriptions, we found that the proportions of RefCOCO, RefCOCO+, and RefCOCOg are 47.2%, 0.1%, and 16.2%, respectively. By incorporating this positional bias, our Ref-Diff/ds demonstrates a noticeable improvement on RefCOCO (33.64 vs. 32.85 in Table 7). A similar positive effect is observed in RefCOCOg. This indicates that the positional bias can be used to

Table 7 Ablation study on positional bias. Bold numbers indicate the best performance.

Method	Position	RefCOCO		RefCOCO+		RefCOCOg	
		oIoU	mIoU	oIoU	mIoU	oIoU	mIoU
Ref-Diff/ds		32.85	33.45	**34.43**	33.19	34.59	38.32
Ref-Diff/ds	✓	**33.64**	**35.27**	**34.43**	**35.72**	**34.82**	**41.82**

Table 8 Ablation study on different α. Bold numbers indicate the best performance.

Method	α	RefCOCO		RefCOCO+		RefCOCOg	
		oIoU	mIoU	oIoU	mIoU	oIoU	mIoU
Ref-Diff/ds	0.00	33.64	35.27	34.43	35.72	34.82	41.82
Ref-Diff+	0.05	35.12	36.54	34.00	35.04	37.87	40.06
Ref-Diff+	0.10	**35.16**	**37.21**	**35.56**	**37.29**	**38.62**	**44.02**
Ref-Diff+	0.20	34.50	35.14	35.16	37.06	37.63	42.66
Ref-Diff+	0.50	29.79	33.36	31.45	34.58	34.12	36.41
Ref-Diff/gs	1.00	26.04	29.82	26.68	30.06	26.84	30.73

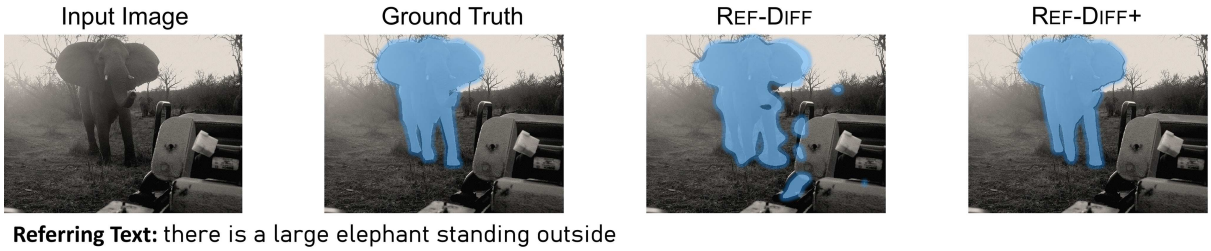

Figure 2 (Color online) Effectiveness of the generative model in segmentation capability. Ref-Diff is capable of segmenting the correct content even without the assistance of the pre-trained segmentor and CLIP. Combined with the pre-trained segmentor, Ref-Diff+ achieves precise segmentation of the expected regions.

enhance the RIS performance on datasets including positional descriptions.

***α* for combining generative and discriminative models.** We observe that the hyperparameter α that determines the balance between discriminative and generative models is crucial and directly affects the model's overall performance. As shown in Table 8, the model achieves optimal performance when α is set to 0.1, while performance tends to deteriorate when α is either lower or higher. These results highlight the critical role of α in achieving the desired performance of the model.

5 Discussion

5.1 Segmentation without a pre-trained segmentor

Segmentation results. From Figure 2, it can be observed that even without the assistance of the segmentor and CLIP, our Ref-Diff is able to accurately segment the corresponding content. This is primarily attributed to the attention projection of the generative model onto the relevant visual content (please refer to Subsection 5.4). Despite the presence of two huge instances in the image, our Ref-Diff is still capable of performing accurate segmentation. This highlights the significant potential of the generative model, as it exhibits segmentation capabilities comparable to segmentation models and categorization capabilities comparable to discriminative models.

Proposals from generative model. Based on the proposals from the attention information of the generative model, our proposed Ref-Diff is capable of performing the RIS task without relying on an external segmentor. Figure 3 demonstrates the proposals extracted from the generative model. In comparison to the approach involving an external segmentor, the generative model, with its language conditioning information, selectively extracts proposals relevant to the referring text, thereby ignoring the numerous additional irrelevant objects present in the image. This proposal-extraction approach enhances the efficiency and relevance of the RIS task within our Ref-Diff.

5.2 Effect of generative model

Qualitative analysis. The first example shown in Figure 4 indicates that the discriminative model fails to accurately locate the leftmost broccoli. Similarly, in the second example, CLIP fails to eliminate the redundant

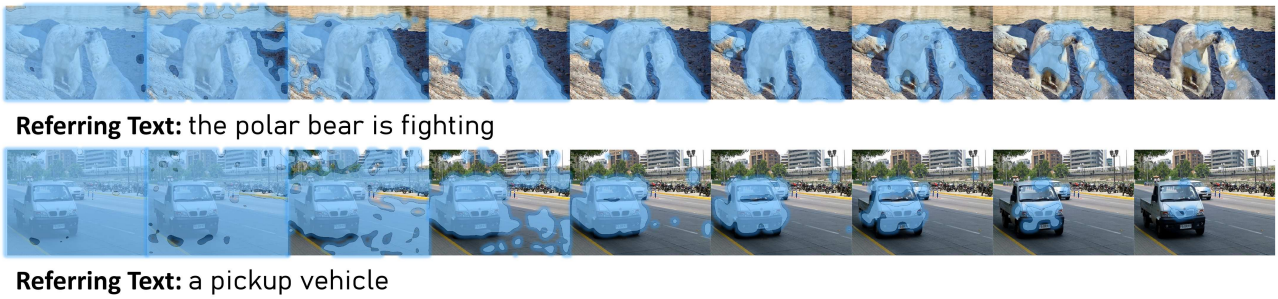

Figure 3 (Color online) Visualization of proposals with different threshold in (7).

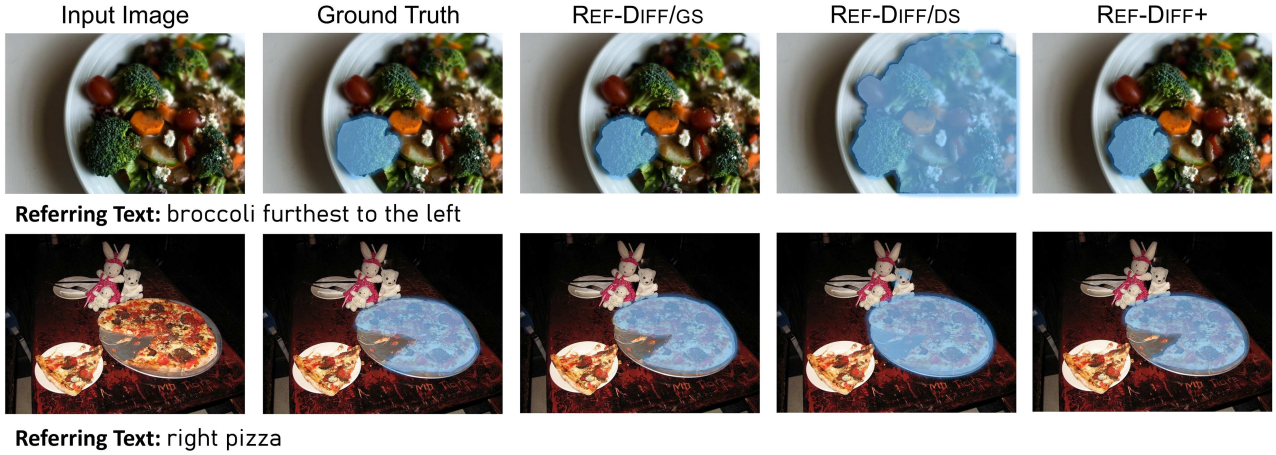

Figure 4 (Color online) Effectiveness of a generative model in localization capability. The discriminative model evaluates whether the image contains text-related content, which mistakenly results in selecting larger regions.

Table 9 Results with different lengths of referring text.

Method	1–2		3–4		5–6		7–8		9+	
	oIoU	mIoU	oIoU	mIoU	oIoU	mIoU	oIoU	mIoU	oIoU	mIoU
Ref-Diff/ds	42.14	40.12	38.64	39.57	32.93	35.39	28.89	32.74	24.36	30.87
Ref-Diff+	43.93	41.94	38.27	39.86	34.29	38.99	33.28	36.70	28.47	32.52

region of the plate. We conclude that this issue may arise due to the inherent limitation of the discriminative model, which is trained to identify whether the given text and the image are well-aligned. Therefore, it lacks the capability of localizing objects within the image. Consequently, relying solely on the discriminative model to discern whether a region contains redundant content can be challenging. However, combining both the generative and discriminative models can provide the best results.

Quantitative analysis. We conduct quantitative experiments on oIoU to explore the advantages of the generative model. We divided RefCOCO into several subsets based on the length of the referring text and calculated their performance separately. Here, Ref-Diff+ refers to the complete network that includes the segmentor, generative model, and discriminative model, while Ref-Diff/ds removes the generative model.

Table 9 shows that Ref-Diff integrating the generative model is more pronounced for longer referring text. This is because the generative model can provide information that the discriminative model struggles to capture in complex referring relationships. Therefore, the advantage of the generative model can be attributed to its stronger localization capability.

5.3 Effect of discriminative model

Qualitative analysis. In the example shown in Figure 5, we observe that the generative model incorrectly segments the screen as a separate object due to its prominence as a significant visual feature of a laptop, attracting more attention from the generative model. The generative model places greater emphasis on the key visual content, resulting in incomplete segmentation. However, this can be effectively mitigated in our full model Ref-Diff+, due

Figure 5 (Color online) Effectiveness of discriminative model. The generative model exhibits higher sensitivity to salient visual features, which can result in partial segmentation when solely relying on the generative model. By integrating the discriminative model with it, we can effectively mitigate such errors and achieve more accurate results.

Table 10 Ablation study on different attention tokens. Bold numbers indicate the best performance.

Method	Token	RefCOCO		RefCOCO+		RefCOCOg	
		oIoU	mIoU	oIoU	mIoU	oIoU	mIoU
Ref-Diff/gs	`CLS`	17.55	22.53	18.40	24.91	16.39	22.25
Ref-Diff/gs	`End`	15.03	26.94	16.82	25.65	16.76	21.48
Ref-Diff/gs	`ROOT`	**26.04**	**29.82**	**26.68**	**30.06**	**26.84**	**30.73**

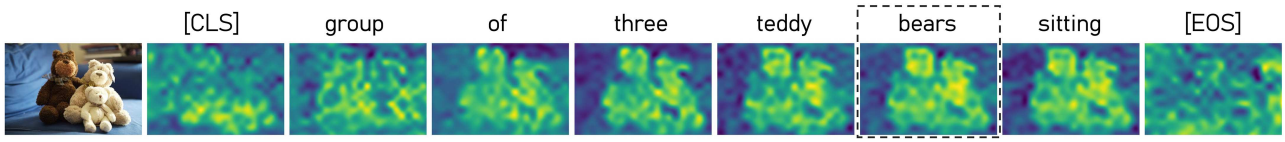

Figure 6 (Color online) Attention from the generative model. The generative model projects attention to different regions of the image based on different tokens, which is the key reason for the effectiveness of Ref-Diff. The dashed box highlights the root token and its corresponding attention map.

to the robust categorization capability of the discriminative model.

Quantitative analysis. We quantify the correct inner area proportion, specifically, the ratio of the predicted segmentation that lies within the ground-truth result. Without using the discriminative model, Ref-Diff/gs obtained 27.7%, and the full model Ref-Diff+ obtained 41.2% of the correct inner area proportion. We found that the generative model has certain drawbacks: its higher sensitivity to salient visual features, which makes it prone to producing partial segmentation results. This is the strength of the discriminative model.

5.4 Attention from the generative model

Attention maps. The region highlighted by the generative model is investigated by visualizing samples along with the attention weights assigned to each token, as depicted in Figure 6. We find that the generative models exhibit contextual understanding capabilities, as they can effectively allocate attention to relevant regions in the image, thereby accomplishing both localization and segmentation tasks. Notably, the attention assigned to the subject token closely aligns with final segmentation results. This finding elucidates why generative models can perform well in zero-shot referring segmentation tasks without relying on a separate segmentor or a discriminative model. In contrast to classification models, where the first token often represents the entirety of the text, we observe that the first token in our generative model does not capture the complete context. However, as discussed earlier, the subject token successfully captures comprehensive information from the text, enabling accurate results.

Attention tokens. We validate that commonly used sentence representations corresponding the `[CLS]` and end-of-sentence tokens are effectively used by referring to the information collected by the generative model. As shown in Table 10, we find that, compared to the root token mentioned in this paper, the performance of commonly used `[CLS]` and end-of-sentence tokens is lower. This also supports our finding that for the generative model, the root token of the referring text contains more comprehensive referential information than other tokens, making it more suitable for referring segmentation tasks.

5.5 Case study

We have shown three examples of varying difficulty in Figure 7 to showcase the effectiveness of our Ref-Diff. In the first example, we observe that Ref-Diff accurately identifies and segments the correct object within similar objects. In the second example, despite the presence of numerous objects in the image and complex spatial relationships,

Figure 7 (Color online) Case studies. Blue indicates the predicted regions. This case is a failed example, where green denotes ground-truth regions. Ref-Diff demonstrates its ability to accurately segment objects based on the provided referring texts, even in the presence of complex spatial relationships in the image.

Table 11 Inference time cost on different models. † denotes using Hyper-SD [51] and EfficientViT-SAM [52] to speed up inference.

Method	Seg.	Gen.	Dis.	Inference time (s)	Multiple
Global-local CLIP [5]	✓		✓	0.68	1.00×
SAM-CLIP	✓		✓	2.70	3.97×
Ref-Diff		✓		1.17	1.72×
Ref-Diff/gs	✓	✓		2.74	4.03×
Ref-Diff/ds	✓		✓	2.64	3.88×
Ref-Diff+	✓	✓	✓	3.18	4.68×
Ref-Diff+†	✓	✓	✓	1.14	1.68×

Ref-Diff identifies the correct objects accurately through the understanding of the generative and discriminative models. In the third example, we encountered a segmentation failure of Ref-Diff due to the presence of some degree of ambiguity in the referring text. Ref-Diff incorrectly identifies the leftmost hand as the first arm, resulting in a segmentation error. Enhancing the robustness of Ref-Diff is an area of future work that deserves further investigation.

5.6 Time complexity

We report the inference time of different variants of Ref-Diff and models in Table 11. When more modules are used, Ref-Diff requires more time for inference. Inspired by existing diffusion and SAM acceleration methods, we have also used an acceleration method for Stable Diffusion, Hyper-SD [51], and a speeding-up method for SAM, EfficientViT-SAM [52], which demonstrate faster inference speed.

6 Broader impact and limitations

Zero-shot RIS has broad applications in industrial and real-world domains, such as image editing, robot control, and human-machine interaction. Our Ref-Diff, through the combination of generative and discriminative models, has successfully demonstrated the feasibility of training-free yet high-quality referring segmentation in various data-scarce scenarios. This significantly reduces the deployment cost of artificial intelligence in related fields.

However, referring texts are sensitive to ambiguity (see Subsection 5.5), which currently results in noticeable segmentation errors. At the same time, because of using a diffusion-based generative model and a large-scale segmentor, SAM enhances localization capability but incurs longer inference times, e.g., multiple DDIM steps for the diffusion model to obtain effective attention information. Although the proposed diffusion-based approach improves localization through iterative refinement, this process introduces substantial time cost. Even with Hyper-SD and EfficientViT-SAM acceleration, our full pipeline (1.14 s) remains slower than non-diffusion and light segmentor baselines like Global-local CLIP (0.68 s), as shown in Table 11. In the future, we will further investigate zero-shot referring segmentation with lower computational costs and higher robustness.

7 Conclusion

In this work, we proposed a novel referring diffusional segmentor (Ref-Diff) for zero-shot RIS, which effectively leverages the fine-grained multimodal information from generative models. We demonstrated that Ref-Diff, using a generative model alone, can outperform state-of-the-art weakly supervised models without any proposal generator in the most popular datasets. Moreover, by combining both generative and discriminative models, Ref-Diff+ outperformed these competing methods by a significant margin. Overall, our work provides a simple yet promising direction for zero-shot RIS by exploiting the potential of generative models, which brought new perspectives for addressing the challenges of this task.

Acknowledgements This work was supported by National Key R&D Program of China (Grant No. 2022YFA1004100).

References

1 Strudel R, Laptev I, Schmid C. Weakly-supervised segmentation of referring expressions. ArXiv:2205.04725
2 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning, 2021. 8748–8763
3 Zhong Y, Yang J, Zhang P, et al. Regionclip: region-based language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 16793–16803
4 Dong X, Bao J, Zheng Y, et al. Maskclip: masked self-distillation advances contrastive language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 10995–11005
5 Yu S, Seo P H, Son J. Zero-shot referring image segmentation with global-local context features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 19456–19465
6 Rombach R, Blattmann A, Lorenz D, et al. High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 10684–10695
7 Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with clip latents. ArXiv:2204.06125
8 Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding. In: Proceedings of the Advances in Neural Information Processing Systems, 2022. 36479–36494
9 Kirillov A, Mintun E, Ravi N, et al. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 4015–4026
10 Xu L, Huang M H, Shang X, et al. Meta compositional referring expression segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 19478–19487
11 Yan B, Jiang Y, Wu J, et al. Universal instance perception as object discovery and retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 15325–15336
12 Yang Z, Wang J, Tang Y, et al. Semantics-aware dynamic localization and refinement for referring image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2023. 3222–3230
13 Liu J, Ding H, Cai Z, et al. Polyformer: Referring image segmentation as sequential polygon generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 18653–18663
14 Ding H, Liu C, Wang S, et al. VLT: vision-language transformer and query generation for referring segmentation. IEEE Trans Pattern Anal Mach Intell, 2022, 45: 7900–7916
15 Yang Z, Wang J, Tang Y, et al. Lavt: language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 18155–18165
16 Zhao W, Rao Y, Liu Z, et al. Unleashing text-to-image diffusion models for visual perception. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 5729–5739
17 Liu C, Ding H, Zhang Y, et al. Multi-modal mutual attention and iterative interaction for referring image segmentation. IEEE Trans Image Process, 2023, 32: 3054–3065
18 Pei G, Yao Y, Shen F, et al. Hierarchical Co-attention propagation network for zero-shot video object segmentation. IEEE Trans Image Process, 2023, 32: 2348–2359
19 Suo Y, Zhu L, Yang Y. Text augmented spatial aware zero-shot referring image segmentation. In: Proceedings of Findings of the Association for Computational Linguistics: EMNLP, 2023. 1032–1043
20 Han Z, Zhu F, Lao Q, et al. Zero-shot referring expression comprehension via structural similarity between images and captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 14364–14374
21 Liu X, Huang S, Kang Y, et al. Vgdiffzero: text-to-image diffusion models can be zero-shot visual grounders. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2024. 2765–2769
22 Wang H, Zhan Y, Liu L, et al. Balanced similarity with auxiliary prompts: towards alleviating text-to-image retrieval bias for CLIP in zero-shot learning. ArXiv:2402.18400
23 Yu S, Seo P H, Son J. Pseudo-ris: distinctive pseudo-supervision generation for referring image segmentation. In: Proceedings of European Conference on Computer Vision. Milano: Springer, 2024. 18–36
24 Wang S, Kim D, Taalimi A, et al. Learning visual grounding from generative vision and language model. ArXiv:2407.14563
25 Lai X, Tian Z, Chen Y, et al. Lisa: Reasoning segmentation via large language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 9579–9589
26 Wei C, Zhong Y, Tan H, et al. Hyperseg: towards universal visual segmentation with large language model. ArXiv:2411.17606
27 Wei Y, Zhang Y, Ji Z, et al. Elite: encoding visual concepts into textual embeddings for customized text-to-image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 15943–15953
28 Ruiz N, Li Y, Jampani V, et al. Dreambooth: fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 22500–22510
29 Gal R, Alaluf Y, Atzmon Y, et al. An image is worth one word: personalizing text-to-image generation using textual inversion. In: Proceedings of the International Conference on Learning Representations, 2023
30 Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 3836–3847
31 Mou C, Wang X, Xie L, et al. T2I-adapter: learning adapters to dig out more controllable ability for text-to-image diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 4296–4304
32 Wu W, Zhao Y, Shou M Z, et al. Diffumask: synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 1206–1217
33 Brempong E A, Kornblith S, Chen T, et al. Denoising pretraining for semantic segmentation. In: Proceedings of Workshops of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 4174–4185
34 Baranchuk D, Rubachev I, Voynov A, et al. Label-efficient semantic segmentation with diffusion models. In: Proceedings of the International Conference on Learning Representations, 2022

35 Amit T, Shaharbany T, Nachmani E, et al. SegDiff: image segmentation with diffusion probabilistic models. ArXiv:2112.00390
36 Cheng Z, Li K, Jin P, et al. Parallel vertex diffusion for unified visual grounding. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2024. 1326–1334
37 Chen S, Sun P, Song Y, et al. Diffusiondet: diffusion model for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 19830–19843
38 Ji Y, Chen Z, Xie E, et al. DDP: diffusion model for dense visual prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 21741–21752
39 Saxena S, Kar A, Norouzi M, et al. Monocular depth estimation using diffusion models. ArXiv:2302.14816
40 Li A C, Prabhudesai M, Duggal S, et al. Your diffusion model is secretly a zero-shot classifier. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 2206–2217
41 Clark K, Jaini P. Text-to-image diffusion models are zero shot classifiers. In: Proceedings of the Advances in Neural Information Processing Systems, 2024
42 Ding H, Liu C, He S, et al. Mevis: a large-scale benchmark for video segmentation with motion expressions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 2694–2703
43 Liu C, Ding H, Jiang X. Gres: generalized referring expression segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 23592–23601
44 Burgert R, Ranasinghe K, Li X, et al. Peekaboo: text to image diffusion models are zero-shot segmentors. ArXiv:2211.13224
45 Ke B, Obukhov A, Huang S, et al. Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 9492–9502
46 Wang C, Li X, Ding H, et al. Explore in-context segmentation via latent diffusion models. ArXiv:2403.09616
47 Xu J, Liu S, Vahdat A, et al. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023. 2955–2966
48 Nagaraja V K, Morariu V I, Davis L S. Modeling context between objects for referring expression understanding. In: Proceedings of the European Conference of Computer Vision, 2016. 792–807
49 Mao J, Huang J, Toshev A, et al. Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. 11–20
50 Wang X, Yu Z, De Mello S, et al. Freesolo: learning to segment objects without annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 14176–14186
51 Ren Y, Xia X, Lu Y, et al. Hyper-sd: trajectory segmented consistency model for efficient image synthesis. ArXiv:2404.13686
52 Zhang Z, Cai H, Han S. Efficientvit-sam: accelerated segment anything model without performance loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 7859–7863