IteRPrimE: Zero-shot Referring Image Segmentation with Iterative Grad-CAM Refinement and Primary Word Emphasis

Yuji Wang*, Jingchen Ni*, Yong Liu, Chun Yuan, Yansong Tang†

Shenzhen International Graduate School, Tsinghua University
*{yuji-wan24, njc24}@mails.tsinghua.edu.cn, †tang.yansong@sz.tsinghua.edu.cn

Abstract

Zero-shot Referring Image Segmentation (RIS) identifies the instance mask that best aligns with a specified referring expression without training and fine-tuning, significantly reducing the labor-intensive annotation process. Despite achieving commendable results, previous CLIP-based models have a critical drawback: the models exhibit a notable reduction in their capacity to discern relative spatial relationships of objects. This is because they generate all possible masks on an image and evaluate each masked region for similarity to the given expression, often resulting in decreased sensitivity to direct positional clues in text inputs. Moreover, most methods have weak abilities to manage relationships between primary words and their contexts, causing confusion and reduced accuracy in identifying the correct target region. To address these challenges, we propose **IteRPrimE** (**Ite**rative Grad-CAM **R**efinement and **Prim**ary word **E**mphasis), which leverages a saliency heatmap through Grad-CAM from a Vision-Language Pre-trained (VLP) model for image-text matching. An iterative Grad-CAM refinement strategy is introduced to progressively enhance the model's focus on the target region and overcome positional insensitivity, creating a self-correcting effect. Additionally, we design the Primary Word Emphasis module to help the model handle complex semantic relations, enhancing its ability to attend to the intended object. Extensive experiments conducted on the RefCOCO/+/g, and PhraseCut benchmarks demonstrate that IteRPrimE outperforms previous SOTA zero-shot methods, particularly excelling in out-of-domain scenarios.

Code — https://github.com/VoyageWang/IteRPrimE

Introduction

Referring Image Segmentation (RIS) requires the model to generate a pixel-level referred object mask based on a textual description, extending the applicability to various tasks such as robot interaction and image editing (Yang et al. 2024, 2022; Liu et al. 2024b; Lai et al. 2024; Luo et al. 2024b). Different from standard semantic segmentation (Wang, Zhao, and Sun 2023; Wang et al. 2024b; Han et al. 2023; Luo et al. 2024a; Bai et al. 2024), RIS necessitates the differentiation of instances within the same category and their relationships with other objects or the scene, which requires high demands on the semantic understanding and spatial perception of the model. However, annotating exact pairs of images, descriptions, and ground-truth masks is both expensive and time-intensive, as the annotation of a query needs a grasp of diverse positional and attributive details within the image (Liu, Ding, and Jiang 2023; Ding et al. 2023; Liu et al. 2019). Recent weakly supervised RIS techniques (Strudel, Laptev, and Schmid 2022; Lee et al. 2023; Xu et al. 2022) have been introduced to mitigate these annotation challenges, yet they still depend on paired data for training purposes and have relatively poor performance. In contrast, a zero-shot approach holds greater value. Leveraging vision-language pre-trained (VLP) models such as CLIP (Radford et al. 2021), this method efficiently generalizes across diverse concepts and unseen categories without further training and fine-tuning.

Existing methodologies to harness the characteristics of being unnecessary to fit training data presented by zero-shot learning often employ a two-stage pipeline, shown in Figure 1 (a). As a discriminator between the images masked by the candidate masks and the expression, CLIP is used to select the instance mask whose similarity score is the highest (Sun et al. 2024; Yu, Seo, and Son 2023; Suo, Zhu, and Yang 2023; Ni et al. 2023). However, we observed that these methods always malfunctioned when encountering text inputs with positional information such as "left" and "right". Due to only a single instance contained in a masked image, the absence of relative spatial perception can be the inherent limitation of these CLIP-based paradigms. Previous pieces of literature alleviate this issue by injecting the human priors or bias that explicitly prompts the CLIP with the given direction clues (Ni et al. 2023; Suo, Zhu, and Yang 2023). To be more specific, they manually design spatial decaying weights from 1 to 0 in the directions consistent with text phrases to make the model aware of positional information, but it can not generalize the scenarios out of predefined directions such as "next to". Additionally, the domain shift for CLIP from the natural image to the masked image can also impact the segmentation performance (Liu et al. 2024a; Ding et al. 2022; Zhu and Chen 2024).

Some researchers (Lee et al. 2023) have leveraged Grad-

*These authors contributed equally.
†Corresponding Author.

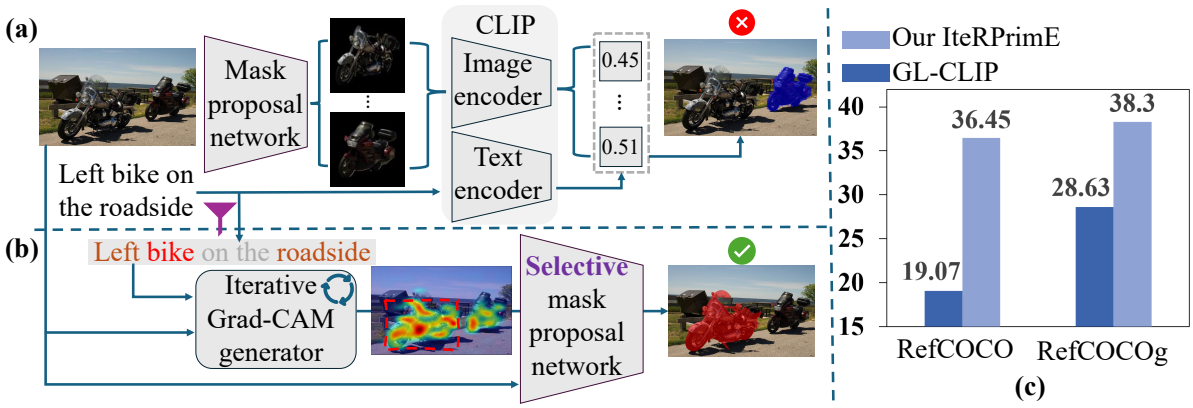

Figure 1: (a) The general pipeline of CLIP-based methods. They lack the perception of spatial relative position due to the masked images. (b) The pipeline of our IteRPrimE with Iterative Grad-CAM Refinement Strategy and Primary Word Emphasis of "bike". (c) This is a comparative experiment of positional phrase accuracy between IteRPrimE and GL-CLIP on RefCOCO and RefCOCOg.

CAM (Selvaraju et al. 2017) and created two specialized loss functions to attenuate the detrimental effects of positional phrases in weakly supervised settings. Although the losses are unsuitable for zero-shot scenarios, Grad-CAM can partially mitigate the deleterious effects associated with masked images. This is because the method maintains the integrity of the model's spatial perception capabilities by delineating the regions with the greatest attention in the original image for localization, shown in Figure 1 (b). Nevertheless, we still find two major problems by analyzing the occurrences and characteristics of Grad-CAM. First, Grad-CAM struggles to discriminate the semantic relations between different noun phrases, due to the lack of a stronger consideration of the primary word than other context words, shown in baseline predictions of Figure 2 (a). Specifically, the model's weak ability to effectively prioritize the main word in complex expressions undermines its overall performance. Second, Grad-CAM is limited to identifying only small areas of the referred object, which consequently results in selecting undesired instance masks.

To overcome these challenges, we propose a novel framework namely, **IteRPrimE** (**Ite**rative Grad-CAM **R**efinement and **Prim**ary word **E**mphasis) utilizing Grad-CAM for zero-shot RIS. First, we implement an iterative refinement strategy to enhance the representational accuracy and enlarge the indicated area of Grad-CAM, progressively improving the model's concentration on the target object with each cycle, shown in Figure 2 (b). Simultaneously, this strategy is particularly beneficial when the referring expression includes positional words, as it offers the model chances of self-correction at each iteration, shown in Figure 2 (c). Second, the Primary Word Emphasis Module (PWEM) plays a crucial role in enhancing the weak abilities to handle the complex semantic relationships between primary words and other contexts. This module is achieved by emphasizing the Grad-CAMs of the main word within the referring expression, from local and global aspects. Finally, a post-processing module is designed to select a high-quality, contiguous instance mask from a mask proposal network, which encapsulates the target object as indicated by Grad-CAM. By addressing the limitations, the IteRPrimE approach achieves superior performance over prior zero-shot state-of-the-art techniques, notably excelling in out-of-domain scenarios and exhibiting robust cross-domain transfer proficiency. Our main contributions include

1. To our best knowledge, we are the first to use Grad-CAM to instruct Segmentors for zero-shot RIS tasks.
2. We propose the Iterative Grad-CAM Refinement Strategy (IGRS) and Primary Word Emphasis Module (PWEM) to enhance the accuracy and representation of Grad-CAM for better localization, shown in Figure 2.
3. Compared to the previous CLIP-based method, our method significantly outperforms it with inputs containing positional information, shown in Figure 1 (c). Additionally, the approach achieves a better performance on the four popular benchmarks, especially for the out-domain datasets.

Related Works

Zero-shot referring image segmentation. For the fully-supervised setting, training a well-specialized model for RIS needs massive paired text-visual annotations, which are sometimes not affordable and accessible (Shah, VS, and Patel 2024; Liu, Ding, and Jiang 2023; Yang et al. 2022; Wang et al. 2022b; Kim et al. 2022; Ding et al. 2021; Jing et al. 2021). Besides, these models have relatively weak ability in out-of-domain scenarios due to the limited data and a domain gap. Therefore, the zero-shot RIS methods are proposed as the alternative. Global- and local-CLIP (GL-CLIP) (Yu, Seo, and Son 2023) is the first proposed to segment the instance given the text input with zero-shot transfer. By interfacing with the mask proposal network FreeSOLO (Wang et al. 2022a), the approach leverages both global

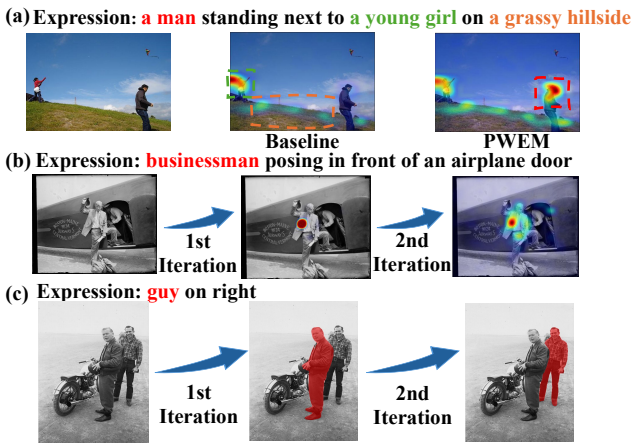

Figure 2: (a) The weak ability of the baseline model to differentiate the semantic relationships between the primary word "man" and the other noun phrases colored green and orange. PWEM can make the model aware of the targeted instance referred to by the main word. (b) The IGRS facilitates the expansion of highlighted areas, surpassing the confined small regions. (c) IGRS offers the model chances of self-correction.

and local textual-image similarity to enhance the discriminative capabilities of the CLIP model. Based on GL-CLIP, some researchers (Wang et al. 2024a) combine the original CLIP similarity score with their proposed Balanced Score with Auxiliary Prompts (BSAP), namely BSAP-H, to reduce the CLIP's text-to-image retrieval hallucination. Ref-Diff (Ni et al. 2023) demonstrates that the text-to-image generative model like Stable Diffusion (Rombach et al. 2022) can generate the intended mask from the cross-attention map, which has considerable performance. TAS (Suo, Zhu, and Yang 2023) mainly depends on another large captioner network BLIP2 (Li et al. 2023) to mine the negative text based on the previous mask proposal network plus discriminator paradigm, which achieves favorable performances. Additionally, SAM (Kirillov et al. 2023) is utilized for better segmentation accuracy. However, these CLIP-based methods struggle to segment the referred subject with positional-described text queries, due to the absence of spatial relationships in the masked image.

Grad-CAM for localization. Grad-CAM (Selvaraju et al. 2017) is proposed to provide explainable clues indicating the regions the model pays attention to for the prediction head. In the context of the Image Text Matching (ITM) objective from any VLP (Li et al. 2022; Xu et al. 2023a,b), Grad-CAM enables the establishment of a modality mapping from the textual to the visual domain, specifically calibrated for the task of visual localization. Many works utilize it to localize the objects with the given text (Shen et al. 2024; Lee et al. 2023; Xu et al. 2022; He et al. 2022; Li et al. 2021). However, these approaches either generate a bounding box annotation or are employed within weakly supervised scenarios. Compared to approaches (Shin, Xie, and Albanie 2022; Zhou, Loy, and Dai 2022; Luo et al. 2024a) that perform zero-shot open vocabulary semantic segmentation with Grad-CAM, we are the first to propose the Grad-CAM for zero-shot RIS to study its behaviors under longer and complex textual inputs instead of a single category noun. To address problems of lacking consideration between main words and the other, the PWEM is proposed to aggregate the Grad-CAM from local-spatial and global-token levels. Secondly, a novel iterative refinement strategy is employed to obtain a better representation of Grad-CAM step by step.

Preliminaries

The generation of Grad-CAM is essential for harnessing it for RIS. Given an image-expression pair (I, E), we can obtain their corresponding embeddings, v and e, by the visual encoder $v = f_I(I)$ and text encoder $e = f_T(E)$, respectively. Then, for multimodal fusion, these two embeddings are fed to the cross-attention layers used to align the visual and textual information (Yu et al. 2022; Vaswani et al. 2017). The resultant attention activation maps, $\mathbf{A}$, can indicate the activated and recognized regions of v concerning each query textual token in e. However, these indication clues are usually scattered and not densely distributed in the relevant regions. Thus, the gradients, $\mathbf{G}$ can be used to sharpen and dilute the effect of non-relevant regions in $\mathbf{A}$, where contribute less to the output objective, y, like Image Text Matching (ITM). The result of this gradient-weighted dilution process is known as Grad-CAM, $\mathbf{H}$.

In the cross-attention layer, the Grad-CAM can be formulated by Equation (1)

$$\mathbf{H} = \mathbf{A} \odot \mathbf{G}, \tag{1a}$$

$$\mathbf{G} = clamp\left(\frac{\partial y}{\partial \mathbf{A}}, 0, \infty\right), \tag{1b}$$

where $clamp$ removes negative gradients, which often represent noise or irrelevant features. Finally, the Grad-CAM used to indicate the image regions, $\mathbf{H}_f$, is the mean over all the number of text tokens $|e|$, as shown in Equation (2)

$$\mathbf{H}_f = E_k\left(\mathbf{H}^k\right), k \in |e|, \mathbf{H}_f \in R^{B\times h\times w} \tag{2}$$

where $\mathbf{H}^k$ denotes the Grad-CAM for the k-th text token, B is the batch size, and $h \times w$ is the size of visual latent space. This averaging process treats every word equally and ignores the importance of the primary word, thereby undermining the performance of RIS.

Method

Overview

Figure 1 (b) demonstrates the entire workflow of our method for zero-shot RIS, IteRPrimE, which can be divided into two parts: an iterative Grad-CAM generator and a selective mask proposal network. First, the Grad-CAM generator is a VLP model with cross-attention layers. The proposed IGRS and PWEM are integrated into the generator. Finally, within the mask proposal network, a post-processing module is designed to select the candidate instance masks, ensuring the accurate and detailed localization of the target object.

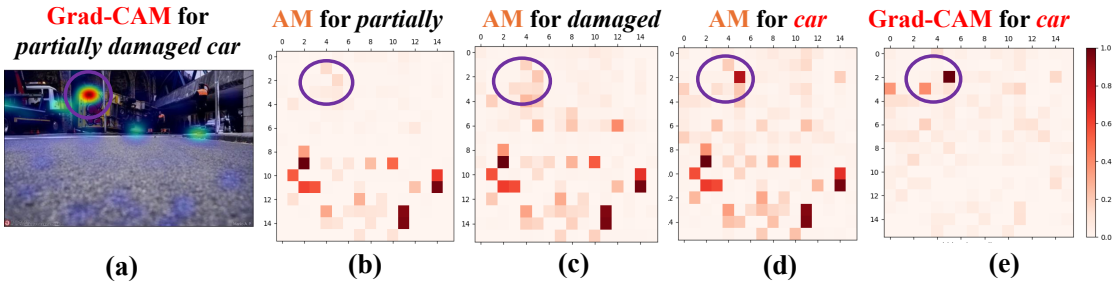

Figure 3: The Grad-CAMs and attention maps (AM) of "partially damaged car". Since the attention map (d) and Grad-CAM (e) of the primary word "car" both contain unique activation areas compared to the others, they can be harnessed from local-spatial and global-token perspectives to enhance the focus on the targeted regions, respectively.

Primary Word Emphasis Module

The PWEM is an essential component of the IteRPrimE, designed to confront the challenge posed by the weak capability of Grad-CAM to manage the semantic relationships in input texts featuring multiple potential referred nouns. This module emphasizes the Grad-CAM of the primary word in the expression, thereby increasing the focus on the main word during the averaging operation. Specifically, we first use an NLP processing toolbox to parse the part-of-speech (POS) tags of each word, filtering out a set of text tokens that includes special $< CLS >$ token of BERT (Devlin et al. 2018), nouns, adjectives, verbs, proper nouns, and numerals. These words are recognized as effective tokens W that can provide distinct semantics and their contextual information. They are composed of primary words W_m and their contexts W_c, where $W = W_m \cup W_c$ and $W_m \cap W_c = \emptyset$. Then, we extract the primary noun from these effective words (e.g. "car" in "partially damaged car" shown in Figure 3) by employing a designed algorithm. It first generates a syntax tree, identifies the leftmost noun phrase (NP), and then finds the rightmost noun (NN) within that NP, which can be detailed in Algorithm 1 in the appendix.

As shown in the right part of Figure 4, we emphasize the effect of the primary word Grad-CAM from two perspectives: local spatial-level and global token-level augmentation. Different from the other contextual effective words W_c, the attention map $\mathbf{A}^{W_m}$ and Grad-CAM $\mathbf{H}_m$ of the primary token holds the unique activated areas that probably indicate the correct localization of Grad-CAM shown in Figure 3. Therefore, to highlight and isolate the specific contribution of the primary word from the local spatial level, we compute the L_2 normalized differences, $\mathbf{A}_{dif}$ between the main word activation map and the other context word activation maps, $\mathbf{A}^{W_c}$. The activation difference is further integrated with gradients from the main word $\mathbf{G}^{W_m}$, forming a spatial modulator to indicate the local spatial importance in the main word Grad-CAM, $\mathbf{H}_m$. Thus, we can obtain the local spatial-level enhanced Grad-CAM of the primary word, $\mathbf{H}_l$, as shown in Equation (3)

$$\mathbf{A}_{dif} = \frac{\mathbf{A}^{W_m} - \mathbf{A}^{W_c}}{||\mathbf{A}^{W_m} - \mathbf{A}^{W_c}||_2}, \tag{3a}$$

$$\mathbf{H}_l = \mathbf{A}_{dif} \odot \mathbf{G}^{W_m} \odot \mathbf{H}_m, \tag{3b}$$

where $\mathbf{H}_m = \mathbf{A}^{W_m} \odot \mathbf{G}^{W_m}$ following Equation (1). Broadcasting occurs when the dimensions do not match.

From the global aspect, we manually add the weight of the main word Grad-CAM $\mathbf{H}_m$ along the token axis during mean operations, which provides additional enhanced focus on the primary token. Therefore, we can obtain the global token-level Grad-CAM $\mathbf{H}_g$ by Equation (4)

$$W^{'} = W \cup \{W_m\} \times N_c, \tag{4a}$$

$$\mathbf{H}_g = \mathbf{A}^{W^{'}} \odot \mathbf{G}^{W^{'}} \tag{4b}$$

where N_c is the number of context tokens and $\{W_m\} \times N_c$ means repeating the main word for N_c times. Finally, the resulting augmented Grad-CAM, $\mathbf{H}_a$, is the mean of concatenated local and global Grad-CAMs , $\mathbf{H}_c$, along the token axis, where $\mathbf{H}_c = [\mathbf{H}_g, \mathbf{H}_l]$. This map significantly improves the model's Grad-CAM localization accuracy, shown in PWEM of Figure 2 (a).

Iterative Grad-CAM Refinement Strategy

Masked Language Modeling (MLM) can be used for bidirectional image generative Transformers such as MasGIT (Chang et al. 2022). The iterative generative paradigm offers self-correction chances for the model to optimize step-by-step in the latent space. Inspired by this, we propose a novel iterative strategy of Grad-CAM to gradually steer the model's attention to the region that the model is not attentive to initially, which brings benefits from two sides. On the one hand, for the circumstances in which Grad-CAM correctly localizes the instance initially, the gradually refined Grad-CAM can be better gathered around the targeted instance region. On the other hand, for the first incorrect localization, the model can attend to other semantic instances to recheck the Grad-CAM prediction, especially for the positional phrase inputs. The overall approach of IGRS is illustrated in the left part of Figure 4.

For simple notification, we use H_t to represent the resultant t-th iteration Grad-CAM from the PWEM $\mathbf{H}_a$. Equation (5) delineates the aggregation and refinement process of Grad-CAM representational updation, which entails the combination with Grad-CAM in the penultimate iteration step $(t-1)$, under the constraint of a zero initial condition

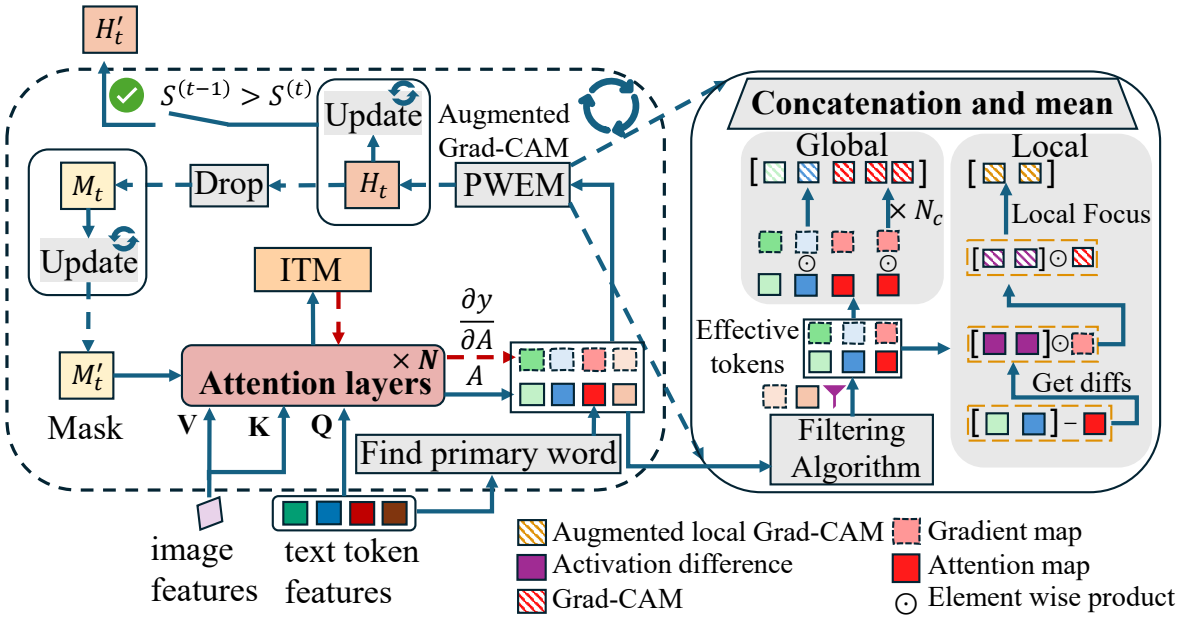

Figure 4: The proposed IGRS (left) and PWEM (right). The mask $M_t^{'}$ is the attention mask for cross-attention layers by dropping the most salient regions of Grad-CAM to zero. PWEM filters the meaningless tokens and augments the Grad-CAM representation from local and global aspects.

$H_0 = 0$.

$$H_t^{'} = \lambda H_{t-1}^{'} + (1-\lambda)\sigma(H_t), \tag{5}$$

where $H_{t-1}^{'}$ and $H_t^{'}$ are the resultant refined heatmaps from the $(t-1)$-th and t-th iterations, $\sigma(.)$ is a sigmoid function to scale the value appropriately, and the hyperparameter λ is a balancing factor. To instruct the model to focus on the region previously not paid attention to, in each iteration, a binary attention mask M_t would be generated from the refined Grad-CAM heatmap H_t by dropping the most attentive region to 0, as shown in Equation (6).

$$M_t = \mathcal{P}(H_t, \theta), \mathcal{P}(H, \theta) = \begin{cases} 0 & \text{if } \sigma(H) \geq \theta \\ 1 & \text{if } \sigma(H) < \theta, \end{cases} \tag{6}$$

where $\mathcal{P}(H, \theta)$ represents the process of applying a sigmoid function to stretch the values and then thresholding the result at θ to create a binary mask. The binary mask M_t is then combined with the previous mask $M_{t-1}^{'}$ by the logical *and* operation, $M_t^{'} = M_{t-1}^{'} \wedge M_t$, where M_0 is a tensor of ones. The $\wedge$ ensures the model can expand to other regions regardless of the places previously focused. This attention binary mask will be fed into the cross-attention layer of a VLP to mask out the visual regions in embedding v, ensuring the text token queries no longer pay attention to the zero regions within the mask.

For an interactive algorithm, the stopping condition is essential. To make the iterative process more flexible, we introduce a dynamic stopping criterion based on the proposed soft ITM score at timestep t, $S^{(t)}$, which is calculated as the product of the ITM, from the VLP model and the relevance score, $S^{(t)} = ITM^{(t)} \cdot R^{(t)}$, where $R^{(t)}$ is defined by:

$$R^{(t)} = \frac{\sum_{x \in X} \sum_{y \in Y} (1 - \widetilde{H}_{t-1}^{'})}{X \times Y}, \tag{7}$$

where $\widetilde{H}_{t-1}^{'}$ is the interpolated Grad-CAM heatmap of $H_{t-1}^{'}$ with the same size as the original image with width X and height Y. This relevance score measures the average overlooked Grad-CAM intensity. A higher $R^{(t)}$ indicates that there are some regions less attentive to previously, guiding the model to focus on these overlooked areas in the next iteration. If the score for the current iteration $S^{(t)}$ is less than the score from the previous iteration $S^{(t-1)}$, the iterative process is terminated. The total iterative times should not exceed ν.

Selective Mask Proposal Network

Through the aforementioned steps, we can employ the Grad-CAM indication clue to instruct the Segmentors to predict the referred instance mask. For a given image, the mask proposal network would predict the N_b masks but they can not autonomously choose which object mask users refer to by the language. Therefore, the selection module within the network is designed to select the mask indicated by the Grad-CAM, which divides the selection procedures into two phases: the filtering phase and the scoring phase.

Assuming that the Grad-CAM has successfully localized the instance, the center point of Grad-CAM should be within the inner part of the object. Based on this hypothesis, the selection mechanism is initiated by a preliminary evaluation that involves two main criteria. First, we identify the set of coordinates, $\mathcal{C}_{max}$, where the heatmap reaches its peak

Methods	RefCOCO				RefCOCO+				RefCOCOg			Average
	val	testA	testB	avg.	val	testA	testB	avg.	val	test	avg.	
Zero-shot methods												
GL-CLIP (Yu, Seo, and Son 2023)	26.7	25.0	26.5	26.1	28.2	26.5	27.9	27.5	33.0	33.1	33.1	28.4
BSAP (Wang et al. 2024a)	27.3	27.0	27.1	27.1	28.7	27.8	28.3	28.3	34.5	34.5	34.5	29.4
Region token (Yu, Seo, and Son 2023)	23.4	22.1	24.6	23.4	24.5	22.6	25.4	24.2	27.6	27.3	27.5	24.7
SAM-CLIP (Ni et al. 2023)	26.3	25.8	26.4	26.2	25.7	28	26.8	26.8	38.8	38.9	38.9	29.6
Ref-Diff (Ni et al. 2023)	37.2	38.4	**37.2**	37.6	37.3	40.5	33	36.9	44	44.5	44.3	39.0
TAS (Suo, Zhu, and Yang 2023)	39.8	41.1	36.2	39.0	43.6	49.1	**36.5**	43.1	**46.6**	**46.8**	**46.7**	42.5
CaR (Sun et al. 2024)	33.6	35.4	30.5	33.0	34.2	36.0	31.0	33.7	36.7	36.6	36.7	34.3
Weakly-supervised methods												
TSEG (Strudel, Laptev, and Schmid 2022)	25.4	-	-	-	22.0	-	-	-	22.1	-	-	-
Chunk (Lee et al. 2023)	31.1	32.3	30.1	31.8	31.3	32.1	30.1	31.2	32.9	-	-	-
IteRPrimE (ours)	**40.2**	**46.5**	33.9	**40.2**	**44.2**	**51.6**	35.3	**43.7**	46.0	45.8	45.9	**42.9**

Table 1: Comparison of different methods on different datasets. "avg." denotes the mean performance across various splits within individual datasets, while the terminal "Average" column represents the composite mean derived from all dataset splits.

values. Then, the m-th candidate mask B^m is examined to determine if it includes at least one activated pixel at any of these coordinates. Second, to ensure the quality of the masks, we apply a connected component labeling technique to constrain the number of connected components in each mask, ensuring that the number of these components does not exceed a predefined threshold of κ. The combined criteria for the preliminary evaluation are defined as follows:

$$\begin{aligned} \mathcal{A} &= \left\{ m \in N_b \mid B^m_{(x,y)} \neq 0, \exists (x,y) \in \mathcal{C}_{max}, \right\}, \\ \mathcal{F} &= \{ m \in N_b \mid g_{cc}(B^m) \leq \kappa \}, \\ \mathcal{D} &= \mathcal{A} \cap \mathcal{F}. \end{aligned} \tag{8}$$

In the above equations, g_{cc} denotes a function that quantifies the number of connected components within the m-th candidate from total N_b masks. The intersection of sets $\mathcal{A}$ and $\mathcal{F}$, denoted as $\mathcal{D}$, yields the subset of candidate masks that fulfill both the activation and mask quality requirements. This evaluation process filers the irrelevant and empty masks to reduce the computational cost and enhance efficiency.

Subsequent to the preliminary filtering phase, we proceed to evaluate each remaining candidate mask through a weighted scoring mechanism that leverages the Grad-CAM heatmap. This involves computing an element-wise product-based score for each mask concerning the heatmap. We define the score for the j-th candidate mask as $Z(j)$ from the set $\mathcal{D}$. The scoring process is formulated below:

$$\begin{aligned} Z(j) &= \sum_{x \in X} \sum_{y \in Y} \left(B^j_{(x,y)} + B^j_{(x,y)} \odot \widetilde{H}'_{(x,y)} \right) \\ \hat{Z}(j) &= \frac{Z(j)}{\sum_{x \in X} \sum_{y \in Y} B^j_{(x,y)}}, \quad j \in \mathcal{D}. \end{aligned} \tag{9}$$

where $\widetilde{H}'_{(x,y)}$ is the final output Grad-CAM of original image size. The final step in our selection process involves identifying the candidate mask with the maximum normalized score, $\hat{Z}(j)$, as the chosen segmentation output:

$$B_{select} = \arg\max_{j \in \mathcal{D}} \hat{Z}(j). \tag{10}$$

Method	Training dataset	All	Unseen
CRIS	RefCOCO	15.5	13.8
	RefCOCO+	16.3	14.6
	RefCOCOg	16.2	13.9
LAVT	RefCOCO	16.7	14.4
	RefCOCO+	16.6	13.5
	RefCOCOg	16.1	13.5
GL-CLIP	N/A	23.6	23.0
TAS	N/A	25.6	-
Ref-Diff	N/A	29.4	-
IteRPrimE (ours)	N/A	**38.1**	**37.9**

Table 2: Comparison of oIoU on PhraseCut for different supervised and zero-shot methods.

This approach ensures that the selected mask aligns with the regions of interest highlighted by the Grad-CAM heatmap, thereby ensuring the precision and efficacy of RIS.

Experiments

Experimental Settings

Datasets and metrics. We employ the RefCOCO (Nagaraja, Morariu, and Davis 2016), RefCOCO+ (Nagaraja, Morariu, and Davis 2016), RefCOCOg (Kazemzadeh et al. 2014; Mao et al. 2016), and PhraseCut datasets (Wu et al. 2020) for evaluating the proposed zero-shot methods. RefCOCO with shorter expressions (average 1.6 nouns, 3.6 words) contains massive positional phrases (50%), especially those with direct direction clues like "left" or "right". In contrast, RefCOCO+ focuses on the attribute phrases with the same average expression length. RefCOCOg is a more challenging benchmark that has longer phrases (average 2.8 nouns, 8.4 words) and complex expressions. To verify the effectiveness of the mode in out-of-domains, we adapt our model to the PhraseCut dataset which contains the additional 1271 categories in the test split based on 80 in COCO. Following (Sun et al. 2024; Yu, Seo, and Son 2023; Han et al. 2024), we utilize the mean Intersection over Union (mIoU) for RefCOCO series, a common metric for RIS. Following (Yu,

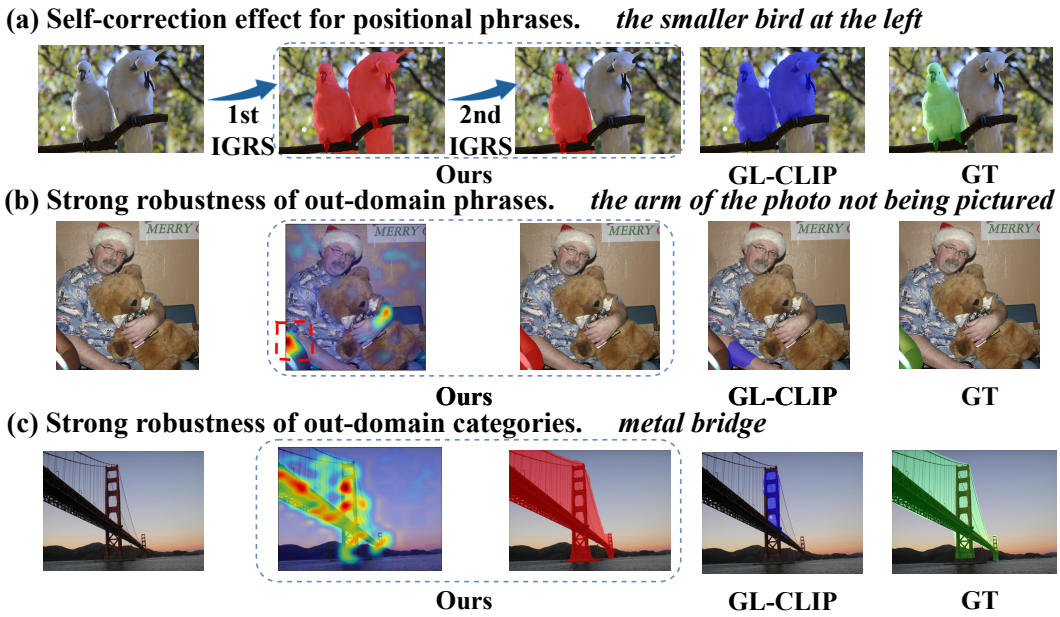

Figure 5: The qualitative comparisons with GL-CLIP. (a) The self-correction effect is brought by our IGRS, especially for positional phrases. (b) For the unseen phrases like "not", our model shows better robustness. (c) shows the gathering effect of IteRPrimE with high confidence to select the whole mask instead of a part like GL-CLIP.

Method	RefCOCO testA	RefCOCOg test
Overall Mean	43.4	41.3
GVLP (Shen et al. 2024)	45.0	41.9
Global Augment	**46.5**	45.7
Local Augment	45.3	42.3
PWEM	**46.5**	**45.8**

Table 3: Comparison of methods with different Grad-CAM generation methods on RefCOCO testA and RefCOCOg test datasets.

Method	RefCOCO testA	RefCOCOg test	RefCOCOg val
Mask image	46.3	45.3	45.2
Mask feature	**46.5**	**45.8**	**46.0**

Table 4: Performance comparison of masking out the salient regions in the image level and feature level (attention mask).

Seo, and Son 2023; Wu et al. 2020), we report the overall Intersection over Union (oIoU) for the PhraseCut dataset.

Implementation details. We use the commonly used mask proposal network, Mask2Former (Cheng et al. 2022; Liang et al. 2023), to obtain 200 instance-level mask proposals. Following (Shen et al. 2024; Lee et al. 2023), we utilize the base model ALBEF to study the Grad-CAM for localization and it is generated in the 8th cross-attention layer. In processing the input text, a prefatory phrase "there is a" is appended. The hyperparameter balancing factor λ, upper connecting limit κ, iterative number ν, and binarization threshold θ are 0.8, 12, 3, and 0.5, respectively. All experiments are conducted on a 24 GB RTX 3090 GPU.

Results

Main results. As shown in Table 1, IteRPrimE almost achieves the best performance on all three datasets, especially in the testA splits of RefCOCO and RefCOCO+. It outperforms the SOTA TAS method with a 0.4% average improvement. For all the splits of RefCOCO and RefCOCO+ rich in short positional phrases, our model obtains an average of 40.2% and 43.7% compared to the 39.0% and 43.1% of TAS, respectively. Therefore, our method is more robust to the positional information compared to the CLIP-based paradigms. However, the model may have the relatively weaker capability of complex expressions shown in RefCOCOg, which can be attributed to the data limitation and gap in the pertaining stage. Additionally, by using the additional captioner of BLIP2 (Li et al. 2023) and SAM (Kirillov et al. 2023), TAS maintains the best performance across some splits, especially for complex phrases, but it has the drawback of low throughput and heavy volumes.

Zero-shot evaluation on unseen domain. Notably, as shown in Table 2, our model has high capabilities of cross-domain zero-shot transfer compared to other zero-shot SOTA and the existing supervised methods CRIS (Wang et al. 2022b) and LAVT (Yang et al. 2022). IteRPrimE significantly outperforms both kinds of methods in the out-domain scenarios. Upon assessment within a subset of categories not present in the RefCOCO datasets (denoted as the "Unseen" column), our model shows the best robustness compared to the supervised methods with huge performance degradation. Notably, the underperformance of the TAS model on this dataset may be attributed to the predominance of complex outdoor scenes within the dataset. In such intricate environments, the reliance on an additional

Method	RefCOCOg test			RefCOCO testA		
	Position	Others	Overall	Position	Others	Overall
GVLP w/o IGRS	33.0	43.6	41.3	34.7	53.2	44.7
GVLP w/ IGRS	33.7	44.3	41.9	35.1	53.6	45.0
PWEM w/o IGRS	36.4	47.5	45.1	36.1	54.8	46.1
PWEM w/ IGRS	**37.4**	**48.2**	**45.8**	**36.5**	**55.0**	**46.5**

Table 5: Ablation studies of the proposed PWEM and IGRS.

captioning model for annotation by TAS could potentially introduce greater noise, thereby compromising the model's performance. However, facing complex environmental contexts, our model's efficacy in localizing pertinent regions is attributed to its retention of spatial perception. Concurrently, the integration of IGRS and PWEM has further bolstered IteRPrimE's proficiency in addressing the complicated interrelationships among objects within the scene, thereby leading to this commendable performance.

Qualitative comparisons. Figure 5 shows the comparisons with GL-CLIP (Yu, Seo, and Son 2023). First, we demonstrate that our IGRS module possesses a self-corrective mechanism, the same answer as GL-CLIP initially before refining its predictions by revisiting initially overlooked regions. In Figure 5 (b), the scarcity of such negative phrases in the training set is offset by our model's robustness. Finally, we address the limited highlighted region of initial Grad-CAM representation by the IGRS, demonstrated in Figure 5 (c). The more gathering of the Grad-CAM, the more likelihood that the correct instance mask will be selected instead of the part.

Ablation Study

Effect of PWEM. According to Equation (2), the mean operation is essential for the generation of Grad-CAM and deeply influences the Grad-CAM representational accuracy. Therefore, Table 3 presents the results of the ablation study, examining the impact of various aggregation configurations for Grad-CAM generation. The "Overall Mean" is the direct mean of all the tokens' Grad-CAM, but the GVLP uses the selected effective tokens for averaging (Shen et al. 2024). The remaining is introduced before as shown in Equation (3) and Equation (4). Compared to the previous methods, the proposed PWEM can significantly improve the performances because it can save the examples that fail due to the weak complex semantic understanding between the main word and the other contexts. Additionally, global augmentation shows stronger potential than local because it could dominate the effect during aggregation.

Effect of mask position in IGRS. Table 4 evaluates the position that the binary mask M applied. "Mask image" means adding the mask into the original image so that the indicated regions are masked out, similar to GL-CLIP. However, this can degrade the performance due to the absence of relative relationships of regions. Our method for attention masking in the cross-attention layer is more robust, with improvement on all three splits.

Effect of our proposed PWEM and IGRS. Table 5 evaluates the performance improvements achieved by integrating different modules within our methodology. The "Position" category encompasses those test samples that explicitly feature positional expressions. Conversely, the "Others" category serves as the complement. These results demonstrate that our modules not only improve general performance but also enhance the model's ability to manage complex semantic and spatial relations, particularly in positional contexts.

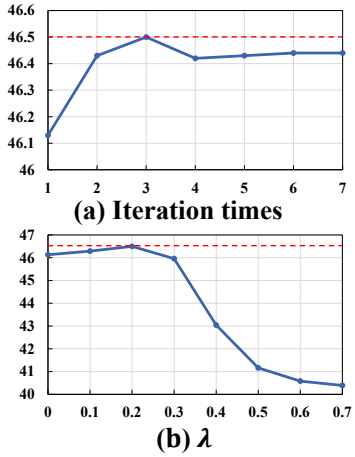

Figure 6: The line charts of two hyperparameters.

Different assembly of iteration times and λ. Figure 6 presents the ablation study of two hyperparameters in IGRS, analyzing the effect of varying iteration times and λ on the RefCOCO testA dataset. Figure 6 (a) shows that as the number of iterations increases from 1 to 3, the metric improves, peaking at 46.5%. However, beyond three iterations, the performance change becomes minimal. Therefore, selecting 3 iterations is optimal for balancing performance and time efficiency. Figure 6 (b) presents another line chart analyzing the impact of λ in the Grad-CAM updation. The metric increases as λ is gradually raised from 0 to 0.2 while exceeding this point, performance declines with higher alpha values. Overall, the optimal value of λ is 0.2.

Conclusion

This paper presents IteRPrimE, a novel framework for Zero-shot Referring Image Segmentation (RIS), addressing the limitations of previous methods in handling positional sensitivity and complex semantic relationships. By incorporating an Iterative Grad-CAM Refinement Strategy (IGRS) and a Primary Word Emphasis Module (PWEM), IteRPrimE enhances the model's ability to accurately focus on target regions and manage semantic nuances. Extensive experiments on RefCOCO/+/g and PhraseCut benchmarks demonstrate that IteRPrimE significantly outperforms previous state-of-the-art zero-shot methods, particularly in out-of-domain contexts. These findings highlight the framework's potential to advance zero-shot RIS by improving model sensitivity to positional and semantic details. Future research endeavors may seek to extend the Grad-CAM-guided RIS paradigm to encompass all segmentation tasks across varying levels of granularity with linguistic directives.

Acknowledgments

This work was supported by Shenzhen Science and Technology Program under Grant CJGJZD20220517142402006.

References

Bai, S.; Liu, Y.; Han, Y.; Zhang, H.; and Tang, Y. 2024. Self-calibrated clip for training-free open-vocabulary segmentation. *arXiv preprint arXiv:2411.15869*.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022. Maskgit: Masked generative image transformer. In *CVPR*, 11315–11325.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 1290–1299.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, H.; Liu, C.; He, S.; Jiang, X.; and Loy, C. C. 2023. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *CVPR*, 2694–2703.

Ding, H.; Liu, C.; Wang, S.; and Jiang, X. 2021. Vision-language transformer and query generation for referring segmentation. In *ICCV*, 16321–16330.

Ding, J.; Xue, N.; Xia, G.-S.; and Dai, D. 2022. Decoupling zero-shot semantic segmentation. In *CVPR*, 11583–11592.

Han, K.; Liu, Y.; Liew, J. H.; Ding, H.; Liu, J.; Wang, Y.; Tang, Y.; Yang, Y.; Feng, J.; Zhao, Y.; et al. 2023. Global knowledge calibration for fast open-vocabulary segmentation. In *CVPR*, 797–807.

Han, Z.; Zhu, F.; Lao, Q.; and Jiang, H. 2024. Zero-shot referring expression comprehension via structural similarity between images and captions. In *CVPR*, 14364–14374.

He, S.; Guo, T.; Dai, T.; Qiao, R.; Wu, C.; Shu, X.; and Ren, B. 2022. VLMAE: Vision-language masked autoencoder. *arXiv preprint arXiv:2208.09374*.

Jing, Y.; Kong, T.; Wang, W.; Wang, L.; Li, L.; and Tan, T. 2021. Locate then segment: A strong pipeline for referring image segmentation. In *CVPR*, 9858–9867.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 787–798.

Kim, N.; Kim, D.; Lan, C.; Zeng, W.; and Kwak, S. 2022. Restr: Convolution-free referring image segmentation using transformers. In *CVPR*, 18145–18154.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *ICCV*, 4015–4026.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *CVPR*, 9579–9589.

Lee, J.; Lee, S.; Nam, J.; Yu, S.; Do, J.; and Taghavi, T. 2023. Weakly supervised referring image segmentation with intra-chunk and inter-chunk consistency. In *ICCV*, 21870–21881.

Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; et al. 2022. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 19730–19742. PMLR.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; and Marculescu, D. 2023. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, 7061–7070.

Liu, C.; Ding, H.; and Jiang, X. 2023. Gres: Generalized referring expression segmentation. In *CVPR*, 23592–23601.

Liu, R.; Liu, C.; Bai, Y.; and Yuille, A. L. 2019. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *CVPR*, 4185–4194.

Liu, Y.; Bai, S.; Li, G.; Wang, Y.; and Tang, Y. 2024a. Open-vocabulary segmentation with semantic-assisted calibration. In *CVPR*, 3491–3500.

Liu, Y.; Zhang, C.; Wang, Y.; Wang, J.; Yang, Y.; and Tang, Y. 2024b. Universal segmentation at arbitrary granularity with language instruction. In *CVPR*, 3459–3469.

Luo, J.; Khandelwal, S.; Sigal, L.; and Li, B. 2024a. Emergent Open-Vocabulary Semantic Segmentation from Off-the-shelf Vision-Language Models. In *CVPR*, 4029–4040.

Luo, Z.; Xiao, Y.; Liu, Y.; Li, S.; Wang, Y.; Tang, Y.; Li, X.; and Yang, Y. 2024b. Soc: Semantic-assisted object cluster for referring video object segmentation. *Advances in Neural Information Processing Systems*, 36.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 11–20.

Nagaraja, V. K.; Morariu, V. I.; and Davis, L. S. 2016. Modeling context between objects for referring expression understanding. In *ECCV*, 792–807. Springer.

Ni, M.; Zhang, Y.; Feng, K.; Li, X.; Guo, Y.; and Zuo, W. 2023. Ref-diff: Zero-shot referring image segmentation with generative models. *arXiv preprint arXiv:2308.16777*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *CVPR*, 618–626.

Shah, N. A.; VS, V.; and Patel, V. M. 2024. LQMFormer: Language-aware Query Mask Transformer for Referring Image Segmentation. In *CVPR*, 12903–12913.

Shen, H.; Zhao, T.; Zhu, M.; and Yin, J. 2024. GroundVLP: Harnessing Zero-Shot Visual Grounding from Vision-Language Pre-training and Open-Vocabulary Object Detection. In *AAAI*, volume 38, 4766–4775.

Shin, G.; Xie, W.; and Albanie, S. 2022. Reco: Retrieve and co-segment for zero-shot transfer. *Advances in Neural Information Processing Systems*, 35: 33754–33767.

Strudel, R.; Laptev, I.; and Schmid, C. 2022. Weakly-supervised segmentation of referring expressions. *arXiv preprint arXiv:2205.04725*.

Sun, S.; Li, R.; Torr, P.; Gu, X.; and Li, S. 2024. Clip as rnn: Segment countless visual concepts without training endeavor. In *CVPR*, 13171–13182.

Suo, Y.; Zhu, L.; and Yang, Y. 2023. Text augmented spatial-aware zero-shot referring image segmentation. *EMNLP*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, H.; Zhan, Y.; Liu, L.; Ding, L.; Yang, Y.; and Yu, J. 2024a. Towards Alleviating Text-to-Image Retrieval Hallucination for CLIP in Zero-shot Learning. *arXiv preprint arXiv:2402.18400*.

Wang, X.; Yu, Z.; De Mello, S.; Kautz, J.; Anandkumar, A.; Shen, C.; and Alvarez, J. M. 2022a. Freesolo: Learning to segment objects without annotations. In *CVPR*, 14176–14186.

Wang, Y.; Zhao, R.; and Sun, Z. 2023. Efficient Remote Sensing Transformer for Coastline Detection with Sentinel-2 Satellite Imagery. In *IGARSS*, 5439–5442. IEEE.

Wang, Y.; Zhao, R.; Wei, S.; Ni, J.; Wu, M.; Luo, Y.; and Luo, C. 2024b. Convolution Meets Transformer: Efficient Hybrid Transformer for Semantic Segmentation with Very High Resolution Imagery. In *IGARSS 2024*, 9688–9691. IEEE.

Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; and Liu, T. 2022b. Cris: Clip-driven referring image segmentation. In *CVPR*, 11686–11695.

Wu, C.; Lin, Z.; Cohen, S.; Bui, T.; and Maji, S. 2020. Phrasecut: Language-based image segmentation in the wild. In *CVPR*, 10216–10225.

Xu, H.; Ye, Q.; Yan, M.; Shi, Y.; Ye, J.; Xu, Y.; Li, C.; Bi, B.; Qian, Q.; Wang, W.; et al. 2023a. mplug-2: A modularized multi-modal foundation model across text, image and video. In *ICML*, 38728–38748. PMLR.

Xu, J.; De Mello, S.; Liu, S.; Byeon, W.; Breuel, T.; Kautz, J.; and Wang, X. 2022. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, 18134–18144.

Xu, X.; Wu, C.; Rosenman, S.; Lal, V.; Che, W.; and Duan, N. 2023b. Bridgetower: Building bridges between encoders in vision-language representation learning. In *AAAI*, volume 37, 10637–10647.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 18155–18165.

Yang, Z.; Wang, J.; Ye, X.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2024. Language-aware vision transformer for referring segmentation. *IEEE TPAMI*.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Yu, S.; Seo, P. H.; and Son, J. 2023. Zero-shot referring image segmentation with global-local context features. In *CVPR*, 19456–19465.

Zhou, C.; Loy, C. C.; and Dai, B. 2022. Extract free dense labels from clip. In *ECCV*, 696–712. Springer.

Zhu, C.; and Chen, L. 2024. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE TPAMI*.