Zero-shot Referring Image Segmentation with Global-Local Context Features

Seonghoon Yu[1] Paul Hongsuck Seo[2] Jeany Son[1]
[1]AI Graduate School, GIST [2]Google Research
seonghoon@gm.gist.ac.kr phseo@google.com jeany@gist.ac.kr

Abstract

Referring image segmentation (RIS) aims to find a segmentation mask given a referring expression grounded to a region of the input image. Collecting labelled datasets for this task, however, is notoriously costly and labor-intensive. To overcome this issue, we propose a simple yet effective zero-shot referring image segmentation method by leveraging the pre-trained cross-modal knowledge from CLIP. In order to obtain segmentation masks grounded to the input text, we propose a mask-guided visual encoder that captures global and local contextual information of an input image. By utilizing instance masks obtained from off-the-shelf mask proposal techniques, our method is able to segment fine-detailed instance-level groundings. We also introduce a global-local text encoder where the global feature captures complex sentence-level semantics of the entire input expression while the local feature focuses on the target noun phrase extracted by a dependency parser. In our experiments, the proposed method outperforms several zero-shot baselines of the task and even the weakly supervised referring expression segmentation method with substantial margins. Our code is available at https://github.com/Seonghoon-Yu/Zero-shot-RIS.

1. Introduction

Recent advances of deep learning has revolutionised computer vision and natural language processing, and addressed various tasks in the field of vision-and-language [4, 19, 27, 28, 36, 43, 50]. A key element in the recent success of the multi-modal models such as CLIP [43] is the contrastive image-text pre-training on a large set of image and text pairs. It has shown a remarkable zero-shot transferability on a wide range of tasks, such as object detection [9, 10, 13], semantic segmentation [7, 12, 59, 63], image captioning [40], visual question answering (VQA) [47] and so on.

Despite its good transferability of pre-trained multi-modal models, it is not straightforward to handle dense prediction tasks such as object detection and image segmentation. A pixel-level dense prediction task is challenging since there is a substantial gap between the image-level contrastive pre-training task and the pixel-level downstream task such as semantic segmentation. There have been several attempts to reduce gap between two tasks [44, 54, 63], but these works aim to fine-tune the model consequently requiring task-specific dense annotations, which is notoriously labor-intensive and costly.

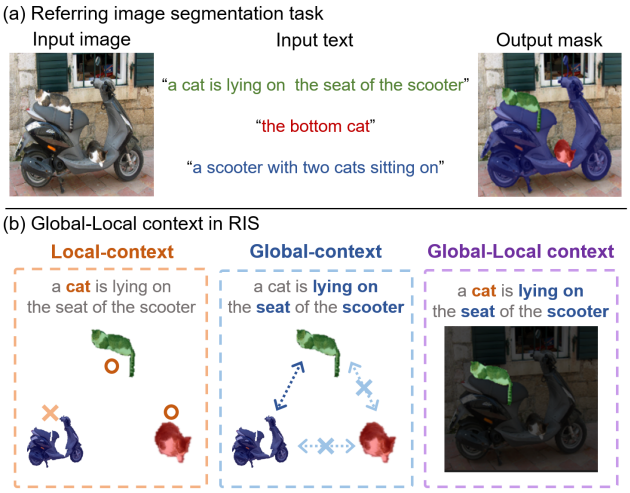

Figure 1. Illustrations of the task of referring image segmentation and motivations of global-local context features. To find the grounded mask given an expression, we need to understand the relations between the objects as well as their semantics.

Referring image segmentation is a task to find the specific region in an image given a natural language text describing the region, and it is well-known as one of challenging vision-and-language tasks. Collecting annotations for this task is even more challenging as the task requires to collect precise referring expression of the target region as well as its dense mask annotation. Recently, a weakly-supervised referring image segmentation method [48] is proposed to overcome this issue. However, it still requires high-level text expression annotations pairing with images for the target datasets and the performance of the method is far from that of the supervised methods. To tackle this issue, in this paper, we focus on zero-shot transferring from the pre-trained knowledge

of CLIP to the task of referring image segmentation.

Moreover, this task is challenging because it requires high-level understanding of language and comprehensive understanding of an image, as well as a dense instance-level prediction. There have been several works for zero-shot semantic segmentation [7, 12, 59, 63], but they cannot be directly extended to the zero-shot referring image segmentation task because it has different characteristics. Specifically, the semantic segmentation task does not need to distinguish instances, but the referring image segmentation task should be able to predict an instance-level segmentation mask. In addition, among multiple instances of the same class, only one instance described by the expression must be selected. For example, in Figure 1, there are two cats in the input image. If the input text is given by *"a cat is lying on the seat of the scooter"*, the cat with the green mask is the proper output. To find this correct mask, we need to understand the relation between the objects (*i.e. "lying on the seat"*) as well as their semantics (*i.e. "cat", "scooter"*).

In this paper, we propose a new baseline of zero-shot referring image segmentation task using a pre-trained model from CLIP, where global and local contexts of an image and an expression are handled in a consistent way. In order to localize an object mask region in an image given a textual referring expression, we propose a mask-guided visual encoder that captures global and local context information of an image given a mask. We also present a global-local textual encoder where the local-context is captured by a target noun phrase and the global context is captured by a whole sentence of the expressions. By combining features in two different context levels, our method is able to understand a comprehensive knowledge as well as a specific trait of the target object. Note that, although our method does not require any additional training on CLIP model, it outperforms all baselines and the weakly supervised referring image segmentation method with a big margin.

Our main contributions can be summarised as follows:

- We propose a new task of zero-shot referring image segmentation based on CLIP without any additional training. To the best of our knowledge, this is the first work to study the zero-shot referring image segmentation task.
- We present a visual encoder and a textual encoder that integrates global and local contexts of images and sentences, respectively. Although the modalities of two encoders are different, our visual and textual features are dealt in a consistent way.
- The proposed global-local context features take full advantage of CLIP to capture the target object semantics as well as the relations between the objects in both visual and textual modalities.
- Our method consistently shows outstanding results compared to several baseline methods, and also outperforms the weakly supervised referring image segmentation method with substantial margins.

2. Related Work

Zero-shot Transfer. Classical zero-shot learning aims to predict unseen classes that have not seen before by transferring the knowledge trained on the seen classes. Early works [3, 14, 34] leverage the pre-trained word embedding [5, 39] of class names or attributes and perform zero-shot prediction via mapping between visual representations of images and this word embedding. Recently, CLIP [43] and ALIGN [19] shed a new light on the zero-shot learning via large-scale image-text pre-training. They show the successive results on various downstream tasks via zero-shot knowledge transfer, such as image captioning [40], video action localization [51], image-text retrieval [1] and so on. Contrary to classical zero-shot learning, zero-shot transfer has an advantage of avoiding fine-tuning the pre-trained model on the task-specific dataset, where collecting datasets is time-consuming. There have been several works that apply CLIP encoders directly with tiny architectural modification without additional training for semantic segmentation [63], referring expression comprehension [49], phrase localization [25] and object localization [17]. Our work is also lying on the line of this research field.

Zero-shot Dense Prediction Tasks. Very recently, with the success of pre-training models using large-scale image-text pairs, there have been several attempts to deal with dense prediction tasks with CLIP, *e.g.* object detection [9, 10, 13, 24, 30, 45], semantic segmentation [22, 29, 37, 42, 44, 58, 59, 63, 64] and so on. These dense prediction tasks, however, are challenging since CLIP learns image-level features not pixel-level fine-grained features. In order to handle this issue, ViLD [13] introduces a method which crop the image to contain only the bounding box region, and then extract the visual features of cropped regions using CLIP to classify the unseen objects. This approach is applied in a wide range of dense prediction tasks which are demanded the zero-shot transfer ability of CLIP [7, 9, 10, 12, 49, 59]. While this method only considers the cropped area, there are several methods [25, 63] to consider the global context in the image, not only just the cropped region. Adapting CLIP [25] proposed the phrase localization method by modifying CLIP to generate high-resolution spatial feature maps using superpixels. MaskCLIP [63] modifies the image encoder of CLIP by transforming the value embedding layer and the last linear layer into two 1×1 convolutional layers to handle pixel-level predictions. In this work, we focus on extracting both global and local context visual features with CLIP.

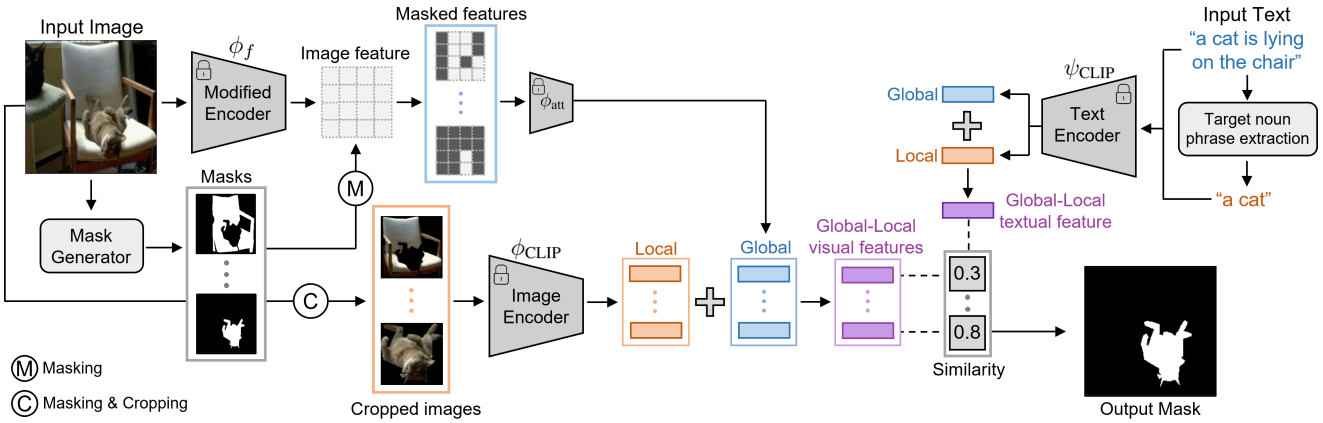

Figure 2. Overall framework of our global-local CLIP. Given an image and an expression as inputs, we extract global-local context visual features using mask proposals, and also we extract a global-local context textual feature. After computing the cosine similarity scores between all global-local context visual features and a global-local context textual feature, we choose the mask with the highest score.

Referring Image Segmentation. Referring image segmentation aims to segment a target object in an image given a natural linguistic expression introduced by [18]. There have been several fully-supervised methods for this task, where images and expressions are used as an input, and the target mask is given for training [2, 20, 33, 55, 60, 62]. Most of works [6, 11, 23, 60, 61] focuses on how to fuse those two features in different modalities extracted from independent encoders. Early works [26, 32] extract multi-modal features by simply concatenating visual and textual features and feed them into the segmentation networks [35] to predict dense segmentation masks. There have been two branches of works fusing cross-modal features; an attention based encoder fusion [11, 57, 60] and a cross-modal decoder fusion based on a Transformer decoder [6, 54, 61]. Recently, a CLIP-based approach, which learns separated image and text transformer using a contrastive pre-training, has been proposed [54]. Those fully supervised referring image segmentation methods show good performances in general, but they require dense annotations for target masks and comprehensive expressions describing the target object. To address this problem, TSEG [48] proposed a weakly-supervised referring image segmentation method which learns the segmentation model using text-based image-level supervisions. However, this method still requires high-level referring expression annotations with images for specific datasets. Therefore, we propose a new baseline for zero-shot referring image segmentation without any training or supervisions.

3. Method

In this section, we present the proposed method for zero-shot referring image segmentation in detail. We first show an overall framework of the proposed method (3.1), and then discuss the detailed methods for extracting visual features (3.2) and textual features (3.3) to encode global and local contextual information.

3.1. Overall Framework

To solve the task of referring image segmentation, which aims to predict the target region grounded to the text description, it is essential to learn image and text representations in a shared embedding space. To this end, we adopt CLIP to leverage the pre-trained cross-modal features for images and natural language.

Our framework consists of two parts as shown in Fig 2: (1) global-local visual encoder for visual representation, and (2) global-local natural language encoder for referring expression representation. Given a set of mask proposals generated by an unsupervised mask generator [52, 53], we first extract two visual features in global-context and local-context levels for each mask proposal, and then combine them into a single visual feature. Our global-context visual features can comprehensively represent the masked area as well as the surrounding region, while the local-context visual features can capture the representation of the specific masked region. This acts key roles in the referring image segmentation task because we need to focus a small specific target region using a comprehensive expression of the target. At the same time, given a sentence of expressing the target, our textual representation is extracted by the CLIP text encoder. In order to understand a holistic expression of the target as well as to focus on the target object itself, we first extract a key noun phrase from a sentence using a dependency parsing provided by spaCy [16], and then combine a global sentence feature and a local target noun phrase feature. Note that, our visual and text encoders are designed to handle both global-context and local-context information in a consistent way.

Since our method is built on CLIP where the visual and textual features are embedded in the common embedding

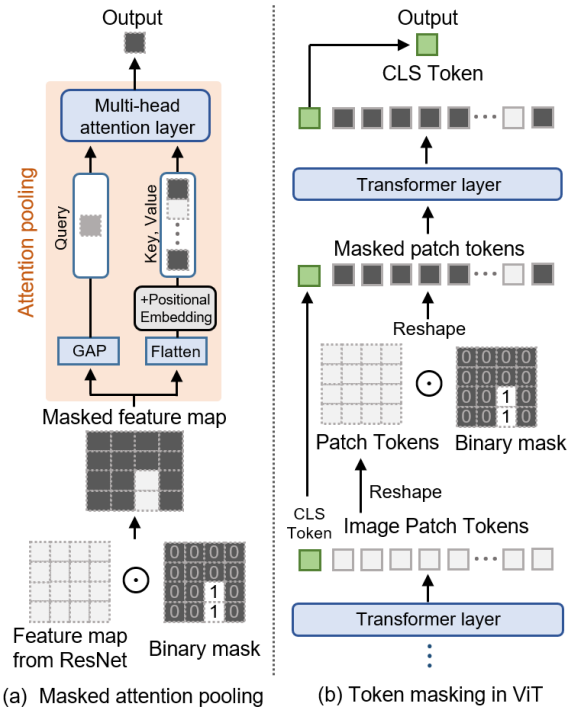

Figure 3. Detailed illustration of our mask-guided global-context visual encoders in ResNet and ViT architectures: (a) Masked attention pooling in ResNet, (b) Token masking in ViT.

space, we can formulate the objective of our zero-shot image referring segmentation task as follows. Given inputs of an image I and a referring expression T, our method finds the mask that has the maximum similarity between its visual feature and the given textual feature among all mask proposals:

$$\hat{m} = \arg \max_{m \in M(I)} \text{sim}(\mathbf{t}, \mathbf{f}_m), \tag{1}$$

where $\text{sim}(\cdot, \cdot)$ is a cosine similarity, $\mathbf{t}$ is the proposed global-local textual feature for a referring expression T, $\mathbf{f}$ is the mask-guided global-local visual feature, and $M(I)$ is a mask proposal set for a given image I.

3.2. Mask-guided Global-local Visual Features

To segment the target region related to the referring expression, it is essential to understand a global relationship between multiple objects in the image as well as local semantic information of the target. In this section, we demonstrate how to extract global and local-context features using CLIP, and how to fuse them.

Since CLIP is designed to learn image-level representation, it is not well-suited for a pixel-level dense prediction such as an image segmentation. To overcome the limitation of using CLIP, we decompose the task into two sub-tasks: mask proposal generation and masked image-text matching.

In order to generate mask proposals, we use the off-the-shelf mask extractor [53] which is the unsupervised instance-level mask generation model. By using mask proposals explicitly, our method can handle fine-detailed instance-level segmentation masks with CLIP.

Global-context Visual Features. For each mask proposals, we first extract global-context visual features using the CLIP pre-trained model. The original visual features from CLIP, however, is designed to generate one single feature vector to describe the whole image. To tackle this issue, we modify a visual encoder from CLIP to extract features that contain information from not only the masked region but also surrounding regions to understand relationships between multiple objects.

In this paper, we use two different architectures for the visual encoder as in CLIP: ResNet [15] and Vision Transformer (ViT) [8]. For the visual encoder with the ResNet architecture, we denote a visual feature extractor without a pooling layer as ϕ_{f} and its attention pooling layer as ϕ_{att}. Then the visual feature, $\mathbf{f}$, using the visual encoder of CLIP, ϕ_{CLIP}, can be expressed as follows:

$$\mathbf{f} = \phi_{\text{CLIP}}(I) = \phi_{\text{att}}(\phi_{\text{f}}(I)), \tag{2}$$

where I is a given image. Similarly, since ViT has multiple multi-head attention layers, we divide this visual encoder into two parts: last k layers and the rest. We denote the former one by ϕ_{att}, and the later one by ϕ_{f} for ViT architectures based on CLIP.

Then given an image I and a mask m, our global-context visual feature is defined as follows:

$$\mathbf{f}_m^G = \phi_{\text{att}}(\phi_f(I) \odot \bar{m}), \tag{3}$$

where $\bar{m}$ is the resized mask scaled to the size of the feature map, and $\odot$ is a Hadamard product operation. We illustrate more details of this masking strategy for each architecture of CLIP in Section 4.1 and Figure 3.

We refer to it as the global context visual feature, because the entire image is passed through the encoder and the feature map at the last layer contain the holistic information about the image. Although we use mask proposals to obtain the features only on masked regions on the feature map, these features already have comprehensive information about the scene.

Local-context Visual Features. To obtain local-context visual features given a mask proposal, we first mask the image and then crop the image to obtain a new image surrounding only an area of the mask proposal. After cropping and masking the image, it is passed to the visual encoder of CLIP to extract our local-context visual feature $\mathbf{f}_m^L$:

$$\mathbf{f}_m^L = \phi_{\text{CLIP}}(\mathcal{T}_{\text{crop}}(I \odot m)), \tag{4}$$

where $\mathcal{T}_{crop}(\cdot)$ denotes a cropping operation. This approach is commonly used in zero-shot semantic segmentation methods [7, 59]. Since this feature focuses on the masked region in the image where irrelevant regions are removed, it concentrates only on the target object itself.

Global-local Context Visual features. We aggregate global- and local-context features over masked regions to obtain one single visual feature that describe a representation of masked regions of the image. The global-local context visual feature is computed as follows:

$$\mathbf{f}_m = \alpha\, \mathbf{f}_m^G + (1-\alpha)\, \mathbf{f}_m^L, \tag{5}$$

where $\alpha \in [0, 1]$ is a constant parameter, m is a mask proposal, $\mathbf{f}^G$ and $\mathbf{f}^L$ are global-context and local-context visual features in Eq. (3) and Eq. (4), respectively. As in Eq. (1), the score for each mask proposal is then obtained by computing similarity between our global-local context visual features and the textual feature of the expression described in the next section.

3.3. Global-local Textual Features

Similar to the visual features, it is important to understand a holistic meaning as well as the target object noun in given expressions. Given a referring expression T, we extract a global sentence feature, $\mathbf{t}^G$, using the pre-trained CLIP text encoder, ψ_{CLIP}, as follows:

$$\mathbf{t}^G = \psi_{\text{CLIP}}(T). \tag{6}$$

Although the CLIP text encoder can extract the textual representation aligning with the image-level representation, it is hard to focus on the target noun in the expression because the expression of this task is formed as a complex sentence containing multiple clauses, *e.g. "a dark brown leather sofa behind a foot stool that has a laptop computer on it"*.

To address this problem, we exploit a dependency parsing using spaCy [16] to find the target noun phrase, NP(T), given the text expression T. To find the target noun phrase, we first find all noun phrases in the expression, and then select the target noun phrase that contains the root noun of the sentence. After identifying the target noun phrase in the input sentence, we extract the local-context textual feature from the CLIP textual encoder:

$$\mathbf{t}^L = \psi_{\text{CLIP}}(\text{NP}(T)). \tag{7}$$

Finally, our global-local context textual feature is computed by a weighted sum of the global and local textual features described in Eq. (6) and Eq. (7) as follows:

$$\mathbf{t} = \beta\, \mathbf{t}^G + (1-\beta)\, \mathbf{t}^L, \tag{8}$$

where $\beta \in [0, 1]$ is a constant parameter, $\mathbf{t}^G$ and $\mathbf{t}^L$ are global sentence and local noun-phrase textual features, respectively.

4. Implementation Details

We use unsupervised instance segmentation methods, FreeSOLO [53], to obtain mask proposals, and the shorter size of an input image is set to 800. For CLIP, the size of an image is set to 224x224. The number of masking layers, k in ViT is set to 3. We set $\alpha = 0.85$ for RefCOCOg, 0.95 for RefCOCO and RefCOCO+, and $\beta = 0.5$ for all datasets.

4.1. Masking in Global-context Visual Encoder

We use both ResNet-50 and ViT-B/32 architectures for the CLIP visual encoder. Masking strategies of the global-context visual encoder for these two architecture are mostly similar but have small differences, described next.

Masked Attention Pooling in ResNet [15]. In a ResNet-based visual encoder of the original CLIP, a global average pooling layer is replaced by an attention pooling layer. This attention pooling layer has the same architecture as the multi-head attention in a Transformer. A *query* of the attention pooling layer is computed by a global average pooling operation onto the feature maps extracted by the ResNet backbone. A *key* and a *value* of the attention pooling layer is given by a flattened feature map. In our masked attention pooling, we mask the feature map using a given mask before computing *query*, *key* and *value*. After masking feature maps, we compute *query*, *key* and *value*, and then they are fed into the multi-head attention layer. The detailed illustration of our masked attention pooling in ResNet is shown in Figure 3a.

Token Masking in ViT [8]. Following ViT, we divide an image into grid patches, and embed patches to a linear layer with positional embeddings to get tokens, and then process those tokens with a series of Transformer layer. To capture global-context of images, we mask tokens in only the last k Transformer layers. The tokens are reshaped and masked by a given mask proposal, and then flattened and applied to the subsequent Transformer layer. As ViT has a class token (CLS), we use the final output feature from this CLS token as our global-context visual representation. The detailed method of our token masking in ViT is also shown in Figure 3b. In our experiments, we use ViT-B/32 architecture for the backbone of our ViT-based visual encoder, and we apply a token masking to the last 3 layers in the visual encoder. We show the performances with respect to the location of token masking layers in the supplementary materials.

5. Experiments

5.1. Datasets and Metrics

We evaluate our method on RefCOCO [41], RefCOCO+ [41] and RefCOCOg [21, 38], where the images and masks in MS-COCO [31] dataset are used to annotate

Table 1. Comparison with Zero-shot RIS baseline methods on three standard benchmark datasets. U: The UMD partition. G: The Google partition. All baseline methods use FreeSOLO as the mask proposal network. † denotes that the model is initialized with the ImageNet pre-trained weights and trained on RIS datasets. FreeSOLO upper-bound is computed between the GT mask and the maximum overlapped FreeSOLO mask with the GT mask.

Metric	Methods	Visual Encoder	RefCOCO			RefCOCO+			RefCOCOg		
			val	test A	test B	val	test A	test B	val(U)	test(U)	val(G)
oIoU	Supervised SoTA method [60]		72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	60.50
	Zero-Shot Baselines										
	Grad-CAM	ResNet-50	14.02	15.07	13.49	14.46	14.97	14.04	12.51	12.81	12.86
	Score map	ResNet-50	19.87	19.31	20.22	20.37	19.65	20.75	18.88	19.16	19.15
	Region token	ViT-B/32	21.71	20.31	22.63	22.61	20.91	23.46	25.52	25.38	25.29
	Cropping	ResNet-50	22.36	20.49	22.69	23.95	22.03	23.49	28.20	27.64	27.47
	Cropping	ViT-B/32	22.73	21.11	23.08	24.09	22.42	23.93	28.69	27.51	27.70
	Global-Local CLIP (ours)	ResNet-50	24.58	23.38	24.35	25.87	24.61	25.61	30.07	29.83	29.45
	Global-Local CLIP (ours)	ViT-B/32	**24.88**	**23.61**	**24.66**	**26.16**	**24.90**	**25.83**	**31.11**	**30.96**	**30.69**
	FreeSOLO upper-bound	-	42.08	42.52	43.52	42.17	42.52	43.80	48.81	48.96	48.49
mIoU	**Zero-Shot Baselines**										
	Grad-CAM	ResNet-50	14.22	15.93	13.18	14.80	15.87	13.78	12.47	13.16	13.30
	Score map	ResNet-50	21.32	20.96	21.57	21.61	21.17	22.30	20.07	20.43	20.63
	Region token	ViT-B/32	23.43	22.07	24.62	24.51	22.64	25.37	27.57	27.34	27.69
	Cropping	ResNet-50	24.31	22.37	24.66	26.31	23.94	25.69	31.27	30.87	30.78
	Cropping	ViT-B/32	24.83	22.58	25.72	26.33	24.06	26.46	31.88	30.94	31.06
	Global-Local CLIP (ours)	ResNet-50	**26.70**	**24.99**	26.48	**28.22**	**26.54**	**27.86**	33.02	33.12	32.79
	Global-Local CLIP (ours)	ViT-B/32	26.20	24.94	**26.56**	27.80	25.64	27.84	**33.52**	**33.67**	**33.61**
	FreeSOLO upper-bound	-	48.25	46.62	50.43	48.28	46.62	50.62	52.44	52.91	52.76
	Weakly-supervised method										
	TSEG [48]	ViT-S/16†	25.95	-	-	22.62	-	-	23.41	-	-

Method	Train dataset	oIoU on PhraseCut	
		All	Unseen
CRIS	RefCOCO	15.53	13.75
	RefCOCO+	16.30	14.62
	RefCOCOg	16.24	13.88
LAVT	RefCOCO	16.68	14.43
	RefCOCO+	16.64	13.49
	RefCOCOg	16.05	13.48
Ours	N/A	**23.64**	**22.98**

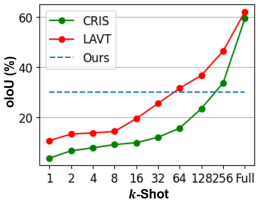

Figure 4. Comparisons to supervised methods in zero-shot setting on PhraseCut (left), and in few-shot setting on RefCOCOg (right). Unseen denotes a subset with classes that are not seen in RefCOCO.

the ground-truth of the referring image segmentation task. RefCOCO, RefCOCO+ and RefCOCOg have 19,994, 19,992 and 26,711 images with 142,210, 141,564 and 104,560 referring expressions, respectively. RefCOCO and RefCOCO+ have shorter expressions and an average of 1.6 nouns and 3.6 words are included in one expression, while RefCOCOg expresses more complex relations with longer sentences and has an average of about 2.8 nouns and 8.4 words. The detailed statistics of those datasets are demonstrated in our supplementary materials.

For the evaluation metrics, we use the overall Intersection over Union (oIoU) and the mean Intersection over Union (mIoU) which are the common metrics for the referring image segmentation task. The oIoU is measured by the total area of intersection divided by the total area of union, where the total area is computed by accumulating over all examples. In our ablation study, we use oIoUs since most of supervised RIS methods [6, 23] adopt it. We also report the mIoUs as in [48], which computes the average IoU across all examples while considering the object sizes.

Table 2. oIoU results of our method and the baselines using COCO instance GT masks. We use a ViT-B/32 model for a visual encoder.

Method	RefCOCO	RefCOCO+	RefCOCOg
Grad-CAM	18.32	18.14	21.24
Score map	23.97	25.50	28.11
Region token	35.59	38.13	40.19
Cropping	36.32	42.07	47.42
Ours	**37.05**	**42.59**	**51.01**

Table 3. oIoU results with different context-level features on the val split of RefCOCOg. We use a ViT-B/32 model for a visual encoder.

Encoder Variants		Textual features		
		Global	Local	Global-Local
Visual features	Global	27.03	27.37	27.60
	Local	28.69	25.23	29.48
	Global-Local	30.18	27.94	**31.11**

5.2. Baselines

We modify some baseline methods extracting dense predictions from CLIP into zero-shot RIS task to compare with our framework, and use FreeSOLO [53] as a mask generator in all baselines.

- **Grad-CAM:** The first baseline is utilizing gradient-based activation map based on Grad-CAM [46] which has been verified in the prior work [17]. After obtaining the activa-

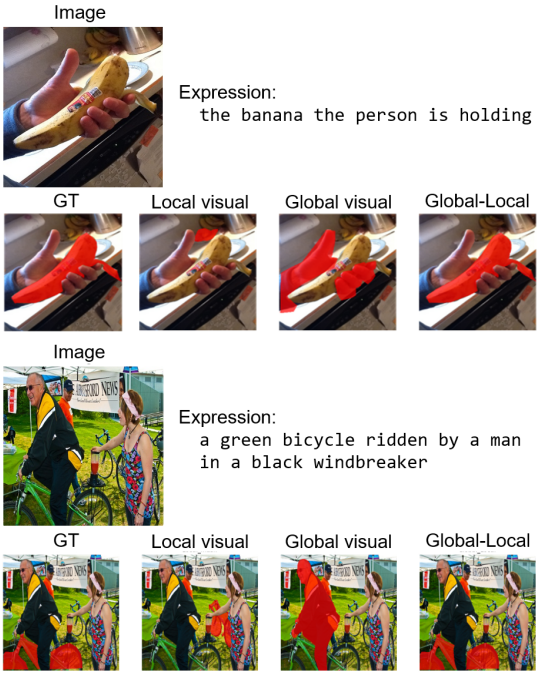

Figure 5. Qualitative results with different levels of visual features. COCO instance GT masks are used as mask proposals to validate the effect of the global-local context visual features.

tion maps using the similarity score of image-text pairs, we mask the maps and aggregate scores for all mask proposals, and select the mask with the highest score.

- **Score Map:** The second baseline is the method extracting a dense score map as in MaskCLIP [63]. As in MaskCLIP, to obtain dense score maps without pooling, a *value* linear layer and the last layer in the attention pooling are transformed into two consecutive 1×1 convolution layers. The feature map extracted from ResNet is forwarded to those two layers to get language-compatible dense image feature map, and then compute a cosine similarity with CLIP's textual feature. After obtaining a score map, we project mask proposals to a score map. The scores in the mask area are averaged and then we select the mask with the maximum score.

- **Region Token in ViT:** The third baseline is a method used in Adapting CLIP [25]. Similar to Adapting CLIP, we use region tokens for each mask proposal for all Transformer layers in CLIP's visual encoder instead of using superpixels. We finally compute the cosine similarity between each class token of a mask proposal and CLIP's textual feature, and then choose the mask with the highest score.

- **Cropping:** The last baseline is our local-context visual features described in Section 3.2. Cropping and masking is a commonly used approach utilizing CLIP for extracting mask or box region feature in a range of zero-shot dense prediction tasks [7, 9, 13, 49, 59]. Therefore, we consider cropping as one of the zero-shot RIS baselines.

Figure 6. Qualitative results with different levels of textual features using COCO Instance GT mask proposals.

5.3. Results

Main Results. We report referring image segmentation performances of our global-local CLIP and other baselines on RefCOCO, RefCOCO+ and RefCOCOg in terms of oIoU and mIoU metrics in Table 1. For a fair comparison, all methods including baselines use FreeSOLO [53] mask proposals to produce the final output mask. The experimental results show that our method outperforms other baseline methods with substantial margins. Our method also surpasses the weakly supervised referring image segmentation method (TSEG) [48] in terms of mIoU[1]. We also show upper-bound performances of using FreeSOLO, where the scores are computed by the IoU between ground-truth masks and its max-overlapped mask proposal. Although there is still a gap compared to the fully-supervised referring image segmentation methods, our method improves performances significantly compared to the baselines with the same upper-bound.

Zero-shot Evaluation on Unseen Domain. To verify the effectiveness of our method in a more practical setting, we report the zero-shot evaluation results with SoTA supervised methods [54, 60] on the test split of PhraseCut [56] in Figure 4 (left). Note that, RefCOCO contains expressions for only 80 salient object classes, whereas PhraseCut covers a variety of additional visual concepts *i.e.* 1272 categories in the test set. Our method outperforms both supervised methods, even though our models were never trained under RIS supervision. When evaluated on a subset of classes that are not seen in the RefCOCO datasets (*Unseen* column), the supervised methods show significant performance degradation, whereas our method works robustly on this subset.

[1]We only compare mIoU scores with TSEG since it reports only mIoU scores in the paper.

Figure 7. Qualitative results of our method with the several baselines. Note that all methods use mask proposals generated by FreeSOLO.

Comparison to supervised methods in few-shot Setting. We also compare our model to two supervised RIS methods [54, 60] in a few-shot learning setting, where the training set includes k instances for each of 80 classes in RefCOCO[2]. Note that the supervised methods use additional forms of supervision in training, whereas our method does not require any form of training or additional supervision; thus this setting is even disadvantageous to our method. Figure 4 (right) shows oIoU while varying k on RefCOCOg. The results clearly show that our method outperforms both supervised methods with large margins when k is small, and the gaps narrow as k gets larger (64 and 256 for LAVT [60] and CRIS [54], respectively). Note that it covers about 10% of the training set when $k = 64$ and the same trends hold for both RefCOCO and RefCOCO+.

5.4. Ablation Study

Effects of Mask Quality. To show the impact of the proposed method without considering the mask quality of the mask generators, we evaluate the performance of our method and the baselines with COCO instance GT masks in Table 2. Our approach has demonstrated superior performance compared to all baselines and has shown a performance improvement of over 3.5%, particularly on RefCOCOg which includes longer expressions. We believe that our method performs well on challenging examples that involve complex expressions, such as those with multiple clauses, which require an understanding of both the language and the scene.

Effects of Global-Local Context Features. We also study the effects of global-local context features in both visual and textual modalities and show the results in Table 3. For this analysis, we use RefCOCOg as it contains more complex expressions with multiple clauses. Among all combinations of two modalities, using both global-local context features in the visual and textual domains leads to the best performance.

Qualitative Analysis. We demonstrate several results that support the effectiveness of our global-local context visual features in Figure 5. To show this effect more clearly, we use COCO instance GT masks as mask proposals. When using only local-context visual features, the predicted mask tends to focus on the instance that shares the same class as the target object. However, when using only global-context visual features, the predicted mask tends to capture the context of the expression but may focus on a different object class. By combining global and local context, our method successfully finds the target mask. We also demonstrate the effectiveness of our global-local context textual features in Figure 6. Furthermore, we compare the qualitative results of our method with baseline methods in Figure 7. Our proposed global-local CLIP outperforms the baseline methods in identifying the target object by taking into account the global context of the image and expression.

6. Conclusion

In this paper, we propose a simple yet effective zero-shot referring image segmentation framework focusing on transferring knowledges from image-text cross-modal representations of CLIP. To tackle the difficulty of the referring image segmentation task, we propose global-local context encodings to compute similarities between images and expressions, where both target object semantics and relations between the objects are dealt in a unified framework. The proposed method significantly outperforms all baseline methods and weakly supervised method as well.

Acknowledgement. This work was supported by the IITP grants (No.2019-0-01842, No.2021-0-02068, No.2022-0-00926) funded by MSIT, the ISTD program (No.20018334) funded by MOTIE, and the GIST-MIT Research Collaboration grant funded by GIST, Korea.

[2]we use object classes in RefCOCO GT annotation. This is to cover all salient objects in the dataset during the few-shot training.

References

[1] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In *CVPR*, 2022. 2

[2] Bo Chen, Zhiwei Hu, Zhilong Ji, Jinfeng Bai, and Wangmeng Zuo. Position-aware contrastive alignment for referring image segmentation. *arXiv preprint arXiv:2212.13419*, 2022. 3

[3] Shiming Chen, Ziming Hong, Yang Liu, Guo-Sen Xie, Baigui Sun, Hao Li, Qinmu Peng, Ke Lu, and Xinge You. Transzero: Attribute-guided transformer for zero-shot learning. In *AAAI*, 2022. 2

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 1

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2019. 2

[6] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, 2021. 3, 6

[7] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, 2022. 1, 2, 5, 7

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 4, 5

[9] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, 2022. 1, 2, 7

[10] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Expand your detector vocabulary with uncurated images. In *ECCV*, 2022. 1, 2

[11] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *CVPR*, 2021. 3

[12] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022. 1, 2

[13] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICML*, 2022. 1, 2, 7

[14] Zongyan Han, Zhenyong Fu, Shuo Chen, and Jian Yang. Contrastive embedding for generalized zero-shot learning. In *CVPR*, 2021. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 5

[16] Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *EMNLP*, 2015. 3, 5

[17] Hsuan-An Hsia, Che-Hsien Lin, Bo-Han Kung, Jhao-Ting Chen, Daniel Stanley Tan, Jun-Cheng Chen, and Kai-Lung Hua. Clipcam: A simple baseline for zero-shot text-guided object and action localization. In *ICASSP*, 2022. 2, 6

[18] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, 2016. 3

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1, 2

[20] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *CVPR*, 2021. 3

[21] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 5

[22] Kwanyoung Kim, Yujin Oh, and Jong Chul Ye. Zegot: Zero-shot segmentation through optimal transport of text prompts. *arXiv preprint arXiv:2301.12171*, 2023. 2

[23] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *CVPR*, 2022. 3, 6

[24] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022. 2

[25] Jiahao Li, Greg Shakhnarovich, and Raymond A Yeh. Adapting clip for phrase localization without further training. *arXiv preprint arXiv:2204.03647*, 2022. 2, 7

[26] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *CVPR*, 2018. 3

[27] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *ACL*, 2021. 1

[28] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 1

[29] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. *arXiv preprint arXiv:2210.04150*, 2022. 2

[30] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint arXiv:2211.14843*, 2022. 2

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[32] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *ICCV*, 2017. 3

[33] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. *arXiv preprint arXiv:2302.07387*, 2023. 3

[34] Lu Liu, Tianyi Zhou, Guodong Long, Jing Jiang, Xuanyi Dong, and Chengqi Zhang. Isometric propagation network for generalized zero-shot learning. In *ICLR*, 2020. 2

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 3

[36] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 1

[37] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2211.14813*, 2022. 2

[38] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 5

[39] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013. 2

[40] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 1, 2

[41] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 5

[42] Prashant Pandey, Mustafa Chasmai, Monish Natarajan, and Brejesh Lall. A language-guided benchmark for weakly supervised open vocabulary semantic segmentation. *arXiv preprint arXiv:2302.14163*, 2023. 2

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2

[44] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, 2022. 1, 2

[45] Hanoona Abdul Rasheed, Muhammad Maaz, Muhammd Uzair Khattak, Salman Khan, and Fahad Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *NeurIPS*, 2022. 2

[46] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017. 6

[47] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? In *ICLR*, 2021. 1

[48] Robin Strudel, Ivan Laptev, and Cordelia Schmid. Weakly-supervised segmentation of referring expressions. *arXiv preprint arXiv:2205.04725*, 2022. 1, 3, 6, 7

[49] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *ACL*, 2022. 2, 7

[50] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 1

[51] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2

[52] Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. *arXiv preprint arXiv:2301.11320*, 2023. 3

[53] Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and Jose M Alvarez. Freesolo: Learning to segment objects without annotations. In *CVPR*, 2022. 3, 4, 5, 6, 7

[54] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, 2022. 1, 3, 7, 8

[55] Zhichao Wei, Xiaohao Chen, Mingqiang Chen, and Siyu Zhu. Learning aligned cross-modal representations for referring image segmentation. *arXiv preprint arXiv:2301.06429*, 2023. 3

[56] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. Phrasecut: Language-based image segmentation in the wild. In *CVPR*, 2020. 7

[57] Jianzong Wu, Xiangtai Li, Xia Li, Henghui Ding, Yunhai Tong, and Dacheng Tao. Towards robust referring image segmentation. *arXiv preprint arXiv:2209.09554*, 2022. 3

[58] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2302.12242*, 2023. 2

[59] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *ECCV*, 2022. 1, 2, 5, 7

[60] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022. 3, 6, 7, 8

[61] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, 2019. 3

[62] Zicheng Zhang, Yi Zhu, Jianzhuang Liu, Xiaodan Liang, and Wei Ke. Coupalign: Coupling word-pixel with sentence-mask alignments for referring image segmentation. In *NeurIPS*, 2022. 3

[63] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022. 1, 2, 7

[64] Ziqin Zhou, Bowen Zhang, Yinjie Lei, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. *arXiv preprint arXiv:2212.03588*, 2022. 2