Curriculum Point Prompting for Weakly-Supervised Referring Image Segmentation

Qiyuan Dai Sibei Yang†
School of Information Science and Technology, ShanghaiTech University
{daiqy2022,yangsb}@shanghaitech.edu.cn

Abstract

*Referring image segmentation (RIS) aims to precisely segment referents in images through corresponding natural language expressions, yet relying on cost-intensive mask annotations. Weakly supervised RIS thus learns from image-text pairs to pixel-level semantics, which is challenging for segmenting fine-grained masks. A natural approach to enhancing segmentation precision is to empower weakly supervised RIS with the image segmentation foundation model SAM. Nevertheless, we observe that simply integrating SAM yields limited benefits and can even lead to performance regression due to the inevitable noise issues and challenges in excessive focus on object parts. In this paper, we present an innovative framework, **P**oint **P**rompt**T**ing (PPT), incorporated with the proposed multi-source curriculum learning strategy to address these challenges. Specifically, the core of PPT is a point generator that not only harnesses CLIP's text-image alignment capability and SAM's powerful mask generation ability but also generates negative point prompts to address the noisy and excessive focus issues inherently and effectively. In addition, we introduce a curriculum learning strategy with object-centric images to help PPT gradually learn from simpler yet precise semantic alignment to more complex RIS. Experiments demonstrate that our PPT significantly and consistently outperforms prior weakly supervised techniques on mIoU by 11.34%, 14.14%, and 6.97% across RefCOCO, RefCOCO+, and G-Ref, respectively.*

1. Introduction

Referring image segmentation (RIS) entails segmenting the referent referred to by a natural language expression in an image [12, 14, 16, 53, 55, 70]. This task delves into pixel-level semantic understanding that aligns with free-form texts, as opposed to pixel classification into closed-set categories in typical semantic segmentation [8, 9, 40, 62].

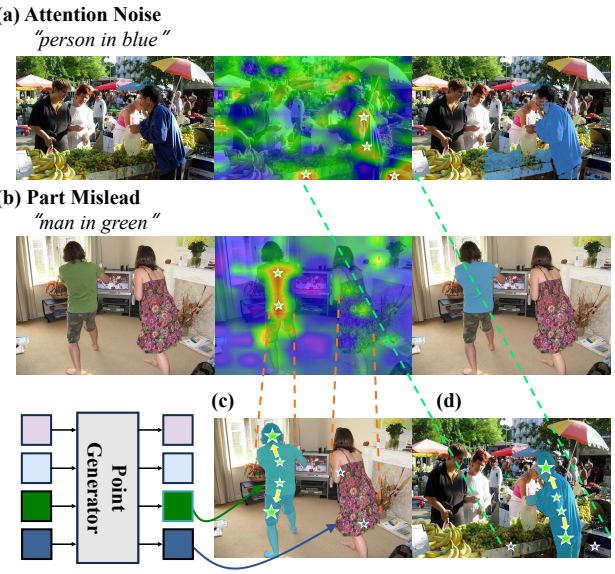

Figure 1. Illustrations of the CLIP attention map on RIS image-text pairs and corresponding mask outputs through SAM. (a) and (b) show background noise issues and excessive focus on object parts that mislead SAM, while (c) and (d) demonstrate the results of our method, which mitigates these issues.

In recent years, RIS has made significant progress, resulting in improved accuracy in the fully supervised learning setting [12, 14, 55, 70] and demonstrating great potential in real-world applications, including text-based human-robot interaction [74, 75] and image editing [3, 78]. However, annotating high-quality pairs of natural language expressions and their corresponding referent masks in images is challenging and time-consuming, limiting the development of RIS which relies on full annotation.

Therefore, our objective is to study RIS through weakly supervised learning in line with some very recent works [18, 27, 37, 53], specifically without the need for any instance-level or pixel-level semantic supervision, thereby mitigating the annotation limitation. In light of the absence of pixel-level semantic annotations, the primary focus of these weakly supervised RIS studies is to transfer the semantic

†Corresponding author

associations from image-text pairs to the pixel level. This is achieved by visual entity discovery and gathering [18], text-to-image response mapping [37], or enhancing Grad-CAM for an improved saliency map [27]. While it is possible to approximate the regions of referents, the segmentation results lack precision.

Lately, the Segment Anything Model (SAM) [22] has demonstrated its proficiency in generating valid image segmentation masks. A naive approach to improving weakly supervised RIS involves employing SAM to refine the coarse localization of referents into precise segmentation masks. Despite relying on annotated masks, SAM's training does not include pixel-level semantic supervision. This supervision aligns with our weakly supervised RIS setting and offers a promising "free lunch" solution as an image segmentation foundation model. However, unexpectedly, this refinement fails to enhance, or even drop, the performance of RIS. We observe that the primary reasons for this are: 1) The pixel-level semantic response obtained from image-text pairs inherently contains noise or activates other objects or attributes mentioned in the expression, yet SAM is not robust when using such noisy responses as input prompts directly. As shown in Figure 1(a), the attention response for the expression in the image indeed localizes the referent. Nevertheless, the presence of certain noisy attention, like that in Figure 1(a), results in SAM predicting a segmentation that encompasses all these responsive regions. 2) More crucially, attending salient rather than comprehensive responses in weakly supervised RIS worsens the problem of segmentation ambiguity in SAM, where input prompts can correspond to multiple valid masks. This is due to the combined effect of SAM's almost edge-oriented, semantic-unaware nature and the image-level semantic supervision inherent to weakly supervised RIS. For example, the green t-shirt in Figure 1(b) naturally elicits a more noticeable response for the expression "man in green". However, using the response as the mask, box, or points to prompt SAM leads to segmenting the t-shirt rather than the man.

In this paper, we propose a novel point prompt learning framework that collaborates with an innovative multi-source curriculum learning strategy to tackle these challenges in weakly supervised RIS. Our framework utilizes frozen CLIP as the text and image encoders, frozen SAM as the mask decoder, and a trainable, lightweight point generator to seamlessly bridge the encoders and the decoder. Specifically, the point generator initializes learnable point queries to represent segmentation masks, which interact with image and text features from the encoders, generating point prompts for SAM to achieve precise mask segmentation. This approach harnesses the inherent advantages of CLIP's pre-trained image-text alignment and SAM's segmentation capabilities, effectively simplifying the RIS problem to learning point queries and selecting the positive query for the referent. Notably, point queries naturally handle noisy responses by distinguishing them as negatives, as illustrated in Figure 1(d).

Moreover, to address the semantic-unaware ambiguity issue, we introduce training on object-centric images like ImageNet [11], encouraging point queries to shift their attention from salient to comprehensive responses. First, ImageNet boasts a large vocabulary with thousands of classes, and its rich semantics can assist point queries in learning semantic-aware point prediction. Additionally, ImageNet predominantly features object-centric images, where the objects are often centrally positioned, making it easier to extract comprehensive class responses using CLIP or unsupervised DINO models [5, 43, 45]. Training on these comprehensive responses can correct the initial point prompts to match those shown in Figure 1(c), resulting in a precise mask of the "person". To the best of our knowledge, we are the first to leverage object-centric images to improve scene-centric RIS.

Ultimately, in essence, RIS remains a scene-centric task. It entails segmenting referents in complex scenes, considering not only class semantics as seen in object-centric scenes but also their location and relationships with other objects. Therefore, we introduce a novel curriculum learning strategy that progressively transitions from class-based segmentation to complex referring image segmentation, incorporating factors such as location and relationships. Additionally, beyond the complexity of expressions, object-centric images in ImageNet and scene-centric images in RIS have distinct data distributions. We also strive to mitigate the domain gap when jointly training on these datasets by introducing a multi-source training strategy.

To evaluate the effectiveness of our PPT and learning strategy, we conduct expressions on common benchmarks in RIS, including RefCOCO [71], RefCOCO+ [71], and G-Ref [41]. In summary, our main contributions are:

- We propose a novel, parameter-efficient point prompting framework (PPT) for weakly-supervised RIS. PPT's core component is a trainable, lightweight point generator to identify the referent and naturally distinguish noise and misleading context as negative prompts, thereby enabling effective integration with frozen CLIP and SAM.
- To the best of our knowledge, we are the first to effectively utilize object-centric images to facilitate scene-centric RIS to learn precise and comprehensive dense semantic alignment between text and image.
- We propose an innovative multi-source curriculum learning strategy to facilitate the gradual learning of the point generator, starting from simpler semantic alignment and progressing to more complex RIS, which meanwhile mitigates the domain mismatch in multi-source training.
- Although straightforward, our PPT consistently outperforms state-of-the-art weakly-supervised RIS by a signif-

icant margin of 11.34%, 14.14%, and 6.97% in terms of mIoU on the three benchmarks, respectively.

2. Related Work

Weakly Supervised RIS: RIS aims to segment the precise mask of the referent through a comprehensive understanding of both the image and the text describing the referent, which has achieved significant performance improvements in the fully supervised setting by employing multi-modal fusion from concatenation operation [16, 30, 36, 51, 64, 67] to recent attention-based approaches [19, 33, 55, 58, 65, 66, 68–70]. However, achieving full supervision demands pixel-level semantic annotations, which entail significant expenses. In response, recent approaches [18, 27, 37] begin exploring weakly supervised RIS, leveraging weak supervisory signals like bounding boxes or image-text pairs, aiming to align pixel-level semantics utilizing coarse-grained supervision. Among them, Kim *et al.* [18] divides image features into several entities and then employs top-down multi-modal attention to select entities combined into the referring segment. Instead of aggregating visual entities, TRIS [37] extracts rough object positions through text supervision as pseudo-labels to train a segmentation network. In contrast, Lee *et al.* [27] focuses on reasoning about word relationships to predict from salient maps of each word. Nevertheless, the masks from these methods are often coarse due to noisy pseudo-labels or lacking dedicated fine segmentation decoding. To address this, we introduce a novel framework, PPT, which uses point prompting to handle noise issues, thus enabling the effective utilization of SAM's segmentation capabilities for high-quality masks.

Curriculum Learning for RIS: Curriculum learning proposed by [1], similar to the natural learning process, gradually increases the complexity of training data during the training to enhance the model's capability. Its effectiveness has been validated in computer vision [6, 52, 72, 76] and has recently been introduced into the vision-language field. For visual reasoning, CLIP-VG [59] is the first to introduce curriculum learning into visual grounding. It iteratively selects high-quality data for training from a pseudo-label set and uses updated weights for the next round of data selection, achieving a progressive effect. For the RIS task, MCRES [61] combines words from expressions into different novel compositions, enabling the model to learn from word-word, word-phrase, and ultimately phrase-phrase relationships. In contrast to previous fully supervised works, we introduce curriculum learning to mitigate the substantial noise present in weakly supervised learning, extending the model's ability from simpler semantic alignment to more complex RIS.

Visual Foundation Models possess substantial knowledge capacity, making them versatile for a wide range of tasks. For example, the CLIP model [45] demonstrates significant performance improvements in open-vocabulary tasks [13, 15, 25, 49, 50, 73]. In image segmentation, the SAM [21] also opens up promising avenues[10, 35]. In fully supervised RIS, while [26, 58, 63] generally treat foundation models merely as weight initialization or auxiliary tools for pseudo-label extraction, under-utilizing their intrinsic powerful representation capabilities. Among these, CRIS [58] builds upon CLIP by adding a vision-language decoder to extend into a segmentation model. Similarly, LISA [26] employs SAM's image encoder and LLaVA [38] to train an additional mask decoder. Recent, other works like Grounding DINO [39] and Grounded SAM [21, 39] require extensive object detection data during training to provide strong prior knowledge for RIS. However, they contradict the weakly supervised setting. In contrast to prior work, our PPT fully harnesses the capabilities of foundation models in both the encoder and decoder stages and concatenates them through a learnable point generator, simultaneously possessing robust image-text understanding and precise mask generation.

3. Method

The framework of our proposed PPT and its corresponding multi-source curriculum learning strategy are shown in Figure 2. First, we introduce PPT, our point prompt learning framework for weakly supervised RIS. Its central focus is on utilizing a trainable, lightweight point generator to transfer the text-image semantic alignment capability from frozen CLIP to SAM, enabling robust and precise mask decoding for referents (see Section 3.1). Next, we propose learning from object-centric images to support the point generator in generating semantic-aware and comprehensive point prompts, as opposed to merely salient ones (see Section 3.2). Finally, we apply curriculum learning to enable the progressive learning of the point generator, moving from simpler class-based segmentation to more complex referring image segmentation, while also mitigating the domain mismatch in multi-source training (see Section 3.3).

3.1. Point Prompt Learning Framework

3.1.1 Point Prompting Architecture

The overall PPT architecture is simple and illustrated in Figure 2. It comprises three primary components: the image and text encoders to extract features, a transformer-based point generator to predict a set of point prompts and their corresponding confidence scores, and a mask decoder that predicts the segmentation from point prompts.

Image Encoder and Text Encoder. Follow previous works [37, 54, 58], we employ CLIP as both the image and text encoders. In contrast to [37] that fully fine-tunes the encoders to adapt weakly supervised RIS, we opt to freeze the encoders. This preserves the pre-trained image-text alignment of CLIP and ensures parameter-efficient tuning for the entire framework. Specifically, given a pair of an image I and

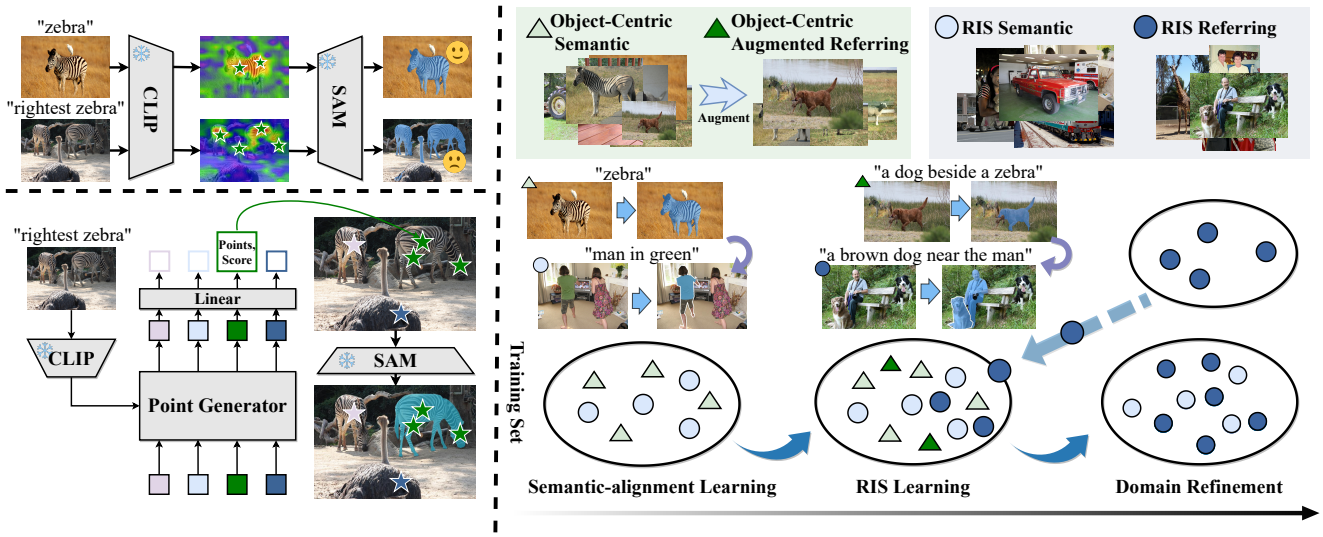

Figure 2. **The overall framework of our PPT and its curriculum learning strategy**. The top left corner demonstrates that a straightforward concatenation of CLIP and SAM exhibits segmentation capability for simple object-centric images but performs poorly on referring expression data. The bottom left corner showcases our point generator, which determines the approximate location of objects by learning a well-distributed set of points, combined with our curriculum learning strategy depicted on the right side of the figure, which transitions from simple object-centric semantics towards the more complex referring domain.

a natural language expression G, we use the encoders to extract visual features $\{V_n\}_{n=1}^{N}$ at N stages and obtain the contextual feature representation at word level as T.

Point Generator. Inspired by the DETR-based object detection methods [4, 28, 77, 79], which represents objects using a set of object queries, we use a set of point queries to represent segmentation masks. These point queries are initially randomly initialized and then interact with image and text features to predict corresponding point prompts, which are subsequently used as input for the mask decoder. First, the interactions are performed across multiple encoding stages because image features at different stages may emphasize different levels of semantic information essential for RIS, ranging from simple shapes to complex semantics. At each stage, such as the n-th stage, the visual features V_n and text features T are initially fused to update their representations to V_n' and T_n following [63]. Then, the set of point queries Q_n alternately query the updated image features V_n' and text features T_n by a classical cross-attention layer [57] as follows,

$$Q_{n+1} = \mathrm{CrossAttn}(\mathrm{CrossAttn}(Q_n, V_n'), T_n). \quad (1)$$

Here, $\mathrm{CrossAttn}$ denotes the cross-attention layer, and the point queries at the first stage are learnable embeddings that are initialized randomly.

Next, for each point query $q_k \in Q_{N+1}$ at the final stage, we regress its point prompt $P_k = \{(p_k^m, r_k^m)\}_{m=1}^{M}$, consisting of M points, and predict its referent confidence score c_k as follows,

$$p_k^m, r_k^m = \mathrm{PointHead}(q_k), c_k = \mathrm{ScoreHead}(q_k), \quad (2)$$

where $\mathrm{PointHead}$ and $\mathrm{ScoreHead}$ are two independent multilayer perceptions for predicting the points and confident scores, respectively. Here, p_k^m denotes the normalized coordinates of the m-th point of the k-th point query, while r_k^m signifies the score for classifying as a positive point.

Mask Decoder. We utilize SAM as the mask decoder and freeze it, similar to our approach with CLIP, to retain its segmentation capability. The mask decoder can be prompted with a set of points, which can be positive or negative, for SAM to infer the area to be segmented and output the corresponding segmentation mask. Specifically, for the k-th point query, we employ the point prompt P_k to prompt the mask decoder, leading to the generation of the segmentation mask s_k within image I.

To summarize, given an input image I and an expression G, our PPT framework outputs a fixed-size of K predictions $\{y_k\}_{k=1}^{K}$, where the k-th query's prediction $y_k = (s_k, P_k, c_k)$ includes the segmentation mask s_k, point prompt $P_k = \{(p_k^m, r_k^m)\}_{m=1}^{M}$, and confidence score c_k, which is formulated as follows,

$$\{y_k\}_{k=1}^{K} = f_\theta(I, G), \text{ where } y_k = (s_k, P_k, c_k) \quad (3)$$

Here, we denote our framework as f_θ, where the θ represents learnable parameters in the point generator.

3.1.2 Objective Function

Inspired by DETR in object detection, we apply the Hungarian algorithm [24] to search for the best bipartite matching between the pseudo labels $\{\hat{y_e}\}_{e=1}^{E}$ and the predictions

$\{y_k\}_{k=1}^{K}$ to determine the best assignments $\sigma(e)$, where $\sigma(e)$ represents the index of the prediction matched with the pseudo label $\hat{y}_e$. The extraction for pseudo labels from image-text pairs will be discussed in Section 3.2 and 3.3. Specifically, for each pair of the pseudo label $\hat{y}_e = (\hat{s}_e, \hat{P}_e, \hat{c}_e)$ and the prediction $y_{\sigma(e)} = (s_{\sigma(e)}, P_{\sigma(e)}, c_{\sigma(e)})$, we define the objective function by considering the segmentation loss related to the point prompt and the mask prediction, as well as the loss for confidence score as the referent. The loss $\ell(y_{\sigma(e)}; \hat{y}_e)$ for the prediction $y_{\sigma(e)}$ given the pseudo label $\hat{y}_e$ is formulated as follows,

$$\begin{aligned} \ell_{\text{seg}}(y_{\sigma(e)}; \hat{y}_e) &= \ell_{\text{mask}}(s_{\sigma(e)}; \hat{s}_e) + \ell_{\text{pt}}(P_{\sigma(e)}; \hat{P}_e), \\ \ell(y_{\sigma(e)}; \hat{y}_e) &= \ell_{\text{seg}}(y_{\sigma(e)}; \hat{y}_e) + \ell_{\text{bce}}(c_{\sigma(e)}; \hat{c}_e), \end{aligned} \tag{4}$$

where ℓ_{mask} represents the DICE loss [31] and binary cross-entropy loss. The point prompt loss ℓ_{pt} is defined as the combination of $L1$ loss for point regression and the binary cross-entropy loss for point classification. Additionally, ℓ_{bce} denotes the binary cross-entropy loss for classifying the prediction as the referent or not.

During inference, we select the segmentation mask $s_{\text{argmax}\{c_e\}}$ corresponding to the highest confidence score as the segmentation prediction for the referent.

3.2. Learning from Object-centric Images

To train our PPT model using image-text pairs, we generate pseudo-labels for referents from these pairs leveraging the pre-trained image-text alignment of CLIP [45]. However, directly extracting pseudo-labels from image-expression pairs in RIS datasets [41, 71] for training does not yield satisfactory results. The main reasons for this are as follows: 1) The pixel-level semantic responses obtained through image-text alignment are inevitably noisy or incomplete (particularly in salient regions), as shown in Figure 1(a)-(b). 2) The complexity of referring expressions adds to the challenge of extracting their responses as effective pseudo-labels, primarily because distinguishing the response corresponding to the referent from other responses becomes arduous. Referring expressions describe rich content, including class, attributes, locations, and relationships. The semantic response to such expressions encompasses information about the referent but also incorporates additional contextual details, such as the "a brown dog near the man" in Figure 2.

We propose to address these challenges by training on the ImageNet [11], as its object-centric characteristics make it particularly suitable for extracting high-quality pseudo-labels. The top left of Figure 2 shows that the pseudo segmentation mask for the object-centric image is precise but noisy for the RIS image. In detail, we construct pseudo datasets for weakly supervised RIS at both the class-semantic level and the more complex relationship level.

Free Data for Simpler Semantic-alignment Learning. We create a simpler semantic alignment dataset, denoted as D_{sem}, from ImageNet for weakly supervised RIS. The dataset is designed to mitigate issues related to noise and partial segmentation by including comprehensive pseudo-masks corresponding to object classes.

Given an image and its class in ImageNet, we first create an image-expression pair (I, G) by randomly applying transformations, such as flipping, to the image and defining the expression as "the [class name] in the middle". Then, we automatically obtain the corresponding pseudo-label $\{\hat{y}_e\}_{e=1}^{E}$, where $\hat{y}_e = (\hat{s}_e, \hat{P}_e, \hat{c}_e)$, and $\hat{s}_e$ and $\hat{P}_e$ denote the pseudo segmentation mask and point prompt, respectively. Note that for semantic-level data, we only obtain the referent's pseudo label, which means $E = 1$ for the dataset D_{sem}, and score $\hat{c}_e = 1$. For the segmentation mask $\hat{s}_e$, we employ the CLIP[32] followed by SAM to directly obtain a high-quality precision mask, benefiting from the object-centric characteristics of the image. For the point prompt $\hat{P}_e$, we randomly and uniformly sample M points within the bounding box of the mask $\hat{s}_e$. Points inside the mask are considered positive, while those outside are considered negative. Building upon this sampling approach, we encourage the model to produce points that are evenly distributed within referents and emphasize the negative points closely outside the boundary of referents, significantly reducing the ambiguity in SAM's point prompting.

Augmented Data for More Complex RIS Learning. Training exclusively on semantic-level data could result in a limited understanding of the more complex relationships needed for RIS. As ImageNet itself does not inherently contain any information about relationships between instances, we propose augmenting the semantic-level dataset D_{sem} to create a referring-level dataset D_{ref}. In dataset D_{ref}, we focus solely on spatial relationships between instances, with the generalization to semantic relationships discussed in Section 3.3. Specifically, we mainly employ two types of relation augmentation as follows: 1) In composition-based augmentation, where multiple images are aggregated into one by scaling and pasing through Mosaic-like [2] augmentation, resulting in crowded images containing multiple instances, along with text that describes absolute positions for the target referent. 2) In fusion-based augmentation, we embed one image within another and create a text for the target instance that encodes their relative relationship.

Through relation augmentation, our dataset D_{ref} not only enriches the instance-level semantic data by introducing more complex image-expression pairs, but it also provides pseudo-labels for context objects other than the referent. These contextual pseudo-labels are crucial for learning to distinguish the referent from other objects mentioned in the expressions. For a sample $(I, G, \{\hat{y}_e\}_{e=1}^{E}) \in D_{\text{ref}}$, its pseudo-labels $\{\hat{y}_e\}_{e=1}^{E}$ are related with E contextual ob-

jects, with each $\hat{y}_e = (\hat{s}_e, \hat{P}_e, \hat{c}_e)$. In these pseudo-labels, the confidence score is set to one for the referent and zero for other context objects.

3.3. Multi-source Curriculum Learning Strategy

In this section, we present a multi-source curriculum learning strategy for jointly training our PPT model f_θ using both multi-source (ImageNet-based and RIS-based) and multi-level (semantic-level and referring-level) datasets. We first introduce the pseudo-label extraction in the RIS dataset. Then, we present the progressively learning stages from the semantic level to the referring level to domain refinement. At each stage, we utilize different sets of data samples from ImageNet-based datasets D_{sem} and D_{ref}, as well as the RIS dataset H.

Pseudo Label Generation. As discussed in Section 3.2, due to the complexity of referring expressions, extracting pseudo-label for the referent directly from the entire expression results in poor label quality. Therefore, for an image-expression pair (I, G), we extract its pseudo-label $\{\hat{y}_e\}_{e=1}^{E}$ by collecting all E candidate referents. This is achieved by extracting candidate pseudo-label separately for each noun phrase in the expression using the same method employed in extracting pseudo-label in D_{sem}. Subsequently, based on the count of candidate referents in one pair, we partition the RIS dataset H into two subsets, denoted as H_{sem} and H_{ref}. In dataset H_{sem}, a sample $(I, G, \{\hat{y}_e\}_{e=1}^{E}) \in H_{\text{sem}}$, with each $\hat{y}_e = (\hat{s}_e, \hat{P}_e, \hat{c}_e)$, contains only one candidate referent, which implies $E = 1$ for the H_{sem} dataset and the score $\hat{c}_1 = 1$ for this referent. For each sample $(I, G, \{\hat{y}_e\}_{e=1}^{E}) \in H_{\text{ref}}$, it contains multiple candidate referents, *i.e.*, $E > 1$. At this stage of pseudo-label extraction, we set the confidence scores $\hat{c}_e = 0$ for all candidate referents because the referent cannot be distinguished from the other candidate referents. Notably, despite the pseudo-labels being generated in datasets H_{sem} and H_{ref}, they are highly noisy and necessitate collaboration with ImageNet-based datasets D_{sem} and D_{ref} for model training.

Semantic-alignment Learning Stage. At this learning stage, our goal is to jointly train the model using multi-source semantic-level datasets, including H_{sem} and D_{sem}, to predict semantic-aware and comprehensive segmentation masks for references. The inclusion of the dataset D_{sem} helps mitigate issues related to noise and partial segmentation pseudo-labels in H_{sem}. The learning objective is formulated as follows,

$$\theta_{\text{sem}} = \text{argmin}_\theta \, \mathbb{E}_{(I,G,\{\hat{y}_e\})\sim H_{\text{sem}}\cup D_{\text{sem}}}[\ell(f_\theta(I,G);\{\hat{y}_e\})], \tag{5}$$

where the loss $\ell(f_\theta(I,G);\{\hat{y}_e\})$ between the model prediction $f_\theta(I,G)$ and the pseudo-label $\{\hat{y}_e\}_{e=1}^{E}$ is defined in Equation 4. And we denote the learned parameters after optimization at this stage as θ_{sem}.

RIS Learning Stage. Furthermore, we enable progressive learning, transitioning from simpler semantic-level segmentation to more complex referring image segmentation with relation modeling. This is achieved through two consecutive learning steps: 1) Training exclusively on the ImageNet-based referring-level dataset D_{ref}, with training samples $(I, G, \{\hat{y}_e\})$ sampled from $H_{\text{sem}} \cup D_{\text{sem}} \cup D_{\text{ref}}$, leads to the learned parameters as θ_{refD}. 2) Updating the pseudo-labels in the referring-level RIS dataset H_{ref} and conducting joint training on all four datasets. Specifically, we select pseudo-referents from the candidate referents in H_{ref} to update it to H'_{ref} using the model f_θ with model parameter θ_{refD}. Notably, even though the augmented ImageNet-based dataset D_{ref} only considers spatial relationships, we observe that the model learned on it can generalize to predict referents in H_{ref} with semantic relationships, thanks to the generalization ability of frozen CLIP encoders. The second training step is formulated as follows,

$$\begin{aligned} &H'_{\text{ref}} \leftarrow \text{Select}(H_{\text{ref}}; f_\theta(\cdot;\theta_{\text{refD}})), \\ &\theta_{\text{ref}} = \text{argmin}_\theta \, \mathbb{E}_{(I,G,\{\hat{y}_e\})\sim D_{\text{all}}}[\ell(f_\theta(I,G;\theta_{\text{refD}});\{\hat{y}_e\})], \end{aligned} \tag{6}$$

where Select represents the selection of pseudo-referents from the candidate referents in H_{ref} according to the model f_θ with model parameter θ_{refD}, the D_{all} is the union of the datasets H_{sem}, D_{sem}, D_{ref}, and H'_{ref}. And the θ_{refD} in $f_\theta(I,G;\theta_{\text{refD}})$ represent the model's parameters are initialized with θ_{refD}. The parameter after optimization at this stage is denoted as θ_{ref}.

Domain Refinement Stage. Despite data augmentation, disparities in data distribution and domain persist between the ImageNet-based and RIS-based datasets. Furthermore, the model $f_{\theta_{\text{ref}}}$ has already acquired point generation capabilities at both the instance and relationship levels through the use of ImageNet-based data in previous stages. Hence, we exclusively fine-tune on the RIS datasets at this stage by continuously adjusting the selected pseudo-referents. The optimization objective is formulated as follows,

$$\begin{aligned} &H^{(i)}_{\text{ref}} \leftarrow \text{Select}(H^{(i-1)}_{\text{ref}}; f_\theta(\cdot;\theta^{(i-1)}),\ H^{(i)} = H^{(i)}_{\text{ref}} \cup H_{\text{sem}}, \\ &\theta^{(i)} = \text{argmin}_\theta \, \mathbb{E}_{(I,G,\{\hat{y}_e\})\sim H^{(i)}}[\ell(f_\theta(I,G;\theta^{(i-1)});\{\hat{y}_e\})]. \end{aligned} \tag{7}$$

Here, $H^{(i)}$ and $\theta^{(i)}$ represent the dataset and learned model parameters after the i-th round of adjustments. At the first round, the referring-level dataset and parameters are H'_{ref} and θ_{ref}, obtained from the RIS learning stage.

4. Experiments

4.1. Datasets and Implementation Details

Datasets. We conduct experiments on three major datasets: RefCOCO[71], RefCOCO+[71], and G-Ref [41]. All the images used in these datasets are from subsets of MSCOCO

Method	Published on	Sup.	Extra Image-Text Pairs	RefCOCO			RefCOCO+			G-Ref
				val	test A	test B	val	test A	test B	val-G
RMI [36]	ICCV '17	F	-	44.33	44.74	44.63	29.91	30.37	29.43	33.11
DMN[42]	ECCV '18	F	-	49.78	54.83	45.13	38.88	44.22	32.29	34.52
Hu *et al.* [17]	CVPR '20	F	-	60.98	62.99	59.21	48.17	52.32	42.11	47.57
GroupViT [60]	CVPR '22	W	CC12M [7],YFCC [56]	12.97	14.98	12.02	13.21	15.08	12.41	16.84
TSEG [53]	arXiv '22	W	ImageNet-21K [11]	25.44	-	-	22.01	-	-	22.05
ALBEF [29]	NeurIPS '21	W	CC [48],SBU [44],COCO [34],VG [23]	23.11	22.79	23.42	22.44	22.07	22.51	24.18
Chunk [27]	ICCV '23	W	CC,SBU,COCO,VG	31.06	32.30	30.11	31.28	32.11	30.13	32.88
Shatter [18]	ICCV '23	W	ImageNet-21K	34.76	34.58	35.01	28.48	28.60	27.98	28.87
TRIS [37]	ICCV '23	W	WebImage Text [45]	31.17	32.43	29.56	30.90	30.42	30.80	36.00
TRIS + SAM	-	-	WebImage Text	25.56	26.56	25.71	26.22	25.75	26.62	29.84
Our PPT	CVPR'24	W	WebImage Text, ImageNet-1K [47]	**46.76**	**45.33**	**46.28**	**45.34**	**45.84**	**44.77**	**42.97**

Table 1. Comparison with state-of-the-art models in weakly supervised RIS on RefCOCO, RefCOCO+ and G-Ref datasets.

Method	Params	P@0.3	P@0.5	P@0.7	mIoU
TRIS	142.09M	46.57	18.69	4.9	31.17
Shatter	-	55.02	24.99	6.35	34.76
Chunk	-	46.12	23.88	9.02	31.06
Ours	20.7M	61.16	50.19	36.22	46.76

Table 2. Comparison of precision metrics. Params denote the number of trainable parameters.

[34], and they respectively contain 142,209, 141,564, and 104,560 sentences. RefCOCO+ does not include absolute directional words, while G-Ref contains longer sentences. Following previous works [55, 70], we divide RefCOCO and RefCOCO+ into the training set, validation set, testA, and testB. As for G-Ref, we use the validation split by Google.

Implementation Details. We employ the CLIP VIT/B-16 as the encoder following [22, 27] and use the smallest ViT-B SAM for mask decoder. We adopt the Adam[20] optimizer with a batch size of 64 and a learning rate of 0.001. The training is conducted for 50 epochs, of which 10, 30, and 10 epochs are for the semantic alignment, RIS learning and domain refinement stages, respectively. We randomly select 2,836 images from the Mini Imagenet [46] as our auxiliary data. We follow previous approaches to use the mIoU (Mean Intersection over Union) and precision at the 0.3, 0.5, and 0.7 thresholds of mIoU as our main evaluation metrics.

4.2. Comparison with the State-of-the-Art Methods

Table 1 compares our PPT with other state-of-the-art approaches across all dataset partitions. Our PPT consistently outperforms all other weakly supervised RIS methods, achieving an average mIoU improvement of 11.34%, 14.14%, and 6.97% on RefCOCO, RefCOCO+, and G-Ref, respectively, over the previous highest performance.

Compared to TRIS [37], which shares a strategy of first extracting pseudo-labels and then training with us, we achieve significant improvements in mIoU on the three datasets, with gains of 15.59%, 14.44%, and 6.97%, respectively. This indicates that our point prompting framework, which employs a progressive curriculum learning strategy from multi-source data, is more effective in discovering and localizing targets. Furthermore, for a fair comparison, we also integrate SAM into open-sourced TRIS. However, such integration leads to a decline in performance due to the inevitable noise issues and challenges in excessive focus on object parts in weakly supervised RIS, which will be detailed in the Appendix. In contrast, our approach outperforms these methods by 20.18%, 19.12%, 16.31%, respectively, reflecting that our point generator effectively eliminates noise by utilizing negative prompts. Compared to Shatter [18], which requires finding various parts of the instance and aggregating them to prediction, similar to our point-prompting design, we outperform it by 11.34%, 16.96%, and 17.1%, respectively. This demonstrates that by leveraging object-centric images from ImageNet, our method can predict precise semantic alignment regions instead of partial ones.

As shown in Table 2, our PPT significantly improves precision at different IoU thresholds compared to other weakly supervised methods. Especially at prec@0.5 and prec@0.7, we improve state-of-the-art methods by 25.2% and 27.2%, respectively, which reflects the accuracy of our point generator in locating referents and the robustness to scattering points. This is achieved by the curriculum learning strategy and augmented object-centric data, contributing to improving fine-grained mask segmentation.

4.3. Ablation Study

We conduct ablation experiments to analyze the effectiveness of our proposed method, as shown in Table 3.

Point Prompting Framework. (1) We evaluate the performance of directly using CLIP + SAM for RIS in a baseline experiment (row 1). We find that it fails to locate objects in a significant number of images, and even when CLIP identifies the correct region, interference caused by the resolution issue of attention map and background noise suppress SAM mask generation performance. (2) Building this foundation, we incorporate the point prompting framework (row 2), achieving a 7.73% mIoU boost. Our point genera-

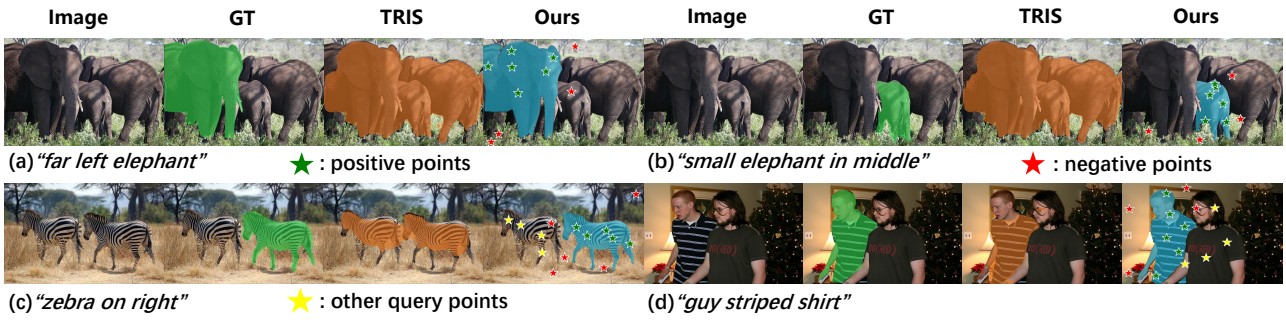

Figure 3. Visualization of our PPT's prediction results and comparison to TRIS [37].

	Method	P@0.3	P@0.5	P@0.7	mIoU
1	CLIP + SAM	28.72	24.84	11.67	26.46
2	Pt Prompt w/o IMG	43.85	25.45	18.35	34.19
3	Semantic Learning	51.69	39.63	28.44	38.91
4	3+RIS Learning	59.42	48.53	35.62	44.14
5	4+Domain Refine (Full)	61.16	50.20	36.22	46.76
6	5 w/o CL	57.93	44.64	34.27	43.28
7	5 w/o Augmented IMG	56.56	45.19	33.03	42.29
8	5 w/o Free IMG	47.29	38.51	27.94	39.33
9	5 w/o IMG	43.78	27.11	18.77	34.47

Table 3. Ablation study on the validation set of RefCOCO. IMG represents Object-centric Images from the ImageNet dataset.

tor directly utilizes high-dimensional features to regress object positions, thereby circumventing the noise introduced by the low resolution of activation maps. Moreover, it includes negative point prompts, which effectively suppress background interference.

Curriculum Learning. (3) We perform first-stage training for our prompting framework and gain 4.72% improvement (row 3), showing the effectiveness of semantic-alignment learning. (4) Next, we continue by augmenting training with RIS data (row 4), which aims at enhancing the learning of relation concepts, achieving a 5.23% mIoU gain. The augmented RIS-stage learning helps acquire rudimentary, more complex relation understanding capability. And the curriculum learning strategy filters out noisy data to enhance the training samples' signal-to-noise ratio. (5) Finally, we introduce the domain refinement stage into our full model, resulting in a further 2.62% mIoU improvement (row 5), which is achieved by eliminating the domain mismatch during the final fine-tuning phase. (6) Row 6 illustrates that without curriculum learning, the mIoU drops by 3.48%, demonstrating the effectiveness of curriculum learning to select the high-quality pseudo-labels progressively.

Learning from Object-centric Images. (7) To assess the impact of object-centric data, we remove augmented data only (row 7), resulting in a 4.47% mIoU drop, underscoring the assistance our constructed dataset provides in the relation understanding for referring scenarios. (8) In row 8, we remove free semantic-alignment data but still keep the augmented data, which resulted in a 7.46% mIoU drop. This is due to the detrimental impact of incorrect pseudo-label cases on training, which our additional data can mitigate. In addition, it demonstrates that a more straightforward semantic alignment is crucial for follow-up referring segmentation. (9) Row 9 demonstrates the results obtained when training without any object-centric images, compared to row 2 and row 5, highlighting that our method achieves maximum gains when both augmented data and curriculum learning are utilized, indicating that curriculum learning can effectively leverage the knowledge from the object-centric images to scene-centric RIS.

4.4. Visualization

Figure 3 visualizes our segmentation results. Expression in (a) requires the segmentation of the leftmost elephant among the three, demonstrating our framework's ability to distinguish instances. In (b), the expression calls for the segmentation of the middle elephant, showcasing our positional understanding capability. In (c), we visualize the points output by other queries as well, and observe that they also locate the regions of interest within their respective instances. In (d), we show that the point query retrieves object position based on textual semantics, rather than relying solely on spatial pronouns.

5. Conclusion

We introduce a novel point prompting framework (PPT) for weakly supervised referring image segmentation. It leverages a point generator to connect frozen CLIP and SAM and incorporates the concept of curriculum learning to this field. The framework utilizes object-centric images to aid in learning dense semantic alignment and relationships between text and images in a weakly supervised setting, which helps naturally mitigate noise issues and excessive focus on object parts.

Acknowledgment: This work was supported by the National Natural Science Foundation of China (No.62206174) and MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech).

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 3

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 5

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 4

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[6] Thibault Castells, Philippe Weinzaepfel, and Jerome Revaud. Superloss: A generic loss for robust curriculum learning. *Advances in Neural Information Processing Systems*, 33:4308–4319, 2020. 3

[7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021. 7

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1

[10] Haixing Dai, Chong Ma, Zhengliang Liu, Yiwei Li, Peng Shu, Xiaozheng Wei, Lin Zhao, Zihao Wu, Dajiang Zhu, Wei Liu, et al. Samaug: Point prompt augmentation for segment anything model. *arXiv preprint arXiv:2307.01187*, 2023. 3

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2, 5, 7

[12] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021. 1

[13] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14093, 2022. 3

[14] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15506–15515, 2021. 1

[15] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 3

[16] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 108–124. Springer, 2016. 1, 3

[17] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4424–4433, 2020. 7

[18] Dongwon Kim, Namyup Kim, Cuiling Lan, and Suha Kwak. Shatter and gather: Learning referring image segmentation with text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15547–15557, 2023. 1, 2, 3, 7

[19] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18145–18154, 2022. 3

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 7

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 7

[24] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 4

[25] 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. 3

[26] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation

via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 3

[27] Jungbeom Lee, Sungjin Lee, Jinseok Nam, Seunghak Yu, Jaeyoung Do, and Tara Taghavi. Weakly supervised referring image segmentation with intra-chunk and inter-chunk consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21870–21881, 2023. 1, 2, 3, 7

[28] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022. 4

[29] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 7

[30] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2018. 3

[31] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*, 2019. 5

[32] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 5

[33] Liang Lin, Pengxiang Yan, Xiaoqian Xu, Sibei Yang, Kun Zeng, and Guanbin Li. Structured attention network for referring image segmentation. *IEEE Transactions on Multimedia*, 24:1922–1932, 2021. 3

[34] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 7

[35] Zheng Lin, Zhao Zhang, Zi-Yue Zhu, Deng-Ping Fan, and Xia-Lei Liu. Sequential interactive image segmentation. *Computational Visual Media*, 9(4):753–765, 2023. 3

[36] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1271–1280, 2017. 3, 7

[37] Fang Liu, Yuhao Liu, Yuqiu Kong, Ke Xu, Lihe Zhang, Baocai Yin, Gerhard Hancke, and Rynson Lau. Referring image segmentation using text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22124–22134, 2023. 1, 2, 3, 7, 8

[38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 3

[39] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[40] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[41] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 2, 5, 6

[42] Edgar Margffoy-Tuay, Juan C Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 630–645, 2018. 7

[43] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[44] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 7

[45] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3, 5, 7

[46] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2016. 7

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 7

[48] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 7

[49] Cheng Shi and Sibei Yang. Edadet: Open-vocabulary object detection using early dense alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15724–15734, 2023. 3

[50] Cheng Shi and Sibei Yang. The devil is in the object boundary: Towards annotation-free instance segmentation using foundation models. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[51] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 38–54, 2018. 3

[52] Yang Shu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Transferable curriculum for weakly-supervised domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4951–4958, 2019. 3

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

[54] Sanjay Subramanian, William Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. *arXiv preprint arXiv:2204.05991*, 2022. 3

[55] Jiajin Tang, Ge Zheng, Cheng Shi, and Sibei Yang. Contrastive grouping with transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23570–23580, 2023. 1, 3, 7

[56] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 7

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[58] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 3

[59] Linhui Xiao, Xiaoshan Yang, Fang Peng, Ming Yan, Yaowei Wang, and Changsheng Xu. Clip-vg: Self-paced curriculum adapting of clip for visual grounding. *IEEE Transactions on Multimedia*, 2023. 3

[60] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 7

[61] Li Xu, Mark He Huang, Xindi Shang, Zehuan Yuan, Ying Sun, and Jun Liu. Meta compositional referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19478–19487, 2023. 3

[62] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 1

[63] Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17503–17512, 2023. 3, 4

[64] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4145–4154, 2019. 3

[65] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4644–4653, 2019. 3

[66] Sibei Yang, Guanbin Li, and Yizhou Yu. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9952–9961, 2020. 3

[67] Sibei Yang, Guanbin Li, and Yizhou Yu. Propagating over phrase relations for one-stage visual grounding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*, pages 589–605. Springer, 2020. 3

[68] Sibei Yang, Guanbin Li, and Yizhou Yu. Relationship-embedded representation learning for grounding referring expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2765–2779, 2020. 3

[69] Sibei Yang, Meng Xia, Guanbin Li, Hong-Yu Zhou, and Yizhou Yu. Bottom-up shift and reasoning for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11266–11275, 2021.

[70] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022. 1, 3, 7

[71] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 2, 5, 6

[72] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Multi-task curriculum framework for open-set semi-supervised learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 438–454. Springer, 2020. 3

[73] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *European Conference on Computer Vision*, pages 106–122. Springer, 2022. 3

[74] B Zhang and H Soh. Large language models as zero-shot human models for human-robot interaction. arxiv 2023. *arXiv preprint arXiv:2303.03548*. 1

[75] Bowen Zhang and Harold Soh. Large language models as zero-shot human models for human-robot interaction. *arXiv preprint arXiv:2303.03548*, 2023. 1

[76] Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021. 3

[77] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 4

[78] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 1

[79] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 4