Boosting Weakly-Supervised Referring Image Segmentation via Progressive Comprehension

Zaiquan Yang, Yuhao Liu[†], Jiaying Lin, Gerhard Hancke[†], Rynson W.H. Lau[†]
Department of Computer Science
City University of Hong Kong
{zaiquyang2-c, yuhliu9-c,jiayinlin5-c}@my.cityu.edu.hk
{gp.hancke, Rynson.Lau}@cityu.edu.hk

Abstract

This paper explores the weakly-supervised referring image segmentation (WRIS) problem, and focuses on a challenging setup where target localization is learned directly from image-text pairs. We note that the input text description typically already contains detailed information on how to localize the target object, and we also observe that humans often follow a step-by-step comprehension process (*i.e.*, progressively utilizing target-related attributes and relations as cues) to identify the target object. Hence, we propose a novel Progressive Comprehension Network (PCNet) to leverage target-related textual cues from the input description for progressively localizing the target object. Specifically, we first use a Large Language Model (LLM) to decompose the input text description into short phrases. These short phrases are taken as target-related cues and fed into a Conditional Referring Module (CRM) in multiple stages, to allow updating the referring text embedding and enhance the response map for target localization in a multi-stage manner. Based on the CRM, we then propose a Region-aware Shrinking (RaS) loss to constrain the visual localization to be conducted progressively in a coarse-to-fine manner across different stages. Finally, we introduce an Instance-aware Disambiguation (IaD) loss to suppress instance localization ambiguity by differentiating overlapping response maps generated by different referring texts on the same image. Extensive experiments show that our method outperforms SOTA methods on three common benchmarks.

1 Introduction

Referring Image Segmentation (RIS) aims to segment a target object in an image via a user-specified input text description. RIS has various applications, such as text-based image editing [17, 13, 2] and human-computer interaction [62, 51]. Despite remarkable progress, most existing RIS works [7, 58, 27, 26, 21, 5] rely heavily on pixel-level ground-truth masks to learn visual-linguistic alignment. Recently, there has been a surge in interest in developing weakly-supervised RIS (WRIS) methods via weak supervisions, *e.g.*, bounding-boxes [9], and text descriptions [54, 18, 30, 4], to alleviate burden of data annotations. In this work, we focus on obtaining supervision from text descriptions only.

The relatively weak constraint of utilizing text alone as supervision makes visual-linguistic alignment particularly challenging. There are some attempts [30, 18, 22, 46] to explore various alignment workflows. For example, TRIS [30] classifies referring texts that describe the target object as positive texts while other texts as negative ones, to model a text-to-image response map for locating potential target objects. SAG [18] introduces a bottom-up and top-down attention framework to discover

[†] Joint corresponding authors.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

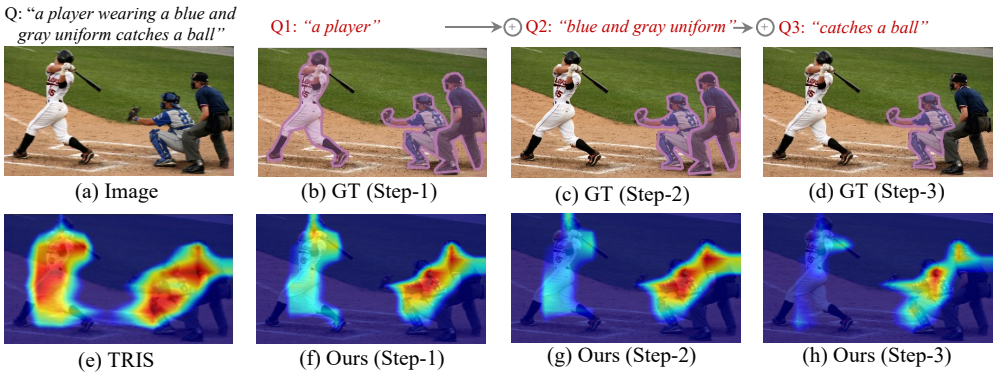


Figure 1: Given an image and a language description as inputs (a), RIS aims to predict the target object (d). Unlike existing methods (*e.g.*, TRIS [30] (e) – a WRIS method) that directly utilize the complete language description for target localization, we observe that humans would naturally break down the sentence into several key cues (*e.g.*, Q1 – Q3) and progressively converge onto the target object (from (b) to (d). This behavior inspires us to develop the Progressive Comprehension Network (PCNet), which merges text cues pertinent to the target object step-by-step (from (f) to (h)), significantly enhancing visual localization. $\oplus$ denotes the text combination operation.

individual entities and then combine these entities as the target of the referring expression. However, these methods encode the entire referring text as a single language embedding. They can easily overlook some critical cues related to the target object in the text description, leading to localization ambiguity and even errors. For example, in Fig. 1(e), TRIS [30] erroneously activates all three players due to its use of cross-modality interactions between the image and the complete language embedding only.

We observe that humans typically localize target objects through a step-by-step comprehension process. Cognitive neuroscience studies [48, 41] also support this observation, indicating that humans tend to simplify a complex problem by breaking it down into manageable sub-problems and reasoning them progressively. For example, in Fig. 1(b-d), human perception would first begin with *"a player"* and identify all three players (b). The focus is then refined by the additional detail *"blue and gray uniform"*, which helps exclude the white player on the left (c). Finally, the action *"catches a ball"* helps further exclude the person on the right, leaving the correct target person in the middle (d).

Inspired by the human comprehension process, we propose in this paper a novel Progressive Comprehension Network (PCNet) for WRIS. We first employ a Large Language Model (LLM) [59] to dissect the input text description into multiple short phrases. These decomposed phrases are considered as target-related cues and fed into a novel Conditional Referring Module (CRM), which helps update the global referring embedding and enhance target localization in a multi-stage manner. We also propose a novel Region-aware Shrinking (RaS) loss to facilitate visual localization across different stages at the region level. ReS first separates the target-related response map (indicating the foreground region) from the target-irrelevant response map (indicating the background region), and then constrains the background response map to progressively attenuate, thus enhancing the localization accuracy of the foreground region. Finally, we notice that salient objects in an image can sometimes trigger incorrect response map activation for text descriptions that aim for other target objects. Hence, we introduce an Instance-aware Disambiguation (IaD) loss to reduce the overlapping of the response maps by rectifying the alignment score of different referring texts to the same object.

In summary, our main contributions are as follows :

- We propose the Progressive Comprehension Network (PCNet) for the WRIS task. Inspired by the human comprehension processes, this model achieves visual localization by progressively incorporating target-related textual cues for visual-linguistic alignment.
- Our method has three main technical novelties. First, we propose a Conditional Referring Module (CRM) to model the response maps through multiple stages for localization. Second, we propose a Region-aware Shrinking (RaS) loss to constrain the response maps across different stages for better cross-modal alignment. Third, to rectify overlapping localizations, we propose an Instance-aware Disambiguation (IaD) loss for different referring texts paired with the same image.

- We conduct extensive experiments on three popular benchmarks, demonstrating that our method outperforms existing methods by a remarkable margin.

2 Related work

Referring Image Segmentation (RIS) aims to segment the target object from the input image according to the input natural language expression. Hu *et al.* [14] proposes the first CNN-based RIS method. There are many follow-up works. Early methods [60, 28, 38] focus on object-level cross-modal alignment between the visual region and the corresponding referring expression. Later, many works explore the use of attention mechanisms [15, 7, 58, 19] or transformer architectures [58, 29] to model long-range dependencies, which can facilitate pixel-level cross-model alignment. For example, CMPC [15] employs a two-stage progressive comprehension model to first perceive all relevant instances through entity wording and then use relational wording to highlight the referent. In contrast, our approach leverages LLMs to decompose text descriptions into short phrases related to the target object, focusing on sentence-level (rather than word-level) comprehension, which aligns more closely with human cognition. Focusing on the visual grounding, DGA [56] also adopts multi-stage refinement. It aims to model visual reasoning on top of the relationships among the objects in the image. Differently, our work addresses the weakly RIS task and aims to alleviate the localization ambiguity by progressively integrating fine-grained attribute cues.

Weakly-supervised RIS (WRIS) has recently begun to attract some attention, as it can substantially reduce the burden of data labeling especially on the segmentation field [25, 57, 63]. Feng *et al.* [9] proposes the first WRIS method, which uses bounding boxes for annotations. Several subsequent works [18, 22, 30] attempt to use weaker supervision signal, *i.e.*, text descriptions. SAG [18] proposes to first divide image features into individual entities via bottom-up attention and then employ top-down attention to learn relations for combining entities. Lee *et al.* [22] generate Grad-CAM for each word of the description and then consider the relations using an intra-chunk and inter-chunk consistency. Instead of merging individual responses, TRIS [30] directly learns the text-to-image response map by contrasting target-related positive and target-unrelated negative texts. Inspired by the generalization capabilities of segmentation foundation models [20, 8, 34], PPT [6] enables effective integration with pre-trained language-image models [43, 33] and SAM [20] by a lightweight point generator to identify the referent and context noise. Despite their success, these methods encode the full text as a single embedding for cross-modality alignment, which overlooks target-related nuances in the textual descriptions. In contrast, our method proposes to combine progressive text comprehension and object-centric visual localization to obtain better fine-grained cross-modal alignment.

Large Language Models (LLMs) are revolutionizing various visual domains, benefited by their user-friendly interfaces and strong zero-shot prompting capabilities [3, 49, 1, 47]. Building on this trend, recent works [42, 55, 53, 45, 64] explore the integration of LLMs into vision tasks (*e.g.*, language-guided segmentation [55, 53], relation [23], and image classification [42]) through parameter-efficient fine-tuning or knowledge extraction. For example, LISA [55] and GSVA [53] utilize LLaVA [32], a large vision-language model (LVLM), as a feature encoder to extract visual-linguistic cross-modality features and introduce a small set of trainable parameters to prompt SAM [20] for reasoning segmentation. RECODE [23] and CuPL [42] leverage the knowledge in LLMs to generate informative descriptions as prompts for different categories classification. Unlike these works, we capitalize on the prompt capability of LLMs to help decompose a single referring description into multiple target object-related phrases, which are then used in our progressive comprehension process for RIS.

3 Our Method

In this work, we observe that when identifying an object based on a description, humans tend to first pinpoint multiple relevant objects and then narrow their focus to the target through step-by-step reasoning [48, 41]. Inspired by this, we propose a Progressive Comprehension Network (PCNet) for WRIS, which enhances cross-modality alignment by progressively integrating target-related text cues at multiple stages. Fig. 2 shows the overall framework of our PCNet.

Given an image $\mathbf{I}$ and a referring expression $\mathbf{T}$ as input, we first feed $\mathbf{T}$ into a Large Language Model (LLM) to break it down into K short phrases $\mathcal{T}_{sub} = \{t_0, t_1, \cdots, t_{K-1}\}$, referred to as target-related text cues. We then feed image $\mathbf{I}$ and referring expression $\mathbf{T}$ and the set of short phrases $\mathcal{T}_{sub}$ into image encoder and text encoder to obtain visual feature $\mathbf{V}_0 \in \mathbb{R}^{H\times W\times C_v}$ and language

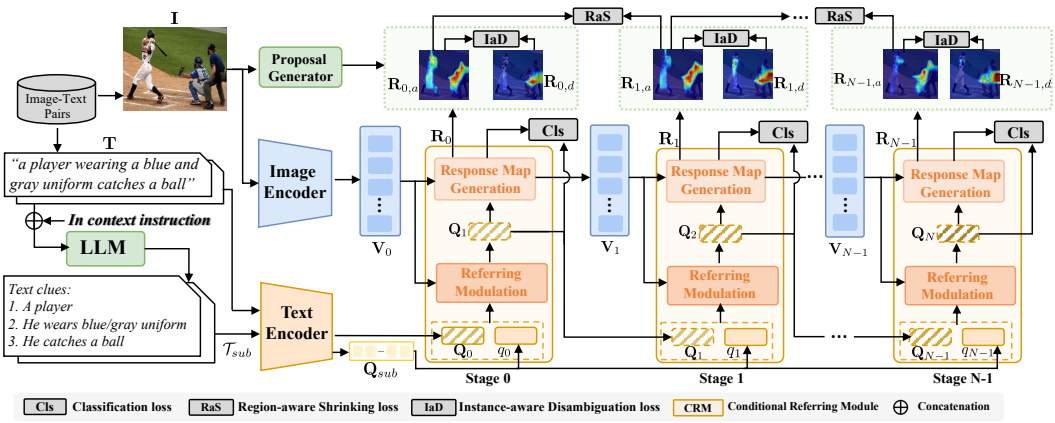


Figure 2: The pipeline of PCNet. Given a pair of image-text as input, PCNet enhances the visual-linguistic alignment by progressively comprehending the target-related textual nuances in the text description. It starts with using a LLM to decompose the input description into several target-related short phrases as target-related textual cues. The proposed Conditional Referring Module (CRM) then processes these cues to update the linguistic embeddings across multiple stages. Two novel loss functions, Region-aware Shrinking (RaS) and Instance-aware Disambiguation (IaD), are also proposed to supervise the progressive comprehension process.

feature $\mathbf{Q}_0 \in \mathbb{R}^{1\times C_t}$, and $\mathcal{Q}_{sub} = \{\mathbf{q}_0, \mathbf{q}_1, \cdots, \mathbf{q}_{K-1}\}$, with $\mathbf{q}_k \in \mathbb{R}^{1\times C_t}$, where $H = H_I/s$ and $W = W_I/s$. C_v and C_l denote the numbers of channels of visual and text features. s is the ratio of down-sampling. We then use projector layers to transform the visual feature $\mathbf{V}_0$ and textual features $\mathbf{Q}_0$ and $\mathcal{Q}_{sub}$ to a unified dimension C, *i.e.*, $\mathbf{V}_0 \in \mathbb{R}^{H\times W\times C}$, $\mathbf{Q}_0$ and $\mathbf{q}_i$ are in $\mathbb{R}^{1\times C}$.

We design PCNet with multiple consecutive Conditional Referring Modules (CRMs) to progressively locate the target object across N stages[1]. Specifically, at stage n, the n-th CRM updates the referring embedding $\mathbf{Q}_n$ into $\mathbf{Q}_{n+1}$ conditioned on the short phrase q_n from the proposed Referring Modulation block. Both $\mathbf{Q}_{n+1}$ and visual embedding $\mathbf{V}_n$ are fed into Response Map Generation to generate the text-to-image response map $\mathbf{R}_n$ and updated visual embedding $\mathbf{V}_{n+1}$. Finally, the response map $\mathbf{R}_{N-1}$ generated by the n-th CRM is used as the final localization result. To optimize the resulting response map for accurate visual localization, we employ the pre-trained proposal generator to obtain localization-matched mask proposals. We also propose Region-aware Shrinking (RaS) loss to constrain the visual localization in a coarse-to-fine manner, and Instance-aware Disambiguation (IaD) loss to suppress instance localization ambiguity.

In the following subsections, we first discuss how we decompose the input referring expression into target-related cues in Sec. 3.1. We then introduce the CRM in Sec. 3.2. Finally, we present our Region-aware Shrinking loss in Sec. 3.3, and Instance-aware Disambiguation loss in Sec. 3.4.

3.1 Generation of Target-related Cues

Existing works typically encode the entire input referring text description, and can easily overlook some critical cues (*e.g.*, attributes and relations) in the description (particularly for a long/complex description), leading to target localization problems. To address this problem, we propose dividing the input description into short phrases to process it individually. To do this, we leverage the strong in-context capability of the LLM [1] to decompose the text description. We design a prompt, with four parts, to instruct the LLM to do this: (1) general instruction $\mathbf{P}_G$; (2) output constraints $\mathbf{P}_C$; (3) in-context task examples $\mathbf{P}_E$; and (4) input question $\mathbf{P}_Q$. $\mathbf{P}_G$ describes the overall instruction, *e.g.*"decomposing the referring text into target object-related short phrases". $\mathbf{P}_C$ elaborates the output setting, *e.g.*, sentence length of each short phrase. In $\mathbf{P}_E$, we specifically curate several in-context pairs as guidance for the LLM to generate analogous outputs. Finally, $\mathbf{P}_Q$ encapsulates the input text description and the instruction words for the LLM to execute the operation. The process of generating

[1]Note that all counts start at 0.

target-related cues is formulated as:

$$\mathcal{T}_{sub} = \{t_0, t_1, \cdots, t_{K-1}\} = \text{LLM}\left(\mathbf{P}_G, \mathbf{P}_C, \mathbf{P}_E, \mathbf{P}_Q\right), \tag{1}$$

where K represents the total number of phrases, which varies depending on the input description. Typically, longer descriptions more likely yield more phrases. To maintain consistency in our training dataset, we standardize it to five phrases (*i.e.*, $K = 5$). If fewer than five phrases are produced, we simply duplicate some of the short phrases to obtain five short phrases. In this way, phrases generated by LLM are related to the target object and align closely with our objective.

3.2 Conditional Referring Module (CRM)

Given the decomposed phrases (*i.e.*, target-related cues), we propose a CRM to enhance the discriminative ability on the target object region conditioned on these phrases, thereby improving localization accuracy. As shown in Fig. 2, the CRM operates across N consecutive stages. At each stage, it first utilizes a different target-related cue to modulate the global referring embedding via a referring modulation block and then produces the image-to-text response map through a response map generation block.

Referring Modulation Block. Considering the situation at stage n, we first concatenate one target-related cue q_n and the L negative text cues obtained from other images [2], to form $\mathbf{q}'_n \in \mathbb{R}^{(L+1)\times C}$. We then fuse the visual features $\mathbf{V}_n$ with $\mathbf{q}'_n$ through a vision-to-text cross-attention, to obtain vision-attended cue features $\hat{\mathbf{q}}_n \in \mathbb{R}^{(L+1)\times C}$, as:

$$\boldsymbol{A}_{v\to t} = \text{SoftMax}\left((\mathbf{q}'_n W_1^{q'}) \otimes (\mathbf{V}_n W_2^V)^\top / \sqrt{C}\right);\ \hat{\mathbf{q}}_n = \text{MLP}(\boldsymbol{A}_{v\to t} \otimes (\mathbf{V}_n W_3^V)) + \mathbf{q}'_n, \tag{2}$$

where $\boldsymbol{A}_{v\to t} \in \mathbb{R}^{(L+1)\times H\times W}$ denotes the vision-to-text inter-modality attention weight. W_*^V and $W_*^{q'}$ are learnable projection layers. $\otimes$ denotes matrix multiplication. Using the vision-attended cue features $\hat{\mathbf{q}}_n$, we then enrich the global textual features $\mathbf{Q}_n$ into cue-enhanced textual features $\mathbf{Q}_{n+1} \in \mathbb{R}^{1\times C}$ through another text-to-text cross-attention, as:

$$\boldsymbol{A}_{t\to t} = \text{SoftMax}\left((\mathbf{Q}_n W_1^Q) \otimes (\hat{\mathbf{q}}_n W_2^{\hat{q}})^\top / \sqrt{C}\right);\ \mathbf{Q}_{n+1} = \text{MLP}(\boldsymbol{A}_{t\to t} \otimes (\hat{\mathbf{q}}_n W_3^{\hat{q}})) + \mathbf{Q}_n, \tag{3}$$

where $\boldsymbol{A}_{t\to t} \in \mathbb{R}^{1\times(L+1)}$ represents the text-to-text intra-modality attention weight. W_*^Q and $W_*^{\hat{q}}$ are learnable projection layers. In this way, we can enhance the attention of $\mathbf{Q}_n$ on the target object by conditioning its own target-related cue features and the global visual features.

Response Map Generation. To compute the response map, we first update the visual features $\mathbf{V}_n$ to $\hat{\mathbf{V}}_n$ by integrating them with the updated referring text embedding $\mathbf{Q}_{n+1}$ using a text-to-visual cross-attention, thereby reducing the cross-modality discrepancy. Note that $\hat{\mathbf{V}}_n$ is then used in the next stage (*i.e.*, $\mathbf{V}_{n+1} = \hat{\mathbf{V}}_n$). The response map $\mathbf{R}_n \in \mathbb{R}^{H\times W}$ at the n-th stage is computed as:

$$\mathbf{R}_n = \text{Norm}(\text{ReLU}(\hat{\mathbf{V}}_n \otimes \mathbf{Q}_{n+1}^\top)), \tag{4}$$

where Norm normalizes the output in the range of [0,1]. To achieve global visual-linguistic alignment, we adopt classification loss $\mathcal{L}_{\texttt{Cls}}$ in [30] to optimize the generation of the response map at each stage. It formulates the target localization problem as a classification process to differentiate between positive and negative text expressions. While the referring text expressions for an image are used as positive expressions, the ones from other images can be used as negative for this image. More explanations are given in appendix.

3.3 Region-aware Shrinking (RaS) Loss

Despite modulating the referring attention with the target-related cues stage-by-stage, image-text classification often activates irrelevant background objects due to its reliance on global and coarse response map constraints. Ideally, as the number of target-related cues used increases across each stage, the response map should become more compact and accurate. However, directly constraining the latter stage to have a more compact spatial activation than the former stage can lead to a trivial

[2] Refer to the Appendix for more details.

solution (*i.e.*, without target activation). To address this problem, we propose a novel region-aware shrinking (RaS) loss, which segments the response map into foreground (target) and background (non-target) regions. Through contrastive enhancement between these regions, our method gradually reduces the background interference while refining the foreground activation in the response map.

Specifically, at stage n, we first employ a pretrained proposal generator to obtain a set of mask proposals, $\mathcal{M} = \{m_1, m_2, \cdots, m_P\}$, where each proposal m_p is in $\mathbb{R}^{H\times W}$ and P is the total number of segment proposals. We then compute a alignment score between the response map $\mathbf{R}_n$ and each proposal m_p in $\mathcal{M}$ as:

$$\mathcal{S}_n = \{s_{n,1}, s_{n,2}, \cdots, s_{n,P}\} \text{ with } s_{n,p} = \max(\mathbf{R}_n \odot m_p), \tag{5}$$

where $\odot$ denotes the hadamard product. The proposal with the highest score (denoted as m_f) is then treated as the target foreground region, while the combination of other proposals (denoted as m_b) is regarded as non-target background regions. With the separated regions, we define a localization ambiguity S_n^{amb}, which measures the uncertainty of the target object localization in the current stage n, as:

$$S_n^{amb} = 1 - (\text{IoU}(\mathbf{R}_n, m_f) - \text{IoU}(\mathbf{R}_n, m_b)), \tag{6}$$

where S_n^{amb} is in the range of $[0, 1]$, and IoU denotes the intersection over union. When the localization result (*i.e.*, the response map) matches the only target object proposal instance exactly, ambiguity is 0. Conversely, if it matches the more background proposals, ambiguity approaches 1.

Assuming that each target in the image corresponds to an instance, by integrating more cues, the model will produce a more compact response map and gradually reduce the ambiguity. Consequently, based on the visual localization results from two consecutive stages, we can formulate the region-aware shrinking objective for a total of N stages as:

$$\mathcal{L}_{\texttt{RaS}} = \frac{1}{N-1}\sum_{n=0}^{N-2} \max\left(0, (S_{n+1}^{amb} - S_n^{amb})\right). \tag{7}$$

By introducing region-wise ambiguity, $\mathcal{L}_{\texttt{RaS}}$ can direct non-target regions to converge towards attenuation while maintaining and improving the quality of the response map in the target region. This enables the efficient integration of target-related textual cues for progressively finer cross-modal alignment. Additionally, the mask proposals can also provide a shape prior to the target region, which helps to further enhance the accuracy of the target object localization.

3.4 Instance-aware Disambiguation (IaD) Loss

Although the RaS loss can help improve the localization accuracy by reducing region-wise ambiguity within one single response map, it takes less consideration of the relation between different instance-wise response maps. Particularly, we note that, given different referring descriptions that refer to different objects of an image, there are usually some overlaps among the corresponding response maps. For example, in Fig. 2, the player in the middle is simultaneously activated by two referring expressions (*i.e.*, the response maps $\mathbf{R}_{*,a}$ and $\mathbf{R}_{*,d}$ have overlapping activated regions), resulting in inaccurate localization. To address this problem, we propose an Instance-aware Disambiguation (IaD) loss to help enforce that different regions of the response maps within a stage are activated if the referring descriptions of an image refer to different objects.

Specifically, given a pair of image $\mathbf{I}_a$ and input text description $\mathbf{T}_a$, we first sample extra N_d text descriptions, $\mathcal{T}_d = \{t_1, t_2, \cdots, t_{N_d}\}$, where the referred target object of each text description t_d is in the image $\mathbf{I}_a$ but is different from the target object referred to by $\mathbf{T}_a$. We then obtain the image-to-text response maps $\mathbf{R}_a$ and $\mathcal{R}_d = \{\mathbf{R}_1, \mathbf{R}_2, \cdots, \mathbf{R}_{N_d}\}$ for $\mathbf{T}_a$ and $\mathcal{T}_d$ through Eq. (4). Here, we omit the stage index n for clarity. Then, based on the Eq. (5), we obtain the alignment scores $\mathcal{S}_a$ and $\{\mathcal{S}_d\}_{d=1}^{N_d}$ for $\mathbf{T}_a$ and $\mathcal{T}_d$. In $\mathcal{S}$, the larger the value, the higher the alignment between the corresponding proposal (specified by the index) and the current text. To disambiguate overlapping activated regions, we constrain that the maximum index of the alignment score between $\mathcal{S}_a$ and each of $\mathcal{S}_d$ must be different from each other (*i.e.*, different texts must activate different objects). Here, we follow [50] to compute the index vector, $y \in \mathbb{R}^{1\times P}$, as:

$$y = \texttt{one-hot}(\texttt{argmax}(\mathcal{S})) + \mathcal{S} - \texttt{sg}(\mathcal{S}), \tag{8}$$

Table 1: Quantitative comparison using mIoU and PointM metrics. "(U)" and "(G)" indicate the UMD and Google partitions. **"Segmentor"** denotes utilizing the pre-trained segmentation models (SAM [20] by default) for segmentation mask generation. † denotes that the method is fully-supervised. "-" means unavailable values. Oracle represents the evaluation of the best proposal mask based on ground-truth. Best and second-best performances are marked in **bold** and underlined.

Metric	Method	Backbone	Segmentor	RefCOCO			RefCOCO+			RefCOCOg		
				Val	TestA	TestB	Val	TestA	TestB	Val (G)	Val (U)	Test (U)
PointM↑	GroupViT [54]	GroupViT	✘	25.0	26.3	24.4	25.9	26.0	26.1	30.0	30.9	31.0
	CLIP-ES [25]	ViT-Base	✘	41.3	50.6	30.3	46.6	56.2	33.2	49.1	46.2	45.8
	WWbL [46]	VGG16	✘	31.3	31.2	30.8	34.5	33.3	36.1	29.3	32.1	31.4
	SAG [18]	ViT-Base	✘	56.2	63.3	51.0	45.5	52.4	36.5	37.3	–	–
	TRIS [30]	ResNet-50	✘	51.9	60.8	43.0	40.8	40.9	41.1	52.5	51.9	53.3
	PCNet$_F$	ResNet-50	✘	59.6	66.6	48.2	54.7	65.0	44.1	57.9	57.0	57.2
	PCNet$_S$	ResNet-50	✘	**60.0**	**69.3**	**52.5**	**58.7**	**65.5**	**45.3**	**58.6**	**57.9**	**57.4**
mIoU↑	LAVT† [58]	Swin-Base	N/A	72.7	75.8	68.7	65.8	70.9	59.2	63.6	63.3	63.6
	GroupViT [54]	GroupViT	✘	18.0	18.1	19.3	18.1	17.6	19.5	19.9	19.8	20.1
	CLIP-ES [25]	ViT-Base	✘	13.8	15.2	12.9	14.6	16.0	13.5	14.2	13.9	14.1
	TSEG [15]	ViT-Small	✘	25.4	–	–	22.0	–	–	22.1	–	–
	WWbL [46]	VGG16	✘	18.3	17.4	19.9	19.9	18.7	21.6	21.8	21.8	21.8
	SAG [18]	ViT-Base	✘	**33.4**	33.5	**33.7**	28.4	28.6	**28.0**	28.8	–	–
	TRIS [30]	ResNet-50	✘	25.1	26.5	23.8	22.3	21.6	22.9	26.9	26.6	27.3
	PCNet$_F$	ResNet-50	✘	30.9	35.2	26.3	28.9	31.9	26.5	29.8	29.7	30.2
	PCNet$_S$	ResNet-50	✘	31.3	**36.8**	26.4	**29.2**	**32.1**	26.8	**30.7**	**30.0**	**30.6**
	CLIP [43]	ResNet-50	✔	36.0	37.9	30.6	39.2	42.7	31.6	37.5	37.4	37.8
	SAG [18]	ViT-Base	✔	44.6	50.1	38.4	35.5	41.1	27.6	23.0	–	–
	TRIS [30]	ResNet-50	✔	41.1	48.1	31.9	31.6	31.9	30.6	38.4	39.0	39.9
	PPT [6]	ViT-Base	✔	46.8	45.3	**46.3**	45.3	45.8	44.8	43.0	–	–
	PCNet$_S$	ResNet-50	✔	**52.2**	**58.4**	42.1	**47.9**	**56.5**	**36.2**	**47.3**	**46.8**	**46.9**
	Oracle	ResNet-50	✔	72.7	75.3	67.7	73.1	75.5	68.2	69.0	68.3	68.4

where $\texttt{sg}(\cdot)$ represents the stop gradient operation. Finally, we denote the index vectors for $\mathcal{S}_a$ and $\{\mathcal{S}_d\}_{d=1}^{N_d}$ as y_a and $\{y_d\}_{d=1}^{N_d}$, and we formulate the IaD loss as:

$$\mathcal{L}_{\texttt{IaD}} = \frac{1}{N_d}\sum_{d=1}^{N_d}\left(1 - ||y_a - y_d||^2\right). \tag{9}$$

By enforcing the constraint at each stage, the response maps activated by different referring descriptions in an image for different instances are separated, and the comprehension of the discriminative cues is further enhanced.

4 Experiments

4.1 Settings

Dataset. We have conducted experiments on three standard benchmarks: RefCOCO [61], RefCOCO+ [61], and RefCOCOg [39]. They are constructed based on MSCOCO [24]. Specially, the referring expressions in RefCOCO and RefCOCO+ focus more on object positions and appearances, respectively, and they are characterized by succinct descriptions, averaging 3.5 words in length. RefCOCOg contains much longer sentences (average length of 8.4 words), making it more challenging than others. RefCOCOg includes two partitions: UMD [40] and Google [39].

Implementation Details. We train our framwork for 15 epochs with a batch size of 36 on an RTX4090 GPU. The total loss for training is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\texttt{Cls}} + \mathcal{L}_{\texttt{RaS}} + \mathcal{L}_{\texttt{IaD}}$. By default, we set the number of stages N to 3, and the number of the additional text descriptions sampled for each image N_d to 1. Without loss of generality, we use FreeSOLO [52] and SAM [20] as the proposal generators to obtain two versions: PCNet$_S$ and PCNet$_F$. Refer to Sec. A for more implementation details.

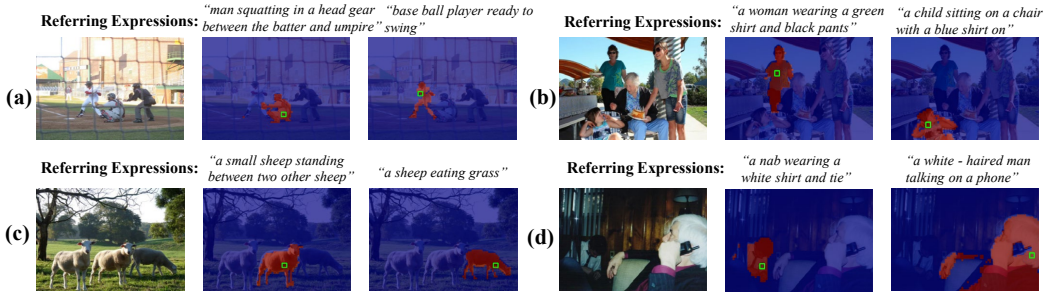


Figure 3: qVisual results of our PCNet. The green markers denote the peaks of the response maps.

Evaluation Metrics. We argue that the key to WRIS is target localization, and the performance evaluation should not rely primarily on pixel-wise metrics. With the accurate localization points, pixel-level masks can be readily obtained by prompting the pre-trained segmentors (*e.g.*, SAM [20]). Thus, following [18, 22, 30], we adopt localization-based metric (*i.e.*, PointM), and pixel-wise metrics (*i.e.*, mean and overall intersection-over-union (mIoU and oIoU) for evaluation. PointM [30] is used to evaluate the localization accuracy, which computes the ratio of activation peaks in the mask region.

4.2 Comparison with State-of-the-Art Methods

Quantitative comparison. Tab. 1 compares our method with various SOTA methods. Specifically, we first compare target localization accuracy using the PointM metric. We evaluate two model variants: PCNet$_F$ and PCNet$_S$, which use different segmentors (*e.g.*, FreeSOLO [52] and SAM [20]) to extract mask proposals for RaS and IaD losses. Even when using FreeSOLO as the proposal generator, our model still significantly outperforms all compared methods. For example, on the most challenging dataset, RefCOCOg, with more complex object relationships and longer texts, PCNet$_F$ achieves performance improvements of 55.2% and 10.3% on the Val (G) set compared to SAG and TRIS[3]. PCNet$_S$ further boosts the performance if we replace FreeSOLO with the stronger SAM.

In addition, we verify the accuracy of the response map through pixel-wise mIoU metric. Results are shown in the middle part of Tab. 1. Our PCNet still achieves superior performances on all benchmarks, against all compared methods. Particularly, PCNet$_F$ and PCNet$_S$ outperform TRIS by an improvement of 10.8% and 14.1% mIoU, respectively, on the RefCOCOg Val (G) set. In the bottom part of Tab. 1, we compare the accuracy of the extracted mask proposals generated using the target localization point (*i.e.*, the peak point of the response map) to prompt SAM. We can see that our PCNet significantly outperforms other WRIS methods. We can also see that higher PointM values correlate with higher mIoU accuracy values of the corresponding mask proposals for different methods. We further tested the mask accuracy using the Ground-Truth localization point (*i.e.*, the last row), and find that its performance even surpasses the fully-supervised method, LAVT [58]. All these results highlight the critical importance of target localization (*i.e.*, peak point) for the WRIS task.

In Fig. 3, we show some visual results of our method across different scenes by using the target localization point (*i.e.*, the green marker) to prompt SAM to generate the target mask. Our method effectively localizes the target instance among other instances within the image, even in complex scenarios with region occlusion (a), multiple instances (b), similar appearance (c), and dim light (d).

4.3 Ablation Study

We conduct ablation experiments on the RefCOCOg dataset and report the results on the Val (G) set from both PCNet$_S$ and PCNet$_F$ in Tab. 2, and from PCNet$_F$ in Tab. 3, Tab. 4, and Tab. 5.

Component Analysis. In Tab. 2, we first construct a single-stage baseline (*1st* row) to optimize visual-linguistic alignment by removing all proposed components and using only the global image-text classification loss $\mathcal{L}_{\texttt{Cls}}$. We then introduce the proposed conditional referring module (CRM) to the baseline to allow for multi-stage progressive comprehension (*2nd* row). To validate the efficacy of the region-aware shrinking loss (RaS) and instance-aware disambiguation loss (IaD), we introduce them separately (*3rd* and *4th* rows). Finally, we combine all proposed components (*5th* row).

[3]For a fair comparison, we remove its 2nd stage as it is used for enhancing pixel-wise mask accuracy.

h

Table 2: Component ablations on RefCOCOg Val (G) set.

$\mathcal{L}_{\texttt{Cls}}$	CRM	$\mathcal{L}_{\texttt{RaS}}$	$\mathcal{L}_{\texttt{IaD}}$	PointM		mIoU		oIoU	
				PCNet_S	PCNet_F	PCNet_S	PCNet_F	PCNet_S	PCNet_F
✓				51.7		25.3		25.1	
✓	✓			53.3		26.8		26.7	
✓	✓	✓		57.7	56.4	29.8	28.5	29.6	28.5
✓	✓		✓	55.3	54.3	28.3	27.7	28.2	27.8
✓	✓	✓	✓	58.6	57.9	30.7	29.8	30.6	30.1

Table 3: Ablation of the number of iterative stages N.

N	mIoU	oIoU	PointM
1	27.4	27.3	55.3
2	29.3	29.4	57.3
3	29.8	30.1	57.9
4	29.5	29.8	56.7

Table 4: Ablation of different modulation strategies in CRM.

Method	mIoU	oIoU	PointM
ADD	28.5	28.4	56.3
TTA	29.3	29.1	57.1
VTA+ADD	29.2	29.1	57.2
VTA+TTA	29.8	30.1	57.9

Table 5: Ablation of the numbers of descriptions N_d in IaD.

N_d	mIoU	oIoU	PointM
0	28.5	28.5	56.4
1	29.8	30.1	57.9
2	29.8	29.7	57.8
3	29.7	29.6	57.7

The results demonstrate that ❶ even using only $\mathcal{L}_{\texttt{Cls}}$, progressively introducing target-related cues through CRM can still significantly enhance target object localization. In particular, PCNet_S achieves improvements of 3.1% on PointM and 5.9% on mIoU; ❷ by using $\mathcal{L}_{\texttt{RaS}}$ to constrain response maps, making them increasingly compact and complete during the progressive comprehension process, the accuracy of target localization is dramatically enhanced, resulting in an improvement of 11.6% on PointM. ❸ although $\mathcal{L}_{\texttt{IaD}}$ can facilitate the separation of overlapping response maps between different instances within the same image and improve the discriminative ability of our model on the target object, the lack of constraints between consecutive stages results in a smaller performance improvement than $\mathcal{L}_{\texttt{RaS}}$; and ❹ all components are essential for our final PCNet, and combining them achieves the best performance. In Fig. 4, we also provide the visual results of the ablation study on two examples. We can see that each component can bring obvious localization improvement.

Number of Iterative Stages. In Tab. 3, we analyze the effect of the number of iterative stages N. When $N = 1$, we can only apply $\mathcal{L}_{\text{Cls}}$ and $\mathcal{L}_{\text{IaD}}$, but not $\mathcal{L}_{\texttt{RaS}}$, resulting in inferior results. Increasing N from 1 to 2 significantly improves the performance due to the progressive introduction of target-related cues. However, the improvement from $N = 2$ to $N = 3$ is less pronounced than from $N = 1$ to $N = 2$, and the performance stabilizes at $N = 3$. At $N = 4$, the performance slightly declines. This is because when the effective short-phrases decomposed by LLM are fewer than the number of stages, we need to repeat text phrases in later stages, which may affect the loss optimization.

Modulation Strategy. In Tab. 4, we ablate different variants of CRM: ❶ directly adding target cue features $\mathbf{q}_n$ and global referring features $\mathbf{Q}_n$ (denoted as ADD); ❷ fusing $\mathbf{q}_n$ and $\mathbf{Q}_n$ using only text-to-text cross-attention (denoted as TTA); ❸ first employing a vision-to-text cross-attention to fuse visual features $\mathbf{V}_n$ and $\mathbf{q}_n$ to obtain vision-attended features $\hat{\mathbf{q}}_n$, and then adding them to $\mathbf{Q}_n$ (denoted VTA+ADD). The results demonstrate that ADD is the least efficient method. TTA

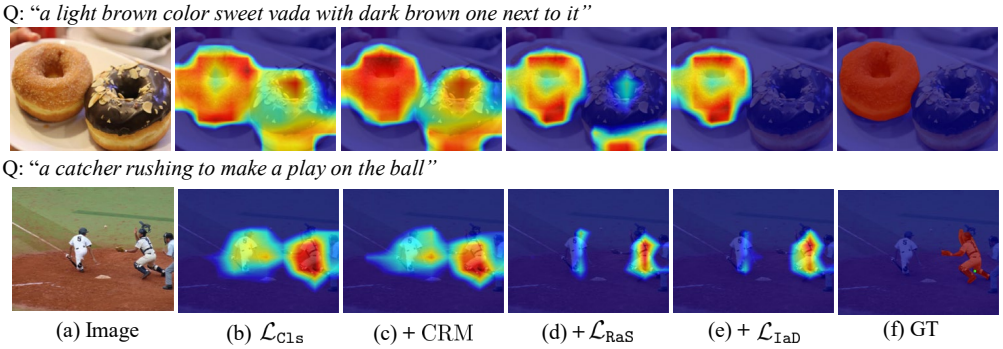


Figure 4: Visualization of the ablation study to show the efficacy of each proposed component.

Q: *"3 teddy bears sitting on a bed"*

(a) Image (b) GT (c) Ours

Figure 5: A failure case of our PCNet. As our model design assumes that there is only one object referred to by the language expression, it usually returns only one object.

outperforms ADD but is less effective than VTA+ADD, verifying the importance of the vision context. Finally, our CRM combines VTA and TTA and achieves the best results.

Number of Referring Texts. In Tab. 5, we analyze the effect of N_d used in $\mathcal{L}_{\text{LaD}}$. The results show that $N_d = 1$ is enough, and the performance deteriorates as N_d increases. This is because an image typically has 2-3 text descriptions, which means N_d should be 1-2. As N_d increases, repeated sampling becomes more frequent, affecting model training and thus leading to poorer results.

5 Conclusion

In this paper, we have proposed a novel Progressive Comprehension Network (PCNet) to perfom progressive visual-linguistic alignment for the weakly-supervised referring image segmentation (WRIS) task. PCNet first leverages a LLM to decompose the input referring description into several target-related phrases, which are then used by the proposed Conditional Referring Module (CRM) to update the referring text embedding stage-by-stage, thus enhancing target localization. In addition, we proposed two loss functions, region-aware shrinking loss and instance-aware disambiguation loss, to facilitate comprehension of the target-related cues progressively. We have also conducted extensive experiments on three RIS benchmarks. Results show that the proposed PCNet achieves superior visual localization performances and outperforms existing SOTA WRIS methods by large margins.

Our method does have limitations. For example, as shown in Fig. 5, when the text description refers to multiple objects, our method fails to return all referring regions. This is because our model design always assumes that there is only one object referred to by the language expression. In the future, we plan to incorporate more fine-grained vision priors [35, 44] and open-world referring descriptions (*e.g.*, camouflaged [12], semi-transparent [31], shadow [36] and *etc.*) into the model design to enable a more generalized solution.

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv:2303.08774 (2023)

[2] Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: CVPR. pp. 18208–18218 (2022)

[3] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. pp. 1877–1901 (2020)

[4] Chen, D., Wu, Z., Liu, F., Yang, Z., Huang, Y., Bao, Y., Zhou, E.: Prototypical contrastive language image pretraining. arXiv preprint arXiv:2206.10996 (2022)

[5] Chng, Y.X., Zheng, H., Han, Y., Qiu, X., Huang, G.: Mask grounding for referring image segmentation. arXiv:2312.12198 (2023)

[6] Dai, Q., Yang, S.: Curriculum point prompting for weakly-supervised referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13711–13722 (2024)

[7] Ding, H., Liu, C., Wang, S., Jiang, X.: Vision-language transformer and query generation for referring segmentation. In: ICCV. pp. 16321–16330 (2021)

[8] Du, Y., Bai, F., Huang, T., Zhao, B.: Segvol: Universal and interactive volumetric medical image segmentation. arXiv preprint arXiv:2311.13385 (2023)

[9] Feng, G., Zhang, L., Hu, Z., Lu, H.: Learning from box annotations for referring image segmentation. IEEE TNNLS (2022)

[10] He, C., Li, K., Zhang, Y., Xu, G., Tang, L., Zhang, Y., Guo, Z., Li, X.: Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. NeurIPS **36** (2024)

[11] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)

[12] He, R., Dong, Q., Lin, J., Lau, R.W.: Weakly-supervised camouflaged object detection with scribble annotations. In: AAAI (2023)

[13] Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-or, D.: Prompt-to-prompt image editing with cross-attention control. In: ICLR (2022)

[14] Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: ECCV. pp. 108–124 (2016)

[15] Huang, S., Hui, T., Liu, S., Li, G., Wei, Y., Han, J., Liu, L., Li, B.: Referring image segmentation via cross-modal progressive comprehension. In: CVPR. pp. 10488–10497 (2020)

[16] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)

[17] Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: CVPR. pp. 6007–6017 (2023)

[18] Kim, D., Kim, N., Lan, C., Kwak, S.: Shatter and gather: Learning referring image segmentation with text supervision. In: ICCV. pp. 15547–15557 (2023)

[19] Kim, N., Kim, D., Lan, C., Zeng, W., Kwak, S.: Restr: Convolution-free referring image segmentation using transformers. In: CVPR. pp. 18145–18154 (2022)

[20] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)

[21] Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. arXiv:2308.00692 (2023)

[22] Lee, J., Lee, S., Nam, J., Yu, S., Do, J., Taghavi, T.: Weakly supervised referring image segmentation with intra-chunk and inter-chunk consistency. In: ICCV. pp. 21870–21881 (2023)

[23] Li, L., Xiao, J., Chen, G., Shao, J., Zhuang, Y., Chen, L.: Zero-shot visual relation detection via composite visual cues from large language models. In: NeurIPS (2024)

[24] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755 (2014)

[25] Lin, Y., Chen, M., Wang, W., Wu, B., Li, K., Lin, B., Liu, H., He, X.: Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In: CVPR. pp. 15305–15314 (2023)

[26] Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: CVPR. pp. 23592–23601 (2023)

[27] Liu, C., Ding, H., Zhang, Y., Jiang, X.: Multi-modal mutual attention and iterative interaction for referring image segmentation. IEEE TIP (2023)

[28] Liu, D., Zhang, H., Wu, F., Zha, Z.J.: Learning to assemble neural module tree networks for visual grounding. In: ICCV. pp. 4673–4682 (2019)

[29] Liu, F., Kong, Y., Zhang, L., Feng, G., Yin, B.: Local-global coordination with transformers for referring image segmentation. Neurocomputing (2023)

[30] Liu, F., Liu, Y., Kong, Y., Xu, K., Zhang, L., Yin, B., Hancke, G., Lau, R.: Referring image segmentation using text supervision. In: ICCV. pp. 22124–22134 (2023)

[31] Liu, F., Liu, Y., Lin, J., Xu, K., Lau, R.W.: Multi-view dynamic reflection prior for video glass surface detection. In: AAAI (2024)

[32] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2024)

[33] Liu, Y., Wang, X., Zhu, M., Cao, Y., Huang, T., Shen, C.: Masked channel modeling for bootstrapping visual pre-training. International Journal of Computer Vision pp. 1–21 (2024)

[34] Liu, Y., Zhu, M., Li, H., Chen, H., Wang, X., Shen, C.: Matcher: Segment anything with one shot using all-purpose feature matching. arXiv preprint arXiv:2305.13310 (2023)

[35] Liu, Y., Ke, Z., Liu, F., Zhao, N., Lau, R.W.: Diff-plugin: Revitalizing details for diffusion-based low-level tasks. In: CVPR (2024)

[36] Liu, Y., Ke, Z., Xu, K., Liu, F., Wang, Z., Lau, R.W.: Recasting regional lighting for shadow removal. In: AAAI (2024)

[37] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2018)

[38] Luo, G., Zhou, Y., Sun, X., Cao, L., Wu, C., Deng, C., Ji, R.: Multi-task collaborative network for joint referring expression comprehension and segmentation. In: CVPR. pp. 10034–10043 (2020)

[39] Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: CVPR. pp. 11–20 (2016)

[40] Nagaraja, V.K., Morariu, V.I., Davis, L.S.: Modeling context between objects for referring expression understanding. In: ECCV. pp. 792–807 (2016)

[41] Plass, J.L., Moreno, R., Brünken, R.: Cognitive load theory. Cambridge University Press (2010)

[42] Pratt, S., Covert, I., Liu, R., Farhadi, A.: What does a platypus look like? generating customized prompts for zero-shot image classification. In: ICCV. pp. 15691–15701 (2023)

[43] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)

[44] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)

[45] Salewski, L., Alaniz, S., Rio-Torto, I., Schulz, E., Akata, Z.: In-context impersonation reveals large language models' strengths and biases. In: NeurIPS (2024)

[46] Shaharabany, T., Tewel, Y., Wolf, L.: What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs. pp. 28222–28237 (2022)

[47] Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., Li, H.: Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. arXiv preprint arXiv:2403.16999 (2024)

[48] Simon, H.A., Newell, A.: Human problem solving: The state of the theory in 1970. American Psychologist (1971)

[49] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv:2302.13971 (2023)

[50] Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning (2017)

[51] Wang, X., Huang, Q., Celikyilmaz, A., Gao, J., Shen, D., Wang, Y.F., Wang, W.Y., Zhang, L.: Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In: CVPR. pp. 6629–6638 (2019)

[52] Wang, X., Yu, Z., De Mello, S., Kautz, J., Anandkumar, A., Shen, C., Alvarez, J.M.: Freesolo: Learning to segment objects without annotations. In: CVPR. pp. 14176–14186 (2022)

[53] Xia, Z., Han, D., Han, Y., Pan, X., Song, S., Huang, G.: Gsva: Generalized segmentation via multimodal large language models. arXiv:2312.10103 (2023)

[54] Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: Groupvit: Semantic segmentation emerges from text supervision. In: CVPR. pp. 18134–18144 (2022)

[55] Yang, S., Qu, T., Lai, X., Tian, Z., Peng, B., Liu, S., Jia, J.: An improved baseline for reasoning segmentation with large language model. arXiv:2312.17240 (2023)

[56] Yang, S., Li, G., Yu, Y.: Dynamic graph attention for referring expression comprehension. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4644–4653 (2019)

[57] Yang, Z., Ke, Z., Hancke, G., Lau, R.: Cross-domain semantic decoupling for weakly-supervised semantic segmentation (2023)

[58] Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: CVPR. pp. 18155–18165 (2022)

[59] Ye, J., Chen, X., Xu, N., Zu, C., Shao, Z., Liu, S., Cui, Y., Zhou, Z., Gong, C., Shen, Y., et al.: A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. arXiv:2303.10420 (2023)

[60] Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: CVPR. pp. 1307–1315 (2018)

[61] Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: ECCV. pp. 69–85 (2016)

[62] Zhu, F., Zhu, Y., Chang, X., Liang, X.: Vision-language navigation with self-supervised auxiliary reasoning tasks. In: CVPR. pp. 10012–10022 (2020)

[63] Zhu, L., Zhou, J., Liu, Y., Hao, X., Liu, W., Wang, X.: Weaksam: Segment anything meets weakly-supervised instance-level recognition. arXiv preprint arXiv:2402.14812 (2024)

[64] Zong, Z., Ma, B., Shen, D., Song, G., Shao, H., Jiang, D., Li, H., Liu, Y.: Mova: Adapting mixture of vision experts to multimodal context. arXiv preprint arXiv:2404.13046 (2024)

A More Implementation Details

A.1 Generation of Target-related Cues

To obtain multiple target-related cues, we leverage the strong in-context capability of the Large Language Model (LLM) [16] to decompose the input referring expression and obtain the target-related textual cues. The Fig. 6 presents the LLM prompting details. The prompt includes four parts: (1) general instruction $\mathbf{P}_G$, (2) output constraints $\mathbf{P}_C$, (3) in-context task examples $\mathbf{P}_E$, and (4) input question $\mathbf{P}_Q$. In part $\mathbf{P}_G$, we define a overall instruction for our task (i.e, decomposing the referring text) Then in part $\mathbf{P}_C$, we elaborate some details about the output (e.g, the sentence length for each cue description). In part $\mathbf{P}_E$, we curate several in-context learning examples as guidance for the LLM to generating analogous output. Considering that the input referring expressions contain various sentence structures, in part $\mathbf{P}_E$ the more examples given, the more reliable the output will be. The part $\mathbf{P}_Q$ instructs the LLM to output the results given the input referring expression.

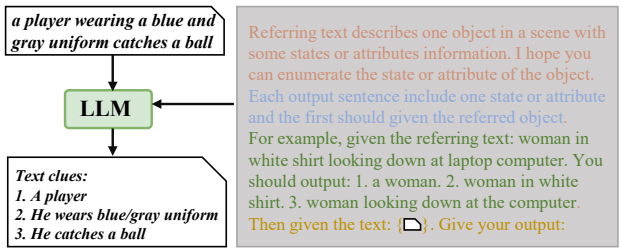


Figure 6: Flow of LLM-based referring text decomposition.

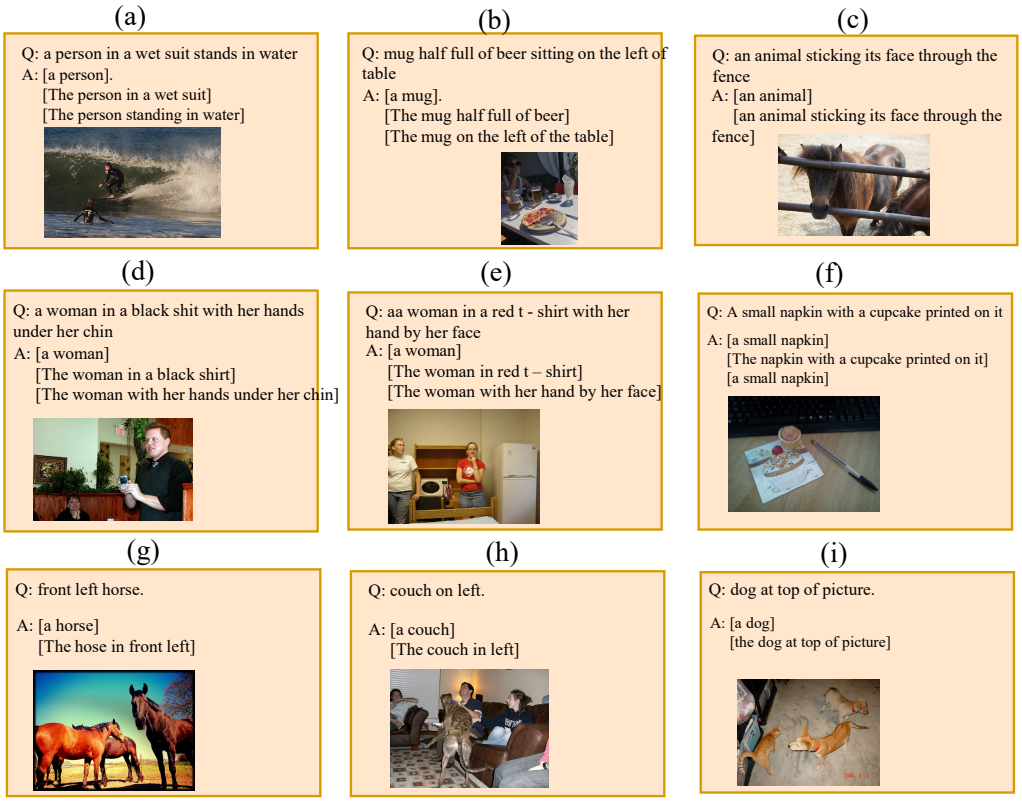


Figure 7: LLM generated examples. We show the LLM generated examples for long language expressions in (a)-(f) and the ones for short language expressions in (g)-(i). "Q" denotes the input language expression and the "A" denotes the output target-related textual cues of LLM.

In Fig. 7, we give some examples of decomposing the referring expressions. The "Q" denotes the input referring content and the "A" denotes the answers of LLM. In most cases, we can obtain reliable target-related textual cues that do not contradict the original text input. Besides, we also notice that there are some cases in which the text is not sufficiently decomposed (e.g, the example (c)) or the LLM outputs redundant results (e.g, the example (f)), which hinders the model to benefit from progressive comprehension to some extent.

A.2 Text-to Image Classification Loss

Our work consists of multiple stages and utilizes $\mathcal{L}_{\texttt{cls}}$ in TRIS [30] at each stage independently for response maps optimization. Here, we omit the index of stage n for clarity. $\mathcal{L}_{\texttt{cls}}$ formulates the target localization problem as a classification process to differentiate between positive and negative text expressions. The key idea of $\mathcal{L}_{\texttt{cls}}$ loss function is to contrast image-text pairs such that correlated image-text pairs have high similarity scores and uncorrelated image-text pairs have low similarity scores. While the referring text expressions for an image are used as positive expressions, the referring text expressions from other images can be used as negative expressions for this image. Thus, given a batch (i.e., B) of image samples , each sample is mutually associated with one positive reference text (i.e., a text describing a specific object in the current image) and mutually exclusive with L negative reference texts (texts that are not related to the target object in the image). Note that the number of batches is equal to the sum of the positive samples and the negative samples (i.e., $B = 1 + L$).

Speficially, in each training batch, B image-text pairs $\{\mathbf{I}_i, \mathbf{T}_i\}_{i=1}^{B}$ are sampled. Through the language and vision encoders, we can get referring embeddings $\mathbf{Q} \in \mathbb{R}^{B\times C}$ and image embeddings $\mathbf{V} \in \mathbb{R}^{B\times H\times W\times C}$. Then, we obtain the response maps $\mathbf{R} \in \mathbb{R}^{B\times B\times H\times W}$ by applying cosine similarity calculation and normalization operation. After the pooling operation as done in TRIS, we further obtain the alignment score matrix $\mathbf{y} \in \mathbb{R}^{B\times B}$. According to the $\mathcal{L}_{\texttt{cls}}$, for i_{th} image in the batch, there is a prediction score $\mathbf{y}[i,:]$, where $\mathbf{y}[i,i]$ predicted by the corresponding text deserves a higher value (i.e, the positive one) and the others deserve lower values (L negative ones). Then classification loss for the i_{th} image from the batch can be formulated as cross-entropy loss:

$$\mathcal{L}_{\texttt{cls},i} = -\frac{1}{B}\sum_{j=1}^{B}\left(\mathbb{1}_{i=j}\log\left(\frac{1}{1+e^{-\mathbf{y}[i,j]}}\right) + (1-\mathbb{1}_{i=j})\log\left(\frac{e^{-\mathbf{y}[i,j]}}{1+e^{-\mathbf{y}[i,j]}}\right)\right),$$

and the classification loss for the batch can be formulated as:

$$\mathcal{L}_{\texttt{cls}} = \frac{1}{B}\sum_{i=1}^{B}\mathcal{L}_{\texttt{cls},i}$$

The i denotes the index for the visual image and the j denotes the index for the referring text.

A.3 Referring Modulation Block

In Sec. 3.2, we have mentioned that the conditional referring module (CRM) utilizes the decomposed textual cues to progressively modulate the referring embedding via a modulation block across N consecutive stages, and then produces the image-to-text response map by computing the patch-based similarity between visual and language embeddings. Specifically, the modulation block is implemented by a vision-to-text and a text-to-text cross-attention mechanism in cascade for facilitating the interaction between cross-modal features. In Fig. 8, we give an overview of the block design. For the block at each stage, We concatenate one target-related cue and the L negative text cues obtained from other images as the conditional text cues and then obtain the vision-attended cue features by the vision-to-text attention. Then by learning the interaction between referring embedding and different textual cue embeddings, the block is expected to enhance the integration of discriminative cues.

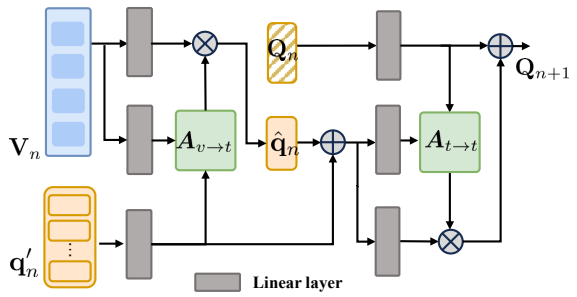


Figure 8: Illustration of referring modulation block.

A.4 Training and Inference

We implement our framework on PyTorch and train it for 15 epochs with a batch size of 36 (*i.e.*, $L+1$) on a RTX4090 GPU with 24GB of memory. The input images are resized to 320×320. We use ResNet-50 [11] as our backbone of image encoder, and utilize the pre-trained CLIP [43] model to initialize the image and text encoders. The down-sampling ratio of visual feature $s = 32$, the channels of vision feature $C_v = 2048$, text features $C_l = 1024$, and the unified hidden dimension $C = 1024$. The network is optimized using the AdamW optimizer [37] with a weight decay of $1e^{-2}$ and an initial learning rate of $5e^{-5}$ with polynomial learning rate decay. For the LLM, we utilize the open-source powerful language model Mistral 7B [16] for referring text decomposition. For the proposal generator, we set the number of extracted proposals $P = 40$ for each image.

B More Quantitative Studies

B.1 Comparisons with other SOTA methods

Table 6: Different criterions for alignment score measurement in $\mathcal{L}_{\texttt{RaS}}$.

Alignment Score	mIoU	PointM	oIoU
Max	29.8	57.9	30.1
Avg	29.1	56.4	29.2

Table 7: Different criterions for alignment score measurement in $\mathcal{L}_{\texttt{IaD}}$.

Alignment Score	mIoU	PointM	oIoU
Max	29.8	57.9	30.1
Avg	28.8	54.9	29.0

In Tab. 6 and Tab. 7, we conduct the ablation studies about the measurement criterion of alignment score. The results demonstrate that the maximum value of the response map in each proposal better represents the alignment level of region-wise cross-modal alignment than the average value. To validate the effectiveness of the modeling the progressive comprehension, we also quantitatively compare the outputs of our method at different stages in Tab. 8. The results show that the localization results gradually improve with more discriminative cues integration, especially in the early stages.

Table 8: Comparison between different stages.

Stage Num.	Stage 0	Stage 1	Stage 2
mIoU	28.6	29.4	29.8
PointM	56.7	57.6	57.9

B.2 More Ablation Studies

Comparison between IaD loss and others. In our IaD loss $\mathcal{L}_{\texttt{IaD}}$, we adopt a hard assignment for deriving the loss function as GroupViT [54]. The motivation is that we aims to get the pseudo mask prediction by the accurate peak value point (i.e., the hard assignment results) instead of relying on whole score distribution $S(\cdot)$ (e.g., S_a, S_d in Sec. 3.4). Thus utilizing the hard assignment to derive the IaD loss well matches our purpose, which helps rectify the ambiguous localization results. If we use the soft assignment (e.g., measuring KL divergence between S_a and S_d), though the equivalent may be simpler, it not only does not match our purpose but also introduces more tricky components for optimization (e.g., extra distribution regularization is required). In order to verify the argument, we conduct a comparison on RefCOCOg(G) val dataset as Tab. 9. The $\mathcal{L}_{\texttt{IaD}}$ even causes a slight decline, while the proposed loss $\mathcal{L}_{\texttt{KL}}$ brings clear improvement on localization accuracy.

Table 9: Comparison between IaD loss and KL loss.

$\mathcal{L}_{\texttt{CLs}}$	$\mathcal{L}_{\texttt{KL}}$	$\mathcal{L}_{\texttt{IaD}}$	PointM	mIoU
✓			51.7	25.3
✓	✓		51.2	24.8
✓		✓	53.1	26.6

Comparison between IaD loss and calibration loss in TRIS. There are essential differences between them. First, the calibration loss in TRIS [30] is used to suppress noisy background activation and thus help to re-calibrate the target response map. In contrast, in our method, we observe that, multiple referring texts corresponding to different instances in one image may locate the same instances (or we say overlapping), due to the lack of instance-level supervision in WRIS.

As for the implementation, the calibration loss adopts the global CLIP score of image-text to implement a simple contrastive learning for revising the response map. Differently, we simultaneously infer the response maps of different referring texts from the same image, and obtain the instance-level localization results by choosing the mask proposal with max alignment score. To further verify the superiority of our loss, we conduct an ablation on the RefCOCO(val) dataset. We use the TRIS without calibration loss as the baseline and then separately introduced these two loss functions for comparison. Both ablations demonstrate that the IaD loss not only refines the response map (mIoU) but also significantly improves the localization accuracy (PointM).

Table 10: Comparison between IaD loss and calibration loss $\mathcal{L}_{\texttt{Cal}}$.

$\mathcal{L}_{\texttt{CLs}}$	$\mathcal{L}_{\texttt{IaD}}$	$\mathcal{L}_{\texttt{Cal}}$	PointM	mIoU
✓			50.3	24.6
✓	✓		51.4	26.4
✓		✓	54.7	26.3

C More Visualization of Localization Results

Progressive Comprehension for Localization. In Fig. 9, we also give visualization of each stage's response map for qualitative analysis. The results show that our proposed CRM module can effectively integrate the target-related textual cues. For example, in the first row, the method produces ambiguous localization result at the first stage. After taking the cue *"with gold necklace"* into consideration, the attention is transferred to the target object at the second stage. Finally, after considering all the cues, the method produces less ambiguous and more accurate localization results.

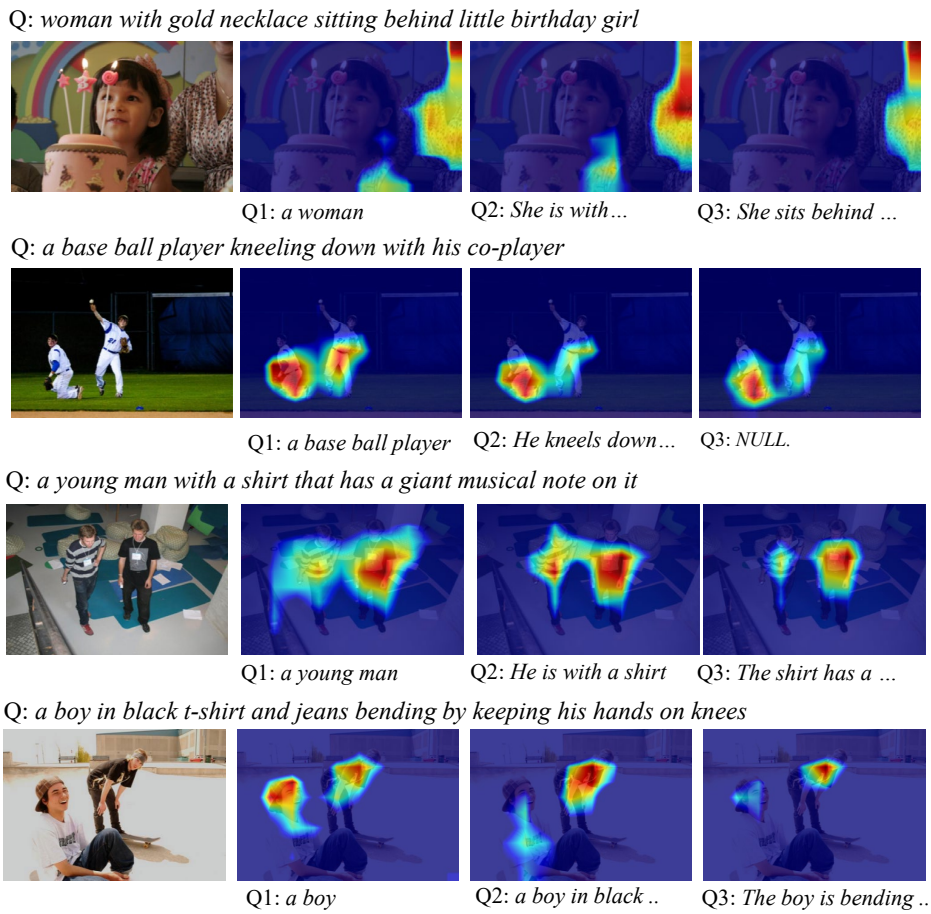


Figure 9: Visualization of progressive localization. With the integration of discriminative cues, the identification of target instance gradually improves.

Qualitative Comparison with Other method. More qualitative comparisons of our method with other methods are shown in the Fig. 10. For the example shown in the fifth row, the query is "a man in a arm striped sweater". The TRIS [30] mistakenly locates the left man as the target regions. In contrast, our PCNet optimizes the response map generation process by continuously modulating the referring embedding query conditioned on the target-related cues instead of a static referring embedding. As a result, our method can obtain more accurate localization result.

D Proposal Generator

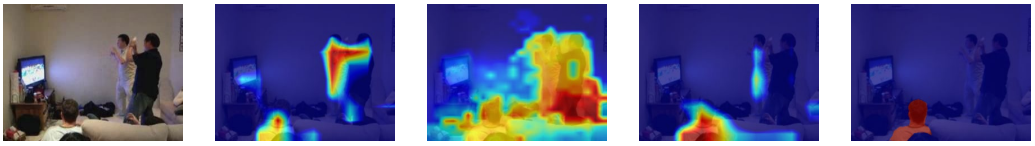
In this work, we adopt the two representative pre-trained segmentors: FreeSOLO [52] and SAM [20] for proposal generation. Specifically, the FreeSOLO [52] is a fully unsupervised learning method that learns class-agnostic instance segmentation without any annotations. The SAM's training utilizes densely labeled data, but it does not include semantic supervision. This supervision does not contradict our weakly supervised RIS setting. More importantly, it offers a promising solution as an image segmentation foundation model and can be used for refining the coarse localization results from weakly-supervised methods into precise segmentation masks as done in the recent works [10, 63]. For the usage of SAM, We adopt the `ViT-H` backbone, the hyperparameter *predicted iou threshold* and *stability score threshold* are set to 0.7, and *points per side* is set to 8.

In Fig. 11 and Fig. 12, we give the generated mask proposals examples by FreeSOLO [52] and SAM [20], respectively. We notice that there are often overlaps among the generated proposals. Thus, we refine the generated proposals by filtering out candidate proposals with small area (the threshold is set as 1000) and then selecting the ones with smaller intersection over union (the threshold is set as 0.8). Considering that the number of proposals generated by the segmentor may be different for different image inputs, in implementation, we maintain consistency by selecting the top 40 proposals with the largest area ($P = 40$). If fewer than 40, we simply complete it with an all-zero mask.

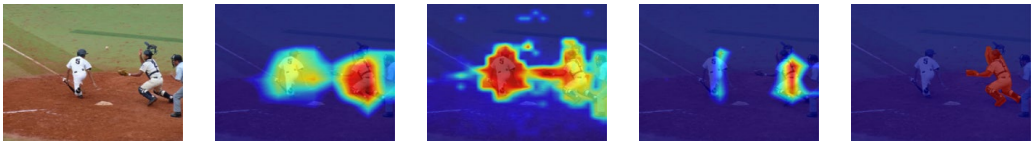
E Broader Impacts

While we do not foresee our method causing any direct negative societal impact, it may potentially be leveraged by malicious parties to create applications that could misuse the segmentation capabilities for unethical or illegal purposes. We urge the readers to limit the usage of this work to legal use cases.

Q: "*the guy wearing white sitting on the couch watching his friends play video games*"

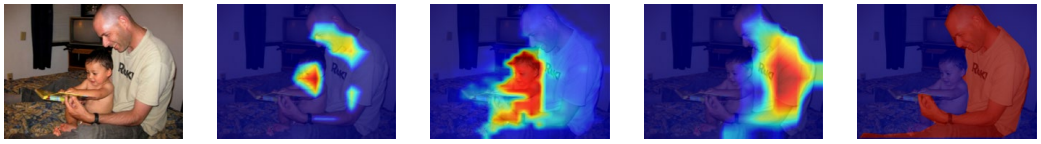


Q: "*a catcher rushing to make a play on the ball*"

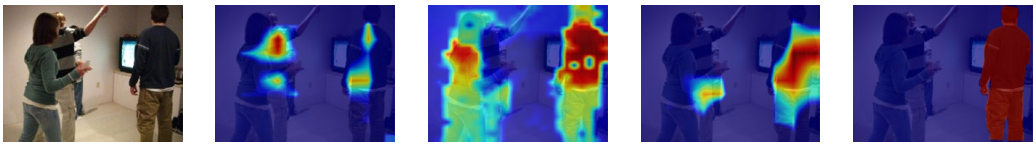


Q: "*a boy in black t - shirt and jeans bending by keeping his hands on knees*"

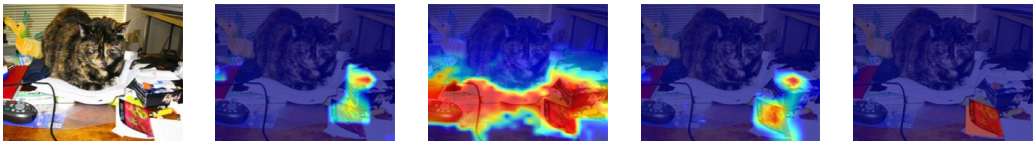
Q: "*a smiling man with a small infant wearing a beige t - shirt and gray jeans*"



Q: "*a man in a arm striped sweater*"



Q: "*a postcard with picture of face of cute girl*"



Q: "*dark haired woman wearing a blue jacket next to a teddy bear*"

(a) Image **(b) TRIS** **(c) SAG** **(d) PCNet** **(e) GT**

Figure 10: More visual comparison between our method with TRIS and SAG for WRIS.

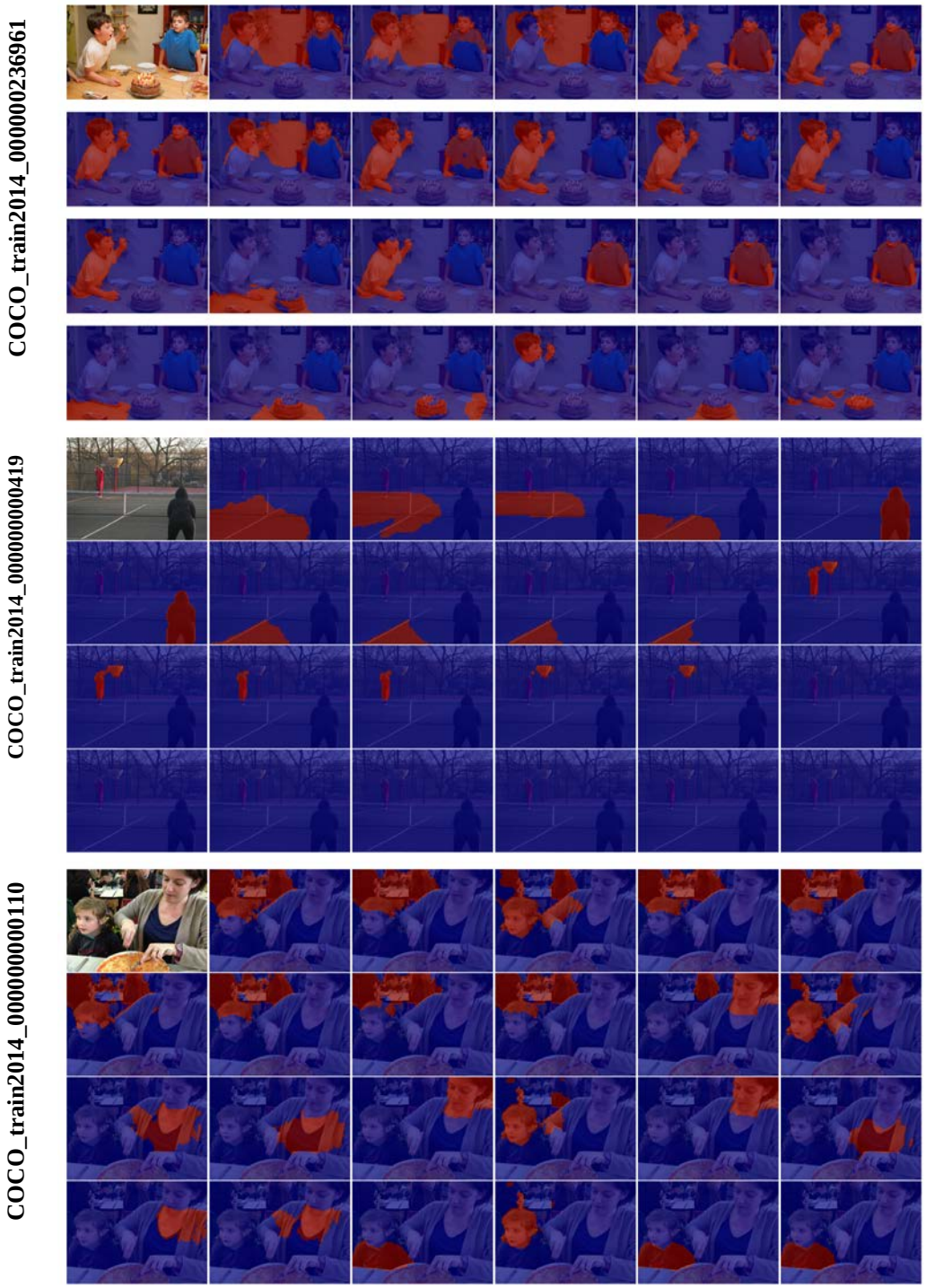


Figure 11: FreeSOLO [52] generated mask proposals examples.

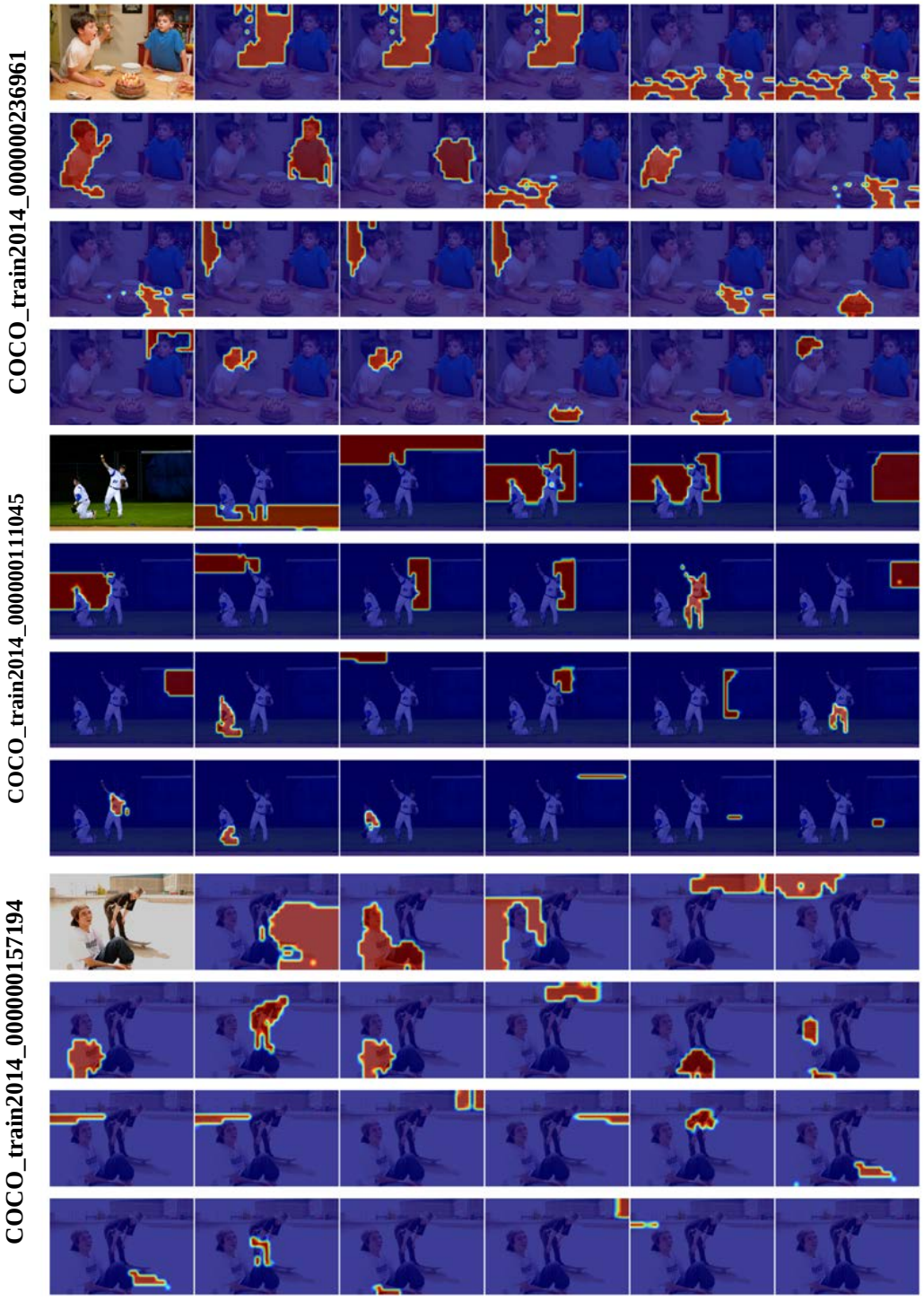


Figure 12: SAM [20] generated mask proposals examples.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In Sec. 1, we make the claims to reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Sec. 5, we discuss the limitations of the work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide implementation details in Sec. 4 and the Appendix for reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We will consider releasing the data and code once the paper is accepted.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide details about the experimental setting (e.g., hyperparameters) in Sec. 4 and the Appendix for reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars due to limited computing resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the information on the computer resources in the implementation details of Sec. 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in the paper conform with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses societal impacts of the work in the Appendix Sec. E.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly cites the existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.