

Region-aware Anchoring Mechanism for Efficient Referring Visual Grounding

Shuyi Ouyang¹ Ziwei Niu¹ Hongyi Wang¹ Yen-Wei Chen^{2*} Lanfen Lin^{1*}
¹Zhejiang University, China ²Ritsumeikan University, Japan
 {ouysy,nzw,whongyi,llf}@zju.edu.cn, chen@is.ritsumei.ac.jp

Abstract

Referring Visual Grounding (RVG) tasks revolve around utilizing vision-language interactions to incorporate object information from language expressions, thereby enabling targeted object detection or segmentation within images. Transformer-based methods have enabled effective interaction through attention mechanisms, achieving notable performance in RVG tasks. However, existing strategies for RVG, which involve direct interaction between visual and linguistic features, face three key challenges: (i) tendency to focus on a single target, (ii) insufficient control over linguistic noise, and (iii) high computational cost. To address these challenges, we propose a Region-aware Anchoring Mechanism (RaAM) that mediates vision-language interactions. In RaAM, region-aware anchors engage in alternating interactions with vision and language modalities, acting as indicators for object presence across different regions within the image. RaAM (i) directs attention to multiple target regions for better localization, (ii) reduces cross-modal redundancy by using anchors as buffers, and (iii) lowers time complexity. In addition, we design region and pixel level loss functions to enhance object presence assessment and edge precision. We evaluate our RaAM-RVG on four benchmark datasets and integrate RaAM into various models by replacing their interaction design. Results show that RaAM outperforms state-of-the-art methods with lower computational cost.

1. Introduction

Referring Expression Comprehension (REC) [51, 56, 62] aims to detect objects within an image that correspond to a given natural language expression, while Referring Expression Segmentation (RES) [54–56] requires precise segmentation of these identified objects. Object localization based on language expressions relies on aligning vision and language modalities. To address more complex real-world needs, recent work has extended RES to cre-

*Corresponding Authors: Lanfen Lin (llf@zju.edu.cn), Yen-Wei Chen (chen@is.ritsumei.ac.jp).

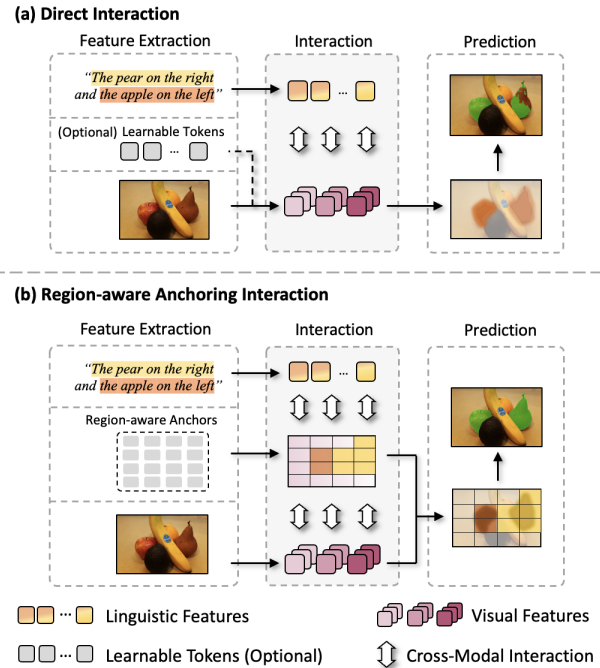


Figure 1. Comparison of the existing architecture (a) for RVG tasks with the proposed RaAM (b).

ate a new dataset, gRefCOCO, and established a benchmark termed Generalized Referring Expression Segmentation (GRES) [23]. The GRES benchmark introduces scenarios where a single sample may contain multiple target objects or none. Since these tasks involve identifying instances with specific attributes from language expressions, we refer to them collectively as *Referring Visual Grounding (RVG)*. The core of RVG lies in leveraging vision-language interactions to integrate object attributes and relationships from language expressions into the visual context, enabling instance localization within images. Additionally, preserving fine-grained details in visual features during interactions is crucial for ensuring accurate boundary delineation.

Transformer-based methods have effectively facilitated vision-language interactions through various attention mechanisms, achieving advanced results across RVG tasks [13, 26, 50]. Most existing studies use *Direct In-*

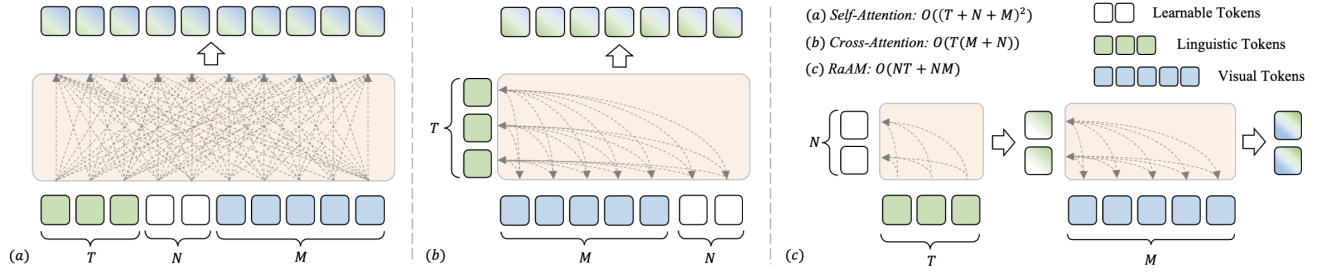


Figure 2. Comparison of existing interaction mechanisms with RaAM in design and complexity: (a) self-attention-like, (b) cross-attention-like, and (c) our RaAM. In (a) and (b), when $N = 0$, methods do not incorporate additional learnable tokens during interaction.

interaction strategies, where information flows directly between vision and language modalities, as shown in Fig. 1(a) [18, 33, 36, 48, 54, 58]. Recent studies have introduced additional learnable tokens to facilitate multimodal comprehension but still relies on direct interaction strategies [23, 39], including the recently emerging Multimodal Large Language Models (MLLMs) [20, 23]. The existing interaction mechanisms can be categorized into the two strategies illustrated in Fig. 2(a)(b). In Fig. 2(a), methods concatenate linguistic tokens, learnable tokens (optional), and visual tokens into a unified sequence for self-attention [43, 59]. Fig. 2(b) incorporates unimodal prompts first, followed by cross-attention for cross-modal interaction [14, 19, 42]. These methods progressively focus on target object locations during multimodal feature learning, facilitating object perception.

However, the direct interaction strategies face three challenges: (i) *Tendency to focus on a single target*: These strategies often prioritize the most prominent target, making it difficult to handle multi-object or no-object scenarios. (ii) *Insufficient control over linguistic noise*: Excessive incorporation of linguistic information may introduce cross-modal redundancy, potentially degrading fine-grained visual representations. (iii) *High computational cost*: As real-world demands for language comprehension grow, the growing length of linguistic feature representations increases computational demands, posing challenges for efficient deployment.

To address these challenges, we propose an efficient **Region-aware Anchoring Mechanism (RaAM)** for RVG, as shown in Fig. 1(b). In RaAM, we introduce learnable region-aware anchors to alternately interact with vision and language modalities, where these anchors serve as indicators of object presence across various regions within the visual environment. RaAM offers three key advantages: (i) *Enhanced complex object localization*: Region-aware anchors act as explicit object indicators, guiding attention to target presence across regions, thereby improving both multi-object localization and no-object discrimination. (ii) *Reduced cross-modal redundancy*: Anchors buffer linguistic noise from visual features. The Language-

to-Vision (L2V) gate further enhances high-frequency details and suppresses low-frequency noise, improving boundary precision. (iii) *Efficient vision-language interaction*: As shown in Fig. 2, unlike existing interaction mechanisms ((a) and (b)), RaAM introduces a region-aware anchoring mechanism (illustrated in (c)) that iteratively alternates between linguistic and visual anchor learning. Compared to (a) self-attention, RaAM reduces time complexity to a linear scale; and relative to (b) cross-attention, RaAM achieves further time complexity reduction due to the greater number of visual feature tokens, where $M \gg T > N$. We develop RaAM-RVG, an RVG model that incorporates RaAM, where vision-language interaction is termed the Region-aware Anchoring Network. Under anchor guidance, RaAM-RVG models visual scenes using linguistic cues, ultimately integrating anchors carrying target location information with visual features. This approach effectively guides predictions through explicit region modeling while reducing computational overhead.

To enhance edge detail discernment and target quantity estimation, we design loss functions. At the region level, we introduce a loss term \mathcal{L}_r to improve the model’s assessment of object presence within each region, refining region-aware anchors’ accuracy in indicating object existence. At the pixel level, we define the loss term \mathcal{L}_p^{seg} and \mathcal{L}_p^{det} . \mathcal{L}_p^{seg} increases the weight of ambiguous boundary areas, enhancing edge precision in complex scenes.

In summary, our contributions are three-folded:

- We propose RaAM-RVG, incorporating our novel interaction mechanism, RaAM, which uses region-aware anchors to guide vision-language interactions. RaAM directs object presence recognition across regions, buffering linguistic noise and reducing time complexity.
- We design loss functions at both the region and pixel levels to improve the accuracy of target presence detection and enhance focus on boundary areas.
- Experimental results show RaAM-RVG surpasses state-of-the-art methods across tasks with lower computational cost. The performance gains when integrating RaAM into other models further confirm its broad applicability.

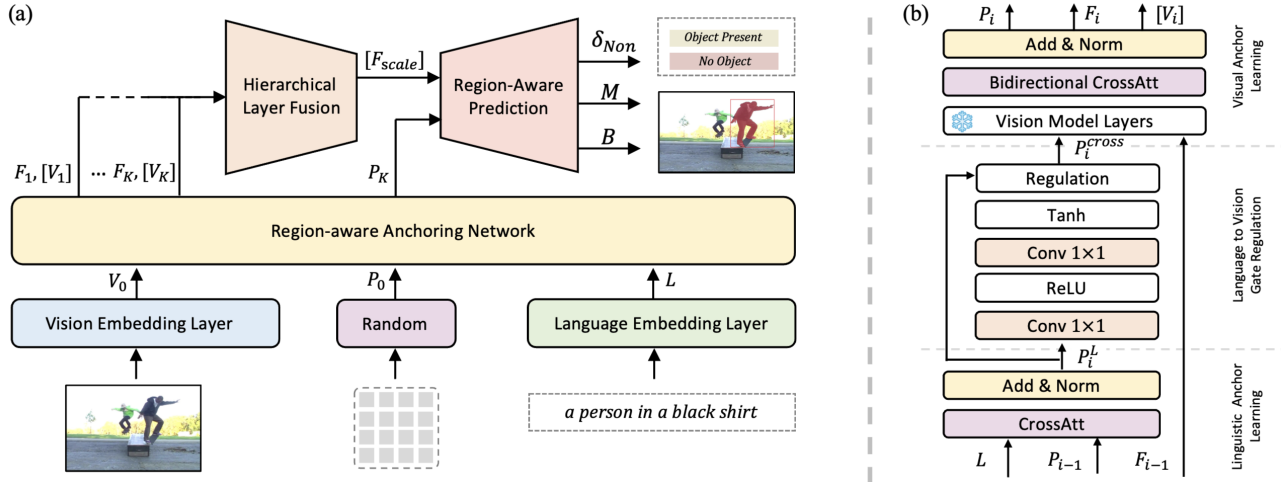


Figure 3. An illustration of RaAM-RVG. (a) Initially, the input image and language expression undergo embedding layers to obtain the visual feature V_0 and linguistic feature L , which are then utilized for the subsequent region-aware anchoring interactions. In addition, we randomly initialize a learnable region-aware anchor P_0 . Subsequently, Region-aware Anchoring Network facilitates interactions between the anchors and linguistic features, as well as visual features. This process yields anchoring fusion features and visual features $F_i, [V_i], i \in \{1, \dots, K\}$ at various stages. These features are then passed to the hierarchical layer fusion block to derive the fused features $[F_{scale}]$, which consists of $[F_{scale}^1, \dots, F_{scale}^K]$. Finally, $[F_{scale}]$ and the final region-aware anchor P_K are integrated for region-aware prediction. (b) illustrates the process of the Region-aware Anchoring Network at each stage, alternating between Linguistic Anchor Learning and Visual Anchor Learning. An L2V gate is introduced to assist in regulating the inter-modal information flow.

2. Related Works

Referring Visual Grounding. The Referring Visual Grounding task comprises Referring Expression Comprehension (REC), Referring Expression Segmentation (RES) and Generalized Referring Expression Segmentation (GRES). The goals of REC and RES are to locate the specific object mentioned in a given language expression in an image and to output the corresponding bounding box or segmentation mask. Early REC approaches [13, 26, 50] often followed a two-stage paradigm. Subsequently, one-stage methods [22, 29, 52] were applied to REC. In recent years, Transformer-based methods TransVG [7] and TRAR [61] have demonstrated promising performance in REC. On the other hand, RES methods typically employ sophisticated attention mechanisms [9, 12] to facilitate the fusion of visual and linguistic information. LAVT [54] and SLViT [33] have incorporated multi-scale architecture designs. Recent studies [49, 62] treat REC and RES as unified point prediction problems, while others [21, 25, 29, 40] propose a multi-task collaborative learning framework to unify REC and RES. To extend RES to more complex real-world scenarios, the GRES task and gRefCOCO dataset [23] have been introduced. GRES research builds on advancements made in traditional RES. For instance, ReLA [23] employed a weight matrix to aggregate masks generated by standard mask classification models, achieving competitive results on the gRefCOCO dataset. Additionally, the success of large language models has driven new advancements in RVG [4, 47],

broadening the potential of GRES in diverse applications.

Vision Language Model. Early vision language models like CLIP aligned images and text, enabling zero-shot classification [35]. Leveraging the Transformer’s strength in both domains, researchers have explored it as a unified vision-language framework [3, 28, 41]. Recently, Transformer-based methods for visual grounding have emerged [21, 33, 40], including EEVG [2], which designs an efficient framework for multi-task grounding with reduced computational cost. Multimodal Large Language Models (MLLMs) provide a universal interface for various tasks, with models aligning image-text features before fine-tuning through instruction [6, 24, 63]. Research like BuboGPT [60] and Shikra [1] highlights MLLMs’ capabilities in fine-grained image understanding and open-world visual grounding. Recent studies have incorporated learnable tokens to facilitate multimodal comprehension, some enhancing unimodal knowledge learning and others introducing additional priors during interaction [14, 43, 59]. The key distinction of RaAM lies in its anchoring designs, which act as cross-modal intermediaries during feature learning and provide explicit region-level object existence guidance.

3. Method

3.1. Overview

The proposed RaAM employs learnable region-aware anchors to guide vision-language interactions, where these an-

chors act as indicators of object presence across various regions in the visual environment. These anchors alternately interact with linguistic and visual features, progressively learning object presence within each region and guiding the refinement of visual feature representations. By incorporating RaAM for vision-language interaction, we construct the RVG model RaAM-RVG. This model regulates multimodal information flow through RaAM, effectively reducing cross-modal redundancy while preserving essential visual details. The illustration is shown in Fig. 3.

Given an image and the reference language expression, our model generates the bounding box or the segmentation mask for specified objects. The input image undergoes the vision embedding layer to obtain $V_0 \in \mathbb{R}^{C_{v_0} \times H_0 \times W_0}$, where H_0 and W_0 are height and width of the feature, and C_{v_0} represents the number of channels. The input language expression is processed through the language embedding layer to obtain $L \in \mathbb{R}^{T \times C_l}$, where T is the length of the feature, C_l represents the number of channels. These embedding layers are implemented using pretrained vision and language models. Subsequently, V_0 , L and initialized region-aware anchors P_0 (Sec. 3.2) are sent to the Region-aware Anchoring Network, which employs a hierarchical architecture with K stages. The anchors interact alternately with linguistic features (Sec. 3.3) and visual features (Sec. 3.4) to obtain optimal region-aware anchor P_K , anchoring fusion features F_i and visual features $[V_i]$, $i \in \{1, \dots, K\}$. Finally, region-aware anchors are integrated with features from each stage for comprehensive processing and prediction (Sec. 3.5).

3.2. Region-aware Anchor Generation

We randomly initialize a trainable tensor $P_i \in \mathbb{R}^{N \times C_l}$ as the region-aware anchor for i -th stage, where N is the length of each anchor, corresponding to N regions within the image. To flexibly incorporate cross-modal interaction at shallow layers and reinforce the cross-modal consistency in deeper layers, we embed the initialized region-aware anchor only in the first J stages (including the initialization before the Region-aware Anchoring Network). The region-aware anchor for stages J to K is derived from the feedback of the preceding stage.

3.3. Linguistic Anchor Learning

Linguistic Interaction. At i -th stage of the Region-aware Anchoring Network, we employ a cross-attention to help the region-aware anchor P_{i-1} learn linguistic knowledge from linguistic feature L . The steps to obtain the activation $Att_i^L \in \mathbb{R}^{N \times T}$ are as follows:

$$Att_i^L = \frac{\omega_{p1}(P_i) \omega_{l1}(L)^\top}{\sqrt{C_l}}, \quad (1)$$

where ω_{p1} , ω_{l1} are projection functions, and ω_{p1} , ω_{l1} is implemented as a 1×1 convolution. Both ω_{p1} and ω_{l1} yield

channels of size C_l . We employ the activation Att_i^L to activate the anchor. The process of obtaining $P_i^L \in \mathbb{R}^{N \times C_l}$ is defined as:

$$P_i^L = \text{Norm}(P_{i-1} + \text{Softmax}(Att_i^L) \omega_{l2}(L)), \quad (2)$$

where ω_{l2} indicates the projection function same as ω_{l1} , and $\text{Norm}(\cdot)$ denotes layer normalization.

Language to Vision Gate Regulation. To regulate the transfer of linguistic knowledge to the vision modality, we introduce a computationally efficient gate unit, referred to as Language to Vision (L2V). The core function of L2V gate is to restrict the flow of linguistic noise. L2V gate learns weight mappings from P_i^L , dynamically rescaling each element in an adaptive manner. The formulation for obtaining $P_i^{cross} \in \mathbb{R}^{N \times C_{v_i}}$ is defined as:

$$P_i^{cross} = \text{Linear}(\gamma(P_i^L) \odot P_i^L + P_{i-1}), \quad (3)$$

where \odot is element-wise matrix multiplication operation, $\text{Linear}(\cdot)$ represents a linear transformation that adjusts the number of channels to C_{v_i} , $\gamma(\cdot)$ is a two-layer perception consisting of sequential 1×1 convolution, ReLU activation, another 1×1 convolution, and Tanh activation.

3.4. Visual Anchor Learning

Given that P_i^{cross} encapsulates knowledge from the language modality, we engage in Visual Anchor Learning on the region-aware anchors and visual features to attain deeply fused multimodal representations. Except for the first stage, F_i^1 is obtained from the anchoring fusion feature F_{i-1} of the preceding stage via pretrained vision model layers $Vision_i(\cdot)$ of the current stage. In this process, $[V_i] = Vision_i(F_{i-1})$, $[V_i]$ denotes the visual features output by all layers within the current stage, and $F_i^1 \in \mathbb{R}^{C_{v_i} \times H_i \times W_i}$ is the final layer's output. Features $[V_i]$ are subsequently used for Hierarchical Layer Fusion, as detailed in Sec. 3.5.

Visual Interaction. We employ a bidirectional cross-attention mechanism to model the global relationship between the anchor P_i^{cross} and visual feature F_i^1 . The steps to obtain the activation $Att_i \in \mathbb{R}^{N \times H_i \times W_i}$ are as follows:

$$Att_i^V = \frac{\omega_{p2}(P_i^{cross}) \text{Flatten}(\omega_{v1}(F_i^1))}{\sqrt{C_{v_i}}}, \quad (4)$$

where ω_{p2} indicates the projection function same as ω_{p1} , ω_{v1} indicates the projection function defined as a 1×1 convolution and an instance normalization, and $\text{Flatten}(\cdot)$ denotes the operation of flattening the two spatial dimensions into a single dimension along the rows. Both ω_{p2} and ω_{v1} yield channels of size C_{v_i} . We employ the activation Att_i^V to activate visual features and region-aware anchors, facilitating bidirectional promotion of visual knowledge and object presence awareness. The process of obtaining anchoring fusion feature $F_i \in \mathbb{R}^{C_{v_i} \times H_i \times W_i}$ and $P_i^V \in \mathbb{R}^{N \times C_{v_i}}$ is

Method	Backbone	Multi-task	RefCOCO			RefCOCO+			G-Ref	
			val	test A	test B	val	test A	test B	val(U)	test(U)
MAttNet [57]	MRCNN-Res101	✓	76.65	81.14	69.99	65.33	71.62	56.02	66.58	67.27
NMTree [16]	MRCNN-Res101	✗	76.41	81.21	70.09	66.46	72.02	57.52	65.87	66.44
LBYL [15]	DarkNet53	✗	79.67	82.91	74.15	68.64	73.38	59.49	-	-
MCN [29]	DarkNet53	✓	80.08	82.29	74.98	67.16	72.86	57.31	66.46	66.01
TransVG [7]	ResNet101	✗	81.02	82.72	78.35	64.82	70.70	56.94	68.67	67.73
TRAR [61]	DarkNet53	✗	-	81.40	78.60	-	69.10	56.10	68.90	68.30
SeqTR [62]	DarkNet53	✓	81.23	85.00	76.08	68.82	75.37	58.78	71.35	71.58
PVD [5]	DarkNet53	✓	82.51	86.19	76.81	69.48	76.83	59.68	68.40	69.57
PVD [5]	Swin-B	✓	84.52	87.64	79.63	73.89	78.41	64.25	73.81	74.13
VG-LAW [40]	ViT-B	✓	86.62	89.32	83.16	76.37	81.04	67.50	76.90	76.96
EEVG [2]	ViT-B	✓	88.08	90.33	85.50	77.97	82.44	69.15	79.60	80.24
UniTAB [53]	Res101	✓	88.59	91.06	83.75	80.97	85.36	71.55	84.58	84.70
PolyFormer [25]	Swin-B	✓	90.38	92.89	87.16	84.98	89.77	77.97	85.83	85.91
RaAM-RVG	Swin-B	✓	91.07	93.13	88.34	85.02	89.79	78.69	86.31	87.26
RaAM-RVG	ViT-B	✓	91.45	93.42	88.71	84.78	89.23	79.24	86.78	87.71

Table 1. Comparison with state-of-the-art methods for REC in terms of Precision@0.5 on three benchmark datasets. U: The UMD partition.

outlined by the following equations:

$$P_i^V = \alpha_1(\text{Softmax}(Att_i^V)\text{Flatten}(\omega_{v2}(V_i))^\top), \quad (5)$$

$$F_i^2 = \text{Unflatten}(\omega_{p3}(P_i^{\text{cross}}))^\top \text{Softmax}(Att_i^V), \quad (6)$$

$$F_i = \text{Norm}(F_i^1 + \alpha_2 F_i^2), \quad (7)$$

where $\text{Unflatten}(\cdot)$ indicates the opposite operation of $\text{Flatten}(\cdot)$, α_1 and α_2 represent learnable coefficients. ω_{v2} and ω_{p3} indicate projection functions same as ω_{v1} and ω_{p2} , respectively.

Propagation of Linguistic Knowledge. The region-aware anchor, interacting with visual features, combines with the P_i^{cross} and is propagated to the next stage. The region-aware anchor $P_i \in \mathbb{R}^{N \times C_{v_i}}$ for the next stage is obtained by the following:

$$P_i = \text{Norm}(P_i^{\text{cross}} + \omega_{p4}(P_i^{\text{cross}})), \quad (8)$$

where $\text{Norm}(\cdot)$ represents a normalization layer, ω_{p4} indicates the function same as ω_{p3} .

3.5. Knowledge Integration and Prediction

Hierarchical Layer Fusion. To integrate visual features from different layers without incurring additional computational costs, we design a straightforward cross-level fusion operation. In our method, the vision backbone is divided into K stages, with the i -th stage comprising Q_i layers. The computation of the fused feature F_{scale}^i for the i -th stage is detailed as the following:

$$F_{scale}^i = \frac{1}{2}(F_i + \frac{1}{Q_i} \sum_{j=1}^{Q_i} V_{i,j}), \quad (9)$$

where $V_{i,j}$ denotes the intermediate feature from $[V_i]$, obtained at the j -th layer within the i -th stage.

Region-aware Prediction. We partition the fused features $[F_{scale}]$ into N regional features and split the anchors into corresponding N anchor representations, denoted as $X_n^i = \text{Split}[F_{scale}^i]$ and $P_n = \text{Split}[P_K]$, where n denotes the n -th region. Within each region, the region-aware anchor interacts with each regional feature through element-wise multiplication, yielding the final regional feature R_n for prediction. The process is formulated as the following:

$$R_n = \text{Concat}[P_n \odot X_n^2, \dots, P_n \odot X_n^K]. \quad (10)$$

The feature from first stage contains an excess of low-level information, which can negatively impact prediction outcomes; therefore, they are excluded from use. We employ separate MLP prediction heads for each region to predict object presence, producing δ_{Non}^n . When δ_{Non}^n is '0' for every region, the model determines that no object matching the language expression. The regional features R_n are concatenated to form R , which is then processed by MLP prediction heads to output the detection results B and segmentation results M . This section focuses on vision backbones such as ViT [11], where the visual feature dimensions remain consistent across all stages. Implementation details for more backbones can be found in the Appendix.

3.6. Multi-task Training

Our method is tailored for RVG, including REC, RES, and GRES. Following previous work [2, 25, 40], we employ a multi-task joint training approach for REC and RES tasks. We design loss functions at both the region and pixel levels to enhance performance across these tasks.

Method	Backbone	Multi-task	RefCOCO			RefCOCO+			G-Ref	
			val	test A	test B	val	test A	test B	val(U)	test(U)
MAttNet [57]	MRCNN-Res101	✗	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61
NMTree [16]	MRCNN-Res101	✗	56.59	63.02	52.06	47.40	53.01	41.56	46.59	47.88
MCN [29]	DarkNet53	✓	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
CRIS [44]	CLIP-ResNet50	✗	69.52	72.72	64.70	61.39	67.10	52.48	59.87	60.36
SeqTR [62]	DarkNet53	✓	67.26	69.79	64.12	54.14	58.93	48.19	55.67	55.64
PVD [5]	DarkNet53	✓	68.87	70.53	65.83	54.98	60.12	50.23	57.81	57.17
LAVT [54]	Swin-B	✗	74.46	76.89	70.94	65.81	70.97	59.23	63.34	63.62
PVD [5]	Swin-B	✓	74.82	77.11	69.52	63.38	68.60	56.92	63.13	63.62
VG-LAW [40]	ViT-B	✓	75.62	77.51	72.89	66.63	70.38	59.89	65.53	66.08
SLViT [33]	SegNext-B	✗	74.02	76.91	70.62	64.07	69.28	56.14	62.75	63.57
PolyFormer [25]	Swin-B	✓	75.96	78.29	73.25	69.33	74.56	61.87	69.20	70.19
P-RIS [39]	ViT-B	✗	76.36	80.37	72.29	67.06	73.58	58.96	64.79	67.16 s
EEVG [2])	ViT-B	✓	78.23	79.27	76.58	69.04	72.65	62.33	69.15	70.01
RaAM-RVG	Swin-B	✓	79.03	80.98	77.30	69.51	75.24	62.77	70.95	71.86
RaAM-RVG	ViT-B	✓	79.35	81.22	77.81	69.54	75.69	63.02	71.30	72.09

Table 2. Comparison with state-of-the-art methods for RES in terms of overall IoU on three benchmark datasets. U: The UMD partition.

Region Level. At the region level, to optimize the representation of target presence within region-aware anchors and enable the model to accurately determine target existence, we employ the following loss function to constrain δ_{Non} :

$$\mathcal{L}_r = \sum_{n=1}^N \mathcal{L}_{ce}(\delta_{Non}^n, \delta_{Non}^{n,gt}), \quad (11)$$

where $\mathcal{L}_{ce}(\cdot)$ denotes the cross-entropy loss.

Pixel Level. At the pixel level, to enhance the accuracy of object boundary discrimination, we increase the loss weight for uncertain regions along object edges. Specifically, we perform dilation and erosion operations on the ground truth label M^{gt} to obtain the boundary region E , defined as $E = dilate(M^{gt}) - erode(M^{gt})$. Here, $dilate(\cdot)$ expands and $erode(\cdot)$ contracts the object’s boundaries. A weight β is then assigned to the edge region where $E^i = 1$. The loss function for the segmentation mask is defined as:

$$\mathcal{L}_p^{seg} = \mathcal{L}_{focal}(M, M^{gt}) + W \odot \mathcal{L}_{dice}(M, M^{gt}), \quad (12)$$

where $W^i = \beta E^i + (1 - E^i)$, \mathcal{L}_{focal} and \mathcal{L}_{dice} are focal loss [38] and dice loss [31]. The loss function employed for the detection task is defined as follows:

$$\mathcal{L}_p^{det} = \mathcal{L}_{smooth-L_1}(B, B^{gt}) + \mathcal{L}_{giou}(B, B^{gt}), \quad (13)$$

where $\mathcal{L}_{smooth-L_1}$ and \mathcal{L}_{giou} are the smooth L1 loss and GIoU loss [37].

Training Loss. Our joint training loss for REC and RES is defined as follows:

$$\mathcal{L}_{Joint} = \lambda_1 \mathcal{L}_r + \lambda_2 \mathcal{L}_p^{det} + \lambda_3 \mathcal{L}_p^{seg}. \quad (14)$$

Methods	Val		TestA		TestB	
	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU
MattNet [57]	47.51	48.24	58.66	59.30	45.33	46.14
LTS [17]	52.30	52.70	61.87	62.64	49.96	50.42
VLT [9]	52.51	52.00	62.19	63.20	50.52	50.88
CRIS [44]	55.34	56.27	63.82	63.42	51.04	51.79
LAVT [54]	57.64	58.40	65.32	65.90	55.04	55.83
CGFormer [34]	62.28	63.01	68.15	70.13	60.18	61.09
ReLA [23]	64.20	65.50	70.78	70.89	60.97	61.05
RaAM-RVG	67.35	70.02	72.98	73.86	64.34	65.77

Table 3. Performance comparison of different methods on GRES task in terms of cIoU and gIoU.

The training loss for GRES is defined as follows:

$$\mathcal{L}_{GRES} = \lambda_4 \mathcal{L}_r + \lambda_5 \mathcal{L}_p^{seg}, \quad (15)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5 are hyperparameters.

4. Experiments

4.1. Dataset and Evaluation

We perform experiments on four widely used benchmark datasets for RVG. The datasets for REC and RES include RefCOCO [56], RefCOCO+ [56], G-Ref [30, 32]. They have 19,994, 19,992, and 26,711 images respectively, containing 50,000, 49,856, and 54,822 references and 142,209, 141,564, and 104,560 reference expressions. For GRES task, we use the gRefCOCO [23], which consists of 278,232 expressions, including 80,022 multi-object and 32,202 no-target samples.

Following existing works [2, 23, 62], we use different evaluation metrics for tasks. For REC, we use Precision at Intersection over Union threshold of 0.5 (Precision@0.5) as

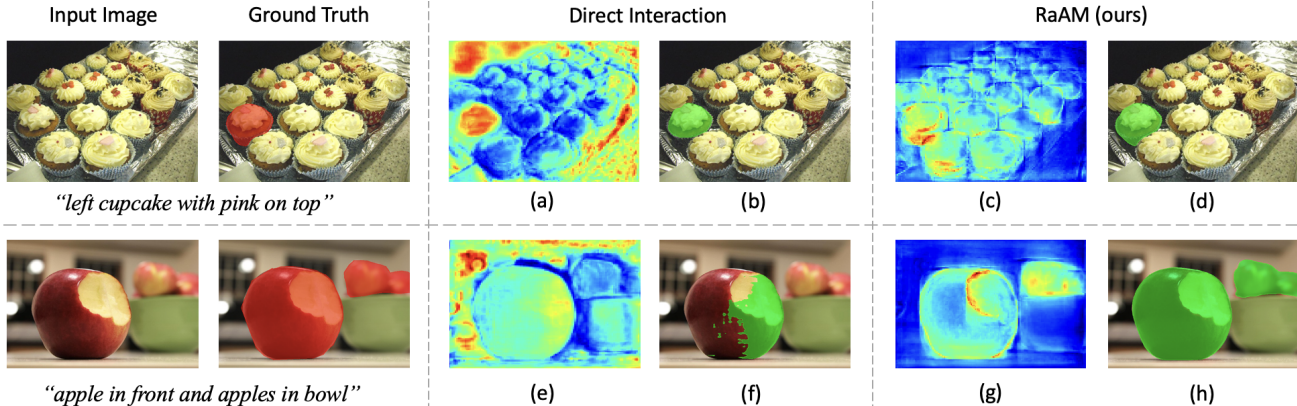


Figure 4. Qualitative results. Each row presents visualizations of feature maps and predicted masks obtained from different methods for the same input. The Direct Interaction method is obtained by replacing the interaction design in RaAM-RVG with that of LAVT [54].

the evaluation metric, measuring how often the predicted bounding box overlaps with the ground truth by at least 50%. For RES, the evaluation metric is Object Intersection over Union (oIoU), which quantifies the overlap between the predicted segmentation mask and the ground truth mask. For GRES, we employ Conditional Intersection over Union (cIoU) and Generalized Intersection over Union (gIoU) to assess the quality of the predicted segmentation, capturing standard overlap and more nuanced spatial alignment between predicted and ground truth regions.

4.2. Implementation Details

We use Vision Transformer-Base [11] and BERT [8, 46] as the default vision and language backbones for RaAM, enabling feature extraction in both the embedding layer and Region-aware Anchoring Network. In RaAM, we set $K = 4$ and $J = 2$, dividing the vision backbone into four stages, with region-aware anchor initialized in the first two stages. We set the default values for key hyperparameters as follows: $\beta = 1.2$, $N = 16$. Additional hyperparameter settings and experiments are detailed in the Appendix.

We use the AdamW optimizer with a weight decay of 0.05. The initial learning rate is $2e-5$, scheduled with polynomial decay (power 0.9). Models are trained for 60 epochs with a batch size of 16. Each reference typically includes 2-3 sentences, and one referring expression per object is randomly selected per epoch. Training is conducted on 2 Nvidia RTX A6000 GPUs.

4.3. Comparison with the State-of-the-Arts

We compare the performance of our proposed RaAM with state-of-the-art (SOTA) methods on four widely-used benchmarks. The tables highlight the best scores in bold, facilitating a straightforward comparison.

REC and RES. Following prior research, REC and RES are treated as multi-task visual grounding tasks [2]. We conduct joint training for these two tasks and evaluate the

Method	Extra Training Data	cIoU	gIoU
RaAM-RVG	✗	67.35	70.02
LAVT [54]	✗	57.64	58.40
w/ RaAM	✗	61.45	62.87
PolyFormer [25]	✗	59.35	61.84
w/ RaAM	✗	63.79	64.47
HyperSeg [45]	✓	62.75	64.48
w/ RaAM	✓	66.68	68.75
GSVA [47]	✓	66.38	70.04
w/ RaAM	✓	68.02	71.53

Table 4. The experiments for the applicability on GRES.

effectiveness of RaAM using the RefCOCO, RefCOCO+, and G-Ref datasets. As shown in Tab. 1 and Tab. 2, RaAM achieves superior performance across all datasets.

GRES. The performance comparison for the GRES task is presented in Tab. 3. RaAM demonstrates strong capability in handling more complex scenarios, achieving SOTA performance across all test sets. Additional results and analyses are provided in the Appendix.

4.4. Applicability Verification

To validate the applicability of our RaAM, we compared it with advanced models: the expert models LAVT [54] and PolyFormer [25] for RVG, and MLLMs HyperSeg [45] and GSVA [47], which use extra training data. For LAVT and PolyFormer, we replaced their cross-modal interaction designs with RaAM and retrained them. For HyperSeg and GSVA, we incorporated RaAM into their vision encoders and fine-tuned them on the gRefCOCO dataset. Evaluation on the gRefCOCO Val set (Tab. 4) shows that RaAM significantly improves performance, demonstrating its strong applicability. In the GRES task, including multi-object and no-object scenes, MLLMs excel due to their powerful comprehension and large training datasets. Our RaAM offers two key advantages: efficiency and broad applicability. It

				cIoU	gIoU
(a) Effectiveness of interaction design					
Direct Interaction				64.77	67.62
RaAM				67.35	70.02
(b) Ablation on design choices of flow control					
α_1	α_2	Linear	L2V		
		✓		63.87	66.46
✓		✓		64.51	67.25
	✓	✓		64.94	67.48
✓	✓	✓		65.74	68.45
✓	✓		✓	67.35	70.02
(c) Stages for anchor generation					
Stage0	Stage1	Stage2	Stage3		
✓				66.38	68.85
✓	✓			67.35	70.02
✓	✓	✓		64.58	66.82
✓	✓	✓	✓	60.22	63.07
(d) Ablation on loss function selection					
	\mathcal{L}_r		\mathcal{L}_p^{seg}		
				65.09	67.63
	✓			65.94	68.70
	✓		✓	67.35	70.02

Table 5. Ablation studies on the gRefCOCO Val set on GRES.

performs well with smaller datasets, and can further enhance the performance of pre-trained MLLMs on RVG.

4.5. Ablation Studies

Effectiveness of the architecture of RaAM. We conducted an ablation study on the gRefCOCO Val set using the GRES task as an example. To validate the effectiveness of the proposed interaction mechanism *RaAM*, we compare the two architectures shown in Fig. 1. In Tab. 5(a), *Direct Interaction* refers to a variant where the vision-language interaction design in *RaAM-RVG* is replaced with the interaction mechanism from LAVT [54]. The results show the advantages of the proposed *Region-aware Anchoring Interaction* strategy.

Ablation on design choices of flow control. We conducted an ablation study on the design of flow control in the Region-aware Anchoring Network. The outcomes are presented in Tab. 5(b). α_1 and α_2 represent the learnable coefficients within Visual Anchor Learning. *Linear* denotes the straightforward linear transformation operation to transmit the anchor to vision modality. *L2V* flexibly regulating the flow of information from language to vision modalities. Upon examination of Tab. 5(b), it is evident that each component for flow control contributes to the performance. Additionally, Tab. 6 shows that L2V significantly improves the metrics at various thresholds, especially at Pr@0.9. This improvement highlights its ability to enhance boundary discrimination by controlling cross-modal redundancy.

Methods	Pr@0.9	Pr@0.8	Pr@0.7	cIoU	gIoU
CGFormer	22.43	56.57	68.93	62.28	63.01
ReLA	23.56	57.01	69.15	64.20	65.50
RaAM-RVG (w/o L2V)	25.86	57.52	70.22	65.74	68.45
RaAM-RVG	27.43	58.97	71.24	67.35	70.02

Table 6. Comparison of precision results on GRES.

Stages for anchor generation. We explore the number of stages J for anchor generation. As shown in Tab. 5(c), the accuracy exhibits a slight increase with the augmentation of anchor generation stages, reaching a peak beyond which a notable decline is observed, particularly after surpassing 2 stages. We adopt $J = 2$ as the default configuration.

Ablation on loss function selection. As shown in Tab. 5(d), the combination of \mathcal{L}_r and \mathcal{L}_p^{seg} losses yields the best performance. Therefore, both loss functions are adopted.

4.6. Visualization Analysis

In Fig. 4, we exemplify the segmentation results and feature maps derived from pairs of RES / GRES inputs. In the first example, comparing the feature maps, (c) demonstrates focused attention on the target “cupcake,” whereas (a) also attends to irrelevant background regions. While the *Direct Interaction* method accurately identifies the correct “cupcake,” its segmentation output lacks the lower part of the object compared to RaAM’s result, indicating that RaAM more effectively captures visual details and accurately delineates object boundaries. In the second example, the *Direct Interaction* method fails to focus on the correct target in the feature map (e), resulting in incomplete segmentation. In contrast, RaAM successfully attends to two key regions within the image and produces correct segmentation outputs, highlighting its superior multi-object localization capability over the *Direct Interaction* strategy.

5. Conclusion

In this work, we present RaAM-RVG, incorporating our novel interaction mechanism, RaAM, which leverages region-aware anchors for guiding vision-language interactions. RaAM effectively directs object presence recognition across regions, reduces cross-modal redundancy, and lowers time complexity. Additionally, we introduce loss functions at region and pixel levels to enhance accuracy. Comprehensive experiments show that RaAM-RVG outperforms state-of-the-art methods across RVG tasks with lower cost. The observed performance gains when integrating RaAM into other models further confirm its broad applicability.

Acknowledgements

This work was supported in part by the National Key Research and Development Project (No. 2022YFC2504605).

References

- [1] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multi-modal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 3
- [2] Wei Chen, Long Chen, and Yu Wu. An efficient and effective transformer decoder-based framework for multi-task visual grounding. *arXiv preprint arXiv:2408.01120*, 2024. 3, 5, 6, 7
- [3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 3
- [4] Shupeng Cheng, Ge-Peng Ji, Pengda Qin, Deng-Ping Fan, Bowen Zhou, and Peng Xu. Large model based referring camouflaged object detection. *arXiv preprint arXiv:2311.17122*, 2023. 3
- [5] Zesen Cheng, Kehan Li, Peng Jin, Siheng Li, Xiangyang Ji, Li Yuan, Chang Liu, and Jie Chen. Parallel vertex diffusion for unified visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1326–1334, 2024. 5, 6
- [6] W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*. 3
- [7] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *ICCV*, pages 1769–1779, 2021. 3, 5
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 7
- [9] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, pages 16321–16330, 2021. 3, 6
- [10] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vlt: Vision-language transformer and query generation for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5, 7, 1
- [12] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *ICCV*, pages 15506–15515, 2021. 3
- [13] Richang Hong, Daqing Liu, Xiaoyu Mo, Xiangnan He, and Hanwang Zhang. Learning to compose and reason with language tree structures for visual grounding. *IEEE transactions on pattern analysis and machine intelligence*, 44(2): 684–696, 2019. 1, 3
- [14] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2, 3
- [15] Binbin Huang, Dongze Lian, Weixin Luo, and Shenghua Gao. Look before you leap: Learning landmark features for one-stage visual grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16888–16897, 2021. 5
- [16] Tianrui Hui, Si Liu, Shaofei Huang, Guanbin Li, Sansi Yu, Faxi Zhang, and Jizhong Han. Linguistic structure guided context modeling for referring image segmentation. In *ECCV*, pages 59–75. Springer, 2020. 5, 6
- [17] Ya Jing, Tao Kong, Wei Wang, Liang Wang, Lei Li, and Tieniu Tan. Locate then segment: A strong pipeline for referring image segmentation. In *CVPR*, pages 9858–9867, 2021. 6
- [18] Seoyeon Kim, Minguk Kang, Dongwon Kim, Jaesik Park, and Suha Kwak. Extending clip’s image-text alignment to referring image segmentation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4611–4628, 2024. 2
- [19] 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. 2
- [20] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 2
- [21] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664, 2021. 3
- [22] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *CVPR*, pages 10880–10889, 2020. 3
- [23] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *CVPR*, pages 23592–23601, 2023. 1, 2, 3, 6
- [24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 3
- [25] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *CVPR*, pages 18653–18663, 2023. 3, 5, 6, 7, 1, 2
- [26] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *CVPR*, pages 1950–1959, 2019. 1, 3
- [27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer:

- Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 1
- [28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3
- [29] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020. 3, 5, 6
- [30] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 6
- [31] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 6
- [32] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, pages 792–807. Springer, 2016. 6
- [33] Shuyi Ouyang, Hongyi Wang, Shiao Xie, Ziwei Niu, Ruofeng Tong, Yen-Wei Chen, and Lanfen Lin. Slvit: scale-wise language-guided vision transformer for referring image segmentation. In *IJCAI*, pages 1294–1302, 2023. 2, 3, 6
- [34] Weize Quan, Pengfei Deng, Kai Wang, and Dong-Ming Yan. Cgformer: Vit-based network for identifying computer-generated images with token labeling. *IEEE Transactions on Information Forensics and Security*, 2023. 6, 2
- [35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 3
- [36] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 2
- [37] Hamid Rezaatofghi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 6
- [38] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017. 6
- [39] Chao Shang, Zichen Song, Heqian Qiu, Lanxiao Wang, Fanman Meng, and Hongliang Li. Prompt-driven referring image segmentation with instance contrasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4124–4134, 2024. 2, 6
- [40] Wei Su, Peihan Miao, Huanzhang Dou, Gaoang Wang, Liang Qiao, Zheyang Li, and Xi Li. Language adaptive weight generation for multi-task visual grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10857–10866, 2023. 3, 5, 6
- [41] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR, 2022. 3
- [42] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *arXiv preprint arXiv:2305.11175*, 2023. 2
- [43] Yikai Wang, TengQi Ye, Lele Cao, Wenbing Huang, Fuchun Sun, Fengxiang He, and Dacheng Tao. Bridged transformer for vision and point cloud 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12114–12123, 2022. 2, 3
- [44] Zhaoping Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, pages 11686–11695, 2022. 6
- [45] Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Zheng Zhao, Jie Hu, and Yujiu Yang. Hyperseg: Towards universal visual segmentation with large language model. *arXiv preprint arXiv:2411.17606*, 2024. 7
- [46] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *EMNLP*, pages 38–45, 2020. 7
- [47] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3858–3869, 2024. 3, 7
- [48] Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xi-ang 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. 2
- [49] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, pages 15325–15336, 2023. 3
- [50] Sibe Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *ICCV*, pages 4644–4653, 2019. 1, 3
- [51] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *CVPR*, pages 4683–4693, 2019. 1
- [52] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive

- sub-query construction. In *ECCV*, pages 387–404. Springer, 2020. [3](#)
- [53] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Unitab: Unifying text and box outputs for grounded vision-language modeling. In *ECCV*, pages 521–539. Springer, 2022. [5](#)
- [54] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, pages 18155–18165, 2022. [1](#), [2](#), [3](#), [6](#), [7](#), [8](#)
- [55] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *CVPR*, pages 10502–10511, 2019.
- [56] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85. Springer, 2016. [1](#), [6](#)
- [57] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018. [5](#), [6](#), [2](#)
- [58] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *Advances in neural information processing systems*, 37:71737–71767, 2024. [2](#)
- [59] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024. [2](#), [3](#)
- [60] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023. [3](#)
- [61] Yiyi Zhou, Tianhe Ren, Chaoyang Zhu, Xiaoshuai Sun, Jianzhuang Liu, Xinghao Ding, Mingliang Xu, and Rongrong Ji. Trar: Routing the attention spans in transformer for visual question answering. In *ICCV*, pages 2074–2084, 2021. [3](#), [5](#)
- [62] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Lijuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *ECCV*, pages 598–615. Springer, 2022. [1](#), [3](#), [5](#), [6](#)
- [63] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. [3](#)