This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DeRIS: Decoupling Perception and Cognition for Enhanced Referring Image Segmentation through Loopback Synergy

Ming Dai[1] Wenxuan Cheng[1] Jiang-jiang Liu[2] Sen Yang[2]
Wenxiao Cai[3] Yanpeng Sun[4] Wankou Yang[1*]
[1]Southeast University [2]Baidu VIS [3] Standford University [4] National University of Singapore

Abstract

Referring Image Segmentation (RIS) is a challenging task that aims to segment objects in an image based on natural language expressions. While prior studies have predominantly concentrated on improving vision-language interactions and achieving fine-grained localization, a systematic analysis of the fundamental bottlenecks in existing RIS frameworks remains underexplored. To bridge this gap, we propose ***DeRIS****, a novel framework that decomposes RIS into two key components: perception and cognition. This modular decomposition facilitates a systematic analysis of the primary bottlenecks impeding RIS performance. Our findings reveal that the predominant limitation lies not in perceptual deficiencies, but in the insufficient multi-modal cognitive capacity of current models. To mitigate this, we propose a Loopback Synergy mechanism, which enhances the synergy between the perception and cognition modules, thereby enabling precise segmentation while simultaneously improving robust image-text comprehension. Additionally, we analyze and introduce a simple non-referent sample conversion data augmentation to address the long-tail distribution issue related to target existence judgement in general scenarios. Notably,* ***DeRIS*** *demonstrates inherent adaptability to both non- and multi-referents scenarios without requiring specialized architectural modifications, enhancing its general applicability. The codes and models are available at* `https://github.com/Dmmm1997/DeRIS`

1. Introduction

Referring Image Segmentation (RIS) [23, 37, 40] is a fundamental yet challenging task in multi-modal understanding. Unlike traditional segmentation tasks [9, 21, 34], which are restricted to predefined and fixed object categories, RIS offers greater flexibility by enabling object segmentation based on free-form textual descriptions. This requires a deeper alignment between diverse linguistic expressions and the visual context. Early RIS tasks [50, 51, 70] typically assumed a one-to-one correspondence between language expressions and target objects. To improve generalization, recent studies [24, 42, 63] have introduced more generalized formulations, enabling segmentation of multiple targets or handling non-referent cases.

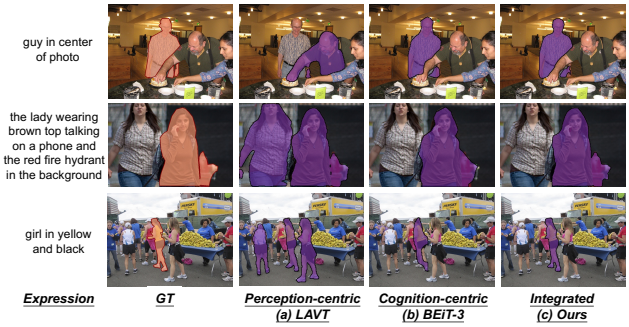

Figure 1. The visualizations of architectures with different emphases are shown in Fig. 2. (a) The *perception-centric* structure emphasizes fine-grained hierarchical features. (b) The *cognition-centric* structure comprehensively understands the image-text context. (c) Our DeRIS integrates the advantages of both.

In this paper, we focus on the core capabilities reflected by model architectures and categorize mainstream RIS methods into two types: *perception-centric* and *cognition-centric*. As illustrated in Fig. 2(a), perception-centric methods [10, 15, 45, 47, 68, 77] rely on hierarchical backbones [17, 20, 48] designed for perceptual tasks, which effectively preserve fine-grained spatial information. However, due to the limited diversity in downstream datasets, multi-modal fusion modules, when trained from scratch, often exhibit weaker content cognition capabilities. As shown in Fig. 1(a), while perception-centric methods excel at retaining fine-grained contour information, they struggle with the contextual understanding of image-text relationships. In contrast, cognition-centric methods [11, 12, 66, 72] (structure is illustrated in Fig. 2(b)) leverage large-scale vision-language pre-trained models [1, 30, 53, 61] to significantly enhance multimodal understanding and alignment capabilities. However, due to the transformer architecture em-

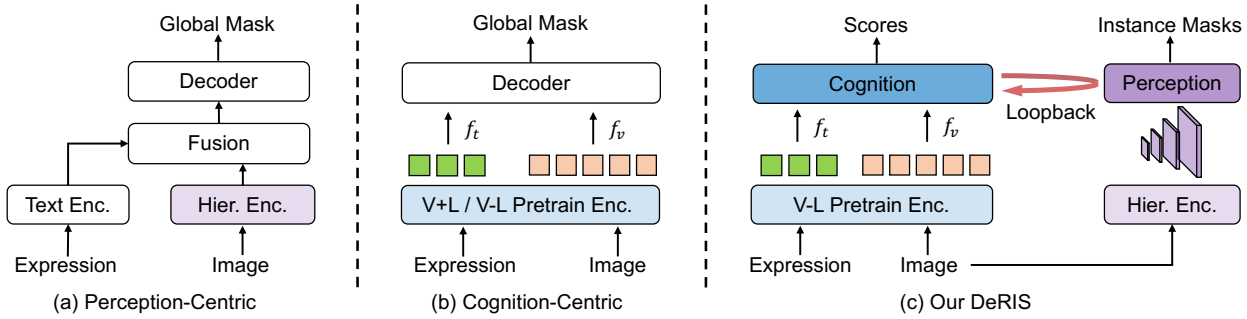

Figure 2. Comparison of architectural paradigms: (a) *Perception-centric* models rely on hierarchical encoders [17, 20, 48] to preserve fine-grained spatial features. (b) *Cognition-centric* models leverage vision-language pre-trained model to enhance multi-modal representation and alignment, where V+L refers to two-stream models [38, 53, 76], and V-L denotes one-stream models [30, 61]. (c) The proposed *DeRIS* framework, which integrates robust cognition and fine-grained perception capabilities.

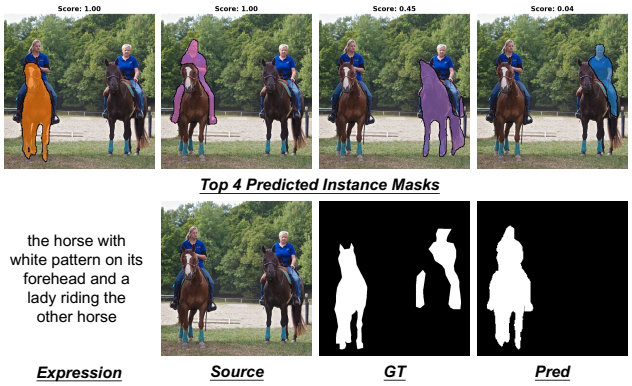

Figure 3. Visualization of strong perception with weak cognition.

ployed, which incurs quadratic computational complexity with respect to input resolution, these methods suffer from a notable limitation: a substantial loss of fine-grained spatial information. This deficiency undermines their performance in high-precision perception tasks. As shown in Fig. 1(b), while cognition-centric approaches can effectively identify corresponding regions, they struggle to delineate precise contour boundaries, highlighting a critical trade-off between comprehension and localization accuracy.

Fundamentally, we argue that RIS inherently involves two critical dimensions: *perception* and *cognition*. Perception focuses on accurately localizing foreground objects, whereas cognition entails a comprehensive understanding of both textual and visual content. Given the significance of both aspects, we propose a novel framework, **DeRIS**, which explicitly decouples perception and cognition, as illustrated in Fig. 2(c). This design effectively integrates the strengths of both perception- and cognition-centric paradigms, enabling robust multi-modal understanding while preserving fine-grained perceptual performance. As shown in Fig. 1(c), DeRIS not only maintains high-precision perception but also enhances contextual comprehension. The remainder of this paper primarily discusses the following three issues.

(1) What constitutes the primary bottleneck in RIS: perception or cognition? DeRIS decouples perception and cognition to independently analyze their contributions to RIS performance. From a quantitative perspective, Table 3 presents the impact of progressively enhancing perception and cognition. Our results show that while improving perception yields only marginal gains (+1.2% cIoU), enhancing cognition leads to a substantial performance boost (+12.9% cIoU). From a qualitative perspective, Fig. 3 visualizes the top 4 instance masks. Although object queries produce accurate masks, referent classification frequently fails due to inadequate contextual and semantic understanding. These findings conclusively demonstrate that *cognition, rather than perception, constitutes the primary bottleneck*. More details are analyzed in Sec. 4.3.1.

(2) How to Effectively Connect Perception and Cognition Components? We propose a loopback architecture to establish strong interactions between the cognition and perception. In this architecture, object queries serve as a medium to generate fine-grained instance-level masks within the perception branch, which are subsequently transferred to the cognition branch to compute referring scores with robust contextual understanding. The process is iteratively reinforced, with the predictions from each iteration being supervised. This progressively information flow forms a powerful loopback synergy, fostering strong connections between the perception and cognition branches. Some synergy forms are compared in Sec. 4.4.1.

(3) How to Improve Non-Referent Judgment Ability? We analyze the gRefCOCO [42] dataset and observe that approximately 9% of the samples are non-referents, leading to a long-tail distribution problem in non-referent classification. To mitigate this issue, we propose a non-referent sample conversion augmentation strategy that dynamically transforms referent-present expressions into non-referent ones, thereby increasing the diversity of non-referent sample pairs. Furthermore, we find that auxiliary supervision for non-referent judgment improves the model's capability to handle non-referent cases, even though it is not explicitly

leveraged during testing. We provide quantitative analyses in Sec. 4.3.3 and qualitative analyses in Appendix F.3.

The main contributions of this paper are as follows:

- We introduce **DeRIS**, a pioneering framework that decouples the perception and cognition components of RIS, seamlessly integrating their respective strengths to achieve high-precision perceptual localization and robust multi-modal contextual understanding.
- We provide in-depth analyses of the inherent limitations in existing RIS frameworks, identifying critical structural bottlenecks and the long-tail distribution challenges. To address these issues, we propose a novel *Loopback Synergy* mechanism that fosters progressively interaction between perception and cognition modules. Furthermore, we develop a simple yet effective non-referent sample conversion strategy to alleviate training instability and enhance model generalization.
- Our proposed approach sets a new state-of-the-art performance on the RefCOCO, RefCOCO+, and RefCOCOg (RIS), as well as the gRefCOCO (GRES) dataset.

2. Related Work

2.1. Referring Image Segmentation (RIS)

Perception-centric RIS methods are typically categorized into two structural paradigms: post-fusion and early-fusion. Post-fusion methods [25, 29, 36, 41, 45, 47, 49, 77] primarily focus on the fusion of visual and linguistic features through specially designed interaction modules that are applied after the visual backbone. In contrast, early-fusion methods [5, 10, 15, 18, 43, 58, 68, 69, 75] emphasize integrating visual and linguistic features during the feature extraction phase via modularized architectural designs, facilitating deeper cross-modal understanding. Cognition-centric RIS methods [11, 12, 66, 72] leverage powerful pre-trained multimodal models [1, 30, 53, 61] to generate aligned or fused multimodal representations, thereby enhancing the model's cognitive capacity for interpreting complex image-text relationships. In this work, we introduce a novel architectural framework that decouples the perception and cognition components, effectively harnessing the strengths of both paradigms to advance RIS performance.

2.2. Vision-Language Understanding

Vision-language pre-trained models can be broadly classified into four primary categories: one-stream, dual-stream, fusion encoder, and multimodal large language models (MLLMs). One-stream models [7, 26, 30, 32] process image and text inputs within a unified stream, concatenating their embeddings to enable cross-modal interactions throughout the feature extraction process. Dual-stream models [28, 38, 53, 76], in contrast, employ separate encoders for each modality, with interactions occurring at a shallow layer between pooled image and text features. Fusion encoder models [1, 35, 57, 61, 74] integrate elements of both one- and dual-stream approaches, facilitating intermediate cross-modal interactions to balance model complexity and performance. Recently, MLLM-based methods [44, 60, 64, 78] harness the advanced reasoning and understanding capabilities of foundational large language models, adapting them for multimodal inputs through tailored training strategies. In this paper, we identify a key deficiency in current models: *inadequate cognitive capability*. To address this, we introduce a loopback synergy strategy that employs object queries as an intermediary to facilitate cascading and progressive transfers of semantic understanding between the target perception and text-image modalities.

2.3. Instance Segmentation

CNN-based instance segmentation methods can be classified into two major paradigms: top-down and bottom-up. The top-down paradigm [2–4, 6, 16, 21] first generates bounding boxes using object detectors, followed by mask segmentation. In contrast, the bottom-up paradigm [13, 19, 46, 52] approaches instance segmentation as a labeling and clustering problem, where pixels are first assigned class labels or embedded into a feature space, then clustered into individual instances. Recently, several transformer-based instance segmentation methods [9, 27, 34, 73] have been proposed. These methods treat instance segmentation as a mask classification task, associating instance categories with a set of predicted binary masks. In this work, we adopt transformer-based architectures and leverage pre-trained weights from Mask2Former [9] to provide the model with prior perception capabilities. This design introduces instance-aware capabilities into the RIS task, enabling more refined spatial discrimination of instances.

3. Methodology

3.1. Overview

Fig. 4 illustrates an overview of the DeRIS architecture, comprising three primary components: *perception branch*, *cognition branch*, and *loopback synergy*. In the cognition branch, we introduce a non-referent sample conversion (NSC) strategy, which dynamically transforms referent image-text pairs into non-referent samples. The low-resolution image $\mathcal{I}_{low}$ and augmented text $\mathcal{T}'$ are then jointly embedded using an image-text encoder, leveraging the vision-language pre-trained (BEiT3 [61]) model. The choice of low resolution is motivated by two factors: (i) mitigating the quadratic computational complexity associated with input resolution, and (ii) the cognition branch primarily focuses on referent classification based on semantic understanding rather than fine-grained percep-

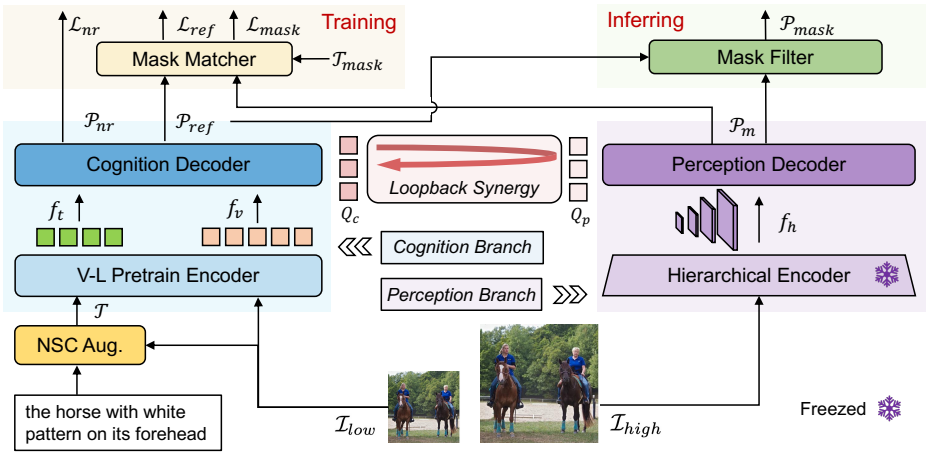

Figure 4. Overview of the proposed DeRIS framework. The RIS task is decoupled into *perception* and *cognition* branches, with a loopback synergy mechanism facilitating iterative information exchange. This design enhances synergy between the two branches, enabling a dynamic and progressive understanding of both perceptual targets and multi-modal semantics.

tion. The perception branch follows a structure similar to Mask2Former [9], employing a hierarchical encoder [48] to extract multi-scale features from high-resolution images $\mathcal{I}_{high}$. These features are then refined using an FPN-like [55] structure to enhance fine-grained visual representations. The loopback synergy adopts a cascade structure, enabling multiple rounds of object query transfer between the perception and cognition branches. This iterative process fosters a dynamic and progressive understanding of both perceptual targets and multi-modal semantics. By refining multi-level features, the model generates high-quality masks while concurrently improving cognitive capabilities. Finally, non-referent samples are identified through query pooling and a binary classifier, formulated as:

$$\mathcal{P}_{nr} = \text{AvgPool}(\text{Linear}(Q_{ck})) \tag{1}$$

where Q_{ck} denotes the cognition queries from the final round of loopback synergy.

3.2. Loopback Synergy

3.2.1. Single-round of Loopback Synergy

The loopback synergy module is designed to establish strong interactions between the perception and cognition branches. As illustrated in Fig. 5, each round of interaction consists of a *cognition layer* and a *perception layer*. Initially, the perception layer provides perceptual queries Q_p and the decoded mask M_p to the cognition layer. Subsequently, the cognition layer interacts with Q_p and the image-text semantic information, producing cognitive query Q_c and a confidence score S_r for each query, which indicates the relevance to the target object. Finally, the fused query Q_f for the next round is obtained by applying the C1 operation between Q_p and Q_r. The C1 operation is defined as:

$$Q_f = \text{C1}(Q_p, Q_r) = \text{MLP}(\text{Concat}(Q_p, Q_r)). \tag{2}$$

This bidirectional information flow forms an effective loopback mechanism, creating robust connections between the perception and cognition branches.

3.2.2. Perception Layer

The operations in the perception layer closely resemble those in Mask2Former [9], as depicted in the right portion of Fig. 5. Initially, initial queries Q_i interact with multi-scale features via deformable cross-attention to capture spatial contextual information. This is followed by a self-attention operation to establish inter-instance relationships, yielding the output Q_p. Subsequently, the predicted mask M_p is derived by performing a dot product between Q_p and the feature map f_m. The feature map f_m integrates both perception-related features f_{h4} and cognition-related features f_v, defined as:

$$f_m = \text{C2}(f_{h4}, f_v) = \text{Conv}(\text{Concat}(f_{h4}, f_v)), \tag{3}$$

where Conv denotes a 1×1 convolution operation. This design primarily aims to bolster the model's capability to generate text-informed proposals by incorporating contextual semantic information from the cognition branch, while simultaneously enhancing its perceptual accuracy.

3.2.3. Cognition Layer

The operations in the cognition layer are depicted in the left portion of Fig. 5. The Instance-Instance Relation aims to establish mutual relationships among Q_p, with the goal of enabling each object to perceive its spatial position. We leverage the mask prediction M_p from the perception layer as a prior. The computation proceeds as follows:

$$f_s = \text{AvgPool}(f_m \times \sigma(M_p)), \tag{4}$$

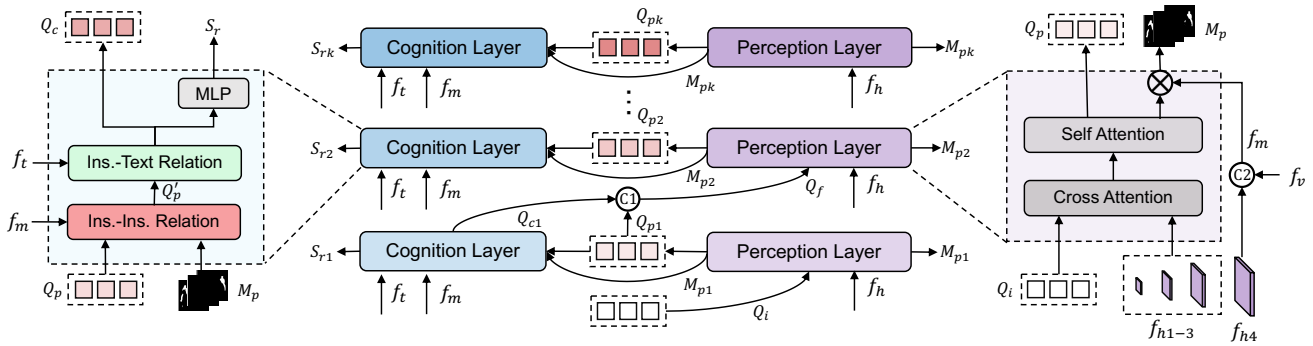

Figure 5. The architecture of the loopback synergy. Each round of interaction consists of a cognition layer and a perception layer. The perception layer provides object queries Q_p and the decoded mask M_p to the cognition layer. The cognition layer interacts with these, and produces cognition queries Q_c and referent scores S_r.

$$Q'_p = \text{Attn}(Q_i, Q_i + f_s, Q_i), \tag{5}$$

where σ denotes the sigmoid operation, and Attn refers to the self-attention operation. The Instance-Text Relation employs Q'_p as the query and the textual features f_t as both the key and value in a cross-attention operation, yielding Q_c, which is closely aligned with image-text semantics. Finally, an MLP is applied to derive the confidence score S_r, representing the probability of each query's relevance to the input expression.

3.3. Non-referent Sample Conversion

We observe that existing models frequently demonstrate suboptimal performance in non-referent judgment, as evidenced by the N-acc metric in general scenarios. Our analysis reveals that the proportion of non-referent samples in the gRefCOCO [42] training set is approximately 9%, highlighting a long-tail distribution challenge. Due to the limited representation of non-referent samples during training, models tend to over-predict the presence of target objects, leading to reduced N-acc scores. To mitigate this issue, we propose a simple non-referent sample conversion augmentation strategy, which dynamically transforms image-text pairs containing targets into non-referent samples by replacing the textual description, thereby increasing the diversity of non-referent sample pairs. However, this approach requires careful design to prevent scenarios where the described target in the modified text might still exist in the corresponding image. We address this challenge through a three-level filtering process:

1. The image corresponding to the selected sentence must be inconsistent with the current image.
2. The length of the selected sentence should be greater than a threshold N_w.
3. The similarity between the selected sentence and the original sentence should be below a threshold T_s.

Here, N_w and T_s are set to 2 and 0.6 by default. The similarity is computed using a fast approximation method:

$$\text{Sim}(s_1, s_2) = \frac{\text{Jac}(s_1, s_2) + \text{Cos}(s_1, s_2)}{2} \tag{6}$$

where Jac and Cos denote the Jaccard and Cosine similarity (See Appendix D.1). Additionally, we introduce an extra hyperparameter R_c, which controls the probability of conversion (See Sec. 4.4.2).

3.4. Training Objectives

The output of DeRIS consists of three components: the segmentation mask $\mathcal{P}_m$, referring classification $\mathcal{P}_{ref}$, and non-referent judgment $\mathcal{P}_{nr}$. Initially, a mask matcher [9] is used to perform Hungarian matching between $\mathcal{P}_m$ and the ground truth mask $\mathcal{T}_{mask}$. The training objective consists of three components:

1. The segmentation loss $\mathcal{L}_{mask}$, associated with each matched $\mathcal{P}_m$, is computed using a combination of BCE and Dice loss to quantify the discrepancy between the predicted mask and the ground truth mask.
2. The referring classification loss $\mathcal{L}_r$, corresponding to each object query, also employs BCE loss to assess whether the query correctly refers to the target object.
3. The non-referent loss $\mathcal{L}_{nr}$ utilizes BCE loss to determine whether the described object exists in the image.

The loss for the i-th loopback round can be expressed as in Eq. 7, and the total loss is given by Eq. 8.

$$\mathcal{L}^i = \lambda_m \mathcal{L}^i_{mask} + \lambda_r \mathcal{L}^i_r + \lambda_{nt} \mathcal{L}^i_{nt}, \tag{7}$$

$$\mathcal{L} = \lambda_{aux} \sum_{i=1}^{N_r - 1} \mathcal{L}^i + \mathcal{L}^{N_r}, \tag{8}$$

where λ_m, λ_r, and λ_{nt} are set to 1.0, 1.0, and 1.0 by default. λ_{aux} controls the weight of the auxiliary losses, which is set to 0.2. N_r denotes the number of interaction rounds, which is set to 3. During inference, the referring classification $\mathcal{P}_{ref}$ is filtered using a threshold $\mathcal{T}_{ref}$. In this paper, we set $\mathcal{T}_{ref} = 0.7$ by default.

Table 1. Comparison with state-of-the-art methods on the RefCOCO/+/g [51, 70] datasets for the RIS task. † indicates the use of the cIoU evaluation metric, ‡ indicates the use of oIoU, while all other methods are evaluated using mIoU.

Method	Visual Encoder	Textual Encoder	RefCOCO			RefCOCO+			RefCOCOg	
			val	test A	test B	val	test A	test B	val(U)	test(U)
MLLM-based Methods:										
LISA-7B† [CVPR'24] [31]	SAM-H + CLIP-L	Vicuna-7B	74.90	79.10	72.30	65.10	70.80	58.10	67.90	70.60
GSVA-13B† [CVPR'24] [65]	SAM-H + CLIP-L	Vicuna-13B	78.20	80.40	74.20	67.40	71.50	60.90	74.20	75.60
GLaMM-7B† [CVPR'24] [54]	SAM-H + CLIP-H	Vicuna-7B	79.50	83.20	76.90	72.60	78.70	64.60	74.20	74.90
SAM4MLLM-8B† [ECCV'24] [8]	SAM-EfficientViT-XL1	Qwen-VL-7B	79.80	82.70	74.70	74.60	80.00	67.20	75.50	76.40
DeRIS-7B† (Ours)	Swin-B + SigLIP	Qwen2-OV-7B	**84.05**	**85.79**	**83.32**	**80.30**	**83.92**	**76.16**	**80.62**	**80.59**
Specialist Methods:										
PVD [AAAI'24] [10]	Swin-B	BERT-B	74.82	77.11	69.52	63.38	68.60	56.92	63.13	63.62
ReMamber‡ [ECCV'24] [67]	Mamba-B	CLIP-B	74.54	76.74	70.89	65.00	70.78	57.53	63.90	64.00
VG-LAW [CVPR'23] [58]	ViT-B	BERT-B	75.62	77.51	72.89	66.63	70.38	59.89	65.53	66.08
CGFormer [CVPR'23] [59]	Swin-B	BERT-B	76.93	78.70	73.32	68.56	73.76	61.72	67.57	67.83
PolyFormer-L [CVPR'23] [45]	Swin-L	BERT-B	76.94	78.49	74.83	72.15	75.71	66.73	71.15	71.17
Prompt-RIS [CVPR'24] [56]	SAM-B + CLIP-B	CLIP-B	78.10	81.21	74.64	71.13	76.60	64.25	70.47	71.29
EEVG [ECCV'24] [5]	ViT-B	BERT-B	79.49	80.87	77.39	71.86	76.67	66.31	73.56	73.47
SimVG-Seg [NeurIPS'24] [11]	BEiT3-B	BEiT3-B	77.78	79.14	76.02	72.21	75.37	67.85	72.19	73.02
C^3VG [AAAI'25] [12]	BEiT3-B	BEiT3-B	81.37	82.93	79.12	77.05	79.61	72.40	76.34	77.10
OneRef-L [NeurIPS'24] [66]	BEiT3-L	BEiT3-L	81.26	83.06	79.45	76.60	80.16	72.95	75.68	76.82
DeRIS-B (Ours)	Swin-S	BEiT3-B	81.99	82.97	80.14	75.62	79.16	71.63	76.30	77.15
DeRIS-L (Ours)	Swin-B	BEiT3-L	**85.72**	**86.64**	**84.52**	**81.28**	**83.74**	**78.59**	**80.01**	**81.32**

Table 2. Comparison with the state-of-the-art methods on gRefCOCO [42] dataset for GRES task.

Method	Visual Encoder	Textual Encoder	Val			TestA			TestB		
			gIoU	cIoU	N-acc.	gIoU	cIoU	N-acc.	gIoU	cIoU	N-acc.
MLLM-based Methods:											
LISA-7B [CVPR'24] [31]	SAM-ViT-H	Vicuna-7B	61.63	61.76	54.67	66.27	68.50	50.01	58.84	60.63	51.91
GSVA-7B [CVPR'24][65]	SAM-ViT-H	Vicuna-7B	66.47	63.29	62.43	71.08	69.93	65.31	62.23	60.47	60.56
SAM4MLLM-8B [ECCV'24] [8]	SAM-EfficientViT-XL1	Qwen-VL-7B	71.86	67.83	66.08	74.15	72.22	63.92	65.29	63.42	59.99
Specialist Methods:											
MattNet [CVPR'18] [71]	ResNet-101	Bi-LSTM	48.24	47.51	41.15	59.30	58.66	44.04	46.14	45.33	41.32
CRIS [CVPR'22] [47]	CLIP-ResNet-101	CLIP-B	56.27	55.34	-	63.42	63.82	-	51.79	51.04	-
LAVT [CVPR'22] [68]	Swin-B	BERT-B	58.40	57.64	49.32	65.90	65.32	49.25	55.83	55.04	48.46
ReLA [CVPR'23] [42]	Swin-B	BERT-B	63.60	62.42	56.37	70.03	69.26	59.02	61.02	59.88	58.40
EEVG [ECCV'24] [5]	ViT-B	BERT-B	62.75	64.04	-	70.93	71.65	-	62.79	62.77	-
HieA2G [AAAI'25] [62]	Swin-B	RoBERTa-B	68.40	64.20	62.80	72.00	70.40	63.40	62.80	60.80	61.00
DeRIS-B (Ours)	Swin-S	BEiT3-B	74.10	68.06	77.03	73.72	71.99	75.98	65.63	64.65	63.44
DeRIS-L (Ours)	Swin-B	BEiT3-L	**77.67**	**72.00**	**82.22**	**75.30**	**73.73**	**78.30**	**67.99**	**67.38**	**66.81**

4. Experiments

4.1. Settings

We conduct extensive experiments on four publicly available datasets: RefCOCO/+/g [51, 70], gRefCOCO [42]. For the perception branch, we initialize the model using pretrained weights from Mask2Former [9] and use an input resolution of 384×384. For the cognition branch, we initialize the model using pretrained weights from BEiT3 [61] and use an input resolution of 224×224. Additional implementation details are provided in the Appendix C.

4.2. Main Results

(1) RIS Results: Table 1 presents a comparative evaluation of DeRIS against recent SOTA methods on the RefCOCO/+/g [51, 70] datasets. Notably, DeRIS surpasses existing advanced models by a substantial margin. Specifically, compared to the recent specialist method OneRef-L [66], DeRIS-L achieves improvements of 4.46%, 4.68%, and 4.33% in mIoU on the validation sets of RefCOCO, RefCOCO+, and RefCOCOg, respectively. Furthermore, we extend DeRIS to a large-model variant by integrating Qwen2-OV-7B [33], replacing BEiT3 [61] with a more powerful multi-modal large language model while employing LoRA [22] for efficient training. Experimental results demonstrate that, at a comparable model scale, DeRIS-7B outperforms SAM4MLLM-8B [8] by 4.25%, 5.70%, and 5.12% in cIoU on the validation sets of RefCOCO/+/g datasets.

(2) GRES Results: Table 2 summarizes the performance of DeRIS on the gRefCOCO [42] dataset, where it consistently outperforms prior methods across nearly all metrics, with a particularly significant improvement in the N-acc metric. These results highlight DeRIS's robustness in handling both

Table 3. The analysis of the perception and cognition capabilities. The results are conducted on the gRefCOCO [42] dataset.

Cognition	Perception	N-acc.	gIoU	cIoU
BERT-B	Swin-S	66.13	64.55	56.26
ViLT-B		68.25 ↑ 2.12	67.12 ↑ 2.57	61.23 ↑ 4.97
BEiT3-B		70.25 ↑ 4.12	70.64 ↑ 6.09	66.31 ↑ 10.05
BEiT3-L		73.16 ↑ 7.03	73.53 ↑ 8.98	69.14 ↑ 12.88
BEiT3-B	Swin-T	70.22	70.13	65.54
	Swin-S	70.25 ↑ 0.03	70.64 ↑ 0.51	66.31 ↑ 0.83
	Swin-B	70.35 ↑ 0.13	70.85 ↑ 0.72	66.78 ↑ 1.20

Table 4. The analysis of cognition component capabilities, including comparisons with MLLM. The metrics are based on the average gIoU across several subsets.

Cognition	Perception	RefCOCO	RefCOCO+	RefCOCOg-U
BEiT3-0.2B [61]	Swin-B	79.06	72.43	73.98
Qwen2-0.5B [33]		81.32 ↑ 2.26	74.63 ↑ 2.20	74.96 ↑ 0.98
Mipha-3B [78]		82.16 ↑ 3.10	79.19 ↑ 6.76	77.38 ↑ 3.40
Qwen2-7B [33]		82.87 ↑ 3.81	79.88 ↑ 7.45	78.30 ↑ 4.32

non-referent and multi-referents scenarios, demonstrating its strong generalization capabilities.

4.3. Quantitative Analyses

In this section, we present a series of analyses from a quantitative perspective. Comprehensive qualitative analyses are provided in Appendix F.

4.3.1. Influence of Cognition and Perception Capability

(1) Cognition Capability is Crucial for RIS. The decoupled structure, which separates the RIS task into cognition and perception branches, provides a robust foundation for analyzing the relative impact of each component on model performance. Specifically, the cognition branch is tasked with determining whether the perceptual output aligns with the referent described in the input sentence by interpreting vision-text semantics. We conduct a fair comparison by systematically varying the model scales of the cognition and perception branches, as presented in Table 3. In the first ablation group of Table 3, we observe a nearly 10-point improvement in cIoU when transitioning from the BERT-B [14] model to the BEiT3-B [61] model. Further scaling the BEiT3 model from base to large yields an additional significant 3-point gain in cIoU. In contrast, the perception branch exhibits minimal performance improvement with increasing model size. These findings lead us to conclude that, for current RIS tasks, cognitive ability plays a more critical role than perceptual ability.

(2) Cognitive Ability as the Primary Bottleneck with Significant Room for Improvement. As shown in Table 4, the model's performance steadily improves as the scale of the cognitive model increases. A similar trend is observed in MLLM [60, 78], further confirming that the cognitive model's capability serves as the primary bottleneck in RIS tasks. This suggests that enhancing multi-modal understanding is the key to further improving RIS performance.

Table 5. The effectiveness and efficiency of different architecture focuses under varying resolutions. Infer Time is measured on a single NVIDIA A100 GPU.

Structure	gIoU	Pr@0.9	Infer Time (ms)
Resolution=224			
Cognition-centric	68.54	8.32	14.7
Perception-centric	64.55 ↓ 3.99	15.79 ↑ 7.47	11.2 ↑ 24%
Integrated	70.64 ↑ 2.10	20.93 ↑ 12.61	15.9 ↓ 8%
Resolution=384			
Cognition-centric	69.87	13.98	28.1
Perception-centric	67.12 ↓ 2.75	23.08 ↑ 9.10	20.0 ↑ 29%
Integrated	71.74 ↑ 1.87	28.39 ↑ 14.41	22.7 ↑ 19%

4.3.2. Effectiveness of Fine-Grained Segmentation

(1) Fine-Grained Segmentation Analysis. DeRIS decouples perception and cognition, effectively leveraging the strengths of both, which leads to improvements in both perceptual and cognitive capabilities. We evaluate the model's fine-grained segmentation ability using the Pr@0.9, with results shown in Table 5. It demonstrates that DeRIS significantly enhances fine-grained segmentation performance.

(2) Efficiency Analysis. Cognition-centric methods typically employ the vision transformer architecture, which has a quadratic computational complexity with respect to the image resolution, making them inefficient for processing high-resolution inputs. In contrast, perception-centric methods use hierarchical architectures, thus enabling more efficient processing of high-resolution inputs. As shown in Table 5, our method maintains high inference efficiency even with powerful performance.

4.3.3. Analysis of Non-referent Judgement

(1) Influence of Non-referent Prediction During Training and Testing. The instance-level prediction framework does not explicitly require non-referent classification, as the model inherently assumes the absence of a target when all object queries yield low confidence scores. However, we observe that incorporating auxiliary supervision for non-referent samples during training significantly improves the model's ability to identify such cases. Moreover, in the post-processing stage, multiplying $\mathcal{P}_{nr}$ with $\mathcal{P}_{ref}$ during inference results in a notable performance gain, as shown in Table 6. This indicates that both implicit target existence learning during training and explicit target presence verification during inference contribute to performance improvements. These findings suggest that the model exhibits an inherent instability in global target existence prediction.

(2) Impact of Non-referent Sample Conversion. In the gRefCOCO [42] dataset, non-referent samples constitute approximately 9% of the total data, making non-referent classification a long-tail distribution problem. To mitigate this issue, we introduce NSC augmentation , a strategy that

Table 6. The impact of non-referent supervision and judgement, as well as the influence of incorporating the NSC strategy.

Train	Test	N-acc	gIoU	cIoU
w/o NSC Augmentation				
		60.19	66.09	63.98
✓		62.71 ↑ 2.52	67.73 ↑ 1.64	65.31 ↑ 1.33
✓	✓	68.80 ↑ 6.09	70.10 ↑ 2.37	66.40 ↑ 1.09
w/ NSC Augmentation				
		75.36 ↑ **15.17**	71.82 ↑ **5.73**	66.33 ↑ **2.35**
✓		79.25 ↑ 3.89	73.77 ↑ 1.95	67.49 ↑ 1.16
✓	✓	79.32 ↑ 0.07	72.78 ↓ 0.99	66.41 ↓ 1.08

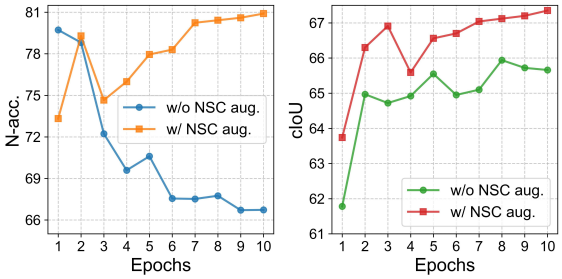

Figure 6. Impact of NSC on training stability and convergence.

dynamically expands the diversity of non-referent samples during training. As shown in Table 6, this approach results in more than 15-point improvement in N-acc. Additionally, NSC enables the model to achieve comparable performance without the need to introduce $\mathcal{P}_{nr}$ during post-processing, further streamlining inference.

(3) Stability Analysis of the Training Process. As shown in Fig. 6, we observe that the N-acc metric fluctuates significantly during training on the gRefCOCO [42] dataset, which leads to instability in other metrics such as cIoU. We attribute this instability primarily to the limited diversity of non-referent image-text samples in the dataset, which causes the model to overfit. However, after introducing the NSC strategy, the N-acc metric becomes more stable throughout the training process. This indicates that the NSC strategy effectively alleviates the issue of non-referent sample scarcity, improving the model's generalization ability.

4.4. Ablation Study

4.4.1. Synergy Forms between Perception and Cognition

The integration of perception and cognition branches is a pivotal aspect of our decoupled framework. To investigate this, we evaluated four interaction forms: P-to-C (Baseline), C-to-P, Hierarchical Combined, and Loopback Synergy, with results presented in Table 7. Our findings indicate that transferring object queries from the cognition branch to the perception branch notably decelerates model convergence. We attribute this to the perception branch's optimization being constrained by the cognition pipeline. Conversely, the reverse flow (P-to-C) supports typical convergence rates. The Hierarchical Combined approach, which

Table 7. The impact of different connection structures.

Structure	gIoU	cIoU	Train Time (h)
Query Transfer: P-to-C	69.98	65.49	2.7
Query Transfer: C-to-P	56.77 ↓ 13.21	54.80 ↓ 10.69	2.7 ↓ 0%
Hierarchical Combined	70.13 ↑ 0.15	66.32 ↑ 0.83	3.2 ↓ 18%
Loopback Synergy	71.37 ↑ 1.39	67.27 ↑ 1.78	2.8 ↓ 3%

Table 8. The impact of non-referent sample conversion rate R_c.

R_c	N-acc.	gIoU	cIoU
0%	70.25	70.64	66.31
5%	74.48 ↑ 4.23	72.49 ↑ 1.85	67.22 ↑ 1.91
10%	75.46 ↑ 0.98	72.50 ↑ 0.01	66.76 ↓ 0.46
15%	79.25 ↑ 3.79	73.77 ↑ 1.27	67.49 ↑ 0.73
20%	79.39 ↑ 0.14	73.67 ↓ 0.10	67.07 ↓ 0.42

fuses f_v with f_{1-3} at the feature level using SimFPN [39], offers modest performance improvements but substantially reduces training efficiency. The proposed loopback synergy strategy, leveraging multi-rounds iterative object query exchanges between branches to enhance both cognitive and perceptual capabilities. This method improves performance by 1-2 points without adding notable computational cost.

4.4.2. Impact of Conversion Rate of Non-referent

The proposed NSC data augmentation strategy transforms image-sentence pairs from referent to non-referent cases. The conversion probability R_c serves as a critical hyperparameter in this process. As shown in Table 8, we observe that increasing R_c leads to a steady improvement in non-referent accuracy (N-acc). However, when R_c exceeds a certain threshold, performance on other metrics, such as cIoU, begins to decline. This occurs primarily because enhancing the diversity of non-referent samples prompts the model to assign higher confidence to predicting the absence of a target, effectively mitigating the low-confidence issue in tail categories stemming from the long-tail distribution of the dataset. In this paper, we set R_c=15% to address the limited diversity and scarcity of non-referent samples.

5. Conclusion

In this paper, we introduce **DeRIS**, a novel framework for RIS that decouples perception and cognition to harness their complementary strengths. Our analyses reveal that while existing models are primarily limited by the inadequate contextual image-text understanding. To address this gap, we propose a multi-round loopback synergy mechanism that iteratively enhances cognitive and perceptual capabilities through query refinement. Additionally, we present a non-referent sample conversion strategy to boost confidence in non-referent judgment, mitigating the effects of long-tail distribution in the dataset that contribute to uncertainty in handling non-referent cases. Extensive experiments demonstrate that our approach significantly enhances performance across both RIS and GRES tasks.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Nos. 62276061 and 62436002. This work is also supported by Research Fund for Advanced Ocean Institute of Southeast University (Major Program MP202404). This work is also supported by the SEU Innovation Capability Enhancement Plan for Doctoral Students (CXJH_SEU 25125).

References

[1] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. In *NeurIPS*, 2022. 1, 3

[2] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *ICCV*, pages 9157–9166, 2019. 3

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *TPAMI*, 43(5):1483–1498, 2019.

[4] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. Blendmask: Top-down meets bottom-up for instance segmentation. In *CVPR*, pages 8573–8581, 2020. 3

[5] Wei Chen, Long Chen, and Yu Wu. An efficient and effective transformer decoder-based framework for multi-task visual grounding. In *ECCV*, 2024. 3, 6

[6] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. In *ICCV*, pages 2061–2069, 2019. 3

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120, 2020. 3

[8] Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In *ECCV*, pages 323–340, 2024. 6

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 1, 3, 4, 5, 6

[10] 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 *AAAI*, pages 1326–1334, 2024. 1, 3, 6

[11] Ming Dai, Lingfeng Yang, Yihao Xu, Zhenhua Feng, and Wankou Yang. Simvg: A simple framework for visual grounding with decoupled multi-modal fusion. In *NeurIPS*, 2024. 1, 3, 6

[12] Ming Dai, Jian Li, Jiedong Zhuang, Xian Zhang, and Wankou Yang. Multi-task visual grounding with coarse-to-fine consistency constraints. *arXiv:2501.06710*, 2025. 1, 3, 6

[13] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. *arXiv:1708.02551*, 2017. 3

[14] 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

[15] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, pages 16301–16310, 2021. 1, 3

[16] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *ICCV*, pages 6910–6919, 2021. 3

[17] Ali Farhadi and Joseph Redmon. Yolov3: An incremental improvement. In *CVPR*, pages 1–6, 2018. 1, 2

[18] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *CVPR*, pages 15506–15515, 2021. 3

[19] Naiyu Gao, Yanhu Shan, Yupei Wang, Xin Zhao, Yinan Yu, Ming Yang, and Kaiqi Huang. Ssap: Single-shot instance segmentation with affinity pyramid. In *ICCV*, pages 642–651, 2019. 3

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 2

[21] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 1, 3

[22] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6

[23] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, 2016. 1

[24] Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. Beyond one-to-one: Rethinking the referring image segmentation. In *ICCV*, pages 4044–4054, 2023. 1

[25] Ziling Huang and Shin'ichi Satoh. Referring image segmentation via joint mask contextual embedding learning and progressive alignment network. In *EMNLP*, pages 7753–7762, 2023. 3

[26] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*, pages 12976–12985, 2021. 3

[27] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *CVPR*, pages 2989–2998, 2023. 3

[28] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 3

[29] Namyup Kim, Dongwon Kim, Suha Kwak, Cuiling Lan, and Wenjun Zeng. Restr: Convolution-free referring image segmentation using transformers. In *CVPR*, pages 18124–18133, 2022. 3

[30] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594, 2021. 1, 2, 3

[31] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, pages 9579–9589, 2024. 6

[32] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *ICLR*, 2020. 3

[33] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6, 7

[34] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *CVPR*, pages 3041–3050, 2023. 1, 3

[35] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 3

[36] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *NeurIPS*, 34, 2021. 3

[37] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *CVPR*, 2018. 1

[38] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2022. 2, 3

[39] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, pages 280–296, 2022. 8

[40] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Recurrent multimodal interaction for referring image segmentation. In *ICCV*, 2017. 1

[41] Chang Liu, Xudong Jiang, and Henghui Ding. Instance-specific feature propagation for referring segmentation. *TMM*, 25:3657–3667, 2022. 3

[42] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: generalized referring expression segmentation. In *CVPR*, pages 23592–23601, 2023. 1, 2, 5, 6, 7, 8

[43] Chang Liu, Henghui Ding, Yulun Zhang, and Xudong Jiang. Multi-modal mutual attention and iterative interaction for referring image segmentation. *TIP*, 32:3054–3065, 2023. 3

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3

[45] 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. 1, 3, 6

[46] Shu Liu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. Sgn: Sequential grouping networks for instance segmentation. In *ICCV*, pages 3496–3504, 2017. 3

[47] Sun'ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. CARIS: context-aware referring image segmentation. In *ACM MM*, pages 779–788, 2023. 1, 3, 6

[48] 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, 2, 4

[49] 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 *CVPR*, pages 10034–10043, 2020. 3

[50] 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. 1

[51] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, pages 792–807, 2016. 1, 6

[52] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NeurIPS*, 2017. 3

[53] 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, 2021. 1, 2, 3

[54] 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. *CVPR*, pages 13009–13018, 2024. 6

[55] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *TPAMI*, 39(6):1137–1149, 2017. 4

[56] Chao Shang, Zichen Song, Heqian Qiu, Lanxiao Wang, Fanman Meng, and Hongliang Li. Prompt-driven referring image segmentation with instance contrasting. In *CVPR*, pages 4124–4134, 2024. 6

[57] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, pages 15638–15650, 2022. 3

[58] 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 *CVPR*, pages 10857–10866, 2023. 3, 6

[59] Jiajin Tang, Ge Zheng, Cheng Shi, and Sibei Yang. Contrastive grouping with transformer for referring image segmentation. In *CVPR*, pages 23570–23580, 2023. 6

[60] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191*, 2024. 3, 7

[61] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *CVPR*, pages 19175–19186, 2023. 1, 2, 3, 6, 7

[62] Yaxian Wang, Henghui Ding, Shuting He, Xudong Jiang, Bifan Wei, and Jun Liu. Hierarchical alignment-enhanced adaptive grounding network for generalized referring expression comprehension, 2025. 6

[63] Jianzong Wu, Xiangtai Li, Xia Li, Henghui Ding, Yunhai Tong, and Dacheng Tao. Towards robust referring image segmentation. *TIP*, 33:1782–1794, 2024. 1

[64] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv:2412.10302*, 2024. 3

[65] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *CVPR*, pages 3858–3869, 2024. 6

[66] Linhui Xiao, Xiaoshan Yang, Fang Peng, Yaowei Wang, and Changsheng Xu. Oneref: Unified one-tower expression grounding and segmentation with mask referring modeling. In *NeurIPS*, 2024. 1, 3, 6

[67] Yuhuan Yang, Chaofan Ma, Jiangchao Yao, Zhun Zhong, Ya Zhang, and Yanfeng Wang. Remamber: Referring image segmentation with mamba twister. *ECCV*, 2024. 6

[68] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. LAVT: language-aware vision transformer for referring image segmentation. In *CVPR*, pages 18134–18144, 2022. 1, 3, 6

[69] Jiabo Ye, Junfeng Tian, Ming Yan, Xiaoshan Yang, Xuwu Wang, Ji Zhang, Liang He, and Xin Lin. Shifting more attention to visual backbone: Query-modulated refinement networks for end-to-end visual grounding. In *CVPR*, pages 15502–15512, 2022. 3

[70] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016. 1, 6

[71] 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 *CVPR*, pages 1307–1315, 2018. 6

[72] Seonghoon Yu, Ilchae Jung, Byeongju Han, Taeoh Kim, Yunho Kim, Dongyoon Wee, and Jeany Son. A simple baseline with single-encoder for referring image segmentation. *arXiv:2408.15521*, 2024. 1, 3

[73] Xiaodong Yu, Dahu Shi, Xing Wei, Ye Ren, Tingqun Ye, and Wenming Tan. Soit: Segmenting objects with instance-aware transformers. In *AAAI*, 2022. 3

[74] Haotian* Zhang, Pengchuan* Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv:2206.05836*, 2022. 3

[75] Zicheng Zhang, Yi Zhu, Jianzhuang Liu, Xiaodan Liang, and Wei Ke. Coupalign: Coupling word-pixel with sentence-mask alignments for referring image segmentation. *NeurIPS*, 35:14729–14742, 2022. 3

[76] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, pages 16793–16803, 2022. 2, 3

[77] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *ECCV*, pages 598–615, 2022. 1, 3

[78] Minjie Zhu, Yichen Zhu, Xin Liu, Ning Liu, Zhiyuan Xu, Chaomin Shen, Yaxin Peng, Zhicai Ou, Feifei Feng, and Jian Tang. Mipha: A comprehensive overhaul of multimodal assistant with small language models. *AAAI*, 2025. 3, 7