Multi-task Visual Grounding with Coarse-to-Fine Consistency Constraints

Ming Dai[1], **Jian Li**[2], **Jiedong Zhuang**[3], **Xian Zhang**[1], **Wankou Yang**[1,4]
[1]School of Automation, Southeast University, China
[2]Youtu Lab, Tencent, China
[3]College of Information Science and Electronic Engineering, Zhejiang University, China
[4]Advanced Ocean Institute of Southeast Univerisity, Nantong, China
mingdai@seu.edu.cn, swordli@tencent.com, zhuangjiedong@zju.edu.cn, 230218200@seu.edu.cn, wkyang@seu.edu.cn

Abstract

Multi-task visual grounding involves the simultaneous execution of localization and segmentation in images based on textual expressions. The majority of advanced methods predominantly focus on transformer-based multimodal fusion, aiming to extract robust multimodal representations. However, ambiguity between referring expression comprehension (REC) and referring image segmentation (RIS) is error-prone, leading to inconsistencies between multi-task predictions. Besides, insufficient multimodal understanding directly contributes to biased target perception. To overcome these challenges, we propose a Coarse-to-fine Consistency Constraints Visual Grounding architecture (C^3VG), which integrates implicit and explicit modeling approaches within a two-stage framework. Initially, query and pixel decoders are employed to generate preliminary detection and segmentation outputs, a process referred to as the Rough Semantic Perception (RSP) stage. These coarse predictions are subsequently refined through the proposed Mask-guided Interaction Module (MIM) and a novel explicit bidirectional consistency constraint loss to ensure consistent representations across tasks, which we term the Refined Consistency Interaction (RCI) stage. Furthermore, to address the challenge of insufficient multimodal understanding, we leverage pre-trained models based on visual-linguistic fusion representations. Empirical evaluations on the RefCOCO, RefCOCO+, and RefCOCOg datasets demonstrate the efficacy and soundness of C^3VG, which significantly outperforms state-of-the-art REC and RIS methods by a substantial margin.

Introduction

Visual grounding is a critical task within the vision-language domain, aimed at establishing a fine-grained correspondence between images and text by grounding a given referring expression within an image (Li et al. 2022b). This task is typically divided into two sub-tasks based on the grounding approach: referring expression comprehension (REC) (Yu et al. 2018; Kamath et al. 2021) and referring image segmentation (RIS) (Kim et al. 2022; Tang et al. 2023). Traditionally, REC and RIS have been treated as separate tasks with distinct technological pathways, necessitating complex, task-specific designs. However, REC and RIS exhibit significant similarities and offer complementary strengths, making

Figure 1: (a) Three examples of inconsistent results between multi-task outputs. (b) Two examples of failure in identifying targets due to insufficient multi-modal understanding.

their unification both logical and advantageous. Recently, multi-task visual grounding has gained prominence as it eliminates the need for task-specific network designs and enables the leveraging of data across both tasks to mutually enhance performance. MCN (Luo et al. 2020) was the first approach to jointly train the REC and RIS tasks, employing a learnable method to establish consistency in attention maps. Recent research has primarily focused on enhancing the interaction across different modalities (Li and Sigal 2021; Su et al. 2023) and exploring auto-regressive approaches to achieve both detection and segmentation (Zhu et al. 2022; Cheng et al. 2024; Liu et al. 2023a). In this paper, we address two overlooked issues: **1)** *How to effectively leverage the complementarity of multi-task predictions to mitigate inconsistencies in results.* **2)** *How to overcome the challenge of insufficient multimodal understanding to enhance perception in complex image-text scenarios.*

Inconsistent predictions between multi-task primarily arise due to the lack of effective constraints linking different tasks. This issue can be exemplified by three scenarios

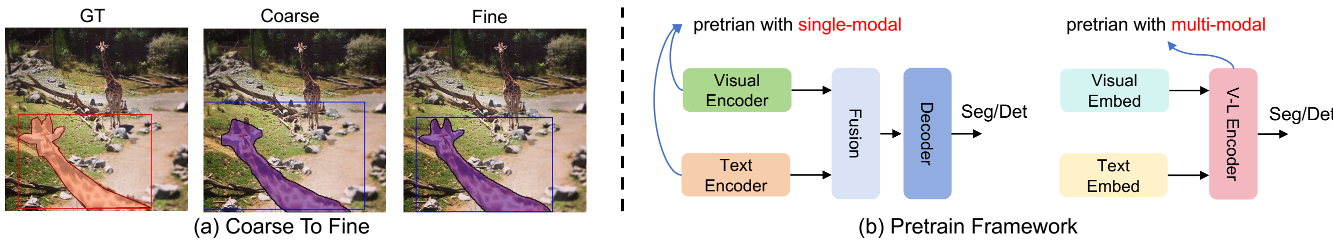

Figure 2: (a) Examples of the intermediate process in the proposed coarse-to-fine consistency constraint framework. (b) Two pretraining architectures: the left diagram illustrates separate encodings for image and text modalities followed by fusion, using single-modal pretraining; the right diagram shows a fused encoding architecture with multimodal pretraining.

depicted in Fig. 1(a): (1) accurate segmentation but erroneous detection, (2) inaccurate segmentation but correct detection, and (3) both segmentation and detection being incorrect yet providing complementary information. The traditional REC is a one-to-one detection task. When uncertainties arise during optimization, the detected result tends to be positioned between potential targets, leading to local optima. Conversely, the RIS task, involving finer-grained pixel-level predictions, can more precisely identify the target but often lacks sufficient spatial awareness. Thus, it becomes essential to introduce a multi-task consistency constraint to guide the model in supplementing information, thereby enhancing recognition in ambiguous situations. To this end, we propose a coarse-to-fine architecture for multi-task visual grounding, named C^3VG. The structure is shown in Fig. 3. Initially, we employ a pixel decoder and a query decoder to independently generate coarse foreground semantics and localization regions in the Rough Semantic Perception (RSP) stage. Subsequently, the Refined Consistency Interaction (RCI) stage refines them and enforces consistency across the multi-task outcomes. Within the RCI stage, we introduce a Mask-guided Interaction Module (MIM) to implicitly integrate the multi-task results from the RSP stage. Furthermore, we apply a bidirectional consistency constraint loss to explicitly enforce consistency across tasks. As illustrated in Fig. 2(a), the RSP stage delivers coarse localization and semantic results. Building on these priors, the RCI stage applies consistency constraints to produce higher-quality predictions.

Insufficient multimodal understanding primarily manifests as an inability to effectively capture the semantic associations between modalities in downstream tasks, particularly when data is limited. Fig. 1(b) shows two instances of identification errors caused by inadequate multimodal understanding: (1) The model incorrectly identifies ‘egg cup’ by focusing only on ‘cup’; (2) The model misinterprets ‘iMac’ due to the absence of prior knowledge. Recently, SimVG (Dai et al. 2024) has confirmed the importance of employing a pretrained multi-modality encoder for improving referential understanding. However, this paper aims to further extend this structure from a single detection task to a multi-task learning framework to validate its broader effectiveness. As shown on the left side of Fig. 2(a), previous methods typically utilize single-modal pretrained models as feature encoders and rely on limited downstream data to learn vision-language fusion representations. Recently, SimVG (Dai et al. 2024) has decoupled the downstream multimodal fusion process and incorporated it into upstream pretraining, resulting in significant performance improvements for the REC task. Fig. 2(b) illustrates the direct integration of the two modalities during the upstream pretraining process, leveraging advances in vision-language pretraining research (Kim, Son, and Kim 2021; Wang et al. 2023). This paper extends the conclusions of SimVG (Dai et al. 2024), demonstrating that the integration of multimodal pretrained models significantly enhances both convergence speed and accuracy in RIS and multi-task visual grounding tasks.

Our main contributions are summarized as follows:

1. We introduce an innovative and efficient coarse-to-fine architecture, C^3VG, specifically designed for multi-task visual grounding.
2. We design a mask-guided interaction module and a bidirectional consistency constraint loss to address the challenge of multi-task prediction inconsistency. These components facilitate implicit interaction and provide explicit supervision for multi-task predictions, respectively.
3. We extend the pretrained multi-modality encoder from a single-task setting to a multi-task joint training framework and validate its impact on addressing the issue of inadequate multimodal understanding.
4. The proposed C^3VG framework significantly outperforms state-of-the-art methods on RefCOCO/+/g datasets for both REC and RIS tasks, while requiring only half or fewer training epochs.

Related Work

Visual Grounding

Referring Expression Comprehension (REC) (Liu et al. 2019; Yang et al. 2020, 2024; Su et al. 2024; Zhuang et al. 2025) predicts a bounding box that tightly encompasses the target object in an image based on a referring expression. **Referring Image Segmentation** (RIS) (Yang et al. 2022; Zhang et al. 2022; Liu et al. 2023c) aims to provide pixel-level localization of a target object in an image based on a referring expression. **Multi-task Visual Grounding** seeks to localize and segment referring expressions using a single, integrated model. MCN (Luo et al. 2020) introduces a consistency energy maximization loss, which constrains the feature activation maps in both REC and RIS to be similar. Some Transformer-based methods (Li and Sigal 2021; Chen, Chen, and Wu 2024) seek more comprehensive multimodal modeling approaches to enhance the performance of multi-task visual grounding. SeqTR (Zhu et al. 2022) and PolyFormer (Liu et al. 2023a) employ a sequential transformer model that processes visual and textual data in a unified manner, enhancing performance on multi-task visual grounding by sequentially refining predictions. Recently, LLM-based methods (Lai et al. 2024; Xia et al. 2024) leverage the capabilities of LLM to enforce rule-based serialization of predictions, effectively integrating the REC and RIS tasks into a unified framework. Our work follows the paradigm of MCN, which primarily explores and investigates consistency constraints. However, our proposed C^{3}VG further enhances model consistency prediction through implicit interactions and explicit supervision.

Vision Language Pre-Training (VLP)

Existing VLP models can be broadly categorized into three types. One-stream models (Chen et al. 2020; Lan et al. 2020; Huang et al. 2021) process both image and text inputs in a single stream. They concatenate image and text embeddings and interact cross-modality information throughout the entire feature extraction process. Dual-stream models (Radford et al. 2021; Jia et al. 2021; Li et al. 2022c) employ separate encoders for each modality. These models do not concatenate modalities at the input level; instead, the interaction between pooled image and text vectors occurs at a shallow layer. Dual-stream models with fusion encoder (Li et al. 2022a; Bao et al. 2022; Singh et al. 2022) combine aspects of both one-stream and dual-stream models. They facilitate intermediate interaction between modalities, potentially striking a balance between complexity and performance. Visual grounding fundamentally constitutes one of the downstream task of VLP. CRIS (Wang et al. 2022) and Dynamic MDETR (Shi et al. 2023) apply dual-stream vision-language pre-training models to leverage their feature alignment and enhanced modality representation capabilities. SimVG (Dai et al. 2024) decouples the concept of multimodal mutual understanding from downstream tasks with limited data to the pre-training phase, achieving significant performance improvements in REC tasks. This paper further addresses the issue of insufficient multimodal understanding for multi-task joint training by employing multimodal fusion representations pre-training method (Wang et al. 2023).

The Proposed C^{3}VG

Architecture Overview

Fig. 3 provides an overview of the C^{3}VG architecture. Initially, the image and text modalities are independently embedded and processed through a multi-modality encoder (MME) for vision-language encoding and fusion, positioning the joint representation of multimodal fusion upstream. A learnable object token is also utilized as the feature representation for the REC task. The framework then advances through the RSP and RCI stages, ultimately yielding high-quality predictions.

Multi-Modality Encoder. The input to C^{3}VG consists of an image $I \in \mathbb{R}^{3\times H\times W}$ and a caption text $T \in \Omega^M$, where Ω denotes the vocabulary set. The image is initially downsampled to 1/16 of its original size using a visual embedding, resulting in $P_i = \{p^1, p^2, \ldots, p^{N_i}\}$. The text is then tokenized into $L_t = \{l^1, l^2, \ldots, l^{N_t}\}$. Additionally, we define a learnable object token T_o as the target feature for the REC branch. The inputs of MME can be expressed as:

$$\mathrm{T} = \{T_o, p^1, p^2 \ldots, p^{N_i}, l^1, l^2, ..., l^{N_t}\}. \tag{1}$$

The MME architecture leverages the pre-trained weights of the BEiT-3 (Wang et al. 2023) model. The output of the MME comprises three components: $T_o \in \mathbb{R}^{B\times 1\times C}$, $T_t \in \mathbb{R}^{B\times N_t\times C}$, $T_i \in \mathbb{R}^{B\times N_i\times C}$.

Rough Semantic Perception Stage. The RSP stage aims to generate a rough localization and semantic outline, serving as priors for the RCI stage. Initially, the outputs of the MME are projected to a common dimension via three unshared linear layers:

$$T_o^{'} = \mathrm{OP}(T_o), \quad T_t^{'} = \mathrm{TP}(T_t), \quad T_i^{'} = \mathrm{IP}(T_i). \tag{2}$$

For the REC branch, the process begins with a query decoder, which enhances the representation of the object token by interacting with text and image tokens. The query decoder is defined as:

$$\begin{aligned} T_c = \mathrm{MCA}(\mathrm{MLP}(\mathrm{Concat}(T_o^{'}, \mathrm{MCA}(T_o^{'} + Q_{init}, \\ T_t^{'} + pos_{1d}))), T_i^{'} + pos_{2d}), \end{aligned} \tag{3}$$

where MCA(A_1, A_2) denotes the multi-head cross attention mechanism, with A_1 serving as the query and A_2 as the key and value. Subsequently, an MLP is employed to regress and predict the REC output $P_b^c \in \mathbb{R}^{B\times 4}$. For the RIS branch, we adopt a text-to-pixel correlation strategy similar to CRIS (Wang et al. 2022) to generate the predicted mask $P_s^c \in \mathbb{R}^{B\times H\times W}$. However, instead of using a 3×3 convolution with padding, we compress the text using a 1×1 convolution without additional padding.

Refined Consistency Interaction Stage. The Refined Consistency Interaction (RCI) stage is designed to harmonize the outputs from the RSP stage, ensuring multi-task consistency through both implicit interactions and explicit constraints. We first introduce a mask-guided interaction module (MIM) that adaptively and implicitly aligns the consistency between the detection and segmentation predictions. Additionally, an auxiliary bidirectional consistency constraint loss is incorporated to explicitly enforce alignment at the result level. In the REC branch, an MLP layer is utilized to regress object features at the RCI stage. In the RIS branch, we integrate SimFPN (Li et al. 2022d) to capture multi-level structures, followed by a UNet-style (Ronneberger, Fischer, and Brox 2015) decoder that performs multi-level fusion and a pixel decoder, consistent with the methodology employed in the RSP stage.

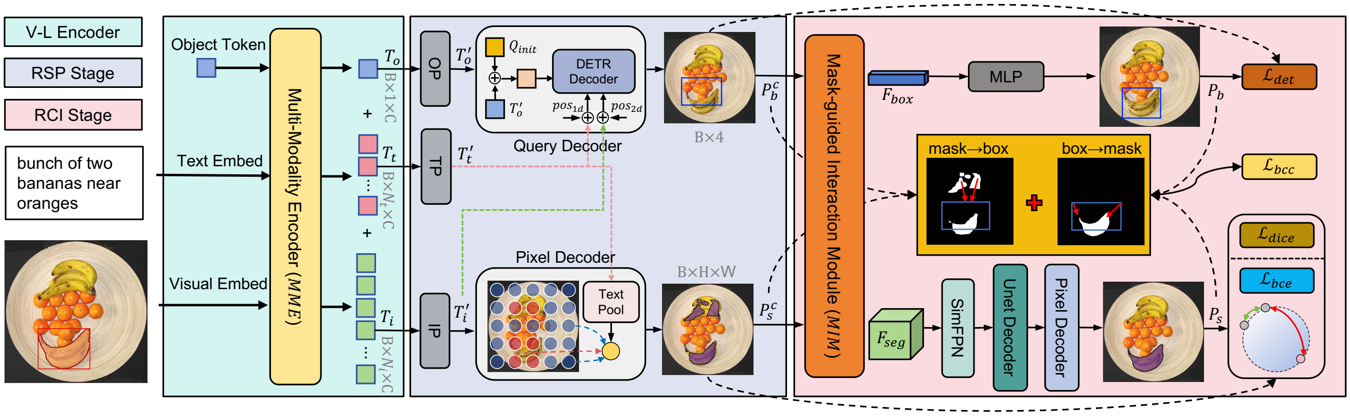

Figure 3: The overall framework of the proposed C[3]VG. First, the image and text features are fused and encoded using a multi-modality encoder. In the RSP stage, the pixel decoder and query decoder generate coarse segmentation and detection results. In the RCI stage, these multi-task priors are further refined through interaction and consistency constraints.

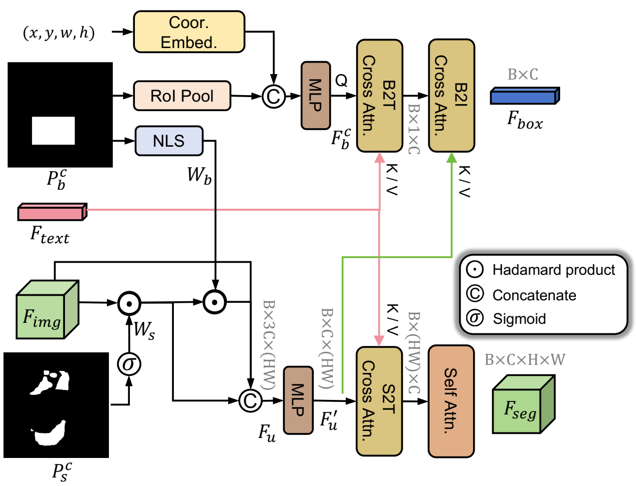

Figure 4: Architecture of the Mask-guided Interaction Module (MIM). "Coor. Embed" denotes a linear layer that maps coordinate positions into the hidden space.

Mask-guided Interaction Module

The RSP stage provides spatial prior information for the RCI stage, while the MIM is designed to implicitly model the relationships between the multi-task results from the RSP stage in a learnable manner. In the REC branch, based on the detection results from the RSP stage $P_b^c \in \mathbb{R}^{B\times 4}$, which are represented as (x, y, w, h), two operations are performed. (1) The results are used as the ROI to pool features from F_{img}. (2) Coordinate representations are obtained through coordinate embedding (CoE). The RSP stage box feature F_b^c is then computed as follows:

$$F_b^c = \text{MLP}(\text{Concat}(\text{RoIP}(P_b^c, F_{img}) + \text{CoE}(P_b^c))). \quad (4)$$

where RoIP denotes the RoI Pooling operation as in Faster R-CNN (Ren et al. 2015). To enable the bounding box to utilize the structural information from the RIS branch and ensure consistent predictions, we interact F_b^c with both textual and visual features. The final interacted object feature F_{box} is expressed as:

$$F_{box} = \text{MCA}(\text{MCA}(F_b^c, F_{text}), F_u^{'}), \quad (5)$$

where the calculation of $F_u^{'}$ is detailed in Eq. 10.

In the RIS branch, we apply the concept of background suppression and foreground enhancement by leveraging the results of both the REC and RIS branches on F_{img}. First, P_b^c is converted to the top-left and bottom-right format by rounding to integers as follows:

$$\begin{aligned} x_1 &= (x - 0.5w) \times w, \;\; y_1 = (y - 0.5h) \times h, \\ x_2 &= (x + 0.5w) \times w, \;\; y_2 = (y + 0.5h) \times h, \end{aligned} \quad (6)$$

$$\{x_1^{'}, y_1^{'}, x_2^{'}, y_2^{'}\} = \{\lfloor x_1 \rfloor, \lfloor y_1 \rfloor, \lceil x_2 \rceil, \lceil y_2 \rceil\}, \quad (7)$$

where $\lfloor * \rfloor$ denotes the floor function, and $\lceil * \rceil$ denotes the ceiling function. The NLS generates a weight mask W_b of the same dimensions as F_{img}, calculated as follows:

$$W_b = \begin{cases} w_1, & \text{if } x_i \in [x_1^{'}, x_2^{'}] \wedge y_j \in [y_1^{'}, y_2^{'}] \\ 1, & \text{otherwise}, \end{cases} \quad (8)$$

where $\forall x_i \in [0, w]$ and $\forall y_j \in [0, h]$. w_1 is set to default values of 0.1, respectively. We then apply a sigmoid function to the predicted mask from the RSP stage to generate the weighted mask $W_s = \sigma(P_s^c)$. The weights W_b and W_s are applied to F_{img} to obtain the box and mask-constrained feature F_u:

$$\begin{aligned} F_s &= W_s \odot F_{img}, \\ F_u &= \text{Concat}(F_s, W_b \odot F_s, F_{img}). \end{aligned} \quad (9)$$

Next, an MLP reduces the channel dimension from $3 \times C$ back to the original C, yielding the fused image representation $F_u^{'}$, which incorporates the predictions from the RSP stage. This process implicitly provides the RCI stage with prior spatial attention information derived from detection and segmentation predictions. As illustrated in Fig. 5, the

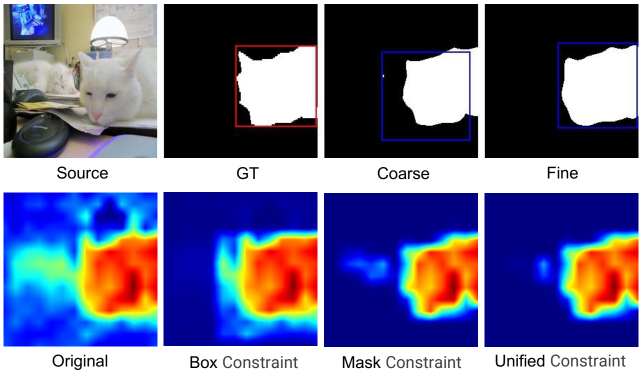

Figure 5: Visualization of intermediate model processes. First row: original image, GT, RSP stage, and RCI stage results. Second row: original, box-constrained, mask-constrained, and unified-constrained heatmaps.

presence of two cats results in divergent attention predictions, leading to suboptimal adjustments of the bounding box prediction during the RSP stage. The MIM mitigates this issue by imposing constraints on the regions of high response within the image space, thereby reducing the model's focus on irrelevant targets and enabling more precise target identification. Furthermore, the fused image representation is interacted with the text, followed by a multi-head self-attention (MSA) layer to further learn consistent semantic associations. This process is expressed as follows:

$$\begin{aligned} F_u^{'} &= \mathrm{MLP}(F_u), \\ F_{seg} &= \mathrm{MSA}(\mathrm{MCA}(F_u^{'}, F_{text})). \end{aligned} \tag{10}$$

Bidirectional Consistency Constraint Loss

To complement the implicit interactions facilitated by the MIM across multi-task outputs, we propose an explicit bidirectional consistency constraint loss, denoted as $\mathcal{L}_{bcc}$. First, $\mathcal{L}_{m2b}$, is designed to enforce the segmentation mask to be contained within the predicted bbox:

$$\mathcal{L}_{m2b} = 1 - \frac{\sum(M_s \odot M_b)}{\sum M_s}, \tag{11}$$

$$M_s = \begin{cases} 1, & \text{if } p_{i,j}^s > t \\ 0, & \text{otherwise} \end{cases}, M_b = \begin{cases} 1, & \text{if } (x_i, y_j) \in P_b \\ 0, & \text{otherwise} \end{cases}, \tag{12}$$

where $p_{i,j}^s$ denotes the pixel values of the predicted segmentation mask after applying the sigmoid function, with $\forall i \in [0, w]$ and $\forall j \in [0, h]$. t is set to 0.5. P_b represents the bounding box prediction. Second, the loss term $\mathcal{L}_{b2m}$ is defined as follows:

$$\mathcal{L}_{b2m} = 1 - \frac{|P_b^s \cap P_b|}{|P_b^s \cup P_b|}, \tag{13}$$

where P_b^s represents the minimal bounding box that encloses the segmentation mask M_s, and P_b denotes the predicted bounding box. This loss is quantified using the Intersection over Union (IoU) metric, which measures the degree of overlap between the bounding box derived from the segmentation mask and the predicted bounding box. It ensures that the predicted bounding box encapsulates the segmentation mask as comprehensively as possible. Finally, the overall consistency constraint loss is defined as $\mathcal{L}_{bcc} = \lambda_1 \mathcal{L}_{b2m} + \lambda_2 \mathcal{L}_{m2b}$, with the weighting coefficients λ_1 and λ_2 set to 1 and 3, respectively.

Training Objectives

The primary optimization loss for the multi-task visual grounding is comprised of two main components: REC and RIS, which are defined as follows:

$$\begin{aligned} \mathcal{L}_{rec} &= \sigma_{l1} \mathcal{L}_{l1} + \sigma_{giou} \mathcal{L}_{giou}, \\ \mathcal{L}_{ris} &= \sigma_{dice} \mathcal{L}_{dice} + \sigma_{bce} \mathcal{L}_{bce}, \end{aligned} \tag{14}$$

where the weighting factors σ_{l1} and σ_{giou} are set to 0.5 and 0.2, respectively, while σ_{dice} and σ_{bce} are both set to 1.0 by default. Both $\mathcal{L}_{rec}$ and $\mathcal{L}_{ris}$ include two-stage components and are augmented by the bidirectional consistency constraint loss, $\mathcal{L}_{bcc}$. The total loss is formulated as:

$$\begin{aligned} \mathcal{L}_{total} = \lambda_c(\lambda_{rec} \mathcal{L}_{rec}^c + \mathcal{L}_{ris}^c) + \\ (\lambda_{rec} \mathcal{L}_{rec}^f + \mathcal{L}_{ris}^f) + \lambda_{bcc} \mathcal{L}_{bcc} \end{aligned} \tag{15}$$

where λ_{rec}, λ_{bcc}, and λ_c are set to 0.5, 0.1, and 0.3, respectively. Here, $\mathcal{L}_{rec}^c$ denotes the REC loss in the RSP stage, while $\mathcal{L}_{ris}^f$ corresponds to the RIS loss in the RCI stage.

Experiments

Experimental Setup

We evaluate the proposed model in RefCOCO (Yu et al. 2016), RefCOCO+ and RefCOCOg (Nagaraja, Morariu, and Davis 2016) datasets. The maximum sentence length is set to 20. The images are resized to 320×320. Based on previous works (Zhu et al. 2022), mIoU and Prec@0.5(Acc(REC) in ablation study) are adopted to evaluate the performance of methods. We train our models for 30 epochs with a batch size of 16. Adam (Kingma and Ba 2014) is adopted as our optimizer. All experiments are conducted on a system with dual NVIDIA 4090 GPUs. Further details will be provided in the supplementary materials.

Main Results

Referring Expression Comprehension. The single-task part presented in Tab. 1 showcase a comparison between our method and prior advanced REC approaches. In comparison to Dynamic MDETR, which utilizes ViT-B as its backbone, C^{3}VG achieves a remarkable improvement of +5.78%-17.98% in Acc(REC). Furthermore, when compared to GroundingDINO (Liu et al. 2023b), which is trained on large-scale data, C^{3}VG delivers a gain of +2.72%-6.71% in Acc(REC) while also reducing inference latency by 58%.
Referring Image Segmentation. The single-task part presented in Tab. 2 compare our C^{3}VG with previous advanced RIS methods. Our C^{3}VG demonstrates an absolute improvement of 9.75%-18.72% over the Transformer-based CRIS (Wang et al. 2022) model. Additionally, it achieves +1.72%-8.15% in mIoU compared to the latest SOTA model Prompt-RIS (Shang et al. 2024), under the same ViT-B backbone conditions.

Method	Publication	Backbone	Data Size	RefCOCO			RefCOCO+			RefCOCOg		Time
				val	test A	test B	val	test A	test B	val(U)	test(U)	(ms)
Single-task												
MDETR (Kamath et al. 2021)	ICCV2021	EfficientNet-B3	200K	86.75	89.58	81.41	79.52	84.09	70.62	81.64	80.89	108
Dyn.MDETR (Shi et al. 2023)	T-PAMI2023	ViT-B	-	85.97	88.82	80.12	74.83	81.70	63.44	72.21	74.14	-
GroundingDINO (Liu et al. 2023b)	ECCV2024	Swin-T	200K	89.19	91.86	85.99	81.09	87.40	74.71	84.15	84.94	120
SimVG (Dai et al. 2024)	NeurIPS2024	BEiT3-ViT-B	174K	<u>90.59</u>	92.80	87.04	83.54	88.05	77.50	85.38	<u>86.28</u>	**44**
Multi-task												
SeqTR (Zhu et al. 2022)	ECCV2022	DarkNet53	174K	81.23	85.00	76.08	68.82	75.37	58.78	71.35	71.58	<u>50</u>
PolyFormer (Liu et al. 2023a)	CVPR2023	Swin-B	174K	89.73	91.73	86.03	83.73	88.60	76.38	84.46	84.96	152
EEVG (Chen, Chen, and Wu 2024)	ECCV2024	ViT-B	174K	90.47	92.73	<u>87.72</u>	81.79	87.80	74.94	85.19	84.72	117
Generalist Models												
Ferret (You et al. 2023)	ICLR2024	Vicuna-7B	> 8M	87.49	91.35	82.45	80.78	87.38	73.14	83.93	84.76	-
LION-12B (Chen et al. 2024)	CVPR2024	FlanT5-11B	3.6M	89.80	<u>93.02</u>	85.57	<u>83.95</u>	<u>89.22</u>	<u>78.06</u>	<u>85.52</u>	85.74	-
C^3VG	AAAI2024	ViT-B	28K	**92.51**	**94.60**	**88.71**	**87.44**	**90.69**	**81.42**	**87.68**	**88.31**	51

Table 1: **Main results** on REC datasets. **Bold** denotes the best performance. <u>Underline</u> denotes the second best performance.

Method	Publication	Backbone	Data	FT	RefCOCO			RefCOCO+			RefCOCOg	
					val	test A	test B	val	test A	test B	val(U)	test(U)
Single-task												
LAVT (Yang et al. 2022)	CVPR2022	Swin-B	RefC	✘	74.46	76.89	70.94	65.81	70.97	59.23	63.34	63.62
ReLA (Liu, Ding, and Jiang 2023)	CVPR2023	Swin-B	RefC	✘	73.82	76.48	70.18	66.04	71.02	57.65	65.00	65.97
Prompt-RIS (Shang et al. 2024)	CVPR2024	CLIP-ViT-B	Com-RefC	-	78.10	81.21	74.64	71.13	76.60	64.25	70.47	71.29
OneRef (Xiao et al. 2024)	NeurIPS2024	BEiT3-ViT-B	Com-RefC	✔	<u>79.83</u>	<u>81.86</u>	76.99	<u>74.68</u>	<u>77.90</u>	<u>69.58</u>	<u>74.06</u>	<u>74.92</u>
Multi-task												
PolyFormer (Liu et al. 2023a)	CVPR2023	Swin-B	Com-RefC	✔	75.96	77.09	73.22	70.65	74.51	64.64	69.36	69.88
PVD (Cheng et al. 2024)	AAAI2024	Swin-B	Com-RefC	✔	74.82	77.11	69.52	63.38	68.60	56.92	63.13	63.62
EEVG (Chen, Chen, and Wu 2024)	ECCV2024	ViT-B	Com-RefC	-	79.49	80.87	<u>77.39</u>	71.86	76.67	66.31	73.56	73.47
Generalist Models												
LISA (Lai et al. 2024)	CVPR2024	Vicuna-7B	-	✔	74.90	79.10	72.30	65.10	70.80	58.10	67.90	70.60
GSVA (Xia et al. 2024)	CVPR2024	Vicuna-7B	-	✔	77.20	78.90	73.50	65.90	69.60	59.80	72.70	73.30
C^3VG	AAAI2024	BEiT3-ViT-B	Com-RefC	✘	**81.37**	**82.93**	**79.12**	**77.05**	**79.61**	**72.40**	**76.34**	**77.10**
C^3VG-oIoU	AAAI2024	BEiT3-ViT-B	Com-RefC	✘	80.89	83.18	77.86	74.68	77.96	68.95	74.43	76.39

Table 2: **Main Results** on RIS Datasets. **Bold** indicates the best performance, and <u>underline</u> indicates the second-best performance. RefC represents training on a single dataset, while Com-RefC refers to the union of the RefCOCO, RefCOCO+, and RefCOCOg training sets. FT denotes whether fine-tuning is performed on the specific dataset.

Multi-Task Visual Grounding. The multi-task results presented in Tab. 1 and Tab. 2 provide a comparative analysis between the proposed C^3VG and existing multi-task visual grounding approaches. Compared to PolyFormer (Liu et al. 2023a), our C^3VG demonstrates marked improvements, surpassing it by margins of +2.09%-5.04% in Acc(REC) and +5.10%-7.76% in mIoU. Furthermore, our method exhibits inference efficiency comparable to that of SeqTR, nearing real-time performance.

Generalist Models. Multimodal Large Language Models (Jin et al. 2024) have also expanded into the visual grounding domain, with their results listed under the generalist models part in Tab. 1 and Tab. 2. These models are distinguished by enormous parameters and extensive pretraining on vast datasets, providing strong generalization capabilities. However, our method demonstrates strong competitiveness compared to these generalist models.

Ablation Studies

Basic Improvement Setting. We implement several techniques to enhance the performance of our baseline model, with the experimental outcomes presented in Tab. 3. The baseline architecture leverages the ViT-B and BERT models as the visual and textual encoders, respectively, with VGTR head. First, we observe a substantial performance boost by incorporating multimodal fusion representation pretraining (BEiT-3), which yields an increase of +5.11% in Acc(REC) and +5.28% in oIoU. This improvement can be attributed to the fact that prior methods often rely on limited downstream data to learn multimodal representations, resulting in inadequate multimodal comprehension. Given the complex and rich semantics inherent in text, pretraining multimodal representations is essential for achieving sophisticated multimodal understanding. Furthermore, the joint training of REC and RIS has shown a mutually beneficial effect, leading to an improvement of +1.68% in Acc(REC) and +1.47% in oIoU. Finally, the integration of SimFPN, which facilitates comprehensive interaction across multi-level features, further enhances oIoU by an additional +1.05%.

Query / Pixel Decoder. The query decoder is designed to integrate guidance from both textual and visual modalities into the tokens utilized by the detection branch, thereby improving localization accuracy. As demonstrated in Tab. 3, the incorporation of the query decoder leads to a +2.26% in-

Method	Acc (REC)	Acc (RIS)	oIoU (RIS)
Baseline	75.33	74.21	62.21
+ MM Pretrain	80.44 $\uparrow$ 5.11	80.08 $\uparrow$ 5.87	67.49 $\uparrow$ 5.28
+ Multi-Task	82.12 $\uparrow$ 1.68	81.78 $\uparrow$ 1.70	68.96 $\uparrow$ 1.47
+ SimFPN	82.25 $\uparrow$ 0.13	82.42 $\uparrow$ 1.36	70.01 $\uparrow$ 1.05
+ Query Decoder	84.51 $\uparrow$ 2.26	82.12 $\downarrow$ 0.30	69.81 $\downarrow$ 0.20
+ Pixel Decoder	84.35 $\downarrow$ 0.16	83.23 $\uparrow$ 1.11	70.81 $\uparrow$ 1.00

Table 3: Ablation study on basic improvement settings.

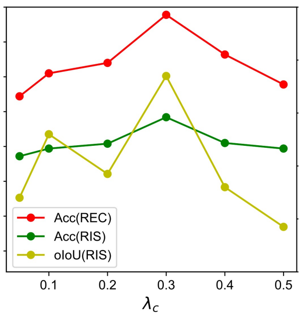

Figure 6: Ablation study on weight of RSP stage λ_c.

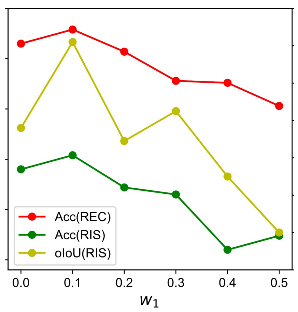

Figure 7: Ablation study on w_1 in MIM.

crease in Acc(REC). The pixel decoder, on the other hand, estimates the confidence of each pixel belonging to the foreground through text-pixel contrastive learning. This addition strengthens the supervision within the segmentation branch, resulting in a +1.00% enhancement in oIoU.

Consistency Constraint Loss. This paper introduces two directions of consistency constraint losses for optimization: mask→box ($\mathcal{L}_{m2b}$) and box→mask ($\mathcal{L}_{b2m}$). The purpose of $\mathcal{L}_{m2b}$ is to align the RIS-predicted mask distribution with the REC-predicted bounding box. In contrast, $\mathcal{L}_{b2m}$ is designed to ensure that the REC-predicted bounding box encompasses the RIS-predicted mask while concurrently suppressing extraneous predictions in non-relevant regions. As demonstrated in Tab. 4, both the $\mathcal{L}_{m2b}$ and $\mathcal{L}_{b2m}$ constraints positively influence performance in both REC and RIS tasks. Moreover, integrating them to establish bidirectional consistency constraints results in further performance enhancements, yielding +1.20% in Acc(REC) and +1.95% in oIoU.

Mask-guided Interaction Module. As illustrated in Tab. 5, MIM introduces a coarse-to-fine learning paradigm, where the RCI stage demonstrates significant improvements over the RSP stage, particularly in segmentation-related metrics. Moreover, the integration of text interaction further enhances Acc(REC) by +0.60%, with minimal impact on RIS metrics. The term 'Coor. Embed.' pertains to encoding the

$\mathcal{L}_{b2m}$	$\mathcal{L}_{m2b}$	Acc (REC)	Acc (RIS)	oIoU (RIS)
		84.35	83.23	70.81
✔		85.12 $\uparrow$ 0.77	83.96 $\uparrow$ 0.73	72.13 $\uparrow$ 1.32
	✔	85.01 $\uparrow$ 0.76	84.38 $\uparrow$ 1.15	72.24 $\uparrow$ 1.43
✔	✔	**85.55** $\uparrow$ **1.20**	**84.59** $\uparrow$ **1.36**	**72.74** $\uparrow$ **1.93**

Table 4: Ablation study on consistency constraint loss.

Method	Acc (REC)	Acc (RIS)	oIoU (RIS)
RSP Stage	84.30	83.00	70.48
RCI Stage	84.54 $\uparrow$ 0.34	84.15 $\uparrow$ 1.15	71.95 $\uparrow$ 1.47
REC branch			
+Text Attn.	85.14 $\uparrow$ 0.60	84.09 $\downarrow$ 0.06	71.83 $\downarrow$ 0.12
+Coor. Embed	85.41 $\uparrow$ 0.27	84.08 $\downarrow$ 0.01	71.98 $\uparrow$ 0.15
Interaction type			
Box	85.93 $\uparrow$ 0.52	84.22 $\uparrow$ 0.14	71.89 $\downarrow$ 0.09
Mask	85.31 $\downarrow$ 0.10	84.63 $\uparrow$ 0.55	72.71 $\uparrow$ 0.73
Unified	**86.05** $\uparrow$ **0.64**	**84.80** $\uparrow$ **0.72**	**72.98** $\uparrow$ **1.00**

Table 5: Ablation study on unified interaction module.

RSP stage's prediction results (x, y, w, h), which results in a 0.27% increase in Acc(REC). In the RIS branch, we conducted ablation studies to assess the introduction of various prior information from the coarse stage, as detailed in the 'Interaction type' section of Tab. 5. These studies reveal that incorporating box interaction further strengthens the REC branch. This enhancement is attributed to the interaction between two stages, wherein the RCI stage imposes more stringent requirements on the prediction box generated by the RSP stage. Additionally, the effect of the background weight w_1 in NLS is depicted in Fig. 7, with $w_1 = 0.1$ employed as the default value in this study. Similarly, utilizing the mask prior from the RSP stage further improves segmentation performance in the RCI stage. Finally, unified interaction improves performance by concurrently integrating positional and semantic priors from the RSP stage. By leveraging the complementary information from both sources, it constructs a consistent multi-task representation. As evidenced by the visualization in Fig. 5, this implicit constraint functions as a foreground feature extraction mechanism. Unlike the post-processing employed in MCN (Luo et al. 2020), MIM utilizes an implicit, learnable modeling approach to interact with multi-task results, thereby achieving consistent representations. Fig. 6 illustrates the impact of the weight proportion of the coarse stage on the loss calculation. Ultimately, $\lambda_c = 0.3$ is adopted as the default value.

Conclusion

In this paper, we present C^3VG, a coarse-to-fine architecture designed for multi-task visual grounding, aimed at addressing issues of prediction inconsistency and inadequate multimodal comprehension. Initially, during the Rough Semantic Perception (RSP) stage, we extract coarse spatial locations and semantic boundaries using query and pixel decoders. Subsequently, we introduce a mask-guided interaction module to implicitly refine predictions from the RSP stage, while a bidirectional consistency constraint loss explicitly enforces coherence during the Refined Consistency Interaction (RCI) stage. Furthermore, to address the challenge of insufficient multimodal understanding, we validate the effectiveness of extending the multimodal encoder from a single-task setting to a multi-task joint training framework. Empirical evaluations substantiate the efficacy and soundness of C^3VG, which outperforms the existing advanced REC and RIS methods by a remarkable margin.

Acknowledgments

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).

References

Bao, H.; Wang, W.; Dong, L.; Liu, Q.; Mohammed, O. K.; Aggarwal, K.; Som, S.; Piao, S.; and Wei, F. 2022. VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chen, G.; Shen, L.; Shao, R.; Deng, X.; and Nie, L. 2024. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26540–26550.

Chen, W.; Chen, L.; and Wu, Y. 2024. An Efficient and Effective Transformer Decoder-Based Framework for Multi-Task Visual Grounding.

Chen, Y.-C.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision (ECCV)*.

Cheng, Z.; Li, K.; Jin, P.; Li, S.; Ji, X.; Yuan, L.; Liu, C.; and Chen, J. 2024. Parallel vertex diffusion for unified visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 1326–1334.

Dai, M.; Yang, L.; Xu, Y.; Feng, Z.; and Yang, W. 2024. SimVG: A Simple Framework for Visual Grounding with Decoupled Multi-modal Fusion. *Advances in Neural Information Processing Systems (NeurIPS)*.

Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; and Fu, J. 2021. Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139, 4904–4916.

Jin, Y.; Li, J.; Liu, Y.; Gu, T.; Wu, K.; Jiang, Z.; He, M.; Zhao, B.; Tan, X.; Gan, Z.; et al. 2024. Efficient multimodal large language models: A survey. *arXiv preprint arXiv:2405.10739*.

Kamath, A.; Singh, M.; LeCun, Y.; Synnaeve, G.; Misra, I.; and Carion, N. 2021. MDETR-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1780–1790.

Kim, N.; Kim, D.; Lan, C.; Zeng, W.; and Kwak, S. 2022. ReSTR: Convolution-free Referring Image Segmentation Using Transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 18145–18154.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning (ICML)*, 5583–5594.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9579–9589.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International conference on machine learning (ICML)*.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022b. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10965–10975.

Li, M.; and Sigal, L. 2021. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in Neural Information Processing Systems (NeurIPS)*, 34.

Li, Y.; Liang, F.; Zhao, L.; Cui, Y.; Ouyang, W.; Shao, J.; Yu, F.; and Yan, J. 2022c. Supervision Exists Everywhere: A Data Efficient Contrastive Language-Image Pre-training Paradigm. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Li, Y.; Mao, H.; Girshick, R.; and He, K. 2022d. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision (ECCV)*, 280–296. Springer.

Liu, C.; Ding, H.; and Jiang, X. 2023. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 23592–23601.

Liu, D.; Zhang, H.; Wu, F.; and Zha, Z.-J. 2019. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4673–4682.

Liu, J.; Ding, H.; Cai, Z.; Zhang, Y.; Satzoda, R. K.; Mahadevan, V.; and Manmatha, R. 2023a. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18653–18663.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Liu, S.-A.; Zhang, Y.; Qiu, Z.; Xie, H.; Zhang, Y.; and Yao, T. 2023c. CARIS: Context-Aware Referring Image Segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia (ACMMM)*.

Luo, G.; Zhou, Y.; Sun, X.; Cao, L.; Wu, C.; Deng, C.; and Ji, R. 2020. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10034–10043.

Nagaraja, V. K.; Morariu, V. I.; and Davis, L. S. 2016. Modeling context between objects for referring expression understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 792–807.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, 8748–8763.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 28.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241.

Shang, C.; Song, Z.; Qiu, H.; Wang, L.; Meng, F.; and Li, H. 2024. Prompt-Driven Referring Image Segmentation with Instance Contrasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4124–4134.

Shi, F.; Gao, R.; Huang, W.; and Wang, L. 2023. Dynamic mdetr: A dynamic multimodal transformer decoder for visual grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. FLAVA: A Foundational Language And Vision Alignment Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Su, W.; Miao, P.; Dou, H.; and Li, X. 2024. ScanFormer: Referring Expression Comprehension by Iteratively Scanning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13449–13458.

Su, W.; Miao, P.; Dou, H.; Wang, G.; Qiao, L.; Li, Z.; and Li, X. 2023. Language adaptive weight generation for multi-task visual grounding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 10857–10866.

Tang, J.; Zheng, G.; Shi, C.; and Yang, S. 2023. Contrastive grouping with transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 23570–23580.

Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O. K.; Singhal, S.; Som, S.; and Wei, F. 2023. Image as a foreign language: BEiT pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Z.; Lu, Y.; Li, Q.; Tao, X.; Guo, Y.; Gong, M.; and Liu, T. 2022. CRIS: Clip-driven referring image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 11686–11695.

Xia, Z.; Han, D.; Han, Y.; Pan, X.; Song, S.; and Huang, G. 2024. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3858–3869.

Xiao, L.; Yang, X.; Peng, F.; Wang, Y.; and Xu, C. 2024. OneRef: Unified One-tower Expression Grounding and Segmentation with Mask Referring Modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yang, L.; Wang, Y.; Li, X.; Wang, X.; and Yang, J. 2024. Fine-grained visual prompting. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Yang, Z.; Chen, T.; Wang, L.; and Luo, J. 2020. Improving one-stage visual grounding by recursive sub-query construction. In *European Conference on Computer Vision (ECCV)*, 387–404.

Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; and Torr, P. H. 2022. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 18155–18165.

You, H.; Zhang, H. A.; Cao, L.; Gan, Z.; Zhang, B.; Wang, Z.; Du, X.; Chang, S.-F.; and Yang, Y. 2023. FERRET: Refer and Ground Anything Anywhere at Any Granularity. In *Internatioanl Conference of Learning Representations (ICLR)*.

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018. MAttNet: Modular Attention Network for Referring Expression Comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *European Conference on Computer Vision (ECCV)*, 69–85.

Zhang, Z.; Zhu, Y.; Liu, J.; Liang, X.; and Ke, W. 2022. Coupalign: Coupling word-pixel with sentence-mask alignments for referring image segmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 14729–14742.

Zhu, C.; Zhou, Y.; Shen, Y.; Luo, G.; Pan, X.; Lin, M.; Chen, C.; Cao, L.; Sun, X.; and Ji, R. 2022. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision (ECCV)*, 598–615. Springer.

Zhuang, J.; Hu, J.; Mu, L.; Hu, R.; Liang, X.; Ye, J.; and Hu, H. 2025. FALIP: Visual Prompt as Foveal Attention Boosts CLIP Zero-Shot Performance. In *European Conference on Computer Vision (ECCV)*, 236–253.