CMIRNet: Cross-Modal Interactive Reasoning Network for Referring Image Segmentation

Mingzhu Xu, *Member, IEEE*, Tianxiang Xiao, Yutong Liu, Haoyu Tang, *Member, IEEE*, Yupeng Hu, *Member, IEEE*, and Liqiang Nie, *Senior Member, IEEE*

***Abstract*— Referring Image Segmentation (RIS) aims to semantically segment the target object (referent) in alignment with the provided natural language query. Existing works still suffer from that the non-referent was segmented mistakenly, which can be attributed to the insufficient comprehension of vision and language. To tackle this problem, we propose a Cross-Modal Interactive Reasoning Network (CMIRNet) to explore semantic information that consistently existed between vision and language. Specifically, we first devise a novel Text-Guided Multi-Modality Joint Encoder (TGMM-JE), where the key expression can be extracted and the important visual features will be encoded under the continuous guidance of language expression. Then, we design a Cross-Graph Interactive Positioning (CGIP) module to locate the key pixels of the referent object in deepest layer. The multi-modality graph data is constructed between visual and linguistic features, and the important pixels can be positioned from cross-graph interaction and intra-graph reasoning. Finally, a novel Cross-Modal Attention Enhanced DEcoder (CMAE-DE) is dedicated to refine the referent object mask from coarse to fine progressively, where hybrid cross modal attentions are explored to enhance the representation of referent object. Extensive ablation studies validate the efficacy of our key modules and comprehensive experimental results show the superiority of our proposed model over 22 state-of-the-art (SOTA) models.**

***Index Terms*— Referring image segmentation, vision and language, cross modal reasoning, graph neural network.**

I. Introduction

GIVEN an image and a natural language query, Referring Image Segmentation (RIS) aims to predict the segmentation mask of the object referenced in the query [1], [2]. Unlike traditional semantic segmentation [3], which assigns pixels to predefined categories, RIS poses a greater challenge due to the unrestricted language expressions. It requires a holistic understanding of language and vision, locating the referent from complex scenes, and predicting a segmentation mask based on key linguistic cues. Despite its promising potential in human-robot interaction [4], [5] and image editing [6], [7], RIS remains a challenging yet unsolved multi-modal task.

Recent advances in deep learning have led to significant progress in RIS. Earlier works [1], [2], [8] propose to extract visual and linguistic features separately, and fuse them by conducting simple concatenation or dot product, then generate the final mask using the fused features. Jiao et al. [9] and Yang et al. [10] fused cross-modal feature using dot products, refining multi-modality features with various levels of visual information. But this basic fusion method struggles to align semantics across modalities. Luo et al. [11] solely relied on sentence-level semantics to produce text-related visual features, neglecting word-level analysis and thus struggling to effectively fuse multi modality features. Yang et al. [12] introduced unidirectional fusion, incorporating original text information into visual features, but may not fully exploit visual richness for unique text generation. Ding et al. [13] proposed vision-guided multi-text generation, relying solely on unidirectional visual cues, lacks effective cross-modal alignment. Bidirectional cross-modal attentions [14], [15], [16] model the dependencies between visual and linguistic information, and generate the discriminative features by vision-language mutually guided learning. However, these cross attentions parallelly integrate vision and text, yet inadequately use new visual or new textual cues to bolster multi-modal feature alignment, and own high complexity for multiple cross attention operations. As visual examples shown in Fig.1, despite their impressive performance in simple scenes, these advanced models still struggle with accurately localizing referents, leading to segmentation errors. It can be attributed to the insufficient comprehension of vision and language information.

To tackle the aforementioned challenges, we propose a Cross-Modal Interactive Reasoning Network (CMIRNet), designed to unveil the consistent semantic connections between vision and language. CMIRNet comprises three pivotal modules: Text-Guided Multi-Modality Joint Encoder (TGMM-JE), Cross-Graph Interactive Positioning (CGIP) module, and Cross-Modal Attention Enhanced DEcoder (CMAE-DE). Initially, TGMM-JE processes pure visual and linguistic features, leveraging visual cues to extract pivotal language expressions. Subsequently, the important visual features are encoded under the continuous guidance of the new key language expression. Then, we design a CGIP module to locate the key pixels of the referent object in deepest

Received 23 May 2024; revised 25 August 2024 and 8 October 2024; accepted 20 November 2024. Date of publication 29 November 2024; date of current version 7 April 2025. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62206157, Grant 62276155, and Grant 62206156; and in part by the Natural Science Foundation of Shandong Province under Grant ZR2022QF047. This article was recommended by Associate Editor R. Du. *(Corresponding author: Haoyu Tang.)*

Mingzhu Xu, Tianxiang Xiao, Yutong Liu, Haoyu Tang, and Yupeng Hu are with the School of Software, Shandong University, Jinan, Shandong 250101, China (e-mail: xumingzhu@sdu.edu.cn; txxiao1216@gmail.com; yutongliu0112@gmail.com; tanghao258@sdu.edu.cn; huyupeng@sdu.edu.cn).

Liqiang Nie is with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, Guangdong 518055, China (e-mail: nieliqiang@gmail.com).

Digital Object Identifier 10.1109/TCSVT.2024.3508752

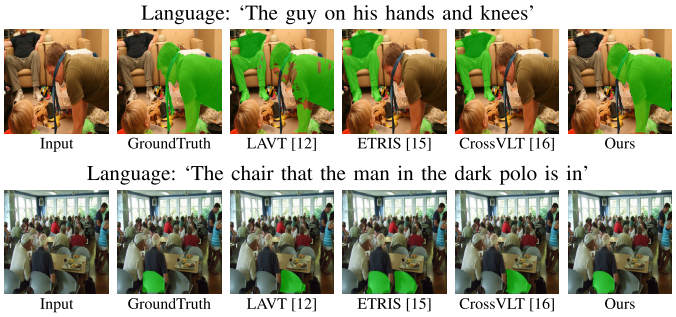


Fig. 1. The visual examples of some non-referent object was segmented mistakenly by several existing advanced methods.

layer. The multi-modality graph data is constructed between visual and language features, and the important pixels can be positioned through cross-graph interaction and intra-graph reasoning. Finally, a novel CMAE-DE is dedicated to refine the referent object mask from coarse to fine progressively. It generates cross-modal spatial attention and cross-modal channel attention to enhance the representation of referent at different levels.

We outline our pivotal contributions as follows:

1. We propose a novel Text-Guided Multi-Modality Joint Encoder (TGMM-JE), which can enhance the language expression incorporating the understanding of visual information. Subsequently, it encodes important visual features under the continuous guidance of the new key language expression.

2. We propose a novel Cross-Graph Interactive Positioning (CGIP) module, where a multi-modality graph is constructed using enhanced visual feature and linguistic feature. This module facilitates the localization of crucial referent pixels by leveraging cross-graph interaction and intra-graph reasoning.

3. We propose a novel Cross-Modal Attention Enhanced DEcoder (CMAE-DE), where hybrid cross-modal attentions are explored to enhance the representation of referent object from deep layers to shallow layers progressively.

4. We have conducted comprehensive ablation studies and experimental analyses on four challenging datasets. The ablation experiments effectively validate the efficacy of our core modules, while exhaustive experimental results confirm the superiority of our CMIRNet against twenty-two SOTA models.

II. Related Works

This section will concisely review the most pertinent works that align with our approach, focusing on semantic segmentation, referring image segmentation, and cross-modal fusion.

A. Semantic Segmentation

Semantic segmentation is a classical task in computer vision, where the objective is to assign each pixel to a predefined category. Most advanced models adopt the Fully Convolution Network (FCN) architecture [17], [18], which is dominant in pixel-level dense prediction tasks. The down-sampling operation in feature encoding stage may result into the coarse boundaries in segmentation task. To tackle this problem, the encoder-decoder structure [18], [19] is adopted to propagate the semantic information from deep layers to shallow layers progressively, reducing the loss of detail information. To extract global contextual information, a multi-scale pyramid pooling module [20] is devised to perceive scale-variant contextual information. Moreover, atrous spatial pyramids [21] are applied in parallel to effectively extract multi-scale information, by enlarging the visual receptive field. In addition, many advanced weakly supervised [22], semi-supervised [23] or unsupervised semantic segmentation [24] methods are explored in depth to reduce time-consuming manual annotation. In general, these progresses in semantic segmentation have provided crucial foundations for RIS task.

B. Referring Image Segmentation

The objective of Referring Image Segmentation (RIS) is to identify the referent object and generate an accurate segmentation mask for it, guided by a linguistic expression [1]. It is a challenging task, which requires to holistically understand the vision and language information. Most advanced RIS models jointly process visual and linguistic features to obtain a distinguished visual feature representation, which is adopted to determine whether each pixel belongs to the referent foreground or non-referent background. Recent advancement in deep learning boosts the performance of RIS. The general process of RIS can be decoupled into three stages, which are feature extraction, cross-modal feature interaction and multi-modality feature analysis. For feature extraction, Jiao et al. [9] enriched visual feature by retrieving an external data pool to enrich the visual cues. Ding et al. [13] enhanced language expression by generating a series of query vectors. For cross-modal feature interaction or multi-modality feature analysis, the critical issue is to enhance the response of relevant regions and suppress the irrelevant regions [25], [26], [27]. The dependency tree [10], [28], [29] derived from the language expression is devised to guide the context modeling. The cross-modal attentions [11], [13], [30], [31] are explored to analyze the correlation between visual and linguistic feature, aiming to pinpoint image regions referenced by the language.

To accommodate arbitrary text queries, the field has embraced the challenge of open-vocabulary image segmentation, as detailed in seminal works such as [32], [33], [34], and [35]. The pivotal aspect of this task lies in overcoming the tendency to collapse on known categories while simultaneously boost the model's ability to segment objects belonging to unknown classes. Furthermore, to unify different image segmentation tasks under a single framework, Zou et al. [36] proposed a universal segmentation decoder that adapts to various task prompts (such as points, boxes, text, etc.). Liu et al. [37] went a step further by reorganizing the diverse distributions of different tasks into a unified text data format, enabling object segmentation at various granularities according to different semantic granularities.

C. Cross-Modal Fusion

In the task of RIS, cross modal fusion is the key process, which targets to align the semantic information commonly existed in visual-linguistic features. Concretely, the classic fusion methods include tensor-based operations, attention

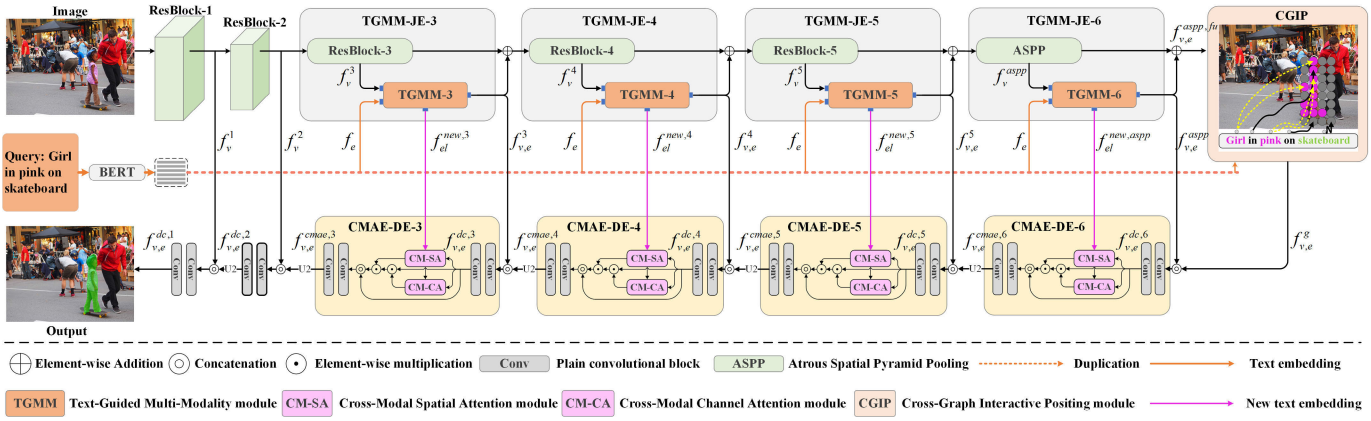


Fig. 2. Architecture of our proposed CMIRNet. The vision-language co-embedding is first encoded in text-guided multi-modality joint encoder (TGMM-JE). Then, the key pixels belonging to the referent object can be located in cross-graph interactive positioning (CGIP) module at deepest layer. Finally, the hybrid cross modal attentions in cross-modal attention enhanced decoder (CMAE-DE) can refine the referent object mask from coarse to fine progressively.

mechanism and graph-based modeling. Tensor-based operations include concatenation [38], multiplication [39] and addition [40]. These simple operations have shown their effectiveness. The attention mechanism [11], [41], [26] is explored to discover the intrinsic connections between specific visual regions and specific words. Deng et al. [42] integrated attention weights from other modalities to update current modal feature, enhancing multi-modality features via multi-round accumulated attention. Tang et al. [43] leverages learnable query tokens to represent referred objects from both textual and visual information. Tan et al. [44] decomposed text into different queries, and calculates the relevance between visual features and each query to improve the accuracy of reasoning. The parallel bidirectional Word-Pixel-Alignment (WPA) [15], [45], [46] is employed to enhance visual-text co-embedding. For a specific modality feature, they treat it as the query and use the features of other modalities as keys and values, aggregating information from other modalities through cross-modal attention. Kim et al. [47] conducts self-attention on concatenated visual and text tokens. Liu et al. [48] innovates with sequential bidirectional attention, leveraging visual context to refine text and leveraging new text to refine visual representations.

Graph model [49], [50], [51] is dominate in analyzing the relationship of entities, which are organized in the form of irregular structure. The linguistic graph structure [28], [29], [52], [53] or the multi-modal visual feature graph structure [27] have been exploited to model global context from language or visual aspect, respectively. Zeng et al. [54] built a graph structure for action units based on temporal context and semantic information, enhancing the performance of action localization. Chen and Li [55] built a multimodal graph, fusing text into proposal boxes, progressively refining them towards ground truth. Huang et al. [56] and Liu et al. [57] propose to model the relationship between pixels and words, using text as a medium to ensure favourable cross-modal interaction. However, most existing methods still suffer from the inaccuracy localization of referent, and the sufficient vision-language mutual guided learning should be further explored.

III. Proposed Method

This section elaborates our proposed Cross-Modal Interactive Reasoning Network (CMIRNet). We start by presenting the overall architecture of CMIRNet in Section III-A. Then, we delve into three key modules: the Text-Guided Multi-Modality Joint Encoder in Section III-B, the Cross-Graph Interactive Positioning module in Section III-C, and the Cross-Modal Attention Enhanced DEcoder in Section III-D, sequentially. Finally, the loss function is briefly described in Section III-E.

A. Overall Architecture

The overarching architecture is depicted in Fig.2. Given an image and natural language expression, we initially extract visual features utilizing ResNet-101 [58] and linguistic features employing BERT [59], respectively. To enhance the perception of multi-scale features, we supplement an Atrous Spatial Pyramid Pooling (ASPP) following the stage-5 encoder. Considering the computational complexity of our network and the higher consistency between deeper visual features and language features, we have opted to commence the interaction between visual and language information from stage-3 to stage-6. To achieve the initial cross-modal alignment in early encoding phase, a novel Text-Guided Multi-Modality Joint Encoder (TGMM-JE) is applied to multiple encoding stages from shallow to deep layers, expecting to generate vision-language co-embedding. To mitigate non-referent interference and enhance referent localization, a novel Cross-Graph Interactive Positioning (CGIP) module is deployed in the deepest layer, precisely pinpointing pivotal pixels/regions of the referent object. Finally, to refine the referent mask from coarse to fine progressively, a novel Cross-Modal Attention Enhanced DEcoder (CMAE-DE) is dedicated, where hybrid cross modal attentions are explored to enhance the representation of referents. We will introduce them in detail next.

B. Text-Guided Multi-Modality Joint Encoder (TGMM-JE)

To achieve cross-modal alignment during early encoding phase, we have devised an unique sequential bidirectional

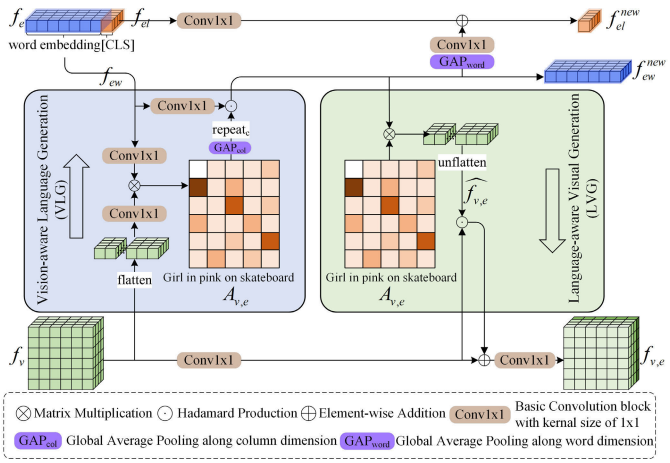


Fig. 3. Internal structure of text-guided multi-modality (TGMM) module.

vision-language joint encoder. While several existing works have introduced attention-based cross-modal alignment methods with effective outcomes, they each possess limitations. For instance, ReSTR [47] conducts self-attention on concatenated visual and text tokens, which maybe only applicable to Transformer framework. LAVT [12] is an unidirectional information fusion and primarily integrates original text information into visual information. CoupAlign [45] utilizes parallel bidirectional attention, yet it update new visual feature relies solely on original text features, neglecting the potential of new text representations in enhancing new visual representation. CARIS [48] innovates with sequential bidirectional attention, leveraging visual context to refine text representations and leveraging new text information to refine visual representations, but its multiple cross-attention and self-attention operations cause higher computational complexity. Different from these approaches, we introduce a simpler sequential bidirectional attention-like method, which only involves one attention calculation and has lower complexity. It not only leverages extensive visual context information to enrich text feature, but also utilizes these new updated text feature to refine visual features in turn. This dual-modal encoding offers a more efficient and effective means of achieving cross-modal feature alignment during the early encoding phase.

As depicted in Fig.2, the TGMM-JE comprises two submodules. The one is basic ResBlock-i ($i \in \{3, 4, 5\}$ is the index of encoder stage) or ASPP submodule in stage-6, which extracts basic visual features denoted as $f_v^i \in \mathbb{R}^{C_v^i \times H^i \times W^i}$. The other one is our new TGMM-i, which generates vision-language co-embedding denoted as $f_{v,e}^i \in \mathbb{R}^{C \times H^i \times W^i}$. The values of H^i, W^i vary with the stages due to the downsampling operations in different encoder blocks. For brevity and writing efficiency, we'll omit the index i from our subsequent symbols. Fig.3 illustrates the internal structure of TGMM. It contains two paths that the Vision-aware Language Generation (VLG) is devised to produce the new language expression incorporating the comprehension of visual information, and the Language-aware Vision Generation (LVG) is dedicated to produce the new visual feature guided by the new language expression. Specifically, we take the original visual feature $f_v \in \mathbb{R}^{C_v \times H \times W}$ and original linguistic feature $f_e \in \mathbb{R}^{C_t \times (T+1)}$ as input. Note that f_e contains word embeddings $f_{ew} \in \mathbb{R}^{C_t \times T}$ and entire sentence embedding $f_{el} \in \mathbb{R}^{C_t \times 1}$ ($[CLS]$).

1) Vision-aware Language Generation (VLG): To emphasize crucial vision-related word features, we initially devise the VLG path, enabling us to derive an enriched language representation that aligns more closely with visual semantics. Given the visual feature f_v and word vector f_{ew}, we compute the affine matrix $A_{v,e} \in \mathbb{R}^{HW \times T}$ between the pixels of feature map (each pixel actually possesses a patch-level information in the original image) and the words, using Eq.(1),

$$A_{v,e} = (Conv_{1\times1}(flatten(f_v)))^T (Conv_{1\times1}(f_{ew})) \quad (1)$$

Each element in $A_{v,e}$ represents the correlation between each pixel and word. The $flatten(\cdot)$ represents the operation that flatten a two-dimensional matrix into an one-dimensional vector. The $Conv_{1\times1}$ represents the basic convolution block for aligning channels, which consists of 1×1 convolution, activation function, and batch normalization operations. To evaluate the importance of each word, the vision-related word attention vector $Att_{word} \in \mathbb{R}^{1\times T}$ can be derived by applying global average pooling to the pixel-word correlation $A_{v,e}$ along the column orientation. Then, a newly vision-aware word embedding $f_{ew}^{new} \in \mathbb{R}^{C\times T}$ can be obtained by re-weighting the importance of each word vector. The operations can be formulated in Eq.(2)-(3),

$$Att_{word} = GAP_{col}(A_{v,e}) \quad (2)$$

$$f_{ew}^{new} = repeat_c(Att_{word}) \odot Conv_{1\times1}(f_{ew}) \quad (3)$$

where $GAP_{col}(\cdot)$ refers to the Global Average Pooling along columns, and the correlation of all pixels to the current word can be pooled into an importance weight. Each element in Att_{word} can be used to identify important vision-related word features and rescale the word features according to the importance. The $repeat_c(\cdot)$ is used to realize the dimension expansion by copying data in channel dimensions. The $Conv_{1\times1}(\cdot)$ is basic convolution block for aligning channels. Given our new vision-aware word embedding $f_{ew}^{new} \in \mathbb{R}^{C\times T}$, we propose to generate the new vision-aware sentence vector $f_{el}^{new} \in \mathbb{R}^{C\times1}$. It is generated by combing the new word embedding f_{ew}^{new} and original f_{el}, as formulated in Eq.(4),

$$f_{el}^{new} = Conv_{1\times1}(GAP_{word}(f_{ew}^{new})) + Conv_{1\times1}(f_{el}) \quad (4)$$

where $GAP_{word}(\cdot)$ refers to Global Average Pooling along the word dimension, and the semantic of all new words can be pooled into a new global sentence semantic. The f_{el}^{new} represents our enhanced vision-aware sentence embedding incorporating an understanding of visual information, which is utilized to refine the referent mask in the decoder stages.

2) Language-aware Vision Generation (LVG): To emphasize crucial text-related visual features, we in turn design the LVG path, empowering us to focus on the text-referential visual regions. Given above new word embedding f_{ew}^{new} and pixel-word correlation $A_{v,e}$, the text-related visual attention map $\widehat{f_{v,e}}$ can be generated using Eq.(5), which aggregates all the words vectors information to each pixel according to the

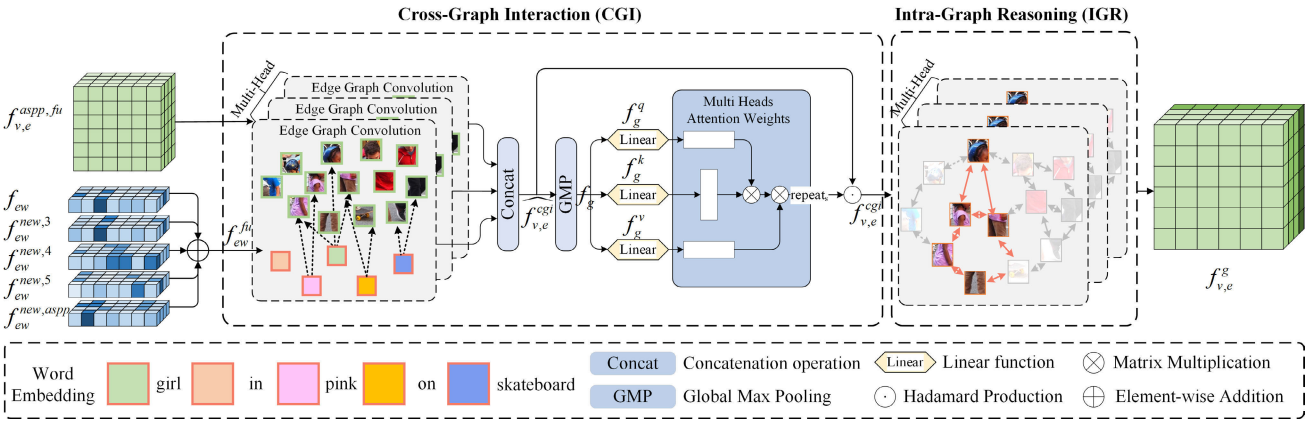


Fig. 4. Internal structure of cross-graph interactive positioning (CGIP) module.

pixel-based correlations between pixels and words.

$$\widehat{f_{v,e}} = unflatten(f_{ew}^{new} A_{v,e}^{T}) \tag{5}$$

The $unflatten(\cdot)$ is the inverse operation of $flatten(\cdot)$, used to convert a one-dimensional vector into a two-dimensional matrix. Using the original affine matrix $A_{v,e}$ serves two purposes: a good estimation for the new one and lower computation by reducing affine matrix calculations. The final new language-aware visual features $f_{v,e}$ can be encoded by incorporating original visual features, as formulated in Eq.(6),

$$f_{v,e} = Conv_{1\times1}(\widehat{f_{v,e}} \odot Conv_{1\times1}(f_v) + Conv_{1\times1}(f_v)) \tag{6}$$

where $\odot$ means the element-wise multiplication. The $Conv_{1\times1}$ is used for aligning channels. The visualization of input (f_v^i) and output ($f_{v,e}^i$) of TGMM-i shown in Fig.10 verify that TGMM can enhance the model to focus on the referent object.

C. Cross-Graph Interactive Positioning Module (CGIP)

The above vision-language joint encoding can significantly reduces interference from non-referential backgrounds, yet it still falls short of precisely locating the referent object. To overcome this limitation, we introduce an innovative Cross-Graph Interactive Positioning (CGIP) module, designed to locate the critical pixels/regions that belong to the target object at the deepest layer. Fig.4 illustrates the internal structure of our CGIP. It consists of two graph inference stages that Cross-Graph Interaction (CGI) is devised to locate the key pixels based on the pivotal linguistic information, and the Intra-Graph Reasoning (IGR) is dedicated to improve the integrity of referents globally. Specifically, we take the visual feature $f_{v,e}^{aspp,fu}$ generated from the ASPP encoder stage and the fused word embedding f_{ew}^{fu} as input. Note that the f_{ew}^{fu} is combined from the original word embedding f_{ew} and four new word embeddings in our previous TGMM modules, using Eq.(7),

$$f_{ew}^{fu} = lin(f_{ew}) + f_{ew}^{new,3} + f_{ew}^{new,4} + f_{ew}^{new,5} + f_{ew}^{new,aspp} \tag{7}$$

where $lin(\cdot)$ is a linear function for aligning channels.

1) Cross-Graph Interaction (CGI): To locate the pivotal pixels/regions of the target, we devise the CGI, expecting to select most representative text information for each pixel. A cross-modal graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ is built by treating each pixel of $f_{v,e}^{aspp,fu}$ as one kind of graph vertex $\mathcal{V}_v$ and each word of f_{ew}^{fu} as another kind of graph vertex $\mathcal{V}_{ew}$. Similar to existing semantic segmentation [60], [61] or salient object detection [62] methods, we justify pixels as graph nodes for two benefits: fewer pixels reduce complexity; region-level interactions boost performance. For each pixel vertex $\mathcal{V}_v^p$, we select and aggregate the most representative vision-text edge features $(\mathcal{V}_v^p - \mathcal{V}_{ew}^m)$ from its K nearest neighbors, expecting to reduce irrelevant information and enhance accuracy, as formulated in Eq.(8),

$$\widehat{\mathcal{V}_v^p} = MH(\max_{m\in\mathcal{N}_K^p}(conv(concat(\mathcal{V}_v^p, \mathcal{V}_v^p - \mathcal{V}_{ew}^m)))) \tag{8}$$

where $p \in \{1,2,\ldots,HW\}$ index the pth pixel in visual feature and m index current pixel's mth nearest neighboring word vector in word embedding. The $concat(\cdot)$ is the concatenation operation, and $conv(\cdot)$ refers to the basic convolution block. The max is used to select and aggregate the most representative features. $\mathcal{N}_K^p$ refers to the K nearest neighbors (K is set to 9 in default) of pth visual vertex. MH represents multi-head information aggregation, which is utilized to explore and aggregate the most representative information in different subspaces, and can enhance the feature diversity. Then, a new set of $\widehat{\mathcal{V}_v^p}$ form a new visual feature $\widehat{f_{v,e}^{cgi}} \in \mathbb{R}^{head\times C_{head}\times H\times W}$ from above graph inference ($head$ is set to 8 in default).

To better assess the importance of different heads, a novel multi-head attention weight is dedicated to further improve the distinctiveness. The multi-head global information $f_g \in \mathbb{R}^{head\times C_{head}}$ is generated by applying global max pooling on the visual feature $\widehat{f_{v,e}^{cgi}} \in \mathbb{R}^{head\times C_{head}\times H\times W}$, using Eq.(9),

$$f_g = squ(GMP(\widehat{f_{v,e}^{cgi}})) \tag{9}$$

where $GMP(\cdot)$ refers to the global max pooling in spatial dimension, and $squ(\cdot)$ is the squeeze operation in spatial dimension. To analyze head correlations and capture global

information from different heads, we compute multi-head attention weights $W_{MH} \in \mathbb{R}^{head \times C_{head}}$ using Eq.(10)-Eq.(11),

$$f_g^k, f_g^q, f_g^v = lin_k(f_g), lin_q(f_g), lin_v(f_g) \tag{10}$$

$$W_{MH} = softmax(f_g^q (f_g^k)^T) f_g^v \tag{11}$$

where $lin_k(\cdot), lin_q(\cdot), lin_v(\cdot)$ denotes the linear function, and $softmax(\cdot)$ refers to the softmax normalization. The final visual feature $f_{v,e}^{cgi}$ output from CGI can be obtained by weighting each head as Eq.(12),

$$f_{v,e}^{cgi} = repeat_s(unsqu(W_{MH})) \odot \widehat{f_{v,e}^{cgi}} \tag{12}$$

where $unsqu(\cdot)$ is the unsqueeze operation in spatial dimension. The $repeat_s(\cdot)$ is used to realize the dimension expansion by copying data in spatial dimensions. The visualization of $f_{v,e}^{cgi}$ shown in Fig.10 demonstrate that CGI successfully pinpoints pivotal regions of the target, albeit sparsely.

2) Intra-Graph Reasoning (IGR): Although the CGI module has located the key pixel areas of the object, it is sparse and incomplete. To improve the object integrity, we design the IGR, expecting to make the target more complete. Specifically, a complete graph is constructed by treating each pixel of $f_{v,e}^{cgi}$ as graph vertex, and the relationship among each pixel-pair can be generated by using Eq.(13),

$$A_{v,v} = (conv(f_{v,e}^{cgi}))^T (conv(f_{v,e}^{cgi})) \tag{13}$$

where $A_{v,v} \in \mathbb{R}^{HW \times HW}$ is the adjacency matrix of the complete graph. $conv(\cdot)$ is the basic convolution block. Then, we perform intra-graph convolution operation to aggregate important information across the complete graph using Eq.(14),

$$\begin{aligned} f_{v,e}^{g} &= reshape(GCN_{MH}(f_{v,e}^{cgi})) \\ &= reshape(flatten(f_{v,e}^{cgi}) A_{v,v} w_{igr}) \end{aligned} \tag{14}$$

where $w_{igr} \in \mathbb{R}^{HW \times HW}$ is learnable parameters, and $GCN_{MH}(\cdot)$ is the multi-head graph (*head* is set to 8 in default) convolution operation. The final visual feature $f_{v,e}^{g} \in \mathbb{R}^{C \times H \times W}$ can be obtained from our IGR module. The visualization of $f_{v,e}^{g}$ shown in Fig.10 demonstrate that IGR can effectively improve the object's completeness.

D. Cross-Modal Attention Enhanced DEcoder (CMAE-DE)

The above positioning module is capable of roughly pinpointing the referent object at the deepest layer, yet the localization appears coarse. In order to further enhance the discrimination and optimize the referent mask from coarse to fine gradually, we propose a novel CMAE-DE, which is applied to multi-level features from deep layers to shallow layers. Fig.5 depicts the internal structure of our CMAE-DE, which contains Cross-Modal Spatial Attention (CM-SA), and Cross-Modal Channel Attention (CM-CA). The CM-SA is devised to produce a spatial attention map, highlighting pixels strongly correlated with the language expression. The CM-CA analyzes the relationship between spatial attention map and visual feature maps, assigning channel-wise weights to aggregate crucial visual features. Hence, CMAE-DE refines

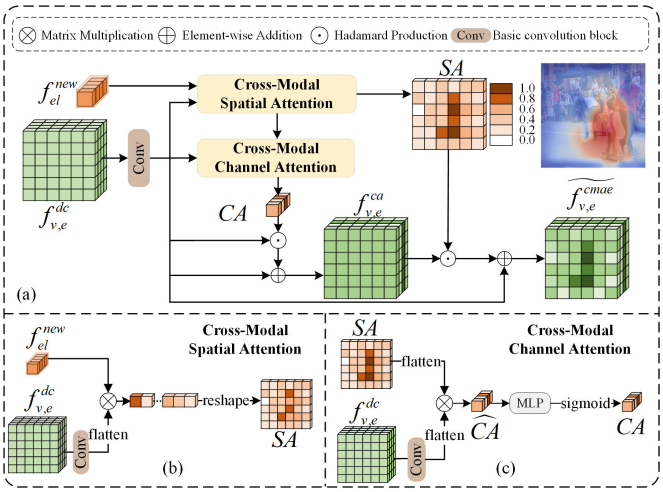


Fig. 5. Internal structure of cross-modal attention enhanced (CMAE) module.

referent features in both spatial and channel dimensions, guided by our new language expression.

Specifically, as shown in Fig.2, for the current CMAE-DE-i ($i \in \{3, 4, 5, 6\}$), we take the multi-modality visual feature $f_{v,e}^{dc,i}$ and the newly obtained sentence vector $f_{el}^{new,i}$ as input. The multi-modality visual feature $f_{v,e}^{dc,i}$ is obtained by combing the output of previous decoder part ($f_{v,e}^{cmae,i+1}$) and current encoder stage ($f_{v,e}^{i}$). The current multi-modality visual features $f_{v,e}^{dc,i}$ can be obtained using the following formula,

$$f_{v,e}^{dc,i} = \begin{cases} conv(Concat(f_{v,e}^{aspp}, f_{v,e}^{g})), & i = 6 \\ conv(Concat(Up_2(f_{v,e}^{cmae,i+1}), f_{v,e}^{i})), & i \in \{3, 4, 5\} \\ conv(Concat(Up_2(f_{v,e}^{cmae,i+1}), f_{v}^{i})), & i \in \{2\} \\ conv(Concat(Up_2(f_{v,e}^{dc,i+1}), f_{v}^{i})), & i \in \{1\} \end{cases} \tag{15}$$

where the $Concat(\cdot)$ denotes the concatenation operation, and $f_{v,e}^{aspp}$ represents the visual features generated from the ASPP submodule, $Up_2(\cdot)$ refers to 2 times up-sampling operation. Additionally, two basic convolution blocks are adopted to adjust the channel dimensions. For the CMAE, as shown in Fig.5(a), it consists of CM-SA and CM-CA, which are devised to progressively refine multi-modality features in both spatial and channel dimensions. For brevity and writing efficiency, in the subsequent discussions, we will omit the index 'i' for the current level of CMAE-DE-i.

1) Cross-Modal Spatial Attention (CM-SA): As depicted in Fig.5(b), to emphasize the crucial spatial locations mentioned in the language expression, we introduce a CM-SA submodule. This submodule generates a spatial attention map $SA \in \mathbb{R}^{1 \times H \times W}$, highlighting the corresponding spatial positions in the visual feature $f_{v,e}^{dc}$ that are strongly related to the updated language feature f_{el}^{new}. The SA signify the likelihood of a pixel belonging to the object, facilitating subsequent target enhancement operations. It is formulated in Eq. (16),

$$SA = reshape((f_{el}^{new})^T flatten(conv(f_{v,e}^{dc}))) \tag{16}$$

where the $conv(\cdot)$ is adopted to align channels. The $flatten(\cdot)$ operation flattens a two-dimensional matrix into a one-dimensional vector. The $reshape(\cdot)$ operation reshape the

Fig. 6. The visual examples of the cross model spatial attention maps. (a) Language: 'first case on left.' (b) Language: 'white car on the left.' (c) Language: 'man in middle.'

one-dimensional vector to two-dimensional matrix. Visual examples in Figs.5 and Figs.6 demonstrate the spatial attention map's effectiveness in highlighting crucial locations.

2) Cross-Modal Channel Attention (CM-CA): Furthermore, as depicted in Fig.5(c), to enhance referent features, we devise CM-CA that weighs their channel importance. The channel-wise vector $\widehat{CA} \in \mathbb{R}^{1\times C}$ can be first generated by analyzing the channel-wise relationship between the spatial attention map SA and the visual feature $f_{v,e}^{dc}$. It is formulated in Eq.(17),

$$\widehat{CA} = (flatten(SA))(flatten(conv(f_{v,e}^{dc})^T)) \tag{17}$$

Then, a set of weights $CA \in \mathbb{R}^{1\times C}$ is learned by applying Multi-Layer Perception (MLP), as formulated in Eq. (18),

$$CA = sig(MLP(\widehat{CA})) \tag{18}$$

where $MLP(\cdot)$ consists of two linearity operations, and $sig(\cdot)$ represents the sigmoid function.

So far, the channel attention CA and the spatial attention SA have been generated. These weights are then used to continuously filter visual feature $f_{v,e}^{dc}$ in both channel and spatial dimensions, removing irrelevant information and accentuating the referent effectively. As formulated in Eq.(19-20),

$$f_{v,e}^{ca} = CA \odot conv(f_{v,e}^{dc}) + conv(f_{v,e}^{dc}) \tag{19}$$

$$\widetilde{f_{v,e}^{cmae}} = SA \odot f_{v,e}^{ca} + conv(f_{v,e}^{dc}) \tag{20}$$

where $\widetilde{f_{v,e}^{cmae}}$ is a cross-modal attention enhanced visual feature. Finally, we further decode the refined features as follows,

$$f_{v,e}^{cmae} = conv(concat(\widetilde{f_{v,e}^{cmae}}, f_{v,e}^{dc})) \tag{21}$$

and it also will be used to guide the next decoder layer's output $f_{v,e}^{cmae,i-1}$ for further refinement and enhancement. For the outputs $f_{v,e}^{cmae,i}, i \in \{3,4,5,6\}$, and $f_{v,e}^{dc,i}, i \in \{1,2\}$, we use a 1×1 convolution layer to project them into two class score maps $P_i \in \mathbb{R}^{2\times H\times W}, i \in \{1,2,3,4,5,6\}$. Among them, we choose P_1 as the final segmentation result. The visualization of $f_{v,e}^{cmae,i}$ and $f_{v,e}^{dc,i}$ shown in Fig.10 demonstrate that CMAE can progressively refine the referent mask.

E. Loss Function

We exercise supervision over all six-layer outputs of the decoder through the employment of a straightforward binary cross-entropy (BCE) loss, which is formulated as follows,

$$\mathcal{L}_{mask} = \frac{1}{6}\sum_{i=1}^{6}\mathcal{L}_{bce}(P_i, G) \tag{22}$$

where P_i is the predicted mask of the i th decoder layer, and G stands for ground truth. $\mathcal{L}_{bce}(\cdot)$ is binary cross-entropy loss.

IV. Experiments

This part introduces the experimental setup with challenging datasets, implementation details, and evaluation metrics in section IV-A. Then, in section IV-B, a comprehensive comparison between our method and twenty-two advanced models is discussed to confirm the superiority of our CMIRNet. The sufficient ablation studies have been performed in section IV-C to validate the efficacy of our proposed pivotal modules.

A. Experimental Settings

1) Datasets: The experiments are conducted on four widely used benchmark datasets, including RefCOCO [1], RefCOCO+ [1], RefCOCOg [63], and ReferIt [64]. All the images in these datasets have been gathered from MS COCO [65] or IAPR TC-12 [66] and are annotated with corresponding language expressions. Each image is accompanied by a diverse set of unrestricted natural language expressions, posing significant challenges and encompassing a wide range of scenarios.

RefCOCO and RefCOCO+ are benchmark datasets widely used for RIS task. The RefCOCO dataset comprises 142,209 referring expressions, referring to 50,000 objects across 19,994 images. RefCOCO+ serves as an enhanced version of RefCOCO, encompassing 141,564 referring expressions for 49,856 objects in 19,992 images. Notably, RefCOCO+ excludes certain attribute words pertaining to the absolute localization of objects, rendering it a more demanding dataset.

RefCOCOg is another widely adopted benchmark dataset, encompassing 104,560 language expressions referring to 54,822 objects in 26,711 images. In comparison to RefCOCO and RefCOCO+. G-Ref stands out as the most challenging due to its average 8.4-words per expression, surpassing RefCOCO and RefCOCO+'s average 3.5-words. The UMD [63] divides the entire dataset into training, validation, and testing sets, and our reported results adhere to this partitioning.

ReferIt comprises an extensive dataset of 130,525 referring expressions for 96,654 objects within 19,894 natural images. It contains more abstract concepts expression, such as 'bottom right', 'Top', which refers to special background regions.

2) Implementation Details: Our approach utilizes PyTorch and runs on an NVIDIA Tesla A40 GPU. We've crafted four versions of CMIRNet, each equipped with unique visual encoders: Res50 and Res101 [58], pre-trained on ImageNet-1K, and SwinB and SwinL [75], pre-trained on ImageNet-22K. The text encoder, a 12-layer BERT [59] with 768 hidden channels, is initialized with official pre-trained weights. The channel C in section III-B and section III-D is set to 512 in default. The input images are resized to 480×480. The maximum sentence length N_l for BERT is set to 20 for all datasets. Both the visual encoder and BERT encoder undergo fine-tuning during training. In the experiments, we adopt two training strategies. One is the standard strategy: we train the model separately using the training set of each dataset and evaluate it on each dataset. The model is trained for 30 epochs with a batch size of 32, utilizing the AdamW optimizer. The initial learning rate is 5e-5 and subsequently declines with

polynomial decay. The other one is Pretrain+Finetune strategy: we first pretrain the model using the combined Referring Expression Comprehension (REC) datasets (Visual Genome [76], RefCOCO, RefCOCO+, and RefCOCOg), leveraging large-scale image-text bounding box annotations. Then the model is finetuned on RIS datasets. Notebly, in Pretrain stage, we enforce the decoder output to regress bounding boxes, and then generate pixel-level segmentation masks during the finetuning phase. The pretraining is excuted for 10 epochs and finetuning is 30 epoches, with a batch size of 32.

3) Evaluation Metrics: Like the existing works [8], [47], [12], we adopt two evaluation metrics, including mask IoU and precision@X. Mask IoU contains overall IoU (oIoU) and mean IoU (mIoU), where oIoU measures total intersection regions over total union regions, and mIoU quantifies the average of intersection regions over union regions. The precision@X ($\mathbf{X} \in \{0.5, 0.7, 0.9\}$) assesses the percentage of IoU scores exceed a specified threshold value of X. This metric primarily assesses the targeting capability of the method, emphasizing its accuracy in locating and segmenting the desired objects.

B. Comparison With State-of-the-Arts Models

In this section, we undertake a comparative analysis of our model against twenty-two SOTA RIS models on four benchmark datasets using both the overall IoU (oIoU) and mean IoU (mIoU). These SOTA models includes fifteen CNN-based models that adopt CNN backbone as visual encoder, and seven transformer-based models that adopt the transformer backbone as visual encoder. As presented in Table I, the quantitative comparison reveals that our proposed method outperforms these advanced models across four benchmark datasets, demonstrating its superior performance.

1) Quantitative Comparison: As evident from Table I, our method (Res101) demonstrates superior performance when compared to the model utilizing CNN as the visual encoder. Specifically, we achieved the top-1 performance in seven out of nine indicators and two second-best ranking, evaluated based on oIoU. While our method obtains a slightly inferior to GLIPN model on testB set of RefCOCO+, our method surpass GLIPN on the other val and testA sets of RefCOCO+. Similarly, although our method trailed behind DMMI on the val subset of RefCOCO-g, it surpasses DMMI on the test subset of RefCOCO-g. Our method (Res50) has also achieved a higher performance when compared with those models utilizing same backbone (e.g. FSFI (Res50)). We also conducted a comparative analysis between our method and six CNN-based models using mIoU metric. As shown in Table I, our approach (Res101, Res50) generally outperforms other advanced models across three datasets, underscoring its superiority.

Moreover, our model (SwinB) outperforms some advanced transformer-based models in both oIoU and mIoU, as shown in Table.I. Even when compared to recent SwinB-based methods like CrossVLT [16] and FSFI [72], our model retains its advantages, confirming the effectiveness of our approach. Compared to the advancd ReLA [73], our method achieves the higher performance on RefCOCO dataset, and falls behind ReLA on RefCOCO+ and RefCOCOg. It is attributed that ReLA may benefit from its advanced explicit relationship modeling. Despite this, we also achieved the second best performance on RefCOCO+ and RefCOCOg, highlighting its competitiveness. PolyFormer [74], the recent advanced model achieving most SOTA performance, was pre-trained on REC datasets, and fine-tuned using RIS datasets. To keep consistent with PolyFormer, we adopted the same training strategy. Similar with PolyFormer, we trained two versions: swinB and swinL. As shown in the Table.I(b), it can be seen that under the same training strategy, our model's oIOU and mIOU performance are both higher than those of PolyFormer's swinB and swinL versions, which also verifies the effectiveness and advancement of our method.

2) Qualitative Comparison: We also conducted a comparative analysis of visual examples against several cutting-edge approaches, including PolyFormer [74], CrossVLT [16], ETRIS [15], and LAVT [12]. Notably, all these methods are open-source with readily available training parameters. As shown in Fig.7, when comparing Res101-based methods, Ours(Res101) exhibits superior performance in precisely locating referent objects and predicting segmentation masks closer to the ground truth than ETRIS(Res101). Specifically, for the case of an elephant in the third row, ETRIS(Res101) fails to detect the actual target, mistakenly highlighting non-referent regions, whereas Ours(Res101) method successfully identifies and segments the referent target. Overall, the models leveraging the swinB version display superior performance. Upon close examination of the visual outputs from LAVT, ETRIS, CrossVLT, and PolyFormer, it became apparent that they all more or less erroneously highlight some non-referent targets. In contrast, our method consistently and accurately locate the referent target, segmenting it with higher target integrity and more complete target masks. These quantitative and qualitative comparisons underscore the effectiveness and superiority of our method.

3) Performance Evaluation of Concurrent Mask and Bounding Box Generation: Furthermore, we employed different supervision labels (mask or/and bounding box) to finetune our foundation model with RIS datasets. For simplicity, we chose the foundation model with swinB as the visual backbone and pre-trained on REC datasets as our starting point. As shown in Fig.8, the decoder outputs the segmentation mask directly, and subsequently, the bounding box is extracted from the contour of the mask by analyzing the coordinates of the top, bottom, left, and right boundaries of the connected regions. To evaluate the impact of these labels on final segmentation performance, we trained three network variants: one is finetuned only with mask labels, one is finetuned with both mask and bounding box labels, and one is finetuned only with bounding box labels. Notably, we utilized $\mathcal{L}_{mask}$ from Eq.(22) for mask supervision and classic IOU loss $\mathcal{L}_{bbox}$ for bounding box supervision. As reported in Table.II, the model (row No.3 and No.6) finetuned only with bounding box labels has the worst segmentation performance, as the mask tends to overfit the bounding box without mask supervision. The model (row No.1 and No.4) finetuned only with mask owns the second-best performance. The model (row No.2 and No.5) finetuned with both mask and bounding box achieves the best performance,

TABLE I

THE PERFORMANCE OF OUR PROPOSED METHOD IS ASSESSED BY COMPARING ITS OVERALL IOU (OIOU) AND MEAN IOU (MIOU) SCORES AGAINST THOSE OF STATE-OF-THE-ART APPROACHES ON FOUR EXTENSIVELY UTILIZED BENCHMARK DATASETS. THE OPTIMAL RESULTS ARE HIGHLIGHTED IN BOLD FACE, AND THE SECOND-BEST SCORES ARE UNDERLINED. THE SYMBOL "-" DENOTES NO DATA AVAILABLE

	Methods	Year	Visual Encoder	Text Encoder	RefCOCO			RefCOCO+			RefCOCOg		ReferIt
					val	testA	testB	val	testA	testB	val	test	test
(a) Standard: Training on the training set of three datasets, respectively.													
mIOU	MCN [25]	2020CVPR	DN53	GloVe	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40	-
	CGAN [11]	2020MM	DN53	GRU	64.86	68.04	62.07	51.03	55.51	44.06	51.01	51.69	-
	LTS [8]	2021CVPR	DN53	GRU	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25	-
	VLT [13]	2021ICCV	DN56	GRU	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65	-
	CRIS [67]	2022CVPR	CLIP	CLIP	70.47	73.18	66.10	62.27	**68.08**	53.68	59.87	60.36	-
	PKS [68]	2023TCSVT	Res101	Bi-GRU	70.87	74.23	65.07	60.09	66.21	51.35	60.38	60.98	-
	CMIRNet (Ours)		Res50	Bert	71.26	74.51	68.15	62.14	67.68	54.23	60.19	61.33	63.10
	CMIRNet (Ours)		Res101	Bert	**72.28**	**74.73**	**69.23**	**62.36**	67.74	**54.69**	**61.41**	**62.46**	**64.00**
	LAVT [12]	2022CVPR	SwinB	Bert	74.46	76.89	70.94	65.81	70.97	59.23	63.34	63.62	-
	CMIRNet (Ours)		SwinB	Bert	**76.03**	**78.14**	**73.14**	**68.20**	**72.98**	**61.28**	**66.69**	**66.49**	**68.12**
oIOU	CMPC [56]	2020CVPR	DRes101	LSTM	61.36	64.53	59.64	49.56	53.44	43.23	-	-	65.53
	BusNet [10]	2021CVPR	DRes101	GloVe	62.56	65.61	60.38	50.98	56.14	43.51	-	-	-
	EFN [26]	2021CVPR	Res101	GRU	62.76	65.69	59.67	51.50	55.24	43.01	-	-	66.70
	SANet [28]	2022TMM	DRes101	GloVe	61.84	64.95	57.43	50.38	55.36	42.74	-	-	-
	ISF [27]	2022TMM	DN53	GloVe	65.19	68.45	62.73	52.70	56.77	46.39	52.67	53.00	-
	MMMAI [69]	2023TIP	DN53	GRU	67.88	70.82	65.02	56.98	61.26	50.11	54.79	58.21	-
	GLIPN [70]	2023TNNLS	DN53	GRU	68.51	71.06	65.15	58.43	63.15	**51.67**	57.87	58.13	-
	DMMI [71]	2023ICCV	Res101	Bert	68.56	71.25	63.16	57.90	62.31	50.27	**59.02**	59.24	-
	FSFI [72]	2023TNNLS	Res50	Bert	68.02	71.03	64.38	57.47	61.34	48.31	55.16	55.71	-
	FSFI [72]	2023TNNLS	Res101	Bert	69.83	72.88	64.73	58.51	63.67	49.35	56.10	56.64	-
	CMIRNet (Ours)		Res50	Bert	69.15	72.79	65.21	58.69	64.34	49.72	56.83	58.95	70.44
	CMIRNet (Ours)		Res101	Bert	**70.12**	**73.19**	**66.72**	**59.05**	**64.42**	50.40	58.07	**59.71**	**71.00**
	ReSTR [47]	2022CVPR	ViT-B-16	GloVe	67.22	69.30	64.45	55.78	60.44	48.27	54.48	-	70.15
	LAVT [12]	2022CVPR	SwinB	Bert	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09	-
	ETRIS [15]	2023ICCV	ViT-B-16	CLIP	70.51	73.51	66.63	60.10	66.89	50.17	59.82	59.91	-
	FSFI [72]	2023TNNLS	SwinB	Bert	71.23	74.34	68.31	60.84	66.49	53.24	61.51	61.78	-
	CrossVLT [16]	2023TMM	SwinB	Bert	73.44	76.16	70.15	63.60	69.10	55.23	62.68	63.75	-
	ReLA [73]	2023CVPR	SwinB	Bert	73.82	76.48	70.18	**66.04**	**71.02**	**57.65**	**65.00**	**65.97**	-
	CMIRNet (Ours)		SwinB	Bert	**73.98**	**77.10**	**71.28**	64.15	70.42	56.41	63.52	63.91	**74.16**
(b) Pretrain+Finetune: Pretraining on the combined REC datasets and finetuning on the RIS datasets.													
mIOU	PolyFormer [74]	2023CVPR	SwinB	Bert	75.96	77.09	73.22	70.65	74.51	64.64	69.36	69.88	-
	CMIRNet (Ours)		SwinB	Bert	78.51	80.21	76.04	71.18	75.26	65.03	71.48	72.81	70.15
	PolyFormer [74]	2023CVPR	SwinL	Bert	76.94	78.49	74.83	72.15	75.71	66.73	71.15	71.17	-
	CMIRNet (Ours)		SwinL	Bert	**78.98**	**80.73**	**76.56**	**72.50**	**77.00**	**66.86**	**72.60**	**73.18**	**70.85**
oIOU	PolyFormer [74]	2023CVPR	SwinB	Bert	74.82	76.64	71.06	67.64	72.89	59.33	67.76	69.05	-
	CMIRNet (Ours)		SwinB	Bert	77.72	79.82	74.42	68.26	74.06	61.14	69.17	71.89	75.25
	PolyFormer [74]	2023CVPR	SwinL	Bert	75.96	78.29	73.25	69.33	74.56	61.87	69.20	70.19	-
	CMIRNet (Ours)		SwinL	Bert	**78.24**	**80.44**	**75.22**	**69.82**	**75.33**	**62.07**	**70.36**	**72.09**	**75.66**

Fig. 7. Visual examples comparison between our proposed method and several advanced models. The notation enclosed in parentheses designates the backbone employed within the model, while the symbol ‡ signifies that the model was pre-trained utilizing REC datasets, and finetuned on RIS datasets.

as bounding box supervision aids in more precise object localization, while mask supervision refines the segmentation mask, ultimately contributing to better segmentation performance together.

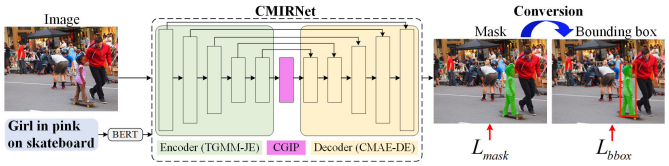


Fig. 8. Illustration of converting the generated mask into a bounding box, and jointly supervising both the mask and bounding box outputs.

TABLE II

DIFFERENT RESULTS ARE OBTAINED UTILIZING DIFFERENT SUPERVISION LOSSES. THE OIOU AND MIOU RESULTS ARE REPORTED ACROSS ALL FOUR DATASETS, WITH THE BEST RESULTS HIGHLIGHTED IN BOLD

	No.	$\mathcal{L}_{mask}$	$\mathcal{L}_{bbox}$	RefCOCO			RefCOCO+			RefCOCOg		ReferIt
				val	testA	testB	val	testA	testB	val	test	test
mIOU	1	✓		78.51	80.21	76.04	71.18	75.26	65.03	71.48	72.81	70.15
mIOU	2	✓	✓	**78.72**	**80.68**	**76.10**	**71.99**	**76.48**	**65.92**	**71.89**	**72.92**	**70.36**
mIOU	3		✓	52.48	51.04	55.07	46.44	46.48	46.43	46.34	46.87	49.54
oIOU	4	✓		77.72	79.82	74.42	68.26	74.06	61.14	69.17	71.89	75.25
oIOU	5	✓	✓	**77.84**	**80.16**	**74.59**	**69.01**	**74.99**	**61.36**	**69.25**	**71.96**	**75.65**
oIOU	6		✓	50.48	49.74	52.37	44.49	45.39	43.66	44.61	45.87	53.75

C. Ablation Study

In this part, we perform ablation studies on RefCOCO to verify key components' efficacy in improving segmentation. We train various network variants with different configurations on RefCOCO training set, and report the results on RefCOCO-val, testA, and testB, respectively. For brevity and convenience, we mainly employ Res101 as the primary visual backbone. Sequential ablation experiments examine the interactions between different modules and within each module.

1) Ablation Experiments Between Different Modules: To validate the efficacy of our three key modules, we trained eight network variants, each equipped with different key modules.

a) On the Effectiveness of Our TGMM Module: As illustrated in the row No.1 and row No.2 of Table III, for the RefCOCO validation set, compared with the 'baseline' network (row No.1), the 'base+TGMM' network (row No.2) acquires 2.41% and 3.45% performance improvement in terms of oIoU and mIoU, respectively, and acquire approximate 4.17%, 5.11%, 3.01% absolute point improvement in terms of P@0.5, P@0.7, and P@0.9. The performance improvement can also be observed on the testA and testB set. Additionally, when comparing 'base+CGIP' (row No.3) with 'base+TGMM+CGIP' (row No.6), we observe a performance enhancement achieved by 'base+TGMM+CGIP' over 'base+CGIP' on three data subsets. The consistent performance improvements are also evident when comparing 'base+TGMM+CMAE' (row No.7) with 'base+CMAE' (row No.4). All these results can demonstrate the efficacy of our TGMM, which helps to encode multi-modal features from shallow to deep layers, by understanding the visual and linguistic information comprehensively.

As seen in Fig.9 (columns 5-10), the 'baseline' yields the poorest segmentation, while 'base+TGMM' improves but still misidentifies non-referent areas. Incorporating TGMM into 'base+CGIP' and 'base+CMAE' as 'base+TGMM+CGIP' and 'base+TGMM+CMAE' respectively, leads to more precise referent masks, demonstrating TGMM's effectiveness in enhancing multi-modality co-embedding.

b) On the effectiveness of our CGIP module: As shown in the row No.1 and No.3 of Table III, for the RefCOCO-val set, compared with the 'baseline' (row No.1), the 'base+CGIP' (row No.3) acquires 0.38% and 0.56% performance improvement in terms of oIoU and mIoU, and acquires approximate 0.76%, 1.51%, 0.15% absolute point improvement in terms of P@0.5,0.7,0.9. The performance improvement can also be observed on the testA and testB set. Additionally, when comparing 'base+CMAE' (row No.4) with 'base+CGIP+CMAE' (row No.5), we observe a general performance enhancement achieved by 'base+CGIP+CMAE' over 'base+CMAE' on three data subsets. The general consistent performance improvements are also evident when comparing 'base+TGMM+CGIP' (row No.6) with 'base+TGMM' (row No.2). When comparing 'base+TGMM+CMAE' (row No.7) with 'base+TGMM+CGIP+CMAE' (the 'Full model', row No. 8), we observe that the 'Full model' has achieved a consistent performance improvement of 0.08-0.9 across all 15 metrics. All these experiment results showcase the effectiveness of our CGIP module, which can help to locate the key pixels of the referent, and improve the segmentation accuracy.

As evident in Fig.9 (columns 3-10), 'base+CGIP' precisely locates the referent object, surpassing 'baseline'. Adding CGIP to 'base+TGMM' and 'base+CMAE' forms 'base+TGMM+CGIP' and 'base+CGIP+CMAE' both improving accuracy. Comparing 'base+TGMM+CMAE' (column 5) with 'Full model' (the 'base+TGMM+CGIP+CMAE', column 3), the latter eliminates interference and accurately highlights the referent object. All these visual examples validate CGIP's positive impact on localization.

c) On the effectiveness of our CMAE module: As shown in the row No.1 and row No.4 of Table III, for the RefCOCO validation set, compared with the 'baseline' network (row No.1), the 'base+CMAE' network (row No.4) acquires 1.53% and 1.2% performance improvement in terms of oIoU and mIoU, and acquires approximate 1.37%, 1.52%, and 0.56% absolute point improvement in terms of P@0.5, P@0.7, and P@0.9. The performance improvement can also be observed on the testA and testB set. Additionally, when comparing 'base+CGIP' (row No.3) with 'base+CGIP+CMAE' (row No.5), we observe a significant performance enhancement achieved by 'base+CGIP+CMAE' over 'base+CGIP' on three data subsets. The consistent performance improvements are also evident when comparing 'base+TGMM+CMAE' (row No.7) with 'base+TGMM' (row No.2). All these experimental results indicate that our proposed CMAE module is effective in enhancing the representation of referent features from spatial and channel dimensions.

As shown in Fig.9 (columns 4,5,7-10), 'base+CMAE' yields a more comprehensive mask, albeit with some non-referent refinement errors. Integrating CMAE into 'base+TGMM' and 'base+CGIP' results in 'base+TGMM+CMAE' and 'base+CGIP+CMAE', producing more precise object masks. All these highlight CMAE's positive impact on refining the referent mask based on our new language expression.

d) The visualization of feature maps from different modules: As depicted in Fig.10, we initially visualize the feature

TABLE III

ABLATION RESULTS BETWEEN DIFFERENT MODULES ON THE REFCOCO VALIDATION, TESTA, AND TESTB. THE BEST RESULTS ARE IN BOLD

No.	Model Variants	TGMM	CGIP	CMAE	RefCOCO-val					RefCOCO-testA					RefCOCO-testB				
					P@0.5	P@0.7	P@0.9	oIoU	mIoU	P@0.5	P@0.7	P@0.9	oIoU	mIoU	P@0.5	P@0.7	P@0.9	oIoU	mIoU
1	baseline				77.39	66.77	27.39	66.88	67.99	81.53	71.82	27.49	70.12	70.97	71.74	60.08	27.48	62.98	64.21
2	base+TGMM	✓			81.56	71.88	30.40	69.29	71.44	84.90	75.29	29.98	71.92	73.80	76.27	65.69	31.99	65.57	68.33
3	base+CGIP		✓		78.15	68.28	27.54	67.26	68.55	81.81	72.37	28.18	70.48	71.51	71.97	61.35	27.79	63.07	64.53
4	base+CMAE			✓	78.76	68.29	27.95	68.41	69.19	84.09	74.65	30.02	71.20	73.08	75.25	64.59	30.68	64.59	67.25
5	base+CGIP+CMAE		✓	✓	81.04	71.34	30.78	68.77	71.02	84.23	74.54	29.95	71.48	73.33	75.74	65.16	31.13	65.01	67.63
6	base+TGMM+CGIP	✓	✓		81.83	71.84	30.79	69.63	71.68	85.33	75.78	30.28	72.22	73.99	76.88	66.18	31.23	65.88	68.48
7	base+TGMM+CMAE	✓		✓	82.25	72.60	30.85	69.94	72.11	85.63	77.04	30.69	72.66	74.43	77.92	67.16	32.37	66.48	68.91
8	Full model	✓	✓	✓	**82.41**	**72.96**	**31.44**	**70.12**	**72.28**	**85.93**	**77.41**	**31.59**	**73.19**	**74.73**	**78.00**	**67.40**	**32.50**	**66.72**	**69.23**

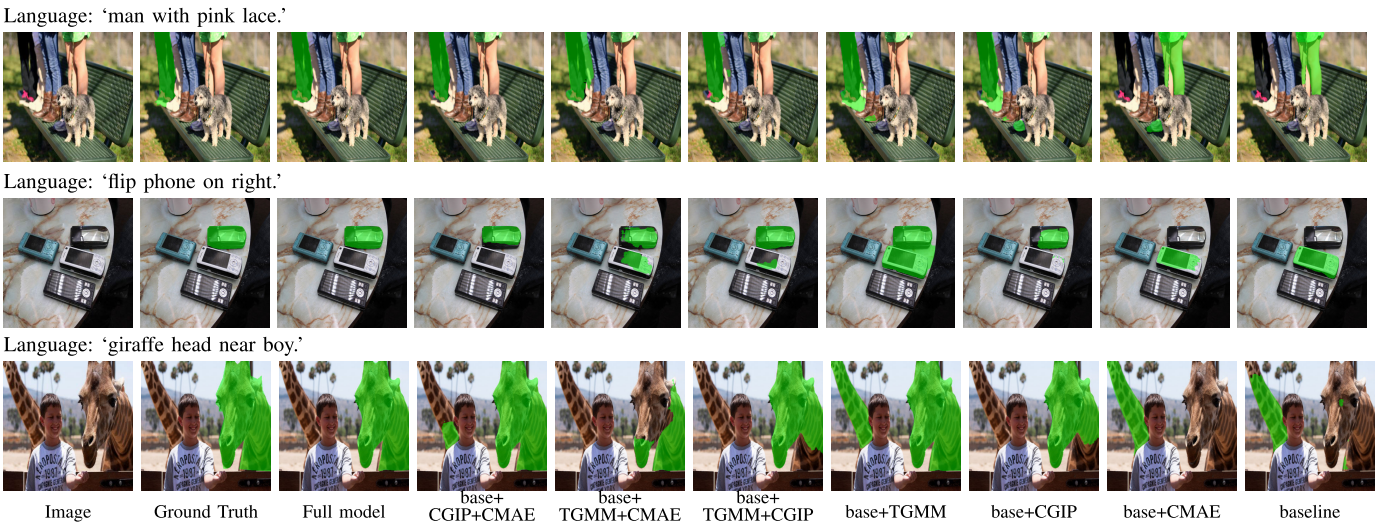


Fig. 9. The visual examples of the segmentation mask generated by models equipped with different modules on RefCOCO validation set.

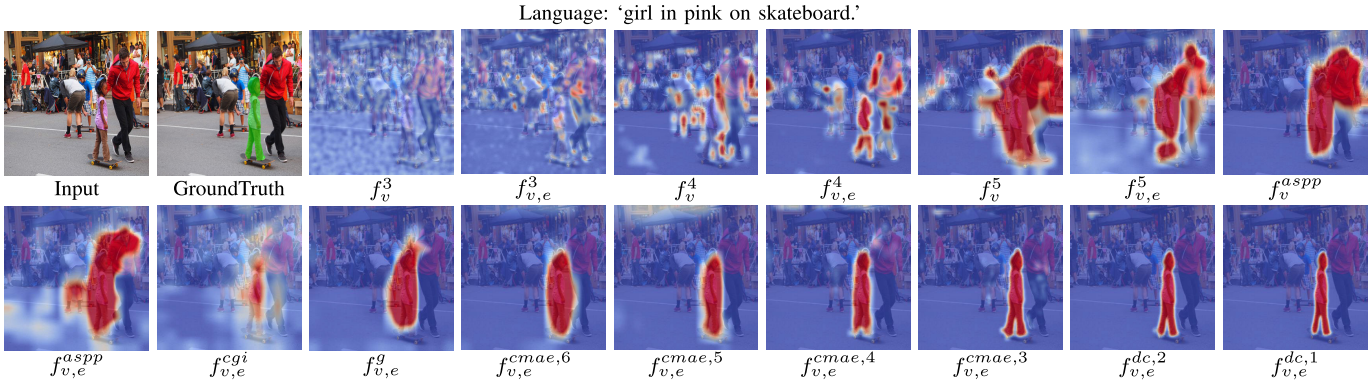


Fig. 10. Visual comparison of feature maps from different modules. The f_v^i and $f_{v,e}^i$ ($i \in \{3, 4, 5, aspp\}$, and $i = aspp$ is equal to $i = 6$) represent the input and output of the ith TGMM in encoder stage, respectively. The $f_{v,e}^{cgi}$ and $f_{v,e}^g$ separately represent the outputs of CGI and IGR within CGIP. The $f_{v,e}^{cmae,i}$ ($i \in \{3, 4, 5, 6\}$) and $f_{v,e}^{dc,i}$ ($i \in \{1, 2\}$) represent the outputs of the ith decoder stage.

maps for the TGMM module. Comparing the input (f_v^i) and output ($f_{v,e}^i$) features of each TGMM-i reveals an enhanced focus on the referent object, attesting to its effectiveness. Notably, the feature maps evolve from dispersed to increasingly target-centric as layers deepen, from stages 3 to 6. Furthermore, the feature maps of the CGIP module are also inspected. Here, $f_{v,e}^{cgi}$ and $f_{v,e}^g$ denote the outputs of CGI and IGR, respectively. As seen in $f_{v,e}^{cgi}$, CGI successfully pinpoints pivotal regions of the target, albeit sparsely. The subsequent IGR improves the object's completeness, as seen in $f_{v,e}^g$, confirming CGIP's efficacy in both localization and integrity enhancement. Lastly, the CMAE module's feature maps, denoted by $f_{v,e}^{cmae,i}$ and $f_{v,e}^{dc,i}$ from stages 6 to 1, demonstrate its ability to refine the target object progressively and eliminate non-target interferences during decoding. This gradual refinement process ensures the achievement of more precise and robust image segmentation outcomes, prominently showcasing CMAE's capability to progressively enhance discrimination and refine the referent mask from coarse to fine. All these visual examples validate the effectiveness of three proposed modules.

e) The statistical assessment of performance gains: To ascertain the credibility of performance increment from individual modules, we conducted Wilcoxon signed rank tests on all data pairs in Table III. For TGMM's performance boost, we first paired 'baseline' (row No.1) with 'base+TGMM'

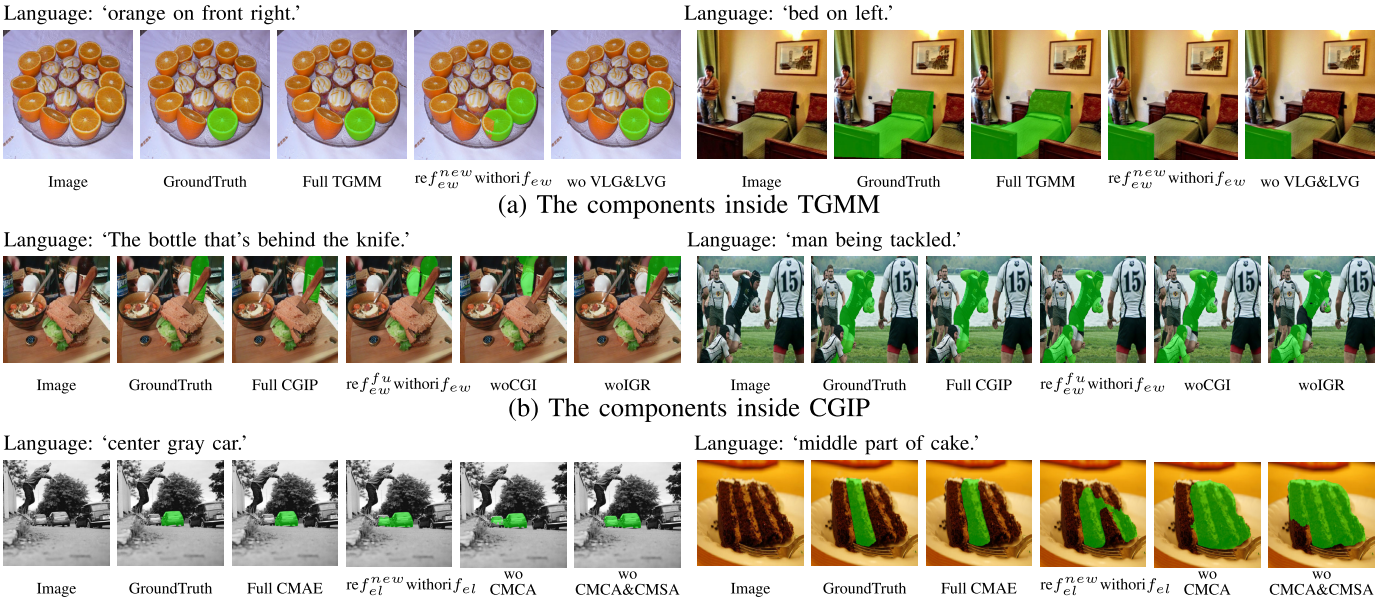


Fig. 11. The visual examples of the segmentation mask generated by models equipped with different components in each module on RefCOCO validation set, where (a) is inside TGMM, (b) is inside CGIP, and (c) is inside CMAE.

TABLE IV
THE STATISTICAL ASSESSMENT OF PERFORMANCE GAINS IN ABLATION STUDIES ACROSS MODULES USING WILCOXON SIGNED-RANK TESTS

Module	Data-pairs	Model variant pairs	P-value
TGMM	No.(1,2)	(baseline, base+TGMM)	0.000061
	No.(3,6)	(base+CGIP, base+TGMM+CGIP)	0.000061
	No.(4,7)	(base+CMAE, base+TGMM+CMAE)	0.000061
	No.(5,8)	(base+CGIP+CMAE, Full model)	0.000061
CGIP	No.(1,3)	(baseline, base+CGIP)	0.000061
	No.(4,5)	(base+CMAE, base+CGIP+CMAE)	0.000305
	No.(2,6)	(base+TGMM, base+TGMM+CGIP)	0.010254
	No.(7,8)	(base+TGMM+CMAE, Full model)	0.000061
CMAE	No.(1,4)	(baseline, base+CMAE)	0.000061
	No.(3,5)	(base+CGIP, base+CGIP+CMAE)	0.000061
	No.(2,7)	(base+TGMM, base+TGMM+CMAE)	0.000061
	No.(6,8)	(base+TGMM+CGIP, Full model)	0.000061

(row No.2) as data pair No.(1,2), and obtain the p-value as 0.000061. Similar analyses for other data-pairs of each module, and all the p-values are reported in Table IV, which indicate all p-values are below the significance level of $\alpha = 0.05$. The majority of the p-value results being 0.000061 is due to the consistent improvement in all indicators for the corresponding data pairs. This conclusively rejects the null hypothesis of no statistically significant difference, confirming that each module's performance gains are genuine and statistically significant.

2) Ablation Experiments Within Each Module: To assess the individual effects of key components in TGMM, CGIP, and CMAE, respectively, we train ten additional network variants, by either removing or replacing the key components.

a) The impact of different components in TGMM: In Section III-B, TGMM comprises VLG and LVG. VLG generates new language embeddings (the new word embedding f_{ew}^{new} and the new sentence embedding f_{el}^{new}) informed by visual content, while LVG encodes language-aware visual features using newly obtained word embedding f_{ew}^{new}. To evaluate f_{ew}^{new}'s effectiveness, we replace it with the original f_{ew} in LVG (denoted as 're f_{ew}^{new} with ori f_{ew}'). Results in Table.V (a) show performance degradation consistently in terms of fourteen among fifteen evaluation metrics on three sub-datasets, indicating f_{ew}^{new}'s importance in filtering irrelevant language information. Removing both VLG and LVG ('wo VLG&LVG') led to significant performance drops across all metrics, emphasizing their positive impact. Visual examples in Fig.11(a) also demonstrate the superiority of the full TGMM model, with the best mask generation. When we replace f_{ew}^{new} with original f_{ew}, some non-referent object masks are mistakenly highlighted. The similar phenomenon can also be observed in network variant (wo VLG&LVG). All these comparisons highlight the effectiveness of its components.

b) The impact of different components in CGIP: In Section III-C, we sequentially execute cross-graph interaction (CGI) and intra-graph reasoning (IGR). CGI aims to pinpoint crucial pixels of the referent, while IGR aims to improve integrity of the referent. For CGI's language input, we integrate original word embedding f_{ew} with fused new embeddings f_{ew}^{fu} for enhanced language precision. To validate their benefits, we trained four network variants: 'Full CGIP' (full model), 're f_{ew}^{fu} with ori f_{ew}' (the fused language f_{ew}^{fu} is replaced with original word embedding f_{ew}), 'woCGI' (without CGI), and 'woIGR' (without IGR). Results in Table.V(b) show that 're f_{ew}^{fu} with ori f_{ew}' consistently underperforms 'Full CGIP' on three subsets, confirming the new embeddings' positive effect. Comparing 'Full CGIP' with 'woCGI' and 'woIGR', we observed that 'woCGI' and 'woIGR' exhibited a consistent decline across 14 out of the 15 metrics evaluated. These analyses confirm CGI and IGR's roles in pinpointing key pixels and enhancing object integrity. To evaluate CGIP versus Cross-Attention and the significance of multi-head attention weights, we trained another two variants: 'w Cross-Attention' (CGIP is replaced by Cross-Attention) and 'wo multi-head weights' (multi-head weights are removed from CGI). Table.V(b) shows consistent performance drops for

TABLE V
ABLATION RESULTS IN INTERNAL MODULES ON REFCOCO VALIDATION, TESTA, AND TESTB SUBSETS. THE BEST RESULTS ARE REMARKED IN BOLD

Model variants	RefCOCO-val					RefCOCO-testA					RefCOCO-testB				
	P@0.5	P@0.7	P@0.9	oIoU	mIoU	P@0.5	P@0.7	P@0.9	oIoU	mIoU	P@0.5	P@0.7	P@0.9	oIoU	mIoU
(a) The components in TGMM-JE.															
Full TGMM	82.41	**72.96**	**31.44**	**70.12**	**72.28**	**85.93**	**77.41**	**31.59**	**73.19**	**74.73**	**78.00**	**67.40**	**32.50**	**66.72**	**69.23**
re f_{ew}^{new} with ori f_{ew}	**82.58**	72.36	31.00	69.79	72.19	85.38	75.96	30.46	72.14	74.12	77.21	66.79	31.70	65.42	68.89
wo VLG&LVG	81.04	71.34	30.78	68.77	71.02	84.23	74.54	29.95	71.48	73.33	75.74	65.16	31.13	65.01	67.63
(b) The components in CGIP.															
Full CGIP	82.41	**72.96**	**31.44**	**70.12**	**72.28**	**85.93**	**77.41**	**31.59**	**73.19**	**74.73**	**78.00**	**67.40**	32.50	**66.72**	**69.23**
re f_{ew}^{fu} with ori f_{ew}	82.09	72.42	31.08	69.89	71.96	85.47	76.35	30.81	72.47	74.27	76.68	66.63	32.33	65.29	68.51
wo CGI	82.38	71.89	30.99	69.32	72.02	85.35	76.45	31.52	72.50	74.40	77.29	66.85	**32.86**	65.95	69.04
wo IGR	**82.51**	72.41	30.43	69.96	72.05	85.50	76.19	31.24	71.95	73.87	76.74	66.58	32.03	65.53	68.44
w Cross-Attention	81.15	71.83	30.99	69.38	71.37	85.86	76.74	31.43	72.68	74.47	77.25	66.85	32.15	66.13	68.76
wo multi-head weights	82.18	72.24	31.43	69.86	72.01	85.75	77.30	31.57	72.71	74.71	77.39	67.07	32.48	66.28	68.97
(c) The components in CMAE-DE.															
Full CMAE	**82.41**	**72.96**	**31.44**	**70.12**	**72.28**	**85.93**	**77.41**	**31.59**	**73.19**	**74.73**	**78.00**	**67.40**	**32.50**	**66.72**	**69.23**
re f_{el}^{new} with ori f_{el}	81.82	72.25	31.40	69.36	71.74	85.05	76.05	30.78	71.95	73.74	76.17	66.30	31.78	65.10	68.05
wo CMCA	82.34	72.64	30.76	70.01	72.06	85.70	76.70	31.25	72.93	74.47	77.35	66.81	33.07	66.01	68.81
wo CMSA&CMCA	81.83	71.84	30.79	69.63	71.68	85.33	75.78	30.28	72.22	73.99	76.88	66.18	31.23	65.88	68.48

both variants compared to 'Full CGIP', indicating that CGIP outperforms Cross Attention in resolving target localization, and multi-head weights can effectively differentiate head importance, enhancing overall model performance. Fig.11(b) visualizes that 'Full CGIP' yields the optimal mask. Substituting f_{ew}^{fu} with f_{ew} results in misidentified non-referent areas. Removing CGI or IGR leads to missed or incomplete referent masks. These comparisons affirm CGI's role in pinpointing key referent pixels and IGR's contribution to semantic completeness.

c) The impact of different components in CMAE: In Section III-D, CMAE comprises CMSA and CMCA. CMSA highlights spatial locations, while CMCA filters channels, guided by our new language embedding. To prove their efficacy, we tested variants: Full CMAE' (full model), 're f_{el}^{new} with ori f_{el}' (our new language sentence guidance f_{el}^{new} is replaced with original sentence embedding f_{el}), 'wo CMCA' (without CMCA), 'wo CMSA&CMCA' (without CMSA&CMCA). Results in Table.V(c) show 'Full CMAE' outperforms all, validating the effectiveness of our embedding's guidance. Removing CMCA or both CMSA&CMCA degrades performance, confirming CMSA&CMCA's effectiveness in focusing spatial attention and filtering channels. Fig.11(c) visualizes 'Full CMAE's' superior masks. Substituting f_{el}^{new} with original f_{el} misidentifies non-referent objects. Removing CMCA&CMSA causes misses or incomplete masks, emphasizing CMAE's ability to pinpoint spatial locations and critical channels.

3) The Analysis of TGMM in Different Encoding Stages: To verify our design choice, we also apply TGMM to different encoding stages. As shown in the Table.VI, 'TGMM-1∼6' represents employing TGMM from stage-1 to stage-6, and the same true for the others. From Table.VI, we can see that 'TGMM-3∼6' achieves the best performance, while 'TGMM-4∼6' achieves suboptimal performance overall, which can be attributed to the better consistency between deep visual features and text features, and the advantage of cross-modal interaction at a deep level. Therefore, we empirically configure TGMM from stage-3 to stage-6.

TABLE VI
ABLATION RESULTS OF CONFIGURING TGMM IN DIFFERENT ENCODER STAGES ON THE REFCOCO VALIDATION. THE BEST RESULTS ARE IN BOLD

No.	Model variants	RefCOCO-val				
		P@0.5	P@0.7	P@0.9	oIoU	mIoU
1	TGMM-1∼6	81.30	71.75	30.64	68.88	71.29
2	TGMM-2∼6	81.82	71.82	30.93	69.12	71.49
3	TGMM-3∼6	**82.41**	**72.96**	**31.44**	**70.12**	**72.28**
4	TGMM-4∼6	82.02	71.98	31.38	69.47	71.88
5	TGMM-5∼6	80.44	70.33	29.22	68.05	70.35
6	TGMM-6	79.05	69.03	29.61	67.13	69.53

TABLE VII
ABLATION RESULTS OF CGIP WITH DIFFERENT MULTI-HEADS ON THE REFCOCO VALIDATION. THE BEST RESULTS ARE IN BOLD

No.	Multi-heads	RefCOCO-val				
		P@0.5	P@0.7	P@0.9	oIoU	mIoU
1	2	82.10	72.45	31.35	69.85	71.92
2	4	82.27	72.84	**31.88**	69.71	72.10
3	8	**82.41**	**72.96**	31.44	**70.12**	**72.28**
4	16	81.96	72.29	31.24	69.52	71.91

4) The Analysis of Multi-Heads in CGIP Module: In CGIP, the multi-heads are utilized to explore the most representative information in different subspaces, and enhance the feature diversity. To verify the impact of multi-heads number on performance, we trained serval CGIP variants configured with different head numbers (such as 2, 4, 8, 16). As shown in Table.VII, when multi-heads number is 8, the overall performance is optimal. The multi-heads is conducted by splitting feature channels according to the number of multi-heads and calculate them separately. This can lead to a low computational complexity. Therefore, we empirically select the number of multi-heads as 8, which has optimal performance and moderate low computational complexity.

5) Comparison of Different Bidirectional Attentions: To compare different bidirectional attentions for vision-language joint-encoding, we substituted TGMM with parallel bidirectional Word-Pixel-Alignment (WPA) from CoupAlign [45] and Sequential Vision-Language Attention (SVLA) from CARIS [48]. As shown in Table.VIII, SVLA outperforms WPA,

TABLE VIII
ABLATION RESULTS AND COMPUTATION COMPLEXITY ANALYSIS ON DIFFERENT BIDIRECTIONAL ATTENTION METHODS IN ENCODER STAGE. WE UTILIZE OUR SWIN-B VERSION FOR THIS ASSESSMENT. THE BEST RESULTS ARE IN BOLD

Methods	RefCOCO-val					Flops	Params
	P@0.5	P@0.7	P@0.9	oIoU	mIoU		
WPA [45]	84.61	76.93	37.06	73.02	75.06	8.74G	20.94M
SVLA [48]	85.31	77.20	37.11	73.21	75.76	56.59G	95.99M
TGMM	**86.26**	**78.30**	**37.18**	**73.98**	**76.03**	3.90G	18.51M

TABLE IX
COMPUTATIONAL COMPLEXITY ANALYSIS OF MODEL COMPONENTS, AND A COMPARISON WITH SEVERAL ADVANCED MODELS. TO MAINTAIN CONSISTENCY, WE UTILIZE OUR SWIN-B VERSION FOR THIS ASSESSMENT

Models	FLOPs	Params
(a) Analysis of model component complexity.		
baseline	221.39 G	132.03 M
base+TGMM	225.29 G (↑ 3.90)	150.54 M (↑ 18.51)
base+TGMM+CGIP	228.33 G (↑ 3.04)	159.67 M (↑ 9.13)
base+TGMM+CGIP+CMAE	266.00 G (↑ 37.67)	191.27 M (↑ 31.60)
(b) Comparison of complexity with serval advanced models.		
ReSTR [47]	52.30 G	122.90 M
LAVT [12]	248.90 G	118.70 M
ETRIS [15]	105.70 G	85.30 M
CrossVLT [16]	200.40 G	212.7 M
ReLA [73]	109.60 G	195.70 M
PolyFormer [74]	104.50 G	277.20 M
CMIRNet (Ours)	266.00 G	191.27 M

which can be attributed to the sequentially re-encoding visuals with new language expression. But it owns the highest complexity due to multiple cross-attention and self-attention operations. Our TGMM achieves best performance with minimal complexity. This can be attributed to the comprehensive understanding of visual and language in our TGMM, and requires a single affine matrix computation. This highlights the efficacy and efficiency of our novel bidirectional attention mechanisms. This substitution also inspires us that our three visual-language fusion methods can be integrated into some larger models, such as EVF-SAM [77], to further boost task performance. Consequently, our method can provide practical and positive contributions to the development of this field.

D. Computational Complexity Analysis

We employ Flops and network parameters as metrics to quantify model complexity. As evident from Table IX (a), the baseline architecture constitutes the primary source of computational complexity. Notably, our proposed TGMM, CGIP, and CMAE only exhibit lower complexities, with TGMM at 3.9GFlops and 18.51M parameters, CGIP at 3.04GFlops and 9.13M parameters, and CMAE at 37.67GFlops and 31.60M parameters. These modules are universal, capable of enhancing cross-modal alignment capabilities and improving performance without substantially augmenting the overall complexity of the model. Moreover, as illustrated in Table IX (b), our model is compared with existing advanced models, such as PolyFormer, ReLA, CrossVLT, and LAVT, showcasing comparable levels of complexity. Notably, while our model surpasses ETRIS and ReSTR in terms of complexity, it significantly excels in accuracy performance. All these analyses verify that our method possesses appropriate computational complexity with high accuracy performance.

V. CONCLUSION

In this paper, we propose a Cross-Modal Interactive Reasoning Network (CMIRNet) for referring image segmentation. This network is comprised of three key modules, including Text-Guided Multi-modality Joint Encoder (TGMM-JE), Cross-Graph Interactive Positioning (CGIP) module, Cross-Modal Attention Enhanced DEcoder (CMAE-DE). TGMM-JE is first devised to extract the key expression and encode the important visual features under the continuous guidance of language expression. Then, the CGIP is designed to locate the key pixels of the referent object in deep layer, by conducting cross-graph interaction and intra-graph reasoning. Finally, the CMAE-DE is dedicated to refine the object mask from coarse to fine progressively, where hybrid cross-modal attentions are explored to enhance the representation of referent object. Extensive experiments on four benchmark datasets showcase our model's superiority over twenty-two state-of-the-art models.

REFERENCES

[1] R. Hu, M. Rohrbach, and T. Darrell, "Segmentation from natural language expressions," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 108–124.
[2] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, and A. Yuille, "Recurrent multimodal interaction for referring image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1271–1280.
[3] Z. Yang, H. Yu, Y. He, W. Sun, Z.-H. Mao, and A. Mian, "Fully convolutional network-based self-supervised learning for semantic segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 132–142, Jan. 2024.
[4] X. Wang et al., "Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6629–6638.
[5] L. Chen, M. Li, M. Wu, W. Pedrycz, and K. Hirota, "Coupled multimodal emotional feature analysis based on broad-deep fusion networks in human–robot interaction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 9663–9673, Jul. 2024.
[6] J. Chen, Y. Shen, J. Gao, J. Liu, and X. Liu, "Language-based image editing with recurrent attentive models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8721–8729.
[7] X. Chai, F. Shao, Q. Jiang, and Y.-S. Ho, "Roundness-preserving warping for aesthetic enhancement-based stereoscopic image editing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1463–1477, Apr. 2021.
[8] Y. Jing, T. Kong, W. Wang, L. Wang, L. Li, and T. Tan, "Locate then segment: A strong pipeline for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9858–9867.
[9] Y. Jiao et al., "Two-stage visual cues enhancement network for referring image segmentation," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1331–1340.
[10] S. Yang, M. Xia, G. Li, H.-Y. Zhou, and Y. Yu, "Bottom-up shift and reasoning for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11261–11270.
[11] G. Luo et al., "Cascade grouped attention network for referring expression segmentation," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1274–1282.
[12] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, and P. H. S. Torr, "LAVT: Language-aware vision transformer for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18155–18165.
[13] H. Ding, C. Liu, S. Wang, and X. Jiang, "Vision-language transformer and query generation for referring segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16301–16310.

[14] Z. Hu, G. Feng, J. Sun, L. Zhang, and H. Lu, "Bi-directional relationship inferring network for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4423–4432.

[15] Z. Xu, Z. Chen, Y. Zhang, Y. Song, X. Wan, and G. Li, "Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 17457–17466.

[16] Y. Cho, H. Yu, and S.-J. Kang, "Cross-aware early fusion with stage-divided vision and language transformer encoders for referring image segmentation," *IEEE Trans. Multimedia*, vol. 26, pp. 5823–5833, 2024.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[18] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Munich, Germany, 2015, pp. 234–241.

[20] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2881–2890.

[21] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.

[22] J. Qin, J. Wu, X. Xiao, L. Li, and X. Wang, "Activation modulation and recalibration scheme for weakly supervised semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, Jun. 2022, pp. 2117–2125.

[23] J. Wu, H. Fan, Z. Li, G.-H. Liu, and S. Lin, "Information transfer in semi-supervised semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 2, pp. 1174–1185, Feb. 2024.

[24] D. Zhang, C. Li, H. Li, W. Huang, L. Huang, and J. Zhang, "Rethinking alignment and uniformity in unsupervised image semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, Jun. 2023, pp. 11192–11200.

[25] G. Luo et al., "Multi-task collaborative network for joint referring expression comprehension and segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10034–10043.

[26] G. Feng, Z. Hu, L. Zhang, and H. Lu, "Encoder fusion network with co-attention embedding for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15501–15510.

[27] C. Liu, X. Jiang, and H. Ding, "Instance-specific feature propagation for referring segmentation," *IEEE Trans. Multimedia*, vol. 25, pp. 3657–3667, 2022.

[28] L. Lin, P. Yan, X. Xu, S. Yang, K. Zeng, and G. Li, "Structured attention network for referring image segmentation," *IEEE Trans. Multimedia*, vol. 24, pp. 1922–1932, 2022.

[29] T. Hui et al., "Linguistic structure guided context modeling for referring image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 59–75.

[30] D. Li et al., "You only infer once: Cross-modal meta-transfer for referring video object segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, Jun. 2022, pp. 1297–1305.

[31] C. Shang et al., "Cross-modal recurrent semantic comprehension for referring image segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 7, pp. 3229–3242, Dec. 2023.

[32] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–11.

[33] K. Han et al., "Global knowledge calibration for fast open-vocabulary segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 797–807.

[34] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, "SAN: Side adapter network for open-vocabulary semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15546–15561, Dec. 2023.

[35] Y. Liu, S. Bai, G. Li, Y. Wang, and Y. Tang, "Open-vocabulary segmentation with semantic-assisted calibration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 3491–3500.

[36] X. Zou et al., "Segment everything everywhere all at once," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., Red Hook, NY, USA: Curran Associates, 2023, pp. 19769–19782.

[37] Y. Liu, C. Zhang, Y. Wang, J. Wang, Y. Yang, and Y. Tang, "Universal segmentation at arbitrary granularity with language instruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 3459–3469.

[38] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10494–10503.

[39] M. Bellver, C. Ventura, C. Silberer, I. Kazakos, J. Torres, and X. Giro-I-Nieto, "RefVOS: A closer look at referring expressions for video object segmentation," 2020, *arXiv:2010.00263*.

[40] T. Y. Lin, P. Dollàr, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2117–2125.

[41] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.

[42] C. Deng, Q. Wu, Q. Wu, F. Hu, F. Lyu, and M. Tan, "Visual grounding via accumulated attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 3, pp. 1670–1684, Mar. 2022.

[43] J. Tang, G. Zheng, C. Shi, and S. Yang, "Contrastive grouping with transformer for referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 23570–23580.

[44] M. Tan, Z. Wen, L. Fang, and Q. Wu, "Transformer-based relational inference network for complex visual relational reasoning," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 20, no. 1, pp. 1–23, Aug. 2023.

[45] Z. Zhang, Y. Zhu, J. Liu, X. Liang, and W. Ke, "Coupalign: Coupling word-pixel with sentence-mask alignments for referring image segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Red Hook, NY, USA: Curran Associates, 2022, pp. 14729–14742.

[46] Z. Luo et al., "SOC: Semantic-assisted object cluster for referring video object segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36. Red Hook, NY, USA: Curran Associates, 2023, pp. 26425–26437.

[47] N. Kim, D. Kim, S. Kwak, C. Lan, and W. Zeng, "ReSTR: Convolution-free referring image segmentation using transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18145–18154.

[48] S.-A. Liu, Y. Zhang, Z. Qiu, H. Xie, Y. Zhang, and T. Yao, "CARIS: Context-aware referring image segmentation," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 779–788.

[49] J. Bruna, W. Zaremba, A. Szlam, and Y. Lecun, "Spectral networks and locally connected networks on graphs," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2014, pp. 1–9.

[50] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 3844–3852.

[51] B. Jiang, Z. Zhang, D. Lin, J. Tang, and B. Luo, "Semi-supervised learning with graph learning-convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11305–11312.

[52] S. Yang, G. Li, and Y. Yu, "Propagating over phrase relations for one-stage visual grounding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham, Switzerland: Springer, 2020, pp. 589–605.

[53] D. Liu, H. Zhang, Z.-J. Zha, M. Wang, and Q. Sun, "Joint visual grounding with language scene graphs," 2019, *arXiv:1906.03561*.

[54] R. Zeng et al., "Graph convolutional module for temporal action localization in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6209–6223, Oct. 2022.

[55] S. Chen and B. Li, "Multi-modal dynamic graph transformer for visual grounding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15534–15543.

[56] S. Huang et al., "Referring image segmentation via cross-modal progressive comprehension," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10488–10497.

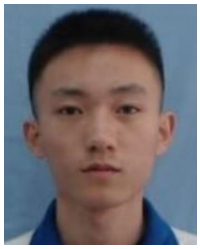
[57] S. Liu, T. Hui, S. Huang, Y. Wei, B. Li, and G. Li, "Cross-modal progressive comprehension for referring segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4761–4775, Sep. 2022.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[59] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[60] X. Li, Y. Yang, Q. Zhao, T. Shen, Z. Lin, and H. Liu, "Spatial pyramid based graph reasoning for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8947–8956.
[61] B. Zhang, J. Xiao, J. Jiao, Y. Wei, and Y. Zhao, "Affinity attention graph neural network for weakly supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8082–8096, Nov. 2022.
[62] X. Zeng, M. Xu, Y. Hu, H. Tang, Y. Hu, and L. Nie, "Adaptive edge-aware semantic interaction network for salient object detection in optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5617416.
[63] V. K. Nagaraja, V. I. Morariu, and L. S. Davis, "Modeling context between objects for referring expression understanding," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 792–807.
[64] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "ReferItGame: Referring to objects in photographs of natural scenes," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 787–798.
[65] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2014, pp. 740–755.
[66] H. J. Escalante et al., "The segmented and annotated IAPR TC-12 benchmark," *Comput. Vis. Image Understand.*, vol. 114, no. 4, pp. 419–428, Apr. 2010.
[67] Z. Wang et al., "CRIS: CLIP-driven referring image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11686–11695.
[68] H. Li, M. Sun, J. Xiao, E. G. Lim, and Y. Zhao, "Fully and weakly supervised referring expression segmentation with end-to-end learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 10, pp. 5999–6012, Jun. 2023.
[69] C. Liu, H. Ding, Y. Zhang, and X. Jiang, "Multi-modal mutual attention and iterative interaction for referring image segmentation," *IEEE Trans. Image Process.*, vol. 32, pp. 3054–3065, 2023.
[70] J. Liu, H. Tan, Y. Hu, Y. Sun, H. Wang, and B. Yin, "Global and local interactive perception network for referring image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 11, 2024, doi: 10.1109/TNNLS.2023.3308550.
[71] Y. Hu et al., "Beyond one-to-one: Rethinking the referring image segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4044–4054.
[72] J. Yang, L. Zhang, and H. Lu, "Referring image segmentation with fine-grained semantic funneling infusion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 10, pp. 14727–14738, Oct. 2024.
[73] C. Liu, H. Ding, and X. Jiang, "GRES: Generalized referring expression segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2023, pp. 23592–23601.
[74] J. Liu et al., "PolyFormer: Referring image segmentation as sequential polygon generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 18653–18663.
[75] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[76] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, pp. 32–73, Feb. 2017.
[77] Y. Zhang et al., "EVF-SAM: Early vision-language fusion for text-prompted segment anything model," 2024, *arXiv:2406.20076*.

Mingzhu Xu (Member, IEEE) received the B.S., M.Sc., and Ph.D. degrees from Harbin Institute of Technology (HIT), Harbin, China, in 2013, 2015, and 2021, respectively. He is currently an Assistant Professor with the School of Software, Shandong University, Jinan, China. His research interests include computer vision, multimedia computing, and information retrieval. He is also an Invited Reviewer for prestigious journals, including IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, *Information Science*, ACM MM, NeurIPS, and ICML.



Tianxiang Xiao is currently pursuing the B.S. degree with the School of Software, Shandong University, Jinan, China. His research interests include computer vision and referring image segmentation.

Yutong Liu is currently pursuing the B.S. degree with the School of Software, Shandong University, Jinan, China. Her research interests include computer vision, referring image segmentation, and visual saliency analysis.

Haoyu Tang (Member, IEEE) received the B.S. and Ph.D. degrees from Xi'an Jiaotong University, China, in 2016 and 2021, respectively. He is currently an Assistant Professor with the School of Software, Shandong University. His research interests include machine learning and multimedia retrieval.

Yupeng Hu (Member, IEEE) received the Ph.D. degree in software engineering from Shandong University. He is currently an Associate Professor with the School of Software, Shandong University. His research interests include information retrieval and data mining. Various parts of his work have been published in famous journals and forums, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, *Science China Information Sciences*, and ACM Multimedia. He has served as a PC Member for ACM MM, ACL, and AAAI; and a reviewer for IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING and IEEE TRANSACTIONS ON MULTIMEDIA.

Liqiang Nie (Senior Member, IEEE) received the B.Eng. degree from Xi'an Jiaotong University and the Ph.D. degree from the National University of Singapore (NUS). He is currently the Dean of the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen Campus). He has co-authored more than 100 CCF-A articles and five books, with 20k plus Google Scholar citations. His research interests lie primarily in multimedia content analysis and information retrieval. He is a fellow of AAIA and IAPR. He is a member of ICME Steering Committee. He has received many awards over the past three years, like ACM MM, and SIGIR best paper honorable mention in 2019, the AI 2000 most Influential Scholars 2020, SIGMM Rising Star in 2020, MIT TR35 China 2020, DAMO Academy Young Fellow in 2020, SIGIR Best Student Paper in 2021, First Price of the Provincial Science and Technology Progress Award in 2021 (rank 1), and Provincial Youth Science and Technology Award in 2022. Some of his research outputs have been integrated into the products of Alibaba, Kwai, and other listed companies. Meanwhile, he is also the Regular Area Chair or the SPC of ACM MM, NeurIPS, IJCAI, and AAAI. He is an Associate Editor (AE) of IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, ACM ToMM, and *Information Science*.