Mask prior generation with language queries guided networks for referring image segmentation

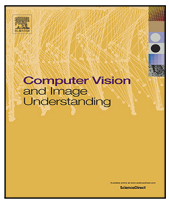
Zhou, J.; Xiao, G.; Lew, M.S.K.; Wu, S.

Computer Vision and Image Understanding 253 (2025) 104296

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



Mask prior generation with language queries guided networks for referring image segmentation

Jinhao Zhou [a], Guoqiang Xiao [a], Michael S. Lew [c], Song Wu [a,b,*]

[a] College of Computer and Information Science, Southwest University, Chongqing 400715, China
[b] Yibin Academy of Southwest University, Yibin, Sichuan 644000, China
[c] LIACS Media Lab, Leiden University, Leiden, Netherlands

ARTICLE INFO

Keywords:
Referring image segmentation
Bidirectional spatial alignment
Channel attention fusion gate
Mask prior generator

ABSTRACT

The aim of Referring Image Segmentation (RIS) is to generate a pixel-level mask to accurately segment the target object according to its natural language expression. Previous RIS methods ignore exploring the significant language information in both the encoder and decoder stages, and simply use an upsampling-convolution operation to obtain the prediction mask, resulting in inaccurate visual object locating. Thus, this paper proposes a Mask Prior Generation with Language Queries Guided Network (MPG-LQGNet). In the encoder of MPG-LQGNet, a Bidirectional Spatial Alignment Module (BSAM) is designed to realize the bidirectional fusion for both vision and language embeddings, generating additional language queries to understand both the locating of targets and the semantics of the language. Moreover, a Channel Attention Fusion Gate (CAFG) is designed to enhance the exploration of the significance of the cross-modal embeddings. In the decoder of the MPG-LQGNet, the Language Query Guided Mask Prior Generator (LQPG) is designed to utilize the generated language queries to activate significant information in the upsampled decoding features, obtaining the more accurate mask prior that guides the final prediction. Extensive experiments on RefCOCO series datasets show that our method consistently improves over state-of-the-art methods. The source code of our MPG-LQGNet is available at https://github.com/SWU-CS-MediaLab/MPG-LQGNet.

1. Introduction

The Referring Image Segmentation (RIS) is an emerging vision and language cross-modal semantic understanding task, which aims to generate a pixel-level mask to segment the target object in an image according to the given textual description of the target (Hu et al., 2016). Compared to the traditional single-modal visual segmentation tasks that rely on fixed categories, the RIS task has to handle free natural language expressions. The target object in this task is inferred according to unconstrained text descriptions, which include words and phrases conveying concepts such as the presented target entity, its actions, attributes, and location. Thus, the RIS has the advantage of using natural language as an interactive interface for applications of language-based human–computer interaction (Ahn et al., 2018), image editing (Ling et al., 2021; Chen et al., 2018), autonomous driving (Codevilla et al., 2019; Toromanoff et al., 2020), etc.

Recently, an increasing amount of research has been dedicated to the RIS task. Considering the randomness and openness of the referring text descriptions, its length may vary from a few words to long sentences, and as the text becomes longer, the difficulty of semantic analysis between vision and language modalities increases significantly. Thus, the key challenge of RIS task is effectively aligning the semantic representations between image and text modality domains. With the advantage of learning long-range dependencies, the recent attention mechanism (Vaswani et al., 2017) has become an attractive architecture in both Natural Language Processing (NLP) and Computer Vision (CV) tasks. The attention mechanism in the Transformer has shown great potential in cross-modal alignment, enabling fine-grained interaction between inter-modalities and intra-modalities. Thus, for the RIS task, various cross-modality attention mechanisms have been proposed to align and fuse the learned features of images and text in the encoding or decoding stage (Ding et al., 2021; Wang et al., 2022; Yang et al., 2022; Wu et al., 2024b; Liu et al., 2023), and the evaluation has demonstrated that the alignment and fusion operations in the early encoding stage have the advantage to explore detailed information in the low-level cross-modalities feature representations. Fig. 1(a) shows

* Corresponding author at: College of Computer and Information Science, Southwest University, Chongqing 400715, China.
E-mail addresses: zjh670105603@email.swu.edu.cn (J. Zhou), gqxiao@swu.edu.cn (G. Xiao), m.s.k.lew@liacs.leidenuniv.nl (M.S. Lew), songwuswu@swu.edu.cn (S. Wu).

https://doi.org/10.1016/j.cviu.2025.104296
Received 20 May 2024; Received in revised form 15 January 2025; Accepted 16 January 2025
Available online 29 January 2025

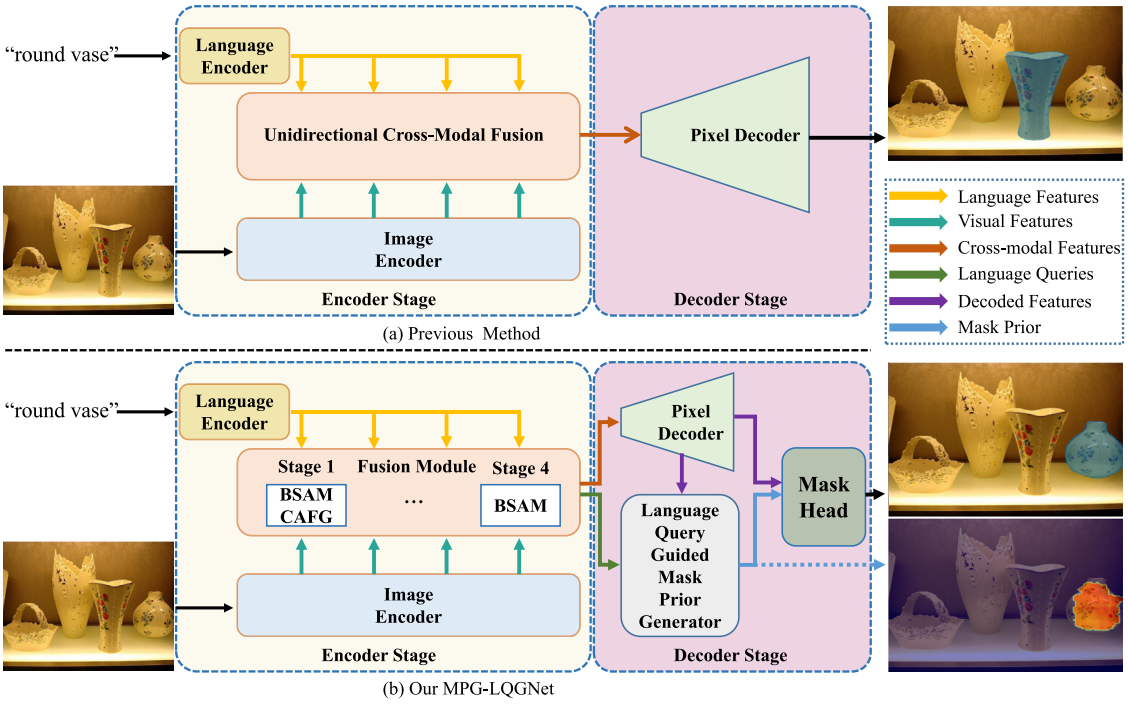


Fig. 1. Illustration of our method and previous state-of-the-art method using early fusion. Our approach utilizes the mask prior generated during the decoding stage to enhance the ability to locate the target. BSAM: Bidirectional Spatial Alignment Module. CAFG: Channel Attention Fusion Gate.

that most recent methods perform the fusion operation in the encoder to generate cross-modality features; however, in the subsequent decoding process, the commonly utilized pixel-based decoder with step-by-step upsampling operation results in imprecise or even completely non-target results. This is mainly because the upsampling-convolution segmentation head is usually employed to obtain mask predictions from low-resolution visual feature maps, which are primarily dominated by visual information but limit the inter-understanding capability between the image and text modalities (Yang et al., 2022; Kim et al., 2022; Li et al., 2021); meanwhile, the loss of detailed information during upsampling further exacerbate the inaccurate results.

Therefore, a novel framework named Mask Prior Generation with Language Queries Guided Network (MPG-LQGNet) is proposed in this paper. The proposed MPG-LQGNet aims to enhance image-text alignment and interaction capability in both the encoding and decoding stages to generate more accurate mask predictions. Specifically, although previous methods provide an effective fusion strategy based on the cross-modality attention mechanism in the encoding stage (Yang et al., 2022; Ding et al., 2021), it is a unidirectional alignment like Fig. 1(a) shows (from text to vision). We argue that generating fused features based on the cross-attention mechanism from vision to text at each intermediate layer of the image encoder can also contain rich cross-modal information. Therefore, in the encoder of our MPG-LQGNet, an effective Bidirectional Spatial Alignment Module (BSAM) is designed to realize the alignment and fusion both from text to vision and from vision to text at different scale stages. Meanwhile, the fused features from vision to text can be used as query vectors in the attention mechanism of the decoding stage; thus, we named them as language query vectors. This process is similar to a comprehensive understanding mechanism, as the language queries obtained in the low-level stage are more inclined towards locating ability, while the language queries from the high-level stage can obtain more semantic understanding about the target objects. Moreover, in the encoding stage, a Channel Attention Fusion Gate (CAFG) is designed to weigh the channels from a global feature perspective to enhance the representation of cross-modal fusion features for the target object. Meanwhile, in the decoder of our MPG-LQGNet, a novel module named Language Query Guided Mask Prior Generator (LQPG) is designed, which generates a mask prior. The mask prior represents a rough localization of the target, which does not fully encompass the entire object along the boundaries like the final prediction does. However, the mask prior serves as a restrictive activation for the target area, guiding the model to focus more on this region during final prediction and mitigating the attention given to the background.

Overall, the main contributions of our MPG-LQGNet are summarized as follows:

- The encoder stage incorporates a Bidirectional Spatial Alignment Module (BSAM) to integrate text and image features and generates language queries. Additionally, a Channel Attention Fusion Gate (CAFG) is implemented to enhance the discrimination of the cross-modal features.
- In the decoding process, we employ a Language Query Guided Mask Prior Generator (LQPG) to highlight target regions using generated language queries for mask prior acquisition.
- We evaluate our method on three popular datasets for the referring image segmentation task, namely RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016), and G-ref (Nagaraja et al., 2016). The results demonstrate that our method surpasses the performance of state-of-the-art RIS methods by a considerable margin. Particularly, our method achieves an average improvement of around 2% on the challenging RefCOCO+ evaluation datasets, as measured by the overall IoU metric.

2. Related work

2.1. Referring image segmentation

The RIS task aims to categorize pixels into target regions or backgrounds according to the given text. SNLE (Hu et al., 2016) is the pioneer work of referring image segmentation. It firstly extracts features from image and text using Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM), respectively, then fuses them and applies a Fully Convolutional Network (Long et al., 2015) on the

fused feature. Similarly, Li et al. (2018) used the Convolutional-LSTM Network to obtain richer multi-modal features by adding image features when encoding each word. In the early stages, the fusion of features primarily involved concatenation-convolution. Although it is simple, it ignores the intricate language expression and structure in the referring text. Graph-based RIS methods were proposed to handle the interactions between complex textual descriptions and visual objects. Huang et al. (2020) identified all the entities and utilized graph edges to model the information flows among entities (graph vertexes). Yang et al. (2021) also used a graph structure for more interpretable reasoning. Later, additional methods utilized the attention mechanism and multi-scale feature aggregation for modeling. Specifically, Ye et al. (2019) used self-attention to better understand the semantics of both modality data. Hu et al. (2020) constructed a bidirectional attention relationship using both the influence of two modality information on each other. Some efforts also capitalized on the multi-scale information of the network. Jain and Gandhi (2021) proposed a novel hierarchical cross-modal aggregation module to exchange contextual information across visual hierarchies. Feng et al. (2021) designed a co-attention mechanism to use language to refine the multi-scale visual features progressively. In addition, LTS (Jing et al., 2021) and MaIL (Li et al., 2021) employed position prior and instance masks, respectively, to achieve more accurate segmentation results. Moreover, VLT (Ding et al., 2021) introduced a query generation module based on the transformer framework to handle the diversity of language comprehension. Other works also incorporated contrastive learning in the referring image segmentation. For example, CRIS (Wang et al., 2022) transferred knowledge from the large model Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) and performed contrastive learning between pixel and textual features. Unlike Ding et al. (2021), Wang et al. (2022), which performs cross-modal fusion in the decoder stage, LAVT (Yang et al., 2022) proposed a hierarchical language-aware visual encoding scheme to perform feature fusion in the encoder stage. Based on the early fusion strategy, Zhang et al. (2022) simultaneously utilized word-pixel alignment and sentence-mask alignment to fully integrate language information. Most recently, Tang et al. (2023) proposed a Group Transformer to capture object-level information by grouping visual features into different regions. Furthermore, both Yang et al. (2023) and Xu et al. (2023) proposed a method based on the existing models to enhance their performance. Yang et al. (2023) strengthened the target-related feature representation through dynamic convolution and iterative progression. Xu et al. (2023) changed the training approach from a meta-learning perspective by constructing a virtual test set with new combinations of text to update the model and improve its generalization performance.

However, problems such as inaccurate location still exist in previous approaches. For example, MaIL (Li et al., 2021) uses instance masks from Mask Region-based Convolutional Neural Network (Mask R-CNN) (He et al., 2017) as an additional modality. These masks do not incorporate textual information, resulting in imprecise predictions. Additionally, LTS (Jing et al., 2021) propose a position prior as a coarse segmentation mask to guide the final prediction. However, this method lacks fine-grained cross-modal interaction, and the feature fusion only occurs in the high-level stage, resulting in less accurate prior knowledge. Considering these issues, we design both spatial and channel dimension feature fusion in the encoder stage to enhance the representation of cross-modal features. We also calculate cross-attention maps in the decoder stage using the generated language queries that involve visual information, achieving a more accurate and comprehensive mask prior to understanding language semantics better, leading to more precise prediction results.

2.2. Transformer

Transformer initially emerged as a multi-head self-attention-based deep neural model that was primarily used for Natural Language Processing (NLP) (Vaswani et al., 2017). Its ability to handle long-range dependencies and capture global relationship patterns makes it highly effective in these tasks. Due to its potential advantages, transformer has also been applied to various Computer Vision (CV) and multi-modal tasks (Wu et al., 2024a; Liu et al., 2024), including object detection (Carion et al., 2020; Zhu et al., 2020), semantic segmentation (Strudel et al., 2021; Zheng et al., 2021) and visual question answering (Mashrur et al., 2024; Jiang et al., 2024).

Here are also some other multi-modal models that use transformers to process text and image information. For instance, CLIP (Radford et al., 2021) employs contrastive learning on both the visual and language features. ViLT (Kim et al., 2021) proposes a pre-trained model for vision-language tasks without convolution. However, these works cannot be directly applied to the RIS task that requires fine-grained (pixel-level) interaction. Moreover, previous works that use transformers only fused language and visual information at specific stages (Yang et al., 2022; Kim et al., 2022; Jing et al., 2021), and the subsequent operations only use the obtained multi-modal features for convolution and other processing to generate the final predictions. In our work, we also take full advantage of the potential of attention mechanisms and implement cross-modal interactions in both the encoder and decoder phases. By incorporating a fine-grained fusion, our model can understand cross-modal information comprehensively.

3. Methodology

3.1. Overview

Fig. 2 illustrates the detailed pipeline of our MPG-LQGNet, which is designed based on the encoder–decoder architecture. During the encoder stage, firstly, the features of the input image are extracted in each intermediate layer of the image encoder, and the feature of the input text is also abstracted from the language encoder. Then, the extracted cross-modal features are fused using the designed Bidirectional Spatial Alignment Module (BSAM) to generate the fused cross-modal feature and language query vector for each encoder stage. Moreover, the Channel Attention Fusion Gate (CAFG) is designed to weight and fuse the cross-modal features based on the channel attention mechanism. During the decoder stage, firstly, the pixel decoder upsamples the pixel-level feature maps and obtains decoding features at four different scales. Then, in the designed Language Query Guided Mask Prior Generator (LQPG), we interact the decoding features from the first three scales with the language queries through cross-attention operation to obtain a mask prior and further refine the language queries for the next layer. It is worth noting that we supervise the mask prior generated at each layer using the ground-truth mask. Finally, in the mask head architecture, we concatenate the decoding features at the highest resolution with the mask prior and use coordinate convolution to obtain the final segmentation results.

3.2. Encoder

3.2.1. Visual and language feature extraction

Following the previous works (Yang et al., 2022; Tang et al., 2023), we adopt Swin Transformer (Liu et al., 2021) and BERT (Kenton and Toutanova, 2019) as our visual encoder and language encoder respectively for fair comparison. Given a language expression contains N_t words, we extract the textual embedding of each word using the language representation model BERT without the last pooling layer, which is denoted as $L \in R^{N_t \times C_t}$, where C_t denote the number of channels. For an input image $I \in R^{H \times W \times 3}$, we use the multi-scale visual features from the four stages of Swin Transformer, denoted as $V_i \in R^{H_i \times W_i \times C_i}$, $i \in \{1, 2, 3, 4\}$, and H_i, W_i, C_i are the height, width and channel dimension of V_i at the ith stage, respectively. Furthermore, in the encoder, we propose incorporating the designed Bidirectional Spatial Alignment Module (BSAM) and Channel Attention Fusion Gate (CAFG) to selectively focus on both fine-grained and global features, respectively, thereby enhancing the discrimination of the fused cross-modal feature representation for the visual target, as shown in Fig. 2(a). The details of each module will be introduced in the following sections.

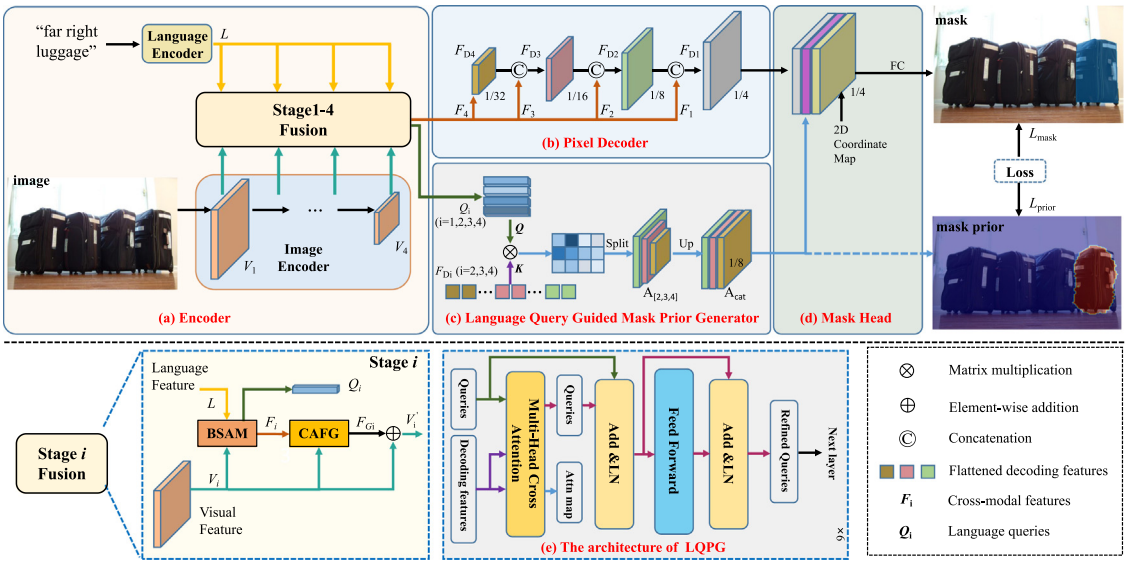


Fig. 2. The detailed flowchart of our proposed MPG-LQGNet. The fusion module in the bottom left corner illustrates the process of generating cross-modal features and language queries at each stage. The part (c) mainly shows the process of generating mask prior at each layer in our proposed Language Query Guided Mask Prior Generator (LQPG). The inputs of the LQPG module are language queries and flattened pixel features concatenated along the spatial dimension from the pixel decoder. In the mask head module, we concatenate the decoding feature F_{D1} and mask prior and use coordinate convolution to get the final segmentation result. The part (e) shows the overview of LQPG (Section 3.3.2). BSAM: Bidirectional Spatial Alignment Module (Section 3.2.2). CAFG: Channel Attention Fusion Gate (Section 3.2.3).

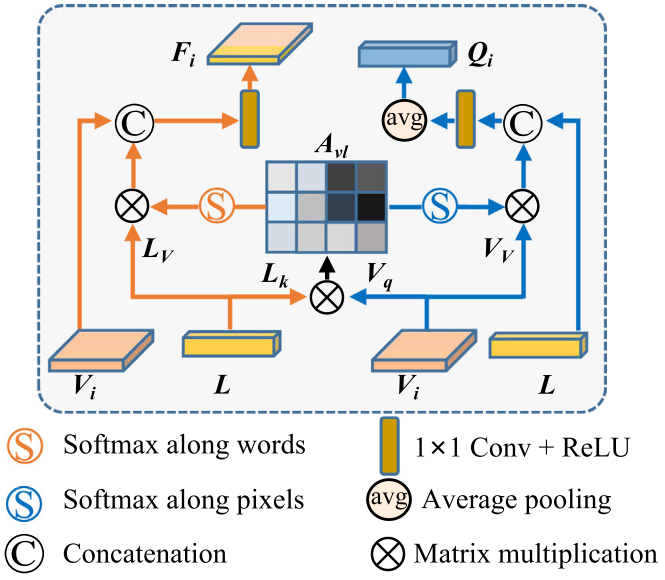


Fig. 3. Pipeline of the proposed Bidirectional Spatial Alignment Module (BSAM).

3.2.2. Bidirectional spatial alignment module (BSAM)

Due to the presence of segmentation targets with significant scale differences in the referring segmentation task (an object or a region), it is difficult to accurately perceive the target area by solely using high-level features extracted from the visual encoder. Meanwhile, the higher downsampling rate in visual feature extraction also leads to a large number of missing object details and outlines in the high-level features, especially in terms of positional information, resulting in inaccurate predictions for the target's edge regions. Previous works, such as Yang et al. (2022), Zhang et al. (2022), have made full use of hierarchical visual transformer (Liu et al., 2021) and conducted cross-modal fusion during the encoding phase to address the referring segmentation problem. Inspired by this, in contrast to previous works, we designed a novel Bidirectional Spatial Alignment Module (BSAM) architecture to fully leverage multi-modal information from different scales and obtain additional language queries from each stage for subsequent processing.

Fig. 3 depicts the schematic representation of our proposed architecture of BSAM. The BSAM aims to achieve bidirectional cross-modal fusion through three main steps. Firstly, to obtain the spatial attention map between pixels and words, we need to transform the channel dimension of language features to match the channel quantity of image features. This is achieved through linear projections, and the specific steps are as follows:

$$V_q = norm(W_{vq}(V_i)), i \in \{1,2,3,4\}, \tag{1}$$

$$L_k = W_{lk}(L), \tag{2}$$

$$V_v = norm(W_{vv}(V_i)), \tag{3}$$

$$L_v = W_{lv}(L), \tag{4}$$

$$A_{vl} = \frac{V_q^T L_k}{\sqrt{C_i}}, \tag{5}$$

where W_{vq}, W_{lk}, W_{vv} and W_{lv} are all projection functions with C_i number of output channels. V_i and L are the original visual features and language features at the ith stage. Here, *norm* indicates instance normalization. Secondly, for the obtained attention weight map $A_{vl} \in R^{HW \times N_t}$ for pixels-words, we proceed to perform vision-language bidirectional alignment by incorporating the value embeddings (V_v and L_v) from one modality into another modality:

$$V_{fi} = [V_i; softmax(A_{vl})L_v], \tag{6}$$

$$L_f = [L; softmax(A_{vl}^T)V_v], \tag{7}$$

where [;] denotes concatenation along the channel dimension. After applying the softmax function to calculate the probability values for each word and pixel, respectively, we can obtain the updated fused visual and textual features V_{fi} and L_f. Finally, we can obtain the output cross-modal aligned and fused feature and language query at each stage as follows:

$$F_i = W_{ov}(V_{fi}), \tag{8}$$

$$Q_i = avg(W_{ol}(L_f)), \tag{9}$$

where W_{ov} and W_{ol} are both linear layer followed by the ReLU activation (Nair and Hinton, 2010) and the *avg* represents the average

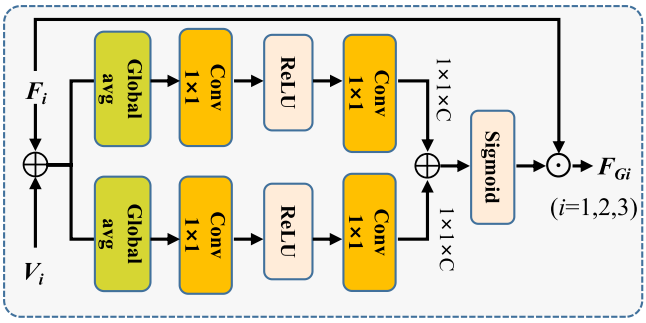


Fig. 4. Pipeline of the proposed Channel Attention Fusion Gate (CAFG).

pooling operation along words. Note that we transform the channel dimension of the cross-modal feature $F_i \in R^{H_i \times W_i \times C_i}$ and the language query $Q_i \in R^{1 \times C_q}$ into C_i and C_q, respectively, for subsequent computations. Here, C_i and C_q represent the original number of visual feature channels and the hidden layer size in the Language Query Guided Mask Prior Generator (LQPG).

So far, we have obtained four sets of aligned and fused cross-modal features and language queries from different stages of the encoder. These features retain both the localization and semantic information from low-level to high-level phases, which can guide the model to generate more accurate segmentation results during the decoding stage. Furthermore, subsequent experiments have also demonstrated the effectiveness of obtaining language queries from the high-resolution image scale.

Although similar methods using bidirectional attention, such as co-attention (Feng et al., 2021) and mutual attention (Liu et al., 2023), have been utilized in the past, these approaches also treated the two modalities of data as value embeddings. However, they only produce a single output, namely cross-modal features for subsequent encoding or decoding stages. Additionally, in the Feng et al. (2021), Yang et al. (2022) methods that also employ encoder fusion strategies, the cross-modal features obtained through a cross-modality attention mechanism are further fed into the next visual encoding block as a form of visual feature maps. Conversely, our designed BSAM is mainly aimed at obtaining additional language queries at each stage for subsequent decoding phases, thereby enhancing the comprehensive utilization of language information.

3.2.3. Channel attention fusion gate (CAFG)

In the process of obtaining fused cross-modal features F_i, the value embeddings are projected from L, so F_i indicates a set of word features for representing pixel features. Typically, after this step, it is common practice to incorporate image features V_i and apply residual processing to preserve the original visual information. The previously proposed Language Gate (LG) mechanism (Yang et al., 2022) aims to prevent the fused cross-modal features F_i from overwhelming the original visual signals V_i. The LG learns a set of spatial weight maps to rescale each element in F_i. Drawing inspiration from Hu et al. (2018), Dai et al. (2021), in contrast to previous methods, we learn a set of weights from the channel dimension of the feature maps to perform channel-wise feature scaling. This is because different channel features carry varying degrees of crucial information. To reduce redundancy and enhance the representation of relevant channel features during skip connection, we introduce a Channel Attention Fusion Gate (CAFG) mechanism that learns from a global perspective, as depicted in Fig. 4. Additionally, taking the advantages of multi-scale channel attention mechanism (Dai et al., 2021) that can learn two sets of weights at different scales, considering the previous implementation of spatial vision-language alignment, to preserve the cross-modal alignment information of finely fused features in the spatial dimension, we design the CAFG mechanism to have two pathways that both learn global channel weights, as described below:

$$CA = norm(Conv_2(ReLU(norm(Conv_1(GAP(X)))))), \tag{10}$$

$$\begin{aligned} F_{Gi} = & F_i \odot \sigma(CA_1(F_i + V_i) \oplus \\ & CA_2(F_i + V_i)), i \in \{1, 2, 3\}, \end{aligned} \tag{11}$$

where σ is the sigmoid function, and GAP is the global average pooling. And CA represents the channel attention algorithm. $\oplus$ denotes the element-wise addition and $\odot$ denotes broadcasting element-wise multiplication. Here, *norm* is the Batch Normalization, and *Conv* is the projection function with 1×1 kernel size to change the number of channels. $F_{Gi} \in R^{H_i \times W_i \times C_i}$ are the output feature map, which will add $V_i \in R^{H_i \times W_i \times C_i}$ to obtain the input feature map $V_i' \in R^{H_i \times W_i \times C_i}$ of the next stage in the visual encoder.

The mechanism of channel attention was widely used after it was proposed in Hu et al. (2018), and its main purpose is to enhance the representation ability of target features. Subsequent ablation experiments also indicate that, compared to learning spatial weights through language gate (Yang et al., 2022), our method of learning weights only in the channel dimension is a more effective approach. Therefore, based on the standard channel attention, we designed such a gating mechanism to better control cross-modal information, so that it can continue visual encoding after being combined with pure visual information.

3.3. Decoder

3.3.1. Pixel decoder

In the decoder stage, we initially obtain progressively upsampled multi-scale decoding feature maps F_{Di} with resolutions of $\frac{1}{32}$, $\frac{1}{16}$, $\frac{1}{8}$, and $\frac{1}{4}$ of the original spatial size, respectively. The four scales of the decoding features obtained will be used for generating mask prior and the final segmentation results. Following the process in Yang et al. (2022), we take the corresponding scale cross-modal features F_i from the encoder as the input of the decoder, as depicted in Fig. 2(b), and can further obtain the pixel-level decoded features based on the following described recursive function:

$$\begin{cases} F'_{D4} & = F_4 \\ F'_{Di} & = Up(W_{di}([F'_{Di+1}; F_i])), i = 3, 2, 1, \end{cases} \tag{12}$$

where [;] denotes concatenation and *Up* denotes 2× bilinear upsampling. W_{di} is a two-layer 3×3 convolution network followed by batch normalization and ReLU activation function (Nair and Hinton, 2010).

In order to calculate the features obtained at this stage with the language queries $Q_i \in R^{1 \times C_q}$ generated by the encoder, we further use linear projection w_q to change the channel dimension of the decoding features F_{Di} to C_q:

$$F_{Di} = norm(W_q(F'_{Di})), i \in \{1, 2, 3, 4\}, \tag{13}$$

where *norm* is the group normalization operation.

3.3.2. Language query guided mask prior generator

In previous methods (Ding et al., 2021; Yang et al., 2022; Kim et al., 2022), the most common approach to restore the size of the feature map and generate segmentation results was only based on the operations of convolution and upsampling. However, considering that the supervision signal for the referring image segmentation, like semantic segmentation, is also the mask of the ground-truth label, these previous methods based on a fully convolutional network are mainly guided by visual information. Additionally, the process of upsampling to change the size is a fixed design that may result in the loss of detailed information about the target objects, such as their contours. To enable the model to comprehensively understand cross-modal semantic information and fully utilize the textual expression in the referring image segmentation task, we designed a novel module of Language

Query Guided Mask Prior Generator (LQPG) to allow the model to perceive the visual target objects more accurately.

As shown in Fig. 2(e), the mask prior generator consists of several layers, with each layer comprising of Multi-Head Cross Attention (MHCA) layer (Vaswani et al., 2017) and a Feed-Forward Network. For each layer, as illustrated in Fig. 2(c), the key embeddings and value embeddings are projected from the decoder features $\{F_{Di} \in R^{H_i \times W_i \times C_q}\}_{i=2}^{4}$ with lower resolution. Query embeddings are projected from the language queries $Q \in R^{4 \times C_q}$. Before implementing the projection functions, we flatten the decoder features F_{Di} to transform the spatial dimension and then calculate with Q to fetch the attention map $A \in R^{4 \times N_p \times h}$, as follows:

$$F_k = Concat(flatten(F_{D2}, F_{D3}, F_{D4})) + e_p, \tag{14}$$

$$A, Q' = MHCA(F_k, Q), \tag{15}$$

where h is the number of attention heads, $e_p \in R^{N_p \times C_q}$ denotes the sinusoidal positional embedding and N_p indicates the number of flattened pixel features F_{Di}. $Concat()$ is the concatenation along the spatial dimension. After obtaining the output queries $Q' \in R^{4 \times C_q}$ from the middle layer, we feed them into a Feed-Forward network followed by a residual operation and Layer Normalization (LN) to derive the refined queries, which will be fed into the next layer.

In Li et al. (2022), the segmentation result is generated based on the attention weights. Inspired by this, in order to obtain the mask prior to better guide the model in accurately locating the target region, we first split and unflatten the attention map A into attention maps A_2, A_3 and A_4, which have the same spatial resolution as F_{D2}, F_{D3} and F_{D4}. Later, we take the obtained attention maps of the three scales and process them through a linear layer. Specifically, we apply a 2× upsampling on attention map A_3 and a 4× upsampling on attention map A_4. In this way, we get the attention maps are all recovered to $\frac{1}{8}$ of the initial image size, and they are concatenated along the channel dimension to get A_{cat}. Finally, through passing the fused attention maps A_{cat} through a fully connected layer and applying a 2× upsampling, we can generate the mask prior $m_p \in R^{\frac{H}{4} \times \frac{W}{4} \times 2}$, which is a two-class score feature map.

Based on the cross-attention operation, we utilize language queries to further perceive the multi-modal information of the decoded features. By continuously learning through the mask prior generator at each layer, we are able to query and activate the relevant regions while suppressing the background or irrelevant parts. Higher cross-attention scores reflect the localization information of the target regions and enable the recovery of fine-grained details of the target object, such as contours and edge information. Additionally, we adopt a deep supervision learning strategy, which means that the mask prior of each layer is supervised by the ground truth mask. This allows the module to focus on the useful information for segmentation in earlier stages and progressively refine the generated mask prior at each layer. The mask prior obtained from the final layer is used as the ultimate output of this module.

Unlike the random initialization of query used in Zhang et al. (2022) and Li et al. (2022), the language queries we propose have already carried multi-modal information representing the target object after the encoding stage, extracting relevant regions from visual features at different scales. Specifically, while Li et al. (2022) proposed to first detect the bounding box of the target to provide its location information, our proposed mask prior is generated through attention weights, resulting in irregular polygon activation map that can incorporate finer and more accurate localization information. The obtained mask prior is coarser compared to the final mask, potentially leading to poorer performance at the boundaries than the final predicted result. However, it primarily represents a restrictive activation area of the target object and utilizes coordinate convolution mentioned later to guide the final prediction.

3.3.3. Mask head and loss function

As illustrated in Fig. 2(d), the aim of the mask head is to predict the final segmentation mask. We take the feature map $F_{D1} \in R^{\frac{H}{4} \times \frac{W}{4} \times C_q}$ from pixel decoder and mask prior $m_p \in R^{\frac{H}{4} \times \frac{W}{4} \times 2}$ as input. The prior information can guide the decoded feature to generate segmentation results with more accurate localization. Following Li et al. (2021), Liu et al. (2018), we utilize a 2-D spatial coordinate map $F_o \in R^{\frac{H}{4} \times \frac{W}{4} \times 2}$ to perceive the spatial information:

$$Y = W_m(Concat(F_{D1}, m_p, F_o)), \tag{16}$$

where $Concat()$ is the concatenation along channel dimension and W_m denotes 1×1 convolution. The final output feature map $Y \in R^{\frac{H}{4} \times \frac{W}{4} \times 2}$ is also a two class score feature map.

As mentioned in the above section, we supervise the mask prior obtained at each layer in the mask prior generator and calculate the average loss value. The output Y and mask prior m_p are both supervised by the cross-entropy loss. The total loss is defined as:

$$L = \lambda \frac{1}{N_l} \sum L_{prior} + L_{mask}, \tag{17}$$

where N_l is the number of layers in the mask prior generator. λ is a hyper-parameter for weighting the mask prior loss L_{prior}. L_{mask} denotes the final segmentation loss function based on Y. When performing inference, we first reshape Y back to the initial image size and apply the argmax operation along the channel dimension to obtain the final mask prediction.

4. Experiments

4.1. Settings

Datasets: We train and evaluate our MPG-LQGNet on three widely-used datasets: RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016) and G-Ref (Nagaraja et al., 2016). The images of the three datasets are all based on the MS COCO dataset (Lin et al., 2014). RefCOCO consists of 142,209 referring expressions for 50,000 targets in 19,994 images, and RefCOCO+ contains 141,564 referring expressions for 49,856 targets in 19,992 images. Meanwhile, G-Ref has 104,560 referring expressions for 54,822 targets in 26,711 images. The language expressions in RefCOCO and RefCOCO+ are more concise than the languages used in the G-Ref, which contain 3.5 words on average. Conversely, expressions in G-Ref are more complex, with an average of 8.4 words, which makes the dataset the most challenging. Significantly, the expressions in RefCOCO+ hardly contain the descriptions of the location information of the targets, such as “on the left”, etc, which also makes the dataset more challenging.

Metrics: We evaluate the performance by three commonly used metrics: the overall intersection-over-union (oIoU), the mean intersection-over-union (mIoU), and precision@X (P@X). The oIoU is measured as the ratio between the total intersection area and the total union area of all test samples. Following previous works (Yang et al., 2022; Xu et al., 2023), we use oIoU as the default metric to compare our method with other works. The mIoU is the average of the IoU over all the test samples. The precision@X score computes the percentage of predictions that have IoU scores higher than X, where $X \in \{0.5, 0.7, 0.9\}$.

Implementation Details: We utilize the official pre-trained Swin-Transformer-Base (Liu et al., 2021) and BERT-Base (Kenton and Toutanova, 2019) models as the visual encoder and language encoder, respectively, for a fair comparison. The BERT model has 12 layers with a hidden size of 768. Our visual encoder layers are initialized with the classification weights pre-trained on ImageNet-22K (Deng et al., 2009). The channel number C_q is set to 256, and the number of layers N_l is 6 in our proposed mask prior generator. The head number is 8 for all layers. We adopt the AdamW (Kingma and Ba, 2014) optimizer with a weight decay of 0.01 and set the learning rate to 0.000015

Table 1
Comparison with state-of-the-art methods in terms of overall IoU on three benchmark datasets. U: The UMD partition.

Methods	Visual backbone	Language backbone	RefCOCO			RefCOCO+			G-Ref	
			val	testA	testB	val	testA	testB	val(U)	test(U)
RRN (Li et al., 2018)	DeepLab-R101	LSTM	55.33	57.26	53.93	39.75	42.15	36.11	–	–
CMSA (Ye et al., 2019)	DeepLab-R101	–	58.32	60.61	55.09	43.76	47.60	37.89	–	–
QRN (Shi et al., 2020)	DeepLab-R101	LSTM	59.75	60.96	58.77	48.23	52.65	40.89	–	–
BRINet (Hu et al., 2020)	DeepLab-R101	LSTM	60.98	62.99	59.21	48.17	52.32	42.11	–	–
MCN (Luo et al., 2020)	Darknet53	GRU	62.44	64.20	59.71	50.62	54.99	44.69	49.22	49.40
CMPC (Huang et al., 2020)	DeepLab-R101	LSTM	61.36	64.53	59.64	49.56	53.44	43.23	–	–
SANet (Lin et al., 2021)	DeepLab-R101	LSTM	61.84	64.95	57.43	50.38	55.36	42.74	–	–
EFN (Feng et al., 2021)	ResNet101	GRU	62.76	65.69	59.67	51.50	55.24	43.01	–	–
BUSNet (Yang et al., 2021)	DeepLab-R101	Self-Attention	63.27	66.41	61.39	51.76	56.87	44.13	–	–
LTS (Jing et al., 2021)	Darknet53	GRU	65.43	67.76	63.08	54.21	58.32	48.02	54.40	54.25
VLT (Ding et al., 2021)	Darknet53	GRU	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65
ReSTR (Kim et al., 2022)	ViT-B	Transformer	67.22	69.30	64.45	55.78	60.44	48.27	54.48	–
CRIS (Wang et al., 2022)	CLIP-R101	CLIP	70.47	73.18	66.10	62.27	68.08	53.68	59.87	60.36
LAVT (Yang et al., 2022)	Swin-B	BERT	72.73	75.82	68.79	62.14	68.38	55.10	61.24	62.09
CoupAlign (Zhang et al., 2022)	Swin-B	BERT	74.70	77.76	70.58	62.92	68.34	56.69	62.84	62.22
BKINet (Ding et al., 2023)	CLIP-R101	CLIP	73.22	76.43	69.42	64.91	69.88	53.39	64.21	63.77
SADLR (Yang et al., 2023)	Swin-B	BERT	74.24	76.25	70.06	64.28	69.09	55.19	63.60	63.56
CGFormer (Tang et al., 2023)	Swin-B	BERT	74.75	77.30	70.64	64.54	71.00	57.14	**64.68**	65.09
MCRES (Xu et al., 2023)	Swin-B	BERT	74.92	76.98	70.84	64.32	69.68	56.64	63.51	64.90
MPG-LQGNet	Swin-B	BERT	**75.52**	**78.25**	**71.85**	**66.44**	**73.18**	**57.61**	64.49	**65.16**

with a polynomial learning rate decay schedule. Our MPG-LQGNet is trained for 35 epochs with batch size 8. The images are resized to 448 × 448 without specific data augmentation. The maximum length of the referring expression is set to 30, and the weight of the mask prior loss λ is 0.1.

4.2. Comparison with state-of-the-art methods

In Table 1, we compare our MPG-LQGNet with some advanced approaches on three datasets, RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016) and G-Ref (UMD partition) (Nagaraja et al., 2016). On the RefCOCO and RefCOCO+ datasets, our proposed MPG-LQGNet outperforms previous methods on all evaluation datasets. Specifically, compared with the recent methods (Tang et al., 2023; Xu et al., 2023), our method achieves better performance with absolute oIoU increases of 0.6%–1.2% on the three subsets of RefCOCO. Especially for RefCOCO+, which is more challenging without the location words in referring expressions, and our MPG-LQGNet also outperforms the second-best method (Tang et al., 2023) by 1.9%, 2.18%, and 0.47% on the validation, testA, and testB sets of RefCOCO+, respectively. This indicates that our MPG-LQGNet can better locate the targets based on the semantic information of the text. On the most challenging G-Ref dataset, our method is competitive to or better than previous state-of-the-art methods, which achieves suboptimal performance on the validation set and surpasses all the compared methods on the test set of G-Ref.

Table 2
Ablation studies of our proposed key components. L_{prior} is the mask prior loss.

No.	Model	oIoU	mIoU	P@0.5	P@0.7	P@0.9
1	Baseline (tile-and-concat)	66.71	63.95	73.82	61.92	20.51
2	Baseline + CAFG	67.03	65.36	75.34	63.29	21.57
3	Baseline (BSAM)	68.77	69.27	79.29	70.18	**24.20**
4	Baseline (BSAM) + CAFG	71.34	70.00	80.81	70.53	23.29
5	Baseline (BSAM) + LQPG	71.63	69.68	80.86	69.01	22.43
6	Ours w/o L_{prior}	72.22	69.89	81.22	69.82	22.38
	Ours	**73.18**	**71.87**	**83.49**	**71.75**	23.29

4.3. Ablation study

We also conducted several ablation studies to evaluate the effectiveness of each designed module in our MPG-LQGNet on the testA split of the RefCOCO+, which is a more challenging dataset with complex scenarios.

Effectiveness analysis of our proposed key components:
Table 2 presents the experimental results of removing different modules based on our MPG-LQGNet framework. In the baseline model, both the Channel Attention Fusion Gate (CAFG) and Language Query Guided Mask Prior Generator (LQPG) are removed, and we use the sentence feature vector that is globally pooled from all words and fused equally to each position of the visual feature map using the "tile-and-concatenate" operation as a replacement for the Bidirectional Spatial Alignment Module (BSAM). Compared to the baseline model, the inclusion of the CAFG (model No. 2) and the use of the BSAM (model

Table 3
Ablation studies of the language queries from different encoding stages.

Language queries selection	oIoU	mIoU	P@0.5	P@0.7	P@0.9
stage1 + stage2	72.78	70.98	82.33	70.78	22.58
stage3 + stage4	69.93	68.58	79.54	67.65	20.66
stage1	71.88	69.78	81.42	68.51	22.33
stage2	70.66	68.99	80.81	69.52	21.62
stage3	70.83	69.47	80.96	68.96	21.37
stage4	71.46	69.06	80.10	69.52	22.23
Ours	**73.18**	**71.87**	**83.49**	**71.75**	**23.29**

No. 3) result in improvements of 0.32% and 2.06% in oIoU, and 1.41% and 5.32% in mIoU, respectively. Additionally, there is a significant improvement in precision at all three threshold settings. In model No. 4, which includes both the BSAM and the CAFG, the oIoU is further increased by 2.57% and 4.31%, respectively, compared to adding each module separately. This demonstrates that fine-grained fusion in the spatial dimension and attention to important channel information in the global dimension effectively enhance the model's performance. Note that in the model No. 3 and No. 4, the BSAM is only used to obtain the cross-modal features F_i without generating language queries. In the model No. 5, the inclusion of LQPG after the model No. 3 leads to improvements of 2.86% and 0.41% in oIoU and mIoU, respectively. Finally, when all three proposed modules are used together, there is a further improvement in the final performance. Specifically, compared to the model No. 4 without the LQPG, the full model increases precision at a threshold of 0.5 by 2.68%. This indicates that the proposed LQPG significantly improves the model's location ability and enables a more comprehensive understanding of cross-modal information, as samples with an IoU of less than 0.5 can be considered as results of location failures. From the last two rows, it can be seen that if we do not use L_{prior} to supervise the generation of the mask prior (model No. 6), there will be a decrease of 0.96% and 1.98% in oIoU and mIoU, respectively. Therefore, adding a supervisory signal in LQPG can enhance the ability of the mask prior in target localization and reduce the activation of redundant information. Furthermore, we observe a slight decrease in the effectiveness of our method in generating high-quality masks (23.29% vs. 24.20% in precision@0.9).

Effectiveness analysis of the language query:

In this section of the ablation experiment, we compared the impact of using language queries from different encoding stages on the model's final predictions. As shown in Table 3, we initially only selected language queries generated from the first two or last two stages, which resulted in a decrease of 0.40% and 3.25% in oIoU, and 0.89% and 3.29% in mIoU, respectively. The fact that using queries from only the first two stages resulted in a smaller decrease suggests the effectiveness of early bi-directional spatial alignment strategy. Additionally, we conduct ablation studies where we only use a single language query from different encoding stages for the cross-modal interaction in the decoding stage. We find that using only the query from stage 1 achieves better results in terms of IoU and precision@0.5 compared to queries from other stages. This also indicates the importance of generating language queries with multi-modal information at the low-level stage, as it provides accurate localization information for the targets, while queries from later stages tend to provide semantic information. By using this comprehensive querying approach, we can generate more accurate mask predictions.

Effectiveness analysis of the Channel Attention Fusion Gate (CAFG): In this section, we compare the proposed CAFG module with various variants of the Attention Fusion Gate (AFG) networks. In our work, after obtaining the cross-modal fused features at each stage of the encoder, we design a gate network based on channel attention to adaptively select different channel information, to represent our target objects accurately. The CAFG we designed is a dual-path approach that generates two sets of channel weights to weight the cross-modal features. We analyze different variants of CAFG, including single-path CAFG, spatial-channel AFG (where one path does not undergo average pooling to generate spatial weights), and spatial-spatial AFG (where both paths generate spatial weights), as shown in the first three rows of Table 4. Compared to our designed CAFG, these variants of attention fusion gate show a decrease of 0.40%, 2.81%, and 1.01% in oIoU, and 1.01%, 2.32% and 1.21% in mIoU, respectively. Additionally, there is also a varying degree of decrease in precision at three different thresholds. These all indicate the effectiveness of our CAFG module, which only weights cross-modal features through the channel dimension from a global perspective. We argue that performing scaling with different weights for different spatial positions on the feature map, after aligning in the spatial dimension, may disrupt the fine-grained cross-modal information. Furthermore, we also replace the CAFG in our model with the Language Gate (LG) proposed in the previous method (Yang et al., 2022) as mentioned in Section 3.2.3, and compare their effectiveness. The LG is a two-layer perceptron with ReLU activation after the first layer and hyperbolic tangent activation after the second layer to learn a set of spatial weight maps. From the last two rows of Table 4, it can be observed that our proposed CAFG demonstrates better performance, with improvements of 1.04% and 2.19% in terms of oIoU and mIoU, respectively. This also confirms the effectiveness of using channel attention to control the fused feature information.

Table 4
Ablation studies of the different variants of CAFG.

Variants of Attention Fusion Gate (AFG) Network	oIoU	mIoU	P@0.5	P@0.7	P@0.9
single-path CAFG	72.78	70.86	81.82	70.89	23.24
spatial-channel AFG	70.37	69.55	80.76	69.32	22.28
spatial-spatial AFG	72.17	70.66	81.72	70.48	22.43
Language Gate	72.14	69.68	80.96	68.15	23.00
Ours	**73.18**	**71.87**	**83.49**	**71.75**	**23.29**

Table 5
Ablation studies of the number of layers in the Language Query Guided Mask-Prior Generator.

The number of layers	oIoU	mIoU	P@0.5	P@0.7	P@0.9
1	69.75	68.44	79.59	66.99	20.20
2	72.11	70.49	82.03	70.38	22.94
3	72.39	70.67	82.28	70.58	22.84
4	72.46	71.21	82.99	71.30	22.73
5	72.87	71.16	83.34	70.53	**23.85**
6(ours)	**73.18**	**71.87**	**83.49**	**71.75**	23.29
7	72.35	70.38	81.72	70.58	23.20
8	72.19	70.21	82.33	69.77	22.84

Effectiveness analysis of the number of layers in the Language Query Guided Mask Prior Generator (LQPG): We study the impact of the number of layers in our proposed LQPG by varying the number of layers from 1 to 8. Each layer is built on the Multi-Head Cross-Attention (MHCA) layer, followed by a Feed-Forward network. According to the results presented in Table 5, there is an overall upward trend in both IoU and precision metrics when the number of layers increases. Specifically, using two layers instead of one has shown a significant improvement, with a 2.36% increase in oIoU and a 2.05% increase in mIoU. This suggests that deep supervision effectively provides early supervision to the LQPG module and retains only the most relevant information for segmenting the target. As a result, the subsequent layer can gradually refine and generate a more accurate mask prior after passing through the first layer. In addition, we observe a decline in model performance when the number of layers exceeds 6. Therefore, we set the default number of layers for other experiments to 6.

Effectiveness analysis of the coordinate convolution for final prediction: As mentioned in Section 3.3.3, we concatenate a 2-D spatial coordinate map for the final mask prediction. Compared to

Fig. 5. Qualitative results of our MPG-LQGNet from different datasets.

traditional convolution, the coordinate convolution involves appending two additional channels to the input feature map, one representing the x-coordinate and the other representing the y-coordinate, to capture spatial information. Table 6 demonstrates that without the use of coordinate convolution, there is a decrease in performance of 0.53% and 1.19% in terms of oIoU and mIoU, respectively. This indicates that when combining different input feature maps from two branches, the utilization of coordinate convolution allows for better spatial localization of the target region.

Model efficiency analysis:

Table 7 presents the number of parameters and computational complexity of our model and a comparison with previous state-of-the-art methods. Compared to the LAVT (Yang et al., 2022), which employs the same backbone, our approach demonstrates a 5.93% and 7.09% increase in params and FLOPs respectively. However, in terms of segmentation accuracy measured by the oIoU metric, our model yields a remarkable enhancement of 4.8%. It is worth noting that relative to BKINet (Ding et al., 2023), our method indeed incurs a higher computational expense, which will be a focus for future optimization efforts. The necessity to generate higher-quality mask prior contributes to this additional overhead. Nevertheless, our primary contribution lies in the substantial improvement of segmentation quality.

Table 6
Ablation study of the coordinate convolution.

Methods for final prediction	oIoU	mIoU	P@0.5	P@0.7	P@0.9
Ours w/o CoordConv	72.65	70.68	82.33	70.33	22.94
Ours	**73.18**	**71.87**	**83.49**	**71.75**	**23.29**

4.4. Visualization

Qualitative examples from different datasets: In this section, we select two sets of cases from the three evaluation datasets, as illustrated

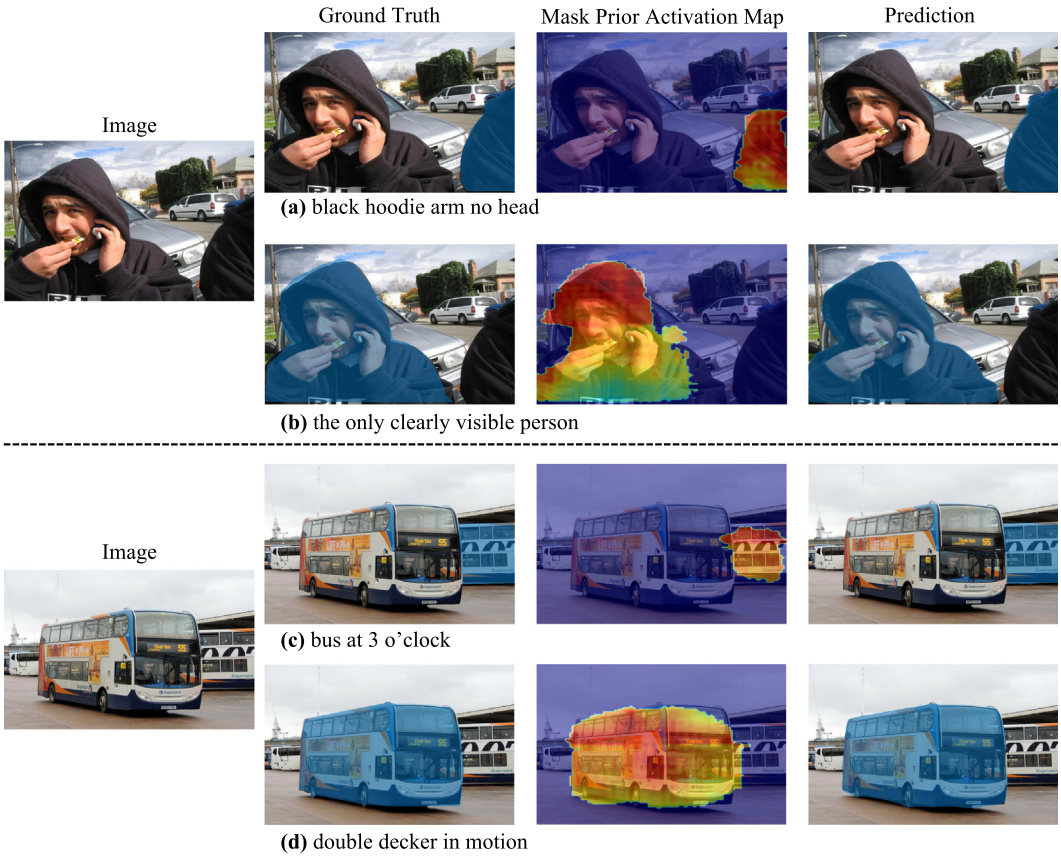


Fig. 6. More qualitative results showing our MPG-LQGNet segmenting different targets in an image.

in Fig. 5. As mentioned earlier, the text expressions in the RefCOCO dataset are simpler and often include some directional words for the object, such as "by the red wall" and "on left side" in examples (a) and (b). Our model is able to accurately understand these auxiliary location phrases and make correct segmentation. In particular, in both of these examples, our approach is capable of distinguishing between targets and irrelevant objects from areas that are relatively far from the camera. As for the RefCOCO+ dataset, its referring text hardly contains similar location expressions, as shown in examples (c) and (d). In the images with multiple people, our model has the capability to locate these targets based on their other attributes, such as color, clothing style, and more. For example, it can locate the person described as "striped yellow and blue" or "black suit". In the G-Ref dataset, the text is longer and has more complex grammatical structures. Example (e) shows the relationship between two entities, "woman" and "tennis ball". Example (f) demonstrates that our model is able to accurately segment smaller objects in the image by understanding the context of certain keywords, such as "small white bowl" and "strawberries". Additionally, we show the activation map of the generated mask prior in the third column, which is a coarse-grained segmentation map and helps provide accurate localization information for the model during final prediction. It can be seen that the mask prior obtained for each case is able to locate the target accurately.

In Fig. 6, we have listed two additional cases (from the RefCOCO+ dataset) to visualize the segmentation results of our model for different targets in the same image. In the first set of examples, there are two people, but we can only see a partial view of one person while the other person's head is clearly visible. Our model demonstrates a thorough understanding of the negation term "no head" and the descriptive phrase "clearly visible" to differentiate between the two targets. In the second set of examples, there are also two entities of the same category. The referring phrase "at 3 o'clock" is a more natural language expression for humans, while "in motion" is used to distinguish the different characteristics of the two buses. Our method successfully identifies the correct targets, further validating the superior understanding of our model in handling flexible natural language.

Table 7
Ablation studies of the model efficiency on the RefCOCO+ testA.

Model	Visual backbone	Language backbone	Params	FLOPs	oIoU
CRIS	CLIP-R101	CLIP	163.48M	78.54G	68.08
LAVT	Swin-B	BERT	118.71M	191.44G	68.38
BKINet	CLIP-R101	CLIP	156.49M	83.14G	69.88
Ours	Swin-B	BERT	125.75M	205.01G	73.18

Qualitative comparison with the baseline model: Fig. 7 shows the visualization results of our model and the baseline model for referring segmentation for some examples. As mentioned in the previous Section 4.3, in the baseline model, we replace BSAM with the "tile-and-concatenate" method for cross-modal fusion and remove our proposed CAFG and LQPG. In example (a), the baseline model is able to understand the word "edge" and identify the object at the edge, but it mistakenly segments out a part of another person's head. Actually, there should be two people at the image edge, and our goal is only to segment the skier's arm. From the last two columns, it can be seen that our model accurately understands the meaning of "closest to us" and distinguishes the desired target among multiple objects. At the same time, the activation map of the mask prior does not activate irrelevant regions, which prevents our model from segmenting incorrect parts. Example (b) shows that our model is able to differentiate the target among three people using the keyword "skateboard", while the baseline model does not understand the specific meaning of the text and produces an ambiguous prediction.

Fig. 7. Qualitative comparison with the baseline model.

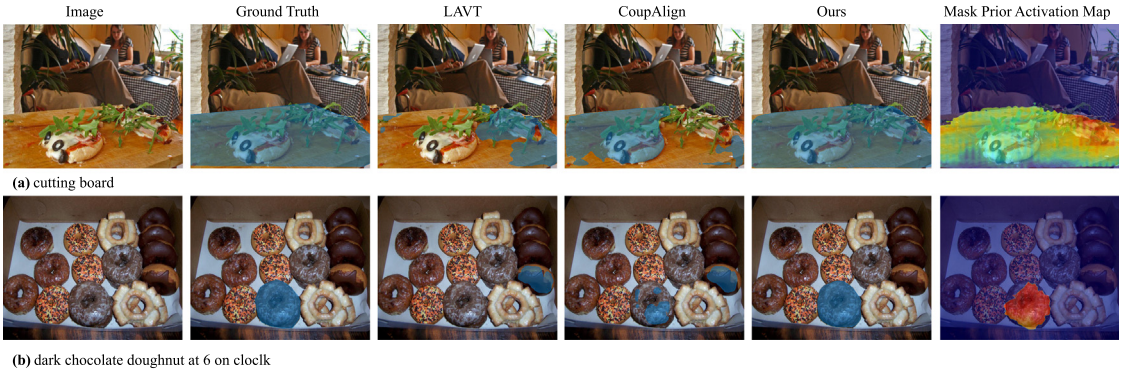


Fig. 8. The qualitative comparisons between LAVT (Yang et al., 2022), CoupAlign (Zhang et al., 2022) and ours.

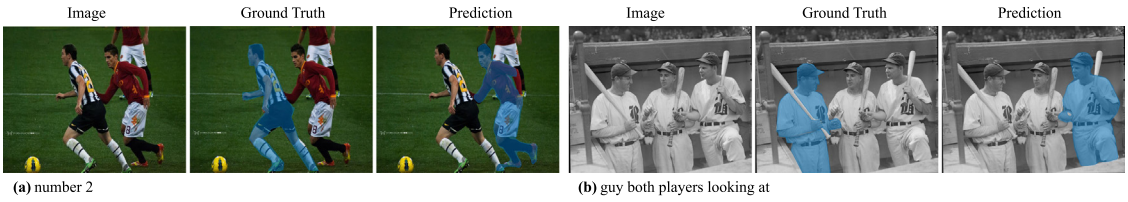


Fig. 9. Visualization of typical failure cases of our method.

Qualitative comparison with some SOTA methods: Fig. 8 presents a visualization comparison between our method and several previous sota methods. We specifically select LAVT (Yang et al., 2022) and CoupAlign (Zhang et al., 2022) for the segmentation performance comparison because both follow a foundational encoder–decoder framework and employ an encoder-fusion strategy, thus making them representative. In example (a), the task is to segment the entire "cutting board," but neither LAVT nor CoupAlign successfully captures the complete target. This suggests that they fail to accurately grasp the detailed meaning of the image and do not effectively implement cross-modal alignment. In the second case, we aim to locate the object at the 6 o'clock position; however, the earlier methods all segment irrelevant targets, indicating insufficient comprehension of high-level textual semantics and leading to incorrect predictions. In contrast, guided by the mask prior, our model precisely localizes the target object, yielding more comprehensive and accurate segmentation results.

Failure Cases: We analyze two typical failure cases, as depicted in Fig. 9. In the first case, we expect to identify the athlete wearing jersey number 2 based on the referring expression "number 2", but our model produces a completely incorrect prediction. We argue that the primary reason for the failure is the model's insufficient representation of textual information. Additionally, the text "number 2" is ambiguous and lacks specificity, as it does not clearly refer to a particular entity. In example (b), there are three individuals, with the two on the right both looking at the one on the far left. Our goal is to segment the target on the far left. But our method demonstrates inadequate high-level understanding of the image, which is reflected in both the semantic representation of the image itself and the modality alignment between the image and the text. Overall, these two failure cases discuss the limitations of our approach in effectively capturing high-level semantic representations of both text and images. We attribute these issues partly to the distribution of the training datasets, which may enable better cross-modal understanding for certain image-text pairs while failing to adequately capture others. Overall, our model achieves a balanced segmentation performance.

5. Conclusion

In this paper, we propose a novel framework of MPG-LQGNet, which shows its high capability of effective interaction and comprehensive understanding of semantic embeddings of vision and language modalities. Specifically, in the encoding stage of our MPG-LQGNet, the designed

Bidirectional Spatial Alignment Module (BSAM) and the Channel Attention Fusion Gate (CAFG) are used to bidirectional fuse vision and language modality features from a spatial fine-grained and global channel perspective, respectively, meanwhile, obtaining language queries from different stages of the encoder to guide the mask prior generation in the decoding stage. In the decoding stage, the proposed Language Query Guided Mask Prior Generator (LQPG) can generate the mask prior based on the activation of the language queries. The mask prior can provide a constraining activation region as a rough localization of the target to more effectively guide the final segmentation. Extensive experiments demonstrate the effectiveness of each proposed module in our MPG-LQGNet; meanwhile, compared to state-of-the-art RIS methods, our MPG-LQGNet shows its high capability to accurately locate visual targets and comprehend the semantic information of the language, achieving state-of-the-art performance on referring image segmentation task. However, our work also has limitations in segmenting smaller targets or obtaining high-quality masks. Future work will focus on more complex real-life scenarios, improving the model's ability to understand longer texts and enhancing the quality of mask generation.

CRediT authorship contribution statement

Jinhao Zhou: Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Guoqiang Xiao:** Writing – review & editing. **Michael S. Lew:** Writing – review & editing. **Song Wu:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Sichuan Science and Technology Program (2025ZNSFSC0482).

Data availability

Data will be made available on request.

References

Ahn, H., Choi, S., Kim, N., Cha, G., Oh, S., 2018. Interactive text2pickup networks for natural language-based human–robot collaboration. IEEE Robot. Autom. Lett. 3 (4), 3308–3315.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: European Conference on Computer Vision. Springer, pp. 213–229.

Chen, J., Shen, Y., Gao, J., Liu, J., Liu, X., 2018. Language-based image editing with recurrent attentive models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8721–8729.

Codevilla, F., Santana, E., López, A.M., Gaidon, A., 2019. Exploring the limitations of behavior cloning for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9329–9338.

Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., Barnard, K., 2021. Attentional feature fusion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3560–3569.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Ieee, pp. 248–255.

Ding, H., Liu, C., Wang, S., Jiang, X., 2021. Vision-language transformer and query generation for referring segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16321–16330.

Ding, H., Zhang, S., Wu, Q., Yu, S., Hu, J., Cao, L., Ji, R., 2023. Bilateral knowledge interaction network for referring image segmentation. IEEE Trans. Multimed..

Feng, G., Hu, Z., Zhang, L., Lu, H., 2021. Encoder fusion network with co-attention embedding for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15506–15515.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2961–2969.

Hu, Z., Feng, G., Sun, J., Zhang, L., Lu, H., 2020. Bi-directional relationship inferring network for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4424–4433.

Hu, R., Rohrbach, M., Darrell, T., 2016. Segmentation from natural language expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer, pp. 108–124.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Huang, S., Hui, T., Liu, S., Li, G., Wei, Y., Han, J., Liu, L., Li, B., 2020. Referring image segmentation via cross-modal progressive comprehension. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10488–10497.

Jain, K., Gandhi, V., 2021. Comprehensive multi-modal interactions for referring image segmentation. arXiv preprint arXiv:2104.10412.

Jiang, Y., Yan, T., Yao, M., Wang, H., Liu, W., 2024. Cascade transformers with dynamic attention for video question answering. Comput. Vis. Image Underst. 242, 103983.

Jing, Y., Kong, T., Wang, W., Wang, L., Li, L., Tan, T., 2021. Locate then segment: a strong pipeline for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9858–9867.

Kenton, J.D.M.-W.C., Toutanova, L.K., 2019. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, Vol. 1. p. 2.

Kim, N., Kim, D., Lan, C., Zeng, W., Kwak, S., 2022. Restr: convolution-free referring image segmentation using transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18145–18154.

Kim, W., Son, B., Kim, I., 2021. Vilt: vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. PMLR, pp. 5583–5594.

Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Li, R., Li, K., Kuo, Y.-C., Shu, M., Qi, X., Shen, X., Jia, J., 2018. Referring image segmentation via recurrent refinement networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5745–5753.

Li, Z., Wang, M., Mei, J., Liu, Y., 2021. Mail: a unified mask-image-language trimodal network for referring image segmentation. arXiv preprint arXiv:2111.10747.

Li, Z., Wang, W., Xie, E., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., Lu, T., 2022. Panoptic segformer: delving deeper into panoptic segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1280–1289.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, pp. 740–755.

Lin, L., Yan, P., Xu, X., Yang, S., Zeng, K., Li, G., 2021. Structured attention network for referring image segmentation. IEEE Trans. Multimed. 24, 1922–1932.

Ling, H., Kreis, K., Li, D., Kim, S.W., Torralba, A., Fidler, S., 2021. Editgan: high-precision semantic image editing. Adv. Neural Inf. Process. Syst. 34, 16331–16345.

Liu, C., Ding, H., Zhang, Y., Jiang, X., 2023. Multi-modal mutual attention and iterative interaction for referring image segmentation. IEEE Trans. Image Process..

Liu, R., Lehman, J., Molino, P., Petroski Such, F., Frank, E., Sergeev, A., Yosinski, J., 2018. An intriguing failing of convolutional neural networks and the coordconv solution. Adv. Neural Inf. Process. Syst. 31.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, S., Xiao, G., Lew, M.S., Gao, X., Wu, S., 2024. Core-attributes enhanced generative adversarial networks for robust image enhancement. Eng. Appl. Artif. Intell. 131, 107799.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

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

Mashrur, A., Luo, W., Zaidi, N.A., Robles-Kelly, A., 2024. Robust visual question answering via semantic cross modal augmentation. Computer Vision and Image Understanding 238, 103862.

Nagaraja, V.K., Morariu, V.I., Davis, L.S., 2016. Modeling context between objects for referring expression understanding. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. Springer, pp. 792–807.

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning. ICML-10, pp. 807–814.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.

Shi, H., Li, H., Wu, Q., Ngan, K.N., 2020. Query reconstruction network for referring expression image segmentation. IEEE Trans. Multimed. 23, 995–1007.

Strudel, R., Garcia, R., Laptev, I., Schmid, C., 2021. Segmenter: transformer for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7262–7272.

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

Toromanoff, M., Wirbel, E., Moutarde, F., 2020. End-to-end model-free reinforcement learning for urban driving using implicit affordances. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7153–7162.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T., 2022. Cris: clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11686–11695.

Wu, S., Shan, S., Xiao, G., Lew, M.S., Gao, X., 2024a. Modality blur and batch alignment learning for twin noisy labels-based visible–infrared person re-identification. Eng. Appl. Artif. Intell. 133, 107990.

Wu, S., Yuan, X., Xiao, G., Lew, M.S., Gao, X., 2024b. Deep cross-modal hashing with multi-task latent space learning. Eng. Appl. Artif. Intell. 136, 108944.

Xu, L., Huang, M.H., Shang, X., Yuan, Z., Sun, Y., Liu, J., 2023. Meta compositional referring expression segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19478–19487.

Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H., 2022. Lavt: language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18155–18165.

Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H., 2023. Semantics-aware dynamic localization and refinement for referring image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, No. 3. pp. 3222–3230.

Yang, S., Xia, M., Li, G., Zhou, H.-Y., Yu, Y., 2021. Bottom-up shift and reasoning for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11266–11275.

Ye, L., Rochan, M., Liu, Z., Wang, Y., 2019. Cross-modal self-attention network for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10502–10511.

Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L., 2016. Modeling context in referring expressions. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer, pp. 69–85.

Zhang, Z., Zhu, Y., Liu, J., Liang, X., Ke, W., 2022. Coupalign: coupling word-pixel with sentence-mask alignments for referring image segmentation. Adv. Neural Inf. Process. Syst. 35, 14729–14742.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al., 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6881–6890.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J., 2020. Deformable detr: deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159.