EHIN: Early-aware Hierarchical Interaction Network for Weakly-supervised Referring Image Segmentation

Hongjun Li[a], Nan Wang[a,*], Anqing Chen[b], Jiang Liu[c], Wanli Ma[d], Weide Liu[e], Yakun Ju[f], Paul L. Rosin[c], Hantao Liu[c], Wei Zhou[c,*]

[a]*College of Computer Science and Cyber Security, Chengdu University of Technology, China*
[b]*State Key Laboratory of Oil and Gas Reservoir Geology and Exploitation, Chengdu University of Technology, China*
[c]*School of Computer Science and Informatics, Cardiff University, UK*
[d]*Department of Engineering, University of Cambridge, UK*
[e]*Harvard Medical School, Harvard University, USA*
[f]*School of Computing and Mathematical Sciences, University of Leicester, UK*

Abstract

Referring image segmentation (RIS) aims to segment target regions in images based on natural language descriptions. Although weakly-supervised RIS frameworks have been proposed to reduce reliance on costly manual annotations, their performance remains limited due to both the low quality of pseudo-labels and the inherent difficulty in achieving effective interaction of visual and textual features. In this paper, we propose a novel weakly-supervised framework named Early-aware Hierarchical Interaction Network (EHIN). The proposed network includes two key components, which are designed to enhance pseudo-labels generation and improve the interaction of visual and textual features for RIS, respectively. First, EHIN incorporates an Early-aware Contrastive Learning Module (ECLM) that enhances feature discrimination by leveraging contrastive learning to distinguish target features from background noise. By integrating the module early into the processing pipeline, ECLM operates on raw image features directly, preserving richer visual details while reducing reliance on labeled data and thus improving the reliability of pseudo-labels. Second, EHIN integrates a Hierarchical Interaction Prompt Module (HIPM) to facilitate comprehensive interaction between visual and textual features and enhance subsequent feature fusion. Extensive experimental results on four benchmark datasets demonstrate that the proposed EHIN outperforms the state-of-the-art RIS. Code is available at https://github.com/CDUT-DBGroup/MFP-TRIS.

Keywords: Referring Image Segmentation, Weakly-supervised, Early-aware Contrastive Learning, Hierarchical Interaction Prompt, Multimodality

1. Introduction

Referring image segmentation (RIS) is a multimodal task that integrates vision and language to segment specific objects in an image based on the input language expression. It has attracted considerable interest among researchers in computer vision and natural language processing, serving as a vital link between human language, machine understanding, and real-world contexts [1]. Therefore, RIS has extensive practical applications, such as autonomous driving, intelligent robotics, medical image analysis, remote sensing, and human-computer interaction, highlighting its broad relevance and real-world impact [2, 3]. RIS distinguishes itself from traditional image segmentation, which generally targets predefined, fixed categories [4, 5], by addressing the segmentation of objects or scenes described in natural language [6]. Such descriptions often convey intricate details about objects, actions, attributes, and spatial relationships within the image using human syntax and vocabulary. The core challenge lies in accurately mapping textual descriptions to corresponding visual elements in the image [7]. This complex task demands a model capable of not only interpreting image content but also comprehending and processing the nuanced semantics and contextual information embedded in the text [8]. Recent research in RIS has grown significantly, with a prevalent approach involving separate

*Corresponding authors
Email addresses: 2023020919@stu.cdut.edu.cn (Nan Wang), zhouw26@cardiff.ac.uk (Wei Zhou)

feature extraction from images and text using independent encoders [9], followed by integration through a cross-modal decoder for final prediction [10, 11]. This strategy harnesses the complementary strengths of visual and linguistic information to enhance the model's accuracy in locating and segmenting target objects. Notably, in many existing approaches, cross-modal interactions are primarily confined to post-encoding stages, where the decoder alone handles the alignment between visual and textual features [12, 13, 14]. However, this methodology still fails to fully leverage the potential for effective interaction of visual and textual information.

Moreover, although weakly-supervised RIS frameworks have been proposed to reduce reliance on costly manual annotations, their performance often suffers when pseudo-labels are noisy or imprecise, especially during early training stages. In such cases, suboptimal supervision can hinder the model's ability to fully exploit rich cross-modal semantic correlations. Therefore, we propose a novel weakly-supervised framework, the Early-aware Hierarchical Interaction Network (EHIN), which comprises two key components. First, we introduce an Early-aware Contrastive Learning Module (ECLM), which incorporates a contrastive learning strategy to reduce reliance on labeled data while improving the quality of pseudo-labels. Specifically, ECLM leverages contrastive learning by aligning image and text embeddings at an early stage using CLIP-ViT. Second, we design a Hierarchical Interaction Prompt Module (HIPM), which enables a more comprehensive interaction between visual and textual features. As illustrated in Fig. 1, our approach effectively identifies target objects in the image and generates precise segmentation maps. The improved accuracy stems from the Early-aware Contrastive Learning Module (ECLM), which enhances feature discrimination by mitigating background noise interference, thereby refining the quality of pseudo-labels. Additionally, the Hierarchical Interaction Prompt Module (HIPM) facilitates effective knowledge transfer from pre-trained models, enabling a more comprehensive interaction between visual and textual features, which further contributes to the precision of the generated segmentation maps.

To evaluate the effectiveness of the proposed method, we conducted comprehensive experiments on several widely used RIS datasets. Our EHIN exhibited superior performance across four key datasets: RefCOCO, RefCOCO+, RefCOCOg, and ReferIt.

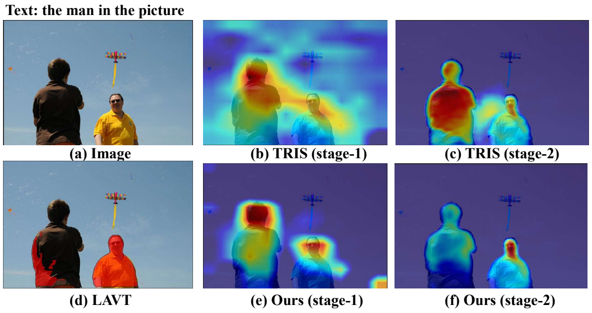
The key contributions of this work are summarized as follows:

- In weakly-supervised referring image segmentation (RIS), the generation of high-quality pseudo-labels during the initial stage is critical for guiding subsequent learning. To address this challenge and enhance the extraction and use of fine-grained visual semantics through early-stage alignment of visual and textual modalities, we propose a novel weakly-supervised Early-aware Hierarchical Interaction Network (EHIN) for RIS. Our methodology emphasizes the critical refinement of pseudo-labels generation quality through hierarchical feature interaction and early-stage semantic awareness.
- We propose an Early-aware Contrastive Learning Module (ECLM) that enhances feature discrimination and representation learning by leveraging contrastive learning to distinguish target features from background noise, thereby reducing dependence on labeled data and improving the reliability of pseudo-labels. Unlike traditional contrastive learning methods that operate on high-level or projected features, our ECLM performs feature fusion directly from the original text and image representations. This early-stage cross-modal alignment strategy improves the quality of pseudo-labels and leads to more effective model learning by preserving richer visual details after CLIP [15] feature extraction.
- We propose a Hierarchical Interaction Prompt Module (HIPM) that enables a more comprehensive interaction of visual-textual features and enhances subsequent feature fusion. The HIPM introduces a hierarchical fusion mechanism that integrates visual and textual features at multiple levels, and incorporates repeated residual connections throughout the fusion process, resulting in more robust and comprehensive multimodal representations.
- We comprehensively evaluate the Early-aware Hierarchical Interaction Network (EHIN), demonstrating its effectiveness on four public referring image segmentation datasets.

The remainder of the paper is organized as follows. Section 2 reviews the most relevant related works. Section 3 describes the proposed method. Sections 4 and 5 delve into experiments and results. Section 6 discusses the limitations of our method and provides further analysis. Finally, Section 7 concludes the paper.

2. Related Work

Weakly-Supervised Referring Image Segmentation While fully supervised RIS methods have



(a) Image (b) TRIS (stage-1) (c) TRIS (stage-2)

(d) LAVT (e) Ours (stage-1) (f) Ours (stage-2)

Figure 1: Given an input image and text query, our method generates high-quality pseudo-labels for target localization in Stage-1 and accurately segments the target object in Stage-2. Our approach achieves higher segmentation accuracy in Stage-1, utilizing only the text query as the supervisory signal. (a) represents the original image, while (e) and (f) illustrate the results of our method at Stage-1 and Stage-2, respectively. (b) and (c) correspond to the segmentation results of TRIS [16] at Stage-1 and Stage-2, respectively, and (d) depicts the segmentation outcome of the LAVT [17] model.

achieved significant progress, their reliance on costly pixel-level annotations limits their scalability and practical applicability. To alleviate the annotation burden, weakly-supervised RIS has been introduced [18], using weaker forms of supervision such as bounding boxes [19], scribbles [20], points [21], class labels [22, 23, 24, 25], and textual expressions [26, 27, 28]. Among these, text-only supervision is highly attractive for RIS due to its low cost and flexibility, but imposes additional challenges on cross-modal feature alignment and pseudo-labels quality. Recent efforts, such as TRIS [16], have explored weakly-supervised frameworks that incorporate CNN-based visual encoders, text-guided decoders, and bilateral prompt mechanisms for cross-modal interaction. While TRIS improves flexibility in multimodal interaction, it still suffers from insufficient feature alignment and suboptimal Stage-1 pseudo-labels—coarse masks produced in the initial weakly-supervised stage for subsequent refinement—thus limiting its segmentation accuracy.

Fusion Methods in Referring Image Segmentation Referring Image Segmentation (RIS) aimed to segment target objects in images conditioned on natural language descriptions [6]. A core challenge in RIS lay in effectively fusing visual and textual features [29]. Fusion methods were commonly categorized into early, mid-level, and late fusion [30]. In early fusion methods, Zhao et al. [31] proposed building a single stream encoder to achieve the early fusion and Hu et al. [6] proposed directly concatenating multimodal inputs. Liu et al. [32] proposed to average multimodal inputs. Mid-level fusion [33] leveraged CNN or Transformer encoders to extract modality-specific features, followed by feature-level interactions. In late-level fusion methods, Owens et al. [34] independently processed each modality and merged their final decisions through decision-level aggregation. Although these fusion strategies advanced multimodal learning, they suffered from common limitations when applied to Referring Image Segmentation (RIS). Specifically, they struggled to establish fine-grained alignment between language and vision due to insufficient cross-modal interaction and inadequate modeling of context-dependent relationships, leading to suboptimal segmentation accuracy. In addition, various approaches employed iterative refinement [35, 36, 32], dynamic filters [27, 37, 38], and attention mechanisms [39, 40, 12, 41, 42, 43, 44, 28, 45, 46] to improve fine-grained feature alignment and emphasize relevant regions. Recently, Transformer-based models [47, 48, 9, 49, 50, 51] demonstrated strong potential for modeling cross-modal dependencies due to their superior sequence modeling capabilities. CLIP [15] proposed a dual-encoder framework that leveraged Transformers for both image and text encoders. Nevertheless, achieving fine-grained alignment and effective interaction between visual and textual modalities remained an open challenge, particularly when handling diverse and complex descriptions.

In addition to the aforementioned methods, various innovative approaches have been proposed to further enhance referring image segmentation. For instance, Watanabe et al. [52] introduced a generative adversarial network incorporating referring image segmentation for text-guided image manipulation, aiming to preserve text-irrelevant regions while modifying text-relevant parts. Ji et al. [53] provided a comprehensive survey analyzing the main challenges in RIS and summarizing existing solutions, highlighting strategies such as multimodal fusion and robustness enhancement. Sun et al. [54] proposed a model based on convolutional nonlinear spiking neural P systems to enhance feature interaction and alignment in RIS. Pu et al. [50] presented an end-to-end method aligning linguistic and visual relationships to improve segmentation accuracy. Zhang et al. [55] introduced a cross-modal transformer with language queries to facilitate deep interaction between vision and language modalities. Dai and Yang [56] introduced Curriculum Point Prompting (PPT), which effectively incorporates SAM and CLIP for weakly-supervised referring image segmentation, demonstrating significant performance gains. While their approach benefits from SAM's strong segmentation priors, it also faces challenges due to noisy point prompts.

Based on this, we propose a novel weakly-supervised

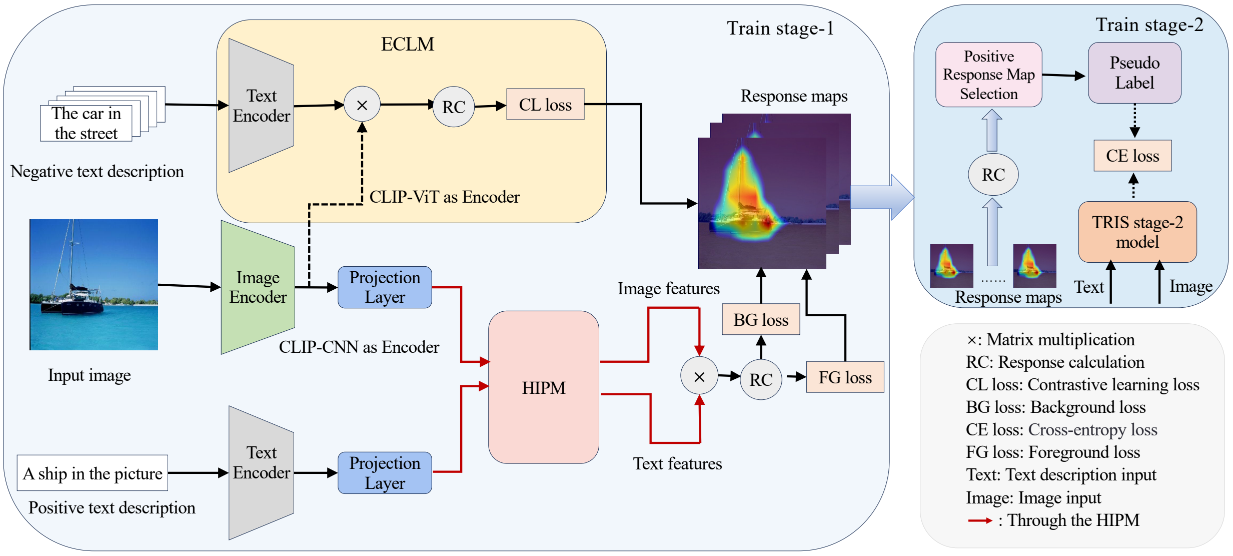


Figure 2: Overview of the EHIN. Stage-1: The model is jointly optimized with the Early-aware Contrastive Learning Module (ECLM) and the Hierarchical Interaction Prompt Module (HIPM), where ECLM provides contrastive supervision and HIPM enables hierarchical fusion. The response maps are obtained through their joint interaction. Both the image and text encoders are from the CLIP model. These additions improve the quality and the reliability of the pseudo-labels. Stage-2: Using the high-quality pseudo-labels generated in Stage-1, we train a segmentation network to produce the final segmentation maps. The proposed framework ensures better pseudo-labels generation and segmentation performance through the synergy of the two stages.

Early-aware Hierarchical Interaction Network (EHIN) for RIS, specifically designed to address the challenges of low-quality pseudo-labels and the difficulty of effectively aligning and integrating visual and textual features. Fig. 1 illustrates the practical application of the EHIN, showcasing its effectiveness in processing multimodal information.

3. Method

The weakly-supervised RIS task aims to link visual content with referring expressions at the pixel level without requiring pixel-level annotations. In real-world images, input descriptions often explicitly identify the target object, occasionally specifying its location. For example, in Fig. 1, “the man in the picture” clearly designates the target object for segmentation. Similarly, multiple valid descriptions for the same image, such as “the kite in the sky” or “the man on the left side of the picture”, can both locate the target object and provide contextual details about the background or other non-target elements.

To effectively leverage such diverse descriptions and address the challenges of feature interaction and pseudo-labels noise, we propose an Early-aware Contrastive Learning Module (ECLM). Specifically, ECLM incorporates a contrastive learning strategy that enhances the feature discrimination and improves the quality of pseudo-labels generation, thereby facilitating more accurate semantic segmentation. The positive textual description focuses on the semantics of a target object in the image, while the negative description, derived from unrelated images, provides clear contrast. By performing contrastive comparisons between images in the ECLM, the effectiveness of contrastive learning can be significantly enhanced.

Fig. 2 outlines the EHIN, which consists of two key stages: Stage-1 and Stage-2. Stage-1 initially employs the CLIP model as an encoder to process both textual and visual inputs, facilitating the interaction between text and image. Then the HIPM integrates multimodal inputs into a shared representation space, allowing deeper cross-modal fusion and improving alignment at both coarse and detailed levels. Negative textual descriptions are incorporated via an ECLM to enhance the model’s precision in distinguishing image details, improving target object localization guided by positive descriptions.

Stage-2 primarily leverages the pseudo-labels generated by Stage-1 to train the segmentation network. These pseudo-labels are derived from the response maps produced in the first stage and approximate true annotations without requiring direct pixel-level labels. By selecting high-confidence regions from model predictions or custom strategies, the pseudo-labels effectively guide the segmentation network to accurately identify target regions from textual descriptions. To validate the quality of the pseudo-labels, we quantitatively evaluate the outputs of the Stage-1 model—prior to pseudo-label generation—against ground-truth annotations using established metrics such as Pointing Metric (PointM) and Intersection-over-Union (IoU). As these intermediate predictions form the basis for the final pseudo-labels (which are further refined through models like IRNet), their strong performance indicates that the generated pseudo-labels are likely to be semantically reliable, thus supporting the overall effectiveness of our two-stage training framework.

3.1. Generation of the Response Map

We propose a CLIP-based method for accurately localizing target entities in images by effectively fusing visual and textual information. Given an input image $I \in \mathbb{R}^{H_1 \times W_1 \times 3}$, where $H_1 = 320$ and $W_1 = 320$ denote the height and width of the image and 3 corresponds to the RGB color channels, along with a textual query of length T, we first extract visual and textual features using the CLIP model. Specifically, the visual encoder E_v generates visual features $V_\epsilon \in \mathbb{R}^{H \times W \times C_v}$, and the textual encoder E_l produces textual features $L_\epsilon \in \mathbb{R}^{1 \times C_l}$, where H and W are the spatial dimensions of the visual feature map directly output by the CLIP visual encoder, and $C_v = 2048$ and $C_l = 1024$ are the numbers of channels in the visual and textual feature representations, respectively. To enable cross-modal interaction, visual features V_ϵ and textual features L_ϵ are projected into a shared hidden dimension $C_d = 1024$ using linear projection layers, producing unified feature representations $V \in \mathbb{R}^{H \times W \times C_d}$ and $L \in \mathbb{R}^{1 \times C_d}$. These features are then normalized to ensure consistent scaling for subsequent operations. Specifically, we apply L2 normalization to both image and text features within the model, immediately after they are extracted from the input data. This normalization is performed during both training and inference, ensuring that all features are projected onto the unit hypersphere.

To minimize the disparity between visual and textual features, we introduce HIPM, which produces the following feature representations:

$$V_F = V + \alpha \Gamma_{\text{Prompt1}}(V, L) + \beta \Gamma_{\text{Prompt2}}(V, L), \quad (1)$$

$$L_F = L + \alpha \Gamma_{\text{Prompt1}}(V, L) + \beta \Gamma_{\text{Prompt2}}(V, L) \quad (2)$$

where V_F and L_F indicate the updated visual and textual features, respectively. $\Gamma_{\text{Prompt}}(V, L)$ implies the new features obtained after the original textual and visual features pass through the Prompt1 and Prompt2 modules. The coefficients α and β are used to weight the alignment between visual and textual features, and correspond to the same parameters later defined in Equations (1), (2), (8), (9) and (12), (13). We set $\alpha = 0.1$ and $\beta = 0.05$ throughout our experiments. A detailed ablation study on the effects of these parameters is provided in Table 7.

The overall process is implemented within our proposed HIPM framework (Section 3.2), where the Prompt1 module [16] serves as the fundamental building block of the feature prompting mechanism.

After deriving the aligned features V_F and L_F, we compute the similarity $R_{i,j}$ between each pixel position and the textual features, defined as:

$$R_{i,j} = \sum_{\theta=1}^{C_d} V_{i,\theta} \cdot L_{j,\theta} \quad (3)$$

where $\theta \in \{1, \ldots, d\}$ indexes the hidden feature dimension of size d, and the dot represents element-wise multiplication. This adjustment highlights the image regions most aligned with the textual description, with higher scores indicating a higher likelihood of containing the target object.

3.2. Hierarchical interaction Prompt Module

To enable effective knowledge transfer from pretrained models [15] and achieve a more comprehensive interaction between visual and textual features, we propose the HIPM, incorporating a bilateral prompt (BP) designed to align domain disparities between visual and textual features. Unlike some conventional prompting approaches that either adjust visual and textual features independently [11] or enhance textual features using visual inputs alone [57, 58, 59], the HIPM tackles the complex challenge of seamlessly integrating images and textual descriptions, thereby strengthening the model's ability to understand multimodal information. To further strengthen cross-modal interaction, HIPM introduces a hierarchical fusion mechanism that progressively integrates visual and textual features across multiple levels of abstraction. Moreover, it incorporates repeated residual connections throughout the fusion layers, which facilitate iterative refinement and lead to more robust and semantically rich multimodal representations. The bilateral prompt enables bidirectional complementary reinforcement between the modalities. This method strengthens the cross-modal

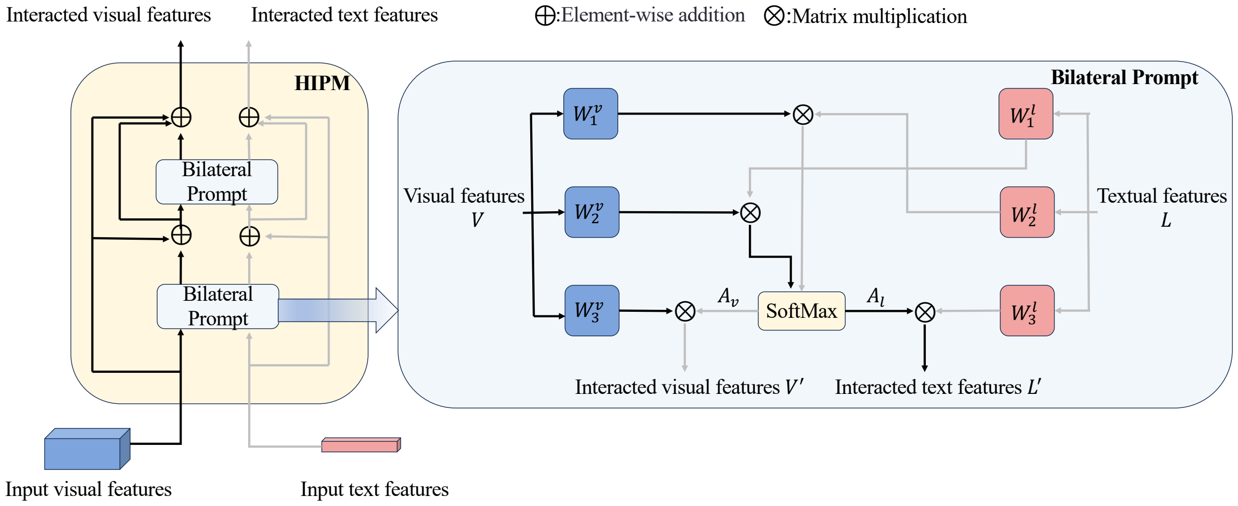


Figure 3: Structure of the HIPM. V represents visual features, and L represents textual features. Black arrows indicate visual features, while gray arrows indicate textual features. W_*^* denote the projection functions.

associations between visual and textual features and enhances the model's effectiveness in vision-language interaction tasks. Unlike conventional single-path skip connections, we introduce a dual-residual integration strategy, in which the original visual (or language) features are incorporated twice across the multi-stage bilateral attention blocks. This design explicitly aims to preserve unimodal representations during iterative cross-modal refinement, thereby preventing excessive drift in the feature space and maintaining the integrity of modality-specific semantics. The hierarchical fusion in HIPM is motivated by the need to capture cross-modal interactions at multiple levels of abstraction, enabling progressive refinement of joint representations. Repeated residual connections are introduced to facilitate information flow, mitigate gradient vanishing, and promote feature reuse, as supported by the theory of residual learning. This design ensures both effective feature integration and information preservation across fusion stages.

Fig. 3 illustrates the architecture of the proposed HIPM. Using the input visual features $V \in \mathbb{R}^{H\times W\times C_d}$ and textual features $L \in \mathbb{R}^{1\times C_d}$, we first generate two attention maps based on the following formula:

$$A_l = \text{SoftMax}\left(\frac{\left(VW_1^v\right)\otimes\left(LW_2^l\right)^T}{\sqrt{C_d}}\right), \tag{4}$$

$$A_v = \text{SoftMax}\left(\frac{\left(LW_1^l\right)\otimes\left(VW_2^v\right)^T}{\sqrt{C_d}}\right) \tag{5}$$

where $A_l \in \mathbb{R}^{1\times HW}$, $A_v \in \mathbb{R}^{HW\times 1}$ signify the mappings from visual features to textual features and from textual features to visual features, respectively, $W_*^v \in \mathbb{R}^{C_d\times C_d}$ and $W_*^l \in \mathbb{R}^{C_d\times C_d}$ indicate the learnable parameters for V and L, respectively, and the symbol $\otimes$ implies the matrix multiplication. Our fusion strategy is based on bilateral cross-attention, inspired by the designs in TRIS [16] and LAVT [17]. This mechanism computes attention weights via matrix multiplication to capture content-dependent semantic relationships across modalities. Compared to concatenation-based fusion, it allows for more expressive and flexible visual-linguistic interactions. The subsequent calculation is performed as follows:

$$L' = A_l^T \otimes \left(VW_3^v\right), \tag{6}$$

$$V' = \text{Reshape}\left(A_v^T \otimes \left(LW_3^l\right), (B, C, N)\right)^{\top(1,2)} \tag{7}$$

where Reshape$(\cdot)$ denotes a tensor reshaping operation that changes the shape without altering data layout, and the superscript $\top_{(1,2)}$ denotes transposition of the second and third dimensions, $L' \in \mathbb{R}^{1\times C_d}$ and $V' \in \mathbb{R}^{H\times W\times C_d}$ denote the visual and textual features after the first enhancement, respectively, and the subsequent L_n' and V_n' imply the visual and textual features after passing through the n-th bilateral prompt, respectively. As shown in Fig. 3, the visual and textual features after the first bilateral prompt in our HIPM are expressed as follows:

$$L_1' = L' + \alpha L, \tag{8}$$

$$V_1' = V' + \alpha V \tag{9}$$

The features obtained after processing through the second bilateral prompt are formulated as follows:

$$L'_2 = BP(L'_1), \tag{10}$$

$$V'_2 = BP(V'_1) \tag{11}$$

The resulting visual and textual features are expressed as follows:

$$L'_F = L'_2 + \alpha L + \beta L'_1, \tag{12}$$

$$V'_F = V'_2 + \alpha V + \beta V'_1 \tag{13}$$

where α and β indicate their respective weight coefficients. This approach effectively minimizes disparities between multimodal features.

3.3. Early-aware Contrastive Learning Module and Pseudo-Labels Generation

Unlike conventional contrastive learning methods that typically operate on high-level or projected features, our Early-aware Contrastive Learning Module (ECLM) performs feature fusion directly on the original image and text representations extracted by a CLIP-ViT encoder [15]. This early-stage fusion strategy enables finer-grained cross-modal alignment, thereby enhancing the semantic consistency between visual and textual features. To leverage this alignment for supervision, the ECLM generates contrastive signals that guide the model to distinguish target-relevant features from background noise under weak supervision. The resulting aligned representations are incorporated into the loss function, improving the reliability of pseudo-labels and facilitating more effective downstream learning. In our experiments, the input image I and a positive text description $R_p \in \mathbb{R}^T$ serve as the positive sample. We then randomly select a set of sample descriptions $R_n \in \mathbb{R}^{N\times T}$ from other image descriptions in the dataset for negative samples to perform contrastive learning. Positive and negative samples are defined at the text level, where positives are textual descriptions paired with their corresponding images, and negatives are descriptions of unrelated images from the same training dataset. This strategy is consistent with prior works [15, 16] on referring expression comprehension and negative sample construction. As illustrated in Fig. 2, we first compute the response map for the input image I and compute the image-level score y_i for each description as follows:

$$y_i = \max_i R^a_{i,j} + \frac{1}{HW}\sum_{i=1}^{HW}\left(R^a_{i,j} + \psi\left(R^a_{\cdot,j}\right)\right) \tag{14}$$

where $\psi\left(R^a_{\cdot,j}\right)$ refers to the regularization term proposed by [60]. We regard the positive query Q_p as the ground-truth description of the target object. To introduce contrastive supervision, we further sample K negative queries $\{F^k_n\}^K_{k=1}$ from the same image, each referring to different objects. The contrastive alignment loss is then defined as:

$$\mathcal{L}_{\text{CL\_loss}} = -\left(\log S\left(I, R_p, Q_p\right) + K\sum_{k=1}^{K}\log\left(1 - S\left(I, R_p, F^k_n\right)\right)\right) \tag{15}$$

Here, $S(\cdot,\cdot,\cdot)$ denotes a similarity function that measures the semantic alignment between a visual region and a textual query, defined as:

$$S\left(I, R_p, Q\right) = \varphi\left(E_v(I \odot u_p(R_p)),\ E_l(Q)\right) \tag{16}$$

where E_v and E_l denote the visual and text encoders of CLIP, respectively. The symbol $\odot$ represents the Hadamard (element-wise) product, and $u_p(\cdot)$ is an upsampling function that resizes the region mask R_p to match the resolution of the image I. The function $\varphi(\cdot,\cdot)$ computes the cosine similarity between the resulting visual and textual embeddings.

During the classification process, we compare the target object with other objects in the same image, treating these as background noise to enhance accuracy. A baseline response map is generated using the input image I and the positive textual description R_p. In addition, K negative sample descriptions are randomly selected from the dataset as negative samples.

The generation of pseudo-labels in our framework is primarily driven by response maps produced in Stage-1, which is enhanced with two specialized modules: the Early-aware Contrastive Learning Module (ECLM) and the Hierarchical Interaction Prompt Module (HIPM). This stage outputs coarse yet semantically consistent response maps for each query, serving as initial pseudo-instance segmentation cues. These initial masks are subsequently refined by the Instance Relationship Network (IRNet) [61], which leverages image-level supervision to improve mask quality and enforce structural consistency.

3.3.1. Response Map Generation

In the initial stage, the IRNet model generates a Class Activation Map (CAM), highlighting image regions most relevant to the predicted class. The CAM provides insights into the model's preliminary focus, identifying critical areas for further refinement and optimization.

Table 1: Experiment datasets (Images = number of images; Ref. Exp. = number of referring expressions).

Dataset	Images	Ref. Exp.	Description	Subsets
RefCOCO	19, 994	142, 209	Emphasizes object positioning properties	–
RefCOCO+	19, 992	141, 564	Focuses on the visual attributes of objects.	–
RefCOCOg	26, 711	104, 560	Includes more extended and intricate descriptions.	Google, UMD
ReferIt	20, 000	120, 072	Features concise and accurate text descriptions, often using single-word labels.	–

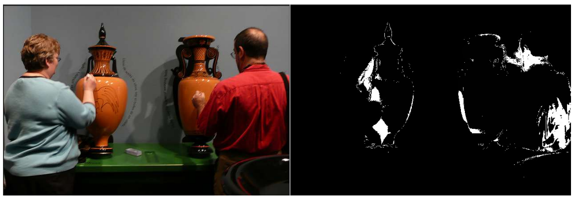


Figure 4: Comparison of outcomes: Original Image and the label from IRNet

3.3.2. Conditional Random Field (CRF) Optimization

To improve the accuracy and coherence of the initial predictions, CRF optimization is applied to the response map, effectively refining sequential and structured data. CRF models relationships between adjacent pixels to ensure consistency of the predicted results, remove isolated outliers, and preserve critical feature boundaries. This process effectively refines the response map, enhancing spatial consistency and prediction accuracy.

3.3.3. Iterative Refinement Label Generation

After CRF optimization, the response map serves as input for the iterative refinement phase, generating the label from IRNet. This refined label acts as a crucial guide for the IRNet model, facilitating detailed analysis and adjustments in subsequent iterations. By linking the initial predictions with the final, high-precision results, the label from IRNet helps the model focus on areas requiring further improvement, progressively enhancing the accuracy of the predictions. Once fully trained, the IRNet model generates pseudo-labels for use in the second stage of training. The original image and its corresponding label from IRNet are shown in Fig. 4.

3.4. Loss Function for Stage-1

To optimize the network in Stage-1, we jointly supervise foreground-text alignment, background suppression, and early-aware contrastive learning signals. The overall loss function is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{FG\_loss}} + \lambda_2 \mathcal{L}_{\text{CL\_loss}} + \lambda_3 \mathcal{L}_{\text{BG\_loss}} \tag{17}$$

Here, λ_1, λ_2, and λ_3 are weighting coefficients that balance the contributions of the foreground alignment loss $\mathcal{L}_{\text{FG\_loss}}$, the contrastive learning loss $\mathcal{L}_{\text{CL\_loss}}$, and the background suppression loss $\mathcal{L}_{\text{BG\_loss}}$, respectively. Following the setting in [16], we set $\lambda_1 = 5$, $\lambda_2 = 2$, and $\lambda_3 = 1$ throughout all experiments. We do not perform ablation studies on these hyperparameters, as they are empirically found to be insensitive to performance within a reasonable range.

The foreground loss $\mathcal{L}_{\text{FG\_loss}}$ encourages alignment between image regions and their corresponding textual phrases:

$$\mathcal{L}_{\text{FG\_loss}} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\text{sim}(I_i, T_i)\right) \tag{18}$$

In this equation, N is the batch size, I_i denotes the i-th image, and T_i is the corresponding textual description. The function $\text{sim}(I_i, T_i)$ computes the cosine similarity between normalized CLIP image and text embeddings. This term penalizes low similarity between correctly matched image-text pairs. The background loss $\mathcal{L}_{\text{BG\_loss}}$ suppresses spurious alignments between image regions and unrelated (negative) textual phrases:

$$\mathcal{L}_{\text{BG\_loss}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{K}\sum_{k=1}^{K} \log\left(1 - \text{sim}(I_i, T^-_{i,k})\right) \tag{19}$$

Here, $T^-_{i,k}$ represents the k-th negative phrase for the i-th image, and K is the number of negative samples per image. The loss penalizes high similarity between the image and negative phrases, thereby improving discrimination between foreground and background concepts. We further elaborate on the contrastive learning loss $\mathcal{L}_{\text{CL\_loss}}$ in Equation (15), as it serves as the core supervision signal in our Early-aware Contrastive Learning Module (ECLM).

3.5. Model Training With Pseudo-labels

In Stage-2, we follow the training procedure described in TRIS [16], where the model is trained in a supervised manner using the pseudo-labels generated in

Stage-1. Specifically, we employ standard segmentation loss (cross-entropy and Dice loss), the Adam optimizer, a batch size of 24, an initial learning rate of 1e-5, and train for 15 epochs, with all hyperparameters kept consistent with TRIS to ensure fair comparison. Distributed data parallelism is used when multiple GPUs are available, and an exponential moving average (EMA) of model weights with a consistency loss (MSE or KL divergence) is optionally applied.

4. Experiments and Results

4.1. Dataset

The datasets used in this study, detailed in Table 1, include RefCOCO, RefCOCO+ [67], RefCOCOg [68], and ReferIt [69]. The first three datasets were derived from MSCOCO [4], where each sample consisted of an image and its corresponding referring expressions, although they differed in several key aspects. Specifically, RefCOCO and RefCOCO+ were collected from interactive games, while RefCOCOg was sourced from non-interactive settings. In addition, RefCOCO emphasizes object positional attributes, whereas RefCOCO+ highlights appearance features of objects. By contrast, RefCOCOg poses a greater challenge with its longer, more complex referring expression. It is further divided into two subsets: Google [68] and UMD [70].

The ReferIt dataset comprises 20,000 images from SAIAPR12, with 54,127 in the train set, 5,842 in the validation set, and 60,103 in the test set, respectively. The text descriptions are concise and precise; for example, a mountain image is labeled simply as 'mountain'. In our experiments, the positive and negative textual descriptions are constructed from RefCOCO, RefCOCO+, RefCOCOg and ReferIt expressions.

4.2. Experimental Setup

In our experiments, both ResNet-50 and ViT served as image encoders, with both the image and text encoders initialized with CLIP. We set the HIPM parameters α and β to 0.1 and 0.05, respectively. The ECLM employed six negative samples for each positive instance; for images lacking sufficient descriptions additional descriptions are randomly sampled from other images to ensure $K = 6$.

All models were implemented using PyTorch and trained with a batch size of 24 for 15 epochs.

4.2.1. Evaluation Metrics

The model's performance on the RIS task was assessed using multiple evaluation metrics. Following the established guidelines [39, 47], we used the intersection over union (IoU) metric to measure the model's accuracy in segmenting the target region. In addition, to evaluate the model's performance in object localization, we utilized PointM [16] and BoxAcc [71, 64] as supplementary metrics.

In the ablation experiments, we incorporated PointIt and mean intersection over union (mIoU) as additional metrics to evaluate the impact of parameter settings on model performance. This approach validated the effectiveness of individual components and provided a deeper understanding of the model's behavior under varying conditions.

4.3. Experimental Results

We assess the effectiveness of the proposed method by benchmarking it against a series of advanced models, as detailed in Table 2. It presents a quantitative comparison across segmentation and localization accuracy using two key evaluation metrics. Our approach outperforms existing methods, underscoring its efficacy and superiority.

Table 2 highlights our model's superior performance in Stage 1 compared to prior state-of-the-art models on the RefCOCO+, RefCOCOg, and ReferIt datasets for both metrics, IoU and PointM. Furthermore, in Stage-2, our model achieves state-of-the-art performance on the vast majority of tested datasets. For the IoU metric, our model achieves a performance improvement of nearly 3% over TRIS on the ReferIt dataset in Stage-1. Moreover, for the Pointing Metric (PointM), the proposed method outperforms TRIS by nearly 14% on the RefCOCO+ dataset in Stage-1. Although our model delivers substantial enhancements on RefCOCOg, RefCOCO+ and ReferIt datasets, the improvement on RefCOCO is comparatively modest. The disparity arises because the RefCOCOg, RefCOCO+ and ReferIt datasets offer more precise descriptive expressions, enabling better alignment of visual and textual features and enhancing model training performance. To better understand the performance gap on the RefCOCO testA split, we attribute it to both dataset-specific linguistic patterns and architectural design differences. RefCOCO testA contains more person-centric descriptions with fine-grained spatial and action-related cues. PCNet is specifically designed to handle such complexity through phrase decomposition via a Large Language Model (LLM), progressive localization using multi-stage mod-

Table 2: Comparison of CNN models performance across four datasets utilizing IoU and PointM as evaluation metrics. "Sup." denotes the supervision type used by the model; T represents the text description labels; (G) and (U) imply different splits of the RefCOCOg dataset, namely UMD and Google; and "-" signifies no available value.

Metric	**Method**	**Sup.**	**Backbone**	**RefCOCO**			**RefCOCO+**			**RefCOCOg**			**ReferIt**
				val	testA	testB	val	testA	testB	val(G)	val(U)	test(U)	val
IoU	AMR+ [62]	T	ResNet-50	14.12	11.69	17.47	14.13	11.47	18.13	15.83	15.46	15.59	18.98
	GbS+ [63]	T	VGG16	14.59	14.60	14.97	14.49	14.49	15.77	14.21	13.75	14.20	14.21
	WWbL+ [64]	T	VGG16	18.26	17.37	19.90	19.85	18.70	21.64	21.84	21.75	21.82	27.68
	TRIS (Stage-1) [16]	T	ResNet-50	25.11	26.47	23.80	22.31	21.61	22.86	26.93	26.62	27.27	33.33
	PCNet$_F$ [65]	T	ResNet-50	30.90	**35.20**	26.30	28.90	31.90	26.50	29.80	29.70	30.20	-
	ICICS [66]	T	ResNet-50	31.06	32.30	30.11	**31.28**	**32.11**	30.13	-	-	-	-
	EHIN (Our, Stage-1)	T	ResNet-50	25.10	25.95	25.11	25.53	23.77	25.18	27.43	27.04	27.50	36.14
	TRIS (Stage-2) [16]	T	ResNet-50	31.17	32.43	29.56	30.90	30.42	30.80	36.00	36.19	36.23	44.57
	EHIN (Our, Stage-2)	**T**	**ResNet-50**	**33.12**	34.49	**30.38**	30.41	30.20	**31.34**	**36.14**	**36.90**	**36.99**	**45.47**
PointM	AMR+ [62]	T	ResNet-50	15.55	5.52	28.91	16.33	5.90	30.27	25.51	24.96	26.14	7.12
	GbS+ [63]	T	VGG16	21.58	19.52	25.95	20.95	18.34	25.96	24.64	24.60	25.38	30.30
	WWbL+ [64]	T	VGG16	31.28	31.15	30.79	34.47	33.30	36.10	29.32	32.13	31.37	42.84
	TRIS (Stage-1) [16]	T	ResNet-50	51.92	60.88	43.02	40.85	40.94	41.13	52.48	51.98	53.29	61.70
	PCNet$_F$ [65]	T	ResNet-50	**59.60**	**66.60**	**48.20**	54.70	65.00	44.10	57.90	57.00	57.21	-
	EHIN (Our, Stage-1)	T	ResNet-50	44.32	43.77	42.73	55.63	54.16	43.47	53.64	51.96	52.41	61.49
	TRIS (Stage-2) [16]	T	ResNet-50	54.72	65.64	43.40	53.72	61.30	45.24	58.01	58.84	58.70	67.00
	EHIN (Our, Stage-2)	**T**	**ResNet-50**	55.31	64.06	43.70	**58.05**	**67.19**	**45.84**	**59.65**	**59.19**	**59.42**	**69.32**

ules, and specialized losses (RaS and IaD) for ambiguity suppression. In contrast, our EHIN framework employs a holistic bilateral attention strategy without explicit language decomposition or multi-stage refinement. While effective across general benchmarks, it may be less optimized for the highly specific linguistic structures found in testA. On the ReferIt dataset, the IoU metric is improved by approximately 3% compared to TRIS in Stage-1, largely due to more accurate annotations. These refined annotations improve the model's ability to identify and localize target objects, ensuring higher data consistency and accuracy. Our approach achieves substantial improvements compared with earlier models, supported by two key elements. First, the newly introduced HIPM effectively facilitates interaction between visual and textual features. Second, our early-aware contrastive learning technique minimizes the effects of background noise and irrelevant features, thereby improving the model's overall performance. Table 2 presents a comparative analysis of our model's performance against other models.

As shown in Figs. 5, 6, the proposed HIPM delivers superior image segmentation performance in Stage-1. Remarkably, in several instances, its segmentation quality rivals the Stage-2 results of the traditional TRIS method. This finding underscores the HIPM method's enhanced capacity to interpret the relationship between image content and its corresponding textual descriptions early in the process, offering a novel approach to effective image segmentation. In Stage-2, leveraging the pseudo-labels generated in Stage-1 enables precise targeting of objects during segmentation, ensuring accurate localization of target regions. As illustrated in Fig. 6, this approach preserves segmentation accuracy while further refining target boundaries, significantly enhancing the model's representational capabilities.

4.4. Ablation Study and Analysis

To evaluate the effectiveness of the key components in the proposed network, we performed a series of ablation experiments on the RefCOCO+ dataset using the validation set (val) for systematic analysis. We investigated the performance of different numbers of bilateral prompt layers within HIPM, along with evaluating the single-layer BP, the inclusion of the HIPM and the ECLM.

The performance of different numbers of bilateral prompt layers within the HIPM is presented in Table 3. Notably, the best performance is achieved when the number of layers is set to two. This finding indicates that increasing the number of layers does not necessarily enhance the interaction of visual and textual features. The underlying reason is that while deeper models generally improve data fitting, the limited availability of textual and visual data constrains the benefits of deeper architectures. This limitation is particularly evident when the number of words does not exceed 20, where additional layers fail to provide further improvements in feature alignment.

To further verify the effectiveness of our dual-residual integration strategy, we conducted an ablation study comparing models with a single residual connection reuse ("Repeat Once") and our proposed dual reuse ("Repeat Twice"). As shown in Table 4, repeating

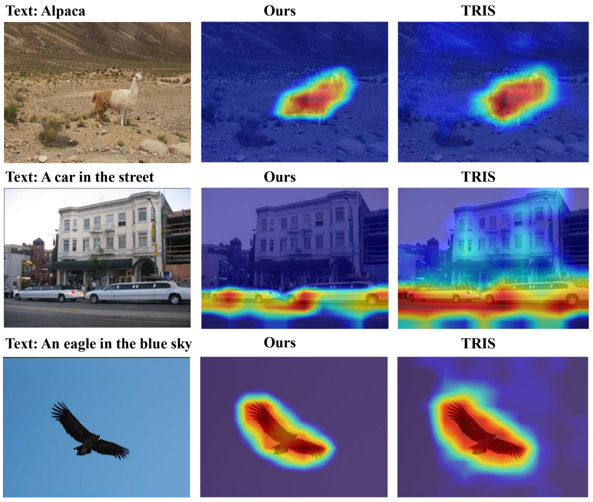


Figure 5: Comparison visualization for Stage-1 models, illustrating the results obtained after Stage-1 and their comparative analysis with TRIS results to evaluate performance improvements and differences.

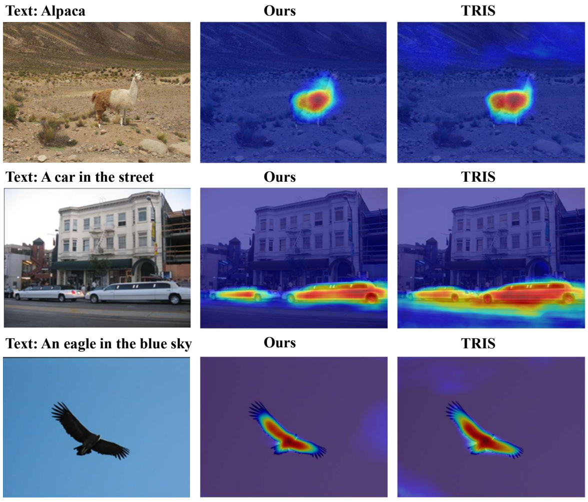


Figure 6: Comparison visualization for Stage-2 models, illustrating the results obtained after Stage-2 and their comparative analysis with TRIS results to evaluate performance improvements and differences.

Table 3: Impact analysis of different numbers of bilateral prompt layers in HIPM on the RefCOCO+ val dataset.

bilateral prompt Layer numbers within HIPM	IoU	mIoU	PointIt	PointM	BoxAcc
1	23.83	23.21	54.77	45.83	19.89
2	**25.53**	**26.77**	**64.02**	**55.63**	23.98
3	23.74	23.91	52.72	44.74	**24.50**
4	23.19	23.57	55.97	47.04	20.12

Table 4: Comparison between single and dual residual reuse strategies on the RefCOCO+ validation set. “Repeat Once” incorporates the original visual/language features once in the HIPM, while “Repeat Twice” follows our proposed dual-residual integration.

Residual Strategy	IoU	mIoU	PointIt	PointM	BoxAcc
Repeat Once	18.81	19.71	56.55	44.47	14.25
Repeat Twice (Ours)	**25.53**	**26.77**	**64.02**	**55.63**	**23.98**

Table 5: Impact Analysis of Loss Components on RefCOCO+ Val Performance ($\mathcal{L}_{\mathrm{BG\_loss}}$: Background Loss, $\mathcal{L}_{\mathrm{FG\_loss}}$: Foreground Loss, $\mathcal{L}_{\mathrm{CL\_loss}}$: Contrastive Learning Loss). We further observe that using only $\mathcal{L}_{\mathrm{CL\_loss}}$ leads to a drastic performance collapse (mIoU $\approx 1\%$), making the model unable to function

Different Loss Components	IoU	mIoU	PointIt	PointM	BoxAcc
$\mathcal{L}_{\mathrm{BG\_loss}}$ & $\mathcal{L}_{\mathrm{FG\_loss}}$ & $\mathcal{L}_{\mathrm{CL\_loss}}$	**25.53**	**26.77**	**64.02**	**55.63**	**23.98**
$\mathcal{L}_{\mathrm{BG\_loss}}$ & $\mathcal{L}_{\mathrm{FG\_loss}}$	24.82	25.74	63.77	55.22	23.93
$\mathcal{L}_{\mathrm{BG\_loss}}$ & $\mathcal{L}_{\mathrm{CL\_loss}}$	22.26	23.56	62.75	51.28	19.38
$\mathcal{L}_{\mathrm{FG\_loss}}$ & $\mathcal{L}_{\mathrm{CL\_loss}}$	23.95	24.92	63.78	53.00	22.64
$\mathcal{L}_{\mathrm{BG\_loss}}$	19.51	18.46	54.65	46.11	14.60
$\mathcal{L}_{\mathrm{FG\_loss}}$	17.22	16.38	52.39	43.50	11.19

the original modality features twice significantly improves both accuracy and IoU, confirming that such a design helps stabilize feature refinement and preserve unimodal representations across stages.

We perform a detailed ablation to assess the contributions of each loss component. As shown in Table 5, removing any single term leads to a clear performance drop, confirming their complementary roles. The background loss ($\mathcal{L}_{\mathrm{BG\_loss}}$) proves most critical, with its absence causing the largest degradation in IoU and mIoU, highlighting its role in modeling spatial context. The contrastive loss ($\mathcal{L}_{\mathrm{CL\_loss}}$) consistently enhances alignment across modalities, while the foreground loss ($\mathcal{L}_{\mathrm{FG\_loss}}$) improves object boundary precision. The full combination achieves the best results across all metrics, validating the effectiveness of our joint loss design. We further observe that using only $\mathcal{L}_{\mathrm{CL\_loss}}$ leads to a drastic performance collapse (mIoU $\approx 1\%$), making the model unable to function. This extreme case is omitted from Table 5 for clarity, but it confirms that contrastive supervision alone is insufficient without dense foreground and background guidance.

For the evaluation of other modules, we first began by removing the HIPM structure, retaining only a single bilateral prompt to assess the performance of the simplified model. Subsequently, we gradually added the HIPM, followed by the ECLM. This sequential approach clarified the contribution of each module to the model's performance. All evaluations of the components were conducted in Stage-1, with results recorded after each modification or addition.

As shown in Table 6, each module impacts the model's performance differently. Combining BP and HIPM in this configuration enhances performance, indicating that their integration during training improves feature extraction and detection accuracy. Adding the ECLM to the HIPM significantly enhances model performance, demonstrating the effectiveness of our approach. This result confirms that integrating multimodal interaction strategies and Early-aware substantially improves referring image segmentation. Fig. 7 presents the comparison results between the Late-aware Contrastive Learning and the Early-aware Contrastive Learning Module (ECLM) on the RefCOCO+ train dataset, demonstrating the effectiveness of our ECLM. The key distinction between Late-aware Contrastive Learning and Early-aware Contrastive Learning is that, in the former, the features undergo HIPM processing prior to being utilized in contrastive learning.

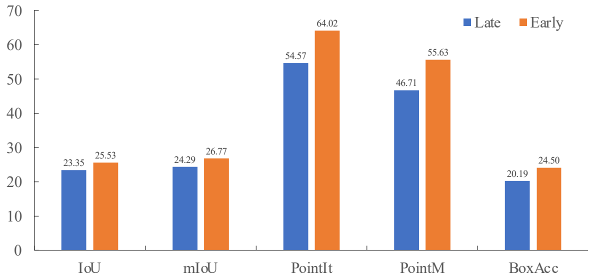


Figure 7: Comparison between the Late-aware Contrastive Learning and the proposed Early-aware Contrastive Learning on the RefCOCO+ val dataset.

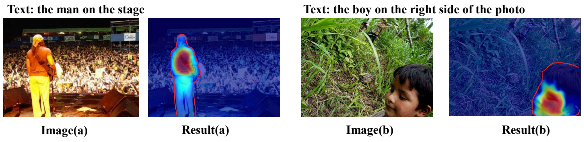
5. Hyperparameter Analysis

Table 7 presents the comparative results for varying α and β values in the HIPM. Analysis of the results reveals that the optimal parameter combination is $\alpha = 0.1$ and $\beta = 0.05$, delivering superior performance across various evaluation metrics. By contrast, increasing α to 0.2 slightly reduces overall performance, particularly in *PointIt* and *BoxAcc*, suggesting that higher weights may cause overfitting and suboptimal feature selection.

Moreover, Table 8 presents the ablation experiment results for selecting the number of text samples K in ECLM and the batch size N during model training. For N experiments, the selected value of K was fixed at 6, and for K experiments, N was set to 24. The table indicates that IoU and mIoU metrics achieve optimal performance when N is set to 24. However, increasing the value of N further degrades performance, as excessive N values reduce the correlation between negative samples, hindering the model's ability to extract meaningful information. With $N = 24$ ECLM effectively captures valuable key features, thus enhancing the overall model performance.

In addition, although there is not much difference in the results when K is set to 1 and 6 on the RefCOCO+ dataset, a slight performance gain is still observed when K is set to 6 in Table 9. Similar patterns are observed on both the RefCOCOg and ReferIt datasets, where $K = 6$ consistently yields the best overall performance across all evaluation metrics. This phenomenon can be attributed to ECLM, where appropriately increasing the K value exposes the model to a broader range of samples, enabling it to learn more diverse and rich information. Consequently, an optimal K value enhances the learning efficiency of the model and improves performance across multiple evaluation metrics.

To evaluate the practical efficiency of our method, we compare the average inference time per sample with TRIS on a single NVIDIA A40 GPU. As shown in Table 10, under identical settings (batch size = 1, input resolution = 320×320), EHIN achieves an average inference time of 2.05 seconds, outperforming TRIS, which takes 2.18 seconds. This demonstrates that EHIN not only improves segmentation performance but also maintains competitive computational efficiency.



Text: the man on the stage
Text: the boy on the right side of the photo
Image(a)
Result(a)
Image(b)
Result(b)

Figure 8: Two failure cases. Both images are from the Referit (U) val set, and the results are from Stage-2. The GT objects are indicated with red contours for visualization.

6. Limitations and Discussion

Nonetheless, the proposed model faces challenges with limited data, highlighting the need for optimization strategies for small datasets. In addition, as illustrated by the failure case in Fig. 8, our method exhibits limitations when segmenting images that require fine-grained detail. For example, in the case of "the boy on the right

Table 6: Ablation study for impact analysis of individual modules on the RefCOCO+ val dataset. EHIN represents our Early-aware Hierarchical Interaction Network, HIPM represents our Hierarchical Interaction Prompt Module, and ECLM represents our Early-aware Contrastive Learning Module.

Methods	**IoU**	**mIoU**	**PointIt**	**PointM**	**BoxAcc**
EHIN without HIPM & ECLM	22.22	22.06	52.14	44.45	18.33
EHIN without ECLM	23.83	23.21	54.77	45.83	19.89
EHIN without HIPM	23.22	23.42	54.30	46.57	19.85
EHIN	**25.53**	**26.77**	**64.02**	**55.63**	**23.98**

Table 7: The influence of parameters α and β in HIPM on model performance, evaluated on the RefCOCO+ val dataset.

α	β	**IoU**	**mIoU**	**PointIt**	**PointM**	**BoxAcc**
0.1	0.02	22.83	23.21	54.77	45.83	19.89
0.1	**0.05**	**25.53**	**26.77**	**64.02**	**55.63**	**23.98**
0.1	0.1	23.05	23.37	52.71	44.24	20.56
0.2	0.02	23.13	23.17	51.18	42.95	18.97
0.2	0.05	23.06	23.00	50.32	42.05	19.72
0.2	0.1	23.22	23.30	50.07	43.13	19.68

Table 8: The impact of varying K and N values on model performance in ECLM is investigated by fixing one parameter at a time, evaluated on the RefCOCO+ val dataset. When evaluating N, the optimal K= 6 is fixed, and different values of N are selected; similarly, when evaluating K, the optimal N= 24 is fixed.

Metric	**N**				**K**			
	6	**12**	**24**	**48**	**1**	**3**	**6**	**9**
IoU	21.05	22.65	**25.53**	23.30	23.21	22.84	**25.53**	23.15
mIoU	22.84	24.76	**26.77**	23.65	23.57	23.06	**26.77**	23.52
PointIt	54.77	56.16	**64.02**	55.00	54.56	54.51	**64.02**	53.52
PointM	51.11	53.78	**55.63**	46.18	45.96	45.94	**55.63**	45.03
BoxAcc	18.49	21.20	**23.98**	20.77	20.10	19.24	**23.98**	20.31

Table 9: To evaluate the robustness of ECLM to the number of negative samples (K), we conduct sensitivity analysis on RefCOCOg and ReferIt (validation splits), with the number of positive samples (N) fixed at 24. Results show that ECLM performs consistently across datasets, with minor variations likely due to differences in linguistic complexity and referential ambiguity.

Metric	**RefCOCOg(K)**				**Referit(K)**			
	1	**3**	**6**	**9**	**1**	**3**	**6**	**9**
IoU	25.10	25.00	**27.43**	25.41	34.23	32.60	**36.14**	33.37
mIoU	25.56	25.60	**27.15**	26.04	33.12	31.57	**35.56**	32.66
PointIt	66.50	**69.19**	67.42	67.65	63.91	69.92	**73.03**	54.45
PointM	52.81	52.91	**53.64**	52.98	53.47	59.29	**61.49**	45.06
BoxAcc	26.88	26.09	**30.63**	27.80	38.17	39.05	**42.27**	37.05

Table 10: Average inference time per sample (in seconds) on a single NVIDIA A40 GPU. Batch size = 1, input resolution = 320×320.

Model	**Inference Time (seconds)**
TRIS	2.18
EHIN (ours)	**2.05**

side of the photo”, the model failed to accurately segment certain elements such as the hair. This indicates that the model may struggle to capture subtle or intricate features in complex scenes. To address this, future research should focus on optimizing the use of textual data. Specifically, expanding and enriching the textual supervision signals may offer a viable solution to improve the model’s ability to align linguistic and visual cues at a finer granularity. By incorporating more detailed, context-aware textual guidance during training, the model could learn to better distinguish subtle differences in visual features, thereby improving segmentation performance in challenging, fine-grained tasks. While recent SAM-based methods such as PPT[56] and Grounded-SAM[72] have extended SAM to accept textual prompts, they still rely on heavy segmentation supervision and pretrained mask decoders. Our method differs fundamentally in that it operates in an end-to-end fashion, trained with only weak supervision from image-text pairs, without relying on any large-scale segmentation masks or pretrained segmentation models. This design allows for better generalization in low-resource settings and significantly reduces computational complexity. Currently, our approach does not incorporate the Segment Anything Model (SAM), as our primary focus is on addressing the challenges posed by low-quality pseudo-labels and the effective alignment and integration of visual and textual features. Nevertheless, we plan to integrate SAM in future work to leverage its strengths in segmentation tasks, particularly in enhancing the model’s generalization and precision in complex scenes.

7. Conclusion

This study introduces the novel Early-aware Hierarchical Interaction Network (EHIN) for RIS, using textual descriptions as the sole supervisory signal. Existing RIS models often struggle to effectively fuse visual and textual features, and their performance is further constrained by the lack of high-quality pseudo-labels for supervision. EHIN is composed of two key modules. The first, the Early-aware Contrastive Learning Module (ECLM), leverages contrastive learning to mitigate dependence on labeled data and improve the quality of pseudo-labels. The second, the Hierarchical Interaction Prompt Module (HIPM), enables comprehensive interaction between visual and textual features and enhances subsequent feature fusion. We evaluate our model on four benchmark datasets: RefCOCO, RefCOCO+, RefCOCOg, and ReferIt. Experiments demonstrate that our model outperforms state-of-the-art models in RIS.

8. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant U24A20591). Supported by High-performance Computing Platform of Chengdu University of Technology.

References

[1] L. Wang, P. Miao, W. Su, X. Li, N. Ji, Y. Jiang, Multimodal referring expression comprehension based on image and text: a review, Journal of Image and Graphics 28 (5) (2023) 1308–1325. doi:10.11834/jig.221024.

[2] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Wang, W. Y. Wang, L. Zhang, Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 6622–6631. doi:10.1109/CVPR.2019.00679.

[3] X. Yang, K. Xu, S. Chen, S. He, B. Y. Yin, R. Lau, Active matting, Advances in Neural Information Processing Systems 31 (2018).

[4] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, e. D. Zitnick, C. Lawrence", T. Pajdla, B. Schiele, T. Tuytelaars, Microsoft COCO: Common objects in context, in: Computer Vision – ECCV 2014, Springer International Publishing, 2014, pp. 740–755. doi:10.1007/978-3-319-10602-1_48.

[5] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, A. Torralba, Semantic understanding of scenes through the ADE20K dataset, International Journal of Computer Vision 127 (2019) 302–321. doi:10.1007/s11263-018-1140-0.

[6] R. Hu, M. Rohrbach, T. Darrell, Segmentation from natural language expressions, in: Computer Vision – ECCV 2016, Springer International Publishing, 2016, pp. 108–124. doi:10.1007/978-3-319-46448-0_7.

[7] S. Huang, T. Hui, S. Liu, G. Li, Y. Wei, J. Han, L. Liu, B. Li, Referring image segmentation via cross-modal progressive comprehension, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10485–10494. doi:10.1109/CVPR42600.2020.01050.

[8] T. Hui, S. Liu, S. Huang, G. Li, S. Yu, F. Zhang, J. Han, Linguistic structure guided context modeling for referring image segmentation, in: Computer Vision – ECCV 2020, Springer International Publishing, 2020, pp. 59–75. doi:10.1007/978-3-030-58607-2_4.

[9] Z. Wang, Y. Lu, Q. Li, X. Tao, Y. Guo, M. Gong, T. Liu, CRIS: CLIP-driven referring image segmentation, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11676–11685. doi:10.1109/CVPR52688.2022.01139.

[10] C. Hazirbas, L. Ma, C. Domokos, D. Cremers, FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture, in: Computer Vision – ACCV 2016, Springer International Publishing, 2017, pp. 213–228. doi:10.1007/978-3-319-54181-5_14.

[11] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, Y. Qiao, CLIP-Adapter: Better vision-language models with feature adapters, International Journal of Computer Vision 132 (2) (2024) 581–595. doi:10.1007/s11263-023-01891-x.

[12] Z. Hu, G. Feng, J. Sun, L. Zhang, H. Lu, Bi-Directional relationship inferring network for referring image segmentation, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4423–4432. doi:10.1109/CVPR42600.2020.00448.

[13] S. Qiu, S. Zhang, T. Ruan, Text-Guided refinement for referring image segmentation, Applied Sciences–Basel 15 (9) (2025) 5047. doi:10.3390/app15095047.

[14] S. Lei, X. Xiao, T. Zhang, H.-C. Li, Z. Shi, Q. Zhu, Exploring fine-grained image-text alignment for referring remote sensing image segmentation, IEEE Transactions on Geoscience and Remote Sensing 63 (2025) 1–11. doi:10.1109/TGRS.2024.3522293.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning, Vol. 139 of Proceedings of Machine Learning Research, PMLR, 2021, pp. 8748–8763.

[16] F. Liu, Y. Liu, Y. Kong, K. Xu, L. Zhang, B. Yin, G. P. Hancke, R. W. H. Lau, Referring image segmentation using text supervision, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22067–22077. doi:10.1109/ICCV51070.2023.02022.

[17] Z. Yang, J. Wang, Y. Tang, K. Chen, H. Zhao, P. H. Torr, LAVT: Language-aware vision transformer for referring image segmentation, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18134–18144. doi:10.1109/CVPR52688.2022.01762.

[18] G. Feng, L. Zhang, Z. Hu, H. Lu, Learning from box annotations for referring image segmentation, IEEE Transactions on Neural Networks and Learning Systems 35 (3) (2024) 3927–3937. doi:10.1109/TNNLS.2022.3201372.

[19] Z. Liang, P. Wang, K. Xu, P. Zhang, R. W. H. Lau, Weakly-supervised salient object detection on light fields, IEEE Transactions on Image Processing 31 (2022) 6295–6305. doi:10.1109/TIP.2022.3207605.

[20] S. Yu, B. Zhang, J. Xiao, E. G. Lim, Structure-consistent weakly supervised salient object detection with local saliency coherence, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Vol. 35, 2021, pp. 1743–1751. doi:10.1609/aaai.v35i4.16434.

[21] S. Gao, W. Zhang, Y. Wang, Q. Guo, C. Zhang, Y. He, W. Zhang, Weakly-supervised salient object detection using point supervision, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Vol. 36, 2022, pp. 1144–1152. doi:10.1609/aaai.v36i1.19947.

[22] N. Araslanov, S. Roth, Single-stage semantic segmentation from image labels, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 4252–4261.

doi:10.1109/CVPR42600.2020.00431.

[23] D. Pathak, E. Shelhamer, J. Long, T. Darrell, Fully convolutional multi-class multiple instance learning, arXiv preprint arXiv:1412.7144 abs/1412.7144 (2014).

[24] X. Tian, K. Xu, X. Yang, B. Yin, R. W. H. Lau, Weakly-supervised salient instance detection, in: Proceedings of the 31st British Machine Vision Conference (BMVC 2020), British Machine Vision Association (BMVA), 2020.

[25] X. Tian, K. Xu, X. Yang, B. Yin, R. W. H. Lau, Learning to detect instance-level salient objects using complementary image labels, International Journal of Computer Vision 130 (3) (2022) 729–746. doi:10.1007/s11263-021-01553-w.

[26] N. Kim, D. Kim, S. Kwak, C. Lan, W. Zeng, ReSTR: Convolution-free referring image segmentation using transformers, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18124–18133. doi:10.1109/CVPR52688.2022.01761.

[27] E. Margffoy-Tuay, J. C. Pérez, E. Botero, P. Arbeláez, Dynamic multimodal instance segmentation guided by natural language queries, in: Computer Vision – ECCV 2018, Springer, 2018, pp. 656–672. doi:10.1007/978-3-030-01252-6_39.

[28] C. Wu, Z. Lin, S. Cohen, T. Bui, S. Maji, PhraseCut: Language-based image segmentation in the wild, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10216–10225. doi:10.1109/CVPR42600.2020.01023.

[29] K. Zhang, F. Wu, G. Zhang, J. Liu, M. Li, BVA-Transformer: Image-text multimodal classification and dialogue model architecture based on Blip and visual attention mechanism, Displays 83 (2024) 102710. doi:10.1016/j.displa.2024.102710.

[30] C. G. M. Snoek, M. Worring, A. W. M. Smeulders, Early versus late fusion in semantic video analysis, in: Proceedings of the 13th Annual ACM International Conference on Multimedia, Association for Computing Machinery, 2005, p. 399–402.

[31] X. Zhao, L. Zhang, Y. Pang, H. Lu, L. Zhang, A single stream network for robust and real-time RGB-D salient object detection, in: Computer Vision – European Conference on Computer Vision (ECCV), Springer International Publishing, 2020, pp. 646–662. doi:10.1007/978-3-030-58542-6_39.

[32] C. Liu, Z. Lin, X. Shen, J. Yang, X. Lu, A. Yuille, Recurrent multimodal interaction for referring image segmentation, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 1271–1280.

[33] D. Lin, G. Chen, D. Cohen-Or, P.-A. Heng, H. Huang, Cascaded feature network for semantic segmentation of RGB-D images, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1320–1328. doi:10.1109/ICCV.2017.147.

[34] A. Owens, A. A. Efros, Audio-Visual scene analysis with self-supervised multisensory features, in: Computer Vision – ECCV 2018, Springer International Publishing, 2018, pp. 639–658. doi:10.1007/978-3-030-01231-1_39.

[35] D.-J. Chen, S. Jia, Y.-C. Lo, H.-T. Chen, T.-L. Liu, See-Through-Text grouping for referring image segmentation, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7453–7462. doi:10.1109/ICCV.2019.00755.

[36] R. Li, K. Li, Y.-C. Kuo, M. Shu, X. Qi, X. Shen, J. Jia, Referring image segmentation via recurrent refinement networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 5745–5753. doi:10.1109/CVPR.2018.00602.

[37] Y.-W. Chen, Y.-H. Tsai, T. Wang, Y.-Y. Lin, M.-H. Yang, Referring expression object segmentation with caption-aware consistency, in: Proceedings of the British Machine Vision Conference (BMVC), 2019.

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

[39] L. Ye, M. Rochan, Z. Liu, Y. Wang, Cross-Modal self-attention network for referring image segmentation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10494–10503. doi:10.1109/CVPR.2019.01075.

[40] S. Liu, T. Hui, S. Huang, Y. Wei, B. Li, G. Li, Cross-Modal progressive comprehension for referring segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (9) (2022) 4761–4775. doi:10.1109/TPAMI.2021.3079993.

[41] G. Feng, Z. Hu, L. Zhang, H. Lu, Encoder fusion network with co-attention embedding for referring image segmentation, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15501–15510. doi:10.1109/CVPR46437.2021.01525.

[42] H. Shi, H. Li, F. Meng, Q. Wu, Key-Word-Aware network for referring expression image segmentation, in: Computer Vision – European Conference on Computer Vision (ECCV) 2018, Springer International Publishing, 2018, pp. 38–54. doi:10.1007/978-3-030-01231-1_3.

[43] G. Luo, Y. Zhou, X. Sun, L. Cao, C. Wu, C. Deng, R. Ji, Multi-task collaborative network for joint referring expression comprehension and segmentation, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10031–10040. doi:10.1109/CVPR42600.2020.01005.

[44] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, T. L. Berg, MAttNet: Modular attention network for referring expression comprehension, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 1307–1315. doi:10.1109/CVPR.2018.00142.

[45] M. Yang, X. Dong, W. Zhang, P. Xie, C. Li, S. Chen, A feature fusion module based on complementary attention for medical image segmentation, Displays 84 (2024) 102811. doi:10.1016/j.displa.2024.102811.

[46] M. Li, M. Ge, Enhanced-Similarity attention fusion for unsupervised cross-modal hashing retrieval, Data Science and Engineering 10 (2) (2025) 258–276. doi:10.1007/s41019-024-00274-7.

[47] H. Ding, C. Liu, S. Wang, X. Jiang, VLT: Vision-language transformer and query generation for referring segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (6) (2023) 7900–7916. doi:10.1109/TPAMI.2022.3217852.

[48] M. Li, L. Sigal, Referring transformer: A one-step approach to multi-task visual grounding, Advances in neural information processing systems 34 (2021) 19652–19664.

[49] C. Zhu, Y. Zhou, Y. Shen, G. Luo, X. Pan, M. Lin, C. Chen, L. Cao, X. Sun, R. Ji, SeqTR: A simple yet universal network for visual grounding, in: Computer Vision – European Conference on Computer Vision (ECCV) 2022, Springer Nature Switzerland, 2022, pp. 598–615. doi:10.1007/978-3-031-19833-5_35.

[50] M. Pu, B. Luo, C. Zhang, L. Xu, F. Xu, M. Kong, Text-Vision relationship alignment for referring image segmentation, Neural Processing Letters 56 (2) (2024) 64. doi:10.1007/s11063-024-11487-2.

[51] F. Liu, Y. Kong, L. Zhang, G. Feng, B. Yin, Local-global coordination with transformers for referring image segmentation, Neurocomputing 522 (2023) 39–52. doi:10.1016/j.neucom.2022.12.018.

[52] Y. Watanabe, R. Togo, K. Maeda, T. Ogawa, M. Haseyama, Text-Guided image manipulation via generative adversarial network with referring image segmentation-based guidance, IEEE Access 11 (2023) 42534–42545. doi:10.1109/ACCESS.2023.3269847.

[53] L. Ji, Y. Du, Y. Dang, W. Gao, H. Zhang, A survey of methods for addressing the challenges of referring image segmentation, Neurocomputing 583 (2024) 127599. doi:10.1016/j.neucom.2024.127599.

[54] S. Sun, P. Wang, H. Peng, Z. Liu, Referring image segmentation with multi-modal feature interaction and alignment based on convolutional nonlinear spiking neural membrane systems, International Journal of Neural Systems 34 (12) (2024) 2450064. doi:10.1142/S0129065724500643.

[55] W. Zhang, Q. Tan, P. Li, Q. Zhang, R. Wang, Cross-modal transformer with language query for referring image segmentation, Neurocomputing 536 (2023) 191–205. doi:10.1016/j.neucom.2023.03.011.

[56] Q. Dai, S. Yang, Curriculum point prompting for weakly-supervised referring image segmentation, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 13711–13722. doi:10.1109/CVPR52733.2024.01301.

[57] Y. Rao, W. Zhao, G. Chen, Y. Tang, Z. Zhu, G. Huang, J. Zhou, J. Lu, DenseCLIP: Language-guided dense prediction with context-aware prompting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18061–18070. doi:10.1109/CVPR52688.2022.01755.

[58] K. Zhou, J. Yang, C. C. Loy, Z. Liu, Learning to prompt for vision-language models, International Journal of Computer Vision 130 (9) (2022) 2337–2348. doi:10.1007/s11263-022-01653-1.

[59] K. Zhou, J. Yang, C. C. Loy, Z. Liu, Conditional prompt learning for vision-language models, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16795–16804. doi:10.1109/CVPR52688.2022.01631.

[60] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2999–3007. doi:10.1109/ICCV.2017.324.

[61] Y. Zhou, H. Chen, J. Xu, Q. Dou, P.-A. Heng, IRNet: Instance relation network for overlapping cervical cell segmentation, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, Springer International Publishing, 2019, pp. 640–648. doi:10.1007/978-3-030-32239-7_71.

[62] J. Qin, J. Wu, X. Xiao, L. Li, X. Wang, Activation modulation and recalibration scheme for weakly supervised semantic segmentation, Proceedings of the AAAI Conference on Artificial Intelligence 36 (2) (2022) 20108–20116. doi:10.1609/aaai.v36i2.20108.

[63] A. Arbelle, S. Doveh, A. Alfassy, J. Shtok, G. Lev, E. Schwartz, H. Kuehne, H. B. Levi, P. Sattigeri, R. Panda, C.-F. Chen, A. Bronstein, K. Saenko, S. Ullman, R. Giryes, R. Feris, L. Karlinsky, Detector-Free weakly supervised grounding by separation, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1781–1792. doi:10.1109/ICCV48922.2021.00182.

[64] T. Shaharabany, Y. Tewel, L. Wolf, What is where by looking: Weakly-supervised open-world phrase-grounding without text inputs, in: Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS 2022), Curran Associates, Inc., 2022, pp. 28222–28237.

[65] Z. Yang, Y. Liu, J. Lin, G. Hancke, R. Lau, Boosting weakly supervised referring image segmentation via progressive comprehension, Advances in Neural Information Processing Systems 37 (2024) 93213–93239.

[66] J. Lee, S. Lee, J. Nam, S. Yu, J. Do, T. Taghavi, Weakly supervised referring image segmentation with intra-chunk and inter-chunk consistency, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21813–21824. doi:10.1109/ICCV51070.2023.01999.

[67] L. Yu, P. Poirson, S. Yang, A. C. Berg, T. L. Berg, Modeling context in referring expressions, in: European conference on computer vision, 2016, pp. 69–85.

[68] J. Mao, J. Huang, A. Toshev, O.-M. Camburu, A. L. Yuille, K. Murphy, Generation and comprehension of unambiguous object descriptions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 11–20. doi:10.1109/CVPR.2016.9.

[69] S. Kazemzadeh, V. Ordonez, M. Matten, T. Berg, ReferItGame: Referring to objects in photographs of natural scenes, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 787–798.

[70] V. K. Nagaraja, V. I. Morariu, L. S. Davis, Modeling context between objects for referring expression understanding, in: Computer Vision – European Conference on Computer Vision (ECCV) 2016, Springer International Publishing, 2016, pp. 792–807. doi:10.1007/978-3-319-46493-0_48.

[71] H. Akbari, S. Karaman, S. Bhargava, B. Chen, C. Vondrick, S.-F. Chang, Multi-Level multimodal common semantic space for image-phrase grounding, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 12468–12478. doi:10.1109/CVPR.2019.01276.

[72] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, L. Zhang, Grounded SAM: Assembling open-world models for diverse visual tasks, arXiv preprint arXiv:2401.14159 (2024).