SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

November 2026, Vol. 69, Iss. 11, 212103:1–212103:22
https://doi.org/10.1007/s11432-025-5107-7

Cross-modal bidirectional interaction model for referring remote sensing image segmentation

Zhe DONG[1,2], Yuzhe SUN[1,2], Tianzhu LIU[1,2] & Yanfeng GU[1,2*]

[1]*School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China*
[2]*Heilongjiang Province Key Laboratory of Space-Air-Ground Integrated Intelligent Remote Sensing, Harbin Institute of Technology, Harbin 150001, China*

Received 3 June 2025/Revised 15 October 2025/Accepted 14 December 2025/Published online 21 October 2026

Abstract Referring remote sensing image segmentation (RRSIS) aims to generate pixel-level masks of target objects in remote sensing images based on natural language expressions. This task is inherently challenging due to the complex geospatial relationships and varying object scales in remote sensing scenarios. To address the aforementioned challenges, a novel RRSIS framework is proposed, termed the cross-modal bidirectional interaction model (CroBIM). Specifically, a context-aware prompt modulation (CAPM) module is designed to integrate spatial positional relationships and task-specific knowledge into the linguistic features, thereby enhancing the ability to capture the target object. Additionally, a language-guided feature aggregation (LGFA) module is introduced to incorporate linguistic cues into multi-scale visual features, with an attention deficit compensation mechanism being applied to ensure robust alignment across object scales. Finally, a mutual-interaction decoder (MID) is developed to align vision and language features through iterative bidirectional refinement, enabling precise segmentation even in complex geospatial contexts. Moreover, we also construct RISBench, a new large-scale benchmark dataset comprising 52472 image-language-label triplets. Extensive benchmarking on RISBench and two other prevalent datasets demonstrates the superior performance of the proposed CroBIM over existing state-of-the-art (SOTA) methods. The source code for CroBIM and the RISBench dataset will be publicly available at https://github.com/HIT-SIRS/CroBIM.

Keywords vision and language, referring remote sensing image segmentation (RRSIS), cross-modal reasoning, benchmark dataset, aerial imagery

Citation Dong Z, Sun Y Z, Liu T Z, et al. Cross-modal bidirectional interaction model for referring remote sensing image segmentation. Sci China Inf Sci, 2026, 69(11): 212103, https://doi.org/10.1007/s11432-025-5107-7

1 Introduction

Deep learning has become a cornerstone for remote sensing applications [1, 2], enabling tasks such as semantic segmentation [3], object detection [4], and change detection [5]. However, these methods often focus on visual comprehension, neglecting the modeling of object relationships and deeper semantic understanding.

Recent advancements in large language models (LLMs) have spurred research into vision-language models (VLMs), enabling tasks such as image captioning [6, 7], image-text retrieval [8, 9], text-based remote sensing image generation [10,11], and visual question answering [12,13]. However, the task of referring remote sensing image segmentation (RRSIS) remains relatively unexplored.

As illustrated in Figure 1, given remote sensing images and language expressions, RRSIS aims to provide pixel-level masks for specific regions or objects based on the content of the images and expressions. Its core principle is to achieve precise object localization and segmentation by matching textual descriptions with image content. RRSIS breaks the boundaries of traditional semantic understanding of remote sensing data, enabling non-expert users to retrieve objects in remote sensing images through human-computer interaction. It has broad application prospects in land use analysis [14], search and rescue operations [15], environmental monitoring [16], military intelligence generation [17], agricultural production [18], and urban planning [19].

Although referring image segmentation in natural scenarios has made some progress, research on RRSIS is still in its infancy. Yuan et al. [20] first introduced the concept of the RRSIS task and proposed a language-guided cross-scale enhancement (LGCE) module based on the language-aware vision Transformer (LAVT) [21] to improve segmentation performance for small and sparsely distributed objects. Furthermore, the rotated multi-scale interaction network (RMSIN) [22] is designed to address the prevalent challenges of complex scales and orientations

* Corresponding author (email: guyf@hit.edu.cn)

© Science China Press 2026

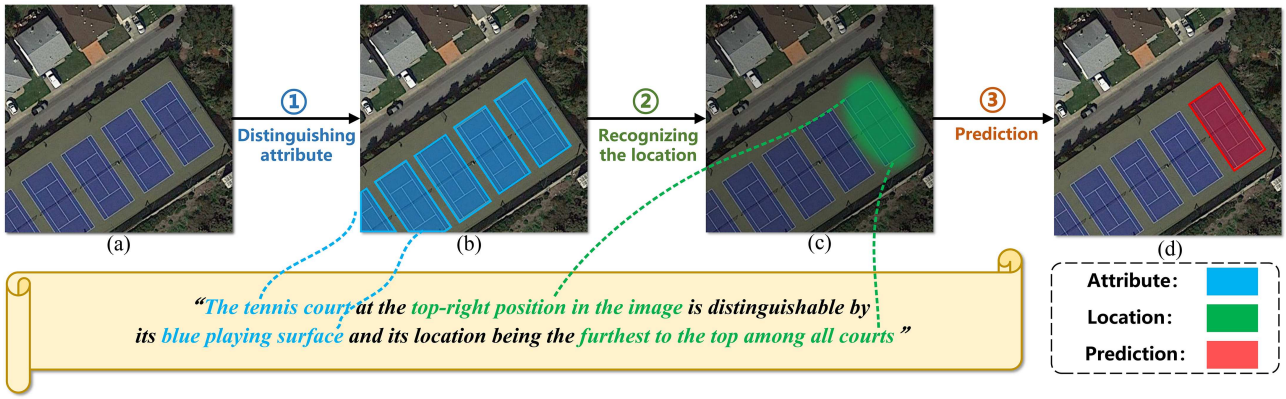


Figure 1 (Color online) Illustration of the RRSIS task. (a) The input consists of a referring expression and an image. (b) The model first identifies all candidate objects described in the expression based on information such as category, color, and shape (e.g., 'tennis court' and 'blue playing surface'). (c) After identifying all potential candidate objects that match the input expression, additional information such as position and size (e.g., 'top-right position', 'furthest to the top among all courts') is utilized to highlight the target object. (d) Through relation-aware reasoning, the final segmentation mask of the predicted object is obtained.

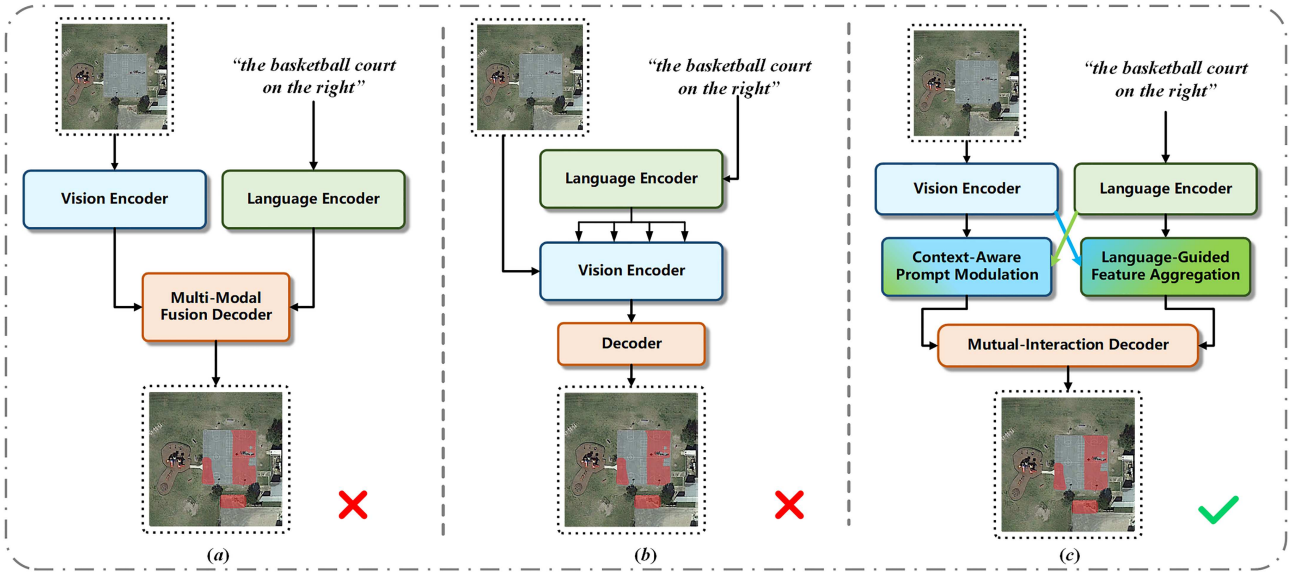


Figure 2 (Color online) Conceptual comparison of RRSIS frameworks. (a) Cross-modal feature fusion during decoding; (b) directly integrating linguistic information into visual features; (c) our CroBIM model.

in RRSIS. To manage cross-scale fine-grained information, the intra-scale interaction module (IIM) and cross-scale interaction module (CIM) are developed. Additionally, adaptive rotated convolution (ARC) is introduced to enhance the model's robustness to rotational variations. The aforementioned methods rely solely on jointly embedding linguistic features during visual encoding to perceive relevant linguistic context at each spatial location. Although these approaches have achieved satisfactory performance, the interrelation and alignment of visual and linguistic features across multiple levels of the encoding process have not yet been thoroughly explored.

Specifically, as shown in Figures 2(a) and (b), failing to consider the underlying correlations between linguistic and visual information and merely fusing cross-modal heterogeneous features at different stages can lead to attention drift, resulting in a mismatch between visual features and the regions described in the query expression. Moreover, compared to natural scene images, the diversity of remote sensing data and the complex geospatial relationships embedded in the corresponding expressions present significant challenges for accurately locating and segmenting the target regions.

In this paper, we introduce a novel cross-modal bidirectional interaction model (CroBIM) for the RRSIS task, addressing the previously identified challenges. As depicted in Figure 2(c), the essence of CroBIM lies in its capability to facilitate bidirectional interaction and correlation between visual and linguistic features throughout both the encoding and decoding phases. This enables precise visual-linguistic alignment during the prediction stage. Specifically, the context-aware prompt modulation (CAPM) module is introduced to enhance the text

feature encoding process by incorporating multi-scale visual contextual information via learnable prompts. This integration enables the model to effectively perceive the spatial structure and relative positioning of target objects described in the referring expressions. Additionally, we propose a language-guided feature aggregation (LGFA) module that fosters interaction between multi-scale visual representations and linguistic features, thereby capturing cross-scale dependencies and addressing complex scale variations. To further enhance feature aggregation, the LGFA incorporates an attention deficit compensation mechanism. Finally, we design a mutual-interaction decoder (MID) to achieve precise vision-language alignment via cascaded bidirectional cross-attention, ultimately generating highly accurate segmentation masks.

To further advance research in RRSIS, we construct a new dataset called RISBench, with images sourced from the DOTA-v2 [23] and DIOR [24] remote sensing object detection datasets. RISBench consists of 52472 image-language-label triplets. The language expressions provide not only basic category information but also details on color, shape, location, size, relative position, and relative size, with an average length of 14.31 words. Additionally, we employed a semi-automated approach to generate pixel-level mask annotations using bounding box prompts from the VRSBench dataset [25] and a customized segment anything model (SAM) [26]. Compared to the RefSegRS and RRSIS-D datasets, our RISBench offers a greater number of triplets, a wider range of spatial resolutions, and a richer diversity of objects within each category.

In summary, the contributions of this work can be summarized in the following four aspects.

(1) We propose CroBIM, a novel framework specifically designed to address the unique challenges of RRSIS, including the diversity of remote sensing data and the complexity of geospatial relationships described in natural language.

(2) We design three task-specific modules to address the core challenges of RRSIS: the CAPM integrates multi-scale visual context with linguistic features using learnable prompts, the LGFA captures cross-scale dependencies to enhance feature aggregation, and the MID ensures precise vision-language alignment through cascaded bidirectional cross-attention, collectively enabling accurate segmentation mask predictions.

(3) We construct RISBench, a large-scale benchmark consisting of 52472 image-language-label triplets with diverse spatial resolutions (0.1–30 m), 26 object categories, and detailed 8-attribute referring expressions, significantly advancing RRSIS research.

(4) Comprehensive experiments on three benchmark datasets demonstrate that CroBIM achieves state-of-the-art (SOTA) performance with significant improvements in both accuracy and robustness.

The remainder of this paper is structured as follows. Section 2 reviews related work on RRSIS. Section 3 details the construction process of our proposed RISBench dataset and provides an analysis of its key characteristics. In Section 4, we describe the proposed methodology in detail. Section 5 presents a comprehensive set of experiments and in-depth analyses. Section 6 conducts ablation studies to verify the effectiveness of each component. Finally, Section 7 concludes the paper and offers insights into potential future research directions.

2 Related work

2.1 Referring image segmentation

Compared to other multimodal tasks, referring image segmentation is more challenging as it requires effective coordination and reasoning between language and vision to accurately segment the target regions in an image. Multimodal fusion, diversity of expression, and robustness are three critical challenges that need to be addressed in the current state of referring image segmentation tasks [27].

Hu et al. [28] proposed a CNN-LSTM framework for referring image segmentation, effectively extracting visual and linguistic features for precise segmentation. A recurrent refinement network (RRN) [29] was proposed to capture multi-scale semantics in image representations. The RRN iteratively optimized the initial mask using a recursive optimization module to achieve a high-quality pixel segmentation mask.

However, these methods often neglect the modality gap and fail to fully capture the interaction between language and images. Recent studies have introduced attention mechanisms to address these limitations. A cross-modal self-attention (CMSA) module by Ye et al. [30] was proposed to effectively capture long-range dependencies between language and visual features. A cascade-grouped attention network (CGAN) [31] is designed, consisting of cascade-grouped attention (CGA) and instance-level attention loss (ILA). By performing hierarchical reasoning on images and effectively distinguishing different instances, CGAN enhances the correlation between text and images. Besides, Hu et al. [32] introduced a bidirectional relationship inferring network (BRINet) to model cross-modal information dependencies. BRINet utilized a visual-guided language attention module to filter out irrelevant regions and enhance

semantic matching between target objects and expressions.

2.2 Visual grounding for remote sensing data

Similar to RRSIS, visual grounding for remote sensing data (RSVG) specifically entails using a remote sensing image alongside an associated query expression to determine the bounding box for a target object of interest. By localizing objects in remote sensing scenes through natural language guidance, RSVG provides object-level understanding and enhances accessibility. Compared to query expressions in natural images, expressions in RSVG frequently encompass complex geospatial relationships, and the objects of interest are often not visually prominent.

GeoVG [33] is the first RSVG framework, which utilizes a language encoder to learn spatial relationships in geographic space, an image encoder to adaptively attend to remote sensing scenes, and a fusion module to integrate textual and visual features for visual localization. Zhan et al. [34] proposed a large-scale benchmark dataset DIOR-RSVG and designed a Transformer-based multigranularity visual language fusion (MGVLF) module, which addresses the challenges of large-scale variations and cluttered backgrounds in remotely sensed images. By leveraging multiscale visual features and multigranularity textual embeddings, more discriminative representations are learned. Besides, the language-guided progressive visual attention framework (LPVA) [35] utilized a progressive attention module to adjust visual features at multiple scales and levels, enabling the visual backbone to focus on expression-related features. Additionally, a multi-level feature enhancement decoder aggregated visual contextual information, enhancing feature distinctiveness and suppressing irrelevant regions.

2.3 Referring remote sensing image segmentation

Referring image segmentation in the context of remote sensing data has emerged as a novel area of investigation in recent times. Studies pertaining to this specific task are currently in an ascent stage and remain relatively scarce. Yuan et al. [20] first introduced the RRSIS task in the remote sensing domain. To facilitate research on RRSIS, they constructed a benchmark dataset RefSegRS by designing various referring expressions and automatically generating corresponding masks. Specifically, the RefSegRS dataset consists of 4420 image-language-label triplets. Furthermore, to address the challenge of segmenting small and scattered objects in remote sensing images, they devised an LGCE module based on the LAVT [21]. The LGCE module leveraged linguistic features as guidance to improve the segmentation of small objects by integrating deep and shallow features, thereby enhancing the complexity and diversity of the approach. In addition, to address the spatial variations and rotational diversity of targets in aerial images, the RMSIN [22] was proposed. RMSIN introduced the intra-scale interaction module (IIM) and CIM within the LAVT framework, enabling the extraction of detailed features and facilitating comprehensive feature fusion. Moreover, to effectively handle the intricate rotational variations of objects, the decoder of RMSIN integrated the adaptive rotated convolution (ARC). This integration enhances the network's capability to capture and represent complex object rotations, thereby improving the overall performance on the RRSIS task.

Recently, foundation models have shown promising potential in advancing RRSIS tasks. Dong et al. [36] proposed DiffRIS, which harnesses the semantic understanding capabilities of pre-trained text-to-image diffusion models for enhanced cross-modal alignment. DiffRIS introduces a context perception adapter (CP-adapter) that dynamically refines linguistic features through global context modeling and object-aware reasoning, along with a progressive cross-modal reasoning decoder (PCMRD) that iteratively aligns textual descriptions with visual regions for precise segmentation. By bridging the domain gap between general vision-language understanding and remote sensing applications, DiffRIS achieves significant performance improvements across multiple benchmarks. Similarly, Chen et al. [37] introduced RSRefSeg, a referring remote sensing image segmentation foundation model that leverages CLIP for visual and textual encoding. RSRefSeg employs both global and local textual semantics as filters to generate referring-related visual activation features in the latent space, which then serve as input prompts for the SAM to refine segmentation masks. These foundation model-based approaches demonstrate the effectiveness of leveraging large-scale pre-trained models in enhancing multimodal task comprehension for remote sensing applications.

3 Dataset construction and analysis

In this section, we will introduce the construction procedure and statistical analysis of our proposed RISBench dataset in Subsections 3.1 and 3.2.

Table 1 Comparative analysis of RISBench and prior RRSIS datasets.

Dataset	Triplet count	Image size	Spatial resolution	Expression attribute
RefSegRS [20]	4420	512×512	0.13 m	3
RRSIS-D [22]	17402	800×800	0.5–30 m	7
RISBench	52472	512×512	0.1–30 m	8

3.1 RISBench dataset construction

Motivated by SAM [26] and RMSIN [22], we combine bounding box prompts with SAM to generate pixel-level masks using a semi-automatic method, significantly reducing the cost of manual annotation. The steps for generating fine-grained pixel-level annotations for the RRSIS task are as follows.

Step 1. We collected remote sensing images, referring textual descriptions, and corresponding visual grounding boxes from the VRSBench dataset [25]. However, the bounding boxes often exhibited inaccuracies, such as misalignments and inappropriate sizing (either too large or too small). Additionally, there exists a significant domain gap between natural and remote sensing scenes, which exacerbates the problem. Consequently, directly applying these bounding boxes to the SAM model results in unsatisfactory segmentation masks, necessitating further refinement and optimization.

Step 2. To ensure the accuracy and reliability of the generated mask annotations, we employed the PA-SAM method [38] to optimize the SAM model. Specifically, we trained a specialized adapter using a subset of remote sensing images with manually corrected masks, enabling SAM to better handle the unique characteristics of remote sensing imagery, such as complex geospatial relationships and varying object scales. The fine-tuning process involved using bounding box prompts as input, optimizing the mask decoder features with a combination of dice loss and cross-entropy loss, and iteratively refining the model through multiple rounds of training and validation. After generating the initial masks, we conducted a meticulous human verification process, where expert annotators manually reviewed and corrected the masks. This process included initial inspections, detailed corrections using specialized tools, and a consensus process for uncertain cases, ensuring high-quality and objective annotations.

Step 3. To enhance the segmentation masks, we employ a meticulous human verification process. Expert annotators manually review and correct the masks generated by the optimized SAM model. Initially, annotators identify inaccuracies or misalignments based on their expertise and predefined criteria. Detailed inspections are then conducted on flagged masks, where annotators assess boundary precision and object shapes by zooming in on specific image regions. Using specialized tools, they refine contours, correct misalignments, resize segments, and resolve boundary ambiguities. When initial corrections are uncertain, a consensus process is initiated, involving independent reviews by multiple annotators and final decisions through majority agreement or expert discussion. This human-in-the-loop approach ensures high-accuracy mask annotations, bridging the gap between automated segmentation and the nuanced understanding required for remote sensing imagery.

While potential biases may exist in the dataset due to factors such as annotator subjectivity, dataset characteristics, and linguistic variations, we have taken several steps to minimize their impact. During the annotation process, we employed a consensus-based verification approach, where multiple expert annotators independently reviewed and corrected the masks, ensuring objectivity and accuracy. Additionally, the fine-tuning of the SAM model and the use of diverse remote sensing images helped reduce biases related to object scales and spatial relationships. These measures ensure that the dataset is robust and representative, enhancing the model's generalization ability to other datasets and real-world scenarios.

3.2 Dataset statistics

After meticulously filtering out uninformative image-language label triplets, we curated the RISBench dataset, comprising 52472 high-quality image-language label triplets. This dataset is partitioned into a training set with 26300 triplets, a validation set with 10013 triplets and a test set with 16159 triplets, ensuring robust model development and evaluation. Each image in RISBench is uniformly sized at 512×512 pixels, maintaining consistency across the dataset. The spatial resolution of the images spans from 0.1 to 30 m, encompassing a diverse range of scales and details. The semantic labels are categorized into 26 distinct classes, each annotated with 8 attributes, thereby facilitating a comprehensive and nuanced semantic segmentation analysis. As shown in Table 1, compared to the previous RRSIS datasets, our dataset demonstrates significant improvements in both quantity and diversity.

Additionally, the distribution of categories and object sizes is illustrated in Figure 3(b). Moreover, the referring expressions in our dataset have an average length of 14.31 words, and the vocabulary size encompasses 4431 unique words, underscoring the richness and complexity of the language component. The distribution of word lengths

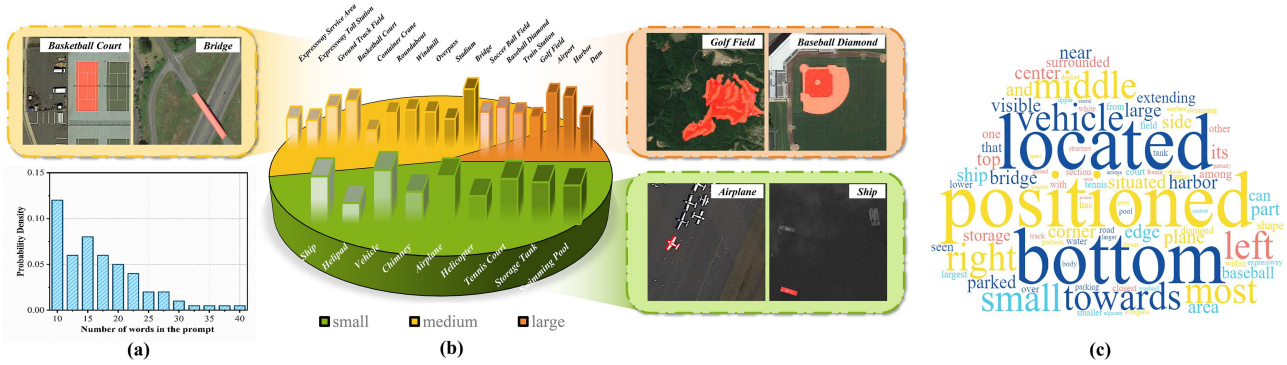


Figure 3 (Color online) Statistical analysis of the constructed RISBench dataset. (a) Distribution of the word length of referring expressions; (b) distribution of the object categories and object size; (c) word cloud for top 50 words within the referring sentences in our RISBench dataset.

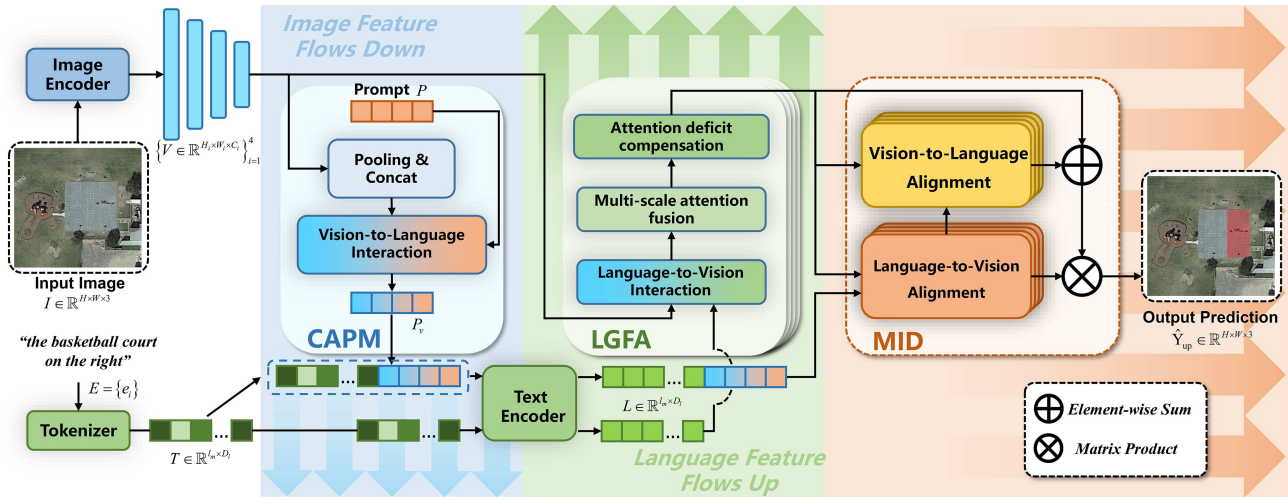


Figure 4 (Color online) Overview of our proposed CroBIM framework, which comprises five key components: an image encoder, a text encoder, a CAPM module, an LGFA module, and an MID.

within these expressions is depicted in Figure 3(a), providing further insight into the linguistic characteristics of the dataset. Figure 3(c) illustrates the word cloud representation of the RISBench dataset.

4 Methodology

The overall architecture of our proposed framework is shown in Figure 4. Our CroBIM framework processes an input image $I \in \mathbb{R}^{H\times W\times 3}$ and a language expression $E = \{e_i\}, i \in \{0,\ldots,N\}$, where H and W denote the height and width of the input image, respectively, and N represents the length of the referring expression. For the image encoder, we employ either Swin Transformer [39] or ConvNeXt [40], which extracts multi-scale visual features $\left\{V \in \mathbb{R}^{H_i\times W_i\times C_i}\right\}_{i=1}^{4}$ from the input image. Here, $(H_i, W_i) = \left(H/2^{i+1}, W/2^{i+1}\right)$ and C_i denote the spatial resolution and channel dimension of the i-th visual feature, respectively. Additionally, we employ BERT [41] as the text encoder to process the referring description by tokenizing and padding it, which generates text tokens $T \in \mathbb{R}^{l_m\times D_l}$. These tokens are then input into the BERT encoder to derive linguistic features $L \in \mathbb{R}^{l_m\times D_l}$. In this context, l_m denotes the maximum token length, while D_l represents the dimension of the linguistic features.

The details of each part will be introduced in the following sections.

4.1 Context-aware prompt modulation

Prompt learning enhances model adaptability to specific tasks by introducing learnable parameters. However, free-form text prompts often lack sufficient contextual information, leading to suboptimal quality of learned representations. To address this issue, we propose a CAPM module. The CAPM module integrates multi-scale visual contextual information during the text encoding process, aiding the model in better perceiving the spatial struc-

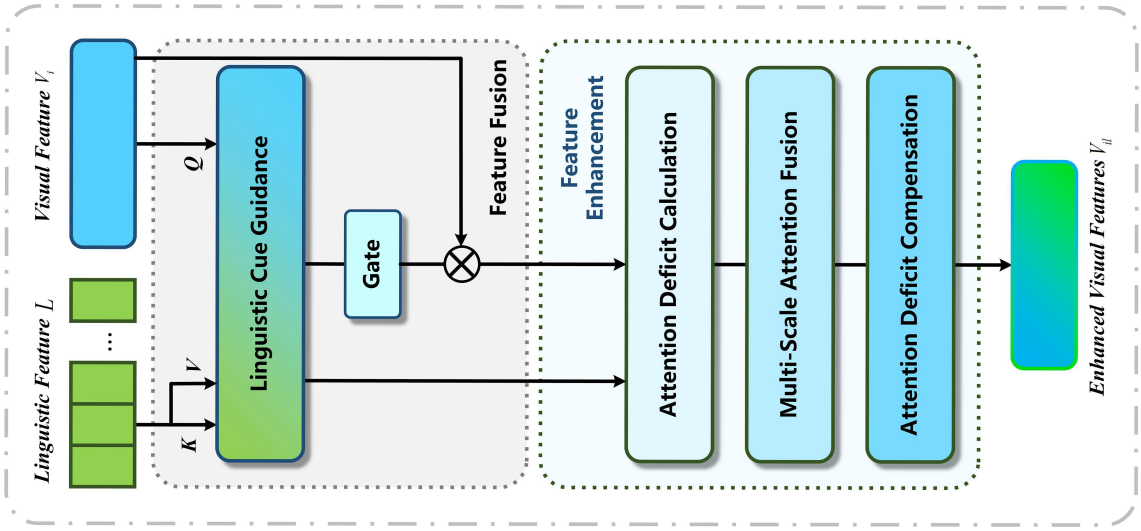


Figure 5 (Color online) An illustration of our proposed LGFA module. After performing cross-modal feature fusion, attention deficit compensation is introduced to refine multi-scale visual features using textual constraints, thereby ensuring that attention across different stages is focused on the same target region.

ture and relative positioning of target objects, thereby improving its ability to capture and identify these objects effectively.

Given the multi-scale visual features $\{V_i\}_{i=1}^4$ produced by the image encoder, we first apply adaptive average pooling to extract cross-scale contextual information. Subsequently, the pooled features from each scale are concatenated and flattened to form a multi-scale context embedding V_e:

$$V_e = \text{Flatten}\left(\text{Concat}\left(\{\text{Pool}_{s\times s}\left(V_i\right)\}_{i=1}^4\right)\right) \in \mathbb{R}^{4s^2\times C_{\text{total}}}, \tag{1}$$

where $\text{Concat}(\cdot)$ denotes the channel concatenation operation, $\text{Flatten}(\cdot)$ converts a multidimensional tensor into a one-dimensional vector, and $\text{Pool}_{s\times s}$ represents adaptive average pooling with an output size of $s \times s$. Here, $C_{\text{total}} = \sum_{i=1}^4 C_i$, and s is set to 1 in this work.

Furthermore, learnable textual prompts $P \in \mathbb{R}^{N_p\times D_l}$ are introduced as supplementary inputs to guide the model in incorporating domain-specific knowledge pertinent to the RRSIS task into the learning process, where N_p is set to 4. The learnable textual prompts P are initialized using Xavier uniform initialization to ensure stable training dynamics. Specifically, the prompt parameters are drawn from a uniform distribution $\mathcal{U}(-a, a)$, where $a = \sqrt{6/(N_p + D_l)}$. With $N_p = 4$ and $D_l = 768$, this results in initialization values sampled from $\mathcal{U}(-0.0881, 0.0881)$. This initialization scheme maintains consistent activation variance across network layers and facilitates effective gradient flow during backpropagation, thereby enabling the prompts to efficiently incorporate multi-scale visual context into the text encoding process. This enhancement aims to improve the model's capability to comprehend and generate responses relevant to the current task. To achieve image-to-text cross-modal interaction, we introduce a cross-attention mechanism to integrate multi-scale context into the learnable textual prompts P with V_e:

$$P_v = \text{CrossAttn}(V_e, P) = \text{Softmax}\left(P\omega_q\left(V_e\omega_k\right)^{\text{T}}\right)V_e\omega_v, \tag{2}$$

where $\omega_q, \omega_k, \omega_v$ are the projection matrices, and P_v represents the context-aware textual prompts.

The context-aware textual prompts P_v are then concatenated with the text tokens T and jointly input into the BERT encoder to obtain linguistic features $L_v \in \mathbb{R}^{(l_m+N_p)\times D_l}$ that incorporate visual context.

4.2 Language-guided feature aggregation

To forge a robust visual-linguistic synergy and seamlessly integrate dependable linguistic features of referred objects into multi-scale visual representations, we introduce a sophisticated LGFA module. As shown in Figure 5, the LGFA module adeptly captures and models the intricate interdependencies between visual and linguistic modalities.

Initially, the visual feature $V_i \in \mathbb{R}^{H_i\times W_i\times C_i}$ undergoes processing through a projection function ω_{iq}. Subsequently, it is spatially expanded, as delineated by

$$V_{iq} = \text{Flatten}(\omega_{iq}(V_i)), \tag{3}$$

where ω_{iq} denotes the projection function for the visual feature and $V_{iq} \in \mathbb{R}^{l_m \times (H_i \times W_i)}$ represents the projected and spatially expanded visual feature.

Subsequently, the linguistic feature $L \in \mathbb{R}^{l_m \times D_l}$ undergoes transformation through projection functions ω_{ik} and ω_{iv}:

$$L_{ik}, L_{iv} = \omega_{ik}(L), \omega_{iv}(L), \tag{4}$$

where ω_{ik} and ω_{iv} function as projection mechanisms for the linguistic features, with L_{ik} and L_{iv} representing the linguistic features utilized for computing the attention scores and generating the final attention output, respectively.

Following this, we compute the attention score matrix S_i between the visual and linguistic features:

$$S_i = V_{iq}^{\mathrm{T}} L_{ik} \in \mathbb{R}^{(H_i \times W_i) \times D_l}. \tag{5}$$

Afterward, the attention score matrix S_i is normalized using the Softmax function, multiplied by L_{iv}^{T}, and finally, gated cross-modal activation is obtained by applying a Gate operation subsequent to the unflatten operation:

$$\mathrm{Att}_i = \mathrm{Gate}\left(\mathrm{Unflatten}\left(\mathrm{Softmax}\left(\frac{S_i}{\sqrt{l_m}}\right) L_{iv}^{\mathrm{T}}\right)\right), \tag{6}$$

where $\mathrm{Gate}(\cdot)$ represents the application of a 1×1 convolution followed by a GELU activation function, while $\mathrm{Unflatten}(\cdot)$ indicates the inverse operation of $\mathrm{Flatten}(\cdot)$.

Finally, the input visual feature V_i is reweighted by integrating the attention weights Att_i, resulting in the integrated cross-modal feature map $V_{li} \in \mathbb{R}^{H_i \times W_i \times C_i}$:

$$V_{li} = \mathrm{Conv}_{1\times1}\left(\mathrm{Att}_i\right) \odot V_i, \tag{7}$$

where $\odot$ denotes element-wise matrix multiplication, and $\mathrm{Conv}_{1\times1}$ represents the 1×1 convolution function.

To align the cross-modal correlations between adjacent stages and refine the aggregated multi-scale features, we resample the attention maps $\{S_i\}_{i=1}^{4}$ of different scales to a uniform size (H_4, W_4):

$$S_i' = \mathcal{I}(S_i, (H_4, W_4)), i \in \{1, 2, 3, 4\}, \tag{8}$$

where $\mathcal{I}$ denotes the resampling operation.

Subsequently, we introduce an attention deficit compensation mechanism that identifies regions where attention diverges across different scales and explicitly enhances these areas by integrating multi-scale visual features guided by linguistic cues. Specifically, we calculate the attention deficit map $\mathcal{M} \in \mathbb{R}^{H_4 \times W_4}$ between cross-scale attentions and select the top K regions with the highest correlation differences:

$$\mathcal{M} = \sum_{i=1}^{3} \left| S_i' - S_{i+1}' \right|, \tag{9}$$

$$\{r_k\}_{k=1}^{K} \leftarrow \mathrm{TopK}\left(\mathcal{M}, K\right). \tag{10}$$

For each attention deficit region r_k, the multi-scale features $\{V_{li}^{r_k}\}_{i=1}^{4}$ corresponding to that region are projected to a uniform channel dimension $C_{\hat{v}}$ and concatenated to yield the cross-scale feature representation $V_l^{r_k}$:

$$V_l^{r_k} = \mathrm{Concat}\left(\overrightarrow{\mathrm{Proj}}\left(\mathcal{I}\left[V_{l1}^{r_k}, V_{l2}^{r_k}, V_{l3}^{r_k}, V_{l4}^{r_k}\right]\right)\right), \tag{11}$$

where $\mathrm{Concat}(\cdot)$ denotes the concatenation operation, and $\overrightarrow{\mathrm{Proj}}(\cdot)$ indicates the channel projection layer.

We then proceed to characterize cross-scale dependencies with the following steps:

$$\widetilde{V}_l^{r_k} = \mathrm{MSA}\left(\mathrm{LN}\left(V_l^{r_k}\right)\right) + V_l^{r_k}, \tag{12}$$

where $\mathrm{LN}(\cdot)$ denotes the layer normalization operator, and $\mathrm{MSA}(\cdot)$ signifies the multi-head self-attention mechanism.

Following this, the enhanced sequence is meticulously restructured back into its original patch configuration, precisely adhering to the initial order of concatenation, ensuring a seamless and coherent transformation:

$$\left[\widetilde{V}_{l1}^{r_k}, \widetilde{V}_{l2}^{r_k}, \widetilde{V}_{l3}^{r_k}, \widetilde{V}_{l4}^{r_k}\right] = \mathcal{I}'\left(\mathrm{Split}\left(\overleftarrow{\mathrm{Proj}}(\widetilde{V}_l^{r_k})\right)\right), \tag{13}$$

where $\overleftarrow{\mathrm{Proj}}(\cdot)$ and $\mathcal{I}'$ represent the inverse operations of $\overrightarrow{\mathrm{Proj}}(\cdot)$ and $\mathcal{I}$, and $\mathrm{Split}(\cdot)$ represents the channel separation operation.

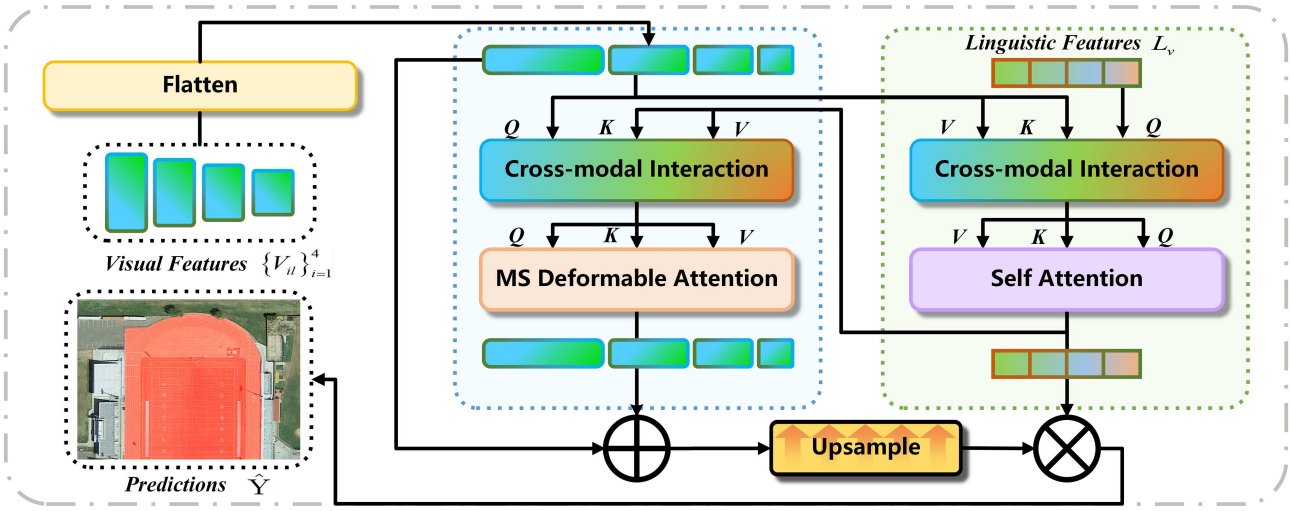


Figure 6 (Color online) Illustration of the MID, which incorporates visual context into language features via vision-to-language alignment, followed by aligning these enriched language features with individual visual pixels through language-to-vision alignment. The aligned visual-linguistic features are then used to accurately segment the target objects.

4.3 Mutual-interaction decoder

Integrating the complementary information between cross-modal features is a fundamental challenge in the RRSIS task. Cross-modal features often encompass inconsistent information, and without considering the intermediate interactions and thorough alignment between different modalities, it is impossible to ensure the discriminative power of the learned representations. To address the aforementioned challenges, we have meticulously designed the MID, as illustrated in Figure 6, aiming to achieve more comprehensive vision-language alignment and precise pixel prediction.

The MID utilizes visual features $\{V_{li}\}_{i=1}^{4}$ and linguistic features L_v as inputs to predict the mask of the referred object. Prior to executing vision-language alignment, a series of operations are performed to harmonize the dimensions of the visual and linguistic features.

$$V_{\text{ms}} = \text{Flatten}\left(\text{Concat}\left(\text{Proj}_v\left(\{V_{li}\}_{i=1}^{4}\right)\right)\right) \in \mathbb{R}^{N\times D}, \tag{14}$$

where Proj_v projects the multi-scale visual features into a hidden dimension $D = 256$, and $N = \sum_{i=1}^{4} H_i W_i$.

Specifically, given the linguistic features L_v and visual features V_{ms}, the language-to-vision interaction is facilitated through cross-attention, self-attention, and feed-forward networks (FFN). These mechanisms are employed to update the linguistic features L_v using V_{ms}. Each layer of cross-attention, self-attention, and FFN is followed by a residual connection and layer normalization, ensuring a coherent and stable transformation of the features.

$$\hat{L}_v = \text{FFN}(\text{SelfAttn}(\text{CrossAttn}(L_v, V_{\text{ms}}))). \tag{15}$$

Subsequently, the refined linguistic features $\hat{L}_v$ are meticulously aligned with the visual features V_{ms} on a pixel-by-pixel basis through a sophisticated vision-to-language interaction mechanism:

$$\hat{V}_{\text{ms}} = \text{FFN}(\text{MSDeformAttn}(\text{CrossAttn}(V_{\text{ms}}, \hat{L}_v))), \tag{16}$$

where $\text{MSDeformAttn}(\cdot)$ denotes the multiscale deformable attention [42].

Upon executing mutual interactions through bidirectional cross-modal feature alignment, we derive the harmonized linguistic and visual features, denoted as $\hat{L}_v$ and $\hat{V}_{\text{ms}}$.

Finally, the enhanced visual features $\hat{V}_{\text{ms}}$ are combined with the original visual features V_{ms} and subsequently mapped through a 1×1 convolutional layer followed by spatial resampling to obtain the mask embedding $V_{\text{out}} \in \mathbb{R}^{H_1\times W_1\times D}$. V_{out} is then element-wise multiplied with the [CLS] token $L_{\text{out}} \in \mathbb{R}^{D}$ of the linguistic features $\hat{L}_v$, resulting in the final prediction mask $\hat{Y} \in \mathbb{R}^{H_1\times W_1}$.

$$\hat{Y}^{(i,j)} = V_{\text{out}}^{(i,j)} \cdot L_{\text{out}}, \tag{17}$$

where (i, j) represents the pixel position in a two-dimensional space. $\hat{Y}$ is then upsampled to the same spatial resolution (H, W) as the input image via bilinear interpolation.

Unlike existing methods, which typically rely on unidirectional or static interactions that may lead to attention drift or misalignment between modalities, our proposed MID introduces a bidirectional refinement process. This process enables continuous and iterative alignment between visual and linguistic modalities throughout the decoding stage. Such bidirectional interaction ensures that both modalities are dynamically optimized, resulting in more precise feature integration and robust segmentation performance, particularly in complex remote sensing scenarios characterized by diverse geospatial relationships.

4.4 Training objective

In the RRSIS task, object mask prediction is typically framed as a pixel-wise binary classification problem. Due to the significant class imbalance in remote sensing images, where target pixels are relatively scarce compared to background pixels, a conventional cross-entropy loss function may lead to a model that prioritizes learning from background pixels, adversely affecting the performance in detecting target regions. To address this issue, we employ a combined loss function consisting of cross-entropy loss and dice loss [43] as our training objective:

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{\text{cross-entropy}}(\hat{Y}_{\text{up}}, Y) + (1-\lambda) \cdot \mathcal{L}_{\text{dice}}(\hat{Y}_{\text{up}}, Y), \tag{18}$$

where λ is a hyperparameter to balance the two loss functions and is set to 0.9 in the paper, $\hat{Y}_{\text{up}}, Y \in \mathbb{R}^{H\times W}$ represent the upsampled prediction $\hat{Y}$ and the ground truth, respectively.

The cross-entropy loss is computed as

$$\mathcal{L}_{\text{cross-entropy}}\big(\hat{Y}_{\text{up}}, Y\big) = -\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\Big[Y^{(i,j)}\log\big(\hat{Y}_{\text{up}}^{(i,j)}\big) + \big(1-Y^{(i,j)}\big)\log\big(1-\hat{Y}_{\text{up}}^{(i,j)}\big)\Big]. \tag{19}$$

The dice loss, which directly optimizes the overlap between the predicted mask and ground truth, is formulated as

$$\mathcal{L}_{\text{dice}}\big(\hat{Y}_{\text{up}}, Y\big) = 1 - \frac{2\sum_{i=1}^{H}\sum_{j=1}^{W}\hat{Y}_{\text{up}}^{(i,j)}\,Y^{(i,j)} + \epsilon}{\sum_{i=1}^{H}\sum_{j=1}^{W}\big(\hat{Y}_{\text{up}}^{(i,j)}\big)^2 + \sum_{i=1}^{H}\sum_{j=1}^{W}\big(Y^{(i,j)}\big)^2 + \epsilon}, \tag{20}$$

where (i,j) denotes the pixel coordinates, $\hat{Y}_{\text{up}}^{(i,j)}$ and $Y^{(i,j)}$ represent the predicted probability and ground truth label at position (i,j), respectively, and ϵ is a small smoothing constant (set to 10^{-5} in our implementation) to ensure numerical stability and avoid division by zero.

By combining both losses, our training objective leverages the complementary strengths of pixel-wise classification accuracy and region-level overlap optimization, resulting in robust and accurate segmentation performance for the RRSIS task.

5 Experiments

In this section, we perform comprehensive experiments to assess the efficiency and effectiveness of our proposed RRSIS framework.

5.1 Dataset and evaluation metrics

We conduct experiments on three datasets, including two publicly available datasets and our constructed RISBench dataset. The detailed information for these three datasets is provided as follows.

RefSegRS [20] comprises 4420 image-language-label triplets across 285 scenes. The dataset is divided into a training set with 151 scenes and 2172 referring expressions, a validation set with 31 scenes and 431 expressions, and a test set with 103 scenes and 1817 expressions. The images are sized at 512×512 pixels, with a spatial resolution of 0.13 m.

RRSIS-D [22] contains a diverse dataset of 17402 images, each paired with corresponding masks and referring expressions. The dataset is organized into three subsets: a training set with 12181 image-language-label triplets, a validation set containing 1740 triplets, and a test set with 3481 triplets. All images are standardized to a resolution of 800×800 pixels. Additionally, the semantic annotations cover 20 categories and include 7 attributes, thereby enriching the semantic context of the referring expressions.

RISBench includes a total of 52472 image-language-label triplets. It is divided into three subsets: a training set consisting of 26300 triplets, a validation set consisting of 10013 triplets and a test set comprising 16159 triplets. All images are uniformly formatted to a resolution of 512×512 pixels, with spatial resolutions ranging from 0.1 to 30 m. The dataset's semantic labels are divided into 26 unique classes, with each class further annotated by 8 attributes.

Following the prior study [44], overall intersection-over-union (oIoU), mean intersection-over-union (mIoU), and precision at different threshold values $X \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$ (Pr@X) are selected as evaluation metrics.

oIoU is computed as the ratio of the cumulative intersection area to the cumulative union area across all test samples, thereby giving greater emphasis to larger objects:

$$\text{oIoU} = \left(\sum_t I_t\right) \Big/ \left(\sum_t U_t\right). \tag{21}$$

In contrast, mIoU is calculated by averaging the IoU values between predicted masks and ground truth annotations for each test sample, treating small and large objects equally:

$$\text{mIoU} = \frac{1}{M} \sum_t I_t / U_t, \tag{22}$$

where t denotes the index of the image-language-label triplets and M indicates the total size of the dataset. I_t and U_t represent the intersection and union area between the predicted and ground-truth regions.

5.2 Experimental setup

We employ Swin Transformer [39] and ConvNeXt [40] as the visual backbones in our approach. The Swin Transformer backbone is initialized with classification weights from the Swin-Base model pre-trained on ImageNet-22K [45]. The ConvNeXt backbone, specifically the ConvNeXt-Base and ConvNeXt-Large variants, is pre-trained on the GeoSense dataset, which comprises approximately 9 million diverse remote sensing images, using our proposed SMLFR self-supervised learning framework [46]. This pretraining on GeoSense enables the ConvNeXt backbone to learn robust and generalizable feature representations tailored for remote sensing image interpretation. For the language backbone, we use the base BERT model with 12 layers and a hidden size of 768, as available in the HuggingFace library [47]. The maximum sequence length for text descriptions is set to 20 tokens.

Our method is implemented using the PyTorch framework, and we employ the AdamW optimizer [48] with a weight decay of 0.01. During training, we adopt a differential learning rate strategy: the backbone networks are fine-tuned with a learning rate of 0.000005, while the task-specific modules (CAPM, LGFA, and MID) are trained with a learning rate of 0.00005. This strategy allows the pre-trained backbones to adapt gradually to the RRSIS task while preventing catastrophic forgetting of previously learned representations. The learning rate is decayed polynomially throughout training using a polynomial decay schedule with a power of 0.9. We use a batch size of 32, and each model is trained for 40 epochs on eight NVIDIA A800 GPUs (80 GB memory each). During both training and testing phases, images are resized to 480×480 pixels using bilinear interpolation. No data augmentation or post-processing techniques are applied to ensure fair comparison with baseline methods.

For all comparison methods, we utilize their officially released implementations with default hyperparameters when available. Methods employing ResNet-101 as the visual encoder (RRN, CMSA, LSCM, CMPC, BRINet, CMPC+, BKINet, ETRIS, and CRIS) are initialized with weights pre-trained on ImageNet-1K (ILSVRC2012), obtained from the PyTorch model zoo. Methods leveraging Swin Transformer-Base (LGCE, LAVT, RMSIN, CrossVLT, RIS-DMMI, robust-ref-seg, MAFN, and CARIS) employ identical ImageNet-22K pre-trained weights as our CroBIM with Swin-B variant, ensuring a fair comparison under the same initialization conditions. All baseline methods are trained using their reported optimal configurations to ensure reproducibility and fairness in performance evaluation.

5.3 Comparison with state-of-the-art methods

5.3.1 *RRSIS-D*

To evaluate the effectiveness of our proposed method, we conducted experiments on the RRSIS-D dataset. The comparison results are presented in Table 2. We compared CroBIM with several LSTM-based, CLIP-based, and BERT-based methods from classical to SOTA, i.e., RRN [29], CMSA [30], LSCM [49], CMPC [50], BRINet [32], CMPC+ [51], BKINet [52], ETRIS [53], CRIS [54], LGCE [20], LAVT [21], RMSIN [22], CrossVLT [55], RIS-DMMI [56], robust-ref-seg [57], CARIS [58], MAFN [59]. Among the compared methods, LGCE, RMSIN, and

Table 2 Comparison with SOTA methods on the RRSIS-D dataset. The optimal and sub-optimal performances for each metric are marked in bold and underlined, respectively.

Method	Visual encoder	Text encoder	Pr@0.5		Pr@0.6		Pr@0.7		Pr@0.8		Pr@0.9		oIoU		mIoU	
			Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
CMSA [30]	ResNet-101	Word embedding	55.68	55.32	48.04	46.45	38.27	37.43	26.55	25.39	9.02	8.15	69.68	69.43	48.85	48.54
RRN [29]	ResNet-101	LSTM	51.09	51.07	42.47	42.11	33.04	32.77	20.80	21.57	6.14	6.37	66.53	66.43	46.06	45.64
LSCM [49]	ResNet-101	LSTM	57.12	56.02	48.04	46.25	37.87	37.70	26.35	25.28	7.93	7.86	69.28	69.10	50.36	49.92
CMPC [50]	ResNet-101	LSTM	57.93	55.83	48.85	47.40	36.94	35.28	25.25	25.45	9.31	9.20	70.15	69.41	51.01	49.24
BRINet [32]	ResNet-101	LSTM	58.79	56.90	49.54	48.77	39.65	38.61	28.21	27.03	9.19	8.93	70.73	69.68	51.41	49.45
CMPC+ [51]	ResNet-101	LSTM	59.19	57.95	49.41	48.31	38.67	37.61	25.91	24.33	8.16	7.94	70.80	70.13	51.63	50.12
BKINet [52]	ResNet-101	CLIP	58.79	56.90	49.54	48.77	39.65	39.12	28.21	27.03	9.19	9.16	70.78	69.89	51.14	49.65
ETRIS [53]	ResNet-101	CLIP	62.10	61.07	53.73	50.99	43.12	40.94	30.79	29.30	12.90	11.43	72.75	71.06	55.21	54.21
CRIS [54]	ResNet-101	CLIP	56.44	54.84	47.87	46.77	39.77	38.06	29.31	28.15	11.84	11.52	70.98	70.46	50.75	49.69
LGCE [20]	Swin-B	BERT	68.10	67.65	60.61	61.53	51.45	51.42	42.34	39.62	23.85	22.94	76.68	76.33	60.16	59.37
LAVT [21]	Swin-B	BERT	65.23	63.98	58.79	57.57	50.29	49.30	40.11	38.06	23.05	22.29	76.27	76.16	57.72	56.82
RMSIN [22]	Swin-B	BERT	68.39	67.16	61.72	60.36	52.24	50.16	41.44	38.72	23.16	22.81	**77.53**	75.79	60.23	58.79
CrossVLT [55]	Swin-B	BERT	67.07	66.42	59.54	59.41	50.80	49.76	40.57	38.67	23.51	23.30	76.25	75.48	59.78	58.48
RIS-DMMI [56]	Swin-B	BERT	70.40	68.74	63.05	60.96	54.14	50.33	41.95	38.38	23.85	21.63	77.01	76.20	61.70	60.25
robust-ref-seg [57]	Swin-B	BERT	64.22	66.59	58.72	59.58	50.00	49.93	35.78	38.72	24.31	23.30	76.39	**77.40**	58.92	58.91
MAFN [59]	Swin-B	BERT	71.58	71.41	64.08	63.17	53.29	52.47	41.91	39.64	23.26	22.41	76.06	76.28	62.49	61.71
CARIS [58]	Swin-B	BERT	71.61	71.50	64.66	63.52	54.14	52.92	42.76	40.94	23.79	**23.90**	77.48	77.17	62.88	62.12
CroBIM (ours)	Swin-B	BERT	74.20	**75.00**	66.15	66.32	54.08	54.31	41.38	41.09	22.30	21.78	76.24	76.37	63.99	64.24
CroBIM (ours)	ConvNeXt-B	BERT	**74.94**	74.58	**67.64**	**67.57**	**57.18**	**55.59**	**44.66**	**41.63**	**24.60**	23.56	76.94	75.99	**65.05**	**64.46**

MAFN are the only three approaches specifically designed for the RRSIS task in the remote sensing domain. All other methods included in our comparison are SOTA algorithms from the computer vision domain. Although not specifically tailored for remote sensing, these methods represent the current best practices in referring image segmentation and have been widely recognized within the computer vision community.

It can be observed that LSTM-based methods generally perform worse than CLIP-based and BERT-based methods. This is because, in the context of referring image segmentation, where the textual descriptions may require understanding of nuanced spatial relationships and contextual cues, LSTMs may struggle to fully capture the necessary semantics and syntactic variations due to their sequential processing nature. The combination of a Swin Transformer-based visual encoder and a BERT-based text encoder has become the mainstream solution for current referring image segmentation tasks.

Among all the compared methods, our CroBIM utilizes two different visual encoders (Swin-B and ConvNeXt-B) and a text encoder (BERT), achieving optimal or sub-optimal performance across multiple metrics. Notably, CroBIM with ConvNeXt-B as the visual encoder demonstrates outstanding performance in the majority of metrics, particularly achieving the best results in Pr@0.5, Pr@0.6, Pr@0.7, Pr@0.8, Pr@0.9, and mIoU. Specifically, CroBIM with ConvNeXt-B achieved 74.94% (Val) and 74.58% (Test) for Pr@0.5, and 67.64% (Val) and 67.57% (Test) for Pr@0.6, significantly outperforming other methods. Meanwhile, CroBIM with Swin-B achieved 74.20% (Val) and 75.00% (Test) for Pr@0.5, and 66.15% (Val) and 66.32% (Test) for Pr@0.6, also demonstrating strong performance. Notably, CroBIM (ConvNeXt-B) achieved the highest mIoU scores of 65.05% (Val) and 64.46% (Test), which are 2.17% (Val) and 2.34% (Test) higher than the closest competing method, CARIS. Since the proposed CroBIM focuses on enhancing the model's ability to achieve precise segmentation of target categories, especially in complex remote sensing backgrounds, the model may attain higher local accuracy when handling challenging and less frequent target categories. This emphasis, however, may lead to suboptimal performance in oIoU. We also conduct a qualitative comparison between our model and the baseline to offer a comprehensive understanding of the results, as demonstrated in Figure 7.

5.3.2 *RefSegRS*

We further conduct experiments on the RefSegRS dataset to validate the superiority of our proposed framework, and the performance on both the validation and test sets is reported in Table 3. First, in terms of the Pr@0.5 and Pr@0.6 metrics, CroBIM consistently outperforms other SOTA methods on both the validation and test sets. Notably, CroBIM with ConvNeXt-B as the visual encoder achieves the highest accuracy of 75.89% on the test set for Pr@0.5, significantly surpassing the second-best method, RIS-DMMI, which attained 63.89%. This result demonstrates that CroBIM effectively segments the target regions under lower overlap thresholds. Additionally,

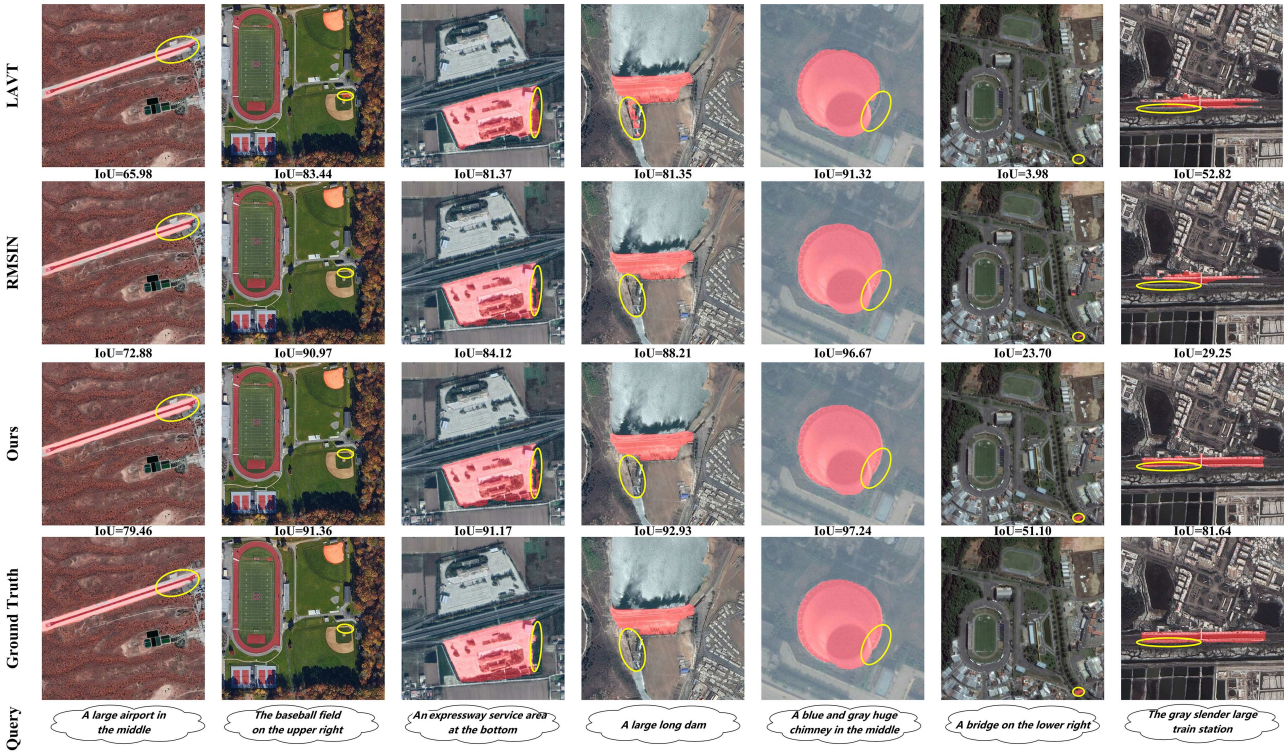


Figure 7 (Color online) Visualization of segmentation results for CroBIM and comparison methods on the RRSIS-D dataset test set, with yellow circles highlighting areas of improved performance and corresponding IoU scores displayed.

Table 3 Comparison with SOTA methods on the RefSegRS dataset. The optimal and sub-optimal performances for each metric are marked in bold and underlined, respectively.

Method	Visual encoder	Text encoder	Pr@0.5		Pr@0.6		Pr@0.7		Pr@0.8		Pr@0.9		oIoU		mIoU	
			Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
CMSA [30]	ResNet-101	Word embedding	39.24	26.14	38.44	18.52	20.39	10.66	11.79	4.71	1.52	0.69	63.84	62.11	43.62	38.72
RRN [29]	ResNet-101	LSTM	55.43	30.26	42.98	23.01	23.11	14.87	13.72	7.17	2.64	0.98	69.24	65.06	50.81	41.88
LSCM [49]	ResNet-101	LSTM	56.82	31.54	41.24	20.41	21.85	9.51	12.11	5.29	2.51	0.84	62.82	61.27	40.59	35.54
BRINet [32]	ResNet-101	LSTM	36.86	20.72	35.53	14.26	19.93	9.87	10.66	2.98	2.84	1.14	61.59	58.22	38.73	31.51
MAttNet [60]	ResNet-101	LSTM	48.56	28.79	40.26	22.51	20.59	11.32	12.98	3.62	2.02	0.79	66.84	64.28	41.73	33.42
BKINet [52]	ResNet-101	CLIP	52.04	36.12	35.31	20.62	18.35	15.22	12.78	6.26	1.23	1.33	75.37	63.37	56.12	40.41
ETRIS [53]	ResNet-101	CLIP	54.99	35.77	35.03	23.00	25.06	13.98	12.53	6.44	1.62	1.10	72.89	65.96	54.03	43.11
CRIS [54]	ResNet-101	CLIP	53.13	35.77	36.19	24.11	24.36	14.36	11.83	6.38	2.55	1.21	72.14	65.87	53.74	43.26
RMSIN [22]	Swin-B	BERT	68.21	42.32	46.64	25.87	24.13	14.20	13.69	6.77	3.25	1.27	74.40	68.31	54.24	42.63
CrossVLT [55]	Swin-B	BERT	67.52	41.94	43.85	25.43	25.99	15.19	14.62	3.71	1.87	1.76	76.12	69.73	55.27	42.81
RIS-DMMI [56]	Swin-B	BERT	86.17	63.89	74.71	44.30	38.05	19.81	18.10	6.49	3.25	1.00	74.02	68.58	65.72	52.15
CARIS [58]	Swin-B	BERT	68.45	45.40	47.10	27.19	25.52	15.08	14.62	7.87	3.71	1.98	75.79	69.74	54.30	42.66
robust-ref-seg [57]	Swin-B	BERT	81.67	50.25	52.44	28.01	30.86	17.83	17.17	9.19	5.80	2.48	77.74	71.13	60.44	47.12
LGCE [20]	Swin-B	BERT	79.81	50.19	54.29	28.62	29.70	17.17	15.31	9.36	5.10	2.15	78.24	71.59	60.66	46.57
LAVT [21]	Swin-B	BERT	80.97	51.84	58.70	30.27	31.09	17.34	15.55	9.52	4.64	2.09	78.50	71.86	61.53	47.40
CroBIM (ours)	Swin-B	BERT	87.24	64.83	75.17	44.41	44.78	17.28	19.03	9.69	**6.26**	2.20	**78.85**	72.30	65.79	52.69
CroBIM (ours)	ConvNeXt-B	BERT	**93.04**	**75.89**	**87.70**	**61.42**	**66.13**	**34.07**	**26.91**	**12.99**	5.80	**2.75**	77.95	**72.33**	**71.93**	**59.77**

CroBIM continues to exhibit a marked advantage in the Pr@0.6 metric, achieving an accuracy of 61.42% on the test set, which is substantially higher than RIS-DMMI's 44.30%. As the overlap threshold increases, CroBIM maintains robust performance. Specifically, at the more stringent Pr@0.9 threshold, CroBIM attains an accuracy of 6.26% on the test set, further confirming its effectiveness in high-precision segmentation tasks.

CroBIM also demonstrates superior performance in two key metrics: oIoU and mIoU. For oIoU, CroBIM achieves test scores of 72.33% (ConvNeXt-B) and 72.30% (Swin-B), ranking first and second, respectively, significantly outperforming competing methods such as LAVT and RIS-DMMI. Similarly, in terms of mIoU, CroBIM achieves



Figure 8 (Color online) Visualization of segmentation results for CroBIM and comparison methods on the RefSegRS dataset test set, with yellow circles highlighting areas of improved performance and corresponding IoU scores displayed.

59.77% (ConvNeXt-B) and 52.69% (Swin-B) on the test set, once again outperforming all other methods. These results highlight that CroBIM not only excels at lower overlap thresholds but also delivers significant improvements in overall segmentation accuracy. Figure 8 presents a qualitative analysis contrasting our model with the comparison methods, providing insights into the performance differences.

5.3.3 *RISBench*

In our constructed RISBench dataset, the proposed CroBIM model also demonstrates significant performance advantages over competing methods, achieving either the best or second-best results across multiple evaluation metrics, as shown in Table 4. Notably, CroBIM outperforms the current SOTA models on both the validation and test sets across various threshold levels, including Pr@0.5, Pr@0.6, Pr@0.7, Pr@0.8, and Pr@0.9. These results further validate the effectiveness and robustness of our model. Specifically, the CroBIM model, utilizing the ConvNeXt-B visual encoder and BERT text encoder, achieves the highest test set precision at Pr@0.5, reaching 77.55%, outperforming all other methods. Additionally, at Pr@0.6, Pr@0.7, and Pr@0.8, the model attains precisions of 72.83%, 66.38%, and 55.93%, respectively, significantly surpassing existing methods. Particularly, at the higher threshold of Pr@0.9, CroBIM leads other methods with a precision of 34.07%, demonstrating the model's strong capability to handle more complex scenarios that demand higher precision.

In contrast, although models such as RIS-DMMI and CARIS exhibit competitive performance on certain metrics—CARIS achieves 75.10% on oIoU and RIS-DMMI reaches 70.58%—these models fail to maintain consistent performance under higher precision thresholds. CroBIM, on the other hand, exhibits competitive results across both comprehensive metrics, oIoU and mIoU. The ConvNeXt-B variant of CroBIM achieves a test set mIoU of 69.33%, surpassing all comparison methods. This indicates that our model excels not only in individual precision metrics but also in overall segmentation accuracy.

Compared to other models utilizing the Swin-B and BERT combinations, CroBIM significantly enhances vision-language alignment between visual and textual features through improvements in feature extraction. By leveraging an innovative cross-modal bidirectional interaction mechanism, our model more effectively captures key information in the image that corresponds to the textual description. This capability is particularly beneficial in remote sensing imagery where targets are complex and the background is often noisy. Even under such challenging conditions, CroBIM consistently achieves excellent segmentation results. The visualization results in Figure 9 further highlight the superiority of our approach.

Table 4 Comparison with SOTA methods on the RISBench dataset. The optimal and sub-optimal performances for each metric are marked in bold and underlined, respectively.

Method	Visual encoder	Text encoder	Pr@0.5		Pr@0.6		Pr@0.7		Pr@0.8		Pr@0.9		oIoU		mIoU	
			Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
RRN [29]	ResNet-101	LSTM	54.62	55.04	46.88	47.31	39.57	39.86	32.64	32.58	11.57	13.24	47.28	49.67	42.65	43.18
LSCM [49]	ResNet-101	LSTM	55.87	55.26	47.24	47.14	40.22	40.10	33.55	33.29	12.78	13.91	47.99	50.08	43.21	43.69
BRINet [32]	ResNet-101	LSTM	52.11	52.87	45.17	45.39	37.98	38.64	30.88	30.79	10.28	11.86	46.27	48.73	41.54	42.91
MAttNet [60]	ResNet-101	LSTM	56.77	56.83	48.51	48.02	41.53	41.75	34.33	34.18	13.84	15.26	48.66	51.24	44.28	45.71
CMPC [50]	ResNet-101	LSTM	54.89	55.17	47.77	47.84	40.38	40.28	32.89	32.87	12.63	14.55	47.59	50.24	42.83	43.82
CMPC+ [51]	ResNet-101	LSTM	57.84	58.02	49.24	49.00	42.34	42.53	35.77	35.26	14.55	17.88	50.29	53.98	45.81	46.73
ETRIS [53]	ResNet-101	CLIP	59.87	60.98	49.91	51.88	35.88	39.87	20.10	24.49	8.54	11.18	64.09	67.61	51.13	53.06
CRIS [54]	ResNet-101	CLIP	63.42	63.67	54.32	55.73	41.15	44.42	24.66	28.80	10.27	13.27	66.26	69.11	53.64	55.18
LAVT [21]	Swin-B	BERT	68.27	69.40	62.71	63.66	54.46	56.10	43.13	44.95	21.61	25.21	69.39	74.15	60.45	61.93
RMSIN [22]	Swin-B	BERT	70.05	71.01	64.64	65.46	56.37	57.69	44.14	45.50	21.40	25.92	69.51	74.09	61.78	63.07
LGCE [20]	Swin-B	BERT	68.20	69.64	62.91	64.07	55.01	56.26	43.38	44.92	21.58	25.74	68.81	73.87	60.44	62.13
CrossVLT [55]	Swin-B	BERT	70.05	70.62	64.29	65.05	56.97	57.40	44.49	45.80	21.47	26.10	69.77	74.33	61.54	62.84
CARIS [58]	Swin-B	BERT	73.46	73.94	68.51	68.93	60.92	62.08	48.47	50.31	24.98	29.08	70.55	**75.10**	64.40	65.79
RIS-DMMI [56]	Swin-B	BERT	71.27	72.05	66.02	66.48	58.22	59.07	45.57	47.16	22.43	26.57	**70.58**	74.82	62.62	63.93
robust-ref-seg [57]	Swin-B	BERT	67.42	69.15	61.72	63.24	53.64	55.33	40.71	43.27	19.43	24.20	69.50	74.23	59.37	61.25
CroBIM (ours)	Swin-B	BERT	76.59	75.75	71.73	70.34	64.32	63.12	53.18	51.12	28.53	28.45	69.08	73.61	67.52	67.32
CroBIM (ours)	ConvNeXt-B	BERT	**77.41**	**77.55**	**72.62**	**72.83**	**66.74**	**66.38**	**55.92**	**55.93**	**32.17**	**34.07**	69.12	73.04	**68.70**	**69.33**

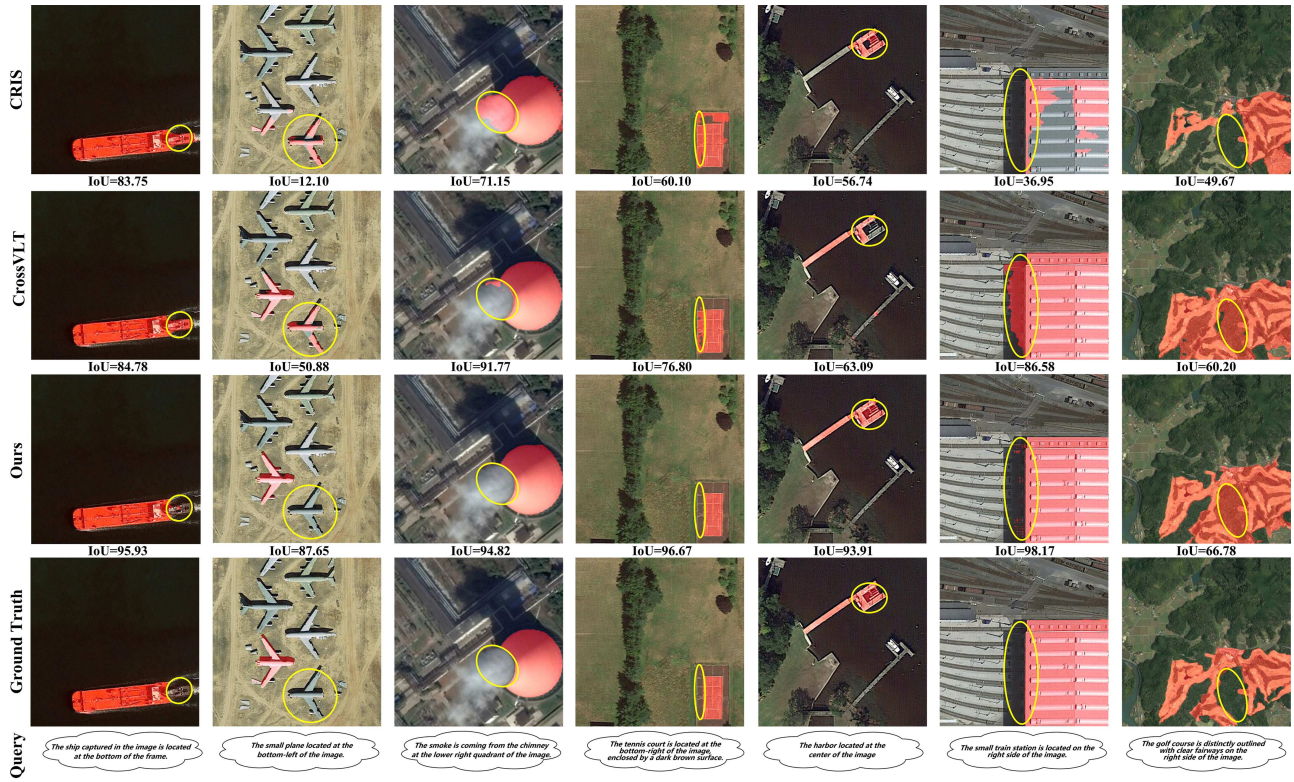


Figure 9 (Color online) Visualization of segmentation results for CroBIM and comparison methods on the RISBench dataset test set, with yellow circles highlighting areas of improved performance and corresponding IoU scores displayed.

Besides, the experimental results demonstrated that our proposed CroBIM framework not only outperformed the existing RRSIS-specific methods (LGCE, RMSIN, and MAFN) but also surpassed the adapted SOTA methods from computer vision. This highlighted the effectiveness and robustness of our approach in handling the unique challenges of remote sensing imagery, such as complex spatial relationships and diverse object scales.

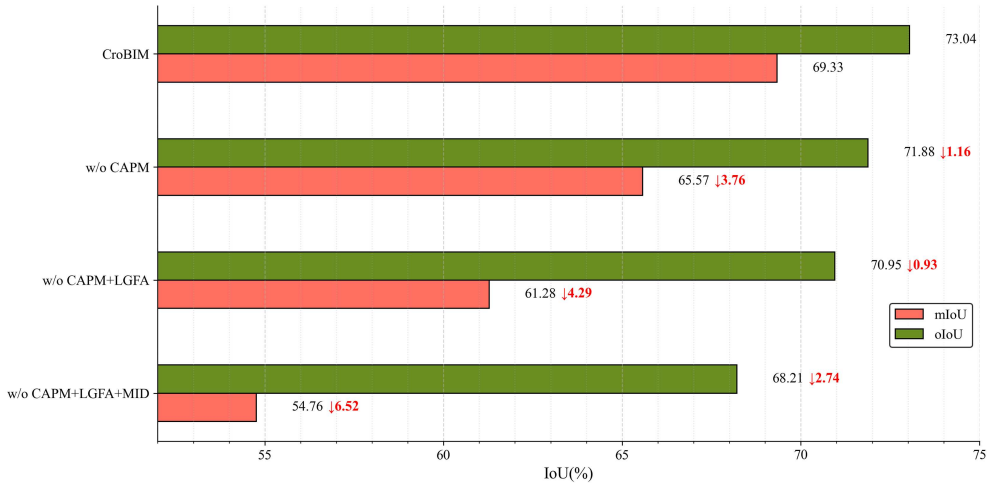


Figure 10 (Color online) Performance comparison of CroBIM and its ablated variants (w/o CAPM, LGFA, and MID) on the RISBench test set.

6 Ablation study

To assess the effectiveness of our designs within the CroBIM framework, we perform comprehensive experiments using our constructed RISBench dataset and present the quantitative outcomes obtained from the test set analysis.

6.1 Effectiveness of each component

To evaluate the effectiveness of the three key components of CroBIM (CAPM, LGFA, and MID), we conducted ablation experiments on the RISBench test dataset. The experimental results are presented in Figure 10.

When the CAPM module is removed from the complete CroBIM framework, the mIoU drops from 69.33% to 65.57%, representing a significant performance reduction of 3.76%. Similarly, the oIoU decreases from 73.04% to 71.88%. These results underscore the critical role of CAPM in integrating multi-scale spatial context into linguistic features, thereby enabling precise localization of target objects in remote sensing images. In the absence of CAPM, the model struggles to capture the spatial positional relationships described in the referring expressions, leading to a noticeable decline in segmentation accuracy.

When the LGFA module is further removed, the performance degradation becomes more pronounced. The mIoU decreases to 61.28% (a reduction of 4.29%) and the oIoU drops to 70.95% (a reduction of 0.93%) compared to the prior configuration. This highlights the pivotal role of LGFA in aggregating visual and linguistic features across multiple scales and compensating for attention deficits. Without LGFA, the model loses its ability to effectively handle objects of varying scales, which is a fundamental challenge in remote sensing scenarios.

Finally, replacing the MID module with a simple decoder [21] results in mIoU and oIoU dropping further to 54.76% and 68.21%, respectively. These findings clearly demonstrate the importance of MID in ensuring precise alignment between visual and linguistic features through iterative bidirectional refinement. The absence of MID significantly impairs the model's ability to resolve the complex cross-modal relationships required for accurate segmentation, particularly in remote sensing datasets characterized by diverse geospatial relationships and linguistically rich referring expressions.

In summary, the above results conclusively demonstrate that all three components are indispensable for achieving precise segmentation in complex scenarios. Each module contributes uniquely to the overall performance of CroBIM, ensuring its robustness and effectiveness in addressing the challenges inherent in RRSIS tasks.

6.2 Design of CAPM

To further validate the effectiveness of the CAPM module, we conduct a detailed analysis of its context-aware prompt design. First, we perform ablation studies on the combinations of visual features at different scales, with the results presented in Table 5(a). As can be observed, the combination incorporating all four visual features, $\{V_1, V_2, V_3, V_4\}$, achieves the best performance in terms of prediction accuracy (Pr@0.5, Pr@0.7, Pr@0.9) and mIoU. This demonstrates that integrating visual features from multiple scales significantly enhances the model's prediction accuracy and overall segmentation quality. The improved performance suggests that the comprehensive

Table 5 Ablation studies on options design of CAPM. The best result for each metric is marked in bold.

Option	Pr@0.5	Pr@0.7	Pr@0.9	oIoU	mIoU
(a) Combination of visual features at different scales					
$\{V_1, V_2, V_3\}$	77.22	66.28	33.96	72.95	69.04
$\{V_1, V_2, V_4\}$	77.15	66.25	33.82	72.88	69.11
$\{V_2, V_3, V_4\}$	77.26	66.24	33.89	**73.16**	69.15
$\{V_1, V_2, V_3, V_4\}$	**77.55**	**66.38**	**34.07**	73.04	**69.33**
(b) Design of pooling					
Max pool	76.38	65.04	33.52	72.47	68.67
Average pool	77.23	66.15	33.61	72.84	69.10
Adaptive average pool	**77.55**	**66.38**	**34.07**	**73.04**	**69.33**

Table 6 Ablation studies on attention deficit compensation of LGFA. The best result for each metric is marked in bold.

S1	S2	S3	S4	Pr@0.5	Pr@0.7	Pr@0.9	oIoU	mIoU
			✓	76.64	65.72	33.61	72.18	68.55
		✓	✓	77.01	65.93	33.85	72.61	68.78
	✓	✓	✓	77.14	66.21	**34.12**	72.87	69.02
✓	✓	✓	✓	**77.55**	**66.38**	34.07	**73.04**	**69.33**

utilization of features across different levels of granularity contributes to a more robust and precise representation, thus improving the model's ability to capture complex visual cues.

We further investigate the impact of different pooling strategies in the dimensionality reduction of multi-scale visual features, as shown in Table 5(b). It can be observed that, compared to max pooling and average pooling, adaptive average pooling consistently achieves superior performance across all evaluation metrics. The superior performance of adaptive average pooling can be attributed to two key advantages. First, unlike standard average pooling, which applies uniform pooling across all spatial locations, adaptive average pooling dynamically adjusts the pooling kernel size and stride to achieve the desired output dimensions, enabling more flexible integration of features across different scales while preserving both global and local information. Second, compared to max pooling, which selects only maximum activation values and may discard subtle contextual cues, adaptive average pooling aggregates information from all spatial positions within each pooling region, providing a more comprehensive representation of multi-scale visual context. This comprehensive feature aggregation is particularly crucial in remote sensing scenarios where target objects exhibit complex spatial relationships and varying scales, ultimately facilitating more effective cross-modal interaction in the CAPM module. These results suggest that the adaptive mechanism is more effective in flexibly integrating features across different scales, thereby preserving more information that is crucial for accurate predictions. Moreover, by dynamically adjusting the weighting of features based on their respective scales, adaptive average pooling enables a more comprehensive retention of both global and local information during the dimensionality reduction process, ultimately enhancing segmentation performance.

6.3 Design of LGFA

We further validated the effectiveness of attention deficit compensation within the LGFA module. The ablation study results presented in Table 6 demonstrate the efficacy of applying attention deficit compensation at different stages (S1–S4) in improving the aggregation and alignment of visual and linguistic features. As attention deficit compensation is progressively applied from a single stage (S4) to multiple stages, a significant performance improvement across all metrics can be observed. Notably, when attention deficit compensation is applied at all stages (S1–S4), the model achieves the best segmentation results: Pr@0.5 of 77.55%, Pr@0.7 of 66.38%, and mIoU of 69.33%. The most prominent gains are observed in Pr@0.5 and oIoU, indicating that the complete attention deficit compensation strategy enhances the precision of feature alignment and overall segmentation accuracy. These findings underscore the importance of applying attention deficit compensation at multiple stages, as it leads to more robust integration of visual and linguistic features, thereby improving the segmentation performance.

To investigate the impact of K (the number of attention deficit regions) in the LGFA module, we also conduct systematic ablation experiments by varying K from 0 to 128. As shown in Table 7, when $K = 0$ (without attention deficit compensation), the model achieves baseline performance of 72.31% oIoU and 68.62% mIoU. Introducing compensation with $K = 8$ immediately improves performance to 72.59% oIoU (+0.28%) and 68.94% mIoU (+0.32%), validating the effectiveness of the proposed mechanism. As K increases to 32, both metrics show consistent im-

Table 7 Ablation results of the number K of attention deficit regions in the LGFA module, evaluated on the RISBench test set using the ConvNeXt-B backbone. The values in bold correspond to K = 32, which is adopted as the default configuration in all experiments since it provides the best trade-off between segmentation accuracy and computational cost.

K value	oIoU (%)	mIoU (%)	FLOPs (G)
0	72.31	68.62	82.32
8	72.59	68.94	82.36
16	72.84	69.25	82.42
32	**73.04**	**69.33**	82.53
64	73.06	69.32	82.75
128	73.07	69.35	83.17

Table 8 Ablation studies on attention mechanisms of MID. The best result for each metric is marked in bold.

Attention	Pr@0.5	Pr@0.7	Pr@0.9	oIoU	mIoU
PWAM [21]	72.64	63.87	30.59	70.28	65.54
WPA [61]	75.83	64.87	32.71	72.19	67.02
CBAM (ours)	**77.55**	**66.38**	**34.07**	**73.04**	**69.33**

provement, reaching optimal performance of 73.04% oIoU and 69.33% mIoU. These represent substantial gains of 0.73% and 0.71% over the baseline, respectively. Beyond $K = 32$, performance plateaus with only marginal improvements ($K = 128$: 73.07% oIoU, 69.35% mIoU), suggesting that the most critical attention deficit regions have been effectively captured. From a computational efficiency perspective, the overhead remains minimal across all configurations. At $K = 32$, FLOPs increase from 82.32 to 82.53 G (only 0.26% increase). Even with $K = 128$, FLOPs reach only 83.17 G (1.03% increase), demonstrating excellent computational efficiency.

Notably, performance remains stable across $K \in [8, 128]$, with mIoU fluctuating within 0.41 percentage points and oIoU within 0.48 percentage points. This robustness indicates that our method is not overly sensitive to K, reducing the hyperparameter tuning burden and enhancing practical applicability across diverse remote sensing scenarios. The diminishing returns beyond $K = 32$ can be theoretically explained by the attention deficit distribution. The top-K mechanism prioritizes regions with the largest cross-scale attention divergence. When $K = 32$, the most significant deficit regions are adequately covered, while additional regions contribute marginal value due to smaller attention inconsistencies. Based on these findings, we adopt $K = 32$ as the default configuration in all experiments, providing an optimal balance between segmentation accuracy and computational efficiency.

6.4 Design of MID

Moreover, we investigate the effectiveness of the cascaded bidirectional attention mechanism (CBAM) employed in our MID, and compare it with two existing approaches: the unidirectional visual-to-language attention module PWAM [21], and the parallel bidirectional attention WPA [61]. The comparative results are presented in Table 8. As observed, the unidirectional attention mechanism in PWAM yielded the lowest segmentation performance. This underperformance can be attributed to its reliance on unidirectional interactions, which fail to adequately capture the joint representation of cross-modal visual-linguistic features.

In contrast, our cascaded bidirectional attention mechanism achieved the best performance, outperforming PWAM and WPA by 3.79% and 2.31% in terms of mIoU, respectively. These results demonstrate that the cascaded bidirectional attention effectively enables deep interaction between linguistic features and multi-scale visual contexts, fostering vision-language alignment in a more efficient manner. This enhanced interaction not only improves the quality of feature representation but also significantly boosts segmentation performance, highlighting the superiority of our proposed approach in addressing the challenges of visual-linguistic cross-modal tasks.

6.5 Qualitative results

In Figure 11, we visualize the attention maps at various stages of the model to analyze CroBIM's vision-language alignment mechanism in greater depth. First, in the encoder's attention maps, we observe that as the network depth increases, CroBIM progressively focuses on more fine-grained target areas. The attention maps from the early stages reveal a broad feature capture, with attention distributed across the global information in the image. This is closely related to the high resolution and complex background of remote sensing imagery—at these early stages, the model must focus on the overall structure to establish a comprehensive understanding of the scene. As the network layers progress, attention increasingly narrows to target regions that are closely aligned with the textual description, indicating that the model effectively filters out background noise and captures the key features of the

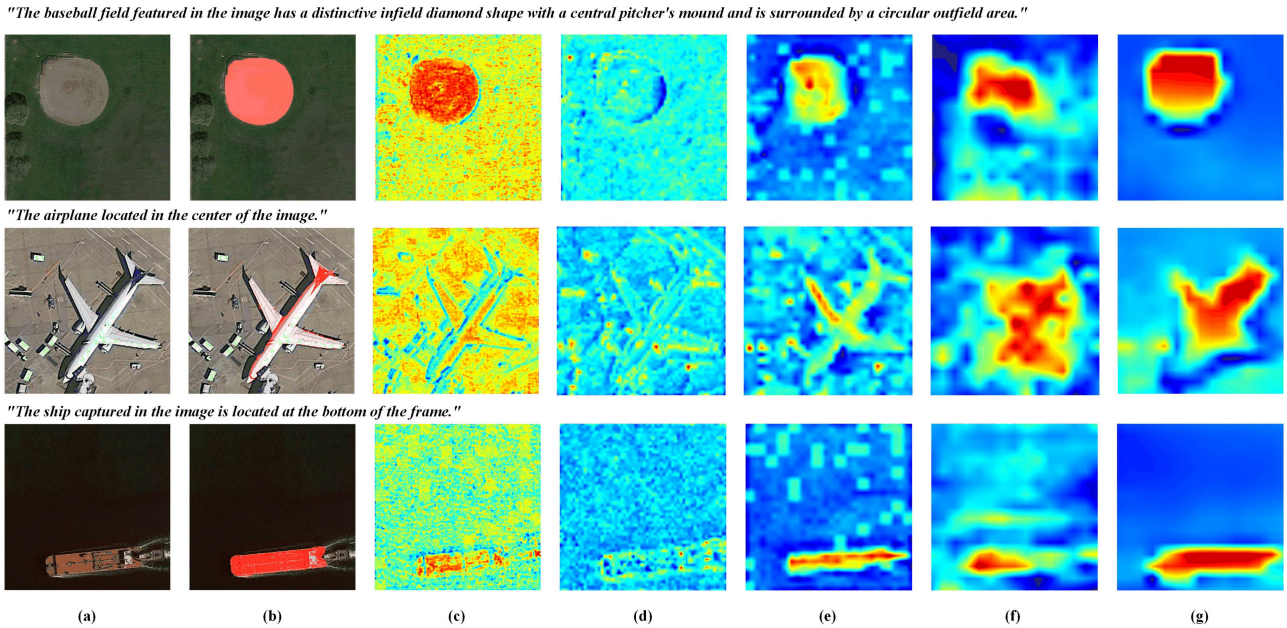


Figure 11 (Color online) Attention map visualization from different stages in CroBIM. (a) Input image; (b) ground truth; (c)–(f) attention maps of stages S1–S4 in the encoder; (g) attention maps of the decoder.

target. For instance, in remote sensing images with complex terrain or buildings, CroBIM is able to accurately localize critical objects such as buildings and roads in the mid-to-late layers. Furthermore, the attention maps in the decoder stage illustrate the model's segmentation capability. In the decoder, the attention maps demonstrate a high degree of refinement, with the model exhibiting more precise attention to the edges and finer details of the target regions. This focus further validates CroBIM's strong performance in processing high-resolution remote sensing images, particularly in tasks requiring precise target segmentation. The decoder's attention maps clearly show that the model can accurately delineate the boundaries between the target and background, especially in areas where textures or colors are highly similar.

To further demonstrate the superiority of CroBIM, we compared its attention maps with those of RMSIN at the final stage of the decoder, as shown in Figure 12. The comparison reveals that RMSIN often suffers from attention drift, failing to focus accurately on the target regions specified in the textual query. In contrast, CroBIM consistently maintains precise attention on the target objects, even in complex scenarios with challenging geospatial relationships or similar foreground-background appearances. This refined attention mechanism enables CroBIM to achieve high-precision segmentation by accurately delineating target boundaries and capturing fine details, particularly in regions with similar textures or colors.

Although our CroBIM effectively models the query expression and aligns visual features with textual embeddings, some failure cases still occur, as shown in Figure 13. These failure cases can be categorized into four types: semantic ambiguity, visual similarity, imprecise annotations, and boundary targets. First, semantic ambiguity arises when the textual description corresponds to multiple objects within the image, leading the model to segment multiple instances instead of the intended target. Second, visual similarity occurs when the foreground object closely resembles the background in appearance, making it difficult for the model to distinguish between them. Third, imprecise annotations refer to errors introduced by inaccurate or inconsistent labeling, which can mislead the model during training and evaluation. Finally, boundary targets present challenges when the object is located near the image boundary, where limited contextual information may cause the model to misinterpret its position and shape, resulting in inaccurate segmentation.

7 Conclusion

In this paper, we introduce the CroBIM framework to address the challenge of referring image segmentation in remote sensing scenarios. By leveraging bidirectional visual-text feature interaction and alignment, CroBIM bridges the gap between visual perception and language understanding, facilitating precise target segmentation. Furthermore, we construct a large-scale benchmark dataset, RISBench, which encompasses a more extensive set of

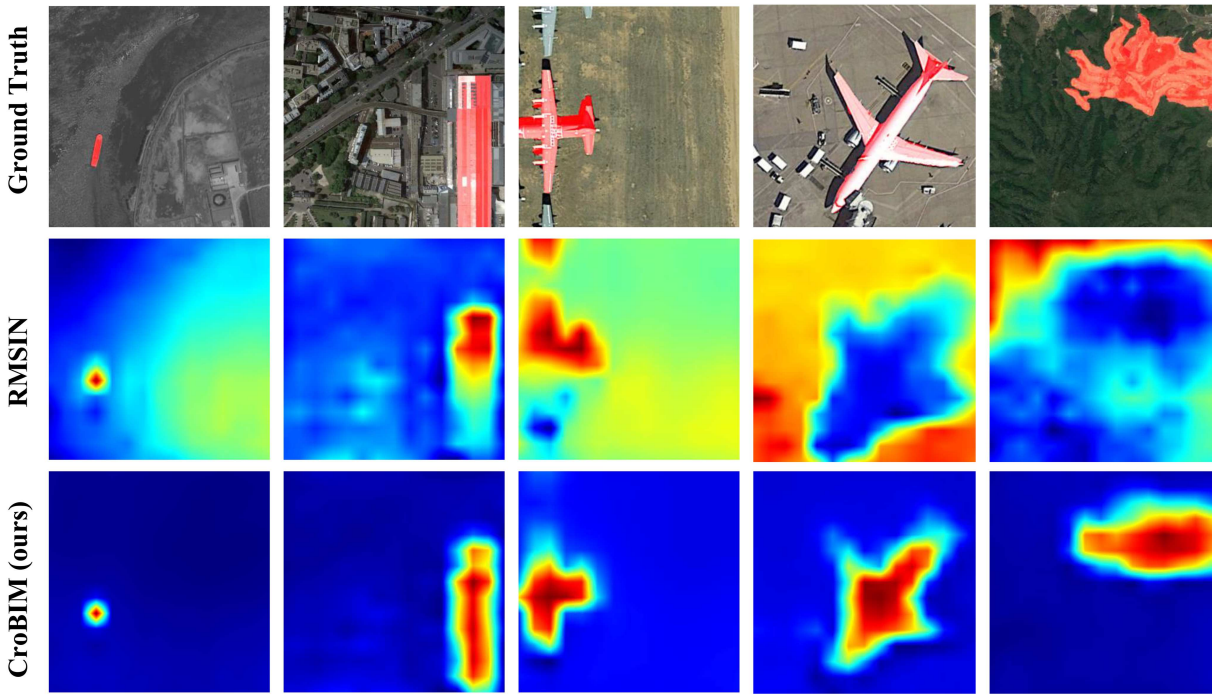


Figure 12 (Color online) Comparison of attention maps between CroBIM and RMSIN at the decoder stage.

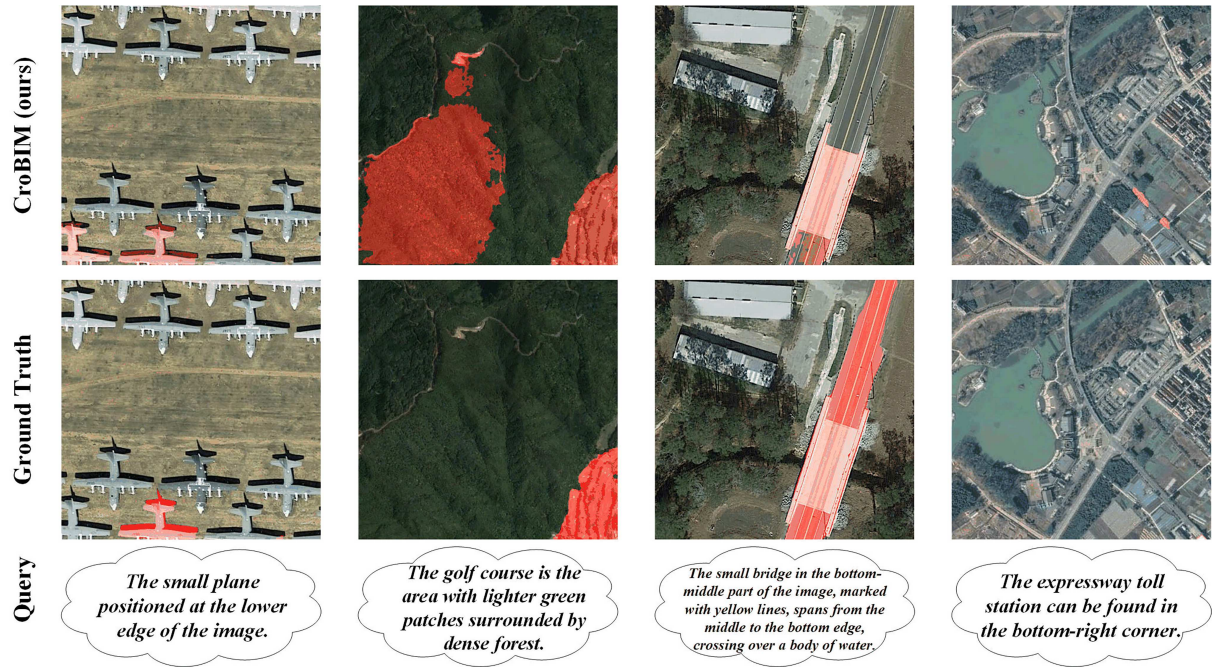


Figure 13 (Color online) Failure cases of our proposed CroBIM on the RISBench test set. From left to right, the four failure types are semantic ambiguity, visual similarity, imprecise annotations, and boundary targets.

image-language-label triplets, richer attribute expressions, a broader range of spatial resolutions, and more detailed textual descriptions. Experimental results demonstrate that the proposed CroBIM framework outperforms SOTA methods across three benchmark datasets, underscoring its efficacy and superiority.

The ability to precisely segment target objects in remote sensing images based on natural language descriptions not only enhances the accessibility of remote sensing data but also opens up new possibilities for interdisciplinary research and practical applications in fields such as disaster response, urban planning, and environmental monitoring. For future work, we aim to embed domain-specific knowledge [62,63] of remote sensing imagery into language models, such as sensor imaging theory, spatial correlations, and spectral characteristics of ground objects, to further enhance

remote sensing data analysis and interpretation. Additionally, another promising research direction is the integration of textual information with remote sensing through geolocation. By incorporating non-traditional geographic data, such as geotagged social media posts [64] and newspaper articles, remote sensing data can be combined with complementary sources. This approach broadens the potential applications [65–67] of remote sensing visual-language foundation models.

Acknowledgements This work was supported by National Natural Science Foundation of China (Special Program) (Grant No. 624B2051).

References

1 Du D, Gu Y, Liu T, et al. Spectral reconstruction from satellite multispectral imagery using convolution and transformer joint network. IEEE Trans Geosci Remote Sens, 2023, 61: 1–15
2 Dong Z, Liu T, Gu Y. Spatial and semantic consistency contrastive learning for self-supervised semantic segmentation of remote sensing images. IEEE Trans Geosci Remote Sens, 2023, 61: 1–12
3 Dong Z, Gao G, Liu T, et al. Distilling segmenters from CNNs and transformers for remote sensing images' semantic segmentation. IEEE Trans Geosci Remote Sens, 2023, 61: 1–14
4 Cai X, Lai Q, Wang Y, et al. Poly kernel inception network for remote sensing detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 27706–27716
5 Chen H, Qi Z, Shi Z. Remote sensing image change detection with transformers. IEEE Trans Geosci Remote Sens, 2022, 60: 1–14
6 Zia U, Mohsin Riaz M, Ghafoor A. Transforming remote sensing images to textual descriptions. Int J Appl Earth Obs, 2022, 108: 102741
7 Zhao R, Shi Z, Zou Z. High-resolution remote sensing image captioning based on structured attention. IEEE Trans Geosci Remote Sens, 2022, 60: 1–14
8 Rahhal M M A, Bazi Y, Alsharif N A, et al. Multilanguage transformer for improved text to remote sensing image retrieval. IEEE J Sel Top Appl Earth Obs Remote Sens, 2022, 15: 9115–9126
9 Yuan Z, Zhang W, Tian C, et al. Remote sensing cross-modal text-image retrieval based on global and local information. IEEE Trans Geosci Remote Sens, 2022, 60: 1–16
10 Xu Y, Yu W, Ghamisi P, et al. Txt2Img-MHN: remote sensing image generation from text using modern Hopfield networks. IEEE Trans Image Process, 2023, 32: 5737–5750
11 Zhao R, Shi Z. Text-to-remote-sensing-image generation with structured generative adversarial networks. IEEE Geosci Remote Sens Lett, 2022, 19: 1–5
12 Yuan Z, Mou L, Wang Q, et al. From easy to hard: learning language-guided curriculum for visual question answering on remote sensing data. IEEE Trans Geosci Remote Sens, 2022, 60: 1–11
13 Zheng X, Wang B, Du X, et al. Mutual attention inception network for remote sensing visual question answering. IEEE Trans Geosci Remote Sens, 2022, 60: 1–14
14 MohanRajan S N, Loganathan A, Manoharan P. Survey on land use/land cover (LU/LC) change analysis in remote sensing and GIS environment: techniques and challenges. Environ Sci Pollut Res, 2020, 27: 29900–29926
15 Wang S, Han Y, Chen J, et al. A deep-learning-based sea search and rescue algorithm by UAV remote sensing. In: Proceedings of the 2018 IEEE CSAA Guidance, Navigation and Control Conference (CGNCC), 2018. 1–5
16 Yuan Q, Shen H, Li T, et al. Deep learning in environmental remote sensing: achievements and challenges. Remote Sens Environ, 2020, 241: 111716
17 Zhang B, Wu Y, Zhao B, et al. Progress and challenges in intelligent remote sensing satellite systems. IEEE J Sel Top Appl Earth Obs Remote Sens, 2022, 15: 1814–1822
18 Weiss M, Jacob F, Duveiller G. Remote sensing for agricultural applications: a meta-review. Remote Sens Environ, 2020, 236: 111402
19 Wellmann T, Lausch A, Andersson E, et al. Remote sensing in urban planning: contributions towards ecologically sound policies? Landscape Urban Plan, 2020, 204: 103921
20 Yuan Z, Mou L, Hua Y, et al. RRSIS: referring remote sensing image segmentation. IEEE Trans Geosci Remote Sens, 2024, 62: 1–12
21 Yang Z, Wang J, Tang Y, et al. LAVT: language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 18155–18165
22 Liu S, Ma Y, Zhang X, et al. Rotated multi-scale interaction network for referring remote sensing image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 26658–26668
23 Ding J, Xue N, Xia G S, et al. Object detection in aerial images: a large-scale benchmark and challenges. IEEE Trans Pattern Anal, 2022, 44: 7778–7796
24 Li K, Wan G, Cheng G, et al. Object detection in optical remote sensing images: a survey and a new benchmark. ISPRS J Photogramm, 2020, 159: 296–307
25 Li X, Ding J, Elhoseiny M. VRSBench: a versatile vision-language benchmark dataset for remote sensing image understanding. 2024. ArXiv:2406.12384
26 Kirillov A, Mintun E, Ravi N, et al. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 4015–4026
27 Ji L, Du Y, Dang Y, et al. A survey of methods for addressing the challenges of referring image segmentation. Neurocomputing, 2024, 583: 127599
28 Hu R, Rohrbach M, Darrell T. Segmentation from natural language expressions. In: Proceedings of the 14th European Conference on Computer Vision–ECCV 2016, Amsterdam, 2016. 108–124
29 Li R, Li K, Kuo Y C, et al. Referring image segmentation via recurrent refinement networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 5745–5753
30 Ye L, Rochan M, Liu Z, et al. Cross-modal self-attention network for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. 10502–10511
31 Luo G, Zhou Y, Ji R, et al. Cascade grouped attention network for referring expression segmentation. In: Proceedings of the 28th ACM International Conference on Multimedia, 2020. 1274–1282
32 Hu Z, Feng G, Sun J, et al. Bi-directional relationship inferring network for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 4424–4433
33 Sun Y, Feng S, Li X, et al. Visual grounding in remote sensing images. In: Proceedings of the 30th ACM International Conference on Multimedia, 2022. 404–412
34 Zhan Y, Xiong Z, Yuan Y. RSVG: exploring data and models for visual grounding on remote sensing data. IEEE Trans Geosci Remote Sens, 2023, 61: 1–13
35 Li K, Wang D, Xu H, et al. Language-guided progressive attention for visual grounding in remote sensing images. IEEE Trans Geosci Remote Sens, 2024, 62: 1–13
36 Dong Z, Sun Y, Liu T, et al. DiffRIS: enhancing referring remote sensing image segmentation with pre-trained text-to-image diffusion models. 2025. ArXiv:2506.18946
37 Chen K, Zhang J, Liu C, et al. RSRefSeg: referring remote sensing image segmentation with foundation models. 2025. ArXiv:2501.06809
38 Xie Z, Guan B, Jiang W, et al. PA-SAM: prompt adapter SAM for high-quality image segmentation. 2024. ArXiv:2401.13051

39 Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 10012–10022
40 Liu Z, Mao H, Wu C Y, et al. A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 11976–11986
41 Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding. 2018. ArXiv:1810.04805
42 Zhu X, Su W, Lu L, et al. Deformable DETR: deformable transformers for end-to-end object detection. 2020. ArXiv:2010.04159
43 Li X, Sun X, Meng Y, et al. Dice loss for data-imbalanced NLP tasks. 2019. ArXiv:1911.02855
44 Wu C, Lin Z, Cohen S, et al. PhraseCut: language-based image segmentation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 10216–10225
45 Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009. 248–255
46 Dong Z, Gu Y, Liu T. Generative ConvNet foundation model with sparse modeling and low-frequency reconstruction for remote sensing image interpretation. IEEE Trans Geosci Remote Sens, 2024, 62: 1–16
47 Wolf T, Debut L, Sanh V, et al. Transformers: state-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020. 38–45
48 Loshchilov I, Hutter F. Decoupled weight decay regularization. 2017. ArXiv:1711.05101
49 Hui T, Liu S, Huang S, et al. Linguistic structure guided context modeling for referring image segmentation. In: Proceedings of the 16th European Conference on Computer Vision–ECCV 2020, Glasgow, 2020. 59–75
50 Huang S, Hui T, Liu S, et al. Referring image segmentation via cross-modal progressive comprehension. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 10488–10497
51 Liu S, Hui T, Huang S, et al. Cross-Modal progressive comprehension for referring segmentation. IEEE Trans Pattern Anal, 2021, 44: 4761–4775
52 Ding H, Zhang S, Wu Q, et al. Bilateral knowledge interaction network for referring image segmentation. IEEE Trans Multimedia, 2024, 26: 2966–2977
53 Xu Z, Chen Z, Zhang Y, et al. Bridging vision and language encoders: parameter-efficient tuning for referring image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 17503–17512
54 Wang Z, Lu Y, Li Q, et al. CRIS: CLIP-driven referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022. 11686–11695
55 Cho Y, Yu H, Kang S J. Cross-aware early fusion with stage-divided vision and language transformer encoders for referring image segmentation. IEEE Trans Multimedia, 2024, 26: 5823–5833
56 Hu Y, Wang Q, Shao W, et al. Beyond one-to-one: rethinking the referring image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 4067–4077
57 Wu J, Li X, Li X, et al. Toward robust referring image segmentation. IEEE Trans Image Process, 2024, 33: 1782–1794
58 Liu S A, Zhang Y, Qiu Z, et al. CARIS: context-aware referring image segmentation. In: Proceedings of the 31st ACM International Conference on Multimedia, 2023. 779–788
59 Shi L, Zhang J. Multimodal-aware fusion network for referring remote sensing image segmentation. IEEE Geosci Remote Sens Lett, 2025, 22: 1–5
60 Yu L, Lin Z, Shen X, et al. MAttNet: modular attention network for referring expression comprehension. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018. 1307–1315
61 Zhang Z, Zhu Y, Liu J, et al. CoupAlign: coupling word-pixel with sentence-mask alignments for referring image segmentation. In: Proceedings of the Advances in Neural Information Processing Systems, 2022, 35: 14729–14742
62 Wei J, Bosma M, Zhao V Y, et al. Finetuned language models are zero-shot learners. 2021. ArXiv:2109.01652
63 Dong Z, Gu Y, Liu T. UPetu: a unified parameter-efficient fine-tuning framework for remote sensing foundation model. IEEE Trans Geosci Remote Sens, 2024, 62: 1–13
64 Zhu X X, Wang Y, Kochupillai M, et al. Geoinformation harvesting from social media data: a community remote sensing approach. IEEE Geosci Remote Sens Mag, 2022, 10: 150–180
65 Häberle M, Hoffmann E J, Zhu X X. Can linguistic features extracted from geo-referenced tweets help building function classification in remote sensing? ISPRS J Photogramm, 2022, 188: 255–268
66 Kruspe A, Häberle M, Kuhn I, et al. Cross-language sentiment analysis of European Twitter messages during the COVID-19 pandemic. 2020. ArXiv:2008.12172
67 Kruspe A, Kersten J, Klan F. Review article: detection of actionable tweets in crisis events. Nat Hazard Earth Syst Sci, 2021, 21: 1825–1845