remote sensing

MDPI

Article

Multimodal Prompt-Guided Bidirectional Fusion for Referring Remote Sensing Image Segmentation

Yingjie Li, Weiqi Jin *, Su Qiu and Qiyang Sun

MOE Key Laboratory of Optoelectronic Imaging Technology and System, Beijing Institute of Technology, Beijing 100081, China; 3220215081@bit.edu.cn (Q.S.)
* Correspondence: jinwq@bit.edu.cn

Abstract: Multimodal feature alignment is a key challenge in referring remote sensing image segmentation (RRSIS). The complex spatial relationships and multi-scale targets in remote sensing images call for efficient cross-modal mapping and fine-grained feature alignment. Existing approaches typically rely on cross-attention for multimodal fusion, which increases model complexity. To address this, we introduce the concept of prompt learning in RRSIS and propose a parameter-efficient multimodal prompt-guided bidirectional fusion (MPBF) architecture. MPBF combines both early and late fusion strategies. In the early fusion stage, it conducts the deep fusion of linguistic and visual features through cross-modal prompt coupling. In the late fusion stage, to handle the multi-scale nature of remote sensing targets, a scale refinement module is proposed to capture diverse scale representations, and a vision–language alignment module is employed for pixel-level multimodal semantic associations. Comparative experiments and ablation studies on a public dataset demonstrate that MPBF significantly outperformed existing state-of-the-art methods with relatively small computational overhead, highlighting its effectiveness and efficiency for RRSIS. Further application experiments on a custom dataset confirm the method's practicality and robustness in real-world scenarios.

Keywords: remote sensing; referring image segmentation; prompt learning; bidirectional fusion

Academic Editor: Andrea Garzelli

Received: 21 March 2025
Revised: 2 May 2025
Accepted: 8 May 2025
Published: 10 May 2025

Citation: Li, Y.; Jin, W.; Qiu, S.; Sun, Q. Multimodal Prompt-Guided Bidirectional Fusion for Referring Remote Sensing Image Segmentation. *Remote Sens.* **2025**, *17*, 1683. https://doi.org/10.3390/rs17101683

1. Introduction

Referring remote sensing image segmentation (RRSIS) [1,2] integrates both linguistic and visual modalities to precisely segment targets based on natural language descriptions. Compared to purely image-based semantic segmentation, incorporating high-level semantic information from the linguistic modality, such as color, size, shape, and spatial relationships, significantly enhances the model's ability to understand complex scenes [3–5]. The integration of text descriptions offers significant advantages by improving the flexibility of target segmentation and enhancing the user-friendliness of human–computer interaction [6–8], which has broad application prospects in applications such as marine target detection and tracking, as well as hazard detection along high-speed rail lines. Consequently, RRSIS has emerged as a key frontier in the field of intelligent remote sensing image analysis.

RRSIS aims to comprehend specific objects referred to by a text description, map the referring relationships to the spatial semantics of the remote sensing image, and accurately locate and segment the target objects. However, there are significant differences between remote sensing images and natural images in terms of spatial semantic representation, primarily reflected in the following two aspects. First, the diversity of land cover and

the distribution complexity of targets introduce intricate spatial relationships [9]. The inherent flexibility of language expression further increases the uncertainty in cross-domain mapping between visual and linguistic information, making it challenging to accurately interpret spatial semantics in remote sensing images. Second, the unique imaging altitude of aerial and satellite remote sensing platforms results in ground targets appearing at small scales. This requires robust cross-modal fine-grained alignment capabilities in models to establish pixel-level semantic associations. Therefore, accurate cross-modal mapping and fine-grained feature alignment between freely expressed text descriptions and high-resolution images remain fundamental challenges in RRSIS.

These challenges can be addressed through the effective fusion of multimodal features. Existing studies can be categorized into early fusion [10–12] and late fusion [3,13–15] approaches based on the stage at which fusion occurs. Earlier research primarily focused on late fusion methods, where encoded linguistic and visual features were integrated by concatenation and bilinear fusion. While simple to implement, late fusion methods are limited in their capacity for deep multimodal interaction [16,17], yielding multimodal features that still retain inherent differences between modalities. With the remarkable success of transformer [18] architectures in natural language processing and computer vision, transformer-based referring segmentation methods have provided a unified framework for handling different modalities [19], advancing research in early fusion. Compared to late fusion, early fusion enables cross-modal interactions at the feature encoding stage, achieving deeper feature integration by introducing additional modality-guided learning during the feature extraction process. However, most existing early fusion methods adopt a unidirectional fusion strategy from language to vision, lacking guidance from the visual modality in the learning of linguistic features, leading to linguistic representations that lack the contextual information provided by the image. To address this issue, CoupAlign [20] introduces precise multimodal alignment through word-pixel and sentence-mask alignment modules, while CrossVLT [21] incorporates a bidirectional fusion module at the encoding stage to enhance interactions between visual and linguistic features. These approaches effectively embed multimodal features using cross-attention mechanisms. However, as cross-attention heavily relies on large-scale matrix multiplications, it significantly increases computational complexity, complicating model optimization.

According to the above analysis, in this paper, we propose a multimodal prompt-guided bidirectional fusion (MPBF) architecture. MPBF simultaneously employs early and late fusion strategies, leveraging multimodal prompts to facilitate efficient feature interaction and semantic alignment, thereby enhancing the deep fusion of multimodal features in RRSIS. In the early fusion stage, we embed learnable prompts into both image and text encoders, and impose a constraint mapping between the prompts from both modalities via coupling functions, which project linguistic and visual features into a shared latent space for deep fusion, enabling cross-modal collaboration with a small amount of additional parameters. In the late fusion stage, visual and linguistic features are further aligned at the pixel level through a scale-aware visual–language fine-grained alignment decoder. To tackle the multi-scale nature of remote sensing targets, we capture diverse scale representations through parallel multi-branch receptive field fusion. Additionally, several visual–language sequential interaction (VLSI) processes are employed to conduct deep multimodal feature interaction. The multimodal prompts generated in the early fusion stage are also integrated into VLSI processes during the late fusion stage. This integration provides prior knowledge of cross-modal coordination from the encoding stage, enhancing the model's ability to perceive fine-grained features.

The main contributions of this article can be summarized as follows:

- We apply multimodal prompting to RRSIS and propose a novel architecture termed multimodal prompt-guided bidirectional fusion (MPBF). This architecture leverages multimodal prompts as a medium for bidirectional fusion of visual and linguistic features. By coupling learnable prompt vectors embedded within image and text encoders, MPBF achieves deep integration of multimodal features in a shared latent space with minimal additional parameters.
- To enhance fine-grained segmentation, we introduce a scale-aware visual–language fine-grained alignment decoder. This decoder incorporates a scale refinement (SR) module that significantly improves the model's ability to learn fine-grained scale representations. Additionally, the VLSI process establishes pixel-level semantic associations between visual and linguistic features, further enhancing the model's robustness in recognizing multi-scale targets.
- Performance evaluations on public datasets demonstrate that our method achieved state-of-the-art (SOTA) results. Validation experiments in real-world application scenarios further confirm that our approach retains high segmentation accuracy, even in complex environments with substantial background noise interference, highlighting its practical value.

2. Related Work

2.1. Referring Image Segmentation

Referring image segmentation (RIS) is an intersecting research area in computer vision and natural language processing (NLP). One of the earliest attempts in this field was based on convolutional neural networks (CNN) and LSTM [22] architectures. Hu et al. [3] were the first to propose and implement the RIS task, employing VGG-16 and LSTM as the image and text encoders, respectively. The extracted visual and linguistic features were concatenated and then fused using a fully convolutional network to generate the segmentation mask. Building on this work, CNN+LSTM became the dominant framework in early studies [23–26]. For example, Liu et al. [23] introduced a multimodal LSTM that recursively models word-level features in sentences; Margffoy-Tuay et al. [25] designed dynamic filtering kernels from linguistic features to refine text-relevant image regions; Chen et al. [26] integrated referring expression generation with segmentation, enforcing consistency between the text and segmentation masks to enhance both linguistic and visual representations.

Despite their simplicity, CNN+LSTM methods suffer from inherent limitations. CNNs struggle to capture long-range dependencies in images, and the recursive structure of LSTMs is inefficient for processing long text sequences. Consequently, these architectures show poor adaptability to long-text descriptions and multi-object scenarios.

With the transformer architectures demonstrating outstanding performance in natural language processing and vision tasks, transformer-based approaches have become the state-of-the-art in RIS. For instance, LAVT [27] employed Swin transformer [28] as the image encoder, integrated BERT [29] for text embeddings, and utilized cross-attention mechanisms for unidirectional fusion from linguistic to visual features. Following this work, a multitude of transformer-based methods have emerged. SADLR [30] introduced semantic-aware dynamic convolutions to enhance target pixel discrimination, while RefSegFormer [12] incorporated target existence prediction, improving robustness against missing target scenarios.

In the field of RRSIS, LGCE [1] utilized a language-guided cross-scale enhancement module to improve segmentation performance for small and sparsely distributed targets. Meanwhile, RMSIN [2] addressed scale variations and rotation challenges in remote sensing images by incorporating multi-scale receptive fields at the encoding stage and employing

multi-scale interaction and rotation convolutions at the decoding stage, effectively handling complex remote sensing scenes.

All the above methods adopt a unidirectional fusion strategy, where linguistic features guide the vision encoder to focus on text-related regions while suppressing irrelevant areas, enabling deep text-driven guidance for image segmentation. However, as these methods only consider unidirectional guidance, they may struggle to capture complex interactions between textual descriptions and visual content. To address this limitation, this paper introduces a bidirectional fusion strategy based on the transformer-based architecture to establish mutual multimodal guidance, enabling the deep integration of multimodal features.

2.2. Prompt Learning

Prompt learning [31] originated in natural language processing as a method that guides pre-trained models to perform downstream tasks using input prompts. By introducing a small amount of additional parameters, prompt learning enables pre-trained models to adapt to new tasks efficiently. Compared to traditional fine-tuning, which modifies the entire model for each specific task, prompt learning significantly reduces memory requirements while mitigating the risk of poor adaptation or loss of generalization when training data are limited.

The success of prompt learning in NLP has accelerated its adoption in computer vision, particularly in multimodal learning [32–34]. To leverage CLIP's strong zero-shot learning capabilities for downstream tasks, CoOp [35] was the first to introduce prompt learning in vision–language models. It incorporated learnable continuous prompts into text descriptions, enabling the model to adaptively learn optimal prompt representations even in few-shot scenarios. However, CoOp exhibited poor generalization to unseen categories. To address this, CoCoOp [36] extended CoOp by integrating sample-specific context extracted from images, allowing the model to dynamically adjust learnable prompts based on the input, thereby improving generalization to novel categories. These studies primarily focused on prompt learning within the text branch of multimodal models, where prompts were embedded in the language modality. In contrast, VPT [37] explored the application of prompt learning to vision models. Based on these advancements, MAPLe [38] introduced multimodal prompts as a bridge to facilitate coordinated learning across modalities.

Inspired by these developments, this paper extends the concept of multimodal prompts to RRSIS. We integrate multimodal prompts into independent vision and language models, utilizing cross-modal prompt coupling to establish deep interactions between the two modalities and enhance segmentation performance.

3. Methodology

In this section, we introduce MPBF, a visual–linguistic feature bidirectional fusion architecture based on multimodal prompts. By integrating early fusion and late fusion strategies, MPBF enables fine-grained alignment of multimodal features, thereby enhancing the segmentation performance of RRSIS. Specifically, we first provide an overview of MPBF. Then, we describe the structure of each component of MPBF in detail.

3.1. Model Overview

The overall architecture of MPBF is illustrated in Figure 1. MPBF consists of three main components: a text encoder, an image encoder, and a scale-aware vision–language fine-grained alignment decoder. In the encoding phase, MPBF enhances deep multimodal fusion while maintaining a low parameter count by inserting learnable prompts into both linguistic and visual features at each stage of feature encoding. This approach also introduces task-

specific knowledge into the pre-trained text and image encoders. To facilitate cross-modal collaboration, a prompt coupling function is employed to establish strongly constrained mappings between the prompts of the two modalities, projecting linguistic and visual features into a shared latent space. In the decoding phase, the multi-scale visual features extracted by the image encoder are processed by an SR module, which enhances the multi-scale representation and enriches the fine-grained scale details in the feature maps. The refined image features are then iteratively aligned with linguistic features through a VLA module, under the guidance of multimodal prompts. This process achieves fine-grained multimodal fusion while maintaining semantic consistency across stages by leveraging prior information in the multimodal prompts.

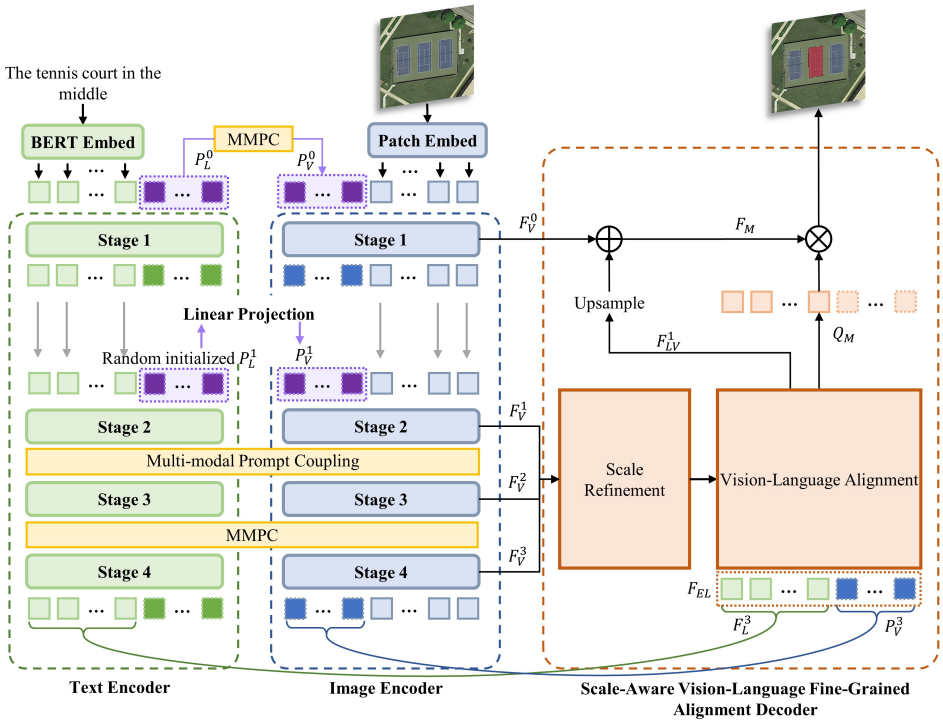


Figure 1. An overview of multimodal prompt-guided bidirectional fusion (MPBF).

3.2. Multi-Stage Image and Text Encoder

Early fusion involves interaction between multimodal features at intermediate states during the encoding phase. Therefore, both the image and text encoders should be divided into multiple feature extraction stages. In this paper, we adopt a Swin transformer with four feature extraction stages as the image encoder. For the input image $I \in \mathbb{R}^{H\times W\times C}$, the Swin transformer performs patch merging at the end of each stage to generate multi-scale feature maps F_V^i $(i = 0, 1, 2, 3)$. Learnable visual prompts are inserted at each stage before the visual features are fed, which interact with visual features in the self-attention module to capture rich contextual information in the image. A pre-trained BERT model is employed as the text encoder, which contains 12 feature extraction layers. To align with the architecture of the image encoder, we divide BERT into four stages with 3 layers per stage. Similar to the image encoder, learnable text prompts are concatenated with the linguistic features before each feature encoding stage.

3.3. Multimodal Prompt-Based Cross-Modal Collaborative Encoding

The fusion method of visual and language prompts is crucial in early fusion, as incorporating information from other modalities in the encoding stage can guide the model to focus on target information during feature learning. To enable mutual guidance between

visual and linguistic features, existing research learns the correlations between these two modalities through cross-attention and embeds relevant information into each modality's feature space, thereby facilitating cross-modal information exchange. However, cross-attention typically relies on multiple linear layers to map features from different modalities into shared subspaces suitable for attention computation. This not only introduces a substantial number of additional parameters but also significantly increases computational complexity, as attention must be computed across all image patches and word tokens during the encoding stage.

To address this issue, we introduce the concept of prompt learning to achieve efficient multimodal feature fusion, offering a novel and parameter-efficient paradigm for modality alignment. Specifically, at each feature extraction stage, we insert a set of learnable prompts for both visual and linguistic features, denoted as $P_V^i \in \mathbb{R}^{N \times C_i^v}$ and $P_L^i \in \mathbb{R}^{N \times C_i^l}$, where $i = 0, 1, \ldots, n$, n represents the encoding stage index, N is the number of learnable prompts (with $N = 5$ in this paper), and C_i^v and C_i^l are the feature dimensions of the image encoder and text encoder, respectively. Prompts at different stages are used to construct feature representations at different levels, meaning that the prompts are not shared among stages. However, within the same feature encoding stage, the prompts are shared across the layers that compose that stage, thereby preserving the correlations among features at the same level.

Multimodal prompt coupling: During the feature encoding process, P_V^i and P_L^i serve as intermediaries for cross-modal interactions. A prompt coupling function is used to achieve collaborative learning of multimodal information, expressed as follows:

$$P_V^i = \mathcal{F}_i(P_L^i), i = 0, 1, \ldots, n, \tag{1}$$

where $\mathcal{F}$ denotes the prompt coupling function, which is implemented as a linear mapping from P_L^i to P_V^i, effectively embedding them into a shared latent subspace. This design implies that P_L^i and P_V^i are inherently linked and co-learned within a common space, ensuring that the resulting cross-modal feature representations are mutually informative and jointly serve both visual and linguistic modalities. This naturally enables multimodal prompts to possess the capability for deep fusion of multimodal features without additional handcrafted designs. Thus, these prompts act as a natural bridge for cross-modal interaction.

During training, the gradient update directions for P_L^i and P_V^i are highly correlated. As a result, effective information extracted during the visual encoding process is backpropagated through the coupling function to refine the language prompts, and vice versa. This bidirectional gradient flow facilitates synchronized information acquisition and exchange between modalities. Therefore, the model forms a consistent joint distribution representation of visual and language modalities at the encoding stage, precluding the modality inconsistency often observed in independently learned representations.

It is important to emphasize that, despite the linear relationship established by the coupling function, the visual and textual prompts are not equivalent. Under the supervision of the segmentation loss, the visual prompts not only encapsulate shared semantic content but also encode purely visual semantics that are either absent from or insufficiently described in the text descriptions.

In terms of parameter overhead, the additional parameters introduced by multimodal prompts primarily originate from two components: the prompt vectors and the coupling function. The parameter count of the prompt vectors scales linearly with the number of prompts N:

$$Params_{prompts} = N \times C_i^l, i = 0, 1, \ldots, n. \tag{2}$$

In this work, a linear projection is employed as the coupling function, whose parameter count is determined by the input and output channel dimensions and can be computed as follows:

$$Params_{coupling} = C_i^l \times C_i^v, i = 0, 1, \ldots, n. \tag{3}$$

Therefore, the total parameter increase introduced by multimodal prompts during the early fusion stage is as follows:

$$Params_{MP} = Params_{prompts} + Params_{coupling}. \tag{4}$$

By contrast, cross-attention introduces significantly more parameters, as separate linear projections are required for the query, key, and value representations. These projections lead to parameter growth that is quadratic with respect to the feature dimension (C_i^l and C_i^v). Consequently, compared to cross-attention-based modality alignment, the proposed multimodal prompts offer a more parameter-efficient alternative. Moreover, the use of a shared latent space imposes strong representational constraints, allowing the coupled prompts to naturally achieve deep multimodal fusion.

3.4. Scale-Aware Vision–Language Fine-Grained Alignment Decoder

The widely distributed multi-scale targets in remote sensing images pose a dual challenge for RIS. On the one hand, such targets require models to learn diverse multi-scale feature representations accurately. On the other hand, due to the significant differences between modalities, the model needs to establish pixel-level semantic associations between text descriptions and images. To further enhance the fine-grained cross-modal alignment of multimodal features, we implement the late fusion of multimodal features in the decoder. Similar to [20,39,40], the decoder is based on the basic structure of the decoder in [41,42]. As shown in Figure 1, the decoder takes multi-level visual features F_V^i, along with the linguistic features F_L^3 and multimodal prompts P_V^3 as inputs.

For visual features, we first select $F_V^1 \in \mathbb{R}^{H_1 \times W_1 \times C_1}$, $F_V^2 \in \mathbb{R}^{H_2 \times W_2 \times C_2}$ and $F_V^3 \in \mathbb{R}^{H_3 \times W_3 \times C_3}$ for linear projection to unify the channel dimension to D. Then, the features from these 3 scales are spatially flattened and concatenated to form a multi-scale feature $F_{MS} \in \mathbb{R}^{S \times D}$, where $S = \sum_{i=1}^{3} H_i W_i, i = 1, 2, 3$. To enhance the model's ability to learn the rich scale variations present in remote sensing images, F_{MS} is processed by an SR module, which captures fine-grained scale representations, yielding fused multi-scale feature representations F_{MR}. For linguistic features, we concatenate F_L^3 output by the text encoder and P_V^3 output by the image encoder to obtain the expanded linguistic features F_{EL}. F_{MR} and F_{EL} undergo fine-grained vision–language alignment through VLA. The aligned visual features F_{LV}^1 are integrated with F_V^0 to produce per-pixel embedded mask features F_M. Concurrently, the aligned linguistic features are used to generate mask queries Q_M. The final predicted segmentation mask is obtained by performing matrix multiplication between F_M and Q_M.

3.4.1. Scale Refinement

The multi-scale characteristics of targets are a key distinguishing factor between remote sensing images and natural images. Due to the image acquisition altitude, objects in remote sensing images span a wide range of scales, with each scale encompassing multiple types of land cover and objects. To enhance the model's ability to perceive and distinguish multi-scale feature representations, we propose an SR module with a multiple receptive field fusion (MRF) mechanism. This mechanism further enriches scale representations in feature maps, enabling the model to better capture fine-grained scale variations in remote sensing images, ultimately improving segmentation accuracy.

The structure of SR is illustrated in Figure 2. First, a linear projection expands the channel dimension of F_{MS} to D', thereby increasing the model's representational capacity and enabling each channel to discriminate features at a finer granularity. The features are then decomposed according to the original concatenation ratios and reshaped along the spatial dimensions to obtain three feature maps $F_V^{i'} \in \mathbb{R}^{H_i \times W_i \times D'}, i = 1, 2, 3$. Here, $F_V^{i'}$ maintains the same spatial scale as F_V^i, ensuring the hierarchical independence of multi-scale features. Subsequently, each decomposed $F_V^{i'}$ undergoes feature enhancement through an MRF mechanism. Specifically, in MRF, $F_V^{i'}$ is further split along the channel dimension into two parts, $F_{V1}^{i'} \in \mathbb{R}^{H_i \times W_i \times D'/2}$ and $F_{V2}^{i'} \in \mathbb{R}^{H_i \times W_i \times D'/2}$, which are processed separately by depthwise convolutions (DWConv) with different kernel sizes. As shown in Figure 2, we employ 3×3 DWConv for $F_{V1}^{i'}$ and 5×5 DWConv for $F_{V2}^{i'}$, yielding feature maps $E_{V1}^{i'}$ and $E_{V2}^{i'}$ that capture complementary contextual cues under distinct receptive fields. These two maps are concatenated along the channel dimension to form $E_V^{i'}$. Each $E_V^{i'}$ is flattened in the spatial dimension and concatenated along the channel dimension. A linear projection is adopted to restore the channel dimension to D. Finally, $F_{MR} \in \mathbb{R}^{S \times D}$ is output, which incorporates fine-grained multi-scale representations.

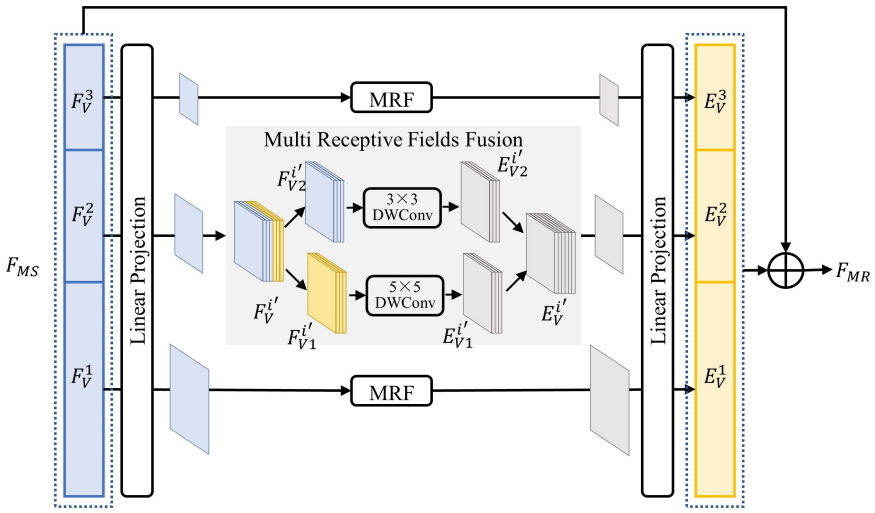


Figure 2. Scale refinement (SR) module.

3.4.2. Vision–Language Alignment

In the late fusion stage, to achieve fine-grained cross-modal alignment, the multi-scale visual features F_{MR} output by the scale refinement module are aligned with linguistic features $F_{EL} \in \mathbb{R}^{(L+N) \times D}$ through the VLA module.

Text descriptions typically focus on image content directly related to the target object. Consequently, the linguistic features $F_L^3 \in \mathbb{R}^{L \times d_L}$, extracted by the text encoder, primarily contain semantic information closely associated with the target. However, other land cover information in the image, besides the target, can also provide valuable context for target identification, and such features are often not included in linguistic features. Therefore, we concatenate multimodal prompts $P_V^3 \in \mathbb{R}^{N \times C_3}$, output by the image encoder, with F_L^3 along the spatial dimension to obtain expanded linguistic features F_{EL}. The channel dimension of F_L^3 and P_V^3 is unified to D by linear projection. Through this process, consistent joint distribution representation of visual and language modalities at the encoding stage, embedded in P_V^3, is propagated to the late fusion stage. Additionally, it incorporates image semantic information beyond textual descriptions into the linguistic features.

The alignment of visual and linguistic features in the VLA module directly impacts the accuracy of mask prediction. As illustrated in Figure 3, to achieve fine-grained align-

ment between F_{EL} and the rich scale representations in F_{MR}, we employ VLSI with a sequential structure.

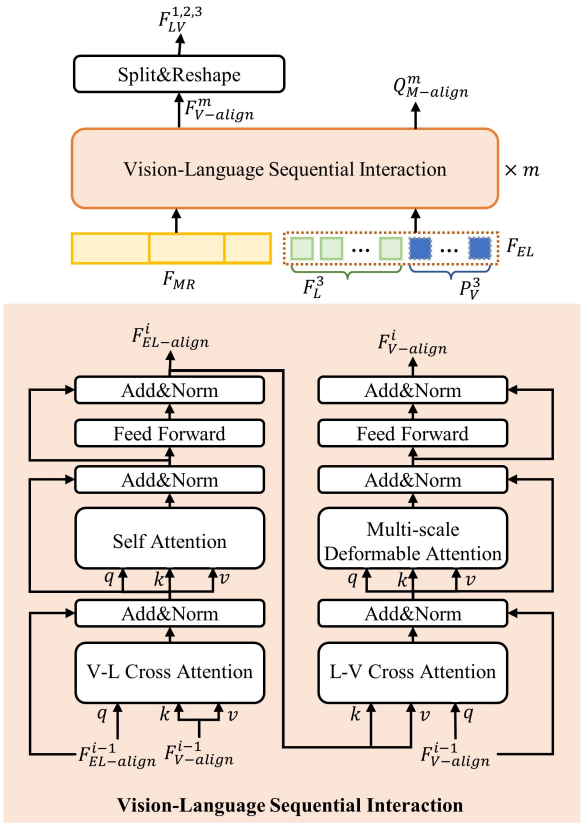


Figure 3. The structure of the vision–language alignment (VLA) module.

VLSI sequentially performs vision-to-language and language-to-vision embedding processes. During the i-th iteration of the VLSI process, for the input $F_{V\text{-}align}^{i-1}$ and $F_{EL\text{-}align}^{i-1}$, the alignment process can be expressed as follows:

$$F_{EL\text{-}align}^{i} = FFN(SA(CrossAttn(F_{EL\text{-}align}^{i-1}, F_{V\text{-}align}^{i-1}))), \tag{5}$$

$$F_{V\text{-}align}^{i} = FFN(MSDA(CrossAttn(F_{V\text{-}align}^{i-1}, F_{EL\text{-}align}^{i}))). \tag{6}$$

VLSI first embeds visual features into linguistic features. To associate linguistic features with the multi-scale representations learned in the SR module, $F_{EL\text{-}align}^{i-1}$ is initially aligned with $F_{V\text{-}align}^{i-1}$ through cross-attention, denoted as $CrossAttn(\cdot)$, from vision to language. In this process, $F_{EL\text{-}align}^{i-1}$ serves as the query (q), while $F_{V\text{-}align}^{i-1}$ serves as the key (k) and value (v) for embedding fine-grained visual features. The linguistic features, after being further processed by self-attention and a feed-forward neural network (FFN), denoted as $FFN(\cdot)$, yield $F_{EL\text{-}align}^{i}$. In the next step, $F_{EL\text{-}align}^{i}$ is fed into the visual feature processing pipeline. During this stage, language-to-vision cross-attention is applied between $F_{EL\text{-}align}^{i}$ and $F_{V\text{-}align}^{i-1}$, embedding fine-grained semantic relationships between the two modalities into visual features. Specifically, $F_{V\text{-}align}^{i-1}$ serves as q, while $F_{EL\text{-}align}^{i}$ serves as both k and v. Since the previous operation has already performed an initial alignment between the two modal features, this step further optimizes the alignment on this foundation. The aligned visual features are then processed by multi-scale deformable attention (MSDA) [43], denoted as $MSDA(\cdot)$. MSDA dynamically adjusts the attention sampling based on target features at different scales, thereby establishing global dependencies across scales. Finally, the refined features are processed by FFN, with the final output represented as $F_{V\text{-}align}^{i}$.

The VLA module consists of six VLSI processes, i.e., $m = 6$. The visual features output by the final VLSI process are re-split and reshaped into three feature maps $F_{LV}^{1,2,3}$ in different scales, which are finely aligned with the linguistic features.

3.4.3. Mask Generation

To generate a segmentation mask with accurate edges, the highest resolution features F_{LV}^{1}, aligned by the VLA module, are combined with the F_{V}^{0} output from the image encoder, yielding mask features F_M with per-pixel embeddings, which are used to generate the segmentation mask. This process is formulated as follows:

$$F_M = upsample(F_{LV}^{1}) + proj(F_{V}^{0}), \tag{7}$$

where $upsample(\cdot)$ denotes bilinear interpolation, and $proj(\cdot)$ represents linear projection. After linear projection, F_{V}^{0} retains the same channel dimensions as F_{V}^{1}.

In the RRSIS task, the prediction of target masks is essentially a binary classification of "foreground" and "background". To predict target masks, the aligned linguistic features $F_{EL-align}^{m}$ are divided into two components: the foreground semantic features $F_{align}^{f} \in R^{L \times D}$ containing target information, and the background semantic features $F_{align}^{b} \in R^{N \times D}$ containing other object-related information. The first layer of F_{align}^{f} serves as a classification ([CLS]) token, representing the overall information of the text description. We select this layer as the foreground query $Q_M^f \in \mathbb{R}^{1 \times D}$. Meanwhile, the background query $Q_M^b \in \mathbb{R}^{1 \times D}$ is obtained from the mean value of $P_{V\text{-}align}^{m}$. Finally, these two queries are concatenated to form the mask queries $Q_M \in \mathbb{R}^{2 \times D}$. The segmentation mask $M \in \mathbb{R}^{H \times W \times 2}$ is obtained by performing matrix multiplication between F_M and Q_M. In this process, each channel of F_M represents a mask prediction for a specific object or part of the image. Through matrix multiplication, Q_M assigns category weights to each predicted mask, and masks of the same category are merged, yielding the segmentation mask for the target.

4. Experiments and Results

The experiments in this section are divided into two parts: first, comparative experiments and ablation studies based on public datasets; second, application experiments using a self-annotated dataset. In the comparative experiments and ablation studies, we conducted quantitative and qualitative analyses against several representative methods, validating the superior performance of our proposed method. To further verify the contribution of each module in the model, we performed detailed ablation studies, analyzing the impact of each module on overall performance and confirming their necessity and effectiveness within the model. In the application experiments, we focused on the practical scenario of hazard detection along high-speed rail lines. We applied our proposed method to a custom high-speed rail hazard multimodal segmentation (HSRMS) dataset, verifying the effectiveness of our method in real-world applications.

4.1. Datasets

4.1.1. RRSIS-D Dataset

In the comparative experiments and ablation studies, we selected the publicly available RRSIS-D dataset [2] for RRSIS as the experimental dataset. Figure 4 shows examples from the RRSIS-D dataset. The text descriptions are below the images, with the red annotations representing the target masks corresponding to the texts.

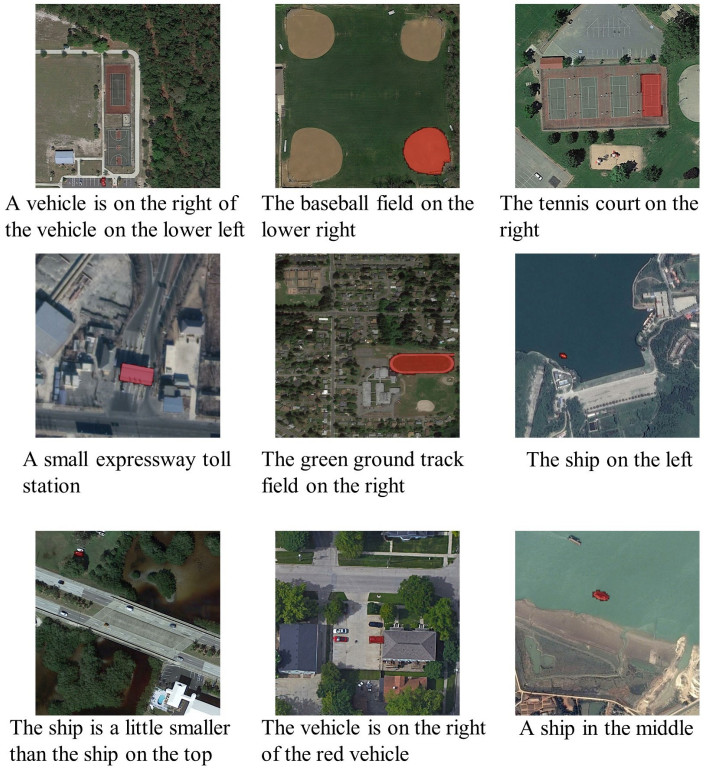


Figure 4. Samples of the RRSIS-D dataset.

The RRSIS-D dataset contains 17,402 text-image pairs, where each textual description corresponds to a single target within the image. The images are uniformly sized at 800 × 800 pixels, with ground sampling distances ranging from 0.5 m to 30 cm. There are 20 distinct categories across all text descriptions. A notable characteristic of RRSIS-D is the high proportion of small targets, with most objects occupying less than 0.1 of the image area. However, the dataset also includes large-scale objects exceeding 400,000 pixels, highlighting significant scale variations among the targets. As a result, the RRSIS-D dataset presents two major challenges for RRSIS: (1) accurate segmentation of a large number of small targets, which are often difficult to distinguish, and (2) handling significant scale variations among objects. These characteristics make RRSIS-D particularly suitable for evaluating the robustness and effectiveness of RRSIS methods.

4.1.2. HSRMS Dataset

In the application experiments, we utilized a custom HSRMS dataset focusing on two typical hazards along high-speed rail lines: plastic greenhouses and color-coated steel sheet (CCSS) roof buildings. The images in the dataset were collected by two satellites: GaoFen-2 and SuperView-1. The ground sampling distance for GaoFen-2 images was 1 m, while for SuperView-1 satellite images, it was 0.5 m. The imagery covered a total area of 896 km^2 along the main line of the Beijing–Zhangjiakou high-speed railway, encompassing a diverse range of natural and man-made environments.

To preserve the authenticity of real-world remote sensing scenes, the dataset construction process did not involve manual filtering or artificial target enhancement. Only unlabeled samples were removed, ensuring that the data distribution closely reflects the complexity and variability encountered in practical remote sensing tasks. The final dataset contains 37,549 text-image pairs, with 18,646 pairs in the training set, 6166 pairs in the validation set, and 12,737 pairs in the test set. The images are uniformly sized at 512 × 512 pixels.

Each textual description typically consists of two key components: the type of hazard and its spatial location within the image. The spatial location is annotated based on the centroid of the connected region corresponding to the target. For images that contain railways or roads, the descriptions often include the relative position of the hazard with respect to the transportation infrastructure. The details of the dataset construction are described in [44]. Table 1 presents the scale distribution of the targets in the dataset. As shown in Table 1, the total area and standard deviation of CCSS roof buildings are both larger than those of plastic greenhouses, suggesting that CCSS roof buildings exhibit greater scale variation. This scale variability is one of the key factors affecting the accuracy of intelligent hazard target extraction. Figure 5 shows examples from the HSRMS dataset for hazards along high-speed rail lines.

Table 1. Hazard target statistics of the HSRMS dataset.

Satellite Model	**SuperView-1**		**GaoFen-2**	
Hazard Type	**CCSS Roof Building**	**Plastic Greenhouse**	**CCSS Roof Building**	**Plastic Greenhouse**
Hazard Number	25,711	7889	21,613	7257
Minimum Area (m^2)	1.14	0.12	1.14	0.12
Maximum Area (m^2)	29,692.80	10,920.70	29,692.80	3609.65
Total Area (m^2)	9,781,844.09	3,786,570.52	8,176,613.10	3,121,702.59
Mean Area (m^2)	380.45	479.98	378.32	430.16
Area Standard Deviation (m^2)	932.36	448.29	994.83	241.81

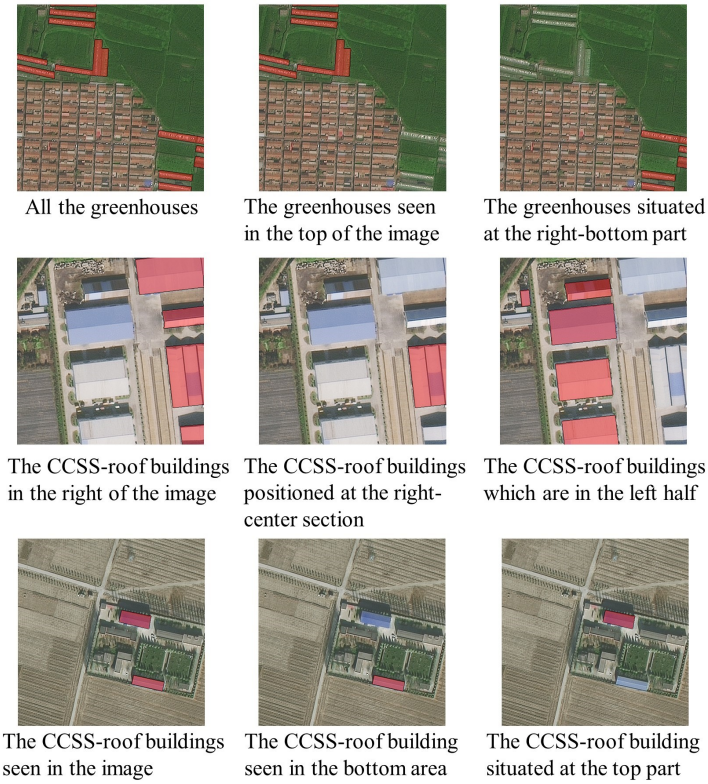


Figure 5. Samples of the high-speed rail hazard multimodal segmentation (HSRMS) dataset.

4.2. Comparison Algorithms

We selected three RIS methods, LAVT, RefSegFormer, and CrossVLT, along with an RRSIS method RMSIN, for comparison. Among these methods, LAVT, RefSegFormer, and RMSIN perform unidirectional fusion from language to vision during the encoding

stage, while CrossVLT adopts a bidirectional fusion strategy between visual and linguistic features, enabling deeper multimodal feature interaction. The quantitative and qualitative results in the comparison experiments are evaluated on the test set and validation set of the dataset.

4.3. Evaluation Metrics

For quantitative analysis, we selected the overall intersection over union (oIoU), the mean intersection over union (mIoU), and Pre@X as evaluation metrics. oIoU measures the intersection over union (IoU) ratio between the predicted segmentation mask and the ground truth by computing the total intersection area divided by the total union area across all test samples. Both the intersection and union areas were accumulated over the entire test (or validation) set. mIoU is the mean IoU across all test samples, averaging the IoU scores for each prediction. Pre@X represents the percentage of test samples, where the IoU score exceeds a threshold X. In our experiments, X was set to $\{0.5, 0.6, 0.7, 0.8, 0.9\}$.

4.4. Experimental Settings

In this study, the base version of the BERT model was employed as the text encoder, initialized with pre-trained weights. BERT comprises 12 layers with a hidden dimension of 768 and a maximum text length set of 20. For the image encoder, the base version of the Swin transformer was utilized, initialized with weights pre-trained on the ImageNet-22K dataset. The model training process employed the AdamW optimizer with a weight decay value of 0.01. The initial learning rate was set to 0.00005, and it was adjusted using the poly learning rate schedule during training. The learning rate was updated according to the following formula:

$$lr = base\_lr \times \left(1 - \frac{epoch}{num\_epoch}\right)^{power}, \tag{8}$$

where lr denotes the updated learning rate, $base\_lr$ is the initial learning rate, and $epoch$ and num_epoch represent the current epoch and total number of epochs, respectively. $power$ controls the decay rate of the learning rate, set to 0.9 in this study. MPBF was trained using cross-entropy loss as the loss function. All experiments were conducted on a single NVIDIA RTX 3090 GPU. The parameter settings for all comparison methods were kept consistent with those reported in the original papers or official code repositories.

4.5. Comparative Experiments

The RRSIS-D dataset contains a variety of types and scales of remote sensing targets, posing higher demands on both the fine-grained alignment capabilities and multi-scale learning abilities of algorithms. Based on the RRSIS-D dataset, we conducted a quantitative and qualitative analysis of the experimental results for all comparison methods, along with an evaluation of their computational complexity.

4.5.1. Quantitative Comparison

The quantitative comparison results of all methods on the test and validation sets are presented in Table 2. The proposed method achieves optimal performance across all evaluation metrics. Specifically, on the test set, our method improves the oIoU and mIoU scores by 0.42% and 0.67%, respectively, compared to the second-best RMSIN. On the validation set, the improvements are 0.8% and 0.43%, respectively, demonstrating a significant performance gain. In addition to the metrics used for overall evaluation, our method also excels in the proportion of samples exceeding five IoU thresholds, indicating a comprehensive improvement in segmentation accuracy across all samples. On the test set, the proportion of high-accuracy samples with IoU > 0.9 (where the ideal value is 1) reaches

26.57%, representing a 2.38% improvement over RMSIN. On the validation set, this metric shows an improvement of 1.44% compared to the second-best CrossVLT.

Table 2. Quantitative comparison results on the RRSIS-D dataset (best values in bold; second-best values underlined).

Method	Dataset Type	Pr@0.5 (%)	Pr@0.6 (%)	Pr@0.7 (%)	Pr@0.8 (%)	Pr@0.9 (%)	oIoU (%)	mIoU (%)
LAVT	Test	69.26	62.28	52.08	40.33	23.64	76.04	60.58
	Validation	70.06	62.59	52.36	42.30	24.89	76.68	61.40
CrossVLT	Test	70.30	63.80	53.81	42.20	25.92	75.92	61.45
	Validation	70.11	63.05	53.51	42.93	<u>25.63</u>	77.44	61.89
RefSegFormer	Test	65.59	59.45	50.89	39.35	23.15	76.39	58.17
	Validation	68.51	60.92	52.76	42.18	25.40	76.50	59.11
RMSIN	Test	<u>75.44</u>	<u>68.31</u>	<u>56.33</u>	<u>43.03</u>	<u>24.19</u>	<u>77.78</u>	<u>64.65</u>
	Validation	<u>75.29</u>	<u>68.22</u>	<u>56.78</u>	<u>44.02</u>	24.60	<u>77.57</u>	<u>65.25</u>
MPBF	Test	**75.90**	**69.18**	**57.84**	**44.04**	**26.57**	**78.20**	**65.32**
	Validation	**75.63**	**69.08**	**58.79**	**46.55**	**27.07**	**78.37**	**65.68**

4.5.2. Qualitative Comparison

To provide an intuitive analysis of the improvements achieved by the proposed method in RRSIS, we conducted a qualitative comparison of MPBF and other methods on the RRSIS-D dataset. Qualitative comparison results are illustrated in Figures 6 and 7, which display representative samples selected from the test set.

As analyzed earlier, complex backgrounds and small-scale targets are key factors that affect segmentation accuracy in RRSIS. In sample (a), the target train station is surrounded by a dense cluster of buildings that share similar geometric features. Moreover, these buildings are closely adjacent to the train station, making it challenging to accurately distinguish the station from the surrounding structures.

As illustrated in Figure 6a, except for the method proposed in this paper, all other methods incorrectly included the buildings adjacent to the train station in their predicted masks. This confusion arose from the unclear semantic distinction between the two types of objects, which ultimately led to a decrease in segmentation accuracy. In sample (b), the text description contains both location information ("lower-right") and target category attributes ("vehicle"). However, since the vehicle in the lower-right corner of the image has significantly fewer distinctive features compared to the surrounding buildings and trees, RefSegFormer, CrossVLT, and RMSIN failed to accurately associate the textual semantics with the target features, thus producing no segmentation results. While LAVT correctly interpreted the positional information "lower-right", it mistakenly identified a white object above the house in the lower-right corner as the target vehicle. In contrast, the method proposed in this study accurately extracted the target mask through fine-grained linguistic and visual feature alignment, demonstrating superior multimodal segmentation performance.

In addition to common positional and categorical attributes, text descriptions often contain a diverse range of complex high-level semantic features, requiring the model to make a comprehensive analysis by integrating these semantic cues. For instance, the text description in sample (c) not only specifies the target category ("ship") but also describes its color attributes ("red and black"). This requires the model to accurately interpret the color description and align it with the corresponding visual features in the image. While our method successfully produced accurate segmentation results, all other comparison methods misidentified the target due to the similar colors of the two ships in the input image. These methods partially segmented the left-side object. The superior performance of the proposed method in this scenario highlights its robust capability in fine-grained feature

matching and semantic understanding. This advantage becomes particularly evident in situations where targets share similar visual characteristics.

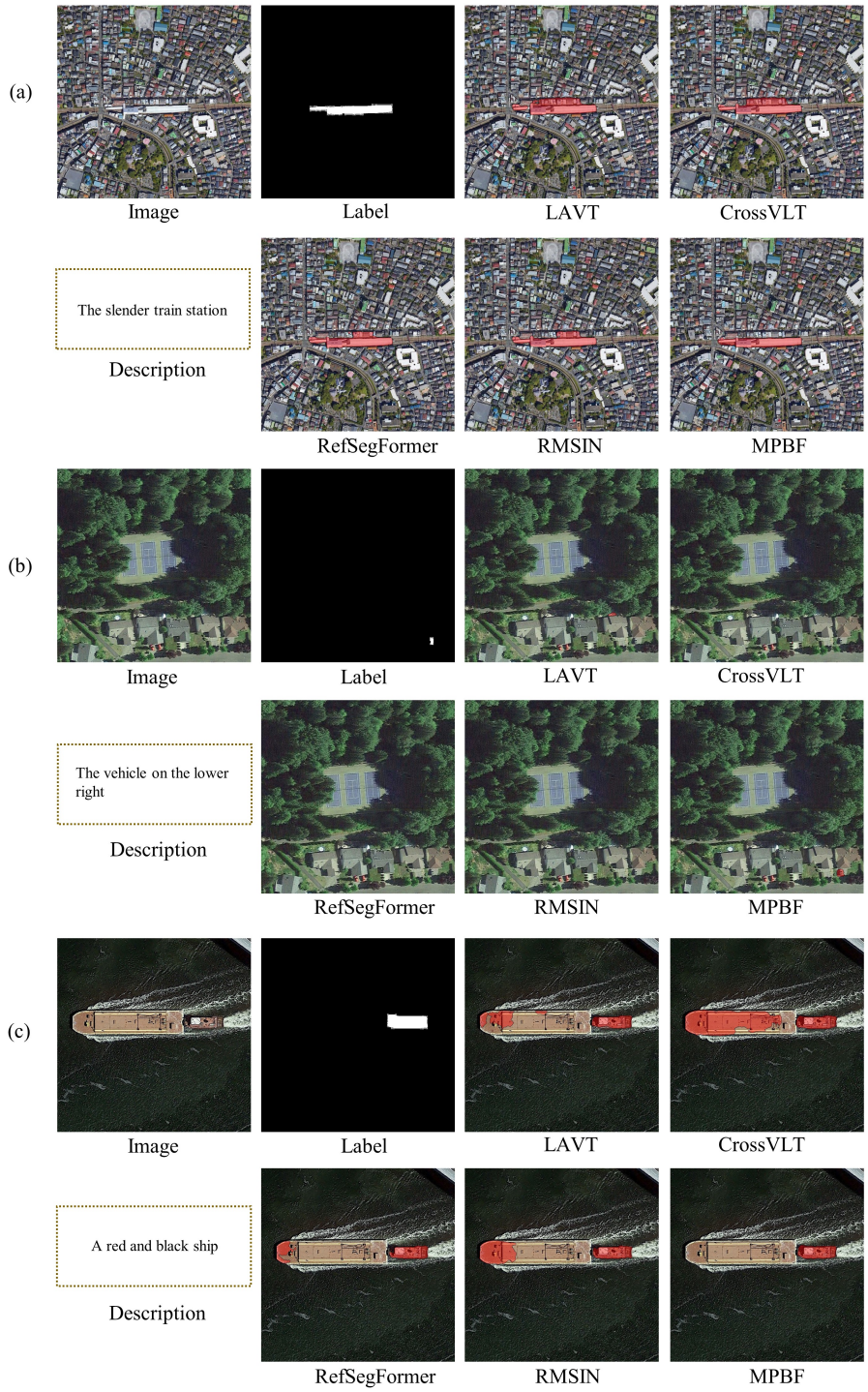


Figure 6. Qualitative comparisons on the samples (**a**–**c**) from the RRSIS-D dataset, which have concise text descriptions. The red annotations representing the predicted masks, the same below.

Compared to the previous samples, samples (a) and (b) in Figure 7 present greater challenges in semantic understanding, as their text descriptions mention multiple objects (both the segmentation target and reference object) along with relative spatial relationships between them. These samples require the model to have a stronger capability for multimodal reasoning.

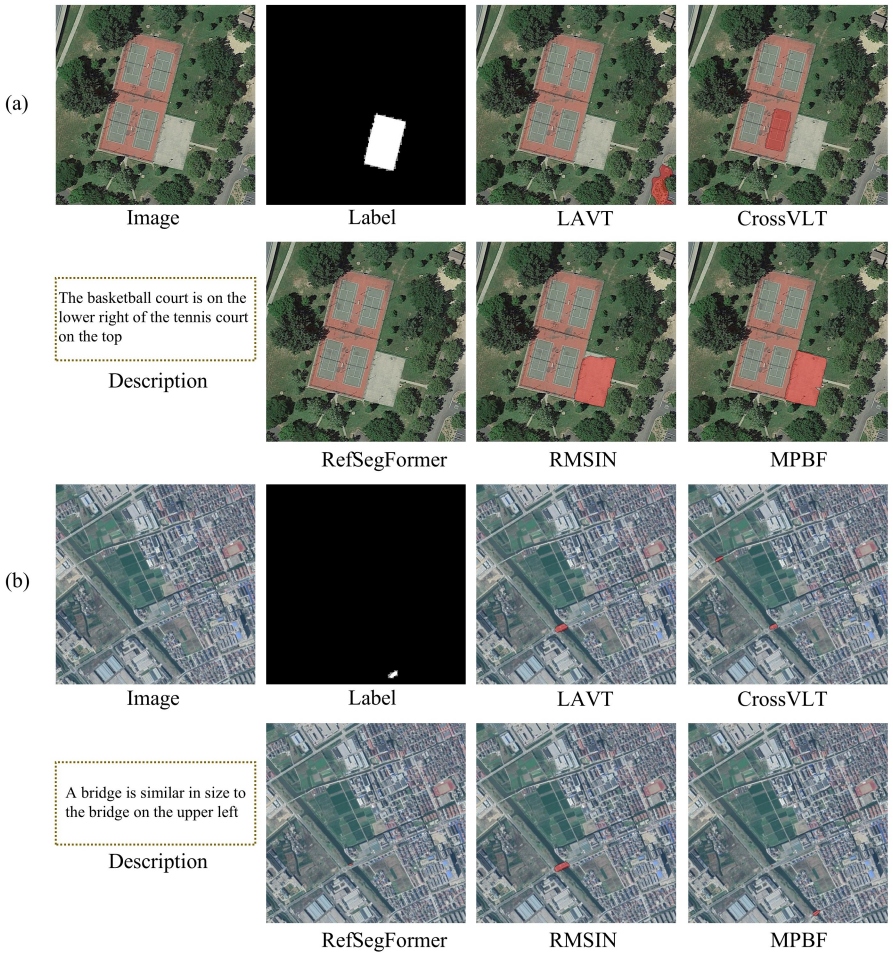


Figure 7. Qualitative comparisons on the samples (**a**,**b**) from the RRSIS-D dataset, which have complicated text descriptions.

For text descriptions involving multiple objects, the model should not only map the object attributes from the text to visual features, but also accurately distinguish the segmentation target from other mentioned objects. Furthermore, the relative spatial position of objects varies depending on the subject being modified. Therefore, the target cannot be directly located based on a single attribute in the description. Instead, a comprehensive analysis of multiple attributes is required to generate an accurate prediction. For example, in sample (a), both the target object and the reference object share the same category attribute. The spatial position (upper-left) modifies the reference object, meaning that the target's position relative to the reference is in the lower-right. Additionally, the text further constrains the target's size, stating that it should be similar to the reference object. Synthesizing these constraints, the bridge in the lower-right of the image satisfies both the width and positional requirements specified in the text. However, LAVT, RefSegFormer, and RMSIN misidentified the target, failing to accurately interpret the size relationships between objects. CrossVLT assigned a higher weight to the spatial attribute, resulting in the segmentation of both bridges to the lower-right of the reference object. In contrast, our proposed method precisely localized the target object by effectively associating linguistic features with visual features, demonstrating superior semantic comprehension.

Similarly, sample (b) contains two types of terrain features: a tennis court and a basketball court, along with two spatial position descriptors: "lower-right" and "top". LAVT accurately interpreted the "lower-right" spatial descriptor but mistakenly identified the green area in the lower-right as a basketball court. CrossVLT confused the target and

reference object, incorrectly extracting the tennis court in the lower-right instead of the basketball court. RefSegFormer misinterpreted the textual information as referring to a non-existent object and, therefore, did not generate a segmentation mask.

The qualitative analysis results demonstrate that the proposed method, which employs multimodal prompts to fuse visual and linguistic features in a shared latent space, achieves superior cross-modal interaction compared to other methods. The model exhibits robust performance in comprehending spatial semantics within remote sensing images, even when dealing with complex text descriptions and significant background interference. Furthermore, the late fusion strategy implemented during the decoding phase establishes pixel-level semantic associations between linguistic and visual features, enhancing the extraction of fine-grained details and improving segmentation accuracy.

5. Discussion

5.1. Failure Case Analysis

Through an in-depth analysis of the prediction results, we also identified several typical failure cases. These are mainly reflected in inconsistent semantic granularity, difficulties in spatial relationship parsing, and insufficient understanding of relative scale. As shown in Figure 8, in sample (a), MPBF incorrectly segmented a large area corresponding to the entire airport in response to the referring expression "airport". The resulting mask extended beyond the annotated ground truth, leading to a false positive. This issue can be attributed to two main causes: first, the annotations for "airport" in the dataset are inconsistent across samples, i.e., some samples annotate the entire airport area, while others only annotate the runway (as in sample (a)), resulting in semantic mismatches between text and image during both training and inference. Second, the runway and the surrounding ground in the image have similar textures and low contrast, which results in insufficient discriminative information for MPBF at the target boundaries. These observations suggest the need for greater consistency in annotation granularity and the incorporation of feature enhancement strategies to improve the model's ability to recognize detailed structures.

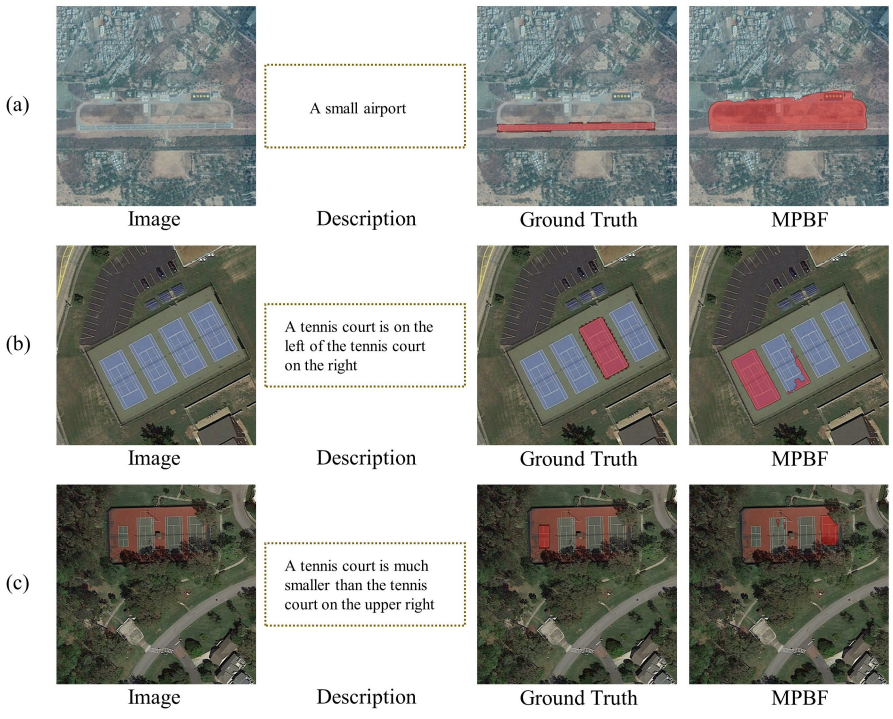


Figure 8. Failure case visualization. (**a**–**c**) are three typical failure samples.

Samples (b) and (c) reveal the model's deficiencies in handling relationships among multiple similar objects. In sample (b), MPBF failed to correctly interpret spatial relational expressions such as "left of" and "on the right", resulting in ambiguity in relative positioning. To address this, we can consider introducing graph-based modeling to explicitly encode spatial relationships and support relational reasoning. In sample (c), the textual description involves a compound reference that includes both relative size ("much smaller") and spatial orientation ("upper-right"). While MPBF demonstrates a reasonable understanding of directional cues, it struggles to interpret comparative relationships related to scale. Therefore, it is advisable to incorporate auxiliary modules or tasks into the language processing pipeline that explicitly handle size comparisons, thereby enhancing the model's ability to resolve relational semantics involving relative size.

5.2. Complexity Analysis

To evaluate the complexity of MPBF, we analyzed the floating point operations (FLOPs), total number of parameters (Params), and frames per second (FPS) across all methods. The comparison results of model complexity, FPS, and the corresponding oIoU values on the test set of the RRSIS-D dataset are listed in Table 3, with their distribution depicted in Figure 9, where the radius of the circle represents the FLOPs value. Although MPBF performs multimodal feature interaction and fusion in both the encoder and decoder, the number of additional parameters introduced by multimodal prompts is significantly lower than that of cross-attention operations in CrossVLT. This allows MPBF to efficiently achieve optimal performance while maintaining a low parameter count, demonstrating its effectiveness in balancing accuracy and computational efficiency. While MPBF has a lower FPS than other methods, this trade-off is justified by the superior accuracy in segmentation, which is often the more critical metric in many applications, especially in remote sensing image analysis tasks where precision is key.

Table 3. Comparison results of model complexity, FPS, and oIoU.

Method	Params(M)	FLOPs(G)	oIoU(%)	FPS(fps)
LAVT	203.70	193.97	76.04	14.10
CrossVLT	212.68	200.39	75.92	14.04
RefSegFormer	195.00	103.64	76.39	11.95
RMSIN	196.65	149.51	77.78	13.71
MPBF	159.12	146.94	78.20	11.82

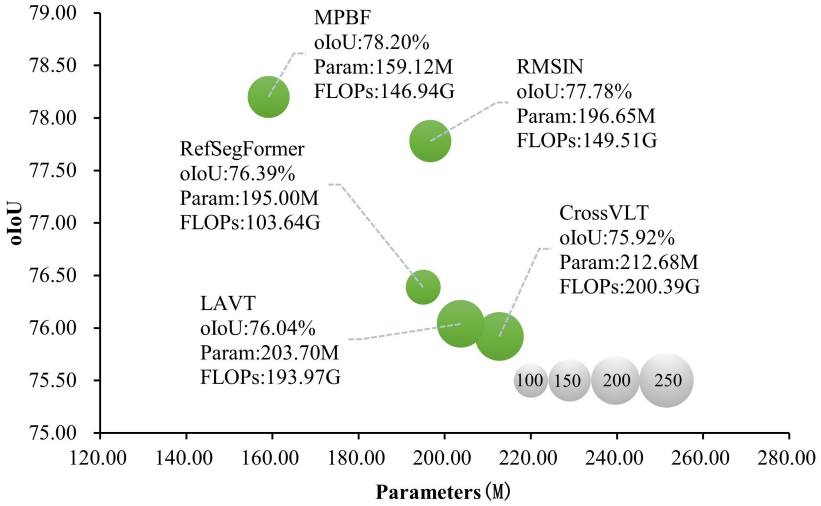


Figure 9. Comparison of model complexity and oIoU values on the test set, where the radius of the circle represents the FLOPs.

5.3. Ablation Study

The ablation studies are conducted on the RRSIS-D dataset to validate the effectiveness of the proposed module designs by comparing and analyzing the impact of each key module in our method.

5.3.1. Ablation Evaluation of Early Fusion

In the early fusion stage, our method adopts a deep prompt setting, where multimodal prompts are inserted at every stage of both the image encoder and text encoder to facilitate deep multimodal feature fusion. To evaluate the effectiveness of the current multimodal prompt setting, we compared two prompt insertion strategies: deep prompt and shallow prompt. In the shallow prompt setting, multimodal prompts were inserted only at the first stage of both the image encoder and text encoder, while all other configurations remained the same as in the deep prompt. As shown in Table 4, the quantitative metrics on the RRSIS-D test set decreased noticeably when using shallow prompts. Specifically, the mIoU value dropped by 1.55%, while the oIoU value decreased by 0.83%. These results indicate that deep prompts provide more flexibility for feature interaction at deeper layers, thereby enhancing the collaborative learning of visual and linguistic features during the encoding stage, which is also consistent with the findings reported in [37,38].

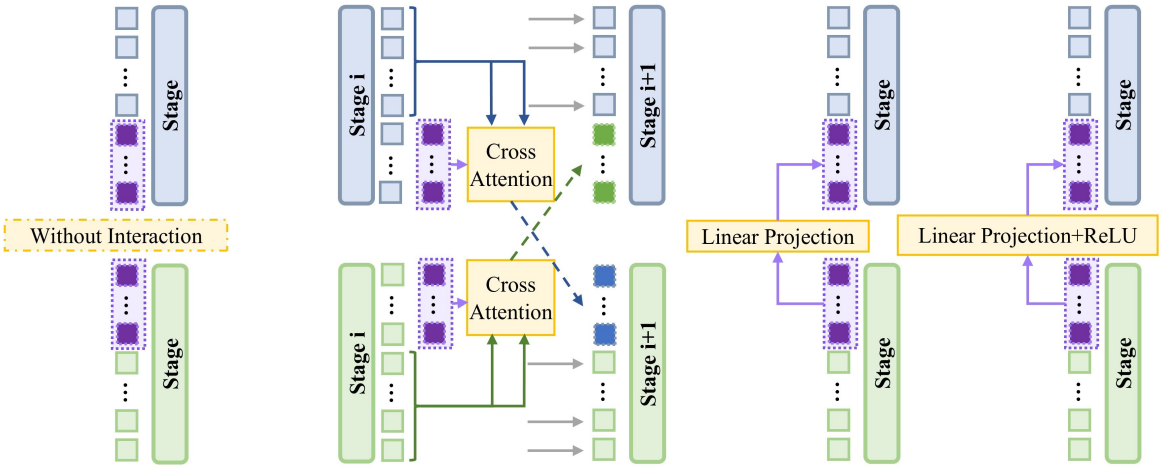
Table 4. Quantitative results of early fusion stage ablation studies.

Method	Fusion Setting	Pr@0.5 (%)	Pr@0.6 (%)	Pr@0.7 (%)	Pr@0.8 (%)	Pr@0.9 (%)	oIoU (%)	mIoU (%)
Prompt Insertion Position	Shallow Prompting	73.28	66.70	56.45	43.00	25.28	77.37	63.77
Prompt Interaction Mode	Independent Vision–Language Prompts	75.50	68.54	57.60	44.01	25.83	78.09	65.01
	Prompt-RIS-Style Interaction	74.23	67.62	56.94	44.10	26.72	77.63	64.53
	Nonlinear Coupling	75.06	68.60	58.26	44.15	26.80	78.22	65.19
Number of Prompts	0	74.78	68.11	57.28	44.07	26.66	77.28	64.53
	3	74.23	68.00	56.85	44.15	26.66	77.66	64.73
	4	75.09	68.77	58.49	44.87	26.34	77.85	65.29
	6	75.75	68.80	57.68	44.84	26.77	78.14	65.33
	10	75.84	69.29	58.80	44.79	27.09	78.19	65.30
MPBF	Deep Prompting/Linear Coupling/5	75.90	69.18	57.80	44.04	26.57	78.20	65.32

Then, concerning the interaction modes of prompts in different modalities, we compared the following four approaches: (a) independent vision–language prompts: no interaction between prompts in the image encoder and text encoder. (b) Prompt-RIS-style [45] interaction: in the image encoder, cross-attentions were used to embed visual context into randomly generated visual prompts, which were then inserted into the text encoder and fused with linguistic features, and vice versa. (c) Linear coupling: the linear projection-based prompt coupling adopted in our method. (d) Nonlinear coupling: this is an extension of the above approach, adding ReLU activation after linear projection. A simplified illustration of these four approaches is provided in Figure 10.

The experimental results in Table 4 show that compared to independent prompts without interaction, linear coupling increased the oIoU value and the mIoU value by 0.11% and 0.31%, respectively. Indicating that linear coupling of multimodal prompts effectively reduces the modality gap between heterogeneous features. The introduction of Prompt-RIS-style interaction led to the degradation of model performance. This may be because for the encoder in our architecture, directly embedding visual (or linguistic) modal features into prompts and then combining them with linguistic (or visual) features still leaves a significant modal disparity between the prompts and the features, which ultimately interferes with the feature learning process rather than facilitating better multimodal fusion.

Nonlinear coupling and linear coupling exhibit similar performance, but compared with nonlinear coupling, linear coupling achieved a 0.13% increase in mIoU.



(a) Independent Vision-Language Prompts (b)Prompt-RIS-Style Interaction (c) Linear Coupling (d) Nonlinear Coupling

Figure 10. Schematics of the different ways multimodal prompts interact.

Finally, to analyze the impact of the number of prompts, we compared the model's performance under different prompt quantity settings. When $N = 0$, the model only applied late fusion, meaning that no prompts were inserted during the encoding stage. In this case, the prompts fed into the VLA module were randomly generated. It should be noted that since the multimodal prompts input to the VLA module are extracted from the last-stage output of the image encoder, the number of multimodal prompts in VLA remains consistent with the prompts used in the early fusion stage, and both change synchronously.

As shown in Table 4, reducing the number of prompts limited the flexibility of deep feature learning, leading to a decrease in segmentation accuracy. However, increasing the number of prompts did not yield significant performance improvements, while the additional prompts introduced higher parameter complexity. Therefore, to balance model performance and computational complexity, we set the number of prompts to 5 in our method.

5.3.2. Ablation Evaluation of Late Fusion

In the late fusion stage, our method first employs the SR module to extract diverse scale representations from image features. Then, in the VLA module, fine-grained alignment between linguistic and visual features is iteratively achieved through VLSI processes. The multimodal prompts generated in the early fusion stage serve as prior information, which is concatenated with linguistic features and jointly input into the VLA module, ensuring semantic consistency across fusion stages.

To verify the effectiveness of the SR module in the late fusion stage, we conducted an ablation study by evaluating the model's performance after removing the SR module. As shown in Table 5, removing the SR module resulted in a 0.82% drop in mIoU and a 0.41% decrease in oIoU, indicating that integrating fine-grained scale representations significantly enhances the model's ability to detect multi-scale targets in remote sensing images.

Next, we analyzed the impact of three different prompt settings in the VLA module: randomly generated prompts, multimodal prompts derived from the text encoder output (P_L^3), and multimodal prompts derived from the image encoder output (P_V^3), as proposed in this study. The results in Table 5 show that multimodal prompts derived from the encoding stage outperformed randomly generated prompts. These multimodal prompts preserved the contextual dependencies of the modality interaction process, ensuring semantic consistency between the encoding and decoding stages. Furthermore, P_V^3 outperformed

P_L^3 by incorporating additional image semantics beyond textual descriptions, making it particularly valuable for fine-grained multimodal alignment in the late fusion stage.

Finally, regarding the design of the vision–language alignment process, we compared the parallel structure proposed in MARIS [46] and the sequential structure adopted in this paper. Experimental results indicate that, due to the inheritance between alignment steps, the sequential structure achieved better visual–language alignment than the parallel one, and established a more accurate mapping between text semantics and image content.

Table 5. Quantitative results of late fusion stage ablation studies.

Method	Fusion Setting	Pr@0.5 (%)	Pr@0.6 (%)	Pr@0.7 (%)	Pr@0.8 (%)	Pr@0.9 (%)	oIoU (%)	mIoU (%)
w/o SR Module	-	74.00	67.42	56.36	43.84	26.66	77.79	64.50
Prompt Source in VLA	P_L^3	75.41	68.46	58.35	44.30	26.77	78.09	65.18
	Random Prompting	74.83	68.20	57.02	44.01	25.94	77.69	64.86
Vision–Language Alignment Mode	Parallel Mode	75.27	68.74	57.54	44.33	26.69	77.88	64.95
MPBF	P_V^3/Serial Mode	75.90	69.18	57.80	44.04	26.57	78.20	65.32

5.4. Application Experiments

Considering that images in public datasets often present idealized conditions, which typically contain relatively simple scenes, clean backgrounds, and highly salient targets, these datasets may not fully reflect the challenges encountered in real-world applications. We evaluated our method in a practical application scenario: hazard detection along high-speed rails. Using satellite remote sensing images collected from different satellites with varying ground sampling distances, we validated the robustness and applicability of our approach in real-world conditions, demonstrating its ability to handle complex, noisy, and diverse environments.

5.4.1. Quantitative Comparison

The quantitative comparison results of all methods on the HSRMS dataset are listed in Table 6. Our proposed method achieves the best performance across all evaluation metrics. Specifically, compared to the second-best RefSegFormer, our method improves oIoU by 0.69% on the test set and 0.91% on the validation set. Similarly, the mIoU increased by 0.95% and 1.72%, respectively. Regarding high-accuracy samples, in the test set, the proportion of samples with IoU > 0.9 reaches 46.25%, representing a 0.54% improvement over RefSegFormer. In the validation set, this metric improves by 1.3% compared to RefSegFormer.

Table 6. Quantitative comparison results on the HSRMS dataset (best values in bold; second-best values underlined).

Method	Dataset Type	Pr@0.5 (%)	Pr@0.6 (%)	Pr@0.7 (%)	Pr@0.8 (%)	Pr@0.9 (%)	oIoU (%)	mIoU (%)
LAVT	Test	89.69	84.87	78.88	67.37	38.74	80.21	79.05
	Validation	88.05	82.81	75.51	63.43	35.58	78.99	77.16
CrossVLT	Test	90.15	85.95	79.78	69.88	43.79	82.30	80.29
	Validation	88.94	84.28	78.77	67.60	41.44	80.42	79.33
RefSegFormer	Test	90.22	<u>86.54</u>	<u>80.36</u>	<u>70.33</u>	<u>45.71</u>	<u>82.34</u>	<u>80.48</u>
	Validation	89.39	84.98	<u>79.14</u>	<u>68.86</u>	<u>42.78</u>	<u>81.27</u>	<u>79.39</u>
RMSIN	Test	<u>90.61</u>	86.49	79.78	68.17	40.57	81.58	80.14
	Validation	<u>89.85</u>	<u>85.14</u>	78.45	67.11	39.65	80.32	79.35
MPBF	Test	**90.83**	**86.79**	**80.87**	**70.80**	**46.25**	**83.03**	**81.43**
	Validation	**91.00**	**87.01**	**80.10**	**69.98**	**44.08**	**82.18**	**81.11**

5.4.2. Qualitative Comparison

Figure 11 visualizes the segmentation results of all methods for hazards along high-speed rails. These examples are selected from the test set of the HSRMS dataset.

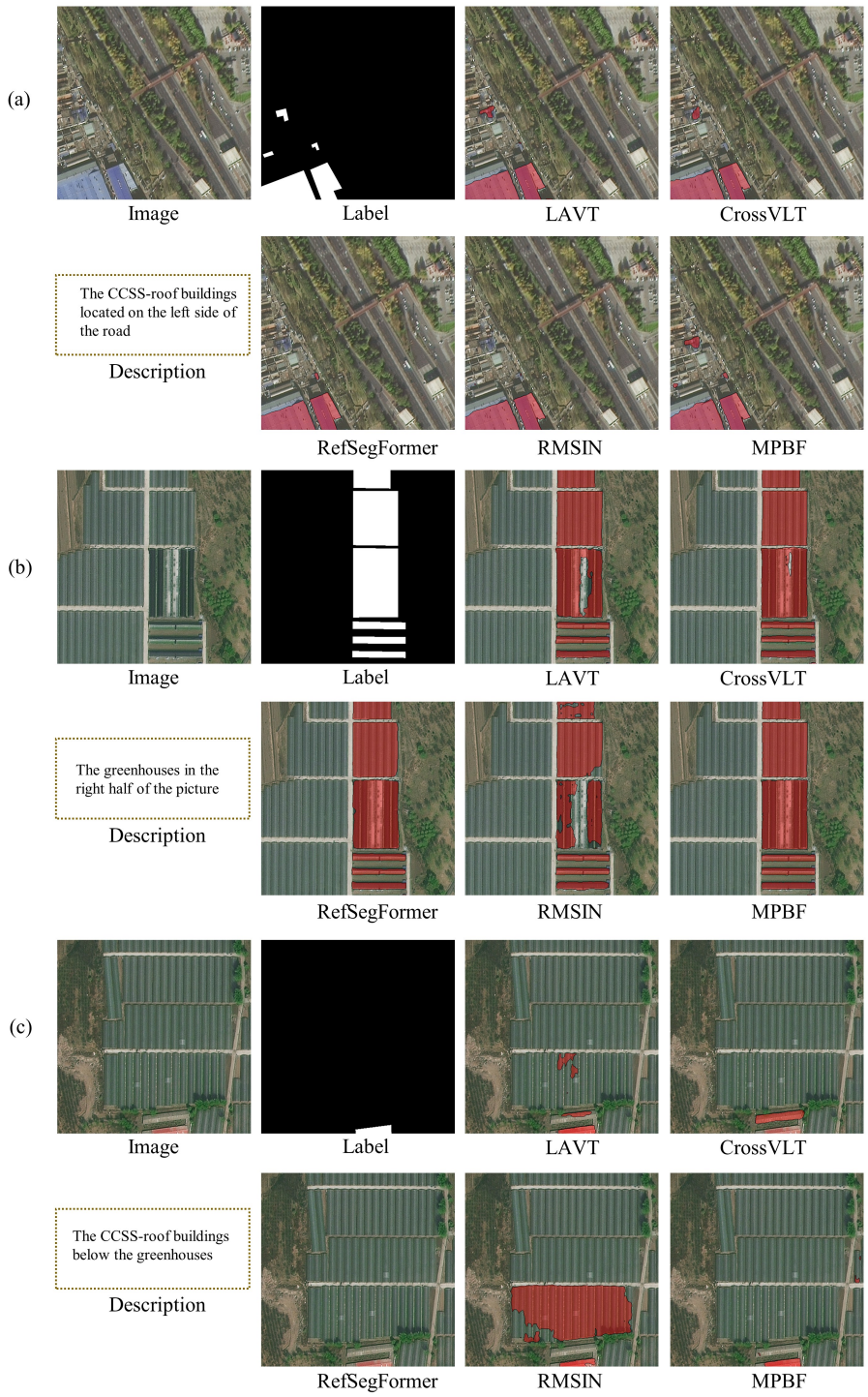


Figure 11. Qualitative comparisons on the HSRMS dataset, (**a**–**c**) are the samples.

As shown in Figure 11, significant scale variation is a typical characteristic of CCSS roof buildings along high-speed rail lines. In sample (a), both large-scale and small-scale CCSS roof buildings appear on the left side of the road. RMSIN completely failed to detect the small-scale CCSS roof buildings, while LAVT, CrossVLT, and RefSegFormer missed some of them. In contrast, our proposed method correctly located all CCSS roof buildings, demonstrating its robust segmentation performance across varying scales.

In remote sensing images, plastic greenhouses typically appear as densely arranged rectangular structures. Due to the different types of covering materials used, plastic greenhouses within the same area may exhibit different geometric and spectral characteristics. In sample (b), for the plastic greenhouses located on the right side, LAVT, CrossVLT, and RMSIN failed to detect some of the plastic greenhouses with different spectral characteristics. In contrast, RefSegFormer and our method achieved higher accuracy, successfully extracting all plastic greenhouses.

The text description in sample (c) includes two types of hazard targets: plastic greenhouses and CCSS roof buildings. As shown in Figure 11c, plastic greenhouses occupy a much larger proportion of the image compared to CCSS roof buildings, creating a significant imbalance in the features of the two types of targets. RMSIN correctly detected the CCSS roof building, but due to interference from other category words and location cues in the text, it mistakenly classified the large plastic greenhouses above as CCSS roof buildings. Moreover, similar spatial characteristics also led to confusion between the red CCSS roof building and the plastic greenhouses above it in terms of texture and geometric structure. LAVT and CrossVLT misclassified the plastic greenhouses above as CCSS roof buildings. RefSegFormer failed to recognize the target at the bottom, returning no valid prediction.

Experimental results in practical application scenarios demonstrate that our proposed method maintains high segmentation accuracy even in remote sensing images of complex scenes. Benefiting from the deep fusion of visual and linguistic features, our method enables accurate comprehension of image and text semantics, establishing a precise mapping between the two modalities. This enables the accurate extraction of fine-grained features. Furthermore, by fully exploiting diverse scale representations in remote sensing images, our method significantly improves its ability to extract multi-scale targets, effectively mitigating the issue of small-scale object omission.

6. Conclusions

In this study, we proposed MPBF for RRSIS. MPBF combines early fusion and late fusion strategies. In the early fusion stage, the model achieves deep multimodal feature fusion with a low parameter overhead using multimodal prompts. In the late fusion stage, MPBF employs an SR module to learn multi-scale feature representations and establishes pixel-level associations between visual and linguistic features through the VLSI processes. Comparative and ablation studies on public datasets demonstrate that our method accurately interprets complex textual descriptions and establishes fine-grained associations between text semantics and image content, achieving SOTA results. Application experiments on the custom-built HSRMS dataset further validate the effectiveness of this method in practical scenarios. When integrated with appropriate hardware platforms, the research findings in this paper will significantly enhance the intelligence and user-friendliness of remote sensing image analysis systems.

Author Contributions: Conceptualization, Y.L. and W.J.; methodology, Y.L.; validation, Y.L. and Q.S. ; formal analysis, Y.L. and Q.S.; investigation, Y.L., W.J. and S.Q.; resources, W.J.; writing—original draft preparation, Y.L.; writing—review and editing, W.J. and S.Q.; visualization, Y.L.; supervision, W.J.; project administration, W.J. and S.Q.; funding acquisition, W.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, Grant No.62171024, and the Ministry of Science and Technology of the People's Republic of China, Grant No.2020YFF0304104.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors are grateful to the researchers at the Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Xiamen University, for providing the RRSIS-D dataset.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yuan, Z.; Mou, L.; Hua, Y.; Zhu, X.X. Rrsis: Referring remote sensing image segmentation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5613312. [CrossRef]
2. Liu, S.; Ma, Y.; Zhang, X.; Wang, H.; Ji, J.; Sun, X.; Ji, R. Rotated multi-scale interaction network for referring remote sensing image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024; pp. 26658–26668.
3. Hu, R.; Rohrbach, M.; Darrell, T. Segmentation from Natural Language Expressions. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N.; Welling, M., Eds.; Springer: Cham, Switzerland, 2016; pp. 108–124.
4. Li, H.; Zhang, X.; Qu, H. DDFAV: Remote Sensing Large Vision Language Models Dataset and Evaluation Benchmark. *Remote Sens.* **2025**, *17*, 719. [CrossRef]
5. Liu, G.; He, J.; Li, P.; Zhong, S.; Li, H.; He, G. Unified Transformer with Cross-Modal Mixture Experts for Remote-Sensing Visual Question Answering. *Remote Sens.* **2023**, *15*, 4682. [CrossRef]
6. Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; Berg, T.L. Mattnet: Modular attention network for referring expression comprehension. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1307–1315.
7. Yang, Z.; Chen, T.; Wang, L.; Luo, J. Improving one-stage visual grounding by recursive sub-query construction. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XIV 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 387–404.
8. 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*, 5604513. [CrossRef]
9. Li, C.; Zhang, W.; Bi, H.; Li, J.; Li, S.; Yu, H.; Sun, X.; Wang, H. Injecting Linguistic Into Visual Backbone: Query-Aware Multimodal Fusion Network for Remote Sensing Visual Grounding. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5637814. [CrossRef]
10. Feng, G.; Hu, Z.; Zhang, L.; Lu, H. Encoder fusion network with co-attention embedding for referring image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15506–15515.
11. Ouyang, S.; Wang, H.; Xie, S.; Niu, Z.; Tong, R.; Chen, Y.W.; Lin, L. SLViT: Scale-Wise Language-Guided Vision Transformer for Referring Image Segmentation. In Proceedings of the IJCAI, Macao, China, 19–25 August 2023; pp. 1294–1302.
12. Wu, J.; Li, X.; Li, X.; Ding, H.; Tong, Y.; Tao, D. Toward Robust Referring Image Segmentation. *IEEE Trans. Image Process.* **2024**, *33*, 1782–1794. [CrossRef]
13. Liu, J.; Ding, H.; Cai, Z.; Zhang, Y.; Satzoda, R.K.; Mahadevan, V.; Manmatha, R. Polyformer: Referring image segmentation as sequential polygon generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18653–18663.
14. Huang, S.; Hui, T.; Liu, S.; Li, G.; Wei, Y.; Han, J.; Liu, L.; Li, B. Referring image segmentation via cross-modal progressive comprehension. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10488–10497.
15. Liu, C.; Jiang, X.; Ding, H. Instance-specific feature propagation for referring segmentation. *IEEE Trans. Multimed.* **2022**, *25*, 3657–3667. [CrossRef]
16. Li, K.; Wang, D.; Xu, H.; Zhong, H.; Wang, C. Language-Guided Progressive Attention for Visual Grounding in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5631413. [CrossRef]
17. Ye, P.; Xiao, G.; Liu, J. Multimodal Features Alignment for Vision–Language Object Tracking. *Remote Sens.* **2024**, *16*, 1168. [CrossRef]
18. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural inform. Process. Syst.* **2017**, *30*, 6000–6010.
19. Xu, P.; Zhu, X.; Clifton, D.A. Multimodal learning with transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12113–12132. [CrossRef]
20. Zhang, Z.; Zhu, Y.; Liu, J.; Liang, X.; Ke, W. Coupalign: Coupling word-pixel with sentence-mask alignments for referring image segmentation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 14729–14742.
21. 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. Multimed.* **2023**, *26*, 5823–5833. [CrossRef]
22. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]

23. Liu, C.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Yuille, A. Recurrent multimodal interaction for referring image segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1271–1280.
24. Li, R.; Li, K.; Kuo, Y.C.; Shu, M.; Qi, X.; Shen, X.; Jia, J. Referring image segmentation via recurrent refinement networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5745–5753.
25. Margffoy-Tuay, E.; Pérez, J.C.; Botero, E.; Arbeláez, P. Dynamic multimodal instance segmentation guided by natural language queries. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 630–645.
26. Chen, Y.W.; Tsai, Y.H.; Wang, T.; Lin, Y.Y.; Yang, M.H. Referring expression object segmentation with caption-aware consistency. *arXiv* **2019**, arXiv:1910.04748 .
27. Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; Torr, P.H. Lavt: Language-aware vision transformer for referring image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18155–18165.
28. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
29. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; Volume 1 (long and short papers); pp. 4171–4186.
30. Yang, Z.; Wang, J.; Tang, Y.; Chen, K.; Zhao, H.; Torr, P.H. Semantics-aware dynamic localization and refinement for referring image segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington D.C., USA, 7–14 February 2023; Volume 37, pp. 3222–3230.
31. Liu, P.; Yuan, W.; Fu, J.; Jiang, Z.; Hayashi, H.; Neubig, G. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* **2023**, *55*, 1–35. [CrossRef]
32. Li, L.; Guan, H.; Qiu, J.; Spratling, M. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 24408–24419.
33. Wu, Z.; Liu, Y.; Zhan, M.; Hu, P.; Zhu, X. Adaptive Multi-Modality Prompt Learning. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, Australia, 28 October—1 November 2024; pp. 8672–8680.
34. Wang, Q.; Yan, K.; Ding, S. Bilateral Adaptive Cross-Modal Fusion Prompt Learning for CLIP. In Proceedings of the 32nd ACM International Conference on Multimedia, Melbourne, Australia, 28 October—1 November 2024; pp. 9001–9009.
35. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Learning to prompt for vision-language models. *Int. J. Comput. Vis.* **2022**, *130*, 2337–2348. [CrossRef]
36. Zhou, K.; Yang, J.; Loy, C.C.; Liu, Z. Conditional prompt learning for vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16816–16825.
37. Jia, M.; Tang, L.; Chen, B.C.; Cardie, C.; Belongie, S.; Hariharan, B.; Lim, S.N. Visual prompt tuning. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 709–727.
38. Khattak, M.U.; Rasheed, H.; Maaz, M.; Khan, S.; Khan, F.S. Maple: Multi-modal prompt learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19113–19122.
39. Liu, S.A.; Zhang, Y.; Qiu, Z.; Xie, H.; Zhang, Y.; Yao, T. CARIS: Context-aware referring image segmentation. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 779–788.
40. Wu, J.; Zhang, Y.; Kampffmeyer, M.; Zhao, X. Prompt-guided bidirectional deep fusion network for referring image segmentation. *Neurocomputing* **2025**, *616*, 128899. [CrossRef]
41. Cheng, B.; Schwing, A.; Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17864–17875.
42. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1290–1299.
43. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
44. Li, Y.; Zuo, D.; Jin, W.; Qiu, S. Intelligent Detection of Hidden Hazards Along High-Speed Rail Based on Optical Remote Sensing Images. *Acta Optica Sinica* **2025**, *45*, 0728004.

45. Shang, C.; Song, Z.; Qiu, H.; Wang, L.; Meng, F.; Li, H. Prompt-driven referring image segmentation with instance contrasting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 4124–4134.
46. Zhang, M.; Liu, Y.; Yin, X.; Yue, H.; Yang, J. MARIS: Referring Image Segmentation via Mutual-Aware Attention Features. *arXiv* **2023**, arXiv:2311.15727.