实体引导和跨模态交互的遥感图像指向分割

贾玉钰，迟凯晨，周情，李强，王琦

西北工业大学 光电与智能研究院, 西安 710072

摘 要：遥感图像指向分割旨在依据文本描述，精准定位并解析特定区域，实现像素级的语义解译。这一任务有效搭建了用户需求与遥感图像智能分析之间的桥梁。然而，遥感图像固有的背景复杂多样、目标与背景对比度低等特性，易导致分割结果的地物混淆。此外，传统基于跨模态注意力机制的指向分割方法在弥合模态差异方面存在不足，难以实现文本描述与地理特征之间的精细对齐。鉴于此，本文提出了一种基于实体引导的跨模态交互方法—Enti-CroM。首先，引入SAM启发的实体引导自推理模块，以提取具备空间结构约束的实体提示，并将其与视觉特征和文本特征融合，构建实体—视觉—文本三模态信息立方体。随后，在跨模态交互阶段，设计了一种层次化模态交互机制：（1）通过无参数的模态互激活，实现跨模态语义信息的高效传播，从而缩小不同模态间的语义鸿沟；（2）在此基础上，引入实体引导下的文本—视觉跨注意力运算，进一步提升模型对不规则地理边界的表征能力。在两个广泛使用的基准数据集RefSegRS与RRSIS-D上的大量实验结果表明，Enti-CroM在mIoU指标上分别实现了1.84%与4.27%的性能提升，性能优于当前多种先进方法，验证了该方法的有效性与优越性。

关键词：遥感图像，指向分割，跨模态交互，SAM，实体感知，注意力机制，空间结构约束

中图分类号：TP751/P2

引用格式：贾玉钰,迟凯晨,周情,李强,王琦.2026.实体引导和跨模态交互的遥感图像指向分割.遥感学报,30(2): 311-324

Jia Y Y, Chi K C, Zhou Q, Li Q and Wang Q. 2026. Entity-guided cross-modal interaction for referring segmentation of remote sensing images. National Remote Sensing Bulletin, 30(2): 311-324[DOI: 10.11834/jrs.20255370]

1 引 言

遥感图像分割作为地理空间分析中的一项核心任务，提供了环境变化监测（佟国峰 等，2015；柳思聪 等，2023）、城市规划（张琳琳 等，2025）和灾害管理（安立强 等，2024）等关键应用所需的重要像素级信息。随着大规模遥感数据集的快速增长及深度学习的革命性进展，该领域衍生出多种新的任务范式，如半监督学习（Huang等，2024；Xuan等，2024）、小样本学习（Jia等，2023；Wang等，2025）和零样本学习（秦楠楠 等，2025），以及面向基础模型的参数高效领域泛化（Zhang等，2025b）和跨域自适应（Zhang等，2025a）等，逐步朝着更加智能和实用的目标迈进。然而，这些任务通常基于预定义类别对所有潜在感兴趣区域进行全面分割，导致它们往往无法有效捕捉图像中不同区域之间的微妙语义关系（Ji等，2024）。此外，在通过自然语言进行人机交互查询时，这些方法也不具备根据用户意图识别特定目标的能力。

在此背景下，一项更具挑战性的任务—指向分割RIS（Referring Image Segmentation）应运而生。RIS旨在根据用户提供的文本描述，精确地定位并分割视觉场景中的特定目标区域，从而缩小人类意图与图像理解之间的语义鸿沟。近年来，借助注意力机制强大的全局建模能力，RIS的研究进展主要集中于单向和双向注意力结构的应用，以实现文本逻辑结构与视觉对象空间关系的有效对齐。Yang等（2022）通过在Swin Transformer的中间层进行语言和视觉特征的早期融合来增强跨

收稿日期：2025-09-05；**预印本：**2025-12-26

基金项目：国家自然科学基金(编号:62471394, U21B2041)

第一作者简介：贾玉钰,研究方向为遥感图像多模态分析、小样本学习的理论和应用。E-mail:jyy2019@mail.nwpu.edu.cn

通信作者简介：王琦,研究方向为计算机视觉、模式识别、遥感图像处理。E-mail:crabwq@nwpu.edu.cn

模态对齐。Ding等（2023）提出查询生成模块，动态生成多组针对输入的特定查询，以表示对语言表达的不同理解。Yang等（2024）利用Mamba高效的线性计算，引入通道空间扭曲机制，实现视觉和文本特征的深度融合。尽管在自然图像领域取得了显著的成果，但将这些技术直接迁移到遥感图像指向分割RRSIS（Remote Sensing Referring Image Segmentation）中仍面临一定的挑战。具体而言，遥感图像呈现出独特的挑战性，例如目标尺度变化极大、地物类别繁杂且背景高度混淆、以及文本描述更依赖复杂的空间上下文关系。因此，为自然图像设计的模态对齐机制难以直接捕捉遥感场景中语言与地理空间特征间的精细对应关系。

近年来，面向RRSIS任务的研究逐渐兴起，大量高质量数据集相继构建，相关算法方法不断涌现。研究者针对遥感场景中目标尺度变化大、背景复杂、模态差异显著等特性，提出了多种具有针对性的技术路线。Yuan等（2024）率先开展了RRSIS任务，并建立RefSegRS数据集。为解决复杂空间尺度和方向的挑战，Liu等（2024）提出大规模的RRSIS-D数据集，并设计尺度内和跨尺度交互模块以及自适应旋转卷积，有效提升分割精度。Dong等（2024）通过上下文感知的提示调制来增强语言特征，并采用语言引导的特征聚合来融合多尺度视觉特征。然而，如图1所示，对上述工作的深入分析表明，它们在应对两个源于遥感数据固有特性和现有方法论局限的核心挑战时仍显不足：从地物特性层面来看，遥感图像与自然图像有着本质的区别，后者通常呈现出显著的前景物体和相对干净的背景。遥感场景由于地物类型之间的光谱相似性以及复杂的背景杂乱，常常呈现出目标与背景之间的低对比度，导致了严重的地物混淆，即分割区域可能包括无关的背景或邻近特征。尽管现有工作尝试融合多尺度特征，但它们仍缺乏显式的空间结构先验来约束分割边界，难以从根本上解决这一问题。从方法论的角度来看，大多数RIS及RRSIS框架依赖于扁平化的交叉注意力机制，将跨模态交互实现为整体特征级的匹配。然而，这些方法在弥合非结构化语言与复杂地理空间特征之间显著的模态鸿沟时，能力依然不足，从而导致粗略对齐，进而降低了目标定位和像素级解释的准确性。

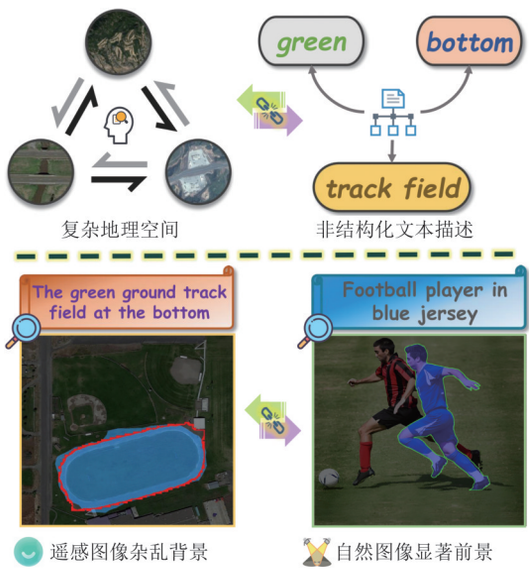


图1 从数据特性和方法论归纳RRSIS任务面临的两个挑战：地物混淆和模态鸿沟

Fig. 1 From the perspectives of data characteristics and methodological paradigms, the RRSIS task encounters two principal challenges: Semantic dispersion and modal gap

为了解决RRSIS中的地物混淆和模态差距这两个关键挑战，本文提出了两个核心思路。首先，针对地物混淆问题，为多模态特征引入显式的空间结构先验。具体而言，本文充分利用SAM（Segment Anything Model）（Kirillov等，2023）的泛化能力和细粒度实体级结构提取能力，引入了基于SAM启发的实体感知引导，在空间层面上约束文本与视觉信息的交互，从而引导模型更加聚焦于语义一致的实体区域。其次，针对传统跨注意力机制在面对显著模态差距时的局限性，本文提出了一种层次化的跨模态交互策略，将模态对齐与空间位置建模过程进行解耦，有效提升了模态间的对齐精度与空间关系建模的灵活性。

融合上述思路，本文提出一种基于实体引导的跨模态交互方法（Enti-CroM）。具体而言，上述两个核心思路被实例化为SAM启发的实体引导自推理模块（SEG）和层次化模态交互机制（HMI）。如图2所示，SEG和HMI都是即插即用的模块，可灵活嵌入至骨干特征提取器的多个阶段，以适配遥感目标的多尺度特性。在实现上，SEG将多个SAM生成的掩码转换为具有丰富空间信息的实体感知提示。鉴于SAM在分割时易将完整目标过度切分为碎片化掩码，SEG通过自推理过程推断实体间的内在关系，并将其细化为连贯、稳

健的实体引导线索。这些线索随后与视觉嵌入和文本嵌入相融合，通过通道级串联构建综合性的三模态特征立方体。接着，HMI利用层次化的交互机制有效缓解模态差距问题：（1）首先设计一种无参数的互激活范式，使不同模态的特征表示能够实现双向、动态的协同与增强，从而弥合跨模态间的固有差异；（2）随后，在实体感知线索的引导下，于空间维度执行跨模态的交叉注意力运算，以精准建模细粒度的跨模态位置依赖关系。

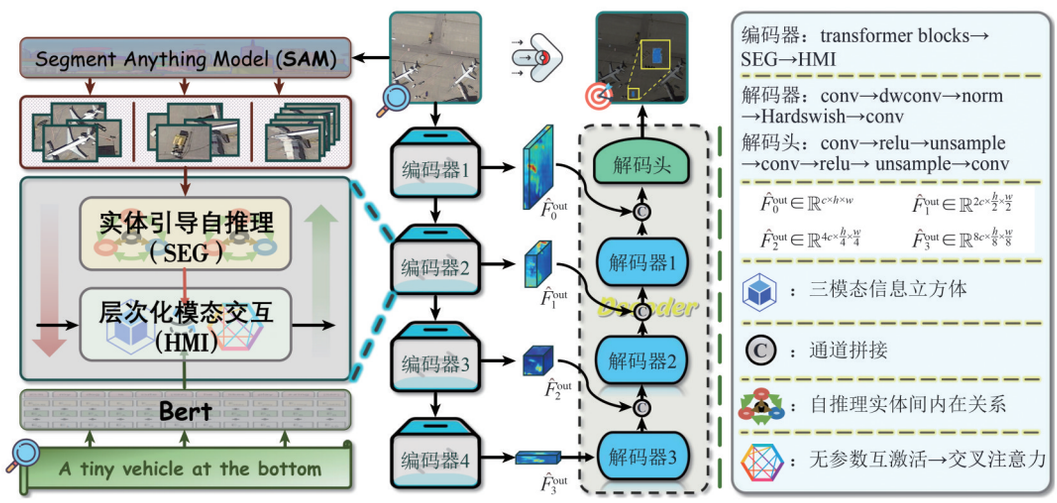


图2 实体引导的跨模态交互框架

Fig. 2 The framework of the entity-guided cross-modal interaction

2 实体引导的跨模态交互方法

在遥感图像指向分割任务中，核心挑战在于实现自然语言文本表达$E=\{e_i|i\in 0,\cdots,N\}$与查询图像$I\in\mathbb{R}^{H\times W\times 3}$之间的细粒度对齐。其中，$N$表示文本描述的长度，$H$与$W$分别表示图像的高度与宽度。查询图像$I$经由包含4个阶段（$t=1,2,3,4$）的分层编码器$\mathcal{E}_t$逐步映射为多尺度视觉特征表示$F_t^{\text{in}}$。为应对上述挑战，本文提出了基于实体引导的跨模态交互方法（Enti-CroM）。该框架的核心在于在编码器的每一个阶段嵌入两个即插即用模块：实体引导自推理模块（SEG），以及层次化模态交互机制（HMI）。具体而言，SEG通过自推理机制生成语义连贯且鲁棒的实体引导，将其与视觉和文本特征嵌入相融合，构建实体—视觉—文本三模态特征立方体，从而在视觉—语言交互过程中引入语义一致性约束。在此基础上，HMI采用分层的跨模态交互策略：首先引入一种无参数的模态互激活范式，实现模态间双向且动态的语义信息协同与增强，有效缩小模态差异；随后，在实体感知线索的引导下，于空间维度执行跨模态交叉注意力运算，以精确建模细粒度的跨模态空间依赖关系，从而提升对复杂地理边界的刻画能力。最终，由多阶段编码器输出的多尺度交互特征被送入解码器模块。该模块在结构设计上参考了ReMamber（Yang等，2024）的核心思想，以充分利用多尺度信息，实现对目标区域的精细化重建，并生成最终的高精度分割结果图。

2.1 实体引导自推理模块

在遥感场景中，由于不同地物类型在光谱反射特性上的高度相似，以及背景环境的复杂干扰，目标与背景之间的对比度常常偏低。这一特性在RRSIS任务中容易引发“地物混淆”现象，即分割结果受到无关背景信息或邻近地物实体的干扰，从而导致语义边界模糊、目标区域受到污染。为缓解这一问题，本文在多模态特征融合之前显式引入空间实体先验。如图3所示，利用SAM在零样本情况下对实体的强感知能力，生成实体感知引导信号，并将其作为先验信息融入跨模态交互过程，以约束模型重点关注语义一致的区域。该设计能够有效抑制地物混淆现象，减少背景干扰，并在此基础上提升分割精度。

在编码器的第t阶段，从编码器$\mathcal{E}_t$提取视觉特征F_t^{in}后，首先将查询图像输入至SAM模型，以生成初始掩码集合$\{m_k\}_{k=1}^{K}$，其中$m_k\in\{0,1\}^{H\times W}$。为

简洁起见，下文描述中省略阶段索引下标t。为避免掩码之间的重叠导致语义歧义，并同时保持每个掩码内部的语义一致性，本文对生成的掩码集合按面积由大到小进行排序。随后，将各掩码之间的交叠区域分配给面积最小的掩码，从而确保整个掩码集合的唯一性与无歧义性。在获得无重叠掩码集合后，我们对其执行掩码均值池化（MAP）操作，将掩码集合映射为实体先验：

$$p_k = \mathcal{MAP}(F^{\text{in}}, m_k) \in \mathbb{R}^{1\times d} \tag{1}$$

式中，d为通道维数，$k \in 1, \cdots, K$。

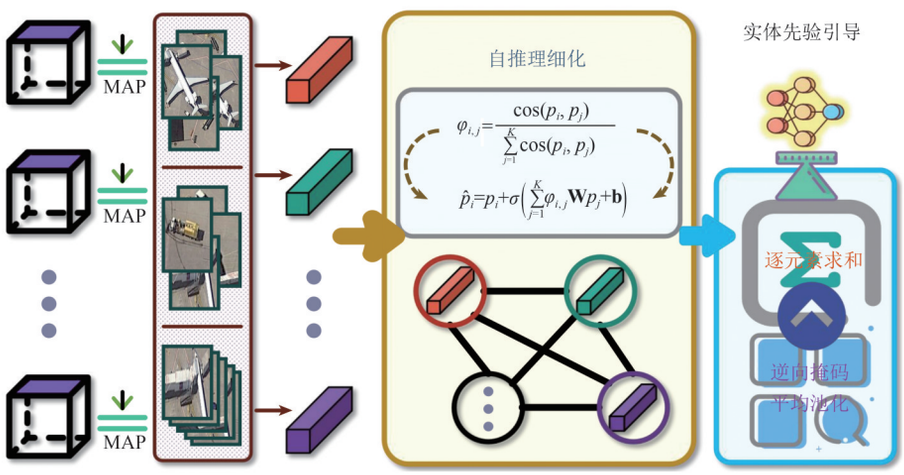


图3 实体引导自推理模块

Fig. 3 Configuration diagram of the SEG module

尽管SAM在生成细粒度掩码方面表现出色，但其输出中常出现碎片化分割现象，即将同一完整的地理空间实体划分为多个较小的掩码候选。若直接将这些由原始掩码生成的实体先验作为引导信息输入后续多模态特征融合模块，可能会引入噪声与语义不一致性，从而影响分割精度。为缓解上述问题，本文提出一种自推理细化机制，利用初始掩码之间固有的空间与语义关联关系，对分割候选进行聚合，生成更具连贯性与鲁棒性的实体感知引导。与图神经网络不同，本文的方法不依赖显式构建图结构，有效避免图卷积操作带来的计算复杂度问题。此外，与传统聚类方法如K-means相比，所提的细化机制不需要预设簇数，也无需通过划分簇来处理噪声和冗余信息。因此，所提方法能够更灵活地处理遥感图像中的复杂地物形态，并在背景干扰较强的情况下保持较高的鲁棒性。具体而言，对于给定的实体先验p_i，其与其他实体先验p_j的亲和系数通过余弦相似度计算，定义为

$$\varphi_{ij} = \frac{\cos(p_i, p_j)}{\sum_{j=1}^{K}\cos(p_i, p_j)} \tag{2}$$

式中，$\cos(\cdot,\cdot)$表示余弦相似度函数，$\varphi_{i,j}$值越大，说明实体先验p_i与p_j之间的关系亲和度越强。基于亲和度系数，执行加权聚合，以促进实体先验间的信息交互：

$$\hat{p}_i = p_i + \sigma\left(\sum_{j=1}^{K}\varphi_{i,j}\boldsymbol{W}p_j + \boldsymbol{b}\right) \tag{3}$$

式中，$\boldsymbol{W} \in \mathbb{R}^{d\times d}$与$\boldsymbol{b} \in \mathbb{R}^{d}$为可学习参数，$\sigma(\cdot)$为非线性激活函数，例如ReLU。最后，采用MAP操作的逆过程，恢复细化后的实体先验的空间信息：

$$g_k = \frac{\sum_{h,w}^{H',W'}\hat{p}_k \otimes m_k(h,w)}{\sum_{h,w}^{H',W'} m_k(h,w)} \tag{4}$$

式中，$k \in 0, \cdots, K$，$\otimes$表示逐元素相乘，(h, w)为空间坐标，(H', W')为当前阶段视觉特征的空间分辨率。最终，可将实体感知引导表示为$G = \sum_{k=1}^{K} g_k \in \mathbb{R}^{H'\times W'\times d}$。

2.2 三模态特征立方体构建

在Enti-CroM框架的跨模态交互过程中，文本描述提供全局语义线索，视觉特征揭示图像内容细节，而由SEG模块生成的实体感知引导则提供关键的空间结构先验。为充分发挥3种模态的互补优势，本文将它们整合至统一表示空间，构建面向多模态交互的三模态特征立方体。前文中，已获得视觉特征F^{in}与实体感知引导G，接下来将重

点描述如何将文本描述E的特征引入其中。

首先，将文本描述E输入文本编码器，得到原始文本特征$T^{\text{ori}} \in \mathbb{R}^{N \times d_t}$，其中$N$为文本长度，$d_t$为通道维数。考虑到在指向分割任务中，文本描述中涉及的颜色、形状、位置等局部属性对于精确定位目标至关重要，本文提出构建局部文本嵌入，以捕捉其与视觉特征间的细粒度对应关系。具体而言，通过矩阵乘法计算局部相关性图：

$$T^{\text{cor}} = F^{\text{in}} \boldsymbol{W}_0^{\text{cor}} \cdot (T^{\text{ori}} \boldsymbol{W}_1^{\text{cor}})^{\text{T}} \in \mathbb{R}^{H' \times W' \times N} \tag{5}$$

式中，$\boldsymbol{W}_0^{\text{cor}} \in \mathbb{R}^{d \times d}$与$\boldsymbol{W}_1^{\text{cor}} \in \mathbb{R}^{d_t \times d}$为可学习参数。为实现高维特征处理且与视觉特征的通道维度对齐，采用单卷积层将T^{cor}投影为$T^{\text{loc}} \in \mathbb{R}^{H' \times W' \times d}$。最终，通过通道拼接方式构建三模态特征立方体：

$$F^{\text{cube}} = [F^{\text{in}} + G; T^{\text{loc}}] \in \mathbb{R}^{H' \times W' \times 2d} \tag{6}$$

该三模态特征立方体融合了不同模态的互补信息，形成统一且表达能力丰富的特征表示，为后续的HMI机制提供充分且坚实的特征基础。

2.3 层次化模态交互机制

三模态特征立方体F^{cube}融合了视觉特征、文本特征与实体感知引导。然而，在RRSIS任务中，要有效弥合语言查询的非结构化特性与地理空间特征复杂且稠密的表示之间存在的显著模态差异，并同时充分利用实体先验实现精确定位，需要设计更加精细且高效的交互机制。传统的交叉注意力机制通常采用“扁平化”的特征块级匹配方式，这种方法忽略了模态间的语义差异，使得交互过程过于粗糙，难以满足RRSIS任务中对细粒度对齐的迫切需求。为解决上述问题，本文提出的HMI机制采用分层的交互策略，将复杂的跨模态交互过程分解为两个依次执行的阶段：跨模态语义传播和实体引导的交叉注意力。在前一阶段融合信息的基础上，通过注意力机制动态地挖掘和强化模态间细微的、与实体相关的对应关系，从而提升模型对精细地理结构的表征能力。该模块的整体结构如图4所示。

模态互激活的灵感来源于神经科学中的跨感官整合机制，其核心思想是让一个模态的特征图能够激活另一个模态中与之相关的语义区域，从而实现一种自下而上、高效的信息交换。本文引入一个无参数的交互范式，通过相互调制实现跨模态特征的双向协同增强。首先，分别从特征立方体F^{cube}中提取代表视觉信息的$F_{\text{vis}}^{\text{cube}} = F^{\text{cube}}[:, :, :, d]$和代表文本信息的$F_{\text{txt}}^{\text{cube}} = F^{\text{cube}}[:, :, d:]$，将二者的空间维度展平并$L2$归一化为

$$\hat{F}_{\text{vis}}^{\text{cube}} \in \mathbb{R}^{M \times d}, \hat{F}_{\text{txt}}^{\text{cube}} \in \mathbb{R}^{M \times d} \tag{7}$$

式中，$M = H' \times W'$。计算位置级别的相似性权重：

$$s = \sigma\left(\hat{F}_{\text{vis}}^{\text{cube}} \cdot \hat{F}_{\text{txt}}^{\text{cube}}\right) \tag{8}$$

式中，$\sigma(\cdot)$表示Sigmoid函数，用于将相似性映射到[0，1]区间，s表示两个模态对应位置的语义激活程度。接着，利用s对两种模态进行互激活：

$$\begin{aligned} F_{\text{vis}}^{\text{aug}} &= s \cdot F_{\text{vis}}^{\text{cube}} + (1 - s) \cdot F_{\text{vis}}^{\text{res}}, \\ F_{\text{txt}}^{\text{aug}} &= s \cdot F_{\text{txt}}^{\text{cube}} + (1 - s) \cdot F_{\text{txt}}^{\text{res}} \end{aligned} \tag{9}$$

式中，$F_{\text{vis}}^{\text{res}}$和$F_{\text{vis}}^{\text{res}}$分别表示与另一模态融合后的残差信息：

$$F_{\text{vis}}^{\text{res}} = F_{\text{vis}}^{\text{cube}} \odot \hat{F}_{\text{txt}}^{\text{cube}}, \quad F_{\text{txt}}^{\text{res}} = F_{\text{txt}}^{\text{cube}} \odot \hat{F}_{\text{vis}}^{\text{cube}} \tag{10}$$

这样实现了基于相似度的无参数特征互激活，促使跨模态中的高相关区域获得更强的响应，而低相关区域则保持模态独立性。

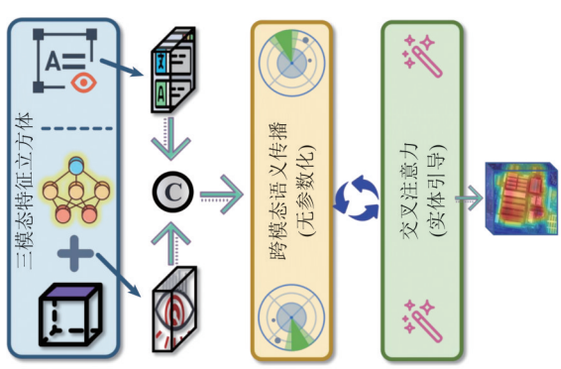


图4 层次化模态交互机制

Fig. 4 Configuration diagram of the HMI module

在前一阶段中，通过无参数模态互激活，实现了语义传播，有效弥合模态差异。为了使注意力机制更聚焦于地理实体边界，本文设计了一种非对称的、实体引导的交叉注意力机制，建立文本表达与视觉区域之间的细粒度空间依赖关系。具体而言，以互激活后的实体引导视觉特征$F_{\text{vis}}^{\text{aug}}$作为查询向量$\boldsymbol{Q}$，文本增强特征$F_{\text{txt}}^{\text{aug}}$分别作为键向量$\boldsymbol{K}$与值向量$\boldsymbol{V}$：

$$\boldsymbol{Q} = F_{\text{vis}}^{\text{aug}} \boldsymbol{W}_Q, \boldsymbol{K} = F_{\text{txt}}^{\text{aug}} \boldsymbol{W}_K, \boldsymbol{V} = F_{\text{txt}}^{\text{aug}} \boldsymbol{W}_V \tag{11}$$

式中，$\boldsymbol{W}_Q$、$\boldsymbol{W}_K$与$\boldsymbol{W}_V$为线性投影矩阵。最后，通过交叉注意力运算得到文本指向的图像特征：

$$F^{\text{out}} = \text{Proj}(CA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + F^{\text{vis}}) \tag{12}$$

式中，$CA(\cdot)$表示交叉注意力运算，Proj$(\cdot)$为1×1卷积投影层。

2.4 基于卷积的解码器

四阶段编码器共同生成一组多尺度的文本指向图像特征$\{F_t^{\text{out}} | t \in \{0, 1, 2, 3\}\}$，并将其输入解码器以生成最终的指向表达分割结果。受到ReMamber（Yang等，2024）的启发，Enti-CroM框架在解码部分采用一种基于卷积的渐进式上采样结构，该结构由4个残差层构成。整体自顶向下的预测过程可归纳为

$$\hat{F}_0^{\text{out}} = R\left(M_0(F_0^{\text{out}})\right) \tag{13}$$

$$\hat{F}_1^{\text{out}} = R\left(M_1\left(\left[\hat{F}_0^{\text{out}}; F_1^{\text{out}}\right]\right)\right) \tag{14}$$

$$\hat{F}_2^{\text{out}} = R\left(M_2\left(\left[\hat{F}_1^{\text{out}}; F_2^{\text{out}}\right]\right)\right) \tag{15}$$

$$\boldsymbol{Y} = N\left(\left[\hat{F}_2^{\text{out}}; F_3^{\text{out}}\right]\right) \in \mathbb{R}^{H \times W \times 2} \tag{16}$$

式中，$\boldsymbol{Y}$表示最终的指向分割掩码，M_*与N均通过卷积实现，其具体结构参数如图2右侧所示。

3 数据结果处理与分析

3.1 数据集介绍

为了严格评估Enti-CroM的性能，本文在两个公开且具有高挑战性的RRSIS数据集上进行了实验。RefSegRS（Yuan等，2024）数据集专为RRSIS任务构建，基于SkyScapes（Azimi等，2019）数据集的图像与像素级注释扩展而成，共包含4420组图像—文本—标签三元组。每个三元组由一幅遥感图像、一条文本指向表达，以及其所指目标的像素级分割掩码组成。这些文本表达经过精心设计，涵盖了丰富的语言要素，包括目标类别、属性及空间关系，充分贴合用户在自然语言中描述对象的习惯。值得注意的是，RefSegRS在目标尺度跨度大、朝向变化多样以及显著性特征往往较弱等方面具有显著挑战性，因此已成为RRSIS领域的一个稳健基准。RRSIS-D（Liu等，2024）数据集是一个大规模的RRSIS基准数据集，其规模与复杂度均明显超过现有的公开数据集，总计包含17402组图像—文本—标签三元组，在空间尺度跨度与目标朝向多样性方面均达到前所未有的水平，对模型在复杂航拍场景中的目标定位与分割能力提出了极高要求。该数据集的构建流程首先利用RSVG数据集的边界框提示，并结合SAM生成初始像素级掩码；随后，通过精细的人工作业对可能存在缺陷的掩码（如出现空洞区域）进行修正，从而确保标注结果的准确性与完整性。

3.2 实验参数细节与评价指标

本文提出的Enti-CroM基于PyTorch框架实现，所有训练与测试均在一张NVIDIA A6000 GPU上完成。所有输入图像在预处理阶段均被调整为480×480分辨率，batch size设为8。模型训练采用AdamW优化器，权重衰减（weight decay）设为0.01，初始学习率为5×10^{-4}，并按照多项式衰减策略逐步降低学习率，共训练40个轮次。在视觉编码器部分，本文采用在ImageNet-22K（Deng等，2009）上预训练的Swin Transformer（Liu等，2021），其包含4个不同的特征提取阶段。文本编码器为12层BERT架构，输出通道维度为768。在SEG模块中，采用SAM模型的基础版本（Kirillov等，2023）。在生成初始掩码时，将输入图像上采样至1024×1024，该步骤为离线处理。为控制计算开销，保留的初始掩码数量最多为40个，超出部分按面积由小到大进行截断。

为了全面评估本文方法的性能，本文采用先前研究中（Yuan等，2024；Liu等，2024）常用的几项评价指标：整体交并比（oIoU）、平均交并比（mIoU）和不同置信度阈值下的精确率（$Pr@X$）。IoU衡量预测分割区域$\boldsymbol{P}$与真实标注区域$\boldsymbol{G}$的重叠程度，定义为

$$\text{IoU} = \frac{|\boldsymbol{P} \cap \boldsymbol{G}|}{|\boldsymbol{P} \cup \boldsymbol{G}|} \tag{17}$$

式中，$|\cdot|$表示集合的基数。mIoU通过对所有类别或感兴趣区域的IoU值取算术平均来计算，为多类别分割任务提供更平衡的性能度量：

$$\text{mIoU} = \frac{1}{C}\sum_{c=1}^{C}\text{IoU}_c \tag{18}$$

式中，C为类别总数，IoU_c为类别c的IoU值。oIoU在全局范围内汇总所有类别与样本的交集与并集，用于评估整体分割质量：

$$\text{oIoU} = \frac{\sum_{c=1}^{C}|\boldsymbol{P}_c \cap \boldsymbol{G}_c|}{\sum_{c=1}^{C}|\boldsymbol{P}_c \cup \boldsymbol{G}_c|} \tag{19}$$

$Pr@X$衡量置信度超过X的预测中，真值例所占比例：

$$Pr@X = \frac{\text{TP}_X}{\text{TP}_X + \text{FP}_X} \tag{20}$$

式中，TP_X和FP_X分别表示在阈值X下的真值例数与伪真例数。

3.3 实验结果对比

表1与表2系统量化比较了Enti-CroM在RefSegRS与RRSIS-D数据集上与多种先进方法的性能表现。实验结果显示，基于LSTM的方法整体表现相对较弱，原因在于LSTM的序列处理特性限制了其对文本描述中细微语境关系与关键语义信息的捕捉能力。在现有方法中，LGCE（Yuan等，2024）、RMSIN（Liu等，2024）、CroBIM（Dong等，2024）和FIANet展示出较优性能，这主要得益于其针对遥感图像特征的定制化结构设计；而其他最初用于自然图像场景的方法，则依据公开源码或已发表文献在遥感数据集上进行了复现与适应性调整。无论采用CLIP还是BERT作为文本编码器，本文提出的Enti-CroM均在各项指标上持续保持领先，充分验证了所设计方法的鲁棒性与通用性。在RRSIS-D数据集上，配备CLIP编码器的Enti-CroM相较CroBIM（Lei等，2025）在mIoU上提升了2.00%，并较CARIS（Liu等，2023）在oIoU上增加了1.16%。同样地，在RefSegRS数据集上，Enti-CroM依然保持显著优势，尤其是在高阈值精确率*Pr*@0.90场景下实现了4.43%的显著提升，凸显了其在高精度指向分割任务中的卓越能力。

表1 RRSIS-D数据集上的定量比较

Table 1 Quantitative comparison on the RRSIS-D dataset

方法	文本编码器	视觉编码器	*Pr*@*X*					oIoU	mIoU
			0.50	0.60	0.70	0.80	0.90		
RRN	LSTM	ResNet-101	51.07	42.11	32.77	21.57	6.37	66.43	45.64
LSCM		ResNet-101	56.02	46.25	37.70	25.28	7.86	69.10	49.92
CMPC		ResNet-101	55.83	47.40	35.28	25.45	9.20	69.41	49.24
BRINet		ResNet-101	56.90	48.77	38.61	27.03	8.93	69.68	49.45
BKINet	CLIP	ResNet-101	56.91	48.78	39.12	27.03	9.16	69.89	49.65
ETRIS		ResNet-101	61.07	50.99	40.94	29.30	11.43	71.06	54.21
CRIS		ResNet-101	54.84	46.77	38.06	28.15	11.52	70.46	49.69
ReMamber		Mamba-B	76.38	**69.25**	55.80	42.13	22.66	76.90	64.27
LGCE	BERT	Swin-B	67.65	61.53	51.42	39.62	22.94	76.33	59.37
LAVT		Swin-B	63.98	57.57	49.30	38.06	22.29	76.16	56.82
RMSIN		Swin-B	67.16	60.36	50.16	38.72	22.81	75.79	58.79
CrossVLT		Swin-B	66.42	59.41	49.76	38.67	23.30	75.48	58.48
CARIS		Swin-B	71.50	63.52	52.92	40.94	23.90	77.17	62.12
CroBIM		Swin-B	75.00	66.32	54.31	41.09	21.78	76.37	64.24
FIANet		Swin-B	74.46	66.96	56.31	42.83	24.13	76.91	64.01
Enti-CroM	CLIP	Swin-B	**78.24**	68.99	**57.69**	44.16	24.25	**78.33**	66.24
Enti-CroM	BERT	Swin-B	77.86	68.91	57.12	**44.80**	**24.52**	78.07	**66.89**

注：加粗字体表示最优结果；下划线字体表示次优结果。

表2 RefSegRS数据集上的定量比较

Table 2 Quantitative comparison on the RefSegRS dataset

方法	文本编码器	视觉编码器	*Pr*@*X*					oIoU	mIoU
			0.50	0.60	0.70	0.80	0.90		
RRN	LSTM	ResNet-101	30.26	23.01	14.87	7.17	0.98	65.06	41.88
LSCM		ResNet-101	31.54	20.41	9.51	5.29	0.84	61.27	35.54
BRINet		ResNet-101	20.72	14.26	9.87	2.98	1.14	58.22	31.51
LSTM-CNN		ResNet-101	15.69	10.57	5.17	1.10	0.28	58.83	24.76
BKINet	CLIP	ResNet-101	36.12	20.62	15.22	6.26	1.33	63.37	40.41
ETRIS		ResNet-101	35.77	23.00	13.98	6.44	1.10	65.96	43.11
CRIS		ResNet-101	35.77	24.11	14.36	6.38	1.21	65.87	43.26
ReMamber		Mamba-B	73.96	61.25	38.78	19.44	4.41	73.26	59.00
LGCE	BERT	Swin-B	73.75	61.14	39.46	16.02	5.45	76.81	59.96
LAVT		Swin-B	71.44	57.40	32.14	15.41	4.51	76.46	57.74
RMSIN		Swin-B	71.60	55.97	31.87	11.72	1.93	71.73	57.78
CrossVLT		Swin-B	71.16	58.28	34.51	16.35	5.06	77.44	58.84
CroBIM		Swin-B	64.83	44.41	17.28	9.69	2.20	72.30	52.69
Enti-CroM	CLIP	Swin-B	**77.10**	**69.26**	44.35	**27.06**	7.82	**78.98**	**63.32**
Enti-CroM	BERT	Swin-B	76.72	69.09	**44.38**	26.49	**8.16**	78.23	63.19

注：加粗字体表示最优结果；下划线字体表示次优结果。

图5和图6直观展示了在多种具有挑战性的遥感场景中，Enti-CroM相较于LGCE与RMSIN所体现的卓越分割性能。对于尺度较大且形状规则的目标，各方法表现差异不大；然而在背景复杂、目标边界不规则、显著性不足等复杂情形下，Enti-CroM的优势尤为突出。得益于实体感知引导（SEG模块），本方法在背景杂乱或低对比度场景中能够生成更加平滑且精确的分割区域；同时，依托高效的跨模态交互设计（HMI机制），Enti-CroM在理解复杂指向描述逻辑方面表现优异，从而能够更为准确地定位目标对象。

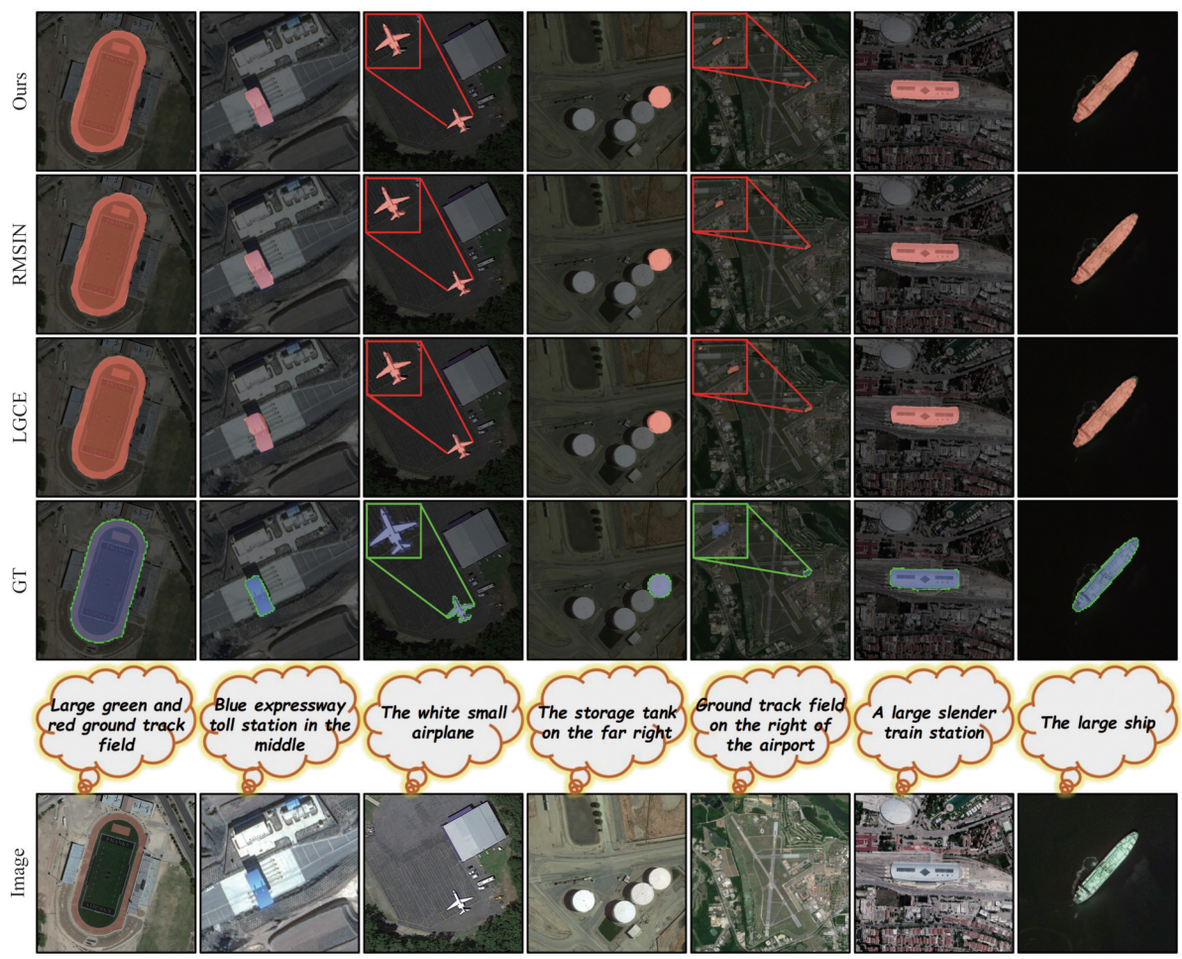


图5 RRSIS-D数据集上的定性比较

Fig. 5 Qualitative comparison of various methods on the RRSIS-D dataset

3.4 Enti-CroM关键模块消融分析

如表3所示，在RRSIS-D数据集上的消融实验结果充分验证了SEG模块与HMI机制在Enti-CroM框架中的关键作用。当仅引入SEG模块时，模型能够利用关键实体先验，有效缓解“地物混淆”现象，从而显著提升分割性能。相比之下，仅使用HMI机制时，性能提升幅度更为显著，这得益于该机制通过分层交互策略，有效弥合了跨模态语义差距。当SEG与HMI联合使用时，二者在实体先验约束与高质量跨模态交互之间形成协同效应，使完整的Enti-CroM相较于基线模型在mIoU指标上实现了7.51%的提升。与此同时，本文对各模块的计算开销进行了量化评估。数据显示，SEG模块的设计极为轻量，仅为模型带来了4 G FLOPs的额外计算量和0.10 ms的推理延迟。HMI机制作为实现细粒度交互的核心，其计算开销相对更高，引入了8 G FLOPs的计算量和0.15 ms的延迟。最终，完整的Enti-CroM模型在实现性能大幅提升的同时，总计算量控制在165 G FLOPs，单图推理时间为0.97 ms。这一结果清晰地表明，本文的方法在可控的计算成本下换取了显著的性能增益，验证了其在精度与效率之间的良好平衡。

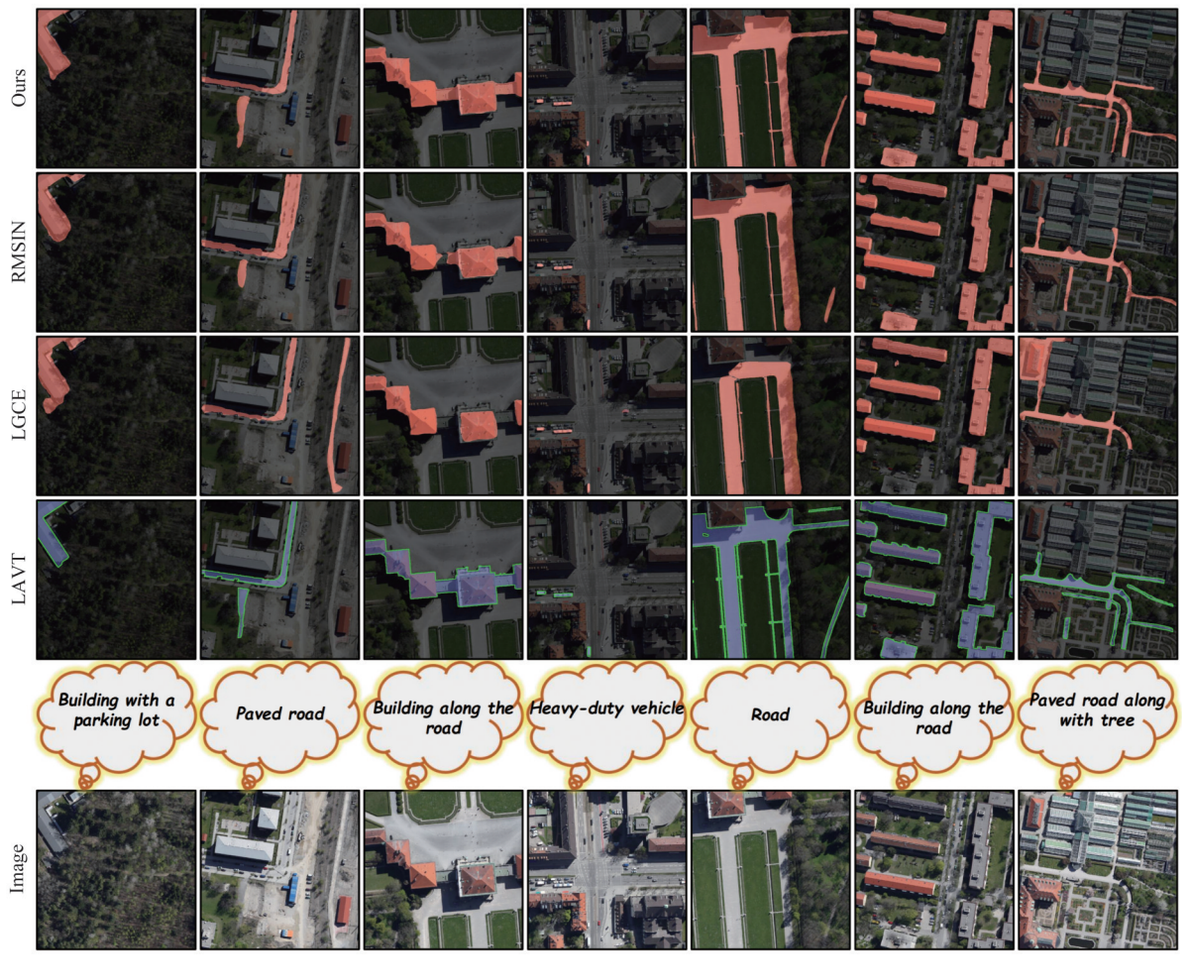


图6　RefSegRS数据集上的定性比较

Fig. 6　Qualitative comparison of various methods on the RefSegRS dataset

表3　RRSIS-D 数据集上核心组件的消融研究分析

Table 3　Ablation study of core model components on the RRSIS-D dataset

SEG	HMI	*Pr*@0.5	*Pr*@0.7	*Pr*@0.9	oIoU	mIoU	Time/ms	FLOPs/G
		69.68	51.92	21.79	74.20	59.38	0.64	154
✓		72.66	53.29	22.46	75.92	62.54	0.74	158
	✓	75.42	55.86	22.93	76.69	64.73	0.79	162
✓	✓	77.86	57.12	24.52	78.07	66.89	0.97	165

图7展示了在SEG与HMI消融配置下，模型在4个特征提取阶段的特征图可视化结果。每一行对应一种不同的模型配置：基线方法表示不包含SEG与HMI的基线模型，仅通过多阶段跨注意力机制进行模态交互；+SEG表示在基线模型中引入SEG模块；+SEG，+HMI表示完整的Enti-CroM，同时包含SEG模块与HMI机制。

由可视化结果可见，引入SEG模块后，模型能够依托实体感知引导较为准确地捕捉目标边界。然而，由于对指向性描述的语义理解仍不充分，其目标定位结果依旧相对粗糙。在此基础上进一步加入HMI机制后，模型通过实现更高质量的跨模态交互，显著增强了对复杂指向语义的理解能力，从而获得了更加精确且语义一致的分割结果。

3.5　SEG模块设计消融分析

为探究实体先验粒度对分割性能的影响，本文针对实体先验提取过程中SAM所生成的初始掩码数量进行了消融实验。如图8所示，在RRSIS-D与RefSegRS两个数据集上，mIoU指标均呈现出明显一致的变化趋势：当掩码数量由10增加至40

时，分割性能显著提升。这是由于更多的掩码能够提供更细粒度且更全面的实体感知引导，从而实现更精准的目标分割。然而，当掩码数量超过40后，性能提升幅度明显减弱并趋于饱和。这表明，在初始掩码数量达到40后，新增掩码所带来的性能增益有限，同时会引入冗余信息并增加计算开销。因此，考虑到分割精度与计算效率之间的平衡，本文在所有实验中将初始掩码数量固定为40，以期在性能和计算成本之间取得最佳平衡。

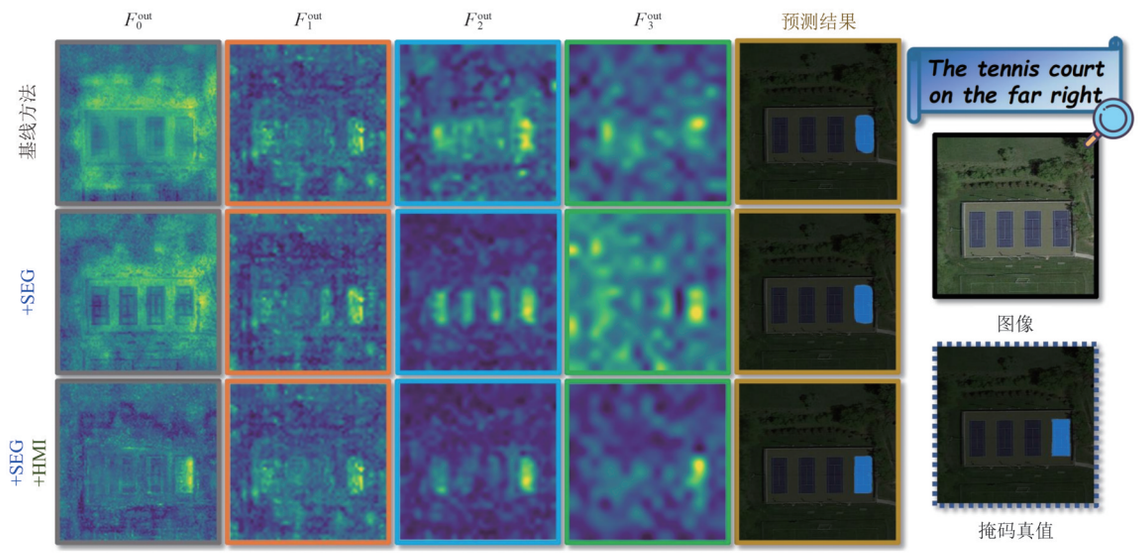


图7 4个阶段的文本指向视觉特征图

Fig. 7 Visualization of text-referred image features across the four stages

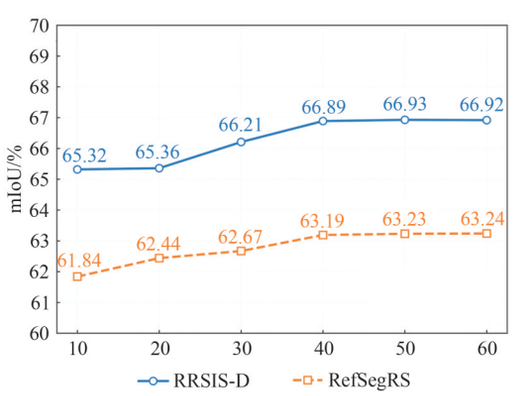


图8 初始掩码数量对性能的影响

Fig. 8 Impact of initial mask numbers

此外，表4的结果进一步从定量角度验证了SEG模块中自推理细化步骤的关键作用。无论是否结合HMI机制，包含自推理流程（w/ SR）的模型在各项评价指标上均优于未包含该流程（w/o SR）的模型。这表明，自推理能够有效利用初始掩码之间固有的空间与语义关系，将其聚合生成更加连贯且鲁棒的实体感知引导；而这种细化后的引导能够稳定提升分割性能，并增强模型在复杂遥感场景中的泛化能力。

表4 RRSIS-D数据集上SEG模块中自推理细化的消融分析

Table 4 Ablation study of self-reasoning refinement in the SEG module on the RRSIS-D dataset

不同策略		*Pr*@0.5	*Pr*@0.7	*Pr*@0.9	oIoU	mIoU
w/o HMI	w/o SR	70.25	52.40	22.16	74.89	60.73
	w/ SR	72.66	53.29	22.46	75.92	62.54
w/ HMI	w/o SR	76.31	56.56	23.27	77.29	65.51
	w/ SR	77.86	57.12	24.52	78.07	66.89

3.6 HMI模块设计消融分析

表5对不同HMI设计策略下的分割性能进行了对比分析。首先给出了跨模态语义传播SP（Semantic Propagation）与实体感知空间注意力扭合SA（Spatial Attention）分别单独作用时的结果。可以观察到，仅使用SP时模型性能显著下降，其原因在于虽然语义传播在一定程度上改善了特征分布并缩小了模态差距，但其缺乏对跨模态空间依赖关系的建模能力。在此基础上，进一步探讨了SP与SA的多种组合策略。其中，“并行”策略同时执行两步操作，并通过逐元素相加融合二者输出。从整体结果来看，按照 “SP→SA”的顺序（即先进行语义传播，再执行空间注意力建模）能

够取得最佳性能。这一现象表明，在完成跨模态语义对齐之后，再建立跨模态的空间位置依赖关系，可以更有效地提升分割结果的精确性与边界刻画能力。

表5 RRSIS-D数据集上HMI模块的设计方式分析

Table 5 Analysis of the design methodology of the HMI module on the RRSIS-D dataset

SP	SA	*Pr*@0.5	*Pr*@0.7	*Pr*@0.9	oIoU	mIoU
✓		69.85	45.24	11.13.16	66.57	51.12
	✓	76.16	56.23	23.27	76.92	65.60
Parallel		77.24	56.84	23.82	77.91	66.44
SA→SP		77.54	57.08	24.36	77.87	66.20
SP→SA		**77.86**	**57.12**	**24.52**	**78.07**	**66.89**

注：加粗字体表示最优结果。

为研究不同的跨模态语义传播策略，本文在两个数据集上对3种方法进行了对比分析。如图9所示，所提无参数模态互激活在性能上优于通道级自注意力。这得益于该方法在空间位置级别直接建模视觉与文本特征的语义匹配程度，并仅在高相关区域进行跨模态信息传递，从而提升了目标区域的语义一致性与边界细粒度表达能力；同时无参数设计有效避免了通道级全局注意力在空间粒度不足时易引入的背景噪声扩散问题，使得跨模态语义传播更加高效与精准。

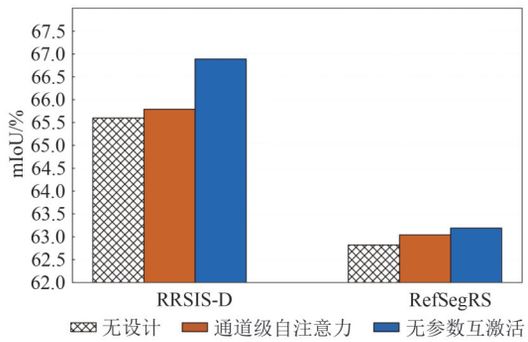
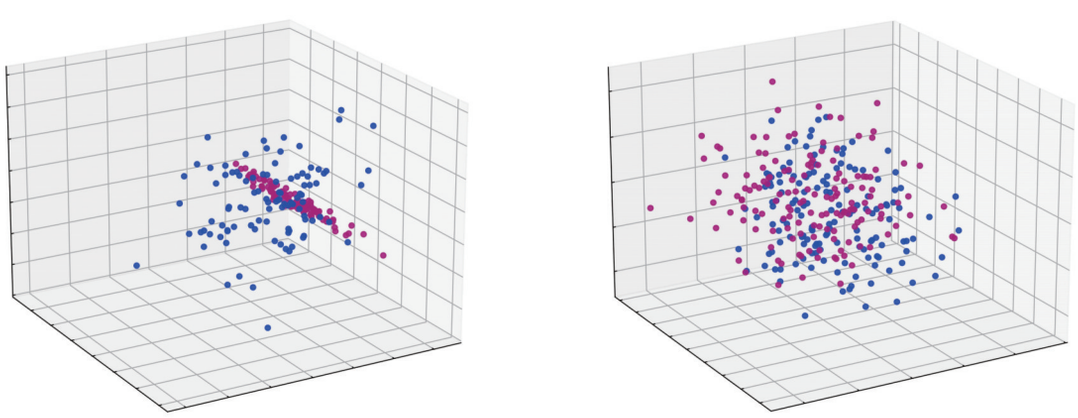


图9 HMI机制中跨模态语义传播策略对比

Fig. 9 Strategy comparison of cross-modal semantic propagation within the HMI mechanism

本文采用主成分分析（PCA）将多模态特征投影至三维空间，以直观展示HMI机制的有效性。其中，视觉特征与文本特征分别以蓝色点和红色点表示。图10（a）展示了原始数据的分布情况。由于文本特征在空间维度上被复制以与图像对齐，因此在三维空间中呈线性分布。然而，文本与视觉特征之间仍存在显著的分布差异，反映出明显的模态间差距。图10（b）展示了经过HMI机制处理后的数据分布。此时，文本与视觉特征在空间中呈现更为分散且相互交错的分布状态，这表明HMI有效促进了跨模态特征的深度融合与信息互补，从而显著缩小了模态间的语义差距。



（a）原始特征分布

（a）Original feature distribution

（b）HMI机制处理后的特征分布

（b）Feature distribution processed by the HMI mechanism

图10 HMI机制对多模态数据分布的影响

Fig. 10 Data distribution after the HMI mechanism

4 结 论

本文针对遥感影像指向分割任务中普遍存在的地物混淆与模态差距问题，提出了一种全新的跨模态分割框架Enti-CroM。该框架面向遥感多模态交互分析的特点，旨在提升文本引导下图像分割的空间精细化表征能力与跨模态语义一致性，从而满足对地观测中对精确边界刻画和高鲁棒性

的需求。

在方法设计上，本文引入了两项核心创新：（1）SAM启发的实体引导模块（SEG）：结合外部可获得的细粒度实体先验，设计自推理机制生成结构化、鲁棒且具备实体感知能力的引导信息，有效提升目标区域与边界的辨识能力；（2）层次化模态交互机制（HMI）：将跨模态语义传播与细粒度空间依赖建模解耦，并在无参数的模态互激活和实体引导跨注意力的双层结构下，突出空间位置对齐、减少噪声传递，从而显著缓解模态差距带来的信息丢失问题。

在两个具有挑战性的RRSIS基准数据集上，本文进行了系统实验与消融分析。结果表明，SEG与HMI模块单独均可带来显著性能提升，二者协同使用时整体分割性能优于当前已知最优方法，特别是在不规则地理边界刻画与跨模态语义一致性方面表现突出。这印证了本文提出的“实体引导 + 层次化交互”设计思路的有效性与适用性。该研究的成果对于提升遥感图像的智能化理解能力和交互式分析水平具有一定的参考意义，并可拓展至其他多模态遥感任务，如跨模态检索、变化检测等。

需要指出的是，本研究仍存在一定局限性：一方面，SEG模块依赖实体先验的准确度和完整性，当先验信息受噪声干扰或缺失时性能会有所下降；另一方面，尽管HMI采用了无参数互激活以降低计算开销，但在超高分辨率影像处理场景中仍存在一定的内存与推理延迟压力。这些问题的产生主要源于模型在保持高空间分辨率表达与高效跨模态交互之间的平衡难度。

未来的研究将围绕以下几个方向展开：（1）探索基于弱监督或无监督机制动态生成更鲁棒的实体先验信息，以减少对外部知识质量的依赖；（2）在保证精度的前提下，引入轻量化结构与特征压缩方法，降低在大幅面遥感影像上的计算成本；（3）扩展框架至更多类型的多模态输入（如高光谱、SAR与文本描述的联合分析），以提升模型的适应性和通用性。

参考文献（References）

An L Q, Zhang J F, Ricardo M and Zhang L. 2024. A review and prospective research of earthquake damage assessment and remote sensing. National Remote Sensing Bulletin, 28(4): 860-884 (安立强, 张景发, Ricardo M, 张磊. 2024. 地震灾害损失评估与遥感技术现状和展望. 遥感学报, 28(4): 860-884) [DOI: 10.11834/jrs.20232093]

Azimi S M, Henry C, Sommer L, Schumannn A and Vig E. 2019. SkyScapes fine-grained semantic understanding of aerial scenes//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE: 7392-7402 [DOI: 10.1109/iccv.2019.00749]

Deng J, Dong W, Socher R, Li L J, Li K and Fei-Fei L. 2009. ImageNet: a large-scale hierarchical image database//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE: 248-255 [DOI: 10.1109/cvpr.2009.5206848]

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

Dong Z, Sun Y Z, Liu T Z, Zuo W M and Gu Y F. 2024. Cross-modal bidirectional interaction model for referring remote sensing image segmentation. arXiv preprint arXiv: 2410.08613 [DOI: 10.48550/arXiv.2410.08613]

Huang W, Shi Y L, Xiong Z T and Zhu X X. 2024. Decouple and weight semi-supervised semantic segmentation of remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing, 212: 13-26 [DOI: 10.1016/j.isprsjprs.2024.04.010]

Ji L X, Du Y L, Dang Y P, Gao W Z and Zhang H. 2024. A survey of methods for addressing the challenges of referring image segmentation. Neurocomputing, 583: 127599 [DOI: 10.1016/j.neucom.2024.127599]

Jia Y Y, Gao J Y, Huang W, Yuan Y and Wang Q. 2023. Exploring hard samples in multiview for few-shot remote sensing scene classification. IEEE Transactions on Geoscience and Remote Sensing, 61: 5615714 [DOI: 10.1109/TGRS.2023.3295129]

Kirillov A, Mintun E, Ravi N, Mao H Z, Rolland C, Gustafson L, Xiao T T, Whitehead S, Berg A C, Lo W Y, Dollár P and Girshick R. 2023. Segment anything//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE: 3992-4003 [DOI: 10.1109/ICCV51070.2023.00371]

Lei S, Xiao X Y, Zhang T L, Li H C, Shi Z W and Zhu Q. 2025. Exploring fine-grained image-text alignment for referring remote sensing image segmentation. IEEE Transactions on Geoscience and Remote Sensing, 63: 5604611 [DOI: 10.1109/TGRS.2024.3522293]

Liu S A, Zhang Y H, Qiu Z F, Xie H T, Zhang Y D and Yao T. 2023. CARIS: context-aware referring image segmentation//Proceedings of the 2023 ACM International Conference on Multimedia. Ottawa: ACM: 779-788 [DOI: 10.1145/3581783.3612117]

Liu S C, Du K C, Zheng Y J, Chen J, Du P J and Tong X H. 2023. Remote sensing change detection technology in the Era of artificial intelligence: inheritance, development and challenges. National Remote Sensing Bulletin, 27(9): 1975-1987 (柳思聪, 都科丞, 郑永杰, 陈晋, 杜培军, 童小华. 2023. 人工智能时代的遥感变化检

测技术: 继承、发展与挑战. 遥感学报, 27(9): 1975-1987) [DOI: 10.11834/jrs.20222199]

Liu S H, Ma Y W, Zhang X Q, Wang H W, Ji J Y, Sun X S and Ji R R. 2024. Rotated multi-scale interaction network for referring remote sensing image segmentation//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE: 26648-26658 [DOI: 10.1109/cvpr52733.2024.02517]

Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, Lin S and Guo B N. 2021. Swin transformer: hierarchical vision transformer using shifted windows//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE: 9992-10002 [DOI: 10.1109/iccv48922.2021.00986]

Qin N N, Kang Z Z, Hui J X, Tao P J, Chen Y P, Guan H Y and Tong X D. 2025. Zero-shot remote sensing interpretation and analysis of transverse aeolian ridges in the "Tianwen-1" landing area. National Remote Sensing Bulletin, 29(2): 460-471 (秦楠楠, 康志忠, 惠建新, 陶鹏杰, 陈一平, 管海燕, 童旭东. 2025. "天问一号"着陆区横向风成脊零样本遥感解译与分析. 遥感学报, 29(2): 460-471) [DOI: 10.11834/jrs.20254328]

Tong G F, Li Y, Ding W L and Yue X Y. 2015. Review of remote sensing image change detection. Journal of Image and Graphics, 20 (12): 1561-1571 (佟国峰, 李勇, 丁伟利, 岳晓阳. 2015. 遥感影像变化检测算法综述. 中国图象图形学报, 20(12): 1561-1571) [DOI: 10.11834/jig.20151201]

Wang Q, Jia Y Y, Huang W, Gao J Y and Li Q. 2025. Embedding generalized semantic knowledge into few-shot remote sensing segmentation. IEEE Transactions on Geoscience and Remote Sensing, 63: 5603413 [DOI: 10.1109/TGRS.2024.3519772]

Xuan W H, Qi H L and Xiao A R. 2024. TSG-Seg: temporal-selective guidance for semi-supervised semantic segmentation of 3D LiDAR point clouds. ISPRS Journal of Photogrammetry and Remote Sensing, 216: 217-228 [DOI: 10.1016/j.isprsjprs.2024.07.020]

Yang Y H, Ma C F, Yao J C, Zhong Z, Zhang Y and Wang Y F. 2024. ReMamber: referring image segmentation with mamba twister// Proceedings of the 18th European Conference on Computer Vision. Milan: Springer: 108-126 [DOI: 10.1007/978-3-031-72684-2_7]

Yang Z, Wang J Q, Tang Y S, Chen K, Zhao H S and Torr P H S. 2022. LAVT: language-aware vision transformer for referring image segmentation//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE: 18134-18144 [DOI: 10.1109/cvpr52688.2022.01762]

Yuan Z H, Mou L C, Hua Y S and Zhu X X. 2024. RRSIS: referring remote sensing image segmentation. IEEE Transactions on Geoscience and Remote Sensing, 62: 5613312 [DOI: 10.1109/TGRS.2024.3369720]

Zhang L L, Wu J H, Meng Q Y, Wang X M, Du H Y and Pan J. 2025. Thermal environmental benefit based on the fine landscape pattern of urban green space. National Remote Sensing Bulletin, 29 (7): 2469-2482 (张琳琳, 吴嘉豪, 孟庆岩, 汪雪淼, 杜弘宇, 潘婧. 2025. 基于城市绿度空间精细景观格局的热环境效益研究. 遥感学报, 29(7): 2469-2482) [DOI: 10.11834/jrs.20254184]

Zhang Y X, Li W, Jia W, Zhang M M, Tao R and Liang S L. 2025a. Cross-domain hyperspectral image classification based on bi-directional domain adaptation. IEEE Transactions on Circuits and Systems for Video Technology, 35(12): 12038-12051 [DOI: 10.1109/tcsvt.2025.3586282]

Zhang Y X, Li W, Zhang M M, Han J W, Tao R and Liang S L. 2025b. SpectralX: parameter-efficient domain generalization for spectral remote sensing foundation models. arXiv preprint arXiv: 2508.01731 [DOI: 10.48550/arXiv.2508.01731]

Entity-guided cross-modal interaction for referring segmentation of remote sensing images

JIA Yuyu, CHI Kaichen, ZHOU Qing, LI Qiang, WANG Qi

School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China

Abstract: Remote Sensing Image Referring Segmentation (RRSIS) aims to accurately locate and delineate specific regions within high-resolution remote sensing imagery on the basis of natural language referring expressions and ultimately achieve pixel-level semantic interpretation. This task critically bridges user demands and intelligent geospatial information analysis. However, compared with natural scene referential segmentation, RRSIS presents two unique challenges.

(1) Relatively low contrast between targets and their surroundings often leads to a semantic dispersion phenomenon, where the segmentation mask covers irrelevant areas.

(2) Substantial cross-modal semantic gaps exist between visual and textual representations. Conventional cross-modal attention mechanisms tend to rely on coarse feature alignments, which are insufficient for fine-grained geographical boundary delineation.

The objective of this study is to design a robust and generalizable framework that can effectively mitigate semantic dispersion, narrow the modality gap, and achieve precise alignment between entity-level textual descriptions and complex geospatial visual features in RRSIS tasks.

The proposed Enti-CroM, an entity-guided cross-modal interaction framework tailored for RRSIS, is adopted in this study.

Entity-Guided Self-Reasoning (SEG) module: Motivated by the Segment Anything Model (SAM), the SEG module injects fine-grained entity priors into the model by leveraging spatial-structural constraints. A self-reasoning process generates robust and coherent entity prompts, which are integrated with visual and textual embeddings to form a trimodal entity–vision–text feature cube. Hierarchical Modality Interaction (HMI) mechanism: Parameter-Free Mutual Activation (PFMA): PFMA is a neuroscience-inspired and spatially aware mutual modulation approach that computes positionwise semantic similarity between modalities without introducing additional learnable parameters. PFMA enables efficient and precise semantic information propagation, suppresses irrelevant background interference, and reduces modality misalignment. Entity-Guided Cross-Attention (EGCA): EGCA incorporates the entity prior as an attention guide to refine the interaction between textual and visual streams and ultimately enhance the ability of the model to represent irregular and fine-grained geographical boundaries. The overall architecture decouples cross-modal semantic propagation from fine-grained spatial dependency modeling to ensure high-level semantic consistency and spatial precision.

Extensive experiments were conducted on two benchmark datasets, namely, RefSegRS and RRSIS-D, which are widely used for RRSIS evaluation. Performance was assessed via the mean intersection-over-union (mIoU) metric. Compared with the strongest existing state-of-the-art method, Enti-CroM achieved absolute mIoU improvements of +3.23% on RefSegRS and +2.62% on RRSIS-D. Ablation studies further confirmed the effectiveness of each component. The SEG module alone significantly improved target localization and robustness to background clutter. The HMI mechanism, particularly PFMA, improved modality alignment and suppression of semantic noise, whereas EGCA improved boundary representation in complex spatial contexts. Qualitative visual comparisons demonstrated that Enti-CroM delivers sharper object boundaries, more accurate correspondence to the referring expressions, and fewer false positive regions, especially in heterogeneous landscapes such as urban areas and agricultural mosaics.

This work addresses two longstanding challenges in RRSIS, namely, semantic dispersion and cross-modal gaps, by integrating entity-guided priors and a hierarchical modality interaction strategy. Incorporating spatially grounded entity cues and explicit, fine-grained semantic alignment allows Enti-CroM to substantially enhance segmentation accuracy and robustness in complex remote sensing scenes. The proposed framework not only sets new benchmarks on two challenging datasets but also offers a general paradigm for entity-aware multimodal analysis in remote sensing. Despite the advantages of the Enti-CroM, it still faces certain limitations, such as reliance on the quality of entity priors and increased computational demand for ultrahigh-resolution imagery. Future work will focus on three aspects: (1) developing adaptive or self-supervised entity prior generation mechanisms to reduce dependency on external annotations; (2) incorporating model compression and acceleration for large-scale deployment; and (3) extending the framework to integrate additional modalities, such as hyperspectral and SAR data, and broaden earth observation applications.

Key words: remote sensing imagery, referring segmentation, cross-modal interaction, SAM, entity awareness, attention mechanisms, spatial-structural constraints

Supported by National Natural Science Foundation of China (No. 62471394, U21B2041)