文章编号　1004-924X(2026)16-2578-17

双向特征调制与文本引导渐进融合的细粒度指代表达分割方法

才　华*，李军龑，寇婷婷，付　强，蔺新博，孙俊喜
（长春理工大学 电子信息工程学院，吉林 长春 130022）

摘要：现有指代表达分割方法在跨模态细粒度理解对齐和文本信息高效利用方面存在显著不足，主要表现为视觉特征的空间冗余与语义模糊；语言描述的指代歧义与词汇级语义忽略；融合模块未能充分利用文本信息。针对上述问题，本文提出双向特征调制与文本引导渐进融合的细粒度指代表达分割方法。采用预训练BEIT3作为图文编码器，设计双向特征调制模块，通过文本引导视觉调制过滤视觉空间冗余特征、视觉引导文本调制降低无关词汇权重。在此基础上，设计文本引导渐进融合模块，利用多层跨模态注意力逐步融合文本特征和图像的高层语义特征，通过多层感知机得到分割掩码。实验表明，本文方法在三个经典指代表达分割数据集上均取得了显著精度提升。其中，在RefCOCO的三个子集上准确率达到77.12%，79.46%，73.45%、在RefCOCO+的三个子集上面达到69.72%，72.86%，62.42%、在RefCOCOg的两个子集上面达到69.12%，69.58%。本文方法具有良好的泛化能力和细粒度跨模态理解对齐能力，能够实现稳定的指代表达分割任务。

关　键　词：指代表达分割；跨模态对齐；双向特征调制；文本引导渐进融合；细粒度理解
中图分类号：TP391.41　　**文献标识码**：A
doi：10.37188/OPE.20263416.2578　　**CSTR**：32169.14.OPE.20263416.2578

Fine-grained referring expression segmentation via bidirectional feature modulation and text-guided progressive fusion

CAI Hua*，LI Junyan，KOU Tingting，FU Qiang，LIN Xinbo，SUN Junxi

（*College of Electronic Information Engineering*，*Changchun University of Science and Technology.*，*Changchun* 130022，*China*）
* *Corresponding author*，*E-mail*：*caihua@cust.edu.cn*

Abstract：Current referring expression segmentation methods had key limitations. They showed spatial redundancy in visual features. Language descriptions had referential ambiguity. Lexical-level semantics were often ignored. Fusion modules failed to fully exploit text information. This paper proposed fine-grained referring expression segmentation via bidirectional feature modulation and text-guided progressive fusion (BFMPF). It used a pre-trained BEIT3 as an image-text encoder. It designed a bidirectional feature modulation module (BFMM). BFMM used global features of one modality to modulate local features of the other. This decoupled strategy suppressed noise before cross-modal interaction. Specifically,

收稿日期：2026-05-13；**修订日期**：2026-06-04.
基金项目：国家自然基金联合基金(No. U2341226)；吉林省自然科学基金(No. 20260102267JC)；2024年空间智能控制技术全国重点实验室开放基金改成光电测量与智能感知中关村开放实验室2024年度开放基金(LabSOMP-2024-14)

text-guided visual modulation filtered visual spatial redundancy. Vision-guided text modulation reduced the weights of irrelevant words. Then a text-guided progressive fusion module (TPFM) was introduced. TPFM embedded convolutions to compensate for the transformer's local modeling. It re-injected text semantics at multiple abstraction levels. It used multi-layer cross-modal attention to progressively fuse text features with high-level image features. A multilayer perceptron finally produced the segmentation mask. Experiments are conducted on three classic datasets. On RefCOCO, the method achieves 77.12%, 79.46% and 73.45% on val, testA and testB. On RefCOCO+, it reaches 69.72%, 72.86% and 62.42% on val, testA and testB. On RefCOCOg, it reaches 69.12% on val-u and 69.58% on test-u. The proposed BFMPF has strong generalization. It enables fine-grained cross-modal alignment. It performs robust referring expression segmentation.

Key words: referring expression segmentation; cross-modal alignment; bidirectional feature modulation; text-guided progressive fusion; fine-grained understanding

1　引　言

指代表达分割[1](Referring Image Segmentation, RIS)是计算机视觉与自然语言处理交叉领域的核心任务,旨在根据自然语言描述从图像中精准分割出目标对象。该任务打破了传统语义分割[2]、实例分割[3]对预定义类别集合的依赖,通过跨模态信息融合实现对自由形式语言表达的理解与视觉目标的精准分割,在人机交互、智能图像编辑、视觉导航及机器人操纵等实际场景中具有重要应用价值。

概括来说,现有指代表达分割方法可以分为3类:直接融合的方法[4-7]、基于注意力机制的方法[8-14]、大规模视觉语言预训练模型的方法[15-18]。直接融合的方法先由卷积神经网络(Convolutional Neural Network, CNN)、循环神经网络(Recurrent Neural Network, RNN)分别提取图像与文本特征,再通过拼接、卷积等简单操作融合,缺少深度跨模态交互。基于注意力机制的方法依托自注意力、交叉注意力等机制精准关联视觉与语言关键特征,并结合多尺度、多粒度策略强化表达,提升分割精度。大规模视觉语言预训练模型的方法基于 Contrastive Language-Image Pre-training (CLIP), Segment Anything Model (SAM)等[19-22]通用预训练特征,通过微调、早融合或任务转换适配分割任务,提高泛化与推理效率。

尽管现有的指代表达分割方法已取得显著进展,但在实现跨模态语义的细粒度对齐与高效特征融合方面仍存在问题。视觉特征存在空间冗余与背景干扰,传统融合难以自适应调制;文本语义挖掘不足,多数方法仅用句子级特征,忽略词汇级信息。此外,文本引导作用未充分贯穿融合过程,难以从粗到细逐步聚焦目标。这导致模型在处理复杂指代表达时,分割精度受限。

针对上述问题,本文提出双向特征调制与文本引导渐进融合的细粒度指代表达分割方法。首先,采用预训练的BEIT3[23]作为图像文本编码器,提取图像与文本特征。然后,设计双向特征调制模块(Bidirectional Feature Modulation Module, BFMM),利用图像和文本的全局表征([CLS] token)作为引导,分别对另一模态的局部特征进行筛选与增强。文本全局特征引导图像调制,抑制视觉空间冗余;图像全局特征引导文本调制,降低无关词汇权重。接着,设计文本引导渐进融合模块(Text-guided Progressive Fusion Module, TPFM),通过多层跨模态注意力将文本特征与图像高层语义逐步融合,实现从目标粗定位到边缘细分割的优化,解决融合模块文本信息利用不足的难题。最后,通过多层感知机生成分割掩码。实验结果表明,与现有方法相比,本文方法在三个经典数据集上均实现了显著的精度提升,验证了所提方法的有效性。

2　相关工作

指代表达分割旨在通过跨模态语义对齐与特征融合,依据自然语言精准定位并分割图像目

标。该任务核心难点在于高效实现视觉特征与文本特征的深度关联及互补融合。近年来研究者围绕多模态信息处理范式提出大量创新方法。本节根据多模态建模逻辑，按照直接融合交互、注意力动态关联、大规模预训练适配三类核心技术路线展开综述，明确各类方法的核心工作与技术创新。

直接融合的方法中，Hu等人[4]首次提出指代表达分割任务，并构建了一个简洁有效的网络结构，该网络首先使用卷积神经网络和长短期记忆网络提取图像和文本特征，然后通过拼接和卷积操作融合这些特征预测分割掩码。随后Liu等人[5]对其进行了改进，实现了视觉信息和语言信息中每个单词的顺序交互，使其更加符合人类视觉定位的行为方式。Li等人[6]利用图像骨干网络中固有的多尺度特征，从粗到细逐步细化分割掩码，增强多模态特征中的结构信息。此类方法融合策略简单，双模态仅浅层交互，无法捕捉细粒度关联；多为后期融合，早期特征无有效交互，信息对齐不充分，复杂场景下分割精度受限。

由于注意力机制[24]出色的全局建模和对齐能力，它被广泛应用于后续的工作中。Restr[8]是第一个无卷积的基于Transformer的模型结构，它将图像和文本特征平行输入到多模态融合编码器，捕捉这两种模态之间的细微关系。受DETR[25]启发，VLT[9]使用交叉注意力从多模态特征中生成查询向量，通过查询向量与图像信息进行交互，具有更强的建模全局上下文的能力。LAVT[10]通过交叉注意力机制将文本特征与视觉骨干网络的中间层特征进行早期融合，挖掘有用的多模态上下文信息。CMSA[11]通过自注意力机制对图像和文本特征之间的长程依赖关系进行建模，关注涉及指代表达分割目标的重要信息。随着研究深入，该类方法进一步聚焦视觉与语言模态的深度均衡交互。具体而言，BKINet[12]通过KLM用图像知识增强文本特征，生成指代核；再由KAM反哺视觉特征，形成闭环双向知识流。CrossVLT[13]在双编码器各阶段引入早期融合，实现跨模态信息的双向增强与对齐。为提升复杂场景下的定位精度。CMIRNet[14]构建CGIP模块，将视觉与语言实体构建为图结构，通过图神经网络进行跨图交互，显式建模实体关系以定位关键像素。尽管上述方法通过双向交互或图网络增强了跨模态关联，但其特征调制过程大多是整体序列间的交互，缺乏显式的、由全局语义主导的局部特征筛选机制，导致视觉背景噪声和文本虚词干扰难以被有效压制。无论是CrossVLT的分阶段双向对齐，还是BKINet的知识学习与知识应用闭环，本质上都是将一个模态的全部局部特征与另一模态的全部局部特征进行无差别混合。这种做法导致视觉特征中的背景噪声和文本特征中的虚词干扰，在缺乏显式筛选机制的情况下被不加区分地传递到对方模态，难以实现真正意义上的细粒度对齐。跨模态交互学习与迭代融合策略已在3D视觉定位[26]和RGBT跟踪[27]等任务中得到探索，通过设计双向信息流与渐进式融合机制有效提升了模态对齐精度，这进一步表明跨模态交互方式的设计对细粒度对齐任务具有关键影响。

近期指代表达分割研究开始充分利用大规模视觉语言预训练模型的强大语义理解与通用特征表征能力，采用预训练加微调的方法，将预训练模型的通用知识迁移至指代表达分割任务中。Huang等人[15]提出DETRIS，通过构建层间密集连接的参数高效微调框架，增强跨模态特征交互，适配未对齐编码器。Chen等人[16]提出SAM4MLLM，将SAM与多模态大语言模型融合，通过查询式方法为SAM生成提示点，实现像素级分割。此类方法模型参数量庞大，难以部署在低算力边缘设备；预训练特征存在天然模态间隙，高分辨率图像易引入空间冗余和背景干扰，传统全局融合难以自适应调制。现有融合策略多采用一次性或早期嵌入的方式，文本的引导作用在特征处理的后期阶段逐渐减弱，难以对分割边界的精细化调整提供持续指导，忽视词汇级语义对细粒度定位的作用，跨模态对齐与任务适配性仍需要优化。

3　本文方法

本文提出的双向特征调制与文本引导渐进融合的细粒度指代表达分割模型（BFMPF）整体网络架构如图1所示。该模型由图像文本编码器、双向特征调制模块（BFMM）、文本引导渐进融合模块（TPFM）、掩码解码器四部分组成。

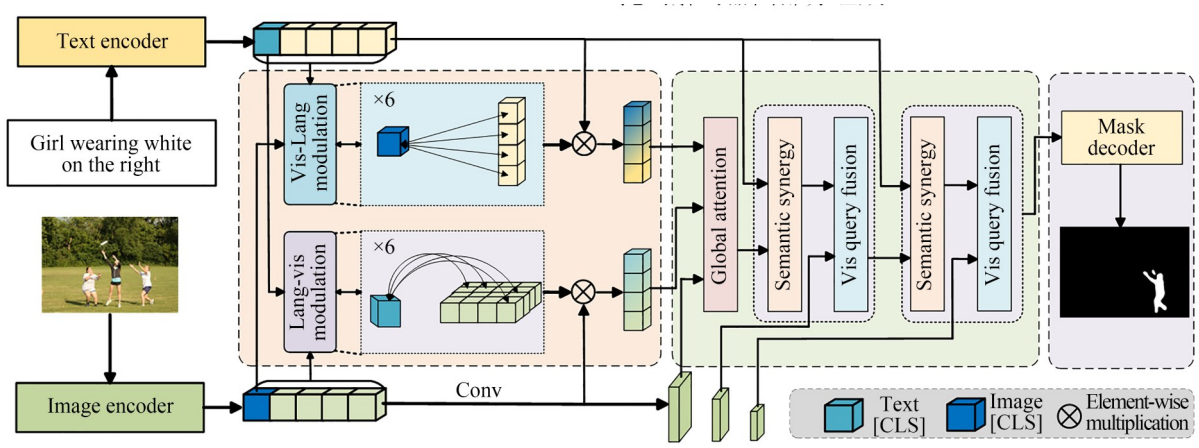

图 1　本文 BFMPF 模型的整体网络架构图

Fig. 1　Overall network architecture of the proposed BFMPF model

图像文本对输入预训练的 BEIT3,分别提取图像特征和文本特征。随后,特征送入 BFMM 进行双向调制,实现空间自适应特征选择和语义消歧。调制后的特征送入 TPFM,结合原始文本特征及卷积浓缩的视觉特征,实现图文特征的渐进式深度融合。最后,融合特征经掩码解码器生成像素级分割掩码。

3.1　基于 BEIT3 的图像文本特征编码

本文使用预训练的 BEIT3[23]作为图像文本编码器,采用双编码器结构对图像文本进行独立编码,其核心结构如图 2 所示。

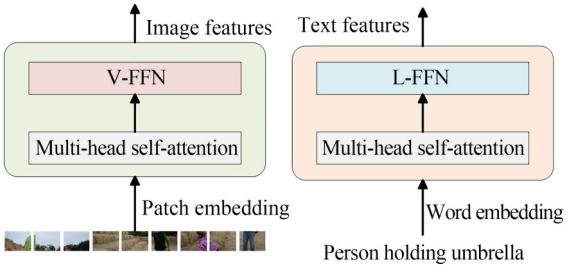

图 2　基于 BEIT3 的图文特征编码

Fig. 2　Image-text feature encoding based on BEIT3

对于输入指代表达文本,首先使用 BEIT3 指定的 SentencePiece 分词器将其转换为离散 token 序列,并添加[CLS]与[SEP]标识,同时通过[PAD]补全到固定长度。基于 BEIT3 的统一嵌入设计,将词嵌入与相对位置嵌入叠加得到原始文本嵌入,并通过掩码矩阵屏蔽[PAD]及随机掩码位置的无效特征。随后将掩码后的文本嵌入输入 BEIT3 文本分支编码器,得到文本序列特征 $F_T \in R^{T\times D}$。其中,T 为文本最大长度,D 为隐藏层维度。

$$F_T = \mathrm{TextEncoder}_{\mathrm{BEIT\text{-}3}}(\mathrm{text}). \tag{1}$$

对于输入图像,首先将其调整为固定尺寸,并采用 14×14 的分块策略将其划分为离散视觉 token 序列,得到固定长度的 Patch 序列。在 Patch 序列前端额外拼接视觉[CLS]Token,得到长度为 $N+1$ 的视觉 token 序列。视觉嵌入由 Patch 嵌入与相对位置嵌入叠加得到,并通过掩码矩阵屏蔽随机掩码位置的无效特征。将掩码后的视觉嵌入输入 BEIT3 图像分支编码器,得到图像序列特征 $F_I \in R^{(N+1)\times D}$:

$$F_I = \mathrm{ImageEncoder}_{\mathrm{BEIT\text{-}3}}(\mathrm{image}). \tag{2}$$

其中:N 为图像 Patch 序列长度,D 为隐藏层维度。

3.2　双向特征调制模块

指代表达分割的核心难点在于文本语义与图像区域的细粒度对齐,要求模型精准捕捉文本中目标类别、方位词等关键信息,挖掘词汇级语义以化解指代模糊、过滤虚词冗余;同时需剔除图像空间冗余、明确模糊语义,规避单模态偏差。

现有跨模态特征交互方案多将图像与文本的全部局部特征进行整体式双向交互,缺乏全局语义对局部特征的显式筛选机制。该模式既无法在传递前过滤视觉冗余与背景噪声,也难以抑制文本虚词干扰,导致细粒度对齐精度受限。针对这一问题,本文设计双向特征调制模块(BFMM),不同于现有工作中将图像与文本的全

部序列特征进行整体式双向交互的做法，BFMM的核心创新在于，通过全局指导局部的解耦策略，区别于现有双向交互方法所普遍采用的序列间整体混合模式。具体而言，每个模态的全局特征（[CLS] token）被提取出来，作为该模态语义的最凝练概括和代理表征，用于有选择性地对另一模态的局部特征进行调制，从而在跨模态信息传递之前即完成冗余信息的筛选与抑制。具体流程如图3所示，双向特征调制模块与已有方法结构对比如图4所示。

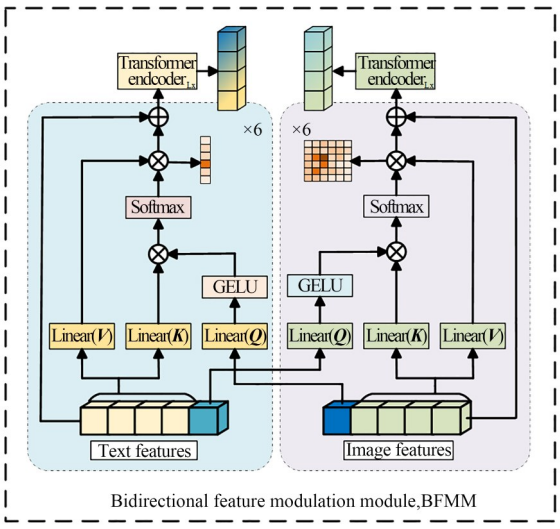

图3　双向特征调制模块结构图

Fig. 3　Structure of bidirectional feature modulation module

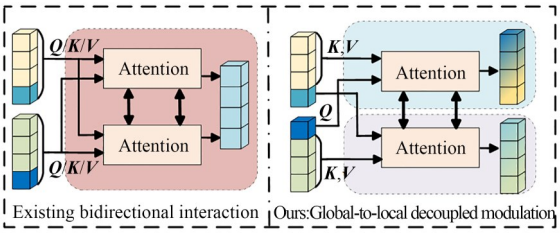

图4　不同跨模态双向交互方法的架构对比

Fig. 4　Comparison of architectures for different cross-modal bidirectional interaction methods

本模块的核心是构建两条独立的跨模态调制链路，分别实现文本全局特征调制图像特征和图像全局特征调制文本特征。具体来说，模块共设置6层调制层，该模块的输入为BEIT3输出的文本特征 F_T 和图像特征 F_I。提取[CLS]token对应的特征作为全局特征，上述过程如式(3)和式(4)所示：

$$F_{I,\mathrm{global}}=F_I[0]\in R^{1\times D}, \tag{3}$$

$$F_{T,\mathrm{global}}=F_T[0]\in R^{1\times D}. \tag{4}$$

对于图像引导文本特征调制，首先，将图像全局特征 $F_{T,\mathrm{global}}$ 和文本局部特征 F_I 分别通过线性层映射为查询 Q_I、键 K_T、值 V_T。上述过程如式(5)所示：

$$\begin{cases} Q_I=GELU(Linear_Q^I(expand(F_{I,\mathrm{global}}))) \\ K_T=Linear_K^{\mathrm{T}}(F_T),\ V_T=Linear_V^{\mathrm{T}}(F_T) \end{cases}, \tag{5}$$

其中，expand()表示将图像全局特征沿第一维度复制以匹配文本特征维度。

以全局图像查询 Q_I 作为引导，与局部文本 K_T 计算点积注意力分数，由于无关词汇对应的文本特征与图像视觉信息的相关性极低，其注意力分数会被自然压低，从而实现对冗余信息的抑制。然后，通过softmax函数对注意力分数进行归一化，得到图像引导的文本注意力权重，并利用该权重对 V_T 加权求和，得到图像全局特征调制后的文本特征 $F_{T\leftarrow I}$ 上述过程如式(6)所示：

$$\alpha_{I\to T}=softmax\left(\frac{Q_IK_T^{\mathrm{T}}}{\sqrt{D}}\right),\ F_{T\leftarrow I}=\alpha_{I\to T}V_T, \tag{6}$$

其中，$\alpha_{I\to T}$ 为图像引导的文本注意力权重。

为保留原始文本特征的基础信息，避免跨模态交互导致的信息丢失，将调制后的特征与原始文本特征进行残差连接，经6层调制后将特征输入Transformer编码器进行深度上下文建模，得到最终的图像调制文本特征 F_T'，上述过程如式(7)所示：

$$F_T'=Enc_T\left(F_T+\sum_{i=1}^{6}F_{T\leftarrow I}^{i}\right)\in R^{T\times D}, \tag{7}$$

其中，Enc_T 表示Transformer编码器层。

文本引导图像调制与图像引导文本特征调制对称，首先，将文本全局特征 $F_{T,\mathrm{global}}$ 和图像特征 F_I 分别通过线性层映射为查询 Q_T、键 K_I、值 V_I。GELU激活函数使生成的查询 Q_T 更具鲁棒性，能够更好地代表文本的全局语义，为后续的图像特征定位提供可靠引导。

$$\begin{cases} Q_T=GELU(Linear_Q^{\mathrm{T}}(expand(F_{T,\mathrm{global}}))) \\ K_I=Linear_K^I(F_I),\ V_I=Linear_V^I(F_I) \end{cases}, \tag{8}$$

其中，expand()表示将文本全局特征沿第一维度复制以匹配文本特征维度。

以全局文本查询 Q_T 为引导，与局部图像键

$\boldsymbol{K}_I$计算点积注意力分数，该分数反映图像各区域与文本语义的匹配程度，目标区域分数显著高于背景区域；经 softmax 归一化得到文本引导的图像注意力权重，随后进行 6 层残差调制并输入 Transformer 编码器建模，得到最终的文本调制图像特征 $F_I' \in \boldsymbol{R}^{(N+1)\times D}$。

$$\begin{cases} \alpha_{T\to I} = Softmax\left(\dfrac{\boldsymbol{Q}_T \boldsymbol{K}_I^{\mathrm{T}}}{\sqrt{D}}\right) \\ F_I' = Enc_I\left(F_I + \sum_{i=1}^{6} F_{I\leftarrow T}^{i}\right) \end{cases}, \tag{9}$$

其中：$\alpha_{T\to I}$为文本引导的图像注意力权重，Enc_I表示 Transformer 编码器层。

3.3　文本引导渐进融合模块

影响指代表达分割精度的关键挑战在于实现文本语义与图像视觉特征间的细粒度、渐进式对齐融合。现有方法通常采用简单的单向注意力机制，文本信息利用不充分，导致文本中的细粒度语义（属性、空间关系）与噪声（虚词）被同等对待，削弱了文本的引导精度。此外现有方法特征融合过程粗糙，常采用一步到位的融合策略，忽略了文本与图像特征在不同抽象层级上的语义鸿沟，难以实现从粗略区域定位到精细边界分割的渐进优化。

为解决上述问题，本文设计文本引导渐进融合模块（TPFM）。本模块的设计核心为卷积增强加文本重注入的渐进式融合思想。首先，本文显式嵌入卷积层，弥补 Transformer 架构在捕获局部空间细节上的固有不足，然后将视觉特征分为多个层级的抽象表达。在此过程中，文本特征并非静态的查询，而是通过语义驱动协同模块被动态更新，并重新注入更高分辨率的视觉特征中，实现文本引导作用的持续激活与深化。其中，语义驱动协同和视觉查询融合的结构如图 5 所示。

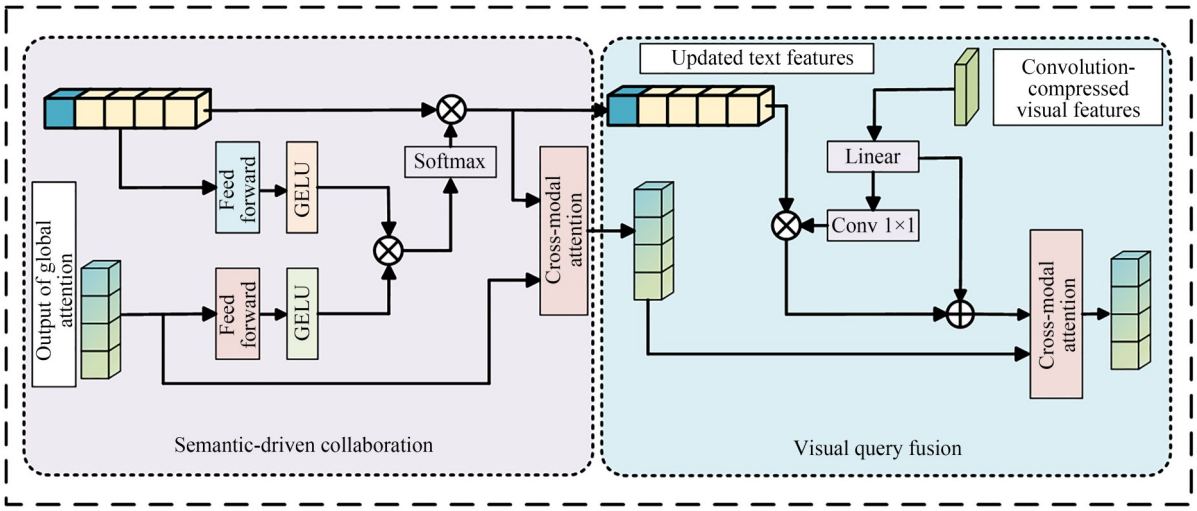

图 5　文本引导渐进融合模块结构图

Fig. 5　Structure of Text-guided Progressive Fusion Module

首先对视觉特征 F_I进行卷积操作，补全局部空间建模的缺失，让特征同时具备全局语义和局部结构信息。分别提取三个不同层次的卷积特征，具体过程如式（10）所示：

$$F_{\mathrm{In}} = GELU\left(Conv_{3\times 3}\left(F_{I(n-1)}\right)\right) \quad (n=1,2,3;\ F_{I0}=F_I), \tag{10}$$

其中，F_{In}表示第 n 个层次的卷积特征。

将经过卷积浓缩后的视觉特征作为查询向量（$\boldsymbol{Q}$），经过调制的文本特征作为键（$\boldsymbol{K}$），经过调制视觉特征作为（$\boldsymbol{V}$），通过全局注意力建立初步关联，得到初步融合特征 F_{q}，上述过程如式（11）和式（12）所示：

$$\begin{cases} \boldsymbol{Q} = Linear_Q(Upsample(F_{I1})) \\ \boldsymbol{K} = Linear_K(F_T'),\ \boldsymbol{V} = Linear_V(F_I') \end{cases}, \tag{11}$$

$$F_q = Attn(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{C_i}}\right)\boldsymbol{V}, \tag{12}$$

其中，C_i表示表示注意力机制中键向量的维度。

然后将得到的初步融合特征 F_q 与文本特征 F_T 一起送入语义驱动协同模块，通过前馈网络和跨模态注意力机制，深化文本与图像特征的交互，进一步增强文本对图像的语义引导。模块共设置两层，权重生成与文本特征更新公式如式(13)和式(14)所示：

$$W_{Tn}=Softmax\left(GELU\left(FFN\left(F_T\right)\right)\right)\#\otimes GELU\left(FFN\left(F_{\text{prev}}\right)\right)(n=1,2), \quad (13)$$

$$F_T^{xn}=F_T W_{Tn}, F_{yn}=CrossAttn\left(F_T^{xn}, F_{\text{prev}}\right), \quad (14)$$

其中：F_{prev} 在 n 为不同值时代表不同的特征，$n=1$ 时，$F_{\text{prev}}=F_q$，$n=2$ 时，$F_{\text{prev}}=F_{s1}$，F_T^{xn} 代表更新的文本特征。W_{Tn} 表示文本图像特征相似性权重矩阵。$\otimes$ 为逐元素乘法。

然后将更新的文本特征和 F_T^{xn}、语义驱动协同的输出特征 F_{yn} 和 $F_{I(n+1)}$ 一起进行视觉查询融合，融入更高层次的语义特征，实现两个模态特征的细粒度对齐。具体来说，首先将 F_{I2} 经过一个线性层和 1×1 卷积，然后与文本特征做矩阵乘法，得到的结果再与经过线性层 F_{I2} 做逐元素乘法，然后在调整维度之后，经过一个残差连接，然后经过跨模态注意力，得到融合后的结果。上述过程如式(15)～式(17)所示：

$$F_{gn}=F_T^{xn}\left(\text{Conv}_{1\times1}\left(\text{Linear}\left(Up\left(F_{I(n+1)}\right)\right)\right)\right), \quad (15)$$

$$F'_{I(n+1)}=Linear\left(Up\left(F_{I(n+1)}\right)\right)+F_{gn}, \quad (16)$$

$$F_{sn}=CrossAttn\left(F_{gn}, F'_{I(n+1)}\right), \quad (17)$$

其中：F_{gn} 表示更新的文本特征与视觉特征的初步融合结果，$F'_{I(n+1)}$ 表示更新的文本特征与视觉特征的最终融合结果，F_{sn} 表示视觉查询融合的输出。Up 表示上采样。

最终得到融合结果：

$$F_f=F_{s2}=CrossAttn\left(F_{g2}, F'_{I3}\right). \quad (18)$$

3.4 掩码解码器

本文使用多尺度特征融合掩码解码器，通过融合不同尺度、不同抽象层级的特征信息，实现目标区域的精准像素级分割。首先将融合特征 F_f 通过 1×1 卷积层进行通道维度统一与特征降维，得到基础特征图 $F_{\text{base}}\in\boldsymbol{R}^{H\times W\times C}$；同时对基础特征图进行步长为 2，4 的下采样，得到两个不同尺度的特征图，用于捕捉多尺度上下文信息：

$$\begin{cases} F_{\text{base}}=Conv_{1\times1}\left(F_f\right) \\ F_{\text{downk}}=MaxPool_{2\times2}^{k}\left(F_{\text{base}}\right)(k=2,4) \end{cases}, \quad (19)$$

其中：F_{base} 表示基础特征图，F_{downk} 表示第 K 次下采样的特征图。

对每个尺度的特征图分别应用 3×3 卷积层与批归一化(Batch Normalization)，增强特征的判别能力。其中，大尺度特征(F_{base})聚焦于目标边界与局部细节，中尺度特征(F_{down2})平衡局部细节与全局上下文，小尺度特征(F_{down4})强化全局语义定位。具体计算如下：

$$F_{\text{downk}}^{+}=BN\left(Conv_{3\times3}\left(F_{\text{downk}}\right)\right)\left(k=1,2,4;\ F_{\text{down1}}=F_{\text{base}}\right), \quad (20)$$

其中：F_{downk}^{+} 表示第 K 层增强后的特征图，$BN()$ 表示批归一化。

将中、小尺度增强特征通过双线性插值上采样至原始图像分辨率，与大尺度增强特征进行通道维度拼接，实现多尺度信息的融合：

$$\begin{cases} F_{\text{up2}}=BilinearUp_{2\times2}\left(F_{\text{down2}}^{+}\right) \\ F_{\text{up4}}=BilinearUp_{4\times4}\left(F_{\text{down4}}^{+}\right) \end{cases}, \quad (21)$$

$$F_{\text{fusion}}=Concat\left(F_{\text{base}}^{+}, F_{\text{up2}}, F_{\text{up4}}\right), \quad (22)$$

其中：$BilinearUp()$ 表示采用双线性插值进行上采样，$Concat()$ 表示拼接。

将融合后的特征图通过 1×1 卷积层映射为单通道输出，经 Sigmoid 激活函数得到最终的像素级分割掩码 $\hat{\boldsymbol{M}}\in[0,1]^{H\times W}$，其中每个像素值表示该位置属于目标区域的概率：

$$\hat{\boldsymbol{M}}=Sigmoid\left(\text{Conv}_{1\times1}\left(F_{\text{fusion}}\right)\right). \quad (23)$$

上述多尺度融合策略与目标语义分层驱动融合的思路[28]具有一致性，即通过不同抽象层次的特征协同互补，实现从全局语义定位到局部细节精化的完整建模。

3.5 损失函数

本文结合二元交叉熵损失(Binary Cross-Entropy，BCE)和 Dice 损失来构造指代表达分割任务的损失函数，兼顾像素级分类准确性与目标区域的整体匹配度，有效缓解分割任务中的类别不平衡问题。设真实分割掩码为 M，预测分割掩码为 $\hat{M}$，$M_{i,j}\in\{0,1\}$ 为真实掩码在像素 (i,j) 处的标签，$\hat{M}_{i,j}\in\{0,1\}$ 为对应位置的预测概率。则指代表达分割的总损失函数定义如

式(24)所示:

$$L_{\text{total}}=\lambda\cdot L_{\text{BCE}}+(1-\lambda)\cdot L_{\text{Dice}}, \quad (24)$$

其中:L_{total}为总损失,λ为平衡系数,经实验验证取$\lambda=0.5$时可最优权衡两类损失的优化效果,L_{BCE}和L_{Dice}分别表示二元交叉熵损失和Dice损失。

4　实验分析

本节详细介绍BFMPF在3个经典指代表达分割数据集上的实验与测试结果,与近三年的经典方法进行全面对比,验证本文模型的性能优势;同时通过消融实验分析各核心模块的有效性与关键参数的合理性,并针对真实场景开展泛化测试,结合可视化结果分析模型的细粒度分割能力。

4.1　指代表达分割数据集

本文采用RefCOCO[29],RefCOCO+[29],RefCOCOg[30]这三个经典的指代表达分割数据集进行实验,这3个数据集的图像均来源于MSCOCO[31]数据集,并配有相应的自然语言指代表达。

RefCOCO[29]数据集包含19 994张图像和142 210个对应文本描述,涵盖约50 000个对象实体。数据集分为四部分,其中训练集(train)有120 624组图像文本对、验证集(val)有10 834组图像文本对、测试集A(testA)有5 657组图像文本对、测试集B(testB)有5 095组图像文本对。

RefCOCO+[29]数据集包含19 992张图像和141 564个对应文本描述,涵盖约49 856个对象实体,文本描述不含绝对方位词。数据集划分与RefCOCO一致,其中训练集(train)有120 191组图像文本对、验证集(val)有10 758组图像文本对、测试集A(testA)有5 726组图像文本对、测试集B(testB)有4 899组图像文本对。

RefCOCOg[30]数据集包含25 799张图像和95 010个对应文本描述,涵盖约49 822个对象实体,但文本描述平均长度大于RefCOCO和RefCOCO+。该数据集有RefCOCOg-google和RefCOCOg-umd两种划分方式,其中RefCOCOg-google划分的训练集(train)有85 474组图像文本对、验证集(val-g)有9 536组图像文本对;RefCOCOg-umd划分的训练集(train)有80 512组图像文本对、验证集(val-u)有4 896组图像文本对、测试集(test-u)有9 602组图像文本对。

4.2　实验设置

本文使用预训练的BEIT3的双编码器结构提取图像文本特征。对于输入图像,将所有数据集的输入图像尺寸$H\times W$统一缩放至420×420,并采用与BEIT3预训练一致的归一化参数。输入文本使用BEIT3官方分词器进行分词。考虑到文本长度分布差异,本文将RefCOCO和RefCOCO+的最大文本长度Qmax设为20;对于平均描述长度更长的RefCOCOg单独将最大文本长度设为40。超过长度的文本将被截断,不足长度的文本则进行填充。本文的模型基于PyTorch深度学习框架实现。所有实验均在2张NVIDIA Tesla V100 32 GB GPU上完成,总批处理大小为32。优化器采用AdamW,$\beta_1=0.9$,$\beta_2=0.999$权重衰减系数设置为1×10^{-4}。初始学习率设为1×10^{-4},在第40个周期衰减至1×10^{-5},总训练轮数为60个周期。

4.3　对比实验

为了验证本文方法的有效性,本实验选取了近三年指代表达分割的8种经典模型,在3个数据集上进行了广泛的对比实验,实验结果如表1所示。

由表1可知,在RefCOCO,RefCOCO+,RefCOCOg三个数据集划分的3个验证集和5个测试集共8个子集上,本文提出的BFMPF模型整体表现优异,与近年基于大模型、多模态基础模型的最新方法形成充分对比。在RefCOCO数据集上,BEIT3-L版本BFMPF在val,testA子集分别取得77.12%,79.46%,相较DETRIS分别高出1.12%,1.26%;testB子集准确率为73.45%,略低于DETRIS的73.50%。RefCOCO+数据集去除了绝对方位描述,模型需依靠物体属性、相对位置完成目标定位。本文BEIT3-L版本BFMPF在该数据集val,testB子集分别达到69.72%,62.42%,取得最优结果。针对文本描述更长、语义关系更复杂的RefCOCOg数据集,BEIT3-L版本BFMPF在val,test子集分别达到69.12%,69.58%,在对比方法中,性能取得最优。

本文方法在RefCOCOg这类长文本场景中,依旧能够保持稳定性能,说明双向特征调制

模块可精准提取文本核心语义，有效抵御长文本中冗余词汇的干扰。与此同时，RefCOCO的testB子集聚集了大量小目标与相似干扰物体，分割难度较大，本文模型在此高难度场景下仍保持接近DETRIS的性能；而RefCOCO+缺失绝对方位词、指代歧义性强，属于典型复杂场景，本文方法在该数据集多个子集上取得最优精度。

表1　RefCOCO，RefCOCO+和RefCOCOg数据集对比实验

Tab. 1　Comparative experiments on RefCOCO, RefCOCO+ and RefCOCOg datasets　(%)

Method	Text encoder	Image encoder	RefCOCO			RefCOCO+			RefCOCOg	
			Val	TestA	TestB	Val	TestA	TestB	Val	Test
LAVT[10]	BERT	SwinB	72.73	75.85	68.79	62.14	68.38	55.10	61.24	62.09
DMMI[32]	Bert	Res101	68.56	71.25	63.16	57.90	62.31	50.27	59.02	59.24
ETRIS[33]	CLIP	ViT-B-16	70.51	73.51	66.63	60.10	66.89	50.17	59.82	59.91
FSFI[34]	BERT	Swin-B	71.23	74.34	68.31	60.84	66.49	53.24	61.51	61.78
CrossVLT[13]	BERT	SwinB	73.44	76.16	70.15	63.60	69.10	55.23	62.68	63.75
BKINet[12]	CLIP	CLIP-R101	73.22	76.43	69.42	64.91	69.88	53.39	64.21	63.77
ERIS[35]	BERT	Swin-B	75.30	78.50	70.10	63.60	69.20	55.10	64.00	64.70
RISCLIP-B[36]	CLIP	ViT-B	75.70	78.00	*72.50*	69.20	73.50	60.70	67.60	68.00
DETRIS[17]	CLIP	DINOv2-B	76.00	78.20	**73.50**	68.90	**74.00**	*61.50*	*67.90*	68.10
AMLRIS[37]	BERT	Swin-B	75.45	78.33	72.12	67.37	73.19	59.51	65.67	66.78
SSP-SAM[38]	SAM-H	CLIP	*77.12*	**80.04**	72.41	66.03	*73.45*	55.29	*68.25*	68.08
BFMPF(Ours)	BEIT3-B	BEIT3-B	76.84	78.63	71.75	*68.94*	70.25	60.33	67.84	*68.24*
BFMPF(Ours)	BEIT3-L	BEIT3-L	**77.12**	*79.46*	*73.45*	**69.72**	72.86	**62.42**	**69.12**	**69.58**

为了更好地证明本文方法的有效性，本文选择参数量、FLOPs和推理平均时间三个指标与经典的DMMI和CGFormer方法进行了比较，实验对比结果如表2所示，输入图像尺寸统一为420×420；推理速度在单张NVIDIA Tesla V100 GPU上以batch size=1进行测试。每次运行统计300次前向传播的平均耗时。具体结果如表2所示。

从表2可以看出，本文提出的BFMPF模型在模型复杂度与分割精度之间取得了良好的平衡。在参数量方面，BFMPF低于DMMI和RISCLIP，与CGFormer和Shared-RIS相比处于中等水平。在计算复杂度方面，BFMPF的FLOPs为476 G，显著低于RISCLIP和CGFormer，仅略高于DMMI和Shared-RIS。推理时间方面，BFMPF的平均处理延迟为84 ms，虽然高于Shared-RIS和DMMI，但远低于RISCLIP和CGFormer。结合表1中的分割精度，BFMPF在RefCOCOval上达到76.84%，高于DMMI、CGFormer等对比方法，证明本文方法以适中的计算开销实现了显著的精度提升。

表2　本文方法与经典模型对比结果

Tab. 2　Comparison results between the proposed method and classic models

模型	骨干网络	参数量/M	FLOPs/G	推理平均时间/ms
RISCLIP	CLIP+CLIP	375	1 380	245
CGFormer	Swin+BERT	252	949	168
DMMI	BEIT3	356	397	72
Shared-RIS	BEiT3	239	155	28
BFMPF	BEIT3	338	476	84

为直观对比不同算法的分割效果,本文选取RefCOCO数据集中子集的多组挑战性样本,将本文BFMPF模型与本领域经典的LAVT和DETRIS算法进行可视化对比,部分结果如图6所示。

由图6可以看出,本文方法在细粒度分割能力上优于对比方法。在边缘细节方面,BFMPF的分割边界更贴合物体真实轮廓,这得益于TPFM模块的多层级渐进融合;在相似物体区分方面,BFMPF能准确理解方位词语义并定位目标,这得益于BFMM中文本引导视觉调制的空间筛选机制。上述定性结果与表1的量化提升相互印证,说明本文方法在复杂指代表达场景下具有更强的跨模态对齐能力。

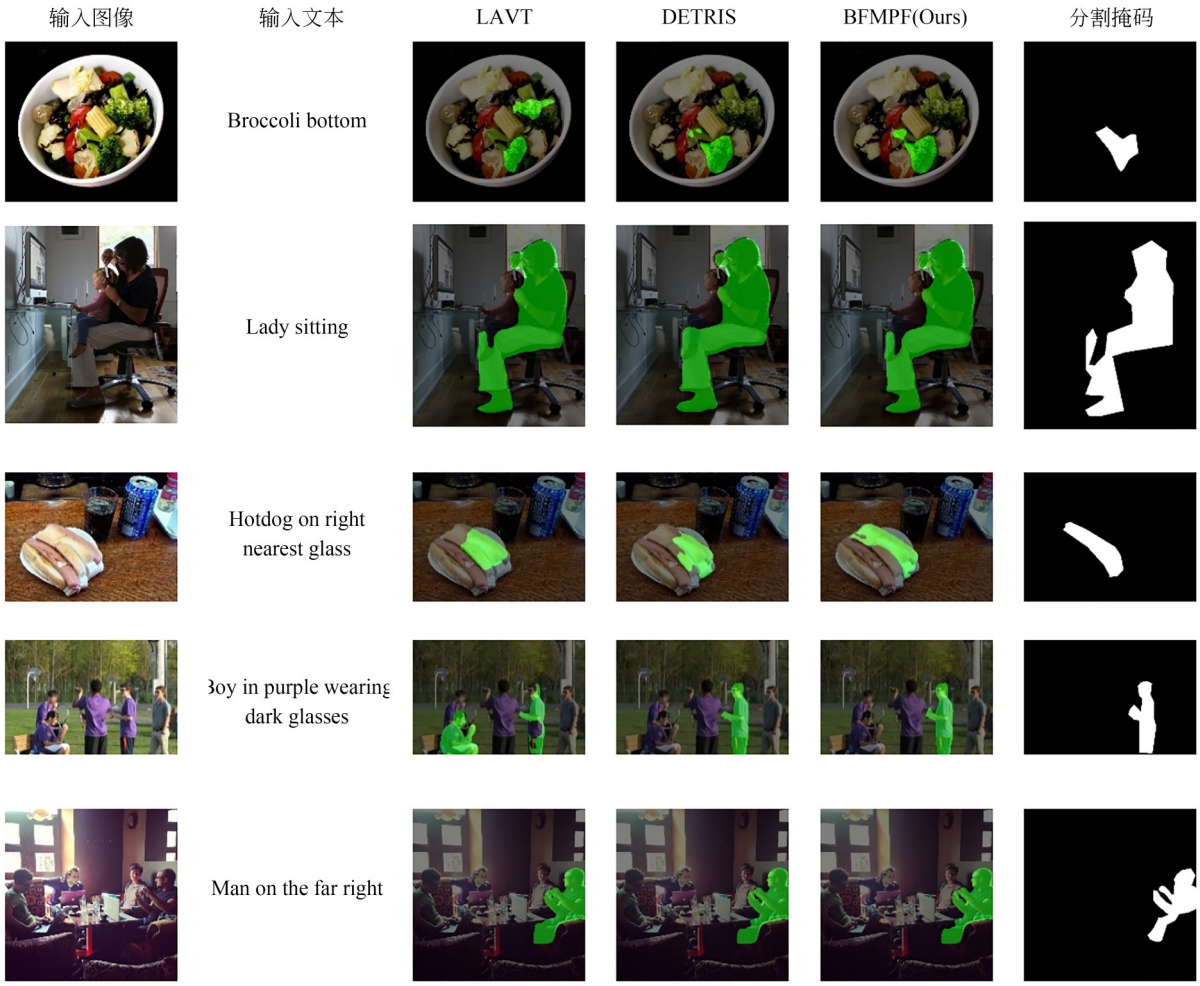

图6　对比实验可视化结果图

Fig. 6　Visualization results of comparative experiments

4.4　消融实验

为了探讨本文方法中各组件对指代表达分割准确率的影响,本节先对双向特征调制模块和文本引导渐进融合做了核心模块有效性验证的消融实验。然后,为了确保实验的全面性,分别对双向特征调制模块的层数、语言引导视觉调制和视觉引导语言调制、融合机制做消融实验。并且通过实验验证模型与骨干网络的可分离性和模型的稳定性。为了降低实验成本,本节所有实验均使用规模最小的RefCOCO数据集。

4.4.1　核心模块有效性验证

为了验证本文网络中各个模块的有效性,共设计了四组实验模型,分别命名为A1,A2,A3,A4。实验中,双向特征调制模块层数设置为6层。实验结果如表3所示。

由表3可知,在RefCOCO数据集的val,tes-

表3　核心模块组合消融实验结果

Tab. 3　Ablation experimental results of core module combination　(%)

Model Naming	BFMM	TPFM	Accuracy		
			Val	Testa	Testb
A1			72.84	76.21	65.31
A2	√		75.65	78.36	69.42
A3		√	75.14	77.88	68.25
A4	√	√	**76.84**	**78.63**	**71.75**

表4　双向特征调制模块层数消融实验结果

Tab. 4　Ablation experimental results of layer number for bidirectional feature modulation module　(%)

Layer number	Accuracy		
	Val	Testa	Testb
2	75.21	77.44	68.12
4	76.02	78.36	69.43
6	**76.84**	**78.63**	**71.75**
8	75.98	77.25	69.36

tA，testB三个划分子集上，对比基线模型A1，仅引入双向特征调制模块的A2模型分割准确率分别提升3.11%，2.15%，4.11%，仅引入文本引导渐进融合模块的A3模型分割准确率分别提升2.30%，1.67%，2.94%，这分别验证了双向特征调制模块在增强跨模态特征交互、文本引导渐进融合模块在精准引导视觉特征融合方面的有效性。

同时引入两个模块的A4模型，相较于A1实现了4.00%，2.42%，6.44%的准确率提升，且性能超越仅使用单一模块的A2与A3模型，这表明双向特征调制模块与文本引导渐进融合模块并非简单的性能叠加，而是存在显著的跨模态协同增益效应，二者结合可更充分地挖掘跨模态细粒度信息，进一步提升分割精度，充分证明了本文方法设计的合理性与优越性。

4.4.2　双向特征调制模块层数分析

双向特征调制模块的交互深度由调制层数决定，其直接影响跨模态特征融合的充分性与模型的泛化能力。层数不足会导致跨模态关联建模不充分，无法精准对齐指代表达与目标区域；层数过多则易造成过拟合与计算开销增加。为验证本文6层调制设计的最优性，本文开展调制层数消融实验，通过对比2，4，6，8层配置下的模型性能，量化分析调制层数对分割结果的影响。实验结果如表4所示。

由表4可知，随着双向特征调制模块层数的增加，分割准确率逐步提高，在6层时达到最佳效果；当层数增加至8层时，模型性能略有下降，原因是调制层数过多导致模型对训练集的特征过度拟合，同时增加了计算开销，降低了模型的泛化能力。因此6层是兼顾特征交互充分性和模型泛化能力的最优配置。

4.4.3　模块与骨干网络可分离性验证

为消除骨干网络BEIT3对本文所提模块有效性的评估偏差，验证BFMM与TPFM模块具备与主流骨干网络解耦的即插即用能力，本文进行了骨干网络替换实验。本文将BFMPF的图像文本编码器分别替换为通用且公开的Bert和SwinB组合、CLIP和ViT-B组合，并在相同的实验设置下于RefCOCO数据集上进行消融实验，其结果如表5所示。

由表中数据可知，三组不同骨干设置下，BFMPF相较各自Baseline均取得显著性能提升，表明BFMM与TPFM模块能够在不同预训练特征空间上稳定发挥跨模态细粒度对齐与融合能力，而非依赖特定骨干网络。在第一组BERT＋SwinB设置下，BFMPF在val，testA和testB上分别达到74.22%，77.15%和70.68%，均超越同骨干的CrossVLT，验证了BFMM中全局指导局部的解耦调制策略相较整体式早期双向融合在抑制视觉冗余与文本噪声方面更具优势。在第二组CLIP＋ViT-B设置下，BFMPF在val和testA上以76.15%和78.82%超越RISCLIP-B，testB上71.45%略低于后者的72.50%。分析认为，RISCLIP-B针对CLIP的图文对比对齐特性进行了深度微调设计，在testB这类密集小目标场景下具备特定优势，而BFMM与TPFM面向通用骨干设计，未对对比学习特征空间做专门适配。综合三组实验结果，无论采用何种骨干网络，BFMM与TPFM模块均能带来稳定且显著的性能增益，证明该方法的核心优势来自模块架构创新本身，当与BEIT3配对时其强大的多模态预训练表征进一步放大了这一优势。

表 5　模块与骨干网络可分离性验证实验结果

Tab. 5　Verification results of separability between module and backbone network　　(%)

Model	Text encoder	Image encoder	Accuracy		
			Val	Testa	Testb
Baseline	BERT	Swin-B	70.52	74.00	67.45
CrossVLT	BERT	SwinB	73.44	76.16	70.15
BFMPF(Ours)	BERT	Swin-B	74.22	77.15	70.68
Baseline	CLIP	ViT-B	71.84	74.85	68.21
RISCLIP-B	CLIP	ViT-B	75.70	78.00	72.50
BFMPF(Ours)	CLIP	ViT-B	76.15	78.82	71.45
Baseline	BEIT3-B	BEIT3-B	72.84	76.21	65.31
BFMPF(Ours)	BEIT3-B	BEIT3-B	76.84	78.63	71.75

4.4.4　TPFM模块渐进融合机制消融实验

为验证TPFM模块渐进融合策略的必要性和有效性,并量化多层级逐步融合相较于单层融合、一次性全局融合的性能差异,本节针对融合方式开展消融实验。本实验共设置三组对比方案:一次性融合,即取消分层交互设计,将调制完成的图文特征直接通过单次跨模态注意力完成全局融合,无多阶段特征迭代与文本重注入操作;单层融合,仅保留TPFM单个融合层级,不再进行多层卷积特征提取与递进式交互,文本语义仅进行一次引导增强;两层渐进融合,采用两级特征交互结构,实现两轮图文特征融合与文本引导;本文方案采用原TPFM多层渐进融合结构,基于三层卷积特征、语义驱动协同与视觉查询融合实现多阶段、由粗到细的渐进式融合。实验结果如表6所示。

由表6可知,模型分割精度随融合层级增加稳步提升,本文三层渐进融合方案相比一次性融合、单层融合、两层渐进融合均取得明显性能优势,在各数据集子集上分别实现不同幅度的精度增长。一次性融合与单层融合因交互次数有限,难以完成图文深度语义对齐与目标精细分割,两层渐进融合的特征融合充分性仍存在不足;而三层渐进融合通过分层提取多尺度视觉特征、多级重注入文本语义,依托多轮跨模态交互逐步弱化背景干扰、弥合模态语义差异,并结合卷积结构弥补Transformer局部建模短板,契合由粗到细的视觉感知规律。实验结果表明,渐进式多层融合是提升跨模态特征对齐效果的关键设计,本文所采用的三层结构性能最优,充分证明了TPFM模块中渐进融合机制的必要性、合理性与有效性。

表 6　TPFM 模块渐进融合机制消融实验结果

Tab. 6　Ablation experimental results of the progressive fusion mechanism of TPFM module　　(%)

融合方式	Val	Testa	Testb
一次性融合	75.71	77.53	69.86
单层融合	76.25	78.01	70.52
两层渐进融合	76.57	78.31	71.14
本文三层渐进融合	76.84	78.63	71.75

4.5　模型稳定性分析

为验证本文BFMPF模型在训练过程中的性能稳定性,消除随机初始化带来的性能波动干扰,同时量化模型输出结果的离散程度。本节选取RefCOCO数据集进行重复实验,实验设置保持不变,仅改变随机数种子,组均从零开始完整训练,不使用热启动与预加载权重。最终计算5组实验结果的均值与样本标准差,综合评估模型稳定性。实验结果如表7所示。

由表7实验结果可知,5组不同随机种子下的独立实验结果差异较小:RefCOCOval,testA,testB子集准确率波动区间分别为77.06%～77.19%,79.37%～79.50%,73.39%～73.52%,对应标准差介于0.05～0.07之间。说明本文模型网络结构设计合理,对训练过程中的随机扰动具备较强的抗干扰能力,训练过程收敛平稳。同

表 7　模型稳定性分析实验结果

Tab. 7　Experimental results of model stability analysis

(%)

随机种子	Val	Testa	Testb
1 324	77.15	79.42	73.41
2 685	77.08	79.50	73.48
3 579	77.19	79.37	73.39
4 681	77.11	79.48	73.52
5 792	77.06	79.41	73.44
均值±标准差	77.12±0.05	79.44±0.07	73.45±0.06

时重复实验均值与原文单次测试结果相吻合，进一步验证了本文算法性能的可靠性，实验结论具备可复现性。

在重复实验的基础上，本文采用配对 t 检验，将 BFMPF 与当前主流算法 DETRIS 开展统计显著性分析。基于 5 组一一对应的实验数据计算可得，RefCOCO 三个子集的检验 p 值均小于 0.01，表明本文方法相对对比模型的精度提升具有高度统计学显著性，证明性能优势并非随机训练误差导致，而是本文模块与网络架构带来的实质性提升。

4.6　泛化测试

为验证本文模型的真实场景泛化能力，对随机拍摄的真实场景图片开展泛化测试，本次测试集包含 60 组图像文本配对样本，图像中 30 张为自然环境下实拍所得，30 张为网络搜集的生活真实照片。覆盖相似目标、位置描述、颜色属性、复杂语义关系四大类典型场景，图像分辨率、光照、拍摄角度均不做人为限定，贴近实际使用环境。从定性和定量两个维度分析模型性能。

所有样本的分割真值参照 RefCOCO 数据集标注规范，由多名研究人员交叉标注、复核修正，确保标签可靠。实验软硬件配置、图像与文本预处理规则与前文实验保持一致，加载 BFMPF 训练权重完成推理。本次采用分割领域通用的交并比(IoU)与像素准确率作为评价指标，两项指标分别用于衡量预测区域与真值的重叠度、整体像素预测正确率。

定量结果显示，模型在 60 组真实场景测试中的平均 IoU 达到 78.2%，平均像素准确率达到 89.5%，在相似目标、位置描述、颜色描述等典型场景下的 IoU 均高于 75%，证明模型在非数据集的真实场景中仍具有良好的分割能力。定性结果如图 7 所示。

图 7　真实场景泛化测试结果

Fig. 7　Real-world scene generalization test results

4.7　可视化实验

图 8 展示了本文方法根据不同类型的文本描述进行指代表达分割的可视化结果,所有的 10 组图像均来自本领域经典数据集。其中,a 和 b 的文本描述仅有一个词,c 和 d 的文本描述不包含绝对位置信息,e 和 f 的文本描述包含绝对位置信息,g 和 f 的文本描述通过语义关系进行定位,i 和 j 为同一目标的不同文本描述。

图 8　本文 BFMPF 模型的指代表达分割可视化结果

Fig. 8　Referring expression segmentation visualization results of BFMPF model

可视化结果表明,本文方法在处理不同类型文本描述时均表现出良好的适应性,无论是单词简单描述还是包含属性、位置、语义关系的复杂描述,无论是含绝对方位信息还是无方位信息的描述,模型均能实现精准的指代表达分割。同时,得益于双向特征调制模块对图文特征的细粒度调制,模型对文本中的复杂语义关系的理解能力得到显著增强,能够精准解析包含属性、位置、语义关联的复杂描述,实现目标区域的精准分割;对于同一目标的不同文本描述,模型能根据描述的语义差异精准匹配对应的目标区域,进一步验证了本文方法的细粒度跨模态对齐能力。

5　结　论

针对现有指代表达分割方法在跨模态细粒度对齐、文本信息深度利用等方面的不足,本文提出了一种双向特征调制与文本引导渐进融合的细粒度指代表达分割方法(BFMPF)。该方法通过双向特征调制模块(BFMM)实现图文模态的相互调制,有效筛选文本中的关键语义信息、强化图像中的目标区域特征,解决了视觉特征空间冗余、语言描述指代歧义的问题;设计的

文本引导渐进融合模块(TPFM)通过卷积补全Transformer局部空间建模短板,结合全局注意力、语义驱动协同、视觉查询融合的多层跨模态注意力架构,实现图文特征的渐进式深度融合,充分发挥了文本特征的引导作用,实现从全局语义关联到局部细粒度对齐的优化。实验结果表明,在RefCOCO,RefCOCO+,RefCOCOg三个经典指代表达分割数据集上,BFMPF(BEIT3-L)在8个测试子集中的6个上取得最优精度,其中RefCOCO的val和testA分别达到77.12%和79.46%,RefCOCO+的val和testB分别达到69.72%和62.42%,RefCOCOg的val-u和test-u分别达到69.12%和69.58%;消融实验表明,双向特征调制模块和文本引导渐进融合模块在RefCOCO testB上分别带来4.11%和2.94%的独立增益,二者联合使用时增益达6.44%,验证了各核心模块的有效性及跨模态协同增益效应;跨骨干网络公平对比实验中,替换为BERT+SwinB和CLIP+ViT-B后平均增益仍分别达3.36%和3.84%,证明模块有效性不依赖于特定预训练骨干;真实场景泛化测试中平均IoU达78.2%,可视化结果表明模型在相似目标区分、位置关系解析、属性描述理解等复杂场景下均能实现精准的细粒度分割,具有良好的泛化能力和跨模态理解对齐能力。

作者贡献声明:

才华:方法设计,论文构思和撰写;
李军龑:实验设计,数据整理与结果可视化;
寇婷婷:论文审核与编辑,实验数据分析;
付强:论文指导,资源支持;
蔺新博:方法可行性评估与分析;
孙俊喜:论文指导与审核,实验数据分析。

参考文献:

[1] JI L X, DU Y L, DANG Y P, *et al*. A survey of methods for addressing the challenges of referring image segmentation[J]. *Neurocomputing*, 2024, 583: 127599.

[2] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 7-12, 2015, *Boston*, *MA*, *USA*. IEEE, 2015: 3431-3440.

[3] HE K M, GKIOXARI G, DOLLAR P, *et al*. Mask R-CNN[C]. 2017 *IEEE International Conference on Computer Vision* (*ICCV*). *October* 22-29, 2017. *Venice*. IEEE, 2017: 2980-2988.

[4] HU R H, ROHRBACH M, DARRELL T. Segmentation from natural language expressions[C]. *Computer Vision-ECCV* 2016. *Cham*: Springer, 2016: 108-124.

[5] LIU C X, LIN Z, SHEN X H, *et al*. Recurrent multimodal interaction for referring image segmentation[C]. 2017 *IEEE International Conference on Computer Vision* (*ICCV*). October 22-29, 2017, Venice, Italy. IEEE, 2017: 1280-1289.

[6] LI R Y, LI K C, KUO Y C, *et al*. Referring image segmentation *via* recurrent refinement networks[C]. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. *June* 18-23, 2018, *Salt Lake City*, *UT*, *USA*. IEEE, 2018: 5745-5753.

[7] MARGFFOY-TUAY E, PÉREZ J C, BOTERO E, *et al*. Dynamic multimodal instance segmentation guided by natural language queries[C]. *Computer Vision-ECCV* 2018. *Cham*: Springer, 2018: 656-672.

[8] KIM N, KIM D, KWAK S, *et al*. ReSTR: Convolution-free referring image segmentation using transformers[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 18124-18133.

[9] DING H H, LIU C, WANG S C, *et al*. VLT: vision-language transformer and query generation for referring segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(6): 7900-7916.

[10] YANG Z, WANG J Q, TANG Y S, *et al*. LAVT: language-aware vision transformer for referring image segmentation[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 18-24, 2022, *New Orleans*, *LA*, *USA*. IEEE, 2022: 18134-18144.

[11] YE L W, ROCHAN M, LIU Z, *et al*. Cross-

modal self-attention network for referring image segmentation[C]. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. *June* 15-20, 2019, *Long Beach*, *CA*, *USA*. IEEE, 2020: 10494-10503.

[12] DING H X, ZHANG S C, WU Q, *et al*. Bilateral knowledge interaction network for referring image segmentation[J]. *IEEE Transactions on Multimedia*, 2024, 26: 2966-2977.

[13] CHO Y, YU H, KANG S J. Cross-aware early fusion with stage-divided vision and language transformer encoders for referring image segmentation [J]. *IEEE Transactions on Multimedia*, 2024, 26: 5823-5833.

[14] XU M Z, XIAO T X, LIU Y T, *et al*. CMIR-Net: cross-modal interactive reasoning network for referring image segmentation[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025, 35(4): 3234-3249.

[15] HUANG J Q, XU Z N, LIU T, *et al*. Densely connected parameter-efficient tuning for referring image segmentation[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025, 39 (4): 3653-3661.

[16] CHEN Y C, LI W H, SUN C, *et al*. SAM4 MLLM: enhance multi-modal large language model for referring expression segmentation[C]. *Computer Vision-ECCV* 2024. *Cham*: Springer, 2025: 323-340.

[17] LI J C, XIE Q, GU R S, *et al*. LGD:Leveraging generative descriptions for zero-shot referring image segmentation[J]. *Pattern Recognition*, 2026, 172: 112549.

[18] TANG W, LIU X, SUN Y, *et al*. Ssp-sam: Sam with semantic-spatial prompt for referring expression segmentation[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[19] RADFORD A, KIM J W, HALLACY C, *et al*. Learning Transferable Visual Models from Natural Language Supervision [EB/OL]. 2021: arXiv: 2103. 00020. https://arxiv. org/abs/2103. 00020

[20] KIRILLOV A, MINTUN E, RAVI N, *et al*. Segment anything[C]. 2023 *IEEE/CVF International Conference on Computer Vision (ICCV)*. *October* 1-6, 2023, *Paris*, *France*. IEEE, 2024: 3992-4003.

[21] HUANG S H, DONG L, WANG W H, *et al*. Language is not all you need: aligning perception with language models[C]. *Advances in Neural Information Processing Systems* 36. *December* 1016, 2023. *New Orleans*, *Louisiana*, *USA*. *Neural Information Processing Systems Foundation*, *Inc.* (*NeurIPS*), 2023: 72096-72109.

[22] 才华, 易亚希, 付强, 等. 基于跨模态引导和对齐的多模态预训练方法[J]. 电子学报, 2024, 52 (10): 3368-3381.
CAI H, YI Y X, FU Q, *et al*. Multimodal pre-training with cross-modal guidance and alignment [J]. *Acta Electronica Sinica*, 2024, 52 (10): 3368-3381. (in Chinese)

[23] WANG W H, BAO H B, DONG L, *et al*. Image as a foreign language: BEIT pretraining for vision and vision-language tasks[C]. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. *June* 17-24, 2023, *Vancouver*, *BC*, *Canada*. IEEE, 2023: 19175-19186.

[24] Xia Z, Pan X, Song S, *et al*. Vision transformer with deformable attention[C]. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022: 4784-4793.

[25] ZHU X Z, SU W J, LU L W, *et al*. Deformable DETR: Deformable Transformers for End-to-end Object Detection [EB/OL]. 2020: arXiv: 2010. 04159. https://arxiv. org/abs/2010. 04159

[26] 才华, 冉越, 张海峰, 等. 跨模态交互学习与迭代融合的3D视觉定位[J]. 光学 精密工程, 2025, 33(24): 3915-3930.
CAI H, RAN Y, ZHANG H F, *et al*. Cross-modal interactive learning and iterative fusion for 3D visual grounding[J]. *Opt. Precision Eng.*, 2025, 33 (24): 3915-3930. (in Chinese)

[27] 邵小强, 李浩, 吕植越, 等. 针对RGBT跟踪的特殊属性的跨模态交互融合网络[J]. 光学 精密工程, 2025, 33(2): 324-336.
SHAO X Q, LI H, LÜ Z Y, *et al*. Special attribute-based cross-modal interactive fusion network for RGBT tracking [J]. *Opt. Precision Eng.*, 2025, 33(2): 324-336. (in Chinese)

[28] 陶侯卓, 罗玉婷, 赵凡. 目标语义分层驱动的红外和可见光图像融合[J/OL]. 液晶与显示, 2025, 1-12. DOI: 10.37188/CJLCD. 2025-0141. CSTR: 32172. 14. CJLCD. 2025-0141.
TAO Y ZH, LUO Y T, ZHAO F. Target semantic hierarchy driven infrared and visible image fusion

[J/OL]. *Chinese journal of liquid crystals and displays*, 2025, 1-12. DOI: 10.37188/CJLCD.2025-0141. CSTR: 32172.14.CJLCD.2025-0141.

[29] YU L C, POIRSON P, YANG S, *et al*. Modeling context in referring expressions[C]. *Computer Vision – ECCV* 2016. *Cham*: Springer, 2016: 69-85.

[30] MAO J H, HUANG J, TOSHEV A, *et al*. Generation and comprehension of unambiguous object descriptions[C]. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). *June* 27-30, 2016, *Las Vegas*, *NV*, *USA*. IEEE, 2016: 11-20.

[31] LIN T Y, MAIRE M, BELONGIE S, *et al*. Microsoft COCO: common objects in context[C]. *Computer Vision-ECCV* 2014. *Cham*: Springer, 2014: 740-755.

[32] HU Y T, WANG Q X, SHAO W Q, *et al*. Beyond One-to-one: rethinking the referring image segmentation[C]. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). *October* 1-6, 2023, *Paris*, *France*. IEEE, 2024: 4044-4054.

[33] XU Z N, CHEN Z H, ZHANG Y, *et al*. Bridging vision and language encoders: parameter-efficient tuning for referring image segmentation[C]. 2023 *IEEE/CVF International Conference on Computer Vision* (*ICCV*). October 1-6, 2023, Paris, France. IEEE, 2024: 17457-17466.

[34] YANG J X, ZHANG L H, LU H C. Referring image segmentation with fine-grained semantic funneling infusion[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, 35(10): 14727-14738.

[35] CUTTANO C, PISTILLI F, CERMELLI F, *et al*. Rethinking cross-modal interaction for efficient referring image segmentation[J]. *IEEE Robotics and Automation Letters*, 2025, 10(8): 7811-7818.

[36] KIM S, KANG M, KIM D, *et al*. Extending CLIP′s image-text alignment to referring image segmentation[C]. *Proceedings of the* 2024 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies* (*Volume* 1: *Long Papers*). *Mexico City*, *MexicoAssociation for Computational Linguistics*, 2024: 4611-4628.

[37] CHEN T F, YANG S, YANG Y G, *et al*. AML-RIS: Alignment-aware Masked Learning for Referring Image Segmentation[EB/OL]. 2026: *arXiv*: 2602.22740. https://arxiv.org/abs/2602.22740

[38] TANG W, LIU X J, SUN Y P, *et al*. SSP-SAM: SAM with semantic-spatial prompt for referring expression segmentation[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026, 36(5): 6374-6389.

通讯作者：

才　华(1977－)，男，吉林长春人，现为长春理工大学电子信息工程学院教授，博士生导师，主要从事机器视觉、自然语言处理与视觉多模态方面的研究。E-mail：caihua@cust.edu.cn

作者简介：

李军龑(2001－)，男，甘肃武威人，现为长春理工大学电子信息工程学院硕士研究生，主要从事自然语言处理与语言多模态方面的研究。E-mail：1905187102@qq.com