Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 8, No. 3, 2026

Referring Image Segmentation via Register-Aware Feature Selection and Adaptive Adaptation

Xiaozhen Gao

Henan Polytechnic University, Jiaozuo, 454000, China

Abstract: Referring Image Segmentation (RIS) aims to segment target regions in an image at pixel level based on natural language descriptions. This paper proposes RAFS (Register-Aware Feature Selection and Adaptation), a parameter-efficient framework built upon DINOv3. Three lightweight adapter modules are designed: MLFS for adaptive multi-level feature aggregation via learnable Gaussian weighting, SCSA for efficient cross-modal fusion via depthwise separable convolutions and cross-modal attention, and RAFF for leveraging register tokens' global context to enhance local features. With only 6.31M additional parameters, RAFS achieves competitive performance on RefCOCO, RefCOCO+, and G-Ref benchmarks.

Keywords: Referring Image Segmentation; Vision Foundation Model; Dinov3; Parameter-Efficient Fine-Tuning; Register Tokens; Cross-Modal Fusion

1. Introduction

Referring Image Segmentation (RIS) is a fundamental task that combines natural language understanding and visual perception. Its goal is to precisely segment the target object in an image based on a given natural language expression [1-3]. Unlike traditional semantic segmentation or instance segmentation, which rely solely on visual information for classification and segmentation, RIS requires the model to simultaneously understand the semantic content of the language description and establish precise pixel-level correspondences in the visual space. For example, given an image containing multiple people and the expression "the person in red on the left," the model needs to accurately comprehend semantic constraints such as "on the left" and "in red," and select the correct individual from multiple candidate targets for pixel-level segmentation. This dual requirement for precise language comprehension and accurate visual localization makes RIS a highly challenging task in cross-modal understanding research. With the rapid development of application scenarios such as human-computer interaction, autonomous driving, intelligent robotics, and computer-aided diagnosis, the practical value of RIS research has become increasingly prominent, garnering widespread attention from both academia and industry [4,5].

In recent years, the rapid advancement of Vision Foundation Models (VFMs) has brought a completely new research paradigm to the RIS task. Self-supervised vision models, represented by the DINO series [6-8], have learned powerful and generalizable visual feature representations through pre-training on massive amounts of unlabeled data, and their features have demonstrated excellent transferability in dense prediction tasks. In particular, DINOv2 [7] has been proven by multiple studies to provide a high-quality feature foundation for various downstream dense prediction tasks, including semantic segmentation, object detection, and depth estimation. DETRIS [9] pioneered the introduction of DINOv2 into the RIS field. By designing an adaptation structure that includes a Dense Aligner and a visual adapter, it achieved outstanding segmentation performance while keeping the backbone network frozen, demonstrating the immense potential of vision foundation models in RIS tasks. The core advantage of this paradigm lies in the fact that, by freezing the large-scale pre-trained backbone network and training only a small number of adaptation parameters, it can fully leverage the rich visual knowledge learned by the foundation model while significantly reducing training costs and the risk of overfitting.

However, existing DINOv2-based RIS methods still have limitations in the following three aspects. First, regarding multi-layer feature utilization, current methods typically adopt a manual strategy to fuse features from fixed layers (e.g., the 4th, 8th, and 12th layers). Different layers of a Vision Transformer encode different levels of information, ranging from low-level textures to high-level semantics. A fixed layer selection strategy cannot adaptively adjust the feature combination method according to the characteristics of different samples and language descriptions. For instance, for expressions describing object colors, shallow color features might be more critical; whereas for expressions describing spatial relationships, deep global semantic features are more important. Second, in terms of cross-modal fusion, although the Dense Aligner in DETRIS is effective, its structure is relatively complex. It includes two sub-modules: a Dense Mixture of Convolutions (D-MoC) and full-scale cross-modal attention, which introduce a considerable amount of extra parameters and, to a certain extent, increase training complexity. Third, DINOv2 and DINOv3 introduce Register Tokens into the Vision Transformer. Studies have shown that these special tokens gather global semantic information during forward propagation [24]. However, existing RIS methods typically discard register tokens directly during the feature extraction stage and solely use patch features, resulting in the waste of valuable global contextual information.

In August 2025, Meta AI released DINOv3 [8] as the latest version of the DINO series of vision foundation models. Compared to DINOv2, DINOv3 features two key improvements. First, it introduces the Gram Anchoring training strategy, which effectively addresses the dense feature degradation problem that occurs during the prolonged training of large-scale models by constraining the Gram matrix of the current model's features to align with early stable checkpoints. This results in a substantial improvement

in the quality of the final model's dense features. Second, it improves the design and training mechanism of register tokens, enabling them to carry global semantic information more effectively while further reducing high-norm artifacts in feature maps. These improvements provide a higher-quality feature foundation for dense prediction tasks based on vision foundation models, which naturally creates conditions for further advancements in the RIS task.

Based on the above analysis, this paper proposes RAFS (Register-Aware Feature Selection and Adaptation), a referring image segmentation framework based on the DINOv3 vision foundation model. RAFS achieves effective adaptation to backbone features through three lightweight, innovative modules while keeping the DINOv3 backbone network completely frozen. The main contributions of this paper are as follows:

(1) We propose the Multi-Layer Feature Selection (MLFS) module, which adaptively aggregates multi-scale feature representations from all 12 intermediate layers of the backbone network through a learnable Gaussian weighting function. This replaces the traditional fixed-layer selection strategy, enabling the model to dynamically adjust the feature combination method based on the characteristics of different samples and language descriptions.

(2) We design a Simplified Cross-Scale Adapter (SCSA) that utilizes depthwise separable convolutions to extract multi-scale local features and achieves efficient fusion of visual and textual features through a lightweight cross-modal attention mechanism. This significantly reduces the number of parameters while maintaining effective cross-modal interaction capabilities.

(3) We propose the Register-Aware Feature Fusion (RAFF) module, which explicitly utilizes the global contextual information carried by DINOv3's register tokens in the RIS task for the first time. Through a gated bottleneck mechanism, global semantics are injected into local patch features, enhancing the model's global perception capability.

(4) We conduct extensive experimental validation and systematic ablation analyses on three standard benchmark datasets: RefCOCO, RefCOCO+, and G-Ref. The results demonstrate that RAFS achieves performance comparable to existing state-of-the-art methods while introducing only 6.31M adapter parameters.

2. Related Work

2.1. Referring Image Segmentation

Referring Image Segmentation (RIS) aims to localize and segment target objects in an image based on natural language expressions, representing one of the core tasks in the field of vision-language understanding. Research in this domain has undergone an evolution through approximately three stages. Early works primarily adopted a two-stage strategy, first employing object detectors to extract candidate regions and then selecting the optimal candidate for segmentation through language matching [10,11]. However, the performance of these methods was inherently limited by the quality of the candidate regions. Subsequently, end-to-end methods based on Fully Convolutional Networks (FCNs) became the mainstream paradigm [1,12]. By fusing visual and linguistic features across different network levels to directly generate segmentation results, these methods successfully bypassed the bottleneck caused by candidate region quality in two-stage approaches.

2.2. Visual Foundation Models

Vision Foundation Models (VFMs), which learn general-purpose visual feature representations with strong generalization capabilities through pre-training on massive data, have become crucial infrastructure in current computer vision research. Based on their training paradigms, vision foundation models can be broadly categorized into three types: supervised pre-trained models (e.g., ImageNet pre-trained ResNet [16]), which acquire robust classification features trained on large-scale annotated data; weakly supervised pre-trained models (e.g., CLIP [17]), which learn vision-language aligned representations using internet-scale image-text pairs based on contrastive learning; and self-supervised pre-trained models, which learn effective visual representations without requiring any annotated data. Self-supervised methods possess the strongest scalability due to their independence from manual annotations, with approaches represented by self-distillation and contrastive learning being particularly successful.

DINO [6] first implemented the self-distillation training paradigm on Vision Transformers. The learned features exhibited remarkable semantic segmentation properties, capable of generating semantically consistent attention maps without any annotations. DINOv2 [7] substantially improved feature quality by scaling the training data to 142 million images and employing a ViT-giant level teacher model for distillation. The recently released DINOv3 [8] introduces two key improvements over DINOv2. First is the Gram Anchoring training mechanism, which effectively addresses the gradual degradation of dense features during the prolonged training of large models by adding a regularization loss for the feature Gram matrix in the later stages of training. Second is the improved training strategy for register tokens, enabling them to more effectively aggregate global semantic information and reduce feature artifacts. DINOv3 has achieved state-of-the-art performance on multiple dense prediction benchmarks, including ADE20k semantic segmentation (63.0 mIoU) and COCO object detection (66.1 mAP). This paper selects DINOv3 ViT-B/16 as the visual backbone network, fully utilizing its high-quality dense features improved by Gram Anchoring and its information-rich register tokens.

2.3. Model Adaptation Strategies

Model adaptation aims to fit large-scale pre-trained models to specific downstream tasks [18]. Compared to standard full-parameter fine-tuning, targeted adaptation methods can significantly reduce the computational cost and storage overhead of training. Furthermore, by focusing updates on specific modules, they also alleviate overfitting issues in few-shot scenarios to some extent. Current mainstream adaptation methods include: Adapter methods [19], which insert small bottleneck networks between Transformer layers and only train the parameters of these bottlenecks; LoRA [20], which approximates the weight update matrix through low-rank matrix factorization to achieve effective fine-tuning without changing the model structure; and Prompt Tuning [21], which adds learnable prompt vectors to the input sequence to guide the model's adaptation to downstream tasks.

In the field of computer vision, VPT [22] first introduced the concept of prompt tuning to Vision Transformers; AdaptFormer [23] designed an Adapter structure parallel to the Transformer's feed-forward network, achieving excellent transfer learning results. In the RIS task, DETRIS [9] designed the Dense Aligner as a visual adapter, achieving

alignment and fusion of multi-layer features through Dense Mixture of Convolutions (D-MoC) and cross-modal attention. The SCSA module in this paper can be regarded as a simplification and improvement of the Dense Aligner. By performing multi-scale feature extraction and cross-modal attention computation in a lower-dimensional space, it maintains effective cross-modal fusion capabilities; meanwhile, MLFS and RAFF are newly designed adaptation modules specifically tailored to the multi-layer feature structure and register token characteristics of DINOv3.

3. Method

3.1. Overall Architecture

The overall architecture of the proposed RAFS framework is illustrated in Figure 1. The entire framework consists of four core components:(1) A frozen DINOv3 ViT-B/16 visual backbone network, responsible for extracting 12 layers of multi-scale visual features and their corresponding register tokens from the input image.(2) A frozen CLIP ViT-B/16 text encoder (with only the adapter layers being trainable), responsible for encoding natural language descriptions into word-level features and sentence-level semantic states.(3) Three trainable lightweight adaptation modules—MLFS, SCSA, and RAFF—which are responsible for the adaptive selection of multi-layer features, vision-text cross-modal fusion, and the injection of global register information, respectively.(4) A shared multimodal decoding head, comprising an FPN-style Neck module, a 6-layer Transformer Decoder, and a convolutional Projector, responsible for progressively decoding the adapted multi-scale features into the final pixel-level segmentation masks.

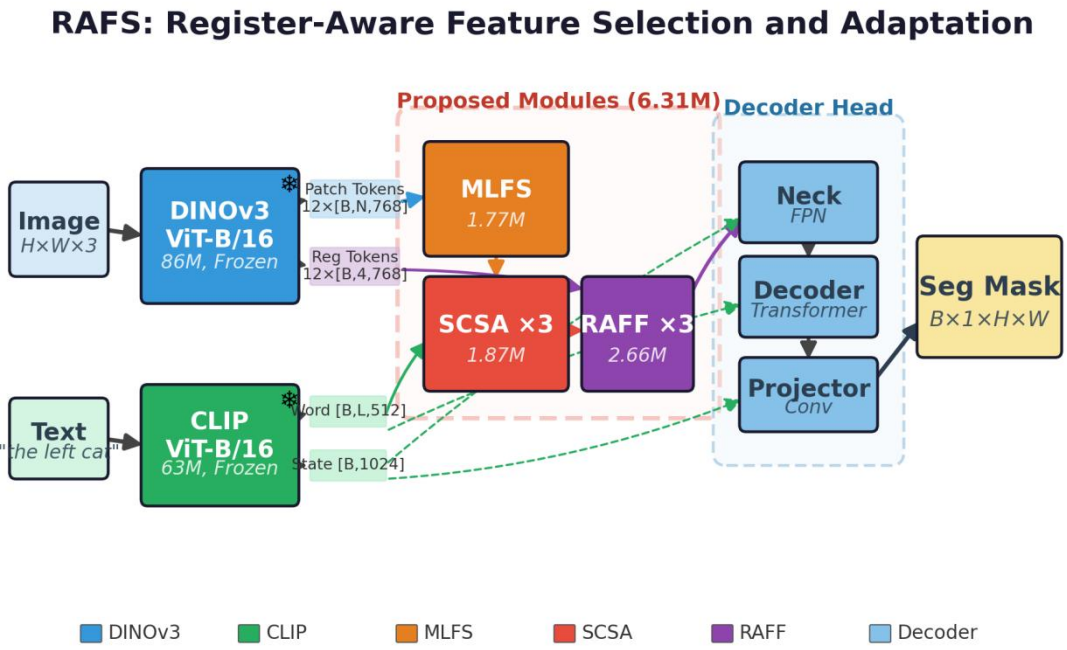

Figure 1: Overview of the proposed RAFS network architecture.

3.2. Multi-Layer Feature Selection(MLFS)

When utilizing the multi-layer features of Vision Transformers, existing methods typically adopt a manual strategy of selecting fixed layers. For instance, DETRIS selects the features from the 4th, 8th, and 12th layers to correspond to shallow, intermediate, and deep multi-scale representations, respectively. However, this hard-coded strategy has two limitations. First, different samples may require features from different layer combinations—for simple scenes, shallow spatial features might be sufficient, whereas complex descriptions of relative positions necessitate deeper semantic features. Second, fixed-layer selection fails to leverage the complementary information between adjacent layers. For example, the features of the 7th and 8th layers might each contain different aspects of information that are valuable for the current sample.

To address this, we propose the MLFS module, which adaptively aggregates multi-scale representations from all 12 layers of features through a learnable soft-selection mechanism. Specifically, MLFS defines K groups of outputs, with each group corresponding to a learnable center position μ_k and a bandwidth parameter σ_k. For the k−th group of outputs, its aggregation weight for the feature of the l−th layer is calculated via a Gaussian function:

$$w_k^l = \frac{\exp\left(-\frac{(l-\mu_k)^2}{2\sigma_k^2}\right)}{\sum_{j=1}^{12}\exp\left(-\frac{(j-\mu_k)^2}{2\sigma_k^2}\right)} \quad (1)$$

where is initialized to the center positions of the 12 layers uniformly partitioned into 3 groups (i.e.,), and is initialized to zero. A normalization operation ensures that the sum of the weights across all layers equals 1. The weighted aggregated features undergo independent linear projections to obtain the final output:

$$F_k = \text{Linear}_k\left(\sum_{l=1}^{12} w_k^l \cdot F_p^l\right) \quad (2)$$

The core advantage of this design lies in its implementation of soft selection for feature layers in a continuously differentiable manner. Through the two parameters of the Gaussian function—center position and bandwidth—each group of outputs can flexibly focus on an arbitrary range of layers: when small, the aggregation concentrates on a few nearby layers, achieving an effect similar to hard selection; when large, the aggregation range expands to more layers,

integrating richer multi-scale information. These parameters are automatically learned during end-to-end training, enabling the network to find the most suitable layer combination strategy for the RIS task. The MLFS module introduces a total of 1.77M parameters (3 sets of linear projection layers).

3.3. Simplified Cross-Scale Adapter(SCSA)

Cross-modal fusion is a core component of the RIS task, and its quality directly determines the model's accuracy in understanding language descriptions and the precision of segmentation boundaries. The Dense Aligner in DETRIS achieves feature alignment through D-MoC (a mixed convolution module containing various kernel sizes) and full-scale cross-modal attention. While effective, its structure is relatively complex. This paper designs SCSA as a simpler and more efficient alternative. The core idea is to simultaneously accomplish multi-scale spatial feature extraction and cross-modal attention fusion within a low-dimensional bottleneck space.

For the input visual features $F\in\mathbb{R}^{B\times N\times D}$ and textual features F_w, the processing pipeline of SCSA consists of four stages:

Dimensionality reduction projection: First, a linear layer is used to reduce the dimensionality of the visual features from D=768 to D/4=192, significantly reducing the complexity of subsequent computations: $X=\mathrm{GELU}\left(\mathrm{Linear}_{\mathrm{down}}(F)\right)$

Multi-scale spatial feature extraction: The dimensionality-reduced features are reshaped into a 2D feature map $X_{2d}\in\mathbb{R}^{B\times d\times H\times W}$, and spatial information from different receptive fields is extracted through 3×3 and 5×5 depthwise separable convolutions, respectively. The 3×3 convolution captures local texture and edge information, while the 5×5convolution covers a broader range of context. The features of the two scales are concatenated along the channel dimension and then fused through a 1×1 convolution to achieve effective integration of information from different scales. The use of depthwise separable convolutions substantially reduces the number of parameters—compared to standard convolutions, the parameter count of depthwise separable convolutions is only at the 1/d level.

Cross-modal attention: The text features are projected into the same 192-dimensional space as the visual features, and then cross-modal interaction is achieved through an 8-head multi-head attention mechanism, utilizing the visual features as the Query, and the text features simultaneously as the Key and Value. This design enables each visual position to selectively enhance or suppress its own features based on the key vocabulary in the text description, thereby achieving language-guided feature modulation.

Residual connection: Finally, the features are projected back to the original 768-dimensional space via a linear layer, and fused with the input features through a learnable scaling factor α (initialized to 0.1) and a residual connection:

$\mathrm{SCSA}(F,F_w)=\mathrm{LayerNorm}(F+\alpha\cdot\mathrm{Linear}_{\mathrm{up}}(X_c))$. A small initial scaling factor ensures that the adapter's output has a minor impact on the original features during the early stages of training, avoiding disruption to the distribution of the pre-trained features and benefiting training stability. As training progresses, α will automatically adjust to an appropriate value. Each SCSA module introduces approximately 0.62M parameters, totaling 1.87M for 3 layers.

3.4. Register-Aware Feature Fusion Module(RAFF)

DINOv3 introduces 4 learnable register tokens into the input sequence of the Vision Transformer. The study by Darcet et al. [24] revealed the important role these register tokens play during the model's forward propagation: they automatically gather global semantic information, functioning similarly to "global memory slots", while effectively improving the quality of the patch feature maps by absorbing high-norm artifacts that would otherwise be dispersed among the patch features. However, existing RIS methods typically only retain patch features and discard both register tokens and CLS tokens when utilizing Vision Transformer features, which means the rich global contextual information embedded in the register tokens is wasted.

This paper proposes the RAFF module, aiming to effectively integrate the global contextual information carried by the register tokens into the patch features through a gated bottleneck mechanism. Its processing pipeline is as follows:

Global context extraction: Mean pooling is performed on the 4 register tokens to obtain a compressed global context vector $g=\frac{1}{4}\sum_{r=1}^{4}F_r^l[r]$. The mean pooling operation aggregates the global information captured individually by the 4 register tokens, resulting in a comprehensive global semantic representation.

Adaptive gating: A dimension-wise gating signal $\mathrm{gate}=\sigma(\mathrm{Linear}_g(g))$ is generated through a linear layer and a Sigmoid activation function. The value of each dimension in the gating signal ranges from 0 to 1, determining the injection strength of the global information on the corresponding dimension. This design enables the model to selectively fuse global information—certain feature dimensions might require stronger global context (such as dimensions encoding spatial relationships), while other dimensions might rely more on local information.

Bottleneck transformation: The global context vector undergoes a non-linear transformation through a bottleneck structure: $\mathrm{ctx}=\mathrm{Linear}_{\mathrm{up}}(\mathrm{GELU}(\mathrm{Linear}_{\mathrm{down}}(g)))$. Its dimensionality is first reduced to D/4, passed through a GELU activation, and then restored to D. This bottleneck design introduces a non-linear transformation while compressing the representation space, enabling the module to learn a more effective global semantic representation.

Gated fusion and normalization: $\mathrm{RAFF}(F_p,F_r)=\mathrm{LayerNorm}(F_p+\mathrm{gate}\odot\mathrm{ctx})$, where $\odot$ denotes element-wise multiplication. Each RAFF module introduces approximately 0.89M parameters, totaling 2.66M for 3 layers.

3.5. Training Objective

RAFS employs a standard binary cross-entropy loss function for end-to-end training:

$$\mathscr{L}=-\frac{1}{\mathrm{HW}}\sum_{i=1}^{\mathrm{HW}}[m_i\log\hat{m}_i+(1-m_i)\log(1-\hat{m}_i)] \quad (3)$$

where m_i and $\hat{m}_i$ are the ground-truth label and predicted probability of the i -th pixel, respectively. During training, the parameters of the DINOv3 and CLIP backbone networks are completely frozen, and only the parameters of the three adaptation modules (6.31M parameters) and the shared decoding head (22.80M parameters) are updated, totaling 29.53M trainable parameters, which accounts for 16.53% of the total model parameters.

4. Experiments

4.1. Datasets

The proposed method is evaluated on three standard RIS benchmark datasets, which are complementary in terms of the types and complexity of their language descriptions:

RefCOCO [25]: Contains 19,994 images and 142,210 referring expressions, partitioned into four subsets: train (120,624), val (10,834), testA (5,657), and testB (5,095). The images in the testA subset mainly contain multiple people, requiring the model to distinguish between different individuals; the images in the testB subset primarily contain multiple objects, requiring the model to distinguish between objects of different categories. The language descriptions in this dataset allow the use of positional words (e.g., "left", "right"), thus imposing higher requirements on spatial reasoning capabilities.

RefCOCO+ [25]: Contains 19,992 images and 141,564 referring expressions, partitioned in the same manner as RefCOCO. The key difference from RefCOCO is that absolute positional words are prohibited. Consequently, the language descriptions rely more heavily on the appearance attributes of objects (such as color, size, and shape), raising higher demands on the model's visual attribute comprehension capabilities.

G-Ref [26]: Contains 26,711 images and 104,560 referring expressions, divided into three subsets under the UMD partition: train (80,512), val (4,896), and test (9,602). The language descriptions in this dataset are significantly longer on average (an average of 8.4 words vs. 3.5 words in RefCOCO) and possess more complex grammatical structures, including more clauses and modifiers. This poses a greater challenge to the model's capabilities in long-text comprehension and complex semantic parsing.

4.2. Implementation Details

Backbone configuration: The visual backbone utilizes DINOv3 ViT-B/16 (86M parameters, patch size 16×16), loading pre-trained weights via the timm library (model name vit_base_patch16_dinov3.lvd1689m) with all parameters completely frozen. The text encoder employs CLIP ViT-B/16 (63M parameters), with all parameters frozen except for the adapter layers.

Training settings: Input images are uniformly resized to a resolution of 448×448 and normalized using the standard ImageNet mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225). The batch size is set to 128 (8 images per GPU, utilizing DistributedDataParallel for distributed training). The AdamW optimizer is employed with an initial learning rate of 1×10^{-4}, and a polynomial learning rate decay strategy (power=0.9) is adopted. The model is trained for a total of 65 epochs. The weight decay rates for the backbone adapters and the decoding head are set to 0.01 and 0.05, respectively, and the gradient clipping threshold is set to 0.5. The main experiments utilize 2-GPU parallel training coupled with Automatic Mixed Precision (AMP) for acceleration.

Evaluation metrics: Two primary metrics are adopted: (1) Mean IoU (mIoU), which calculates the Intersection over Union for each sample and then takes the average, reflecting the model's average segmentation quality across different samples; (2) Overall IoU (oIoU), which calculates the ratio of the total intersection area to the total union area across all samples, emphasizing the segmentation quality of larger objects. Additionally, precision at different thresholds (Pr@50 to Pr@90) is reported as an auxiliary reference.

4.3. Comparison with Existing Methods

4.3.1. Cross-Backbone Comparison with DETRIS-B (DINOv2 Backbone)

Table 1: mIoU (%) comparison with DETRIS-B on RefCOCO, RefCOCO+, and G-Ref

					RefCOCO+					
Method	Backbone	RefCOCO val	testA	testB	val	testA	testB	G-Ref val	test	Avg
DETRIS-B[9]	DINOv2 ViT-B/14	76.0	78.2	73.5	68.9	74.0	61.5	67.9	68.1	71.0
RAFS (Ours)	DINOv3 ViT-B/16	75.85	77.74	72.96	68.30	73.65	61.17	67.81	67.78	70.7

Table 1 presents the mIoU comparison between RAFS and DETRIS-B across the three datasets. Overall, RAFS achieves an average mIoU of 70.7% across all 8 evaluation splits. The gap between this and DETRIS-B's 71.0% is a mere 0.3 percentage points, achieving a highly comparable performance level.

It is particularly worth noting that there is a significant disadvantageous factor for RAFS in this comparison: the difference in spatial resolution. The DINOv3 ViT-B/16 used by RAFS employs a patch size of 16×16, resulting in 28×28=784 patch tokens under an input resolution of 448×448. In contrast, the DINOv2 ViT-B/14 used by DETRIS-B employs a patch size of 14×14, generating approximately 37×37=1369 patch tokens under the same input resolution (DETRIS uses a 526×526 input). More patch tokens equate to a higher feature spatial resolution, which is an inherent advantage for pixel-level segmentation tasks requiring precise boundary localization. Under this disadvantageous condition, RAFS still achieves comparable performance, fully demonstrating that the adaptation modules designed in this paper can effectively compensate for the gap in spatial resolution and successfully leverage the advantages of DINOv3 in feature quality.

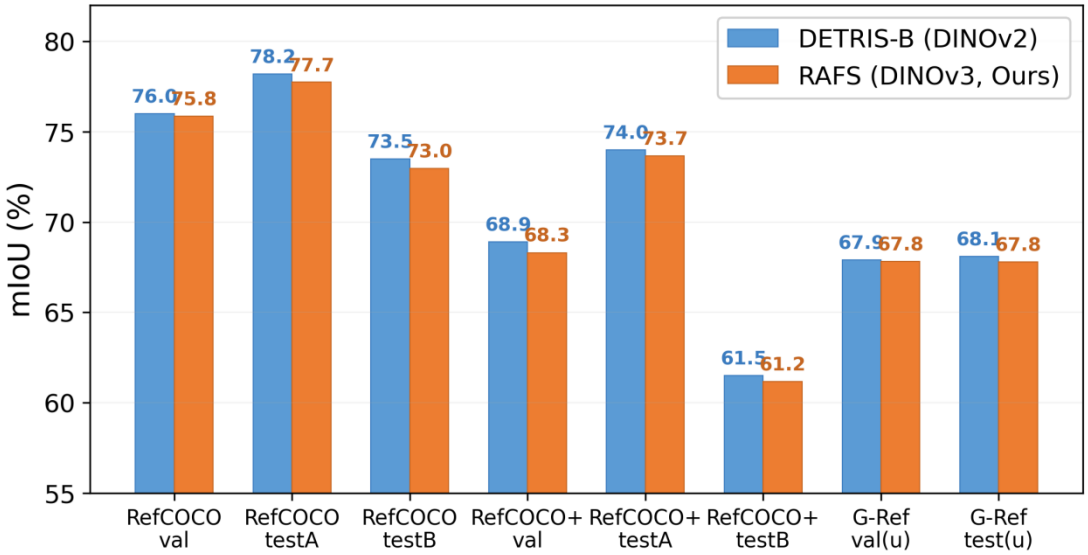

Figure 2: mIoU comparison between RAFS and DETRIS-B across all dataset splits.

The blue bars represent DETRIS-B (DINOv2), while the orange bars represent RAFS (DINOv3). The performance of both models is highly comparable across all splits, with RAFS achieving a comparable performance level despite the disadvantageous condition of lower spatial resolution.

Analyzing the performance across individual datasets, RAFS exhibits the smallest performance gap on RefCOCO (with a difference of only 0.15 on the val split) and a slightly larger gap on RefCOCO+ (a difference of 0.60 on the val split). This discrepancy may arise because RefCOCO+ prohibits the use of positional words, thereby relying more heavily on the fine-grained recognition of visual attributes—a task where higher spatial resolution is inherently more beneficial. On the G-Ref dataset, the performance gap for RAFS is also marginal (a difference of only 0.09 on the val split), indicating that for the comprehension of complex descriptions with long texts, the high-quality features of DINOv3 and the cross-modal fusion module proposed in this paper function effectively.

Table 2: Comparison with DETRIS-B on the oIoU (%) metric

				RefCOCO+			G-Ref		
Method	RefCOCO val	testA	testB	val	testA	testB	val	test	Average
DETRIS-B[9]	74.3	77.6	71.4	65.6	72.0	56.5	65.2	66.4	68.6
RAFS (ours)	74.53	77.08	70.98	64.64	71.42	56.20	66.16	65.82	68.4

Table 2 presents the comparison results for the oIoU metric. The distinction between oIoU and mIoU is that the calculation method of oIoU places greater emphasis on the segmentation quality of large target objects, as large objects account for a larger proportion of the total intersection and total union. From the results, RAFS surpasses DETRIS-B on two splits: on the RefCOCO val split, RAFS achieves 74.53%, exceeding DETRIS-B's 74.3% by 0.23 percentage points; on the G-Ref val split, RAFS achieves 66.16%, exceeding DETRIS-B's 65.2% by 0.96 percentage points. This notable advantage demonstrates that RAFS possesses a stronger capability in processing large target objects within the G-Ref dataset. This can likely be attributed to the dense features of DINOv3—optimized via Gram Anchoring—which maintain better consistency over large-area regions, as well as the global contextual information injected by the RAFF module, which helps preserve segmentation integrity within large target areas.

4.3.2. Gain Analysis of Innovative Modules on DINOv3 Backbone

To more intuitively evaluate the enhancement effects of the three innovative adaptation modules proposed in this paper on DINOv3 backbone features, Table 3 compares the full RAFS method with a baseline configuration that uses only the DINOv3 backbone network plus the shared decoding head (without adding any adaptation modules). This comparison eliminates the interference of backbone network differences and directly reflects the contribution of the innovative modules on the same backbone foundation.

Table 3: Gains of innovative modules on the DINOv3 backbone (RefCOCO, mIoU/%)

Configuration	Adapter Params	val	testA	testB	val Δ	testA Δ	testB Δ
DETRIS-B[9]	—	76.0	78.2	73.5	—	—	—
DINOv3 + Decoder (No Adapter)	0.00M	75.41	78.22	72.38	—	—	—
+ MLFS	1.77M	76.26	78.28	72.94	+0.85	+0.06	+0.56
+ SCSA	1.87M	75.32	77.90	72.20	-0.09	-0.32	-0.18
+ RAFF	2.66M	75.75	78.47	72.62	+0.34	+0.25	+0.24
+ RAFF + SCSA	4.54M	75.57	77.97	72.48	+0.16	-0.25	+0.10
+ SCSA + MLFS	3.65M	75.85	78.30	72.90	+0.44	+0.08	+0.52
+ RAFF + MLFS	4.44M	76.02	78.37	72.80	+0.61	+0.15	+0.42
+ RAFF + SCSA + MLFS(RAFS)	6.31M	75.85	77.74	72.96	+0.44	-0.48	+0.58

Three important findings can be derived from Table 3:

First, the innovative modules bring stable and significant performance improvements on the val and testB splits. Compared to the DINOv3 vanilla backbone baseline, the complete RAFS improves mIoU by 0.44 points on val (75.41%→75.85%) and by 0.58 points on testB (72.38%→72.96%). Since testB primarily contains multi-object discrimination scenarios, these improvements indicate that the innovative modules can effectively enhance the feature representation capability of DINOv3 in complex scenes requiring fine-grained semantic understanding. It is particularly noteworthy that the dual-module combination of RAFF+MLFS achieved the highest value across all configurations on val at 76.02% (+0.61), surpassing the 75.85% of the full three-module combination. This suggests a certain redundancy effect in module combinations, hinting that the potential could be further tapped through more refined module interaction designs in the future.

Second, the performance changes on testA reveal that DINOv3 is approaching saturation in large-target scenarios. The testA split mainly comprises multi-person scenes where large targets (individuals) dominate. The DINOv3 vanilla backbone already achieved a high level of 78.22% on testA; however, after adding the adaptation modules (77.74%), there was a slight decrease (-0.48). This phenomenon suggests that the quality of DINOv3's dense features, optimized via Gram Anchoring, is already exceptionally high for large-target scenarios. Consequently, additional feature transformation modules offer limited marginal gains in such contexts and may even introduce slight informational interference. Nevertheless, the dual-module combinations of SCSA+MLFS (78.30%) and RAFF+MLFS (78.37%) both outperformed the baseline on testA, indicating that appropriate module combinations can still yield benefits in large-target scenarios.

Third, the differentiated contribution patterns of various module combinations demonstrate the functional complementarity of the three innovative modules. SCSA+MLFS achieved the largest improvement on testB (+0.52), reflecting the synergistic effect of cross-modal fusion and multi-layer feature selection for multi-object discrimination. RAFF+MLFS showed the highest gain on val (+0.61), suggesting that the combination of global context injection and multi-layer feature selection is most effective for general scenarios. These differentiated contribution patterns validate the design strategy of adapting across three orthogonal dimensions: feature selection, cross-modal fusion, and global information injection.

4.4. Ablation Study

To comprehensively verify the effectiveness and necessity of each innovative module, systematic ablation experiments were designed on the RefCOCO dataset. These include two sets of experiments: the sequential removal of each module (the "w/o" series) and the retention of only a single module (the "Only" series).

Table 4: Ablation study results of various modules (RefCOCO, mIoU/%)

Configuration	Adapter Params	val	testA	testB
Full (RAFF+SCSA+MLFS)	6.31M	75.85	77.74	72.96
w/o RAFF	3.65M	75.85	78.30	72.90
w/o SCSA	4.44M	76.02	78.37	72.80
w/o MLFS	4.54M	75.57	77.97	72.48
No Adapter (DINOv3)	0.00M	75.41	78.22	72.38
Only RAFF	2.66M	75.75	78.47	72.62
Only SCSA	1.87M	75.32	77.90	72.20
Only MLFS	1.77M	76.26	78.28	72.94

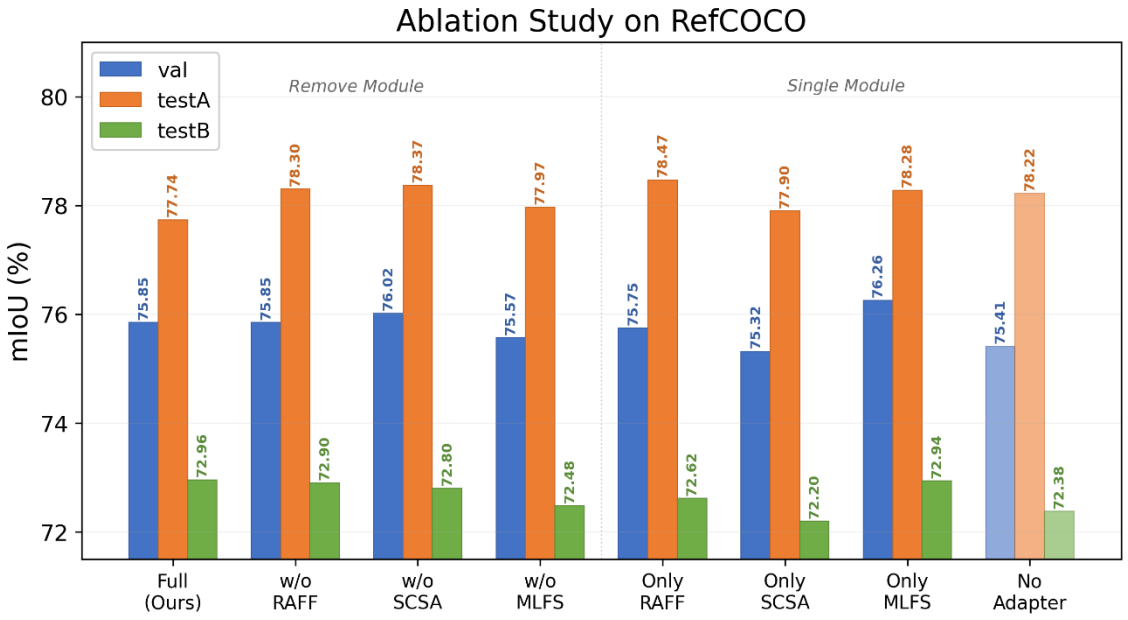

Figure 3: Comparison of ablation study results (RefCOCO dataset).

Blue, orange, and green bars correspond to val, testA, and testB, respectively. The leftmost bars with diagonal hatch

patterns represent the baseline from the DETRIS-B paper, while the rightmost bars represent the DINOv3 No Adapter baseline. Various ablation configurations from the "w/o" and "Only" series are positioned in the middle. "Only MLFS" achieves the highest performance of 76.26% on val, while the "Full" configuration reaches the peak performance of 72.96% on testB.

From the ablation results in Table 4 and Figure 3, the following in-depth analysis can be derived across six dimensions:

(1) The contribution of the MLFS module is the most significant and consistent. Upon removing MLFS, the mIoU dropped from 75.85% to 75.57% (-0.28) on the val split, and from 72.96% to 72.48% (-0.48) on testB, showing a distinct performance decline across both splits. Conversely, performance on testA rose from 77.74% to 77.97% (+0.23), suggesting that MLFS has limited effectiveness in scenarios dominated by large targets. Notably, MLFS provided the largest gain of 0.48 percentage points on testB. Since testB primarily comprises multi-object scenes requiring accurate differentiation between categories, the adaptive multi-layer feature selection of MLFS provides richer multi-scale information for these complex scenarios compared to fixed-layer selection. This result validates the advantage of the learnable Gaussian weighting mechanism in MLFS over hard-coded layer selection: the model can adaptively adjust its focus on features from different layers based on scene complexity.

(2) The SCSA module provides a stable positive contribution on testB. Removing SCSA caused the mIoU on testB to drop from 72.96% to 72.80% (-0.16), indicating that cross-modal attention fusion is essential for multi-object scenes. When multiple candidate targets are present, key attribute words in the language description are crucial for correct localization; the cross-modal attention mechanism in SCSA is key to achieving this language-guided localization. Interestingly, performance slightly improved on val and testA after removing SCSA, potentially because additional cross-modal fusion layers may introduce slight overfitting in relatively simple scenarios.

(3) The RAFF module serves a complementary role for testB. Removing RAFF led to a slight decrease on testB from 72.96% to 72.90% (-0.06), demonstrating that the global contextual information in register tokens provides auxiliary support for multi-object discrimination. The global semantic information injected by RAFF helps the model better understand the overall scene layout, thereby enabling more accurate target localization in multi-object scenarios.

(4) The No-Adapter baseline reveals the powerful feature foundation of DINOv3. Even without any innovative adaptation modules, relying solely on DINOv3's original patch features and the shared decoding head achieved an mIoU of 75.41% on RefCOCO val—only 0.44 percentage points lower than the 75.85% achieved with the full suite of modules. This result is significant on two levels: first, it verifies the exceptional feature quality of DINOv3 as a vision foundation model, whose dense features possess strong representation capabilities even without task-specific adaptation; second, it confirms the rationality of the proposed adaptation modules, as they consistently contribute a 0.44-point improvement on such a robust feature base, effectively enhancing the original features through multi-layer selection, cross-modal fusion, and global information integration.

(5) Single-module experiments reveal the independent gain capabilities of each module. The "Only" series experiments provide a more direct assessment of module value. Only MLFS, using just 1.77M adaptation parameters, reached 76.26% on val (+0.85 vs. No Adapter), surpassing not only the no-adapter baseline but even the "Full" configuration (75.85%) which uses 6.31M parameters. This finding is significant: it indicates that the multi-layer adaptive feature selection mechanism of MLFS is itself a highly efficient feature enhancement strategy, yielding the largest performance gains at a minimal parameter cost. Only RAFF (2.66M) also demonstrated a stable positive contribution, outperforming the no-adapter baseline across all three splits (val +0.34, testA +0.25, testB +0.24) and achieving the highest value among all single-module configurations on testA (78.47%), indicating that global information from register tokens is particularly effective for large target localization. Only SCSA (1.87M) performed relatively weaker, falling slightly below the no-adapter baseline across all splits, suggesting that the cross-modal attention fusion module has limited impact when used in isolation and that its value is primarily realized through synergy with other modules.

(6) Analysis of module combination effects. A comprehensive comparison of the "Only" and "w/o" series results reveals an important combination effect: while Only MLFS outperformed the "Full" configuration on val (76.26% vs. 75.85%), it was closely matched by the "Full" configuration on testB (72.94% vs. 72.96%). This suggests a degree of competitive effect between modules when combined, particularly on the val split. However, the "Full" configuration still achieved the optimal 72.96% on testB, proving that the synergy of the three modules is irreplaceable in complex scenarios requiring fine-grained multi-object discrimination. This observation offers insights for future designs: exploring adaptive gating mechanisms between modules could dynamically adjust the weight of each module based on input characteristics, fully leveraging individual strengths while minimizing redundant interference.

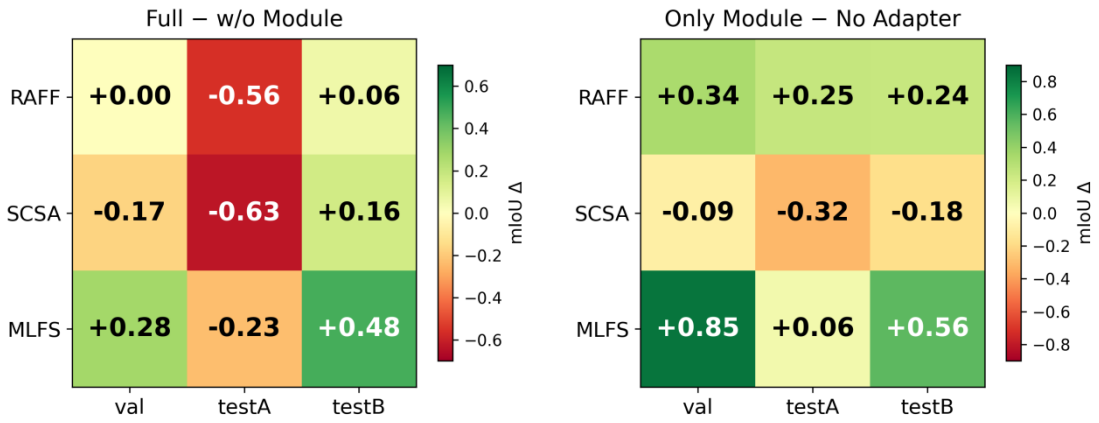

Figure 4: Heatmap of contributions from various modules.

The left panel shows the mIoU difference (Full configuration minus the corresponding "w/o" configuration), reflecting the marginal contribution of each module within the complete system; the right panel shows the mIoU difference (Only single-module configuration minus the No-Adapter baseline), reflecting the independent gain capability of each module. Green indicates a positive contribution, while red indicates a negative effect.

The dual-panel heatmap in Figure 4 reveals the contribution characteristics of each module from two complementary perspectives. In the left panel (w/o analysis), MLFS shows a clear positive contribution on val (+0.28) and testB (+0.48), but its contribution is negative on testA (-0.23); SCSA also exhibits a negative contribution on testA (-0.63). In the right panel (Only analysis), MLFS demonstrates the strongest independent gain capability (val +0.85, testB +0.56), while RAFF provides a stable positive gain on testA (+0.25), and SCSA's independent contribution is relatively weak. The comparison between the two panels reveals an important phenomenon: a module's independent gain (right panel) is not always consistent with its marginal contribution within the complete system (left panel). This indicates the existence of complex interactive effects between modules, suggesting that scene-adaptive module activation strategies could be designed in the future to optimize the combined effect.

4.5. Parameter Efficiency Analysis

Table 4: Parameter statistics for various modules

Component	Parameter Count	Percentage of Total Parameters
RAFF (3 layers)	2.66M	1.49%
SCSA (3 layers)	1.87M	1.05%
MLFS	1.77M	0.99%
Total Adaptation Modules	6.31M	3.53%
Shared Decoding Head (Neck+Decoder+Projector)	22.80M	12.77%
Total Trainable Parameters	29.53M	16.53%
Total Frozen Parameters (DINOv3+CLIP)	149.07M	83.47%

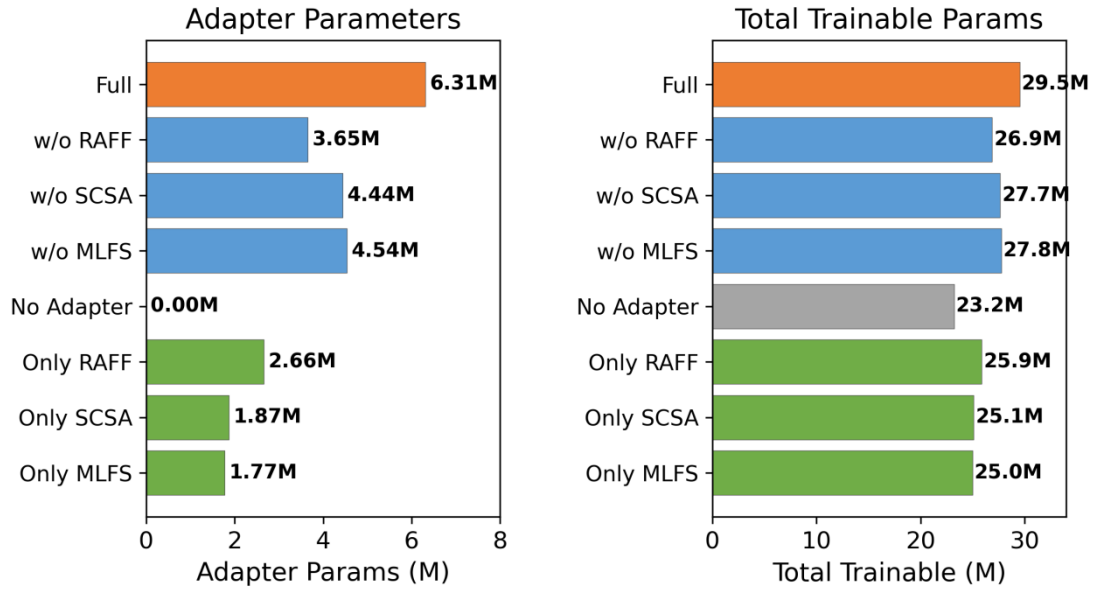

Figure 5: Comparison of parameter counts across configurations.

The left panel shows the parameter count for the innovative adaptation modules, and the right panel shows the total trainable parameters. The innovative modules in the Full configuration account for only 6.31M, representing 3.53% of the total model parameters. Different ablation configurations demonstrate a flexible parameter-performance trade-off by removing or retaining various modules.

Table 4 and Figure 5 detail the distribution of parameter counts for each component of RAFS. The three innovative adaptation modules introduce a combined total of only 6.31M parameters, accounting for 3.53% of the total model parameter count (178.60M), which reflects excellent parameter efficiency. Among the 6.31M parameters for the adaptation modules, RAFF (2.66M) takes the largest share, followed by SCSA (1.87M) and MLFS (1.77M), with the proportions of the three being relatively balanced. This balanced parameter distribution reflects the functional complementarity of the three modules—each of the three aspects (feature selection, cross-modal fusion, and global information integration) requires a certain parameter capacity to achieve effective feature transformation.

As observed in Figure 5, the differences in parameter counts across various ablation configurations are primarily reflected in the innovative module section, while the shared decoding head (22.80M) remains constant across all configurations. This implies that the design of the innovative modules in this paper effectively enhances backbone features at a minimal parameter cost without altering the core decoding process.

4.6. Qualitative Analysis

Figures 6 and 7 show the visualization of segmentation results by RAFS on the RefCOCO validation set. Each row consists of 4 columns: the original image with the text query, the ground-truth annotation mask (green overlay), the model's predicted mask (red overlay, labeled with the IoU value), and an overlay comparison between the ground truth and prediction (green = ground truth, red = prediction).

Figure 6: Visualization of segmentation results by RAFS on the RefCOCO validation set (Group 1).

The first row: "guy right of main guy" (IoU=75.4%), the model correctly understands the relative positional relationship of "right"; the second row: "pizza being cut into" (IoU=62.3%), the model identifies the pizza being operated on; the third row: "batter" (IoU=91.6%), the model accurately segments the batter; the fourth row: "top left dish" (IoU=73.0%), the model localizes the dish at the specified position.

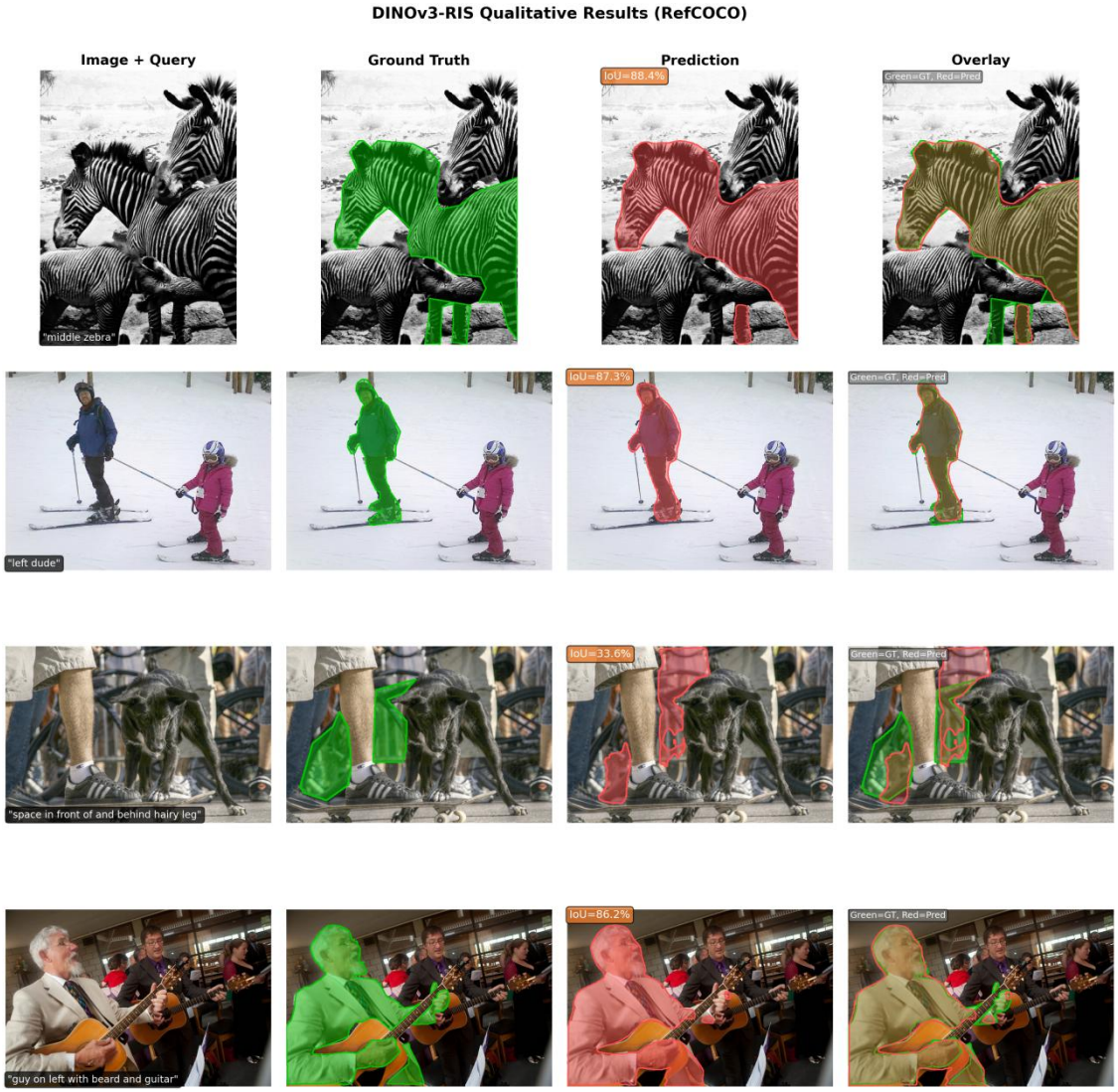

Figure 7: Visualization of segmentation results by RAFS on the RefCOCO validation set (Group 2).

The first row: "middle zebra" (IoU=88.4%), the model accurately identifies the middle zebra among multiple zebras; the second row: "left dude" (IoU=87.3%), the model correctly distinguishes between two individuals in a skiing scene; the third row: "space in front of and behind hairy leg" (IoU=33.6%), this sample demonstrates the model's limitations when facing highly ambiguous and abstract descriptions; the fourth row: "guy on left with beard and guitar" (IoU=86.2%), the model accurately understands a composite description containing multiple attribute constraints.

Several analytical points can be summarized from the visualization results. First, RAFS demonstrates strong comprehension capabilities across various types of referring expressions: for relative positional descriptions (e.g., "right of", "left", "middle"), the model can correctly reason the spatial relationships between the target and reference objects; for attribute descriptions (e.g., "top left", "with beard and guitar"), the model accurately identifies targets satisfying multiple attribute constraints; and for action descriptions (e.g., "batter", "being cut into"), the model localizes the agent or the object of the action based on action semantics. Second, there is a high degree of overlap between the predicted masks and the ground-truth annotations in most test samples, with IoU values generally exceeding 75% and precise boundary localization. However, the results also reveal the current method's limitations in processing highly ambiguous and abstract descriptions (e.g., "space in front of and behind hairy leg"). The target areas indicated by such descriptions are inherently vague, posing extreme challenges to the model's language understanding capabilities.

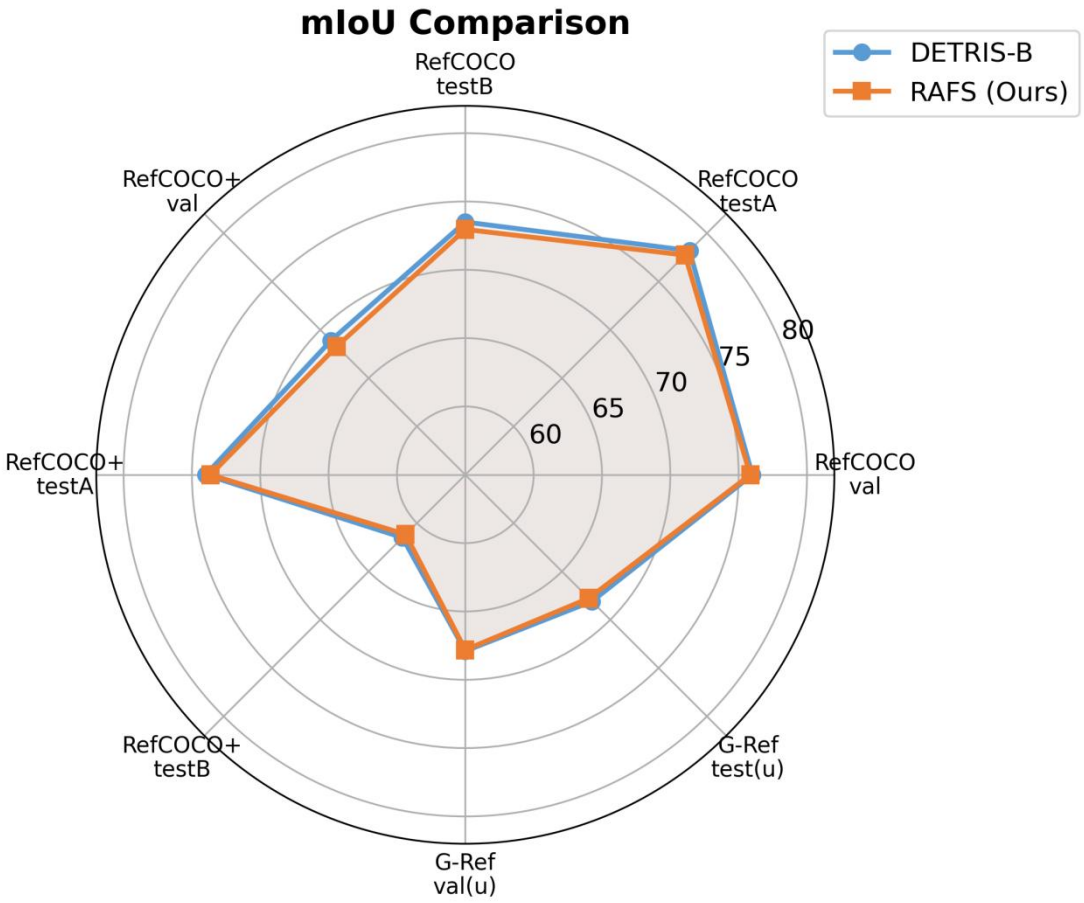

Figure 8: Radar chart comparison of mIoU between RAFS and DETRIS-B across all dataset splits.

The performance profiles of the two methods are highly similar. RAFS is most closely aligned with DETRIS-B on the G-Ref dataset, while a relatively larger gap is observed on RefCOCO+ testB. Overall, the performance differences between the two are distributed uniformly across all splits, with no significant weaknesses identified.

The radar chart in Figure 8 provides a global comparative perspective of the two methods across all dataset splits. The high degree of overlap in their performance profiles further validates that RAFS maintains stable performance comparable to DETRIS-B across different types of linguistic descriptions and varying levels of scene complexity.

5. Conclusion and Future Work

This paper presents RAFS, a parameter-efficient Image Referring Segmentation framework based on the DINOv3 vision foundation model. By designing three lightweight adaptation modules—MLFS, SCSA, and RAFF—efficient adaptation of DINOv3 features is achieved while keeping the backbone network completely frozen. MLFS adaptively aggregates multi-scale representations from all 12 feature layers through a learnable Gaussian weighting mechanism, overcoming the limitations of fixed-layer selection. SCSA implements efficient vision-language fusion in a low-dimensional space using depthwise separable convolutions and cross-modal attention. RAFF explicitly utilizes the global contextual information of register tokens for the first time in RIS tasks to enhance local features.

The three modules collectively introduce only 6.31M trainable parameters. Experimental results on three standard benchmark datasets—RefCOCO, RefCOCO+, and G-Ref—demonstrate segmentation performance comparable to the state-of-the-art method DETRIS-B (average mIoU of 70.7% vs. 71.0%) and surpass DETRIS-B by 0.96 percentage points on the G-Ref val split for the oIoU metric. Systematic ablation experiments verify the effectiveness of each module: MLFS provides the most significant independent gain (reaching 76.26% on val with only 1.77M parameters, surpassing the full model's 75.85%), and RAFF offers stable positive contributions on testA. The three modules exhibit differentiated functional patterns across different scene types.

Despite these achievements, there are limitations and areas for improvement. First, the patch size (16×16) of the DINOv3 ViT-B/16 used in the current method is larger than the 14×14 patch size used by DETRIS-B's DINOv2, resulting in lower spatial resolution of features. This is likely a primary reason for the minor performance gap; future work could explore feature upsampling or sub-patch techniques to mitigate this constraint. Second, ablation experiments suggest that certain modules contribute negatively in specific scenarios, implying that a fixed combination of modules is not globally optimal. Future efforts could involve designing scene-adaptive module activation mechanisms to dynamically decide which modules to enable based on input complexity. Finally, this work only validated the ViT-B scale; subsequent work could extend this to DINOv3 ViT-L (300M parameters) or even larger models to further enhance segmentation precision using a stronger feature foundation.

References

[1] Hu R, Rohrbach M, Darrell T. Segmentation from natural language expressions[C]// ECCV, 2016: 108-124.

[2] Yu L, Poirson P, Yang S, et al. Modeling context in referring expressions[C]// ECCV, 2016: 69-85.

[3] Yang Z, Wang J, Tang Y, et al. LAVT: Language-Aware Vision Transformer for referring image segmentation[C]// CVPR, 2022: 18155-18165.

[4] Kamath A, Singh M, LeCun Y, et al. MDETR: Modulated detection for end-to-end multi-modal understanding[C]// ICCV, 2021: 1780-1790.

[5] Wang Z, Lu Y, Li Q, et al. CRIS: CLIP-driven referring image segmentation[C]// CVPR, 2022: 11686-11695.

[6] [6] Caron M, Touvron H, Misra I, et al. Emerging properties in self-supervised vision transformers[C]// ICCV, 2021: 9650-9660.

[7] Oquab M, Darcet T, Moutakanni T, et al. DINOv2: Learning robust visual features without supervision[J]. TMLR, 2024.

[8] Siméoni O, Vo H V, Seitzer M, et al. DINOv3[R]. arXiv:2508.10104, 2025.

[9] Lin F, Yu S, Han G, et al. DETRIS: Dense Aligner with Text-Rich Features for Referring Image Segmentation[C]// AAAI, 2025.

[10] Hu R, Xu H, Rohrbach M, et al. Natural language object retrieval[C]// CVPR, 2016: 4555-4564.

[11] Mao J, Huang J, Toshev A, et al. Generation and comprehension of unambiguous object descriptions[C]// CVPR, 2016: 11-20.

[12] Li R, Li K, Kuo Y C, et al. Referring image segmentation via recurrent refinement networks[C]// CVPR, 2018: 5745-5753.

[13] Ding H, Liu C, Wang S, et al. VLT: Vision-Language Transformer for referring segmentation[J]. TPAMI, 2023, 45(6): 7900-7916.

[14] Kim D, Kim D, Cho S, et al. ReSTR: Convolution-free referring image segmentation using Transformers[C]// CVPR, 2022: 18145-18154.

[15] Xu M, Wang Y, Liu L, et al. ETRIS: Efficient referring image segmentation[J]. arXiv:2310.12006, 2023.

[16] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]// CVPR, 2016: 770-778.

[17] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]// ICML, 2021: 8748-8763.

[18] Han X, Zhang Z, Ding N, et al. Pre-trained models: Past, present and future[J]. AI Open, 2021, 2: 225-250.

[19] Houlsby N, Giber A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP[C]// ICML, 2019: 2790-2799.

[20] Hu E J, Shen Y, Wallis P, et al. LoRA: Low-Rank Adaptation of large language models[C]// ICLR, 2022.

[21] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning[C]// EMNLP, 2021: 3045-3059.

[22] Jia M, Tang L, Chen B C, et al. Visual prompt tuning[C]// ECCV, 2022: 709-727.

[23] Chen S, Ge C, Tong Z, et al. AdaptFormer: Adapting vision Transformers for scalable visual recognition[C]// NeurIPS, 2022.

[24] Darcet T, Oquab M, Mairal J, et al. Vision Transformers need registers[C]// ICLR, 2024.

[25] Yu L, Lin Z, Shen X, et al. MattNet: Modular attention network for referring expression comprehension[C]// CVPR, 2018: 1307-1315.

[26] Mao J, Huang J, Toshev A, et al. Generation and comprehension of unambiguous object descriptions[C]// CVPR, 2016: 11-20.