TALENT: Target-aware Efficient Tuning for Referring Image Segmentation

Shuo Jin[1,2] Siyue Yu[1*] Bingfeng Zhang[3] Chao Yao[4] Meiqin Liu[5] Jimin Xiao[1]
[1]XJTLU [2]University of Liverpool [3]China University of Petroleum (East China)
[4]University of Science and Technology Beijing [5]Beijing Jiaotong University
shuo.jin@liverpool.ac.uk, {siyue.yu02, jimin.xiao}@xjtlu.edu.cn,
bingfeng.zhang@upc.edu.cn, yaochao@ustb.edu.cn, mqliu@bjtu.edu.cn

Abstract

*Referring image segmentation aims to segment specific targets based on a natural text expression. Recently, parameter-efficient tuning (PET) has emerged as a promising paradigm. However, existing PET-based methods often suffer from the fact that visual features can't emphasize the text-referred target instance but activate co-category yet unrelated objects. We analyze and quantify this problem, terming it the 'non-target activation' (NTA) issue. To address this, we propose a novel framework, TALENT, which utilizes target-aware efficient tuning for PET-based RIS. Specifically, we first propose a Rectified Cost Aggregator (RCA) to efficiently aggregate text-referred features. Then, to calibrate 'NTA' into accurate target activation, we adopt a Target-aware Learning Mechanism (TLM), including contextual pairwise consistency learning and target centric contrastive learning. The former uses the sentence-level text feature to achieve a holistic understanding of the referent and constructs a text-referred affinity map to optimize the semantic association of visual features. The latter further enhances target localization to discover the distinct instance while suppressing associations with other unrelated ones. The two objectives work in concert and address 'NTA' effectively. Extensive evaluations show that TALENT outperforms existing methods across various metrics (*e.g.*, 2.5% mIoU gains on G-Ref val set). Our codes will be released at:* https://github.com/Kimsure/TALENT.

1. Introduction

Referring image segmentation (RIS) aims to segment a specific object within an image guided by a natural language expression, *i.e.*, RIS requires a precise visual-text alignment to build a 'one-to-one' correspondence between the textual expression and the visual regions. This demands fine-grained alignment across diverse object categories, attributes, and spatial relations, making RIS one of the most challenging tasks in vision-language understanding.

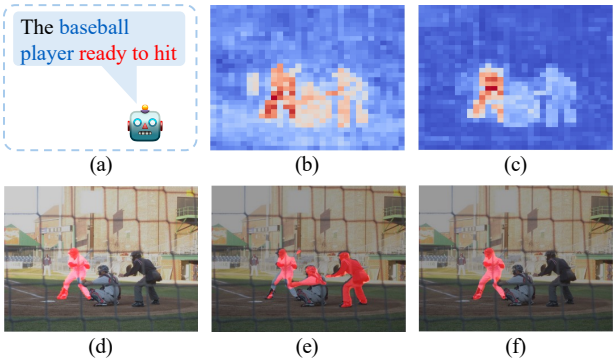


Figure 1. Visual feature activation and segmentation maps. (a) Text descriptions. (b) Visual-text fusion in DETRIS [15], which activates co-category foreground objects. (c) Our TALENT, which emphasizes the text-referred target instance. (d) Segmentation result of the ground truth (GT). (e)-(f) Corresponding segmentation results of DETRIS [15] and our TALENT.

Early studies [13, 16, 38, 54, 58, 59] have demonstrated the effectiveness of parameter-full tuning (PFT), which tunes full parameters in powerful models [9, 32]. Yet, this paradigm introduces a significant training computational overhead as the model size scales up [7, 10, 37, 47, 48, 56, 57]. To reduce the training burden, parameter-efficient tuning (PET) methods are proposed, where image-text adapters [8, 52, 53, 62] are used to jointly finetune the vision and text features extracted from the frozen pretrained backbones for final text-referred segmentation masks.

However, current PET-based frameworks tend to focus on salient, semantically equivalent objects, thereby neglecting the specific target instances designated by the text expressions. As illustrated in Fig. 1, given the textual query "the baseball player ready to hit", the existing method fails to accurately identify the intended target instance ("ready to hit"). Instead, it is misled by the visual prominence

*Corresponding author: siyue.yu02@xjtlu.edu.cn

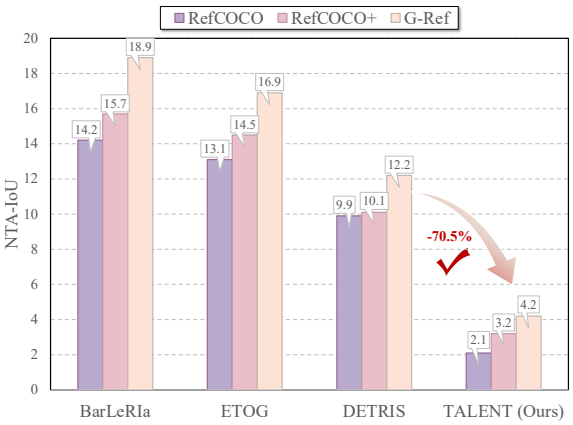


Figure 2. Comparison of 'NTA' influence quantization among various methods on the selected subset of different benchmarks.

of homogeneous objects of the same category ("baseball player"), as shown in Fig. 1 (b). We term this issue 'non-target activation (NTA)', which ultimately leads to an incorrect segmentation result, as depicted in Fig. 1(e).

Further, to quantify the impact of 'NTA', we calculate the proportion of incorrectly predicted regions that fall into other actual non-target instances within the same category. This proportion is named as NTA-IoU to evaluate how strongly the 'NTA' issue affects PET-based methods [15, 39, 51, 60, 61]. Note that a higher NTA-IoU score indicates a stronger influence of 'NTA', whereas a lower score suggests a weaker 'NTA' effect and, therefore, a better segmentation outcome. Fig. 2 reports the NTA-IoU of different methods. Obviously, existing PET-based approaches exhibit higher NTA-IoU scores, where the model tends to segment salient co-category objects rather than the specific text-described target instance.

As described above, our goal is to mitigate the 'NTA' issue to achieve precise target activation and yield more accurate target segmentation. Accordingly, we propose a **T**arget-**A**ware **L**earning framework to suppress 'NTA' for **E**fficie**N**t **T**uning RIS, termed **TALENT**. First, we design a Rectified Cost Aggregator (RCA), which constructs a cost volume to represent the visual-text relationship for text-referred visual feature aggregation. Then, we introduce the Target-aware Learning Mechanism (TLM) to optimize the aggregated features from RCA and thus calibrate the 'NTA' issue, which includes Contextual Pairwise Consistency Learning (CPCL) and Target Centric Contrastive Learning (TCCL).

In detail, CPCL provides a holistic understanding of the referring text by leveraging global text features, along with aggregated RCA features, to construct a text-augmented pairwise feature affinity map. This map guides the model in refining the original RCA feature relationships and learning context-aware semantic associations. However, although the global text feature provides a holistic understanding of one object, it exhibits coarse granularity, causing CPCL to overlook certain fine-grained semantic details. To further enhance the discrimination of the distinct target instance, we introduce TCCL, which strengthens the alignment between visual features and the corresponding target text expression while simultaneously reducing their association with other non-target textual descriptions. These two learning objectives work in concert to enable our target-aware learning, allowing our model to effectively improve feature discriminability and thereby alleviate the 'NTA' issue. As depicted in Fig. 1 (c), TALENT helps the fused feature emphasize the text-referred object and segments the precise instance shown in Fig. 1 (f), which closely aligns with the GT segmentation result in Fig. 1 (d). In addition, TALENT achieves up to triple gains of the NTA-IoU score compared to other PET-based methods, as shown in Fig. 2.

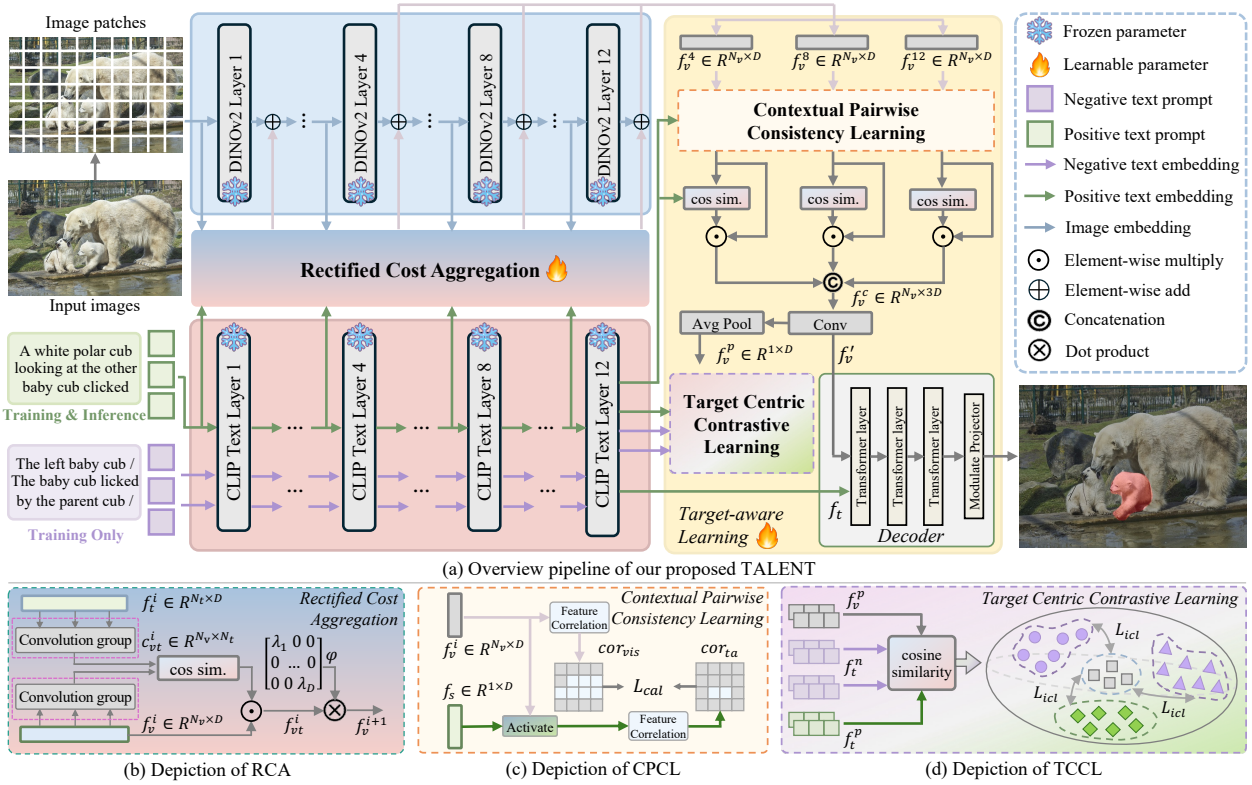
Comprehensive experiments demonstrate that our TALENT outperforms existing approaches, including both PFT-based and PET-based methods, across various benchmarks. In summary, our main contributions are as follows:

- We identify and quantify the 'NTA' issue in PET-based RIS, where visual features attend to salient objects instead of the specific target instance. To address this, we propose TALENT for PET-based RIS through target-aware learning mechanism, where a Rectified Cost Aggregator (RCA) is first introduced for visual-text aggregation.
- We design Target-aware Learning Mechanism (TLM), including Contextual Pairwise Consistency Learning (CPCL) and Target Centric Contrastive Learning (TCCL), to suppress 'NTA': CPCL constrains the RCA feature to learn context-aware semantic relations and TCCL forces it to identify the distinct target instance.
- We conduct comprehensive experiments to demonstrate that TALENT achieves state-of-the-art (SOTA) performance across various benchmarks. (*e.g.*, TALENT achieves a gain of 1.9% mIoU and 3.3% oIoU on the RefCOCO+ testB set over PET-based methods).

2. Related Work

Our work introduces a parameter-efficient tuning approach for RIS. In this section, we summarize two mainstream paradigms and discuss their differences.

Parameter-full Tuning for RIS. RIS is originally introduced by [12], which aims to predict a pixel-level object mask described by a text expression. Early methods [14, 26, 30, 34, 45, 46] typically encode the referring expression into a fixed-length feature vector that is then concatenated with the visual feature. It enabled multimodal interaction through convolutional frameworks like FCN [33]. Subsequent methods [6, 31, 59] have advanced this paradigm by incorporating shallow visual and text features to perform cross-modal fusion at multiple feature levels, which can preserve coarse-to-fine spatial details.



(a) Overview pipeline of our proposed TALENT

(b) Depiction of RCA (c) Depiction of CPCL (d) Depiction of TCCL

Figure 3. Framework pipeline of our TALENT. It contains four main modules: a frozen backbone building upon DINOv2-Reg and CLIP to encode the image and text, a rectified cost aggregator for vision-language interaction, a target-aware learning mechanism to strengthen feature representation, and a transformer decoder for final segmentation.

Some methods [2, 5, 19, 29] also explore the potential of multi-task learning to obtain fine-grained results, where the bounding box of the referring expression can be treated as pseudo labels to enhance the referring segmentation. In addition, multimodal pretraining introduces the advantage of large-scale data. CRIS [54] emphasizes sentence-pixel alignment to leverage multimodal information, while PCAN [1] focuses on position-aware contrastive alignment. However, these methods require PFT, incuring extra computation and limited scalability due to the training burden.

Parameter-efficient Tuning for RIS. PET adjusts only a fraction of the parameters and alleviates the computational burden compared to PFT-based methods. Prominent approaches [8, 18] are designed to integrate lightweight modules into the frozen backbone [20, 40, 41], where only the extra components are updated during tuning. It enables a residual tuning pipeline. Another paradigm for PET is to update only a subset of the original model parameters. Some researchers have explored matrix decomposition methods [11] to reduce learnable parameters by factorizing the weights of pretrained models. Recently, PET-based approaches for RIS have been promisingly improved. ETRIS [55] introduces PET to referring image segmentation by leveraging the bridge module for visual-text fusion. BarLeRIa [51] builds an intertwined adapter along with a normalizing flow mechanism with extremely few trainable parameters. DETRIS [15] employs multiscale convolutional adapters on both vision and text backbones to achieve a significant performance gain. Nevertheless, these PET-based methods don't thoroughly examine whether visual features can effectively respond to text-referred visual regions, *e.g.*, whether the 'NTA' issue can be solved. In contrast, we conduct an in-depth analysis of the 'NTA' issue and propose an effective PET-based solution for RIS.

3. method

3.1. Problem Quantization

As analyzed in Sec. 1, the 'NTA' issue arises from the fact that existing approaches tend to emphasize salient objects within the input image, instead of the text-described specific target instance.

To quantify this issue, we introduce the NTA-IoU metric to evaluate whether the model can accurately distinguish the correct instance from salient objects of the co-category. Specifically, we select a subset of widely used benchmarks:

RefCOCO, RefCOCO+, and G-Ref, which satisfy the following criteria: each image contains multiple text expressions referring to different co-category instances. We define the union of all these GT masks $\{G_1, \ldots, G_n\}$ as the co-category GT G. For a given segmentation map I_1 corresponding to the text-referred object, we exclude the overlap with its matched GT mask G_1, and then compute the IoU between the remaining predicted region and the co-category GT mask. Our NTA-IoU metric can be formulated as:

$$\text{NTA-IoU} = \frac{(I_1 - I_1 \cap G_1) \cap (G - G_1)}{(I_1 - I_1 \cap G_1) \cup (G - G_1)}. \quad (1)$$

Intuitively, NTA-IoU measures the proportion of the incorrectly segmented regions that fall into non-target instances within the same category. A higher NTA-IoU indicates that the model struggles to distinguish between the intended target instance and its non-target objects of the same category, and vice versa.

3.2. Overview

Fig. 3 (a) depicts the overview of our framework, including several modules: a frozen backbone building on DINOv2-Reg [37] and CLIP [42], a Rectified Cost Aggregator (RCA), a Target-aware Learning Mechanism (TLM), including Contextual Pairwise Consistency Learning (CPCL) and Target Centric Contrastive Learning (TCCL), and a transformer decoder. The complete training process is:

1. First, the image is input to the frozen DINOv2-Reg encoder for visual features. Both positive and negative texts are input to the frozen CLIP text encoder for text features. Note that the negative texts are only for TCCL.
2. Instead of directly propagating visual features through DINOv2-Reg layers, the intermediate feature is fused with the corresponding positive text feature through RCA for efficient tuning and then input to the next layer.
3. Then, the visual features from RCA across different layers participate in our CPCL, where the positive sentence-level text feature f_s is utilized to construct a text-augmented pairwise feature affinity map to strengthen context-aware semantic association of visual features.
4. After that, f_s guides each visual feature via a cosine similarity mask to strengthen the text-referred features. Then, the strengthened multi-layer features are aggregated into a single global feature for our TCCL. It further enhances feature discrimination to identify the target instance by leveraging both positive and negative texts.
5. Finally, a transformer decoder generates the final segmentation map. During inference, only the given positive text prompt is used with the image for prediction.

3.3. Rectified Cost Aggregator

As mentioned above, existing PET-based methods can't emphasize the precise target activation. Enlightened by CAT-Seg [4], we propose a Rectified Cost Aggregator (RCA) for visual-text interaction. By constructing vectorized cost-volume, we rectify the model's focus on text-referred regions by suppressing adverse cross-modal interactions.

First, or a tokenized positive text expression $T \in \mathbb{R}^L$ and an input image $I \in \mathbb{R}^{H \times W \times C}$, the frozen CLIP text encoder and DINOv2-Reg encoder are deployed to extract corresponding vectorized features with a fixed token length [42]. Then, our RCA projects them into the same channel dimension. Suppose the projected text feature $f_t^i \in \mathbb{R}^{N_t \times D}$ and the visual feature $f_v^i \in \mathbb{R}^{N_v \times D}$ are the output features of the i-th layer, where N_t and N_v denote the token length of the text and visual features, respectively. Then, RCA uses parallel convolutions to project visual and text features f_v^i and f_t^i into the $\hat{f}_v^i$ and $\hat{f}_t^i$.

Next, cost volume is utilized to localize referring regions through matching text and visual features [4], which is shown in Fig. 3 (b). To suppress irrelevant visual-text interactions in the vectorized cost space, the ReLU activation function is applied to filter out negative matching responses and rectify the positive alignment. It is formulated as,

$$c_{vt}^i = \text{ReLU}\left(\frac{\hat{f}_v^i \cdot \hat{f}_t^i}{\|\hat{f}_v^i\| \cdot \|\hat{f}_t^i\|}\right), \quad (2)$$

where $c_{vt}^i \in \mathbb{R}^{N_v \times N_t}$ denotes the matching cost volume at the i-th layer. It can be treated as a text-referred semantic mask to guide the visual feature as follows,

$$f_{vt}^i = P_{ex}^i(c_{vt}^i) \odot f_v^i, \quad (3)$$

where $P_{ex}(\cdot)^i$ denotes a linear layer to align the dimension of c_{vt}^i and f_v^i , $\odot$ denotes the element-wise multiplication.

To minimize disruption to the feature extraction capability of the frozen vision backbone, we further apply a learnable diagonal matrix [50] to rescale the fused feature as a residual, which is then injected into the next layer:

$$f_v^{i+1} = \nu_{i+1}(f_v^i) + f_{vt}^i * \varphi, \quad (4)$$

where $\nu_i(\cdot)$ denotes the i-th layer of the vision backbone, $\varphi = \text{diag}(\lambda_1, ..., \lambda_D)$ denotes the rescaling diagonal matrix. In this way, the visual feature is guided by the text expression and propagated during the image encoder to generate the text-referred visual feature.

3.4. Target-aware Learning Mechanism

Although our RCA generates text-referred visual features, the 'NTA' issue still persists. It often arises with semantic ambiguity, where the model tends to activate salient co-category objects rather than the distinct text-described target instance. Thus, we propose calibrating 'NTA' through a Target-aware Learning Mechanism (TLM), with two learning objectives: 1) strengthen the context-aware semantic relations of visual features through Contextual Pairwise Con-

sistency Learning (CPCL); 2) facilitate feature discrimination to identify the target instance through Target Centric Contrastive Learning (TCCL).

Contextual Pairwise Consistency Learning. In CPCL, we aim at constructing the text-augmented semantic relationships to refine multi-layer visual features f_v^i from RCA, where $i \in [4, 8, 12]$. As shown in Fig. 3 (c). CPCL applies the holistic, sentence-level text feature $f_s \in \mathbb{R}^{1\times D}$, which is defined as the [EOS] token of the CLIP text encoder [42], to activate the visual feature and obtain the text-augmented feature $f_{ta}^i \in \mathbb{R}^{N_v\times 1}$, formulated as,

$$f_{ta}^i = f_v^i \cdot f_s^\top, \tag{5}$$

where $\top$ denotes the transpose operation. Thereby, f_{ta}^i contains a global coherence between the sentence-level text token and each visual token.

Then, to make visual features learn context-aware semantic relations that are associated with the textual description, we derive the visual and text-augmented feature correlation maps cor_{vis}^i and cor_{ta}^i as follows,

$$\begin{aligned} cor_{vis}^i &= \text{Softmax}\left(\frac{f_v^i \cdot (f_v^i)^\top}{\|f_v^i\|^2}\right), \\ cor_{ta}^i &= \text{Softmax}\left(\frac{f_{ta}^i \cdot (f_{ta}^i)^\top}{\|f_{ta}^i\|^2}\right). \end{aligned} \tag{6}$$

Here, cor_{vis}^i and cor_{ta}^i represent their pair-wise feature relationship of the i-th layer. Since f_{ta}^i is strengthened by text feature, cor_{ta}^i has a strong text-augmented feature relationship. To address 'NTA', we aim to align cor_{vis}^i with cor_{ta}^i, which helps f_v^i accurately activate the plausible visual regions. Hence, we minimize the cosine distance to optimize the feature correlation cor_{vis}^i, where the optimization area is constrained by cor_{ta}^i. In this way, the visual feature enhances the accurate target activation in text-referred regions. This consistency loss $\mathcal{L}_{cpcl}$ is formulated as:

$$\mathcal{L}_{cpcl} = \sum_{i\in[4,8,12]} \|(\boldsymbol{J} - cor_{vis}^i) \odot cor_{ta}^i\|_F^2, \tag{7}$$

where $\boldsymbol{J}$ denotes the all-ones matrix, and $\|\cdot\|_F^2$ denotes the Frobenius norm. Note that cor_{ta}^i is detached as a pseudo label. CPCL optimizes visual features f_v^i to exhibit context-aware feature relations, and enhances the visual-text semantic coherence for target region perception.

Next, f_v^i are multiplied with f_s via cosine similarity maps, which are used to reweight f_v^i for interaction. Finally, they are concatenated as $f_v^c \in \mathbb{R}^{N_v\times 3D}$ for our TCCL.

Target Centric Contrastive Learning. CPCL constrains the visual feature to highlight the correct context region by enhancing semantic associations. However, the enhanced feature tends to exhibit coarse representation due to its optimization with holistic semantic understanding, where the model primarily attends to high-level semantics rather than instance-specific details. To identify the specific instance for fine-grained target perception, we propose TCCL to enhance the feature discrimination by drawing the positive text features closer to the visual representation while repelling negative ones belonging to irrelevant instances. Fig. 3 (d) depicts the detailed processing of our TCCL.

Firstly, to construct instance-level positive and negative pairs, we use the given text that describes the target as the positive sample and collect those referring to another object in the same image as negative samples, as shown on the left of Fig. 3 (a). Then, these text prompts are input to the frozen CLIP text encoder for positive and negative text features, f_s^p and f_s^n, respectively.

Synchronously, we obtain a global prototype $f_v^p \in \mathbb{R}^{1\times D}$ from the fused multi-layer feature f_v^c to align with the sentence-level text features, formulated as,

$$f_v^p = \text{Avg}(\text{conv}(f_v^c)), \tag{8}$$

where $\text{Avg}(\cdot)$ denotes the average pooling operation and $\text{conv}(\cdot)$ denotes a convolution layer to reduce channels. Then, text features are fused with f_v^p via the cosine similarity for contrastive learning. Our $\mathcal{L}_{tccl}$ aims to optimize distances among various features, which is formulated as,

$$\mathcal{L}_{tccl} = -\log \frac{\exp(\text{sim}(f_v^p, f_s^p))}{\exp(\text{sim}(f_v^p, f_s^p)) + \sum_{k=1}^{K} \exp(\text{sim}(f_v^p, f_s^{n_k}))}, \tag{9}$$

where $\exp(\text{sim}(\cdot,\cdot))$ denotes exponential of cosine similarity, and K denotes the number of negative samples. In this way, $\mathcal{L}_{tccl}$ enhances the linkage between visual feature and its textual description for accurate target perception.

Overall, $\mathcal{L}_{cpcl}$ and $\mathcal{L}_{tccl}$ collaboratively support our target-aware learning. $\mathcal{L}_{cpcl}$ refines the pairwise semantic association of visual features, offering a plausible yet coarse target region, while $\mathcal{L}_{tccl}$ further enhances the feature discrimination for fine-grained target instance perception. This joint optimization of TLM enables our model to focus on the text-referred target and mitigate the 'NTA' issue.

3.5. Optimization Objective

Following previous works [15, 54, 55], we utilize a transformer decoder to convert the derived visual feature from our TLM into the segmentation feature f_{seg}. Then, a text-to-pixel discriminative loss [54] is deployed to encourage the initial alignment of textual embeddings with the corresponding visual pixels, defined as:

$$\begin{aligned} \mathcal{L}_{dis}^j\left(f_{seg}^j, f_t\right) &= \begin{cases} -\log\left(\sigma\left(f_{seg}^j \cdot f_t\right)\right), & j \in \mathcal{P} \\ -\log\left(1 - \sigma\left(f_{seg}^j \cdot f_t\right)\right), & j \in \mathcal{N} \end{cases} \\ \mathcal{L}_{dis}\left(f_{seg}, f_t\right) &= \frac{1}{|\mathcal{P}\cup\mathcal{N}|} \sum_{j\in\mathcal{P}\cup\mathcal{N}} \mathcal{L}_{dis}^j\left(f_{seg}^j, f_t\right), \end{aligned} \tag{10}$$

Table 1. Quantitative comparison of our method against other PFT-based and PET-based methods on different benchmarks, evaluated using the oIoU and mIoU metrics. † denotes the re-implemented results, while others are directly sourced from the reported results. u: The UMD partition. g: The Google partition. The best results are marked with **bold**.

Methods	PET	Metirc	RefCOCO			RefCOCO+			G-Ref		
			val	testA	testB	val	testA	testB	val(u)	test(u)	val(g)
LAVT† [59]	✘	oIoU	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1	60.5
DMMI [13]	✘		74.1	77.1	70.2	64.0	69.7	57.0	63.5	64.2	61.2
ReLA [28]	✘		73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0	62.7
CG-former [49]	✘		74.8	77.3	70.6	64.5	71.0	57.1	64.7	65.1	62.5
LISA-Vicuna-13B [25]	✘		71.7	74.7	68.1	59.4	64.2	52.9	65.2	66.1	-
MagNet [3]	✘		75.2	78.2	71.1	66.2	71.3	58.1	65.4	66.2	63.1
LQMFormer [43]	✘		74.2	76.8	71.0	65.9	71.8	57.6	64.7	66.0	63.0
ReMamber [58]	✘		74.5	76.7	70.9	65.0	70.8	57.5	63.9	64.0	-
CoHD [35]	✘		75.1	78.3	71.0	66.8	71.6	58.4	65.8	66.7	-
RISCLIP [23]	✔		73.6	76.5	69.8	65.5	70.6	55.5	64.1	65.1	-
ETOG [61]	✔		71.4	76.1	66.7	62.3	68.5	51.9	61.1	62.8	-
TALENT (Ours)	✔		**75.9**	**78.3**	**72.8**	**66.9**	**72.3**	**58.8**	**65.9**	**66.8**	**64.7**
ETRIS [55]	✔	mIoU	70.5	73.5	66.6	60.1	66.9	50.2	59.8	59.9	57.9
BarLeRIa [51]	✔		72.4	75.9	68.3	65.0	70.8	56.9	63.4	63.8	61.6
RISCLIP [23]	✔		75.7	78.0	72.5	69.2	73.5	60.7	67.6	68.0	-
ETOG [61]	✔		73.4	76.9	69.3	66.0	71.5	56.9	63.8	64.6	-
DETRIS [15]	✔		76.0	78.2	73.5	68.9	74.0	61.5	67.9	68.1	65.9
TALENT (Ours)	✔		**77.8**	**79.4**	**74.8**	**70.1**	**74.9**	**63.4**	**69.7**	**69.1**	**68.4**

where $\mathcal{P}$ and $\mathcal{N}$ denote the class of '1' and '0' in the ground truth, σ is the sigmoid function.

To train the overall network, the final objective loss is

$$\mathcal{L}_{total} = \mathcal{L}_{dis} + \lambda_{cpcl}\mathcal{L}_{cpcl} + \lambda_{tccl}\mathcal{L}_{tccl}, \qquad (11)$$

where λ_{cpcl} and λ_{tccl} are the scaling coefficients. Our TLM further mitigates the 'NTA' issue.

4. experiment

4.1. Datasets and Metrics

Datasets and metrics. We conduct comprehensive experiments on three benchmarks: RefCOCO [22], RefCOCO+ [22], and G-Ref [36], all collected from MS-COCO [27]. To unify the evaluation metrics for a fair comparison, we use overall Intersection-over-Union (oIoU), mean Intersection-over-Union (mIoU), and Precision@X to evaluate segmentation results respectively. The Precision@X metric measures the percentage of test images with an IoU score higher than the threshold X $\in \{0.5, 0.7, 0.9\}$.

4.2. Implementation Details

We build our model using DINOv2-Reg (ViT-B/14) as the vision encoder and CLIP (ViT-B/16) as the text encoder. Our RCA is applied at encoder layers [1,3,5,7,9,11], where the learnable parameter in the rescaling matrix is initialized to 0.2. The coefficients λ_{cpcl} and λ_{tccl} are both set as 0.1. We train our network for 50 epochs using the Adam optimizer with a learning rate of 1e-4, where the input image is resized to 448 $\times$ 448. A learning rate decay is employed at the $35th$ epoch with a decay factor of 0.1. Our model is trained on two RTX4090 GPUs with a batch size of 32.

Table 2. Quantitative comparison using the Precision@X metric on the val-test set of RefCOCO.

Methods	Pr@0.5	Pr@0.7	Pr@0.9
ETRIS [55]	83.4	72.7	17.4
ETOG [61]	83.2	73.0	26.2
CG-former [49]	87.2	78.7	38.8
Prompt-RIS [44]	85.6	76.9	26.2
DETRIS [15]	87.9	80.2	27.5
TALENT (Ours)	**88.6**	**82.3**	**40.1**

4.3. Quantitative Results

Our method is evaluated against existing SOTA methods. As shown in Tab. 1, our method achieves significant gains and sets a new SOTA performance. For example, TALENT outperforms the PET-based method DETRIS [15] by 1.8% mIoU on the RefCOCO val-test set and also surpasses the PFT-based methods ReMamber [58] and CoHD [35], up to 1.9% and 1.8% oIoU on the RefCOCO testB set, with fewer tunable parameters. TALENT even beats LISA-Vicuna-13B [25], which requires a large language model, with 3.6% and 4.7% oIoU gains on RefCOCO testA and testB sets. These results highlight that our method boosts the segmentation performance of the PET-based paradigm.

Additionally, we evaluate our method against several SOTA methods using the Precision@X metric. Tab. 2 shows that TALENT still outperforms existing methods on the im-

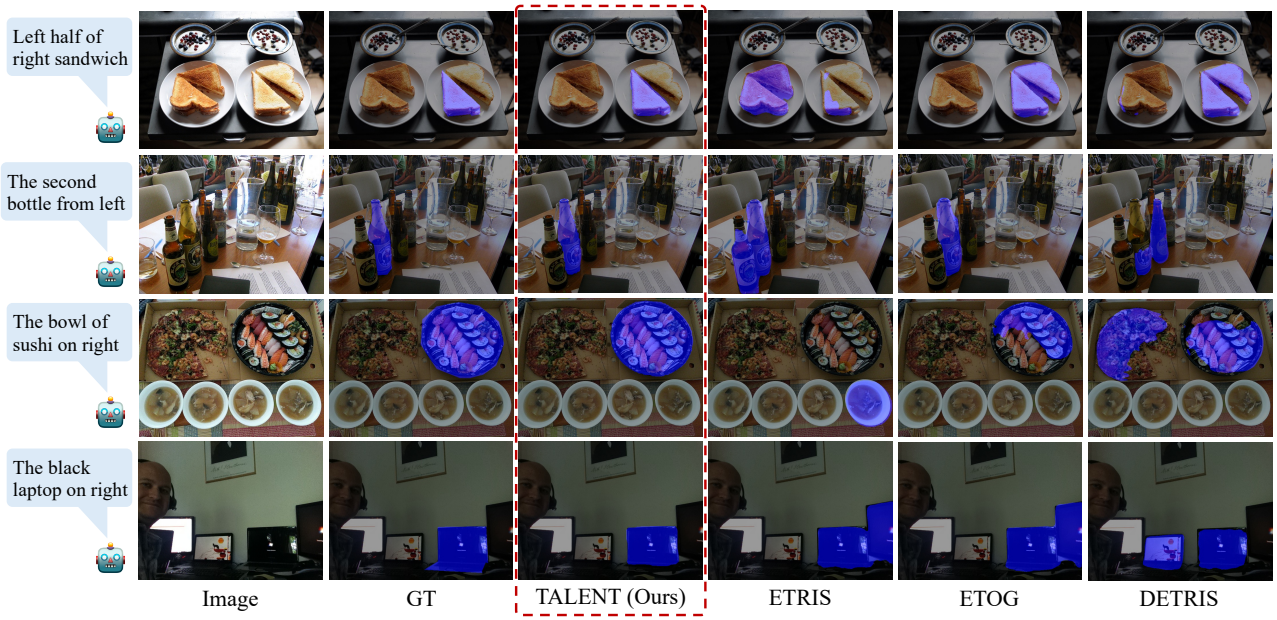


Figure 4. Qualitative comparison. We compare our TALENT with ETRIS [55], ETOG [61], and DETRIS [15]. It's observed that TALENT can accurately localize the target and generate more precise segmentation results.

age percentage with different IoU thresholds. Specifically, TALENT outperforms the PET-based method DETRIS [15] 12.6% and the PFT-based method CG-former [49] 1.3% with the IoU threshold X = 0.9. These results demonstrate that our method can precisely segment the accurate targets. It's noted that we report the result of Prompt-RIS [44] without SAM [24] post-processing for a fair comparison.

4.4. Qualitative Results

In Fig. 4, we compare segmentation results among different PET-based methods [15, 55, 61]. It's observed that previous methods struggle to localize the text-referred target and often segment other salient and visually similar objects, like 'bowl of sushi' in the third row and 'black laptop' in the last row. These results indicate that the issue of 'NTA' is not well addressed. In contrast, TALENT can precisely localize the text-referred target and predict more accurate segmentation maps, which demonstrates that TALENT effectively mitigates the 'NTA' issue. We also visualize the feature activation in Fig. 5. It's shown that TALENT activates the specific target associated with the text expression, while the previous method often activates salient yet unrelated objects. More feature activation and segmentation results are provided in the appendix.

4.5. Ablation Study

Evaluation of tunable parameter efficiency. We compare our TALENT with adapter-based and decomposition-based PET methods to assess the efficiency of model's tunable parameters. Tab. 3 reports the number of tunable parameters for the backbone and the remaining modules. It's observed that our TALENT achieves the best performance with the fewest extra parameters compared with the adapter-based methods. Furthermore, although several decomposition-based methods, including LoRA [11] and Compacter [21], require fewer parameters to finetune the backbone model, the multimodal interaction between visual and textual features still remains insufficient. In contrast, our method achieves a significant improvement of 10% mIoU over these decomposition-based methods with the fewest overall tunable parameters. These results demonstrate both the effec-

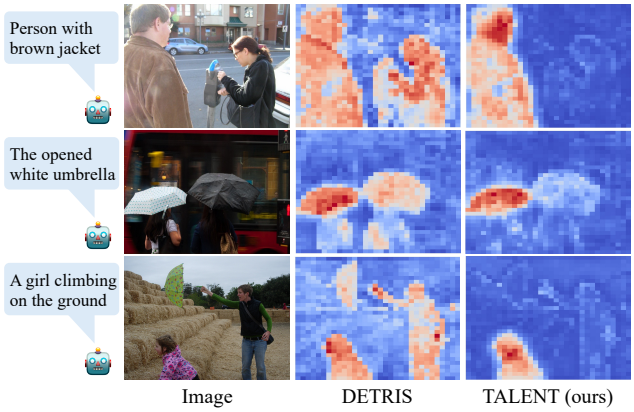


Figure 5. Visualization of feature activation maps. We compare TALENT with the existing SOTA method DETRIS [15].

Table 3. Comparison of parameters and average segment performance on three benchmarks. † indicates the results are calculated from the official implementation of DETRIS [17] and ETRIS [55].

Methods	Params. (backbone)	Params. (other modules)	Params. (overall)	mIoU (Avg.)
CRIS [54]	57.31M	103.94M	161.25M	63.8
LoRA† [11]	**0.03M**	23.98M	24.01M	60.4
Compacter† [21]	0.19M	23.98M	24.17M	61.3
ETRIS† [55]	1.94M	23.98M	25.92M	62.8
BarLeRIa [51]	2.21M	23.98M	26.19M	66.5
DETRIS† [15]	2.71M	22.84M	26.69M	70.4
TALENT (ours)	2.33M	**20.44M**	**22.77M**	**72.0**

tiveness and efficiency of our TALENT.

Effect on various proposed modules. Tab. 4 illustrates comprehensive experiments on each proposed module of our TALENT: RCA, CPCL, and TCCL by adding each one incrementally. The baseline model is constructed by simply combining the vision and text backbones without our proposed modules. Our RCA achieves an average gain of 1.1% mIoU on RefCOCO. Next, our CPCL and TCCL achieve an average gain of 1.0% mIoU and 0.6% mIoU, and further boost the performance gain up to 2.7% mIoU compared with the baseline model. Moreover, it's seen that CPCL and TCCL significantly reduce the NTA-IoU by up to 3.7% mIoU and 4.5% mIoU, indicating that our TLM can effectively address the 'NTA' issue. These improvements enable our TALENT to set the new SOTA performance for PET-based RIS and show that our method can precisely segment the text-referred target.

Table 4. Ablation results on each proposed module to validate the performance impact using the mIoU and NTA-IoU metrics.

RCA	CPCL	TCCL	RefCOCO-mIoU(%)			AVG(%)	NTA-IoU(%)
			val	testA	testB		
✘	✘	✘	74.9	77.1	71.9	74.6	9.9
✔	✘	✘	76.0	77.9	73.3	75.7	8.2
✔	✔	✘	77.2	78.7	74.1	76.7	4.5
✔	✘	✔	76.8	78.4	73.8	76.3	3.7
✔	✔	✔	**77.8**	**79.4**	**74.8**	**77.3**	**2.1**

Effect on feature aggregation architectures for PET. In Tab. 5, we compare different feature aggregation architectures of Eq. (4) and the verification of the rescaling diagonal matrix (φ). It's seen that setting #6 and #4 both surpass setting #5 and #3 by 0.7% mIoU on testB set, indicating that the learnable diagonal matrix contributes to the robust tuning on visual-text interaction. It's also shown that directly adding the residual feature performs better than the concatenation operation. One possible reason is that these two features are mixed in the concatenated channel dimension, which implicitly hinders the tuning process. Moreover, Settings #1 and #2 both suffer from a significant performance drop, which indicates the importance of the propagated vision feature f_v^i. These results show the effectiveness of both feature aggregation strategy and rescaling diagonal matrix.

Table 5. Ablation results on different feature aggregation architectures of Eq. 4. $cat(\cdot)$ means to concatenate the two input features and then use an MLP for channel reduction.

Settings	Architecture	RefCOCO-mIoU(%)		
		val	testA	testB
#1	$\nu_{i+1}(f_{vt}^i)$	71.8	76.0	70.1
#2	$\nu_{i+1}(f_{vt}^i) * \varphi$	71.1	74.5	68.9
#3	$cat(\nu_{i+1}(f_v^i), f_{vt}^i)$	76.1	77.9	73.2
#4	$cat(\nu_{i+1}(f_v^i), f_{vt}^i * \varphi)$	76.6	78.8	73.9
#5	$\nu_{i+1}(f_v^i) + f_{vt}^i$	77.1	78.7	74.1
#6	$\nu_{i+1}(f_v^i) + f_{vt}^i * \varphi$	**77.8**	**79.4**	**74.8**

Effect on sensitivity of the loss scaling coefficients. Tab. 6 evaluates the effects of scaling coefficient λ_{cpcl} and λ_{tccl} on the RefCOCO dataset. Specifically, when λ_{cpcl} and λ_{tccl} are both set to 0.1, TALENT achieves the best performance. It indicates that our progressive context learning serves as an auxiliary loss that will not impact the main loss. These two auxiliary losses improve the segmentation performance and mitigate the 'NTA' issue without extra trainable parameters. More results are summarized in the appendix.

Table 6. Ablation results on each loss scaling coefficient.

λ_{cpcl} / λ_{tccl}	0.3 / 0.1	0.2 / 0.2	0.2 / 0.1	0.1 / 0.3	0.1 / 0.2	0.1 / 0.1	0.1 / 0.0	0.0 / 0.1
val	76.9	77.2	77.5	77.4	77.6	**77.8**	77.2	76.8
testA	78.5	78.6	79.1	78.9	79.1	**79.4**	78.7	78.4
testB	74.3	74.3	74.4	74.2	74.6	**74.8**	74.1	73.8

5. Conclusion

In this paper, we identify 'NTA' issue in PET-based RIS methods and introduce a new metric, NTA-IoU, to quantify its impact. To address 'NTA', we propose TALENT, an efficient tuning framework. First, TALENT introduces a Rectified Cost Aggregator (RCA) for visual-text interaction. Next, TALENT adopts a Target-aware Learning Mechanism (TLM), including Contextual Pairwise Consistency Learning (CPCL) and Target Centric Contrastive Learning (TCCL). CPCL optimizes the visual feature representation guided by a text-augmented affinity map, and TCCL further improves its discriminative ability between the distinct target instance and other irrelevant ones. Extensive experiments indicate that TALENT achieves significant gains over existing methods and effectively mitigates the 'NTA' issue.

6. Acknowledgment

This work was supported by the National Natural Science Foundation of China (No. 62301451, 62301613, 62471405, 62331003), Basic Research Program of Jiangsu (BK20241814), Suzhou Basic Research Program (SYG202316), XJTLU REF-22-01-010 and XJTLU RDF-22-02-066.

References

[1] Bo Chen, Zhiwei Hu, Zhilong Ji, Jinfeng Bai, and Wangmeng Zuo. Position-aware contrastive alignment for referring image segmentation. *arXiv preprint arXiv:2212.13419*, 2022. 3

[2] Silin Cheng, Yang Liu, Xinwei He, Sebastien Ourselin, Lei Tan, and Gen Luo. Weakmcn: Multi-task collaborative network for weakly supervised referring expression comprehension and segmentation. In *CVPR*, pages 9175–9185, 2025. 3

[3] Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask grounding for referring image segmentation. In *CVPR*, pages 26573–26583, 2024. 6

[4] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *CVPR*, pages 4113–4123, 2024. 4

[5] Ming Dai, Jian Li, Jiedong Zhuang, Xian Zhang, and Wankou Yang. Multi-task visual grounding with coarse-to-fine consistency constraints. In *AAAI*, pages 2618–2626, 2025. 3

[6] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, pages 16321–16330, 2021. 2

[7] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, pages 19358–19369, 2023. 1

[8] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, 132(2):581–595, 2024. 1, 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 1

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, pages 1–8, 2022. 3, 7, 8

[12] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, pages 108–124, 2016. 2

[13] Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. Beyond one-to-one: Rethinking the referring image segmentation. In *ICCV*, pages 4067–4077, 2023. 1, 6

[14] Zhiwei Hu, Guang Feng, Jiayu Sun, Lihe Zhang, and Huchuan Lu. Bi-directional relationship inferring network for referring image segmentation. In *CVPR*, pages 4424–4433, 2020. 2

[15] Jiaqi Huang, Zunnan Xu, Ting Liu, Yong Liu, Haonan Han, Kehong Yuan, and Xiu Li. Densely connected parameter-efficient tuning for referring image segmentation. In *AAAI*, pages 3653–3661, 2025. 1, 2, 3, 5, 6, 7, 8

[16] Qihan Huang, Siming Fu, Jinlong Liu, Hao Jiang, Yipeng Yu, and Jie Song. Resolving multi-condition confusion for finetuning-free personalized image generation. In *AAAI*, pages 3707–3714, 2025. 1

[17] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *CVPR*, pages 10488–10497, 2020. 8

[18] Shibo Jie, Zhi-Hong Deng, Shixuan Chen, and Zhijuan Jin. Convolutional bypasses are better vision transformer adapters. In *ECAI*, pages 202–209, 2024. 3

[19] Shuo Jin, Meiqin Liu, Chao Yao, Chunyu Lin, and Yao Zhao. Kernel dimension matters: To activate available kernels for real-time video super-resolution. In *ACMMM*, pages 8617–8625, 2023. 3

[20] Shuo Jin, Siyue Yu, Bingfeng Zhang, Mingjie Sun, Yi Dong, and Jimin Xiao. Feature purification matters: Suppressing outlier propagation for training-free open-vocabulary semantic segmentation. In *ICCV*, pages 20291–20300, 2025. 3

[21] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. In *NeurIPS*, pages 1022–1035, 2021. 7, 8

[22] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 6

[23] Seoyeon Kim, Minguk Kang, Dongwon Kim, Jaesik Park, and Suha Kwak. Extending clip's image-text alignment to referring image segmentation. In *NAACL*, pages 4611–4628, 2024. 6

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 7

[25] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, pages 9579–9589, 2024. 6

[26] Hui Li, Mingjie Sun, Jimin Xiao, Eng Gee Lim, and Yao Zhao. Fully and weakly supervised referring expression segmentation with end-to-end learning. *IEEE T-CSVT*, 33(10): 5999–6012, 2023. 2

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 6

[28] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *CVPR*, pages 23592–23601, 2023. 6

[29] Meiqin Liu, Shuo Jin, Chao Yao, Chunyu Lin, and Yao Zhao. Temporal consistency learning of inter-frames for video super-resolution. *IEEE T-CSVT*, 33(4):1507–1520, 2022. 3

[30] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *IEEE T-PAMI*, 44(9):4761–4775, 2021. 2

[31] Sun-Ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. Caris: Context-aware referring image segmentation. In *ACMMM*, pages 779–788, 2023. 2

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 1

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 2

[34] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*, pages 10034–10043, 2020. 2

[35] Zhuoyan Luo, Yinghao Wu, Tianheng Cheng, Yong Liu, Yicheng Xiao, Hongfa Wang, Xiao-Ping Zhang, and Yujiu Yang. Cohd: A counting-aware hierarchical decoding framework for generalized referring expression segmentation. In *ICCV*, pages 1–8, 2025. 6

[36] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *ECCV*, pages 792–807, 2016. 6

[37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2023. 1, 4

[38] Yuwen Pan, Rui Sun, Yuan Wang, Tianzhu Zhang, and Yongdong Zhang. Rethinking the implicit optimization paradigm with dual alignments for referring remote sensing image segmentation. In *ACMMM*, pages 2031–2040, 2024. 1

[39] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Yu Huang, Yaoming Wang, and Wei Shen. Parameter-efficient fine-tuning in hyperspherical space for open-vocabulary semantic segmentation. In *CVPR*, pages 15009–15020, 2025. 2

[40] Xianglin Qiu, Xiaoyang Wang, Zhen Zhang, and Jimin Xiao. Bias-resilient weakly supervised semantic segmentation using normalizing flows. In *CVPR*, pages 21321–21330, 2025. 3

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 3

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 4, 5

[43] Nisarg A Shah, Vibashan VS, and Vishal M Patel. Lqmformer: Language-aware query mask transformer for referring image segmentation. In *CVPR*, pages 12903–12913, 2024. 6

[44] Chao Shang, Zichen Song, Heqian Qiu, Lanxiao Wang, Fanman Meng, and Hongliang Li. Prompt-driven referring image segmentation with instance contrasting. In *CVPR*, pages 4124–4134, 2024. 6, 7

[45] Mingjie Sun, Jimin Xiao, and Eng Gee Lim. Iterative shrinking for referring expression grounding using deep reinforcement learning. In *CVPR*, pages 14060–14069, 2021. 2

[46] Mingjie Sun, Jimin Xiao, Eng Gee Lim, Si Liu, and John Y Goulermas. Discriminative triad matching and reconstruction for weakly referring expression grounding. *IEEE T-PAMI*, 43(11):4189–4195, 2021. 2

[47] Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening anchor token: Decoding strategy in alleviating hallucinations for mllms. In *ICLR*, pages 1–8, 2025. 1

[48] Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, et al. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In *CVPR*, pages 26147–26159, 2025. 1

[49] Jiajin Tang, Ge Zheng, Cheng Shi, and Sibei Yang. Contrastive grouping with transformer for referring image segmentation. In *CVPR*, pages 23570–23580, 2023. 6, 7

[50] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *ICCV*, pages 32–42, 2021. 4

[51] Yaoming Wang, Jin Li, XIAOPENG ZHANG, Bowen Shi, Chenglin Li, Wenrui Dai, Hongkai Xiong, and Qi Tian. Barleria: An efficient tuning framework for referring image segmentation. In *ICLR*, pages 1–8, 2023. 2, 3, 6, 8

[52] Yuexin Wang, Xiaolei Wang, Yizheng Gong, and Jimin Xiao. Normal-abnormal guided generalist anomaly detection. In *NeurIPS*, pages 1–8, 2025. 1

[53] Yuexin Wang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Tammam Tillo, and Jimin Xiao. Fgpt: Fine-grained prompt tuning for zero-shot anomaly detection. *IEEE T-AI*, pages 1–13, 2026. 1

[54] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, pages 11686–11695, 2022. 1, 3, 5, 8

[55] Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, and Guanbin Li. Bridging vision and language encoders: Parameter-efficient tuning for referring image segmentation. In *ICCV*, pages 17503–17512, 2023. 3, 5, 6, 7, 8

[56] Haochen Xue, Feilong Tang, Ming Hu, Yexin Liu, Qidong Huang, Yulong Li, Chengzhi Liu, Zhongxing Xu, Chong Zhang, Chun-Mei Feng, et al. Mmrc: A large-scale benchmark for understanding multimodal large language model in

real-world conversation. *arXiv preprint arXiv:2502.11903*, 2025. 1

[57] Haolin Yang, Feilong Tang, Linxiao Zhao, Xiang An, Ming Hu, Huifa Li, Xinlin Zhuang, Boqian Wang, Yifan Lu, Xiaofeng Zhang, et al. Streamagent: Towards anticipatory agents for streaming video understanding. *arXiv preprint arXiv:2508.01875*, 2025. 1

[58] Yuhuan Yang, Chaofan Ma, Jiangchao Yao, Zhun Zhong, Ya Zhang, and Yanfeng Wang. Remamber: Referring image segmentation with mamba twister. In *ECCV*, pages 108–126, 2024. 1, 6

[59] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, pages 18155–18165, 2022. 1, 2, 6

[60] Ziqian Yang, Xinqiao Zhao, Xiaolei Wang, Quan Zhang, and Jimin Xiao. Ffr: Frequency feature rectification for weakly supervised semantic segmentation. In *CVPR*, pages 30261–30270, 2025. 2

[61] Houjian Yu, Mingen Li, Alireza Rezazadeh, Yang Yang, and Changhyun Choi. A parameter-efficient tuning framework for language-guided object grounding and robot grasping. *arXiv preprint arXiv:2409.19457*, 2024. 2, 6, 7

[62] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 1