TALENT: Target-aware Efficient Tuning for Referring Image Segmentation

Supplementary Material

A. Overview

In this supplementary material, we provide additional ablation experiments of the target-aware learning mechanism in Sec. B with further discussions, and provide more feature activation and segmentation results in Sec. C.

B. More ablation experiments

B.1. Analysis about target-aware learning mechanism upon visual-text interaction.

To further analyze the effectiveness of our Target-aware Learning Mechanism (TLM), including Contextual Pairwise Consistency Learning (CPCL) and Target-Centric Contrastive Learning (TCCL), we conduct experiments by applying them to another visual-text interaction approach, cross-attention. For a fair comparison, we keep all other components consistent with our Rectified Cost Aggregator (RCA). Note that the cross attention is a standard design, where the image is denoted as the query, while the text is denoted as the key and the value.

As summarized in Table 1, we first discard the visual-text interaction mechanisms between visual and text backbones and take it as the baseline setting. It's observed that this setting can't achieve promising performance. Then, we compare the performance of different visual-text interaction mechanisms. Specifically, when only evaluating RCA against cross-attention in isolation, *i.e.*, without TLM, RCA achieves 1.4% mIoU gains over the baseline and 0.9% mIoU gains over the cross-attention on the RefCOCO testB set. When both CPCL and TCCL are applied, RCA further achieves 2.9% mIoU gains over the baseline and 0.7% mIoU gains over the cross-attention mechanism, with fewer parameter numbers as shown in Tab. 3 of the main paper.

In summary, RCA represents the first attempt to leverage cost-volume analysis for visual–text interaction modeling in PET-based RIS. These results demonstrate that our RCA serves as a more effective method for visual-text interaction compared to the cross-attention mechanism.

Discussion about generalization of TLM. Recall that TLM is designed to refine the multimodal features produced from visual–text interaction modules for mitigating the 'NTA' issue. Therefore, TLM is expected to possess a strong generalization ability to handle multimodal features produced by different visual–text interaction mechanisms. As shown in Tab. 1, TLM also functions effectively when combined with cross-attention instead of RCA. Specifically, when only using cross-attention in isolation, *i.e.*, without TLM, it only achieves an average of 0.4% mIoU gains over the baseline on RefCOCO. When both CPCL and TCCL are applied, it further achieves 2.2% mIoU gains over the baseline model. These results underscore the generalization ability of our TLM, which also works on cross-attention.

Table 1. Ablation experiments on applying CPCL and TCCL to different visual-text interaction mechanisms.

V-L interaction	TLM		RefCOCO			Avg
	CPCL	TCCL	val	testA	testB	
✘	✘	✘	74.9	77.1	71.9	74.6
RCA	✘	✘	76.0	77.9	73.3	75.7
RCA	✔	✔	**77.8**	**79.4**	**74.8**	**77.3**
cross-attention	✘	✘	75.4	77.3	72.4	75.0
cross-attention	✔	✔	77.3	78.9	74.1	76.8

B.2. Effect on sensitivity of loss scaling coefficient

To provide a more comprehensive analysis of the two learning objectives in TLM, we try to fix one coefficient of the two objectives and verify how they influence the segmentation performance. Specifically, we fix the coefficient of CPCL to several representative values and vary the coefficient of TCCL accordingly. The performance curves are plotted in Fig. 1, where each colored line corresponds to a specific CPCL coefficient. It is observed that when λ_{cpcl} is fixed at 0.1, TALENT consistently achieves the best performance on the RefCOCO dataset, regardless of the value of the TCCL coefficient. Moreover, among different values of λ_{cpcl}, TALENT tends to achieve optimal performance when λ_{tccl} is set to 0.1. These results indicate that setting both λ_{cpcl} and λ_{tccl} to 0.1 yields the best performance.

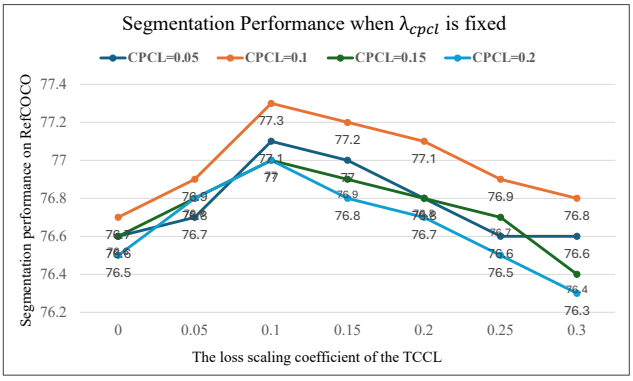

Figure 1. Ablation results of the evaluation of the loss scaling coefficient for our CPCL and TCCL. Each colored line corresponds to a specific CPCL coefficient.

Figure 2. Visualization of segmentation results. We compare our TALENT with existing PET-based methods. It's observed that TALENT can generate more precise segmentation maps.

C. More visualization results

C.1. Segmentation visualization

We compare more visualization results of the segmentation maps with PET-based methods [1–3], which are illustrated in Fig. 2. It's evident that prior approaches have difficulty accurately localizing the target instance described by the text expression. Instead, these methods frequently misidentify other salient and similar objects, like 'apple slices' and 'left corner orange' in the last and the penultimate row of Fig. 2. In contrast, our TALENT can accurately identify and segment the distinct text-referred target instance. These results demonstrate that our proposed target-aware learning mechanism can effectively enhance the feature discrimination ability and reduce the impact of the 'NTA' issue.

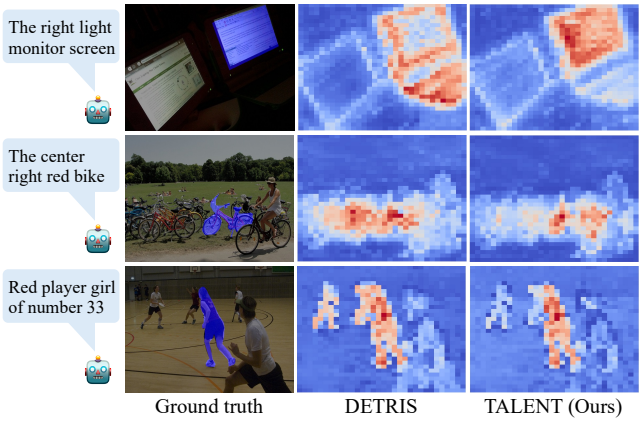

Figure 3. Visualization of feature activation maps. We compare TALENT with the SOTA PET-based method DETRIS [1].

C.2. Feature activation visualization

Fig. 3 compares the feature activation among different PET-based methods to evaluate the effectiveness of mitigating the 'NTA' issue. It's observed that our TALENT can accurately activate the distinct text-referred visual region, whereas the previous SOTA method, DETRIS [1], often activates salient yet unrelated objects. Specifically, given the text expression 'the center right red bike' in the second row, DETRIS [1] often tends to activate other similar objects, *e.g.*, other red bikes. In the third row, DETRIS [1] tends to activate other player girls in red when given the text expression 'red player girl of number 33'. In contrast, our TALENT selectively activates the specific target that consistently aligns with the GT segmentation mask, which underscores the importance of mitigating 'NTA'.

References

[1] 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. 2, 3

[2] 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] 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