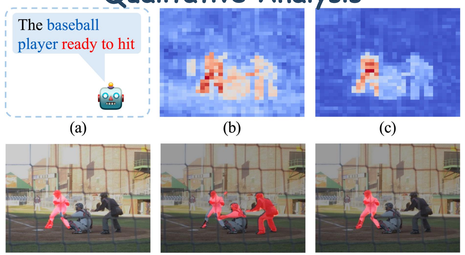
TALENT: Target-aware Efficient Tuning for Referring Image Segmentation

Shuo Jin, Siyue Yu*, Bingfeng Zhang, Chao Yao, Meiqin Liu, Jimin Xiao

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation & Introduction

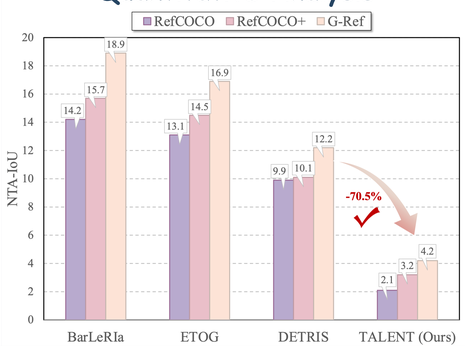
Qualitative Analysis



Non-Target feat Activation

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

Quantitative Analysis

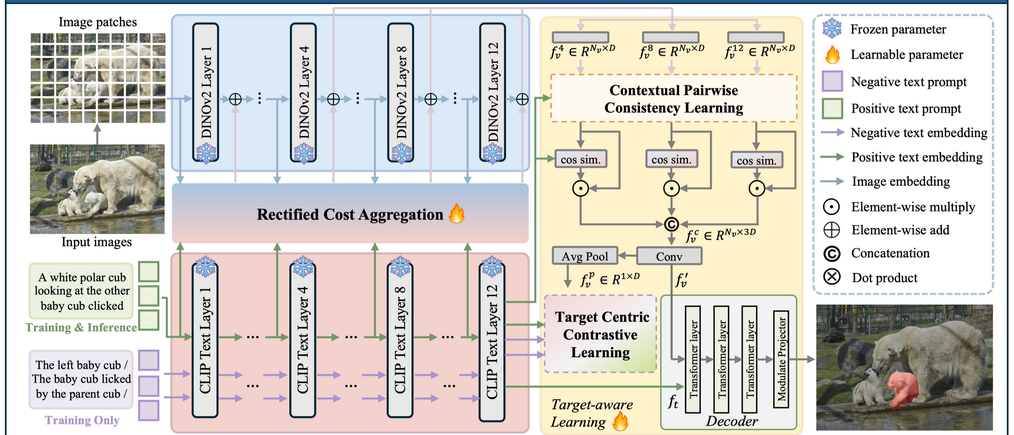


Non-Target Inter-of-Union

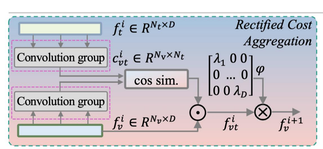
We introduce NTA-IoU to quantify this 'NTA' issue, which measures the proportion of the incorrectly segmented regions that fall into non-target instances within the same category

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

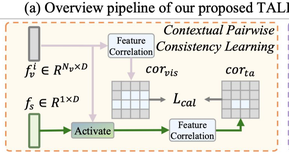
Methodology



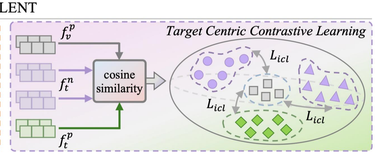
(a) Overview pipeline of our proposed TALENT



(b) Depiction of RCA



(c) Depiction of CPCL



(d) Depiction of TCCL

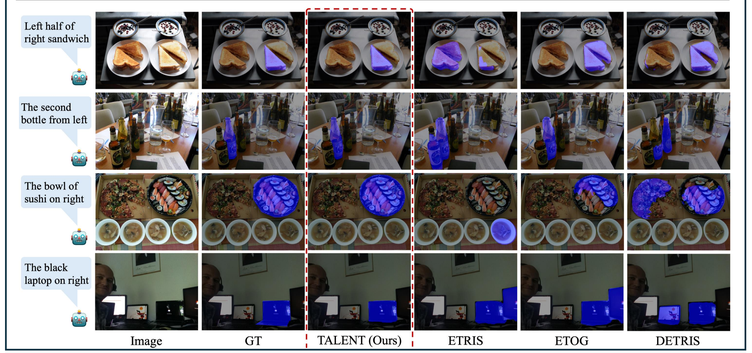
Rectified Cost Aggregation
RCA for efficient tuning with rectification on negative text cues

Contextual Pairwise Consistency Learning
CPCL constrains the visual feature to highlight the correct context region

Target Centric Contrastive Learning
TCCL enhances feature discrimination by aligning positive text features

Experimental Results

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**



Limitations & Future Works

Text Reasoning

Negative correlation isn't necessarily noises. For more abstract text cues, the ability of current mechanism maybe not enough for reasoning

Dual Tower or Single Tower

Recent approaches process image and text tokens with a unified decoder, where NTA issues remain worth exploring.

If interested in our work, CHECK our Paper and Codes HERE

Paper Code

Ablation Studies

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**

Architecture Design for Efficient Tuning

V-L interaction	TLM CPCL	TLM TCCL	RefCOCO val	RefCOCO testA	RefCOCO testB	Avg
✘	✘	✘	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

Target-aware Learning Mechanism as Plug-in Module

Text-activated Feature Visualization