WeakMCN: Multi-task Collaborative Network for Weakly Supervised Referring Expression Comprehension and Segmentation
–Supplementary Material–

Silin Cheng[1*], Yang Liu[2*], Xinwei He[3], Sebastien Ourselin[2], Lei Tan[4†], Gen Luo[5†]

[1]The University of Hong Kong [2]King's College London [3]Huazhong Agricultural University
[4]National University of Singapore [5]OpenGVLab, Shanghai AI Laboratory

In this supplementary material, we present additional qualitative and quantitative results of our proposed WeakMCN. Section 1 includes ablation studies on (1) a comparative analysis between our trainable WRES head and a straightforward SAM-based pipeline, demonstrating the advantages of our approach, (2) the sensitivity analysis of the ISL threshold, (3) the impact of different visual features in DVFE, and (4) the parameter efficiency comparison with existing methods. Section 2 analyzes typical failure cases to identify current limitations and future directions for improvement.

1. Additional Ablation Studies

Comparison with Direct SAM Application. Our method leverages SAM for generating pseudo masks to train the segmentation head. An alternative strategy is to directly employ SAM for mask generation at inference time. To quantitatively evaluate these two approaches, we conducted comparative experiments, with results presented in Table 1: first training a REC model with DVFE for localization (first row), then using its predicted boxes to prompt SAM for mask generation at inference time (second row). While this pipeline achieves competitive performance, achieving 67.36% REC and 53.97% RES on RefCOCO, we observe a notable performance gap compared to our proposed WeakMCN (third row), particularly in RES performance. For instance, on RefCOCO, WeakMCN outperforms this alternative approach by 1.19% and 4.18% in REC and RES metrics respectively. The performance gap highlights two key advantages of our approach: (1) While both methods utilize SAM, ours leverages it only for pseudo mask generation during training, allowing our lightweight WRES head to learn task-specific features, whereas direct SAM application is entirely dependent on the quality of the predicted bounding boxes of WREC head at inference time. (2) Our trainable WRES head enables dynamic feature interaction with the WREC head during training, fostering mutual enhancement between WREC and WRES. These results validate our design choice of using SAM as a teacher model for training rather than as a direct inference tool.

Table 1. **Comparison of replacing the WRES head with the SAM head.**

Model	RefCOCO		RefCOCO+	
	REC	RES	REC	RES
WeakMCN (w/o WRES)	67.36	-	48.94	-
WeakMCN (w/o WRES) + SAM_{head}	67.36	53.97	48.94	37.97
WeakMCN	68.55	58.15	51.48	41.48

Table 2. **Comparison of various hyperparameter thresholds (α) in ISL.**

α	RefCOCO		RefCOCO+	
	REC	RES	REC	RES
0.1	68.03	57.82	50.26	41.58
0.2	68.38	57.91	51.48	41.48
0.3	68.55	58.15	50.49	41.34
0.4	68.64	58.03	50.19	40.57

The impact of the threshold in ISL. Tab. 2 presents the impact of varying hyperparameter thresholds α in ISL. For RefCOCO, the best performance is observed at $\alpha = 0.3$, achieving improvements of 0.61% and 0.19% in the WREC and WRES tasks, respectively, compared to the worst-performing configuration. Similarly, for RefCOCO+, the optimal performance occurs at $\alpha = 0.2$, with gains of 1.29% and 0.91% in the WREC and WRES tasks, respectively. Overall, these results demonstrate that the proposed WeakMCN model exhibits robustness to the choice of α, showing minimal sensitivity to this hyperparameter. In this paper, we adopt $\alpha = 0.3$ for consistency across experiments.

More visual features in visual bank. To investigate the impact of incorporating additional visual features into our

*Equal contribution.
†Corresponding author

Table 3. **Ablation studies of DVFE in WeakMCN.**

$\mathcal{B}$			RefCOCO		RefCOCO+	
V_{dino}	V_{sam}	V_{clip}	REC	RES	REC	RES
✓			67.37	56.14	50.32	40.43
✓	✓		68.55	58.15	51.49	41.47
✓	✓	✓	68.14	57.64	50.98	40.76

Table 4. **The efficiency of DVFE in WeakMCN.**

Features in DVFE			Infrence Speed.	RefCOCO		RefCOCO+	
V_{dark}	V_{dino}	V_{sam}		REC	RES	REC	RES
✓			24.5fps	63.95	46.88	39.84	28.61
✓	✓		20.3fps	67.37	56.14	50.32	40.43
✓	✓	✓	17.7fps	68.55	58.15	51.49	41.47

Table 5. **Comparison of parameters with other weakly-supervised RES or REC methods.** Params denote the number of trainable parameters.Train denote traning hours. Inf denote inference speed.

Model	Multi-task	Params(M)	Train(h)	Inf(fps)	RefCOCO		RefCOCO+	
					REC	RES	REC	RES
RefCLIP [1]	✗	27.50	5	31.3	60.36	-	40.39	-
APL [4]	✗	49.91	7.5	18.2	64.51	-	42.70	-
TRIS [3]	✗	113.56	-	-	-	31.17	-	30.90
Shatter [2]	✗	145.96	25.5	7.51	-	34.76	-	28.48
WeakMCN	✓	34.31	7	17.7	68.55	58.15	51.48	41.48

model, we conduct detailed ablation studies on the Dynamic Visual Feature Encoder (DVFE) as shown in Table 3. We systematically evaluate three visual features: DINO features (V_{dino}), SAM features (V_{sam}), and CLIP features (V_{clip}). Our experiments reveal that while the combination of V_{dino} and V_{sam} achieves strong performance, further incorporating V_{clip} leads to slight performance degradation. For example, on RefCOCO, we observe performance drops of 0.41% and 0.51% for REC and RES tasks respectively when adding V_{clip} to the V_{dino}+V_{sam} combination. We hypothesize that this degradation stems from the redundant information and training noise introduced by excessive visual features, which may contaminate the learned feature representations. This finding emphasizes the crucial importance of maintaining a balanced and efficient visual feature bank rather than merely accumulating features.

The efficiency of DVFE. As shown in Table 4, we conduct ablation studies to analyze the efficiency-performance trade-off of our proposed DVFE. The baseline model with only DarkNet features (V_{dark}) achieves 24.5 FPS but shows limited performance (63.95% REC, 46.88% RES on RefCOCO). By incorporating DINO features (V_{dino}), the inference speed slightly decreases to 20.3 FPS, while bringing substantial improvements in both REC (+3.42%) and RES (+9.26%). The full DVFE implementation with all three features (V_{dark}, V_{dino}, and V_{sam}) further boosts the performance to 68.55% REC (+4.60% over baseline) and 58.15% RES (+11.27% over baseline) on RefCOCO, at the cost of reducing inference speed to 17.7 FPS. Similar performance gains are observed on RefCOCO+, where the full DVFE achieves significant improvements in both REC (+11.65%) and RES (+12.86%) compared to using V_{dark} alone. These results demonstrate that while additional features moderately impact computational efficiency, the performance benefits of our DVFE are substantial and justify the modest decrease in inference speed. The flexible architecture of DVFE enables different feature combinations to meet various speed-accuracy requirements in real-world applications.

Efficiency Comparison with SOTA Methods. The experimental results in Table 5 demonstrate the comprehensive advantages of our WeakMCN in terms of parameter efficiency, training efficiency, and inference speed. From the perspective of model size, with only 34.31M trainable parameters, WeakMCN significantly reduces the number of learnable parameters by 31.3%, 76.5%, and 69.8% compared to APL (49.91M), Shatter (145.96M), and TRIS (113.56M), respectively. In terms of training efficiency, WeakMCN requires only 7 hours for convergence, which is considerably faster than Shatter (25.5h) and comparable to APL (7.5h). For inference speed, WeakMCN achieves 17.7 FPS, showing better real-time capability than APL (18.2 FPS) and significantly outperforming Shatter (7.51 FPS). Despite being more efficient, WeakMCN achieves state-of-the-art performance on both tasks, surpassing RefCLIP (60.36%) by 8.19% and APL (64.51%) by 4.04% in REC accuracy (68.55%), while outperforming Shatter (34.76%) by 23.39% and TRIS (31.17%) by 26.98% in RES performance (58.15%). Particularly noteworthy is that WeakMCN is the only model that simultaneously handles both REC and RES tasks while maintaining competitive efficiency metrics. These results validate the effectiveness of our multi-task learning framework in achieving a superior balance between computational efficiency and performance enhancement.

2. Failure Cases

Fig. 1 illustrates typical failure cases that reveal the current limitations of our approach. Specifically, cases 1-3 demonstrate that WeakMCN tends to produce oversegmented predictions when multiple objects overlap within a single detected bounding box, despite achieving accurate localization. Furthermore, cases 4-6 showcase the model's difficulty in processing complex and lengthy expressions, particularly in terms of precise object localization. These failure cases indicate that there remains substantial room for improvement in WeakMCN's visual reasoning capabilities and scene understanding, especially for handling intricate spatial relationships and complex visual contexts.

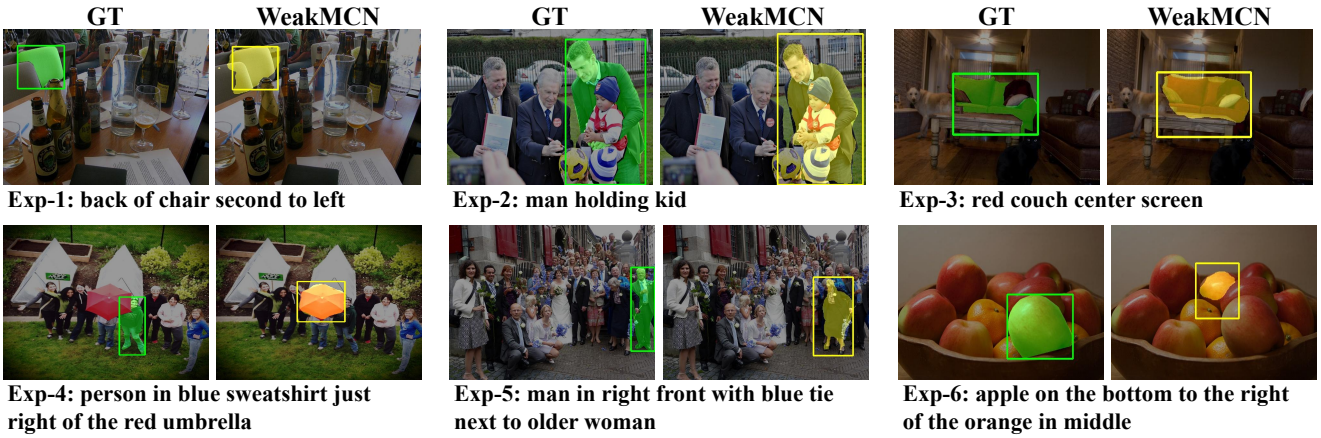

Figure 1. **Failure cases.** The green mask/bounding box is the ground truth, and the yellow one is our prediction.

References

[1] Lei Jin, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. Refclip: A universal teacher for weakly supervised referring expression comprehension. In *CVPR*, 2023. 2

[2] Dongwon Kim, Namyup Kim, Cuiling Lan, and Suha Kwak. Shatter and gather: Learning referring image segmentation with text supervision. In *ICCV*, 2023. 2

[3] Fang Liu, Yuhao Liu, Yuqiu Kong, Ke Xu, Lihe Zhang, Baocai Yin, Gerhard Hancke, and Rynson Lau. Referring image segmentation using text supervision. In *ICCV*, 2023. 2

[4] Yaxin Luo, Jiayi Ji, Xiaofu Chen, Yuxin Zhang, Tianhe Ren, and Gen Luo. Apl: Anchor-based prompt learning for one-stage weakly supervised referring expression comprehension. In *ECCV*, 2024. 2