Appendix of Momentum Pseudo-Labeling for Weakly Supervised Phrase Grounding

Experimental results

We also conducted relevant experiments on RefCOCOg (Mao et al. 2016), and due to space limitations, the results are reported here. The comparison results with CPL (Liu et al. 2023) and RefCLIP (Jin et al. 2023) are shown in the table below.

Method	val-umd	test-umd	val-google
CPL'23	59.19	53.80	53.92
RefCLIP'23	-	-	47.87
ours	55.29	55.83	57.41

Table 1: Experiment results on RefCOCOg

In terms of changing the feature extraction in the visual modality, we tried replacing the backbone with ViT-B (using CLIP's pre-trained parameters), but obtained a lower result as shown in Table 2. This trend is consistent with the phenomenon observed in CYC (Zhang, Wang, and Liu 2023).

Case Study details

We visualize more prediction examples from the RefCOCO+ dataset as Figure 1 shown, comparing our method with the local pseudolabel (CLEM) updating approach. Phrases in RefCOCO+ are typically longer and require contextual understanding for effective grounding, which adds to the complexity.

Experimental details

Dataset Statistics

Flickr30k Entities: Flickr30k Entities is an extended version of the Flickr30k dataset (Young et al. 2014), which further annotates this foundational dataset by adding detailed information about specific entities within the images. The dataset includes about 31k images and 158k corresponding captions, with each image accompanied by five different captions.

RefCOCO/RefCOCO+: The datasets were collected using the ReferitGame (Kazemzadeh et al. 2014), a two-player game. RefCOCO is a dataset for natural language tasks involving object references in images, featuring 142,209 expressions for 50,000 objects across 19,994 images. RefCOCO+ is a variant of RefCOCO that excludes absolute location descriptions, making it more challenging, with 141,564 expressions for 49,856 objects in 19,992 images. In our work, we employ the UNC split (Yu et al. 2016), dividing both datasets into four parts: train, validation, testA, and testB.

Evaluation Mertrics

For each phrase, we take the bounding box with the highest similarity as the prediction for grounding. If the overlap, measured by Intersection Over Union (IOU), between the predicted bounding box and the annotated bounding box, is greater than 0.5, the prediction is considered correct. When a phrase refers to multiple annotated bounding boxes, following the method (Wang et al. 2020), these boxes are merged into one larger golden bounding box. We also use this metric to evaluate the accuracy of pseudo-labels. It is important to note that the fine-grained labels are not visible during the training process.

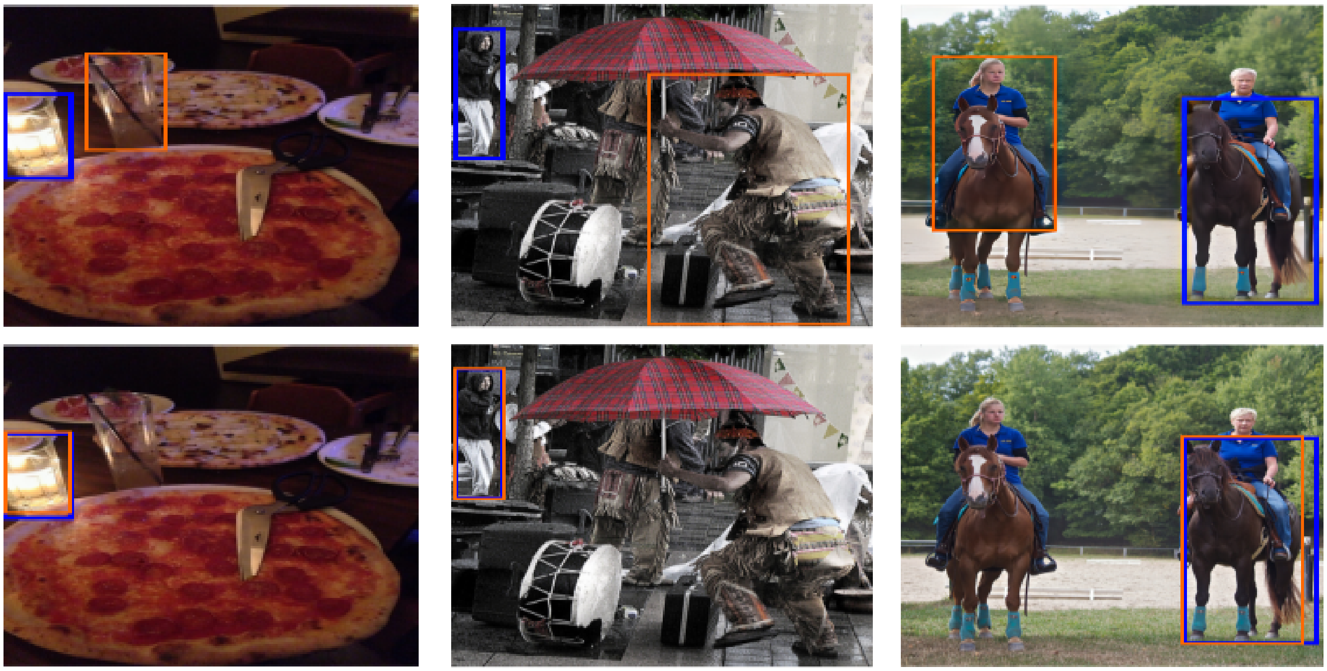
Implementation Details

- **Features Extraction:** For images in the Flickr30K Entities dataset, we used the same image features as in the MAF[1]. Additionally, we extracted features from the COCO dataset using the VOLTA[2] framework for training and testing on the RefCOCO/+ datasets.
- **Hyperparameter Settings:** Our model training employed an SGD optimizer with a momentum coefficient of 0 and a learning rate set at 1e-3. The momentum update coefficient γ for EMA in the model was set to 0.99. For all three datasets, we set the batch size B to 256, meaning each iteration includes 256 image-text pairs. On the Flickr Entities dataset, we initially set the temperature τ_E for pseudo-label generation to 0.3, then gradually decreased it to 0.1 using a moving average. We trained for 45 epochs, with the temperature τ for contrastive loss

[1] https://github.com/qinzzz/Multimodal-Alignment-Framework
[2] https://github.com/e-bug/volta

Method	Prop./Backbone	RefCOCO testA/B	RefCOCO+ testA/B	RefCOCOg val-g
CYC'23	VG/ViT-B	31.27/23.32	31.98/27.04	33.77
CYC'23	VG/RN101	50.80/48.44	43.03/33.40	43.77
ours	VG/ViT-B	46.87/37.72	47.92/38.53	48.70
ours	VG/RN101	70.19/55.74	63.59/45.20	57.41

Table 2: Comparison of experimental results using ResNet101 and ViT-B.



(a) glass with candle in it

(b) far back girl hodded white pants

(c) horse with short hair woman

Figure 1: Visualization of model predictions on the RefCOCO+ validation set: the orange boxes represent the model's predicted regions, and the blue boxes are the annotated results. The caption below each image indicates the phrase that needs to be grounded. The top row displays results from the CLEM method, while the bottom row shows results from our method.

calculation set to 1. For the RefCOCO/+ datasets, we initially set the temperature τ_E to 0.6, then gradually decreased it to 0.1 using a moving average. We trained for 35 epochs, with the temperature $\tau = 1$. Additionally, we set σ to 10 during text encoding.

- **Dual Encoder Settings:** In the visual encoder, we used one linear layer to perform a low-dimensional mapping W_k of the ROI features output by faster-RCNN, which were then fused with region labels extracted by the object detector. This was followed by 6 transformer encoder layers to establish relationships between regions within the image, ultimately outputting the corresponding region features. For text encoding, we used the GloVe[3] from the Wikipedia 2014 + Gigaword 5 version, which contains 6 billion tokens. After applying summing pooling to the hidden states in the phrase, we use a linear layer W_q to align the dimensions of the phrase representation with those of the region features. Additionally, we set the scale parameter σ to 10.
- **Experimental Environment:** All models were trained with one NVIDIA Tesla V100 GPU, running on Ubuntu 18.04, using PyTorch version 1.8.
- **False Negatives Detection:** We also explored various false negative detection schemes, employing different visual models to identify false negative regions. Ultimately, we found that visual features output by the frozen backbone in the original model yielded the best results. Methods based on models like VIT and Unicom (Dosovitskiy et al. 2021; An et al. 2023) were less effective when searching within low-resolution regions.

References

An, X.; Deng, J.; Yang, K.; Li, J.; Feng, Z.; Guo, J.; Yang, J.; and Liu, T. 2023. Unicom: Universal and compact representation learning for image retrieval. *arXiv preprint arXiv:2304.05884*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021.

[3] https://github.com/stanfordnlp/GloVe

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR*.

Jin, L.; Luo, G.; Zhou, Y.; Sun, X.; Jiang, G.; Shu, A.; and Ji, R. 2023. Refclip: A universal teacher for weakly supervised referring expression comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2681–2690.

Kazemzadeh, S.; Ordonez, V.; Matten, M.; and Berg, T. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 787–798.

Liu, Y.; Zhang, J.; Chen, Q.; and Peng, Y. 2023. Confidence-aware Pseudo-label Learning for Weakly Supervised Visual Grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2828–2838.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A. L.; and Murphy, K. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 11–20.

Wang, Q.; Tan, H.; Shen, S.; Mahoney, M. W.; and Yao, Z. 2020. Maf: Multimodal alignment framework for weakly-supervised phrase grounding. *arXiv preprint arXiv:2010.05379*.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2: 67–78.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 69–85. Springer.

Zhang, R.; Wang, C.; and Liu, C.-L. 2023. Cycle-Consistent Weakly Supervised Visual Grounding With Individual and Contextual Representations. *IEEE Transactions on Image Processing*.