Bridging Detection and Language: A Comparative Study of CLIP-Augmented Referring Expression Comprehension

Pinfeng Huang
Department of Statistics
Rice University
ph60@rice.edu

Le Zhang
Department of Computer Science
Rice University
cz81@rice.edu

Video Presentation

https://youtu.be/k72IIbBBwuc

Abstract

Referring Expression Comprehension (REC) tasks require localizing an object in an image given a natural language description. We explore two architectures that integrate the visual-language pretraining power of CLIP with object proposals from Faster R-CNN. The first approach incorporates language guidance directly into the detector by replacing its classification head with a text-conditioned predictor. The second approach decouples detection and matching by scoring cropped proposals using CLIP's pretrained image and text encoders. This report presents both methods, compares their effectiveness, and analyzes how CLIP's semantic alignment capabilities interact with region-level detection in REC tasks.

1. Introduction

Referring Expression Comprehension (REC) involves identifying the region in an image that corresponds to a given textual description. Traditional methods employ object detectors such as Faster R-CNN, often trained on large-scale datasets like COCO, with additional modules for language grounding. In contrast, CLIP [4] demonstrates powerful alignment between full-image and text features, enabling zero-shot classification. This raises a central question: whether CLIP-style alignment can be adapted to region-level detection, especially within the REC setting.

In this work, we investigate two methods that combine CLIP's semantic capabilities with the proposal generation of Faster R-CNN. The first method directly integrates text into the detection pipeline by training a text-conditioned prediction head. The second method takes a decoupled approach: proposals are generated by a fine-tuned detector and then scored via CLIP similarity. Both methods aim to bridge the gap between detection and language grounding, and we compare their effectiveness through empirical evaluation.

Figure 1. Qualitative results from the CLIP-based scoring method. Green boxes indicate ground truth, while red boxes show the top-scoring predicted region. The model generally aligns well with the referring expressions.

2. Related Work

Referring Expression Comprehension (REC) has been studied as a cross-modal localization task, where early methods combined hand-crafted features or CNN embeddings with recurrent language models [7, 3]. Later approaches incorporate object detectors like Faster R-CNN as backbones with attention-based fusion modules for joint reasoning [6, 1]. With the rise of vision-language pretraining, models like CLIP [4] and BLIP [2] have shown strong performance on downstream multimodal tasks. However, most prior work focuses on full-image classification or cap-

tioning, whereas we adapt CLIP for region-level understanding by scoring object proposals. Region-level adaptation of CLIP has been explored in tasks such as dense prediction [5], but these often require retraining or rely on handcrafted prompts. Our method keeps CLIP frozen and leverages pretrained semantics to evaluate detection proposals. To the best of our knowledge, this is one of the first works to compare proposal-based region matching using both detector-derived features and pretrained CLIP features in the context of REC.

3. Methodology

We propose two methods for referring expression comprehension (REC) based on aligning candidate regions with textual descriptions. These methods leverage both a detector-based architecture and a pretrained vision-language model. The overall workflows are illustrated in Figures 2 and 3.

3.1. Method 1: Detector-Based Matching via Dynamic CLIP Predictor

Our first method builds on the Faster R-CNN detector by replacing its default box head with a custom Predictor module, as shown in Figure 2. The model keeps Faster R-CNN only as a region-proposal network: we generate the top-K candidate boxes, crop each patch, and encode it with a frozen CLIP ViT-B/32 image encoder; the user query is expanded with spatial templates that insert a coarse location label (e.g., "top-left") and passed through CLIP's text encoder. For every patch we average its cosine similarities to the prompt embeddings, then blend this score with the RPN objectness via $\hat{S}_i = \lambda\, \bar{S}_i + (1 - \lambda)\, \mathrm{obj}_i$. The highest-scoring box is refined by a two-round discrete search over small translations and scales, re-evaluating CLIP similarity at each step, and the best variant is returned as the final grounding.

3.2. Method 2: Proposal Scoring via CLIP Embeddings

In our second method, we propose a decoupled region scoring strategy illustrated in Figure 3. We first fine-tune a Faster R-CNN model on the referring expression dataset to improve its proposal generation in the RefCOCO context. Then, we extract the top-K high-confidence proposals from each image. These proposals are cropped and resized to 224×224 pixels. Each region is encoded using CLIP's image encoder to obtain a visual embedding, while the referring expression is encoded using CLIP's text encoder. Cosine similarity is computed between each region embedding and the text embedding, and the proposal with the highest similarity is selected as the prediction. This method combines the benefits of a fine-tuned detector with the powerful semantic alignment capabilities of CLIP, enabling effective grounding without requiring a trained matching head.

4. Experimental Settings

We implemented both models using PyTorch and built upon publicly available codebases for Faster R-CNN and CLIP. We used the Faster R-CNN implementation from the torchvision model zoo and the CLIP model from OpenAI's official repository.

For both methods, we used the RefCOCO dataset derived from the MS-COCO images and annotations, where each sample contains an image, a referring expression, and its corresponding bounding box.

For the first method, we fine-tune Faster R-CNN-ResNet-50-FPN (ImageNet pretrained) for 5 epochs using AdamW. The region-proposal network from the same detector is trained, and its box head is replaced by a dynamic CLIP predictor built atop frozen CLIP ViT-B/32 encoders. At inference we keep the top-K RPN boxes, fuse CLIP cosine similarity and RPN objectness with λ and temperature τ, then apply a two-round discrete local search for box refinement.

For the second method, we fine-tuned the Faster R-CNN detector on the RefCOCO dataset using a ResNet-50 backbone pretrained on ImageNet. Region proposals generated by the fine-tuned model were cropped and evaluated using cosine similarity between CLIP image and text embeddings. The CLIP model used was ViT-B/32. The training and inference were conducted in Google Colab using an NVIDIA A100 GPU, with a batch size of 16 used during fine-tuning.

5. Analysis and Discussion

5.1. Performance of Detector-Based Matching

The detector-based matching strategy—swapping the ROI classification head for a custom predictor underperforms on RefCOCO. As Table 1 shows, the baseline yields only 22 % top-1 accuracy and a mean IoU of 0.224. Fine-tuning the new head nudges IoU up to 0.247 but drags accuracy down to 17.7 %, suggesting that reliable text–region alignment remains elusive. One likely bottleneck is proposal quality: the Faster R-CNN backbone was not fully optimized, so many candidate boxes are already off-target. Even so, simply grafting the dynamic CLIP predictor onto imperfect proposals does little; it struggles to reconcile textual cues with the regions it receives.

5.2. Performance of CLIP-Based Scoring

In contrast, the CLIP-based scoring method performed significantly better. Using only CLIP similarity to select among region proposals, the baseline detector reached 41.2% accuracy and 0.407 mean IoU. After fine-tuning the

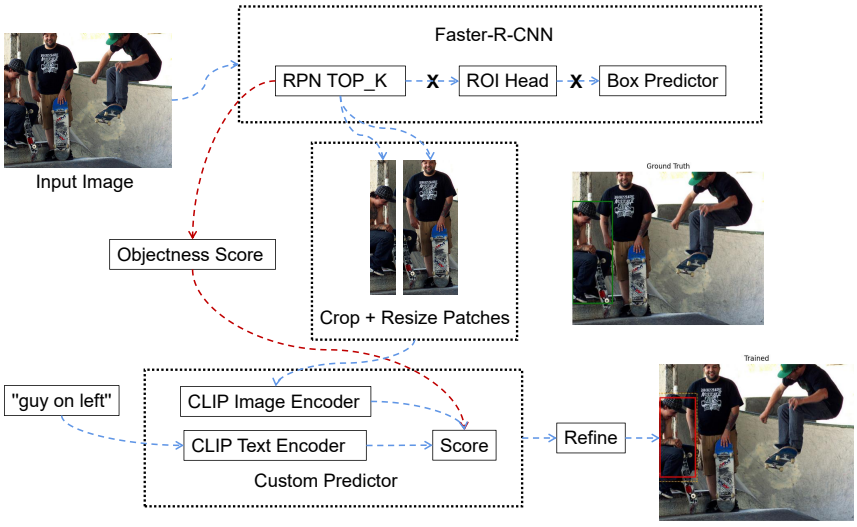

Figure 2. Overview of our CLIP-guided detection approach. We replace Faster R-CNN's box predictor with a CLIP-based scoring module. Region proposals are cropped and compared with a text query using CLIP image and text encoders. The final region is selected based on similarity scores.

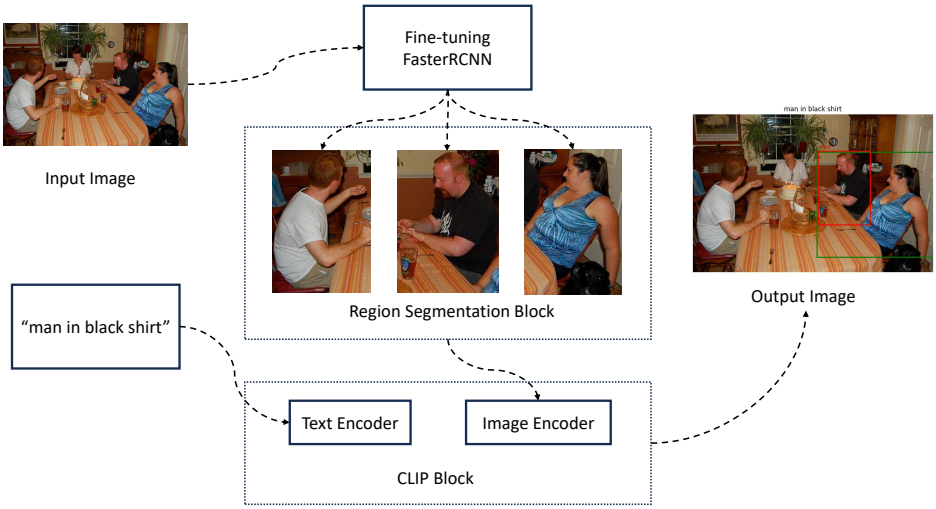

Figure 3. Pipeline of CLIP-based scoring for candidate box selection. Given an image and a text query, Faster R-CNN generates region proposals. CLIP encodes each region and the text, and selects the region with highest similarity. The final output highlights the matched box on the image.

detector for better proposal quality, the performance improved further to 54.7% accuracy and 0.539 mean IoU. These results demonstrate that CLIP's pretrained vision-language alignment can effectively serve as a scoring mechanism for object grounding. Figure 4 illustrates the training loss curve of the fine-tuned detector used in this method.

5.3. Comparative Analysis

Overall, CLIP-based scoring consistently outperformed the detector-based matching method in both accuracy and localization quality. The results suggest that directly leveraging pretrained vision-language embeddings, rather than training a lightweight predictor, may be more effective in REC tasks with limited data.

6. Conclusion

In this project, we investigated two approaches for referring expression comprehension (REC) by integrating region proposals from Faster R-CNN with the visual-language rep-

Method	Model	Train-data	Accuracy	Mean IOU
Detector-Based Matching	Baseline	RefCOCO	0.220	0.224
Detector-Based Matching	Fine-tuning	RefCOCO	0.177	0.247
CLIP-based Scoring	Baseline	RefCOCO	0.412	0.407
CLIP-based Scoring	Fine-tuning	RefCOCO	0.547	0.539

Table 1. Comparison of two approaches on the RefCOCO dataset. CLIP-based scoring outperforms detector-based matching in both accuracy and mean IoU. Fine-tuning the detector further improves results for both methods.

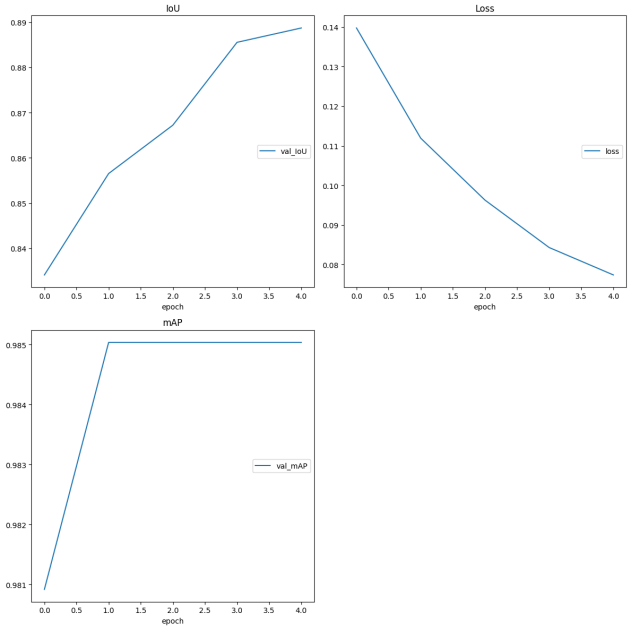

Figure 4. Training loss curve for the fine-tuned Faster R-CNN used in the CLIP-based scoring method.

resentation capabilities of CLIP. One approach modified the detector's classification head to perform text-conditioned prediction, while the other used CLIP to directly score proposals generated by a fine-tuned detector.

Our experiments demonstrate that CLIP-based scoring provides significant advantages in both accuracy and localization performance. While training a text-conditioned head allows tight integration into the detection pipeline, it introduces optimization challenges and underperforms compared to using CLIP's pretrained semantic alignment. Fine-tuning the detector further improved proposal quality and led to the best overall results when combined with CLIP scoring.

This study highlights the effectiveness of combining object detection frameworks with large-scale vision-language models for REC tasks. In future work, we aim to explore joint end-to-end training of both components and experiment with different CLIP architectures. Additionally, incorporating more diverse datasets and evaluating across multiple referring expression benchmarks would provide a broader view of the model's generalization ability.

7. Ethical and Societal Considerations

Our project uses Faster R-CNN and CLIP to match language with image regions. Since both models are trained on general datasets, they may reflect unintended biases or misinterpret certain expressions. We use these methods for research only and recommend human oversight in practical applications. Future work should explore bias mitigation and ensure fair, interpretable results.

References

[1] A. Kamath, M. Singh, N. Carion, G. Synnaeve, F. Massa, A. El-Nouby, N. Goyal, J. Levine, I. Misra, C. Dancette, et al. Mdetr: Modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[2] J. Li, D. Li, C. Xiong, and S. C. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, 2022.

[3] G. Luo, Y. Zhou, X. Sun, L. Cao, C. Wu, C. Deng, and R. Ji. Multi-task collaborative network for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, S. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[5] Y. Rao, W. Zhao, M. Min, J. Lu, and J. Zhou. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[6] S. Yang, G. Yang, Y. Li, and Y. Yu. Cross-modal relationship inference for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[7] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision (ECCV)*, 2016.