

Adaptive Selection based Referring Image Segmentation Supplementary Materials

Anonymous Authors



Figure 1: Visualization of attention maps across the 12 layers of the CLIP ViT-B model, demonstrating how focus areas shift and refine through each layer.

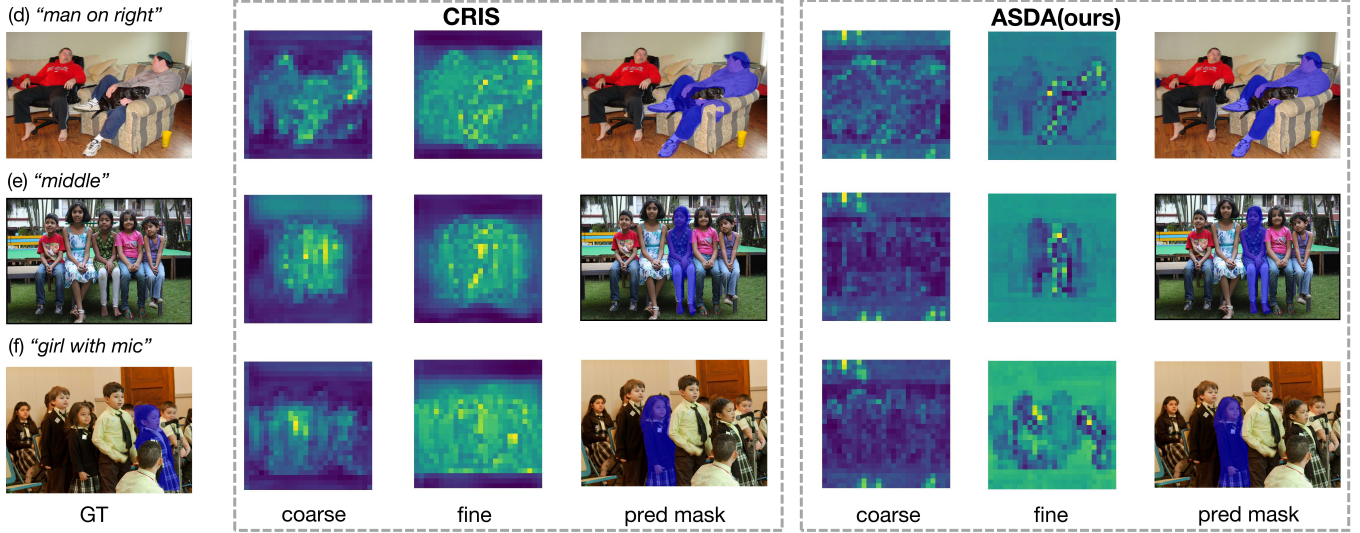


Figure 2: More visualization results of feature maps and final predicted mask in CRIS [1] and ASDA respectively. The coarse feature map represents the features without interaction with word-level language features, while the fine feature map results from the fusion and interaction between word-level features afterwards.

APPENDICES

We provide supplementary information in the following order: erratum in Appendix A, extended experiments in Appendix B, failure cases in Appendix C and additional visualizations in Appendix D.

A ERRATUM

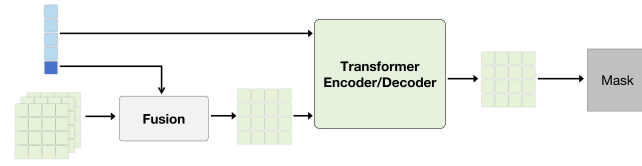
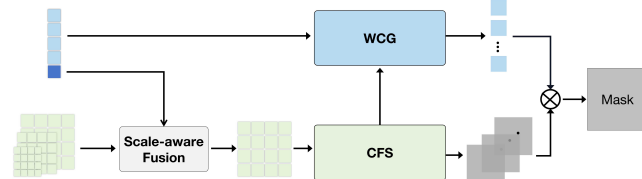
Regarding the Figure 3 included in the submitted paper, there was a typographical error in the spelling of "transformer." It is recommended that readers refer to this correct figure for an accurate understanding of our model's functionality.

B EXTENDED EXPERIMENTS

To assess the effectiveness of our approach under real-world conditions, we perform evaluations using the test split of PhraseCut [2]. While RefCOCO features expressions for only 80 salient object classes, PhraseCut provides a more comprehensive set with 1272 categories in its test set. Regardless of the training dataset, our method ASDA significantly outperformed previous state-of-the-art methods, CRIS [1] and LAVT [3], on the PhraseCut test split. This includes superior performance in categories not seen during training, highlighting ASDA's robust generalization capabilities.

Table 1: phrasecut

| Method | Train dataset | PhraseCut test | |
|------------|---------------|----------------|--------|
| | | all | unseen |
| CRIS | RefCOCO | 15.53 | 13.75 |
| | RefCOCO+ | 16.3 | 14.62 |
| | G-Ref | 16.24 | 13.88 |
| LAVT | RefCOCO | 16.88 | 14.43 |
| | RefCOCO+ | 16.64 | 13.49 |
| | G-Ref | 16.05 | 13.48 |
| ASDA(Ours) | RefCOCO | 20.86 | 18.1 |
| | RefCOCO+ | 20.53 | 17.8 |
| | G-Ref | 20.35 | 17.65 |

(1) single branch**(2) our dual branch****Figure 3: Illustration of single-branch cross-modal alignment in existing RIS methods and our dual alignment that enables linguistic descriptors to directly interact with mask prediction.**

Specifically, when trained on the RefCOCO dataset, ASDA surpasses CRIS by 5.33% and LAVT by 3.98% across all categories respectively, and 4.35% and 3.67% in unseen categories. When trained on other datasets, ASDA exhibited similarly impressive results.

C FAILURE CASES

We illustrate failure cases of our ASDA model in Figure 4. Our model performs well with simple descriptions and in easily distinguishable scenes. However, when descriptions are complex and the scenes contain visually similar objects that are hard to differentiate, we observe that our model can erroneously locate objects.

D ADDITIONAL VISUALIZATIONS

As shown in Figure 1, we visualize the attention maps across the 12 layers of the CLIP ViT-B model using Grad-CAM. This visualization allows us to observe how the focus areas shift and refine through each layer. These insights have inspired us to design an adaptive selection mechanism to selectively focus on language-preferred visual features.

Additionally, we have expanded our analysis to include more visualization results of feature maps and the final predicted masks in both CRIS and ASDA. It is evident that our proposed ASDA's

lounge chair on back left side



a lady sitting in front of a cake with candles on top



bowl behind the others can only see part



Image

GT

Prediction

Figure 4: Failure cases

fine feature map more effectively captures the objects specified in the text. Notably, for item (f) in Figure 2, both CRIS and our ASDA produce incorrect segmentation results due to the presence of two very similar objects in the image. Interestingly, the fine feature map of our ASDA model shows awareness of both confusing objects. Moving forward, we plan to explore more effective mechanisms to address this issue.

REFERENCES

- [1] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. 2022. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11686–11695.
- [2] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhansu Maji. 2020. Phrasecut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10216–10225.
- [3] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. 2022. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18155–18165.