Bridging Vision and Language Encoders:
Parameter-Efficient Tuning for Referring Image Segmentation
Supplemental Material

1. Further Analysis

Effect of Bridger's hidden dim. We list the oIoU results on the RefCOCO test-val set of our proposed method under different dimensions to investigate the impact of Bridger's hyper-parameters on our model's performance, as shown in Table 1. The results indicate that increasing the middle dimension of Bridger leads to minor performance improvements when the dimension is less than or equal to 64. However, a slight decrease is observed when the dimension becomes 128 or larger. These findings demonstrate that our framework is robust to this hyper-parameter.

Dim	Params	oIoU(%)	Pr@0.5	Pr@0.7	Pr@0.9
8	0.22 M	70.55	82.56	70.52	16.60
16	0.45 M	70.78	82.98	70.96	17.01
32	0.92 M	70.92	83.31	71.19	17.42
64	1.94 M	71.06	83.43	72.68	17.40
128	4.28 M	70.39	82.51	71.12	16.99

Table 1: Ablation study of the hidden dimension of Bridger.

Extensibility. In order to further analysis the extensibility of our approach, we have integrated it with a previously established method [1, 2], which contain more details of experiments presented in the paper. As presented in Table 2, the integration of our approach with other methodologies yields a positive impact on the model's performance. This observation serves as evidence of the compatibility of our approach with other methodologies, and underscores the potential for effective integration to further enhance the overall performance of the model.

Method	Trainable Parameters			oIoU(%)	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9
	Backbone	Prompt	Head						
Full-Tuning	120.74 M	0.00 K	23.98 M	70.47	82.62	78.35	71.35	54.47	17.69
Fix Backbone	0.00 M	0.00 K	23.98 M	67.73	79.53	74.42	66.10	46.39	12.41
Adapter [1]	2.39 M	0.00 K	23.98 M	69.46	81.05	76.59	69.69	51.27	16.16
Conv Adapter [2]	1, 20 M	0.00 K	23.98 M	69.33	80.79	76.62	69.80	51.69	15.95
ETRIS (Ours)	1.94 M	0.00 K	23.98 M	71.06	83.43	79.23	72.68	55.39	17.40
Adapter [1]+ETRIS	4.33 M	0.00 K	23.98 M	71.67	84.25	80.12	73.50	56.33	18.43
Conv Adapter [2]+ETRIS	3.14 M	0.00 K	23.98 M	72.11	84.46	80.27	73.92	57.13	18.74

Table 2: Comparison with previous parameter efficient tuning method using Resnet101 as backbone on the oIoU(%) metric on test-val-split of RefCOCO dataset.

Broader Application. We believe that the method could be used for other tasks such as semantic segmentation or non-dense tasks like classification: i) The Bridge architecture facilitates early modal fusion for multi-modal tasks and multi-scale feature aggregation for dense prediction tasks. ii) To achieve this, we propose three transformations: (1) Semantic Segmentation by considering the category name as the text, (2) Object Detection by incorporating an FPN network, and (3) Classification by making minor modifications to the decoder. iii) In practice, for instance, our approach can achieve 88.37% on the visual grounding task on RefCOCO when applying the Bridge to an existing multi-modal detection model (i.e., MDETR). In details, We have added bridgers that connects the visual backbone of MDETR with the text encoder, while fixing the parameters of the dual encoders. Additionally, we have incorporated an FPN (Feature Pyramid Network) to effectively merge feature maps from different stages. The fused feature are then fed forward to the decoding transformer. In anticipation of the future, we aspire to extend the methodology by exploring its applicability to a wider range of tasks, with a particular focus on those in the vision-and-language domain.

2. Limitation

In this section, we conduct failure case analysis to highlight several limitations of this work.

Confusion on visually similar numbers. Figure 1 and Figure 2 presents evidence of erroneous mask predictions produced by our method, which can be attributed to a misinterpretation of digital significance within the image. The model may confuse visually similar numbers, as can be observed from the figure. However, the results also suggest that our method has a certain level of understanding of numerical meanings in both image and text contexts. More results can be seen in Figure 4.

Instability in processing high density of objects. Figure 3 shows our approach's instability of producing a precise mask in scenarios where there is a high density of individuals. In instances where the image contains a multitude of individuals, our approach may yield imprecise mask placement

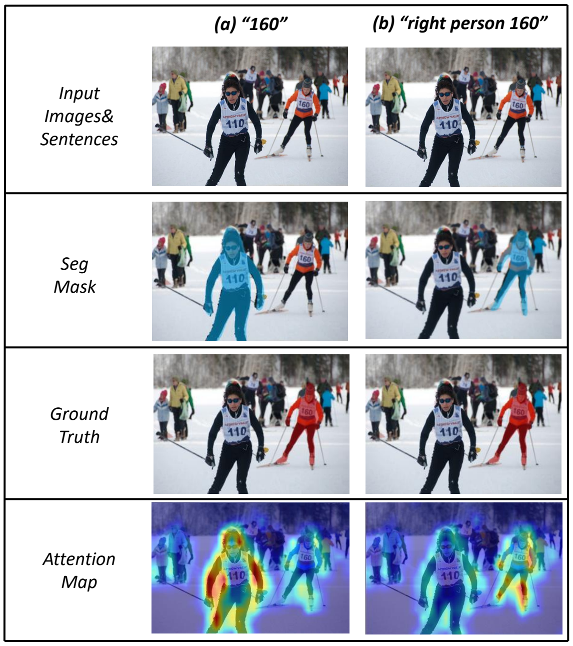

Figure 1: Failure cases when solely describing numbers in the sentences, which show that the model may confuse visually similar numbers.

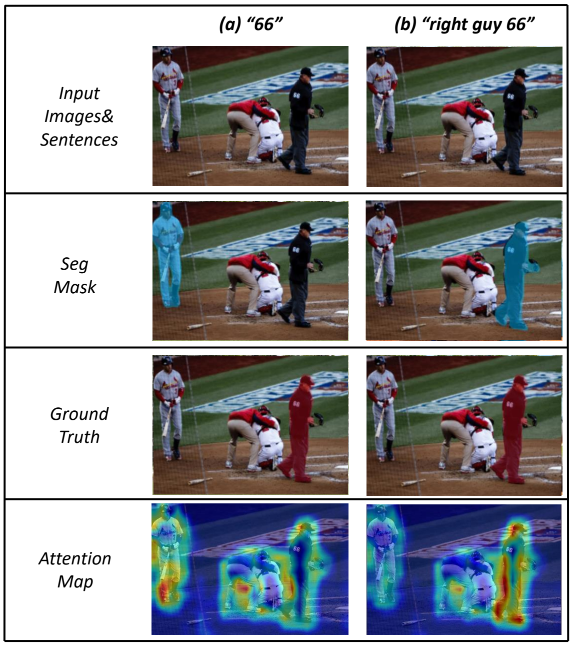

Figure 2: Failure cases when solely describing numbers in the sentences.

during the generation of corresponding masks for occluded persons. More results can be seen in Figure 5.

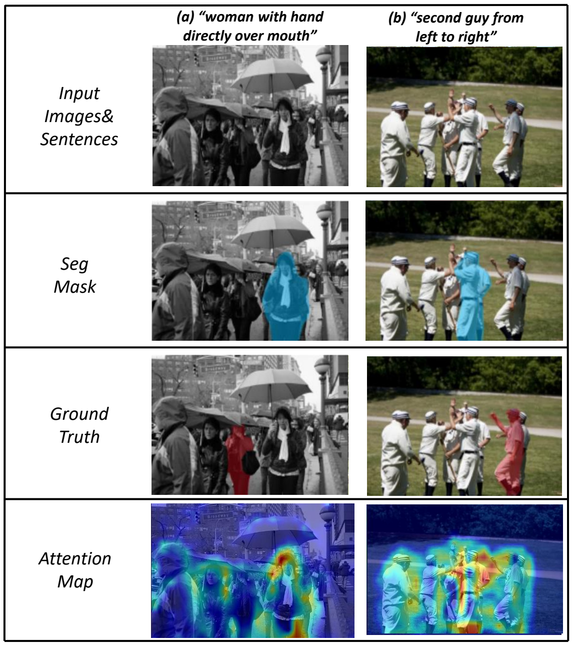
(a) "woman with hand directly over mouth"
(b) "second guy from left to right"
Input Images& Sentences
Seg Mask
Ground Truth
Attention Map

Figure 3: Failure cases when making mask prediction for occluded objects in multi-person scenes.

Given the aforementioned issues, future research endeavors may need to focus on augmenting the model's comprehension of linguistic information and bolstering its resilience to accurately segment occluded objects in a multi-target scene. Such efforts are crucial to improve the efficacy and reliability of computer vision systems, particularly in complex and dynamic environments.

References

[1] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 1

[2] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022. 1

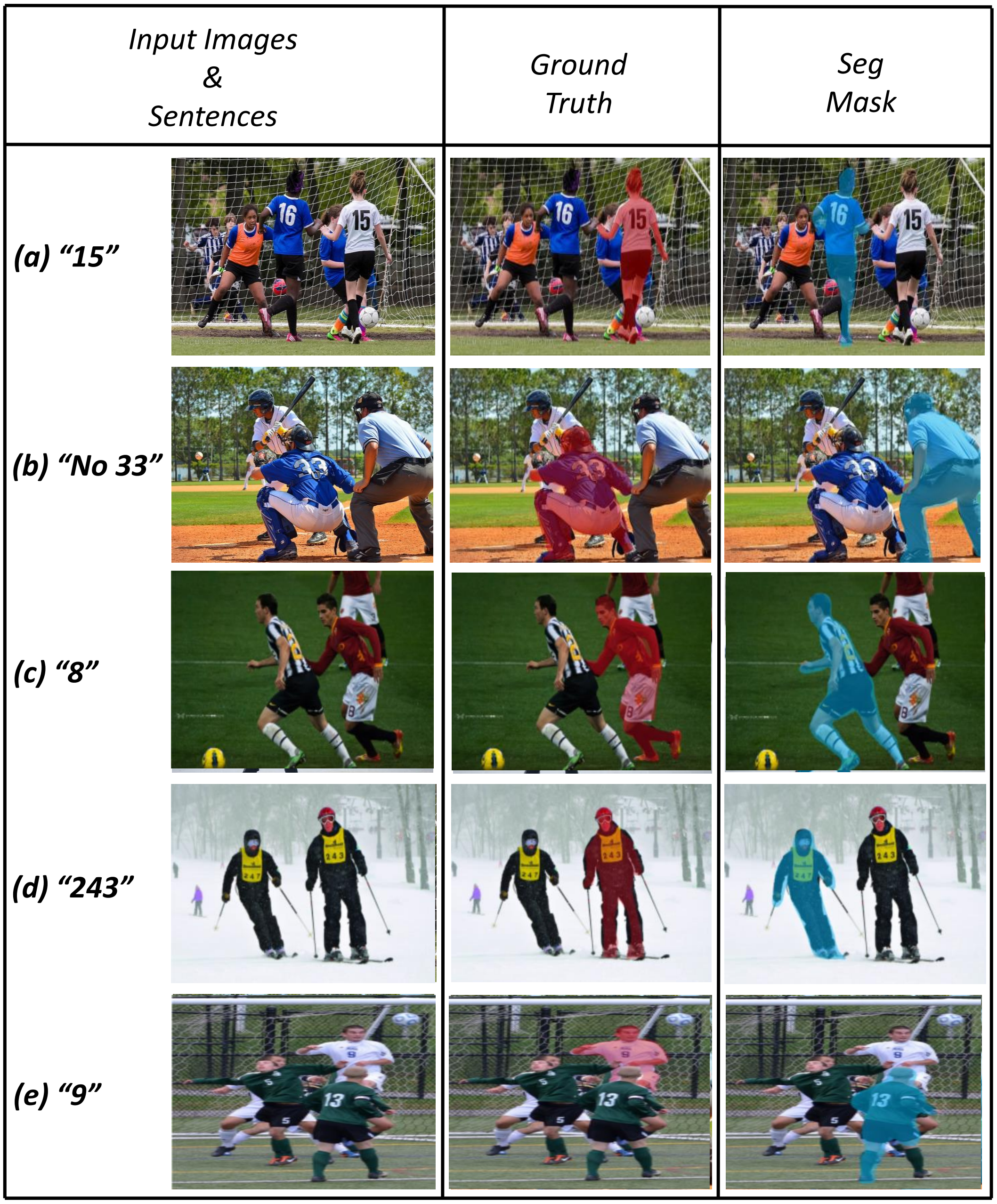

Figure 4: Failure cases when solely describing numbers in the sentences.

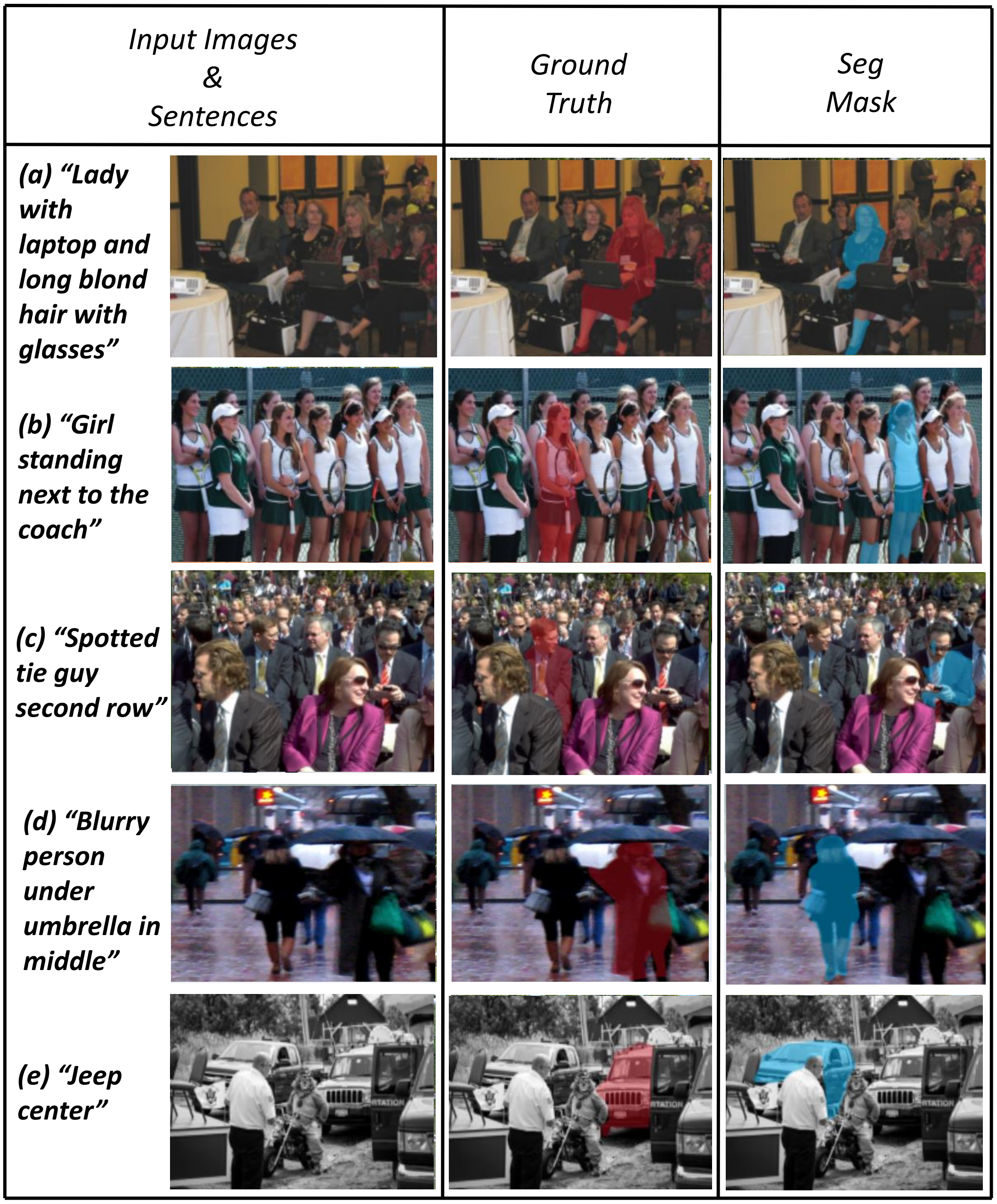
Input Images & Sentences
Ground Truth
Seg Mask
(a) “Lady with laptop and long blond hair with glasses”
(b) “Girl standing next to the coach”
(c) “Spotted tie guy second row”
(d) “Blurry person under umbrella in middle”
(e) “Jeep center”

Figure 5: Failure cases when making mask prediction for occluded objects in multi-person scenes.