Zero-shot Referring Image Segmentation with Global-Local Context Features
- *Supplementary Materials* -

Seonghoon Yu[1] Paul Hongsuck Seo[2] Jeany Son[1]
[1]AI Graduate School, GIST [2]Google Research
seonghoon@gm.gist.ac.kr phseo@google.com jeany@gist.ac.kr

A. Analysis on Global-local Textual Feature

Dataset statistics. The datasets used in our paper, RefCOCO, RefCOCO+ and RefCOCOg, have difference characteristics. As we mentioned in the original manuscript, RefCOCO and RefCOCO+ have shorter expressions and an average of 1.6 nouns and 3.6 words are included in one expression, while RefCOCOg expresses more complex relations with longer sentences and has an average of about 2.8 nouns and 8.4 words. In this supplementary material, we further analyze frequencies with respect to the number or words and nouns in the sentence. As shown in Figure 8, RefCOCOg contains much longer expressions than two other datasets.

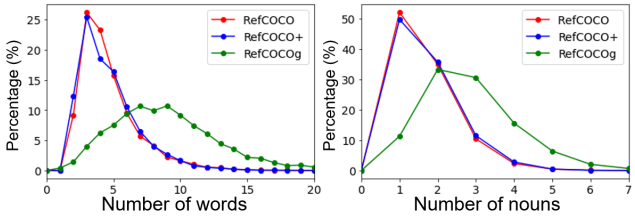


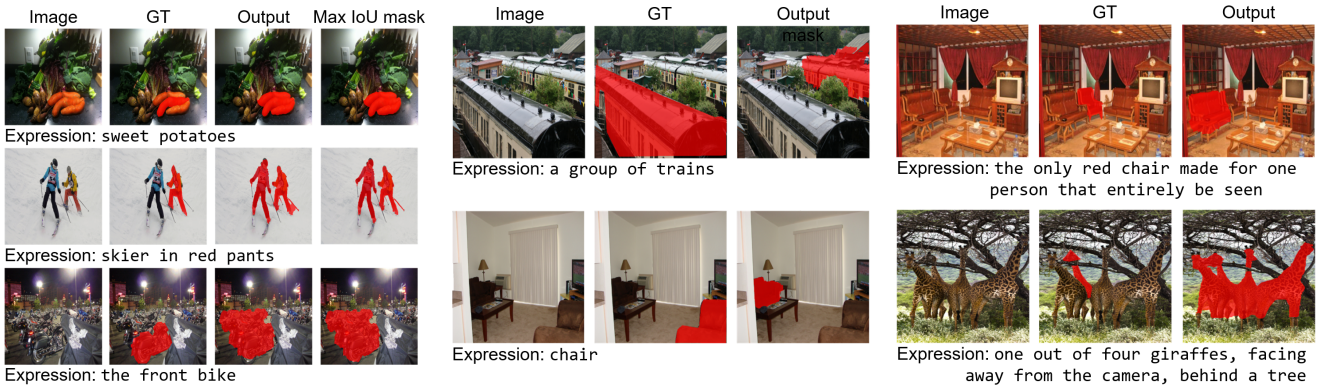
Figure 8. Statistics of datasets. We investigate the number of words and the number of nouns in the sentence on each datasets.

Effects of global-local textual features. Our global-local context textual feature, which is designed to focus on the target noun phrase as well as the whole sentence, is highly effective in longer sentences. To analyze effects of our method with respect to the length of sentences, we show oIoU differences between global-local context textual features and global-context textual features with respect to the length of sentences in Figure 9. It clearly shows that the performances using global-local textual features outperform global-context textual features when the expressions contain more words.

As we illustrated in Figure 4 in the original manuscript, there are no significant difference between global textual features and global-local textual features in RefCOCO and RefCOCO+ datasets. We found that RefCOCO and RefCOCO+ contains shorter sentences where the extracted noun phrase is the same as the sentence. If the target noun phrase is the same as the sentence, our global-local context textual features have no advantage over the global-context one. We measures the percentage of the case that the target noun phrase is the same as the sentence, and there are 48.61% on RefCOCO, 43.43% on RefCOCO+ and 7.46% on RefCOCOg. This explains that the lower performance gains of using our global-local textual features on RefCOCO and RefCOCO+ datasets.

Figure 9. oIoU differences between global-local context textual feature and global context textual feature with respect to length of sentences. We conduct the experiment on RefCOCOg dataset, since this has much longer expressions than two other datasets.

Target noun phrase extraction using SpaCy [1]. To extract local-context textual feature, we need find the target noun phrase in a whole sentence. In this supplementary material, we describe the detailed process of obtaining the target noun phrase using a dependency parsing tool, SpaCy, as follows. SpaCy extracts all noun phrases and the root word of a sentence. The root word, which is also referred to as a head word in SpaCy, is the word that has no dependency with other words, *i.e*. the word that does not have the parent word in the dependency tree. If the root word is a verb, we use the root word's children noun as the root word, and then select the noun phrase containing the root word. If there is no noun phrase containing the root word, we use the whole sentence as the target noun phrase. We show examples of the target noun phrase extracted from the expression in table 4.



(a) Quality of mask proposals (b) Unclear expressions and mask annotations (c) Complicated expressions

Figure 10. Failure cases. (a) Our method depends on the quality of mask proposals. Mask proposals generated by FreeSOLO are sometimes unable to distinguish multiple instances close together. (b) Some labeling noises are included in the datasets. There are multiple possible answers to the given expression. (c) Some expressions are too complex for CLIP to understand.

Table 4. Examples of the target noun phrases in the sentences

Examples:
mom
little girl
near zebra
right sandwich
girl's umbrella
glass of juice in table
yellow baked squash dish
left person with elbow bent
child sitting on womans lap
a cow's ear with a circular tag
flowered quilt on back of couch
a mother giraffe licking her baby
with bruises! okey, closest ugly couch
a black and white dog with pointy ears
that was it ... man in the center up front
the baby boy wearing a red shirt and gray bib
a flat box full of plants labeled wegman's nursery
a man's black tie under all the other ties he is wearing

B. Analysis on hyperparameters α and β.

Our method uses two hyperparmaters, α and β, that combine global and local features for each modality. We present oIoU scores with respect to these parameters for all datasets in Figure 11. For the visual features, we first fix $\beta = 1$ and then choose $\alpha = 0.85$ on RefCOCOg and $\alpha = 0.95$ on RefCOCO and RefCOCO+. With these parameters, we set $\beta = 0.5$ for the text features.

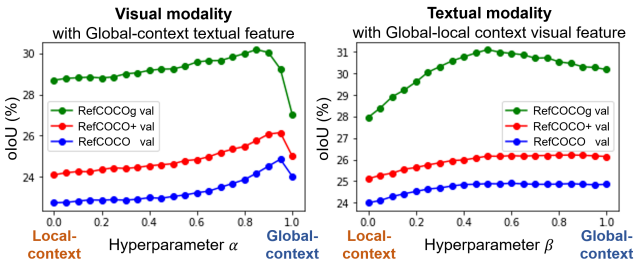


Figure 11. Analysis on hyperparameters α and β for each visual and textual modality.

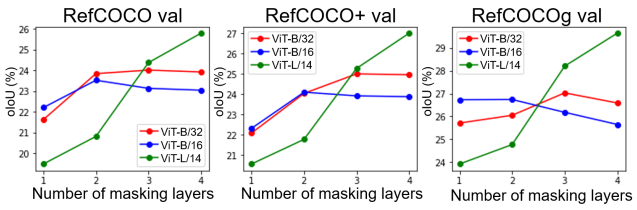


Figure 12. Ablation study on number of masking layers k from the last layer in ViT.

C. Ablation Study on Token Masking in ViT

Token masking is a method for extracting global-context visual features in ViT models. We mask tokens in only the last k Transformer layers to capture global-context of images. Figure 12 shows the oIoU results with respect to the k-th token masking layers from the last layer and the ViT variants of CLIP. We use global-context textual feature to compute the cosine similarity with visual feature and mask proposals from FreeSOLO [2]. We choose $k = 3$ on ViT-B/32, $k = 2$ on ViT-B/16 and $k = 4$ on ViT-L/14, which show the best performances.

Figure 13. Additional qualitative results with different levels of visual features. COCO instance GT masks are used as mask proposals to validate an effect of the global-local context visual features.

D. Additional Qualitative Results

Failure cases. We show failure cases of our Global-Local CLIP in Figure 10. We categorize failure cases into three groups: (a) failure cases due to the quality of FreeSOLO [2]

Figure 14. Additional qualitative results with different levels of textual features using COCO Instance GT mask proposals. Our Global-Local textual feature can predict target objects better than other textual features.

(b) Unclear expression and mask annotations, and (c) too complicated expressions. Since FreeSOLO, which we use to generate mask proposals, is a unsupervised instance segmentation framework, it does not utilize any category labels for training to generate class-agnostic instances masks. Thus it tends to find a cluster of the instances with the same semantic class because it cannot distinguish multiple instances close together. We illustrate the maximum overlapped mask with ground-truth generated by FreeSOLO in the last column of Figure 10a. Moreover, there are a lot of labeling noises in the datasets. Since the task of referring image segmentation is extremely difficult, it is hard to obtain clean annotations. There are multiple possible answers in the image, but the ground-truth mask may contain only one instance as shown in Figure 10b. Furthermore, some expressions are too complex to understand with CLIP, because the text encoder of CLIP is not designed to handle complicated expressions. Therefore our method also has limited ability of understand the complicated natural languages as shown in Figure 10c.

More qualitative results. We demonstrate more results that support the effects of global-local visual and textual context features in Figure 13 and 14. As in Figure 5 and 6 in the original manuscript, we use COCO instance GT masks as mask proposals to show a clear impact of our global-local features. We also illustrate more qualitative results of our method with the several baselines in Figure 15.

Quantitative supports for our qualitative results. We report quantitative results on RefCOCOg that support our qualitative results on the effects of global-local context features in Table 5. To do this, we first compute mask-class accuracy (MC-ACC), the ratio of matching object classes between GT mask and predicted mask. Note that each GT mask is labeled with an object class and the class of a predicted mask is determined by the GT mask with the largest IoU. Then, we compute oIoU on a subset containing only those examples whose predicted mask class is correct; this metric is dubbed as class-conditioned oIoU (CC-oIoU). CC-oIoU captures instance localization performance when predicted mask classes are correct. The results first show that the global feature shows a lower MC-ACC than the local feature. This supports our qualitative finding that global features often select objects of a different class because it is confused by other objects present in the global context. On the other hand, the local feature achieves lower CC-oIoU indicating more incorrect instances due to the lack of global context. In contrast, our global-local feature allows to take advantage of both features and therefore it achieves high MC-ACC and CC-oIoU leading to a significant improvement in the final oIoU.

Table 5. Mask-class accuracy and class-conditioned oIoU with different feature types on RefCOCOg.

Feature type	MC-ACC	CC-oIoU	oIoU
Global	78.90	33.10	27.03
Local	83.36	29.94	25.23
Global-Local	**84.32**	**35.61**	**31.11**

References

[1] Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *EMNLP*, 2015. 1

[2] Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and Jose M Alvarez. Freesolo: Learning to segment objects without annotations. In *CVPR*, 2022. 2, 3

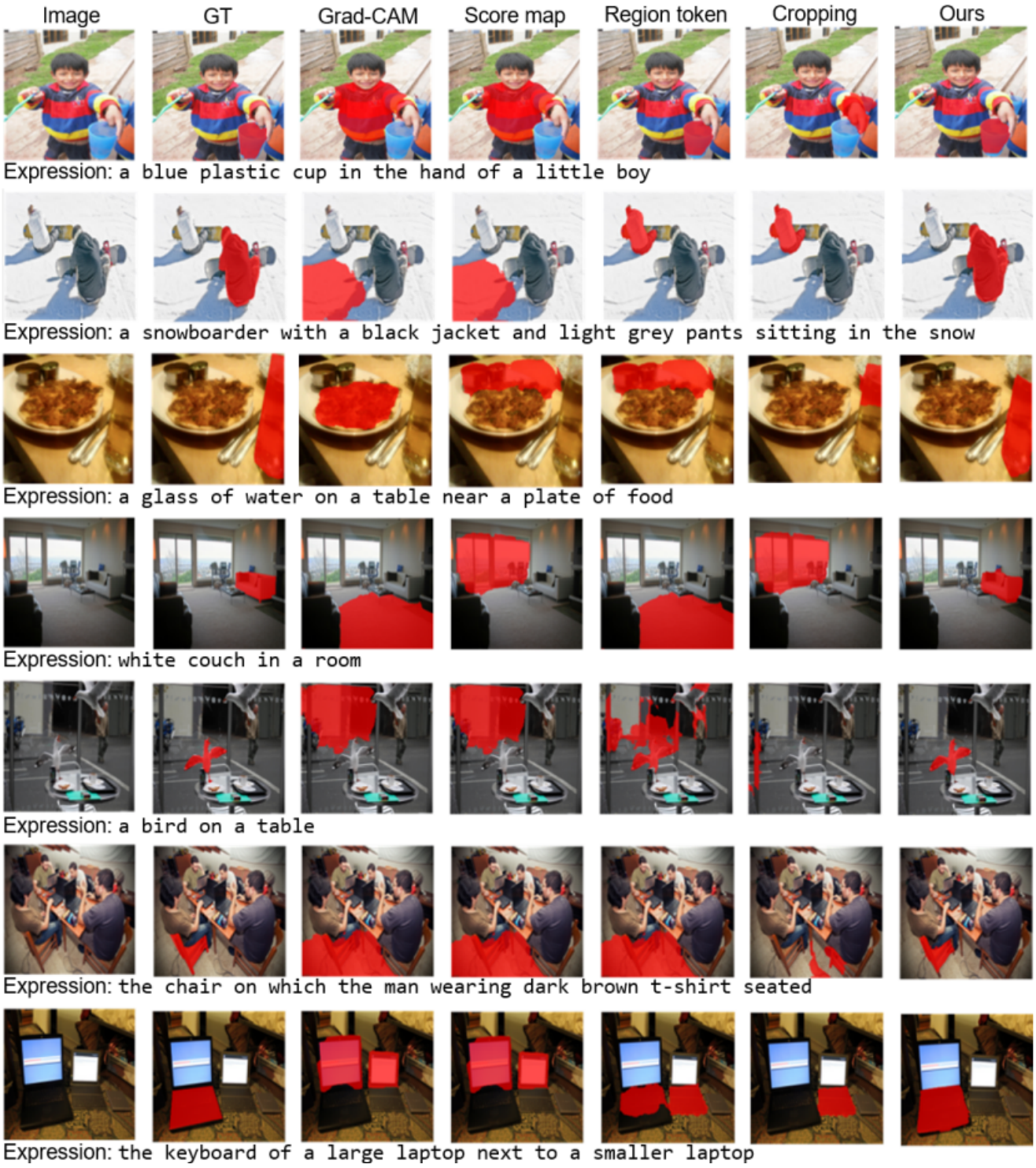


Figure 15. Qualitative results of our method with the several baselines. Note that all methods use mask proposals generated by FreeSOLO.