Unveiling Parts Beyond Objects: Towards Finer-Granularity Referring Expression Segmentation —— Supplementary Material

Wenxuan Wang[1,2,3*] Tongtian Yue[1,2*] Yisi Zhang[4] Longteng Guo[1] Xingjian He[1]
Xinlong Wang[3] Jing Liu[1,2†]
[1] Institute of Automation, Chinese Academy of Sciences (CASIA)
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)
[3] Beijing Academy of Artificial Intelligence (BAAI)
[4] University of Science and Technology Beijing (USTB)
{wangwenxuan2023@ia.ac.cn, wangxinlong@baai.ac.cn, jliu@nlpr.ia.ac.cn}

A. Appendix

In this appendix section, we provide following items:

- (Sec. 1) Potential limitations and future works.
- (Sec. 2) Detailed information about the three benchmark datasets (*i.e*. RefCOCO, RefCOCO+ and RefCOCO+) for classic referring expression segmentation (RES) task.
- (Sec. 3) Implementation details on the three benchmark datasets (*i.e*. RefCOCO, RefCOCO+ and RefCOCO+) for classic RES task and the newly built RefCOCOm benchmark for our multi-granularity RES (MRES) task.
- (Sec. 4) More quantitative analysis about our newly built MRES-32M dataset.
- (Sec. 5) More visualization results of visual comparisons, and the selected samples from our newly built RefCOCOm benchmark and MRES-32M dataset.

A.1. Limitation and Future Work

One potential limitation of this work is that the data scale of our MRES-32M and model capacity of UniRES could be scaled up further to push SOTA performance. Moreover, in this work we mainly focus on the (M)RES task with visual masks as output, which means that although our framework demonstrates the new crucial part-level referring segmentation capabilities, it currently can not produce textual responses and thus can not handle the tasks related to vision-language conversations. However, the integration of large language models could enhance our framework's text comprehension and generation abilities, allowing it to be accordingly adjusted to overcome this limitation. This opens up a future research direction to develop a more general and powerful framework based on multimodal large language models that could interact with user-provided textual and visual inputs across multiple levels of granularity.

A.2. Details about the Benchmark Datasets

RefCOCO [9], stands as one of the largest and frequently utilized datasets derived from MSCOCO [3] for the task of referring expression segmentation. It comprises 142,209 annotated expressions with an average expression length of 3.6 words, labeling 50,000 objects across 19,994 images. The dataset is partitioned into 120,624 training samples, 10,834 validation samples, and two test subsets—test A and test B—containing 5,657 and 5,095 instances, respectively.

RefCOCO+ [9] encompasses 141,564 language expressions with a slightly shorter average expression length of 3.5 words, targeting 49,856 objects within 19,992 images. This dataset follows a similar split as RefCOCO, offering 120,624 training, 10,758 validation, 5,726 test A, and 4,889 test B samples. Unique to RefCOCO+, it omits expressions that use absolute location terms, posing an increased challenge for the classic RES task.

RefCOCOg [4], serves as the third benchmark dataset and contains 104,560 referring expressions, with an significantly longer average length of 8.4 words for 54,822 objects in 26,711 images. The language expressions in this dataset is sourced from Amazon Mechanical Turk, marking a distinction from the previous two datasets. Following previous works, we employ the UMD partition standard [2] for our evaluations in this paper.

A.3. Implementation Details

Experimental Setup. Our work is implemented based on Pytorch [5] and trained with NVIDIA A800 GPUs. Considering the crucial scalability and the ease of implementation, the Vision Transformer [1] is adopted as the image

*Equal technical contribution.
†Corresponding author.

encoder for all the experiments. The text and image encoder are initialized by CLIP [6], while the rest part of model weights are randomly initialized. During training, with 128 and 64 batch size respectively, the AdamW optimizer and a weight decay of 5e-4 are adopted to pre-train and fine-tune our model for 50 and 15 epochs. With a warm-up strategy for 5-epoch pre-training on our MRES-32M and 1-epoch fine-tuning on the specific downstream grounding dataset, the initial learning rate is set to 1e-5 with a cosine decay schedule. Following CRIS [7], due to the extra [SOS] and [EOS] tokens, the input sentences are set with a maximum sentence length of 17 for RefCOCO, our RefCOCOm and RefCOCO+, 22 for RefCOCOg.

During inference, the predicted results by our method is upsampled back to the original image size and binarized with a threshold of 0.35 to the final segmentation result. Any extra post-processing operations or inference tricks can be exploited to further boost the segmentation accuracy but are not included in this work.

A.4. More Analysis about MRES-32M Dataset

Our MRES-32M is composed of 365 object categories and an associated 2,299 part categories. Compared with existing datasets, it covers a wider range of multimodal knowledge and is an important step towards open-world understanding. Among our MRES-32M dataset, the number of referring expressions per objects' category and per parts' category, the word cloud that highlights the head objects' and parts' categories are both presented in Fig. 1.

A.5. More Visualization Results

Visual Comparison with SOTA Methods. In addition, to validate the segmentation quality of our framework, classic RES methods CRIS [7] and LAVT [8], as well as our proposed UniRES are further adopted for qualitative comparison. The provided visualization results in the first and second row of Fig. 2 convinces that our method can better accomplish the classic object-level RES task and generate much better fine-grained segmentation masks of the target objects, greatly reducing the over-segmentation and under-segmentation errors. Similarly, as shown in the third and fourth row in Fig. 2, when facing the more challenging part-level grounding task, our approach can clearly locate and segment the referring target regions more accurately while the other previous methods fail to reach the same level.

Samples of RefCOCOm Benchmark. We have also provided a few more examples in our RefCOCOm benchmark for proposed MRES task in Fig. 3.

Samples of MRES-32M Dataset. A few examples in our newly built MRES-32M dataset for visual grounding task are presented in Fig. 4.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[2] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 108–124. Springer, 2016. 1

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[4] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016. 1

[5] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[7] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 2

[8] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022. 2

[9] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 1

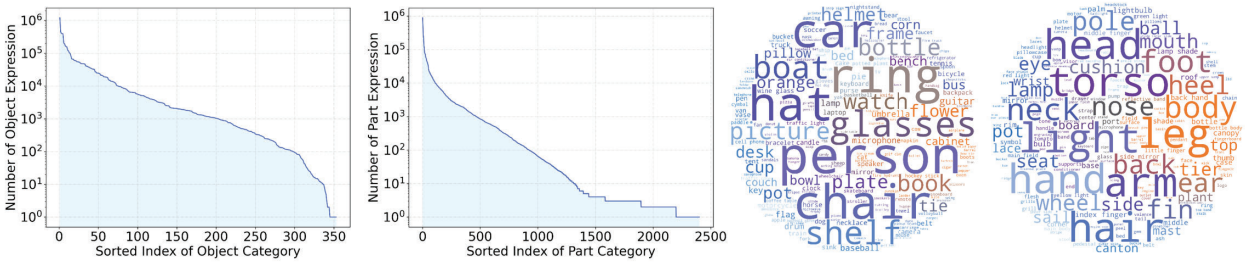

Figure 1. MRES-32M dataset statistics. (a) the number of referring expressions per objects' category in the log scale. (b) the number of referring expressions per parts' category in the log scale. (c) the word cloud highlights the head objects' categories. (d) the word cloud highlights the head parts' categories.

Figure 2. The visual comparison of segmentation results on our RefCOCOm validation set. (a) the input image. (b) CRIS. (c) LAVT. (d) our UniRES. (e) the ground truth.

Figure 3. More selected samples from our proposed RefCOCOm benchmark for multi-granularity RES task.

Figure 4. Selected samples from our built MRES-32M dataset for visual grounding tasks.