CoHD: A Counting-Aware Hierarchical Decoding Framework for Generalized Referring Expression Segmentation

Supplementary Material

1. Additional Details on Experiment Setup

1.1. Datasets

gRefCOCO. It contains 278,232 expressions, which includes 80,022 multiple-target referents and 32,202 empty-target ones. There are 60,287 distinct instances being referred in 19,994 images. The images are split into four subsets: training, validation, test-A, and test-B following the same UNC partition of RefCOCO [6].

Ref-ZOM. Ref-ZOM are selected from COCO dataset [9], which consists of 55,078 images and 74,942 annotated objects. 43,749 images and 58,356 objects are utilized in training, and 11,329 images and 16,586 objects are employed in testing. It is annotated with three different settings, *i.e.*, one-to-zero, one-to-one, one-to-many, each of which corresponds to the empty-target, single-target, and multiple-target in GRES respectively.

R-RefCOCO. There are three different sets in the dataset, R-RefCOCO, R-RefCOCO+, R-RefCOCOg, and each of which is based on the classic RES benchmark, RefCOCO+/g [6]. Only the validation set follows the UNC partition principle and it is officially stated for evaluation. The formulation rule of the dataset is adding negative sentences into the training set at a 1:1 ratio relative to the positive sentences.

RES. RefCOCO [6], RefCOCO+ [6], and RefCOCOg [13] are three standard RES benchmarks, each of which contains 19,994, 19,992, and 26,711 images, with 50,000, 49,856, and 54,822 annotated objects and 142,209, 141,564, and 104,560 annotated expressions, respectively.

1.2. Metrics

For GRES, following [10], we measure the effectiveness of our model by Pr@0.7, gIoU, cIoU and N-acc for gRefCOCO. Meantime, oIoU, mIoU, Acc. used in [4] are adopted for Ref-ZOM. In addition, the standard metrics mIoU, mRR, rIoU are for R-RefCOCO [14]. It is important to note that these metrics are officially specified in each respective benchmark.

Similar to the mean IoU, the Generalized IoU (gIoU) computes the average IoU value for each image across all instances. For the empty-target cases, the IoU values for true positive empty-target instances are considered to be 1 whereas the IoU values for false negative instances are deemed to be 0. cIoU calculates the total intersection pixels over the total union pixels.

mIoU, oIoU is adopted in Ref-ZOM [4], where mIoU is the average IoU value for each image across all cases containing referred objects. oIoU is the same as cIoU. As for R-RefCOCO [14], we use the metric rIoU for quantifying the quality of robust segmentation, which takes the negative sentences into consideration and explicitly assigns the equal weight of the positive one in mIoU calculation. Note that, N-acc. in gRefCOCO, Acc. in Ref-ZOM are in the same formulation where they denote the ratio of the correctly classified empty-target expressions over all the empty-target expressions in the dataset. Similarly, mRR in R-RefCOCO calculates the empty-target expression recognition rate of each image and averages these across the entire dataset.

1.3. Implementation Details

Experiment setup. Our model is implemented with detectron2 [15] in Pytorch. The visual encoder is initialized with the pre-trained weights on ImageNet [1] and the language encoder is an officially pre-trained BERT model [2]. We set the number of Deformable attention layers as 6 following ReLA [10]. There are 3 cascaded semantic decoding modules in HSD for mask generation and query refinement. The weight of $Loss_{mask}$ and $Loss_{count}$ are set as 2 and 0.1 by default. It is worth noting that since $Loss_{exist}$ is trained individually, which has no impact on the main framework, we directly set its weight as 1 for all experiments.

The model is trained with AdamW optimizer with a weight decay of 0.05. The batch size is set to 48. The learning rate is initialed as 2e-5 and scheduled by cosine learning rate decay by default. Following [10, 17], the input images are resized to 480$\times$ 480 and the maximum length of referring expressions is set as 20 for all datasets. Other hyperparameters of the encoding process are the same as ReLA [10]. All experiments are conducted with 8$\times$ A10 and each takes up about one day with 13GB $\sim$ 18GB memory occupied *i.e.*, it depends on the backbone.

Counting labels formulation. We elaborate on the formulation of the counting label. All mentioned datasets adhere to the COCO [9] annotation format. On the one hand, for RES datasets, each annotated expression is accompanied by a label (*category_id*) corresponding to the target category. On the other hand, a list of target categories is incorporated with the additional multi-target scenario for GRES dataset. Consequently, the count of objects can be derived

Method	Backbone	Ref-ZOM Test Set		
		mIoU	oIoU	Acc.
MLLM Methods				
LISA-V-7B [8]	SAM-ViT-H	61.46	60.14	72.58
GSVA-V-7B [16]	SAM-ViT-H	67.98	67.12	82.66
LISA-V-7B [8] (ft)	SAM-ViT-H	65.39	66.41	93.39
GSVA-V-7B [16] (ft)	SAM-ViT-H	68.13	68.29	94.59
Specialist Methods				
MCN [12]	DarkNet-53	54.70	55.03	75.81
CMPC [5]	ResNet-101	55.72	56.19	77.01
VLT [3]	DarkNet-53	60.43	60.21	79.26
LAVT [17]	Swin-B	64.78	64.45	83.11
DMMI [4]	Swin-B	68.21	68.77	87.02
CoHD (Ours)	Swin-B	**69.81**	**68.99**	**93.34**

Table 1. Comparison with state-of-the-art methods on the Ref-ZOM dataset.

λ_{mask}	λ_{count}	gIoU	cIoU	N-acc.	C-acc.
2	0.5	63.23	62.37	54.88	73.28
1	0.5	62.39	61.53	54.81	73.20
1	1	63.06	61.80	55.34	71.88
5	0.1	64.15	62.54	57.24	71.46
2	0.1	**65.89**	**62.95**	**60.95**	**75.45**

Table 2. Results of different ratios of loss.

from the number of categories and each classification label is from the given category of the annotated object. It signifies that the construction of counting ground truth $\mathcal{C}^{gt}$ is straightforward and the additional information extraction is unnecessary. Note that, taking the long-tail distribution of the 80 original COCO categories into consideration where most of the categories are annotated as 0, we instead utilize the 12 super-categories to narrow down the referential search space for providing more precise supervision.

2. Additional Experiments

2.1. Discussion on Inconsistent Performance

As shown in the manuscript, we notice that the performance improvement is inconsistent between RES and GRES. We believe it can be attributed to two folds: 1) ***Restricted effectiveness on the generalized design***: It is obvious that the diversity of referring scenarios in GRES is much more plentiful compared to the RES since RES only includes one-to-one referent case. Due to the inadequate referring semantics between visual and linguistic and the lack of enriched contextual information, *e.g.*, Spatial relationship between instances, counting, or compound structure expressions, it is believed that the great potential of HSD is constrained. Moreover, the effectiveness of object counting mechanism

Supervision Type		gIoU	cIoU	N-acc.
Category	Count			
		63.15	60.67	55.84
✓		64.13	62.36	56.85
✓	✓	65.89	62.95	60.95

Table 3. Effectiveness of category and count-level supervision design of AOC.

Method	Backbone	TFLOPs	Parameters	Inference Time	gIoU
ReLA	Swin-T	0.066T	163M	-	56.87
CoHD	Swin-T	0.068	185M	-	62.95
DMMI	Swin-B	0.392T	341M	95.1 ms/image	62.68
ReLA	Swin-B	0.131T	226M	68.1 ms/image	63.60
CoHD	Swin-B	0.133T	248M	62.5 ms/image	68.42

Table 4. Performance and efficiency comparison with previous SOTA method under different backbones.

in RES is also underestimated. In GRES, it seamlessly integrates all specificities of each scenario by embodying each case into count- and category-level supervision. In terms of RES, the simplification of the formulation limits the potential of object counting, where count number supervision is missing. 2) ***Imbalance dataset scale***: The samples of GRES dataset gRefCOCO is 230,944 samples, while that of each RES dataset is: RefCOCO: 120,624 RefCOCO+: 120,191 RefCOCOg: 80,544. That indicates the specific design and effort on hyper-parameter studies in RES weighs more crucial. Since our main focus is on the generalized referring which accompanies more useful real-world applications, complex hyperparameter studies on RES are not applied.

Considering the hypothesis mentioned above, we believe that enlarging the scale of the dataset can partly alleviate the phenomenon. The results in the joint dataset training [7] demonstrate that our CoHD brings more improvements compared with the single scenario, showing that the potential of our generalized paradigms can be exploited with enriched contextual information.

2.2. Performance on Ref-ZOM Dataset

We report our results on Ref-ZOM benchmark [4] in Tab. 1. As illustrated, our method outperforms all methods under a fair setting, *e.g.*, $+6.3\%$ in Acc., $+1.6\%$ in mIoU. It is worth noting that our method is better than GSVA [16], which utilizes Multi-Modal Large Model (MLLM) [11].

2.3. Additional Ablation Studies

Loss ratio. λ_{mask} and λ_{count} are the coefficients for $Loss_{mask}$ and $Loss_{count}$ respectively. We report the ablation results in Tab. 2. As observed, the appropriate settings of λ_{mask} and λ_{count} help decently integrate the counting ability into hierarchical semantic decoding.

Stacked Layer	gIoU	cIoU	N-acc.
2	66.33	64.58	58.48
3	68.42	65.17	63.68
4	66.70	64.94	60.49

Table 5. Impact on different numbers of stacked layers in the Semantic Decoding Module.

Method	Backbone	gIoU	cIoU	N-acc.
ReLA	Swin-T	56.87	57.73	44.07
ReLA + AOC	Swin-T	60.81	59.34	51.78

Table 6. Compatibility of AOC in previous GRES method.

Aggregation variant	Linear Combination Weights			CoHD
	(1.0, 1.0, 1.0)	(0.5, 0.5, 1.0)	(0.5, 1.0, 0.5)	
gIoU	63.56	64.12	63.66	65.89
cIoU	62.32	62.03	61.53	62.95

Table 7. Impacts of different aggregation methods adopted for Inter-Selection.

Stacked layer in Semantic Decoding Module. We experiment with the impact on the number of stacked layers in the semantic decoding module. As shown in Tab. 5, the insufficient or excessive layers both lead to decreased performance due to incomplete semantic context modeling or over-exaggeration of the redundant contents.

Efficiency *v.s.* performance. We provide detailed comparisons (including model parameters and T-FLOPs.) on the gRefCOCO val set and the results are shown in Tab. 4. It can be seen that our CoHD-B outperforms previous SOTA methods ReLA-B by 4.8% gIoU and 2.7% cIoU with slight parameters increased, *i.e.*, 22M, and even no T-FLOPs costs but faster inference speed. It is worth noting that the performance of CoHD-T is even better than ReLA-B with fewer parameters. In addition, the number of parameters consumed by HSD is 14.7% of the whole model, but with significant performance improvement.

Further discussion on AOC. We provide a further discussion on the effectiveness of AOC as follows: 1) **Supervision type**: We highlight the necessity of the AOC design which embodies each referent scenario into explicit category and count-level supervision in Tab. 3. As indicated, each level of supervision facilitates the precise object perception. 2) **Plug-and-Play**: We also verify the compatibility property of AOC, which can be integrated into the previous GRES method, ReLA, as the plug-in-play module. It can be seen in Tab. 6 that by replacing the misleading binary object existence head with our advanced AOC, the segmentation results can be enhanced. 3) **Strong correspondence to the referent**: The transformation of a given referring expression into the specific supervision type on the semantic query helps enhance the object-awareness by alleviating the impact of implicit and intricate referring scenario such as *"2nd row from bottom 2nd from right"*. The high C-acc. metrics and the precise count prediction reveals that CoHD achieves the category- and count-level understanding of the referent, which eventually benefits the segmentation results.

Further discussion on Inter-Selection. As illustrated in the semantic maps in Fig.3 in the manuscript, the variant of objects in granularity may incur unsatisfied responses at a specific level. It is anticipated that adaptively enhancing the desired regions while suppressing the non-related ones fully unveils the reciprocal benefit of the hierarchical nature. We report the impacts of the variant of aggregation methods in Tab. 7. We speculate that the performance drop potentially stems from the bias accumulation with the fixed weight of each granularity. It is worth noting that the Inter-selection only includes a shared linear layer, where the computation costs can be negligible.

3. Visualizations.

3.1. Multi-granularity Mask Aggregation Visualization

As illustrated in the main paper, we hierarchically aggregate each visual-linguistic correspondence in different granularity for hierarchical semantic decoding. To further prove the rationality of our design, we visualize the aggregated activation map (originally from the semantic map) at each level in Fig. 1. As observed, with progressive mutual modality complementary across granularities, the desired regions of the referent target are fully activated with the increase of the levels.

3.2. Additional Segmentation Comparisons

As illustrated in the main paper, CoHD can better handle the extreme challenges of GRES under multiple/single/empty target scenarios compared with the previous SOTA method ReLA [10]. Here, we incorporate more cases in Fig. 3 to demonstrate our superiority in modeling the intricate referring association between visual and linguistic.

3.3. Segmentation Results of CoHD

Fig. 2 demonstrates the effectiveness of CoHD in complex generalized settings, *e.g.*, complex geometrical relationships between instances, deceptive objects, compound sentence structures, and intricate associations between referring expressions and images.

4. Limitations

Benefiting from the more compatible design, *i.e.*, hierarchical semantic decoding paradigm and the explicit counting ability, our method CoHD sufficiently addresses the limitations of existing GRES and achieves superiority in meeting the challenges of GRES. However, there are some potential limitations. Since the referring sentences in GRES contain multiple-target expressions, the truncation of the input text into 20 may lose detailed descriptions of some of the targets. Although we compensate it with sentence-level textual features, it still remains unsatisfactory in terms of the incompleteness of fine-grained target description. We believe that how to fully utilize the textual expression in GRES is an interesting future direction.

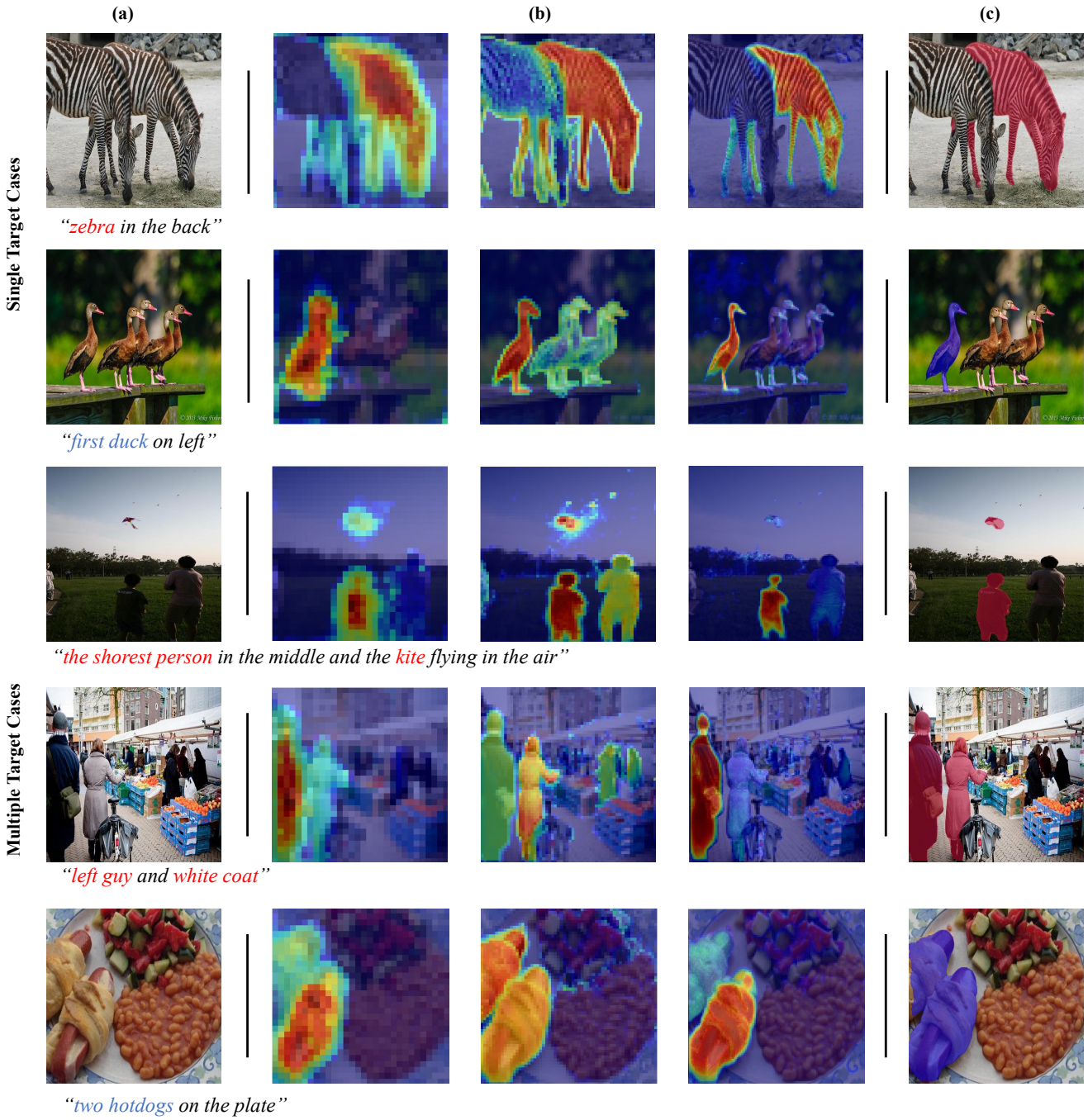

Figure 1. **Visualization in Multi-granularity Mask Aggregation.** (a) and (c) indicate the input image and corresponding segmentation result of our CoHD, respectively. (b) illustrates the aggregated activation map at each granularity.

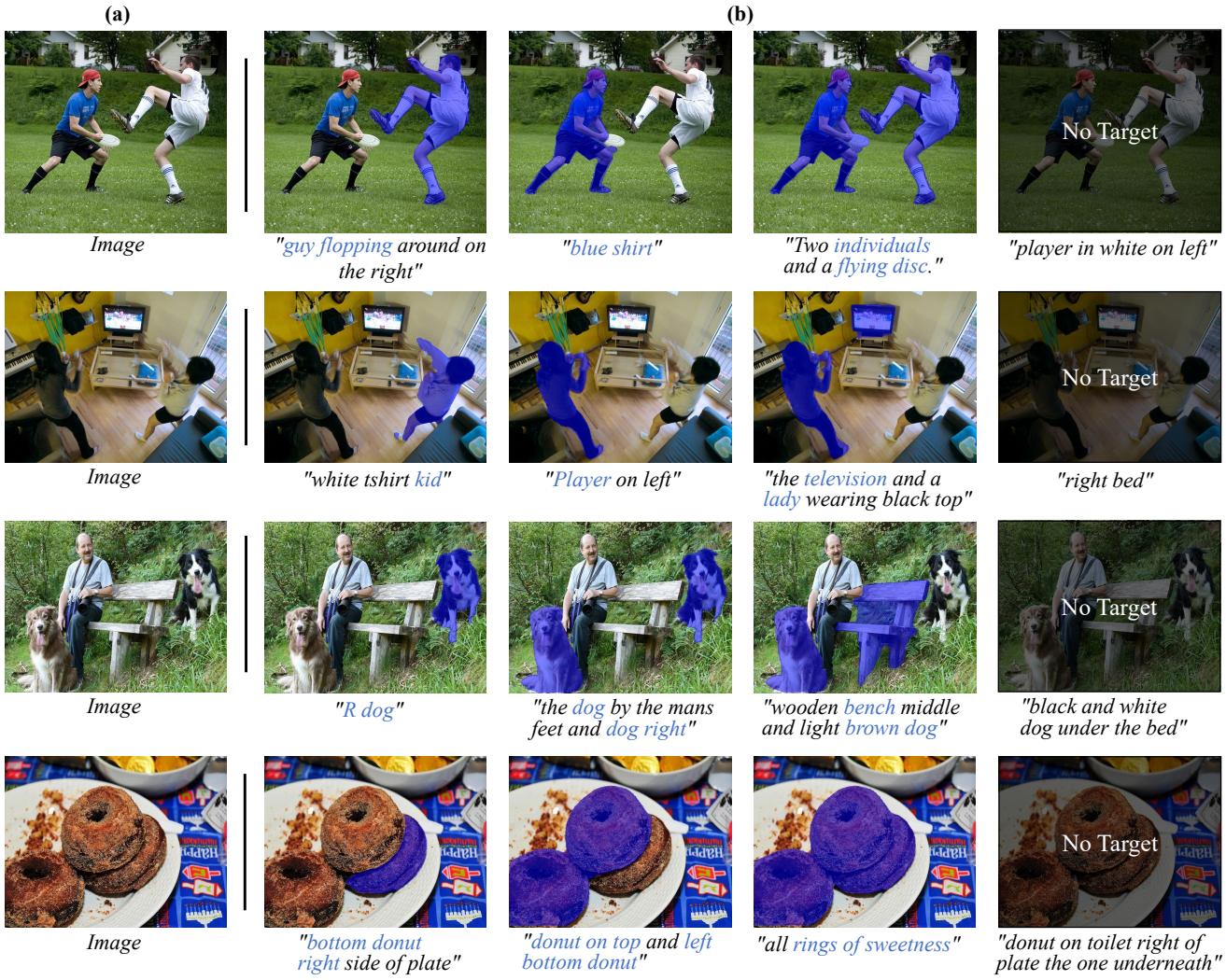

Figure 2. **Segmentation results of CoHD in generalized settings.** (a) denotes the input image, and (b) showcases the segmentation results of CoHD under multiple/single/non-target situations with different referring expressions.

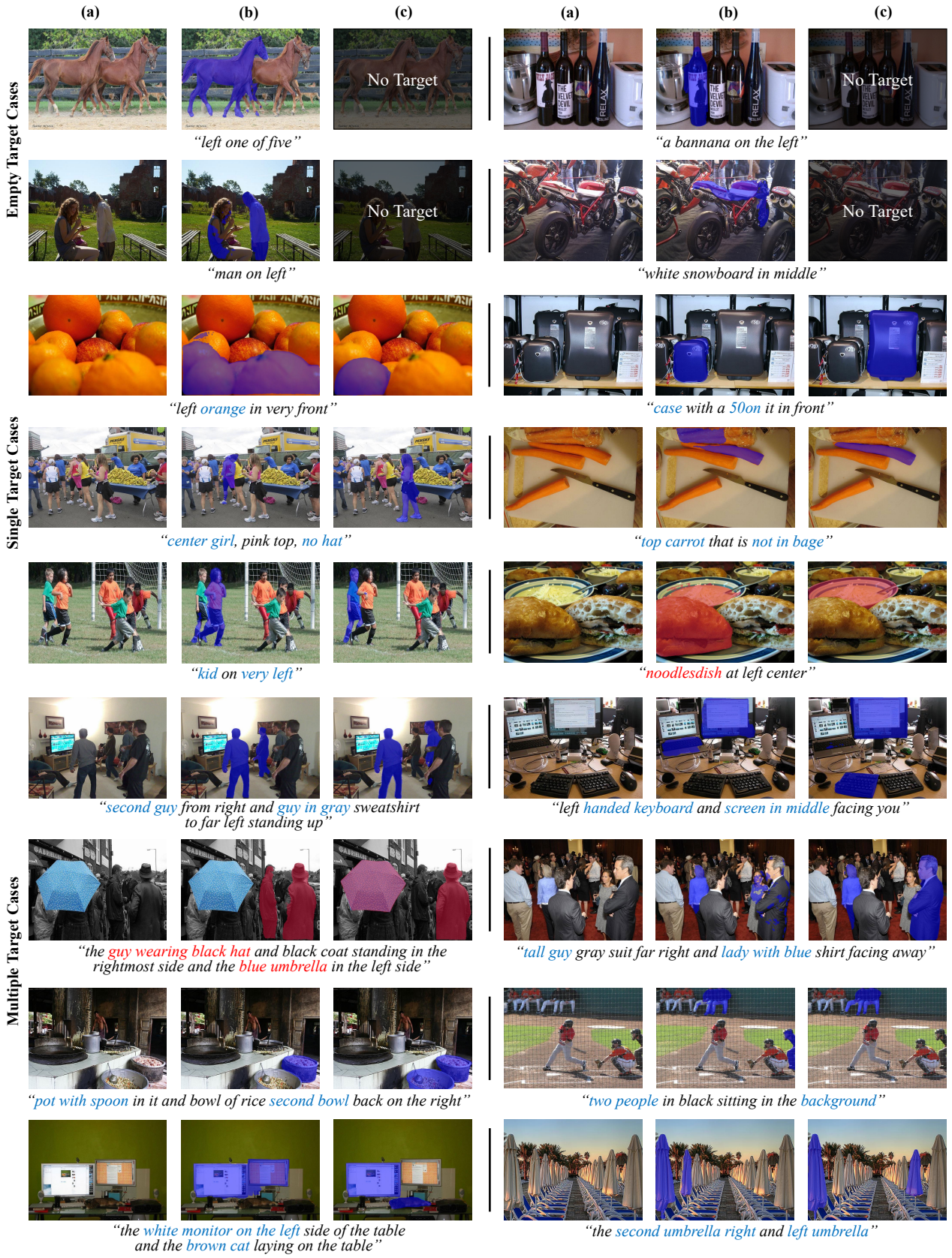

Figure 3. **Segmentation results comparison in the generalized setting.** (a) denotes the input image, (b) and (c) are segmentation results of ReLA and CoHD, respectively. Empty, Single, and Multiple target referent situations with different referring expressions are in top-to-bottom accordance.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 1

[3] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *ICCV*, pages 16301–16310, 2021. 2

[4] Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. Beyond one-to-one: Rethinking the referring image segmentation. In *ICCV*, pages 4044–4054, 2023. 1, 2

[5] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li. Referring image segmentation via cross-modal progressive comprehension. In *CVPR*, pages 10485–10494, 2020. 2

[6] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 1

[7] Seoyeon Kim, Minguk Kang, Dongwon Kim, Jaesik Park, and Suha Kwak. Extending clip's image-text alignment to referring image segmentation. In *NAACL*, pages 4611–4628, 2024. 2

[8] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023. 2

[9] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014. 1

[10] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: generalized referring expression segmentation. In *CVPR*, pages 23592–23601, 2023. 1, 3

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2

[12] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *CVPR*, 2020. 2

[13] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016. 1

[14] Jianzong Wu, Xiangtai Li, Xia Li, Henghui Ding, Yunhai Tong, and Dacheng Tao. Towards robust referring image segmentation. *arXiv preprint arXiv:2209.09554*, 2022. 1

[15] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019. 1

[16] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. *arXiv preprint arXiv:2312.10103*, 2023. 2

[17] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. LAVT: language-aware vision transformer for referring image segmentation. In *CVPR*, pages 18134–18144, 2022. 1, 2