PropVG: Proposal-based Multi-task Collaborative Network for Generalized Visual Grounding

Supplementary Material

A. Datasets

RefCOCO/RefCOCO+ [6] are collected using a two-player game. RefCOCO has 142,209 annotated expressions for 50,000 objects in 19,994 images, and RefCOCO+ consists of 141,564 expressions for 49,856 objects in 19,992 images. These two datasets are split into training, validation, test A and test B sets, where test A contains images of multiple people and test B contains images of multiple instances of all other objects. Compared to RefCOCO, location words are banned from the referring expressions in RefCOCO+, which makes it more challenging.

RefCOCOg [4] is collected on Amazon Mechanical Turk, where workers are asked to write natural language referring expressions for objects. RefCOCOg consists of 85,474 referring expressions for 54,822 objects in 26,711 images. RefCOCOg has longer, more complex expressions (8.4 words on average), while the expressions in RefCOCO and RefCOCO+ are more succinct (3.5 words on average), which makes RefCOCOg particularly challenging. We use the UMD partition for RefCOCOg as it provides both validation and testing sets and there is no overlapping between training and validation images.

gRefCOCO [3] comprises 278,232 expressions, including 80,022 referring to multiple targets and 32,202 to empty targets. It features 60,287 distinct instances across 19,994 images, which are divided into four subsets: training, validation, testA, and testB, following the UNC partition of RefCOCO.

Ref-ZOM [2] is derived from the COCO dataset, consisting of 55,078 images and 74,942 annotated objects. Of these, 43,749 images and 58,356 objects are used for training, while 11,329 images and 16,586 objects are designated for testing. Annotations cover three scenarios: one-to-zero, one-to-one, and one-to-many, corresponding to empty-target, single-target, and multiple-target cases in GRES, respectively.

R-RefCOCO [5] includes three variants: R-RefCOCO, R-RefCOCO+, and R-RefCOCOg, all based on the classic RES benchmark, RefCOCO+/g. Only the validation set adheres to the UNC partition principle, which is officially recognized for evaluation. The dataset formulation incorporates negative sentences into the training set at a 1:1 ratio with positive sentences.

B. Metrics

For GRES [3], we evaluate our model using gIoU, cIoU, and N-acc metrics. For Ref-ZOM [2], we use oIoU and mIoU. R-RefCOCO [5] utilizes mIoU, mRR, and rIoU metrics, as defined in their respective benchmarks. The gIoU is calculated by averaging the IoU across all instances for each image, assigning a value of 1 to true positives in cases of empty targets and 0 to false negatives. The cIoU metric measures the ratio of intersection pixels to union pixels. In Ref-ZOM, mIoU calculates the average IoU for images containing referred objects, while oIoU corresponds to cIoU. For R-RefCOCO, rIoU evaluates segmentation quality, incorporating negative sentences and assigning equal weight to positive instances in the mIoU calculation. N-acc. in gRefCOCO and Acc. in Ref-ZOM both represent the ratio of correctly classified empty-target expressions to the total number of empty-target expressions. Additionally, mRR in R-RefCOCO computes the recognition rate for empty-target expressions, averaged across the dataset.

For GREC [1], we assess the percentage of samples with an F1score of 1, using an IoU threshold of 0.5. A predicted bounding box is considered a true positive (TP) if it overlaps with a ground-truth box with an IoU of at least 0.5. If multiple predictions match, only the one with the highest IoU is counted as TP. Unmatched ground-truth boxes are false negatives (FN), and unmatched predictions are false positives (FP). The F1score for each sample is computed as $\text{F1score} = \frac{2TP}{2TP+FN+FP}$, with a score of 1 indicating a successful prediction. For samples without targets, the F1score is 1 if no predictions are made; otherwise, it is 0.

For REC, we evaluate accuracy based on the grounding results. A predicted region is considered correct if the IoU with the ground truth exceeds 0.5. For RES, we employ mean Intersection over Union (mIoU) between predicted masks and ground truth as the evaluation metric.

C. Additional Methods

C.1. Dataset Construction

To enrich the existing referential datasets with information on foreground objects, we retrieve all corresponding foreground targets from the original COCO dataset based on the `image_id` present in the datasets. Unlike traditional general object detection tasks, our approach focuses solely

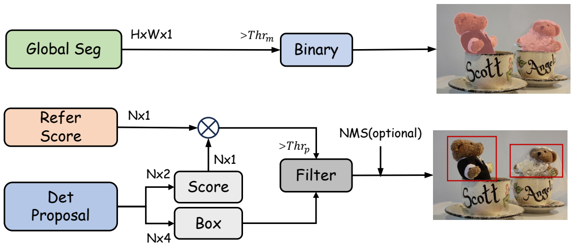


Figure 1. **Post-Process Flowchart.** We combine the Refer Score and Det Score to form the Proposal Referential Score, which is used to filter the referred target via a threshold Thr_p. NMS is optional and not mandatory for DETR-based architectures.

on prominent foreground objects while minimizing the emphasis on weaker or occluded instances. During the dataset construction, we filter out targets in the original COCO dataset where `is_crowd=1`. Additionally, we exclude objects with an absolute area smaller than 100 pixels and retain only those with a relative area greater than 0.05 and less than 0.8. This selection process ensures that the objects in our dataset consist exclusively of significant foreground targets. The primary motivation behind these choices is to accelerate the model's convergence while directing attention away from fine-grained small objects. Instead, we aim to enhance the model's focus on the understanding of referential relationships between text and images.

C.2. Post-process

Post-Processing Procedure. As illustrated in Fig. 1, the post-processing consists of three components: Global Segmentation, Refer Score, and Det Proposal. The Det Proposal branch further decomposes into Det Score and Det Box. For the segmentation branch, we directly apply a threshold Thr_m to binarize each pixel. In the target referential part, we combine the Refer Score with the Det Score, as a valid referred target should possess both high detection confidence and high referential confidence. We filter the combined score using a threshold Thr_p. NMS is optional in this process and not mandatory for DETR-based architectures.

D. Additional Implementation Details

For GVG tasks (GREC, GRES), images are resized to 320 × 320, and models are trained for 12 epochs. For CVG tasks (REC, RES), images are resized to 384 × 384, with training for 30 epochs. We use a unified batch size of 16 and adopt the Adam optimizer. All experiments are conducted on four NVIDIA 4090 GPUs, without using Exponential Moving Average (EMA). The initial learning rate is 5×10^{-5} for the multi-modality encoder and 5×10^{-4} for other parameters. The learning rate decays by a factor of 0.1 at the 7th and 11th epochs. All ablation studies are conducted at a resolution of 224 × 224 and trained for 10 epochs.

E. Additional Ablation Studies

E.1. Impact of the Hyperparameter K in TAS

We analyze the effect of different K values on key metrics, including N-acc., T-acc., F1score, and gIoU. The choice of K directly influences the performance of the model by controlling the number of top-scored segmentation pixels considered by S_{exist}. As shown in Table 1, the value of K determines the extent to which segmentation results contribute to the final target existence score. A higher K suppresses the existence confidence, leading to a reduction in the overall refer score. This intuitively improves N-acc. but decreases T-acc. A crucial aspect to discuss is the optimal number of segmentation pixels to use for weighting the existence score to maximize improvements in F1score and gIoU. Experimental results indicate that when $K = 250$, the framework achieves peak performance across these metrics.

E.2. Impact of the Other Hyperparameter

Table 1 analyzes the impact of varying individual loss weights. First, increasing the weight of $\mathcal{L}_{det}$ (ID 5) improves F1score (69.6) but slightly reduces gIoU, suggesting better detection confidence at the cost of segmentation quality. Second, tuning $\mathcal{L}_{exist}$ (ID 2 vs. ID 1) mainly affects N-acc, with higher weights enhancing classification accuracy but harming F1score. Lastly, adjusting $\mathcal{L}_{ref}$ (ID 6, 7) influences all metrics simultaneously, reflecting its central role in balancing detection, grounding, and segmentation. Overall, the default setting (ID 1) provides the best trade-off.

ID	$\mathcal{L}_{det}$:$\mathcal{L}_{exist}$:$\mathcal{L}_{ref}$	F1score	N-acc	gIoU
1	0.1:0.2:1.0	69.2	71.0	69.9
2	0.1:0.5:1.0	68.4	71.2	69.5
3	0.1:0.0:1.0	67.9	67.8	68.2
4	0.05:0.2:1.0	68.3	71.1	70.4
5	0.2:0.2:1.0	69.6	71.0	68.0
6	0.1:0.2:0.5	69.0	70.3	68.8
7	0.1:0.2:2.0	69.1	70.7	69.8

Table 1. Ablation study on the loss weights in Eq. **??**.

E.3. Impact of the Post-Process

In this section, we analyze two aspects: different scoring strategies and varying post-processing thresholds. For scoring strategies, we compare three methods: using the refer branch score directly, multiplying the refer and detection branch scores, and taking the average of these two scores. As shown in Table 2, the direct use of the refer branch score yields the best results, as this score more accurately reflects the importance of the target. For post-processing,

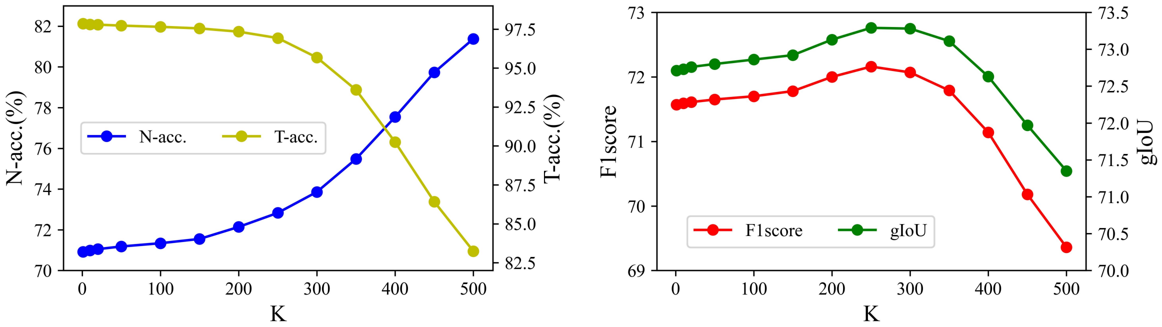


Figure 2. Impact of the Hyperparameter K in TopK Average Scoring (TAS).

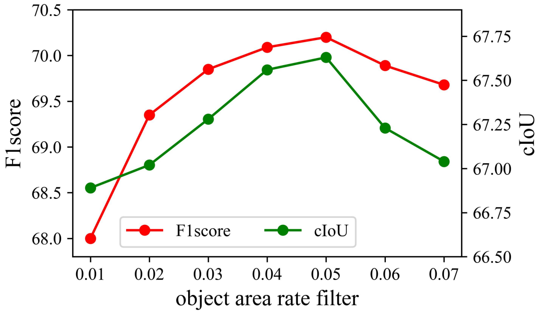


Figure 3. Impact of the lower limit on the relative area of the filtered objects.

we evaluate the performance under different threshold values Thr_p. As presented in Table 3, the model performs best when $Thr_p = 0.9$.

E.4. Impact of Foreground Object Filter

In the training process of the foreground detection branch, we establish a relative area constraint to prevent the model from overly focusing on small targets, which could weaken the optimization effects of other branches. Specifically, we set an upper limit of 0.8 for the relative area of foreground targets and a minimum absolute area of 100. We focus on examining the impact of the lower limit R_{low} on model performance. As shown in Fig. 3, setting R_{low} too low introduces numerous small foreground targets into the training set, leading the model to overemphasize these weaker targets and detracting from the core referential task. Conversely, setting R_{low} too high may result in the neglect of some smaller referential targets during training. Experimental validation indicates that when R_{low} is set to 0.05, the model achieves optimal performance across all evaluation metrics.

F. Analysis

Method	F1score	N-acc.	gIoU	cIoU
S_{refer}	71.59	70.99	72.73	69.18
$S_{\text{refer}} \times S_{\text{det}}$	71.38	72.01	72.83	69.06
Avg($S_{\text{refer}}, S_{\text{det}}$)	71.87	68.55	71.80	68.78
S_{refer} + NMS	71.65	71.34	72.86	69.20

Table 2. Ablation study on post-processing details. The first three experiments compare different scoring strategies for target referencing: S_{refer} refers to using the score from the refer branch directly; $S_{\text{refer}} \times S_{\text{det}}$ denotes the product of scores from the refer and detection branches; and Avg takes the average of the two scores. The final experiment applies Non-Maximum Suppression.

Thr_p	F1score	N-acc.	gIoU	cIoU
0.5	65.14	68.02	66.86	63.91
0.6	65.87	69.19	68.16	65.30
0.7	67.44	69.47	69.10	65.77
0.8	67.98	**70.44**	69.59	66.22
0.9	**68.81**	70.39	**69.85**	**66.24**

Table 3. Effectiveness of Thr_p in post-process.

F.1. Foreground Supervision Analysis

Fig. 4 provides a visual comparison between two regression strategies: direct referring and our proposal-driven referring approach. In the latter, the model first generate the foreground target before subsequently discriminate the proposal's referentiality. For instance, using text (a) as a case study, the direct referring method erroneously identifies a non-referent target with high confidence. In contrast, our DeRIS effectively diminishes the confidence assigned to non-target instances, thereby reducing false positives. Experimental results indicate that the integration of foreground supervision enhances the model's ability to differentiate between foreground and referent ones, leading to a marked reduction in false positives and a substantial improvement in generalization performance.

F.2. Proposal Analysis

By examining the foreground outputs generated for the same image under varying textual descriptions, we observe

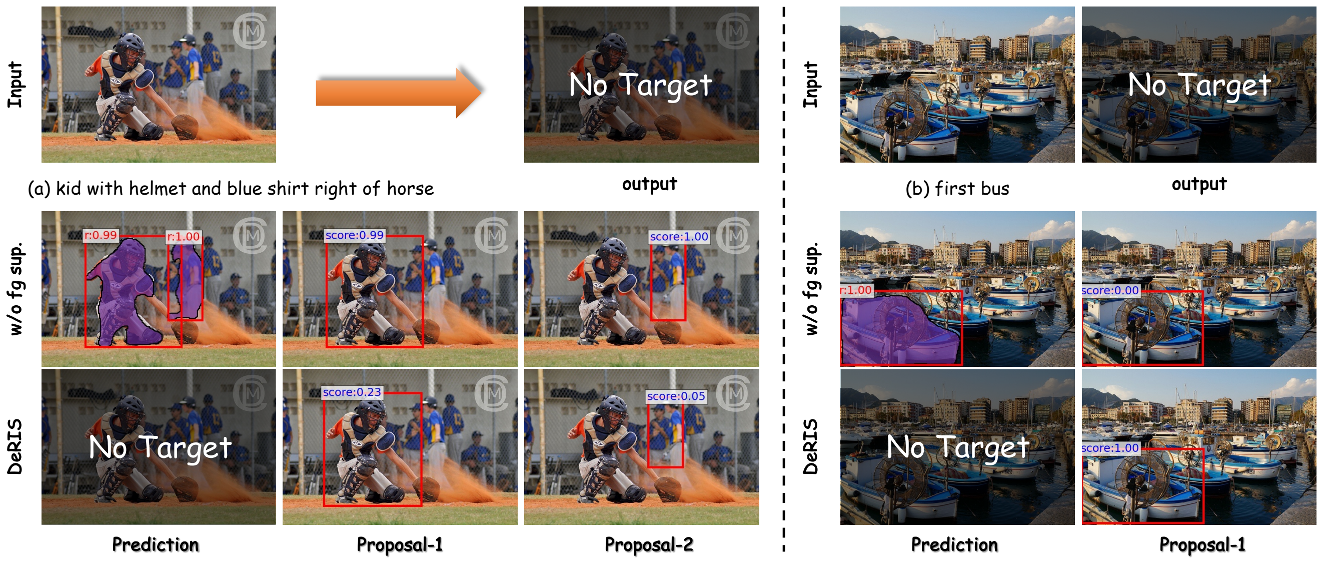


Figure 4. Illustration the effect of incorporating foreground supervision. *w/o fg sup.* denotes the setting where only the referred objects are used as supervision targets, potentially overlooking useful foreground cues.

that the textual input exerts a substantial influence on the generation of foreground targets. Specifically, the model prioritizes targets explicitly referenced in the text as foreground objects. For example, in Fig. 5 (3), where the description 'a guy in an orange tie' is provided, the model accurately identifies and emphasizes the corresponding target as the foreground object. In contrast, in Fig. 5 (2), where no such description is given, the same target is not classified as foreground. Similarly, in Fig. 5 (4), the instruction 'all human beings' prompts the model to designate all humans as foreground targets, including subtle details such as the smaller hand of a person on the left, which is overlooked under alternative descriptions.

Further examples reinforce this trend. In Fig. 5 (5), the phrase 'all of the small pastries' leads the model to detect every pastry in the image as a foreground object, whereas texts omitting 'pastries' exclude them from consideration. In Fig. 5 (7), the model focuses exclusively on objects on the 'left' side, as dictated by the text. In cases such as Fig. 5 (8) and (9), where inclusive terms like 'all' or 'everyone' are used, the model identifies all visible targets as foreground objects. Conversely, as illustrated in Fig. 5 (10), when the text description is misaligned with the image content, the model suppresses foreground generation entirely.

These findings underscore the pivotal role of multimodal understanding approaches, such as BEiT-3, in driving these behaviors. In such frameworks, the interaction between image and text features initiates at the encoding stage, enabling the model to concentrate attention on objects that align with the textual description when a specific reference is provided. This early-stage interplay allows textual information to directly modulate the representation of image features, amplifying responses in regions pertinent to the description while attenuating those in unrelated areas. Consequently, this mechanism facilitates precise and contextually targeted foreground generation.

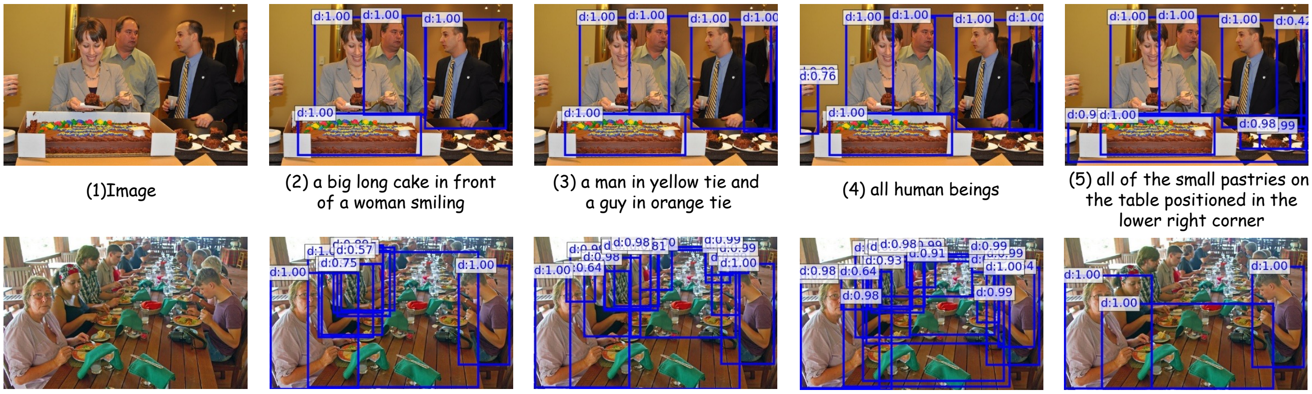
G. Additional Visualization

G.1. Query Visualization

In Fig. 6, we visualize the results of 10 queries per sample, including detection boxes, detection scores, and referential scores. Additionally, we present the corresponding ground truth, predicted proposals, predicted referential targets, and referential masks.

G.2. Detail Visualization

We visualize the detection and segmentation results on multiple datasets, including grefCOCO, RefCOCO/+/, R-RefCOCO/+/, and Ref-ZOM. In Fig. 7, we present the detection and segmentation performance of PropVG on the standard RefCOCO dataset. Fig. 8 demonstrates the model's ability to extract foreground bounding boxes and resolve referential expressions. Fig. 9 highlights PropVG's performance in detection and segmentation under challenging scenarios, showcasing its robustness. Finally, Fig. 10 illustrates the referential capability of PropVG on the Ref-ZOM dataset, including both standard and multi-referent cases. This also demonstrates the model's enhanced detection robustness enabled by its foreground extraction capability.



(1)Image

(2) a big long cake in front of a woman smiling

(3) a man in yellow tie and a guy in orange tie

(4) all human beings

(5) all of the small pastries on the table positioned in the lower right corner

(6) Image

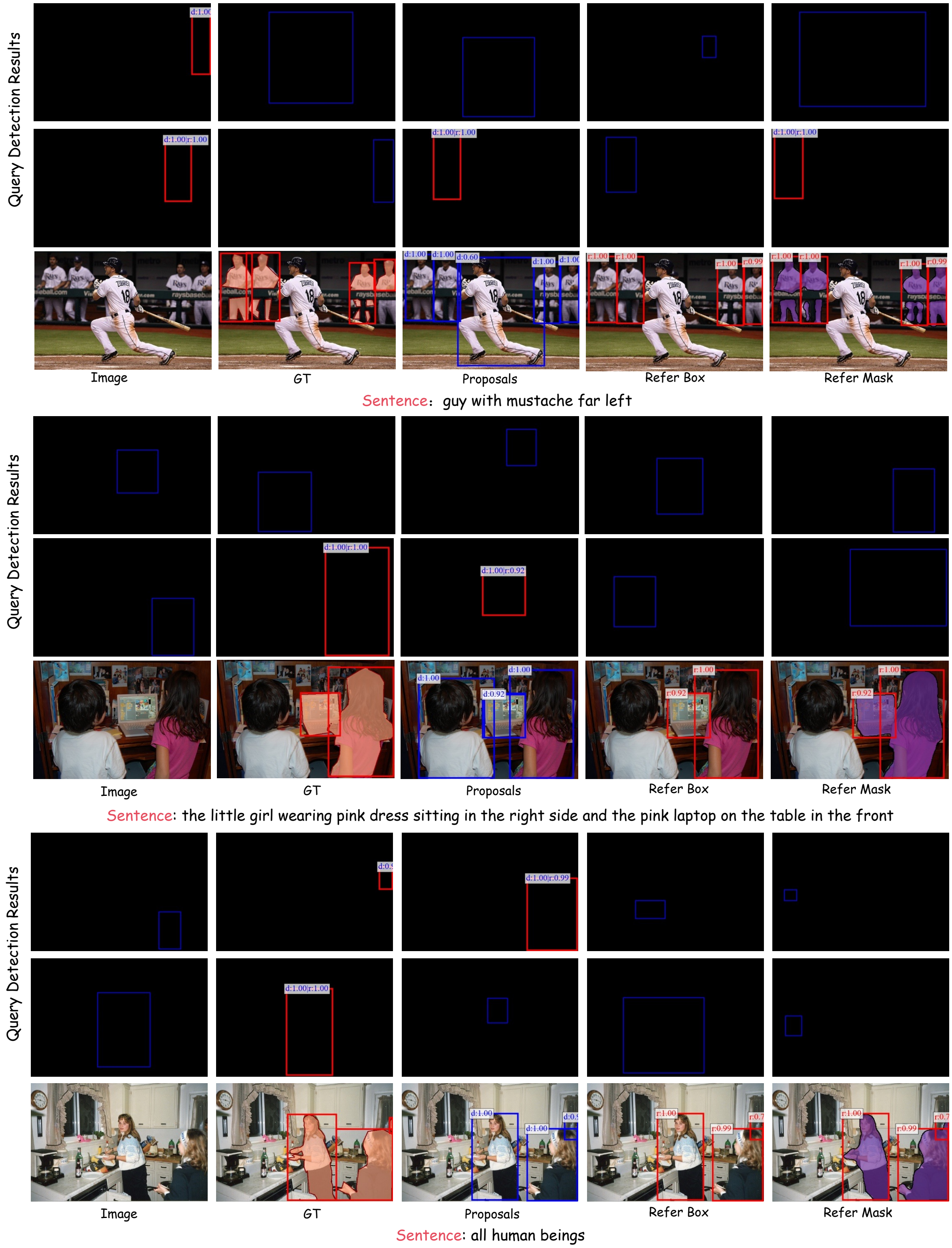
(7) the arrangement of individuals on the left pred

(8) everyone in the image

(9) all beverage holders

(10) table top bottom of pic

Figure 5. Illustration of foreground object prediction results under different texts.

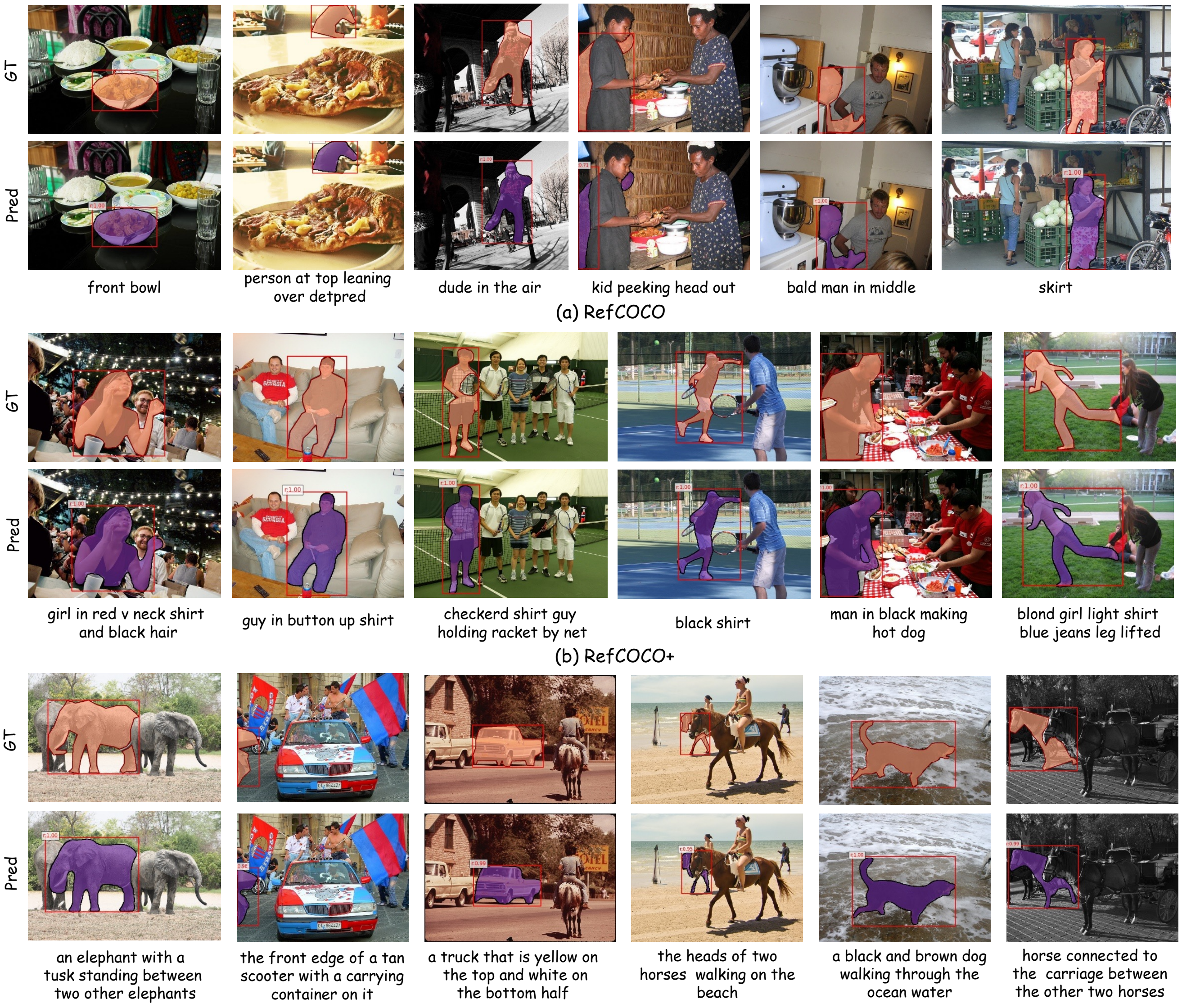


Sentence: guy with mustache far left

Sentence: the little girl wearing pink dress sitting in the right side and the pink laptop on the table in the front

Sentence: all human beings

Figure 6. Visualization of object queries with corresponding detection boxes, detection scores, and referential scores. Red boxes indicate correctly referred objects, while blue boxes represent objects with low referential confidence.


GT
Pred
front bowl
person at top leaning over detpred
dude in the air
kid peeking head out
bald man in middle
skirt

(a) RefCOCO

GT
Pred
girl in red v neck shirt and black hair
guy in button up shirt
checkerd shirt guy holding racket by net
black shirt
man in black making hot dog
blond girl light shirt blue jeans leg lifted

(b) RefCOCO+

GT
Pred
an elephant with a tusk standing between two other elephants
the front edge of a tan scooter with a carrying container on it
a truck that is yellow on the top and white on the bottom half
the heads of two horses walking on the beach
a black and brown dog walking through the ocean water
horse connected to the carriage between the other two horses

(c) RefCOCOg

Figure 7. **Visualization of RefCOCO/+/g dataset.**

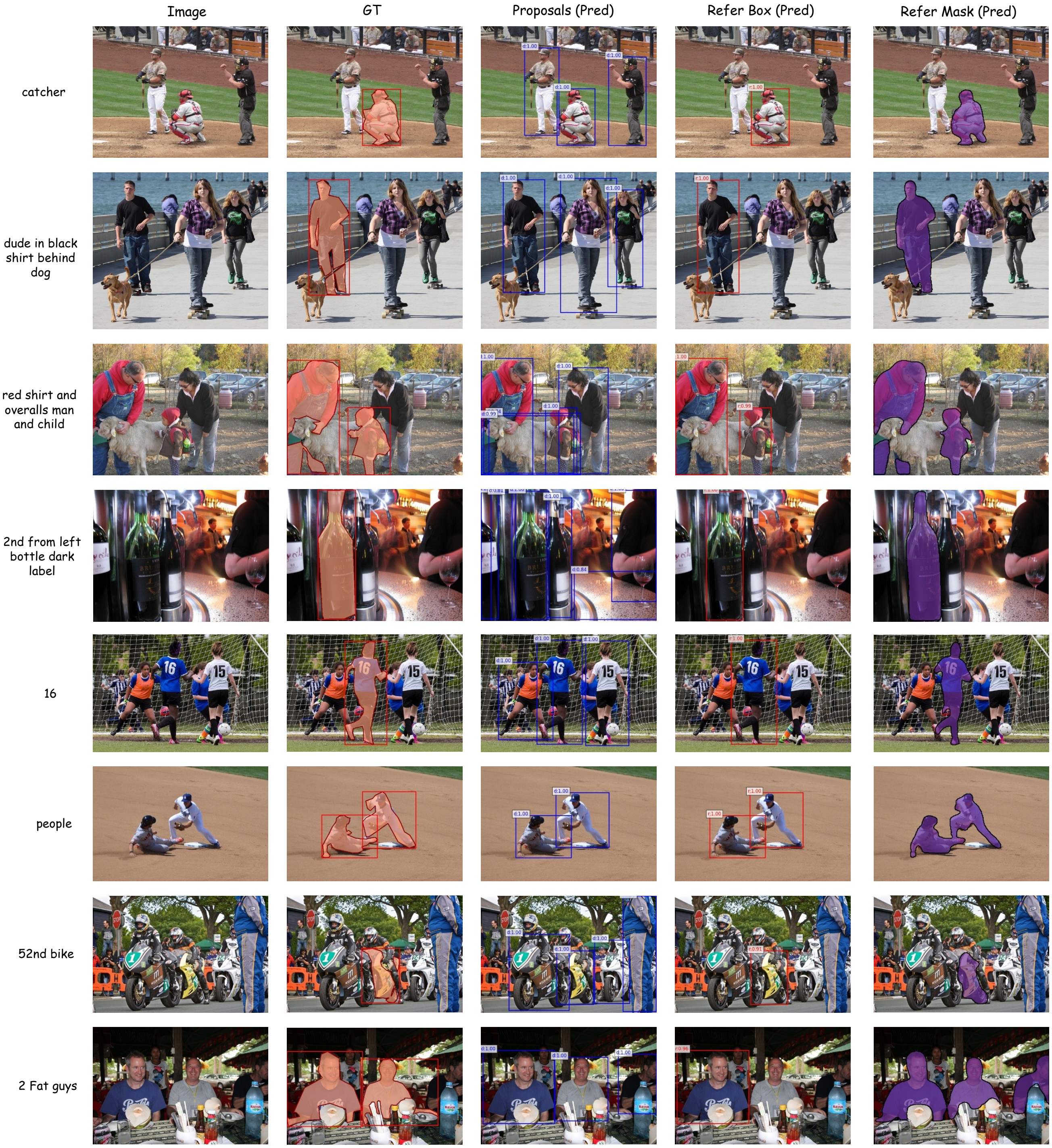
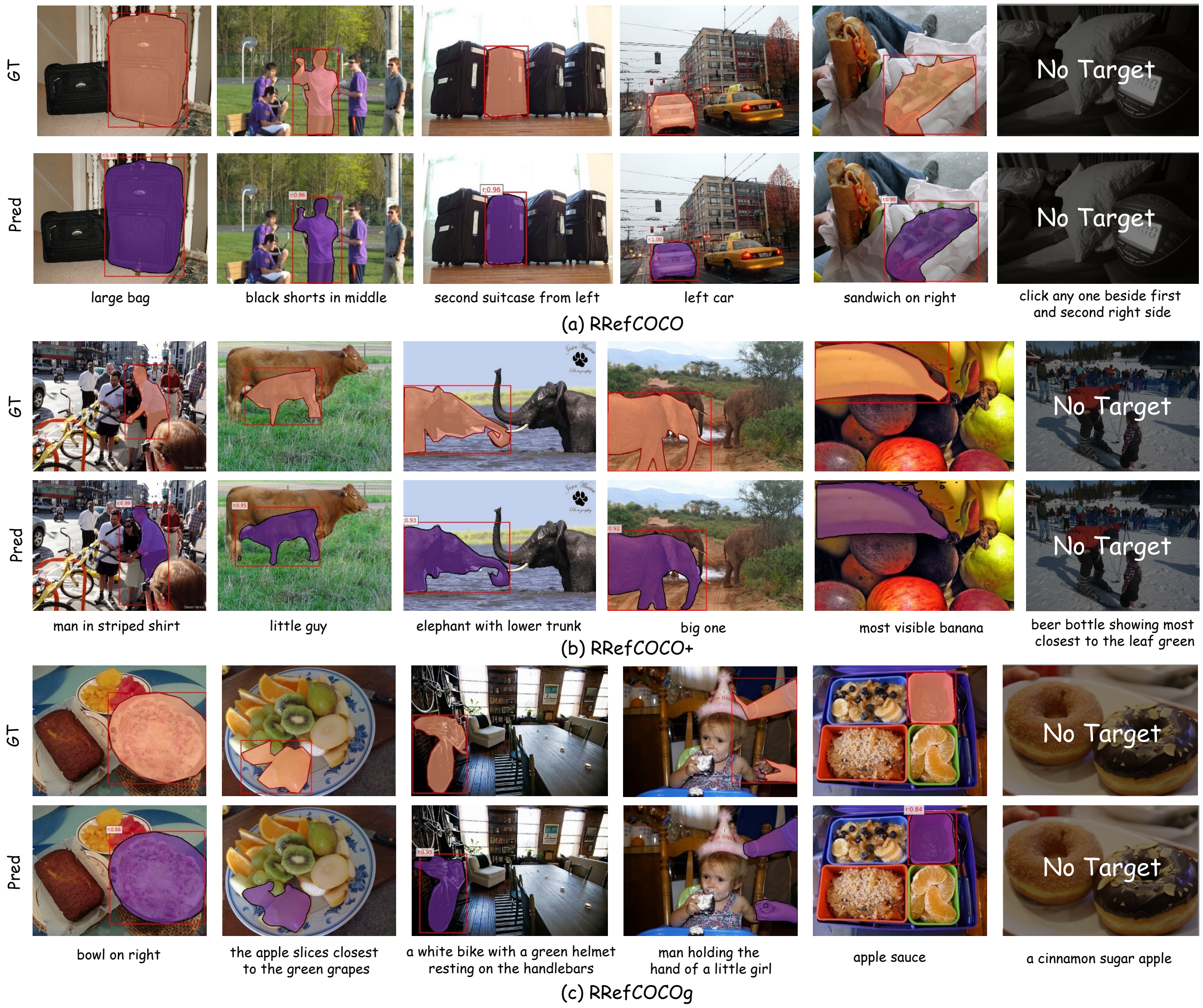


Image
GT
Proposals (Pred)
Refer Box (Pred)
Refer Mask (Pred)
catcher
dude in black shirt behind dog
red shirt and overalls man and child
2nd from left bottle dark label
16
people
52nd bike
2 Fat guys

Figure 8. **Visualization of gRefCOCO dataset.** Proposals(Pred) refers to the predicted proposal boxes, Refer Box(Pred) refers to the predicted refer boxes, and Refer Mask(Pred) refers to the predicted refer masks.

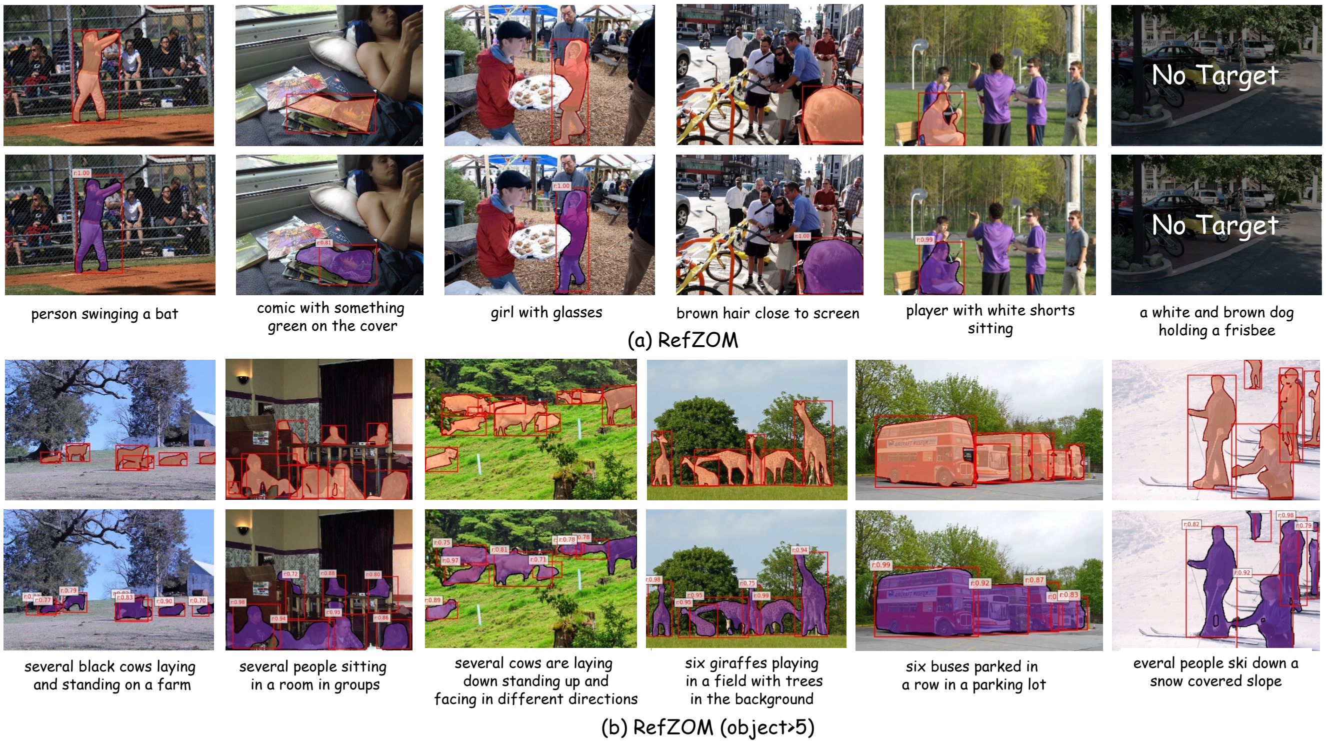

GT
Pred
No Target
No Target
large bag
black shorts in middle
second suitcase from left
left car
sandwich on right
click any one beside first and second right side
GT
Pred
No Target
No Target
man in striped shirt
little guy
elephant with lower trunk
big one
most visible banana
beer bottle showing most closest to the leaf green
GT
Pred
No Target
No Target
bowl on right
the apple slices closest to the green grapes
a white bike with a green helmet resting on the handlebars
man holding the hand of a little girl
apple sauce
a cinnamon sugar apple

(a) RRefCOCO

(b) RRefCOCO+

(c) RRefCOCOg

Figure 9. **Visualization of R-RefCOCO/+/g dataset.**


No Target
No Target
person swinging a bat
comic with something green on the cover
girl with glasses
brown hair close to screen
player with white shorts sitting
a white and brown dog holding a frisbee
several black cows laying and standing on a farm
several people sitting in a room in groups
several cows are laying down standing up and facing in different directions
six giraffes playing in a field with trees in the background
six buses parked in a row in a parking lot
everal people ski down a snow covered slope

(a) RefZOM

(b) RefZOM (object>5)

Figure 10. **Visualization of Ref-ZOM Dataset.** RefZOM (object>5) refers to samples in the Ref-ZOM dataset where the number of objects exceeds five.

References

[1] Shuting He, Henghui Ding, Chang Liu, and Xudong Jiang. GREC: Generalized referring expression comprehension. *arXiv*, 2023. 1

[2] 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

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

[4] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016. 1

[5] Jianzong Wu, Xiangtai Li, Xia Li, Henghui Ding, Yunhai Tong, and Dacheng Tao. Towards robust referring image segmentation. *TIP*, 2024. 1

[6] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85, 2016. 1