# Supplementary Material for "Improving One-stage Visual Grounding by Recursive Sub-query Construction"

Zhengyuan Yang[1] Tianlang Chen[1] Liwei Wang[2] Jiebo Luo[1]

[1]University of Rochester [2]Tencent AI Lab, Bellevue
{zyang39,tchen45,jluo}@cs.rochester.edu, liweiwang@tencent.com

## 1 Qualitative Results

We present additional qualitative results in Figures A, B to visualize the recursive disambiguation process. We highlight the following scenarios.

- **Recursive disambiguation.** The proposed recursive sub-query construction framework improves one-stage visual grounding by addressing the current limitations on grounding long queries. We observe a desired recursive disambiguation process that the text-conditional visual feature step by step generates more accurate and confident predictions.
  As an easy case, better modeling the modifiers of the head noun already corrects a portion of previous failures. For example, in Figure A (a), the peak in the heatmap moves from the tennis player to the referred person in the back, after observing the modifier "watching" in the second round. On the contrary, previous one-stage methods tend to overemphasize the head nouns, without full consideration of the modifiers. Our proposed method also works well on more complex queries via recursive disambiguation, such as in the example "persons head with drill in the middle" from the main paper, and "pony tail lady on the right forefront" in Figure A (b).
- **Challenging regions.** We observe that our method performs well on challenging examples such as Figures A (c) and (d), where the referred target is tiny, the scene contains visually similar distracting objects, and the query includes complex attributes and relationship descriptions.
- **Attributes.** Figures A (e) and (f) include examples that require the correct understanding of attributes such as color and size. For example, in the final round of Figure A (e), the peak moves from the distracting object "trolley" to the referred red suitcase by observing the constructed sub-query "red."
- **Failure cases.** Figures A (g) and (h) show failure cases of our model. The model either misses certain related objects such as the "controller" in Figures A (g) or fails to understand some rarely appeared attributes such as "plaid" in Figures A (h). Therefore, one or more distracting objects remain to have high heatmap responses in the final round, despite the model might still predict the correct bounding box.

We present additional qualitative results in Figure B.

**Table A.** Ablation studies on number of rounds.

| #Rounds | Acc@0.5 | Time(ms) |
|---|---|---|
| $K = 1$ | 58.22 | 23 |
| $K = 2$ | 59.31 | 25 |
| $K = 3$ (Ours) | 60.96 | 26 |
| $K = 4$ | 61.08 | 28 |
| $K = 5$ | 60.80 | 30 |
| $K = 6$ | 61.00 | 32 |

**Table B.** Ablation studies on the modifications in Ours-Large.

| Method | Acc@0.5 | Time(ms) |
|---|---|---|
| Ours-Base | 60.96 | 26 |
| + ConvLSTM | 61.26 | 27 |
| + Size 512 | 61.99 | 34 |
| + Both | 63.12 | 36 |

**Table C.** Performance break-down with attributes.

| *ReferItGame* | Color | Loc. | Size | All |
|---|---|---|---|---|
| Percent | 7.84 | 53.63 | 7.00 | 100.00 |
| One-Stage-BERT | 43.07 | 50.98 | 53.83 | 59.30 |
| Ours-Base | 50.52 | 58.71 | 61.83 | 64.33 |
| **Relative Gain** | 17.30 | 15.16 | 14.86 | 8.48 |

| *RefCOCOg* | Color | Loc. | Size | All |
|---|---|---|---|---|
| Percent | 18.54 | 32.10 | 12.48 | 100.00 |
| One-Stage-BERT | 55.06 | 56.46 | 57.85 | 58.70 |
| Ours-Base | 62.92 | 66.90 | 66.53 | 64.87 |
| **Relative Gain** | 14.28 | 18.49 | 15.00 | 10.51 |

## 2 Ablation Studies

**Number of rounds.** Table A shows the ablation studies on the different number of rounds $K$ on the RefCOCOg-google dataset. We observe that increasing the number of rounds does not lead to an increase in accuracy after a dataset-specific threshold (*e.g.*, $K \geq 3$ on RefCOCOg). Therefore, we select $K = 3$ as the default value in our experiments for a balance between efficiency and accuracy. Although we report all results with $K = 3$ in the main paper's Table 1 for clarity, we note that a different $K$ might slightly improve the accuracy or reduce the inference time on different datasets.

**Ours-Large.** We observe several modules and settings that further improve the grounding accuracy, but meanwhile slightly slow the inference speed or increase the model complexity. Therefore, we list such modifications separately and refer to the corresponding framework "Ours-Large." As shown in Table B, we observe an increase in accuracy with a larger input image size. Furthermore, we improve the accuracy by using all intermediate text-conditional visual feature with a ConvLSTM module. The ConvLSTM module takes feature $\{v^{(k)}\}_{k=1}^{K}$ as the input, and outputs the last hidden state for grounding box prediction.

**Performance break-down with attributes.** The performance break-down study in Section 4.4 shows the effectiveness of our proposed method in modeling and grounding long queries. Other than the improvements on long queries, our method also shows advantages in modeling queries with attribute descriptions, such as color, location, or size. To validate the observation, we construct the attribute subsets from the test set of ReferItGame and RefCOCOg based on the contained attribute keywords, *e.g.*, *"white," "black," "red," "blue,"* etc., for "color;" *"right," "left," "front," "middle,"* etc., for "location;" *"big," "little," "small," "tall,"* etc., for "size."

The first row of Table C shows the experimented dataset and the name of the subset. "Color," "location," and "size" indicate that the query in the subset contains at least one corresponding attribute keywords. "All" reports the performance on the entire dataset. The second row shows the portion of samples in each subset. The remaining rows indicate the grounding accuracy and the relative gain. As shown in the last row, the relative gains on the attribute subsets are around 15% and are higher than the relative gain on the entire dataset of around 10% (*cf.* the middle three and the last column of Table C).

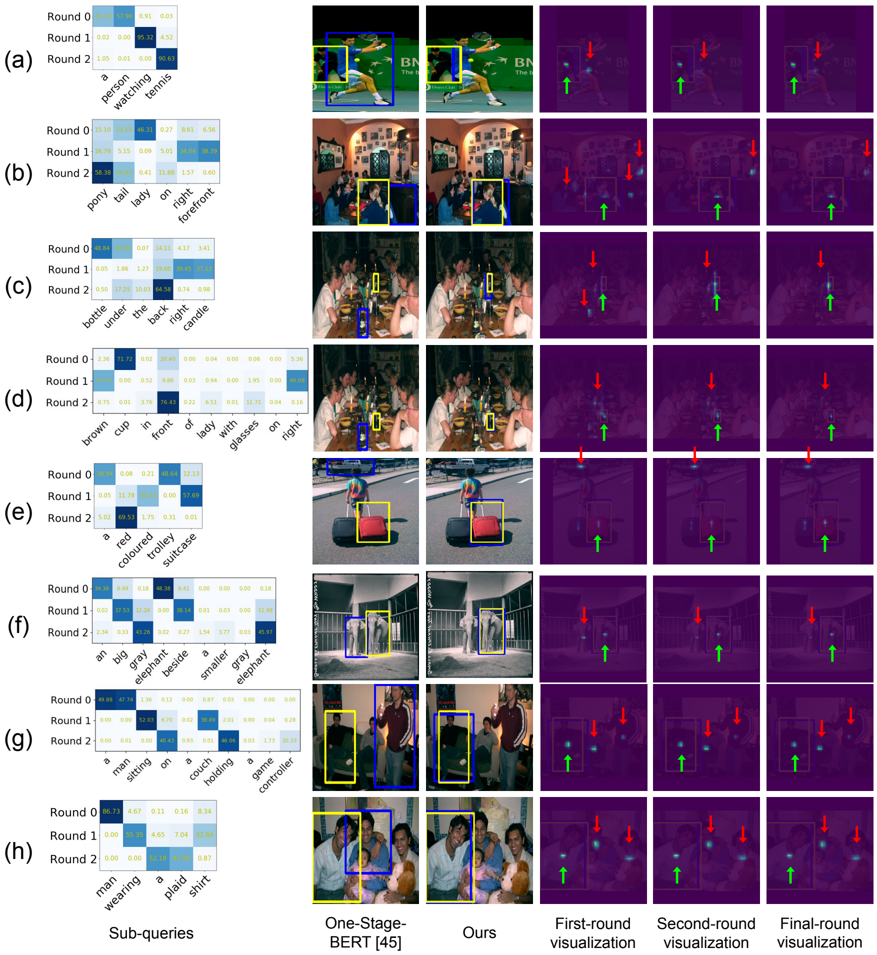

**Fig. A.** Visualizations of the constructed sub-queries and the intermediate text-conditional visual feature at each round. Blue/ yellow boxes are the predicted regions/ ground truths. The green up arrow and the red down arrow highlight the target and the major distracting objects on heatmaps, respectively.

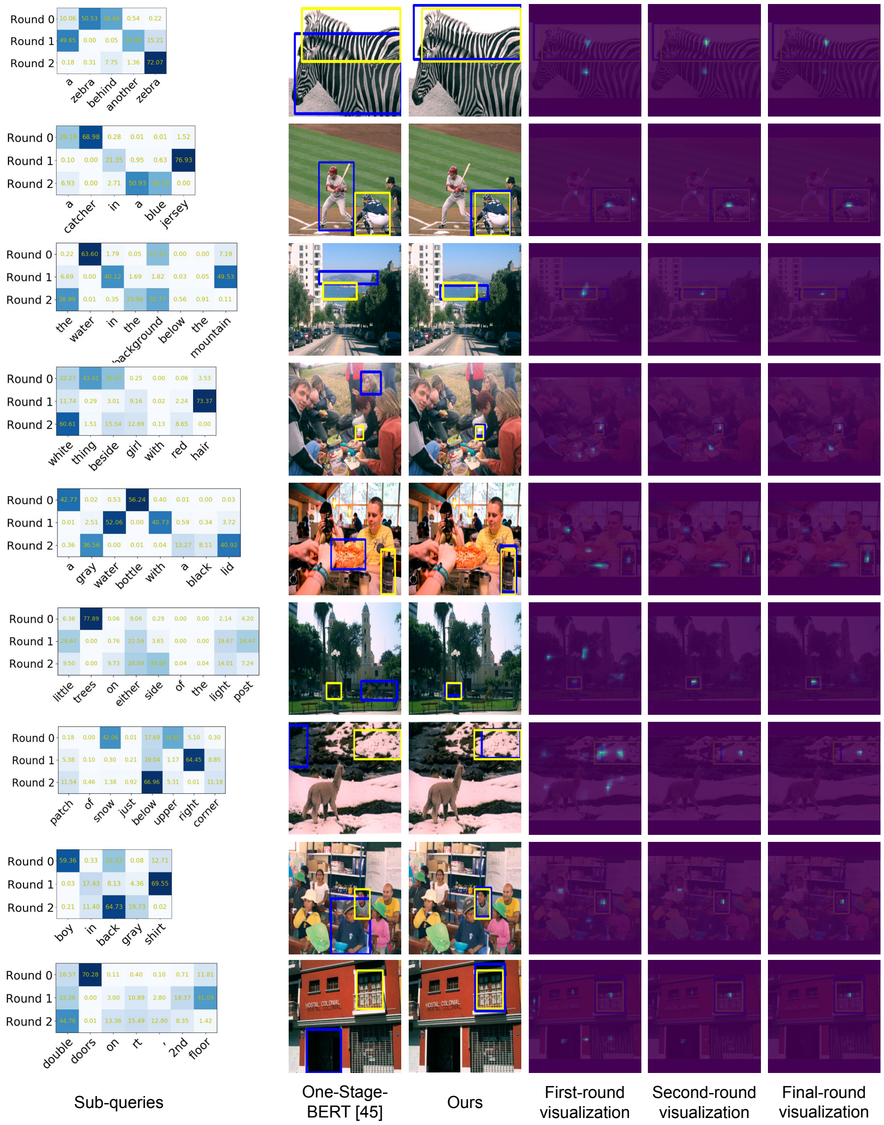
Round 0 10.06 50.53 38.66 0.54 0.22
Round 1 49.65 0.00 0.05 34.96 15.21
Round 2 0.18 0.31 7.75 1.36 72.07
a zebra behind another zebra
Round 0 29.19 68.98 0.28 0.01 0.01 1.52
Round 1 0.10 0.00 21.35 0.95 0.63 76.93
Round 2 6.93 0.00 2.71 50.93 36.74 0.00
a catcher in a blue jersey
Round 0 0.22 63.60 1.79 0.05 27.15 0.00 0.00 7.19
Round 1 6.69 0.00 40.12 1.69 1.82 0.03 0.05 49.53
Round 2 38.99 0.01 0.35 23.96 33.77 0.56 0.91 0.11
the water in the background below the mountain
Round 0 22.27 43.41 30.47 0.25 0.00 0.06 3.53
Round 1 11.74 0.29 3.01 9.16 0.02 2.24 73.37
Round 2 60.61 1.51 15.54 12.69 0.13 8.65 0.00
white thing beside girl with red hair
Round 0 42.77 0.02 0.53 56.24 0.40 0.01 0.00 0.03
Round 1 0.01 2.51 52.06 0.00 40.73 0.59 0.34 3.72
Round 2 0.36 36.56 0.00 0.01 0.04 13.27 8.11 40.92
a gray water bottle with a black lid
Round 0 6.36 77.89 0.06 9.06 0.29 0.00 0.00 2.14 4.20
Round 1 26.67 0.00 0.76 22.59 3.65 0.00 0.00 19.67 26.47
Round 2 9.50 0.00 9.73 26.09 30.88 0.04 0.04 14.01 7.24
little trees on either side of the light post
Round 0 0.18 0.00 42.06 0.01 17.69 34.65 5.10 0.30
Round 1 5.38 0.10 0.30 0.21 19.04 1.17 64.45 8.85
Round 2 11.54 0.46 1.38 0.92 66.96 5.31 0.01 11.19
patch of snow just below upper right corner
Round 0 59.36 0.33 27.52 0.08 12.71
Round 1 0.03 17.43 8.13 4.36 69.55
Round 2 0.21 11.40 64.73 19.73 0.02
boy in back gray shirt
Round 0 16.57 70.28 0.11 0.40 0.10 0.71 11.81
Round 1 23.26 0.00 3.00 10.89 2.80 18.37 41.09
Round 2 44.76 0.01 13.36 15.49 12.80 8.35 1.42
double doors on rt , 2nd floor
HOSTAL COLONIAL
HOSTAL COLONIAL
Sub-queries
One-Stage-BERT [45]
Ours
First-round visualization
Second-round visualization
Final-round visualization

**Fig. B.** Additional qualitative results.