UNIVERSITY of ROCHESTER

Tencent AI Lab

ECCV'20 ONLINE
23-28 AUGUST 2020

Improving One-stage Visual Grounding by Recursive Sub-query Construction

Zhengyuan Yang[1]

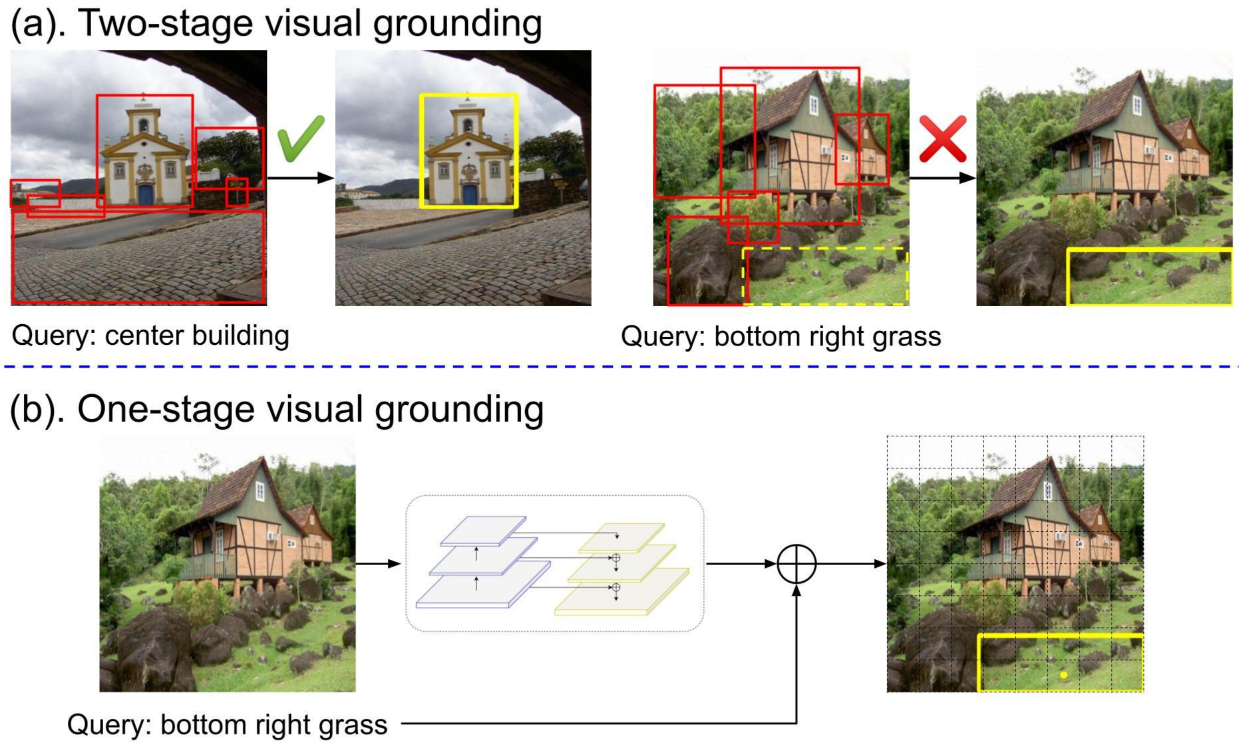
Tianlang Chen[1]

Liwei Wang[2]

Jiebo Luo[1]

Visual Grounding

- Grounding a language query onto a region of the image



(a). Two-stage visual grounding

Query: center building

Query: bottom right grass

(b). One-stage visual grounding

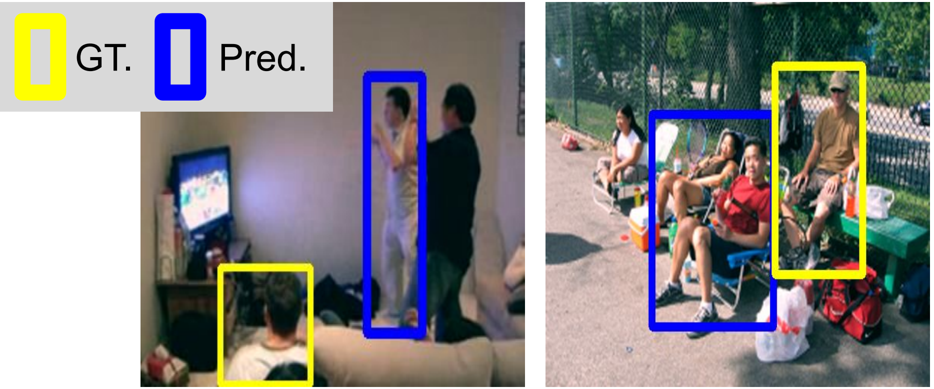
Query: bottom right grass

Figure from Yang, Zhengyuan, et al. "A fast and accurate one-stage approach to visual grounding." *In ICCV* 2019.

One-stage Visual Grounding

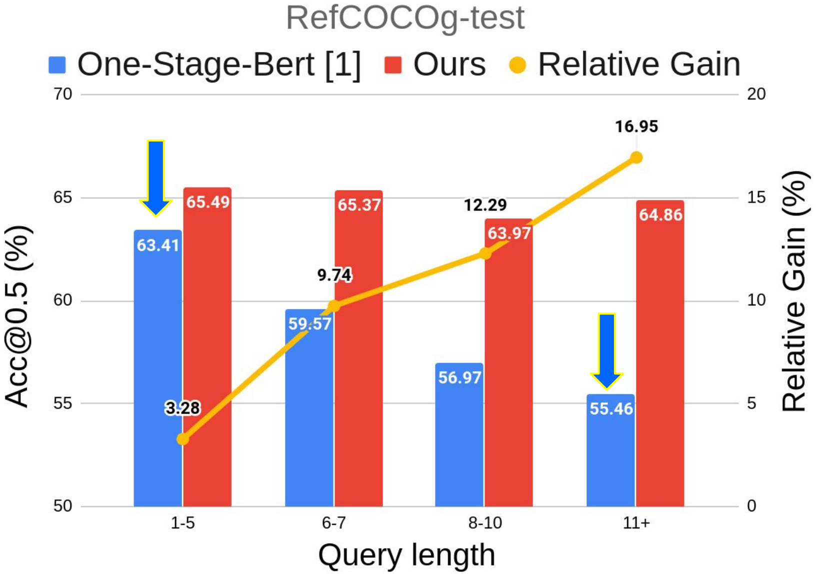
- **Major Limitations**

- Limited performance on long and complicated queries



(a). man sitting on the couch and looking on the tv.

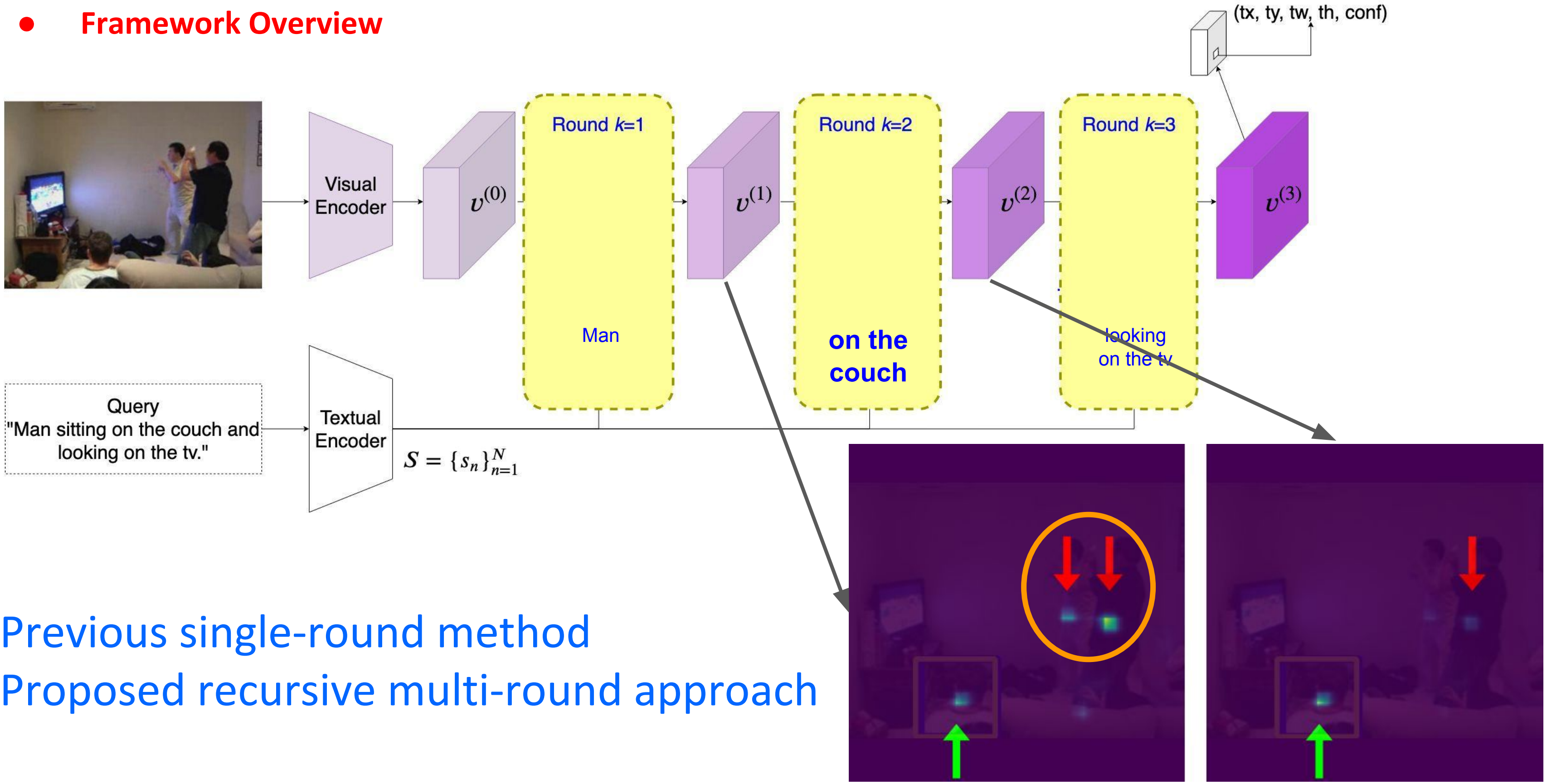
(b). the man in tan shirt in the back.



[1] Yang, Zhengyuan, et al. "A fast and accurate one-stage approach to visual grounding." *In ICCV* 2019.

Method

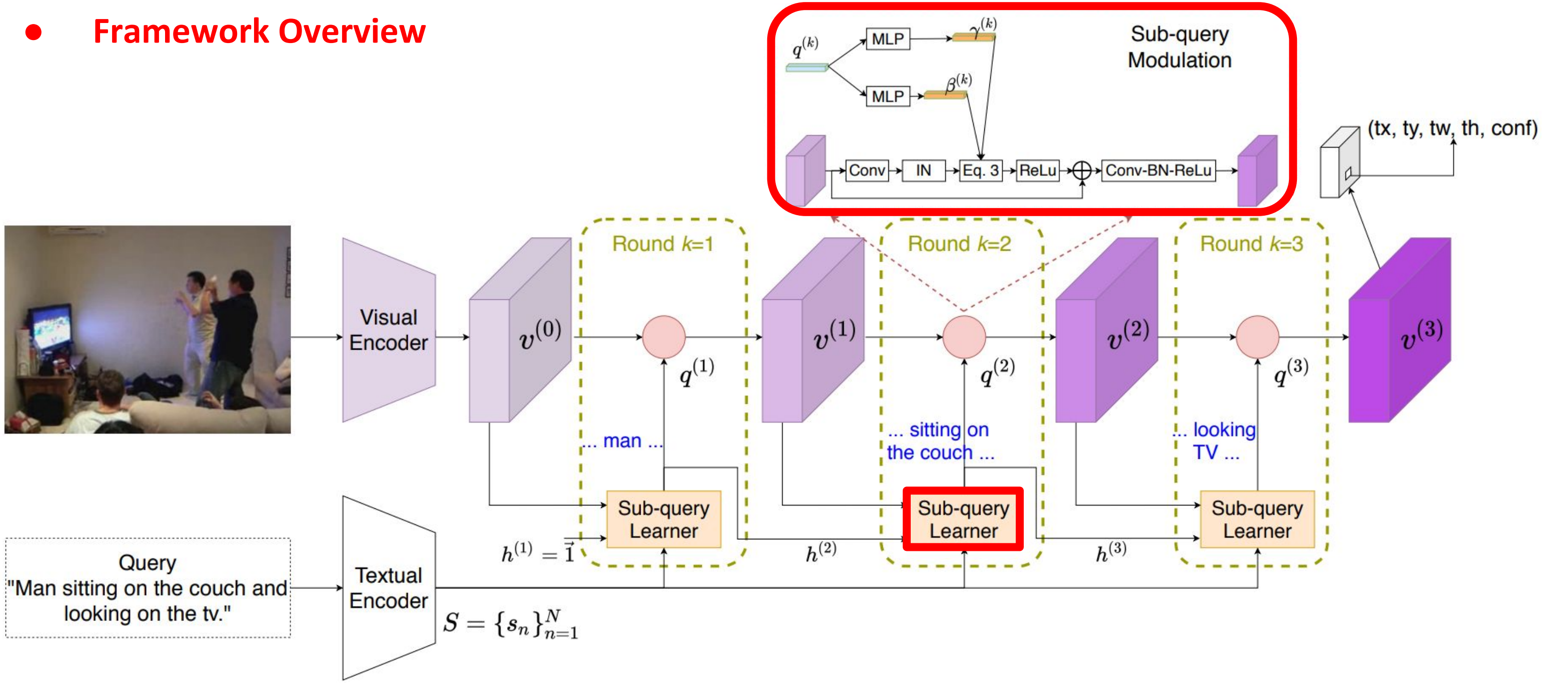
- **Framework Overview**



- Previous single-round method
- Proposed recursive multi-round approach

Method

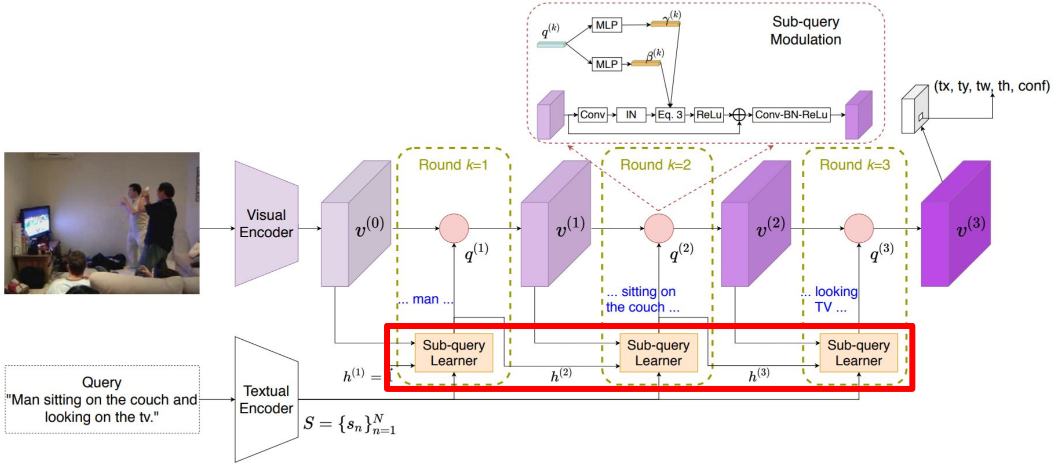
- **Framework Overview**



- Sub-query learner
- Sub-query modulation

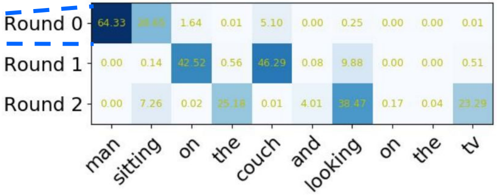
Method

- **Sub-query learner**



Input $\left(\{s_n\}_{n=1}^N, h^{(k)}, v^{(k-1)}\right)$

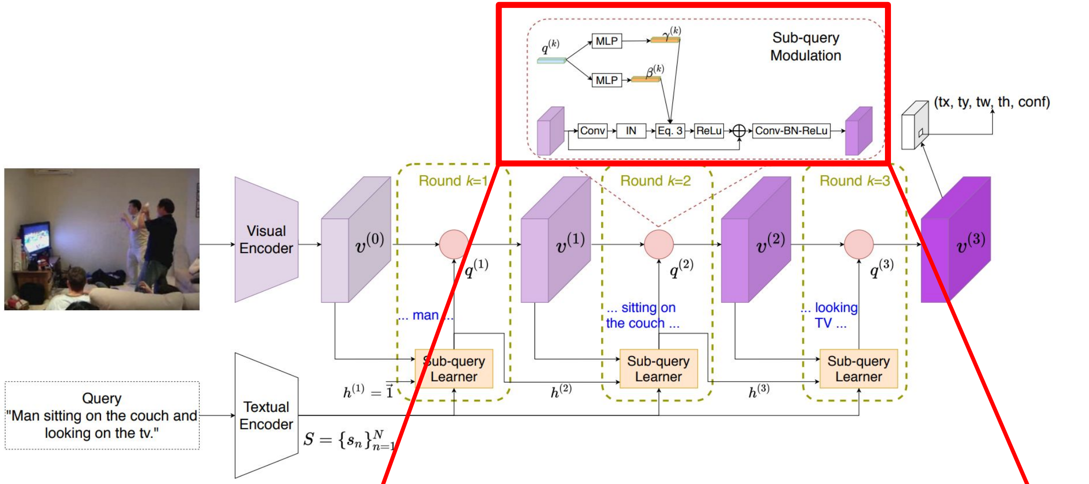
Output $q^{(k)} = \sum_{n=1}^{N} \boxed{\alpha_n^{(k)}} s_n$



- Referring to visual-text featureV_K during sub-query construction

Method

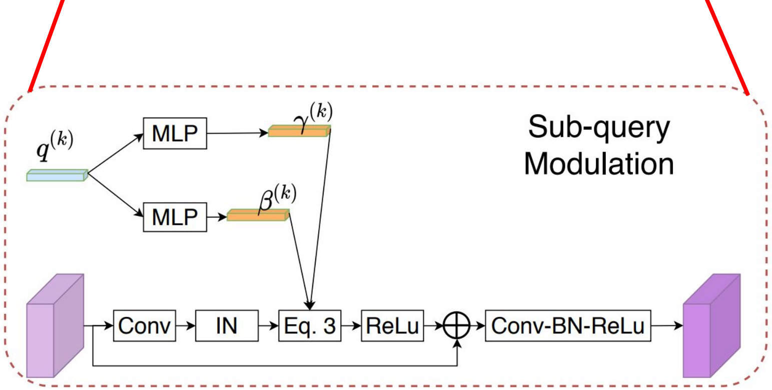
- **Sub-query modulation**



$$\gamma^{(k)} = \tanh\left(W_\gamma^{(k)} q^{(k)} + b_\gamma^{(k)}\right)$$

$$\beta^{(k)} = \tanh\left(W_\beta^{(k)} q^{(k)} + b_\beta^{(k)}\right)$$

$$v^{(k)}(i,j) = v^{(k-1)}(i,j) \odot \gamma^{(k)} + \beta^{(k)}$$



- Scaling and shifting fused feature with new sub-query

Experiments

- **Datasets and metrics**

- Datasets: RefCOCO, RefCOCO+, RefCOCOg, ReferItGame
- Acc@0.5: correct if top-1 IoU>0.5

man sitting on the couch and looking on the tv

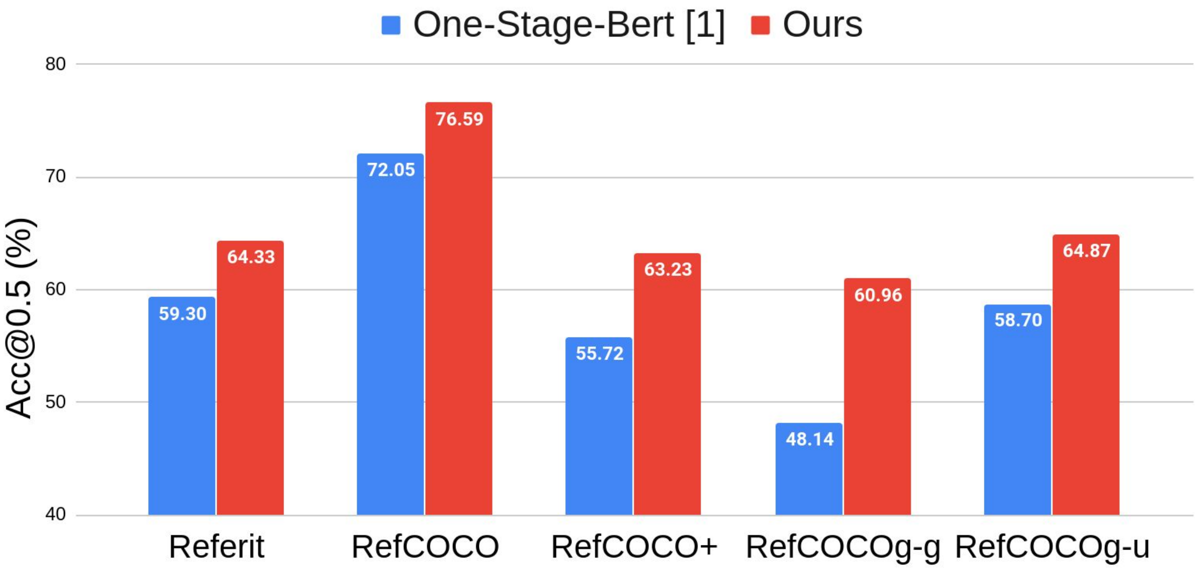
RefCOCO, RefCOCO+,
RefCOCOg

the black backpack on the bottom right

ReferItGame

Experiments

- **Comparison to other methods**

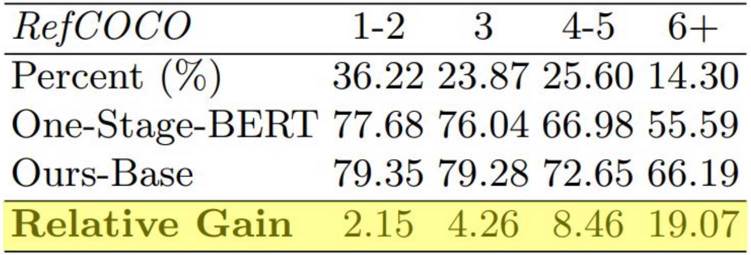


- Over 5% improvements with comparable inference speed

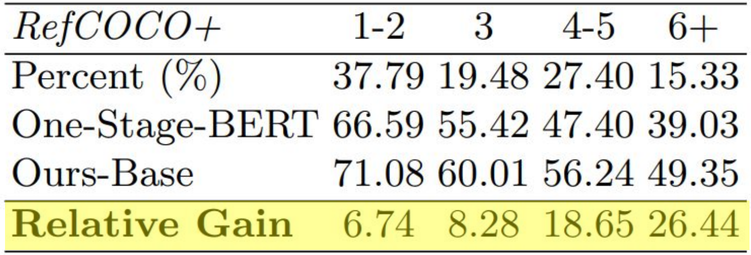
[1] Yang, Zhengyuan, et al. "A fast and accurate one-stage approach to visual grounding." *In ICCV* 2019.

Experiments

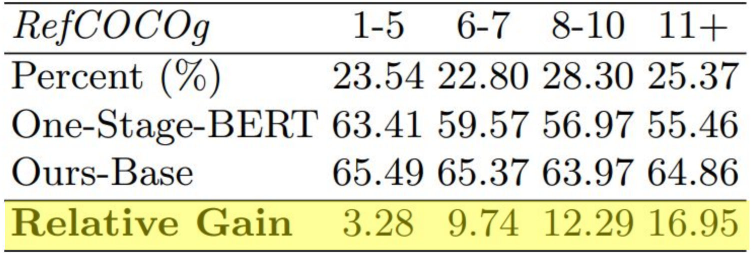
- **Performance break-down with query lengths**



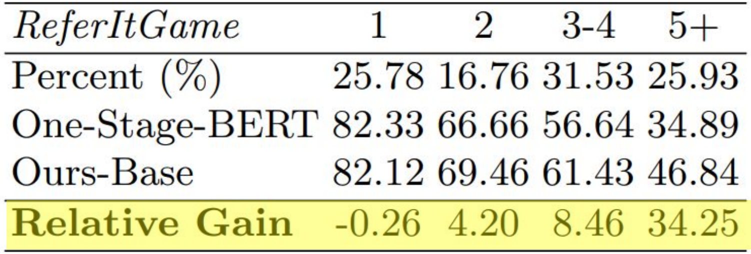
RefCOCO	1-2	3	4-5	6+
Percent (%)	36.22	23.87	25.60	14.30
One-Stage-BERT	77.68	76.04	66.98	55.59
Ours-Base	79.35	79.28	72.65	66.19
Relative Gain	2.15	4.26	8.46	19.07



RefCOCO+	1-2	3	4-5	6+
Percent (%)	37.79	19.48	27.40	15.33
One-Stage-BERT	66.59	55.42	47.40	39.03
Ours-Base	71.08	60.01	56.24	49.35
Relative Gain	6.74	8.28	18.65	26.44



RefCOCOg	1-5	6-7	8-10	11+
Percent (%)	23.54	22.80	28.30	25.37
One-Stage-BERT	63.41	59.57	56.97	55.46
Ours-Base	65.49	65.37	63.97	64.86
Relative Gain	3.28	9.74	12.29	16.95



ReferItGame	1	2	3-4	5+
Percent (%)	25.78	16.76	31.53	25.93
One-Stage-BERT	82.33	66.66	56.64	34.89
Ours-Base	82.12	69.46	61.43	46.84
Relative Gain	-0.26	4.20	8.46	34.25

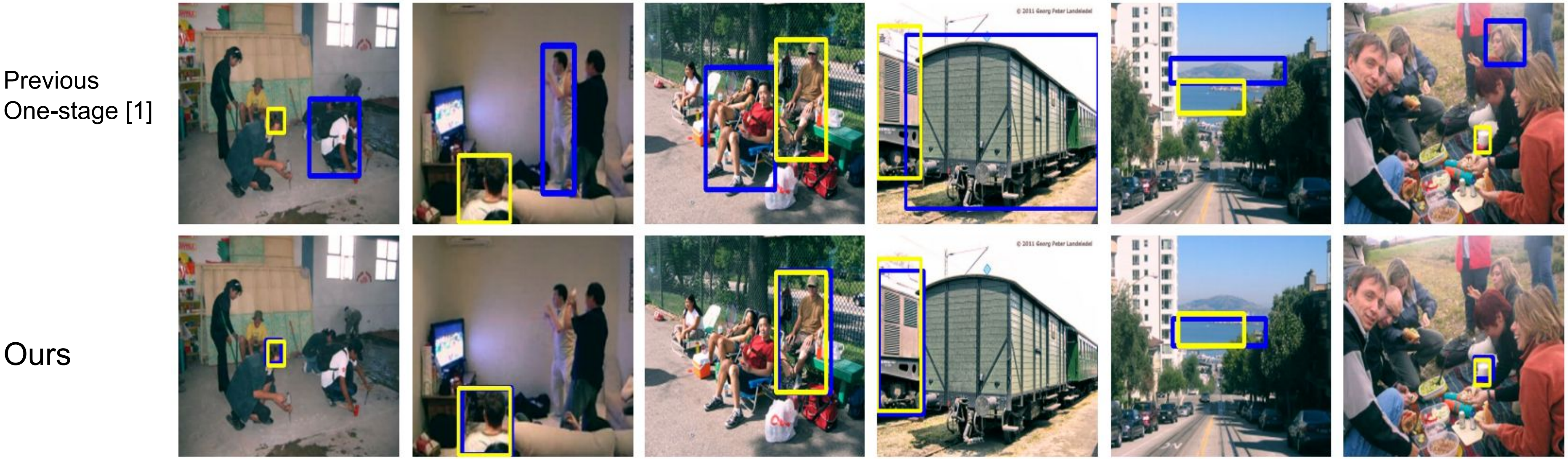
- Better performance on longer queries

[1] Yang, Zhengyuan, et al. "A fast and accurate one-stage approach to visual grounding." *In ICCV* 2019.

Experiments

- **Qualitative results**

🟨 GT. 🟦 Pred.



(a). persons head with drill in the middle.

(b). man sitting on the couch and looking on the tv.

(c). the man in tan shirt in the back.

(d). the rail car on the other track.

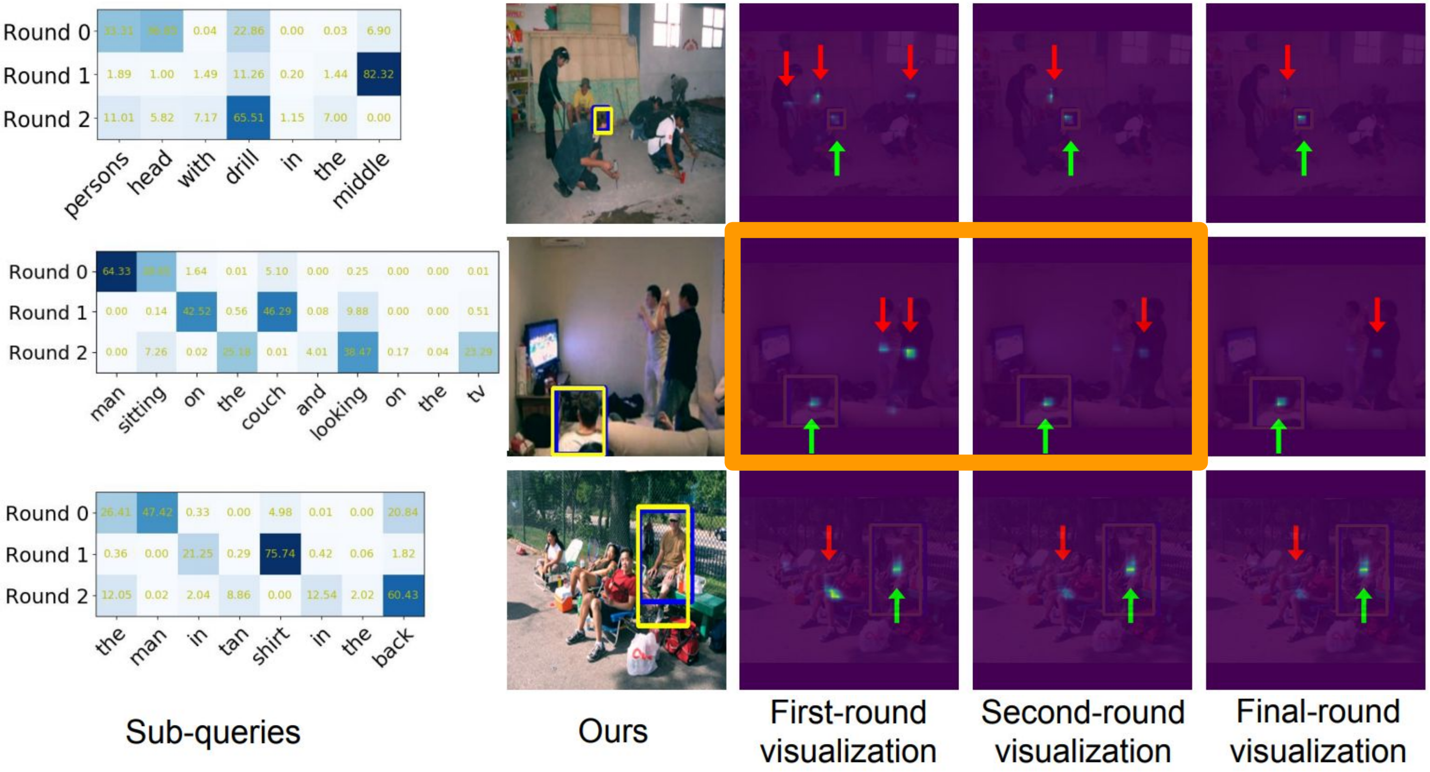
(e). the water in the background below the mountain.

(f). white thing beside girl with red hair.

[1] Yang, Zhengyuan, et al. "A fast and accurate one-stage approach to visual grounding." *In ICCV* 2019.

Experiments

- **Recursive disambiguation**

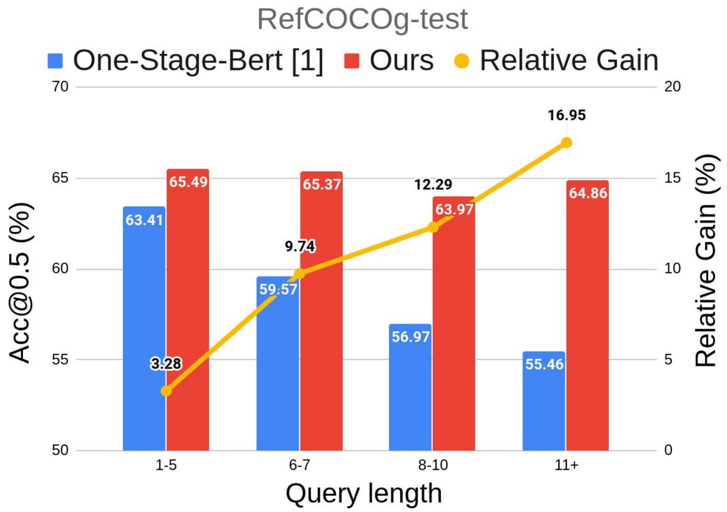


- Recursive dis-ambiguous procedures

Improving One-stage Visual Grounding by Recursive Sub-query Construction

Code & models:
https://github.com/zyang-ur/ReSC

Contact:
zyang39@cs.rochester.edu


RefCOCOg-test
■ One-Stage-Bert [1] ■ Ours ● Relative Gain
Acc@0.5 (%)
Relative Gain (%)
Query length