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]

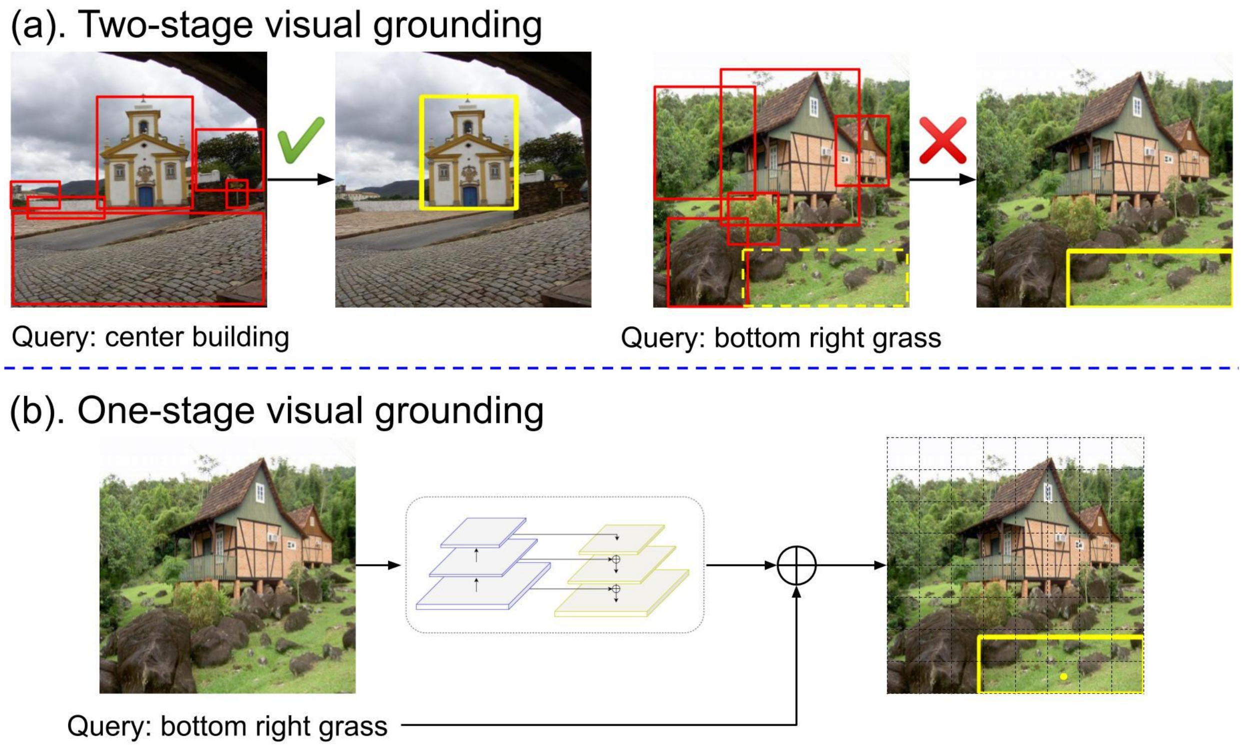
Tianlang Chen[1]

Liwei Wang[2]

Jiebo Luo[1]

Visual Grounding

- Grounding a language query onto a region of the image



(a). Two-stage visual grounding

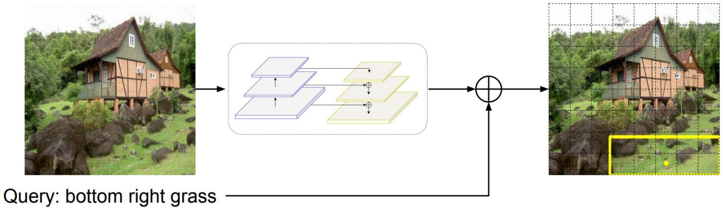
(b). One-stage visual grounding

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

One-stage Visual Grounding

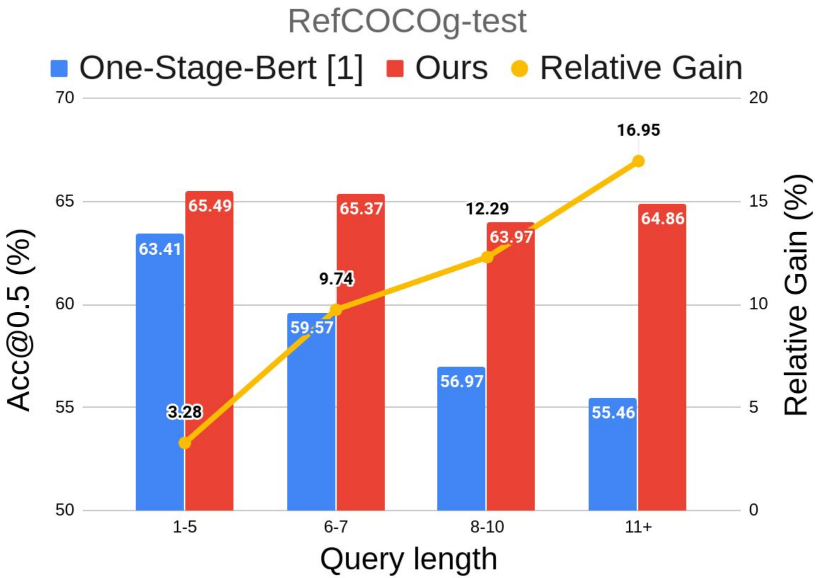
- **Major Limitations**

- Limited performance on long and complicated queries



Single-round fusion:

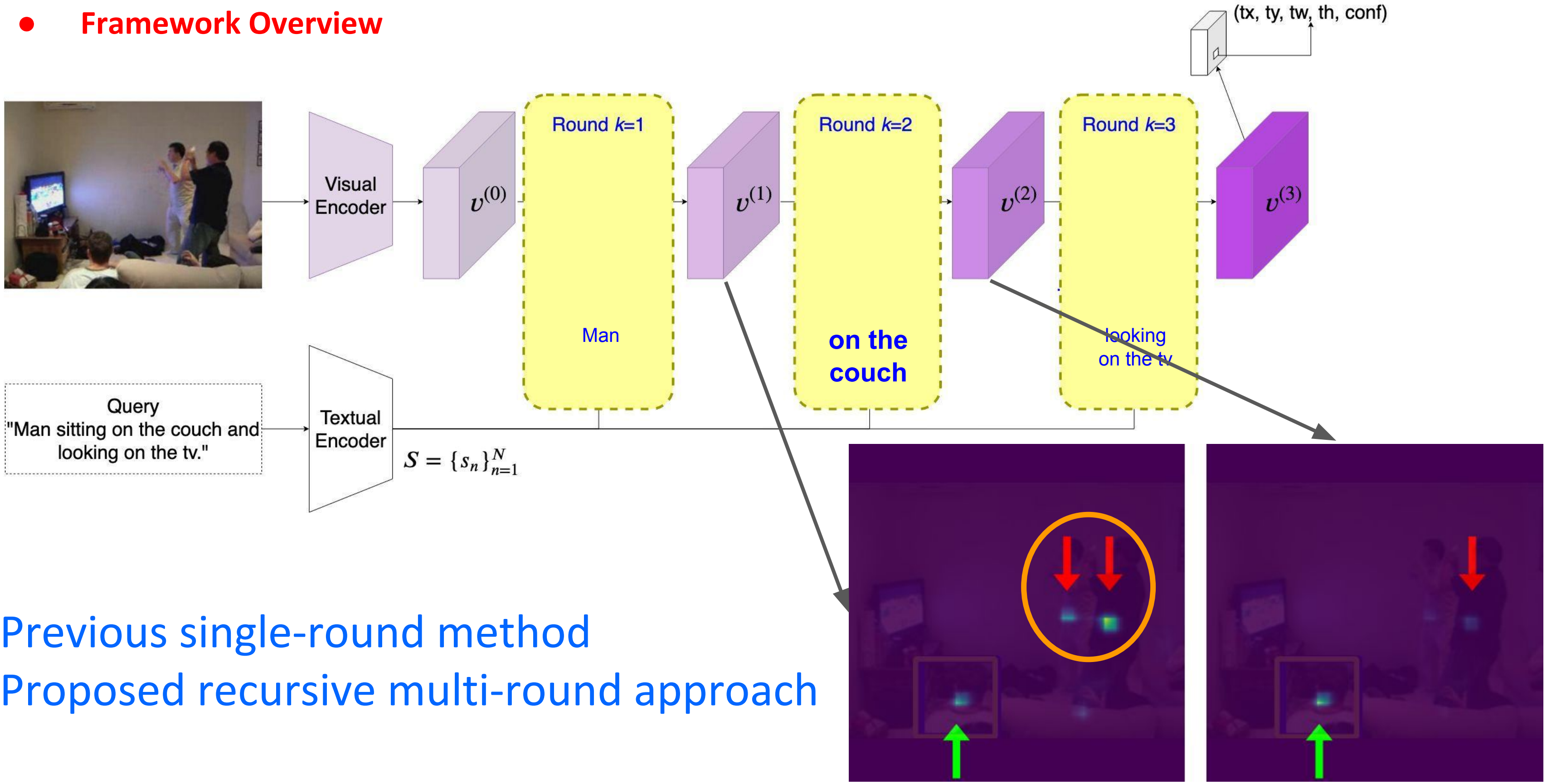
overlooking certain words



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

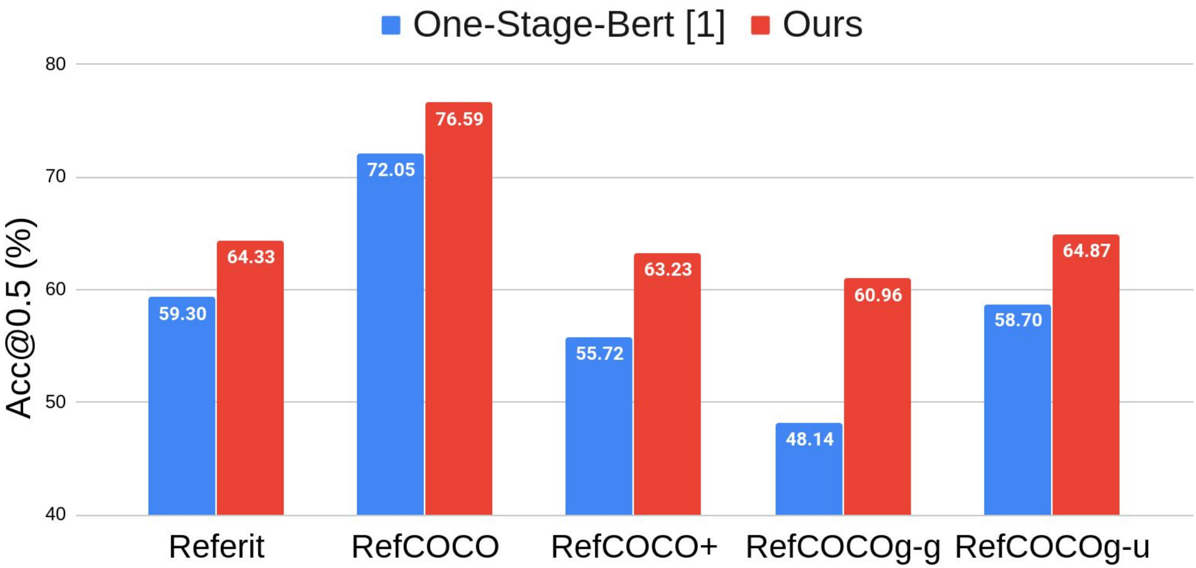
Method

- **Framework Overview**



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

Experiment Results

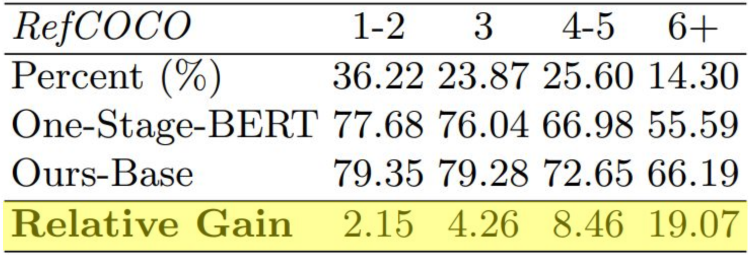


- Significant improvements with comparable inference speed

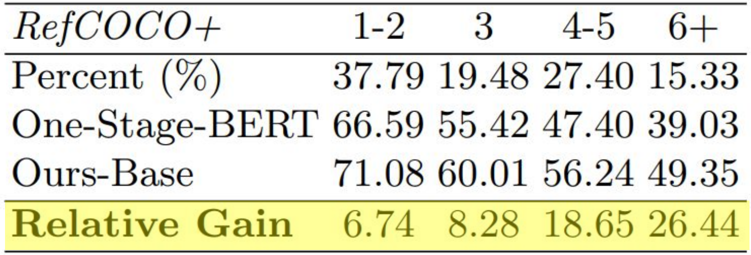
[1] Yang, Zhengyuan, et al. "A fast and accurate one-stage approach to visual grounding." *In ICCV* 2019.

Experiment Results

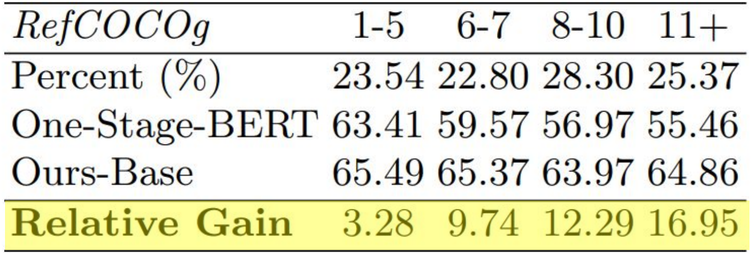
- **Performance break-down with query lengths**



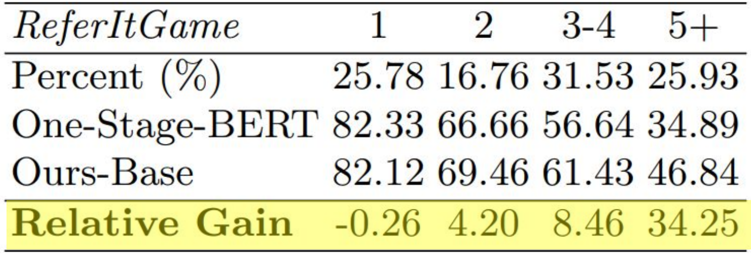
RefCOCO	1-2	3	4-5	6+
Percent (%)	36.22	23.87	25.60	14.30
One-Stage-BERT	77.68	76.04	66.98	55.59
Ours-Base	79.35	79.28	72.65	66.19
Relative Gain	2.15	4.26	8.46	19.07



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

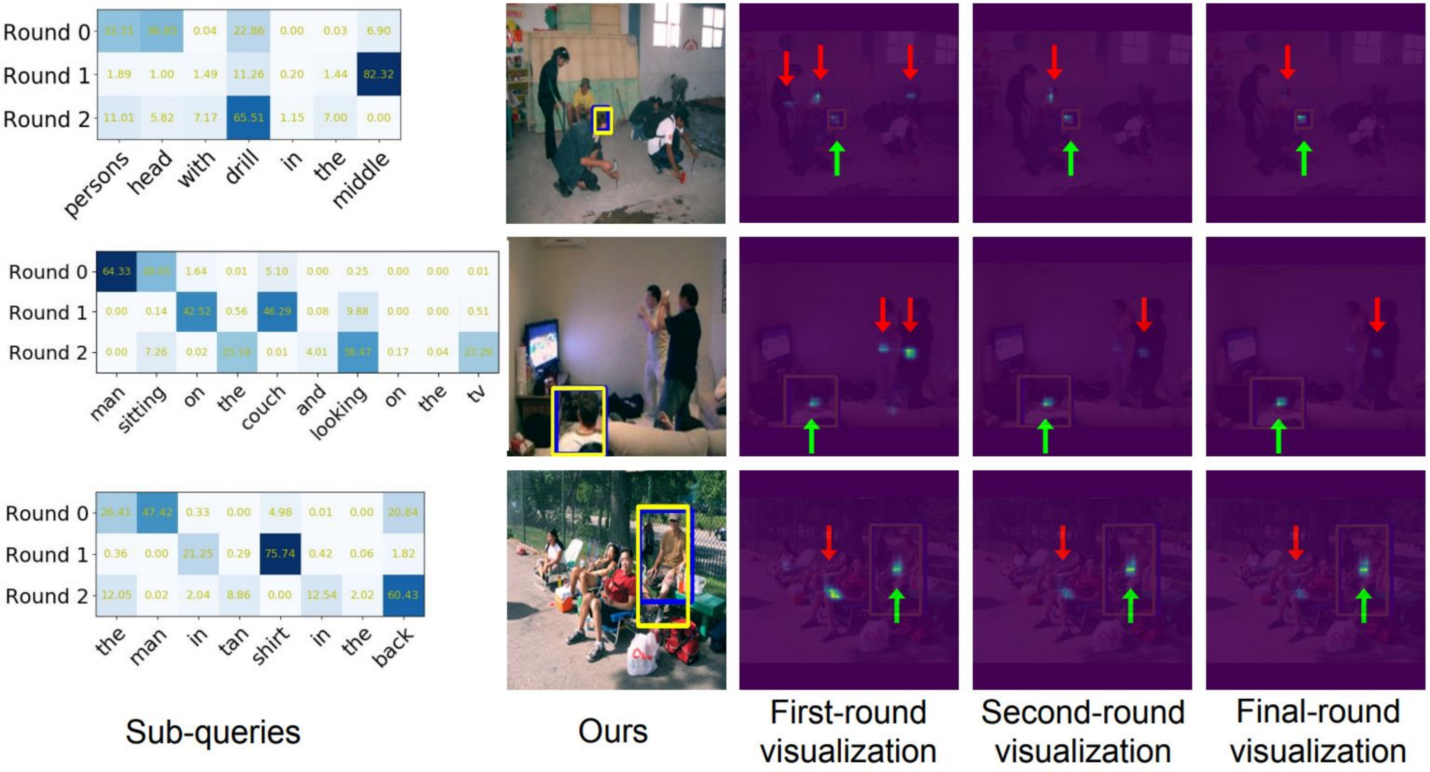


ReferItGame	1	2	3-4	5+
Percent (%)	25.78	16.76	31.53	25.93
One-Stage-BERT	82.33	66.66	56.64	34.89
Ours-Base	82.12	69.46	61.43	46.84
Relative Gain	-0.26	4.20	8.46	34.25

- Better performance on longer queries

Experiment Results

- **Recursive disambiguation**



	persons	head	with	drill	in	the	middle
Round 0	33.31	36.85	0.04	22.86	0.00	0.03	6.90
Round 1	1.89	1.00	1.49	11.26	0.20	1.44	82.32
Round 2	11.01	5.82	7.17	65.51	1.15	7.00	0.00

	man	sitting	on	the	couch	and	looking	on	the	tv
Round 0	64.33	28.65	1.64	0.01	5.10	0.00	0.25	0.00	0.00	0.01
Round 1	0.00	0.14	42.52	0.56	46.29	0.08	9.88	0.00	0.00	0.51
Round 2	0.00	7.26	0.02	25.18	0.01	4.01	38.47	0.17	0.04	23.29

	the	man	in	tan	shirt	in	the	back
Round 0	26.41	47.42	0.33	0.00	4.98	0.01	0.00	20.84
Round 1	0.36	0.00	21.25	0.29	75.74	0.42	0.06	1.82
Round 2	12.05	0.02	2.04	8.86	0.00	12.54	2.02	60.43

Sub-queries

Ours

First-round visualization

Second-round visualization

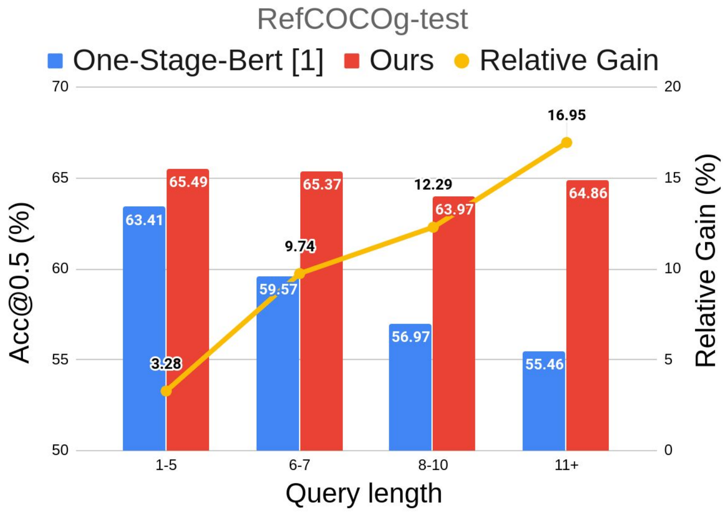
Final-round visualization

- Recursive disambiguous 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