Imperial College London

Phrase Localization Without Paired Training Examples

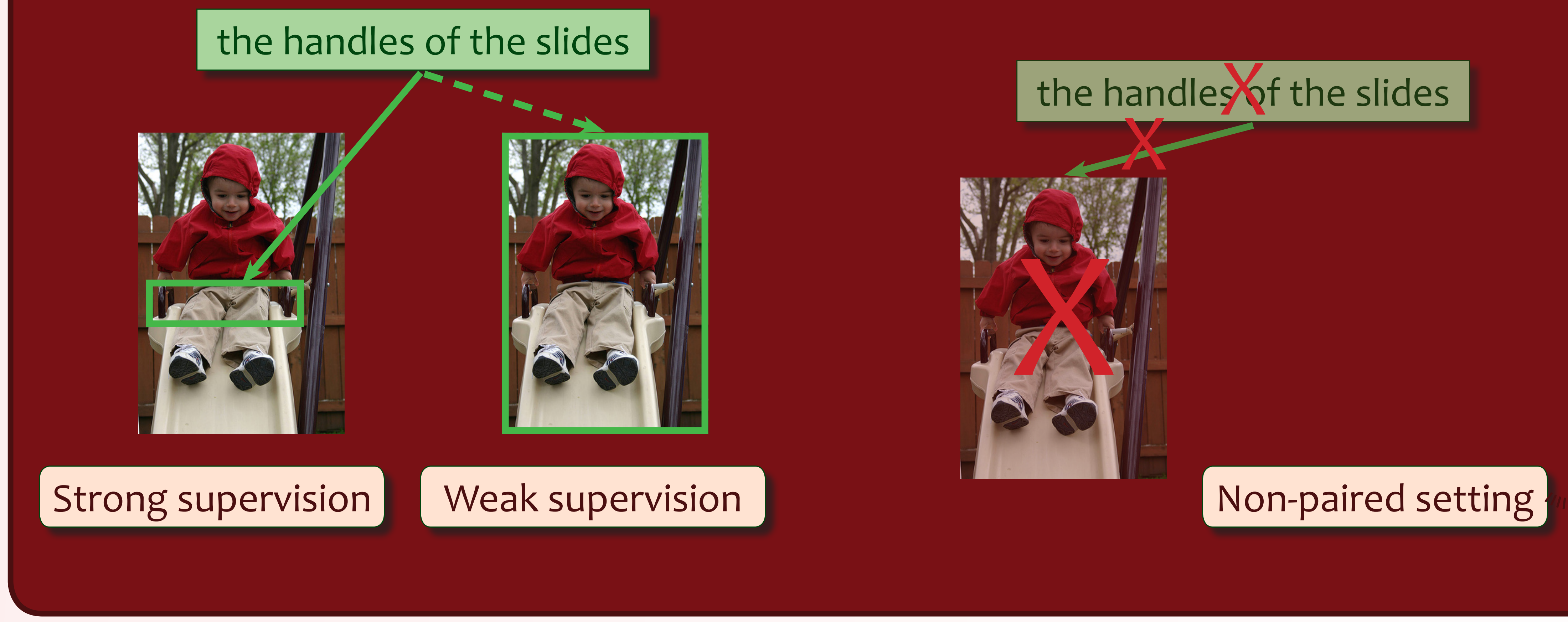
Josiah Wang, Lucia Specia

multimt
Newton Fund

Phrase Localization

- Phrase Localization models are often trained with supervised paired training data.

- What if we don't have such training data?
- Can we still solve the problem?
- Maybe with off-the-shelf tools/model/resources?

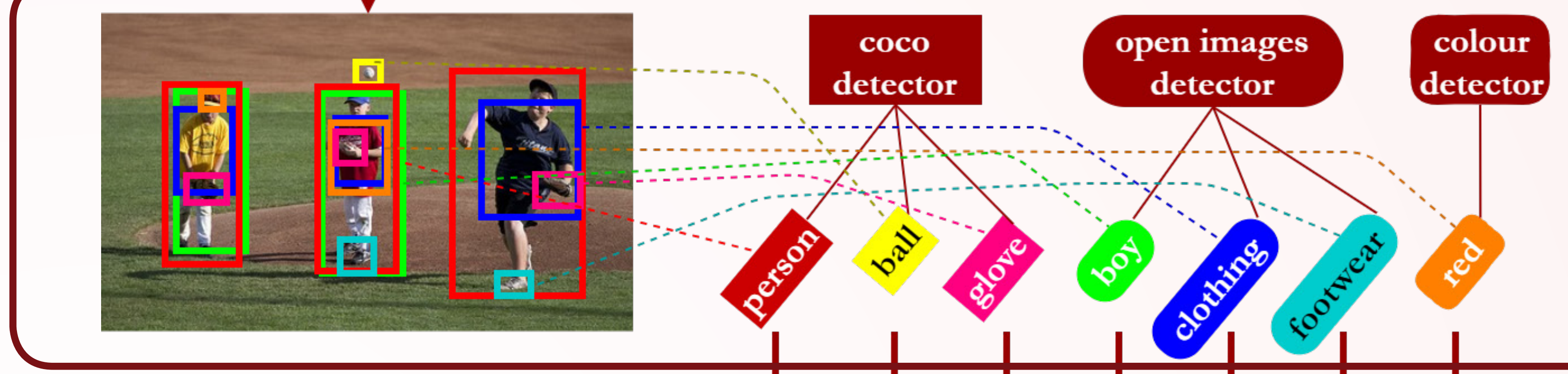


Non-paired Setting Baseline

1 Detect instances

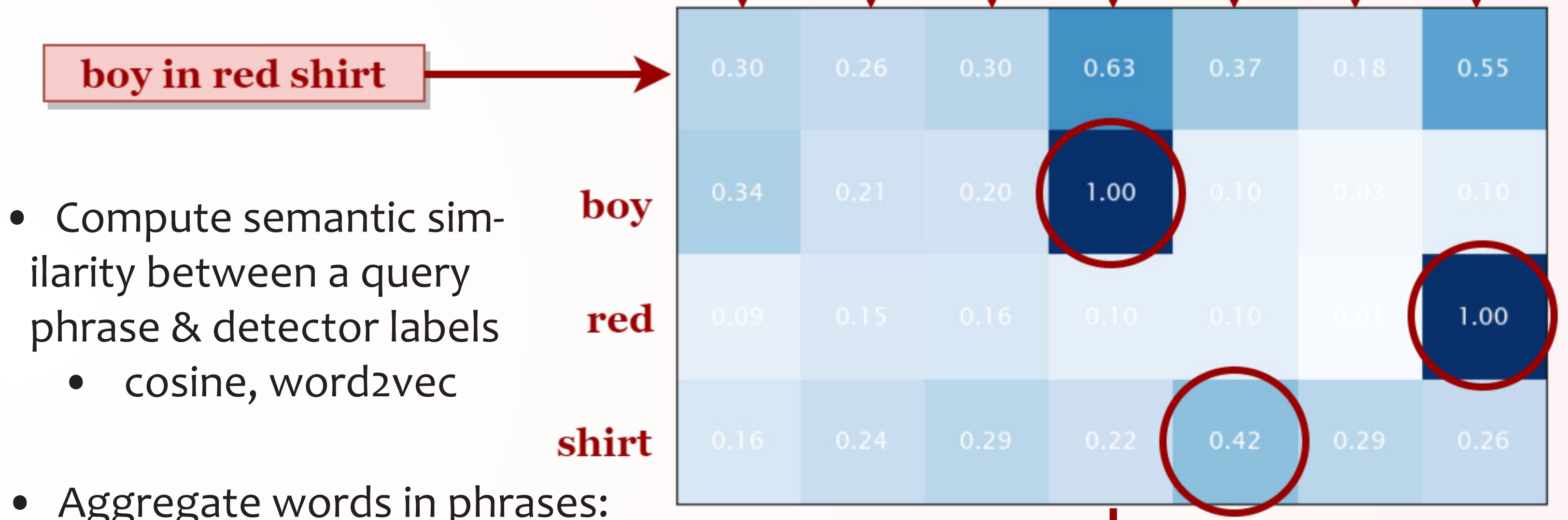
Off-the-shelf pre-trained detectors:

- *CC*: COCO (80 categories)
- *OI*: Open Images (>500 categories)
- *PL*: Places365 (365 categories)
- *YL*: YOLO9000 (>9000 categories)
- *CL*: Colours (11 categories)
- Combinations of above

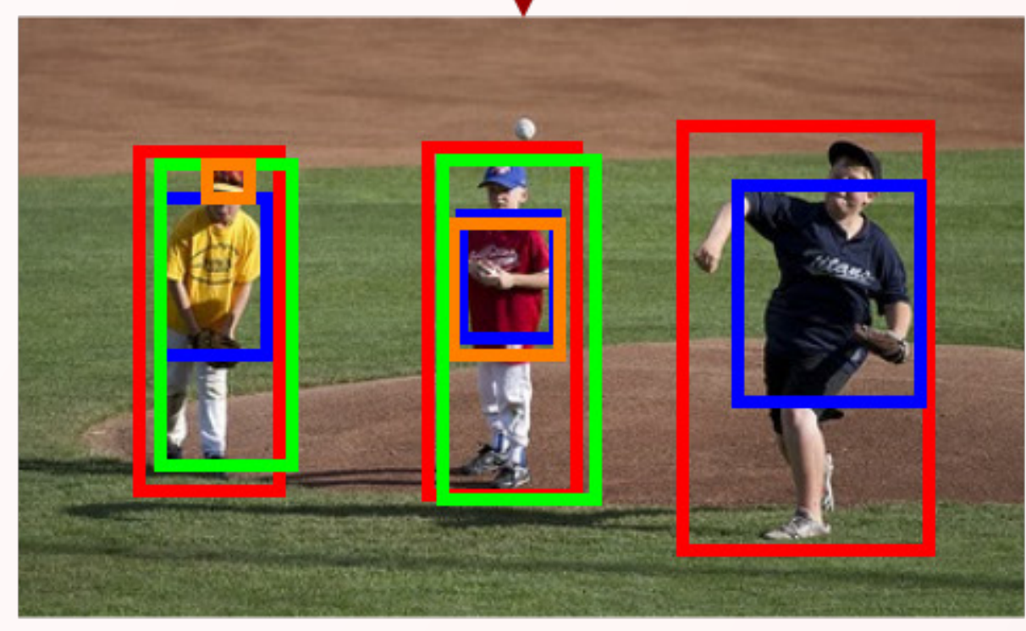


2 Select relevant concepts

- Compute semantic similarity between a query phrase & detector labels
  - cosine, word2vec
- Aggregate words in phrases:
  - average embeddings using one word from phrase
- Output:
  - Ranked/filtered list of bounding boxes

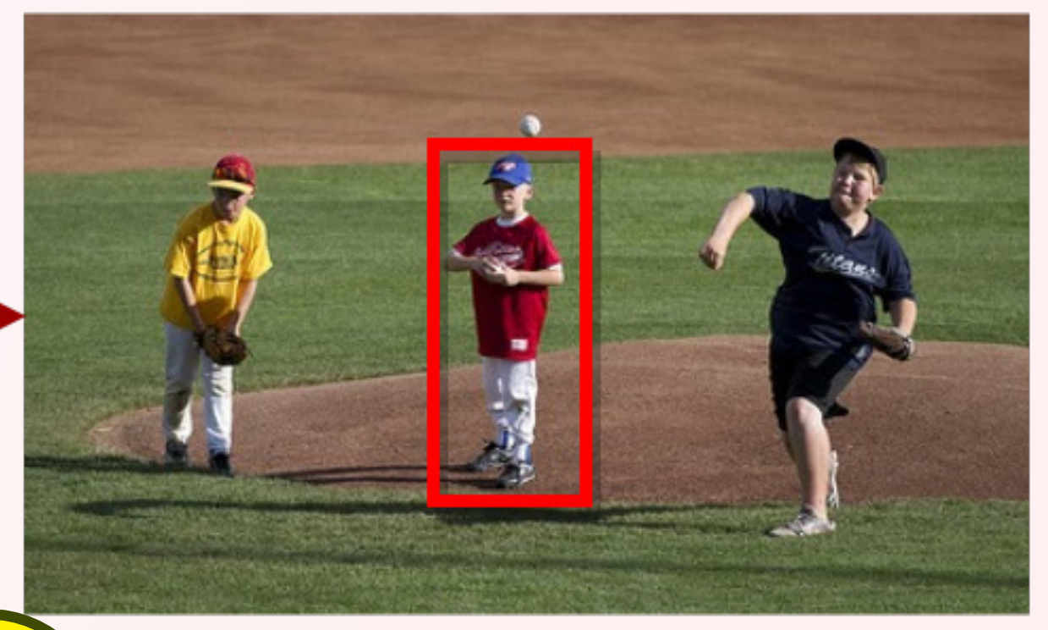


boy in red shirt	person	ball	glove	boy	clothing	footwear	red
	0.30	0.26	0.30	0.63	0.37	0.18	0.55
boy	0.34	0.21	0.20	1.00	0.10	0.03	0.10
red	0.09	0.15	0.16	0.10	0.10	0.01	1.00
shirt	0.16	0.24	0.29	0.22	0.42	0.29	0.26

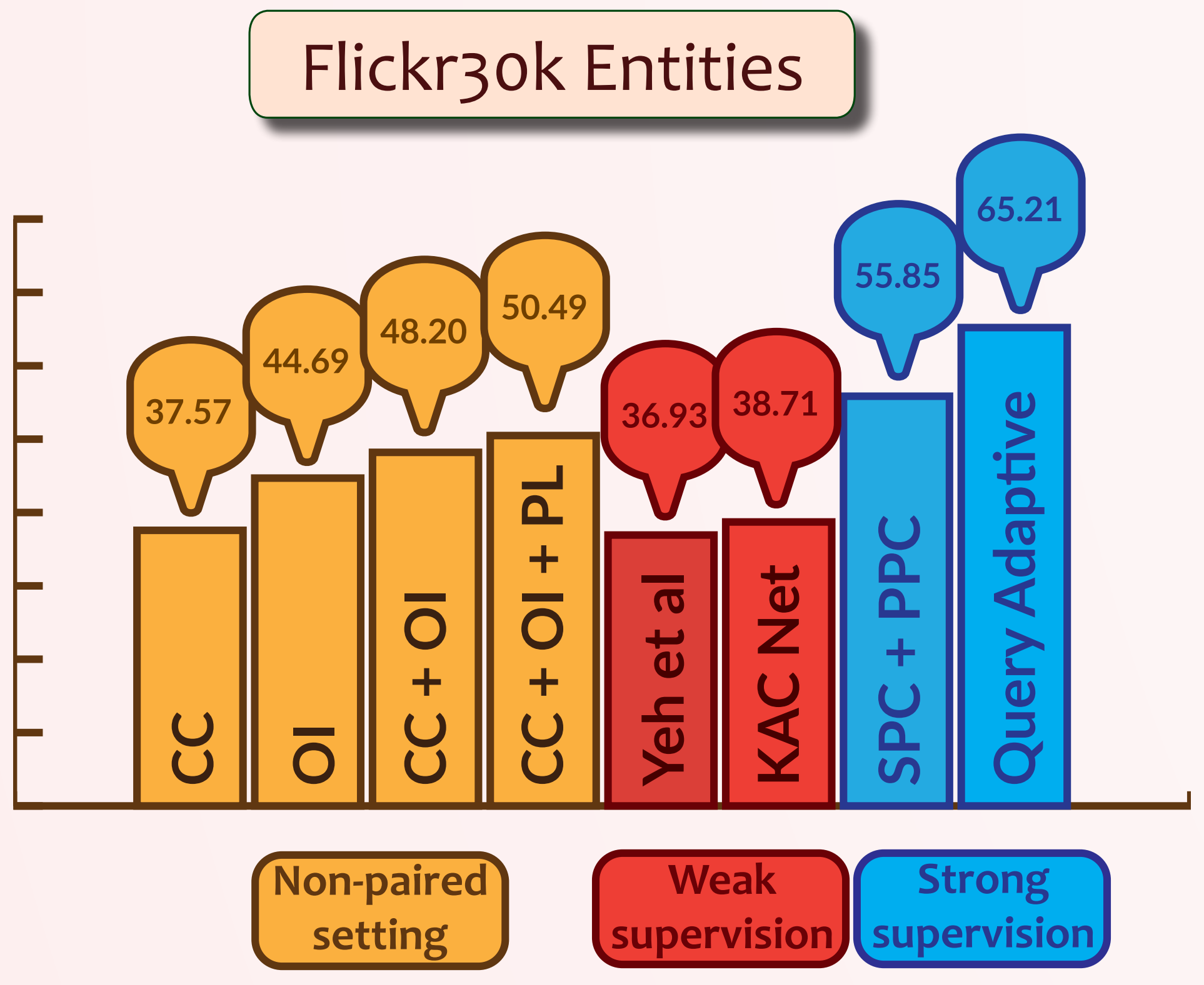


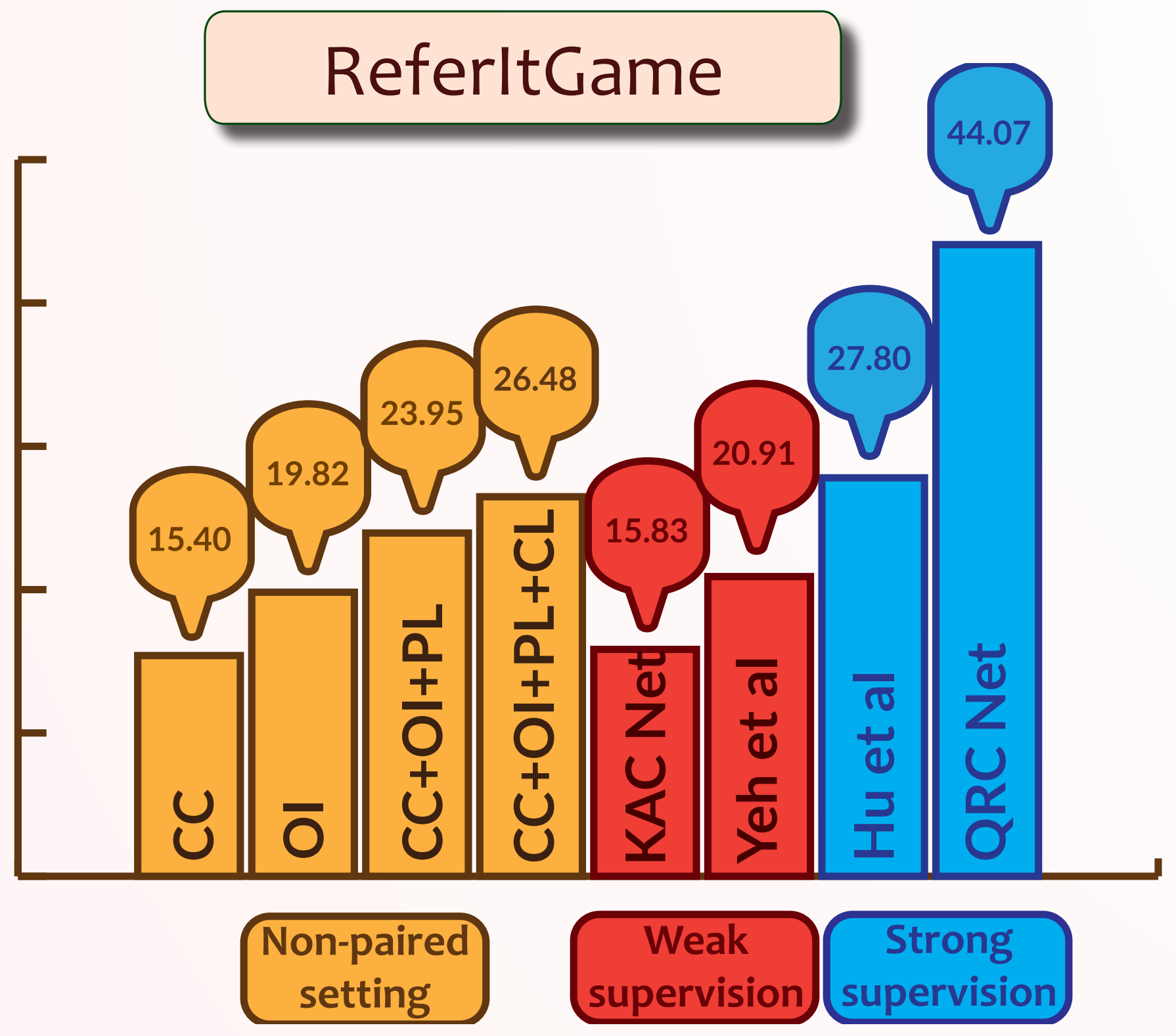
3 Localize phrase

- Select/aggregate from instances most sematically similar to query phrase
- Tie-breakers:
  - Select random
  - Select largest
  - Select most confident
  - Union of bounding boxes
  - Consensus



Experimental Results





Example Ouput

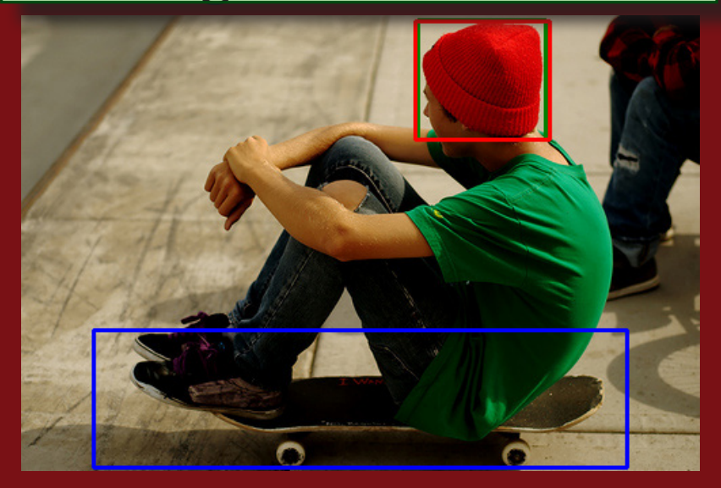
COCO+ OI

vs.

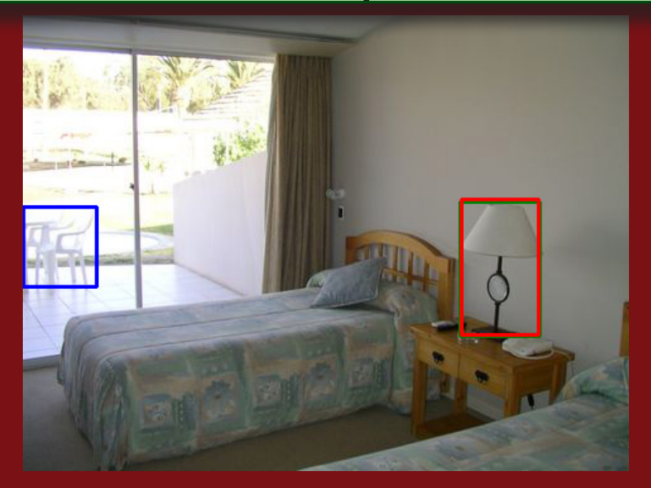
COCO

skyscrapers

bright red beanie

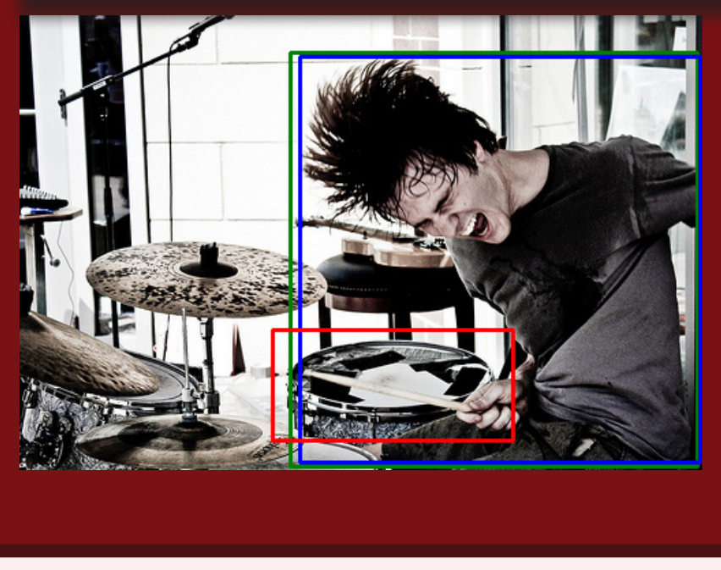


lamp



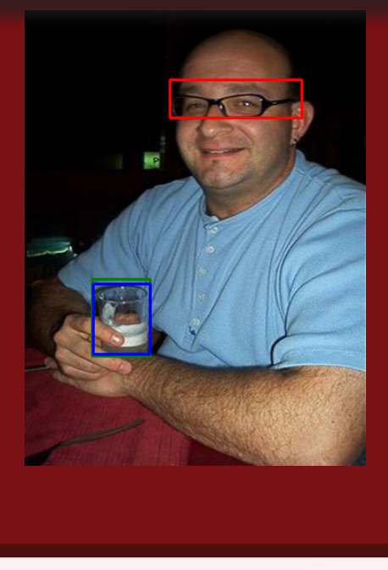
hotel door

A very excited drummer

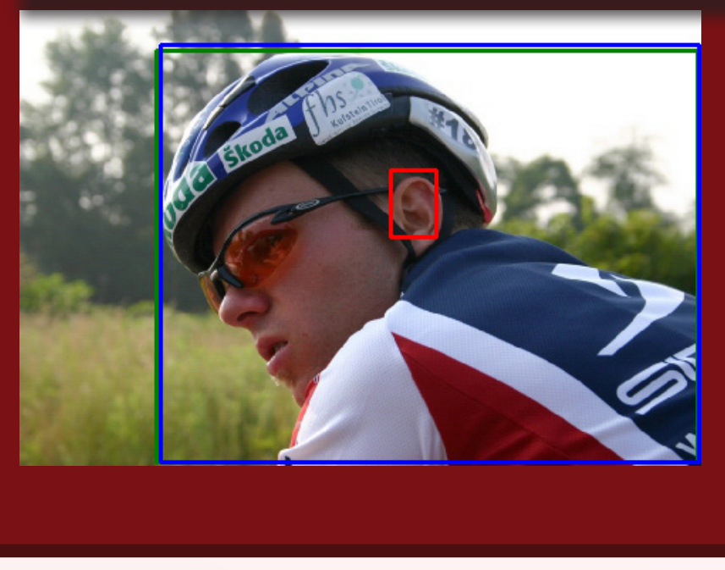


A five member band

glass



pick his nose

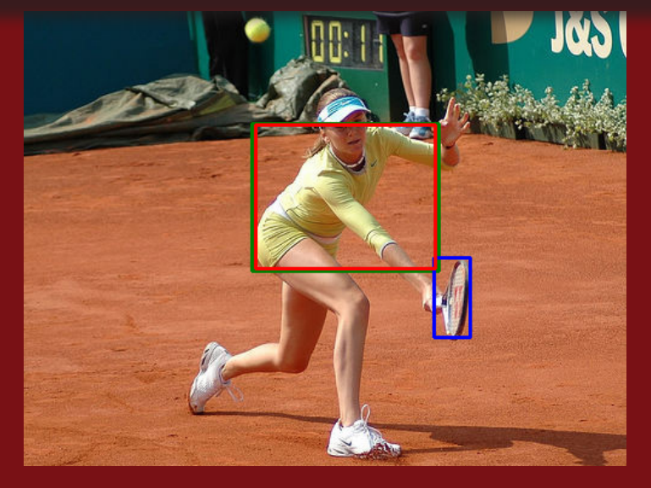


WITH COLOUR

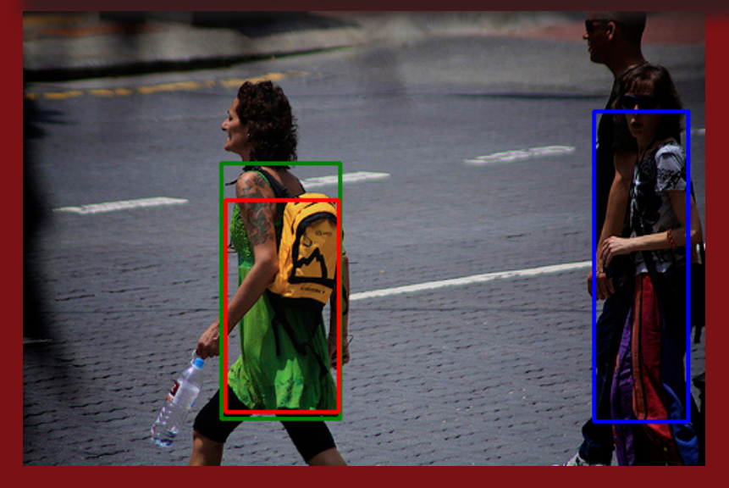
vs.

NO COLOUR

a yellow tennis suit



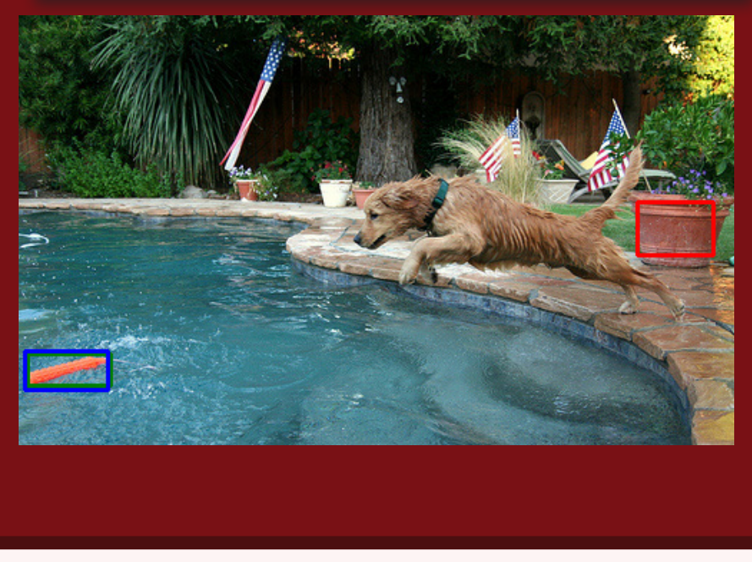
a long green shirt



pink blanket

trees

a red toy



Older male

guy in yellow shirt

bag below women in orange

Discussion

- Non paired setting can be used as a strong baseline for phrase localization (or other V&L tasks)
- Paired data should be used more effectively, *on top of* what can be achieved with simpler methods without paired data
- Need to understand datasets better and not just blindly running complex models
- General, human-like AI: Better generalisation to different tasks