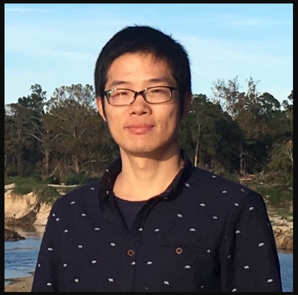
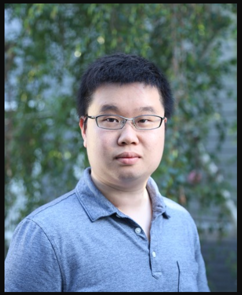
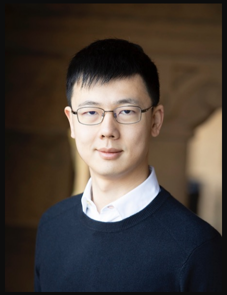
CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

Bongard-HOI: Benchmarking Few-Shot Visual Reasoning for Human-Object Interactions

Huaizu Jiang[1*], Xiaojian Ma[2*], Weili Nie[3], Zhiding Yu[3],
Yuke Zhu[3,4], Anima Anandkumar[3,5]
[1]Northeastern University [2]UCLA [3]NVIDIA [4]UT Austin [5]Caltech

Visual Concept Learning

Bicycle

...

Ride bicycle

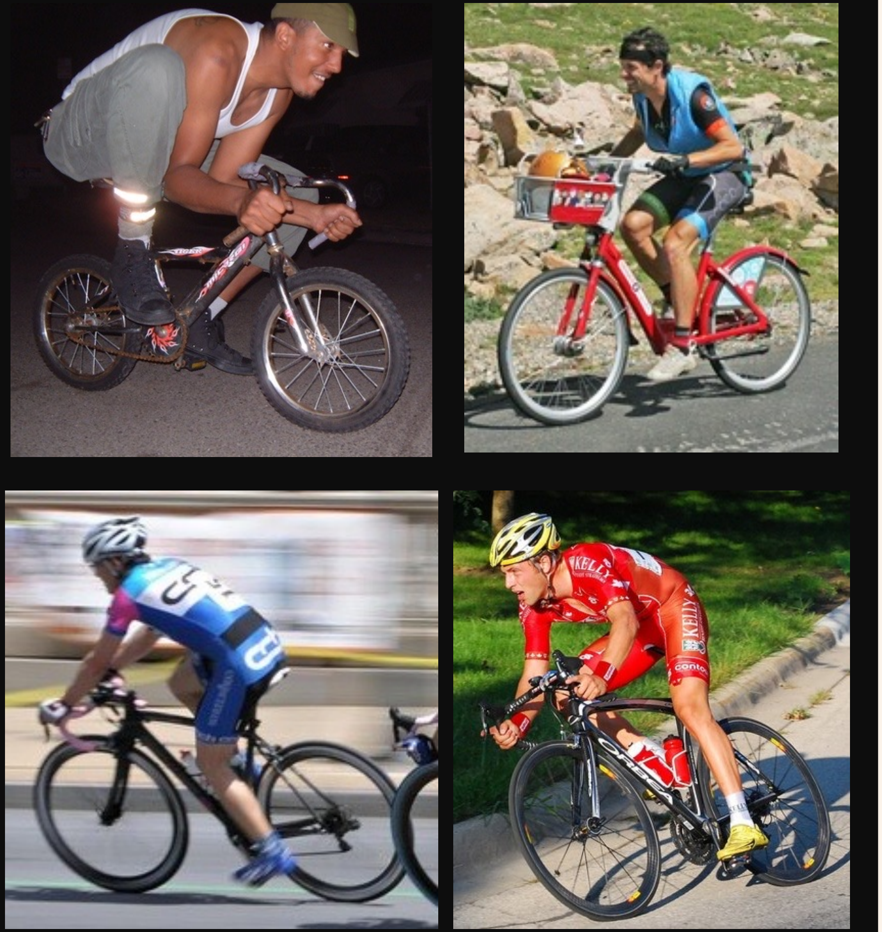
...

- ☹ Need large amount of training data
- ☹ Hard to generalize beyond the training concepts

Bongard-HOI Benchmark

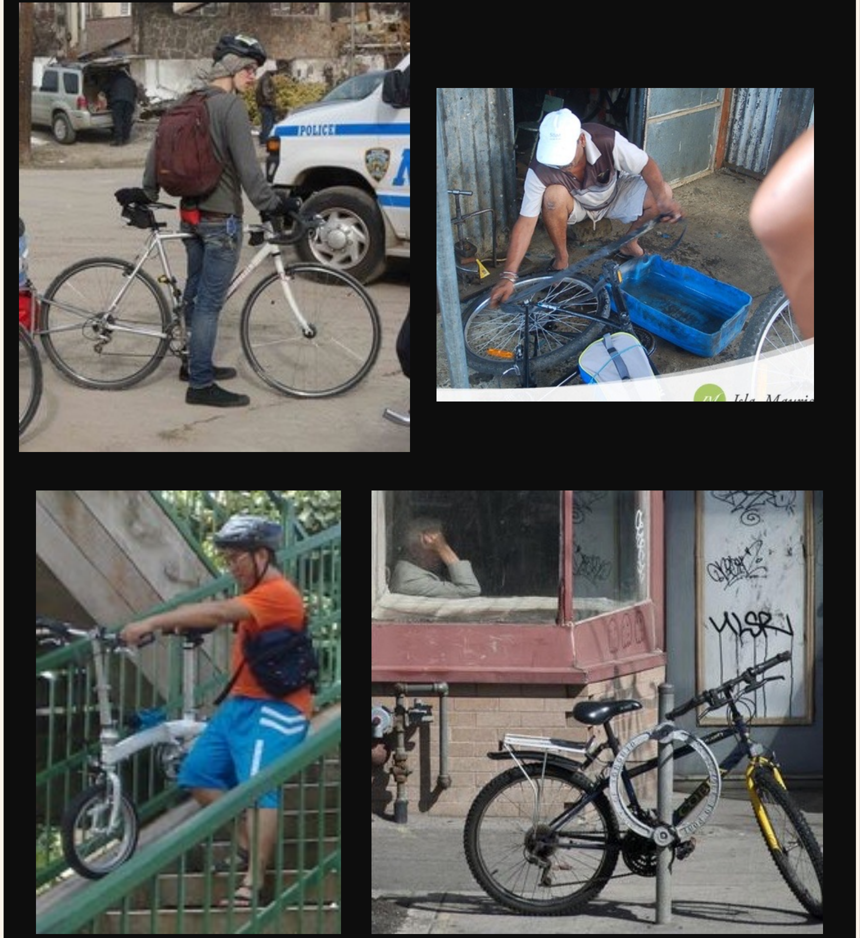
Positive Examples

ride bicycle



Negative Examples

not ride bicycle



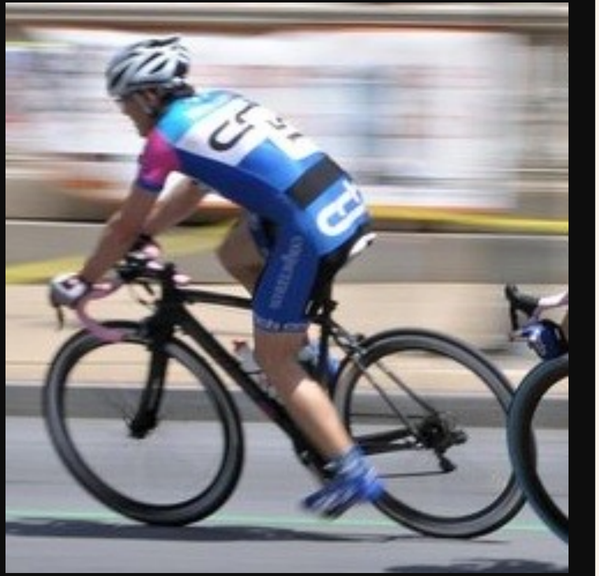
Query Images

positive

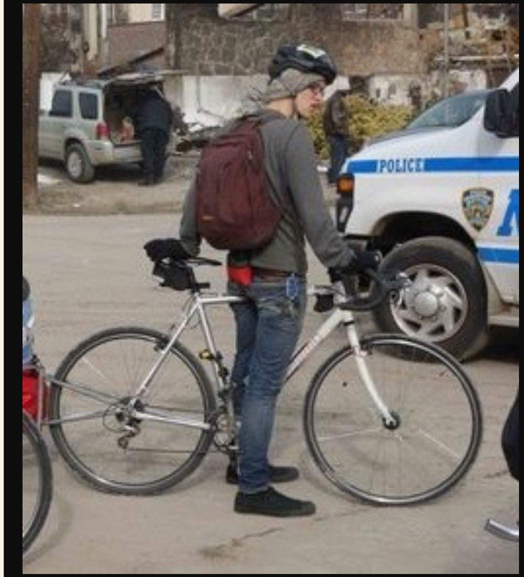
negative

Hard Negatives in Bongard-HOI

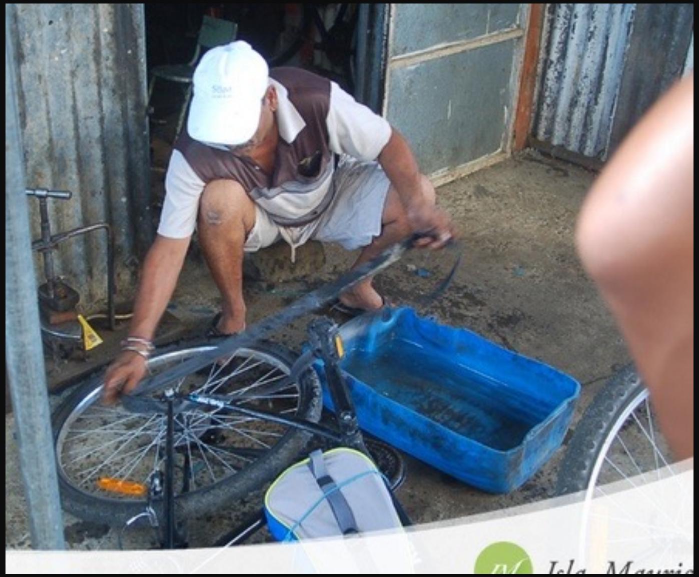
person
ride
bicycle



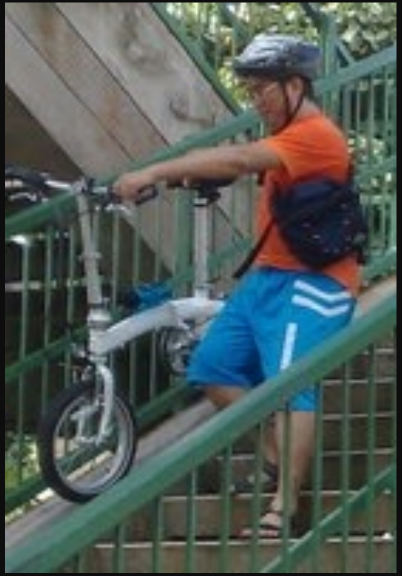
person
straddle
bicycle



person
repair
bicycle

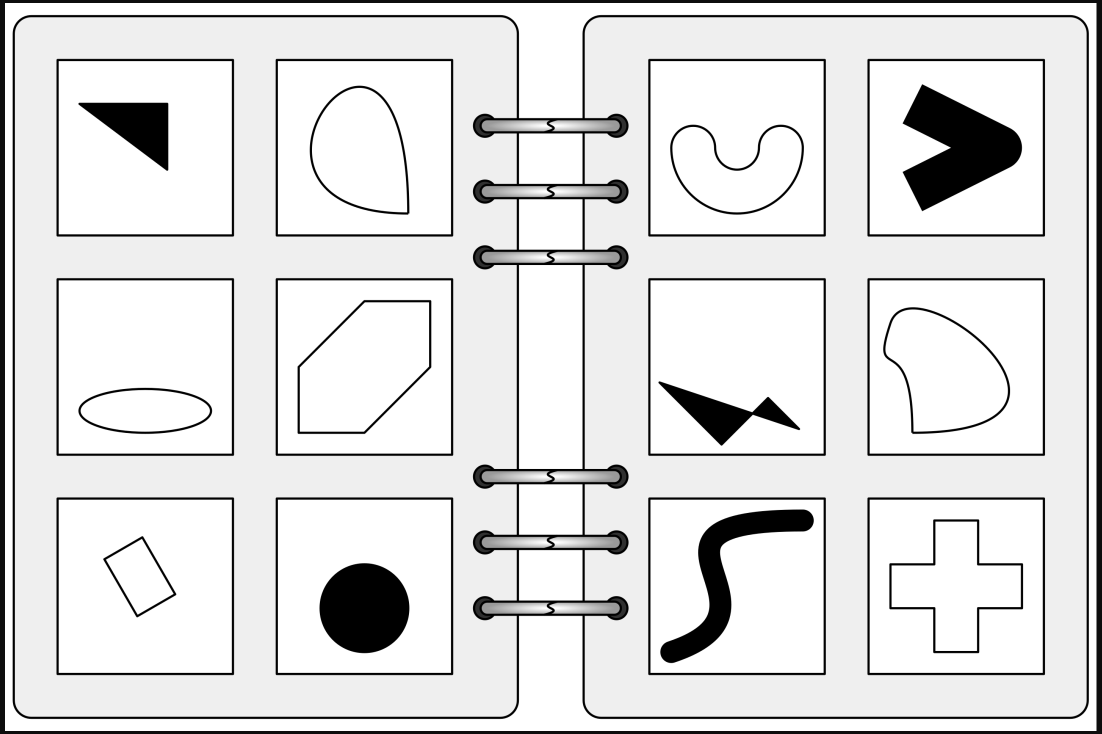


person
walk
bicycle

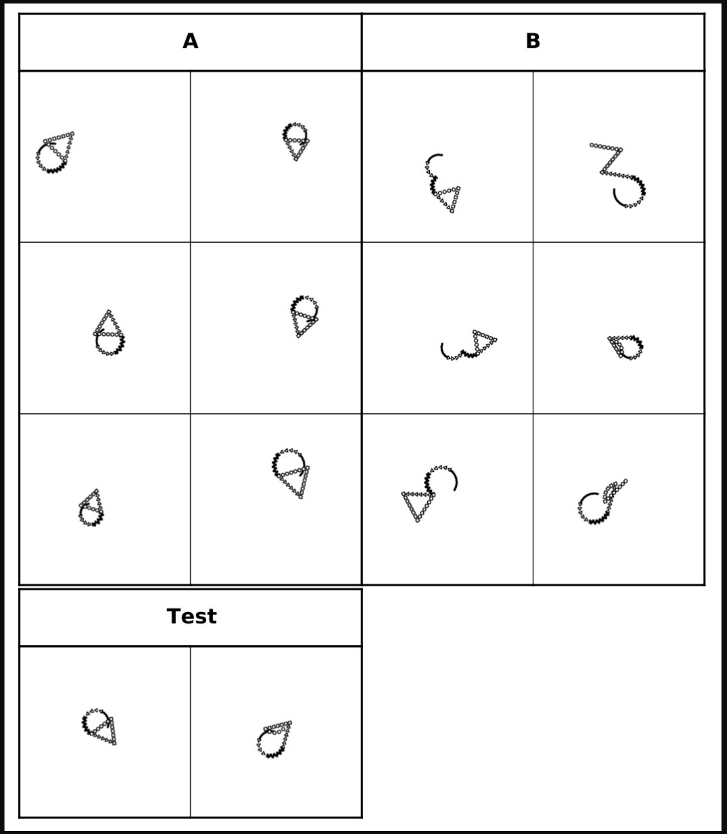


Simple visual recognition is not sufficient.
Visual reasoning (e.g., few-shot learning, context reasoning) of the interactions is needed.

Inspirations from Cognitive Science



Original Bongard problems
[Bongard, 1970.]



Bongard-LOGO
[Nie et al., NeurIPS 2021]

Different Types of Generalization

Training set

sit_on bed

straddle bicycle

hug person

wash car

Test set

wash bicycle

seen action,
seen object

sit_on bench

seen action,
unseen object

greet person

unseen action,
seen object

shear sheep

unseen action,
unseen object

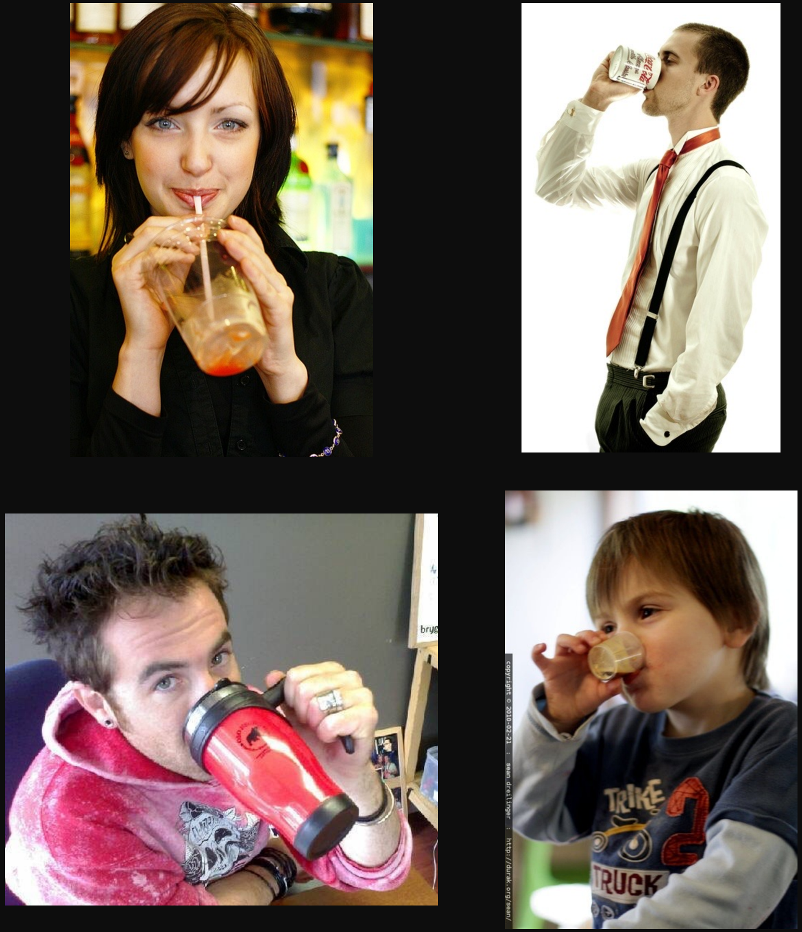
→

Increasing difficulty

Context-Dependent Reasoning

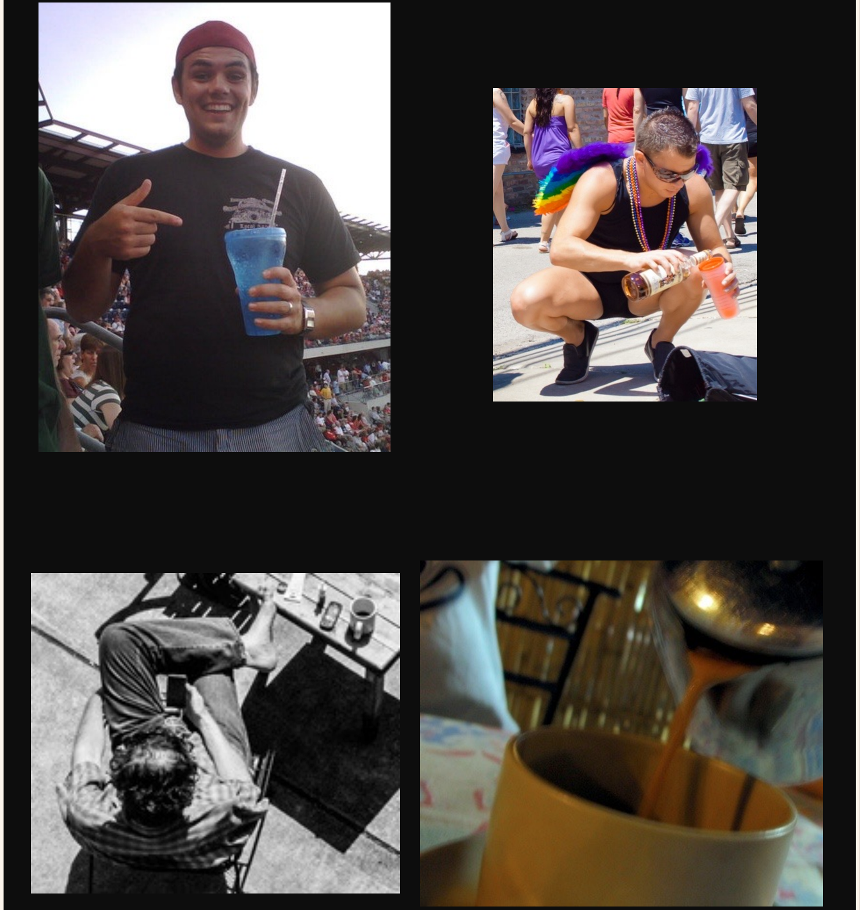
Positive Examples

drink_with cup



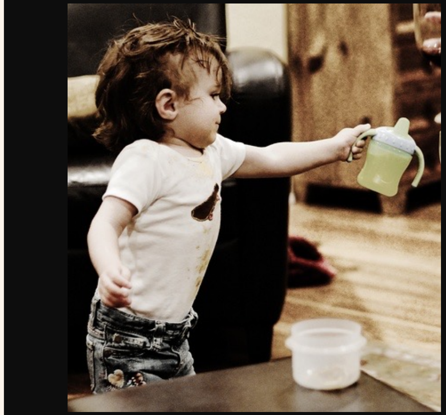
Hard Negative Examples

not drink_with cup



Query Images

negative



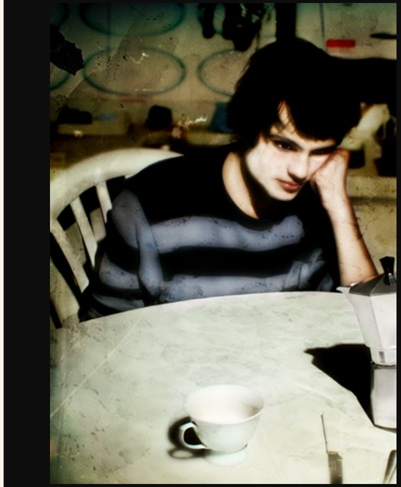
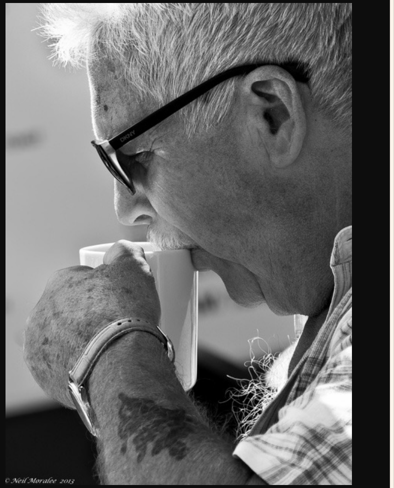
Context-Dependent Reasoning

Positive Examples

hold cup

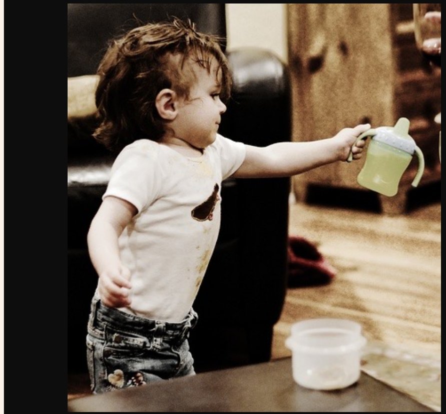
Hard Negative Examples

not hold cup

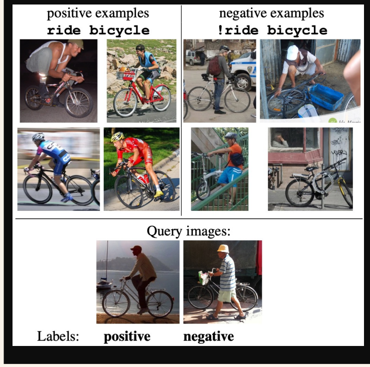



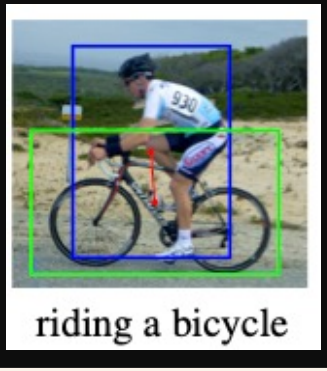
Query Images

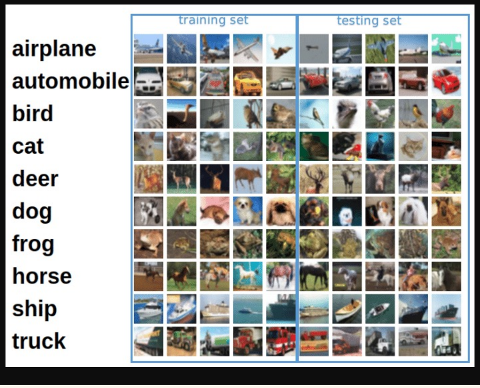
positive

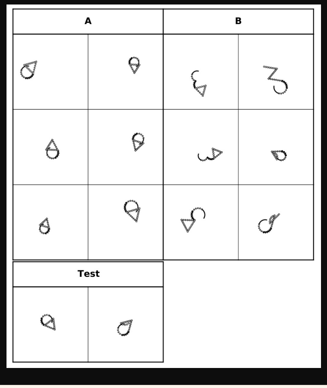


Comparisons with Other Benchmarks



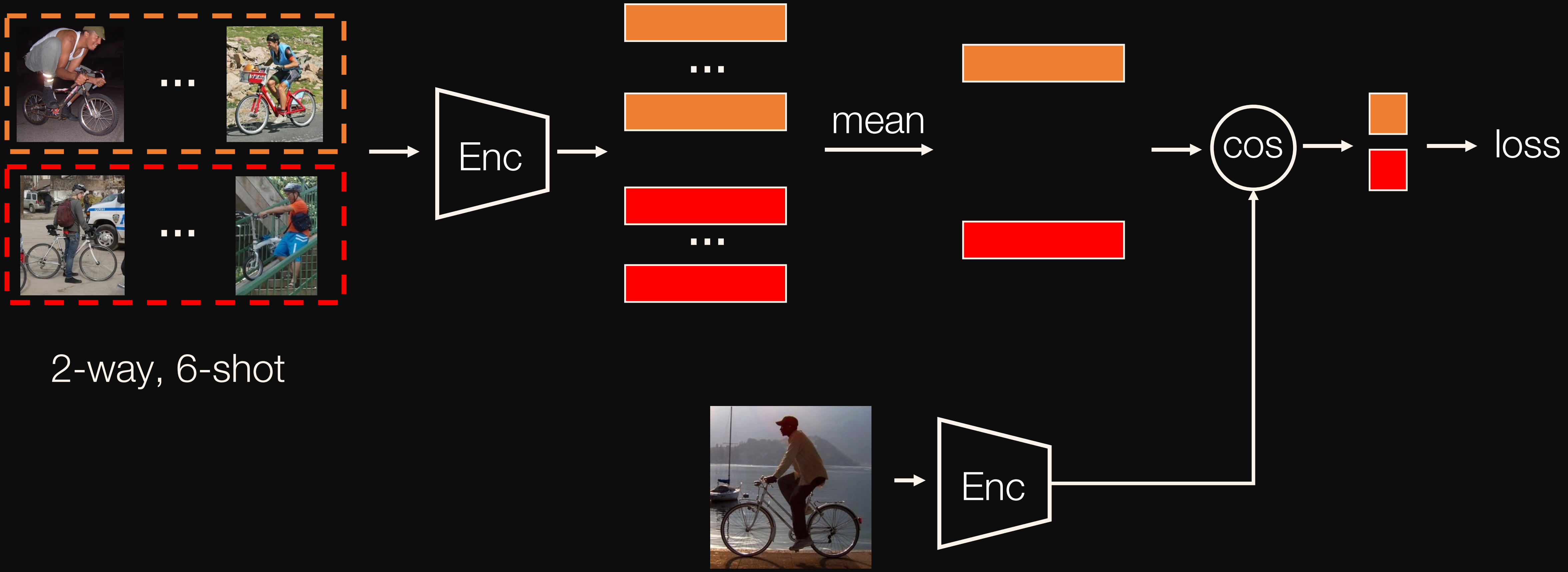






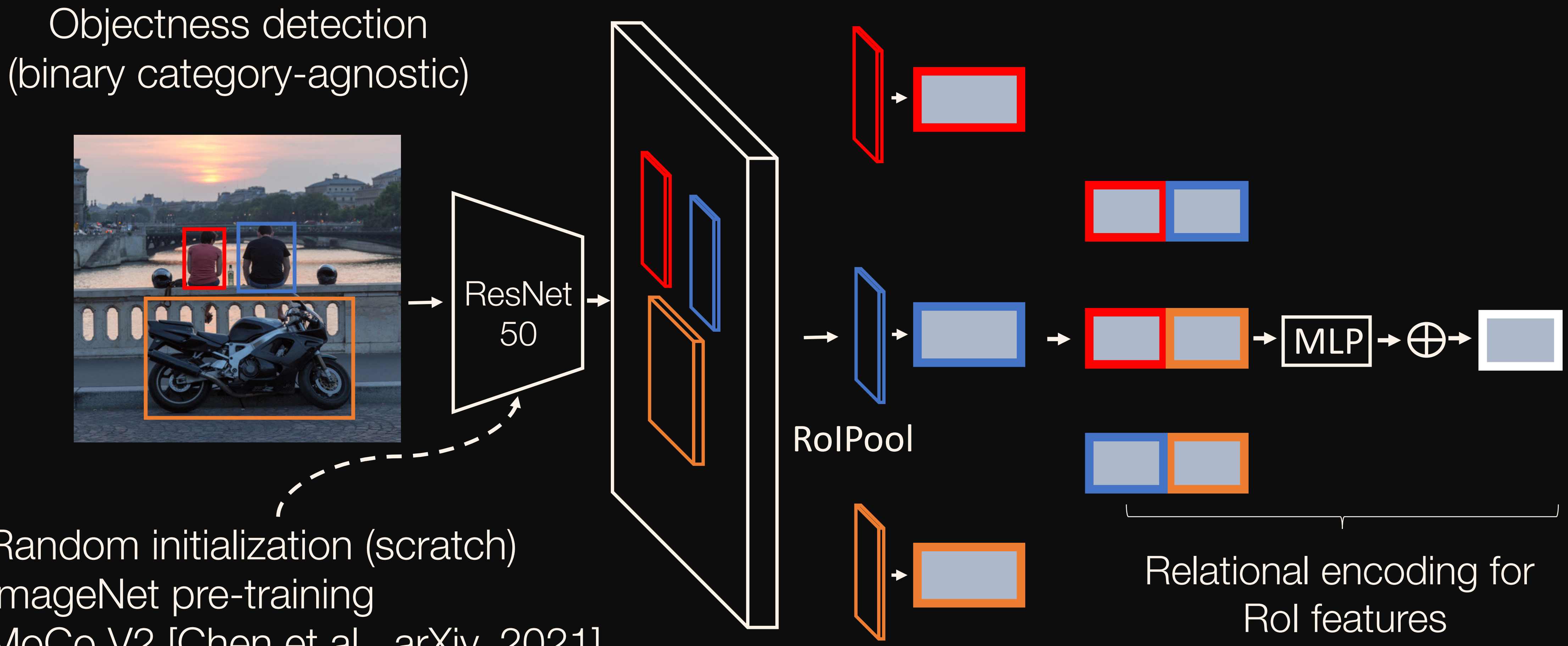
	Bongard-HOI	HOI detection	miniImageNet	Bongard-LOGO
Natural images	✓	✓	✓	✗
Hard negatives	✓	✗	✗	✗
Compositional concept	✓	✓	✗	✓
Few-shot learning	✓	✗	✓	✓
Ctx.-dependent reasoning	✓	✗	✗	✓
Generalization types	✓	✗	✗	✓

Meta-Learning for Bongard-HOI



[Chen et al., Meta-Baseline. ICCV 2021]

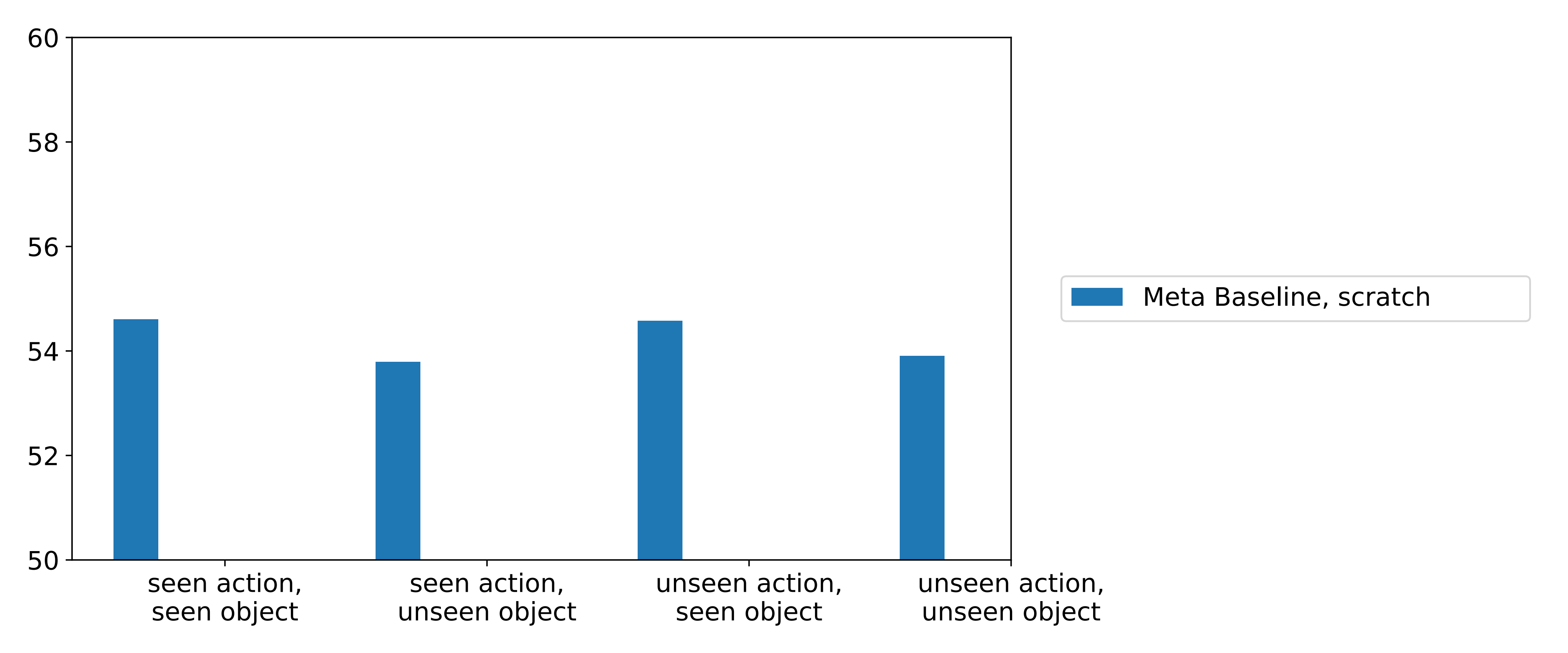
Image Encoding with Relational Network



- Random initialization (scratch)
- ImageNet pre-training
- MoCo V2 [Chen et al., arXiv, 2021]

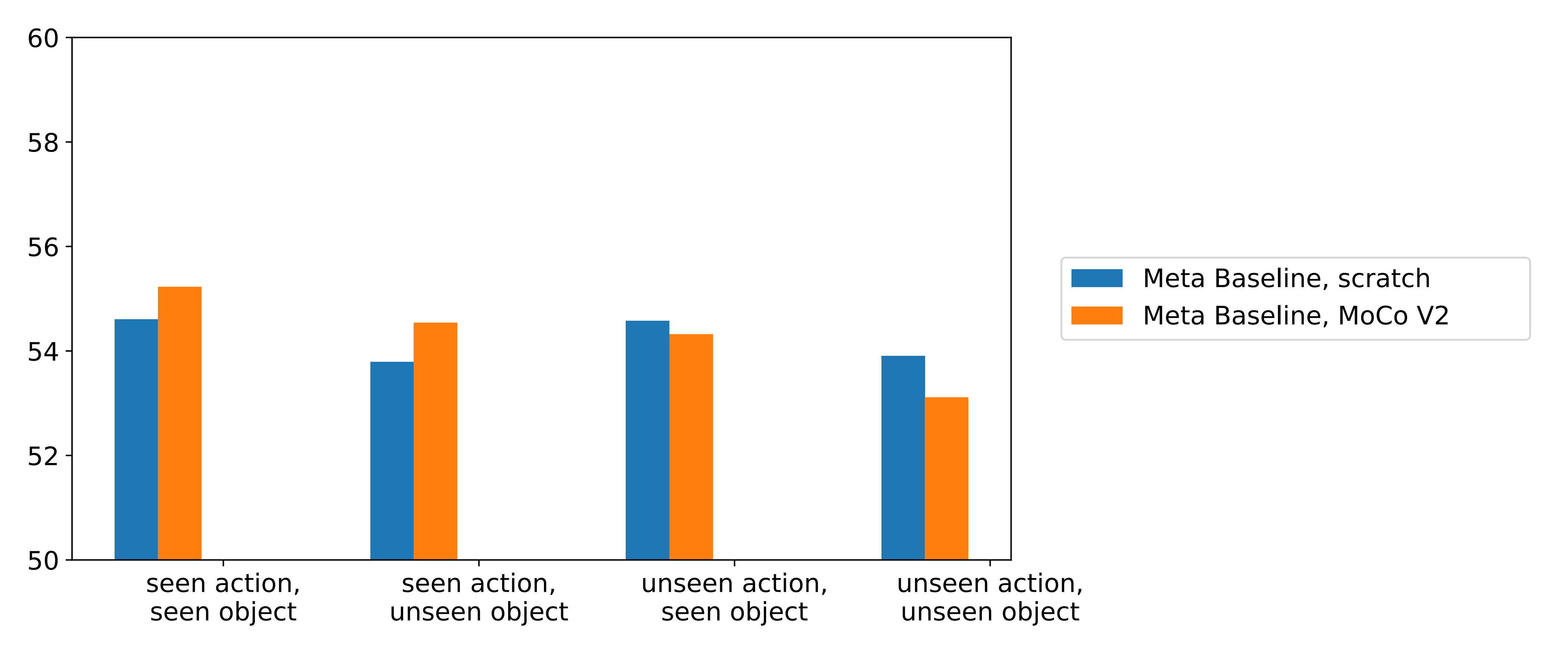
[Santoro et al., NeurIPS 2017]

Comparisons of Meta Learning Models



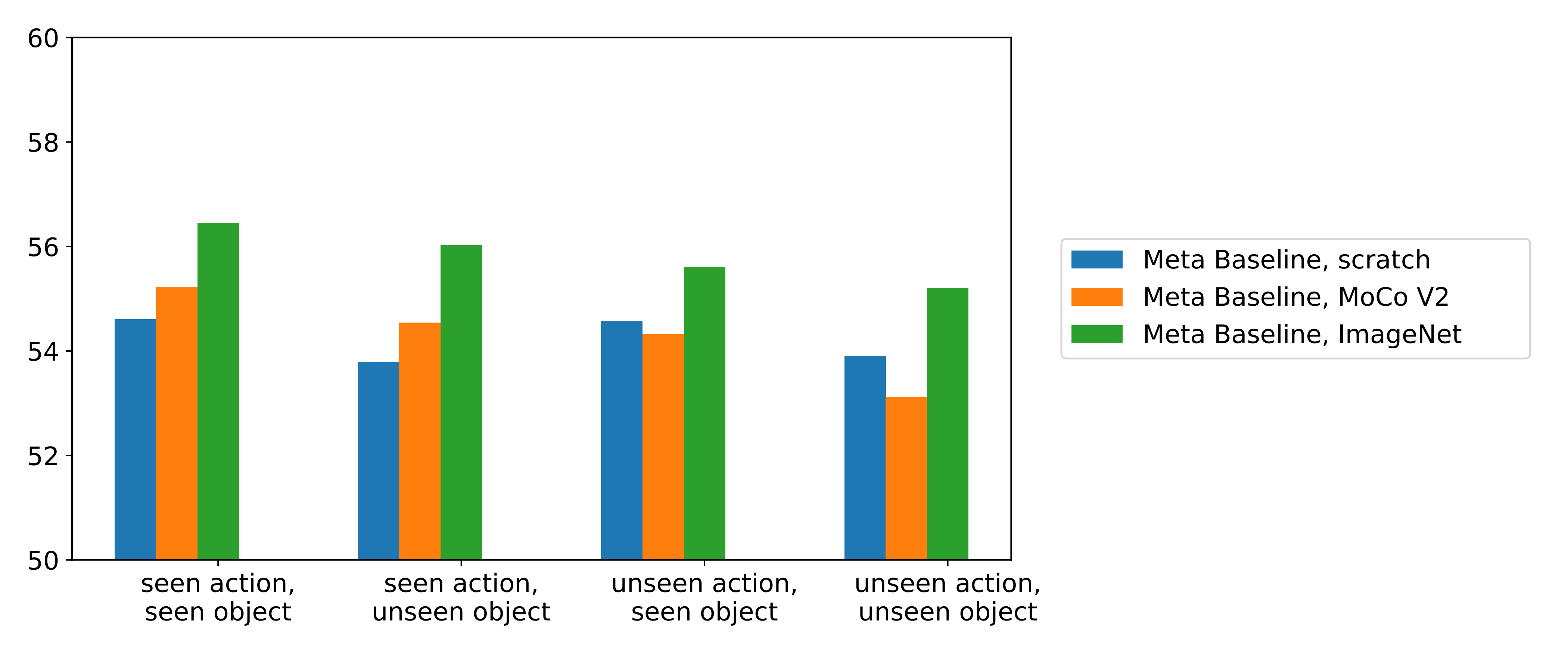
Increasing difficulty

Comparisons of Meta Learning Models



Increasing difficulty

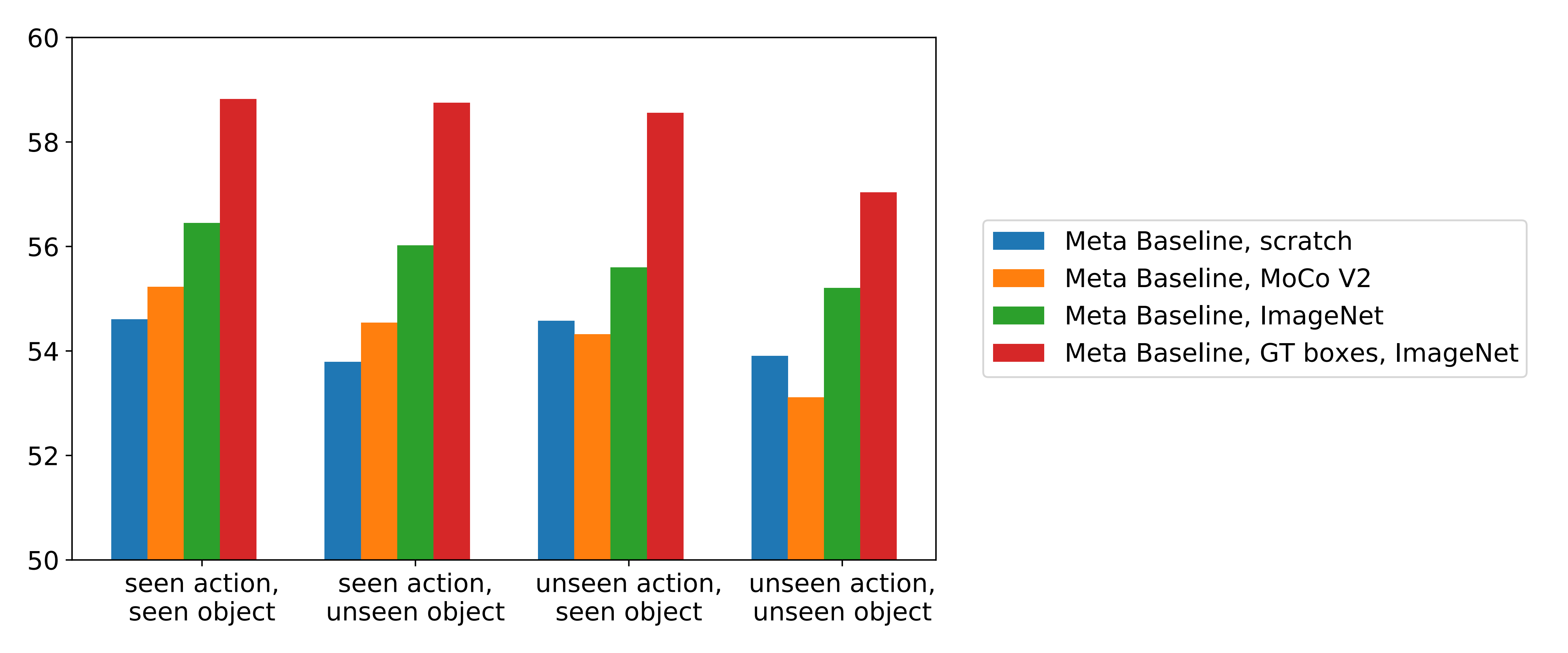
Comparisons of Meta Learning Models



→

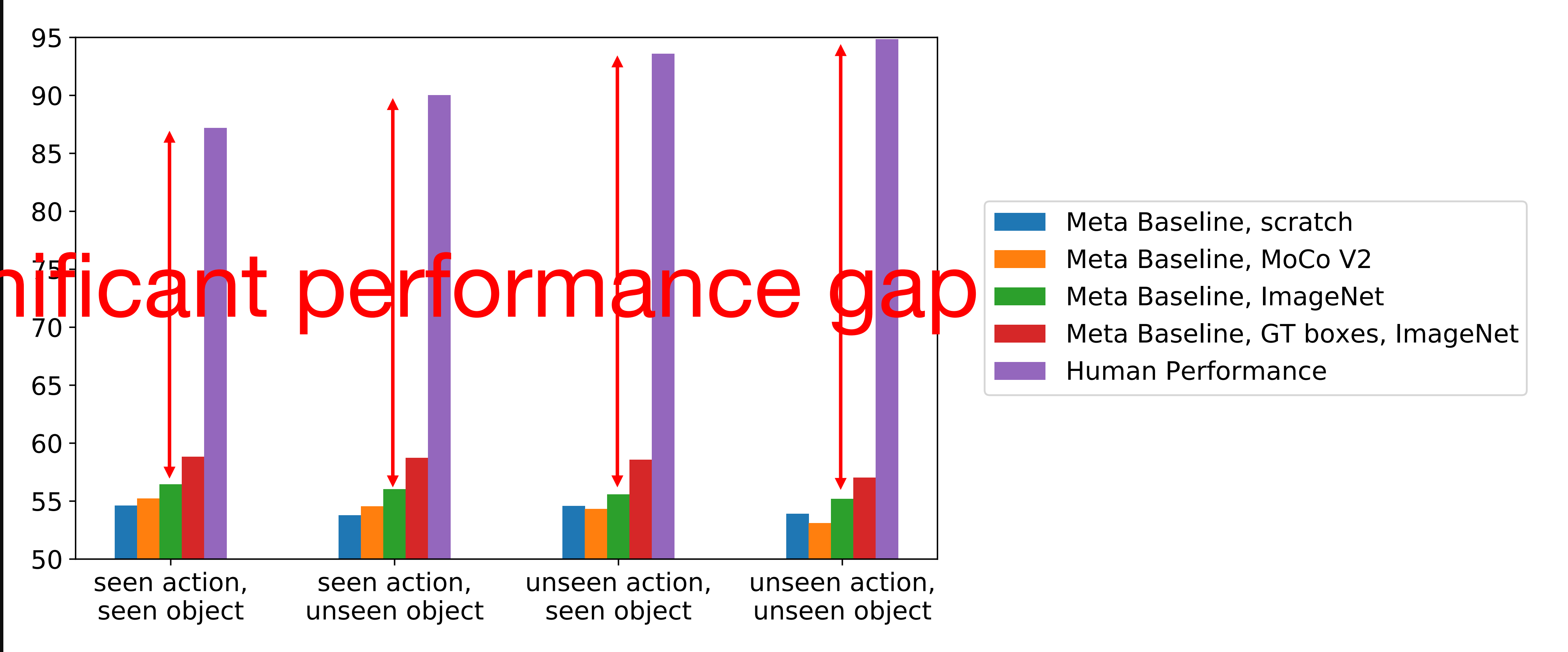
Increasing difficulty

Importance of Holistic Visual Perception and Reasoning



Increasing difficulty

Comparisons with Human Accuracy



significant performance gap

Take-home Messages of Bongard-HOI

- A new benchmark about HOI, highlighting visual reasoning
  - Few-shot learning
  - Context reasoning
  - Generalization beyond training concepts
  - …
- Meta-learning models do not work well enough
  - Pre-training is helpful
  - Holistic visual perception and reasoning is essential
- There exit huge gap w.r.t human performance

Data and Code

Poster: 36b