Image as a single label

Image Source: ImageNet

Image as an object set

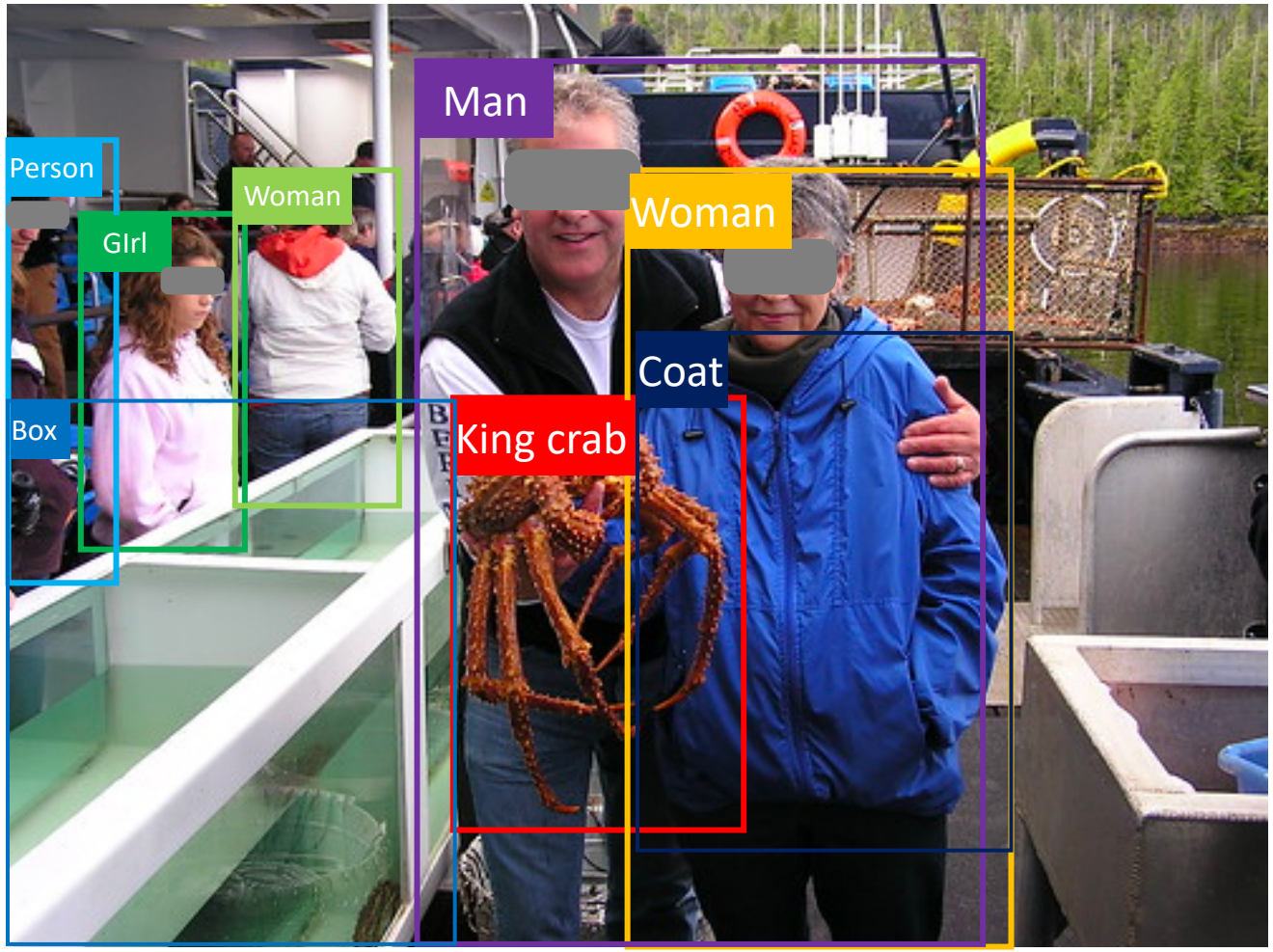


Image Source: ImageNet

Image as a scene graph

Relationships:

"Woman look at box"

"Man hold king crab"

"Woman wear coat"

"Man embrace woman"

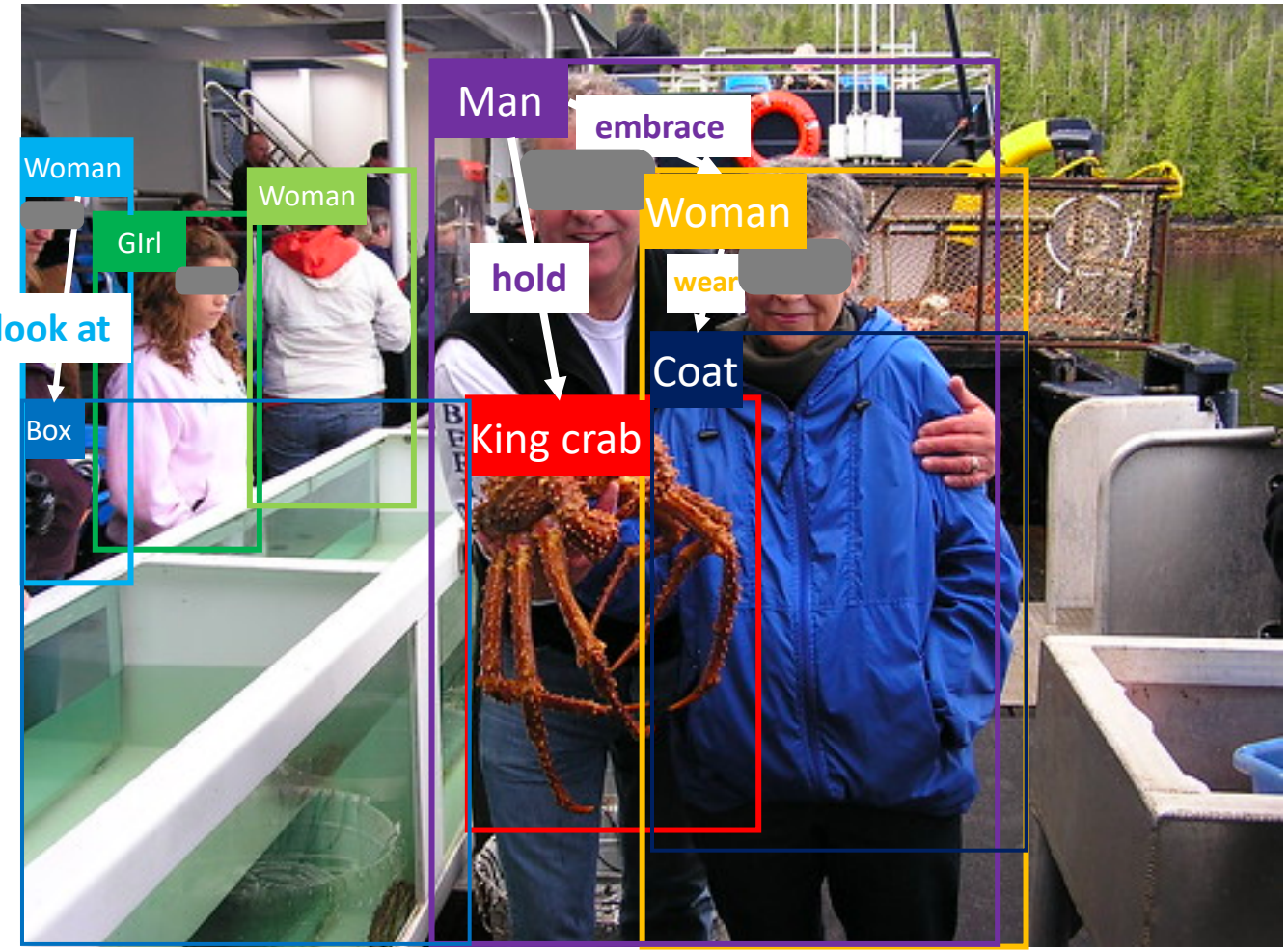


Image Source: ImageNet

Image as a scene graph

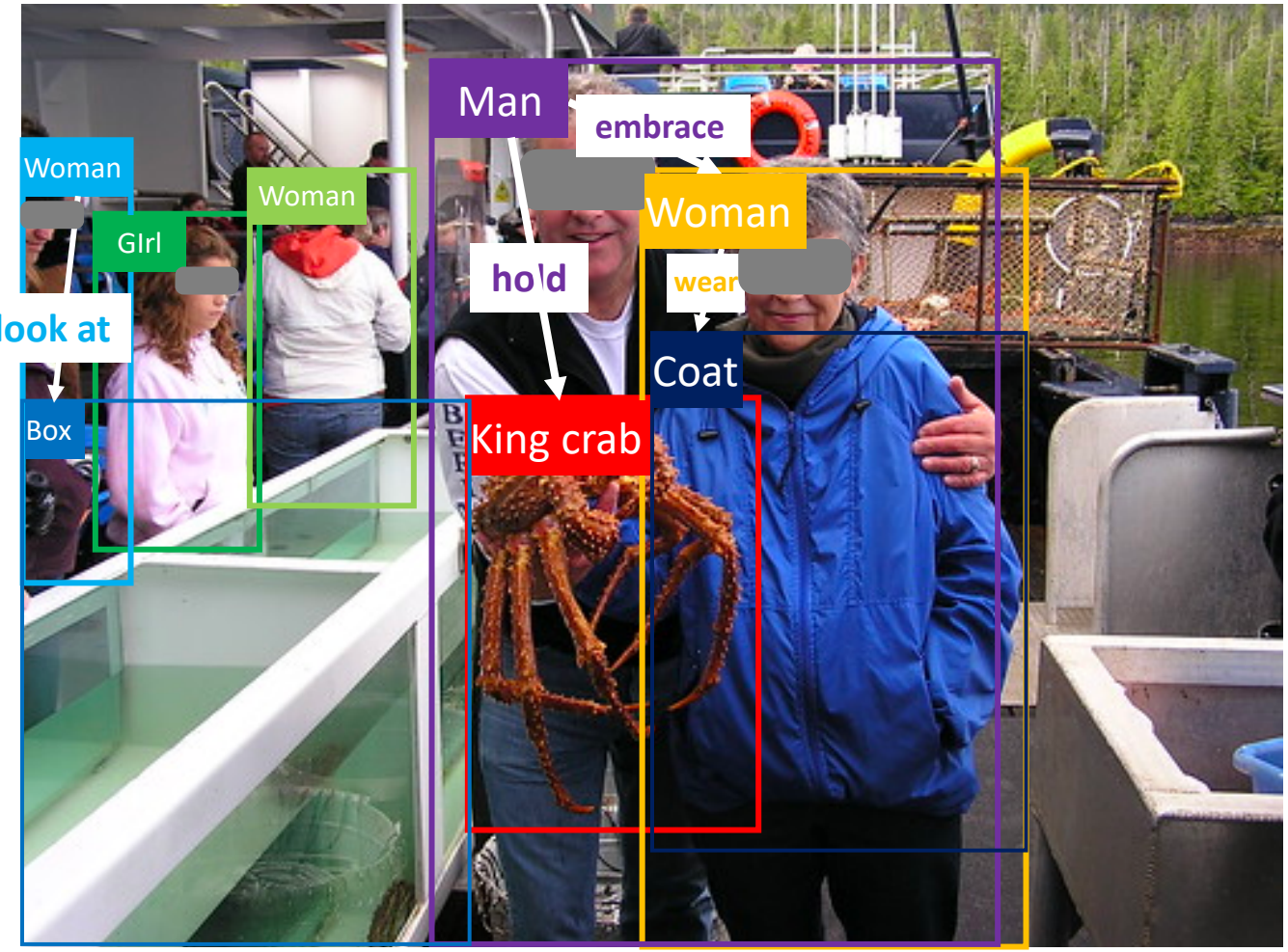
Relationships:

"Woman look at box"

"Man hold king crab"

"Woman wear coat"

"Man embrace woman"



Attributes:

"Red king crab"

"Transparent box"

"Blue coat"

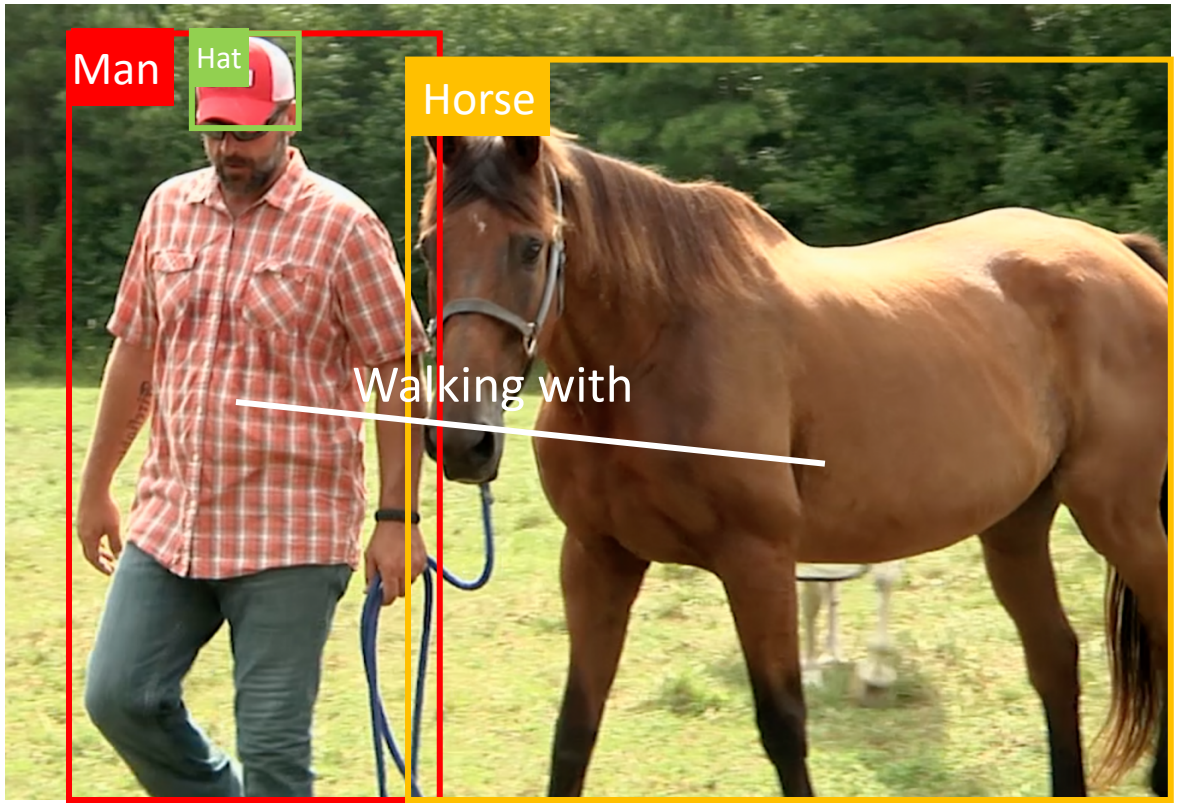
"Smiling woman"

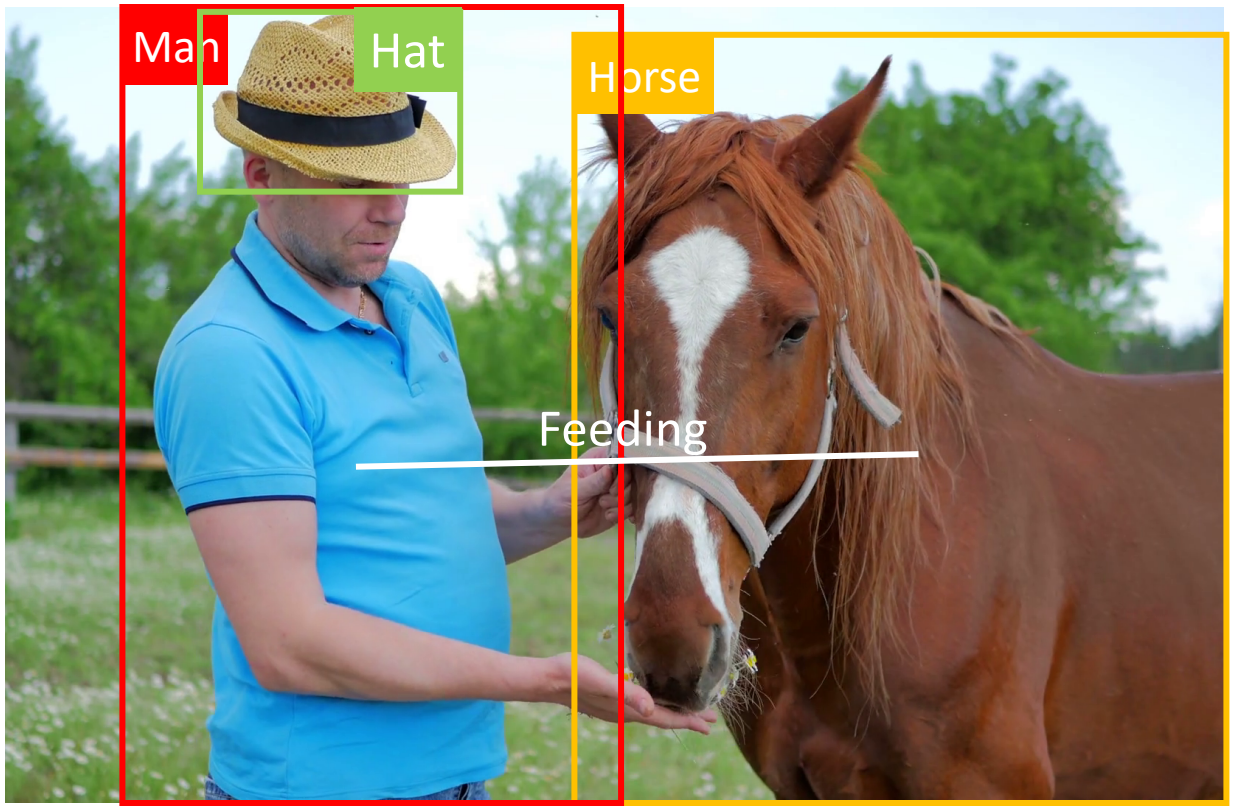
"Smiling Man"

Image Source: ImageNet

Why we need scene graph?

Distinguish images more accurately



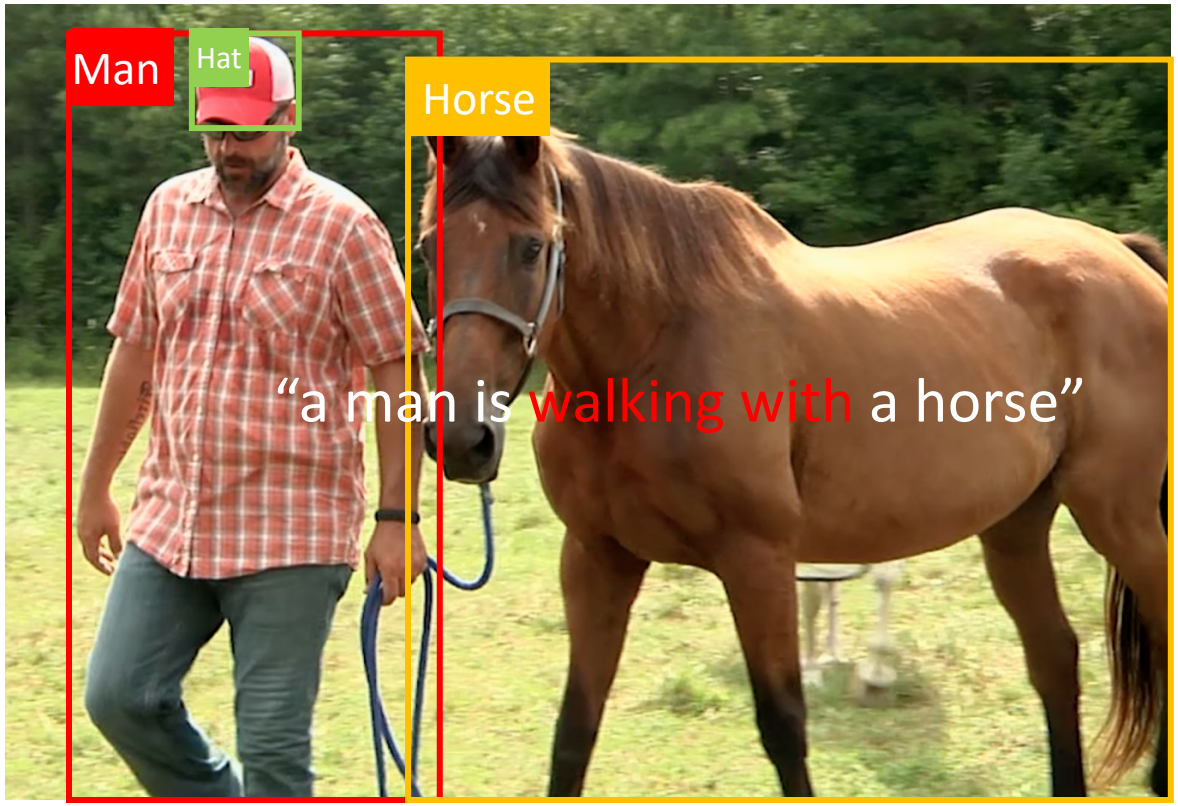


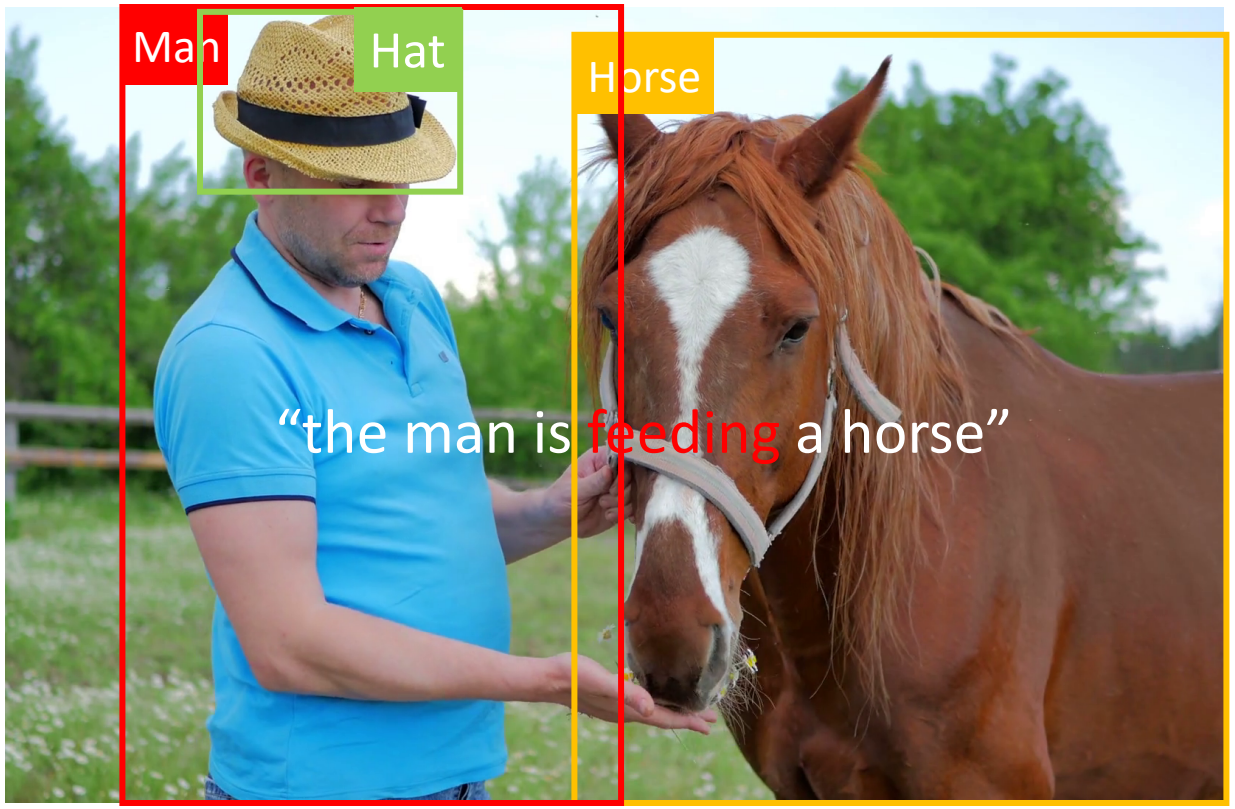
[1] Image Retrieval using Scene Graphs. Johnson et al. CVPR 2015

Left: https://cals.ncsu.edu/wp-content/uploads/2016/08/horse-1500x931.png
Rigth: https://www.videoblocks.com/video/the-man-in-hat-feed-a-brown-horse-with-flowers-on-the-meadow-supmox_3xj0tvkb67

Why we need scene graph?

Describe images more grounding



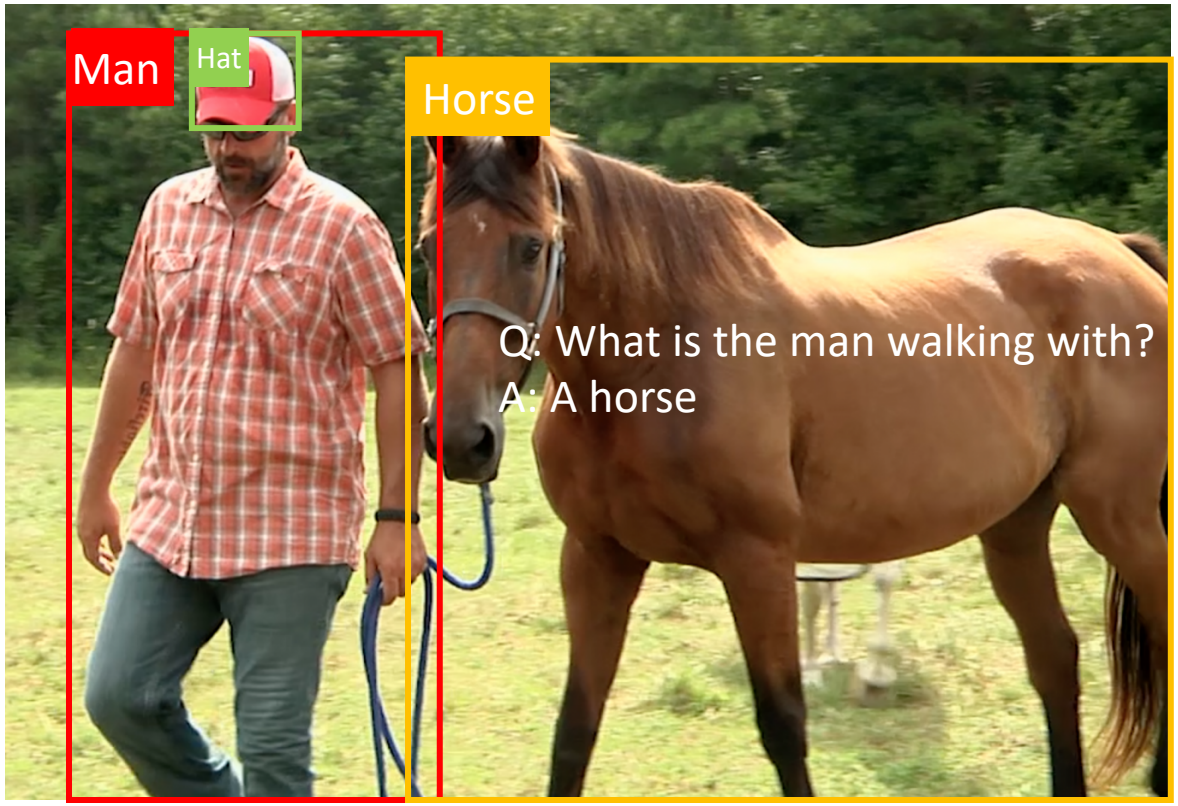


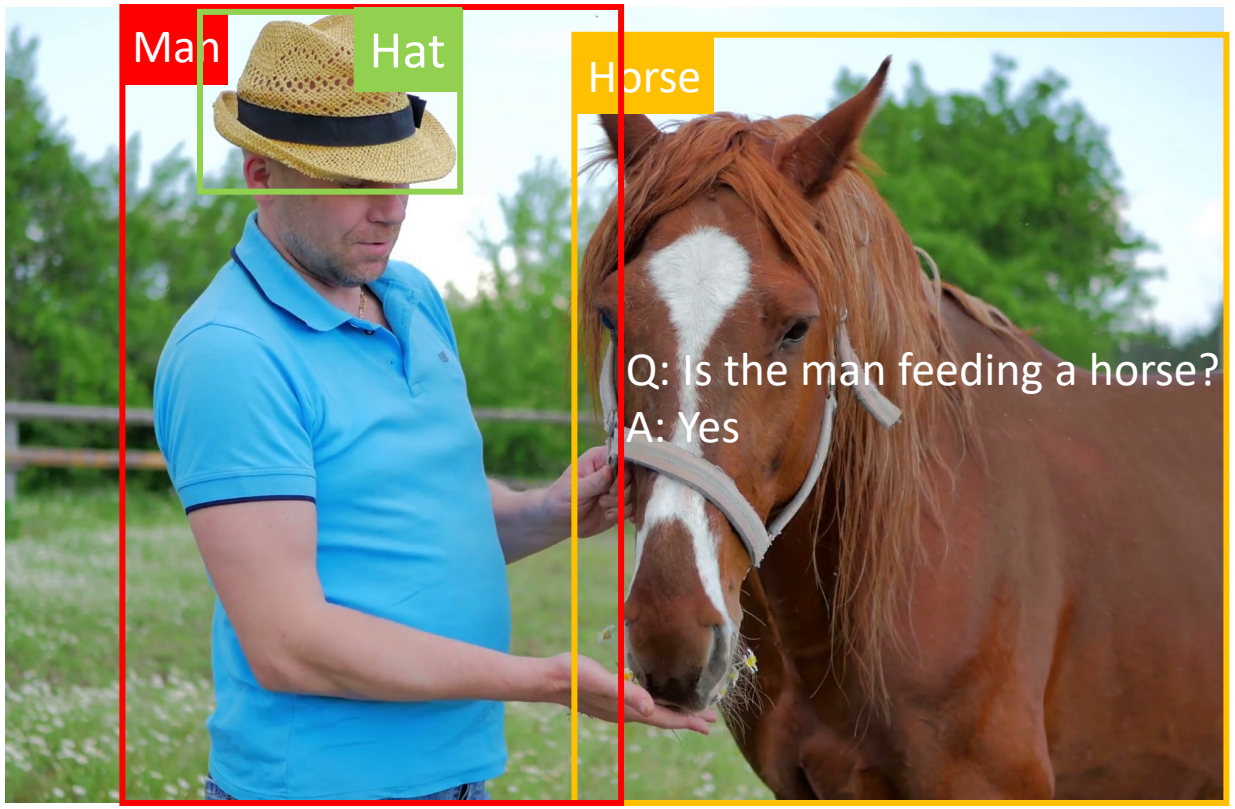
[1]. Auto-Encoding Scene Graphs for Image Captioning. Yang et al. arXiv 2018
[2]. Exploring Visual Relationship for Image Captioning. Yao et al. ECCV 2018

Left: https://cals.ncsu.edu/wp-content/uploads/2016/08/horse-1500x931.png
Rigth: https://www.videoblocks.com/video/the-man-in-hat-feed-a-brown-horse-with-flowers-on-the-meadow-supmox_3xj0tvkb67

Why we need scene graph?

Answer question more precisely





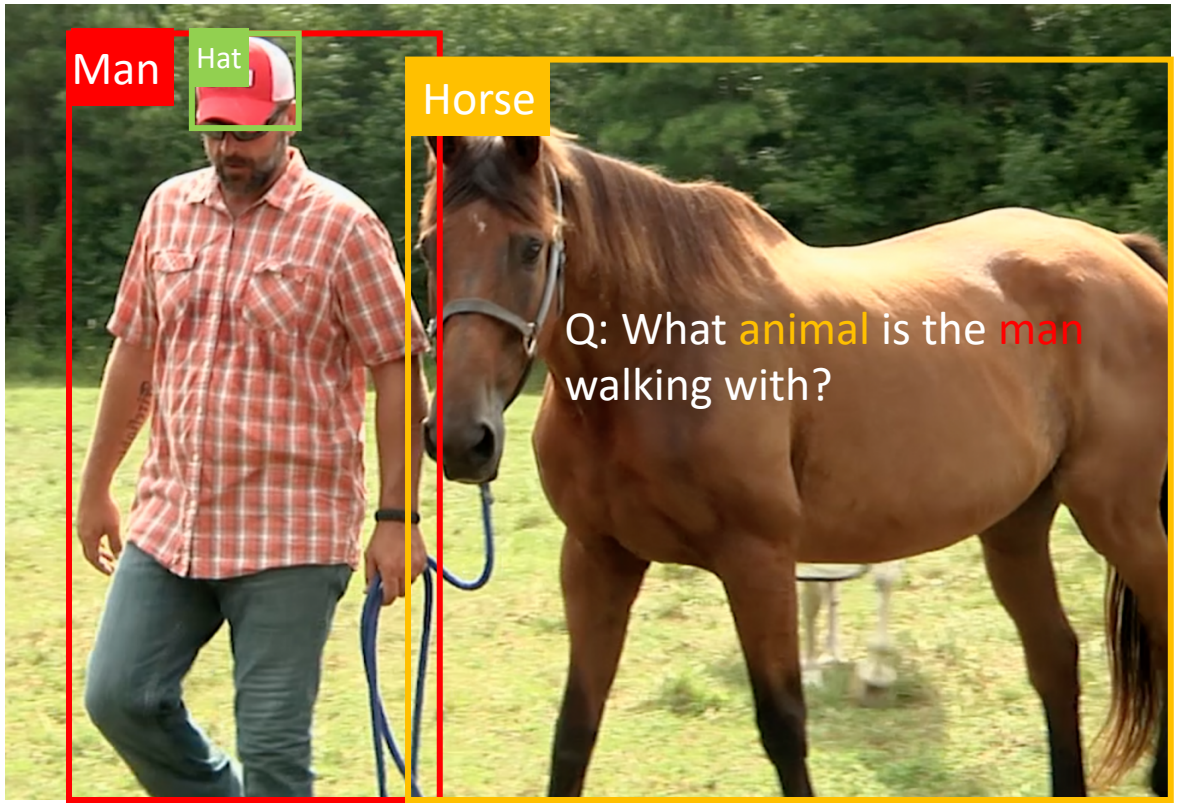
[1] Graph-Structured Representations for Visual Question Answering. Teney et al. CVPR 2017

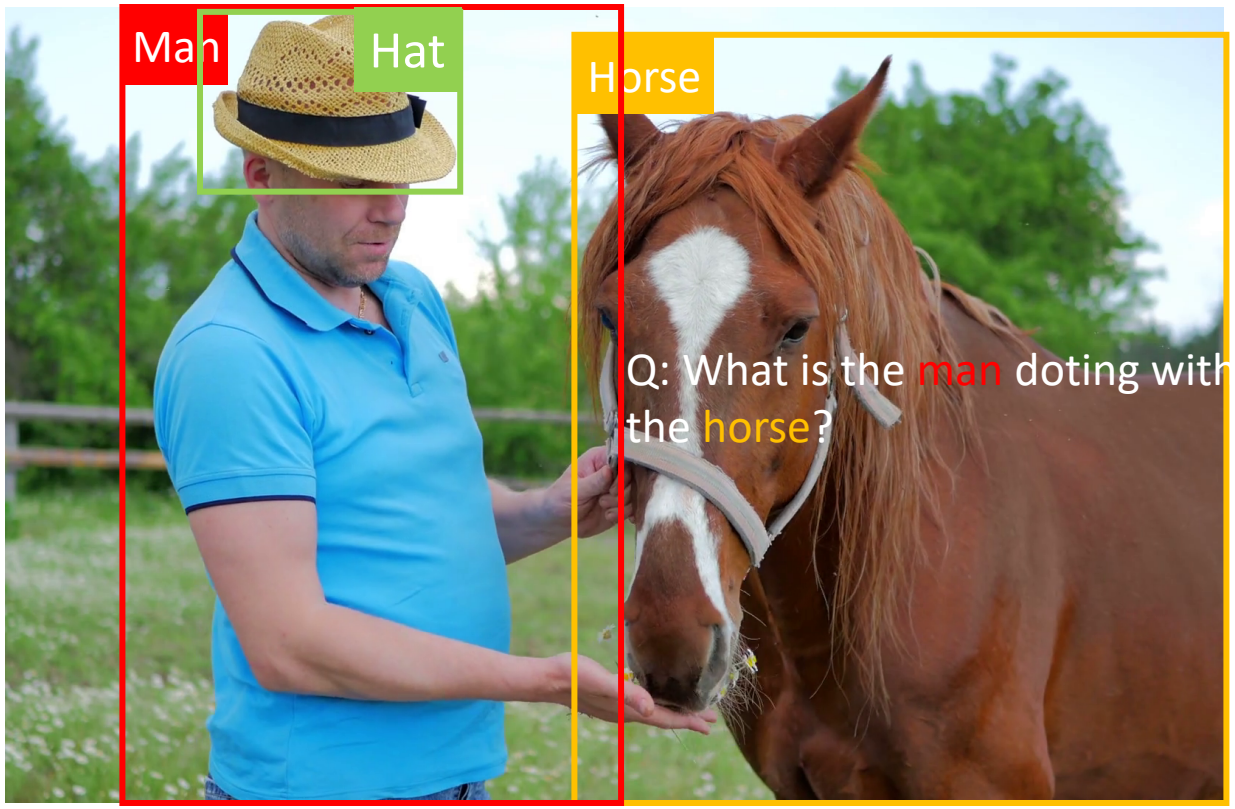
[2] Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. Yi et al. Neurips 2018

Left: https://cals.ncsu.edu/wp-content/uploads/2016/08/horse-1500x931.png
Rigth: https://www.videoblocks.com/video/the-man-in-hat-feed-a-brown-horse-with-flowers-on-the-meadow-supmox_3xj0tvkb67

Why we need scene graph?

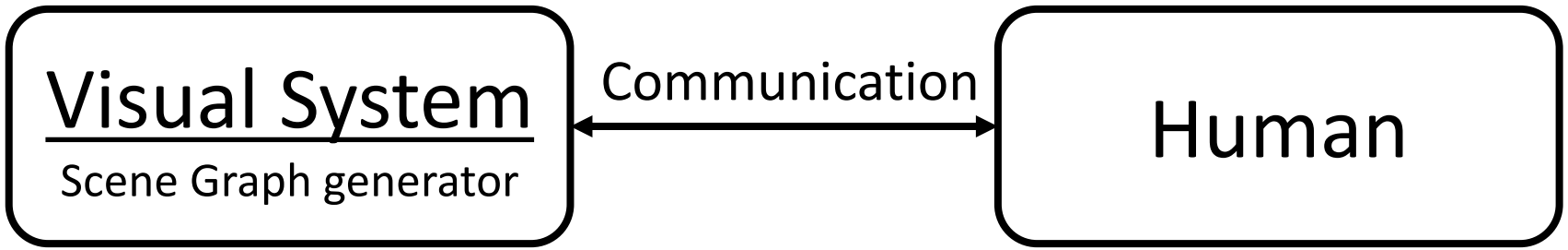
Generate questions more grounding



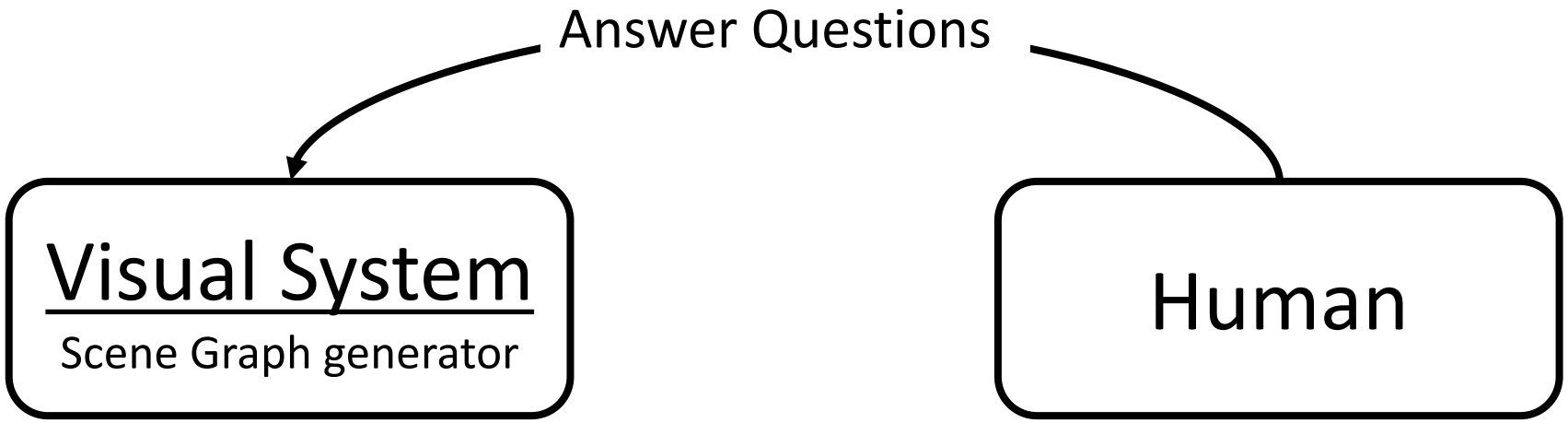


[1] Visual Curiosity: Learning to Ask Questions to Learn Visual Recognition. Yang et al. CoRL 2018
[2] Information Maximizing Visual Question Generation. Krishna et al. CVPR 2019

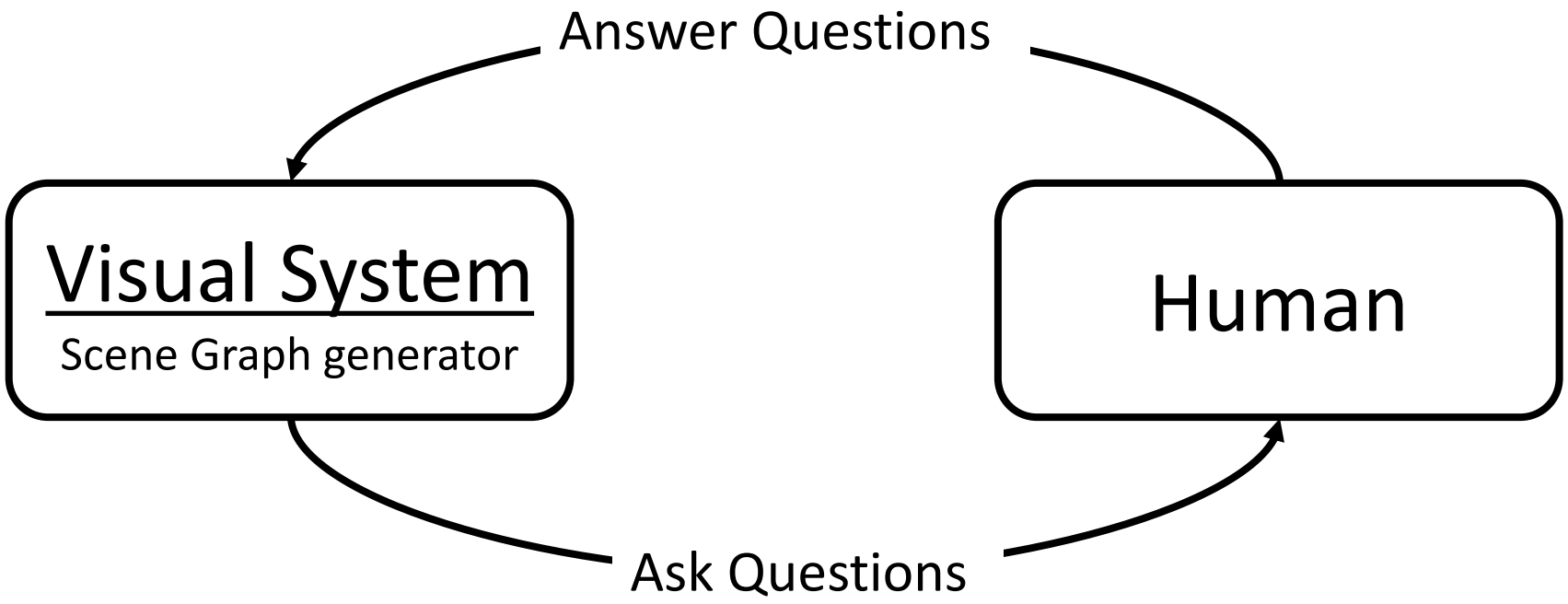
Left: https://cals.ncsu.edu/wp-content/uploads/2016/08/horse-1500x931.png
Rigth: https://www.videoblocks.com/video/the-man-in-hat-feed-a-brown-horse-with-flowers-on-the-meadow-supmox_3xj0tvkb67


Visual System
Scene Graph generator
Communication
Human

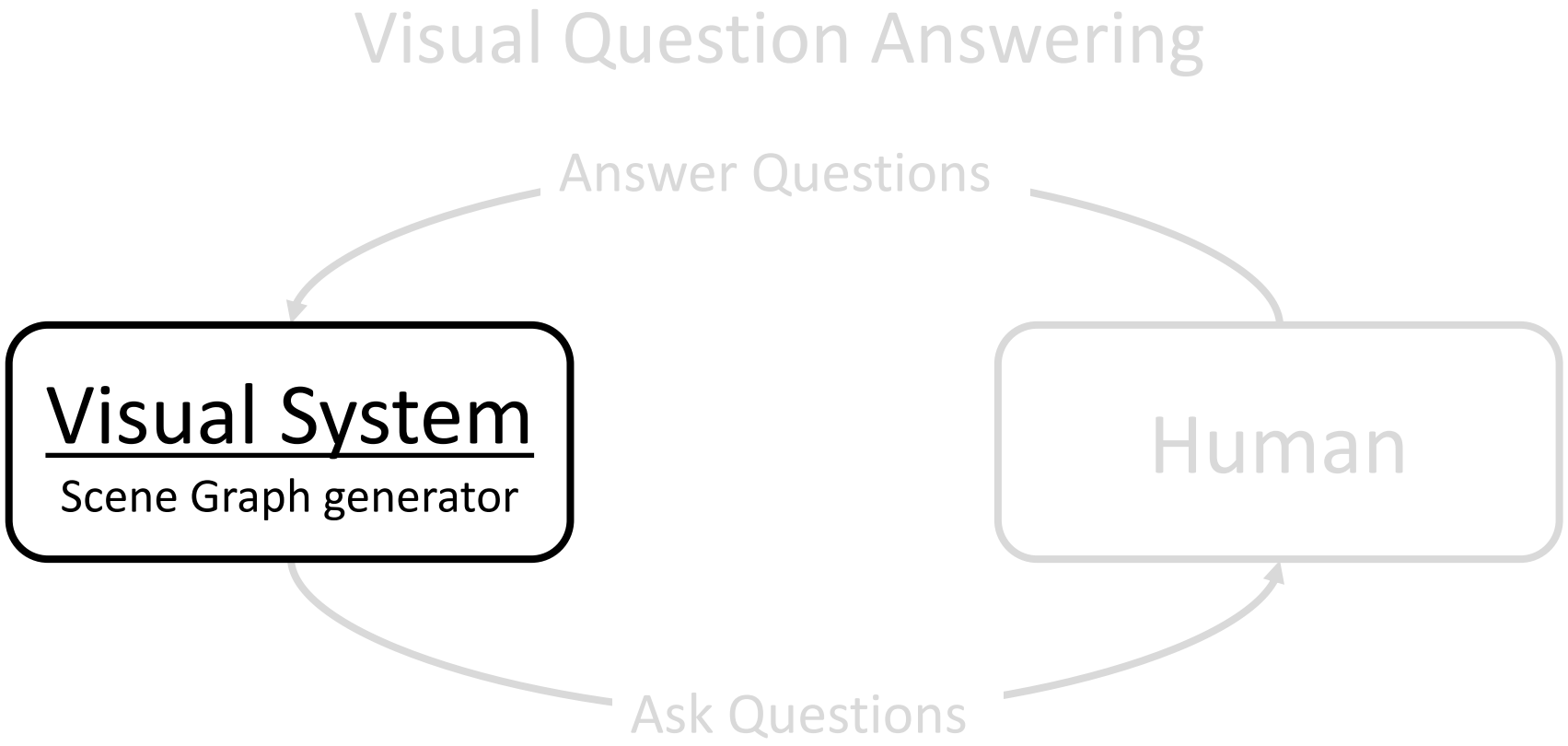
Visual Question Answering



Visual Question Answering



Visual Question Generation

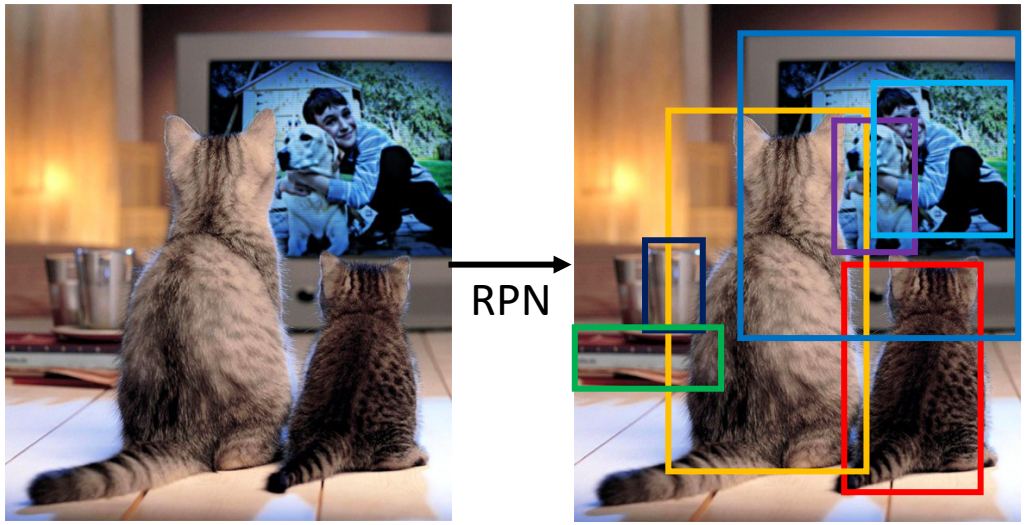

Visual Question Answering
Answer Questions
Visual System
Scene Graph generator
Human
Ask Questions
Visual Question Generation

Skeleton Model

Skeleton Model

Input

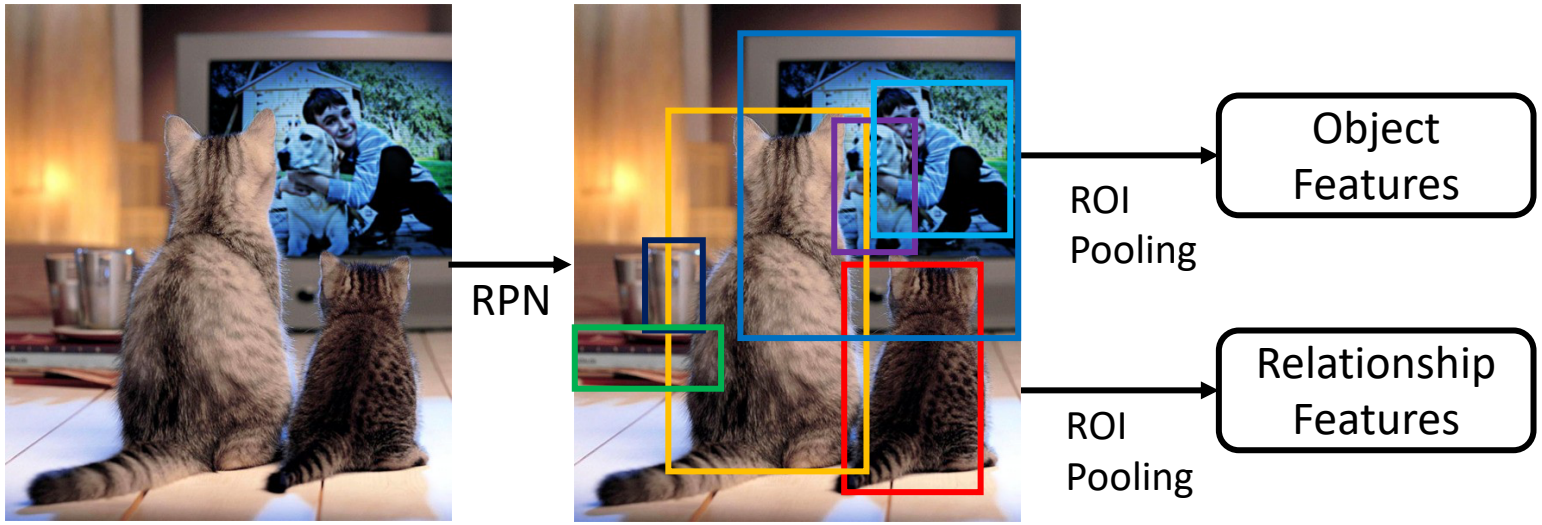
Skeleton Model



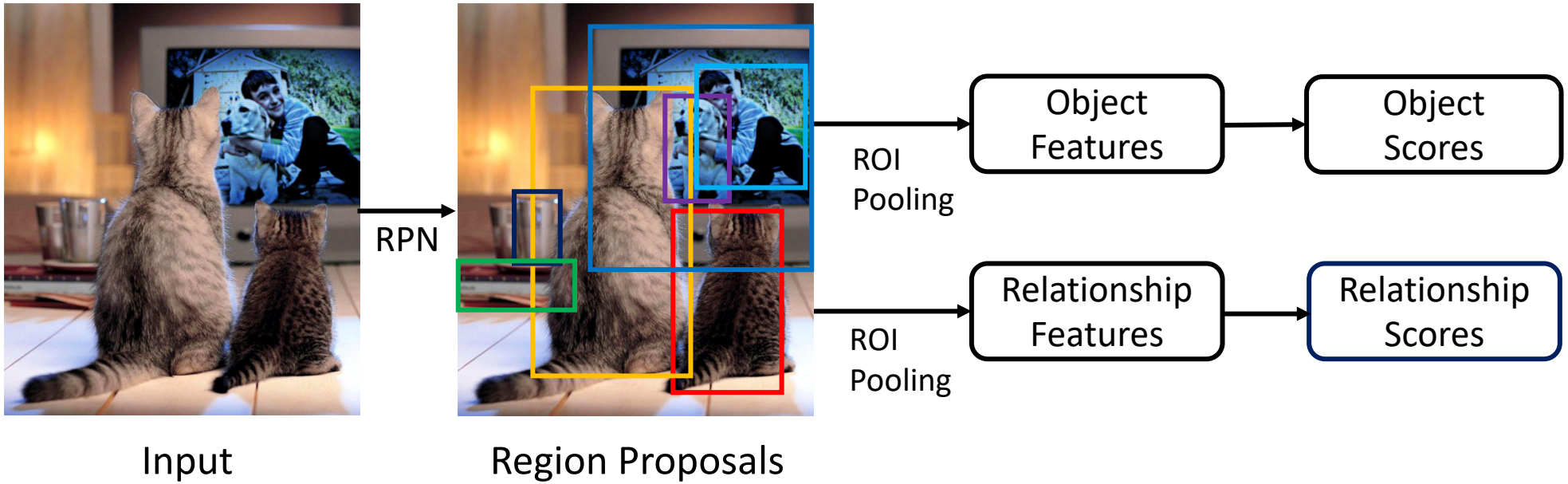
Input

Region Proposals

Skeleton Model



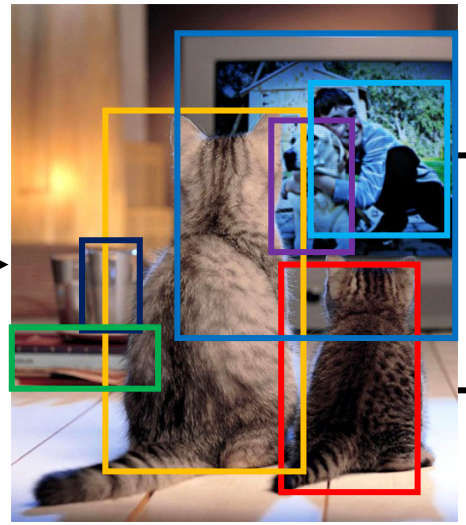
Skeleton Model



Skeleton Model

Input

RPN



Region Proposals

ROI Pooling

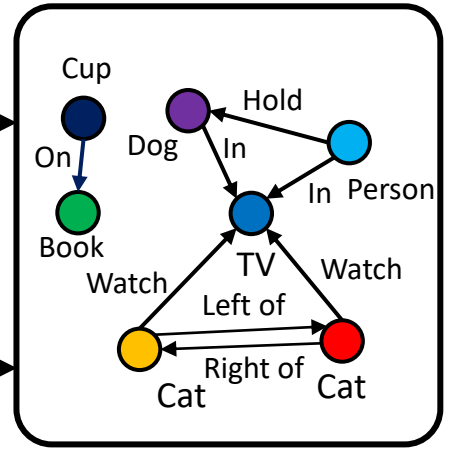
ROI Pooling

Object Features

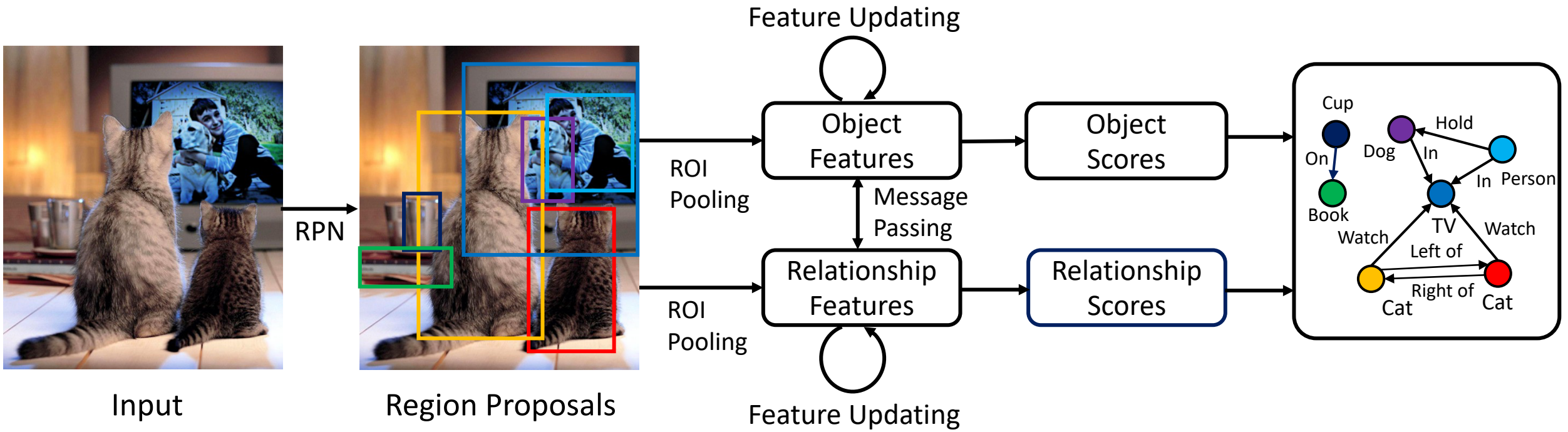
Relationship Features

Object Scores

Relationship Scores

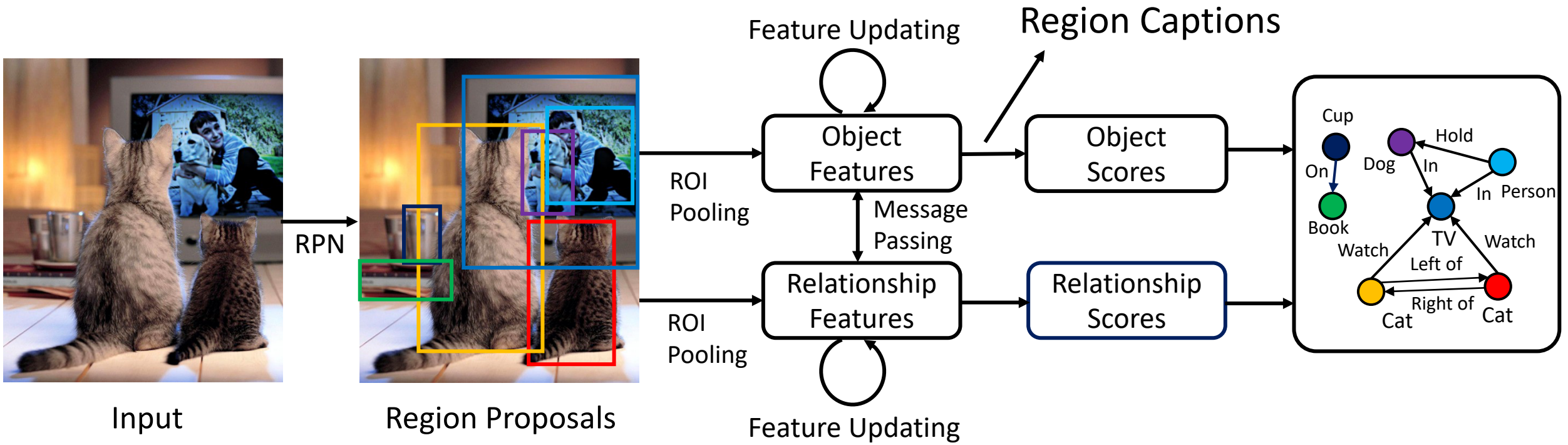


Iterative Message Passing (IMP)



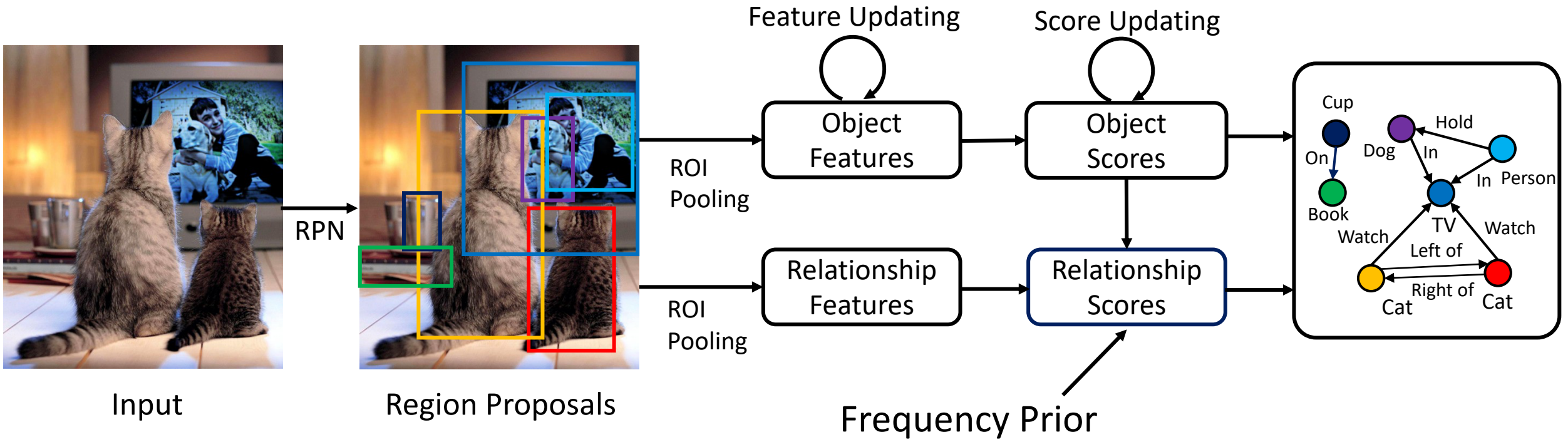
Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

Multi-level Scene Description Network (MSDN)



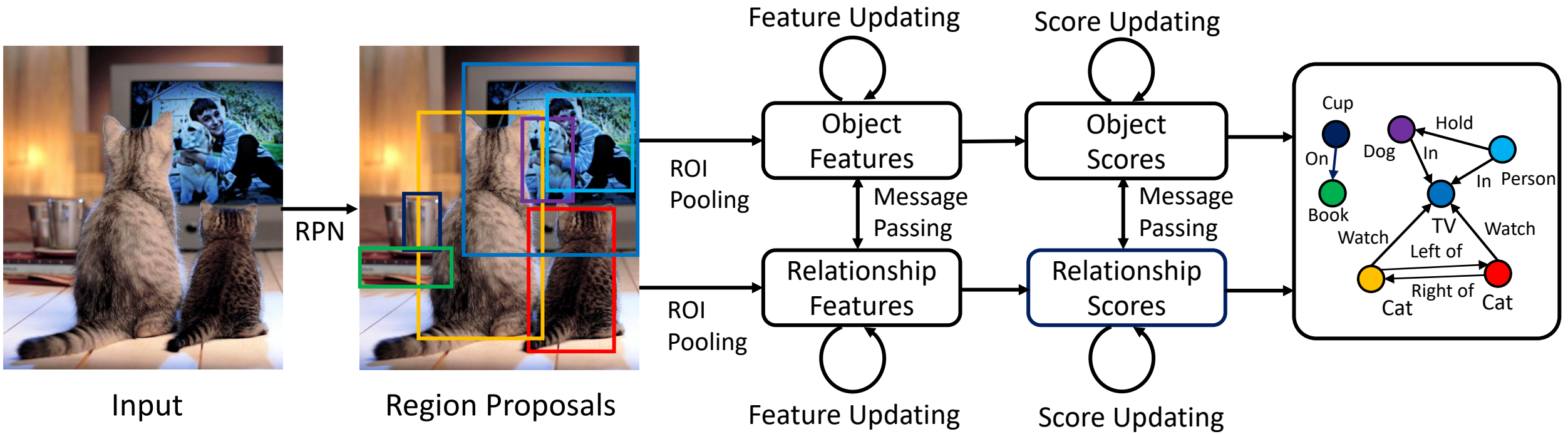
Scene Graph Generation from Objects, Phrases and Region Captions. Li et al. ICCV 2017

Neural Motif Network



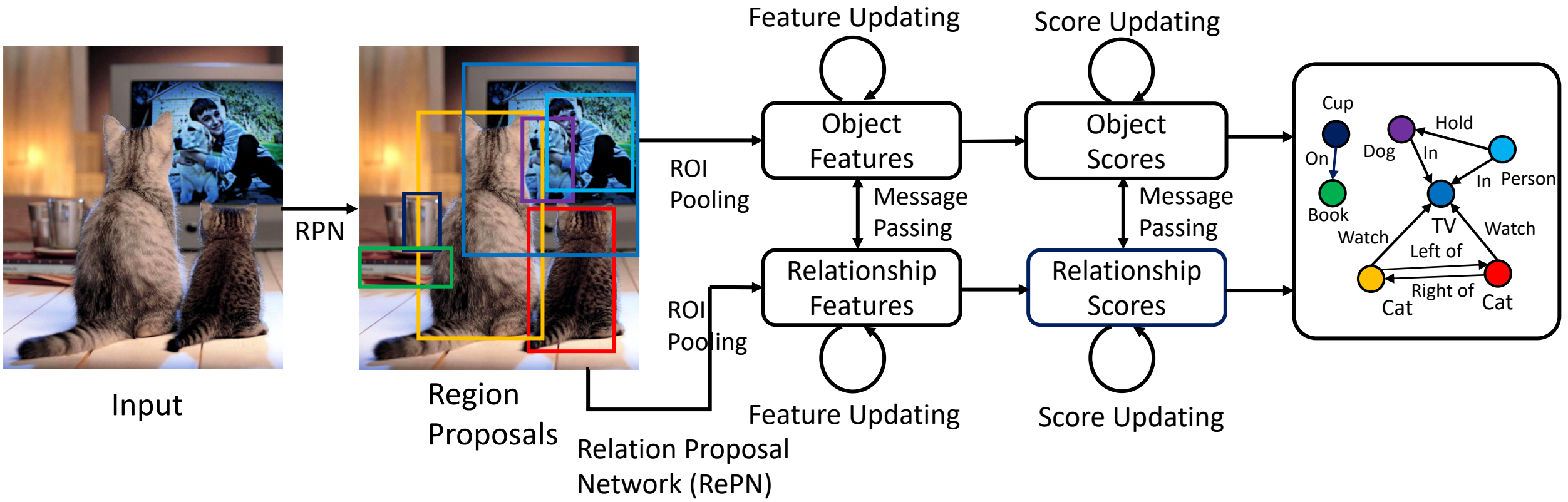
Neural Motifs: Scene Graph Parsing with Global Context. Zellers et al. CVPR 2018

Graph R-CNN (Our work)



Neural Motifs: Scene Graph Parsing with Global Context. Zellers et al. CVPR 2018

Graph R-CNN (Our work)



Jianwei Yang*, Jiasen Lu*, Stefan Lee, Dhruv Batra, Devi Parikh. Graph R-CNN for Scene Graph Generation. ECCV 2018.

Motivations

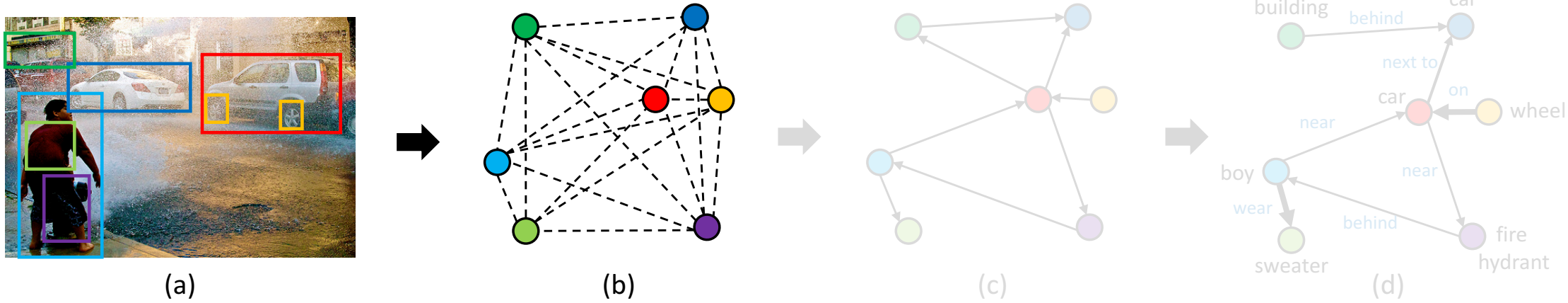
(a)

(b)

(c)

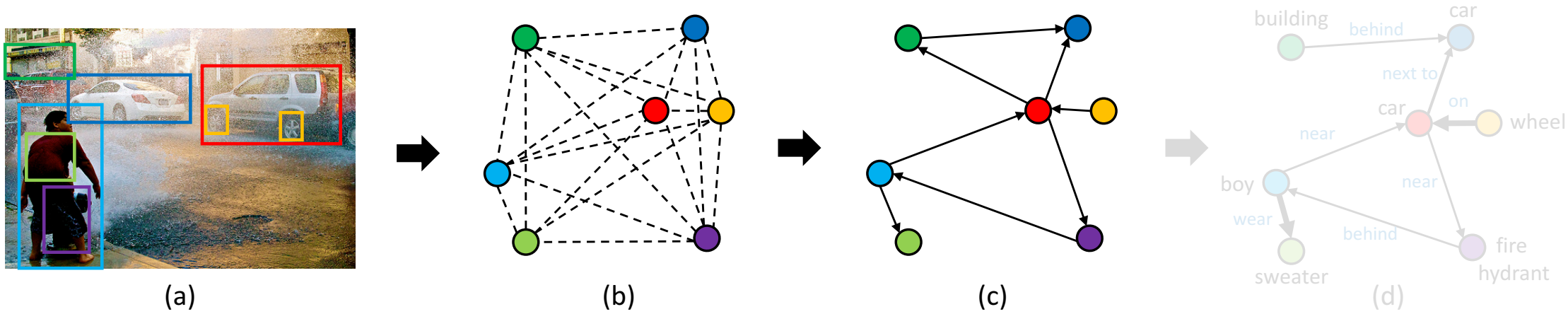
(d)

Motivations



1. Objects in a scene usually have relationships with others;

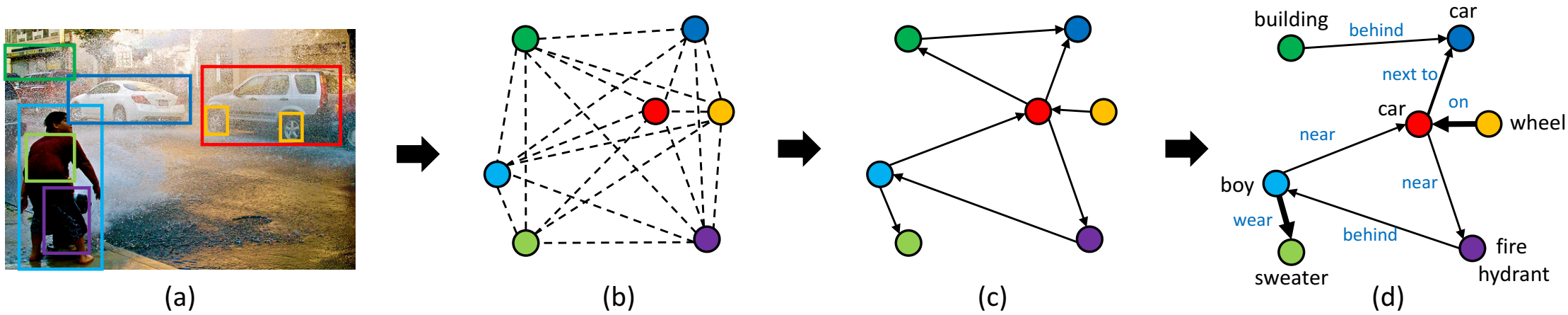
Motivations



1. Objects in a scene usually have relationships with others;

2. Not all object pairs have relationships, the scene graph is usually sparse;

Motivations

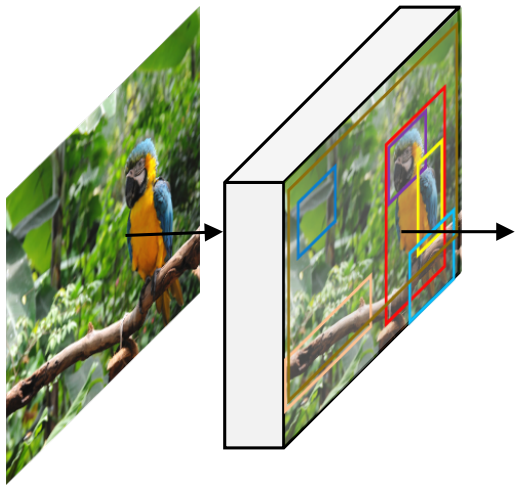


1. Objects in a scene usually have relationships with others;

2. Not all object pairs have relationships, the scene graph is usually sparse;

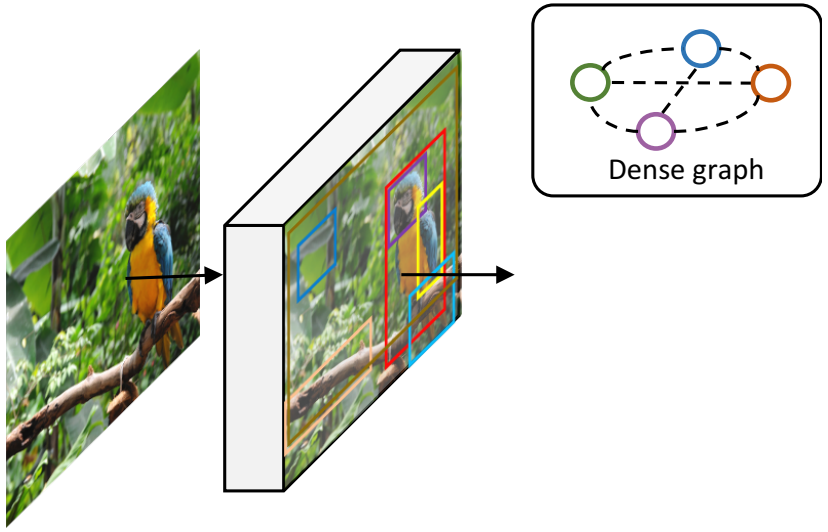
3. Existence of relationships highly depends on the object categories, and type of relationships highly depends on the context.

Framework



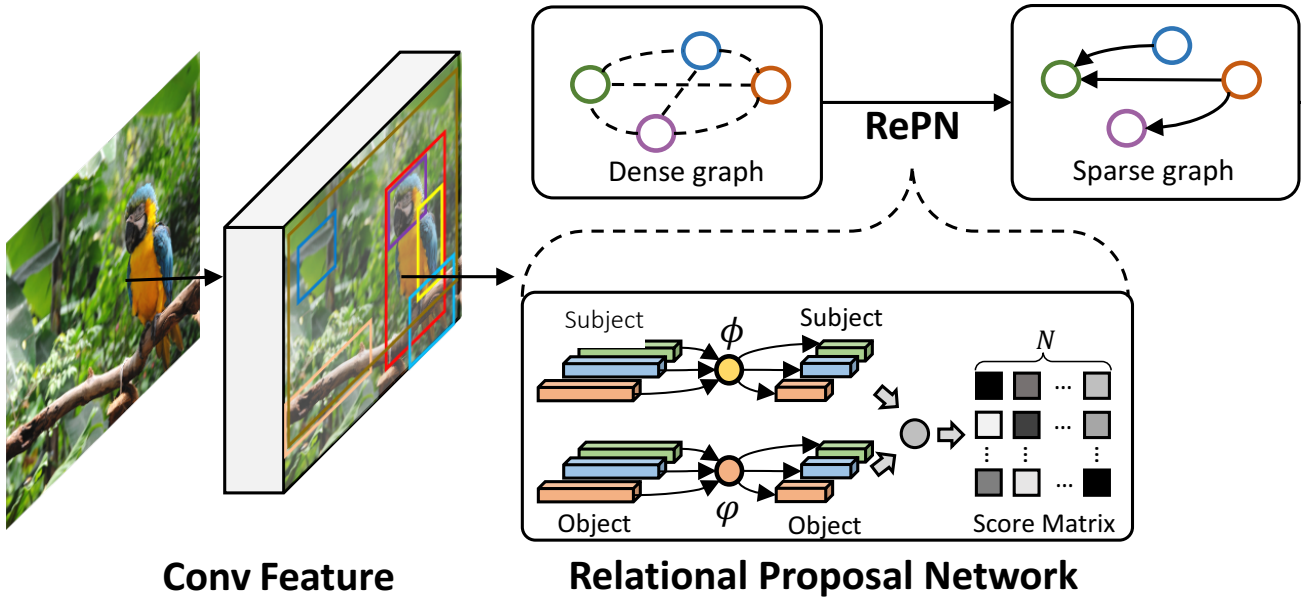
Conv Feature

Framework



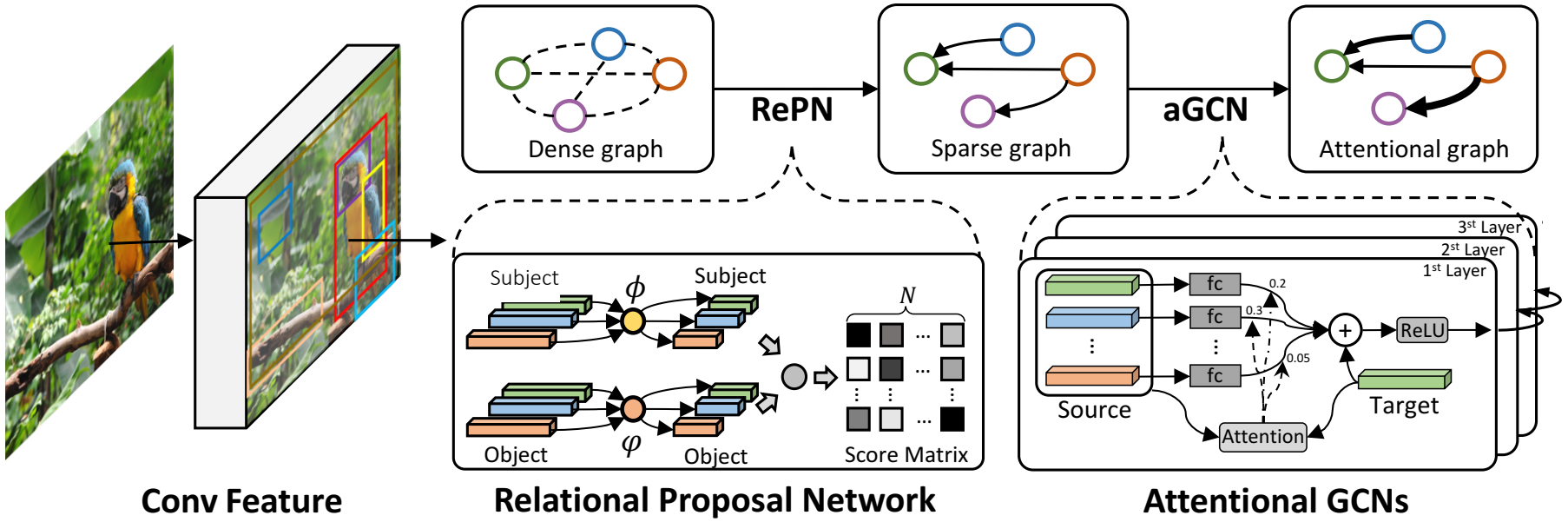
Conv Feature

Framework



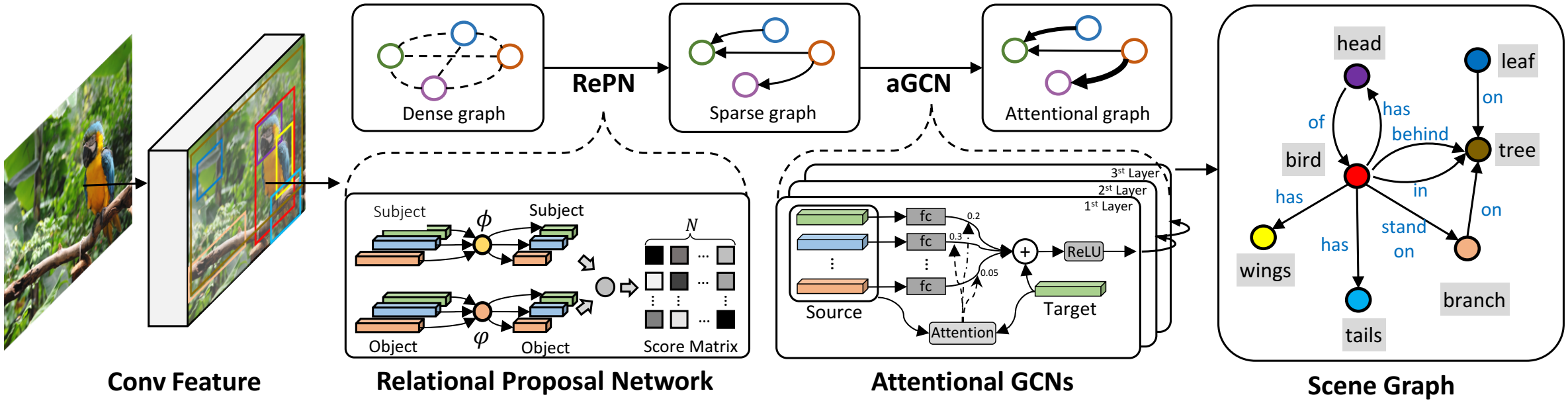
1. Relation proposal network (RePN) to learn to prune the densely connected scene graph;

Framework



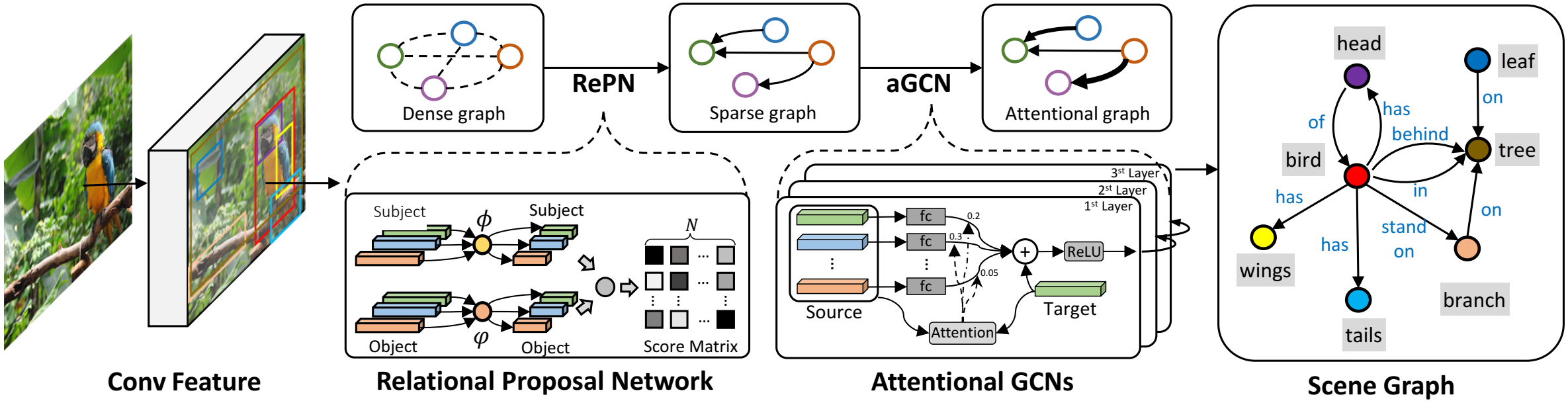
1. Relation proposal network (RePN) to learn to prune the densely connected scene graph;
2. Attentional graph convolutional networks (aGCN) to incorporate the contextual information.

Framework



1. Relation proposal network (RePN) to learn to prune the densely connected scene graph;
2. Attentional graph convolutional networks (aGCN) to incorporate the contextual information.

Framework



I: Input Image; S: Scene graph

V: Scene graph vertices (object)

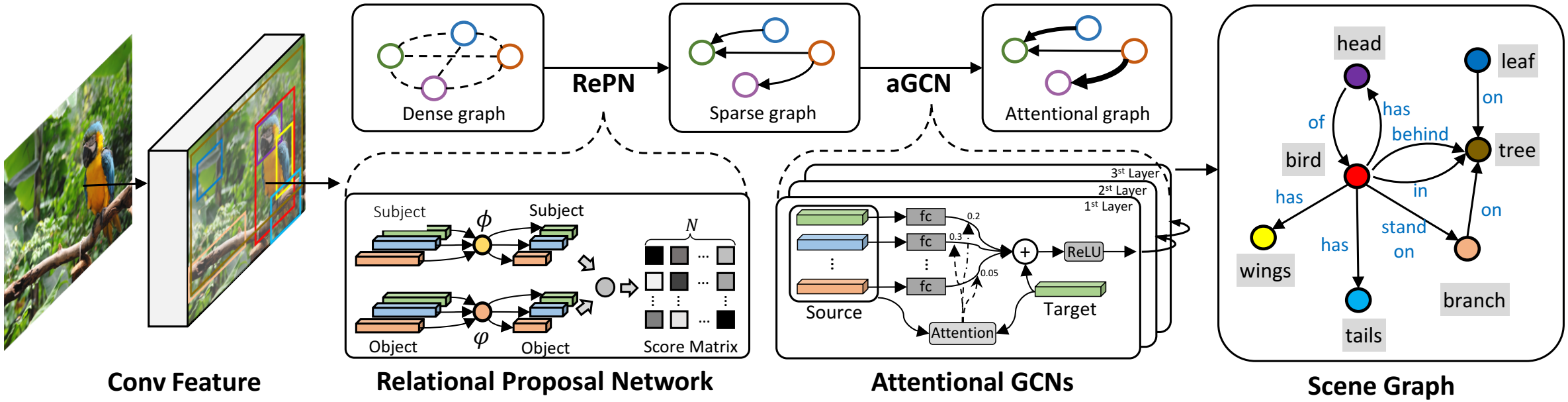
E: Scene graph edges (relationship)

O: Scene graph object labels

R: Scene graph relationship labels

$$P(S|I)$$

Framework



I: Input Image; S: Scene graph

V: Scene graph vertices (object)

E: Scene graph edges (relationship)

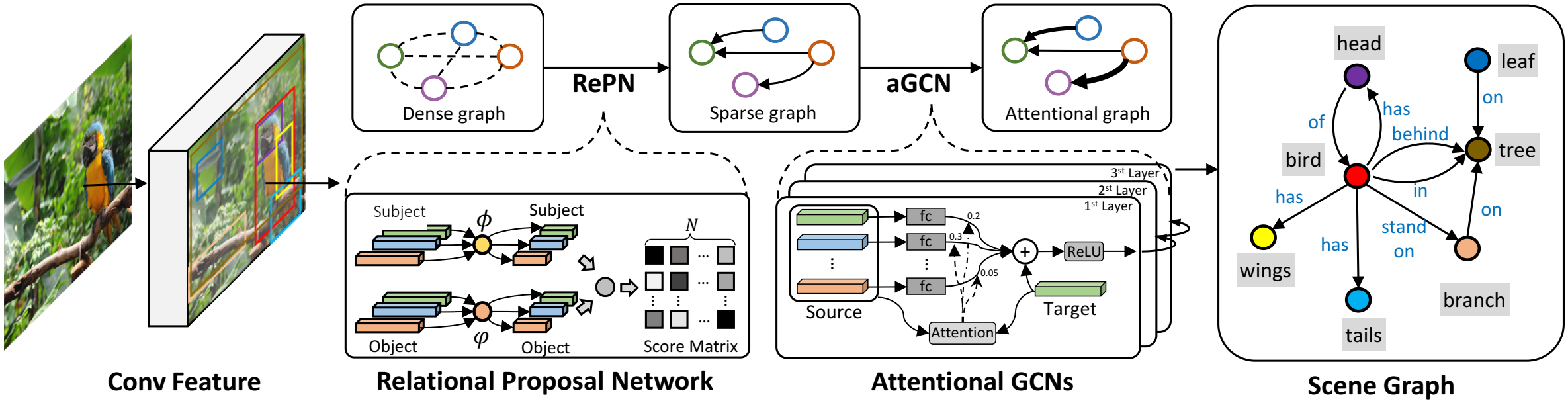
O: Scene graph object labels

R: Scene graph relationship labels

Region Proposal

$$P(S|I) = \overbrace{P(V|I)}$$

Framework



I: Input Image; S: Scene graph

V: Scene graph vertices (object)

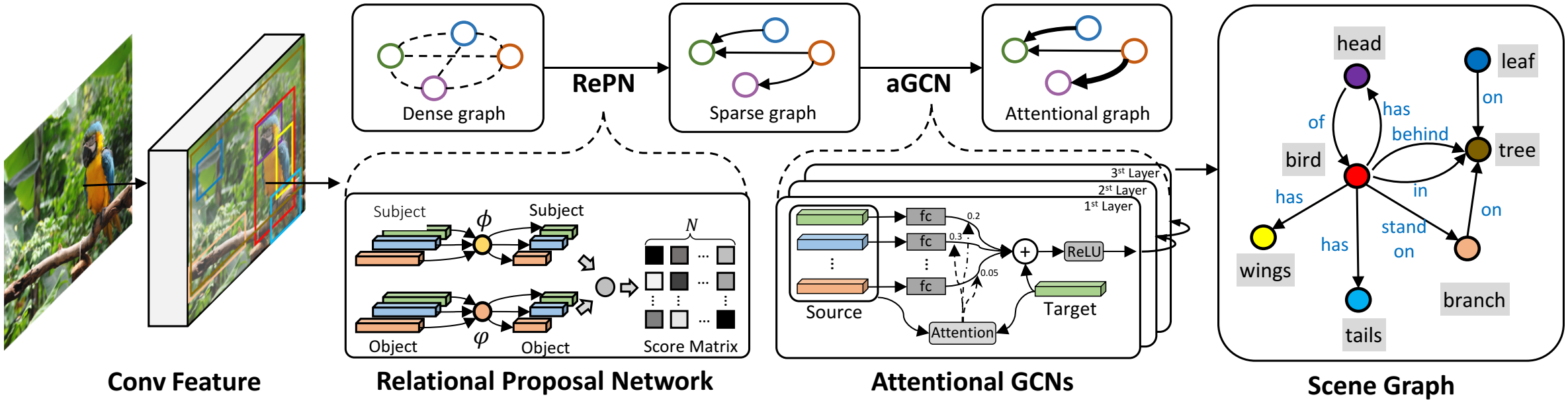
E: Scene graph edges (relationship)

O: Scene graph object labels

R: Scene graph relationship labels

$$P(S|I) = \overbrace{P(V|I)}^{\text{Region Proposal}} \underbrace{P(E|V,I)}_{\text{Relation Proposal}}$$

Framework



I: Input Image; S: Scene graph

V: Scene graph vertices (object)

E: Scene graph edges (relationship)

O: Scene graph object labels

R: Scene graph relationship labels

Region Proposal

Graph Labeling

$$P(S|I) = \overbrace{P(V|I)}\ \underbrace{P(E|V,I)}\ \overbrace{P(R,O|V,E,I)}$$

Relation Proposal

Training

$$P(V|I) \quad P(E|V,I) \quad P(R,O|V,E,I) = P(S|I)$$

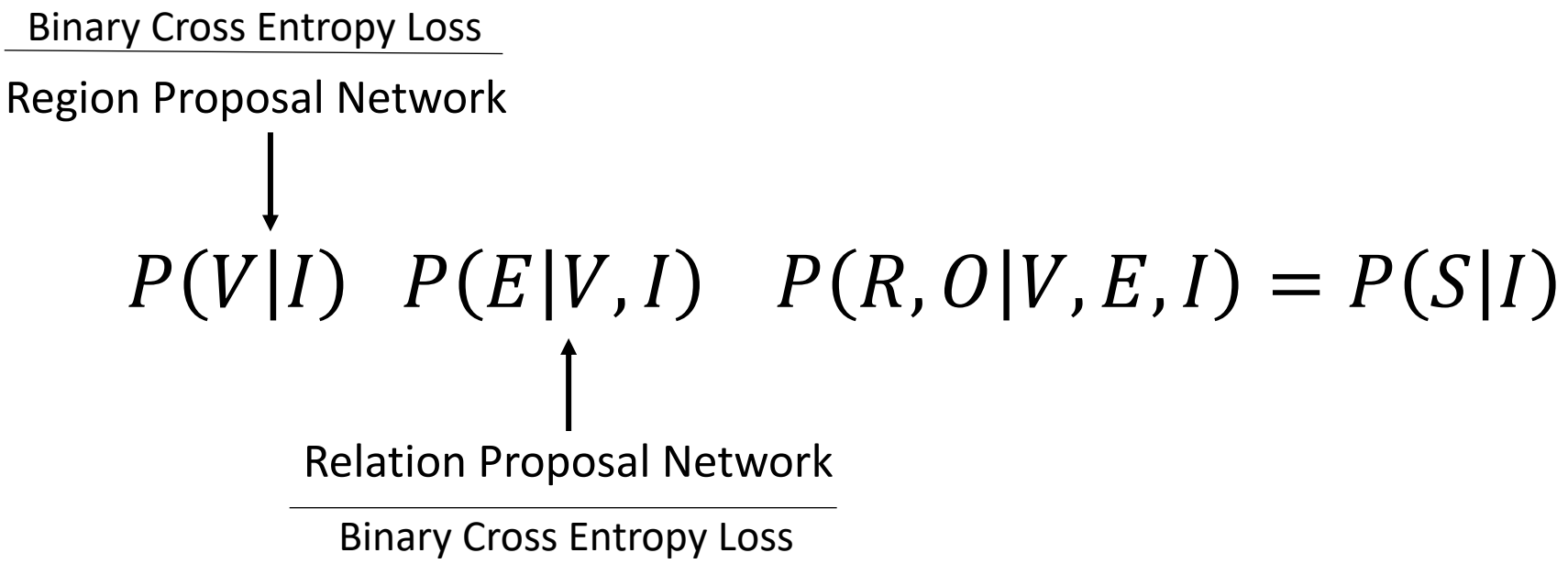
Training

Binary Cross Entropy Loss

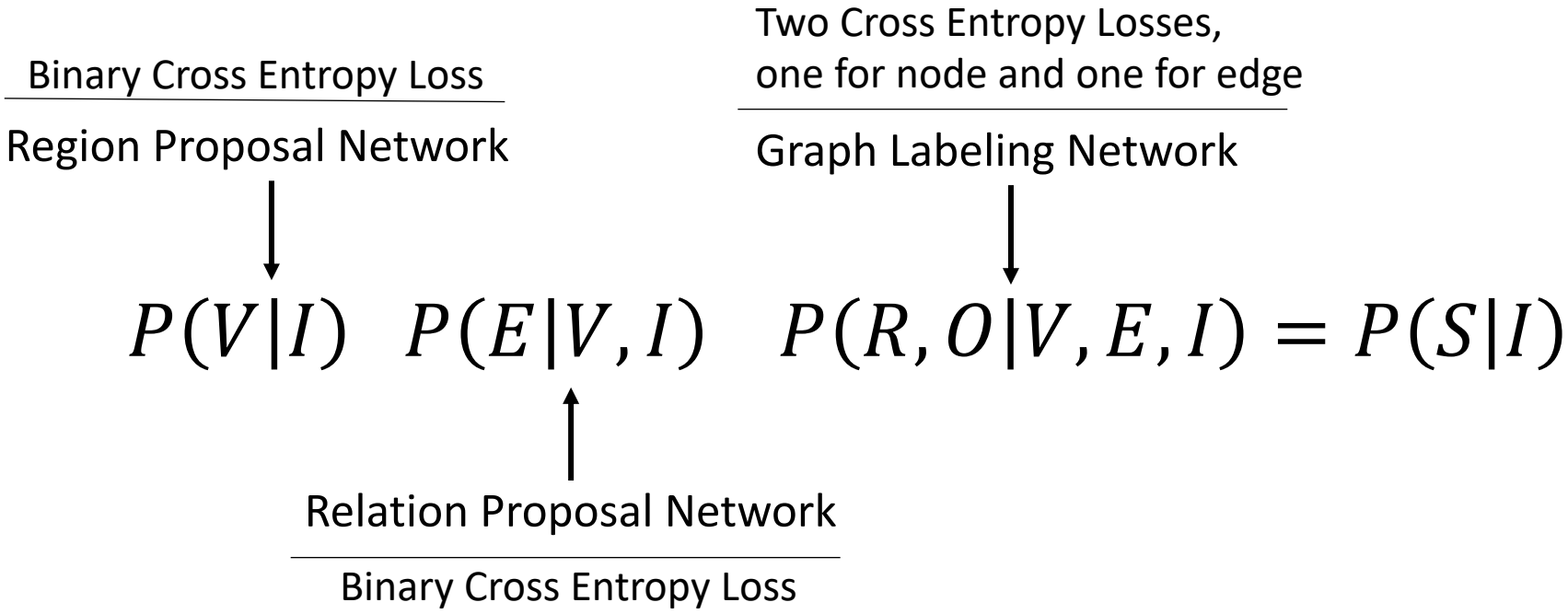
Region Proposal Network

$$P(V|I) \quad P(E|V,I) \quad P(R,O|V,E,I) = P(S|I)$$

Training


Binary Cross Entropy Loss
Region Proposal Network
$P(V|I)$
$P(E|V,I)$
$P(R,O|V,E,I) = P(S|I)$
Relation Proposal Network
Binary Cross Entropy Loss

Training


Binary Cross Entropy Loss
Region Proposal Network
Two Cross Entropy Losses,
one for node and one for edge
Graph Labeling Network
$P(V|I)$
$P(E|V,I)$
$P(R,O|V,E,I) = P(S|I)$
Relation Proposal Network
Binary Cross Entropy Loss

Metrics

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

Metrics

Assume there are N objects extracted from an image, then $N * (N - 1)$ edges

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

Metrics

Assume there are N objects extracted from an image, then $N * (N - 1)$ edges

Step 1: Take maximum for object scores and predicate scores, excluding background class.

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

Metrics

Assume there are N objects extracted from an image, then $N * (N - 1)$ edges

Step 1: Take maximum for object scores and predicate scores, excluding background class.

Step 2: Compute relationship scores: $Rel(i, j) = Subj(i) * Obj(j) * Pred(i, j)$

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

Metrics

Assume there are N objects extracted from an image, then $N * (N - 1)$ edges

Step 1: Take maximum for object scores and predicate scores, excluding background class.

Step 2: Compute relationship scores: $Rel(i, j) = Subj(i) * Obj(j) * Pred(i, j)$

Step 3: Sort the relationship triplets in a descending order.

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

Metrics

Assume there are N objects extracted from an image, then $N * (N-1)$ edges

Step 1: Take maximum for object scores and predicate scores, excluding background class.

Step 2: Compute relationship scores: $Rel(i,j) = Subj(i) * Obj(j) * Pred(i,j)$

Step 3: Sort the relationship triplets in a descending order.

Step 4: Compute the triplet recalls (Recall@50, Recall@100) based on the ground-truth

SGGen: $Recall = \frac{C(T_{pred}\ and\ T_{gt})}{N(T_{gt})}$ IoU > 0.5

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

Metrics

Assume there are N objects extracted from an image, then $N * (N-1)$ edges

Step 1: Take maximum for object scores and predicate scores, excluding background class.

Step 2: Compute relationship scores: $Rel(i,j) = Subj(i) * Obj(j) * Pred(i,j)$

Step 3: Sort the relationship triplets in a descending order.

Step 4: Compute the triplet recalls (Recall@50, Recall@100) based on the ground-truth

SGGen: $Recall = \frac{C(T_{pred}\ and\ T_{gt})}{N(T_{gt})}$ IoU > 0.5

PhrCls: all object locations are known **PredCls**: all object locations and labels are known

[1]. Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

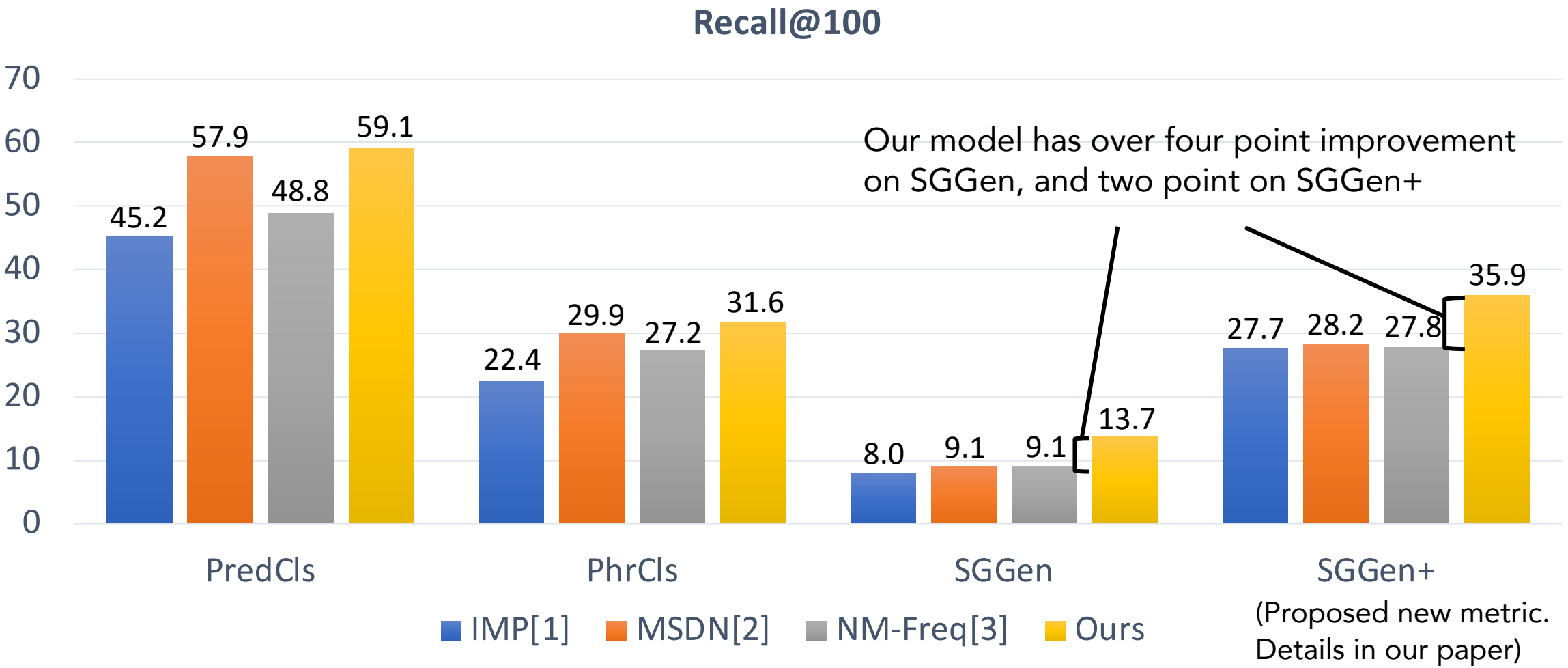
Experiments

Table. Implementation Details.

Dataset	Backbone	#objects	#predicates	Metrics
Visual Genome[1] Train: 75,651 Test: 32,422	VGG-16 Faster R-CNN[2]	150	50	PredCls,SGCls, SGGen,SGGen+, mAP

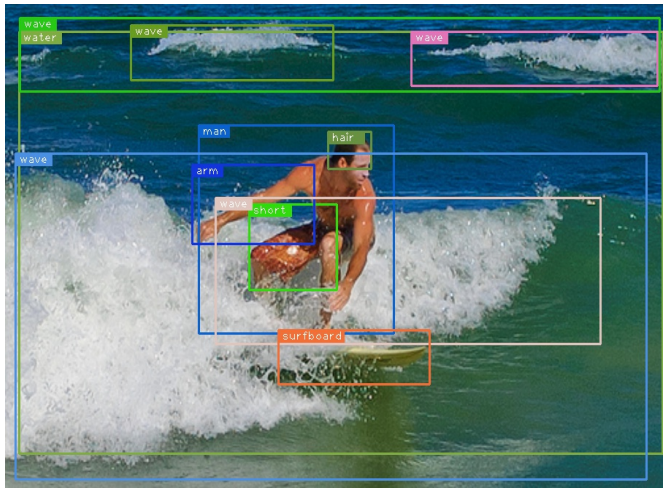
[1] Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. Krishna et al.
[2] A Faster Implementation of Faster R-CNN. Yang and Lu et al.

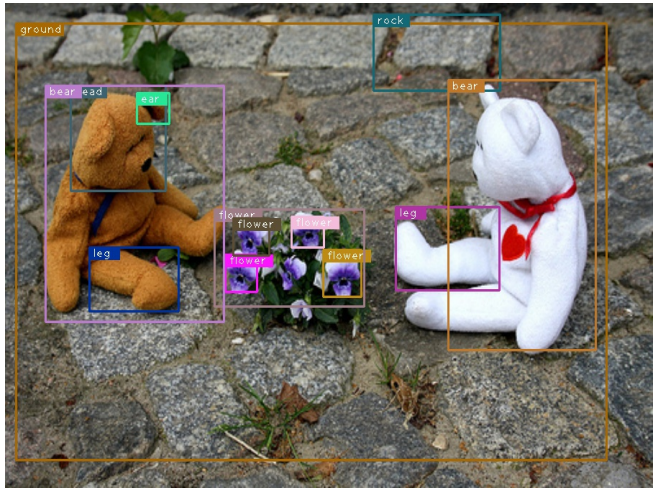
Comparing with Previous Work

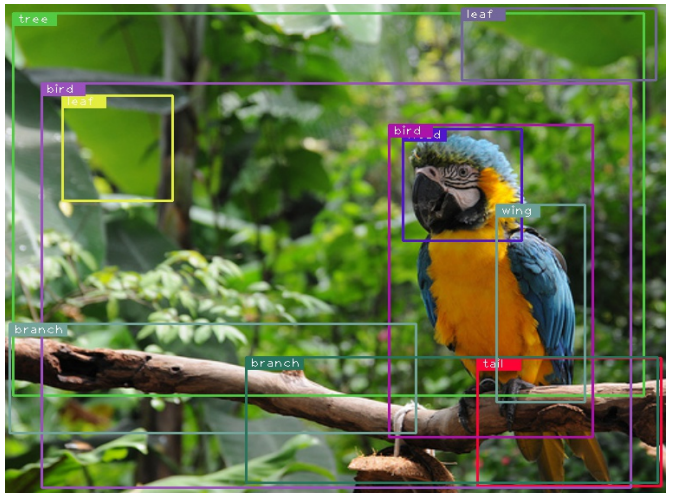


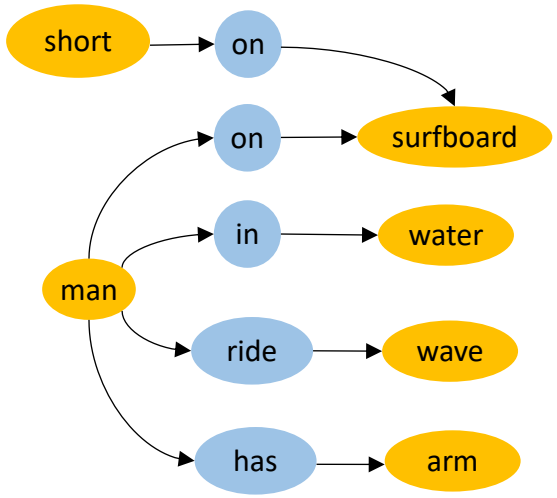
[1] Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017
[2] Scene Graph Generations from Objects, Phrases and Captions. Li et al. ICCV 2017
[3] Neural Motif: Scene Graph Parsing with Global Context. Zellers et al. CVPR 2018

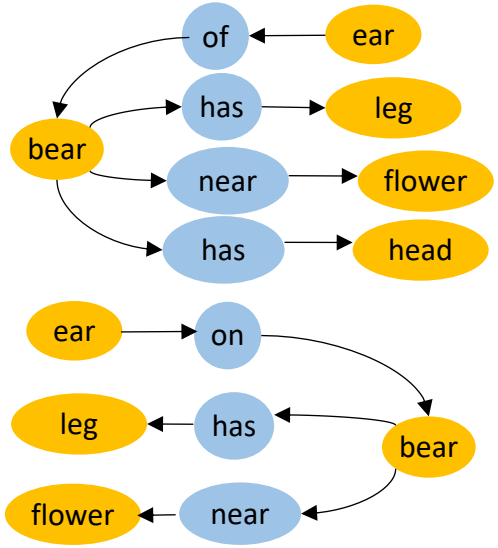
Qualitative Results

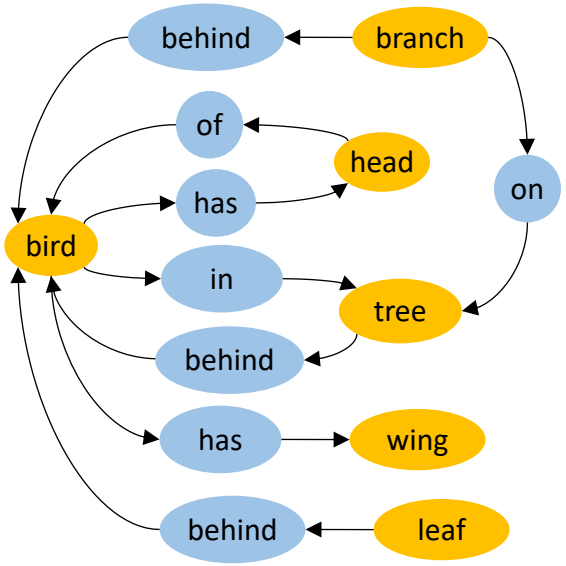




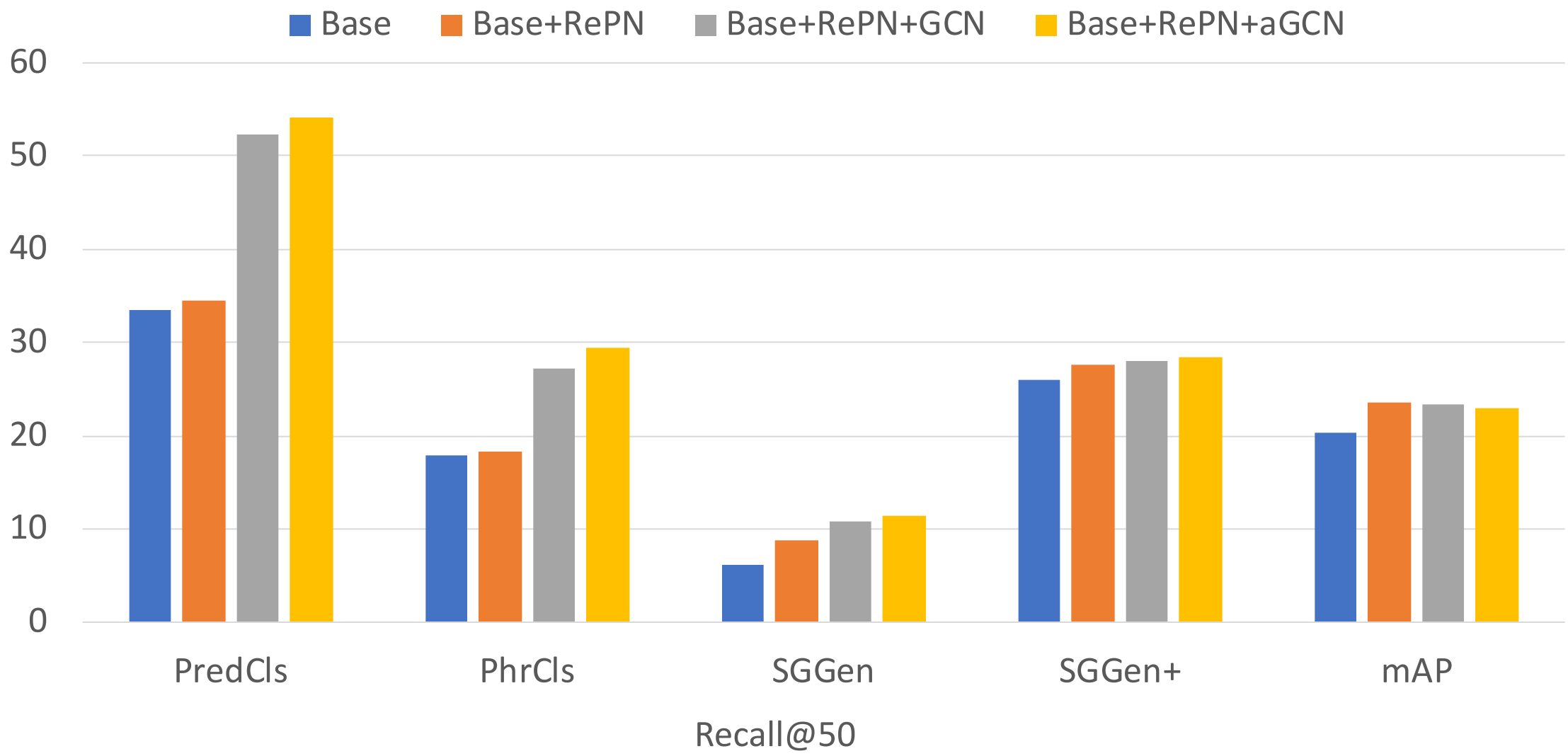




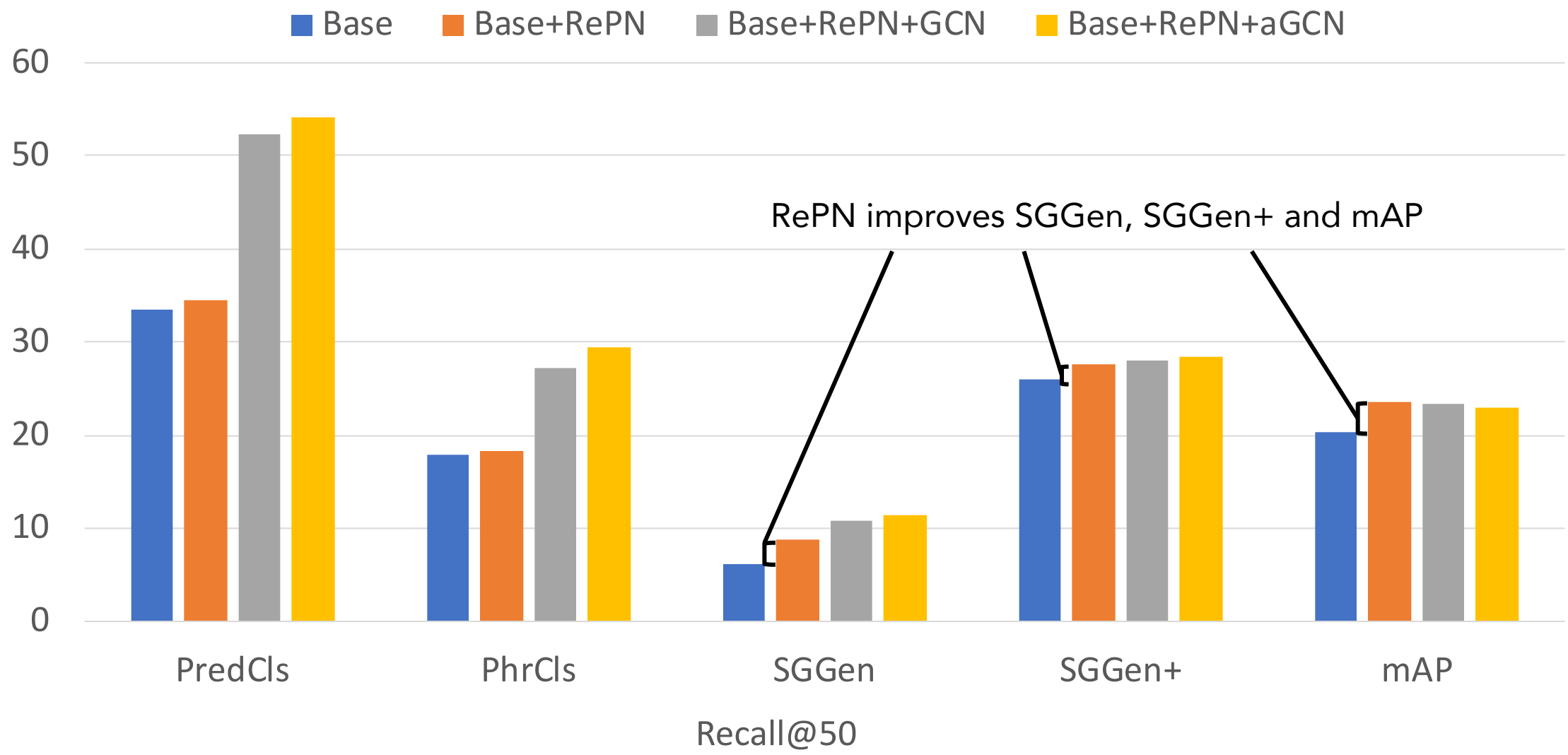




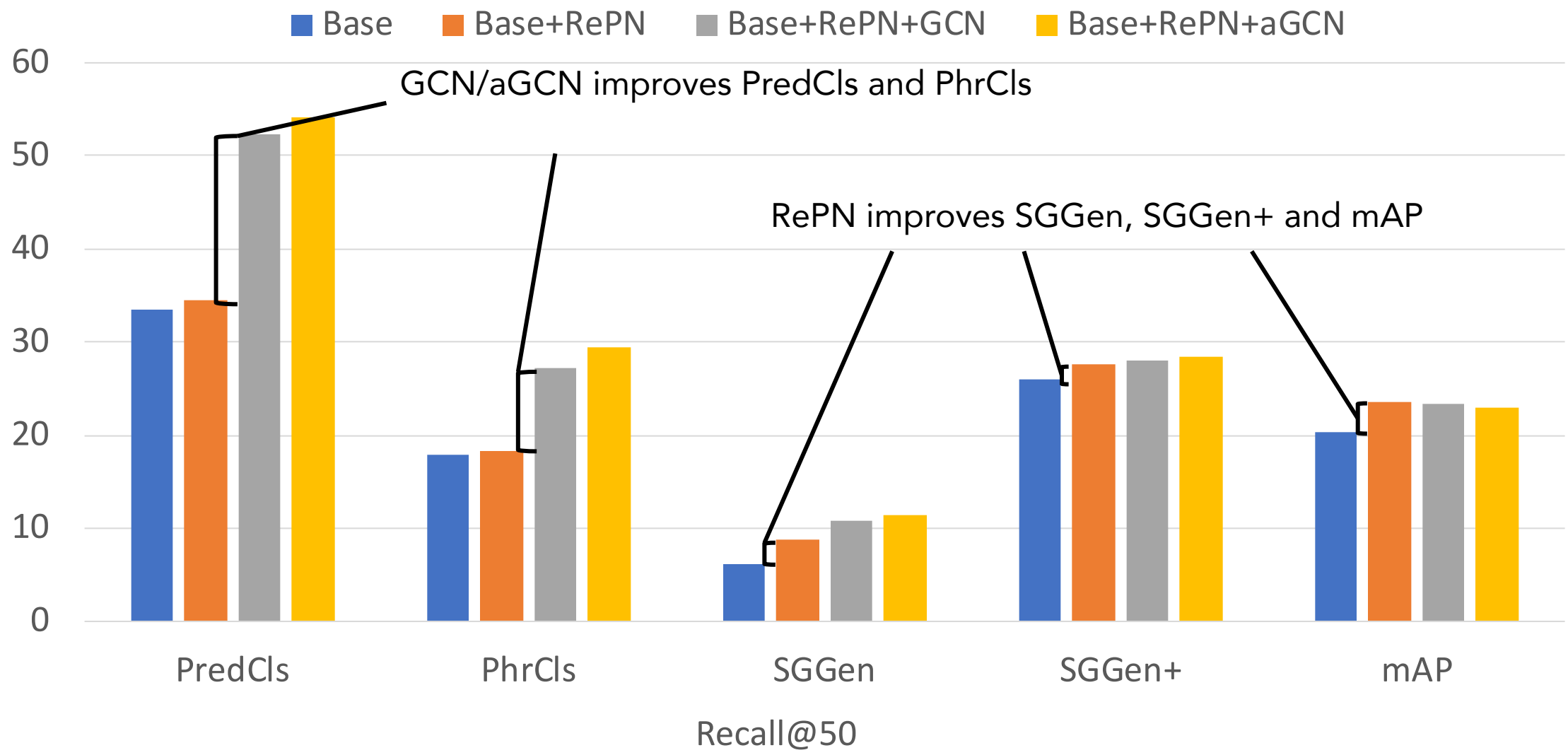
Ablation Study



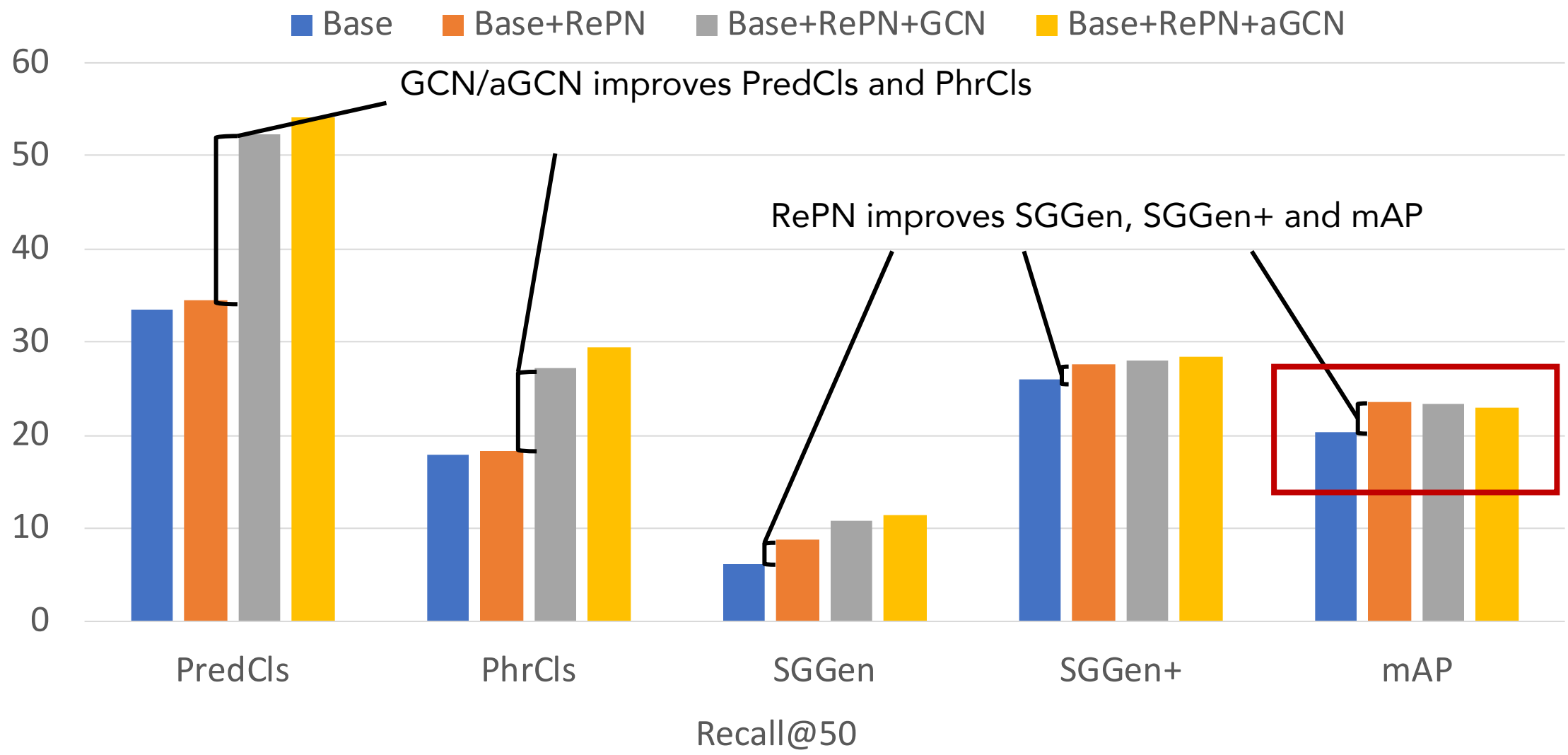
Ablation Study



Ablation Study



Ablation Study



A new codebase for scene graph generation

https://github.com/jwyang/graph-rcnn.pytorch

This project is a set of reimplemented representative scene graph generation models based on Pytorch 1.0, including:

- Graph R-CNN for Scene Graph Generation, our own. ECCV 2018.
- Scene Graph Generation by Iterative Message Passing, Xu et al. CVPR 2017
- Scene Graph Generation from Objects, Phrases and Region Captions, Li et al. ICCV 2017
- Neural Motifs: Scene Graph Parsing with Global Context, Zellers et al. CVPR 2018
- Graphical Contrastive Losses for Scene Graph Generation, Zhang et al, CVPR 2019

The goal of gathering all these representative methods into a single repo is to establish a more fair comparison across different methods under the same settings.

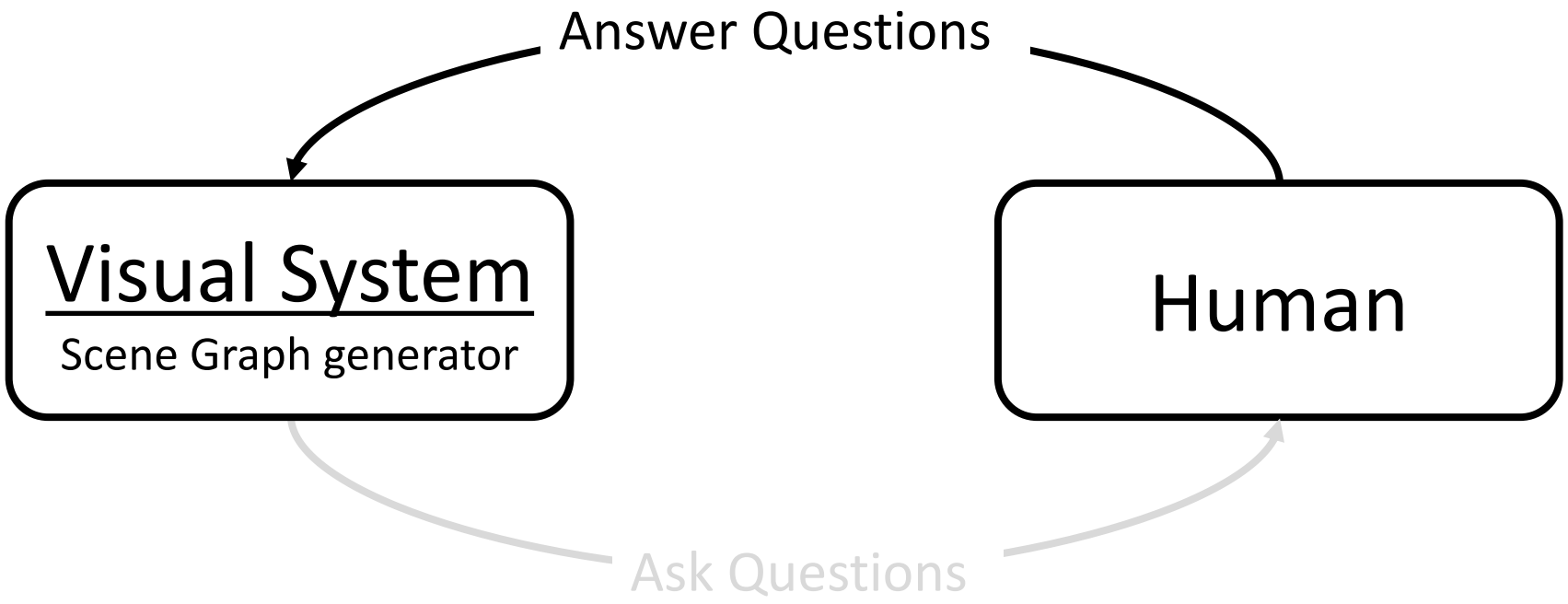
Welcome to contribute!

Summary

- Takeaways:
  - Introducing a general base model for scene graph generation
  - Pruning the fully-connected graph is important for scene graph generation
  - Exploiting the context across objects and predicates is crucial
  - Scene graph generation helps to improve object detection

- Challenges:
  - The dataset is noisy (incomplete and inconsistent annotations)
  - Relationships need more fine-grained categorizing (spatial, semantic, etc)
  - Rare/novel relationship is hard to detect

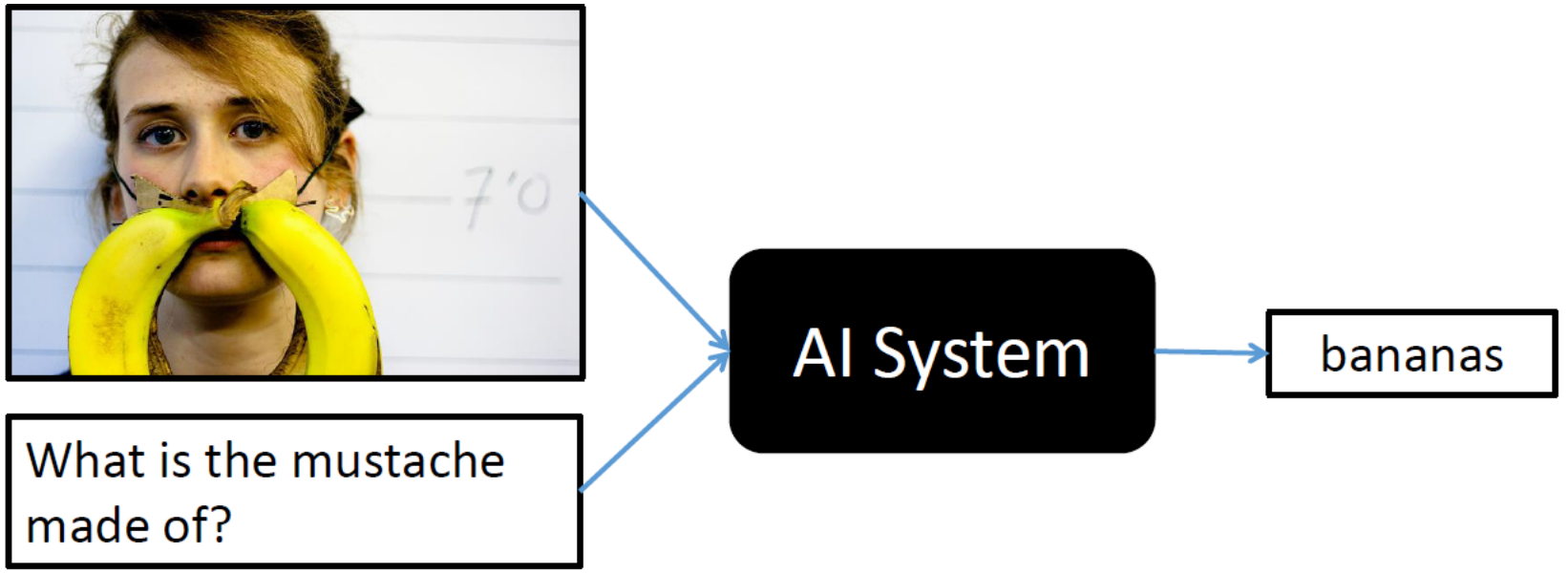
Visual Question Answering



Visual Question Generation

Visual Question Answering

Visual Question Answering is a challenging task that involves **fully visual understanding**, **language understanding** and **reasoning**.



Methodology: Graph Reasoning Machine

- **Visual Understanding**: Scene Graph Generator (extract object, attribute and relationship between objects).
- **Language Understanding**: Program Generator (extract logic reasoning chain in the question)
- **Reasoning**: Learnable Neural-Symbolic Executor (execute programs on s.g.)
- Pros:
  - Make the VQA model more interpretable;
  - Easy to diagnose and analyze the model predictions;
  - Modules are disentangled from each other;
  - Introduces no or few language priors, (probably) better generalization ability;

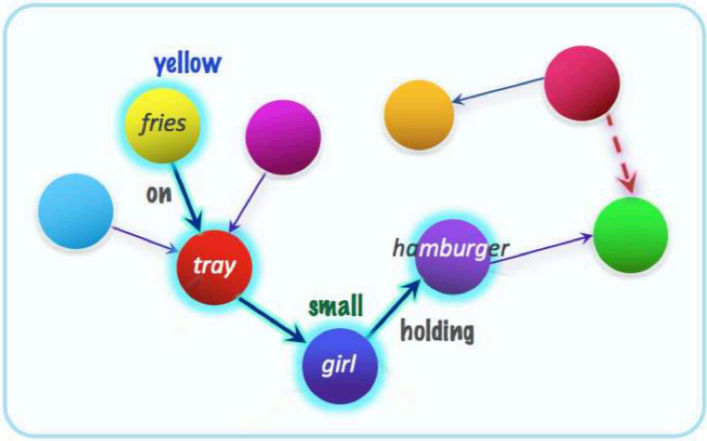
*Joint work with Chuang Gan et al

Compositional Reasoning VQA Dataset

Pattern: What|Which <type> [do you think] <is> <dobject>, <attr> or <decoy>?
Program: **Select:** <dobject> → **Choose** <type>: <attr>|<decoy>
Reference: The food on the red object left of the small girl that is holding a hamburger
Decoy: brown

What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?

Select: hamburger → **Relate:** girl,holding → **Filter** size: small → **Relate:** object, left → **Filter** color: red → **Relate:** food,on → **Choose** color: yellow | brown



Graph Normalization

- Ontology construction
- Edge Pruning
- Object Augmentation
- Global Properties

Question Generation

- Pattern Collection
- Compositional References
- Decoy Selection
- Probabilistic Generation

Sampling and Balancing

- Distribution Balancing
- Type-Based Sampling
- Deduplication

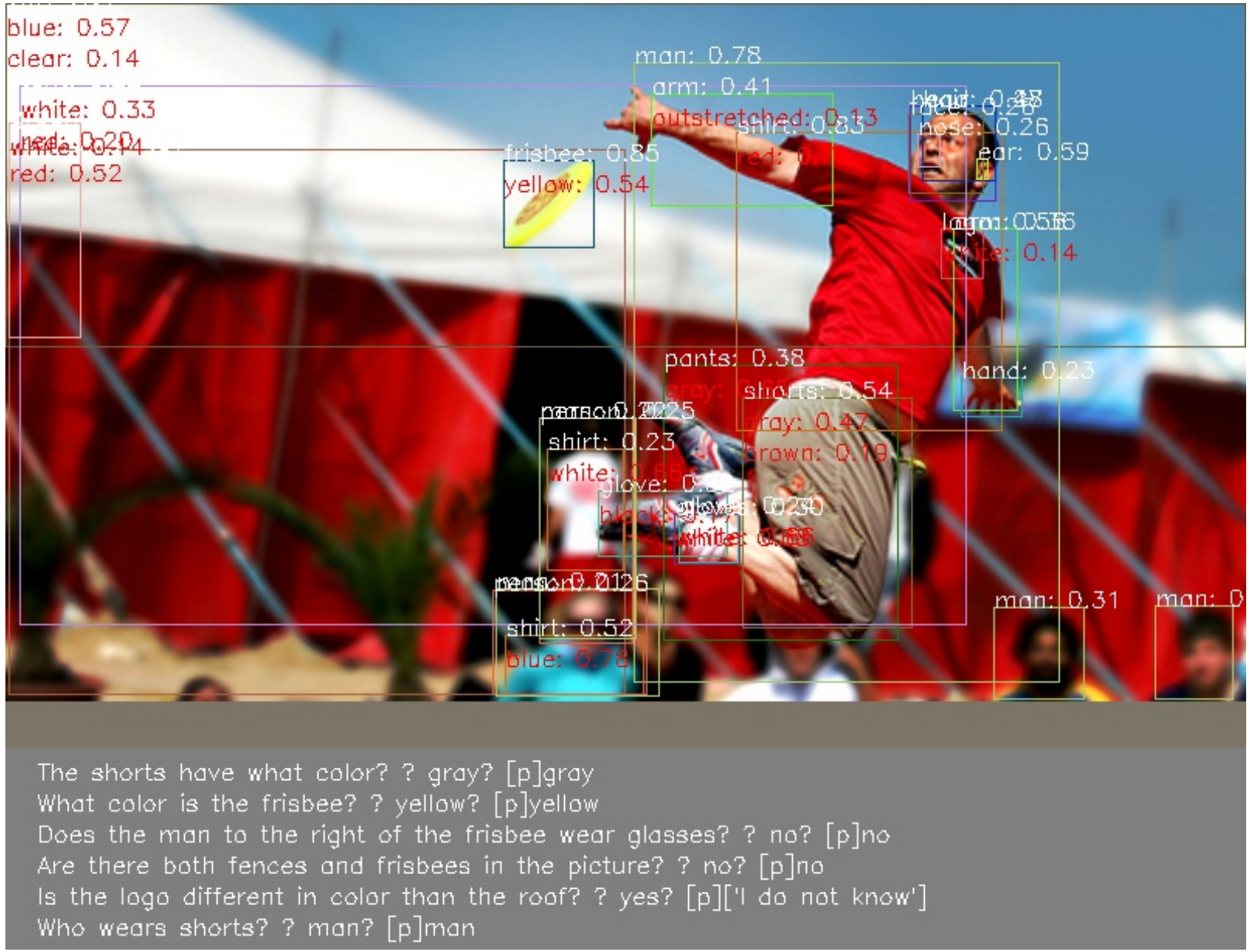
Entailment Relations

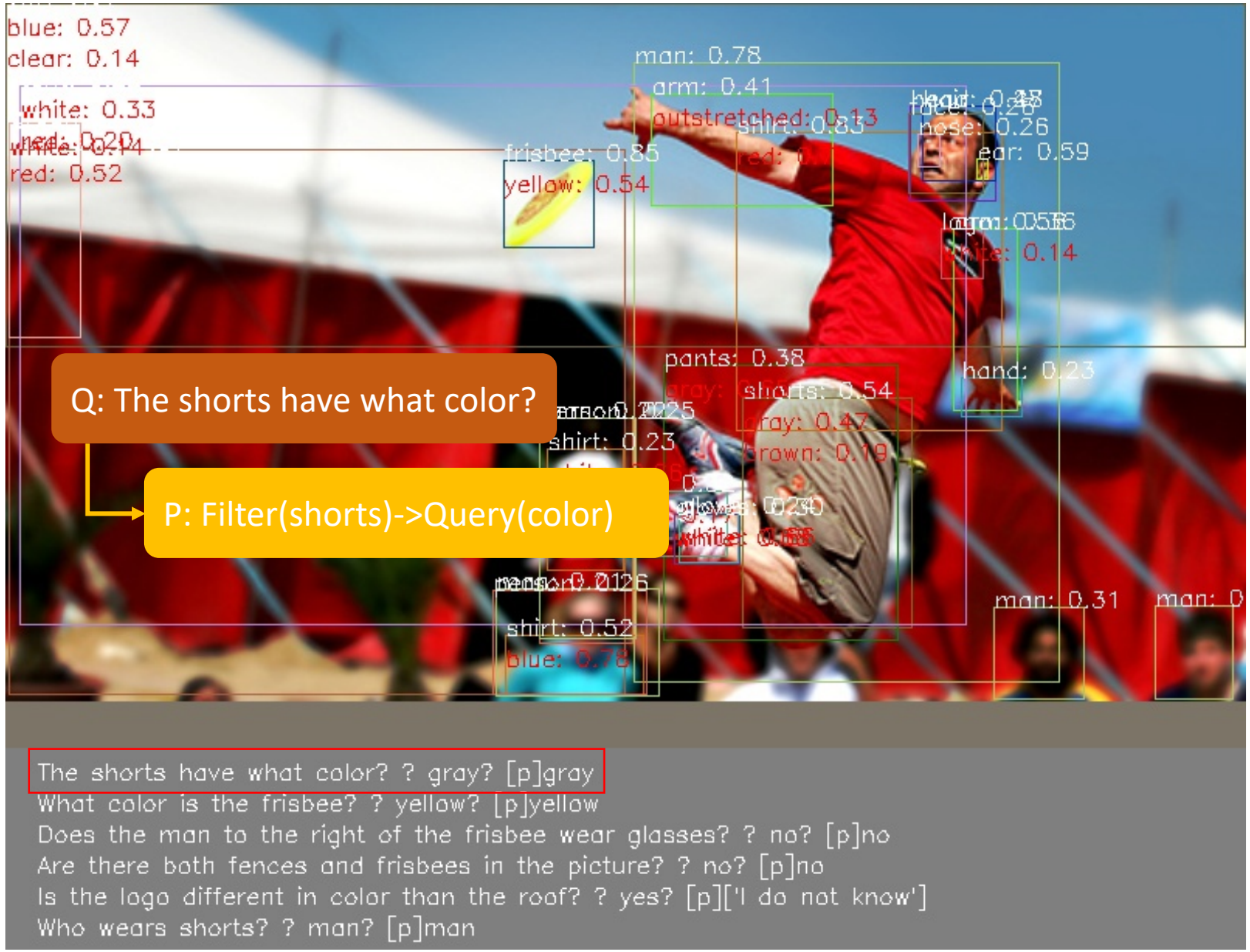
- Functional Programs
- Entailment Relations
- Recursive Reachability

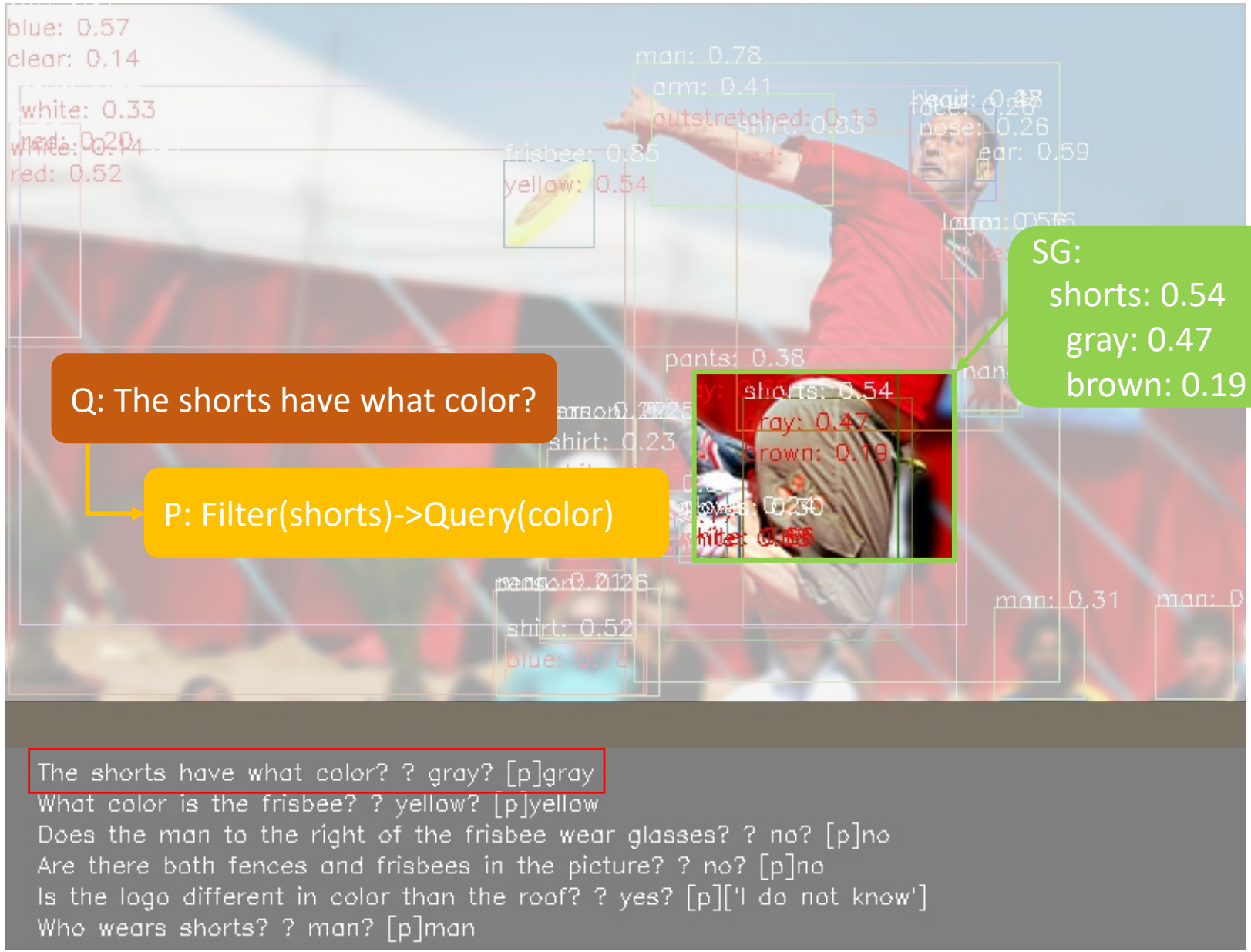
New Metrics

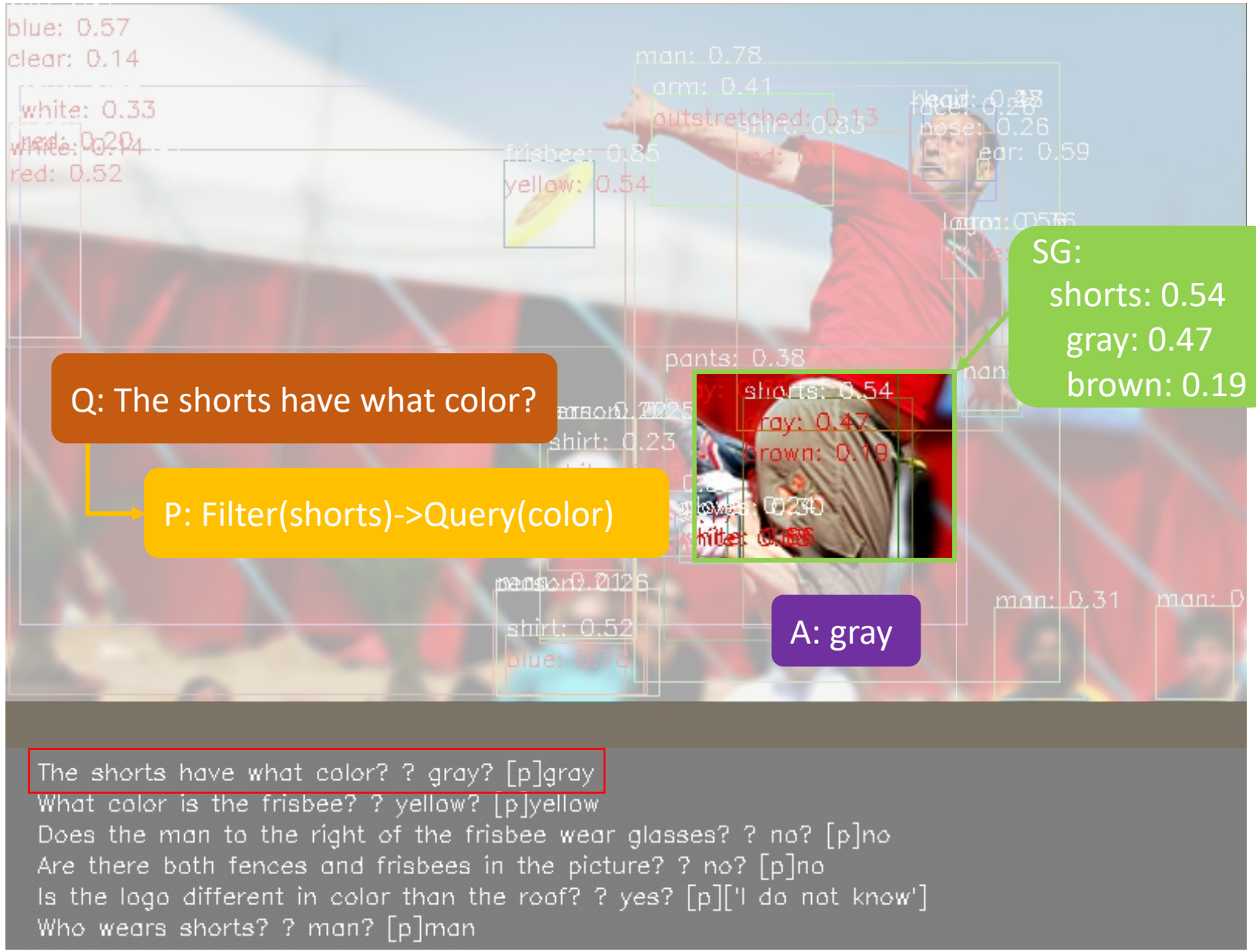
- Consistency
- Validity & Plausibility
- Distribution
- Grounding

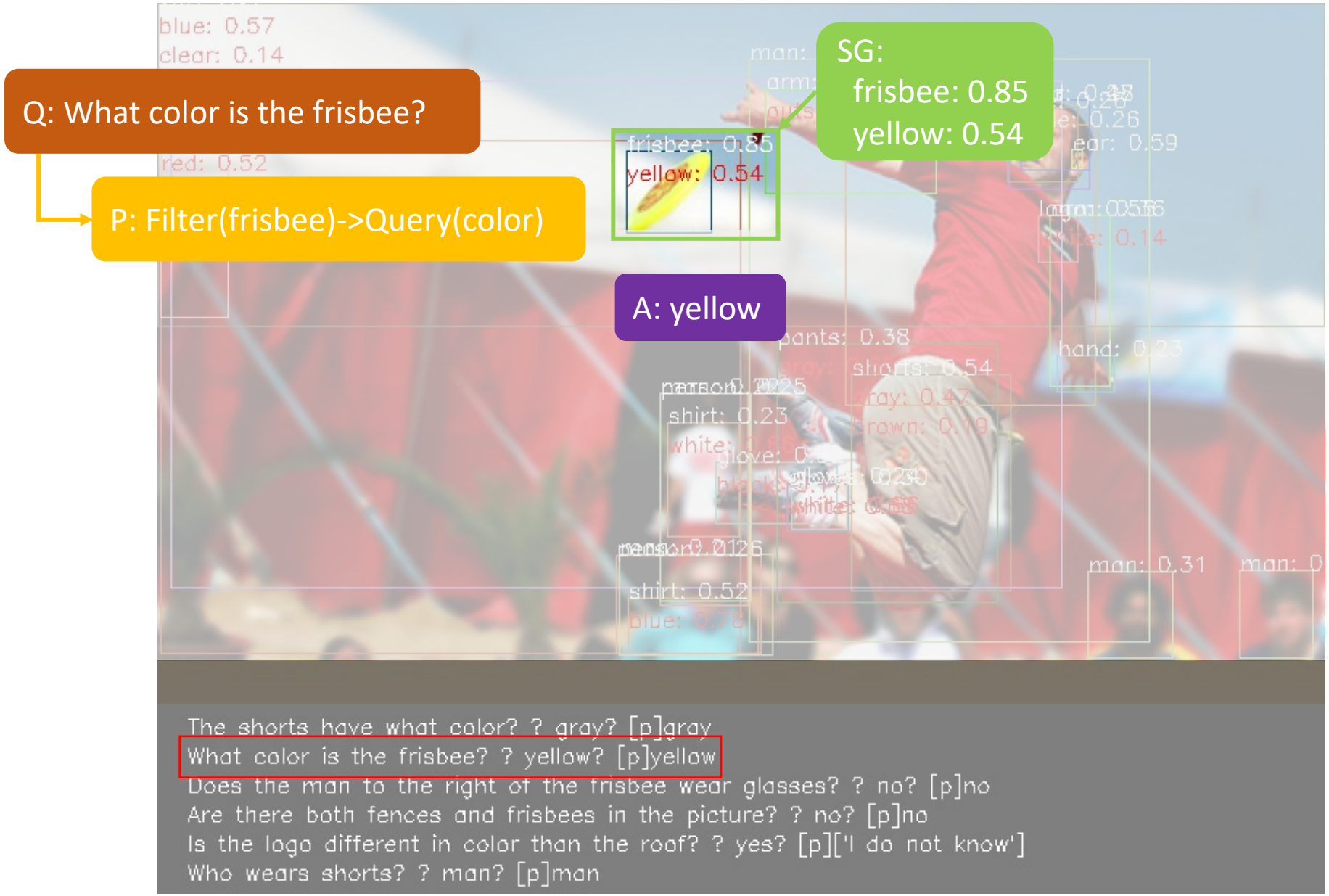
[1] GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. Hudson et al. CVPR 2019

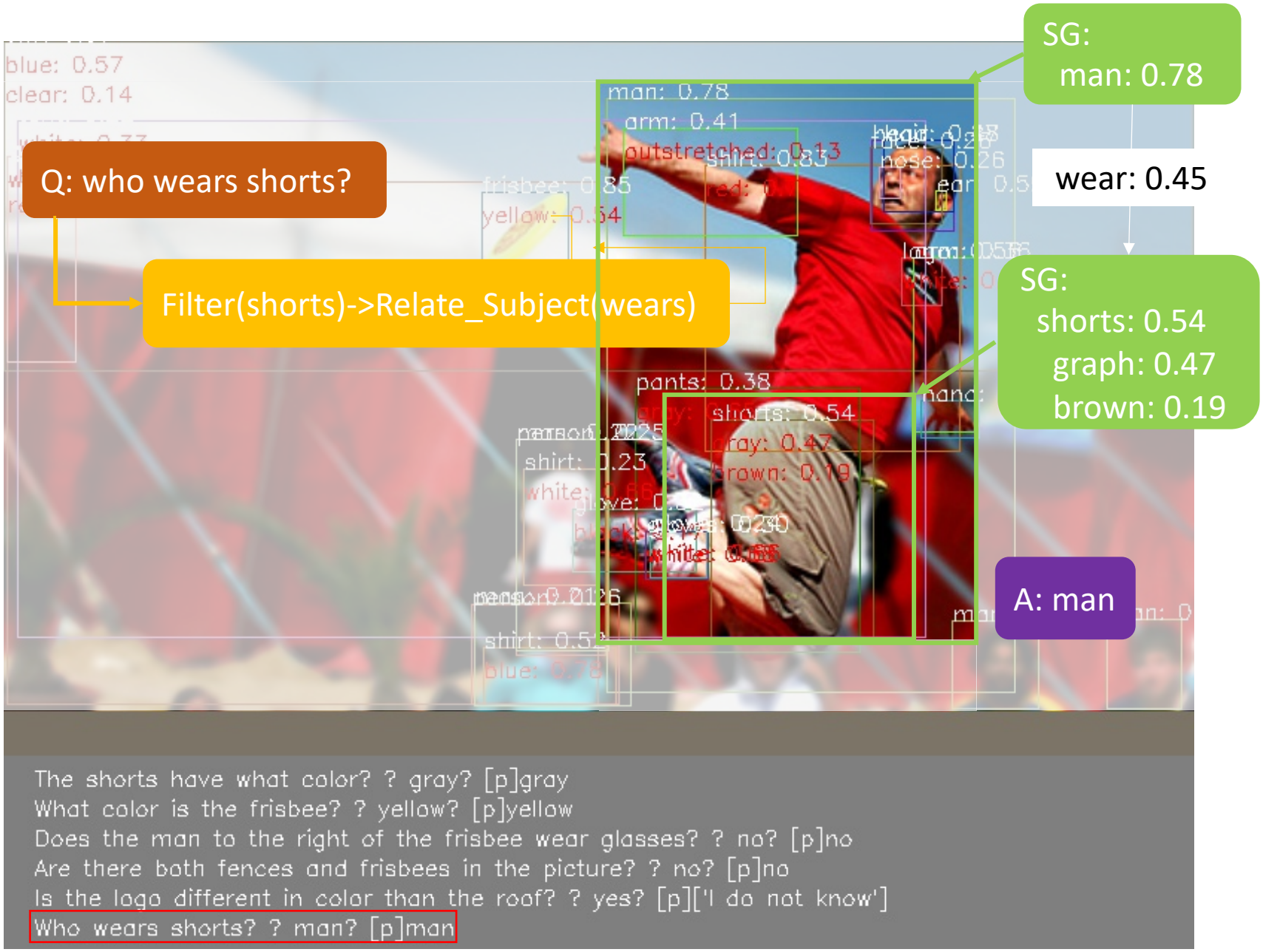

blue: 0.57
clear: 0.14
white: 0.33
red: 0.52
man: 0.78
arm: 0.41
outstretched: 0.13
shirt: 0.83
nose: 0.26
ear: 0.59
frisbee: 0.85
yellow: 0.54
hand: 0.23
pants: 0.38
shorts: 0.54
gray: 0.47
brown: 0.19
shirt: 0.23
shirt: 0.52
blue: 0.78
man: 0.31
The shorts have what color? ? gray? [p]gray
What color is the frisbee? ? yellow? [p]yellow
Does the man to the right of the frisbee wear glasses? ? no? [p]no
Are there both fences and frisbees in the picture? ? no? [p]no
Is the logo different in color than the roof? ? yes? [p]['I do not know']
Who wears shorts? ? man? [p]man


blue: 0.57
clear: 0.14
white: 0.33
red: 0.52
man: 0.78
arm: 0.41
outstretched: 0.13
shirt: 0.83
nose: 0.26
ear: 0.59
frisbee: 0.85
yellow: 0.54
hand: 0.23
pants: 0.38
shorts: 0.54
shirt: 0.23
brown: 0.19
shirt: 0.52
blue: 0.78
man: 0.31
Q: The shorts have what color?
P: Filter(shorts)->Query(color)
The shorts have what color? ? gray? [p]gray
What color is the frisbee? ? yellow? [p]yellow
Does the man to the right of the frisbee wear glasses? ? no? [p]no
Are there both fences and frisbees in the picture? ? no? [p]no
Is the logo different in color than the roof? ? yes? [p]['I do not know']
Who wears shorts? ? man? [p]man

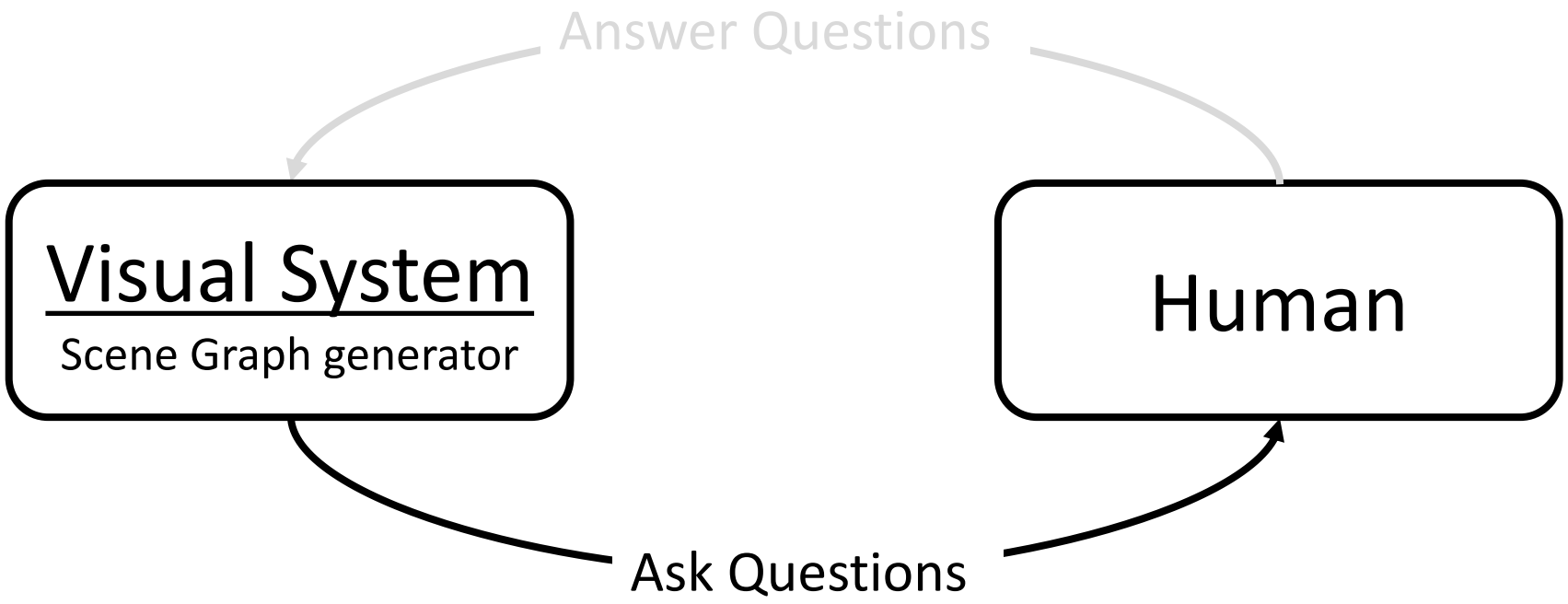

blue: 0.57
clear: 0.14
white: 0.33
red: 0.52
man: 0.78
arm: 0.41
outstretched: 0.13
shirt: 0.83
nose: 0.26
ear: 0.59
frisbee: 0.85
yellow: 0.54
pants: 0.38
shirt: 0.23
shirt: 0.52
man: 0.31
SG:
shorts: 0.54
gray: 0.47
brown: 0.19
Q: The shorts have what color?
P: Filter(shorts)->Query(color)
The shorts have what color? ? gray? [p]gray
What color is the frisbee? ? yellow? [p]yellow
Does the man to the right of the frisbee wear glasses? ? no? [p]no
Are there both fences and frisbees in the picture? ? no? [p]no
Is the logo different in color than the roof? ? yes? [p]['I do not know']
Who wears shorts? ? man? [p]man


blue: 0.57
clear: 0.14
white: 0.33
red: 0.52
man: 0.78
arm: 0.41
outstretched: 0.13
shirt: 0.83
nose: 0.26
ear: 0.59
frisbee: 0.85
yellow: 0.54
pants: 0.38
shirt: 0.23
shirt: 0.52
man: 0.31
SG:
shorts: 0.54
gray: 0.47
brown: 0.19
Q: The shorts have what color?
P: Filter(shorts)->Query(color)
A: gray
The shorts have what color? ? gray? [p]gray
What color is the frisbee? ? yellow? [p]yellow
Does the man to the right of the frisbee wear glasses? ? no? [p]no
Are there both fences and frisbees in the picture? ? no? [p]no
Is the logo different in color than the roof? ? yes? [p]['I do not know']
Who wears shorts? ? man? [p]man


Q: What color is the frisbee?
P: Filter(frisbee)->Query(color)
SG:
frisbee: 0.85
yellow: 0.54
A: yellow
The shorts have what color? ? gray? [p]gray
What color is the frisbee? ? yellow? [p]yellow
Does the man to the right of the frisbee wear glasses? ? no? [p]no
Are there both fences and frisbees in the picture? ? no? [p]no
Is the logo different in color than the roof? ? yes? [p]['I do not know']
Who wears shorts? ? man? [p]man

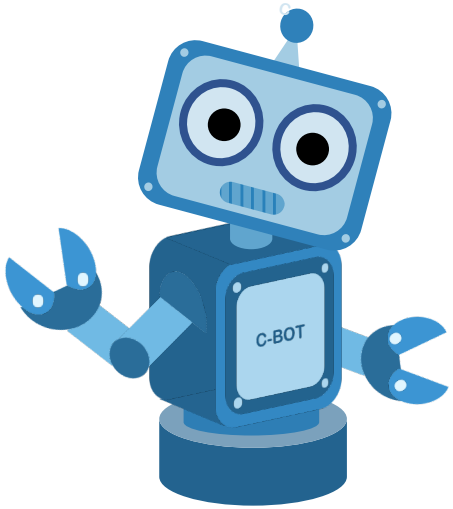

SG:
man: 0.78
wear: 0.45
SG:
shorts: 0.54
graph: 0.47
brown: 0.19
Q: who wears shorts?
Filter(shorts)->Relate_Subject(wears)
A: man
The shorts have what color? ? gray? [p]gray
What color is the frisbee? ? yellow? [p]yellow
Does the man to the right of the frisbee wear glasses? ? no? [p]no
Are there both fences and frisbees in the picture? ? no? [p]no
Is the logo different in color than the roof? ? yes? [p]['I do not know']
Who wears shorts? ? man? [p]man

Visual Question Answering

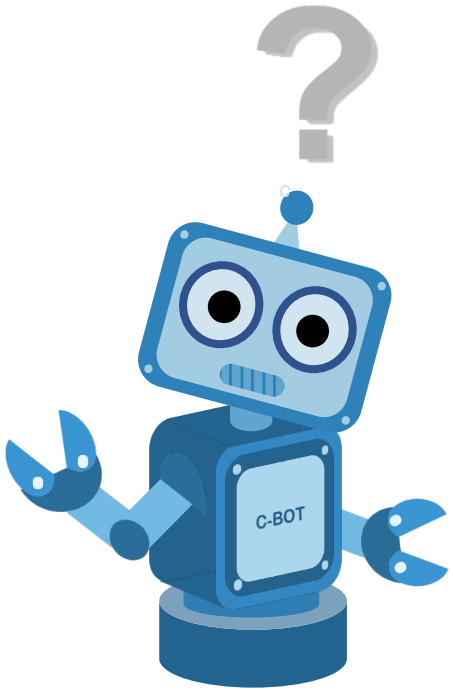


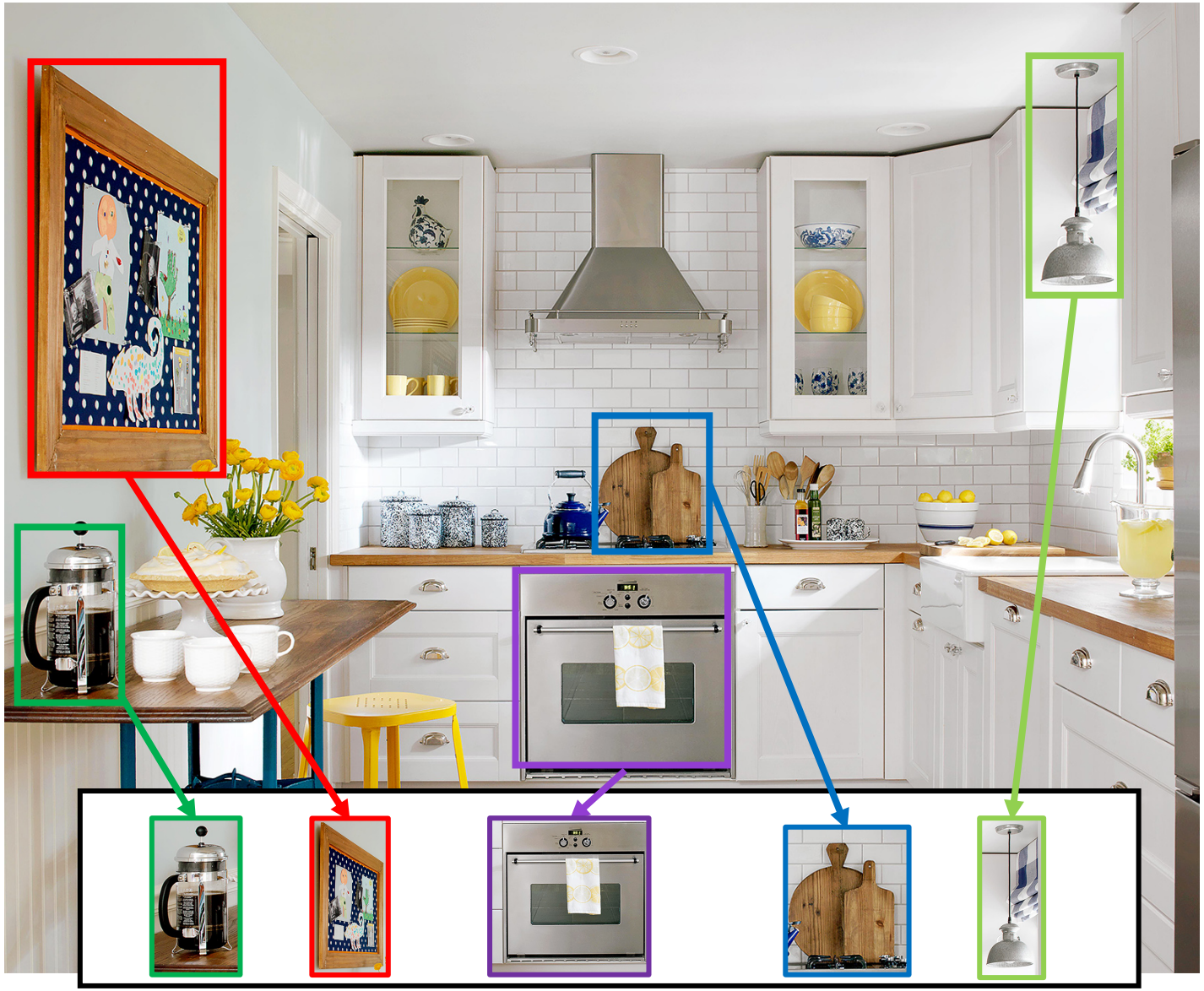
Visual Question Generation

The Open-World Recognition Problem

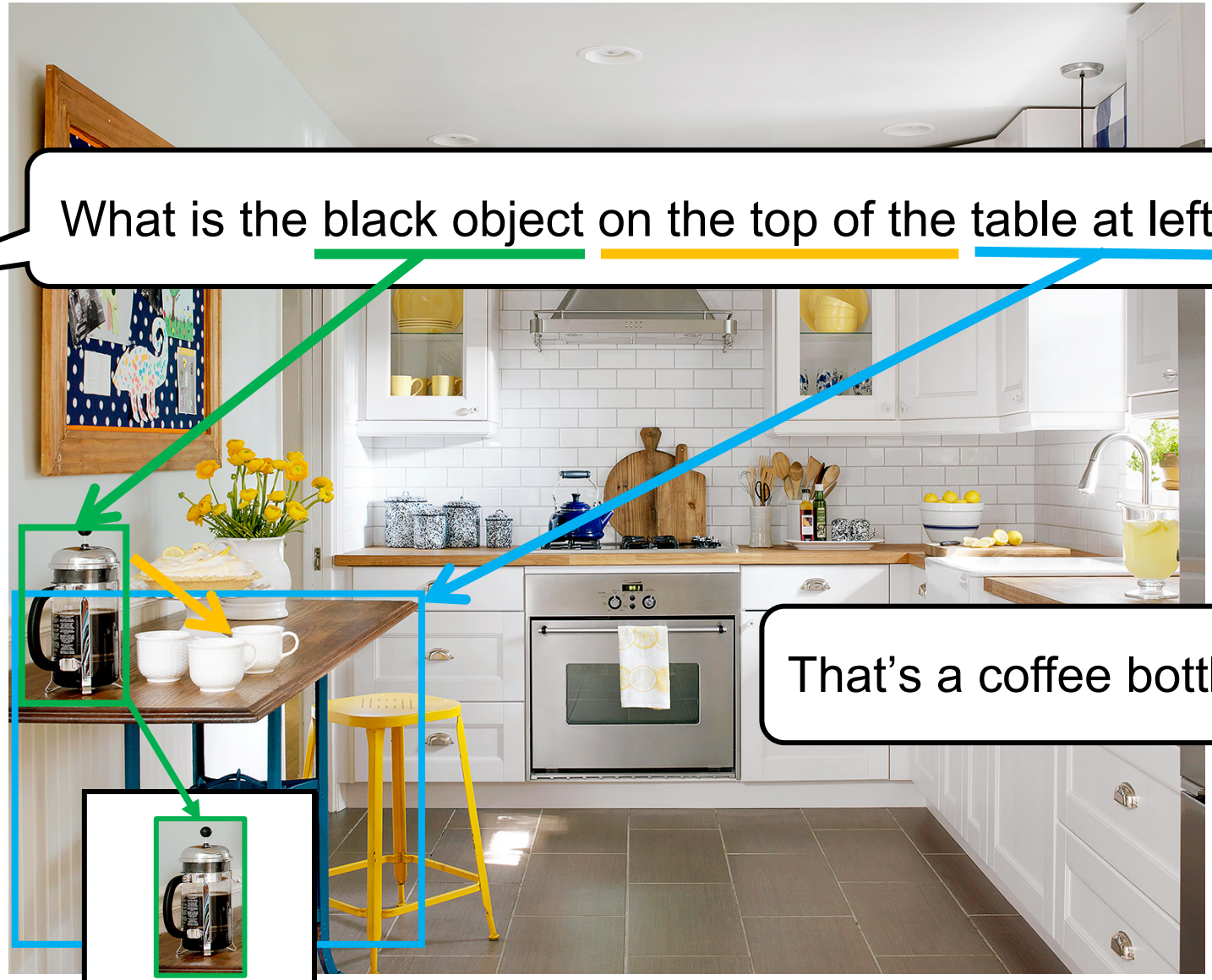


The Open-World Recognition Problem



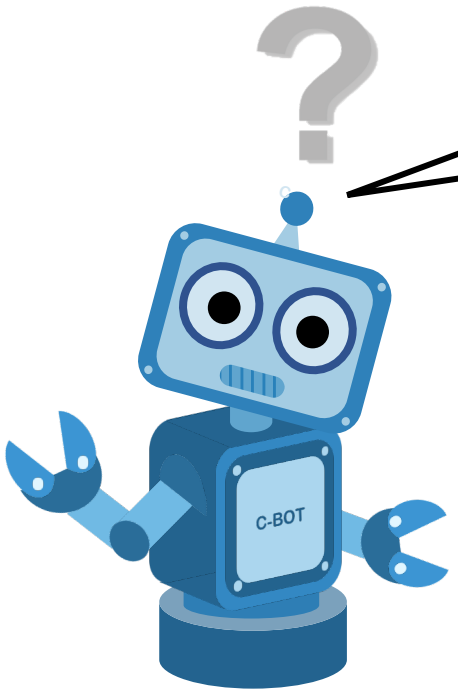


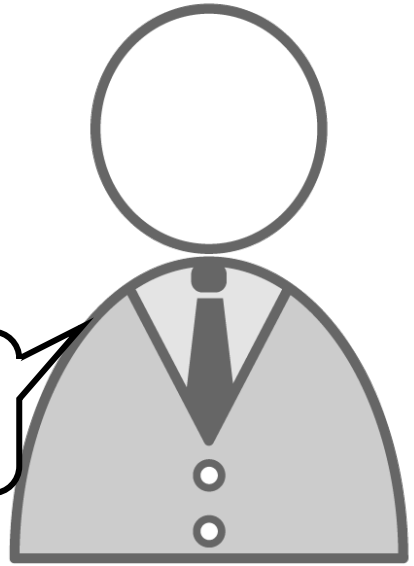
The Open-World Recognition Problem



What is the black object on the top of the table at left side?

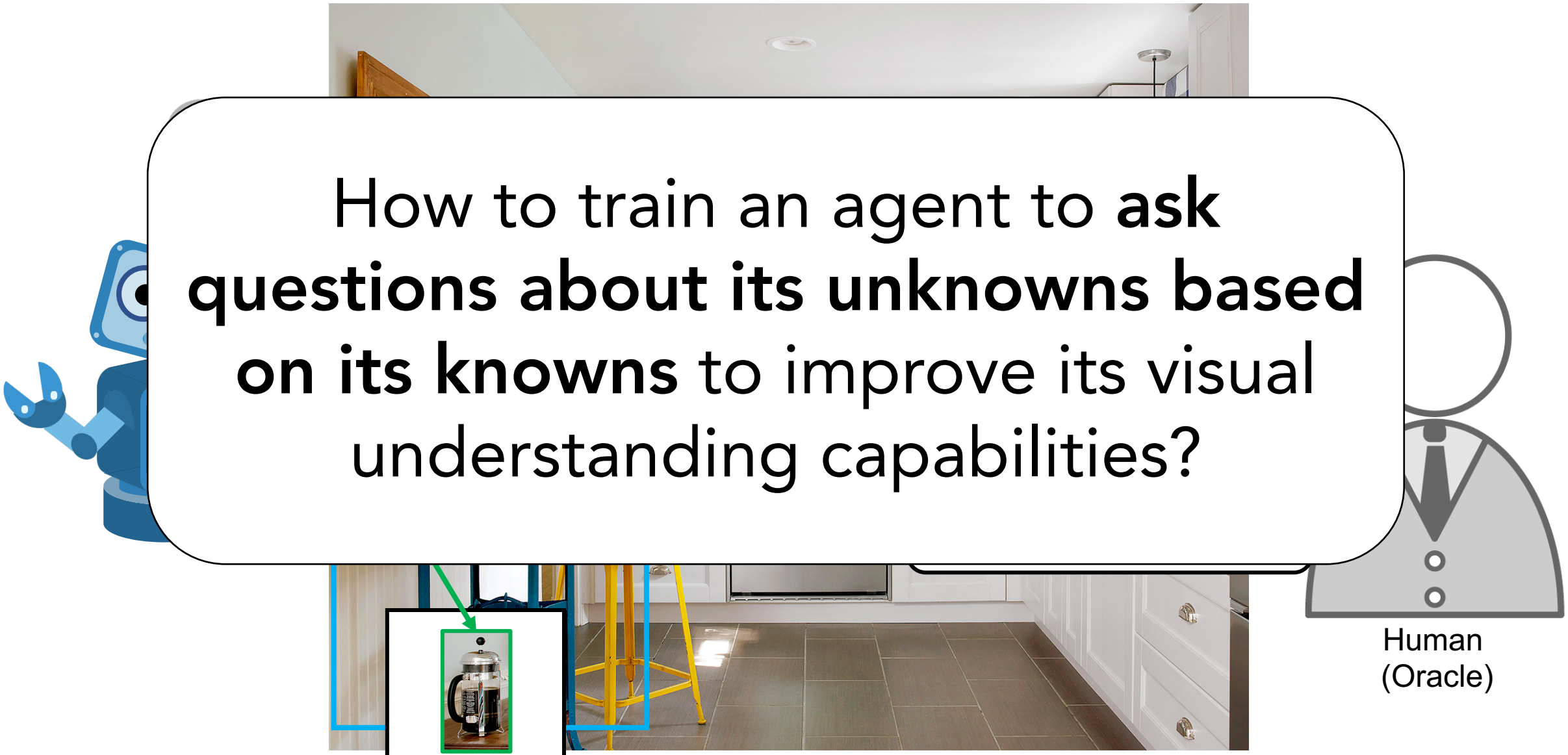
That's a coffee bottle.





Human
(Oracle)

The Open-World Recognition Problem

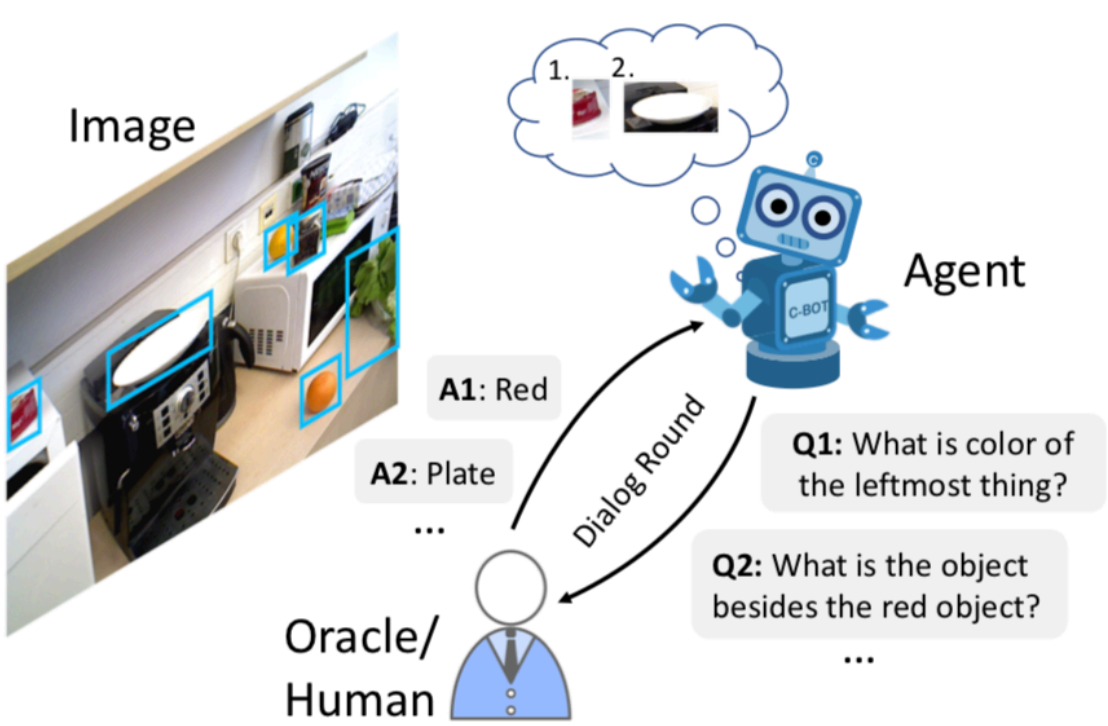


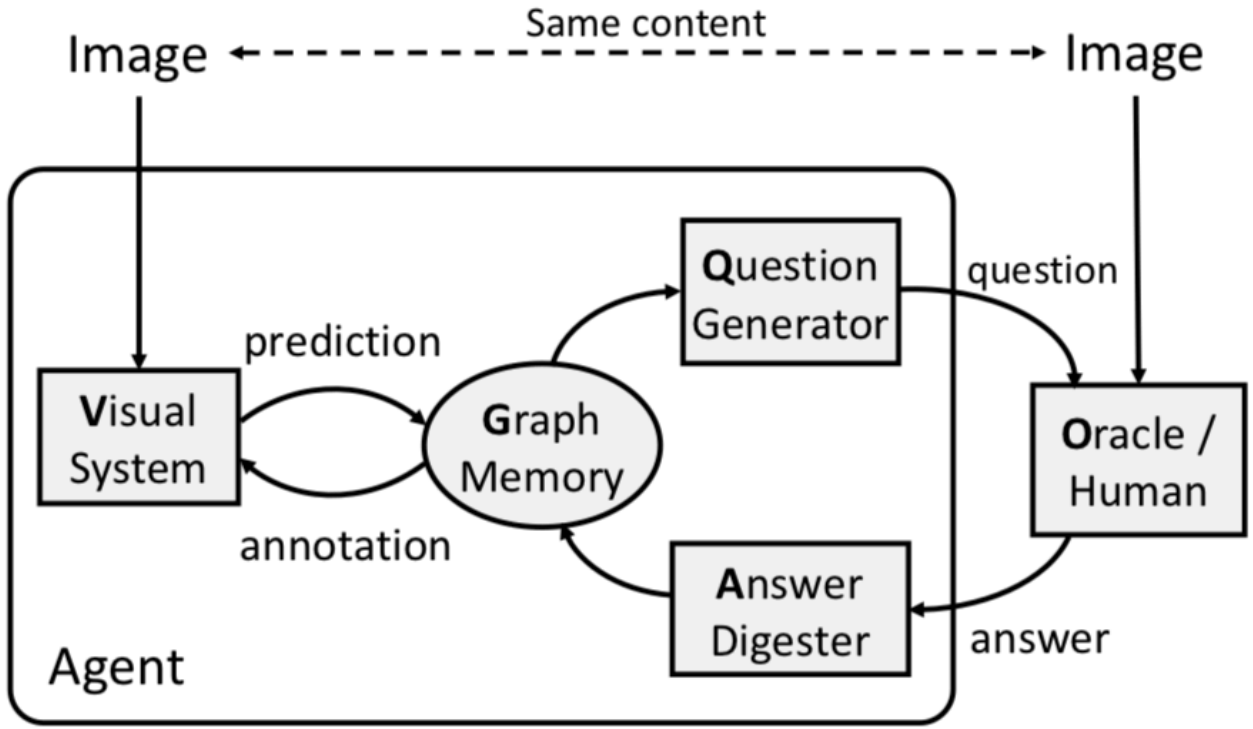
How to train an agent to **ask questions about its unknowns based on its knowns** to improve its visual understanding capabilities?

Human
(Oracle)

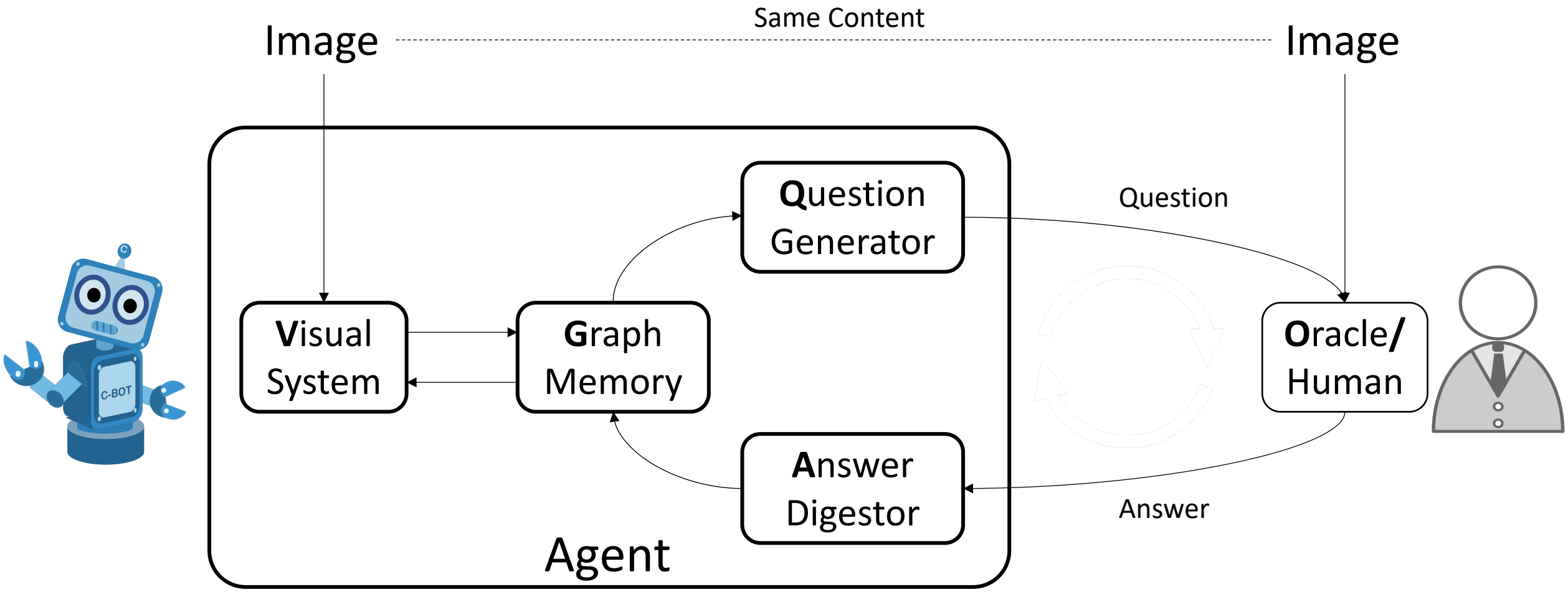
Visual Question Generation: Visual Curiosity

Visual Curiosity: Learning to Ask Questions to Learn Visual Recognition. Yang and Lu et al. CoRL 2018



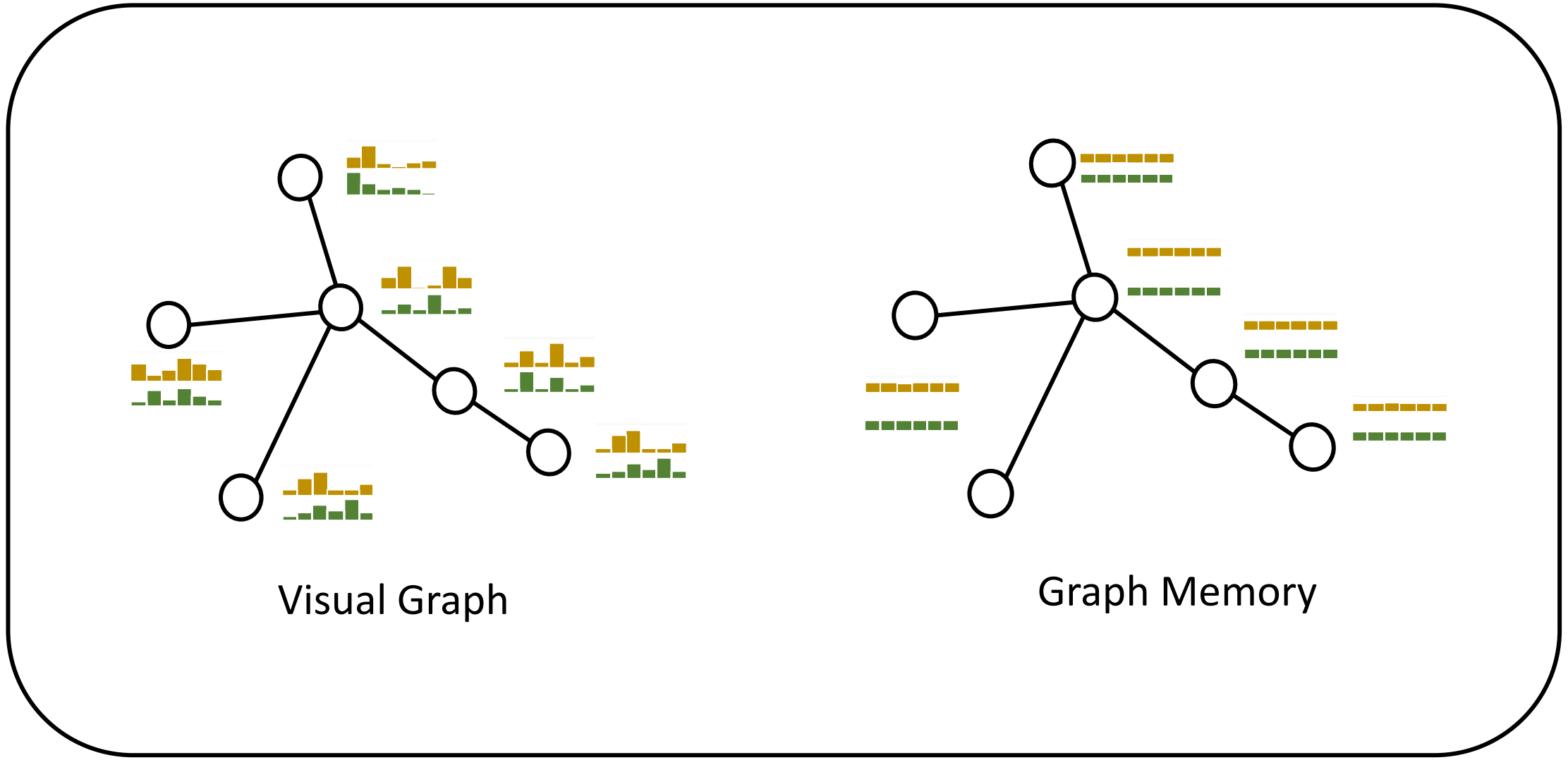


Agent Architecture

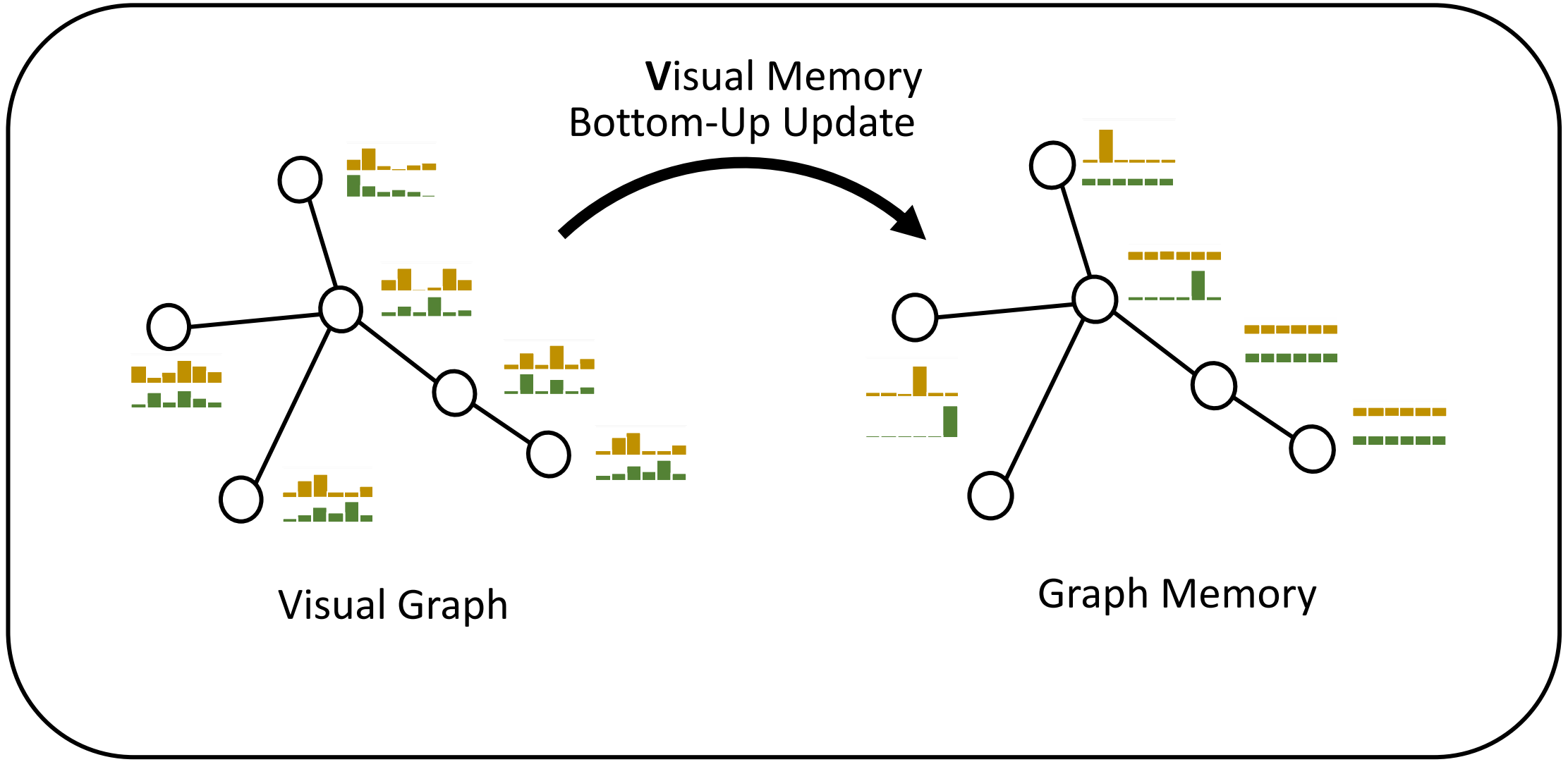


Slide Credit: Stefan Lee

Bottom-up Update

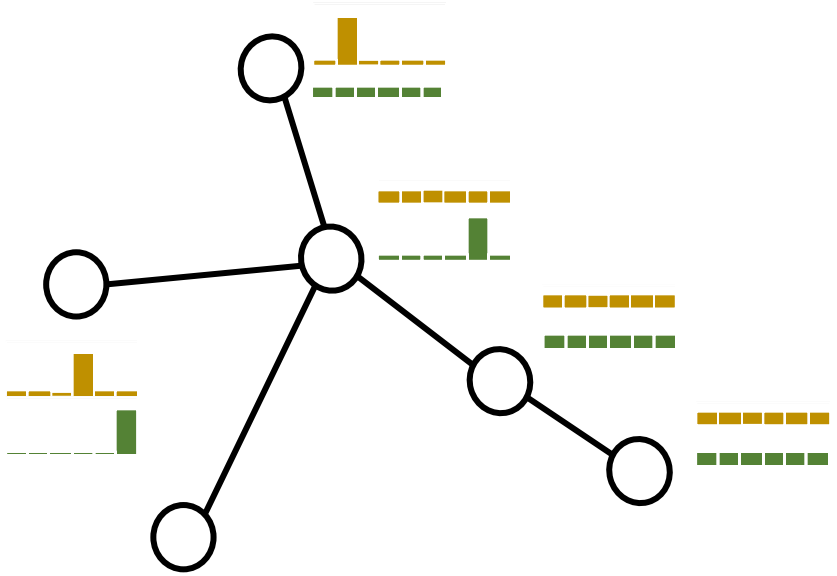


Bottom-up Update



Top-down Update

Visual Memory

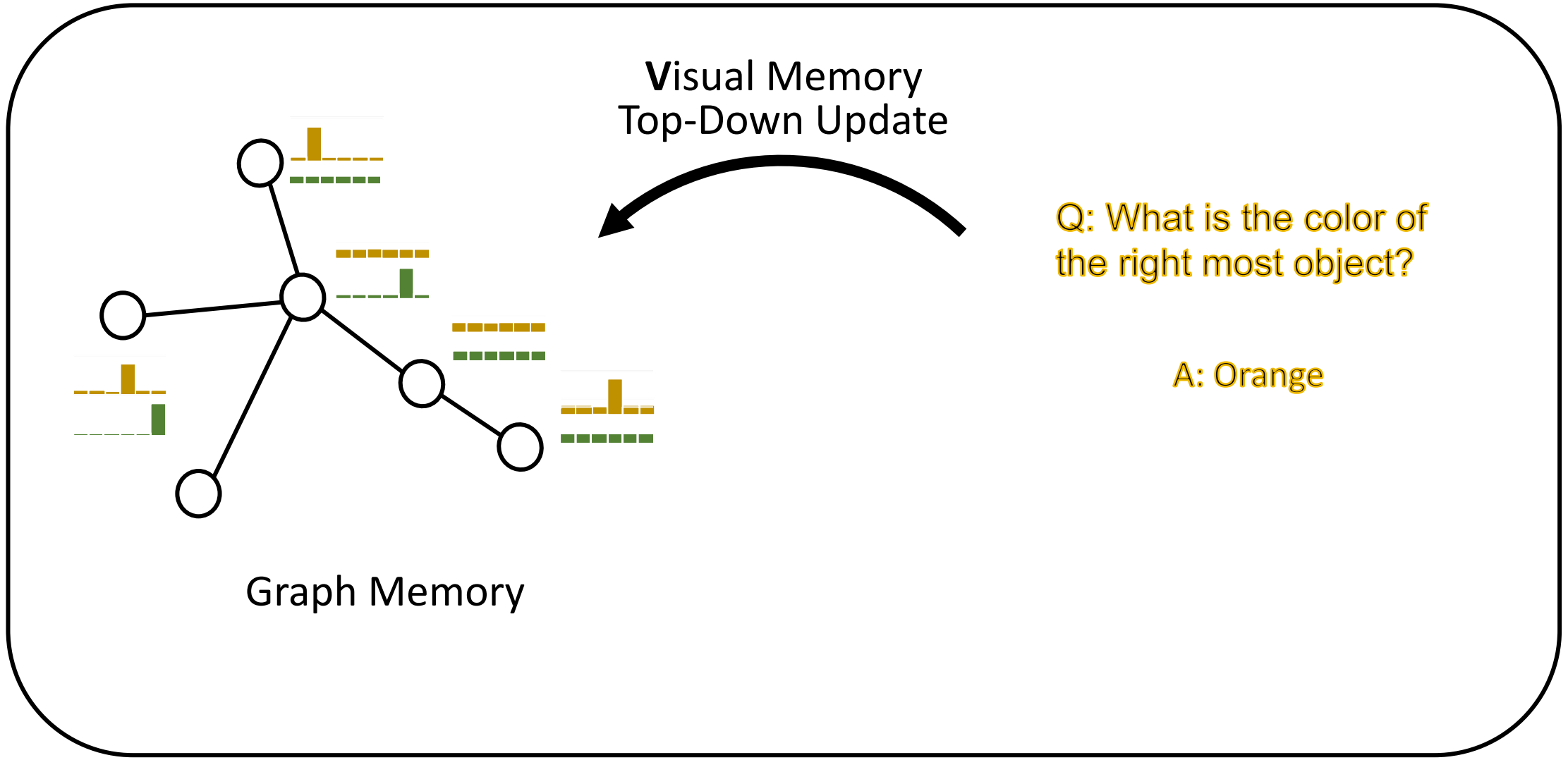


Memory Graph

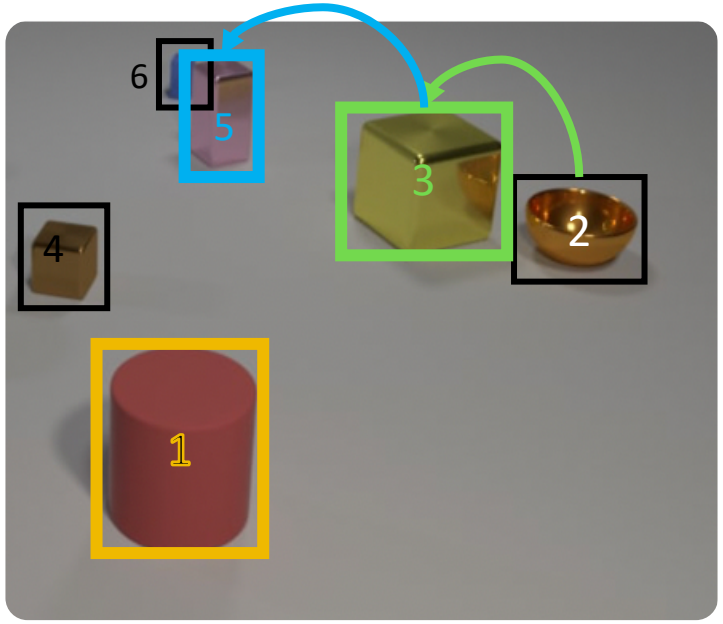
Q: What is the color of the right most object?

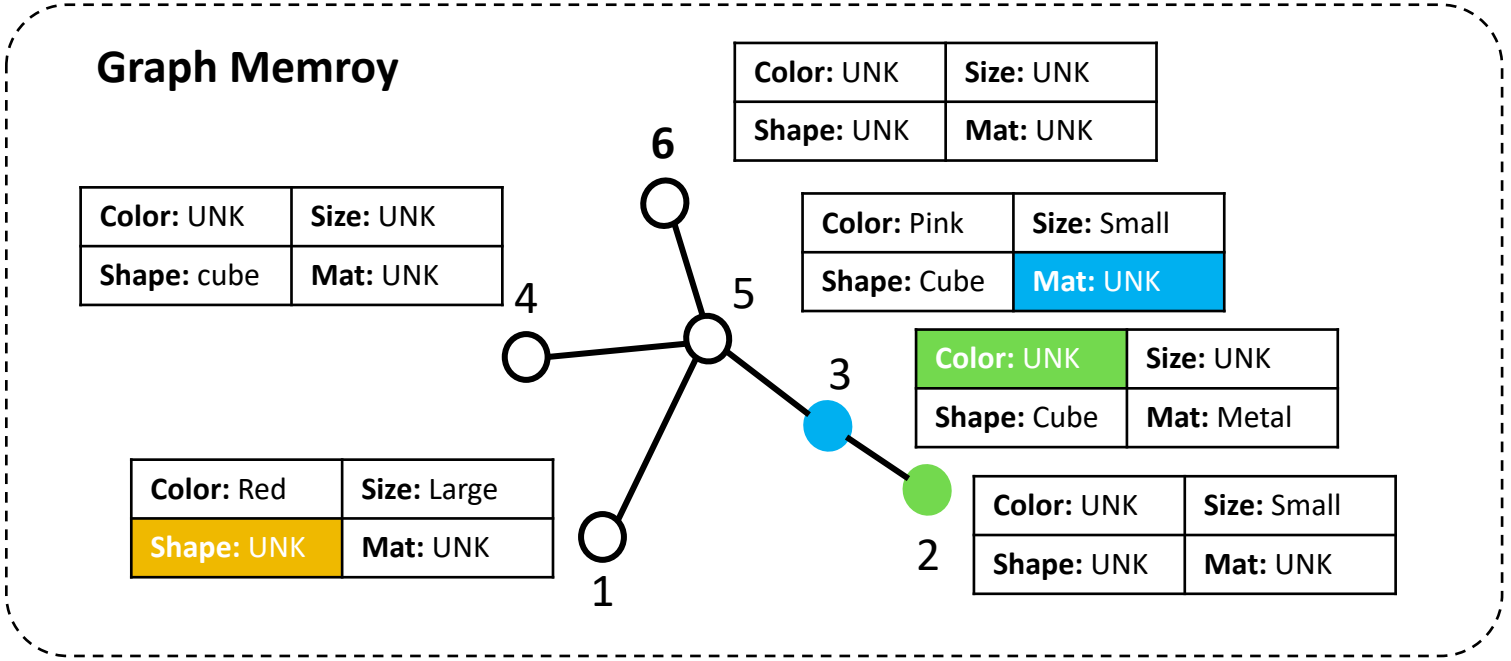
A: Orange

Top-down Update



Question Generator



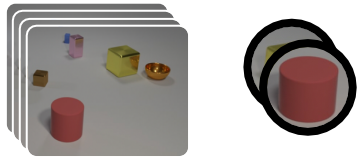


Target	1	3	5
Attribute	Shape	Color	Material
Reference	None	2	3
Question	What is the shape of the front most large red object?	What is the color of the metal cube on the left side of a small object?	What is the material of object at left side of metal cube?

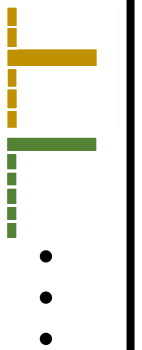
Training Objective: Visual System

Loss: cross-entropy loss between the graph memory and the visual predictions over all images, objects, and attributes

$$\theta_v^* = \arg\min -\sum_{I_i \in I} \sum_{k=1}^{K_i} \log\left(\left[\quad \cdots\right]\right)\begin{bmatrix} \\ \\ \\ \end{bmatrix}$$



Visual attribute predictions



Graph memory

Update visual system after a piece of dialogs with Oracle/Human.

Training Objective: Questioner Policy

Oracle Graph

Reward: $r_i^t = S(G_i^t, G_i^*) - S(G_i^{t-1}, G_i^*)$

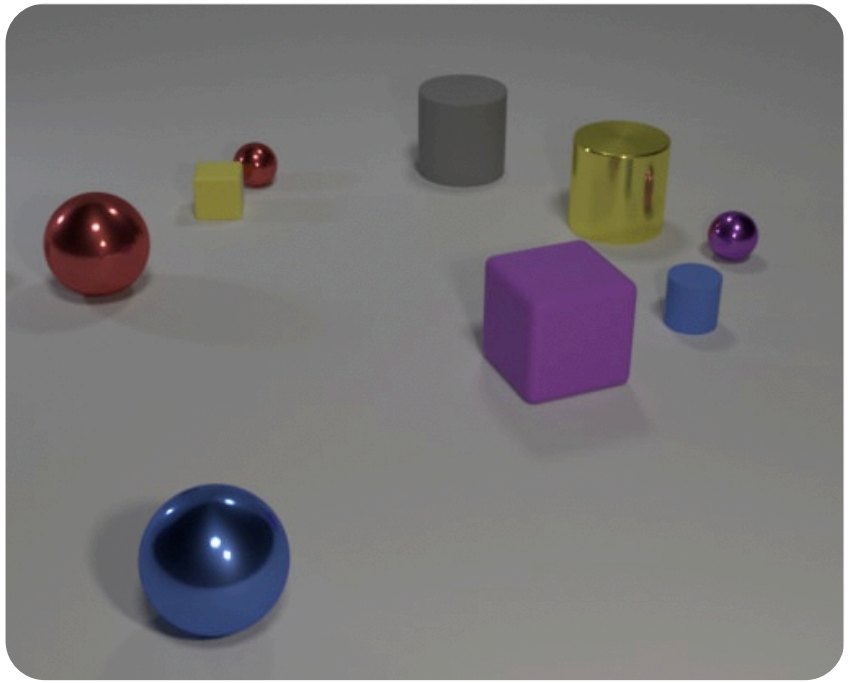
Optimal Policy: $\theta_\pi^* = \arg\max E_V E_{I\sim\mathcal{E}} E_{\pi_q} \left[\sum_{i=1}^{n} \sum_{t=1}^{T} r_i^t (q_i^t \sim \pi_q(h_i^t; \theta_\pi) \right]$

Use A2C and update policy after each episode based on all rounds

Can be improved by:

- Asking unambiguous, informative questions (top-down)
- Improving the visual system quickly (bottom-up)

Experiments: Environments



Synthesized Dataset

Different shapes, colors, materials and sizes. Extended from CLEVR dataset [1]

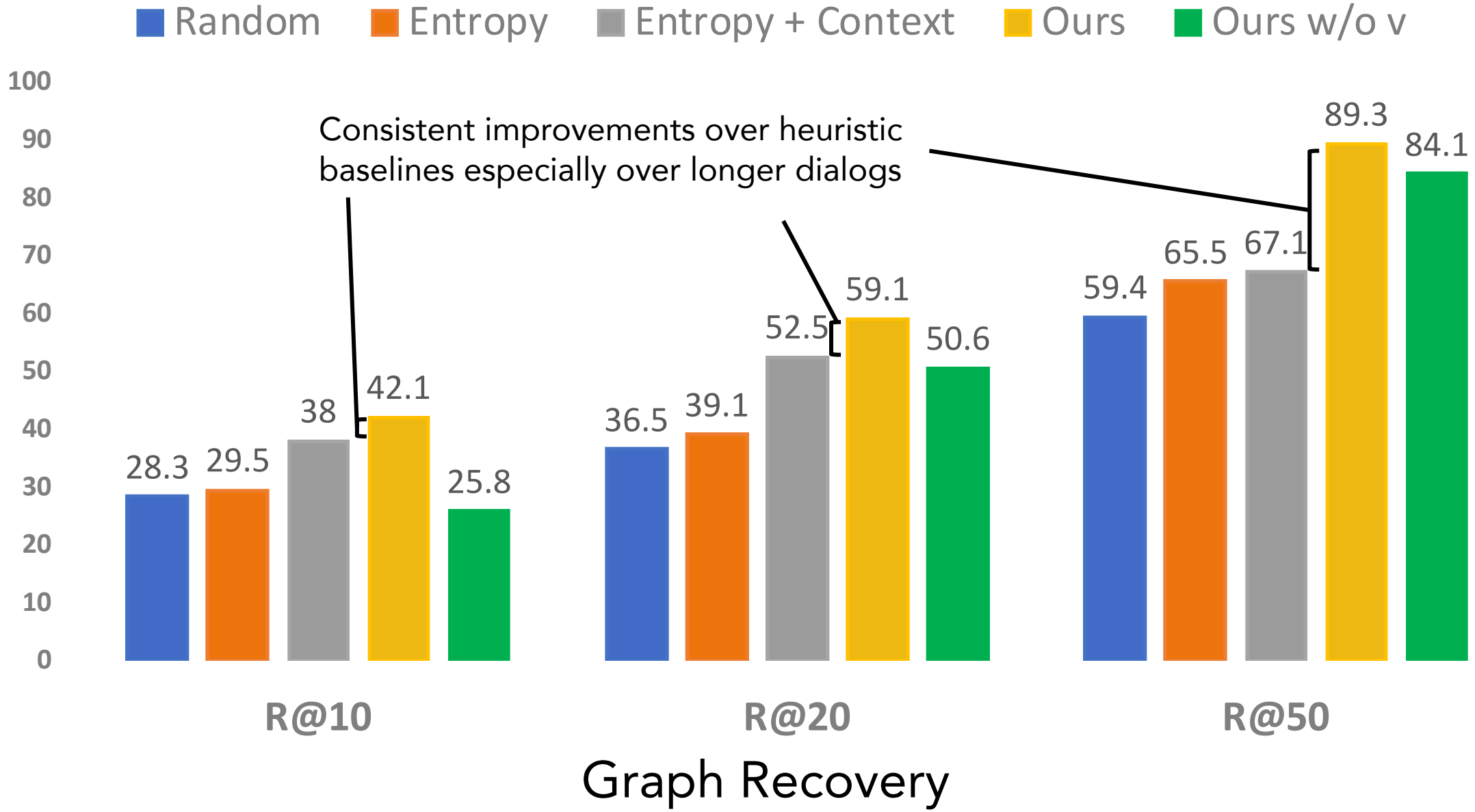
[1] CLEVR. Johnson et al.

Realistic Dataset

Various real indoor scenes.
Annotated based on the ARID dataset [2]

[2] Recognizing Objects In-the-Wild. Loghmani et al.

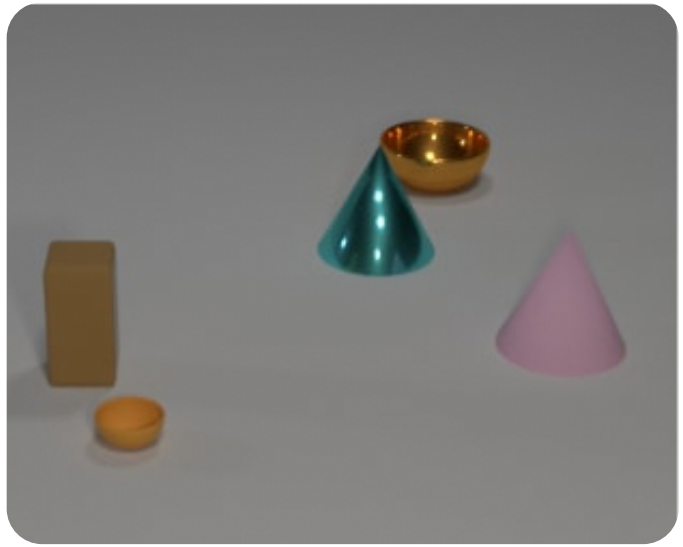
Experiment: Standard Training + Standard Testing



Graph Recovery

Experiments: Novel Object Environments

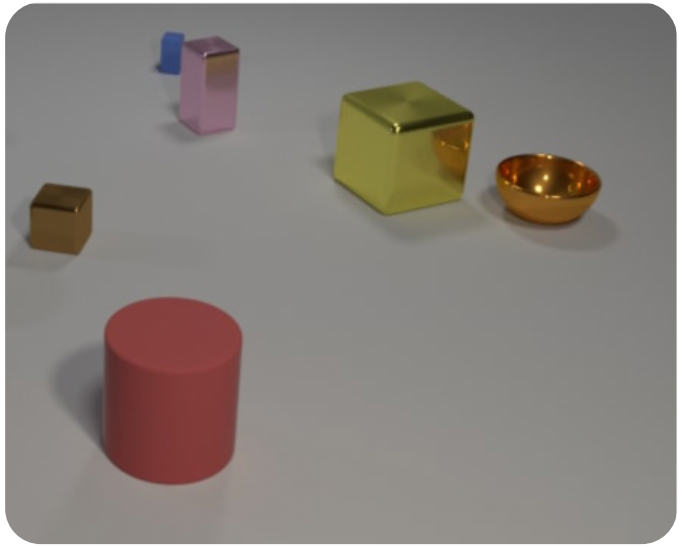
Novel



New colors and shapes

600 images for test (12 episodes)

Mixed



Mix of novel and standard colors and shapes

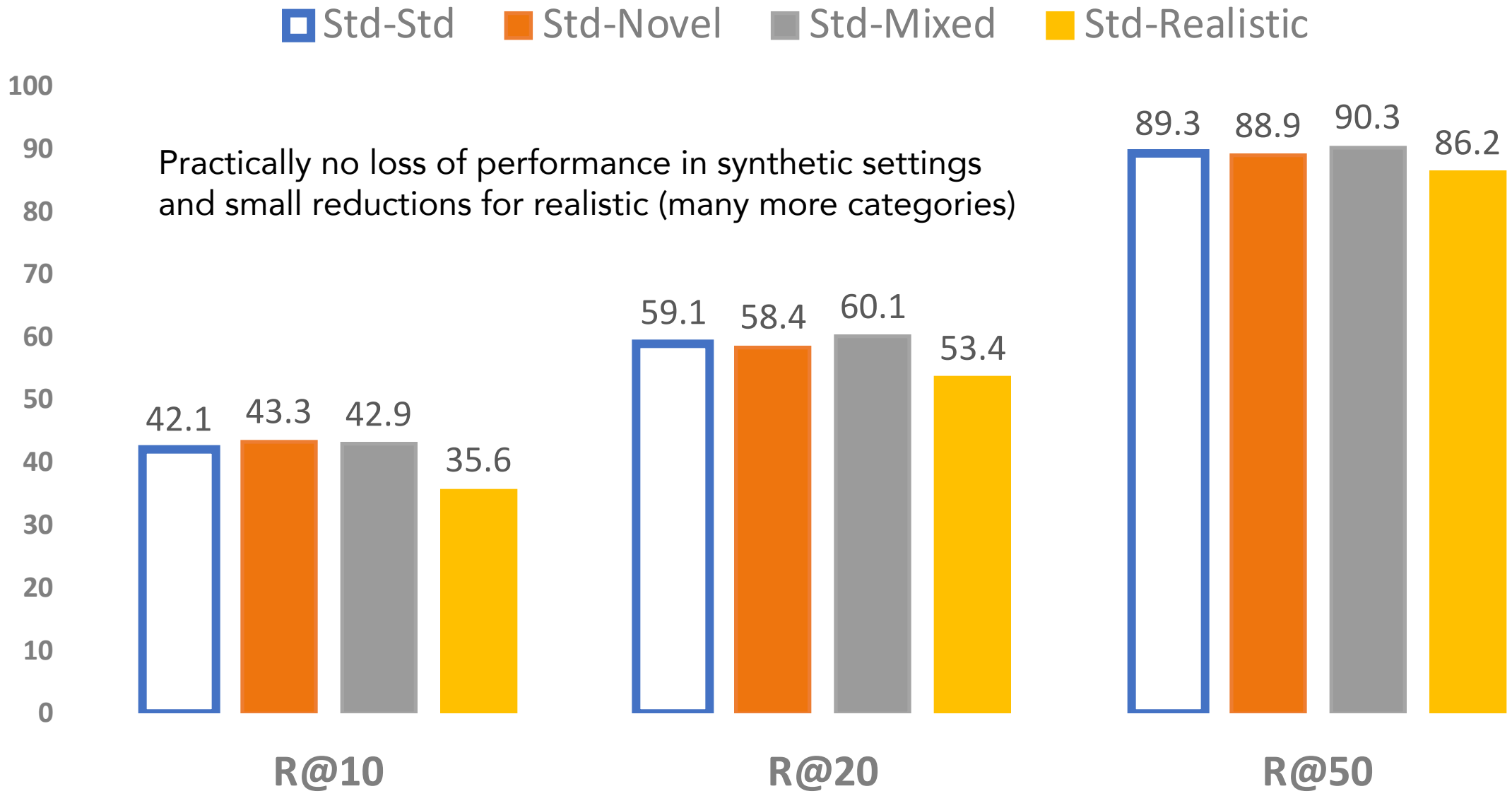
600 images for test (12 episodes)

Realistic

51 categories, 11 colors, 6 materials

1200 images for test (24 episodes)

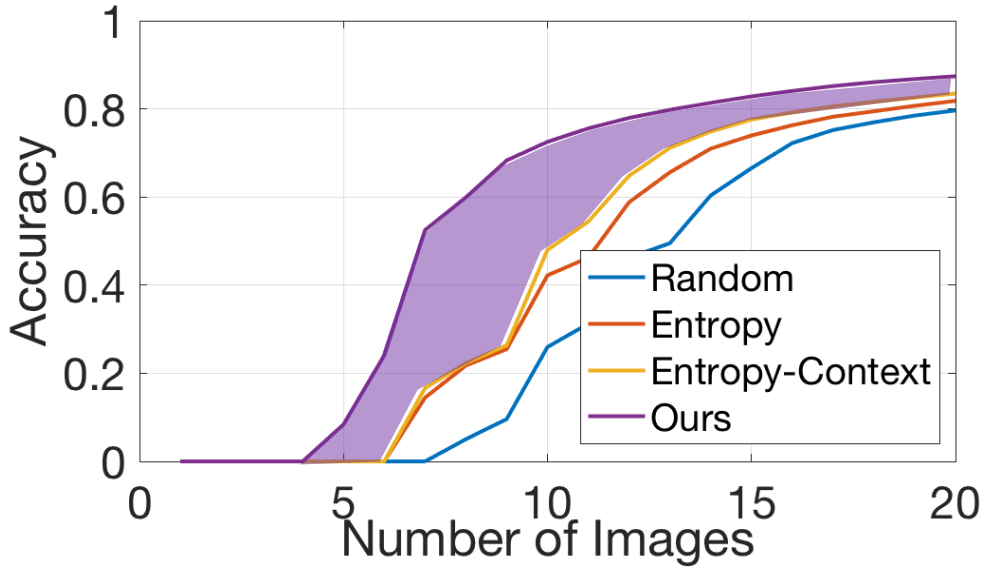
Experiments: Standard Train – New Test Environments



Graph Recovery

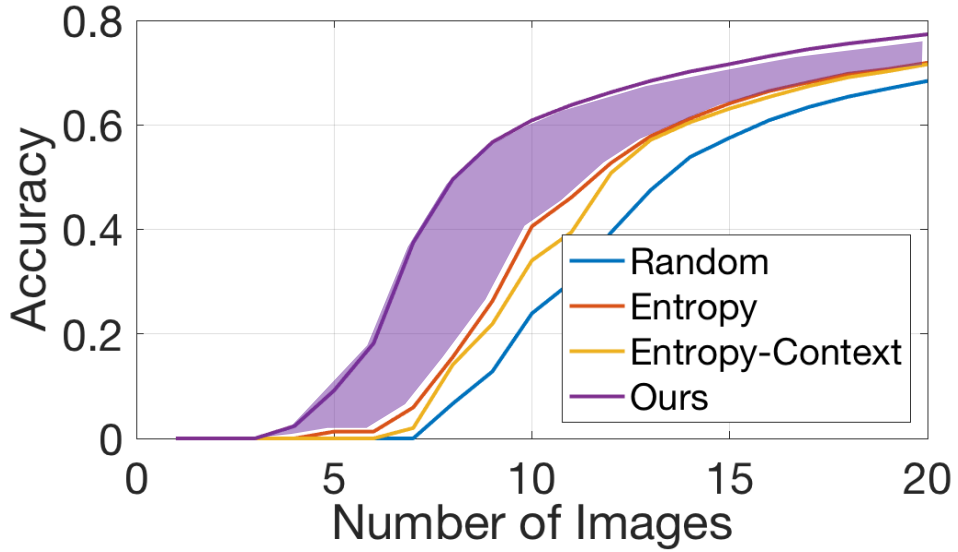
Experiments: Visual System Performance

Standard

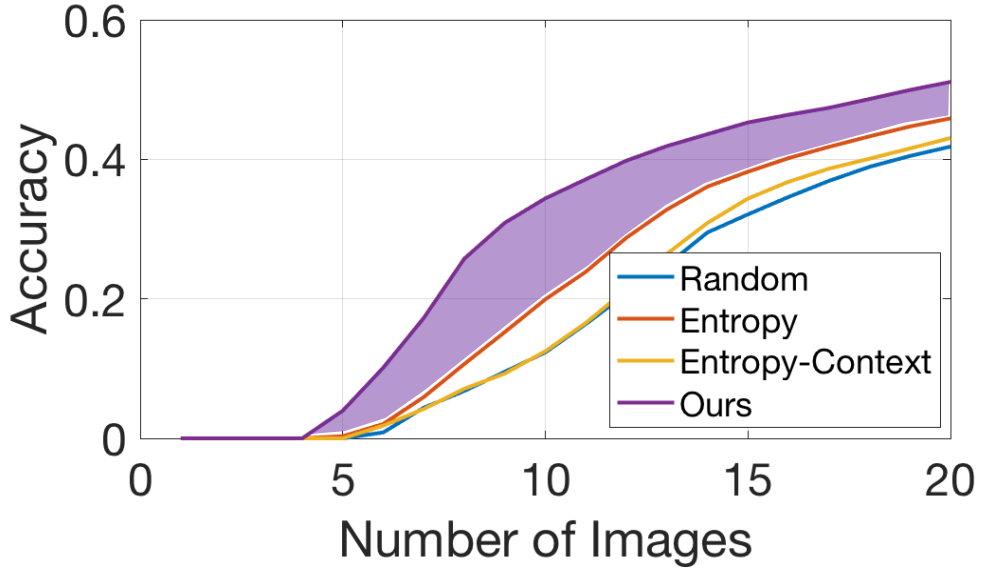


Novel

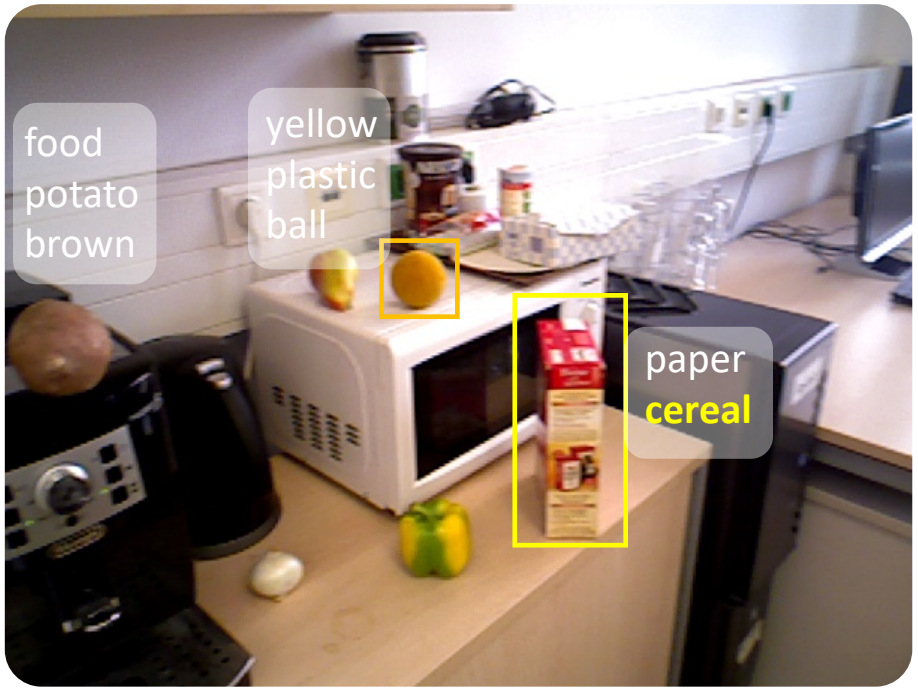
Mixed



Realistic



Experiments: Qualitative Example



What material is the leftmost thing?	food
There is a leftmost object; what is it?	potato
The leftmost object is what color?	brown
⋮	
What is the closest thing that is in front of the yellow plastic ball made of?	paper
What is the closest thing that is in front of the yellow plastic ball?	cereal

Summary for this part

- Takeaways:
  - Scene graph can be used as a comprehensive semantic abstraction of image
  - Scene graph provides grounding information for language-based interaction with human, especially visual question answering and generation
  - Scene graph gives it a chance to make models more interpretable and explainable

- Potential Directions:
  - Leverage scene graph for explicit and effective reasoning on more vision-language tasks, such as expression coreference
  - Language context dependent scene graph generation
  - Combine scene graph and knowledge graph for common sense reasoning

Summary for all

Leveraging external "knowledge" when interpreting images

Specifically, using richer vision, language models/data to improve vision+language models

Representing internal structure in images

Specifically, scene graphs: generating, evaluating and using them for vision+language

Thanks!