Structured Visual Understanding and Interaction with Human and Environment

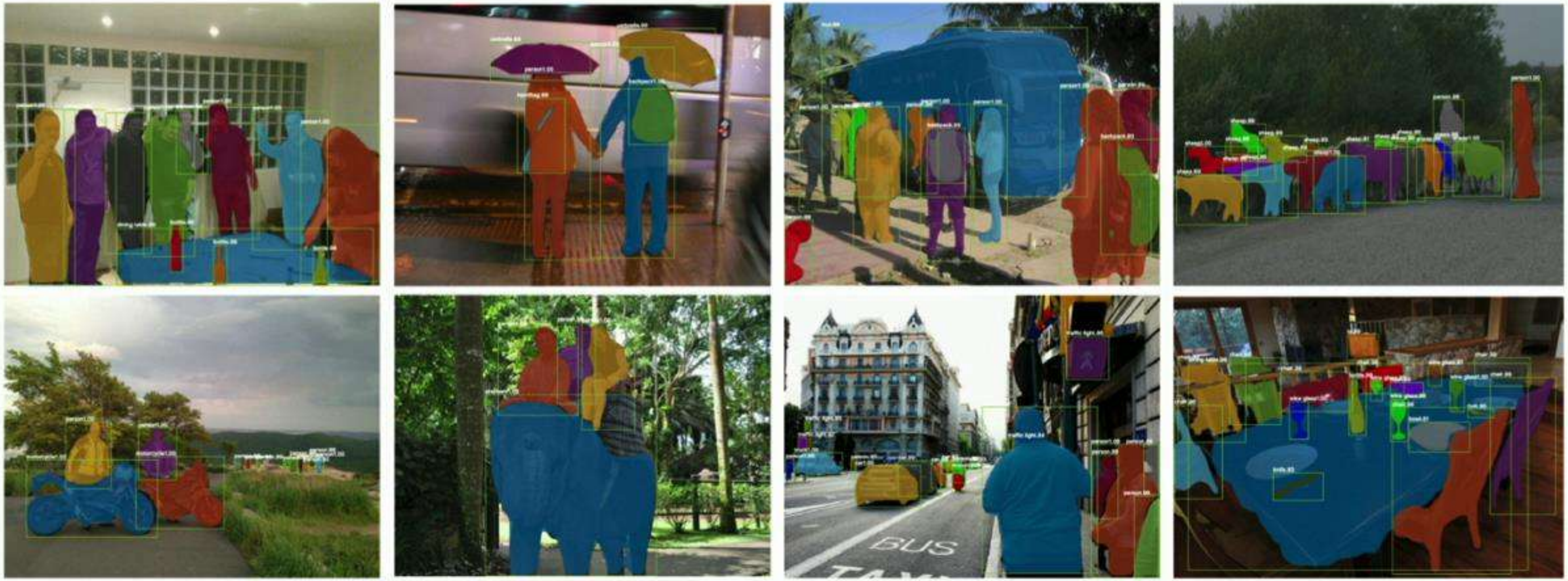
Georgia Tech

Jianwei Yang

09/12/2019

The world around us is highly structured

Images are highly structured



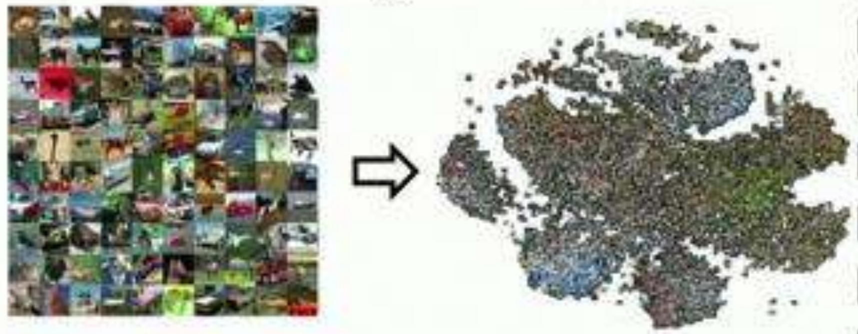
Microsoft COCO: Common Objects in Context. Lin et al. 2014

Images are highly structured

COCO-Stuff: Thing and Stuff Classes in Context. Caesar et al. 2018

My researches:

Cross-Image Structure



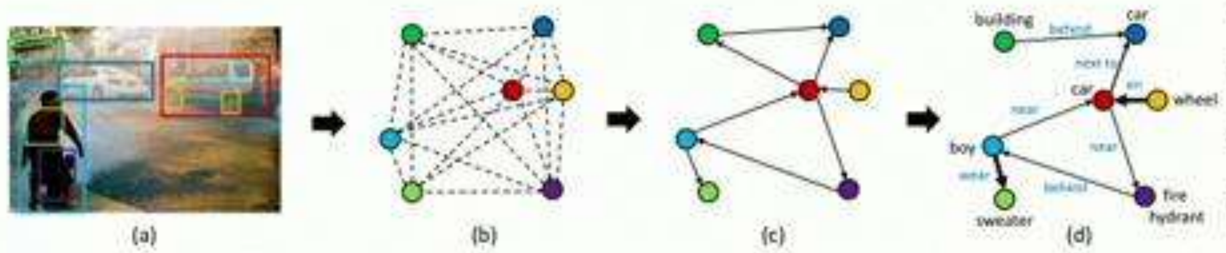
Joint Unsupervised Learning. CVPR 2016

My researches:

Cross-Image Structure

Joint Unsupervised Learning. CVPR 2016

Per-Image Structure



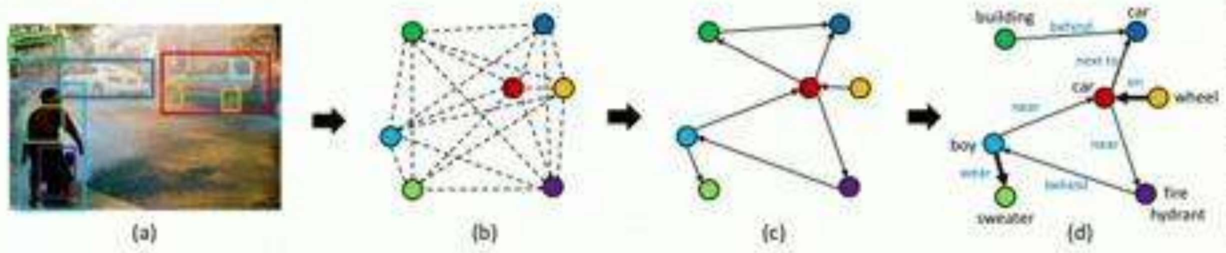
Graph R-CNN for Scene Graph Generation. ECCV 2018

My researches:

Cross-Image Structure

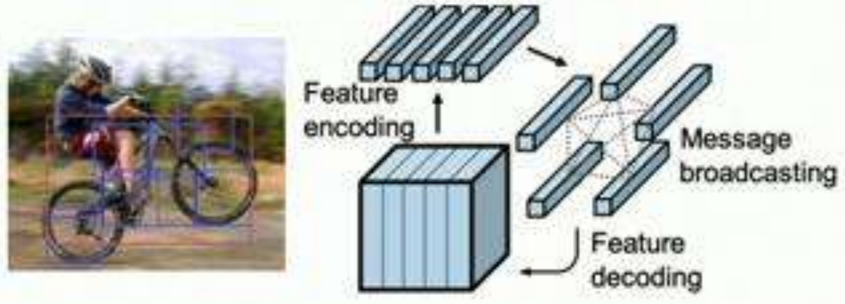
Joint Unsupervised Learning. CVPR 2016

Per-Image Structure



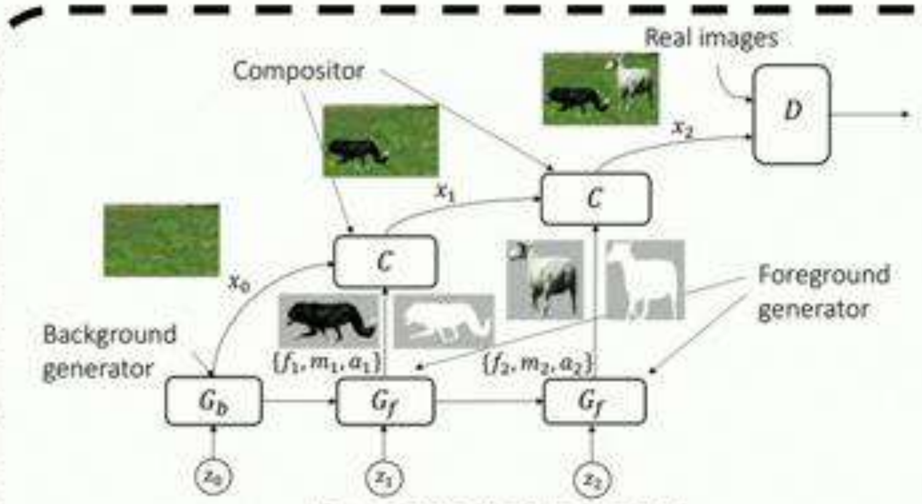
Graph R-CNN for Scene Graph Generation. ECCV 2018

Per-Object Structure



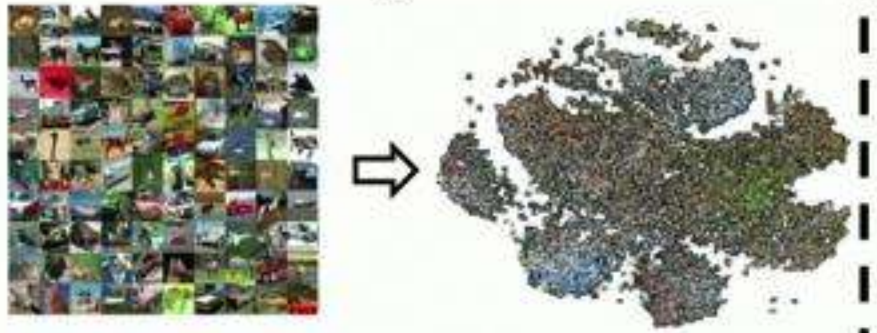
Neuron Communication Networks. NeurIPS 2019

My researches:



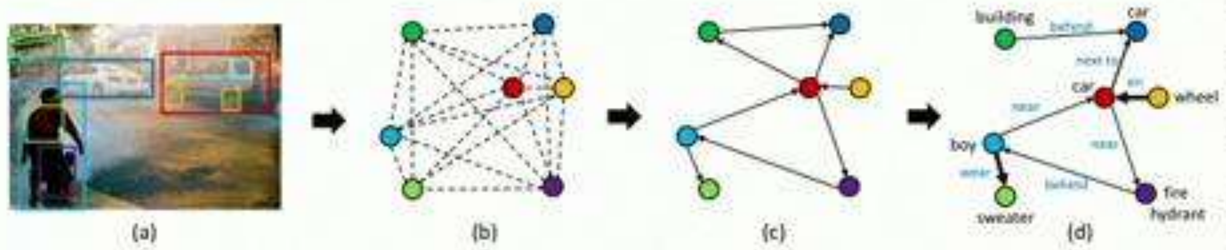
LR-GAN. ICLR 2017

Cross-Image Structure



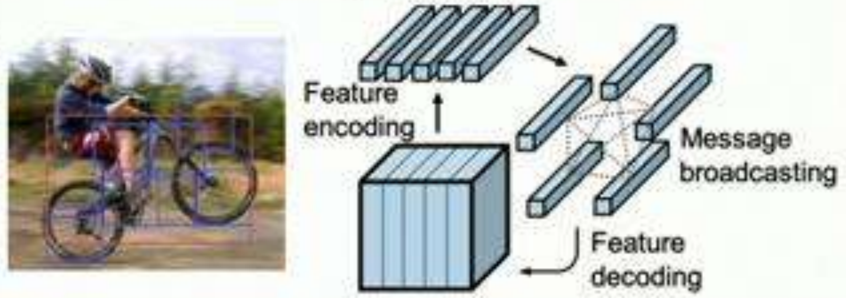
Joint Unsupervised Learning. CVPR 2016

Per-Image Structure



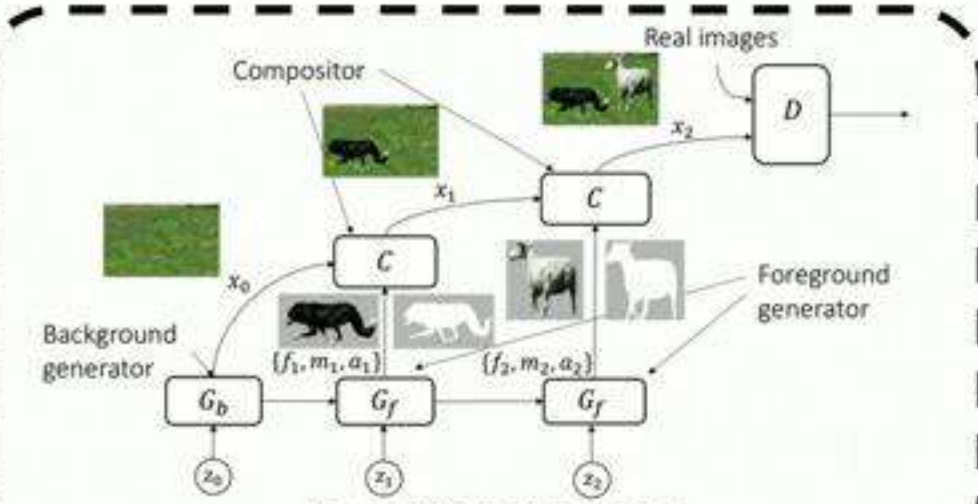
Graph R-CNN for Scene Graph Generation. ECCV 2018

Per-Object Structure



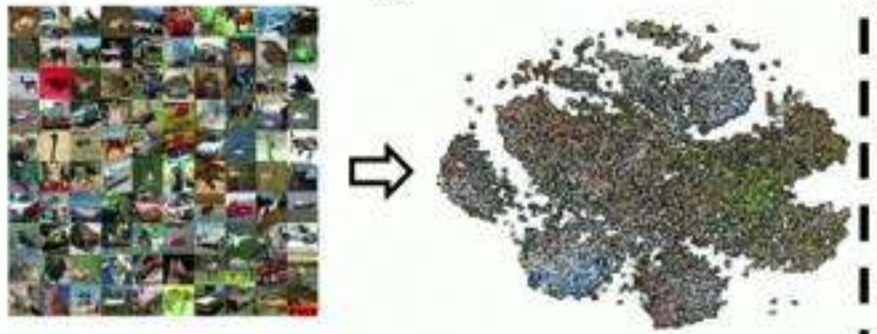
Neuron Communication Networks. NeurIPS 2019

My researches:



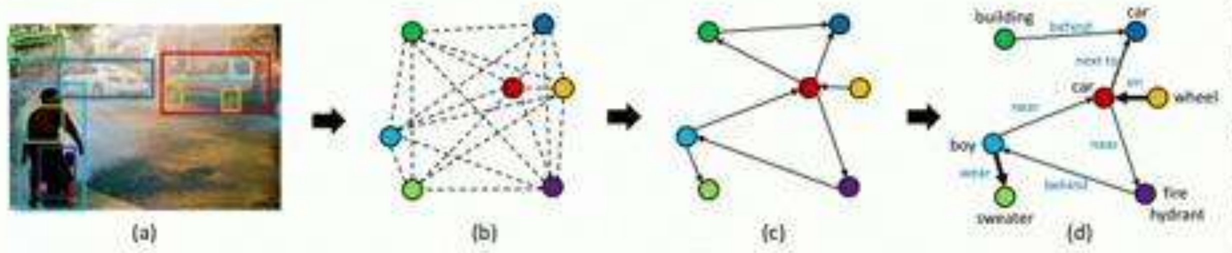
LR-GAN. ICLR 2017

Cross-Image Structure



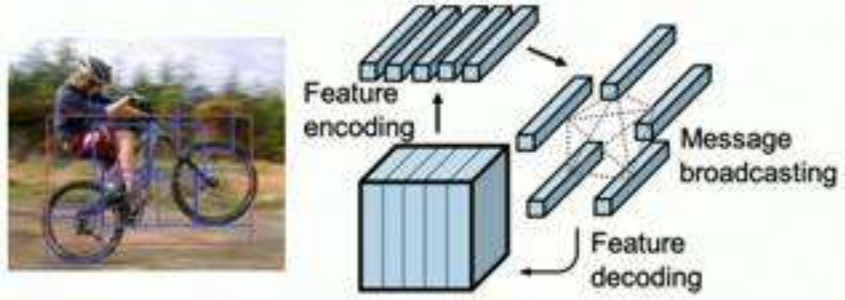
Joint Unsupervised Learning. CVPR 2016

Per-Image Structure

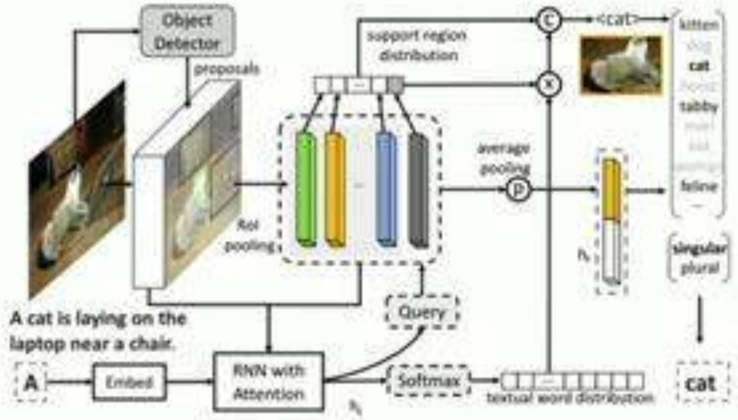


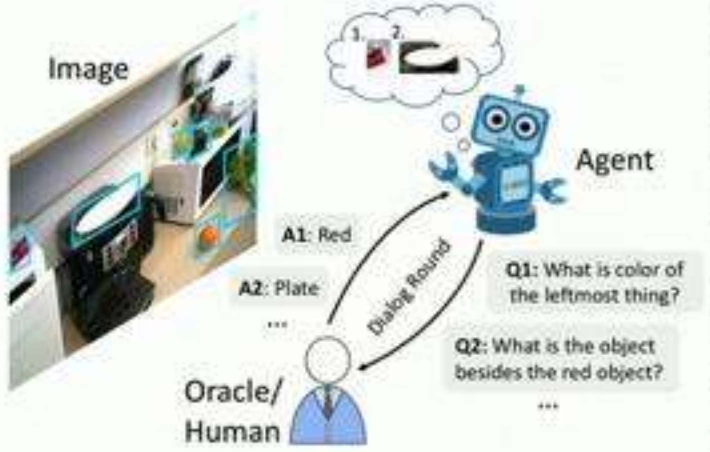
Graph R-CNN for Scene Graph Generation. ECCV 2018

Per-Object Structure



Neuron Communication Networks. NeurIPS 2019



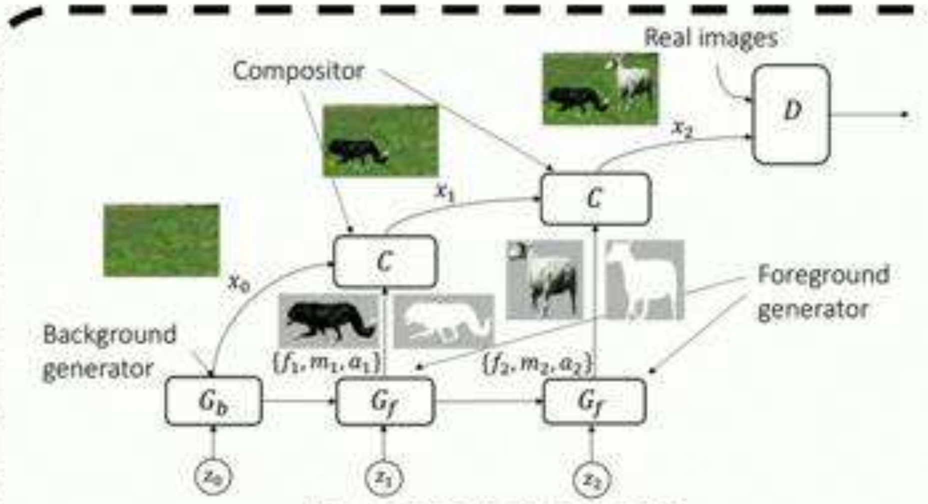


Neural Baby Talk. CVPR 2018

Visual Curiosity. CoRL 2018

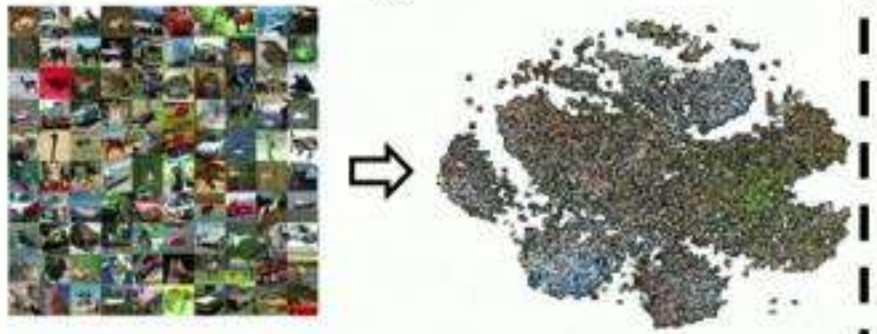
Language

My researches:



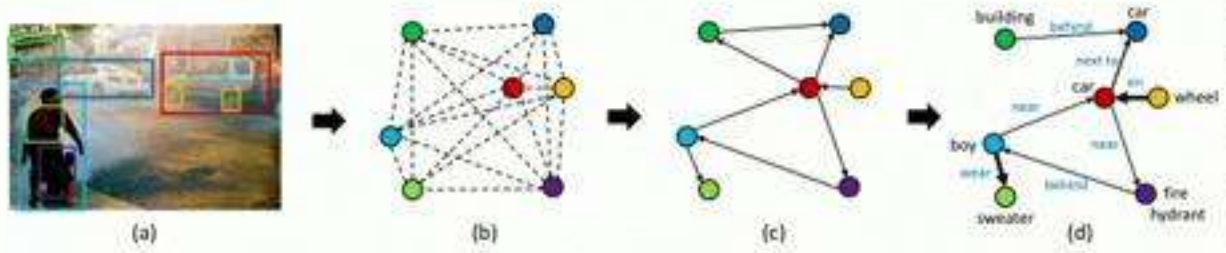
LR-GAN. ICLR 2017

Cross-Image Structure



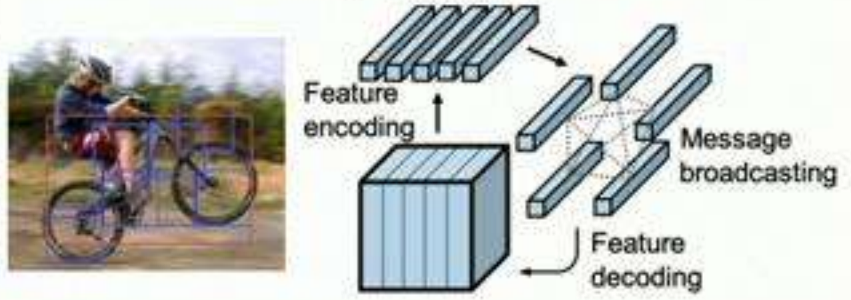
Joint Unsupervised Learning. CVPR 2016

Per-Image Structure

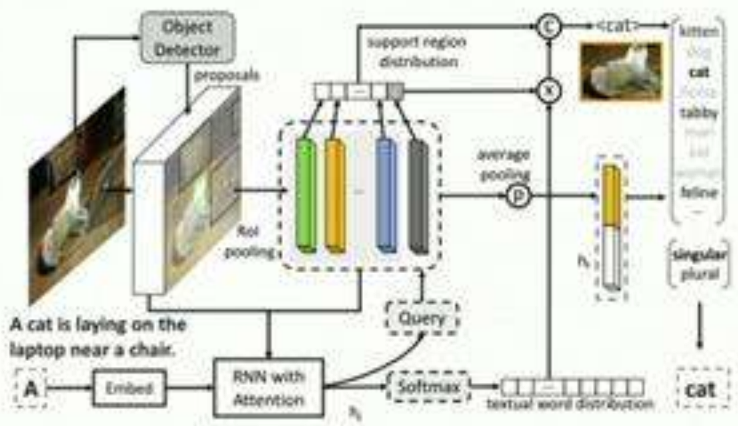


Graph R-CNN for Scene Graph Generation. ECCV 2018

Per-Object Structure



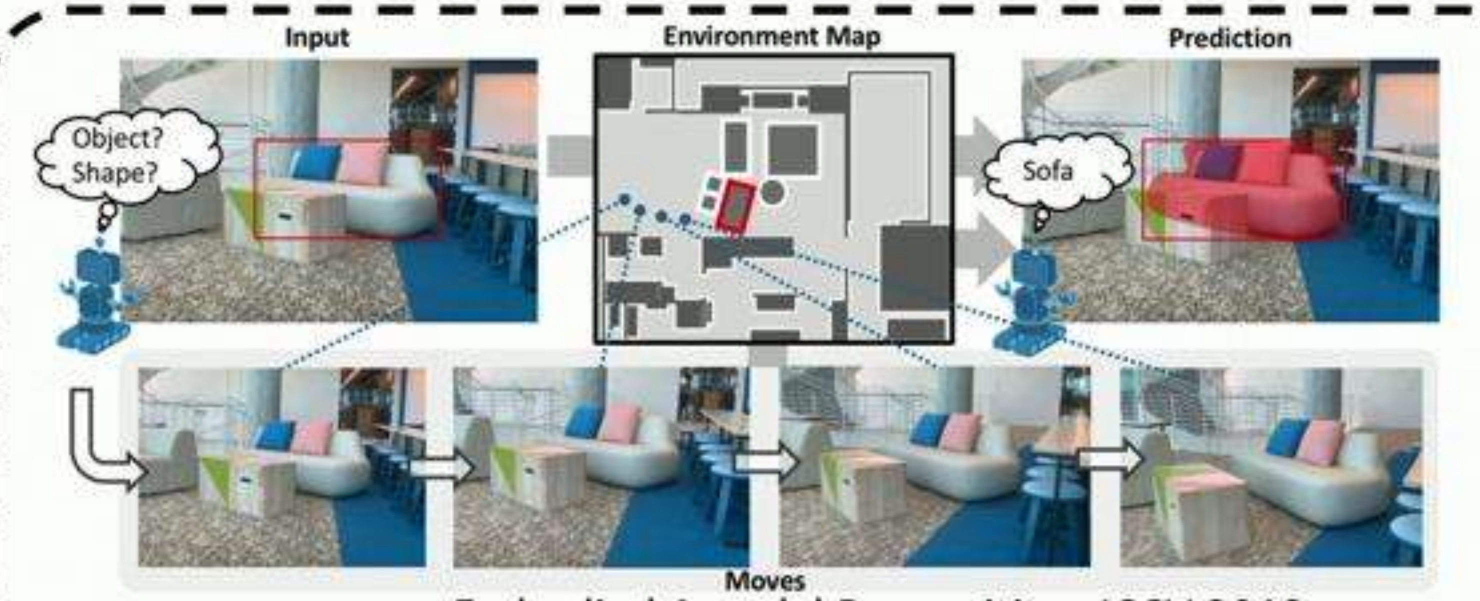
Neuron Communication Networks. NeurIPS 2019



Neural Baby Talk. CVPR 2018

Visual Curiosity. CoRL 2018

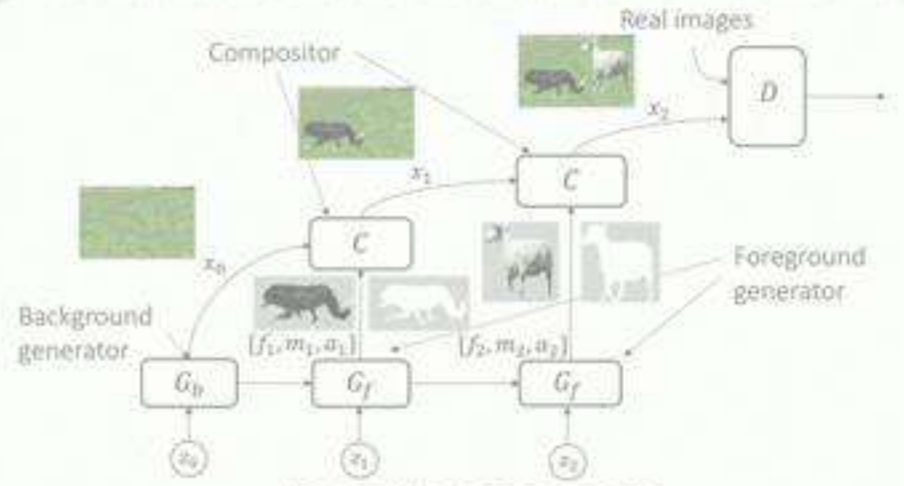
Language



Embodied Amodal Recognition. ICCV 2019

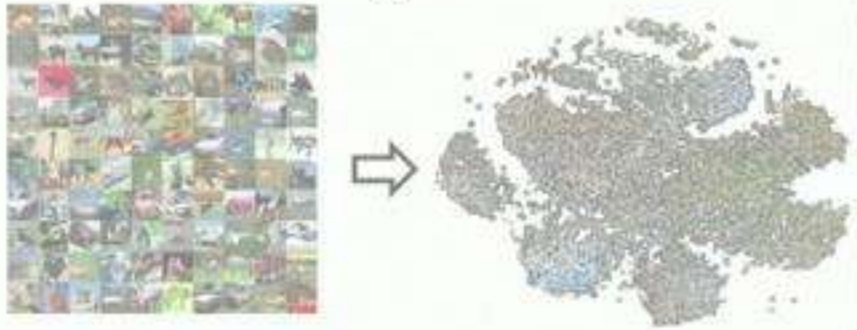
Embodiment

My researches:



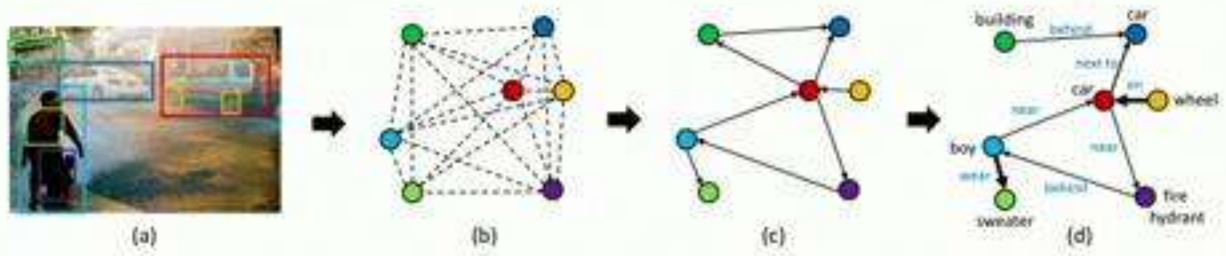
LR-GAN. ICLR 2017

Cross-Image Structure



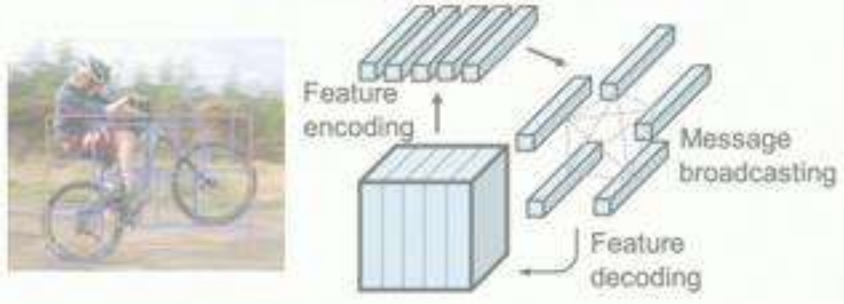
Joint Unsupervised Learning. CVPR 2016

Per-Image Structure

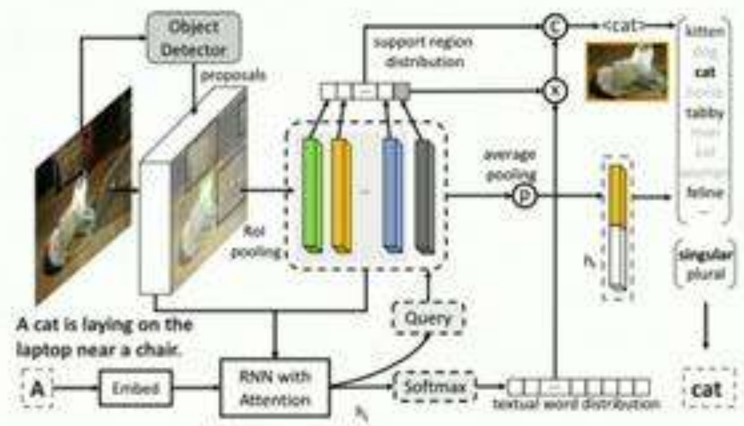


Graph R-CNN for Scene Graph Generation. ECCV 2018

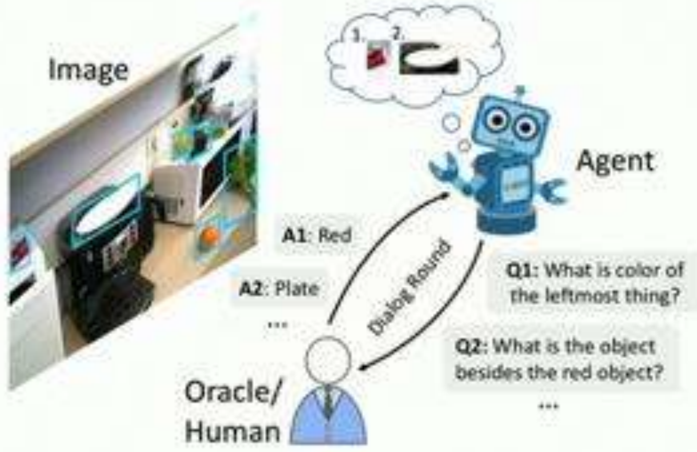
Per-Object Structure



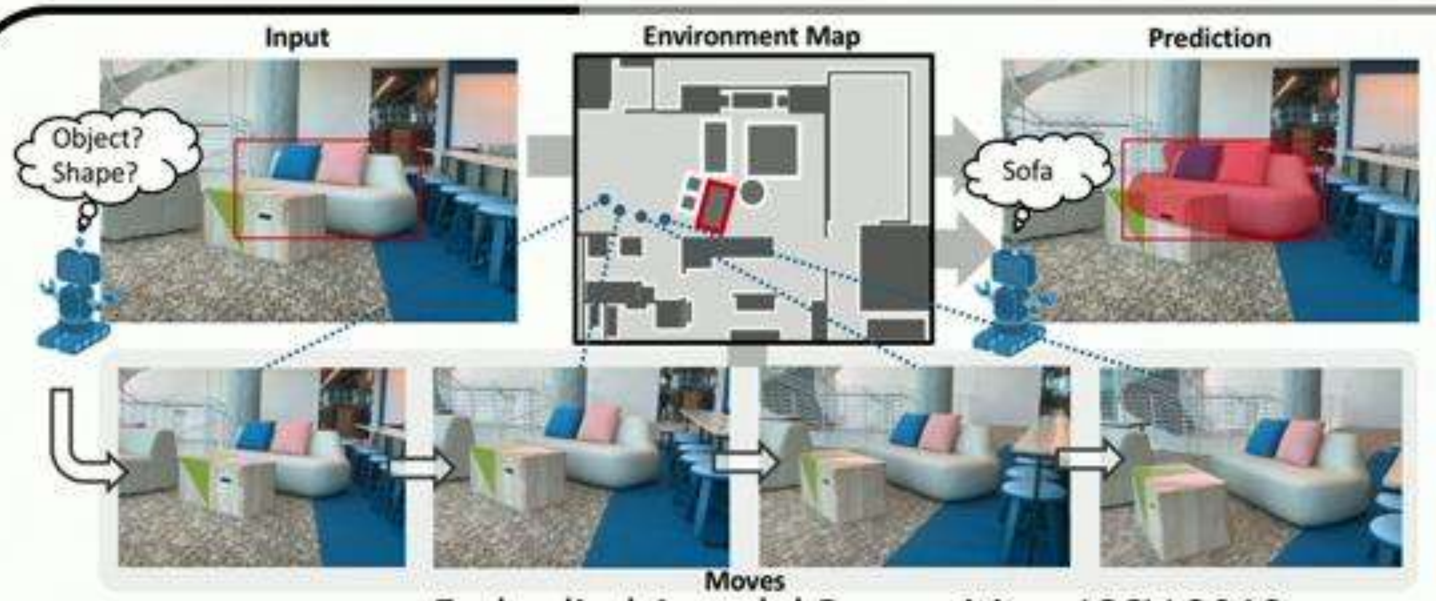
Neuron Communication Networks. NeurIPS 2019



Neural Baby Talk. CVPR 2018

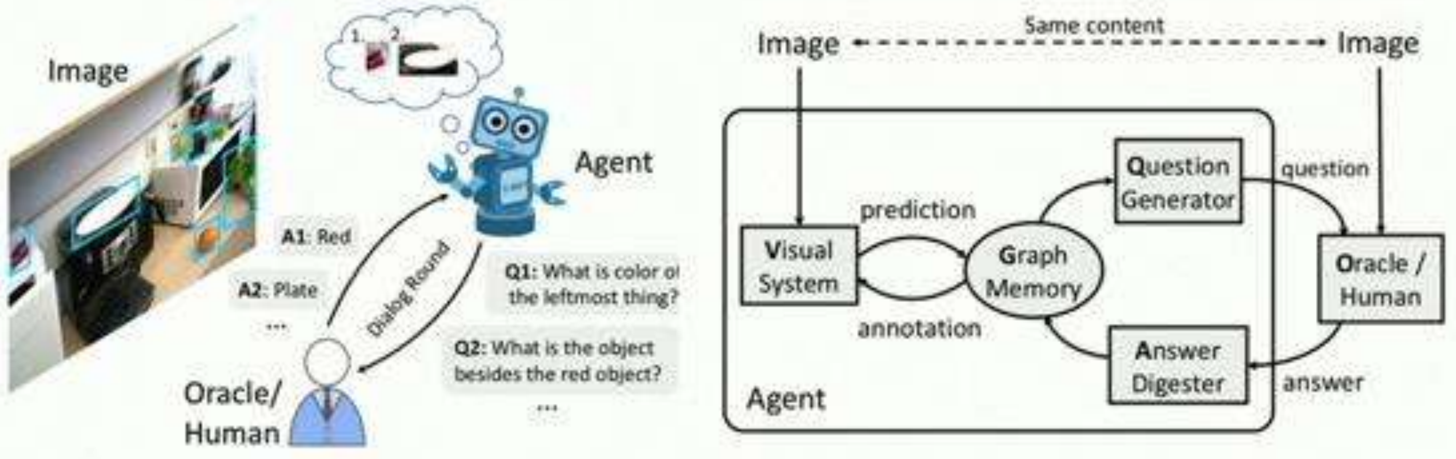


Visual Curiosity. CoRL 2018



Embodied Amodal Recognition. ICCV 2019

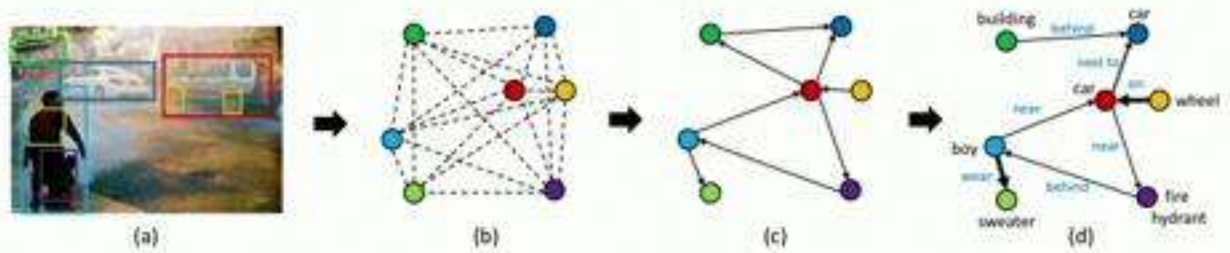
In this talk



Visual Curiosity. CoRL 2018

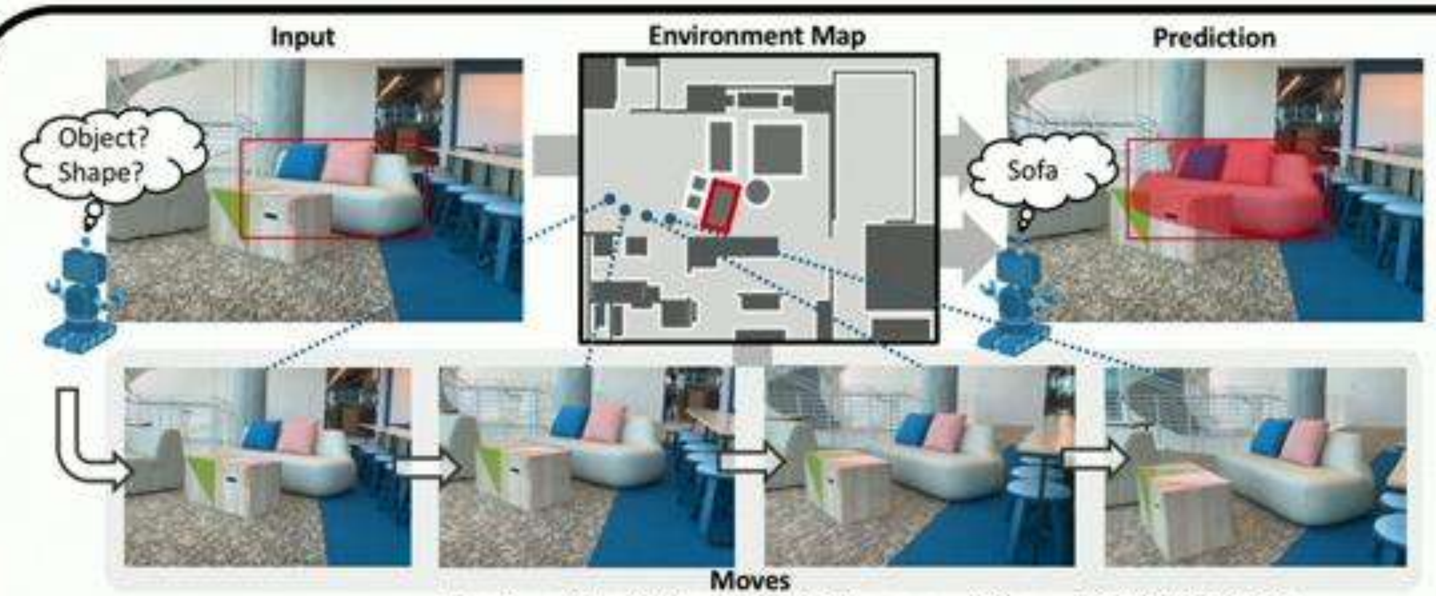
Interact with Human

Per-Image Structure



Graph R-CNN for Scene Graph Generation. ECCV 2018

Structured Visual Understanding



Embodied Amodal Recognition. ICCV 2019

Interact with Environment

Structured Visual Understanding

Graph R-CNN for Scene Graph Generation. ECCV 2018

What is scene graph?

Image as a single label

Image Source: ImageNet

Image as an object set

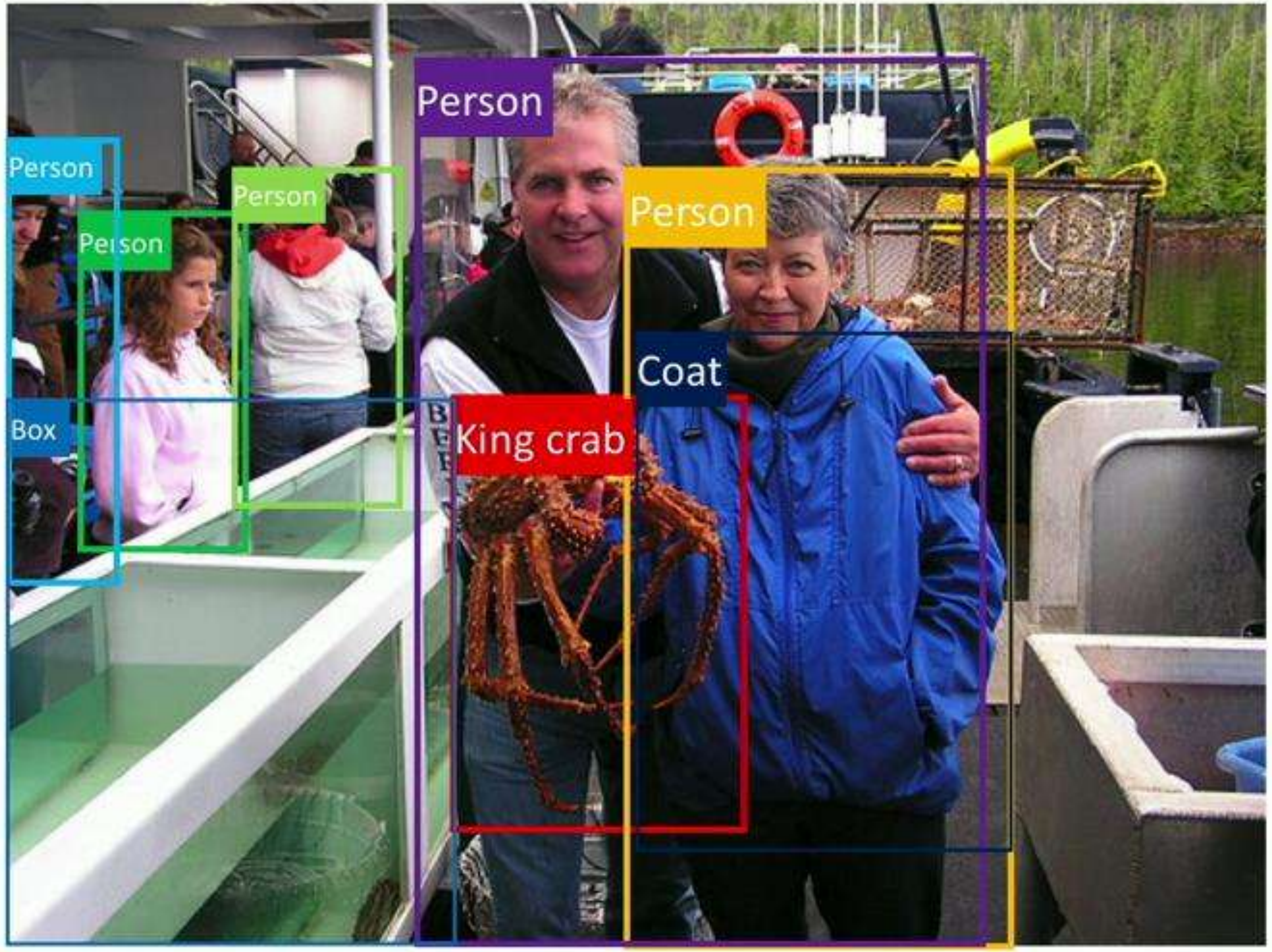
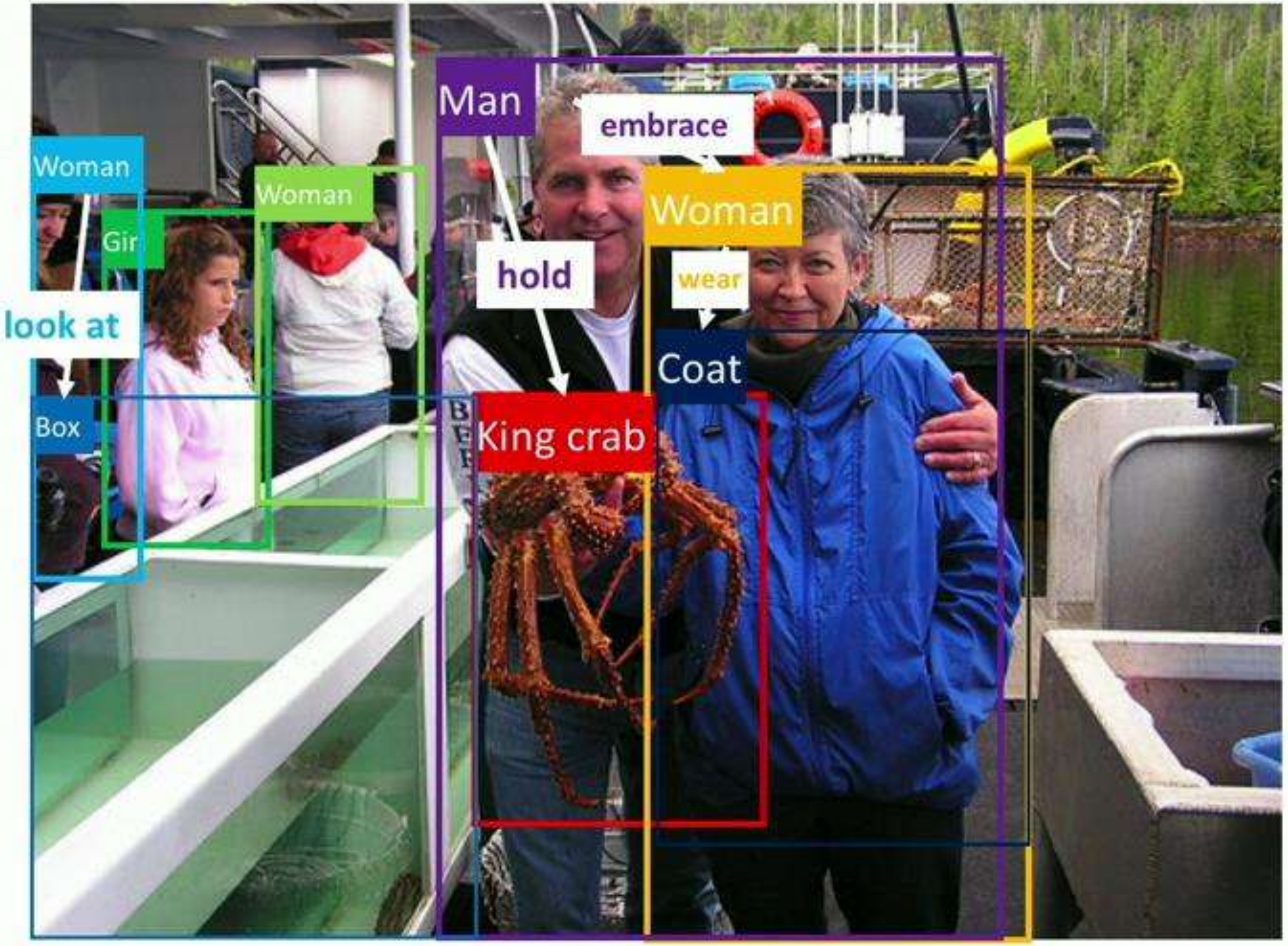


Image Source: ImageNet

Image as a scene graph

"Woman look at box"

"Man hold king crab"



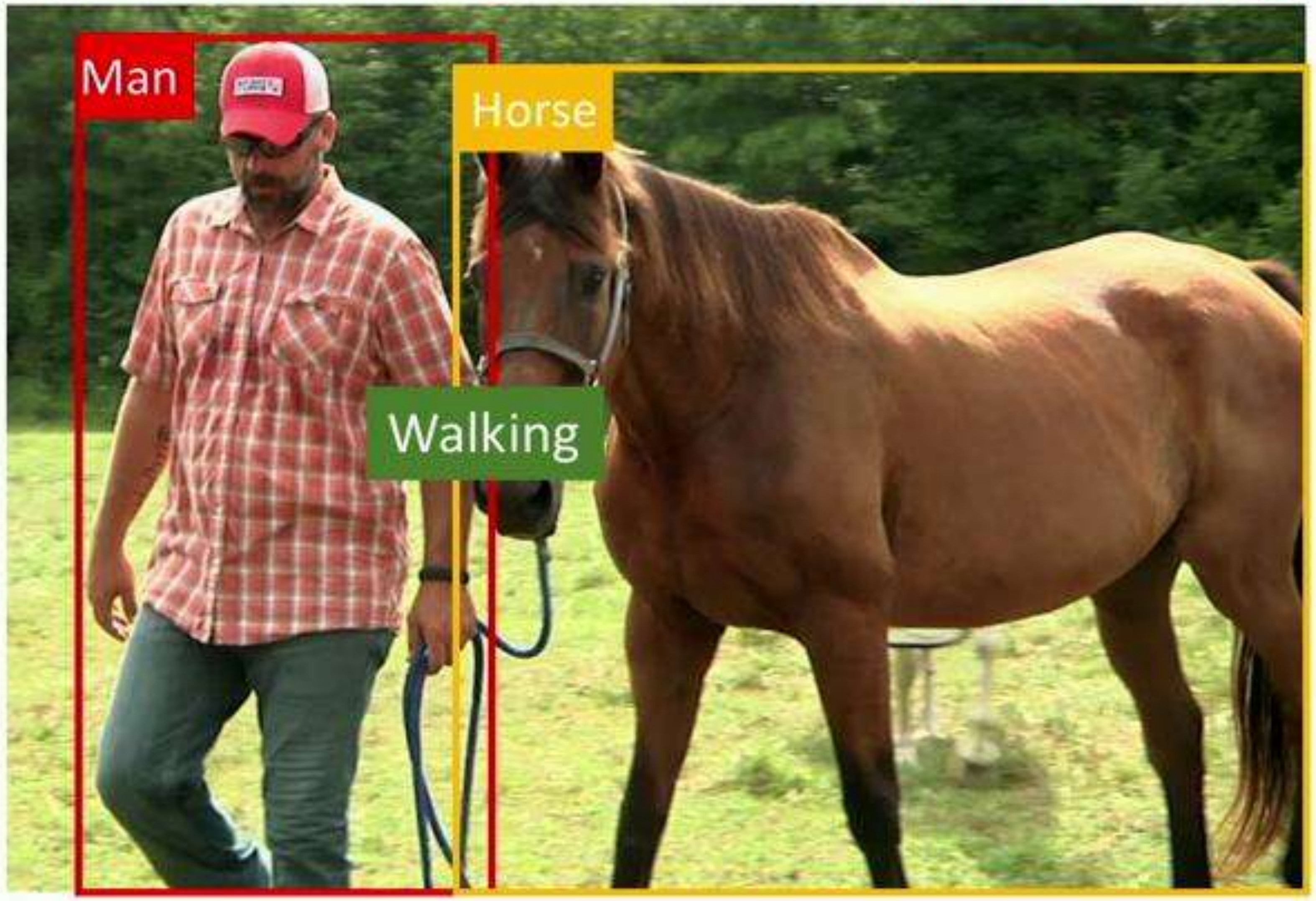
"Woman wear coat"

"Man embrace woman"

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?

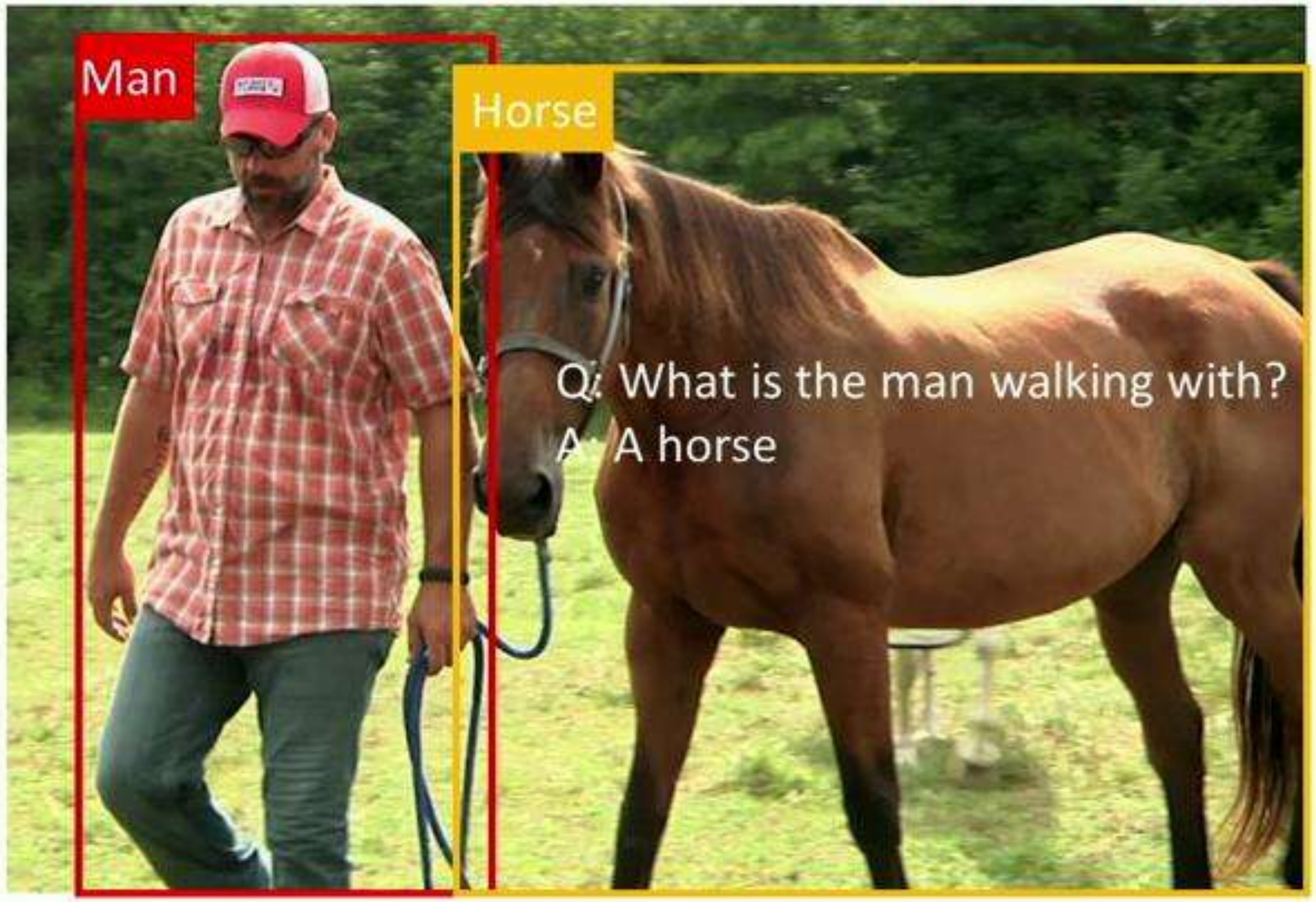
Describe images more grounding

[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

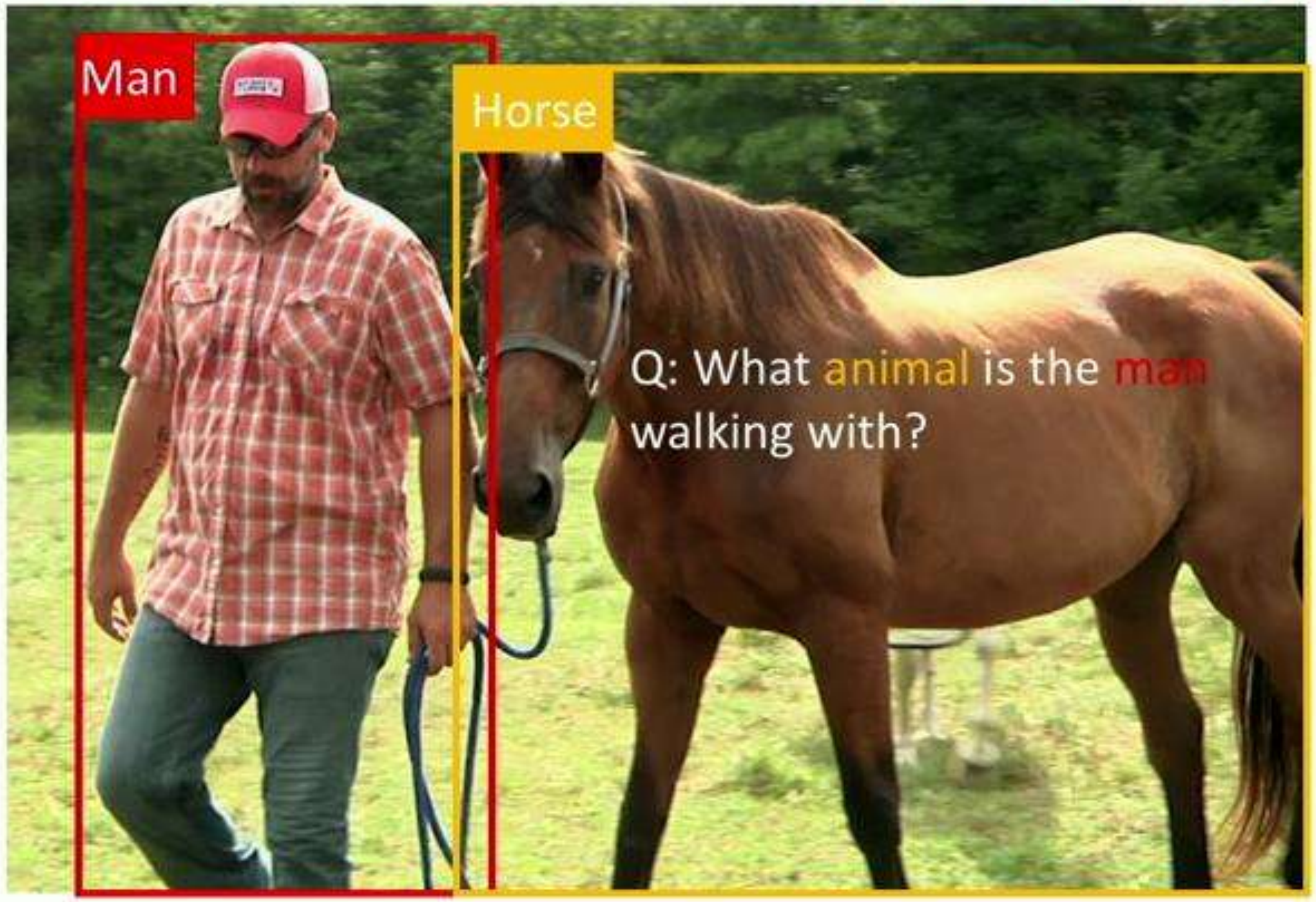


[1] Graph-Structured Representations for Visual Question Answering. Teney et al. CVPR 2017
[2] Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. Yi et al. Neurips 2018

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

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

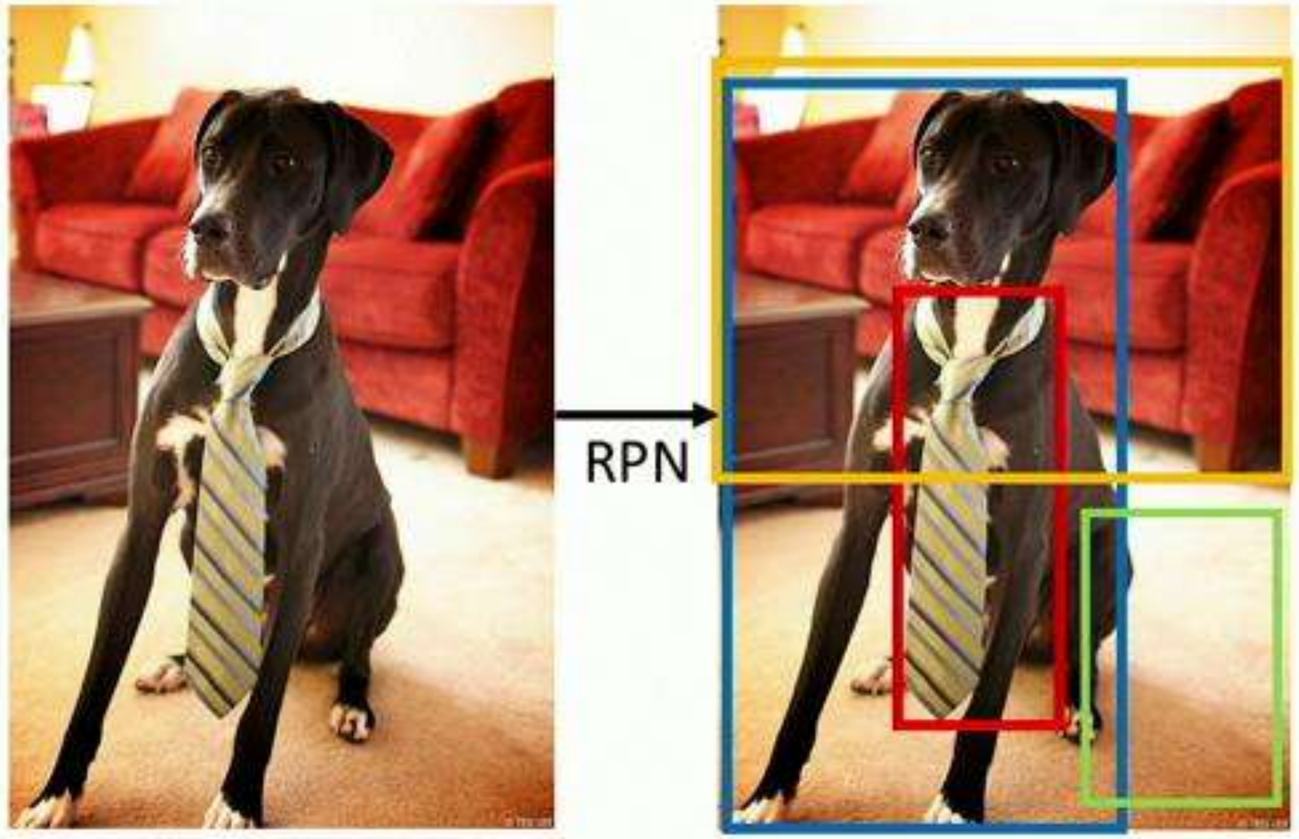
[1] Graph-Structured Representations for Visual Question Answering. Teney et al. CVPR 2017
[2] Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. Yi et al. Neurips 2018

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

General Pipeline

Input

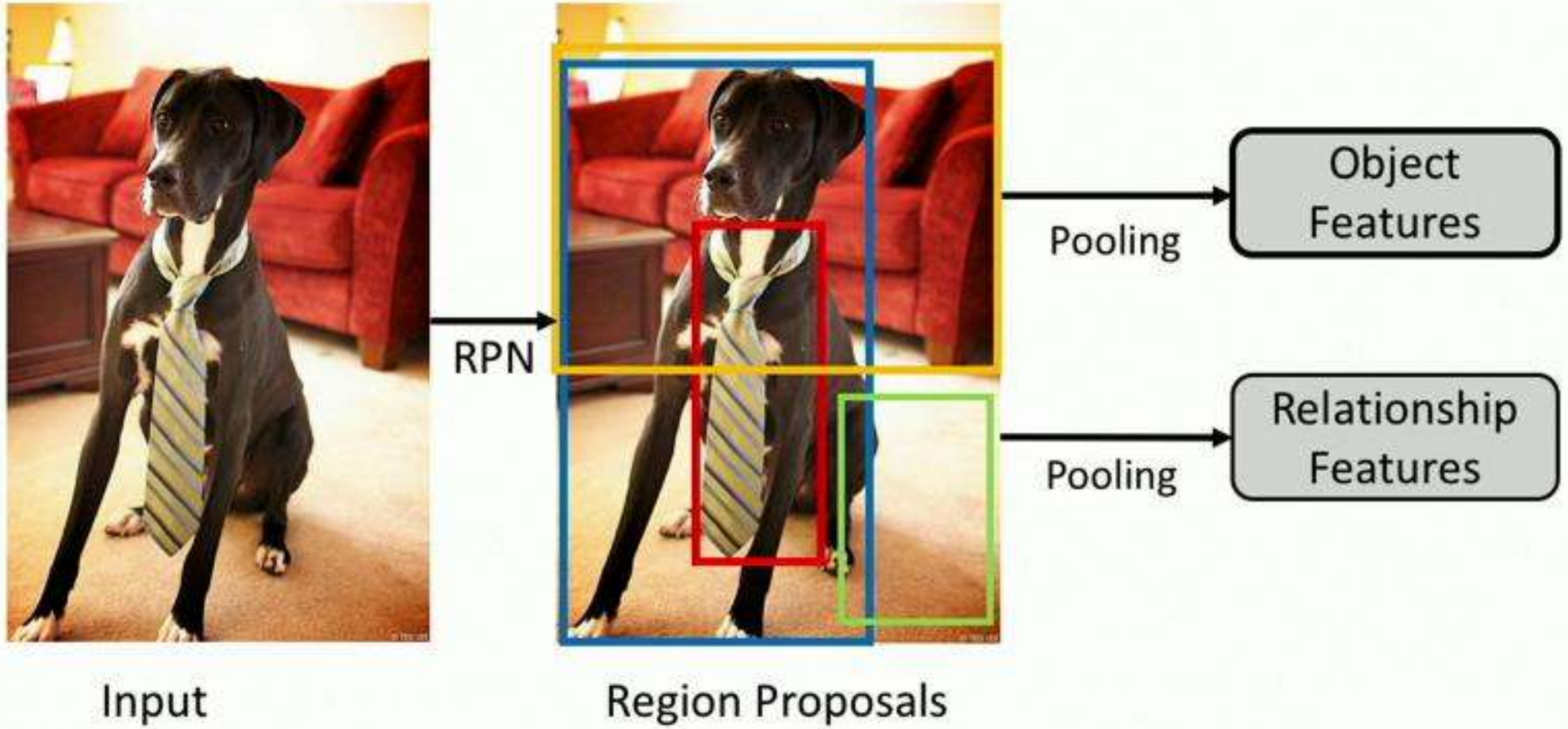
General Pipeline



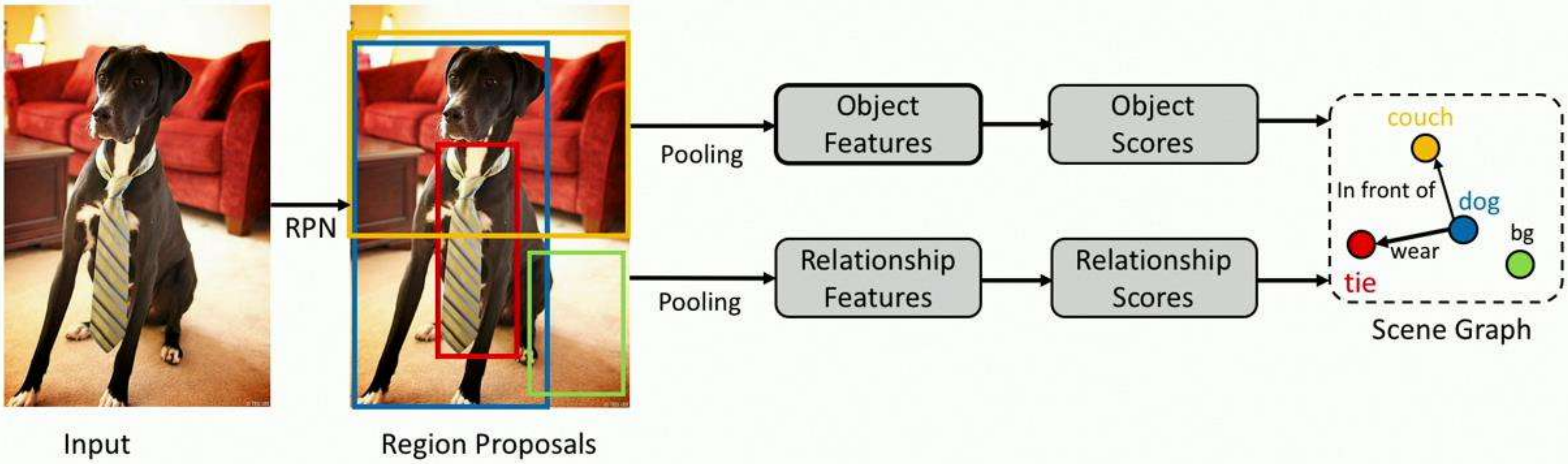
Input

Region Proposals

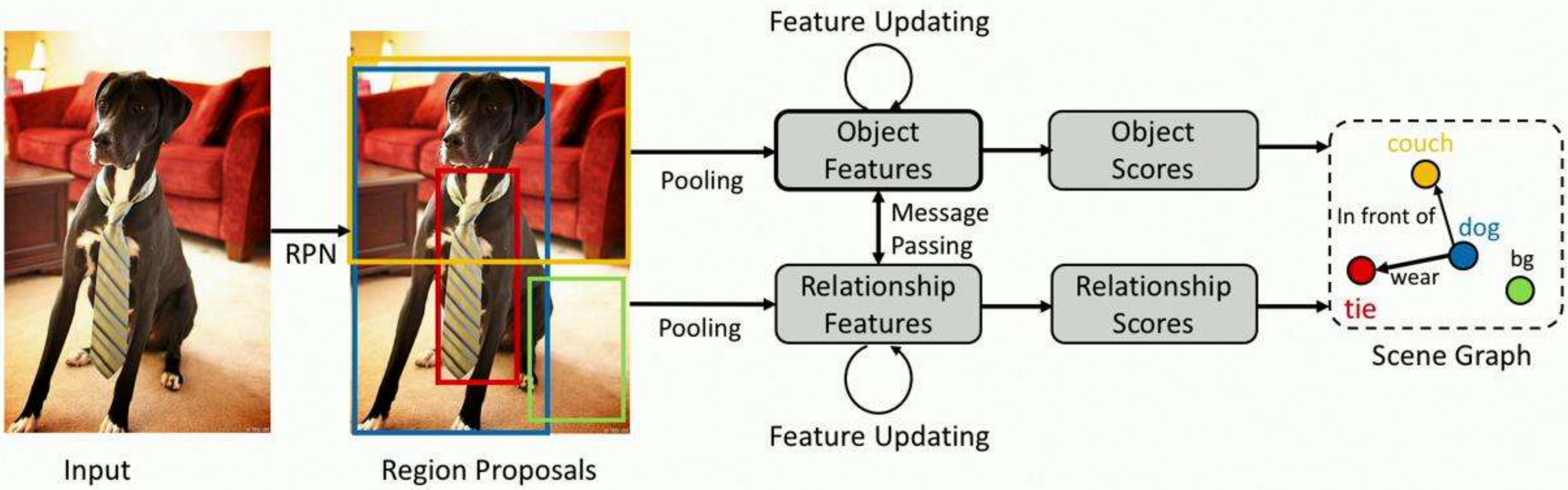
General Pipeline



General Pipeline

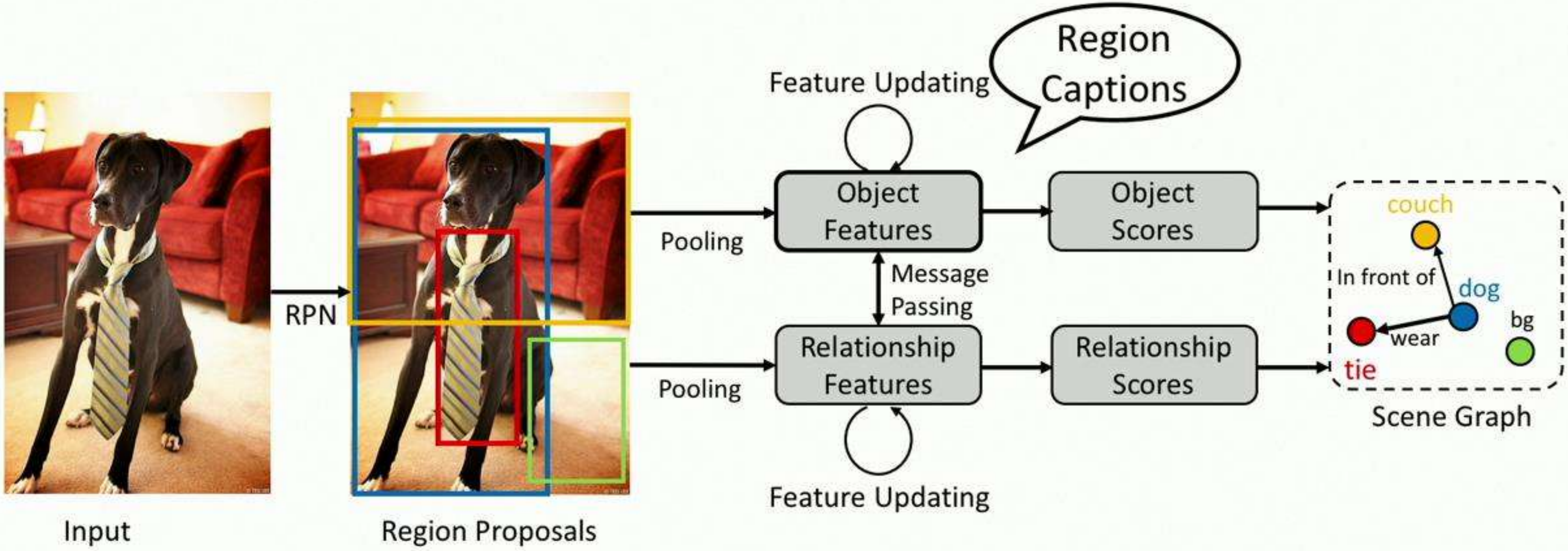


IMP Model



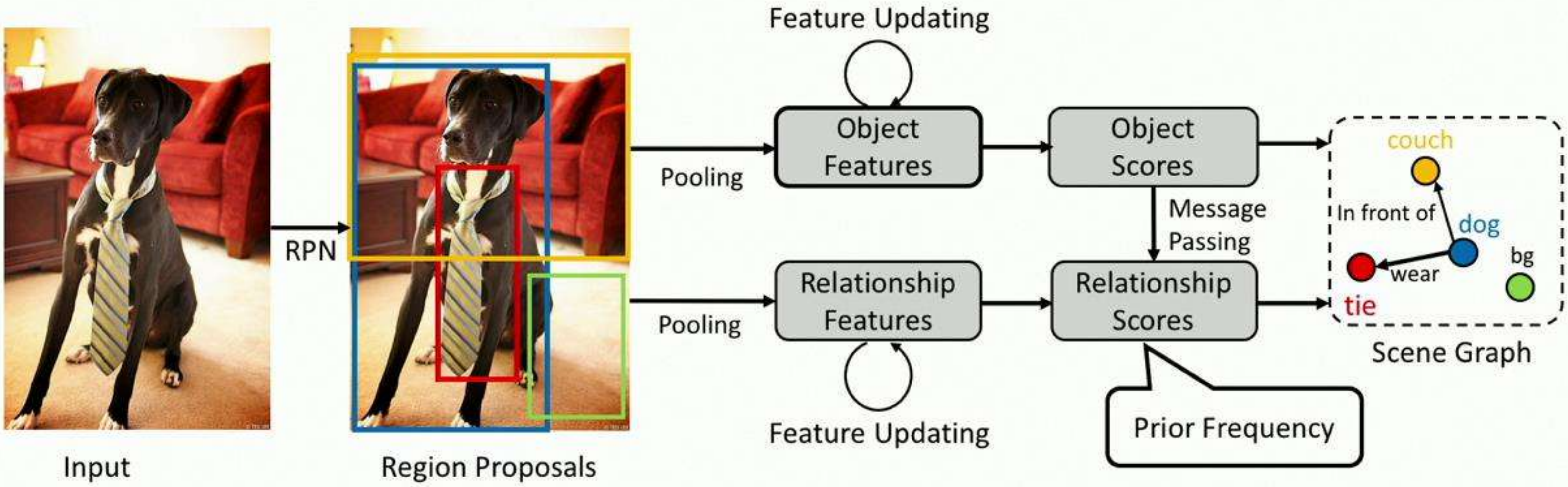
Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017

MSDN Model



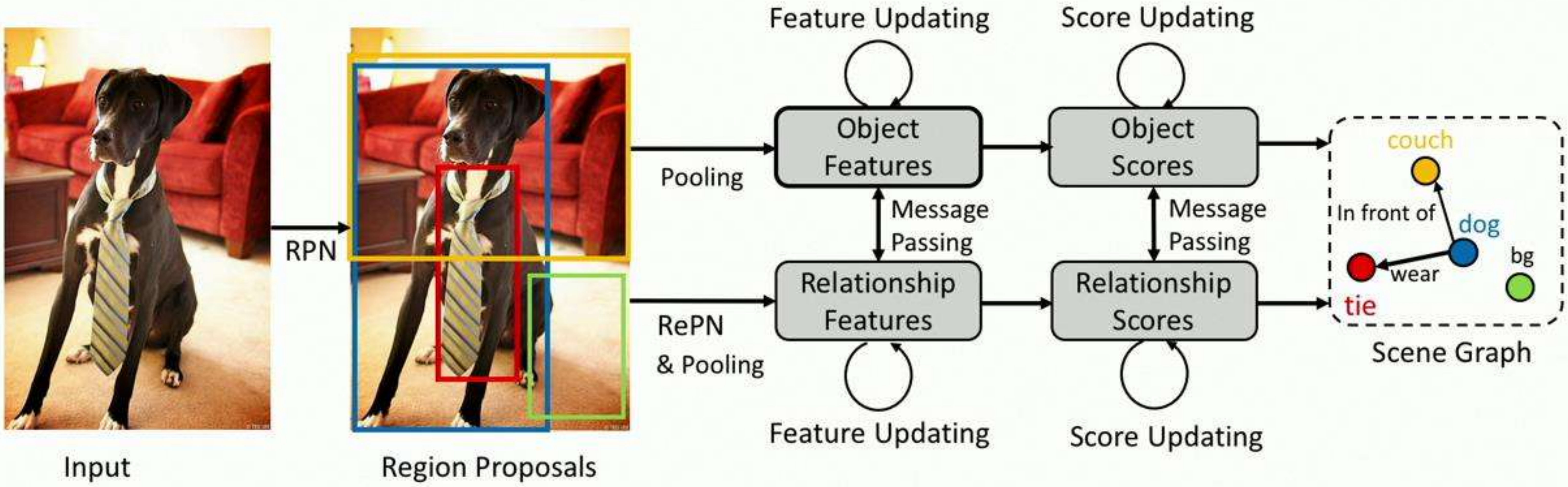
Scene Graph Generation from Objects, Phrases and Region Captions. Li et al. ICCV 2017

Neural Motif Network



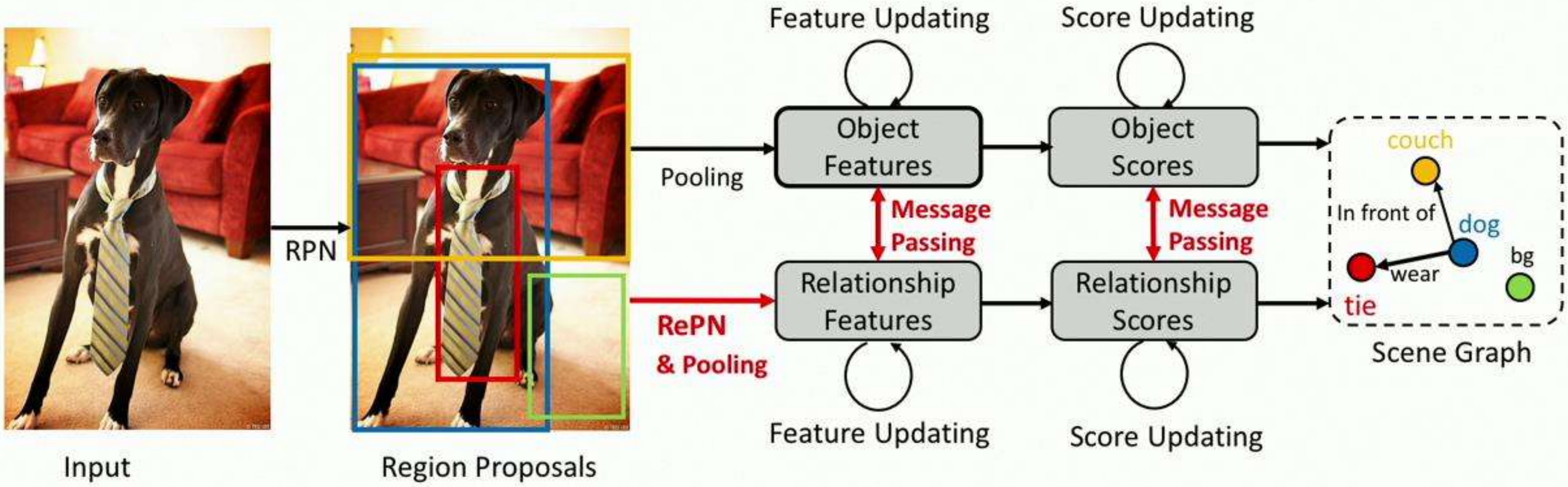
Neural Motifs: Scene Graph Parsing with Global Context. Zellers et al. CVPR 2018

Our model: Graph R-CNN

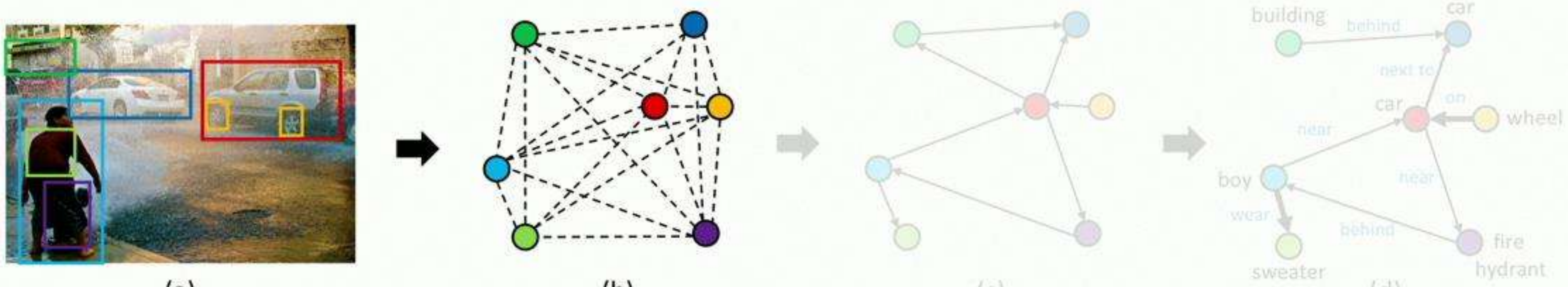


Jianwei Yang*, Jiasen Lu*, Stefan Lee, Dhruv Batra, Devi Parikh. Graph R-CNN for Scene Graph Generation. ECCV 2018.

Our model: Graph R-CNN



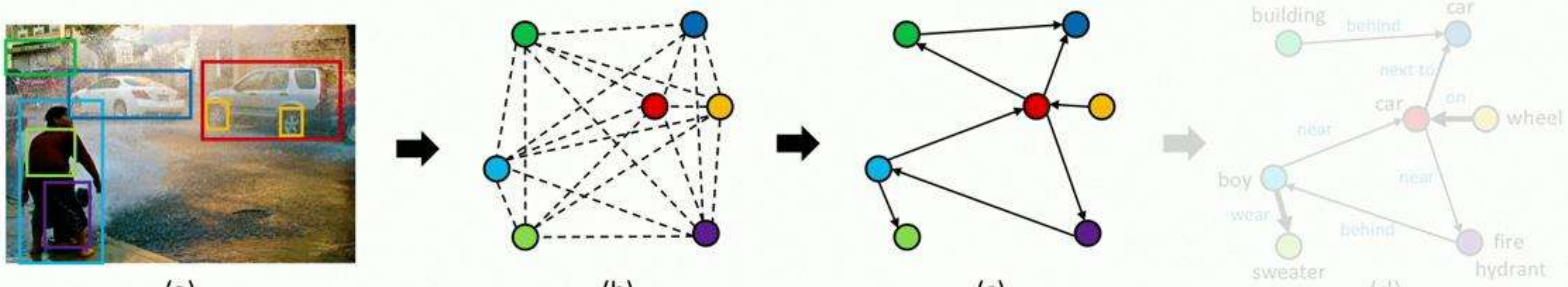
Motivations



(a) (b) (c) (d)

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

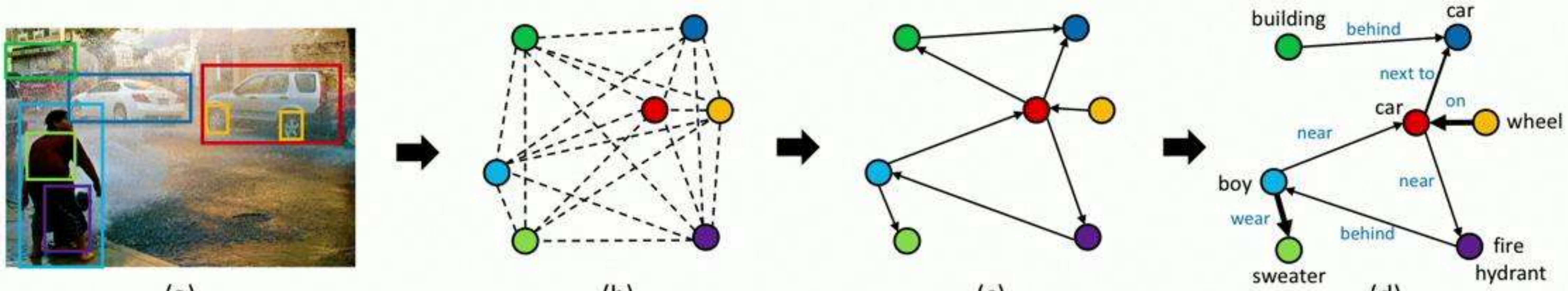
Motivations



(a) (b) (c) (d)

1. Objects in a scene usually have relationships with others;
2. Not all object pairs have relationships, the scene graph is usually sparse,

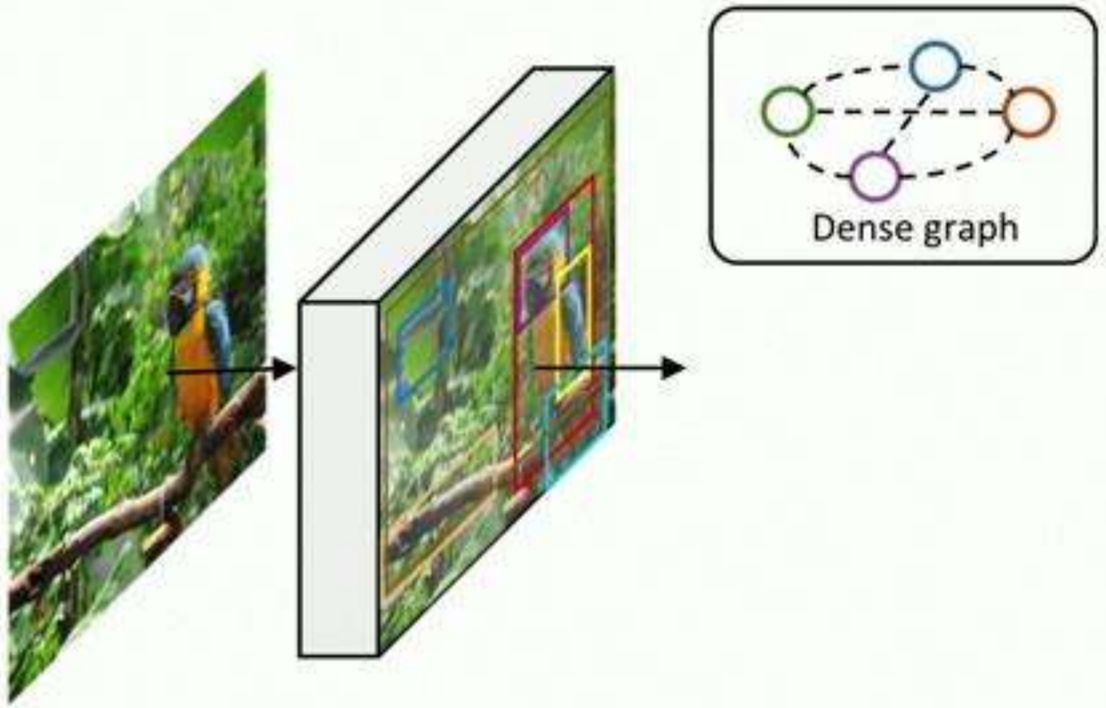
Motivations



(a) (b) (c) (d)

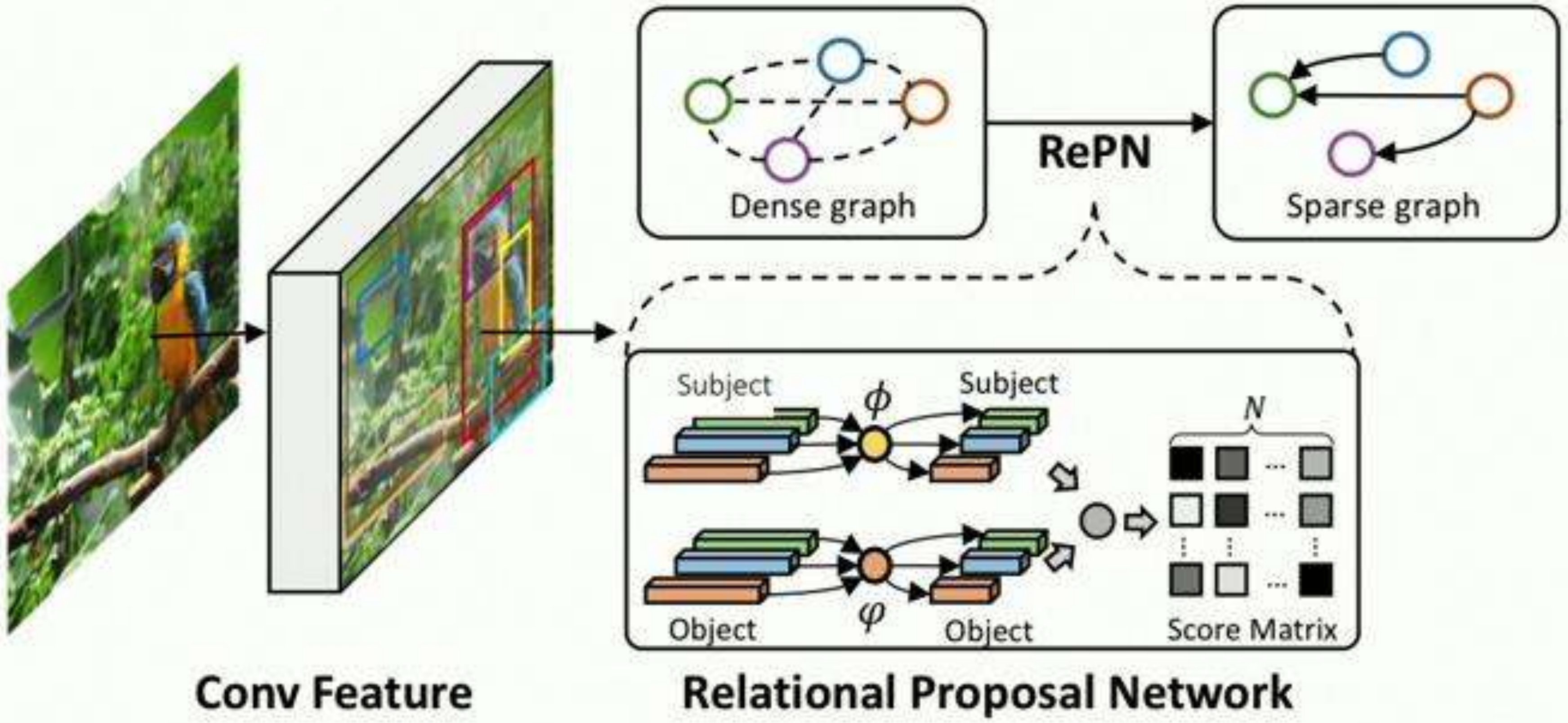
1. Objects in a scene usually have relationships with others;
2. Not all object pairs have relationships, the scene graph is usually sparse,
3. Existence of relationships highly depends on the object categories, and type of relationships highly depends on the context.

Framework



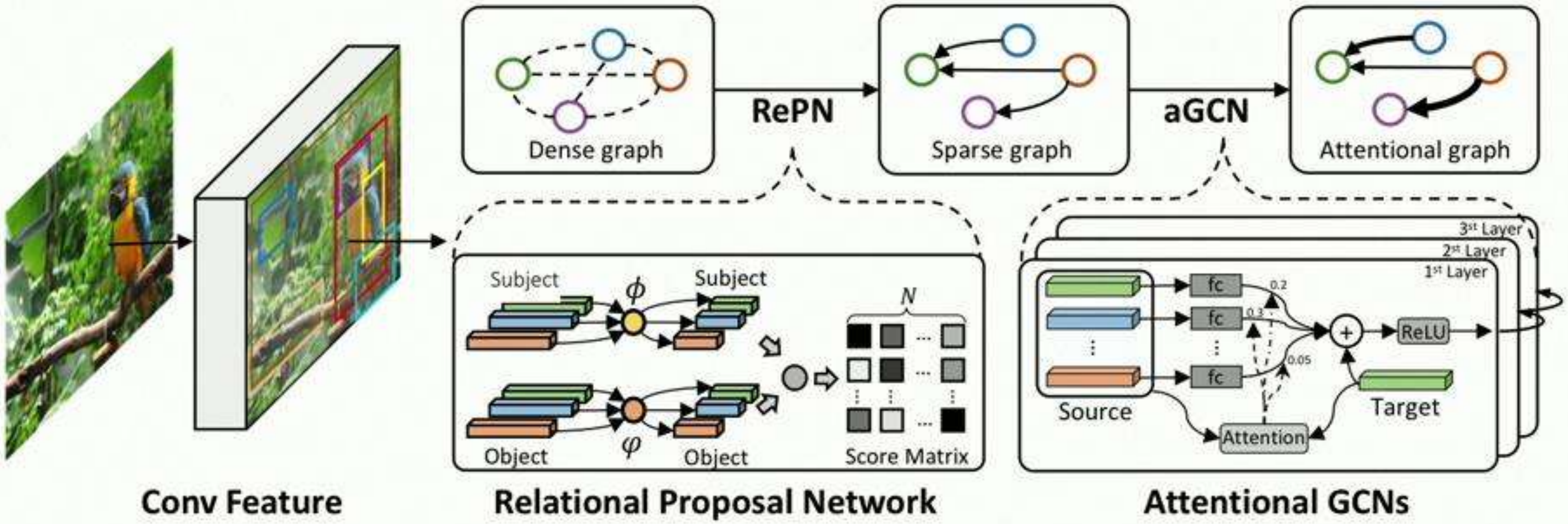
Conv Feature

Framework



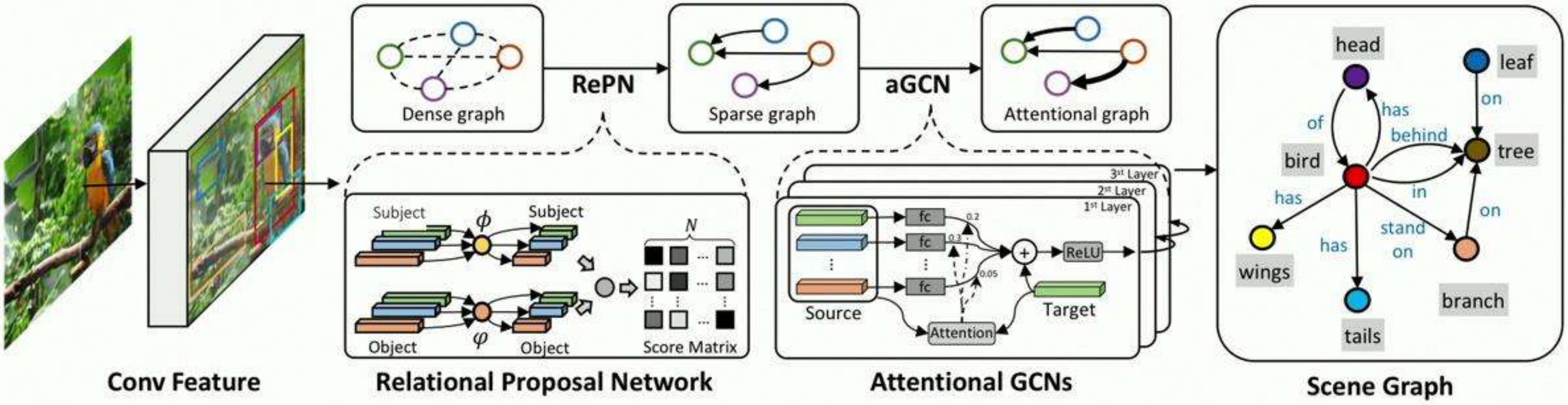
1. Relation proposal network (RePN) to learn to prune the densely connected scene graph;

Framework



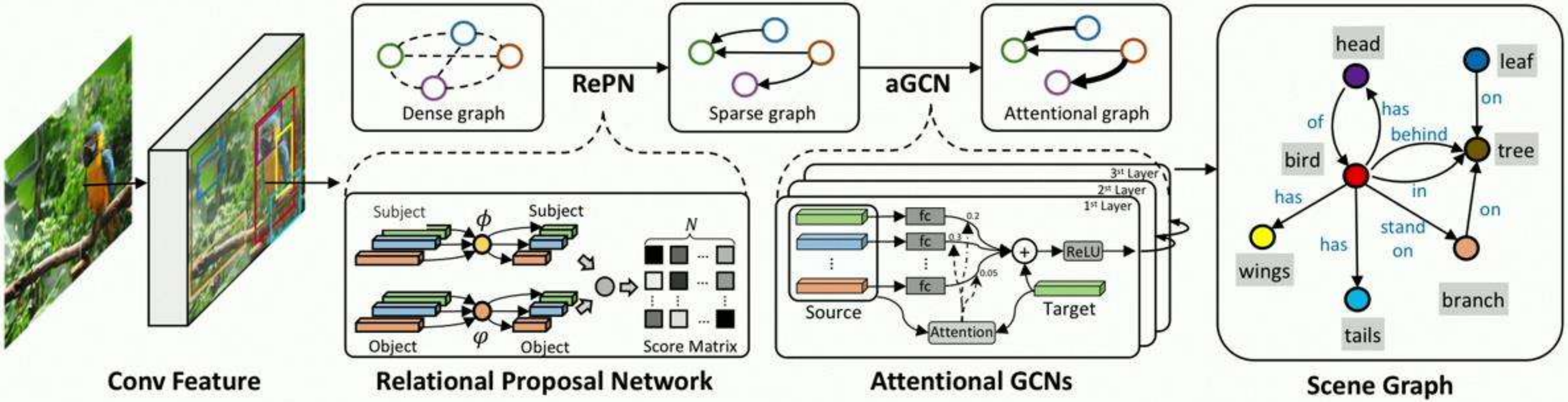
1. Relation proposal network (RePN) to learn to prune the densely connected scene graph;
2. Attentional graph convolutional networks (aGCN) to incorporate the contextual information.

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

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

Relation Proposal Network

Inspired by Region Proposal Network[1]:

Step 1: Compute Relationship-ness between subject and object:

Subj. and obj. rep.　　Kernel functions[2]

$$R(m,n) = f([x_m^o, x_n^o]) =< \phi(x_m^o), \varphi(x_n^o) >$$

Here, we use object prediction scores as the representation.

$$R(p,q) = f\left(\left[x_p^o, x_q^o\right]\right) =< \phi\left(x_p^o\right), \varphi(x_q^o) >$$

Step 2: NMS for object pairs based on pair-wise IoU:

$$IoU\left(\{r_m^o, r_n^o\}, \{r_p^o, r_q^o\}\right) = \frac{I\left(r_m^o, r_p^o\right) + I(r_n^o, r_q^o)}{U\left(r_m^o, r_p^o\right) + U(r_n^o, r_q^o)}$$

[1]. Faster R-CNN. Ren et al. Neurips 2016.
[2]. Non-local Networks. Want et al. CVPR 2018.

Attentional GCN

GCN layer with residual connection[1]:

$$z_i^{(l+1)} = \sigma\left(z_i^{(l)} + \sum_{i \in \mathcal{N}(i)} \alpha_{ij} W z_j^{(l)}\right)$$

$$\downarrow$$

Matrix Computation $z_i^{(l+1)} = \sigma\left(W Z^{(l)} \alpha_i\right)$

Nonlinear function Learnable parameters Inputs from last layer

[1]. Semi-Supervised Classification with Graph Convolutional Networks. Kipf et al. ICLR 2017

Attentional GCN

GCN layer with residual connection[1]:

$$z_i^{(l+1)} = \sigma\left(z_i^{(l)} + \sum_{i \in \mathcal{N}(i)} \alpha_{ij} W z_j^{(l)}\right)$$

Matrix Computation $z_i^{(l+1)} = \sigma\left(W Z^{(l)} \boxed{\alpha_i}\right)$ ← Predetermined Affinities

Nonlinear function Learnable parameters Inputs from last layer

[1]. Semi-Supervised Classification with Graph Convolutional Networks. Kipf et al. ICLR 2017

Attentional GCN

GCN layer with residual connection[1]:

$$z_i^{(l+1)} = \sigma\left(z_i^{(l)} + \sum_{i \in \mathcal{N}(i)} \alpha_{ij} W z_j^{(l)}\right)$$

Matrix Computation $\quad z_i^{(l+1)} = \sigma\left(W Z^{(l)} \alpha_i\right)$

Nonlinear function Learnable parameters Inputs from last layer

Learning the affinities!

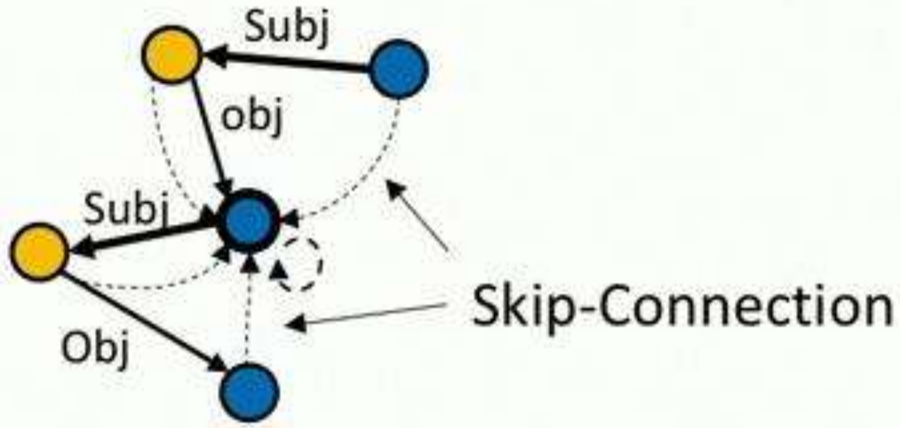
$$u_{ij} = w_h^T \sigma\left(W_a \left[z_i^{(l)}, z_j^{(l)}\right]\right)$$

$$\alpha_i = \mathrm{softmax}(u_i)$$

Attentional GCNs (aGCN) on scene graph:

Update object representations:

$$z_i^o = \sigma\left(W^{\text{skip}} Z^o \alpha^{rs} + W^{sr} Z^r \alpha^{sr} + W^{or} Z^r \alpha^{or}\right)$$



[1]. Semi-Supervised Classification with Graph Convolutional Networks. Kipf et al. ICLR 2017

[2]. Graph Attention Networks. Veličković et al. ICLR 2018

Attentional GCN

GCN layer with residual connection:

$$z_i^{(l+1)} = \sigma\left(z_i^{(l)} + \sum_{i \in \mathcal{N}(i)} \alpha_{ij} W z_j^{(l)}\right)$$

Matrix Computation $$z_i^{(l+1)} = \sigma\left(W Z^{(l)} \alpha_i\right)$$

Nonlinear function   Learnable parameters   Inputs from last layer

Learning the affinities!

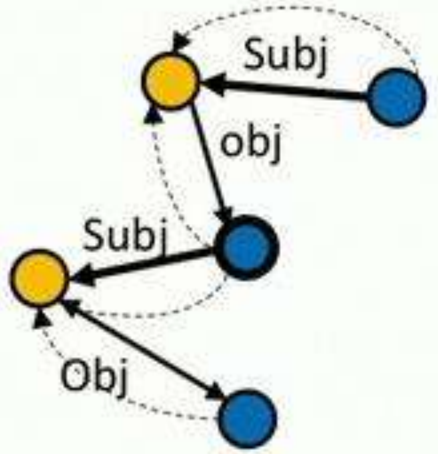
$$u_{ij} = w_h^T \sigma\left(W_a \left[z_i^{(l)}, z_j^{(l)}\right]\right)$$

$$\alpha_i = \mathrm{softmax}(u_i)$$

Attentional GCNs (aGCN) on scene graph:

Update predicate representations:

$$z_i^r = \sigma(z_i^r + W^{rs} Z^o \alpha^{rs} + W^{ro} Z^o \alpha^{ro})$$



[1]. Semi-Supervised Classification with Graph Convolutional Networks. Kipf et al. ICLR 2017
[2]. Graph Attention Networks. Veličković et al. ICLR 2018

Training

Binary Cross Entropy Loss

Region Proposal Network

Two Cross Entropy Losses,
one for node and one for edge

Graph Labeling Network

$$\underbrace{P(V|I)}_{\text{Region Proposal Network}} \quad P(E|V,I) \quad P(R,O|V,E,I) = P(S|I)$$

Relation Proposal Network

Binary Cross Entropy Loss

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: $Recall = \dfrac{C(T_{pred}\ and\ T_{gt})}{N(T_{gt})}$ 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 Train: 75,651 Test: 32,422	VGG-16 Faster R-CNN[1]	150	50	PredCls,SGCls, SGGen, mAP

[1] A Faster Implementation of Faster R-CNN. Yang and Lu et al.

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: $Recall = \frac{C(T_{pred} \text{ and } T_{gt})}{N(T_{gt})}$ 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

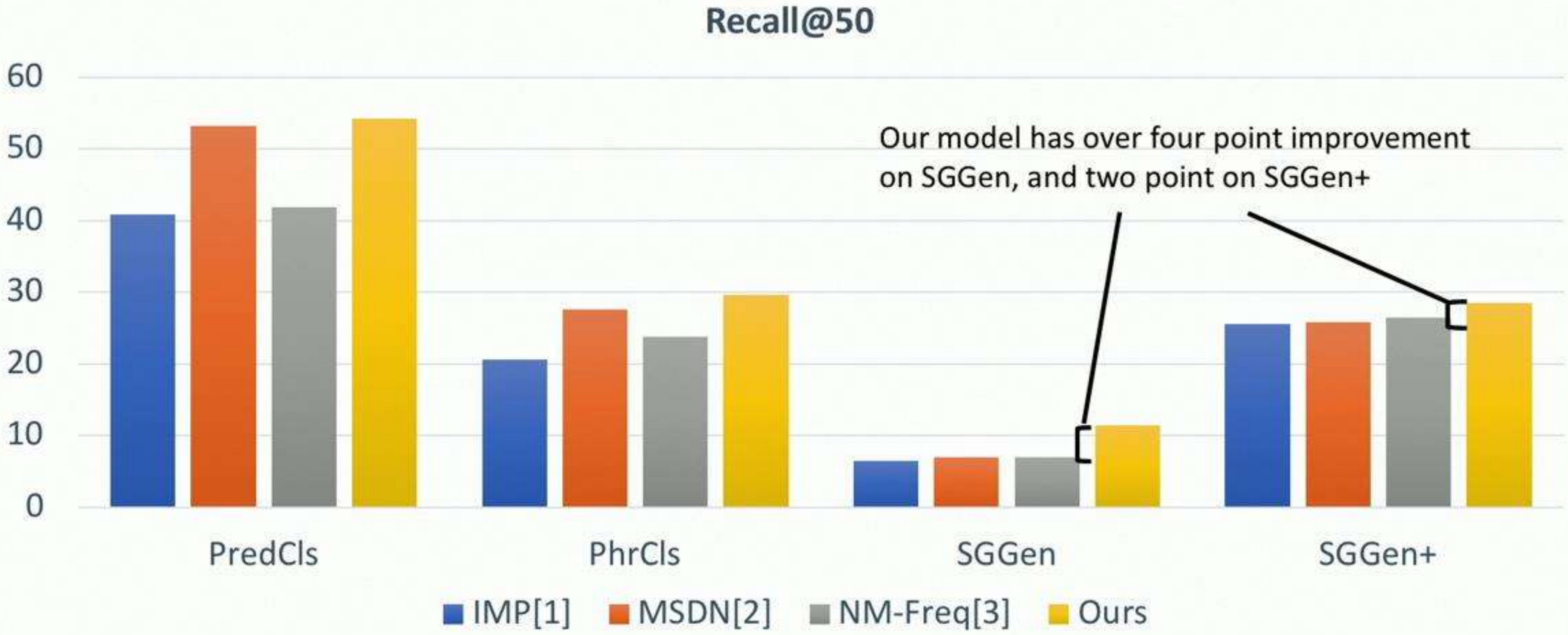
Experiments

Table. Implementation Details.

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

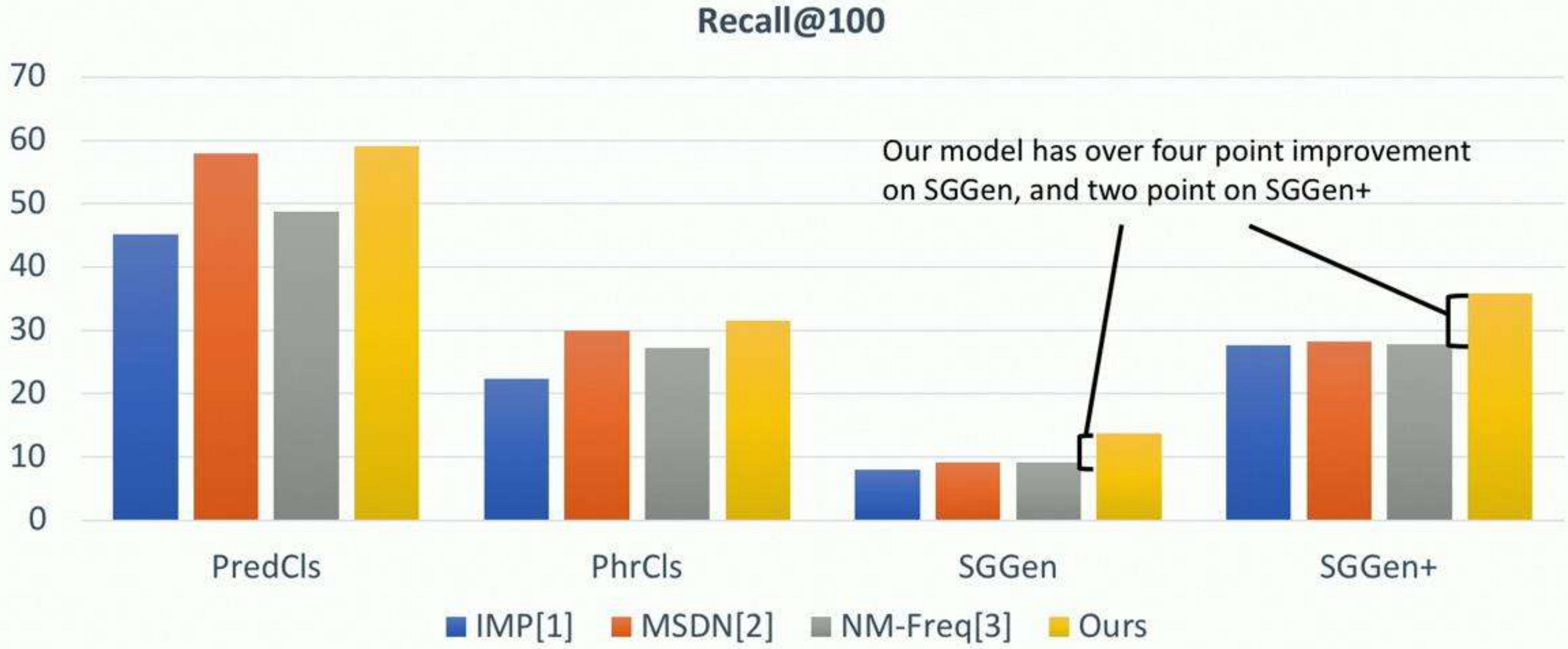
[1] A Faster Implementation of Faster R-CNN. Yang and Lu et al.

Comparing with Previous Work



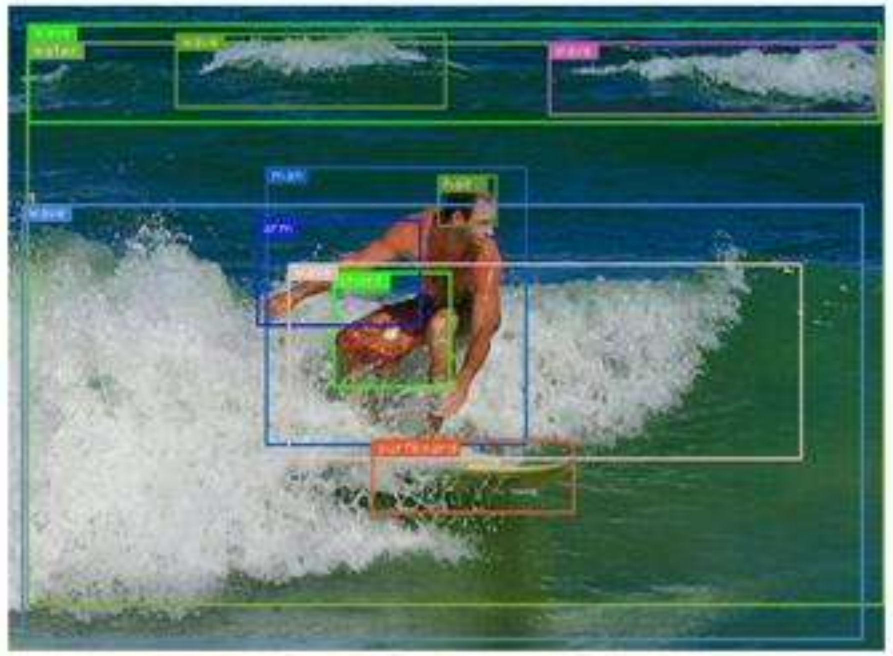
[1] Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017
[2] Scene Graph Generations from Objects, Phrases and Captions. Li et al. ICCV 2017
[3] Neural Motif: Scene Graph Parsing with Global Context. Zellers et al. CVPR 2018

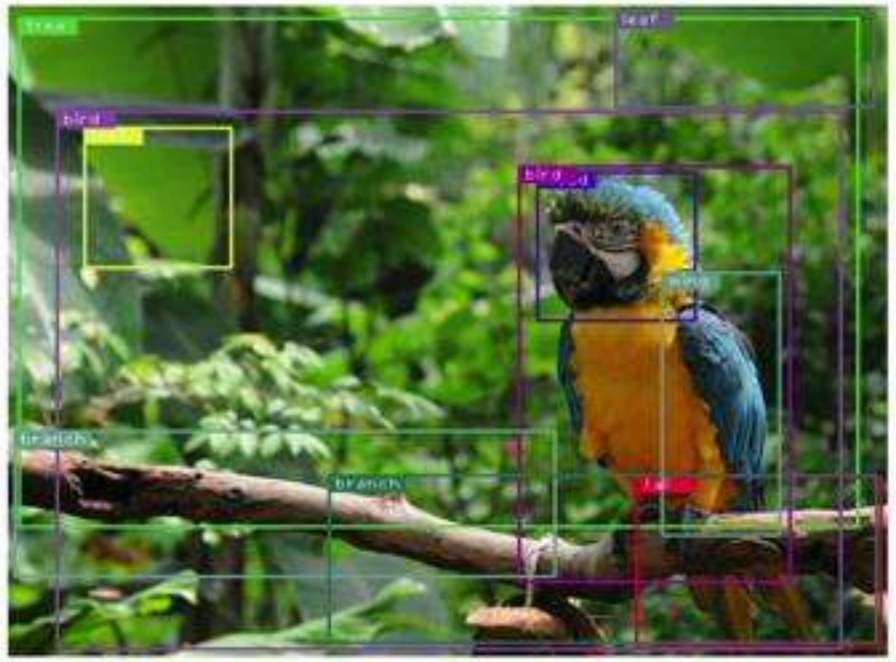
Comparing with Previous Work

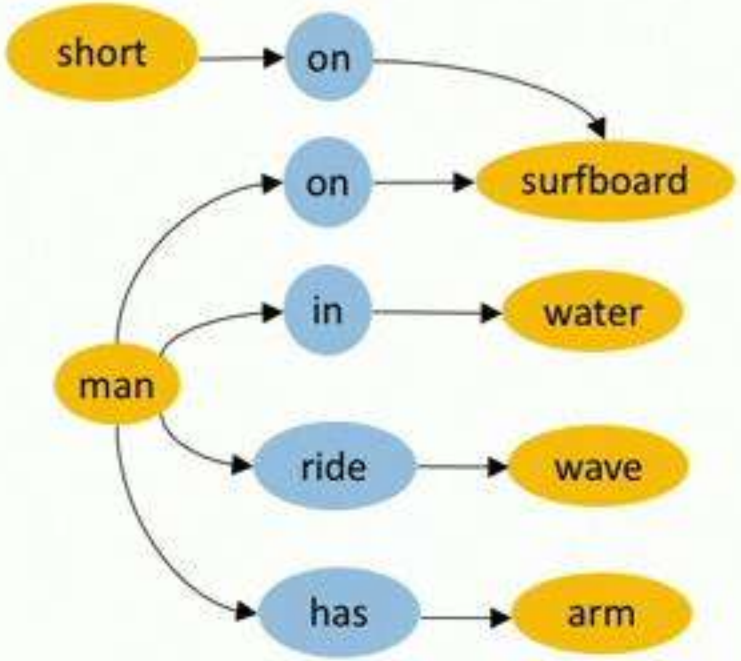


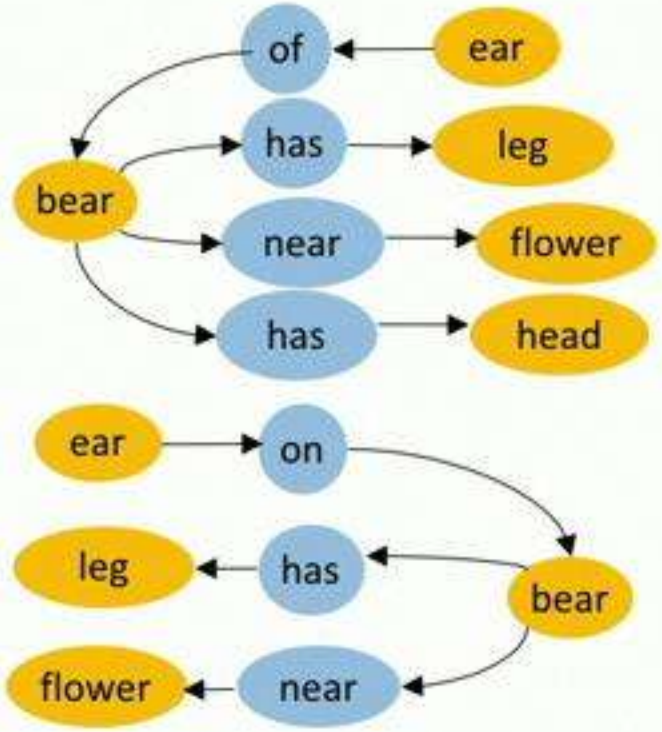
[1] Scene Graph Generation by Iterative Message Passing. Xu et al. CVPR 2017
[2] Scene Graph Generations from Objects, Phrases and Captions. Li et al. ICCV 2017
[3] Neural Motif: Scene Graph Parsing with Global Context. Zellers et al. CVPR 2018

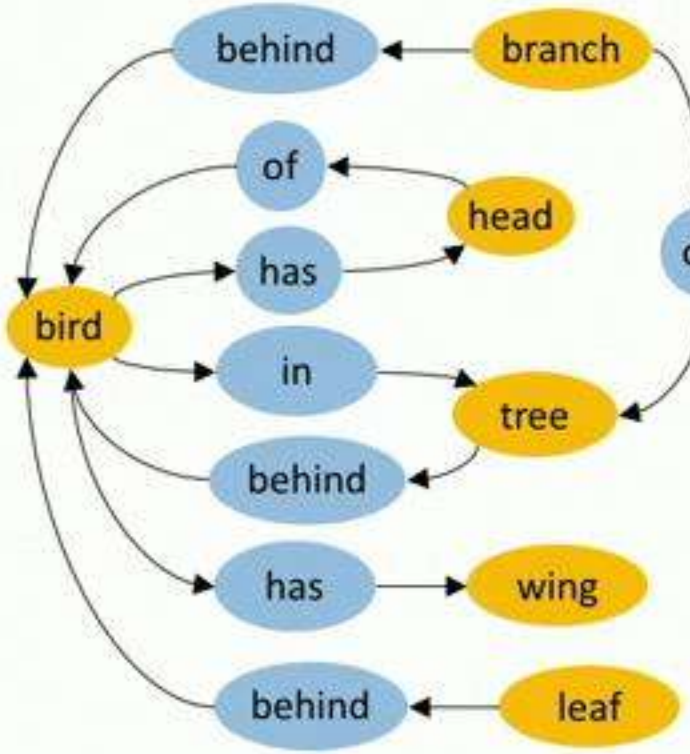
Qualitative Results











Common Sense Emerges

We extract the weights in the score-level aGCN layer, and sort it in descending order.

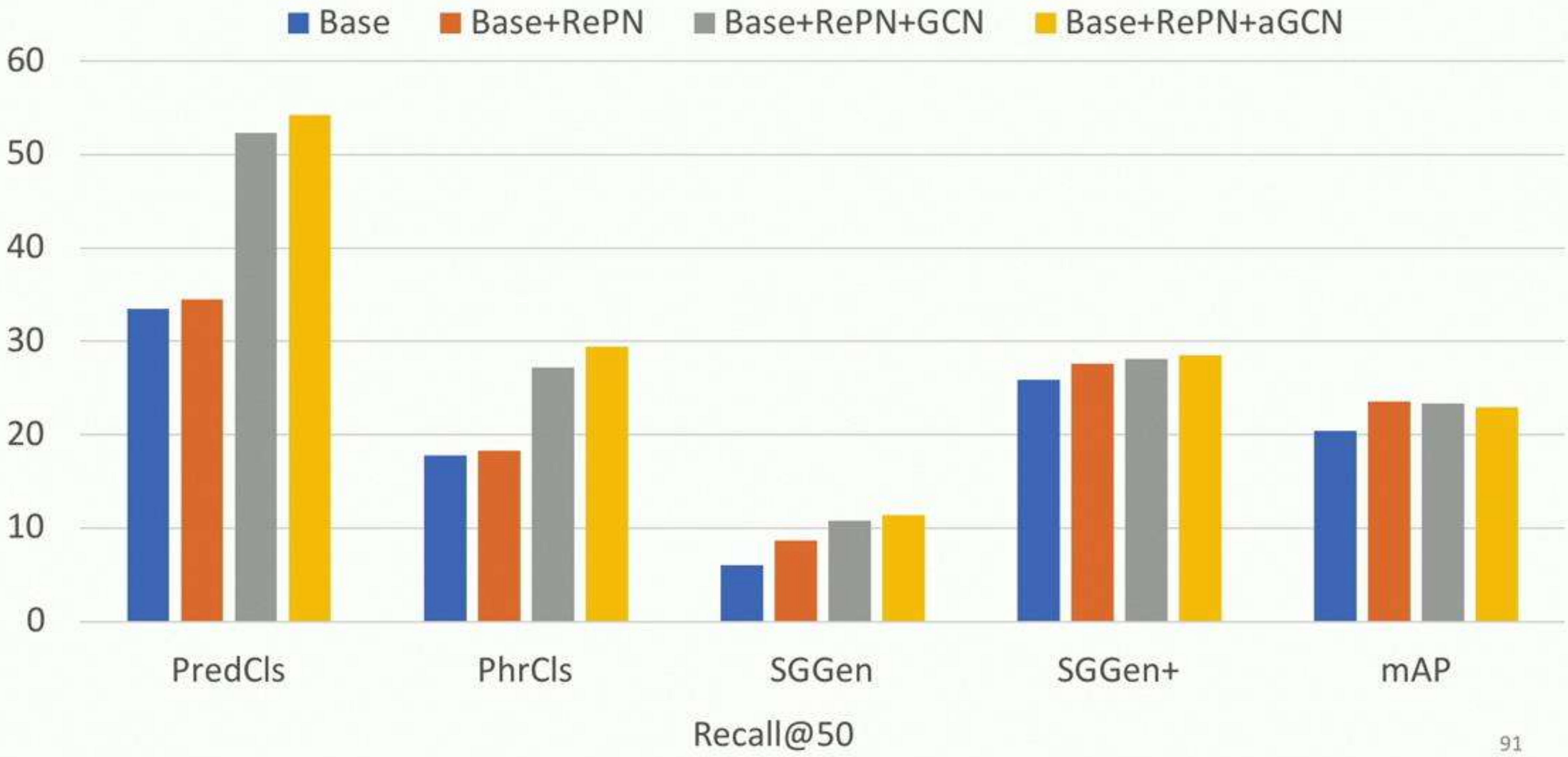
Object-Object Co-Occurrence

Object	Top-1	Top-2	Object	Top-1	Top-2
boat	water	beach	**girl**	woman	hair
plane	wing	tail	**cow**	horse	dog
clock	building	root	**sidewalk**	street	bus
bottle	cup	glass	**handle**	plate	food
bus	truck	vehicle	**snow**	pole	ski

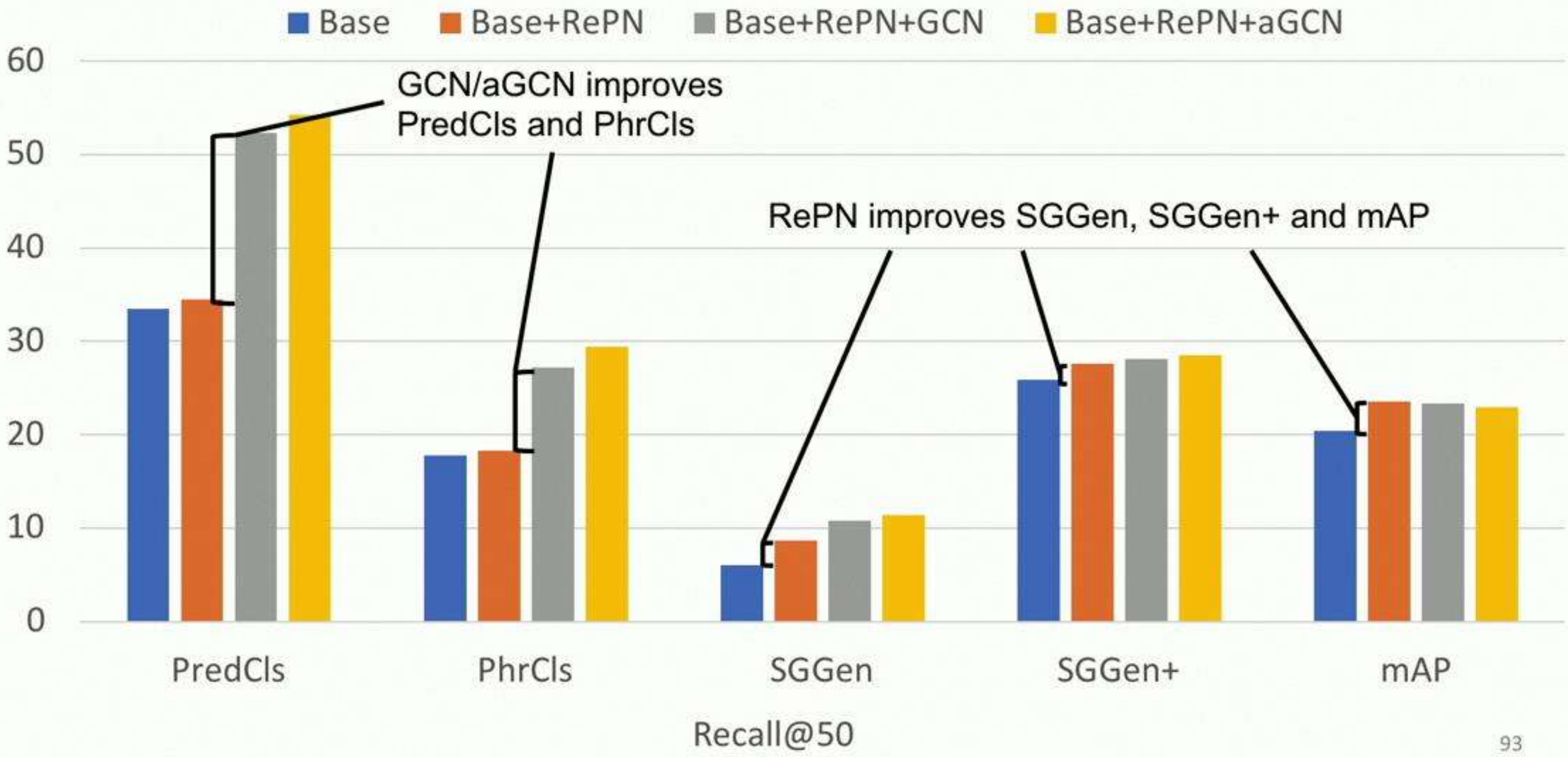
Object-Predicate Co-Occurrence

Object	Top-1	Top-2	Object	Top-1	Top-2
hat	hold	wear	**kite**	watch	look at
boat	in	sit in	**girl**	look at	watch
umbrella	carry	hold	**jacket**	wear	with
track	with	on	**stripe**	on	has
sidewalk	at	walk on	**snow**	on	near

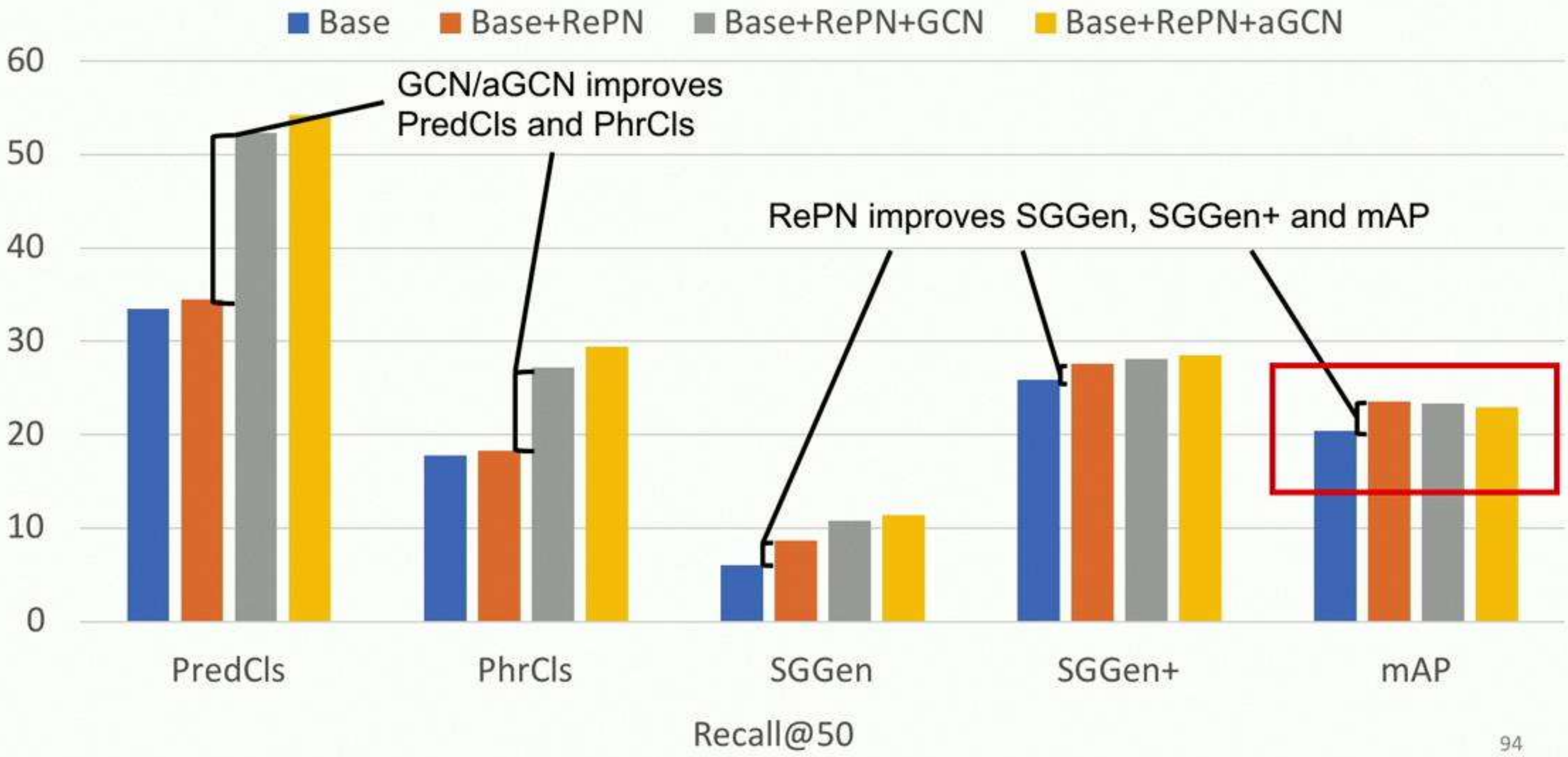
Ablation Study



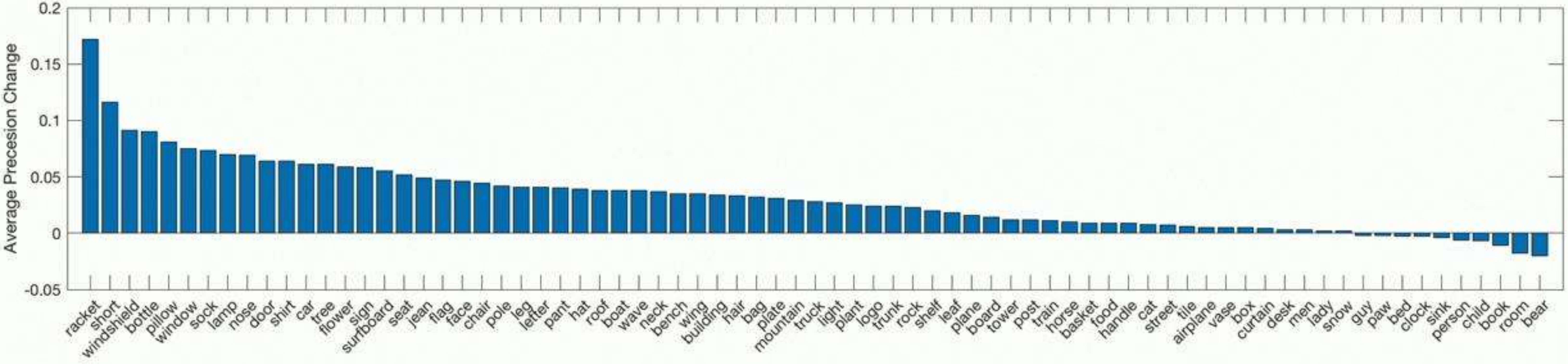
Ablation Study



Ablation Study

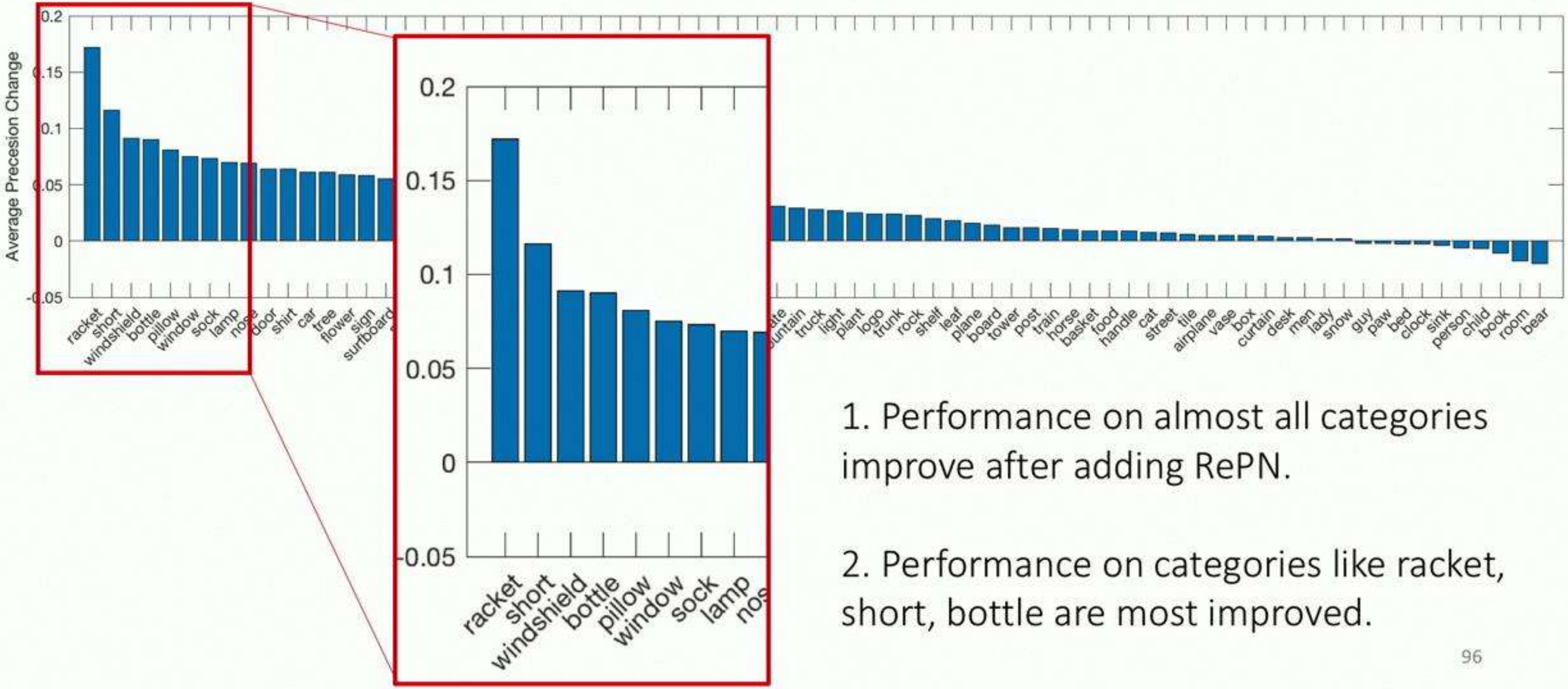


Object Detection Investigation



1. Performance on almost all categories improve after adding RePN.

Object Detection Investigation



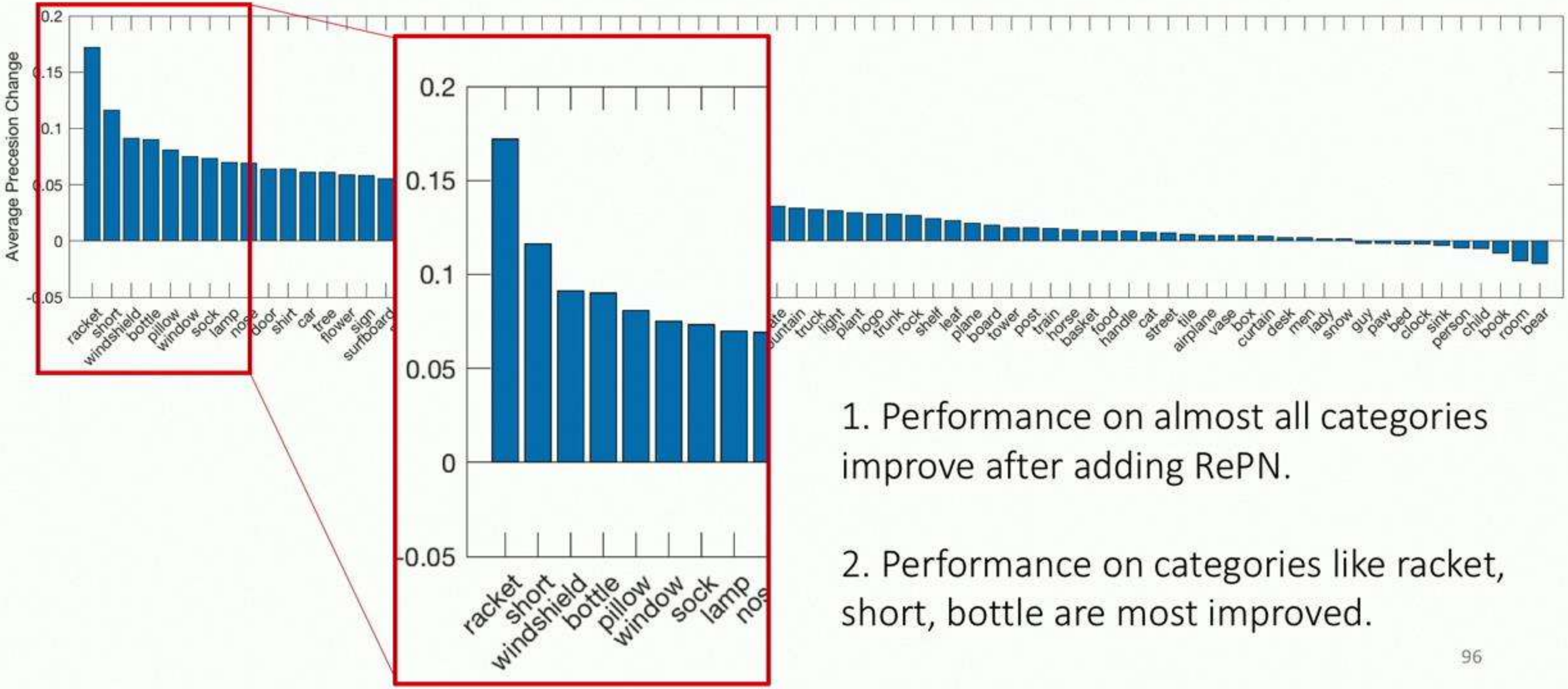
1. Performance on almost all categories improve after adding RePN.

2. Performance on categories like racket, short, bottle are most improved.

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

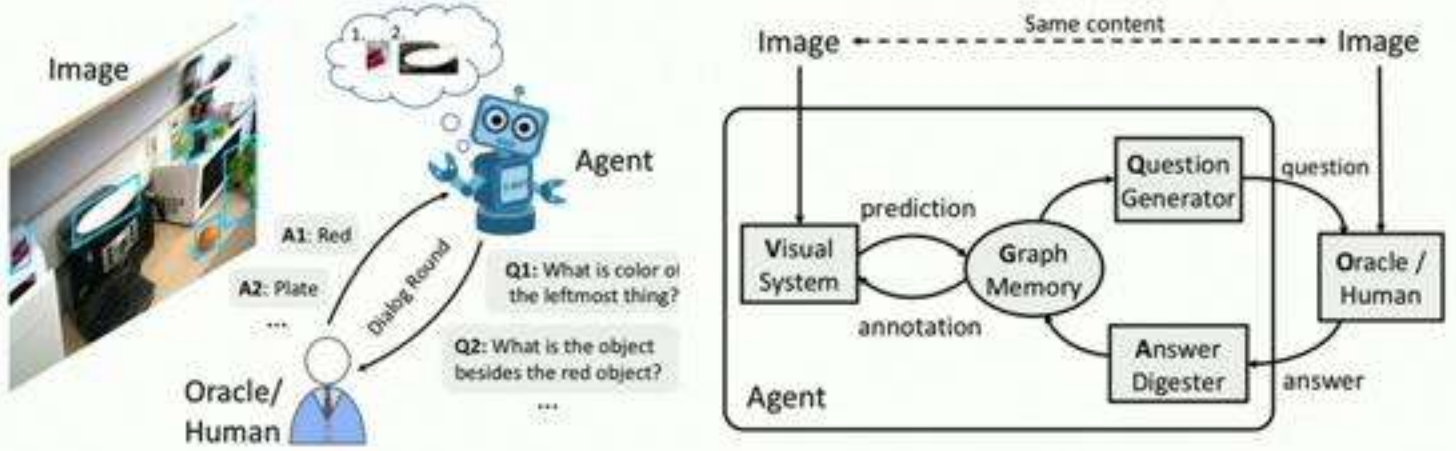
Object Detection Investigation



1. Performance on almost all categories improve after adding RePN.

2. Performance on categories like racket, short, bottle are most improved.

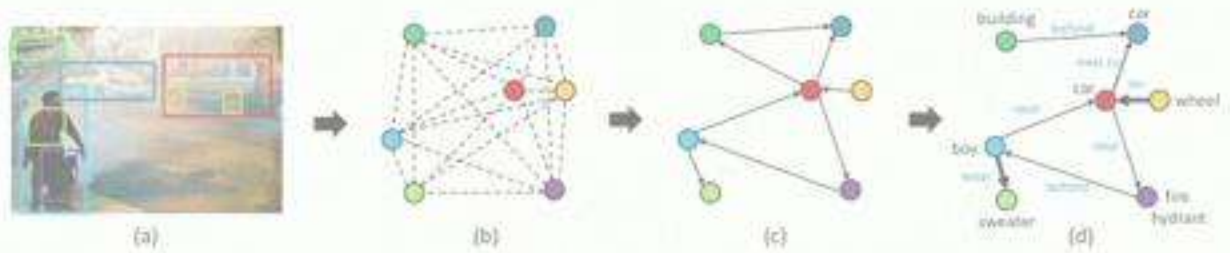
In this talk



Visual Curiosity. CoRL 2018

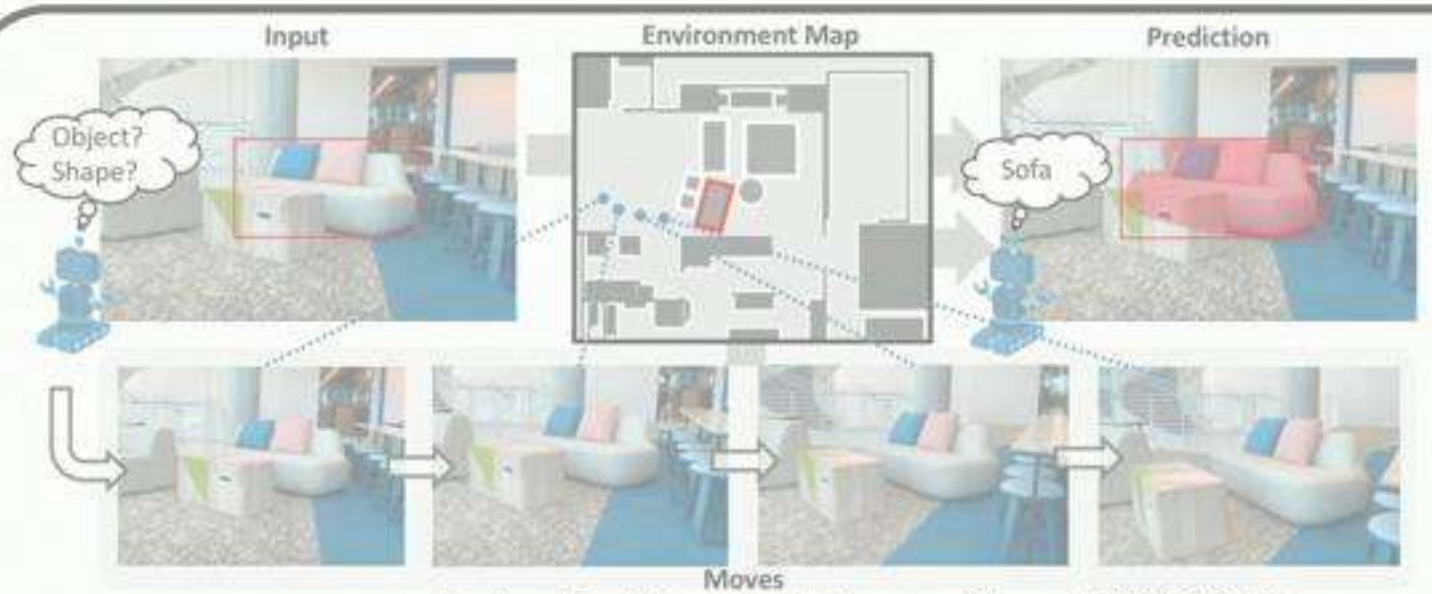
Interact with Human

Per-Image Structure



Graph R-CNN for Scene Graph Generation. ECCV 2018

Structured Visual Understanding



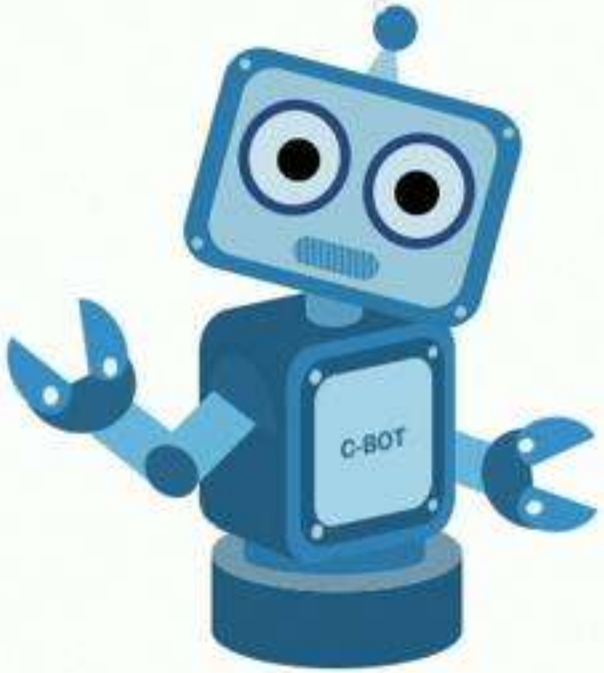
Embodied Amodal Recognition. ICCV 2019

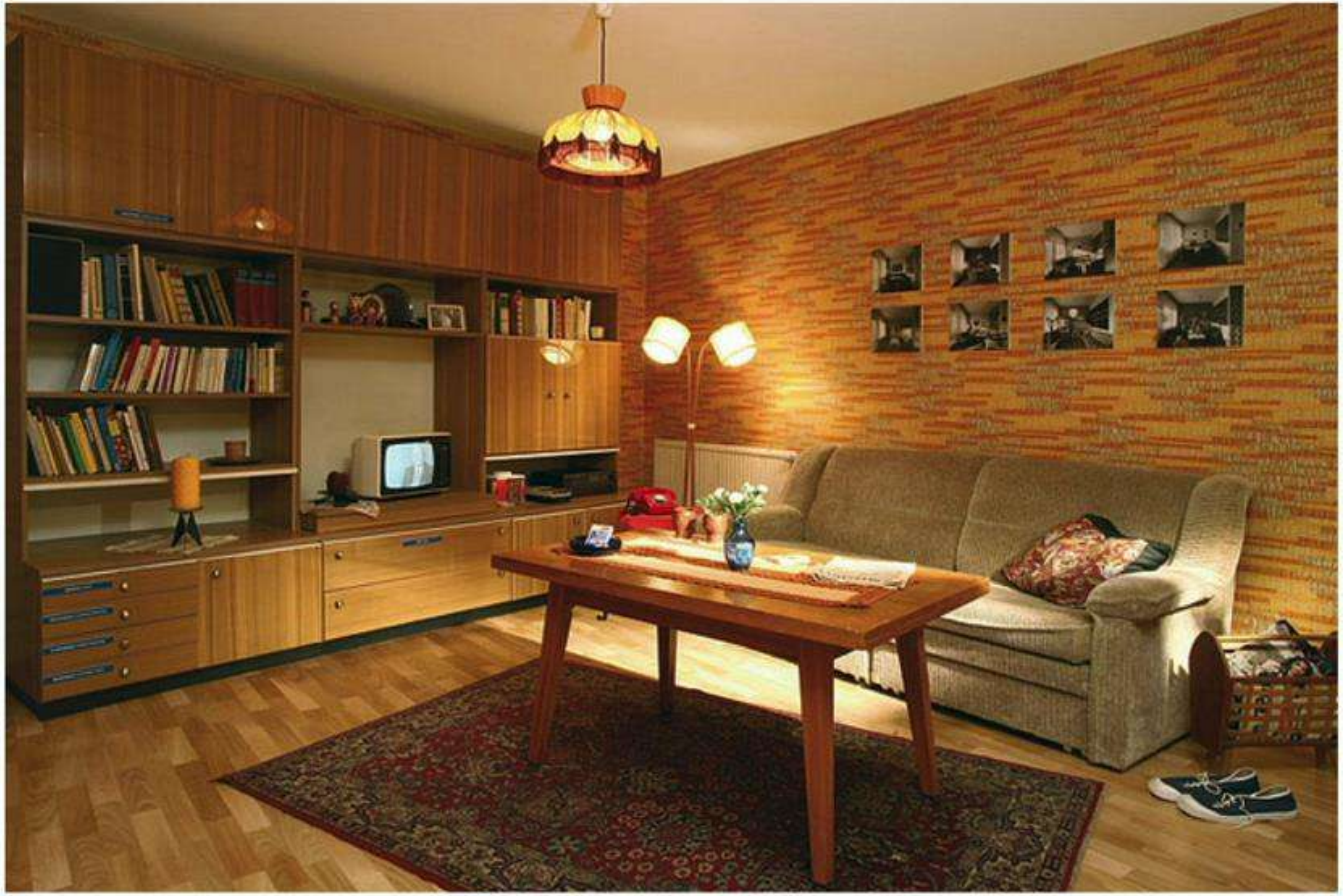
Interact with Environment

Visual Understanding by Asking Questions

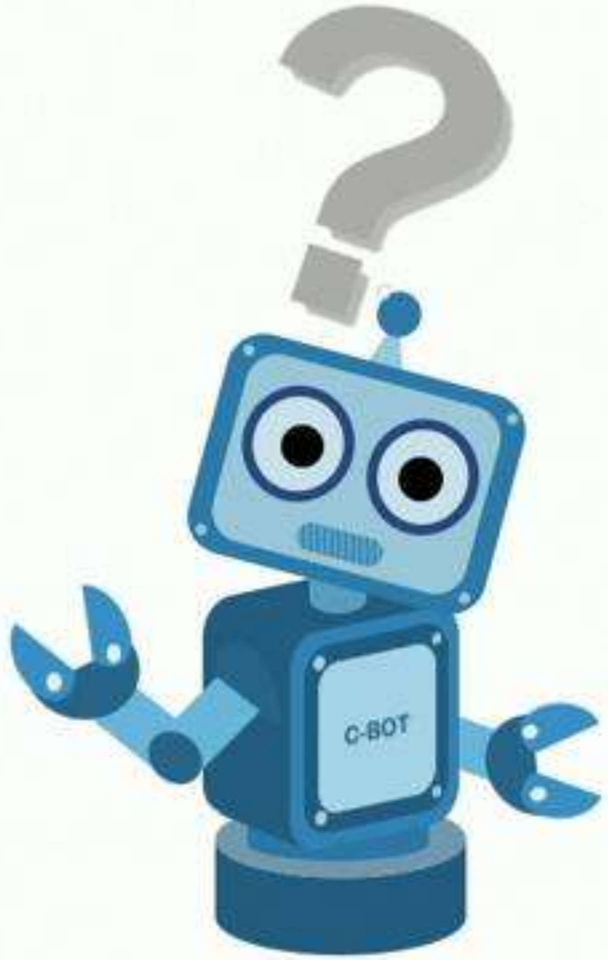
Visual Curiosity: Learning to Ask Questions to Learn Visual Recognition. CoRL 2018.

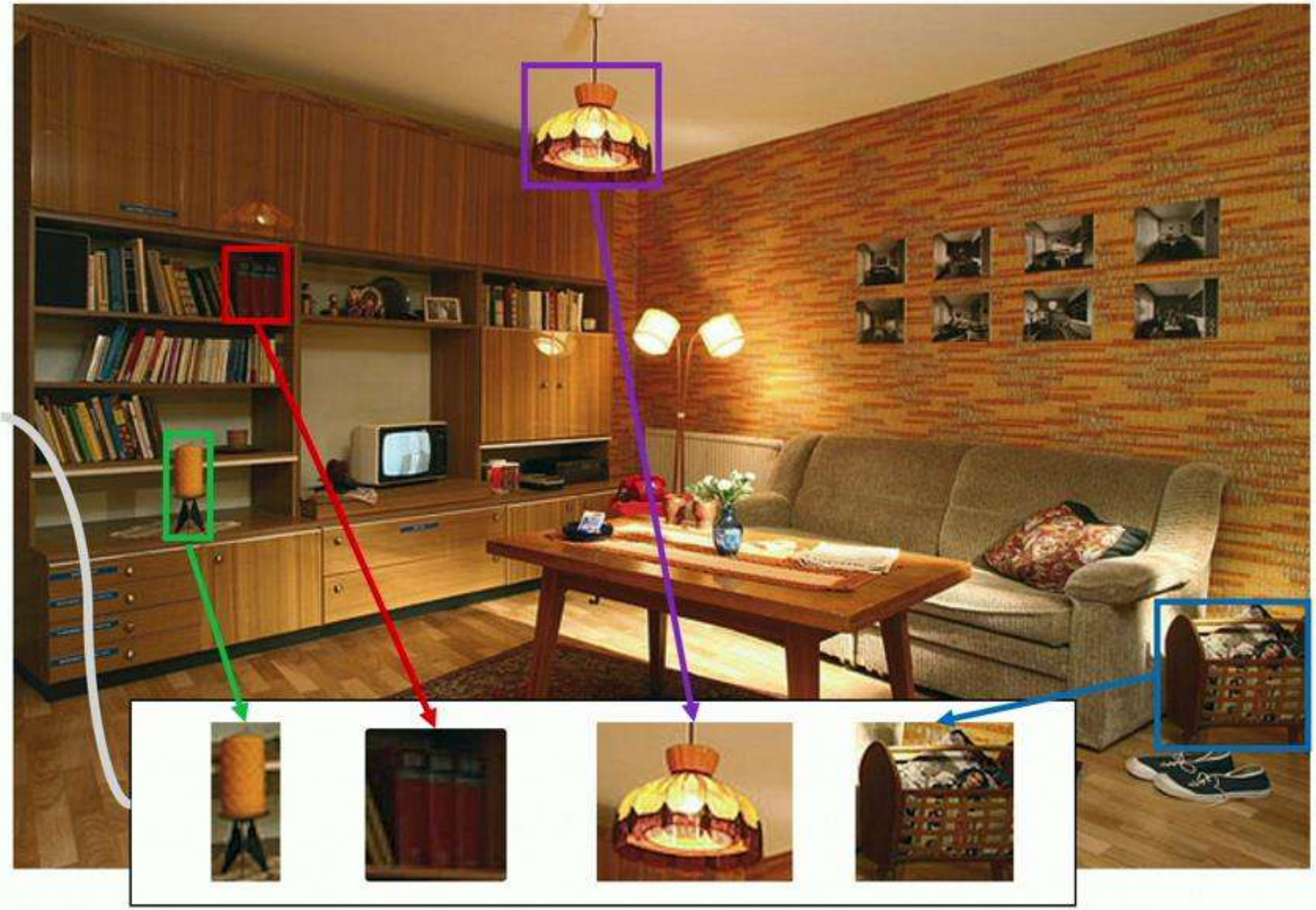
The Open-World Recognition Problem



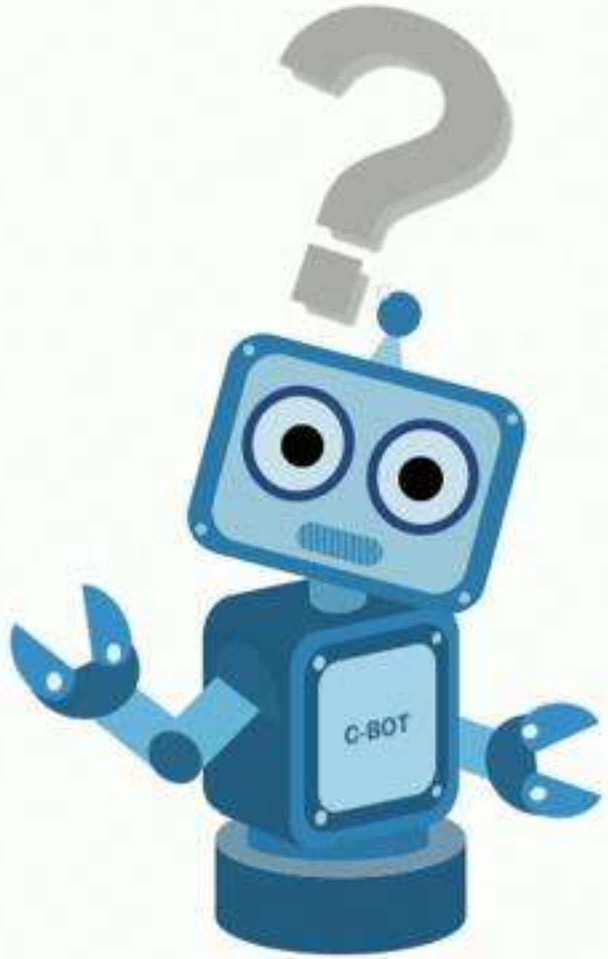


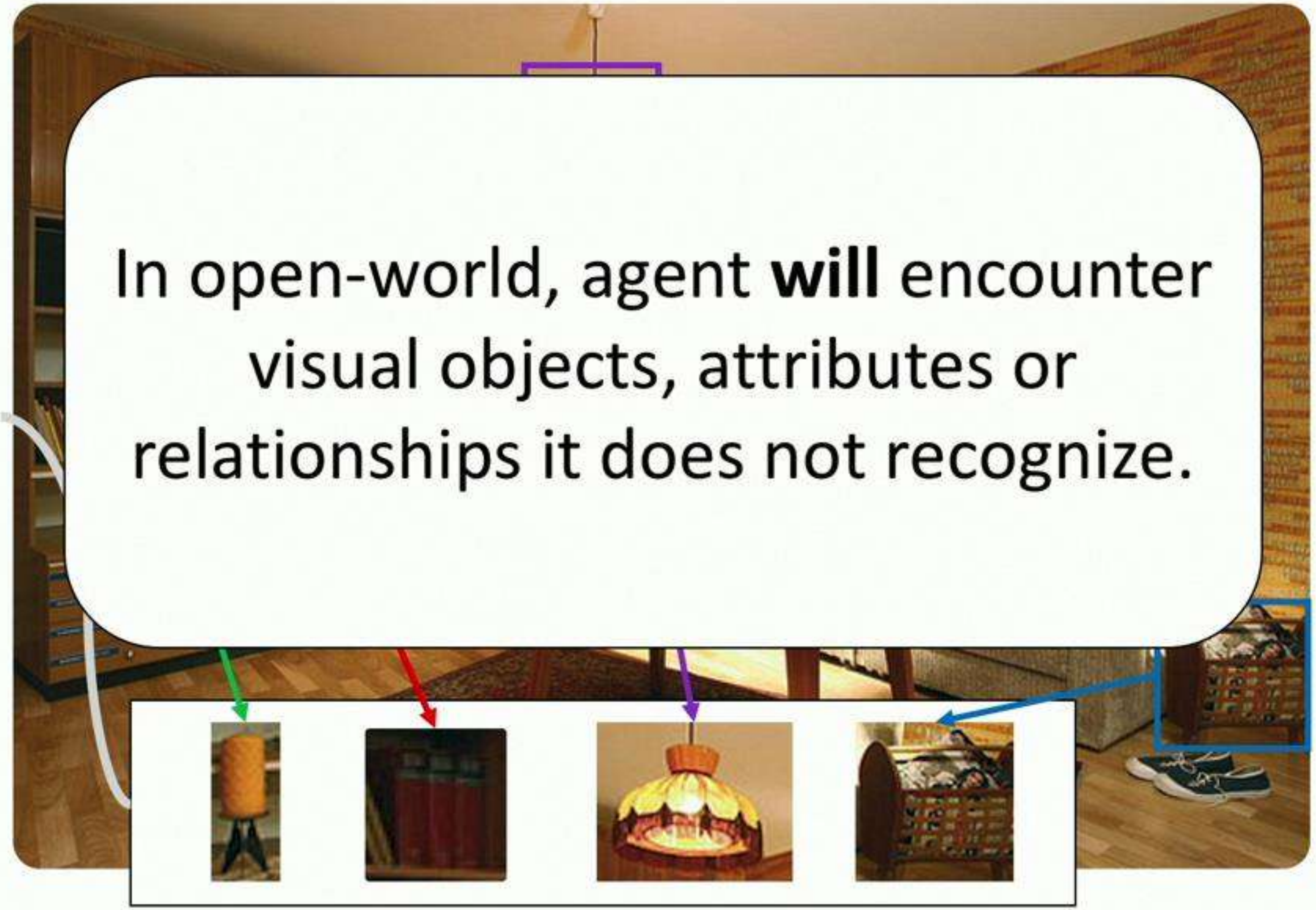
The Open-World Recognition Problem





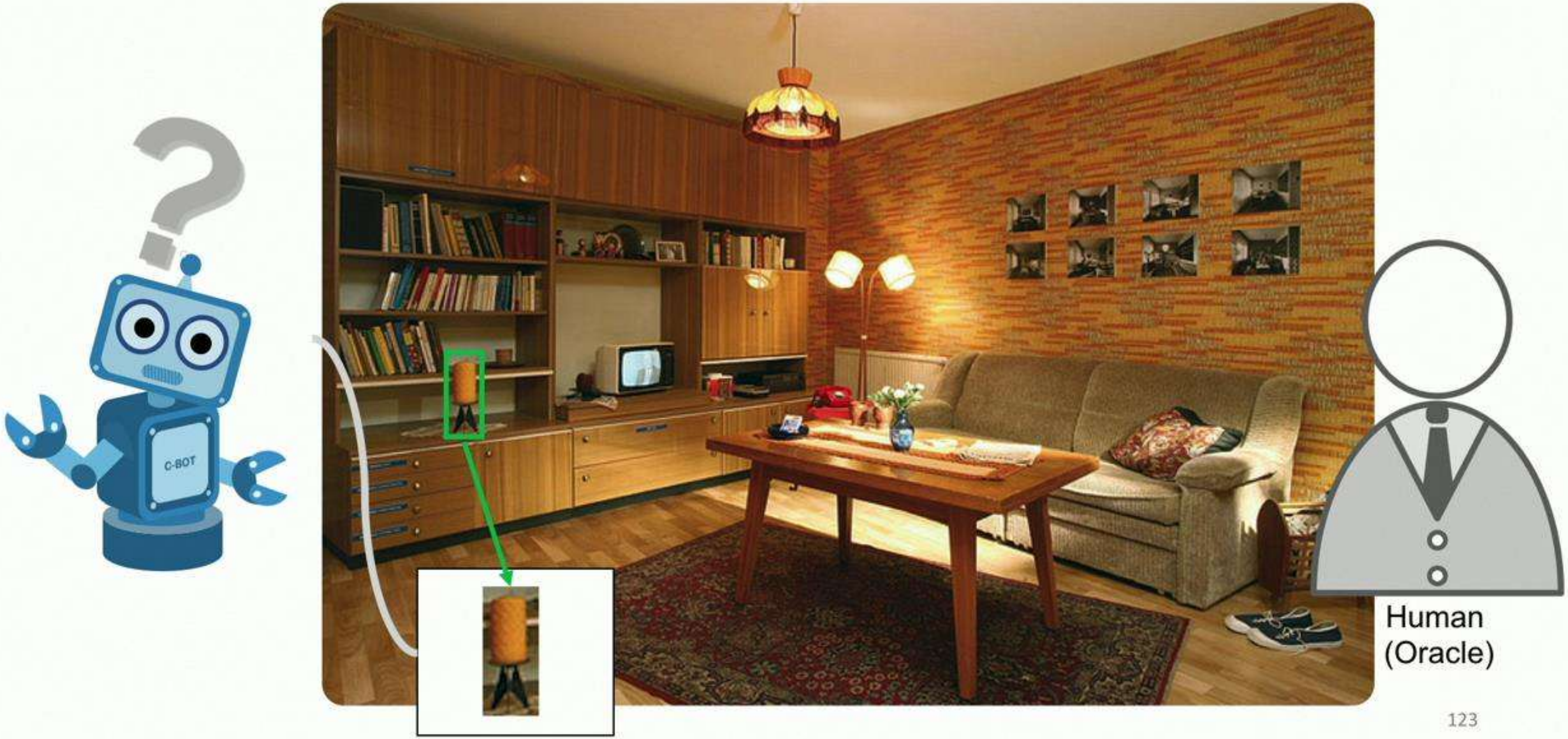
The Open-World Recognition Problem



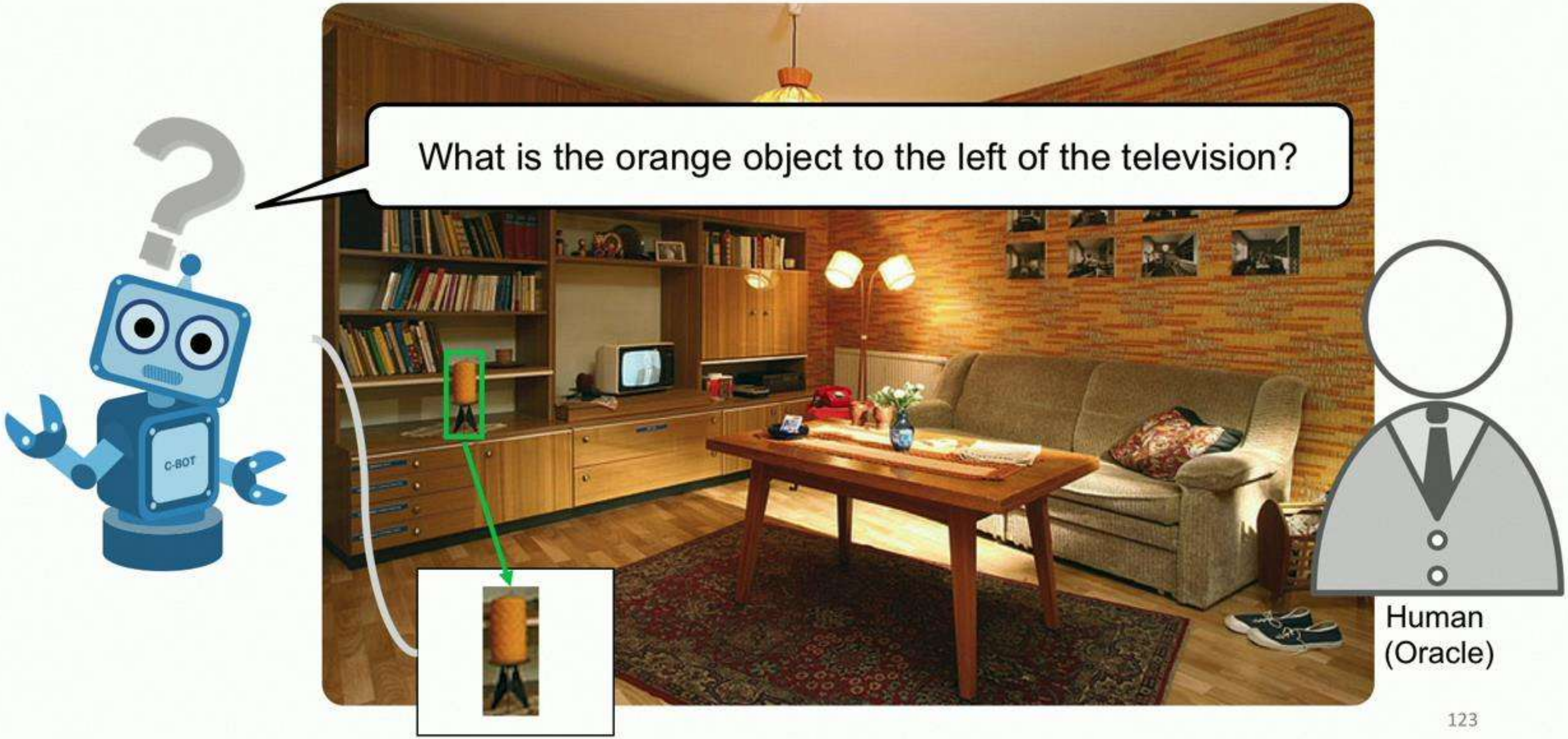


In open-world, agent **will** encounter visual objects, attributes or relationships it does not recognize.

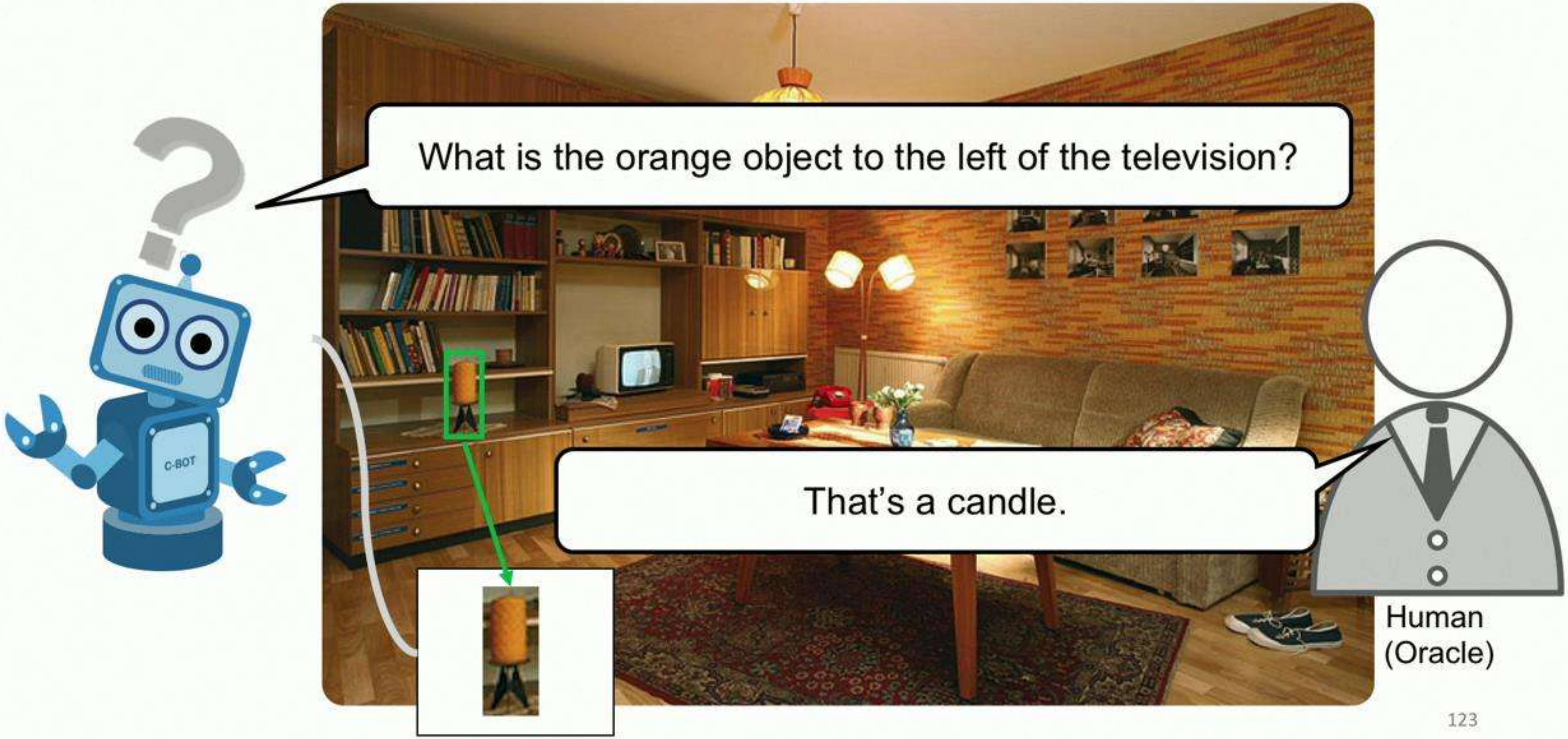
How can an agent learn about these concepts?



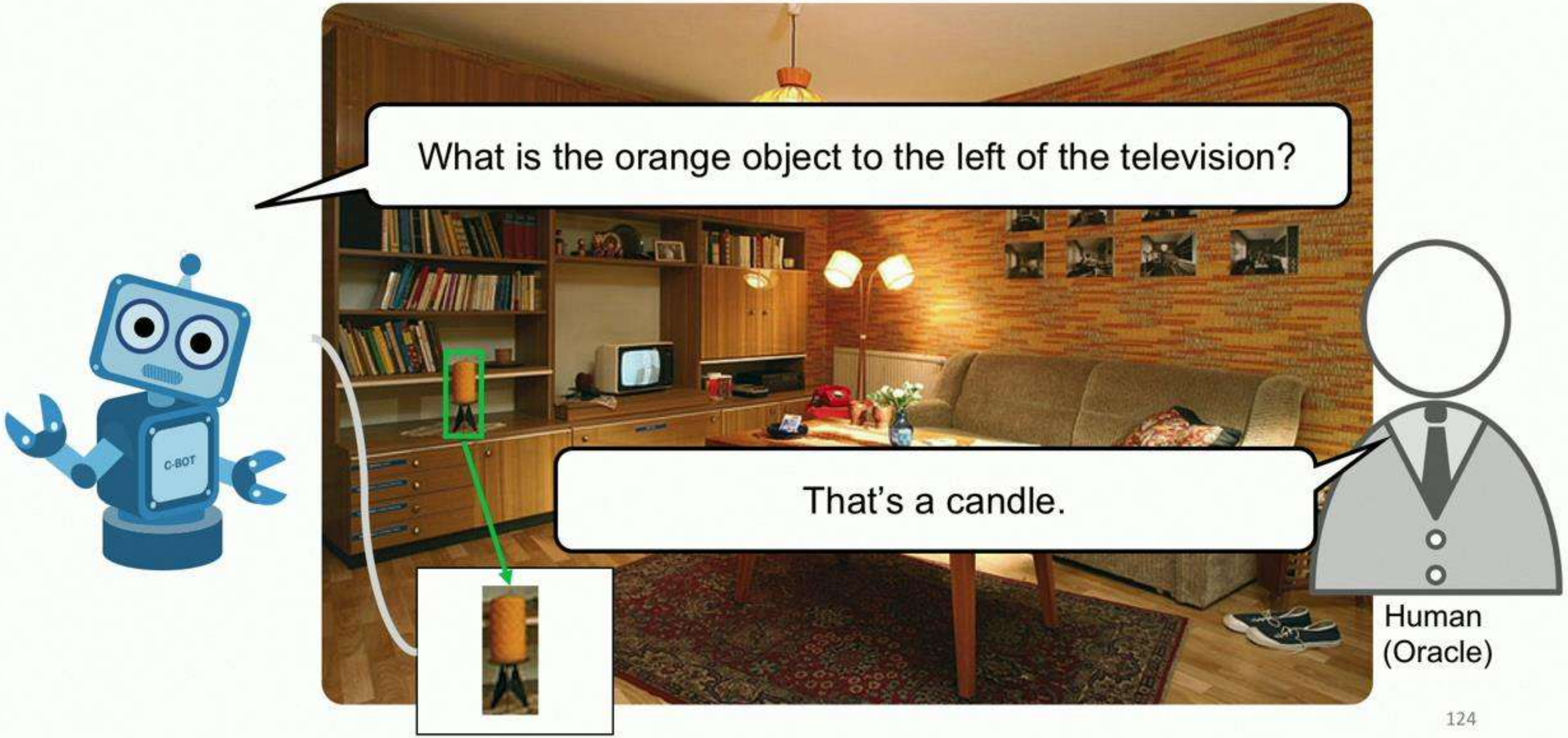
How can an agent learn about these concepts?



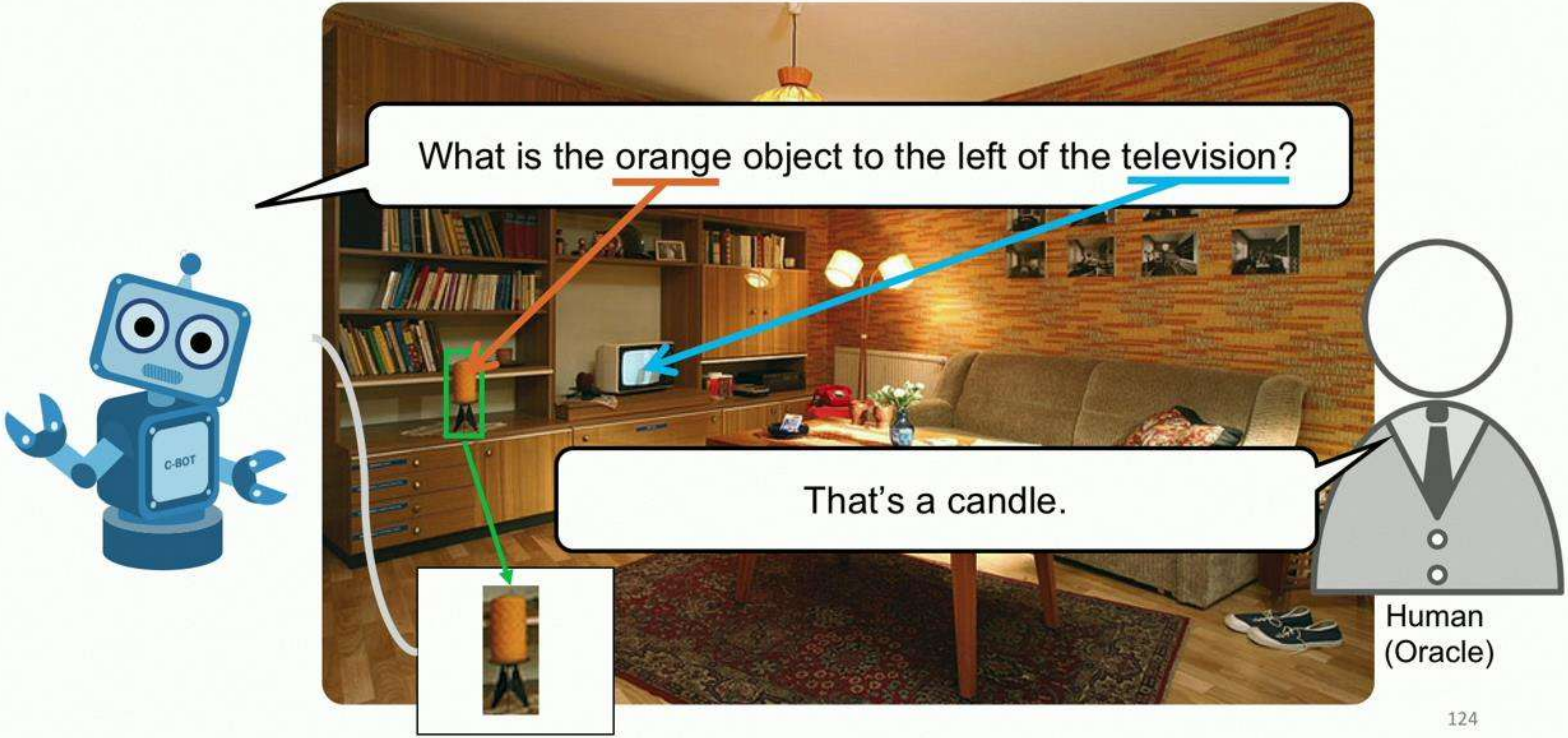
How can an agent learn about these concepts?



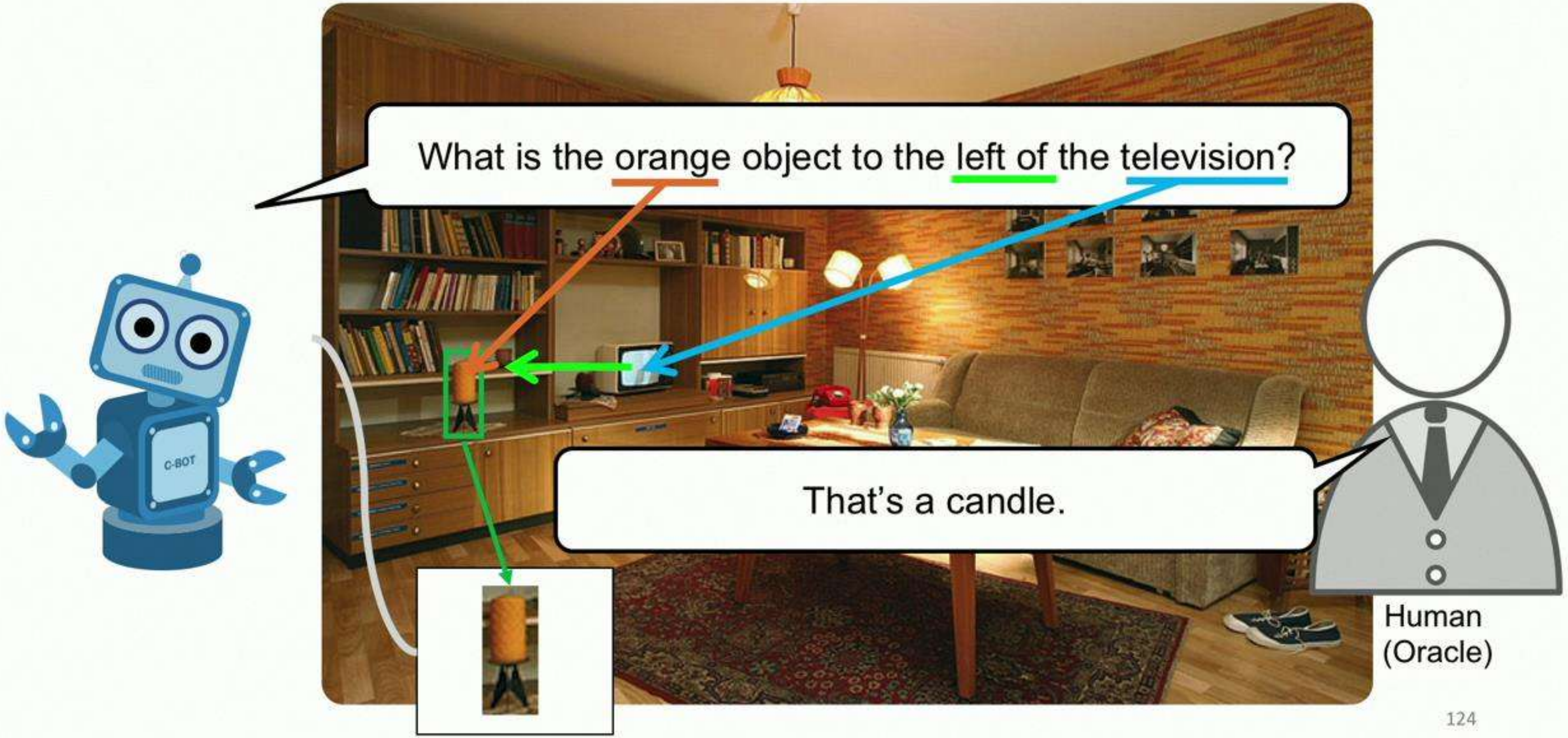
Asking informative questions is challenging!



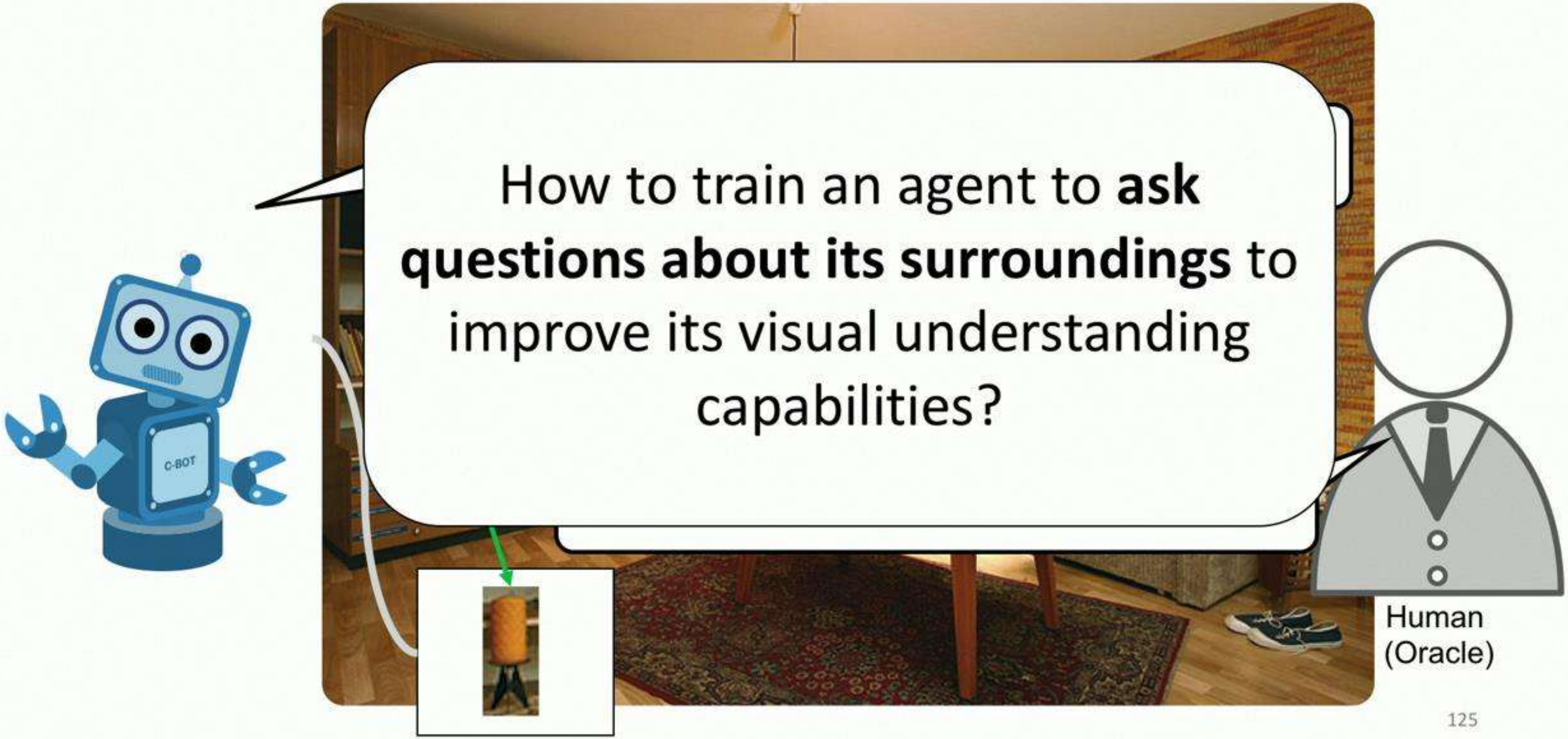
Asking informative questions is challenging!



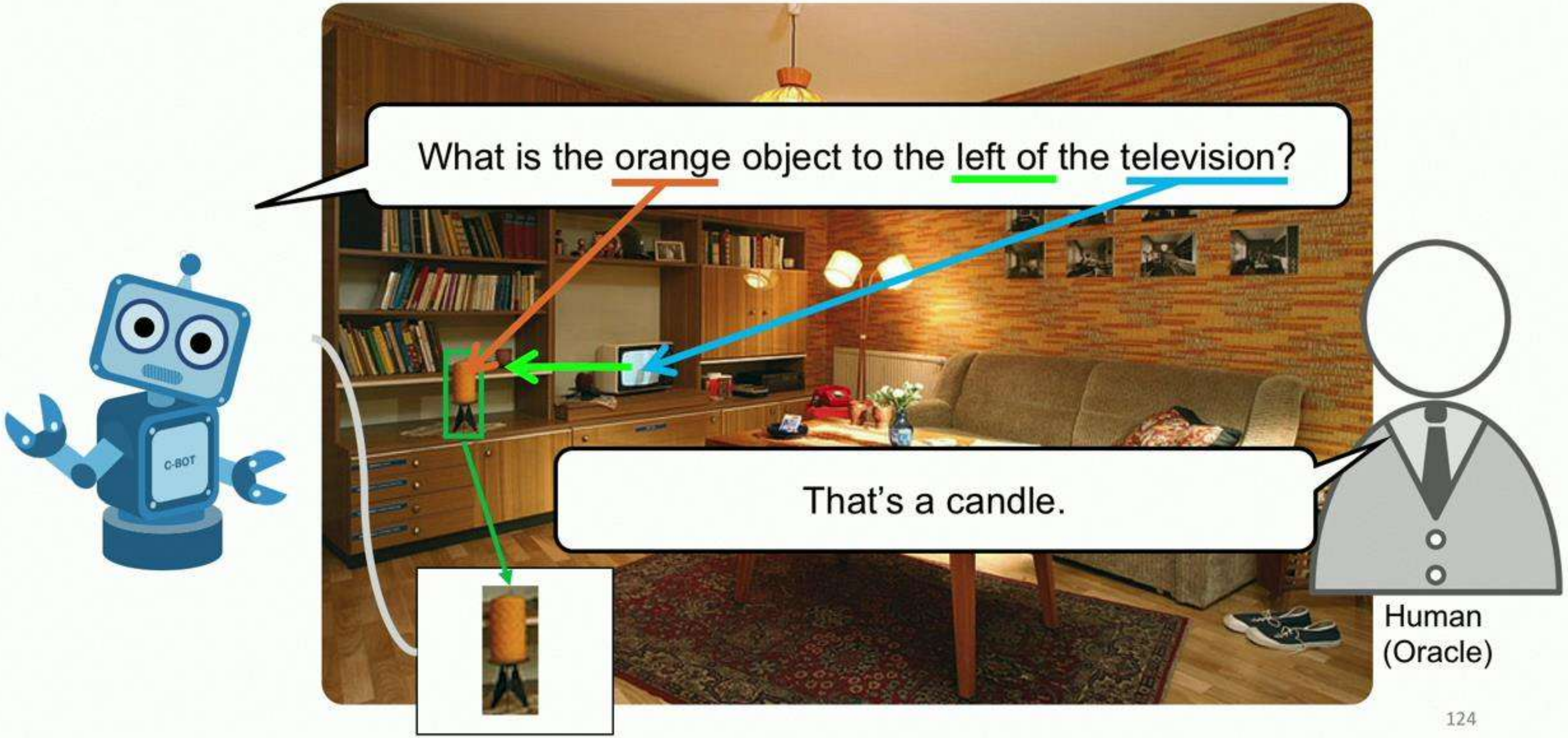
Asking informative questions is challenging!



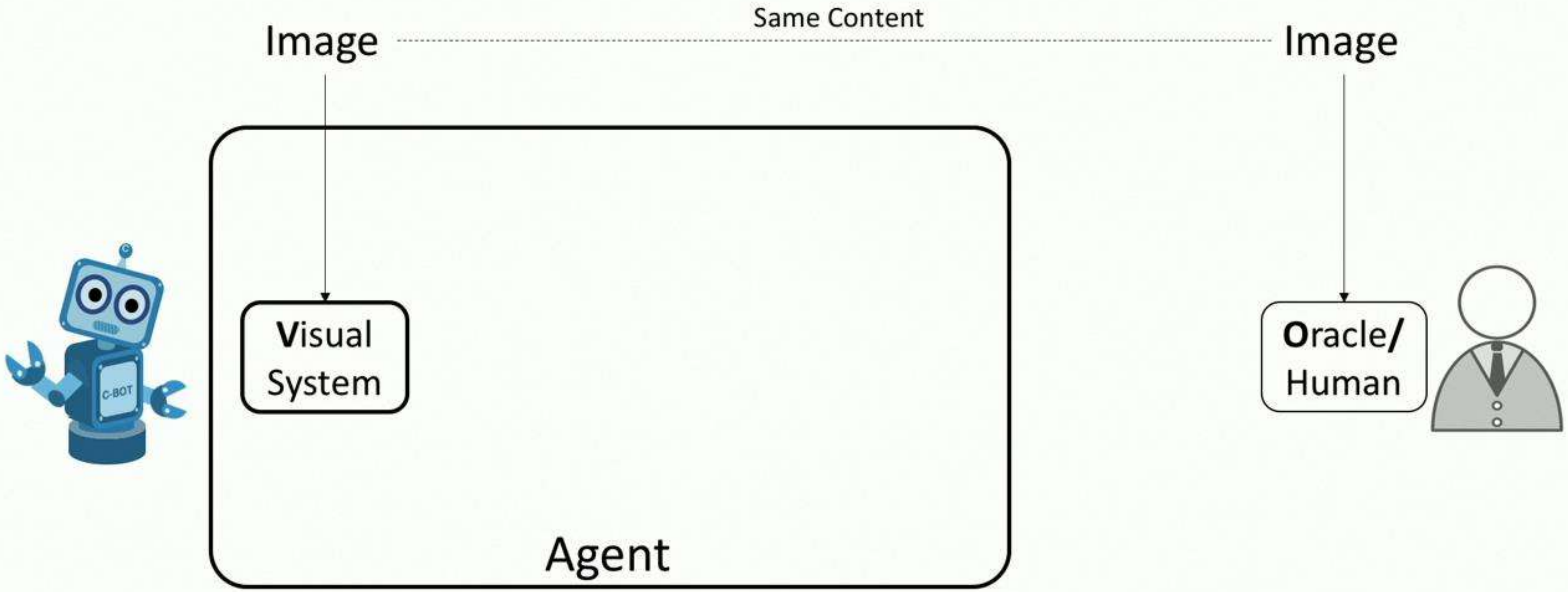
Asking informative questions is challenging!



Asking informative questions is challenging!

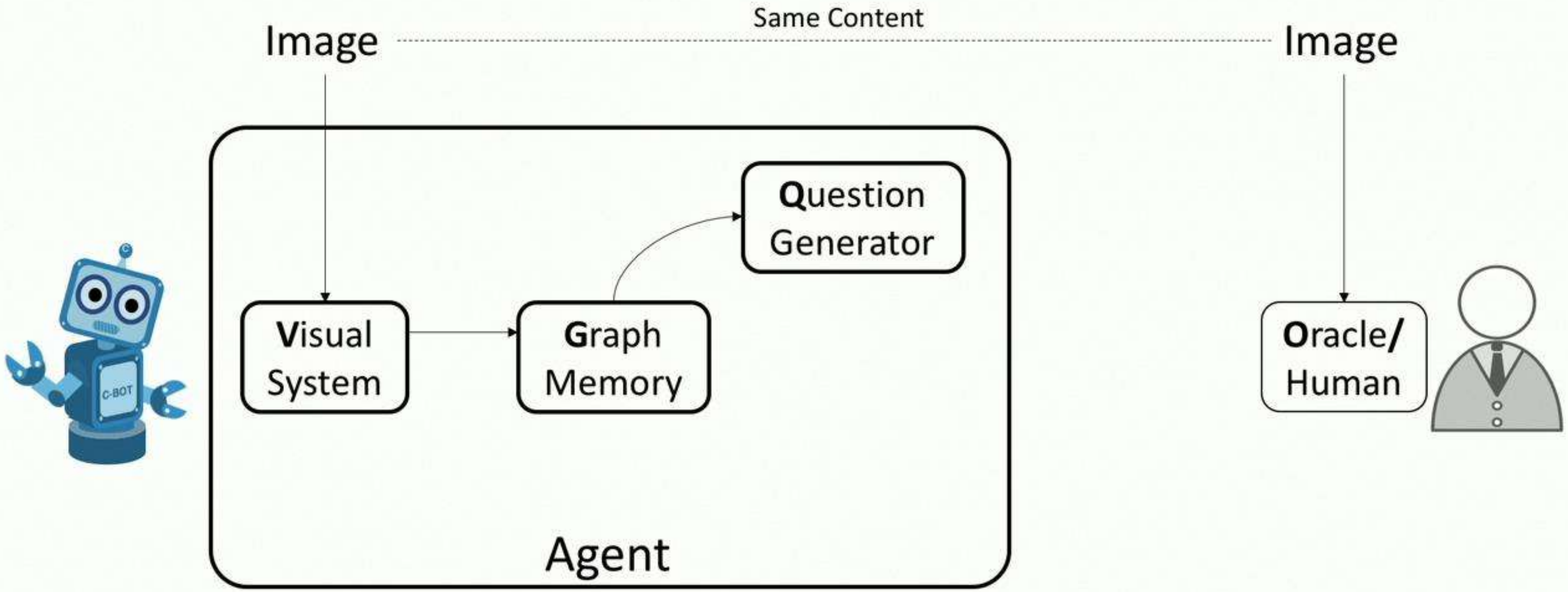


Agent Architecture



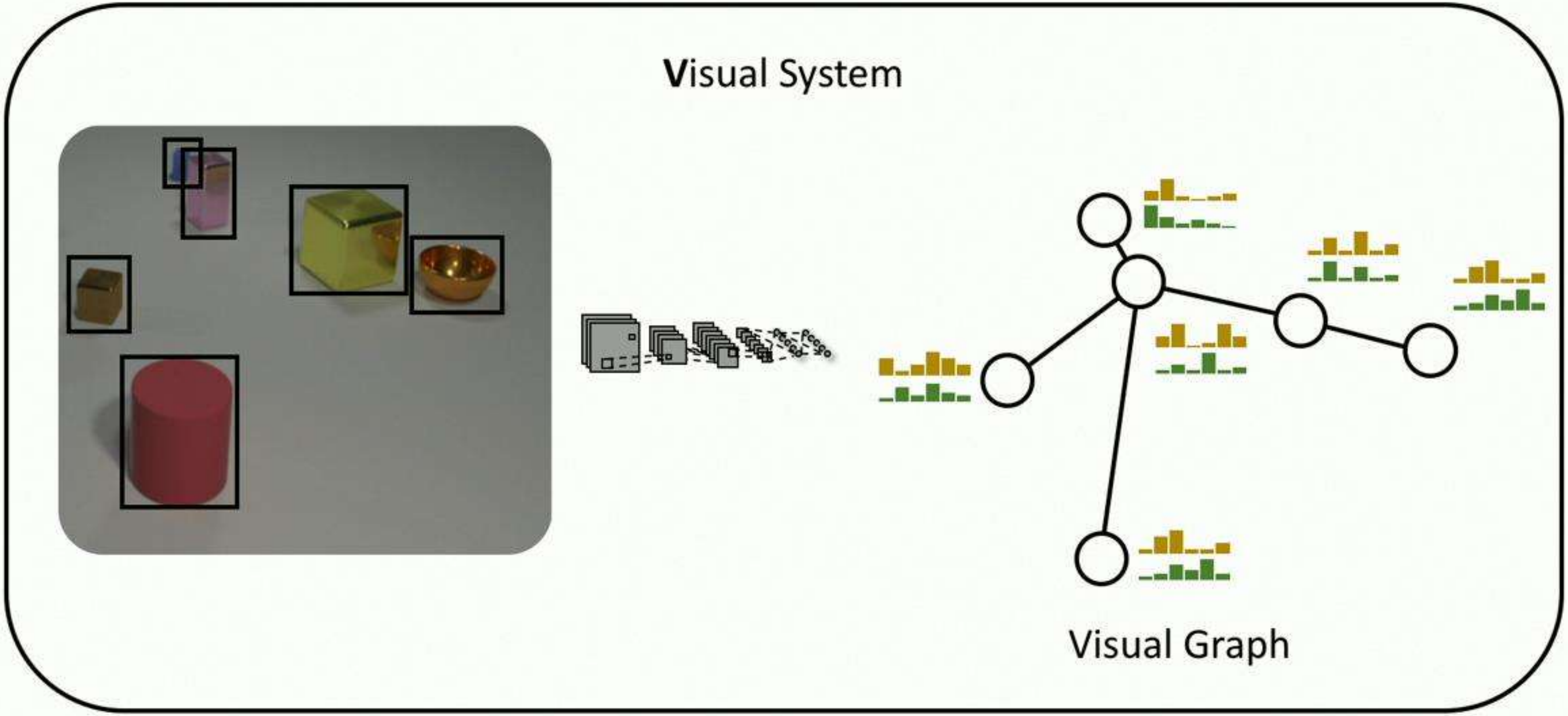
Slide Credit: Stefan Lee

Agent Architecture



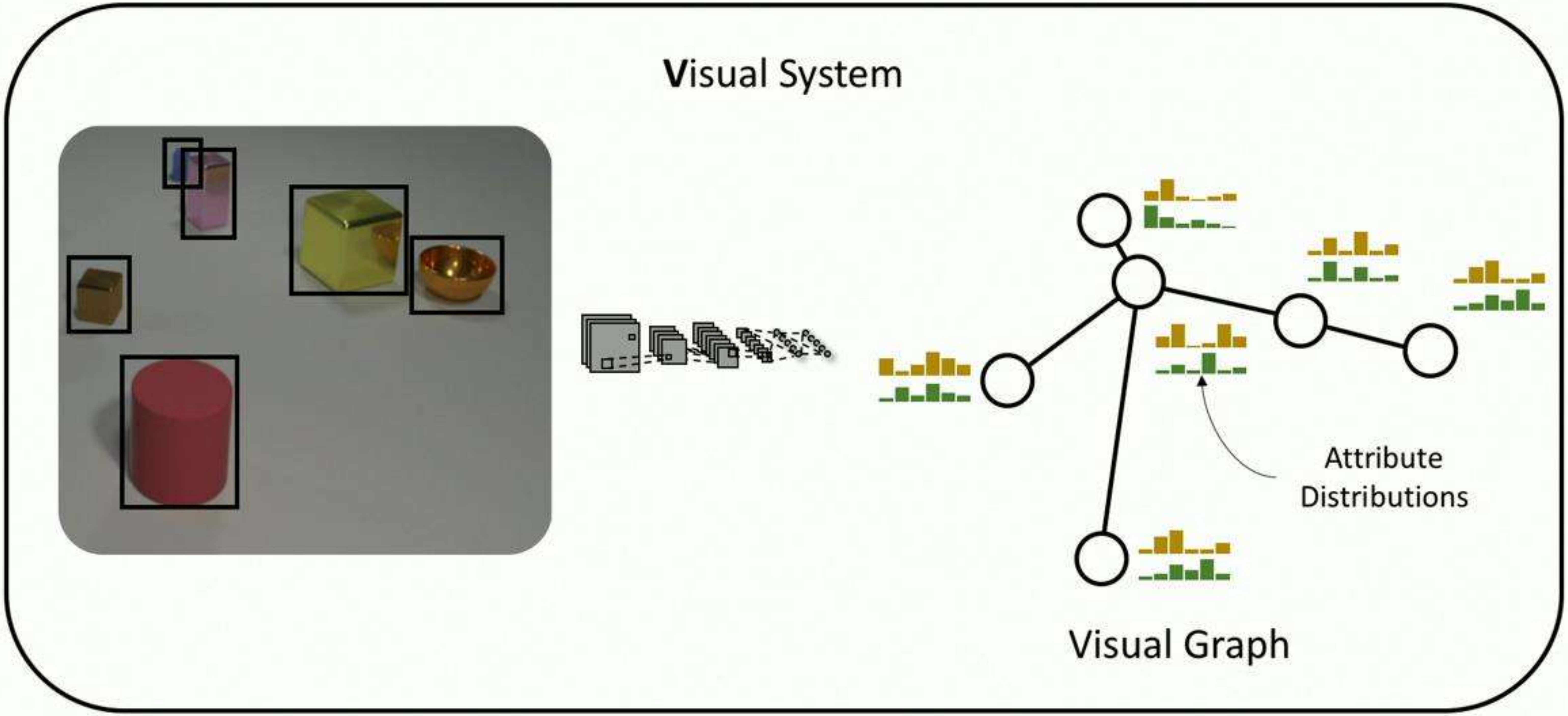
Slide Credit: Stefan Lee

Agent Architecture



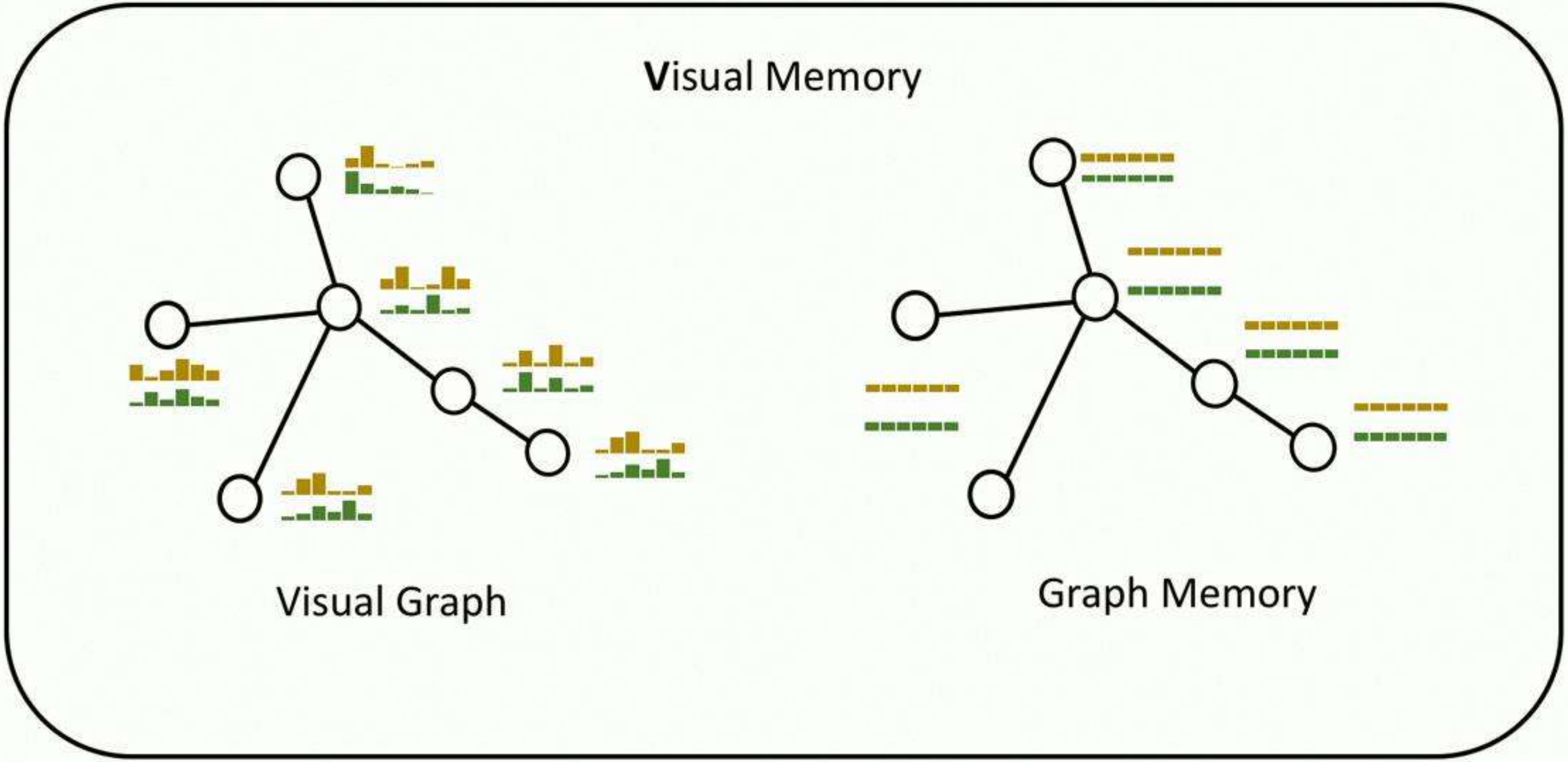
Slide Credit: Stefan Lee

Agent Architecture

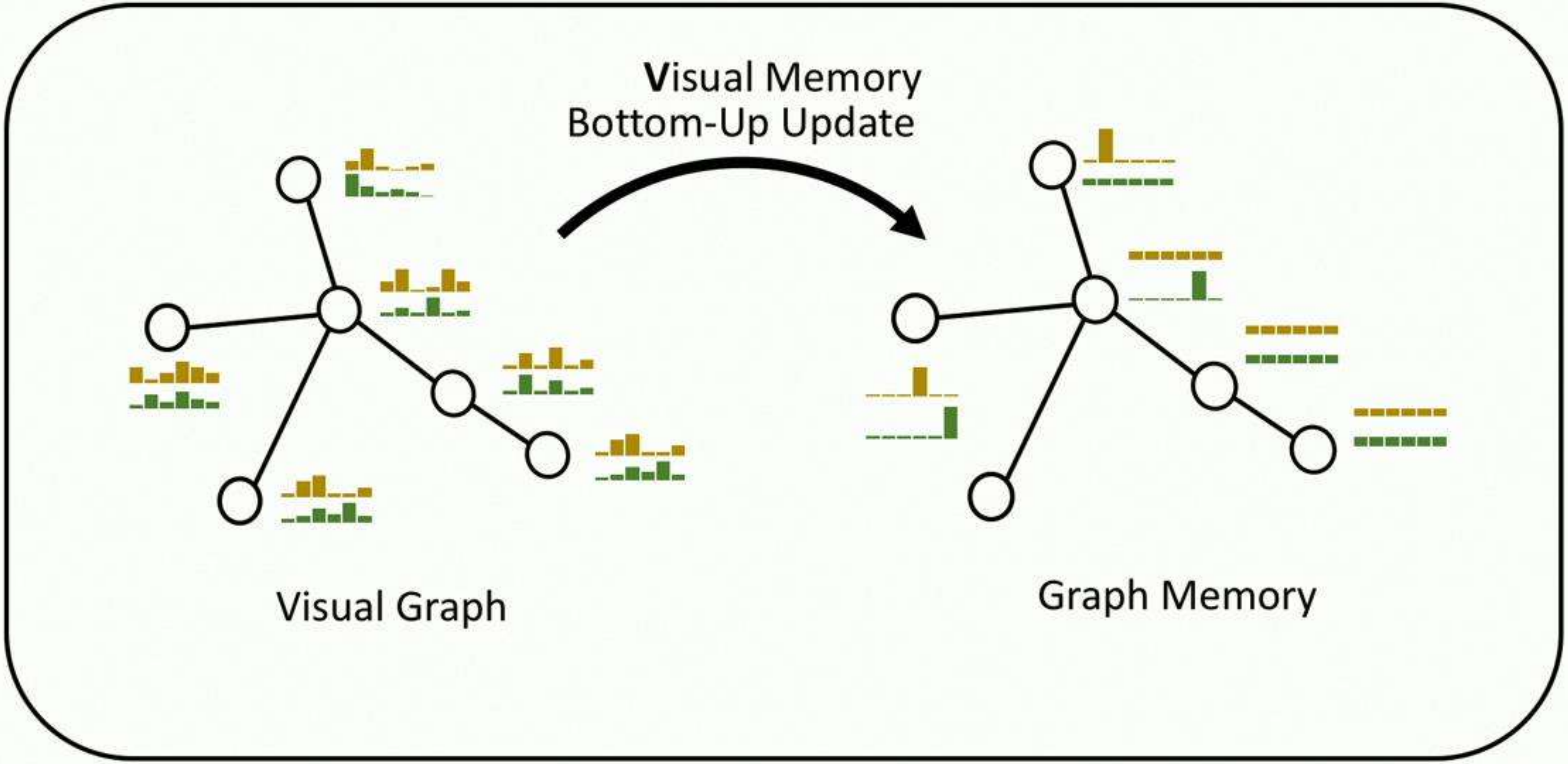


Slide Credit: Stefan Lee

Agent Architecture

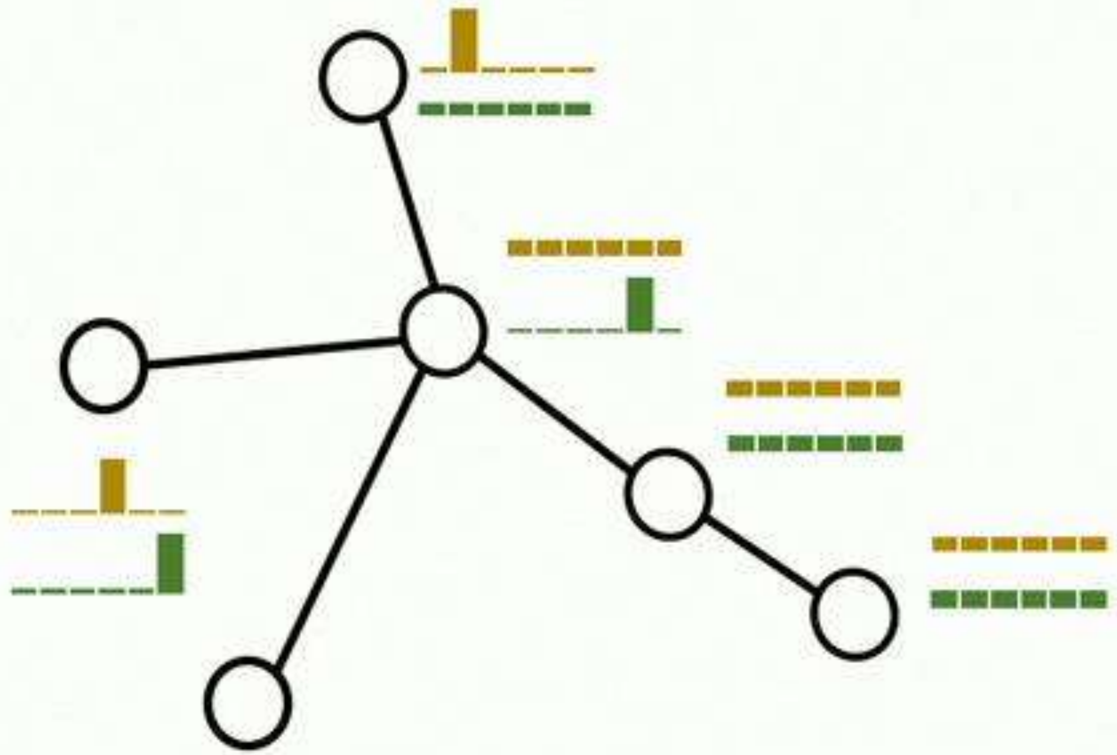


Bottom-up Update



Top-down Update

Visual Memory

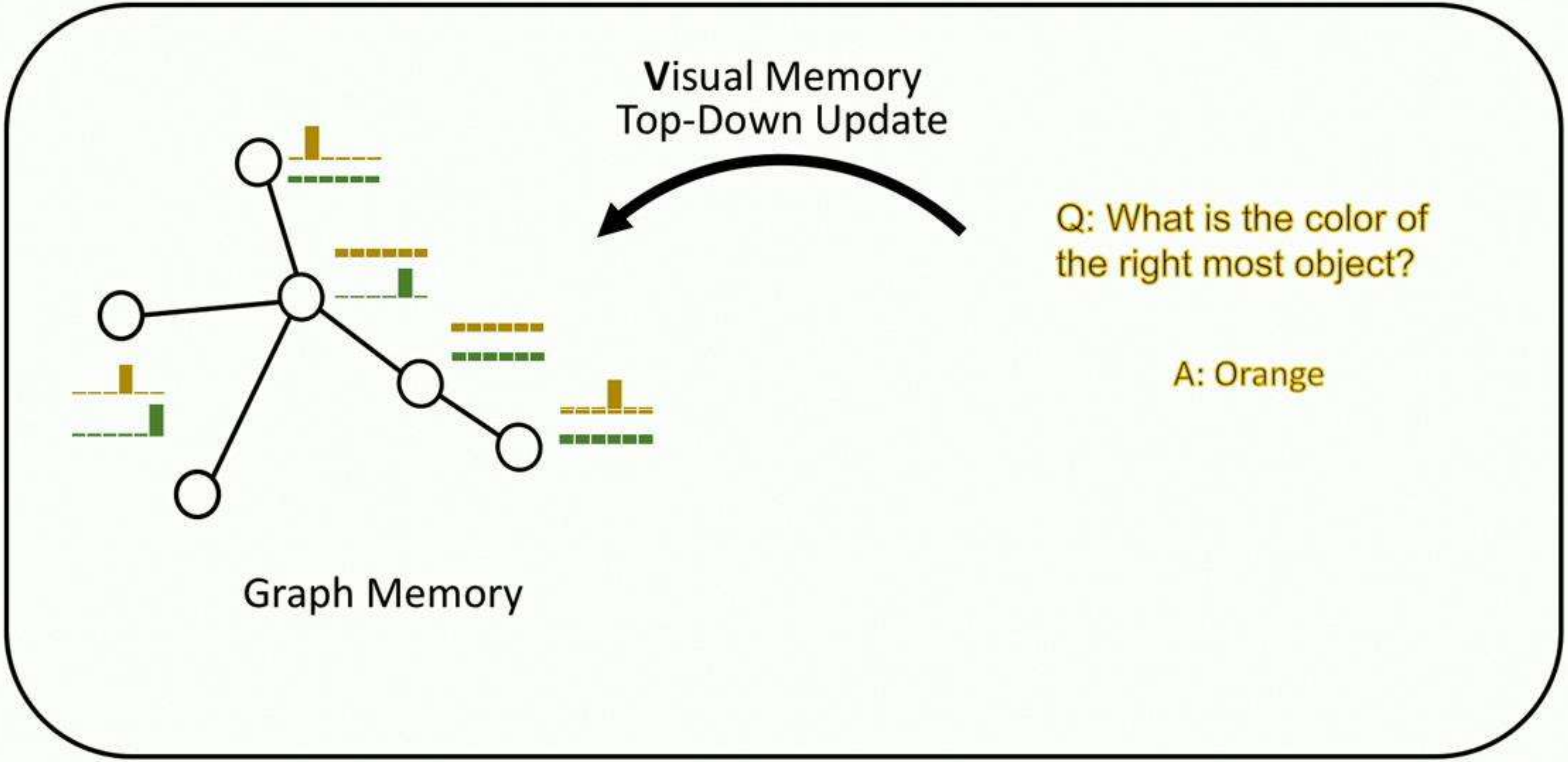


Memory Graph

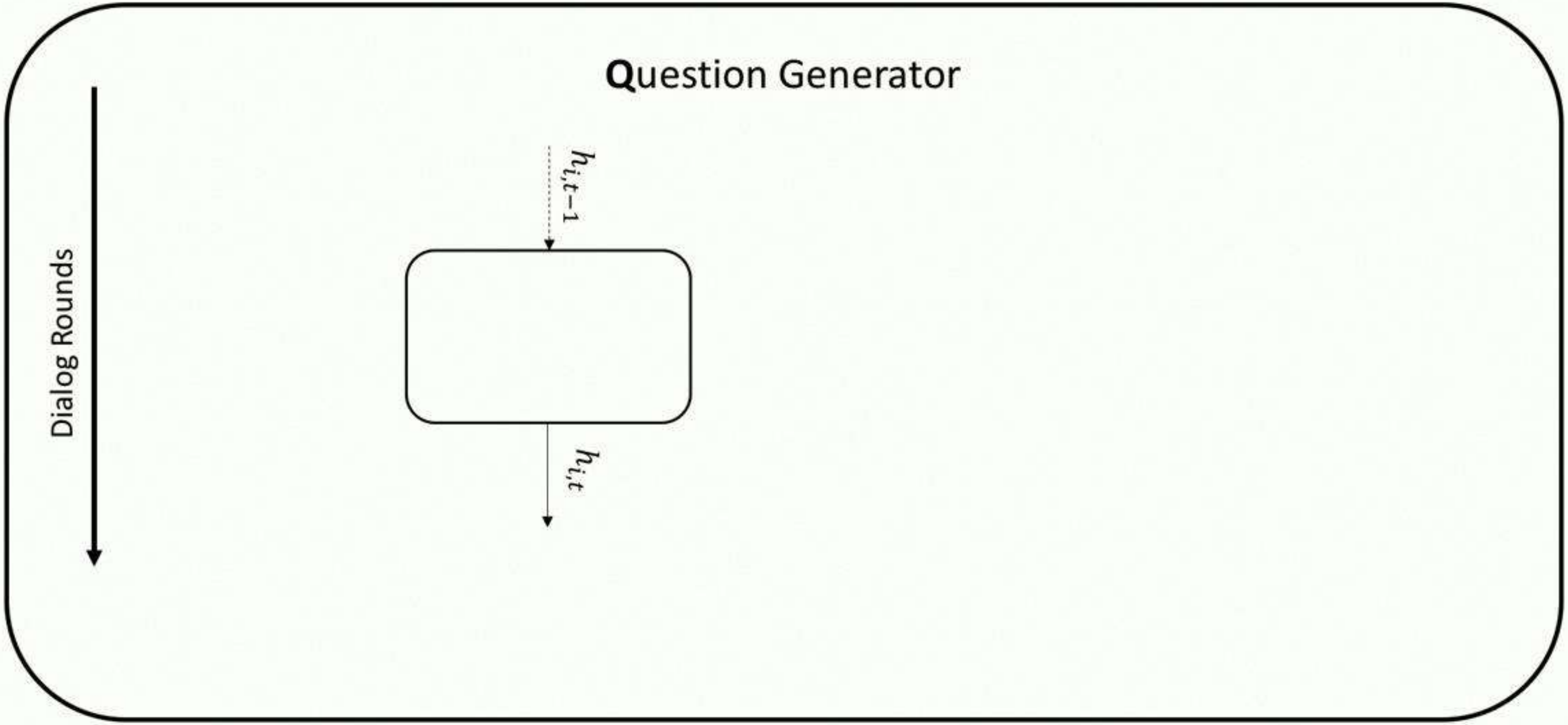
Q: What is the color of the right most object?

A: Orange

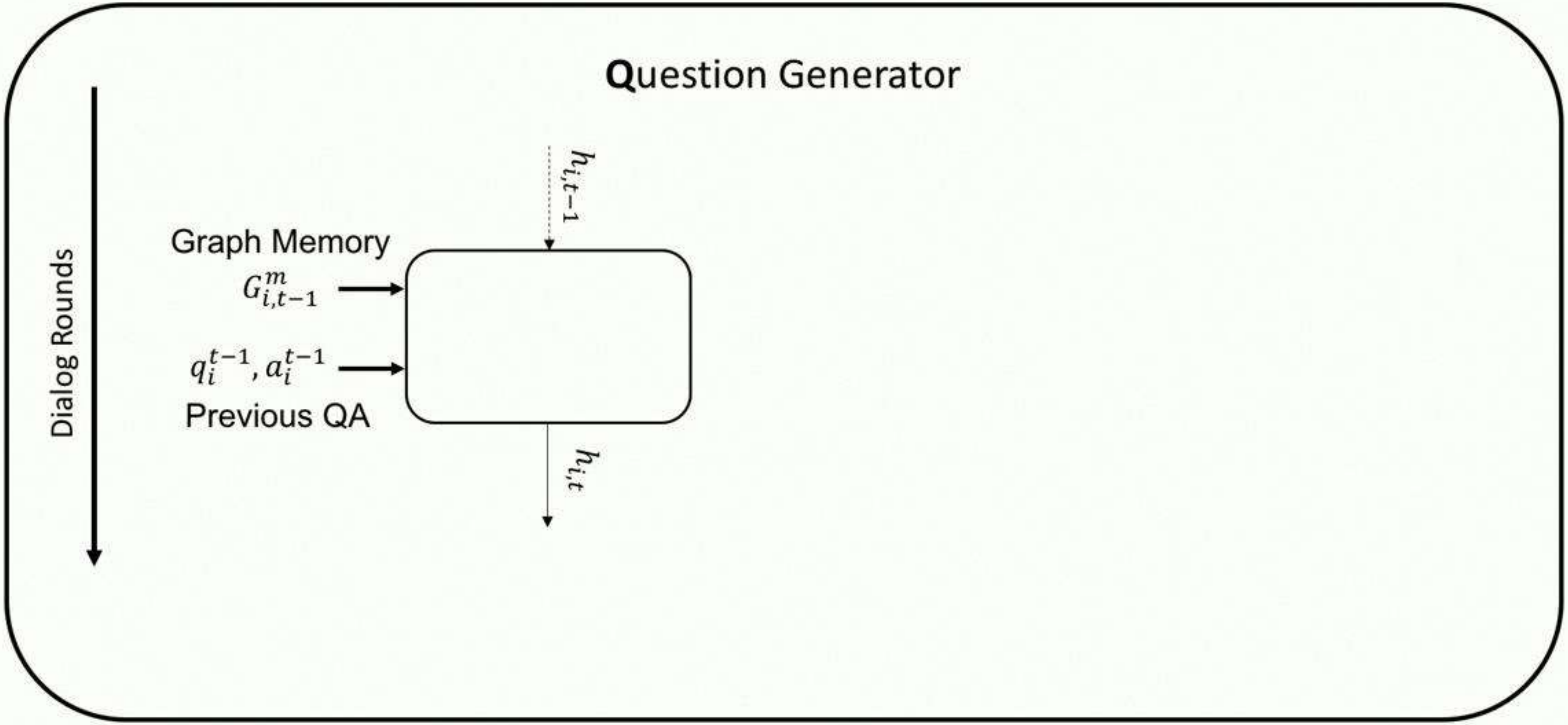
Top-down Update



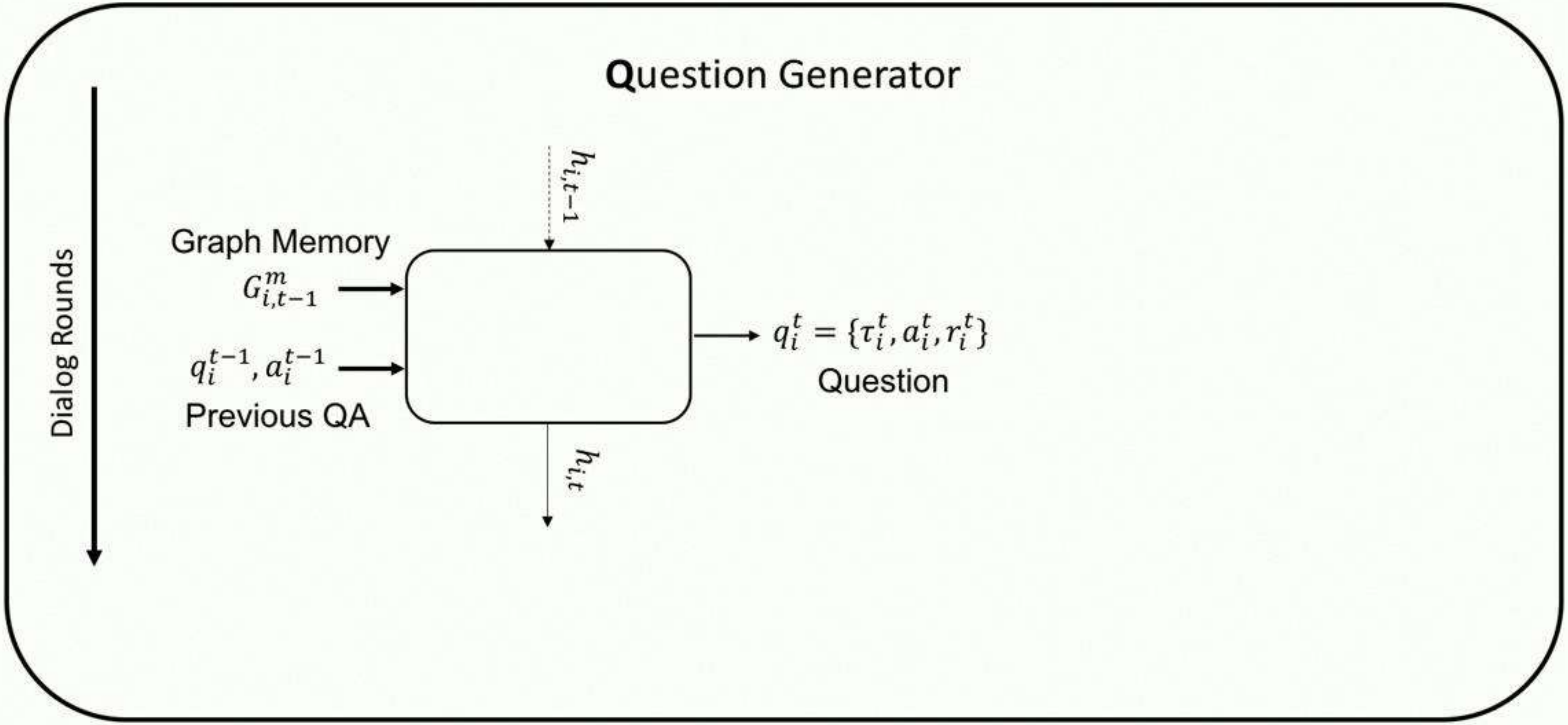
Agent Architecture



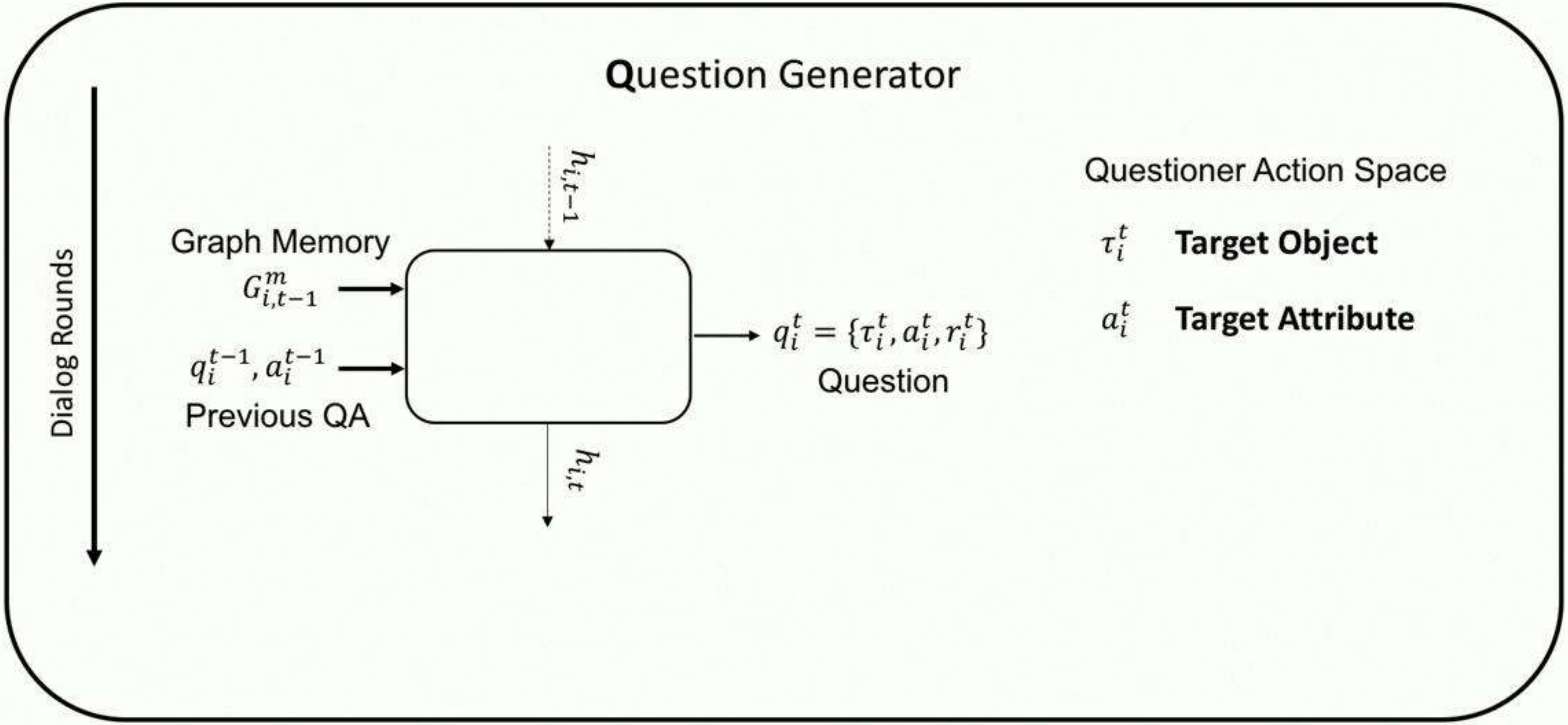
Agent Architecture



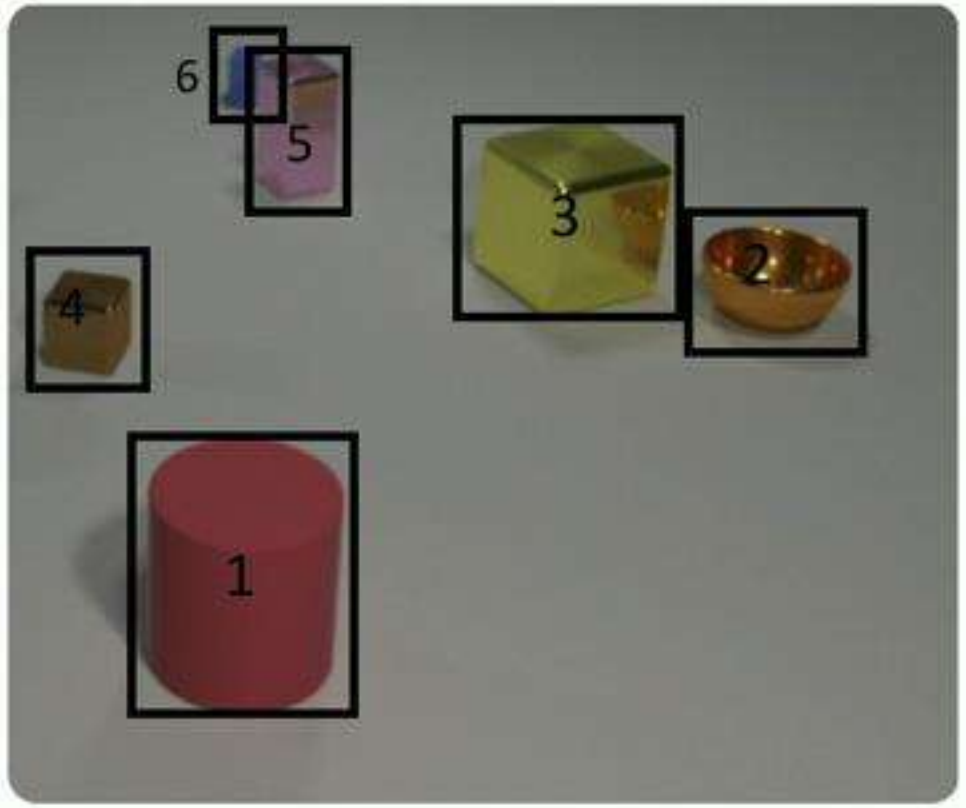
Agent Architecture

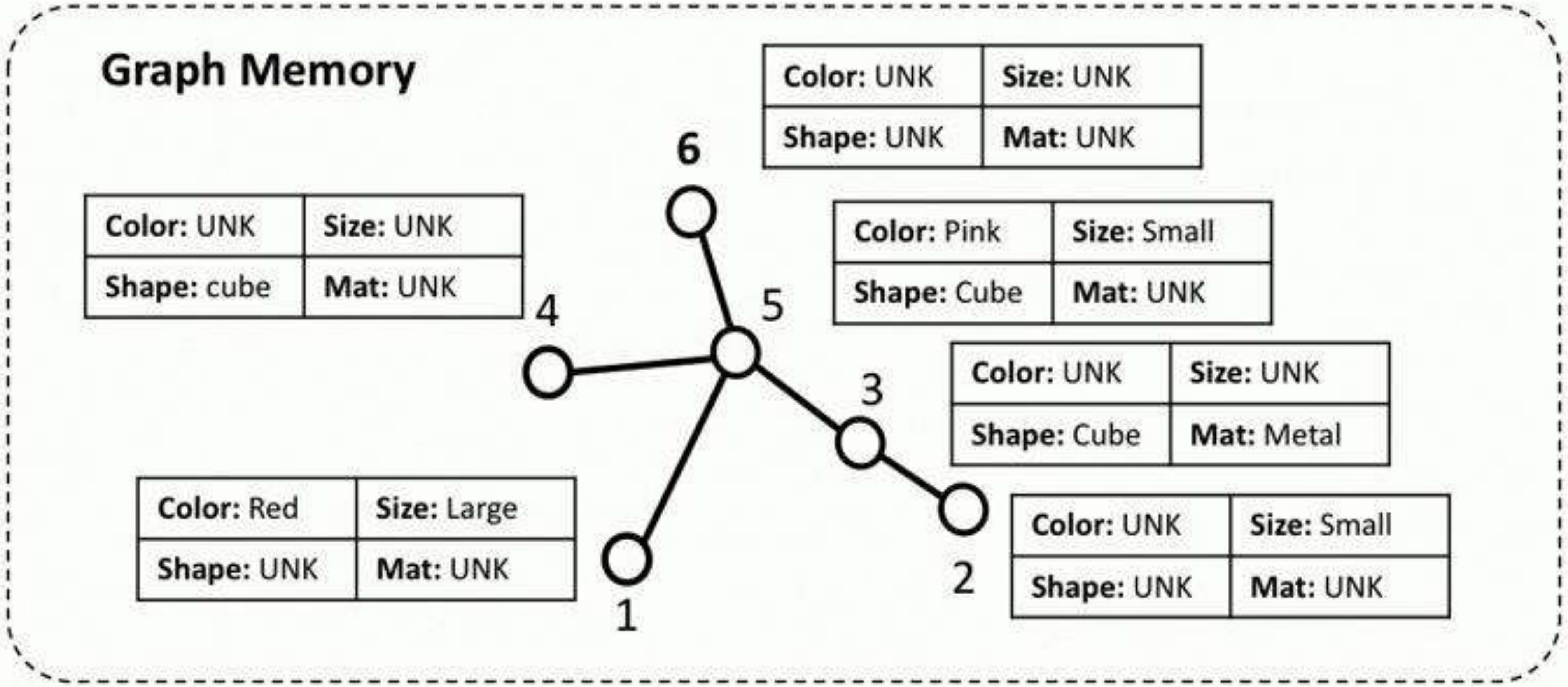


Agent Architecture

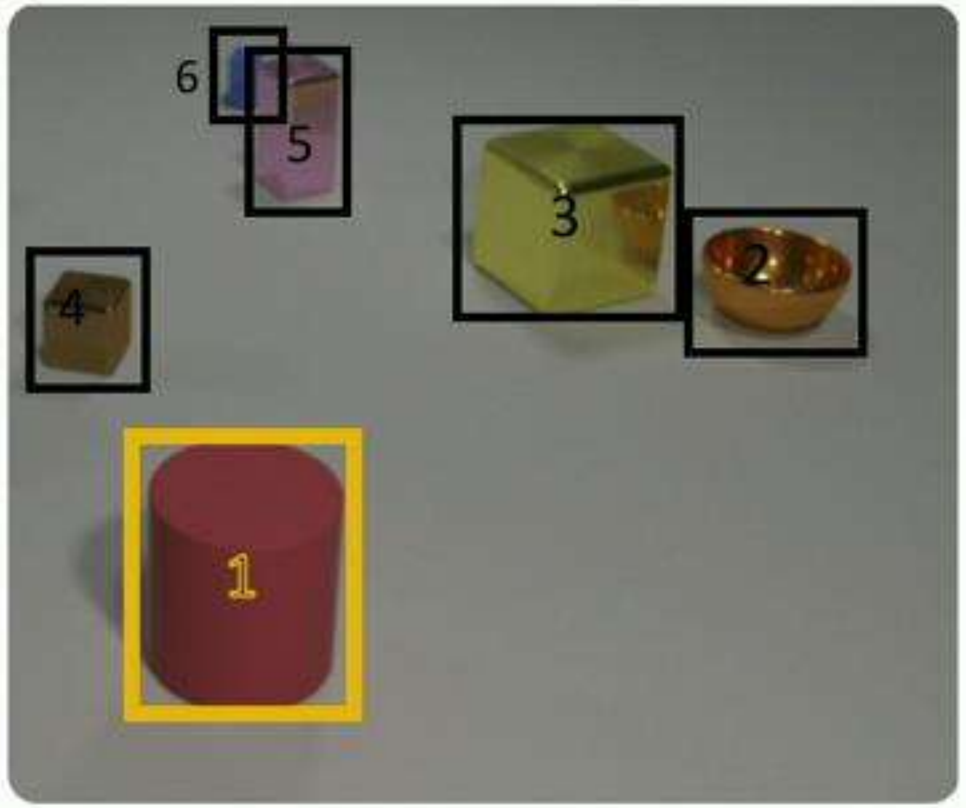


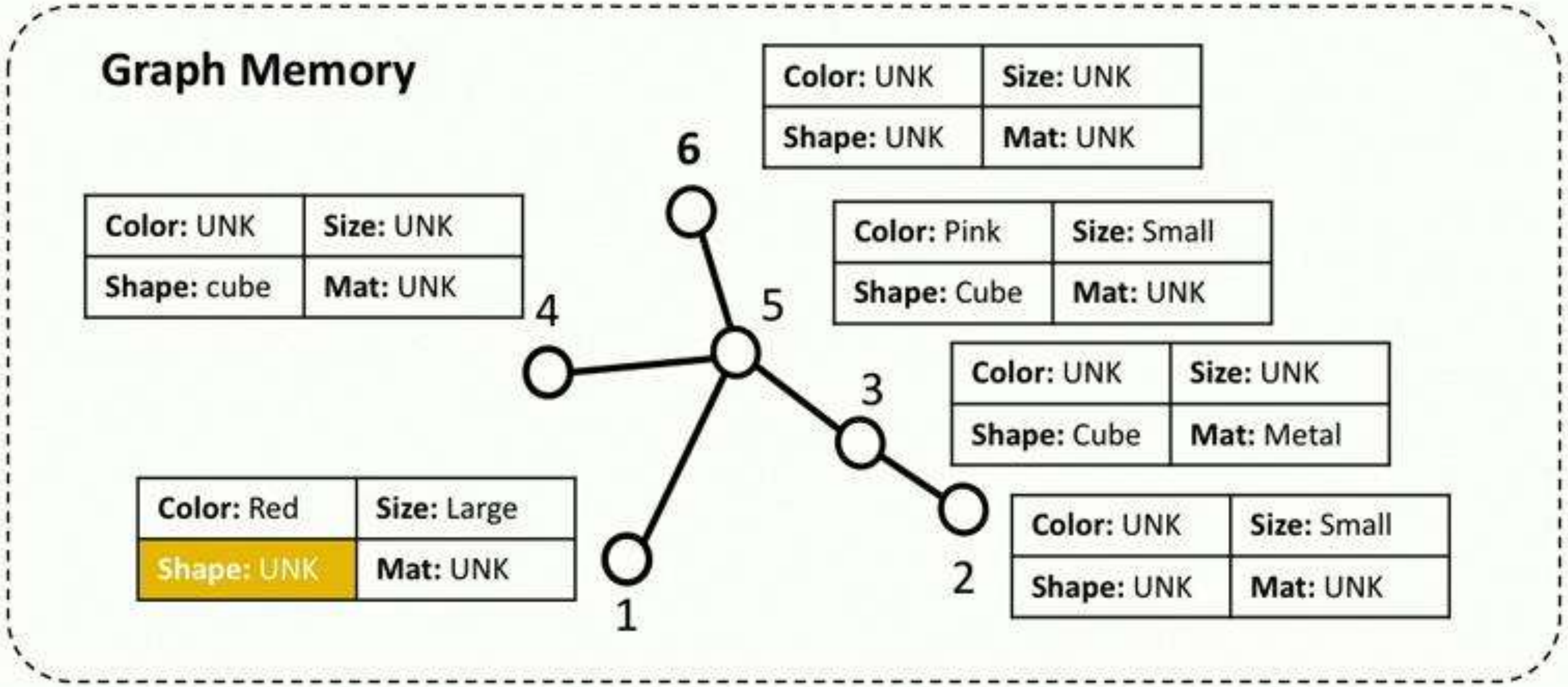
Question Template





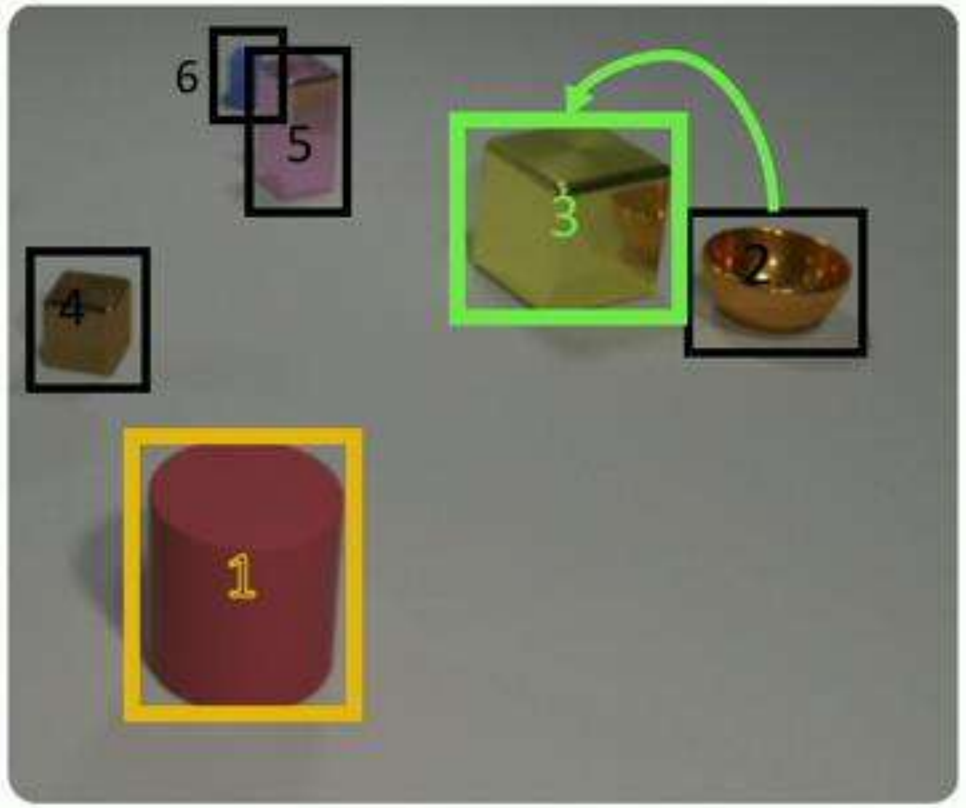
Question Template

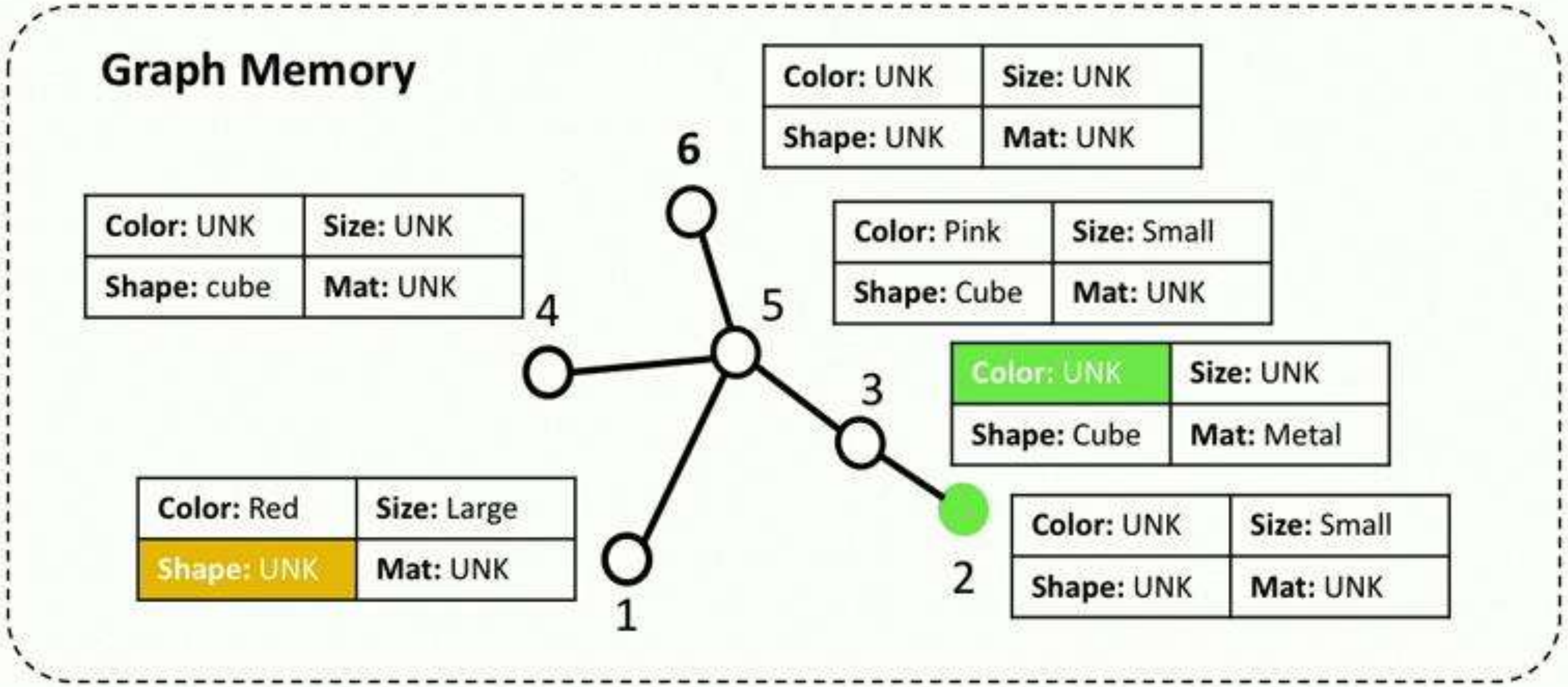




Target	1
Attribute	Shape
Reference	None
Question	What is the shape of the front most large red object?

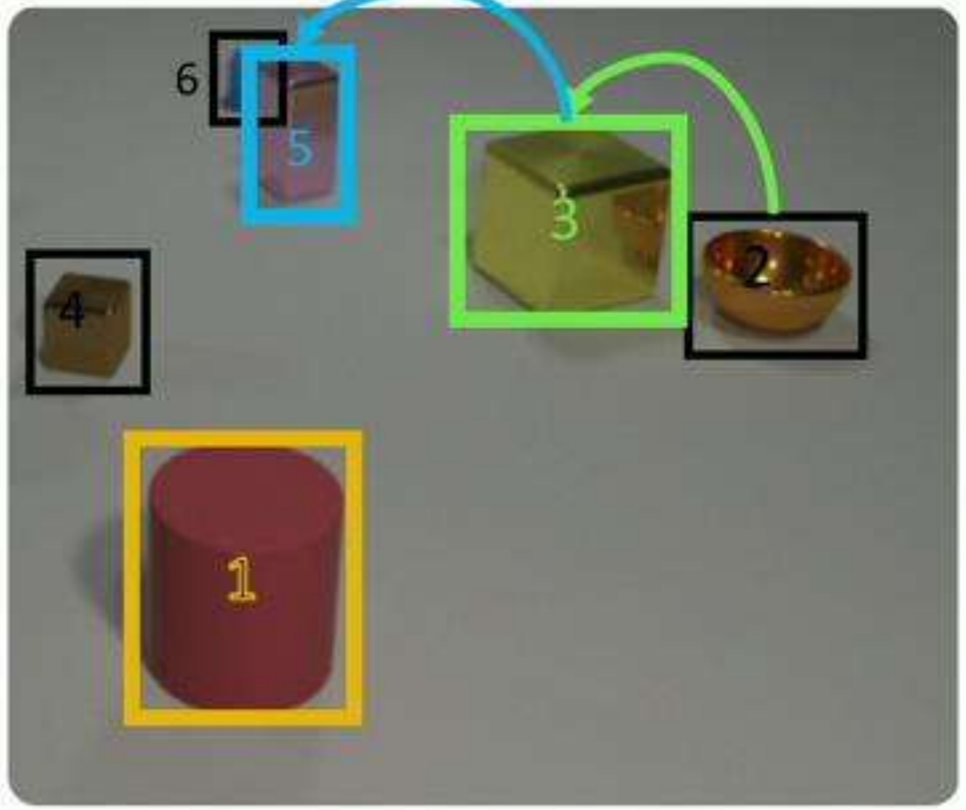
Question Template

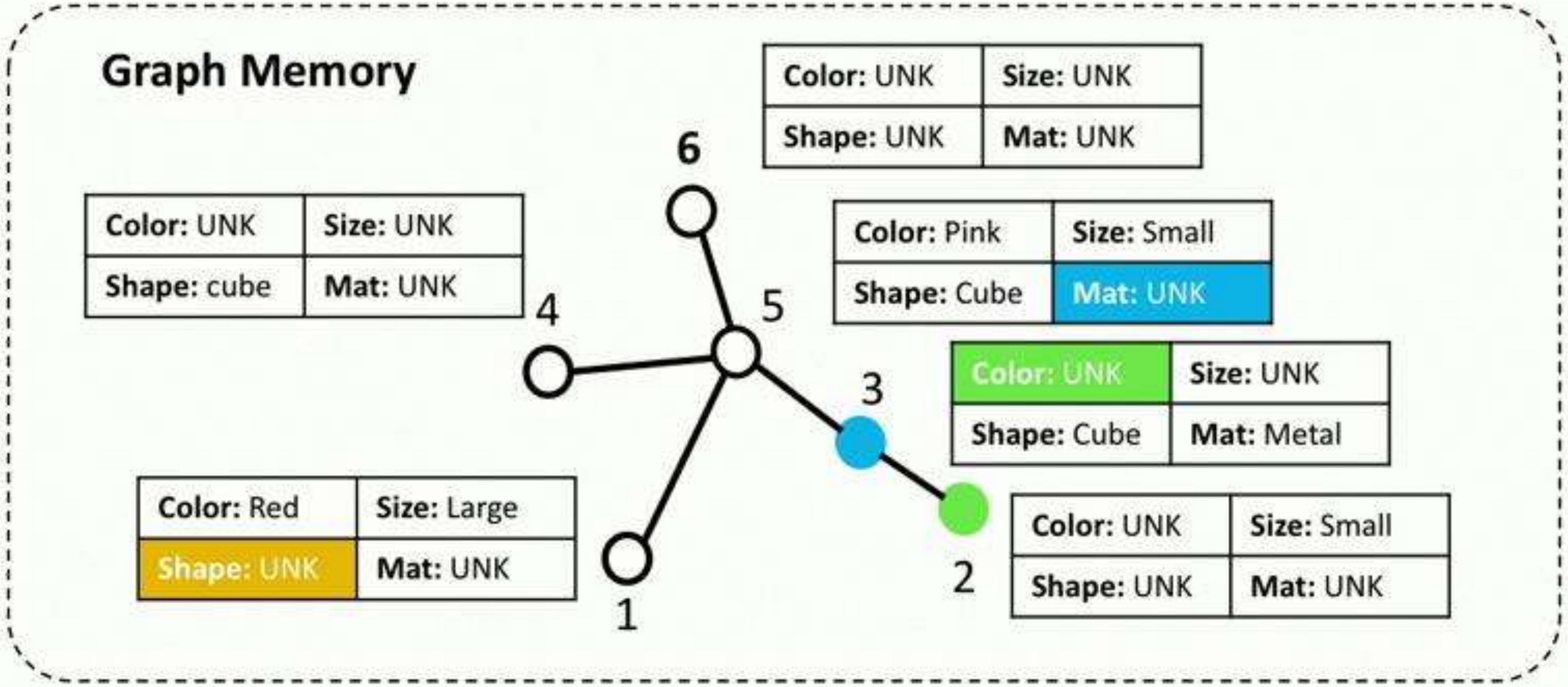




Target	1	3
Attribute	Shape	Color
Reference	None	2
Question	What is the shape of the front most large red object?	What is the color of the metal cube on the left side of a small object?

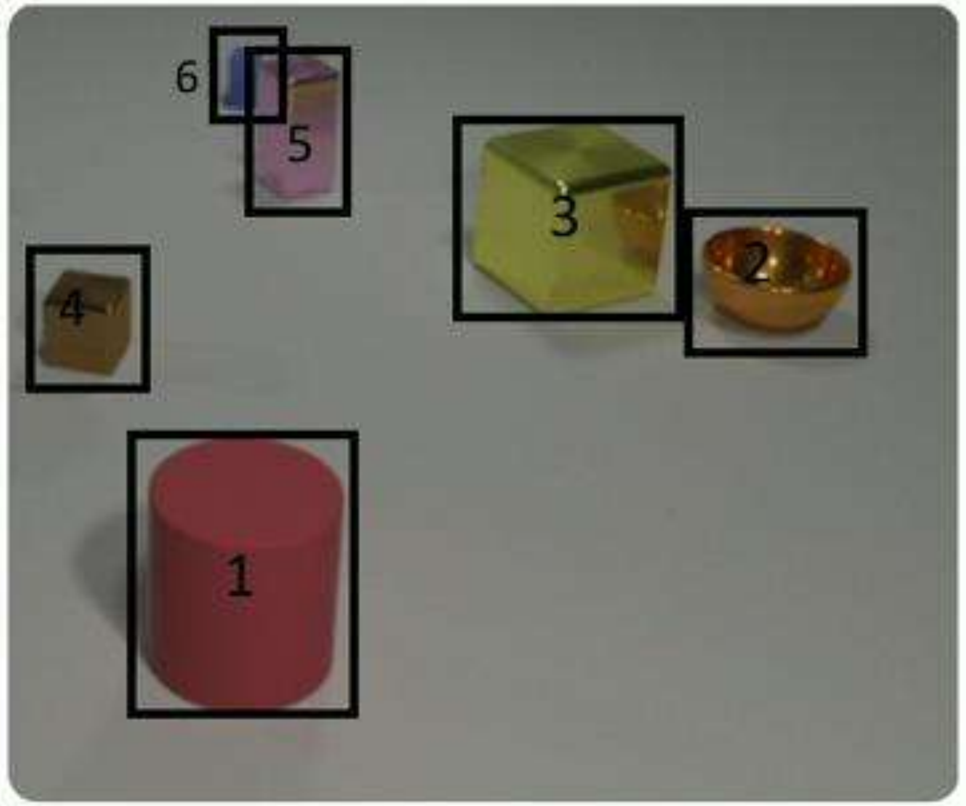
Question Template

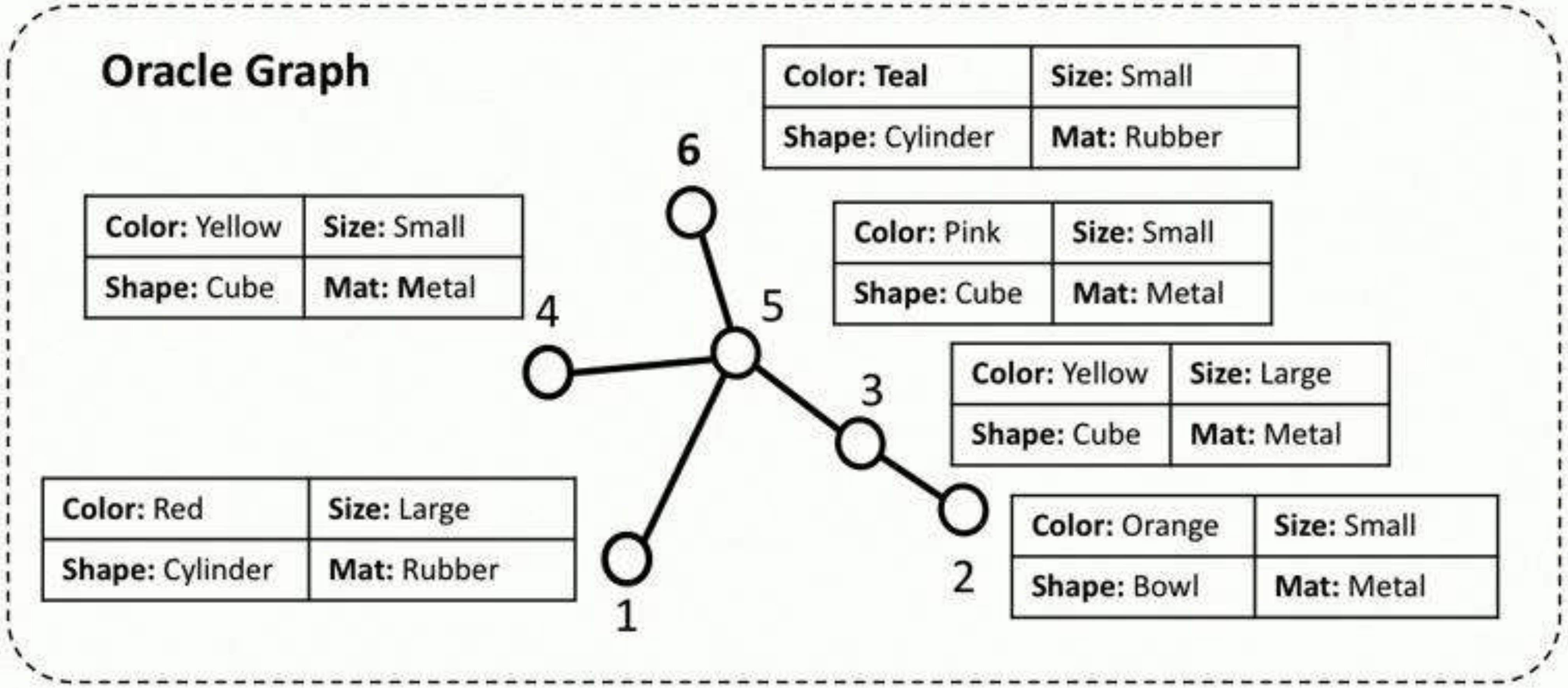




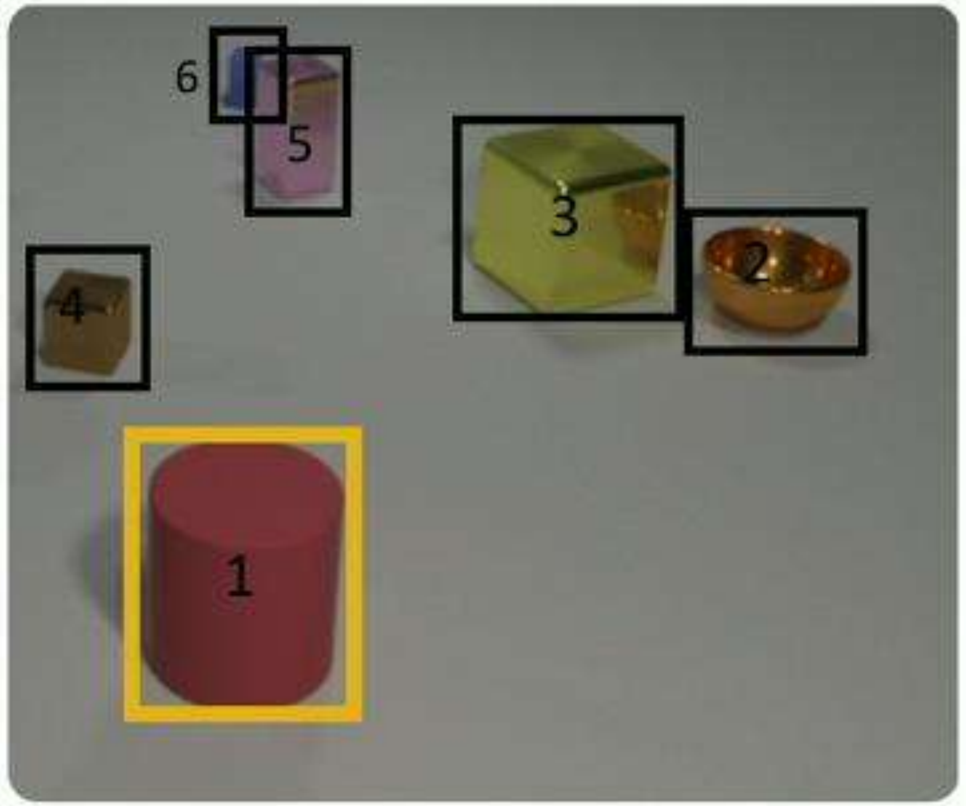
Target	1	3	5
Attribute	Shape	Color	Material
Reference	None	2	3
Question	What is the shape of the front most large red object?	What is the color of the metal cube on the left side of a small object?	What is the material of object at left side of metal cube?

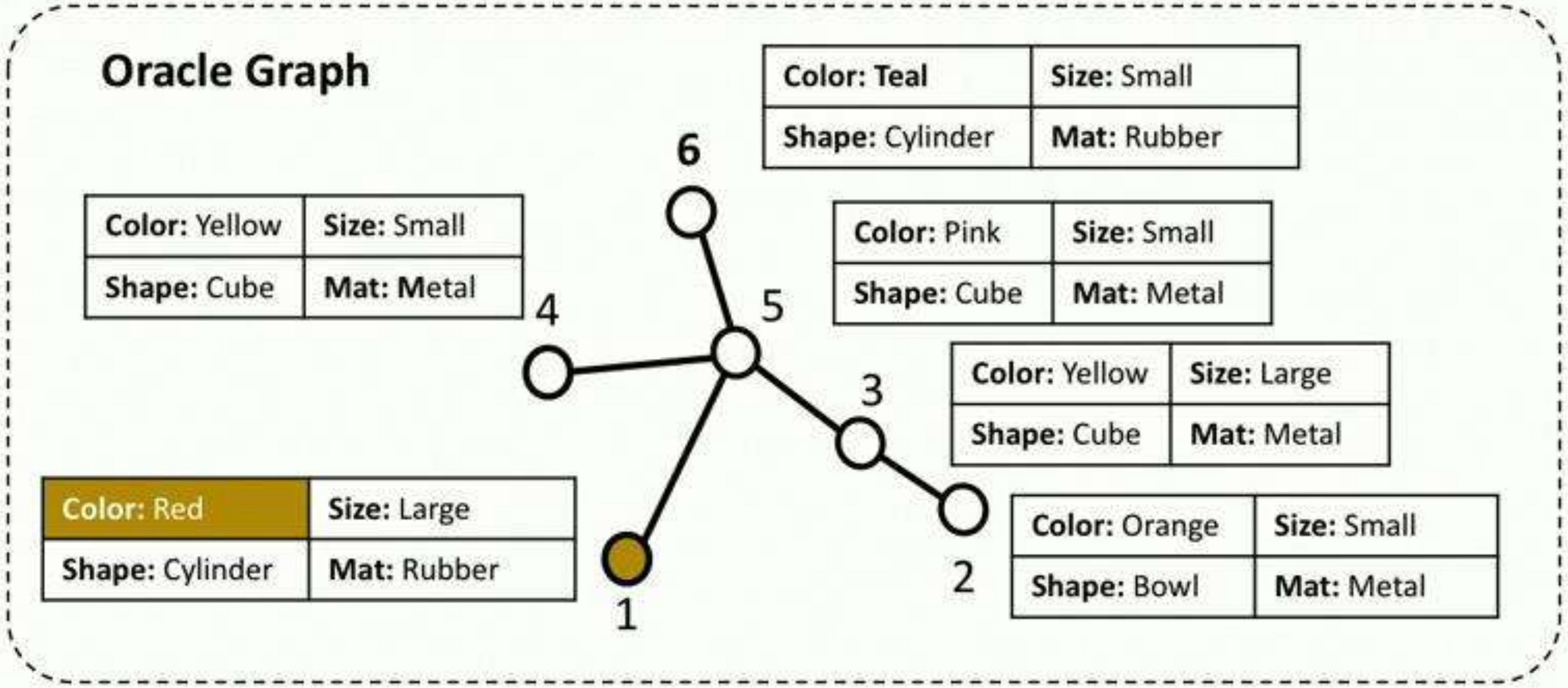
Oracle Question Answering





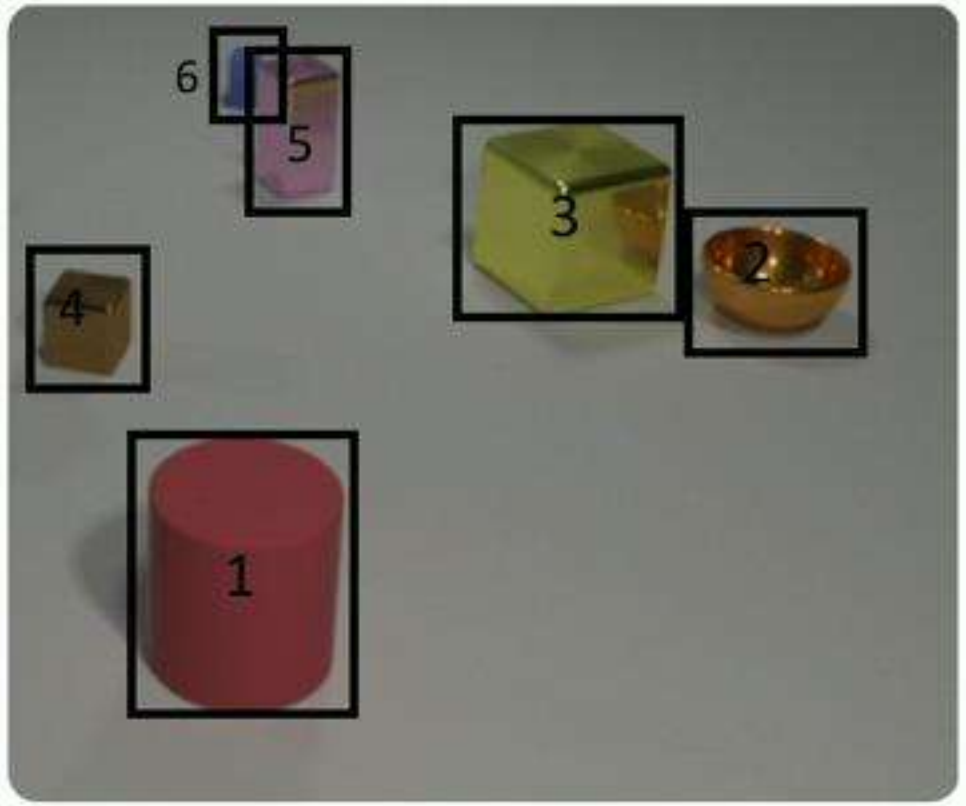
Oracle Question Answering

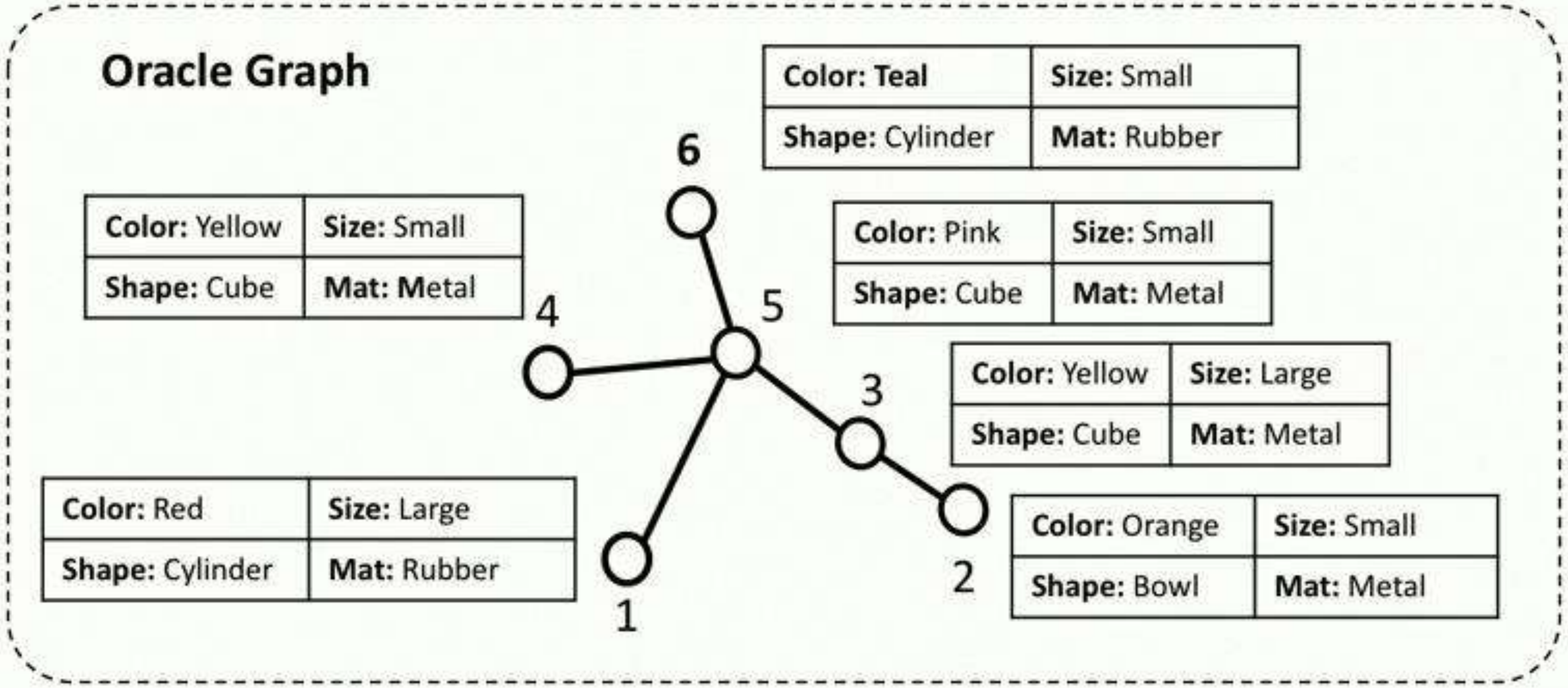




Q: What is the color of the front most object?

Oracle Question Answering

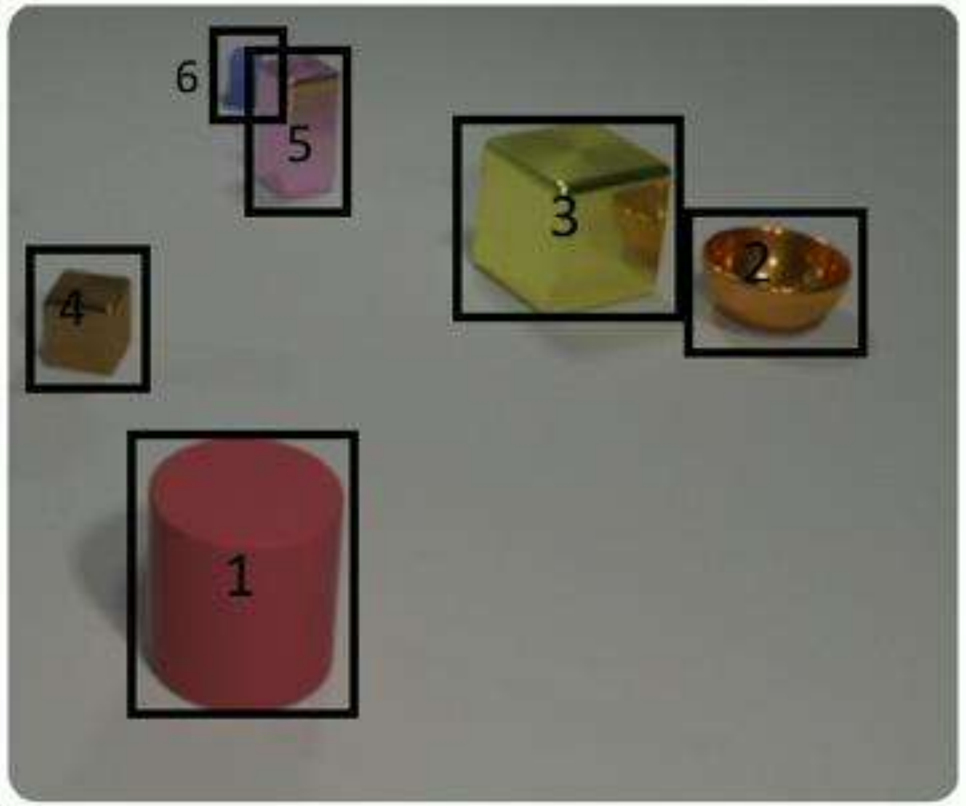


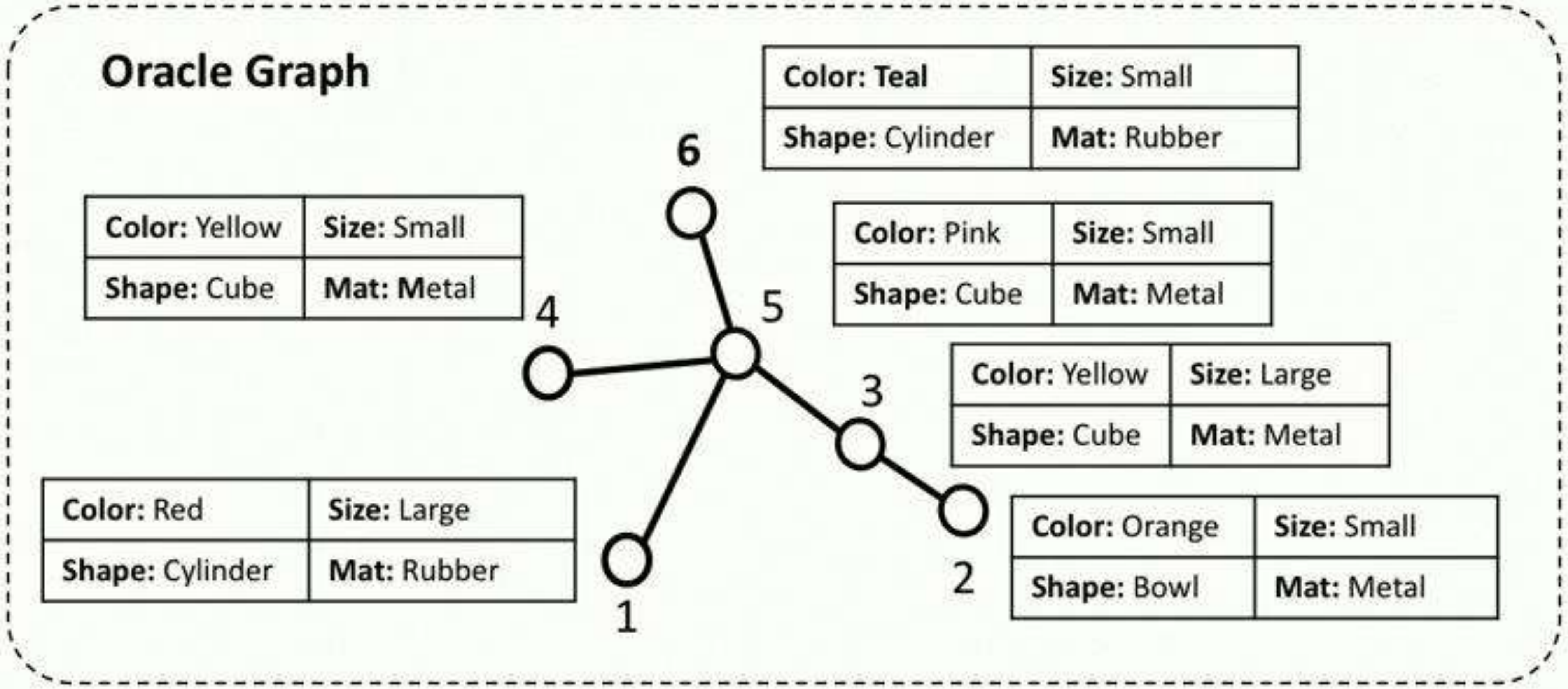


Q: What is the color of the front most object? A: Red

Q: What is the shape of the object left of the green object?

Oracle Question Answering

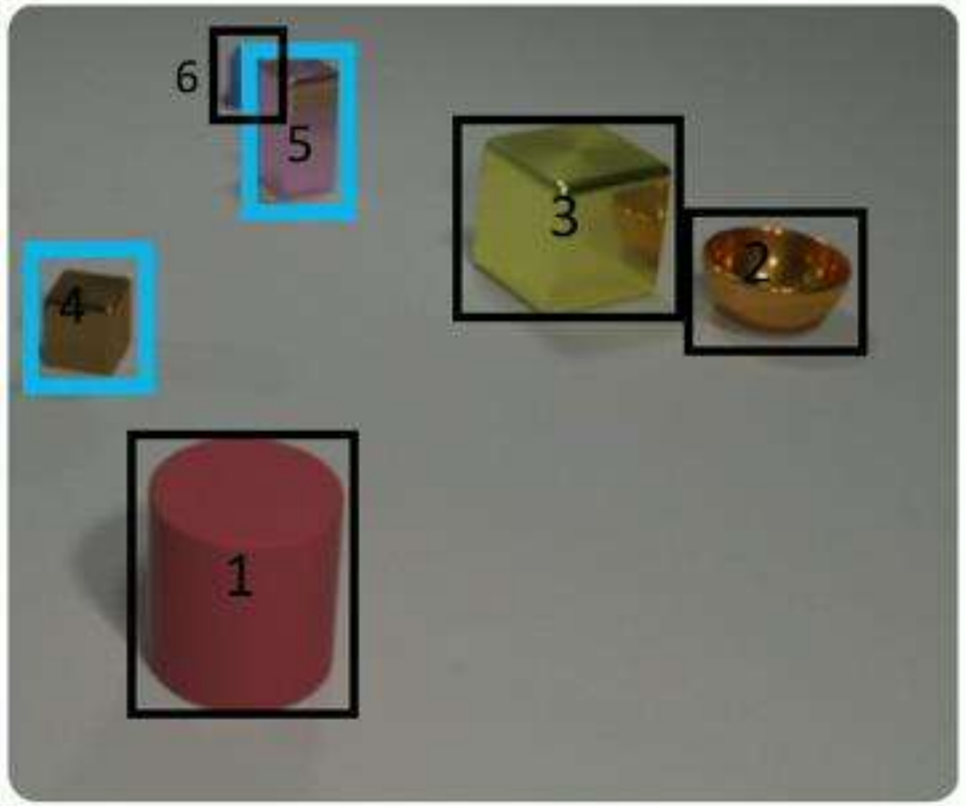


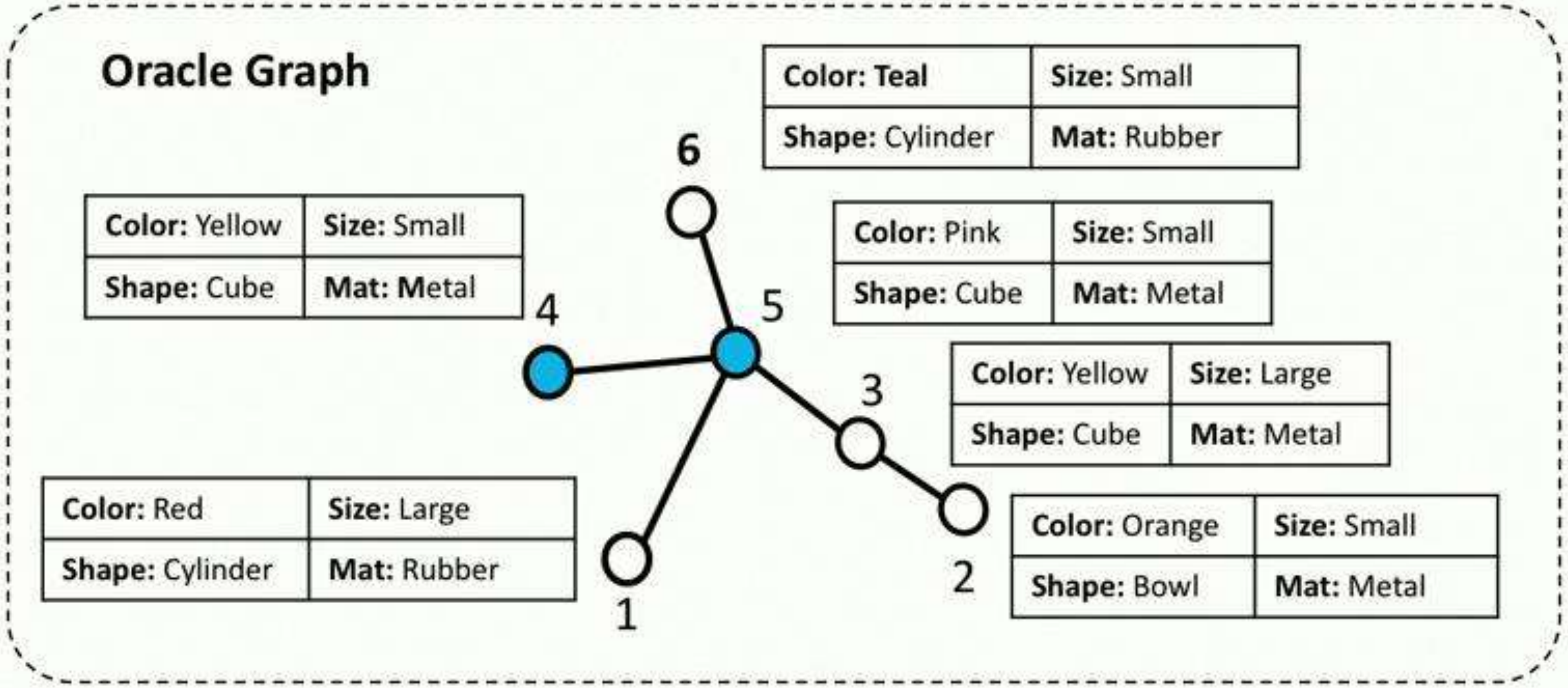


Q: What is the color of the front most object? A: Red

Q: What is the shape of the object left of the green object? A: ⟨ Invalid ⟩

Oracle Question Answering



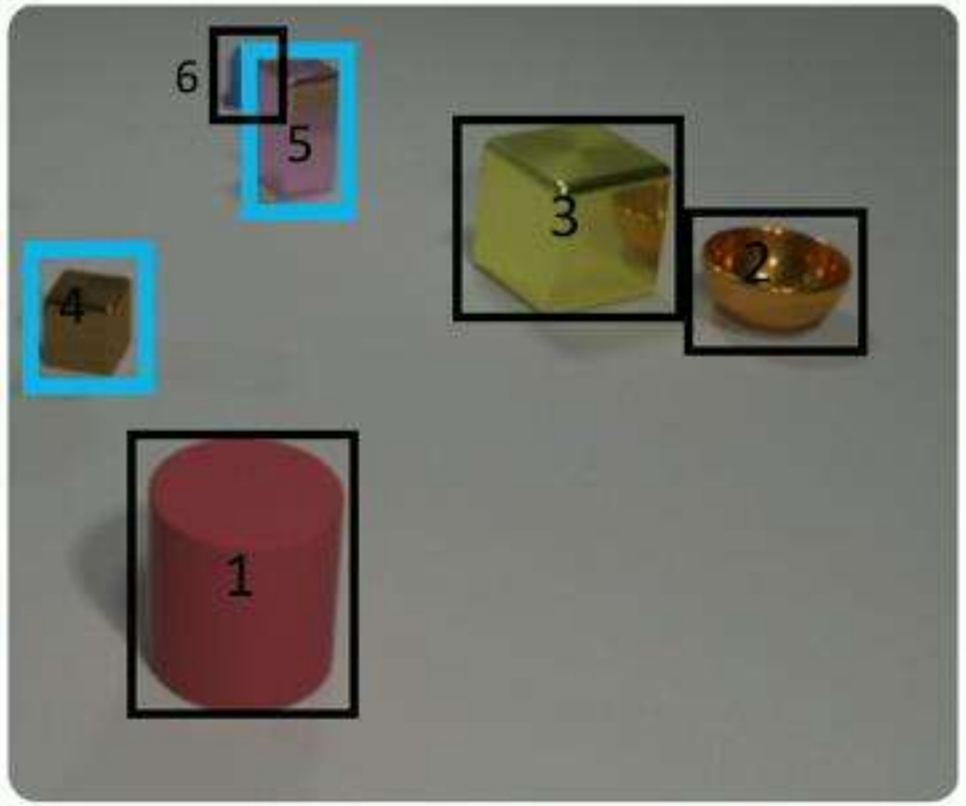


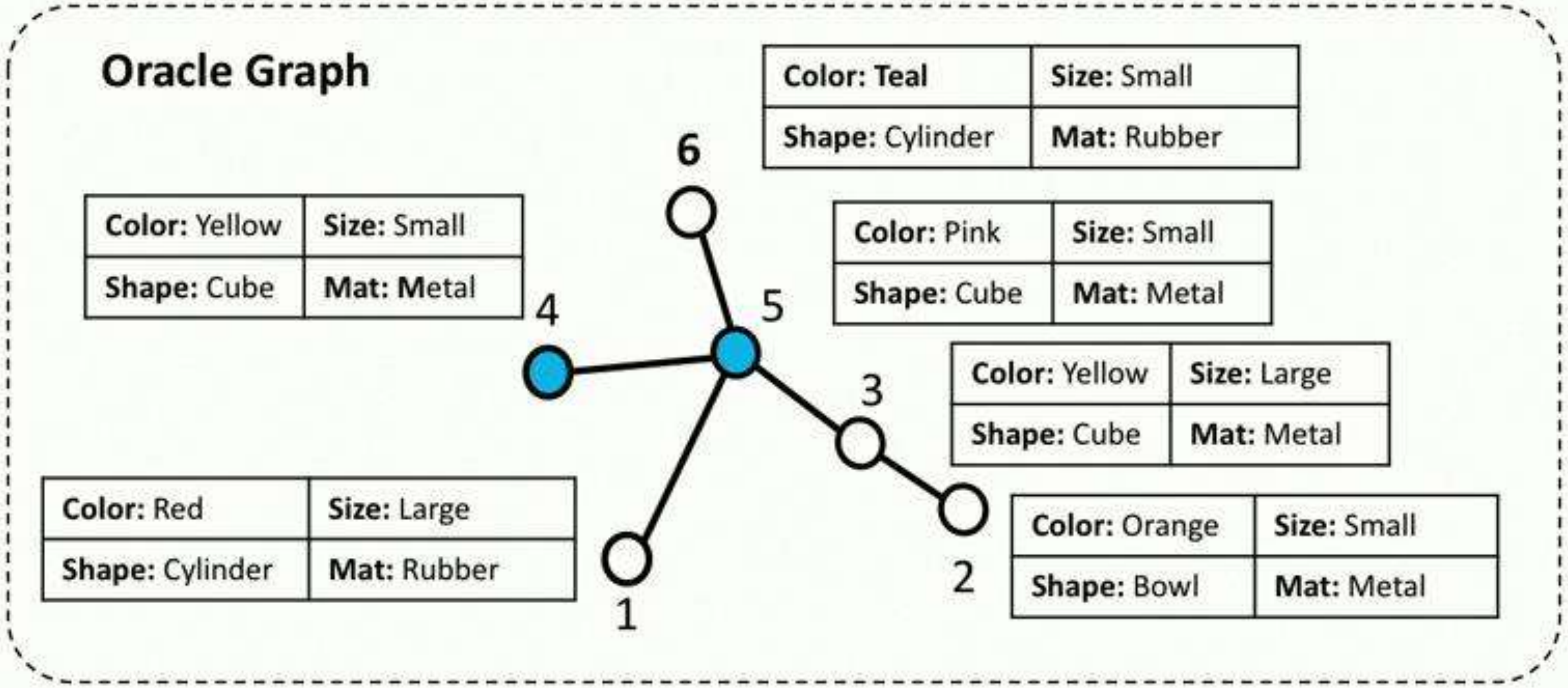
Q: What is the color of the front most object? A: Red

Q: What is the shape of the object left of the green object? A: ⟨ Invalid ⟩

Q: What is color is the small cube?

Oracle Question Answering





Q: What is the color of the front most object? A: Red

Q: What is the shape of the object left of the green object? A: ⟨ Invalid ⟩

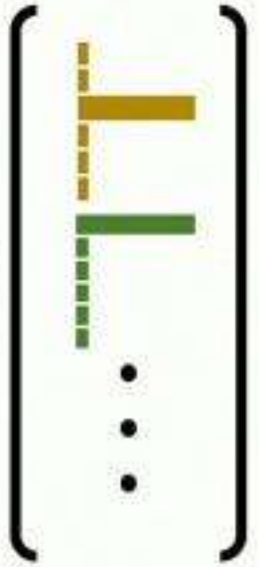
Q: What is color is the small cube? A: ⟨ Ambiguous ⟩

Training Objective: Visual System

Cross-Entropy loss between the graph memory and the visual predictions over all images, objects, and attributes

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

Visual attribute predictions



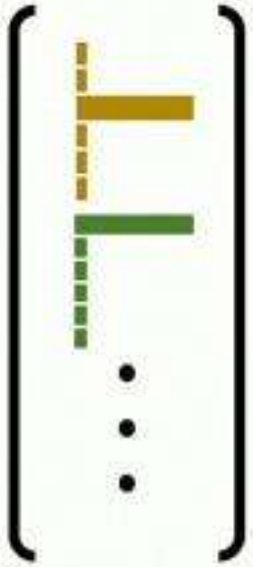
Graph memory

Training Objective: Visual System

Cross-Entropy loss between the graph memory and the visual predictions over all images, objects, and attributes

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

Visual attribute predictions



Graph memory

Update visual system after each dialog with stochastic gradient descent

Training Objective: Questioner Policy

Train the questioner to maximize expected reward over images and dialog rounds in an episode

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

Training Objective: Questioner Policy

Train the questioner to maximize expected reward over images and dialog rounds in an episode

$$\theta_\pi^* = \arg\max \underbrace{E_V E_{I\sim\mathcal{E}}\, E_{\pi_q}}_{\substack{\text{Expectation over visual systems,} \\ \text{episodes, and questions}}} \left[\overbrace{\sum_{i=1}^{n} \sum_{t=1}^{T} r_i^t (q_i^t \sim \pi_q (h_i^t ; \theta_\pi)}^{\substack{\text{Sum of reward over all} \\ \text{images and dialog rounds}}} \right]$$

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

Training Objective: Questioner Policy Reward

Per round change in graph memory accuracy

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

Training Objective: Questioner Policy

Train the questioner to maximize expected reward over images and dialog rounds in an episode

$$\theta_\pi^* = \arg\max E_V E_{I\sim\mathcal{E}} E_{\pi_q} \left[\underbrace{\sum_{i=1}^{n} \sum_{t=1}^{T} r_i^t (q_i^t \sim \pi_q(h_i^t; \theta_\pi)}_{\text{Sum of reward over all images and dialog rounds}} \right]$$

Training Objective: Questioner Policy Reward

Per round change in graph memory accuracy

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

Training Objective: Questioner Policy Reward

Per round change in graph memory accuracy

Oracle Graph

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

Training Objective: Questioner Policy Reward

Per round change in graph memory accuracy

Oracle Graph

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

Graph Memory
at dialog round ***t***

Graph Memory
at dialog round ***t-1***

Training Objective: Questioner Policy Reward

Per round change in graph memory accuracy

Oracle Graph

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

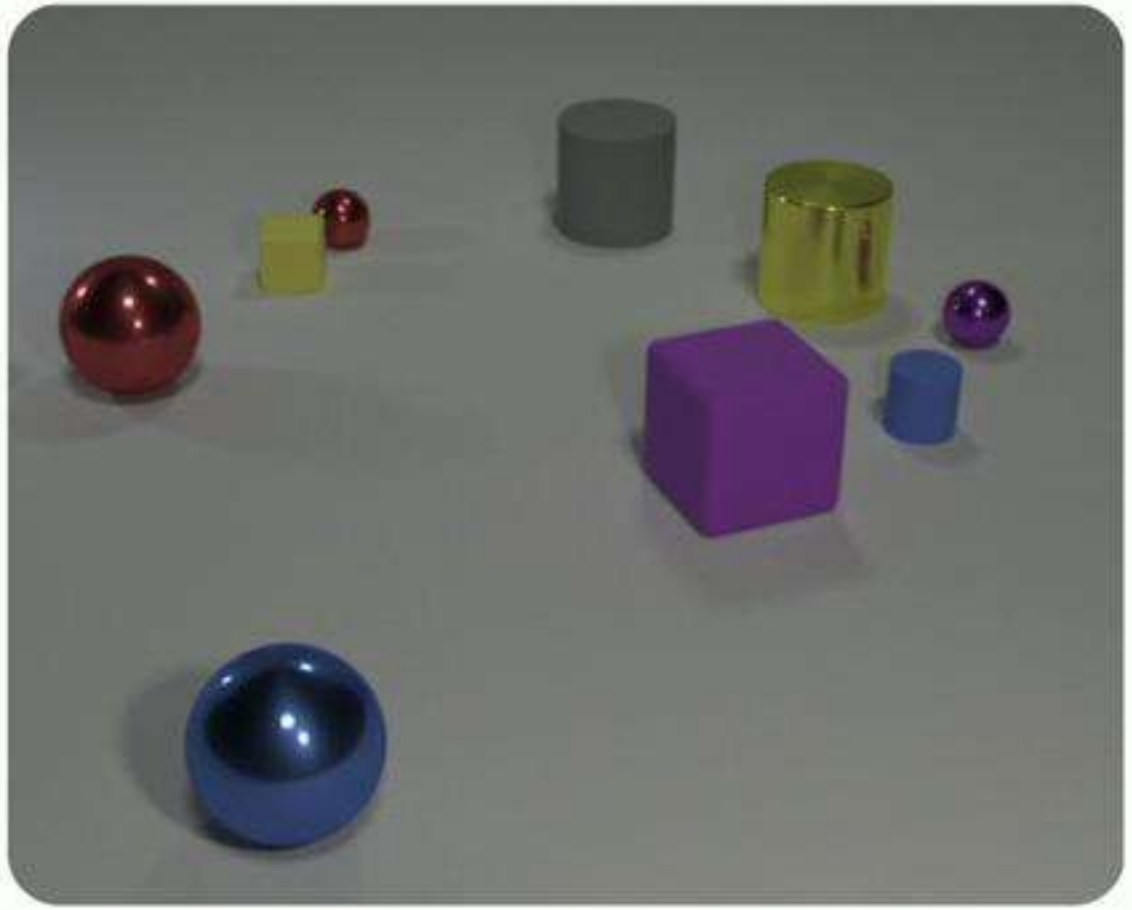
Graph Memory at dialog round ***t***

Graph Memory at dialog round ***t-1***

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.

Experiments: Baselines

Baseline Heuristic Questioners:

Target object

Target attribute

Reference object

Experiments: Baselines

Baseline Heuristic Questioners:

	Random
Target object	Uniform
Target attribute	Uniform
Reference object	Uniform

Experiments: Baselines

Baseline Heuristic Questioners:

	Random	Entropy
Target object	Uniform	Highest entropy
Target attribute	Uniform	Highest entropy
Reference object	Uniform	Lowest entropy

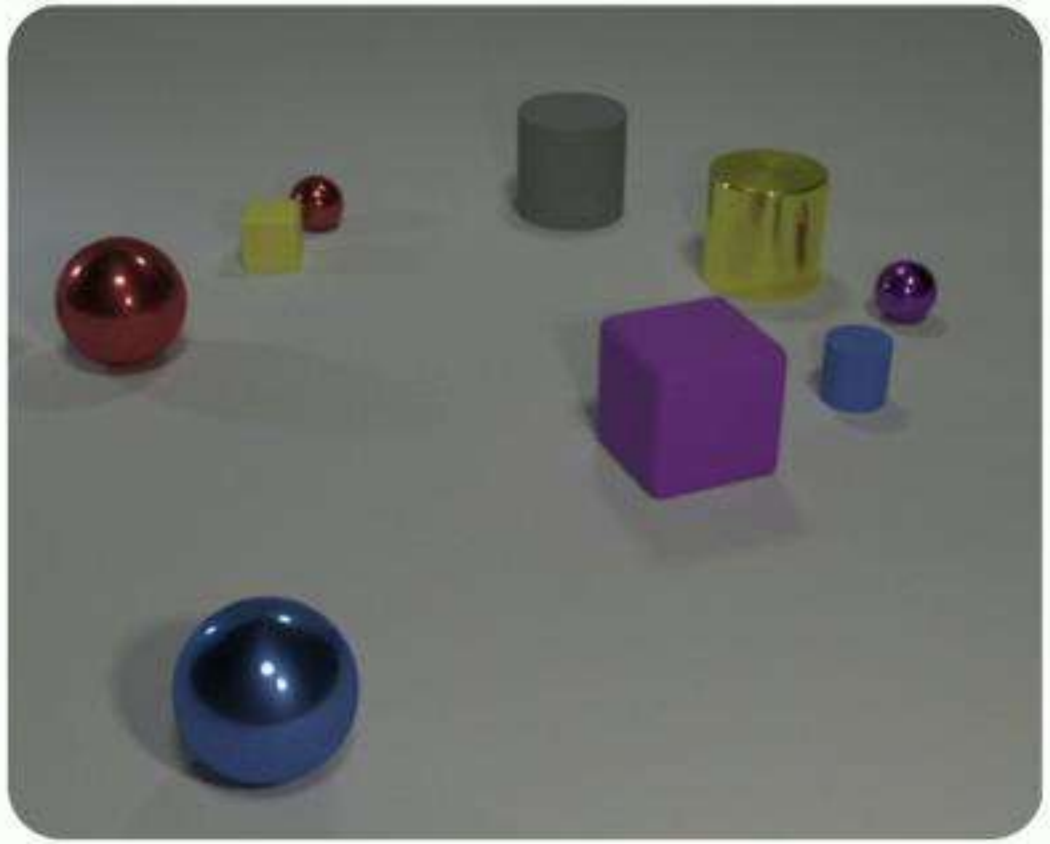
Experiments: Baselines

Baseline Heuristic Questioners:

	Random	Entropy	Entropy + Context
Target object	Uniform	Highest entropy	Highest Entropy + Spatial
Target attribute	Uniform	Highest entropy	Highest Entropy
Reference object	Uniform	Lowest entropy	Lowest Entropy + Spatial

Experiment: Standard Training

Standard Dataset



Shapes: cube, sphere, cylinder

Colors: gray, red, blue, green, yellow, purple

Materials: rubber, metal

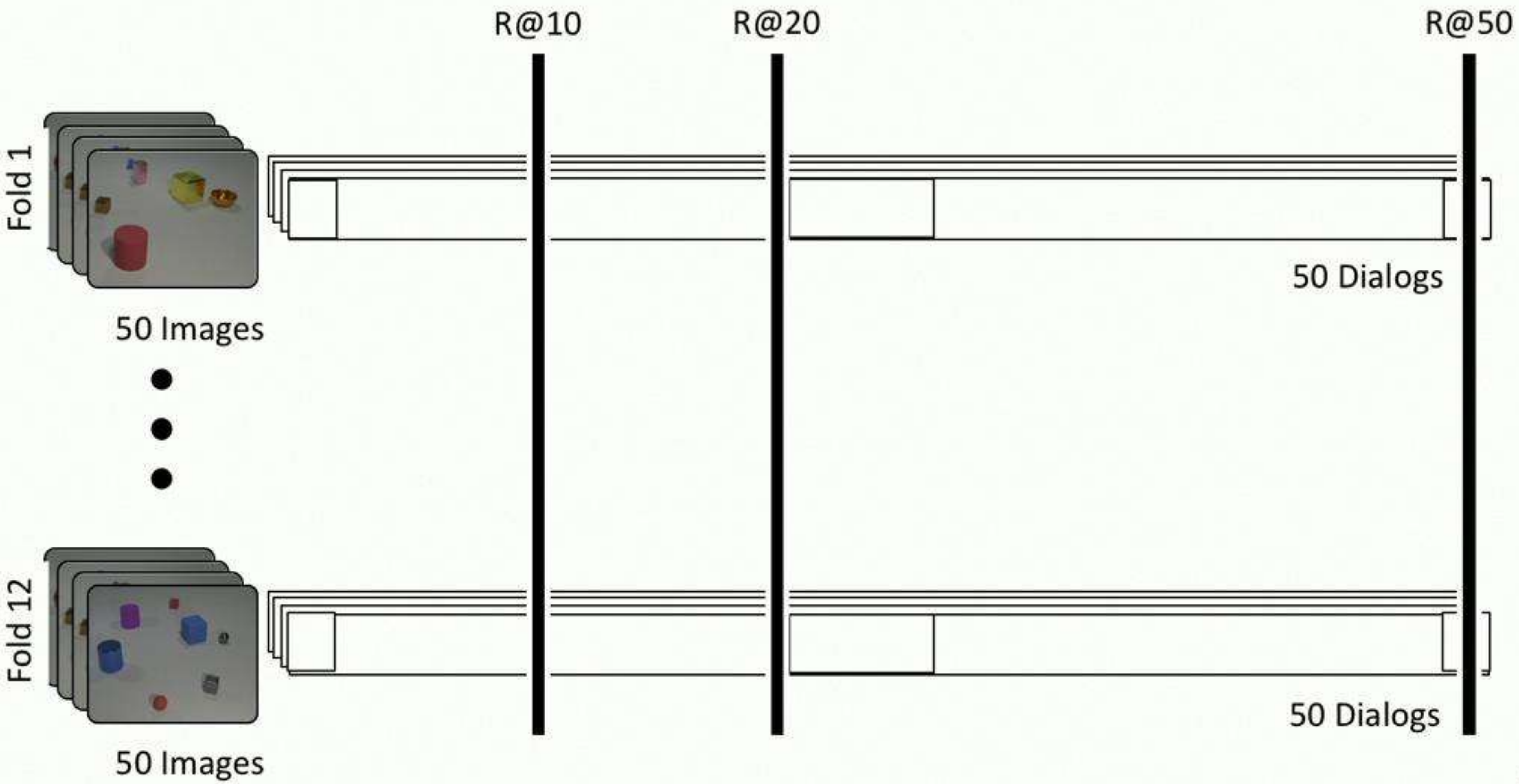
Sizes: small, large

#Objects: 5-10

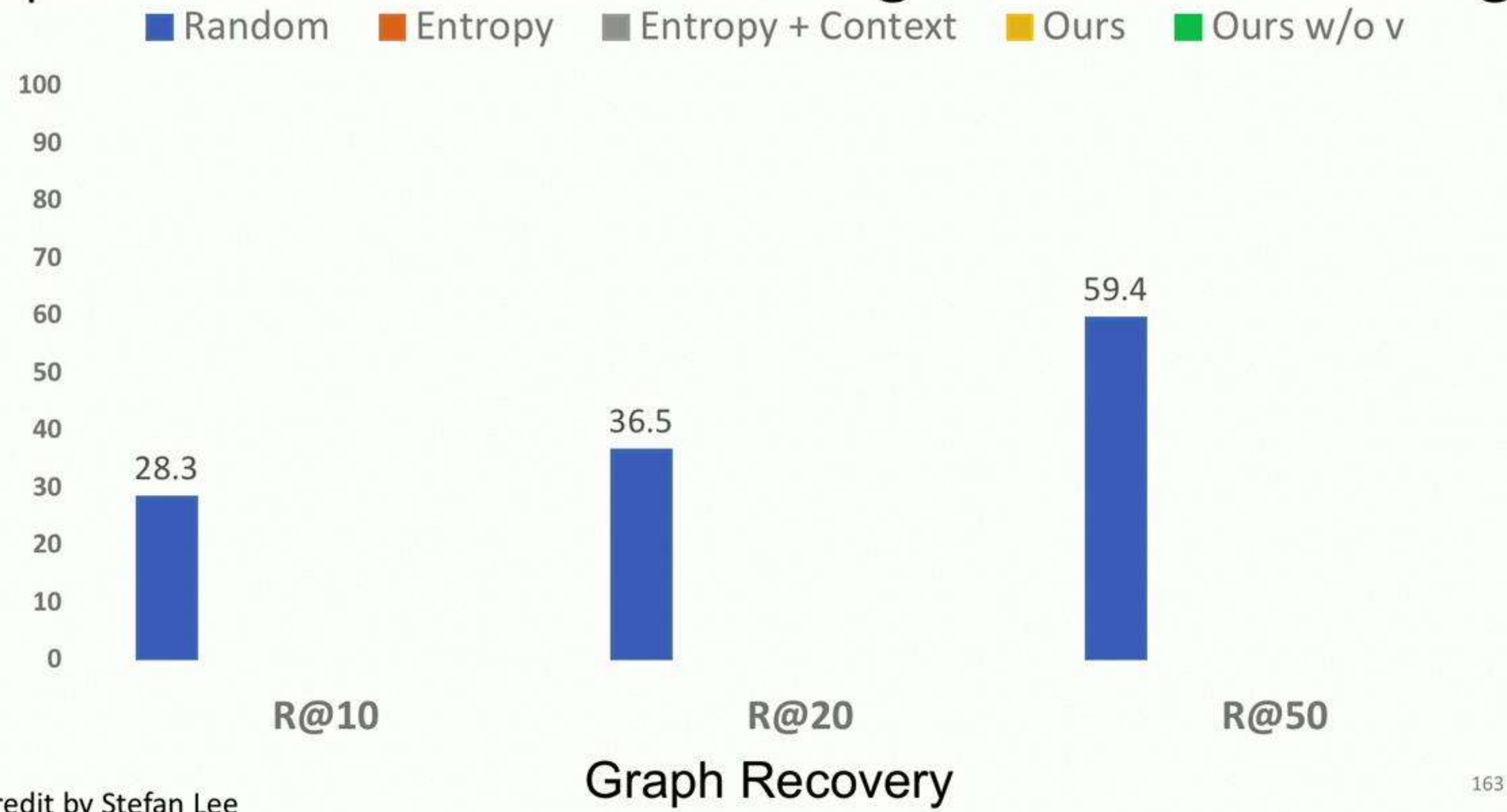
	Train	Val	Test
# Images	900	300	600

Experiments: Metrics

Graph Recovery R@k: Average accuracy of graph memory at dialog round k

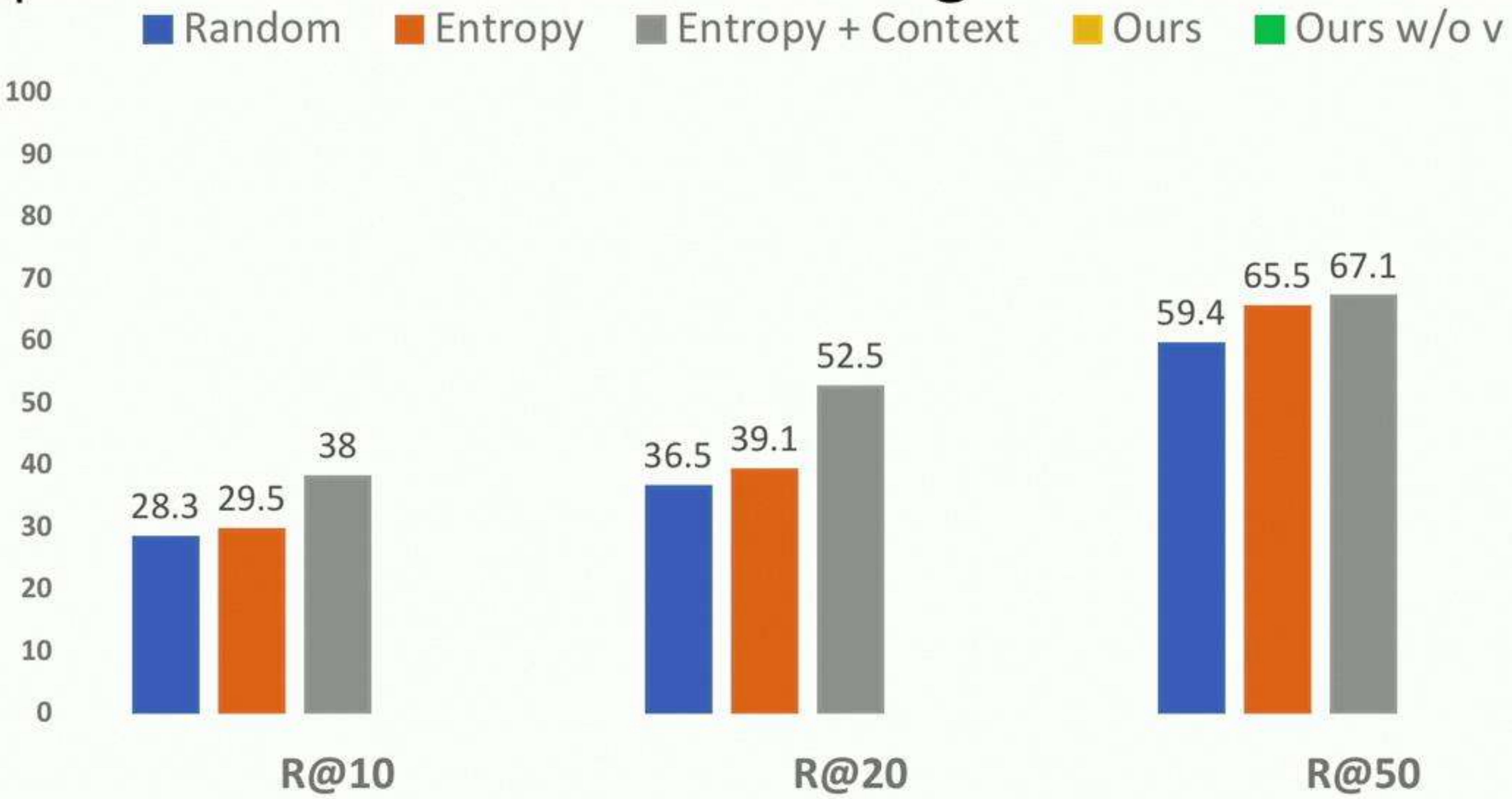


Experiment: Standard Training + Standard Testing



Slide credit by Stefan Lee

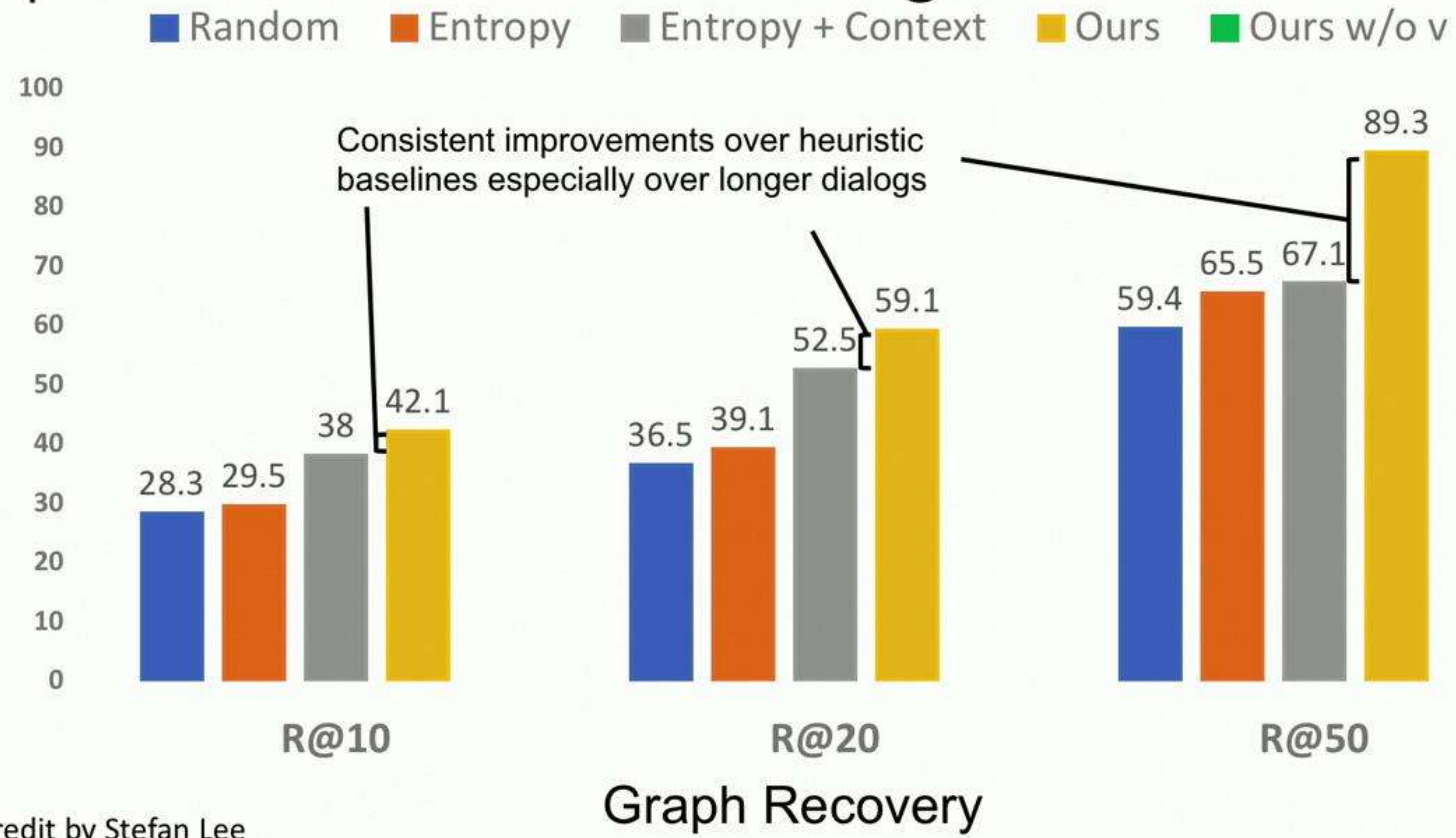
Experiment: Standard Training + Standard Testing



Graph Recovery

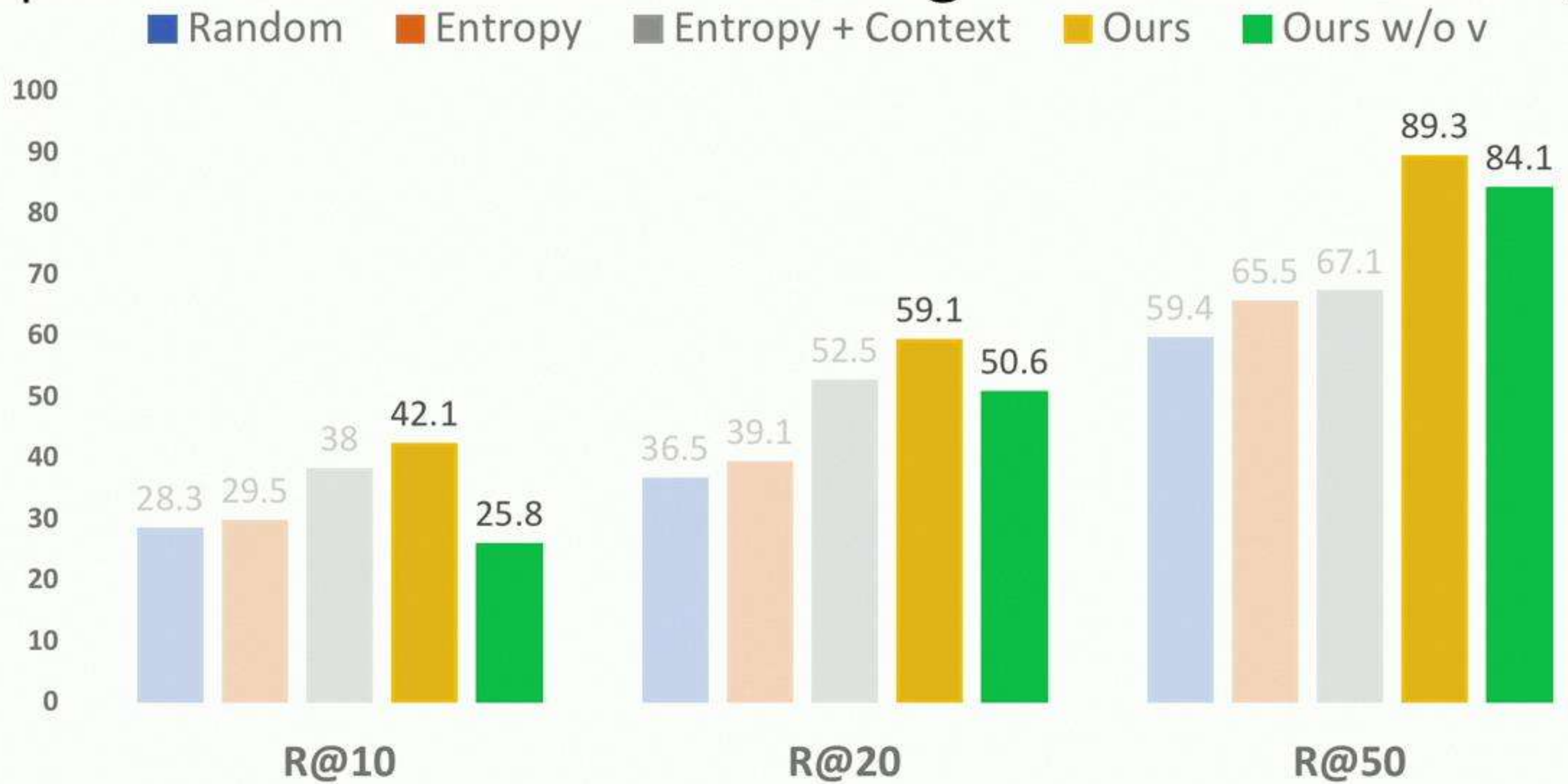
Slide credit by Stefan Lee

Experiment: Standard Training + Standard Testing



Slide credit by Stefan Lee

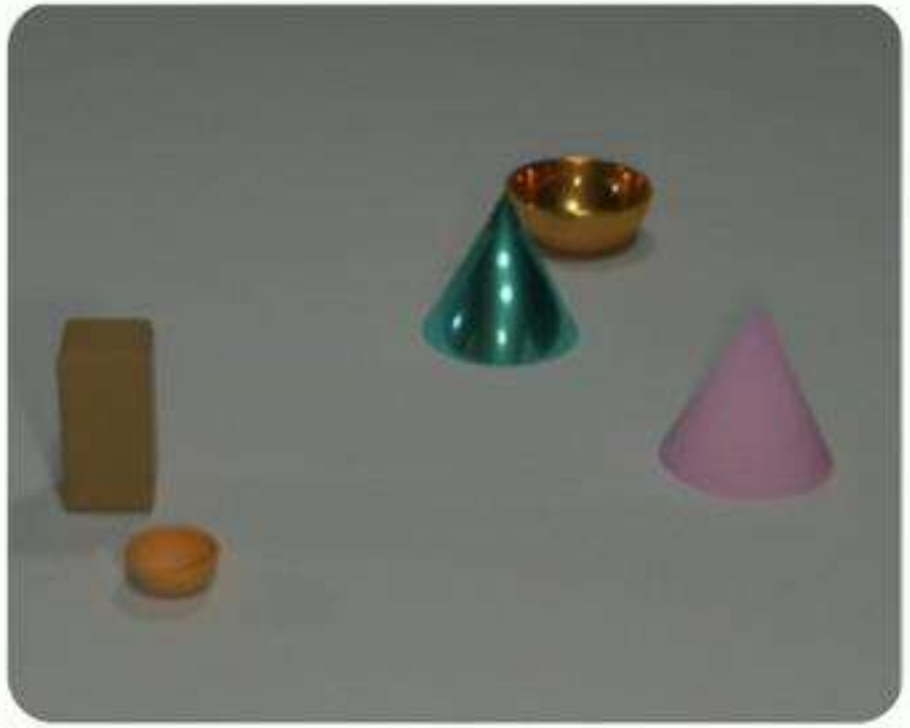
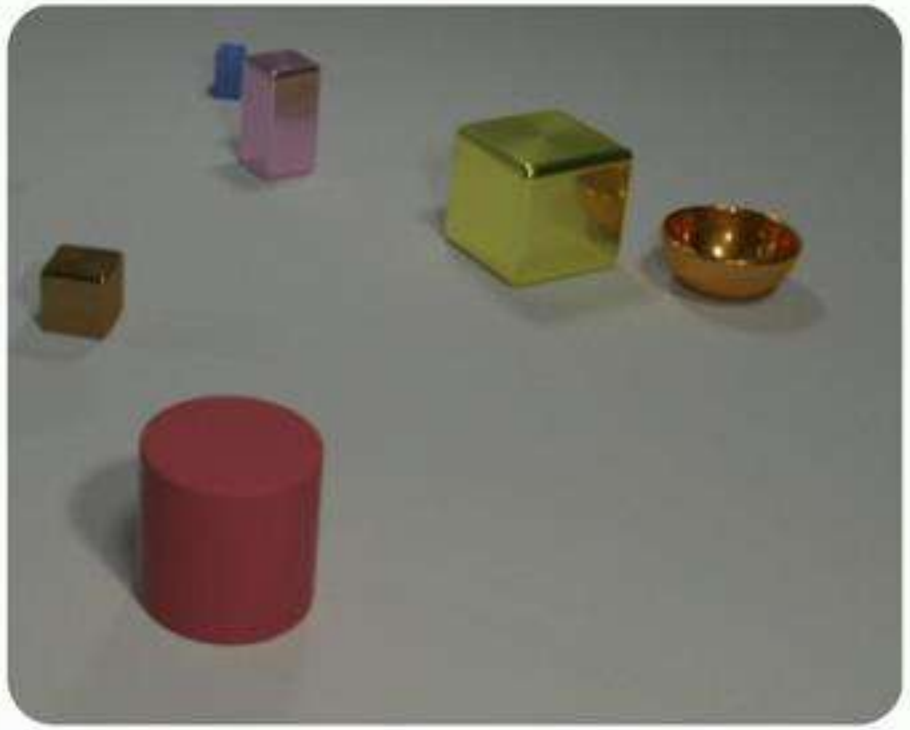
Experiment: Standard Training + Standard Testing



Graph Recovery

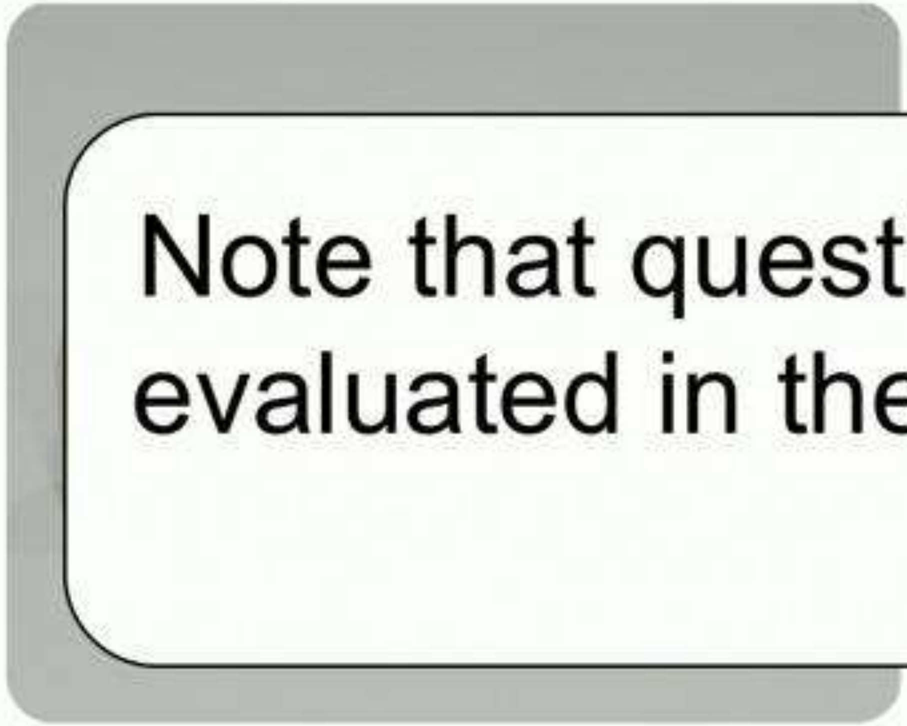
Slide credit by Stefan Lee

Experiments: Novel Object Environments

Novel	Mixed	Realistic
		
New colors and shapes	Mix of novel and standard colors and shapes	51 categories, 11 colors, 6 materials
600 images for test (12 episodes)	600 images for test (12 episodes)	1200 images for test (24 episodes)

Experiments: Novel Object Environments

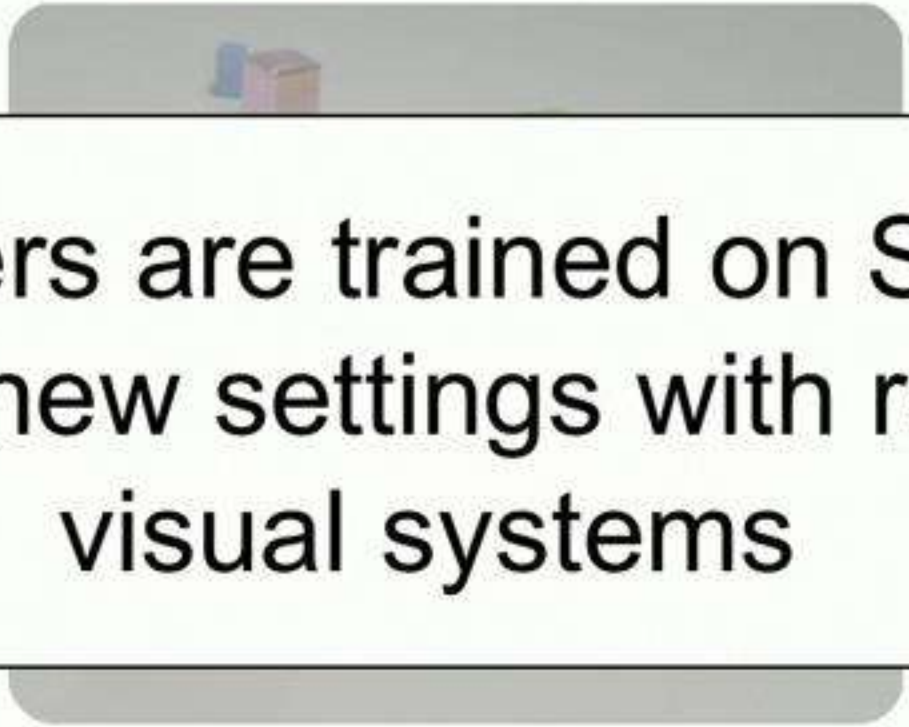
Novel



New colors and shapes

600 images for test
(12 episodes)

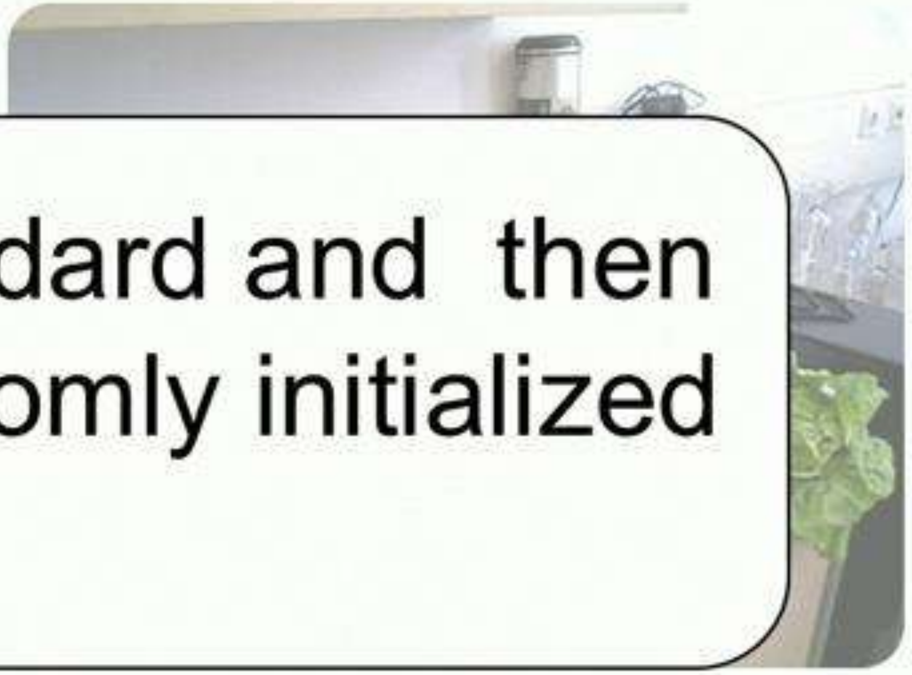
Mixed



Mix of novel and standard colors and shapes

600 images for test
(12 episodes)

Realistic



51 categories, 11 colors, 6 materials

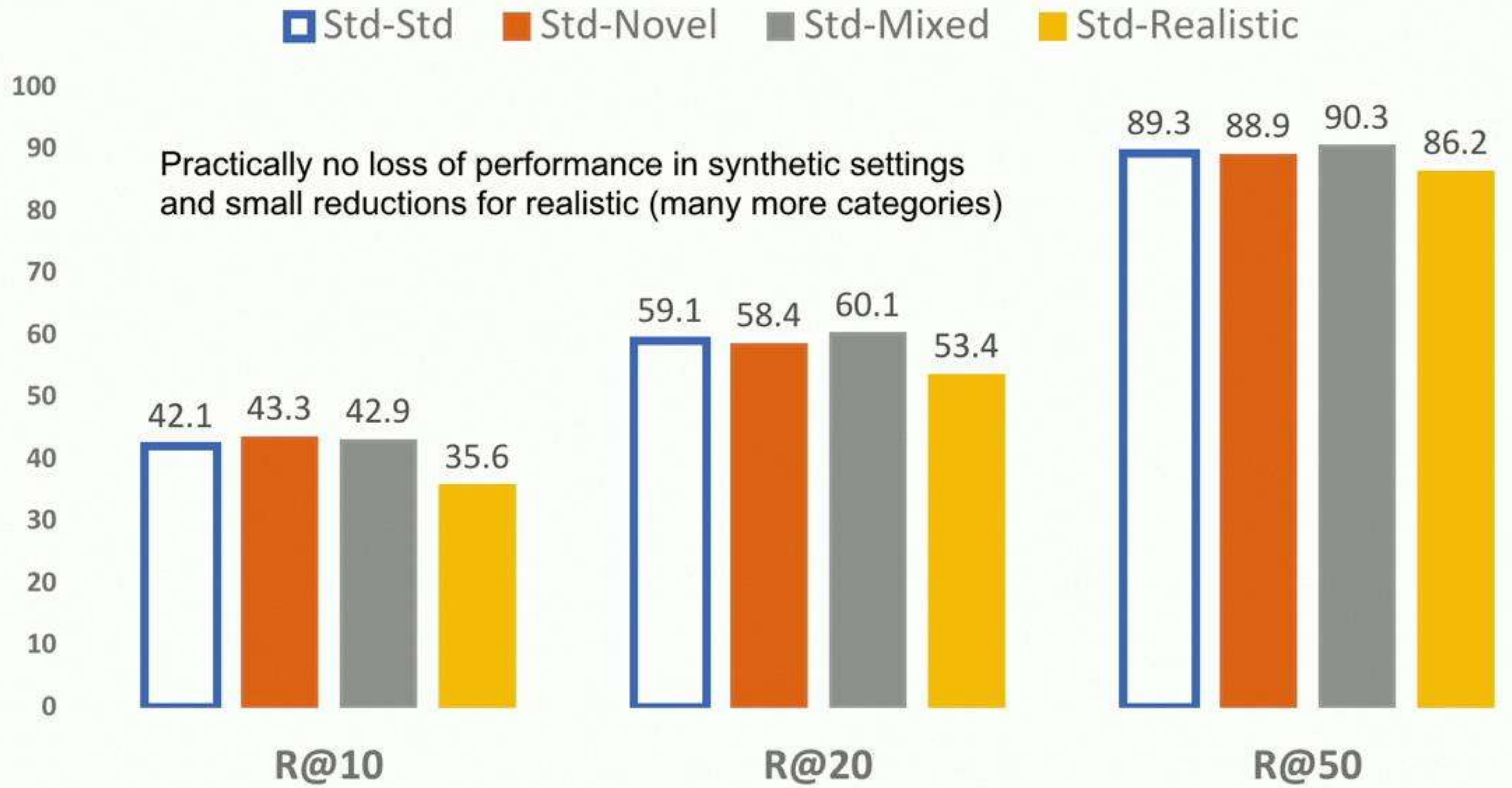
1200 images for test
(24 episodes)

Note that questioners are trained on Standard and then evaluated in these new settings with randomly initialized visual systems

Experiments: Standard Train – New Test Environments

Graph Recovery

Experiments: Standard Train – New Test Environments



Graph Recovery

Experiments: Qualitative Example

What material is the leftmost thing?

Experiments: Qualitative Example

What material is the leftmost thing? food

There is a leftmost object; what is it?

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?	

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?	

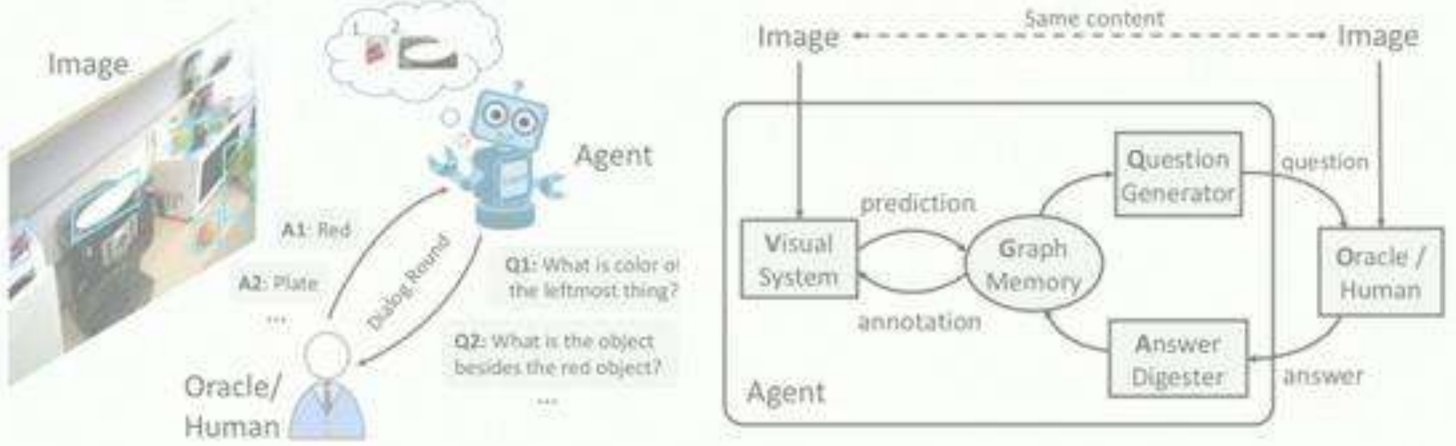
Experiments: Qualitative Example

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

Takeaways

- A new neural-symbolic pipeline is proposed to learn visual curiosity for an agent
- Through interaction with humans (Oracle), the agent improves its visual understanding capacity gradually
- The learned questions generation policy can directly adapt from synthetic dataset to realistic dataset

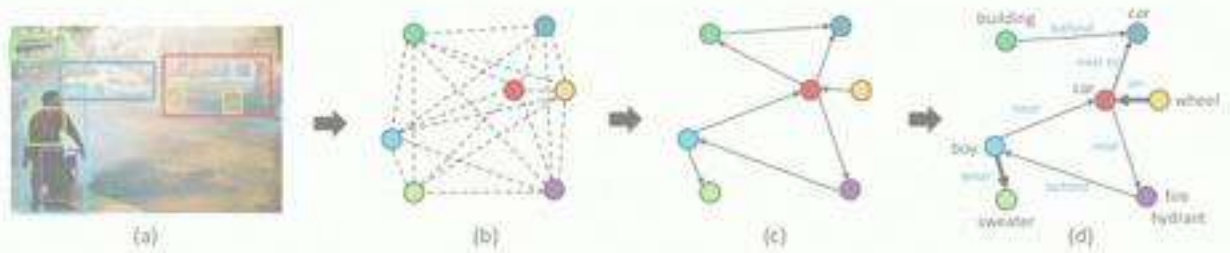
In this talk



Visual Curiosity. CoRL 2018

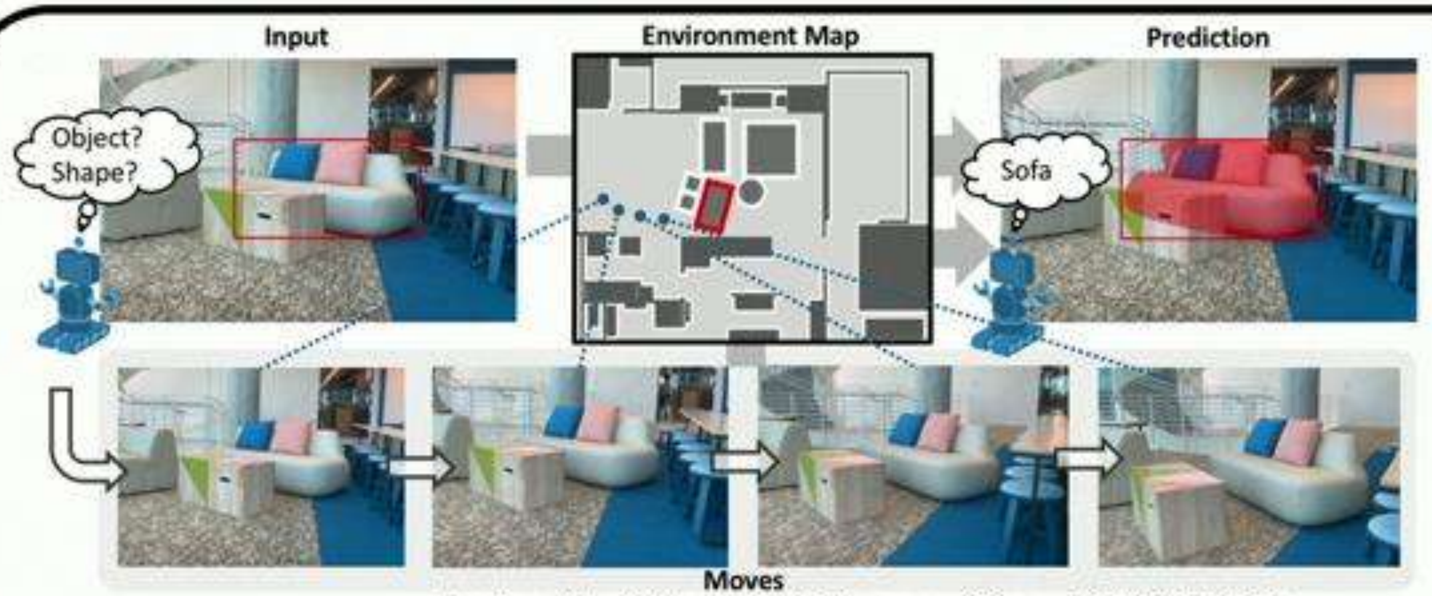
Interact with Human

Per-Image Structure



Graph R-CNN for Scene Graph Generation. ECCV 2018

Structured Visual Understanding



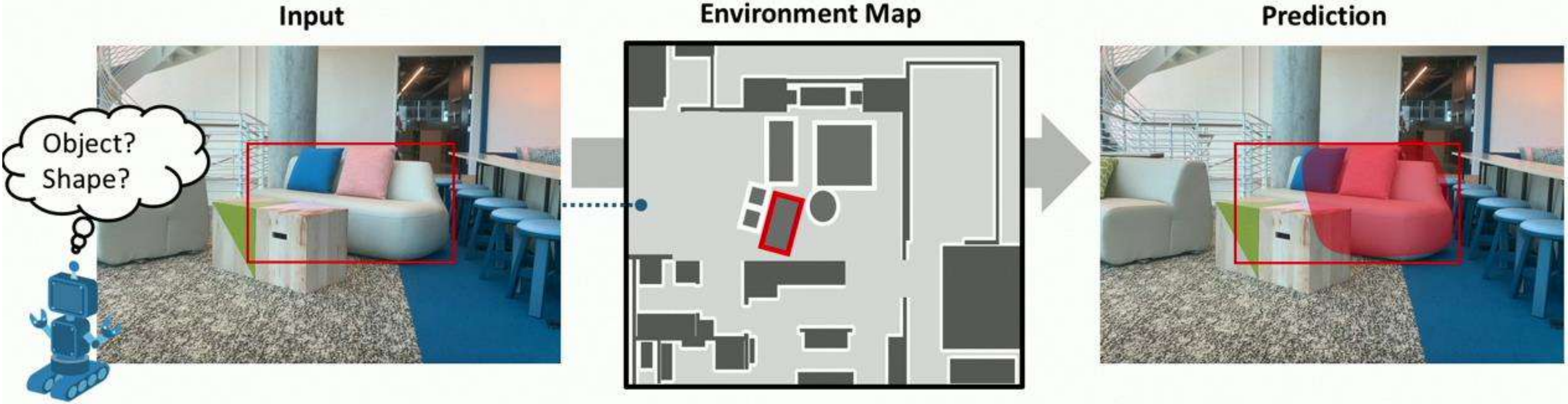
Embodied Amodal Recognition. ICCV 2019

Interact with Environment

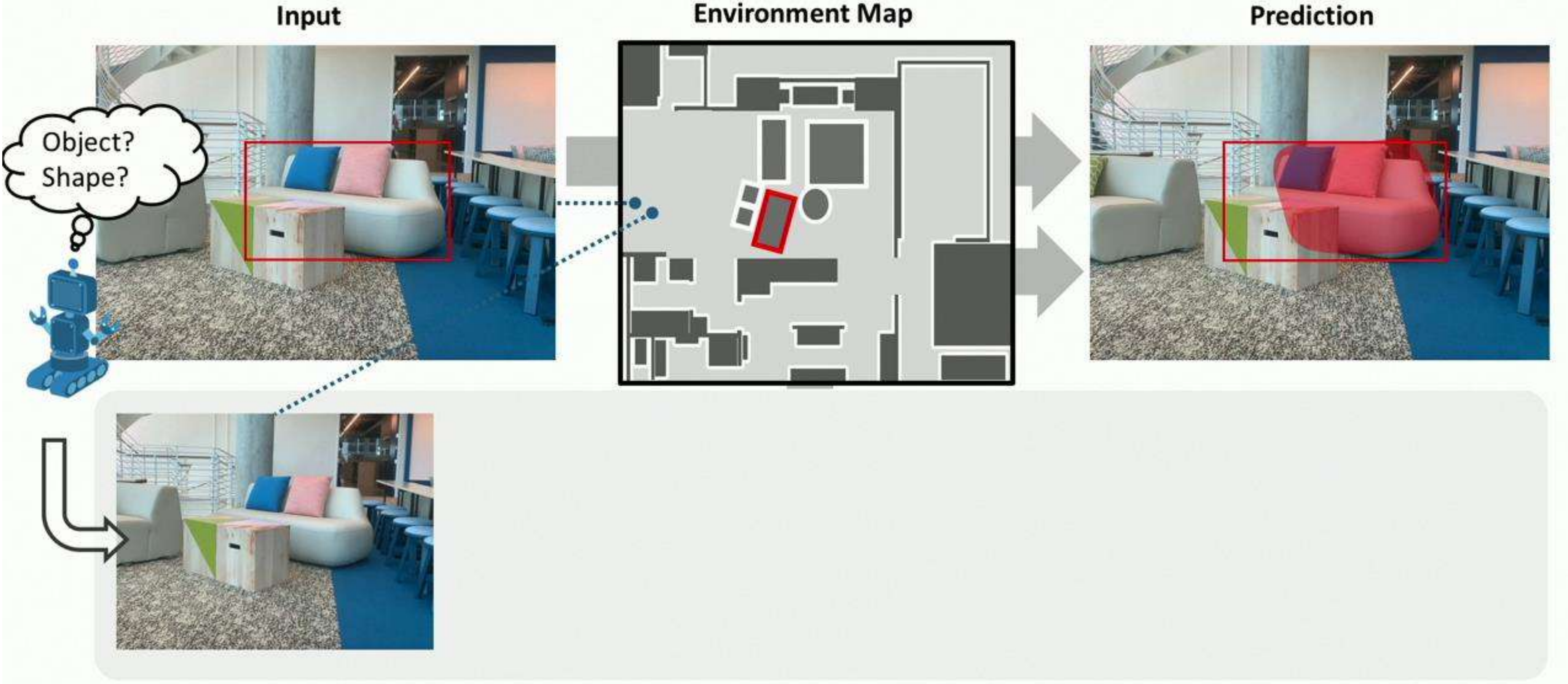
Visual Understanding by Moving in 3D Environment

Embodied Amodal Recognition: Learning to Move to Perceive Object. ICCV 2019

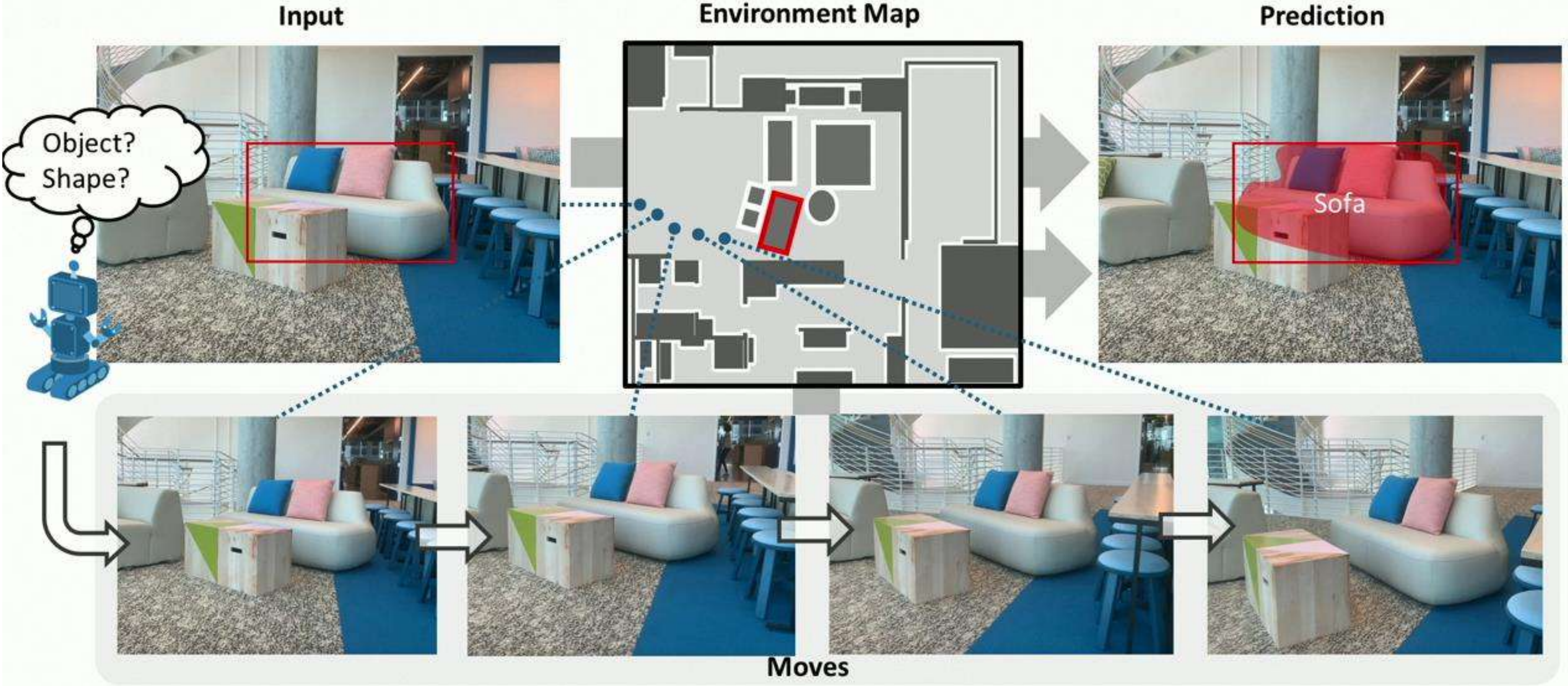
Motivation



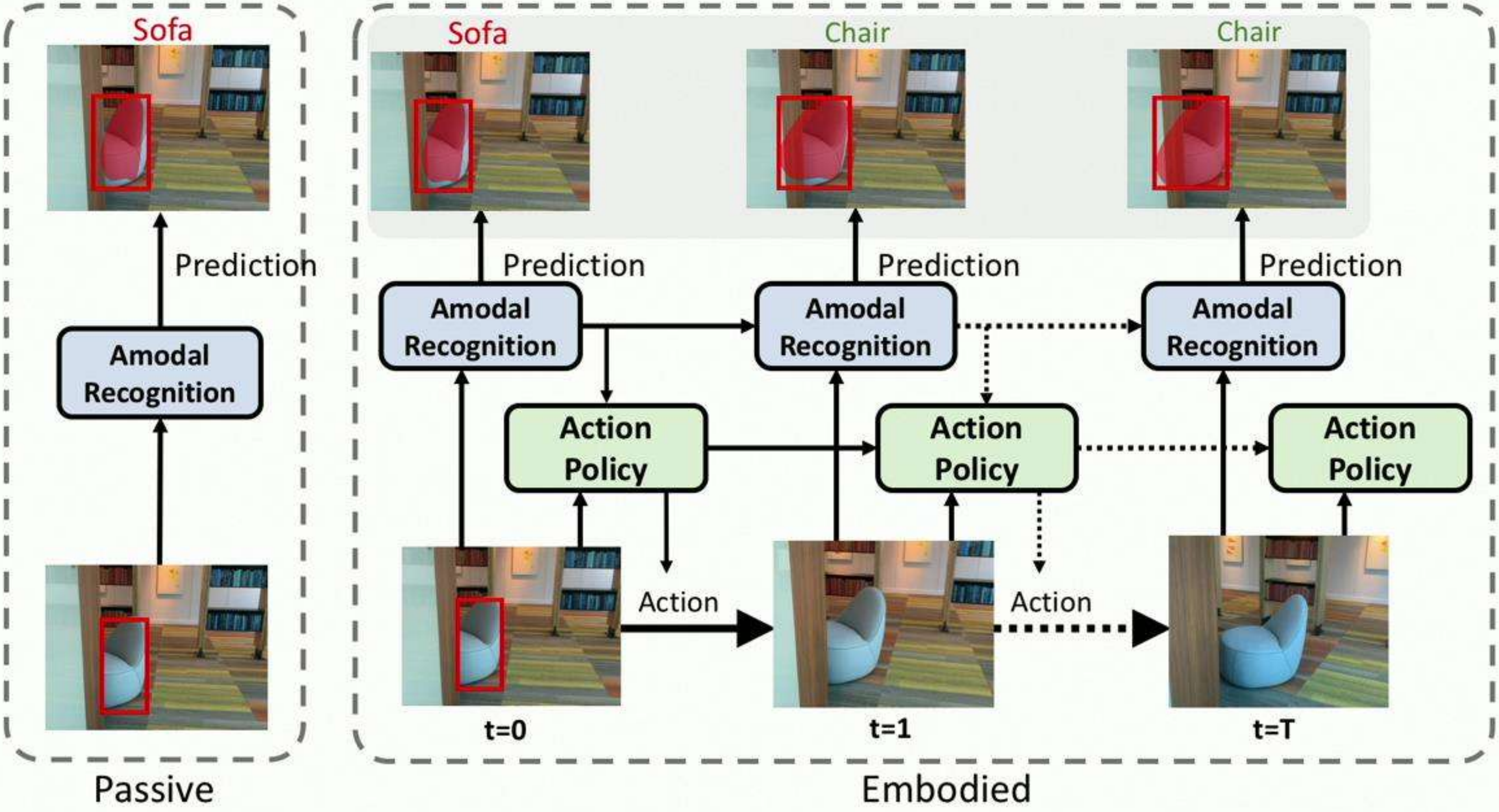
Motivation



Motivation



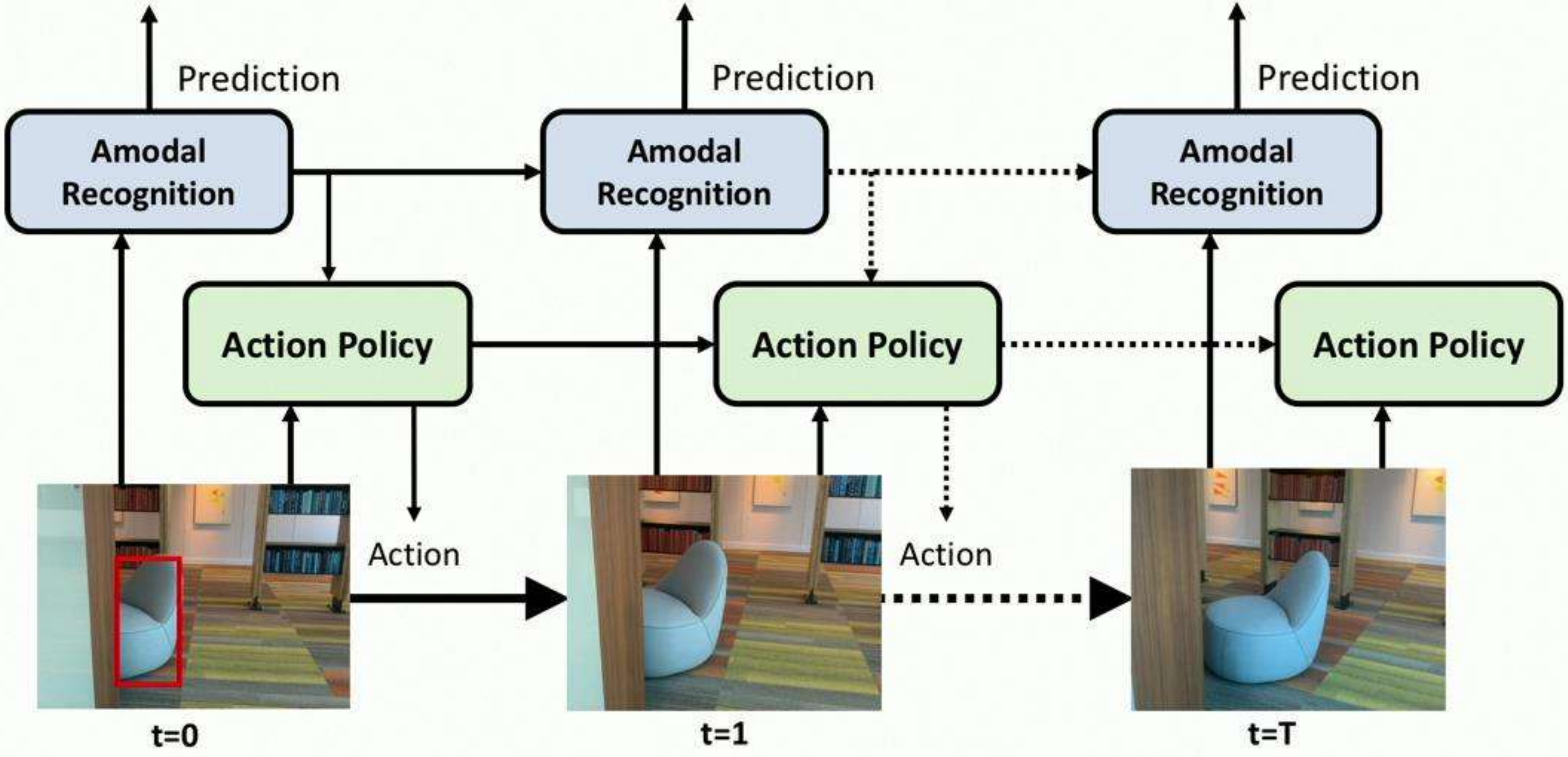
Embodied Amodal Recognition (EAR)



EAR Task

- **Three sub-tasks**
  - Object recognition
  - 2D amodal localization
  - 2D amodal segmentation
- **Single target object**
  - Specify one object as the target
- **Predict for the first frame**

EVR Model


Prediction
Prediction
Prediction
Amodal Recognition
Amodal Recognition
Amodal Recognition
Action Policy
Action Policy
Action Policy
Action
Action
t=0
t=1
t=T

Amodal Recognition

Objective:

Visible Box

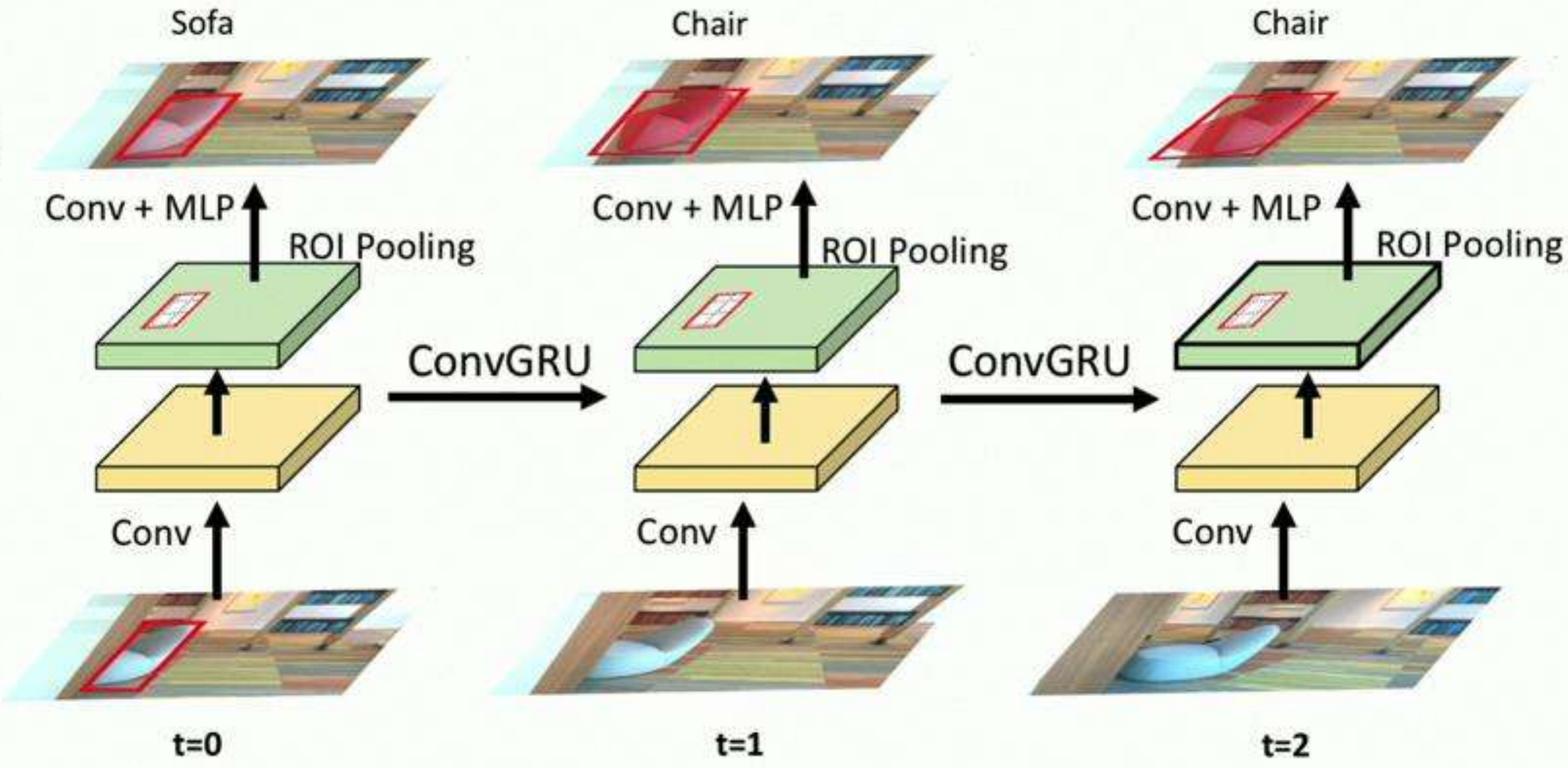
$$y_t = f(b_0, \underbrace{I_0, I_1, \ldots, I_t}_{\text{Observations}})$$

Temporal Aggregation

$$h_t = GRU(x_t, h_{t-1})$$

Three losses:

$$L = L_c + L_b + L_m$$



Learn to Move (Policy)

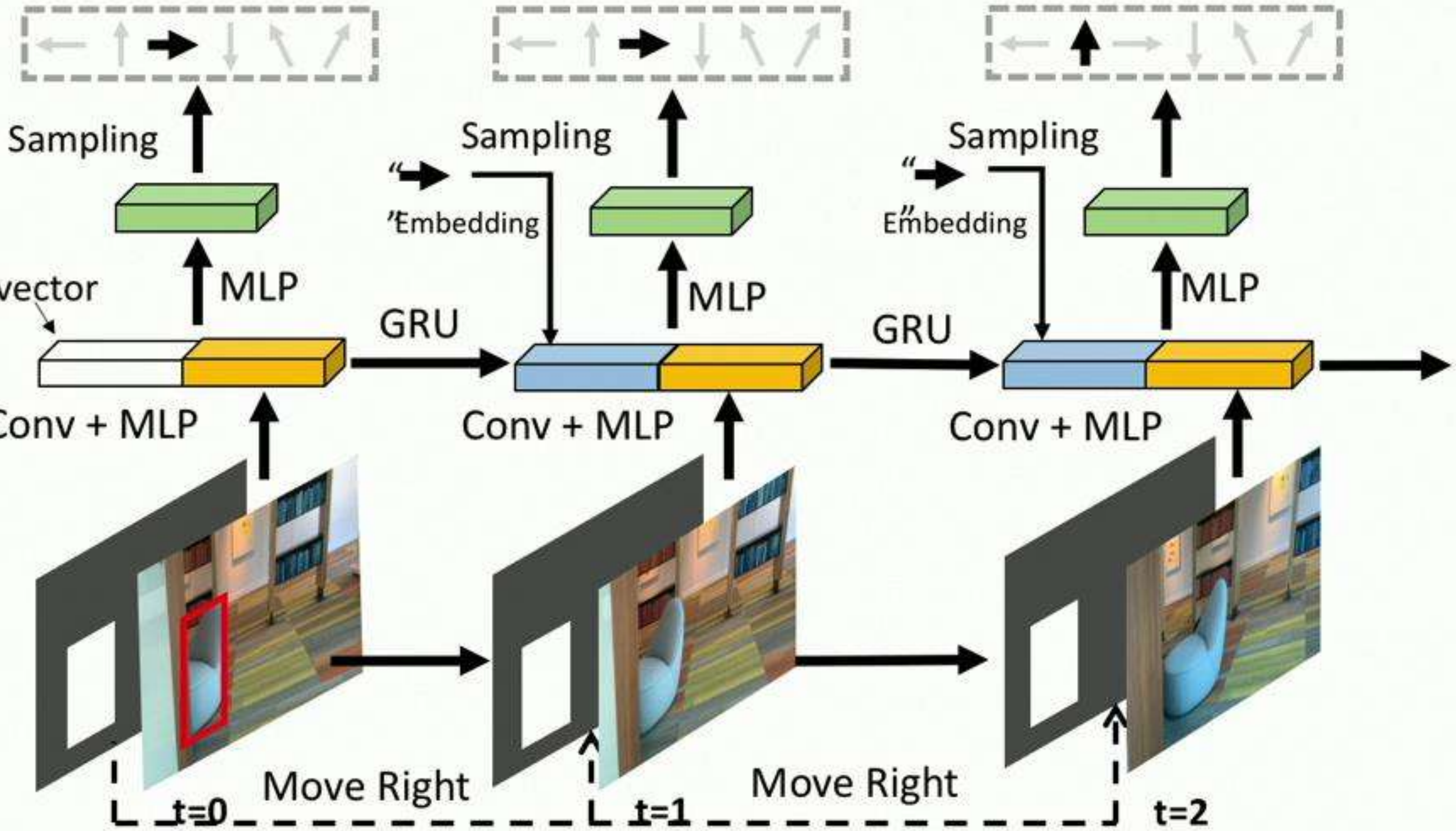
Objective:

$$a_t = \pi(b_0, I_0, I_1, \ldots, I_t)$$

Reward:

$$r_t = \lambda_c Acc_c^t + \lambda_b IoU_b^t + \lambda_m IoU_m^t$$

Reward reshaping
+ REINFORCEMENT



Amodal Recognition

Objective:

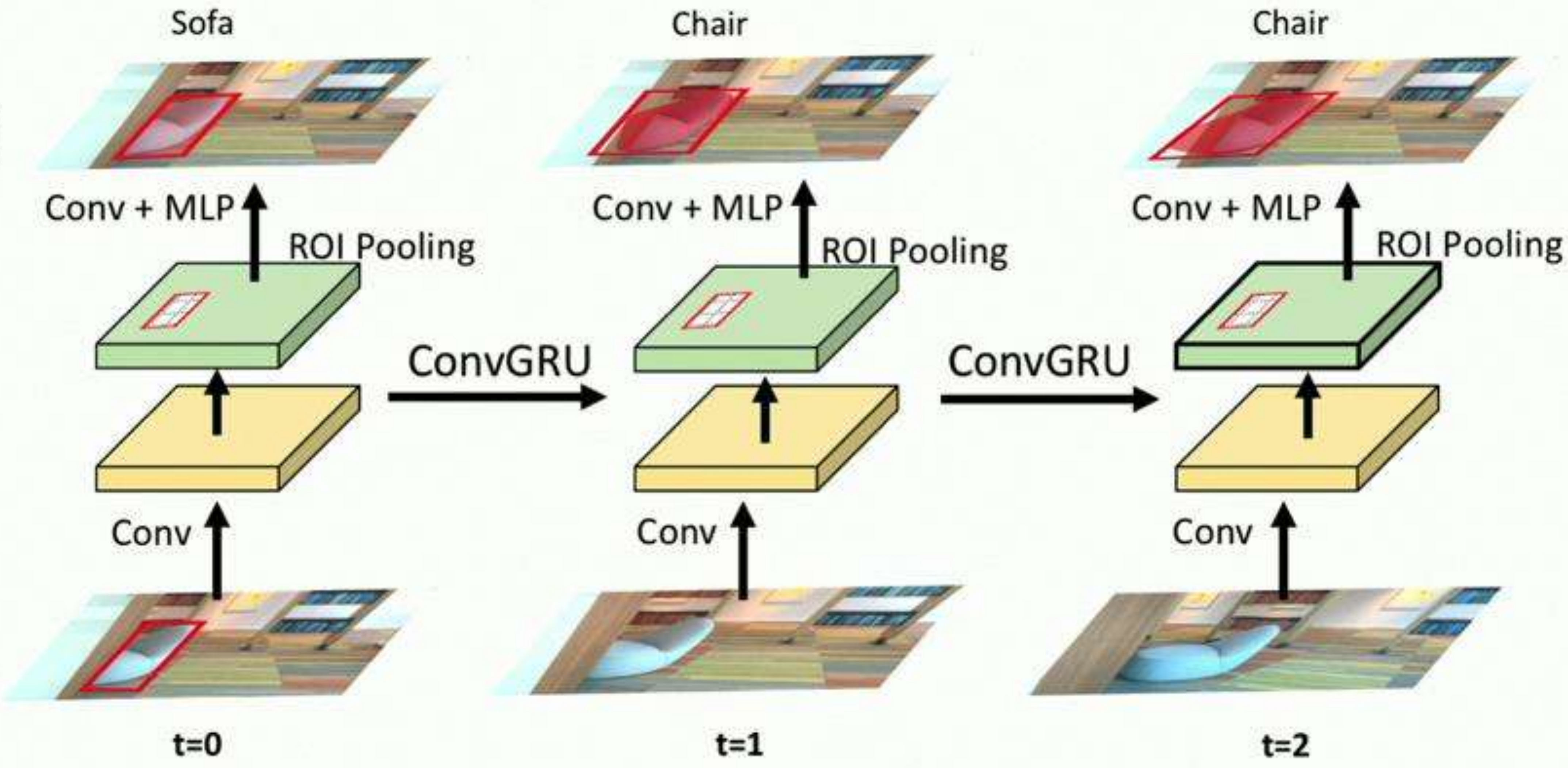
$$y_t = f(\overset{\text{Visible Box}}{b_0}, \underbrace{I_0, I_1, \dots, I_t}_{\text{Observations}})$$

Temporal Aggregation

$$h_t = GRU(x_t, h_{t-1})$$

Three losses:

$$L = L_c + L_b + L_m$$



Learn to Move (Policy)

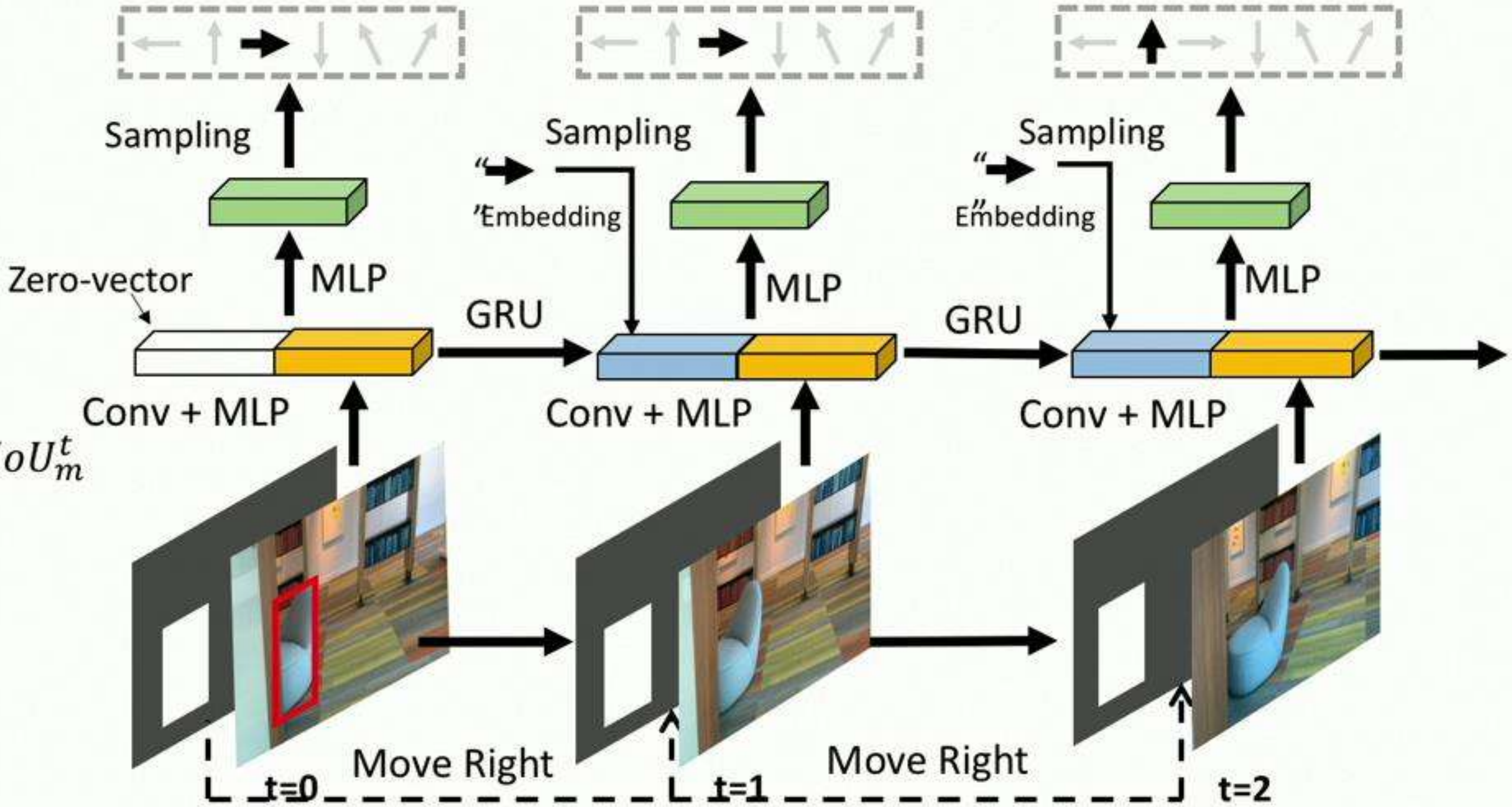
Objective:

$$a_t = \pi(b_0, I_0, I_1, \ldots, I_t)$$

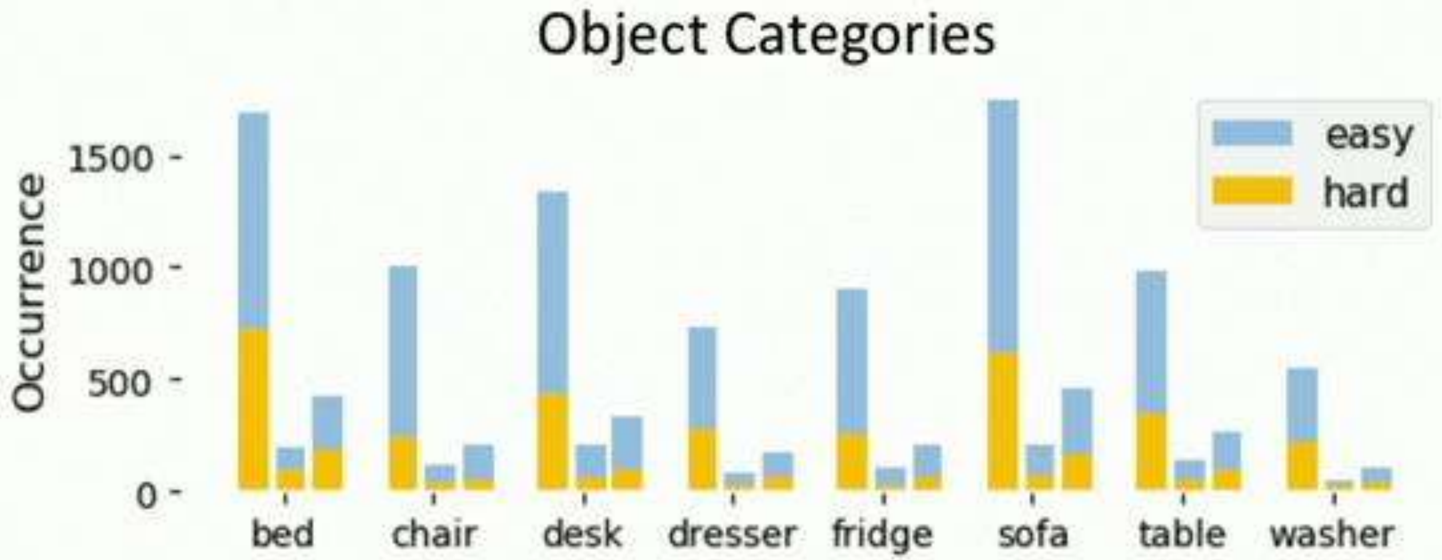
Reward:

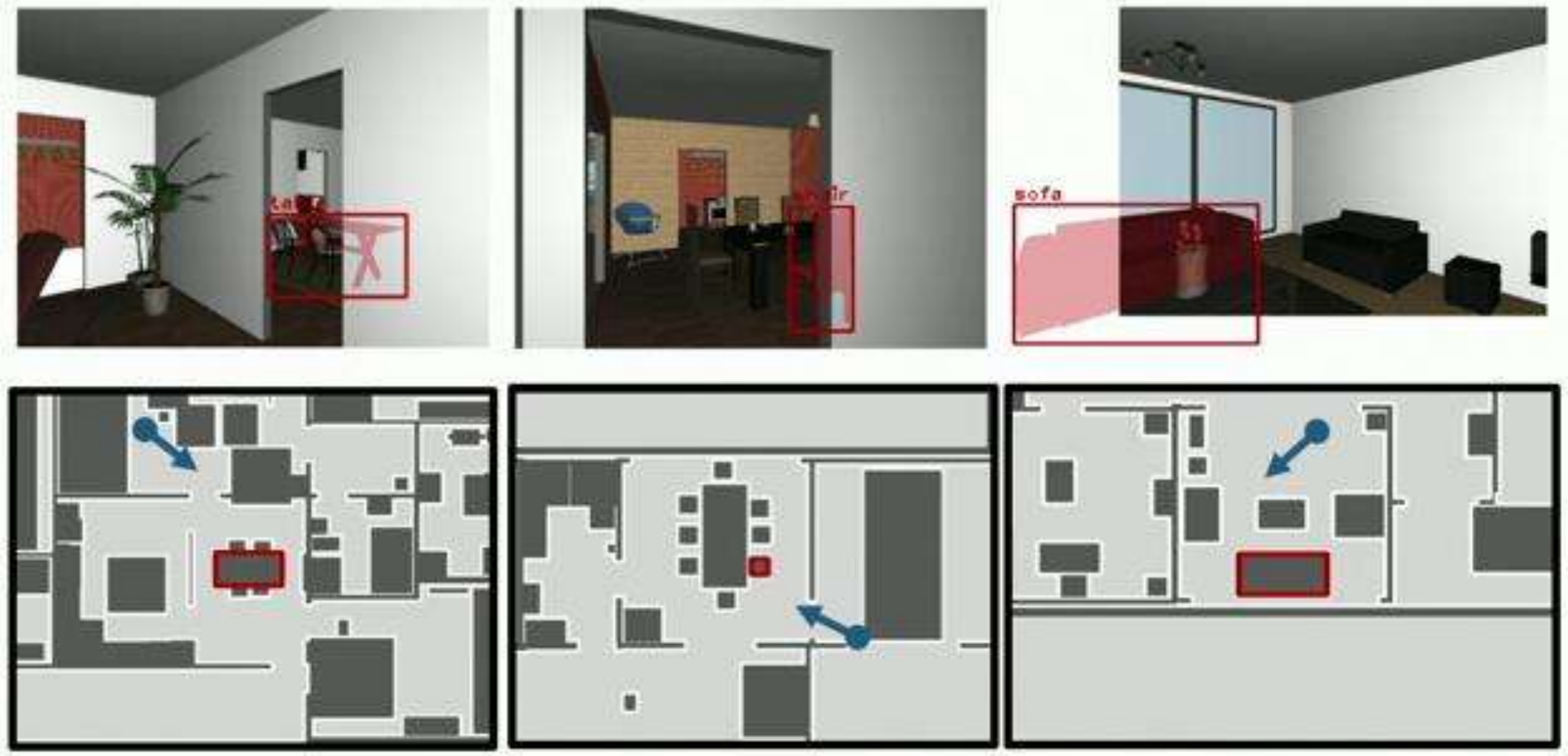
$$r_t = \lambda_c Acc_c^t + \lambda_b IoU_b^t + \lambda_m IoU_m^t$$

Reward reshaping
+ REINFORCEMENT

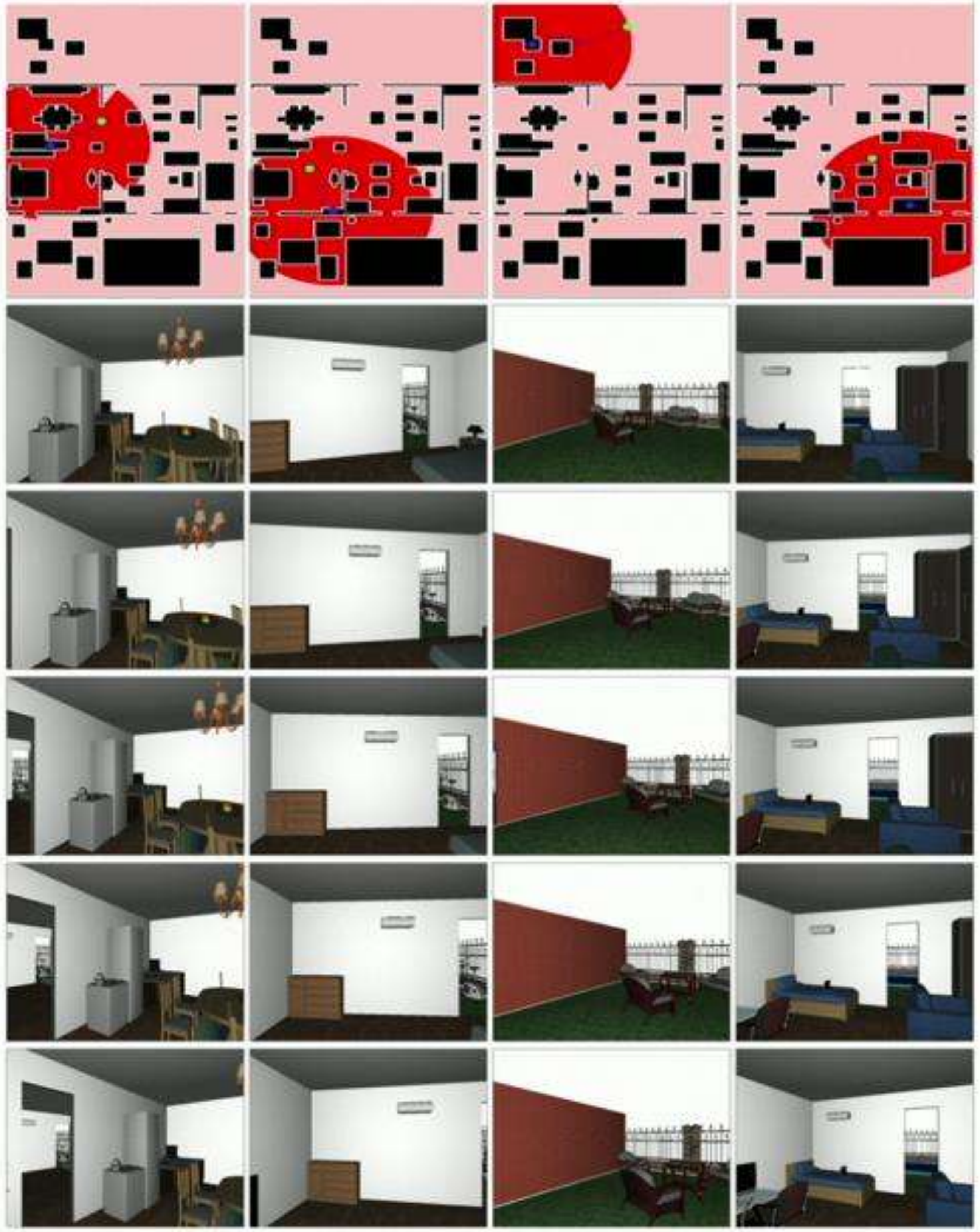


Dataset





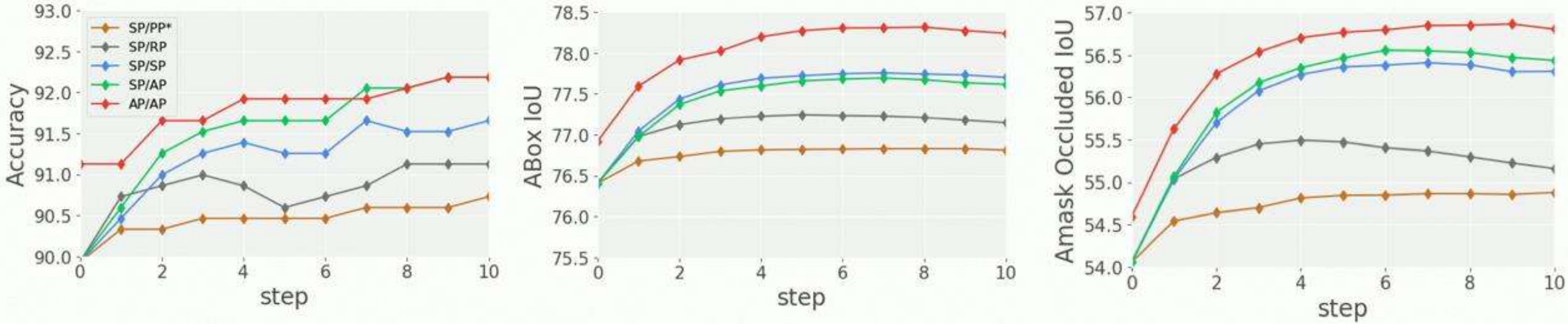
Shortest path toward target



Training

- Stage-wised Training:
  - First train amodal visual recognition with shortest path
  - Then fix amodal visual recognition module, train policy network
  - Afterwards, train amodal visual recognition with learned path

Results



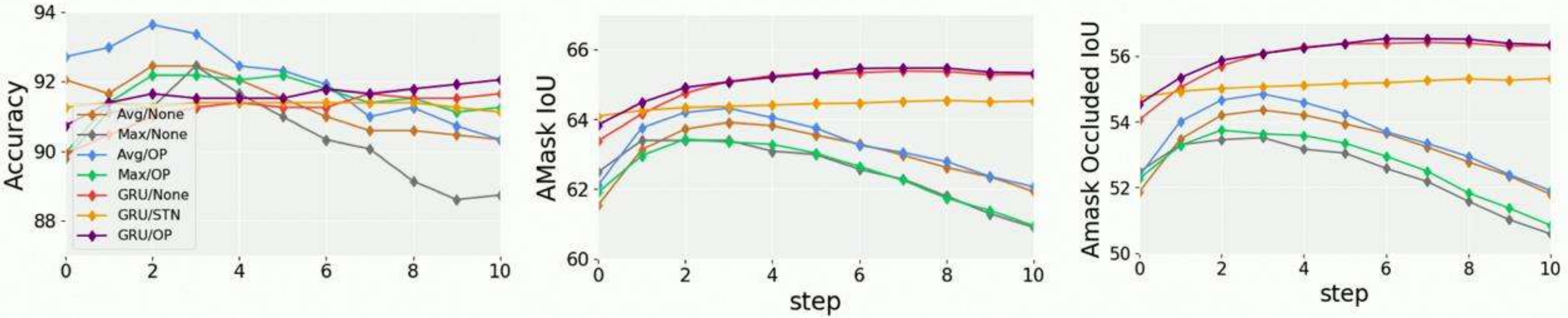
1. Embodiment helps to improve amodal visual recognition performance

2. Our learned moving strategy for agent outperforms other moving strategy and also static visual system.

3. Amodal recognition performance tends to saturate at the end of moving

Ablated Study

Different feature aggregation and feature warping methods:

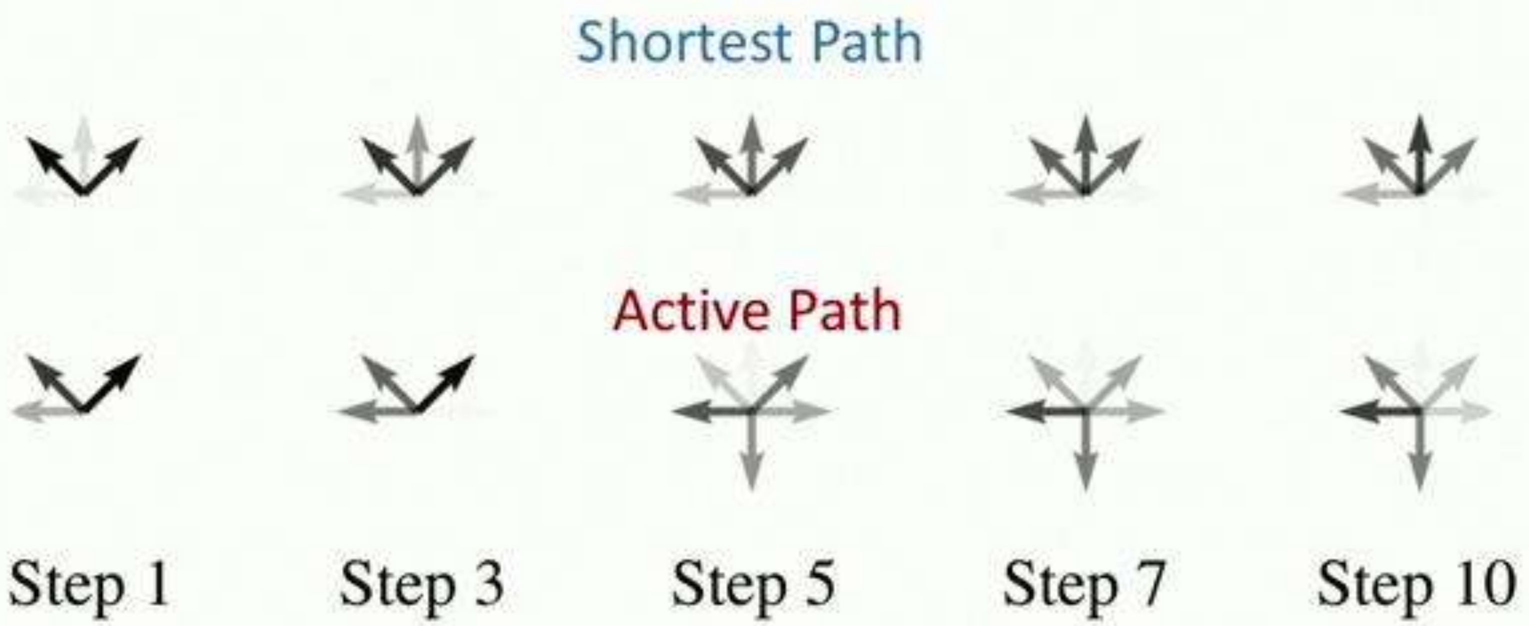


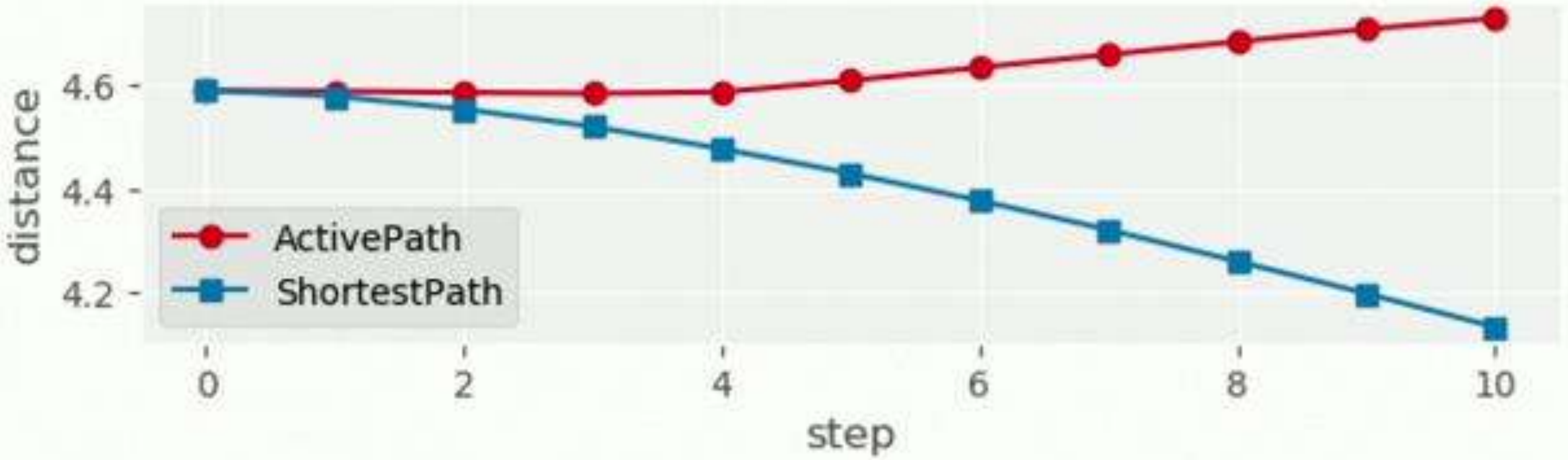
1. GRU works much better than average or max pooling for aggregation

2. Warping feature using optical flow helps to improve the performance

3. Combine GRU and optical flow works slightly better than either

Learned Actions



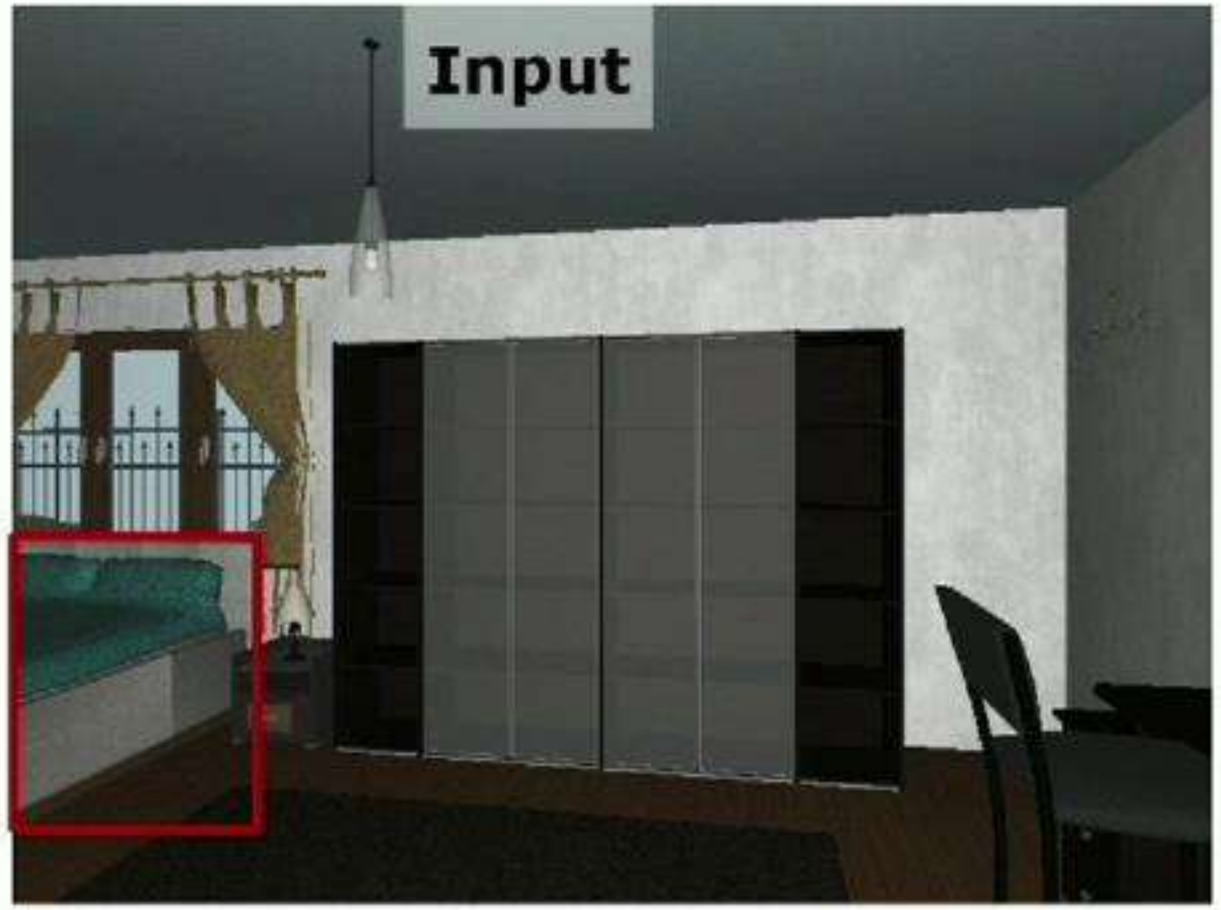


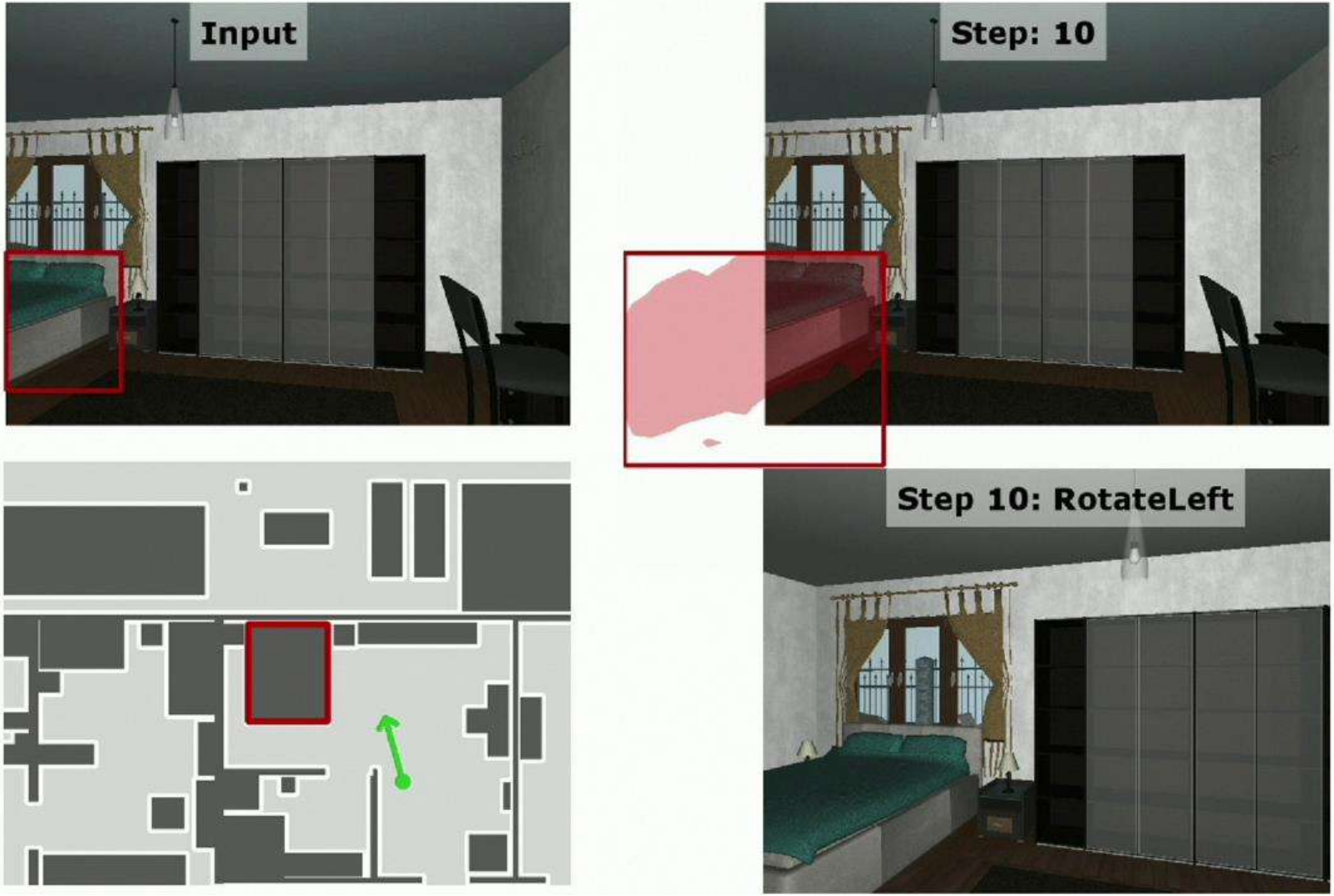
1. Our policy network has learned different moving strategies from shortest path
2. In general, the learned policy keeps the agent in a distance to the target object

Passive Perception vs. Active Perception

Shortest Path vs. Learned Active Path

Shortest-Path


Input


Input
Step: 10
Step 10: RotateLeft

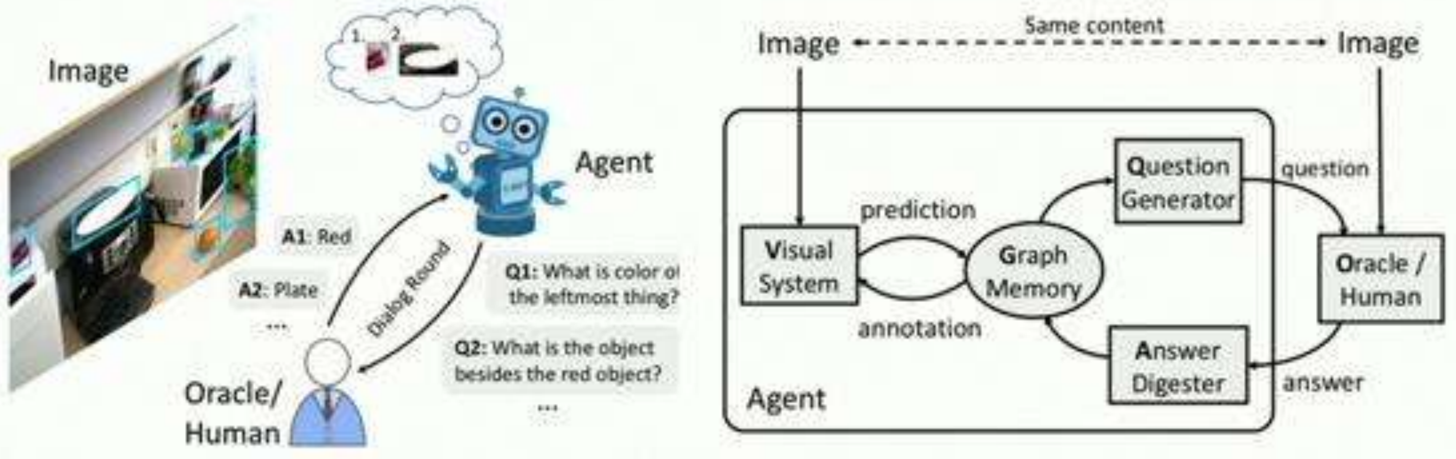
Active Path

Shortest-Path

Takeaways

- A embodied visual recognition system is proposed as an initial step toward an intelligent agent system
- Embodiment helps to get better visual recognition of objects
- Learning a better moving strategy is challenging but helpful to improve visual recognition
- It would be interesting to enable full understanding of the whole environment after moves

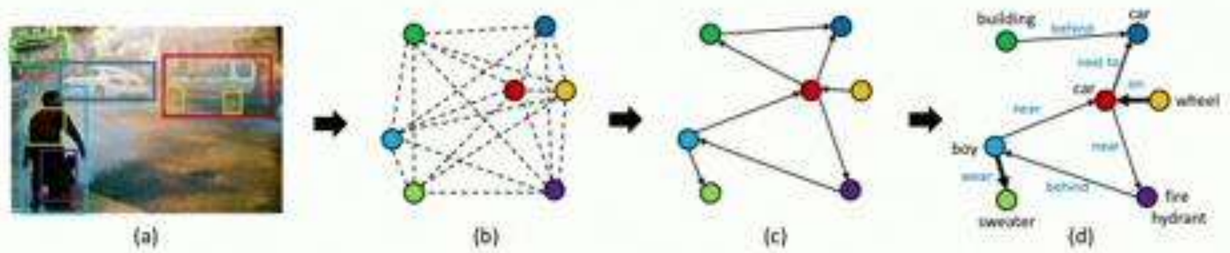
In this talk



Visual Curiosity. CoRL 2018

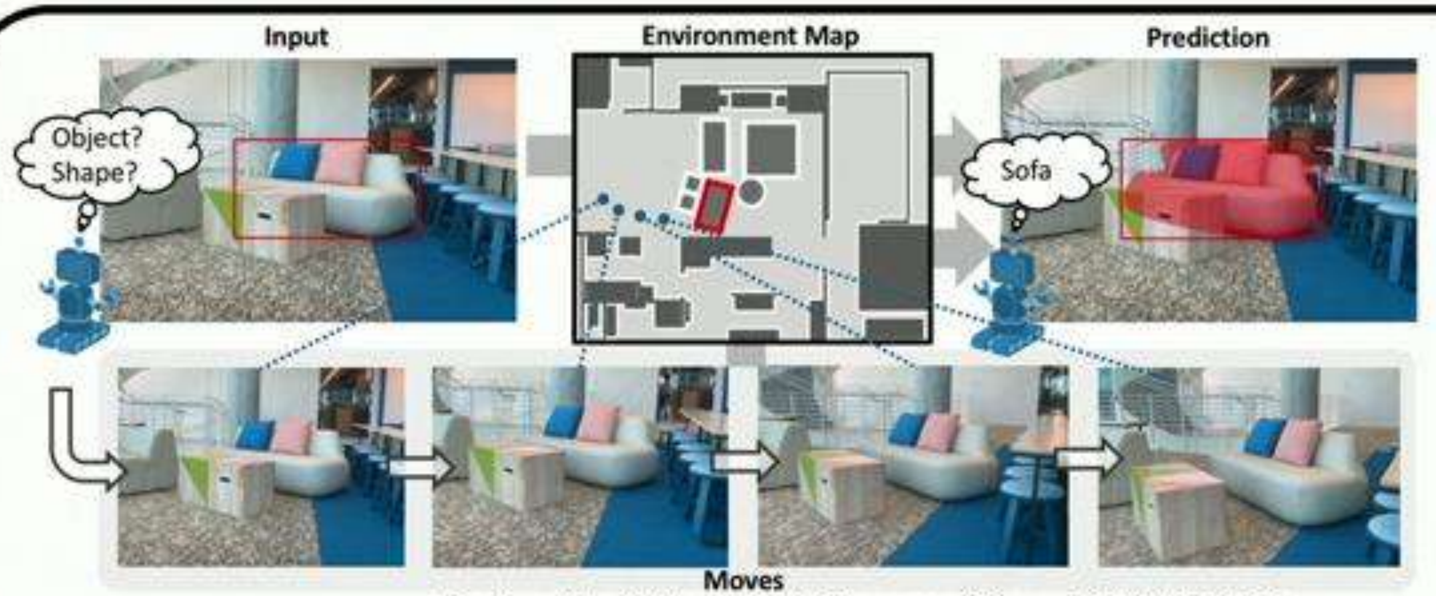
Interact with Human

Per-Image Structure



Graph R-CNN for Scene Graph Generation. ECCV 2018

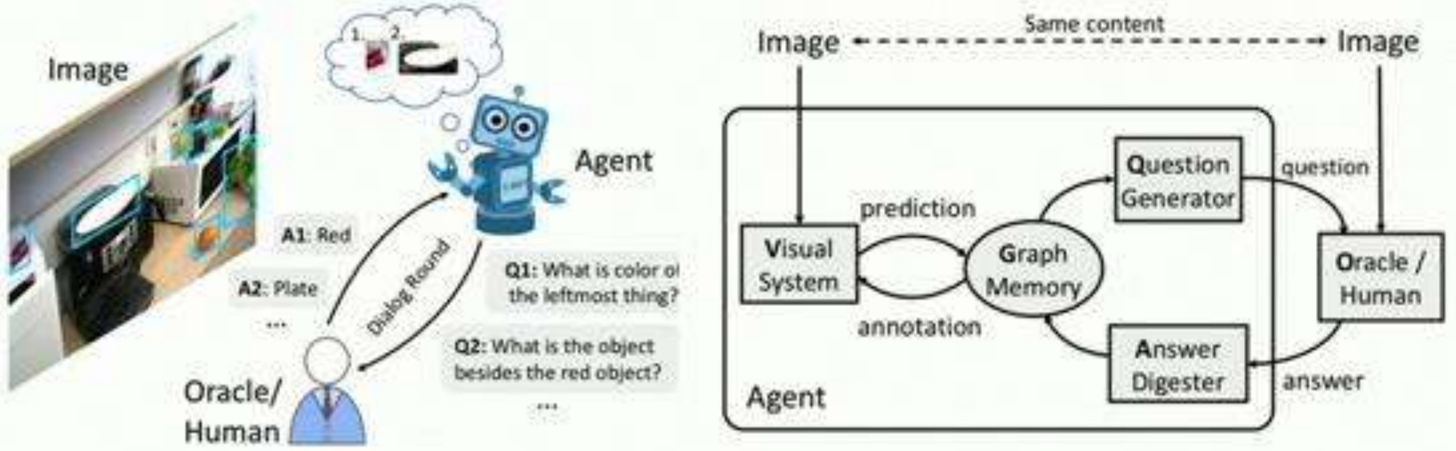
Structured Visual Understanding



Embodied Amodal Recognition. ICCV 2019

Interact with Environment

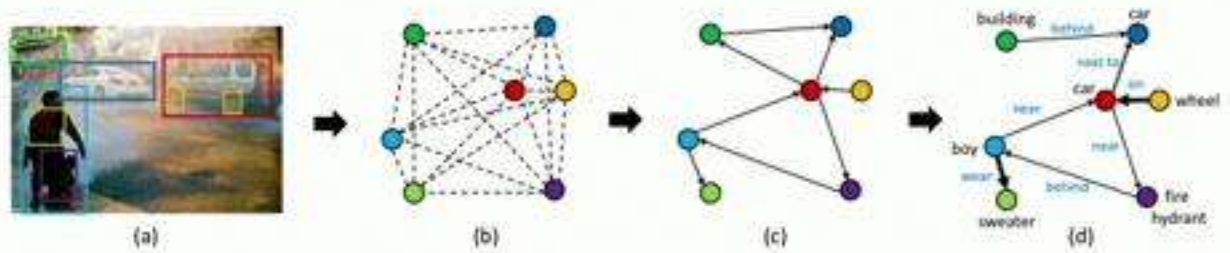
In this talk



Visual Curiosity. CoRL 2018

Interact with Human

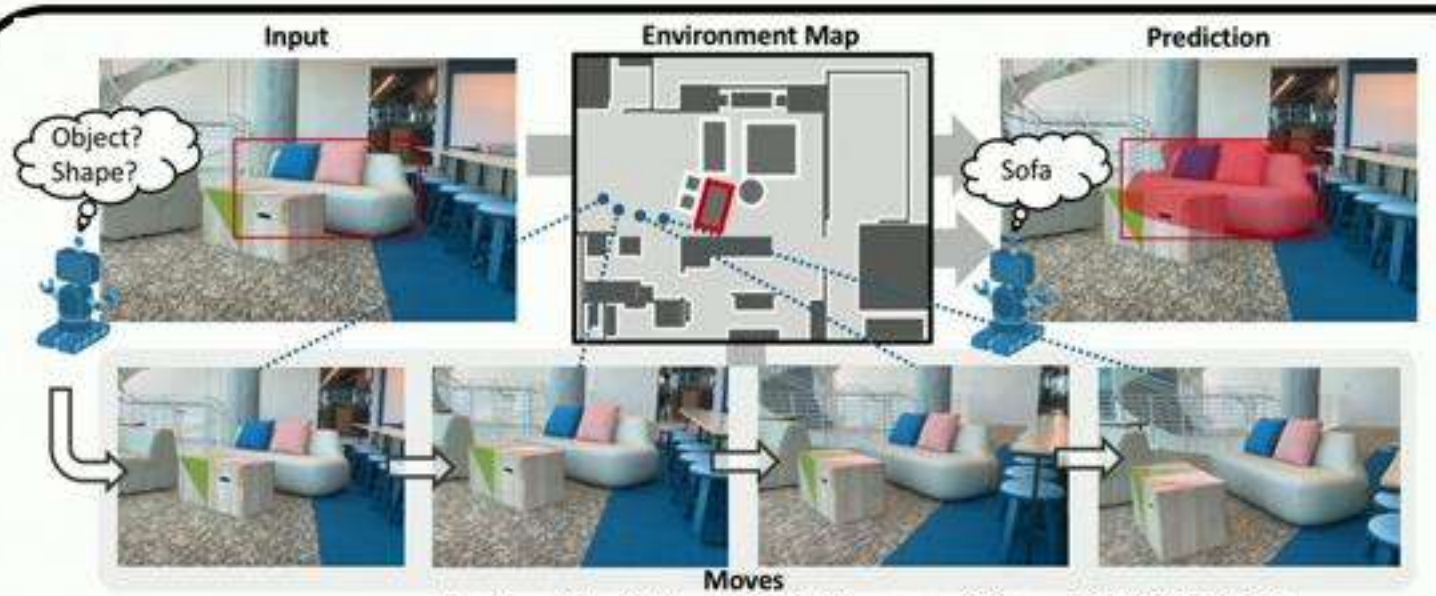
Per-Image Structure



Graph R-CNN for Scene Graph Generation. ECCV 2018

Structured Visual Understanding

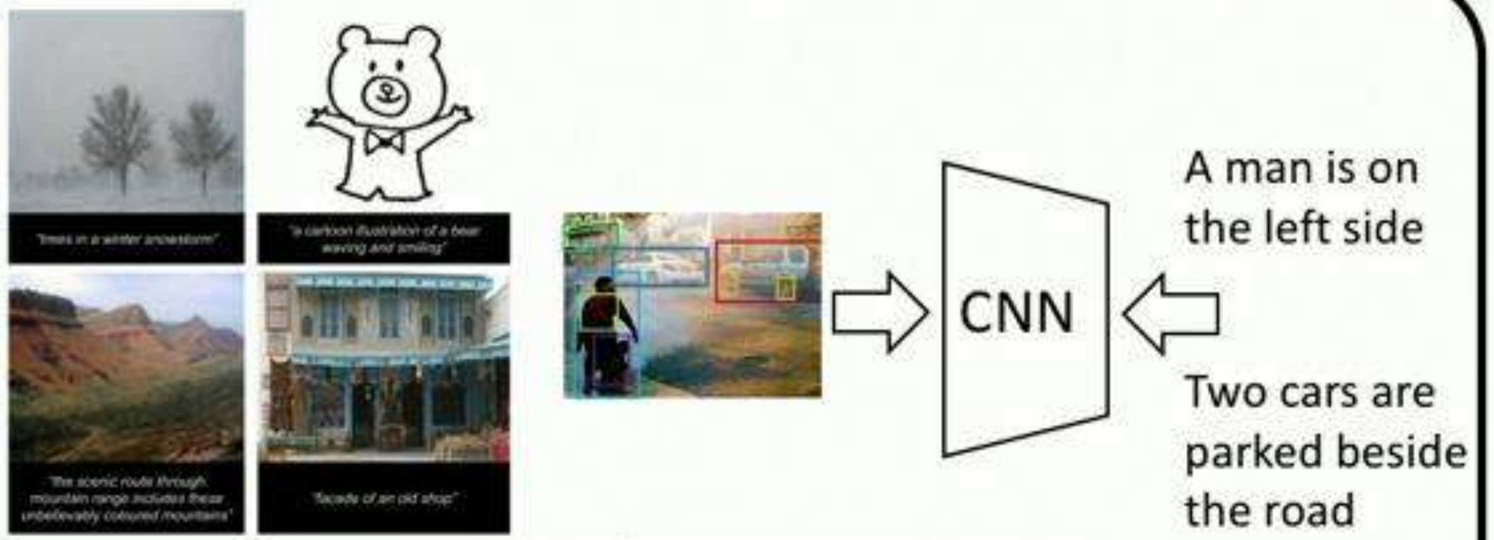
Learn from interactions with human and environment



Embodied Amodal Recognition. ICCV 2019

Interact with Environment

As Future Works

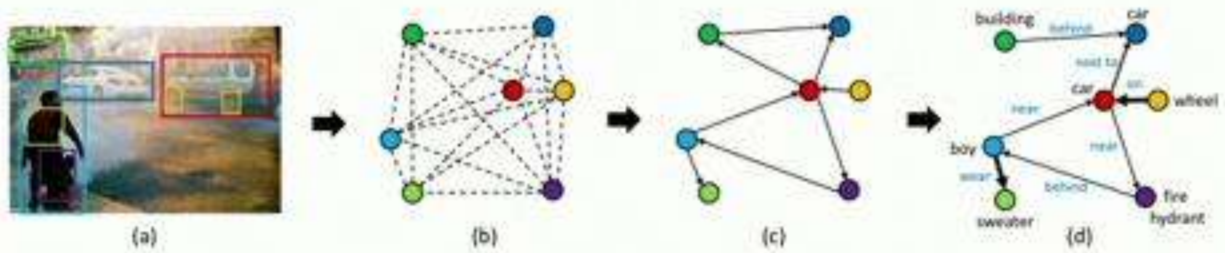


Learning from image corpus

Interact with Human

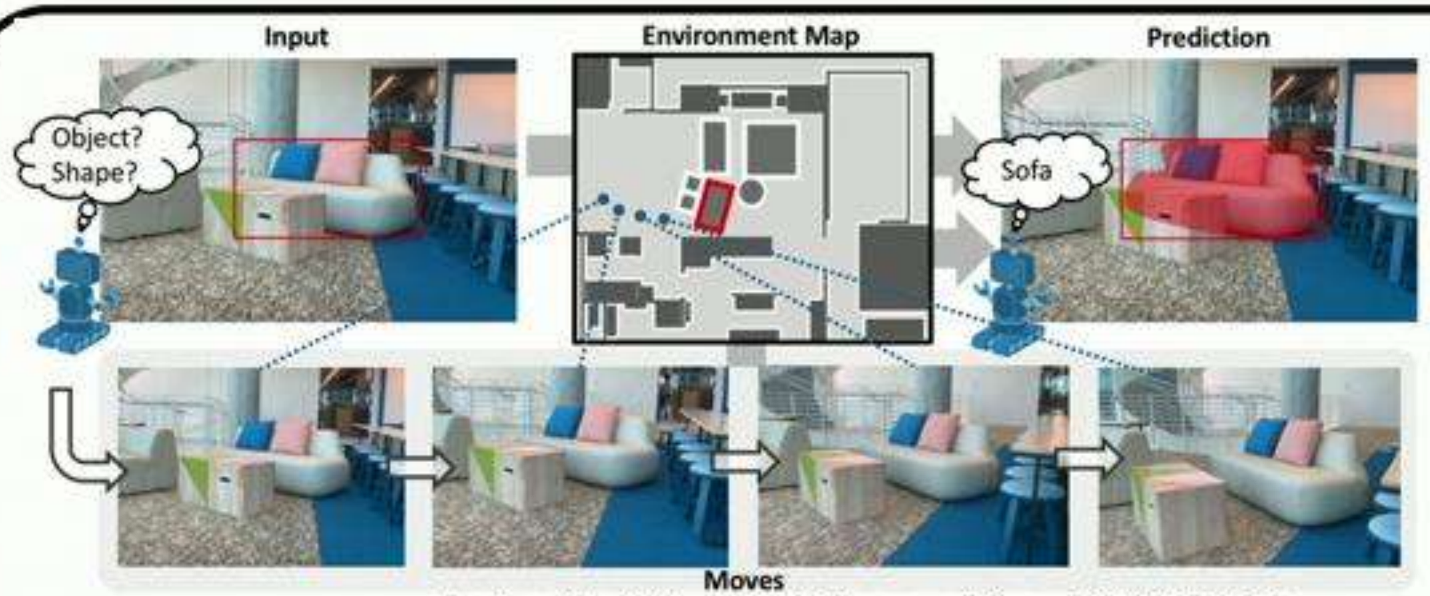
Learn from interactions with human and environment

Per-Image Structure



Graph R-CNN for Scene Graph Generation. ECCV 2018

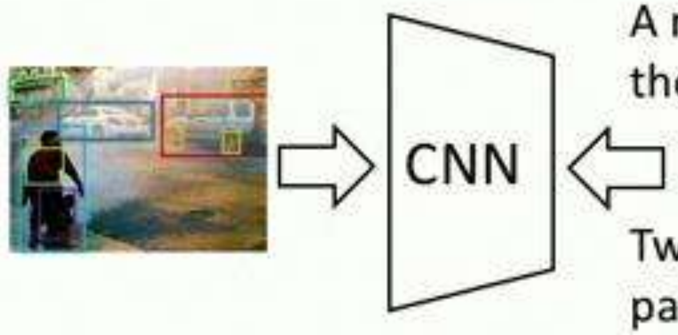
Structured Visual Understanding



Embodied Amodal Recognition. ICCV 2019

Interact with Environment

As Future Works



A man is on the left side

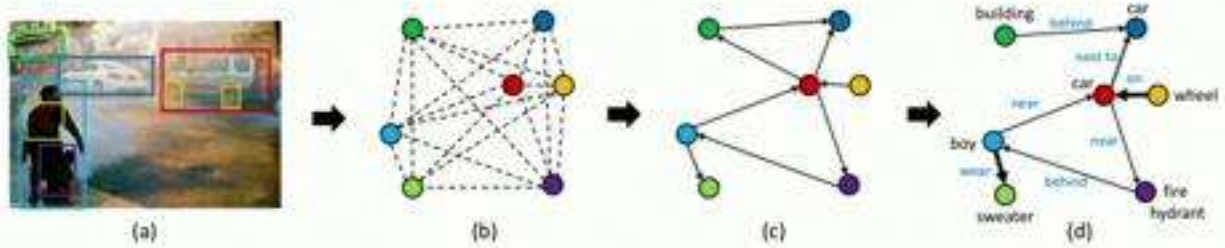
Two cars are parked beside the road

Learning from image corpus

Interact with Human

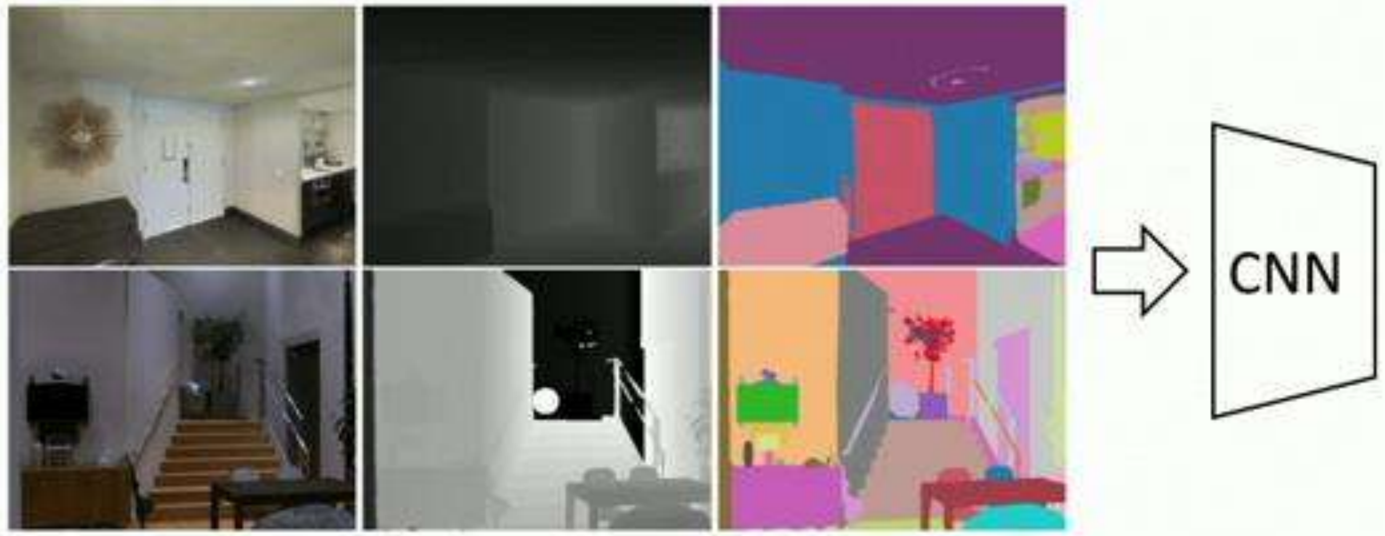
Learn from interactions with human and environment

Per-Image Structure



Graph R-CNN for Scene Graph Generation. ECCV 2018

Structured Visual Understanding



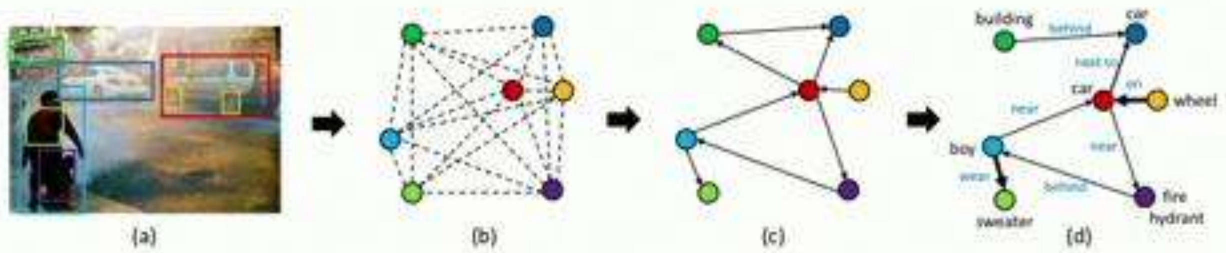
3D structured visual understanding

Interact with Environment

As Future Works

Learn for interactions with human and environment

Per-Image Structure

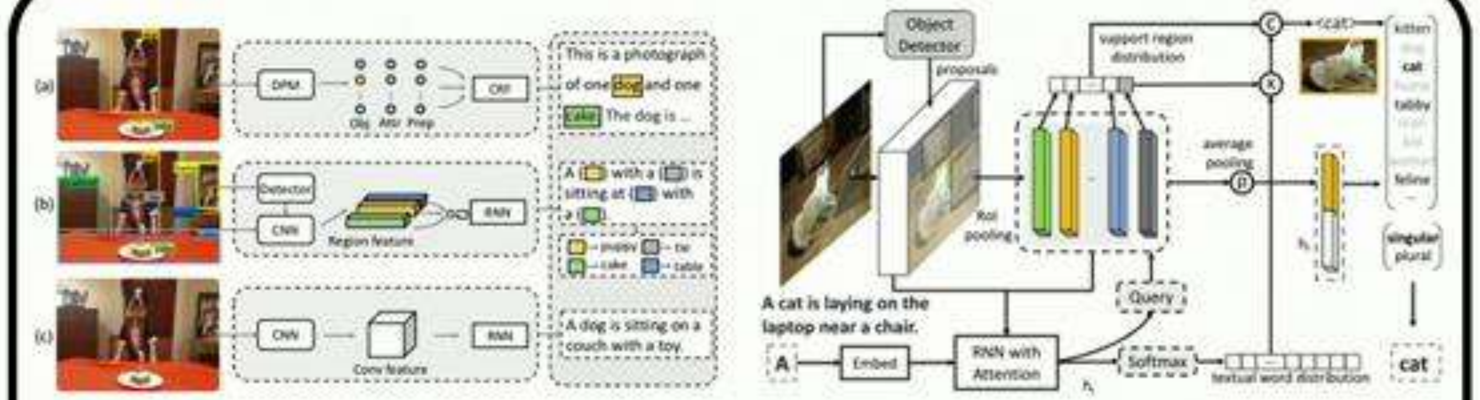


Graph R-CNN for Scene Graph Generation. ECCV 2018

Structured Visual Understanding

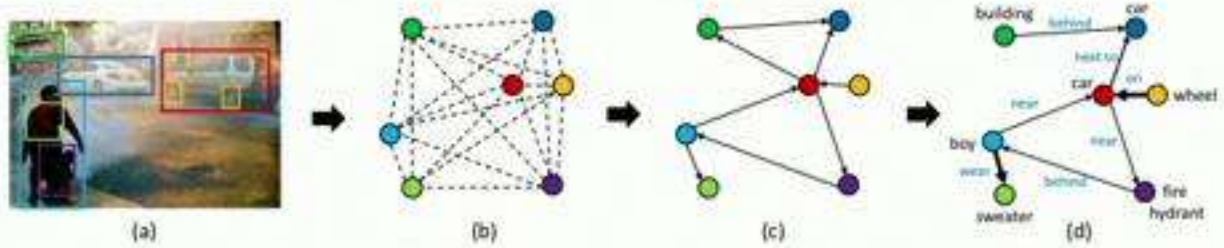
Learr

As Future Works



Neural Baby Talk. CVPR 2018

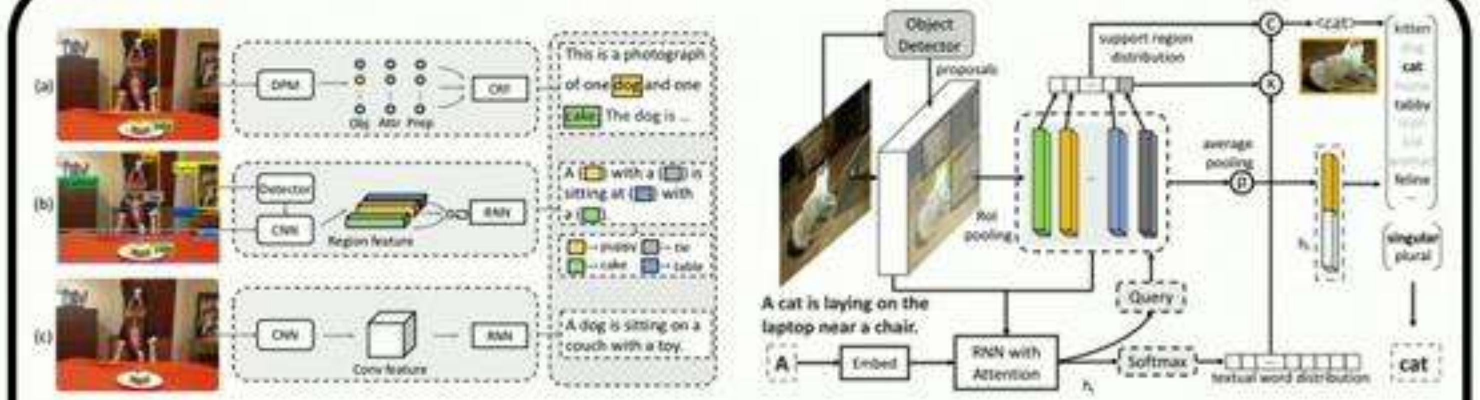
Per-Image Structure



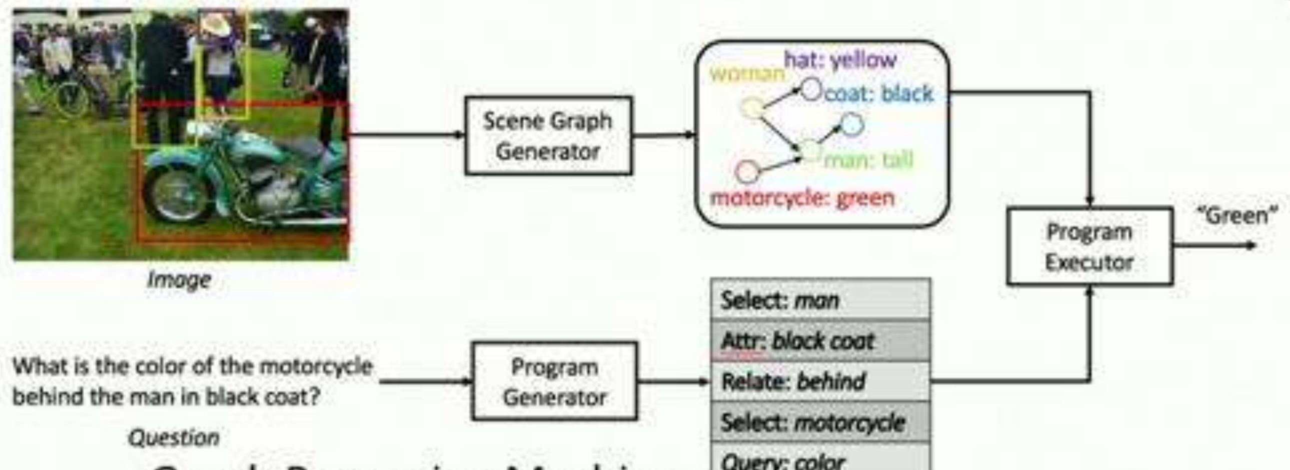
Graph R-CNN for Scene Graph Generation. ECCV 2018

Structured Visual Understanding

As Future Works

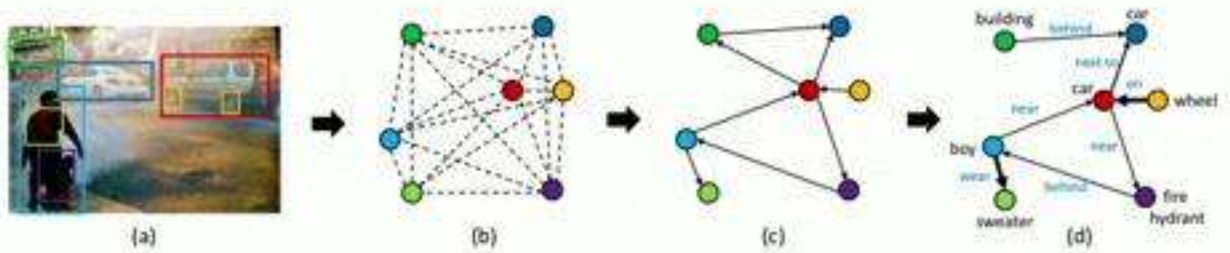


Neural Baby Talk. CVPR 2018



Graph Reasoning Machine

Per-Image Structure

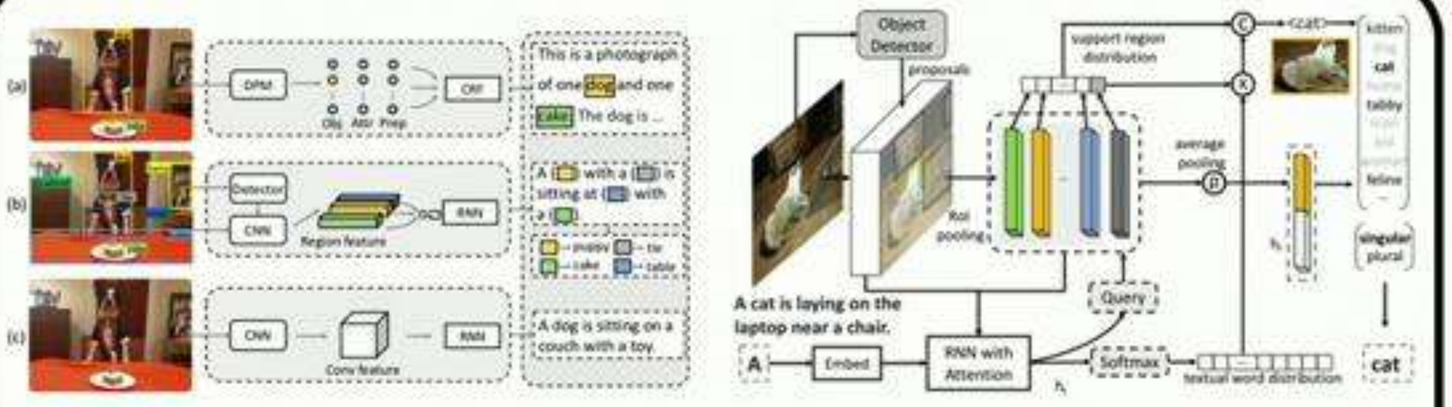


Graph R-CNN for Scene Graph Generation. ECCV 2018

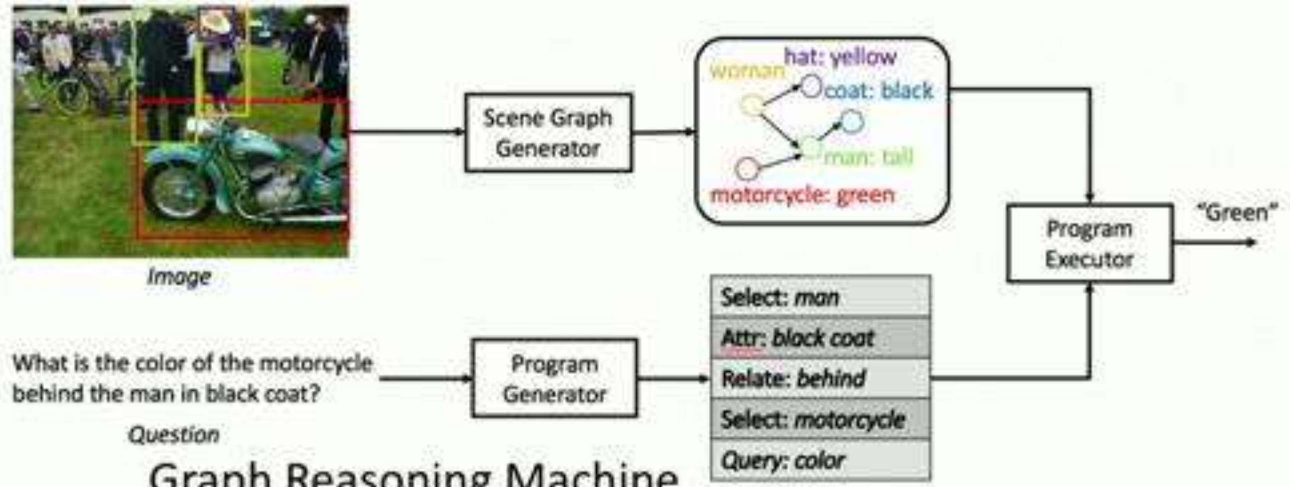
Structured Visual Understanding

Learr

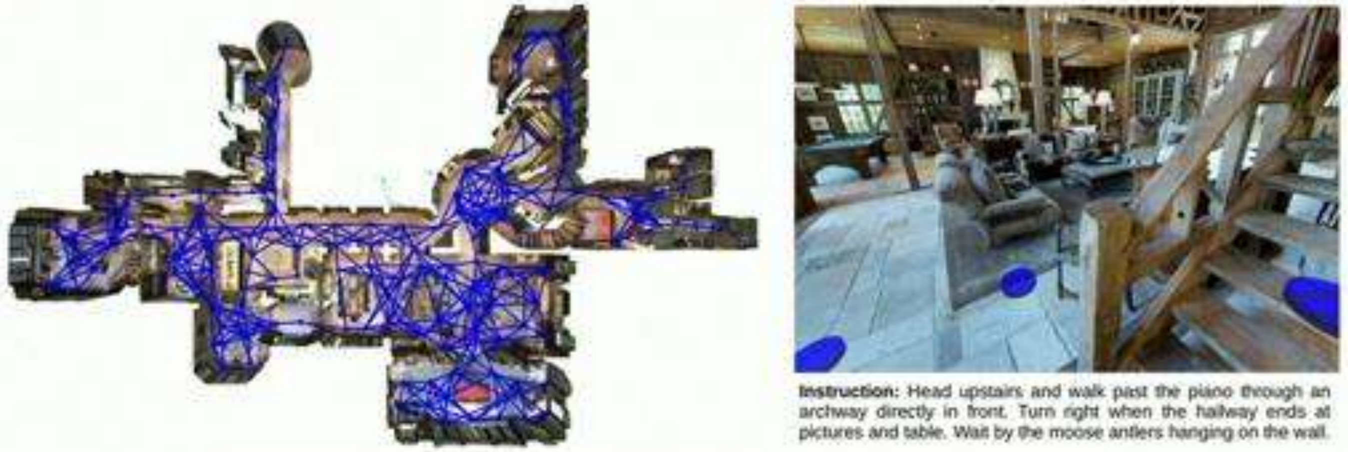
As Future Works



Neural Baby Talk. CVPR 2018

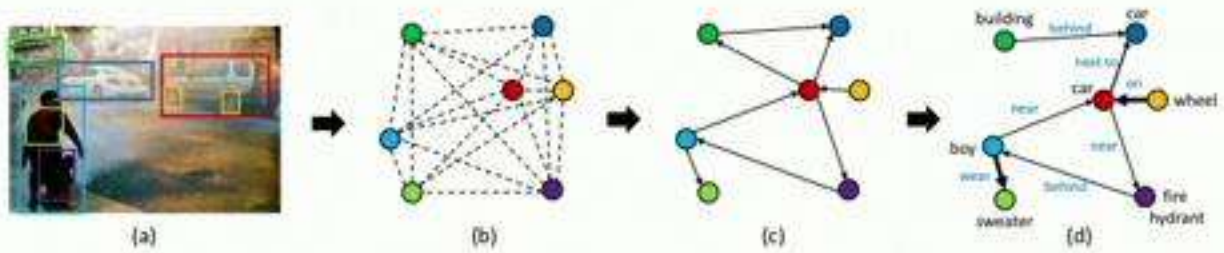


Graph Reasoning Machine



Visual Language Navigation. Anderson et al.

Per-Image Structure

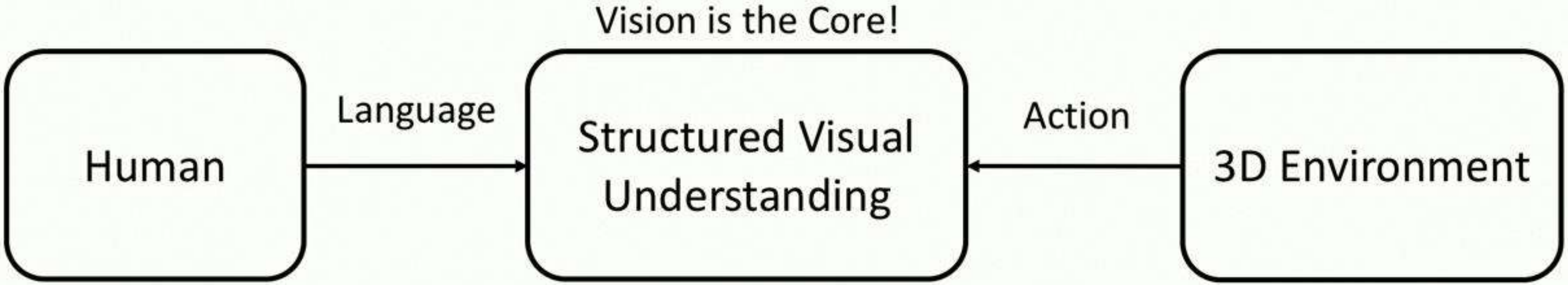


Graph R-CNN for Scene Graph Generation. ECCV 2018

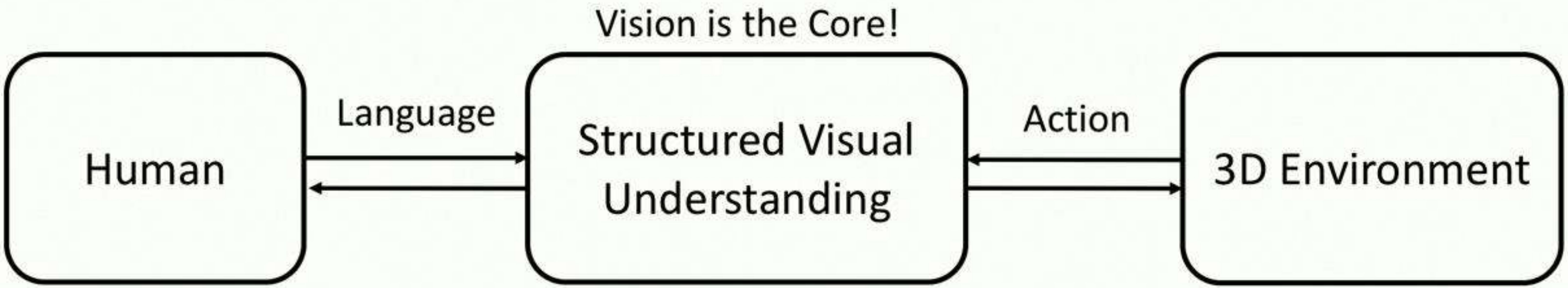
Structured Visual Understanding

Learr

To summarize



To summarize



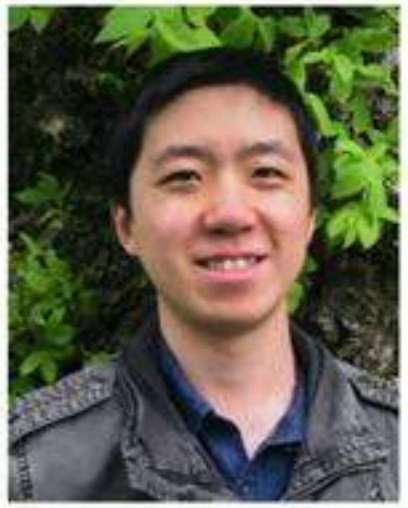
Collaborators

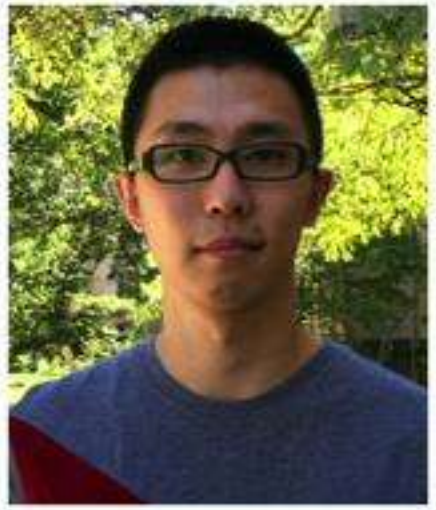

Devi Parikh


Dhruv Batra


Jiasen Lu


Stefan Lee


Zhile Ren


Minze Xu


Anitha Kannan


Xinlei Chen


Ji Lin


Chuang Gan


Hongyuan Zhu


David Crandall

Thanks!