Visual Question Answering (VQA)

Presenter: Thao Minh Le

https://thaolmk54.github.io/

April 17th, 2020

APPLIED ARTIFICIAL INTELLIGENCE INSTITUTE

DEAKIN UNIVERSITY

About Me

- Current a PhD candidate at A2I2, Deakin University.
- Graduated from Tokyo Institute of Technology, Japan (2018) and Hanoi University of Science and Technology, Vietnam (2014).
- Having interests in applications of Machine Learning and Computer Vision.

Agenda

- Introduction to VQA and its applications
- VQA models
- Our contributions to VQA

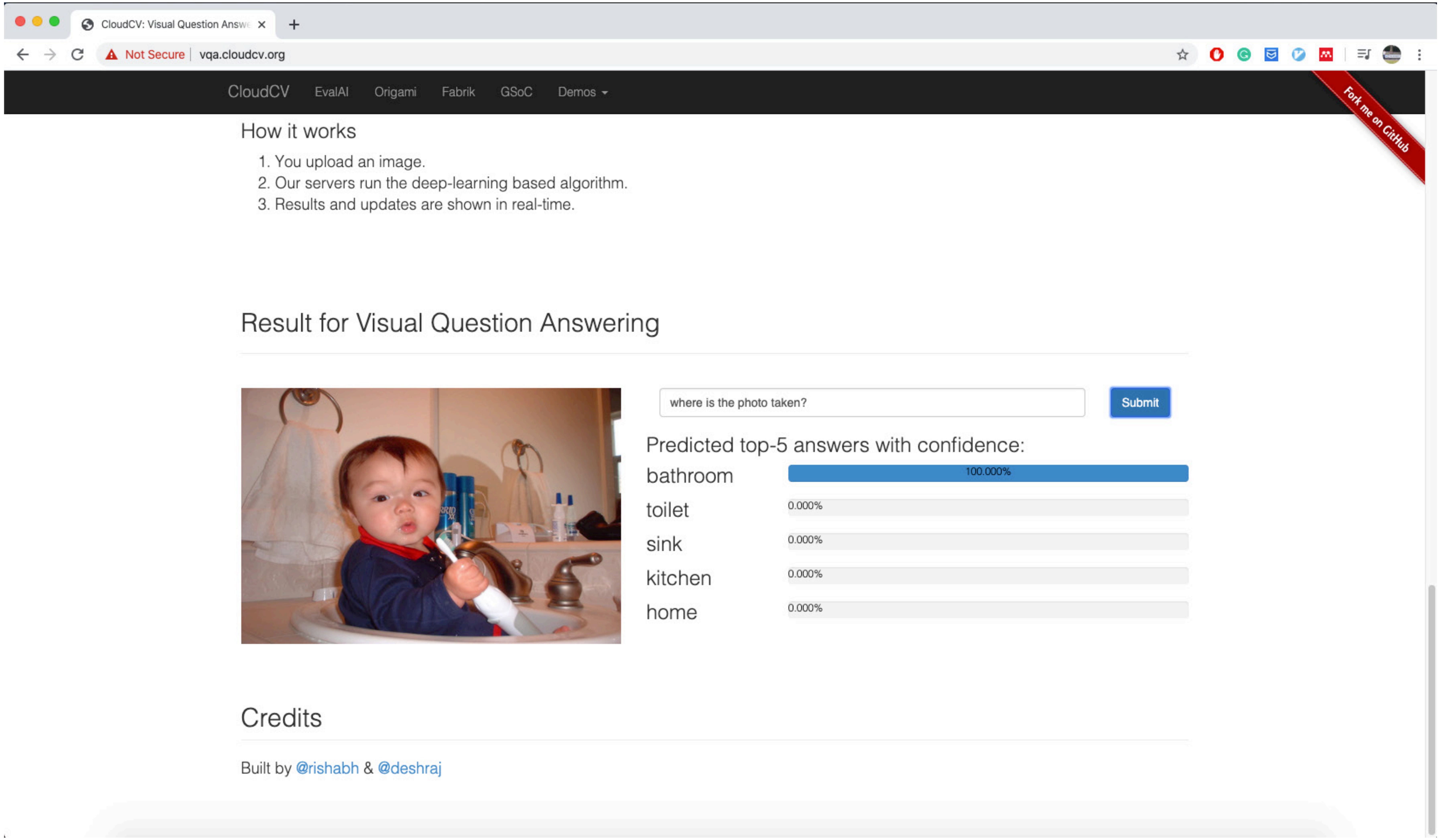
VQA Task

Question
What is the brown animal sitting inside of?

AI System

box

Try VQA yourself



Why Vision + Language?

Pictures/videos are everywhere.

Words are how humans communicate.

Slide credit: Devi Parikh

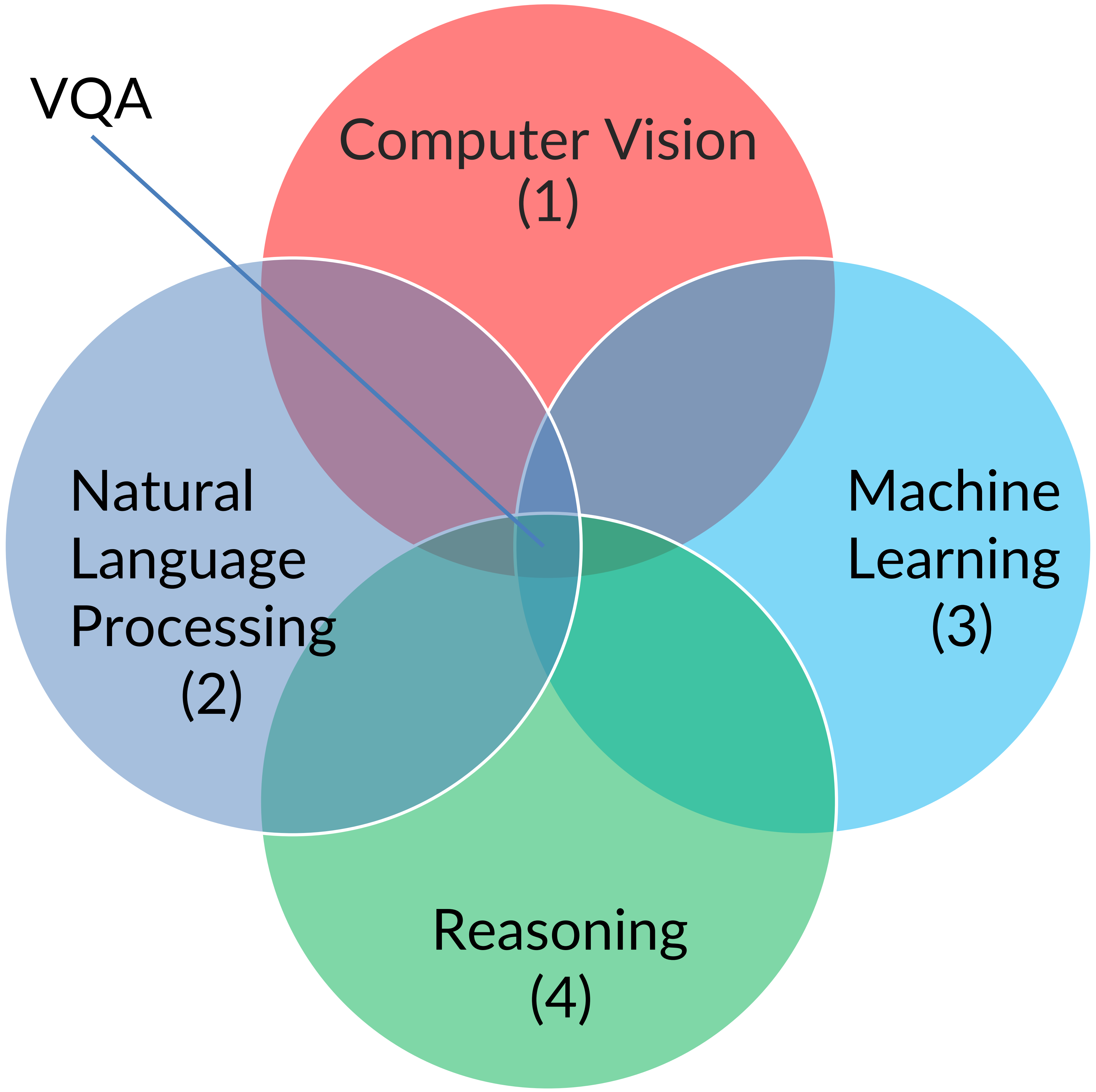
Why VQA Is an AI Testbed?

Question: What can the red object on the ground be used for ? (2)

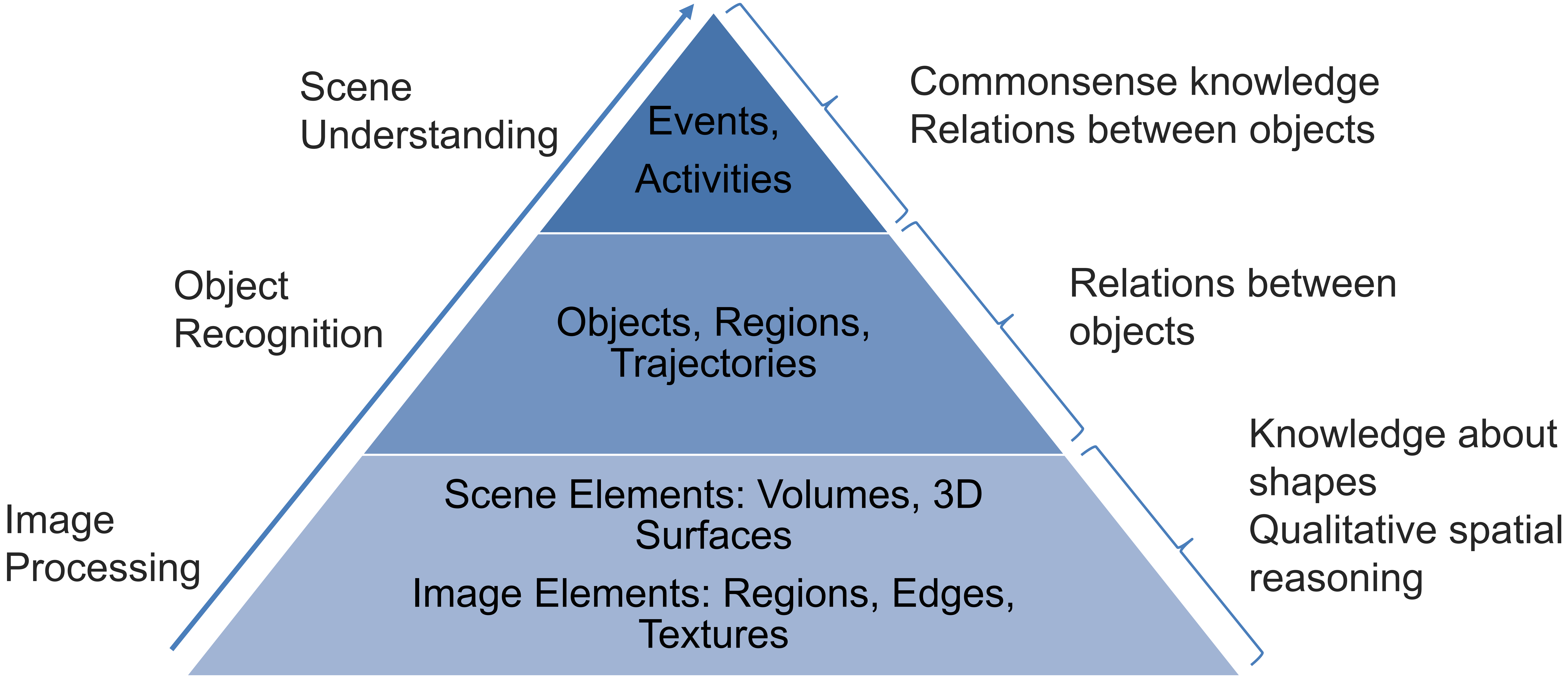
Answer: Firefighting

Support Fact: Fire hydrant can be used for fighting fires. (2, 4)

Wang, Peng, et al. "Fvqa: Fact-based visual question answering." TPAMI 2018



Why VQA Is an AI Testbed?



Adapted from [Somak et al., 2019]

Applications of VQA

- Aid visually-impaired users

Are the any obstacles coming to me?

Image credit: ARIA

Applications of VQA

- Surveillance and visual data summarization

What did the man in red shirt do before entering the building?

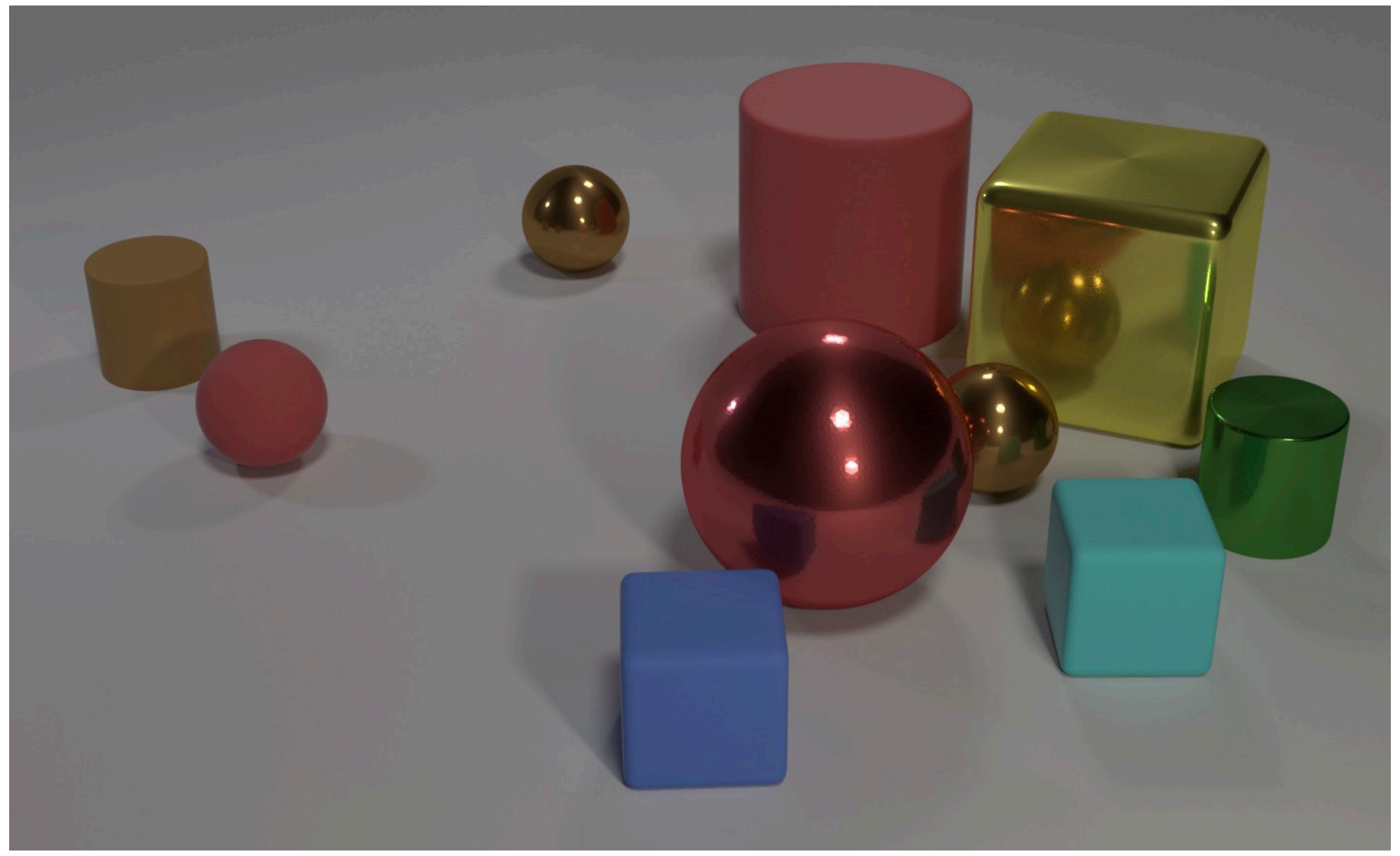
Image credit: journalistsresource.org

VQA Datasets: Image QA

(VQA, Agrawal et al., 2015)

(Q) How many slices of pizza are there?
(Q) Is this a vegetarian pizza?

(CLEVR, Johnson et al., 2017)



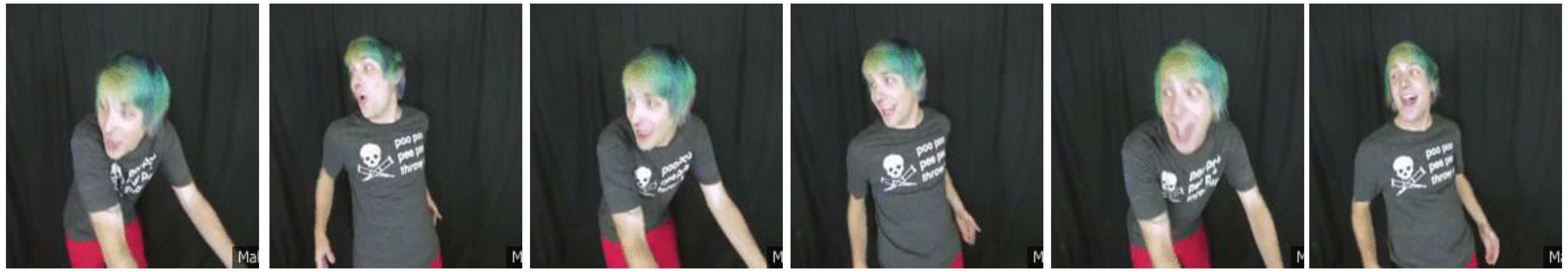
(Q) How many objects are either small cylinders or metal things?
(Q) Are there an equal number of large things and metal spheres?

(GQA, Hudson et al., 2019)

(Q) What is the brown animal sitting inside of?
(Q) Is there a bag to the right of the green door?

VQA Datasets: Video QA

(TGIF-QA, Jang et al., 2018)



Q: What does the man do 5 times?

A: (0) step (3) bounce

(2) sway head (4) knod head

(5): move body to the front

Q: What does the man do before turing body to left?

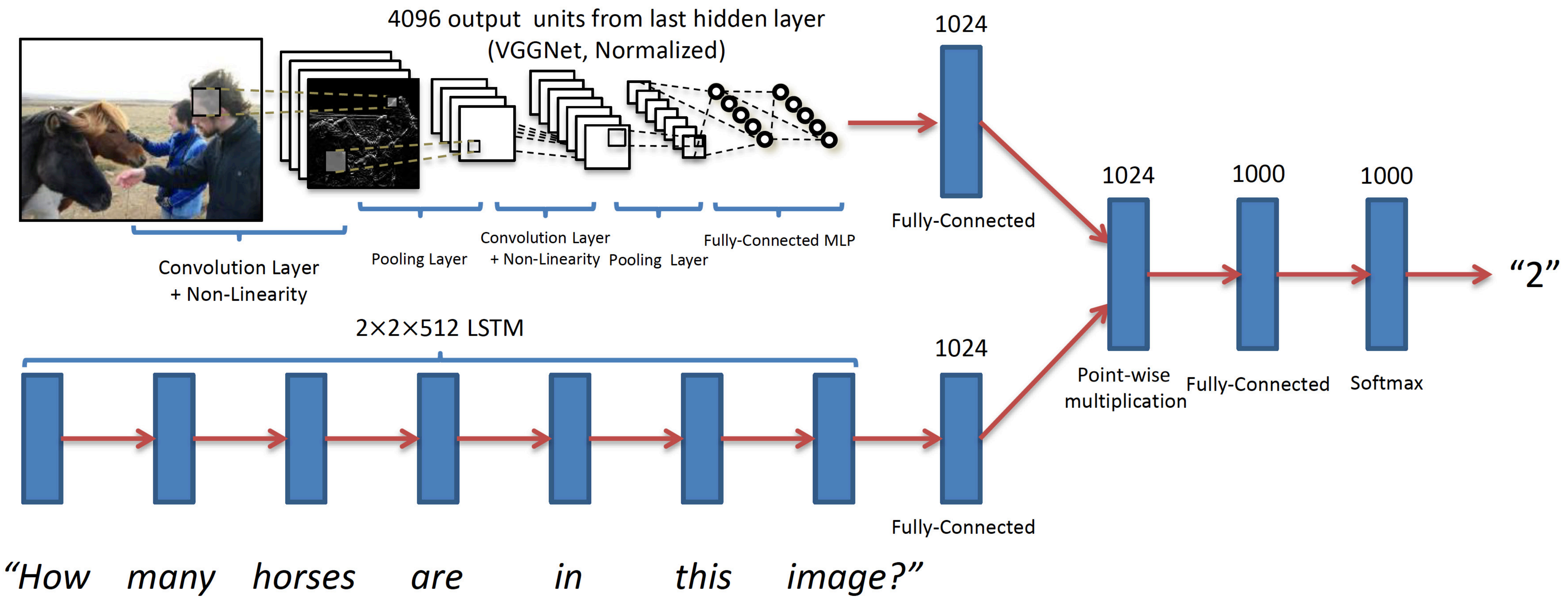
A: (0) run a cross a ring (3) flip cover face with hand

(2) pick up the man's hand (4) raise hand

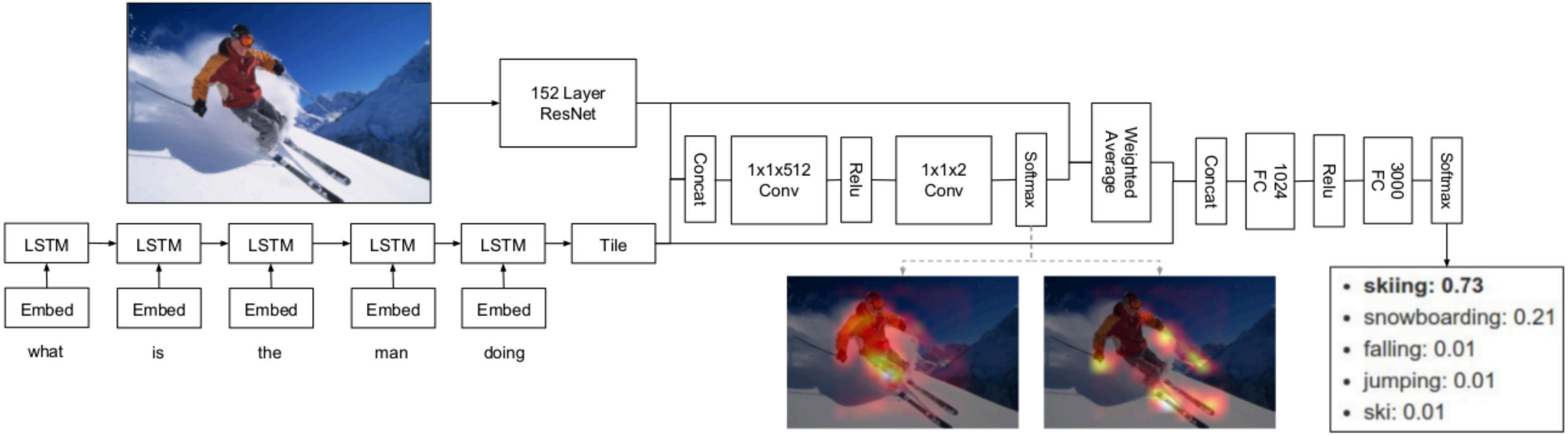
(5): breath

VQA models

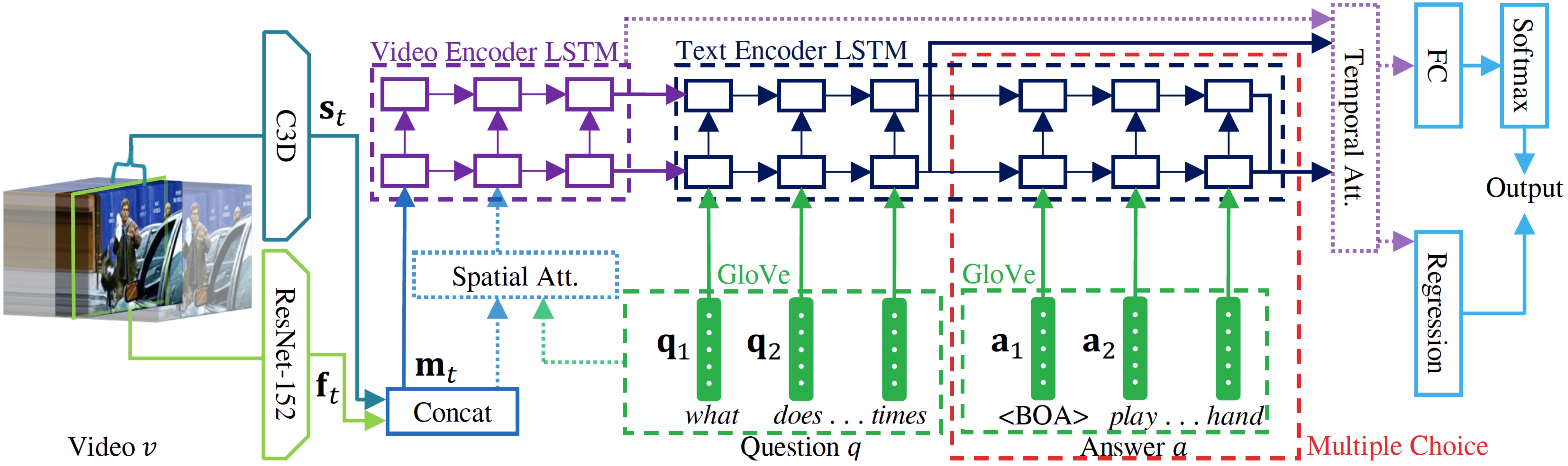
[Image QA, Agrawal et al., 2015]



[Image QA, Kazemi et al., 2017]



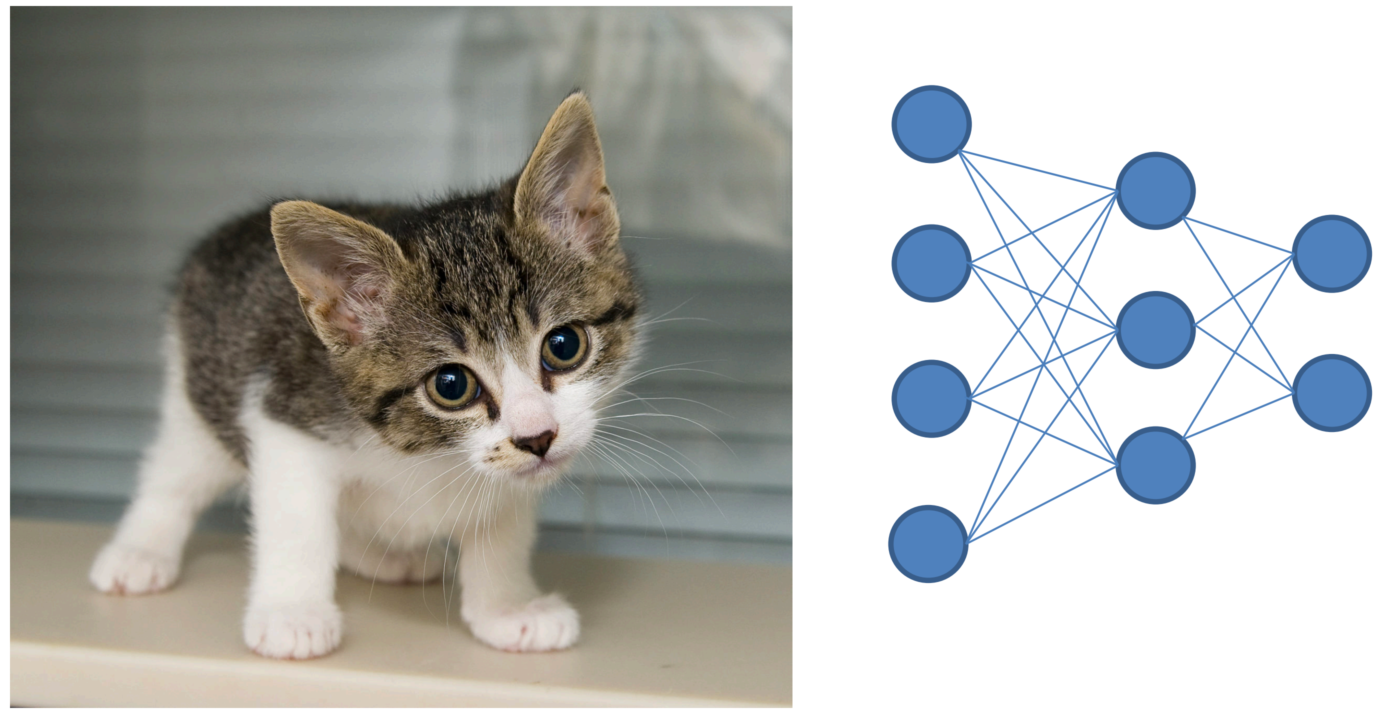
[Video QA, Jang et al., 2018]



Our contributions to VQA

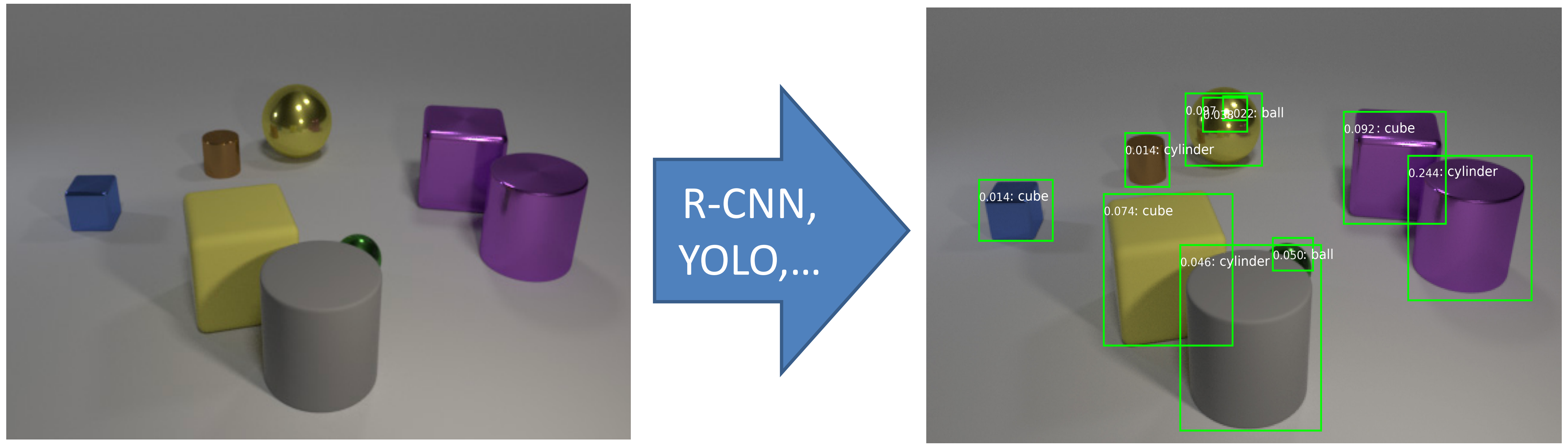
Our Focus: Visual Reasoning

From recognition to visual reasoning



a cat

What is this?



Where do the objects locate?

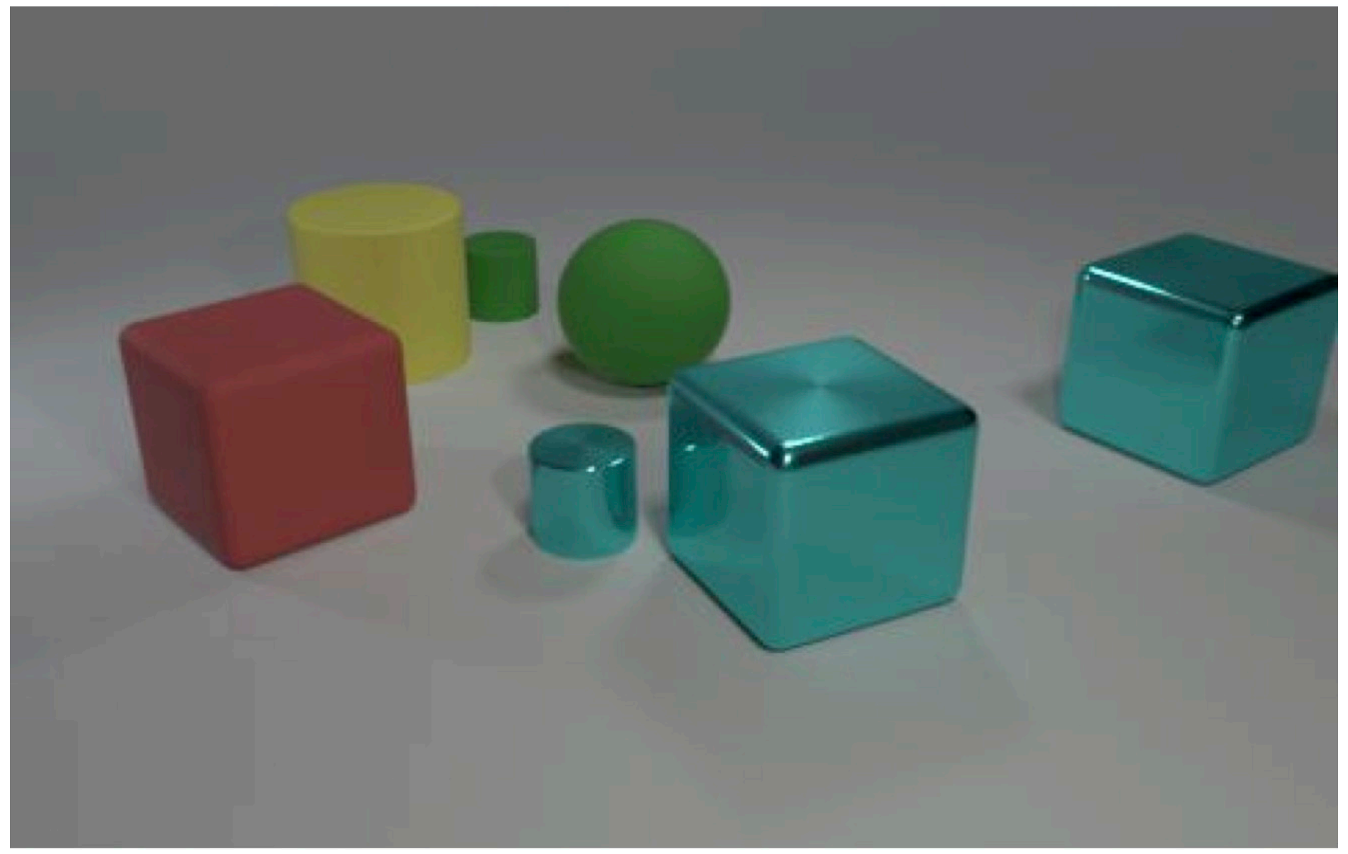
Object recognition

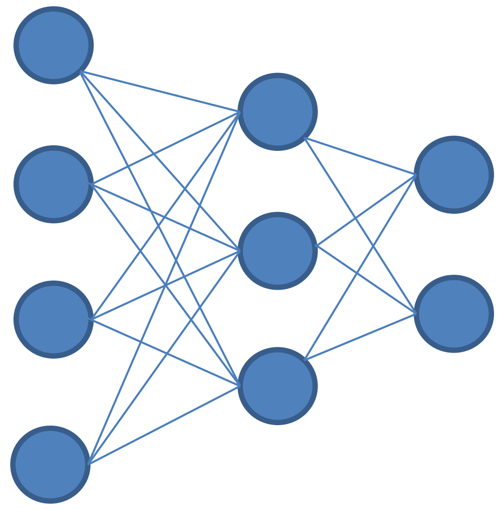
Object detection

Image courtesy: https://dcist.com/

Our Focus: Visual Reasoning

Why things do not go well?





cylinder

What color is the thing with the same size as the blue cylinder?

- The network guessed the most common color in the image.
- Linguistic bias.
- Requires ***multi-step reasoning***: find blue cylinder ➔ locate another object of the same size ➔ determine its color (green).

Reasoning is to deduce knowledge from previously acquired knowledge in response to a query (or a cue) [Roni et al., 1997]

Relational Reasoning in Image QA

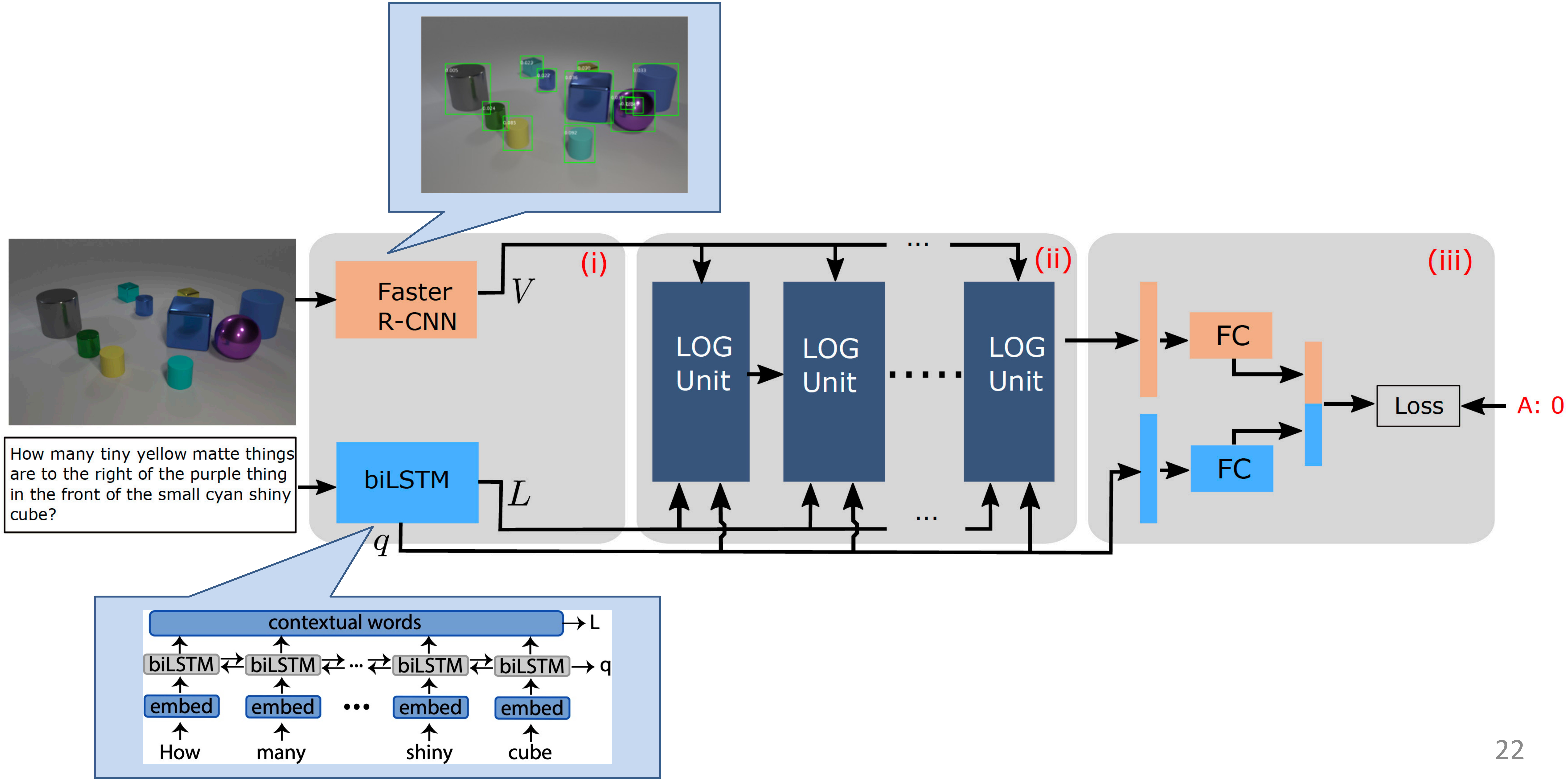
Thao Minh Le, Vuong Le, Svetha Venkatesh and Truyen Tran, “Dynamic Language Binding in Relational Visual Reasoning”, *Under review at IJCAI’20*.

Reasoning with Structured Representation of Spatial Relations

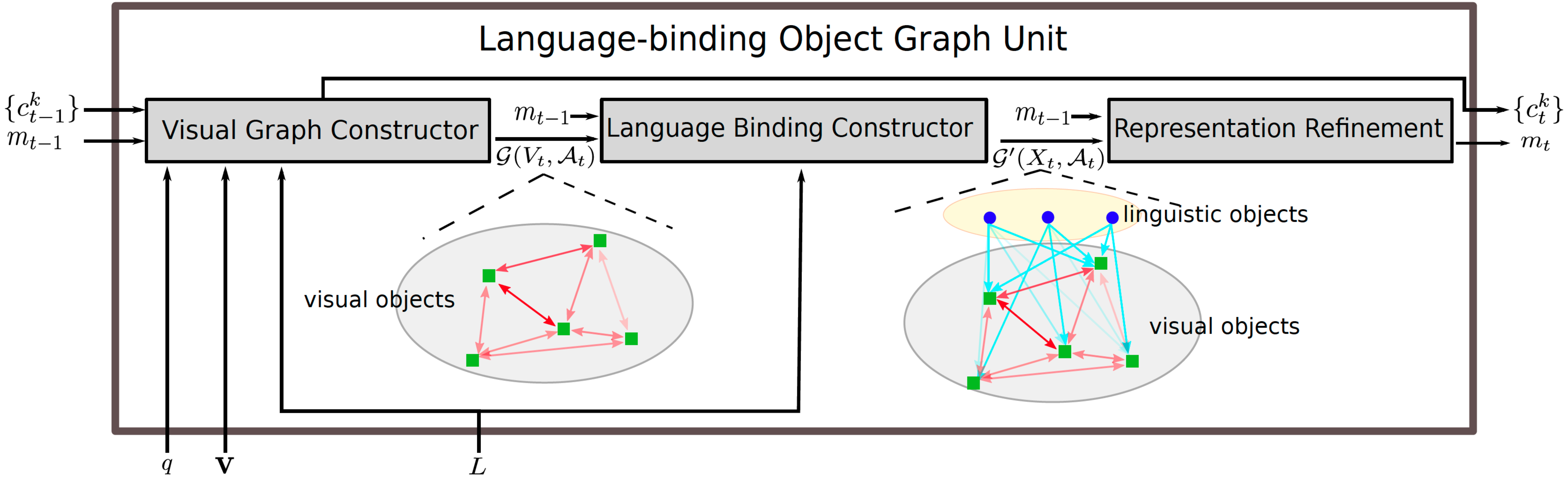
Key insight: *Reasoning is chaining of relational predicates to arrive at a final conclusion*

- Needs to uncover spatial relations, conditioned on query
- Chaining is query-driven
- Objects/language need(s) binding
- Object semantics are query-dependent
- Everything is end-to-end differentiable

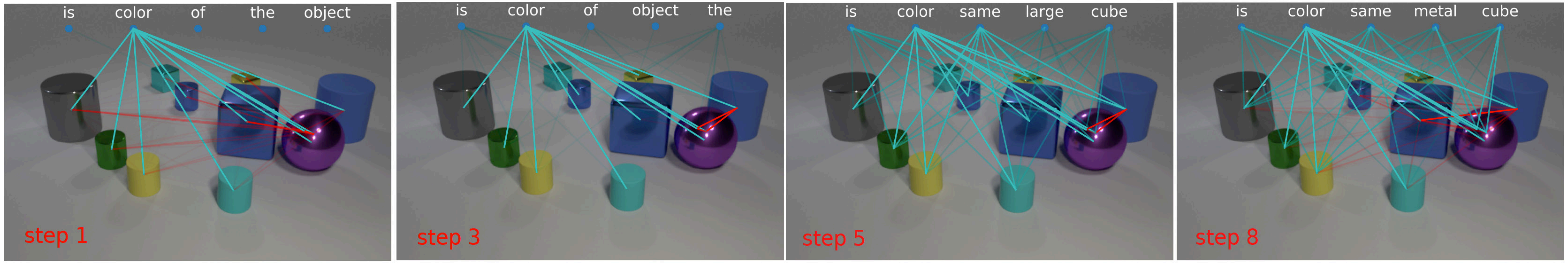
Language-binding Object Graph Model for VQA



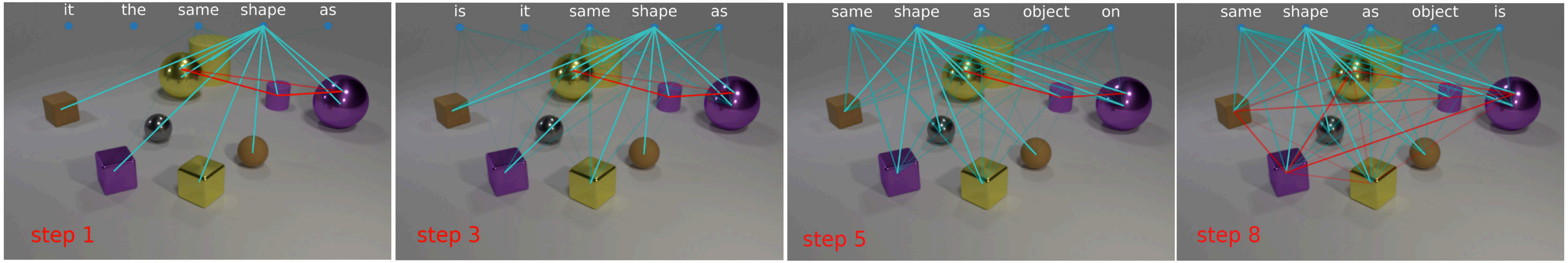
Language-binding Object Graph Unit (LOG)



LOGNet's Output

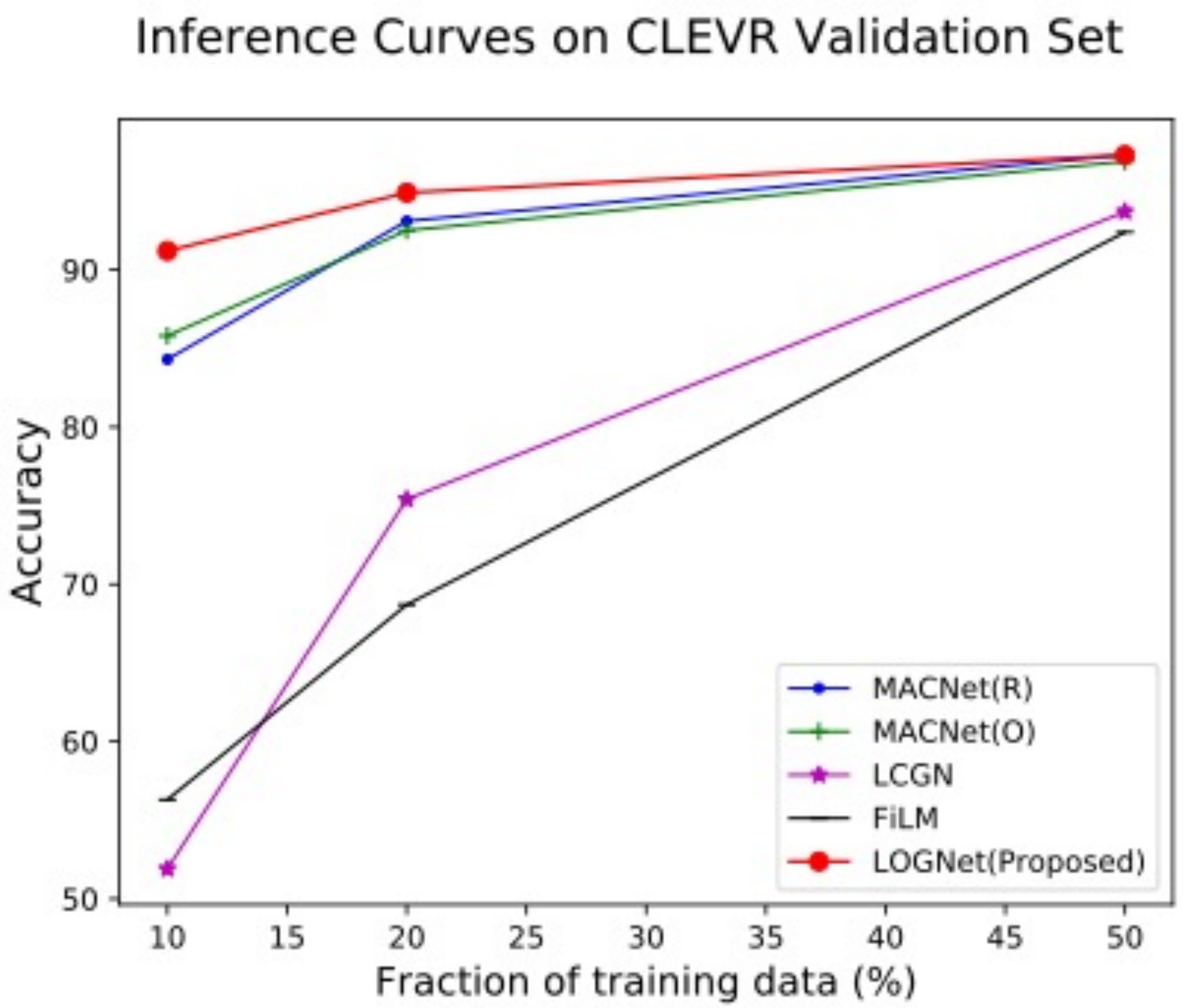


Question: Is the color of the big matte object the same as the large metal cube?
Prediction: yes **Answer**: yes



Question: There is a tiny purple rubber thing; does it have the same shape as the brown object that is on the left side of the rubber sphere?
Prediction: no **Answer**: no

Results



Comparison with SOTAs on CLEVR dataset of different data fractions.

Method	Val. Acc. (%)
FiLM	56.6
MACNet(R)	57.4
LCGN [Hu *et al.*, 2019]	46.3
BAN [Shrestha *et al.*, 2019]	60.2
RAMEN [Shrestha *et al.*, 2019]	57.9
LOGNet	**62.3**

Performance comparison on CLEVR-Human.

Results

Method	Accuracy (%)	
	val	test
Full training data		
CNN+LSTM	49.2	46.6
Bottom-Up [Anderson *et al.*, 2018]	52.2	49.7
MACNet(O)	57.5	54.1
LCGN [Hu *et al.*, 2019]	63.9	56.1
LOGNet	63.3	55.2
Subset 50% training data		
LCGN	60.6	-
LOGNet	60.7	-
Subset 20% training data		
LCGN	53.2	-
LOGNet	55.6	-

Performance on GQA

Method	Val. Acc. (%)
XNM [Shi *et al.*, 2019]	43.4
MACNet(R)	40.7
MACNet(O)	45.5
LOGNet	**46.8**

Performance on

VQA v2 subset of long questions

Relational Reasoning in Video QA

Thao Minh Le, Vuong Le, Svetha Venkatesh and Truyen Tran, “Hierarchical conditional relation networks for video question answering”, *CVPR’20 (Oral)*.

Conditional Relation Network Unit

Motivations:

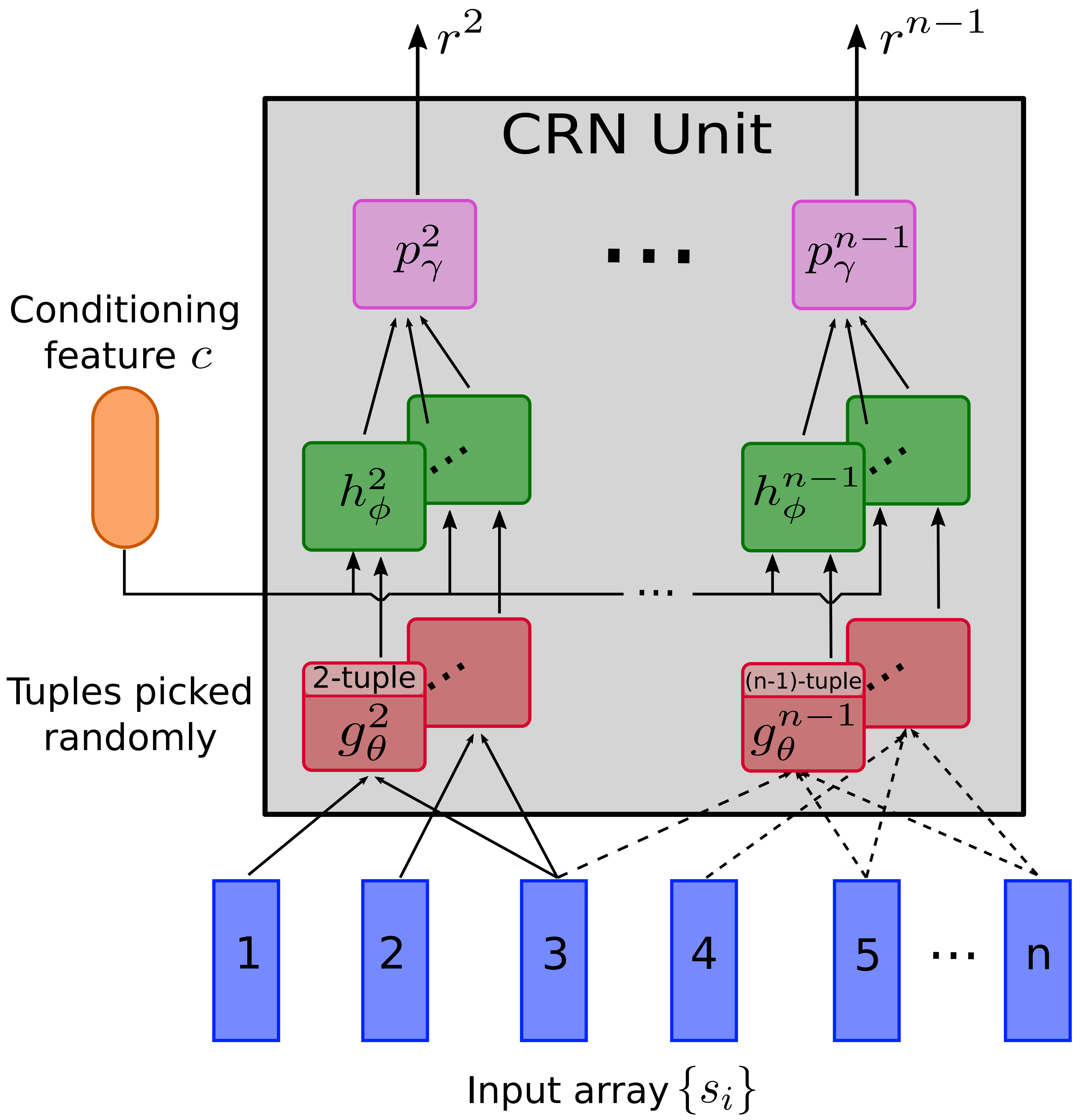
- Lack of a **generic mechanism** in SOTA methods for modelling the **interaction of multimodal inputs**.
- Reflecting the natural **characteristics of videos** (long-short temporal relations, hierarchy, compositionality).

Inputs:

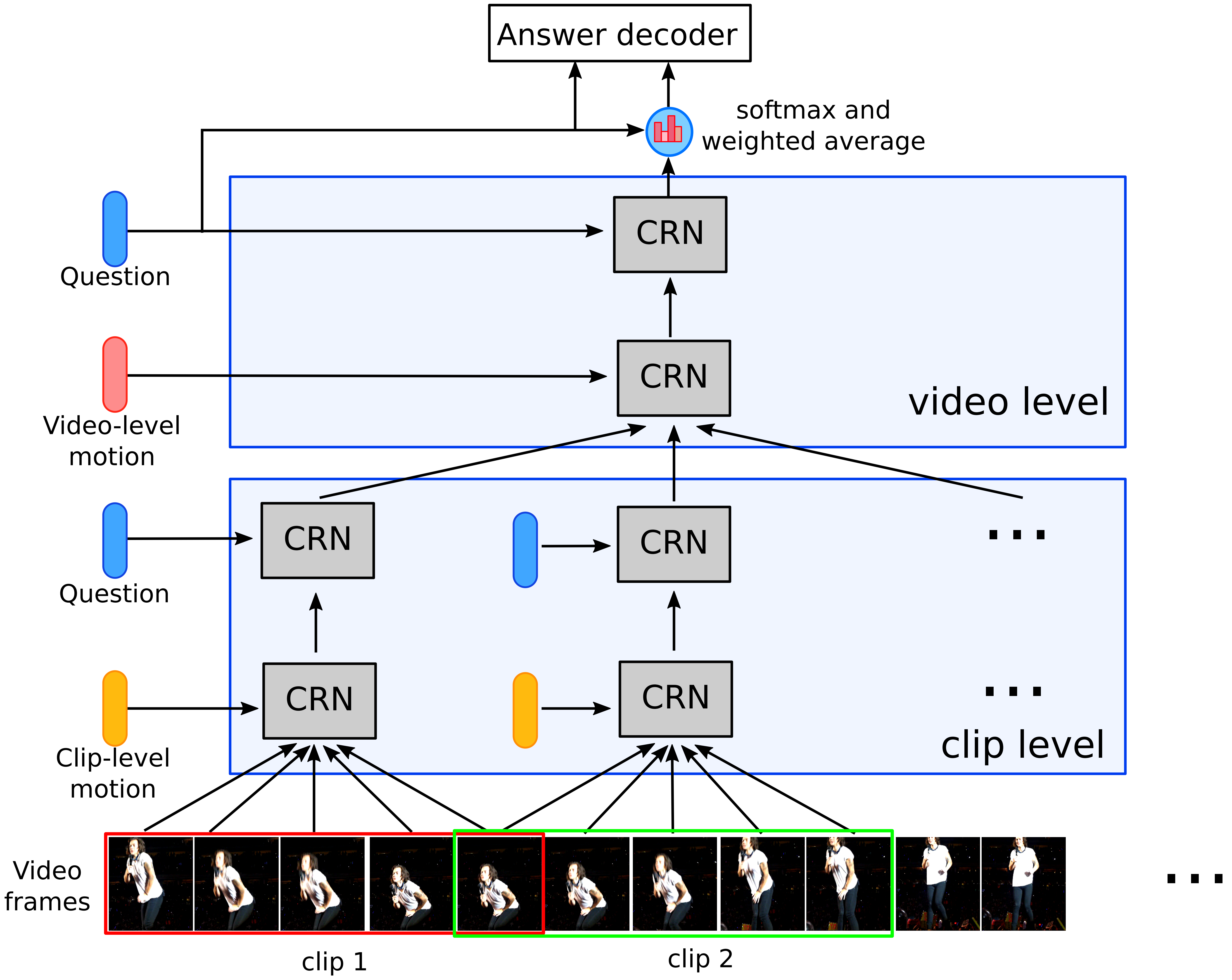
- An array of n objects
- Conditioning feature

Outputs:

- An array of *m* (*m*<*n*) objects



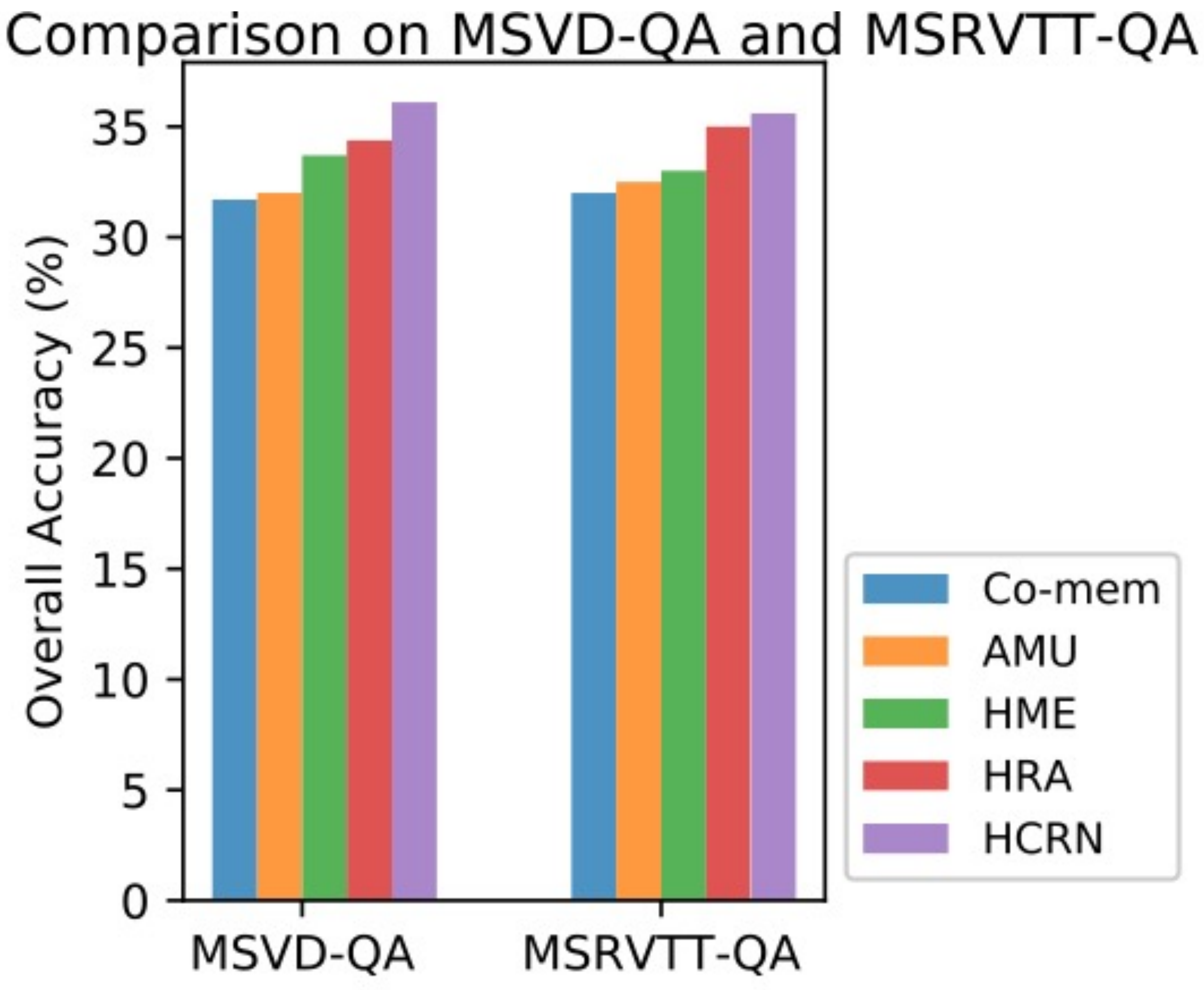
Hierarchical Conditional Relation Networks for Video QA



Results

Model	Action	Trans.	FrameQA	Count
ST-TP	62.9	69.4	49.5	4.32
Co-Mem	68.2	74.3	51.5	4.10
PSAC	70.4	76.9	55.7	4.27
HME	73.9	77.8	53.8	4.02
HCRN	**75.0**	**81.4**	**55.9**	**3.82**

TGIF-QA dataset



Results

Ablation studies on

TGIF-QA dataset

Model	Act.	Trans.	F.QA	Count
Relations (k_{max}, t)				
$k_{max} = 1, t = 1$	65.2	75.5	54.9	3.97
$k_{max} = 1, t = 3$	66.2	76.2	55.7	3.95
$k_{max} = 1, t = 5$	65.4	76.7	56.0	3.91
$k_{max} = 1, t = 9$	65.6	75.6	56.3	3.92
$k_{max} = 1, t = 11$	65.4	75.1	56.3	3.91
$k_{max} = 2, t = 2$	67.2	76.6	56.7	3.94
$k_{max} = 2, t = 9$	66.3	76.7	56.5	3.92
$k_{max} = 4, t = 2$	64.0	75.9	56.2	3.87
$k_{max} = 4, t = 9$	66.3	75.6	55.8	4.00
$k_{max} = \lfloor n/2 \rfloor, t = 2$	73.3	81.7	56.1	3.89
$k_{max} = \lfloor n/2 \rfloor, t = 9$	72.5	81.1	56.6	3.82
$k_{max} = n - 1, t = 1$	75.0	81.4	55.9	3.82
$k_{max} = n - 1, t = 3$	75.1	81.5	55.5	3.91
$k_{max} = n - 1, t = 5$	73.6	82.0	54.7	3.84
$k_{max} = n - 1, t = 7$	75.4	81.4	55.6	3.86
$k_{max} = n - 1, t = 9$	74.1	81.9	54.7	3.87
Hierarchy				
1-level, video CRN only	66.2	78.4	56.6	3.94
1.5-level, clips→pool	70.4	80.5	56.6	3.94
Motion conditioning				
w/o motion	70.8	79.8	56.4	4.38
w/o short-term motion	74.9	82.1	56.5	4.03
w/o long-term motion	75.1	81.3	56.7	3.92
Linguistic conditioning				
w/o linguistic condition	66.5	75.7	56.2	3.97
w/o quest.@clip level	74.3	81.1	55.8	3.95
w/o quest.@video level	74.0	80.5	55.9	3.92
Gating				
w/o gate	74.1	82.0	55.8	3.93
w/ gate quest. & motion	73.3	80.9	55.3	3.90
Full 2-level HCRN	75.1	81.2	55.7	3.88

THANK YOU FOR LISTENING

Q&A