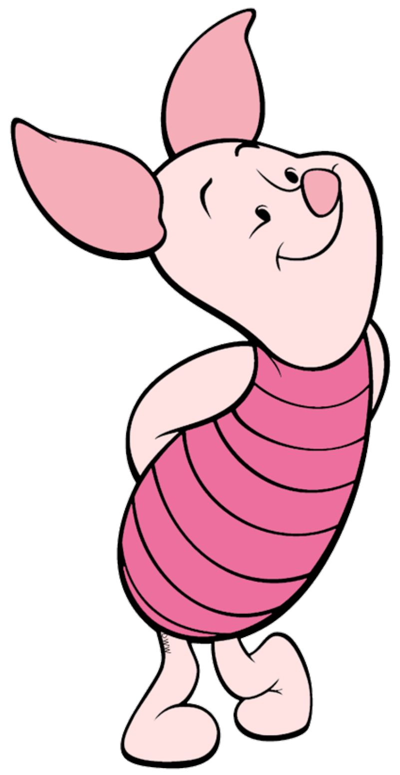
Grounded Causal Commonsense Reasoning

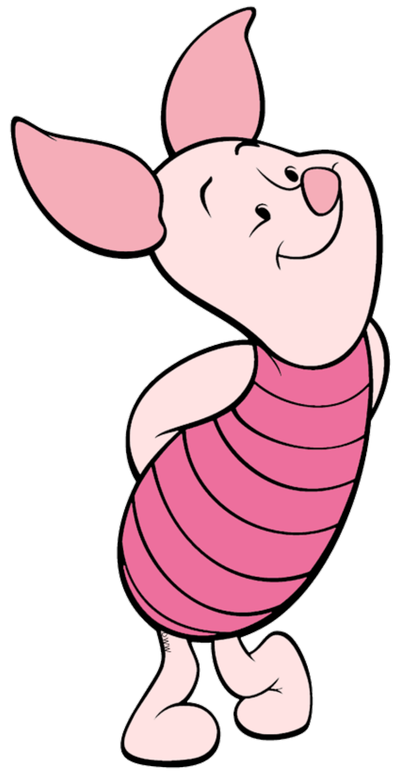


Yejin Choi

Paul G. Allen School of Computer Science & Engineering
University of Washington &
Allen Institute for Artificial Intelligence

Harnad's Symbol Grounding Problem

- Grounding with 3D



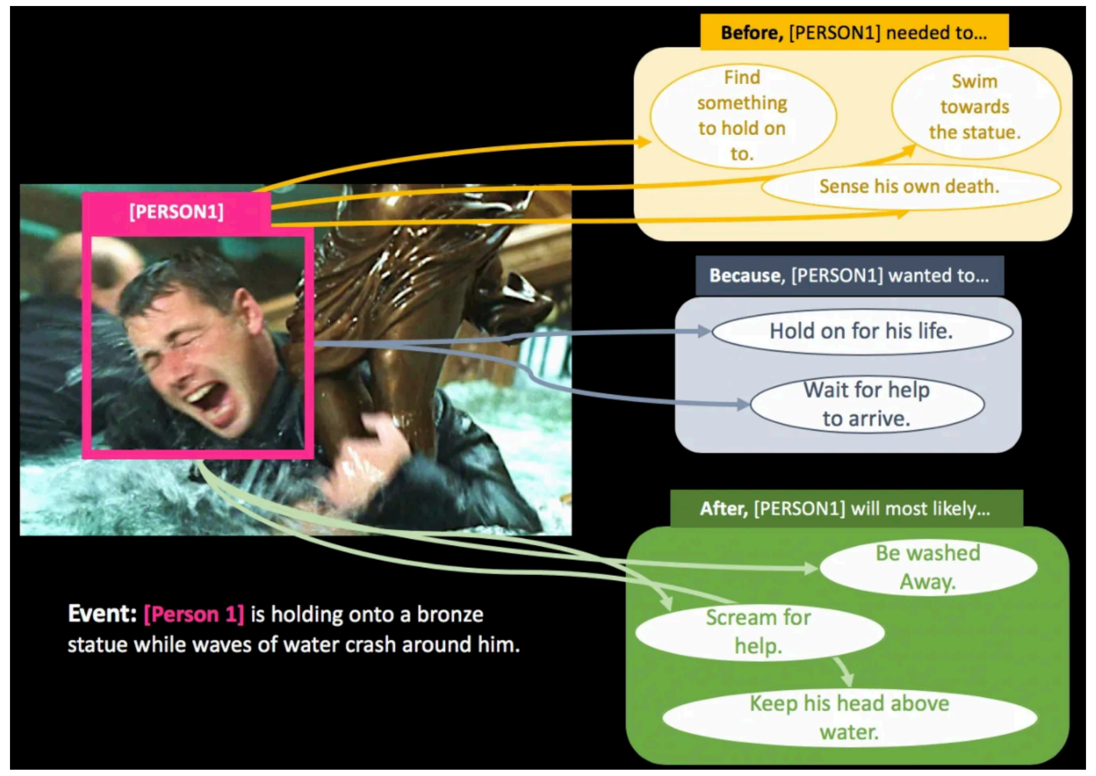
PIGLeT

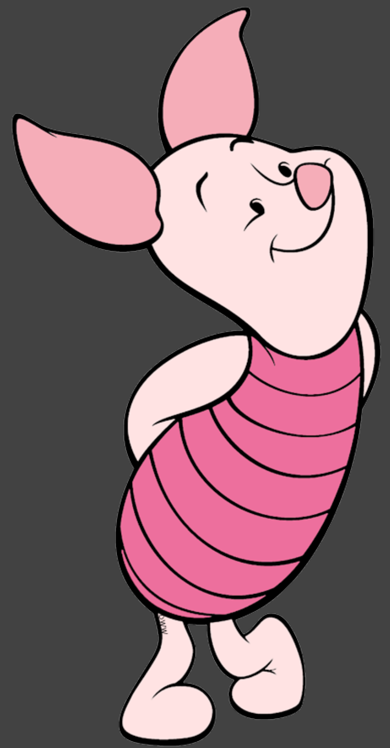
- Grounding with 2D + Time

MERLOT

- Grounding with 2D + KG

Visual Comet





PIGLeT

Language Grounding Through Neuro-Symbolic Interaction in a 3D World

ACL 2021

Rowan Zellers

Ari Holtzman

Matthew Peters

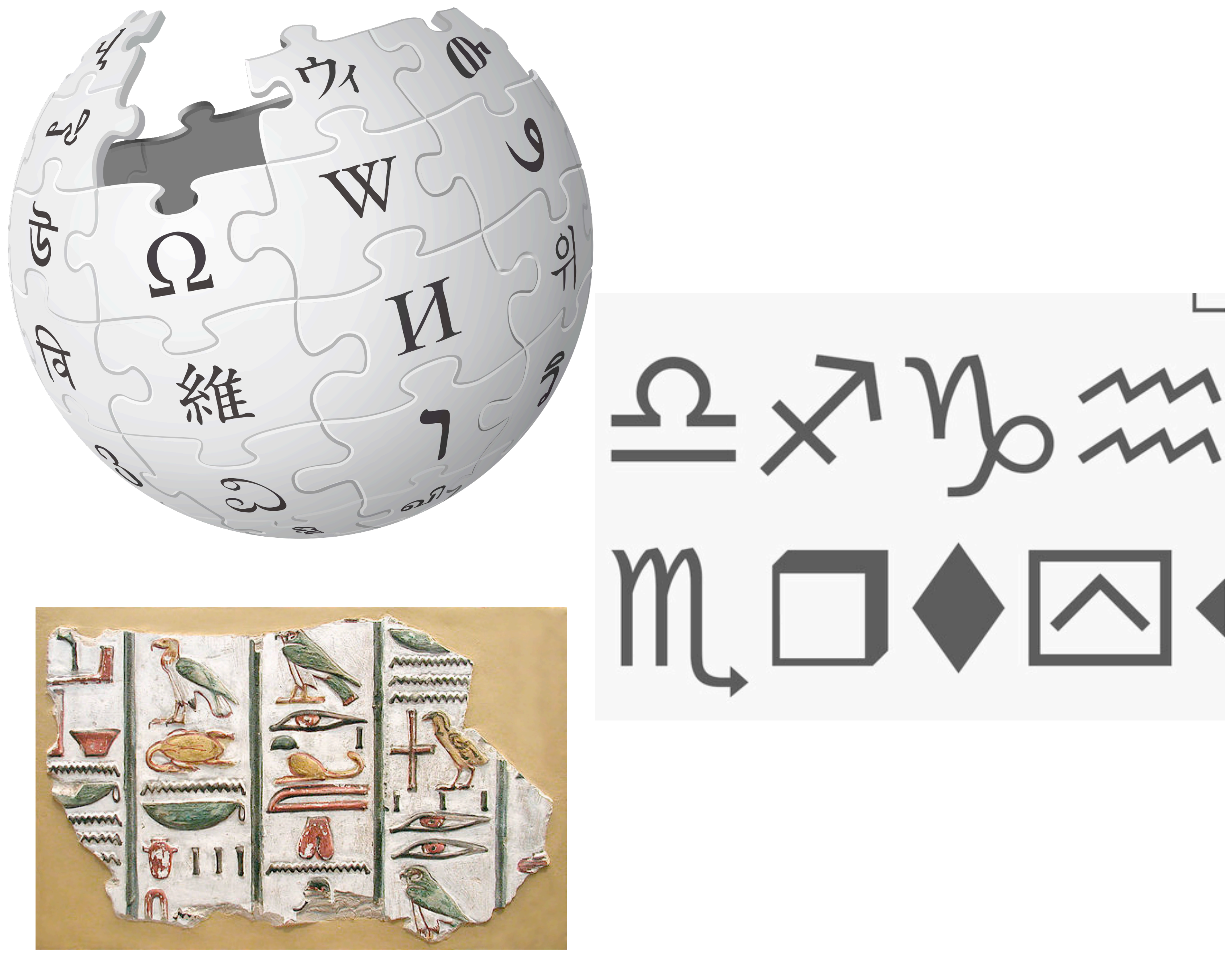
Roozbeh Mottaghi

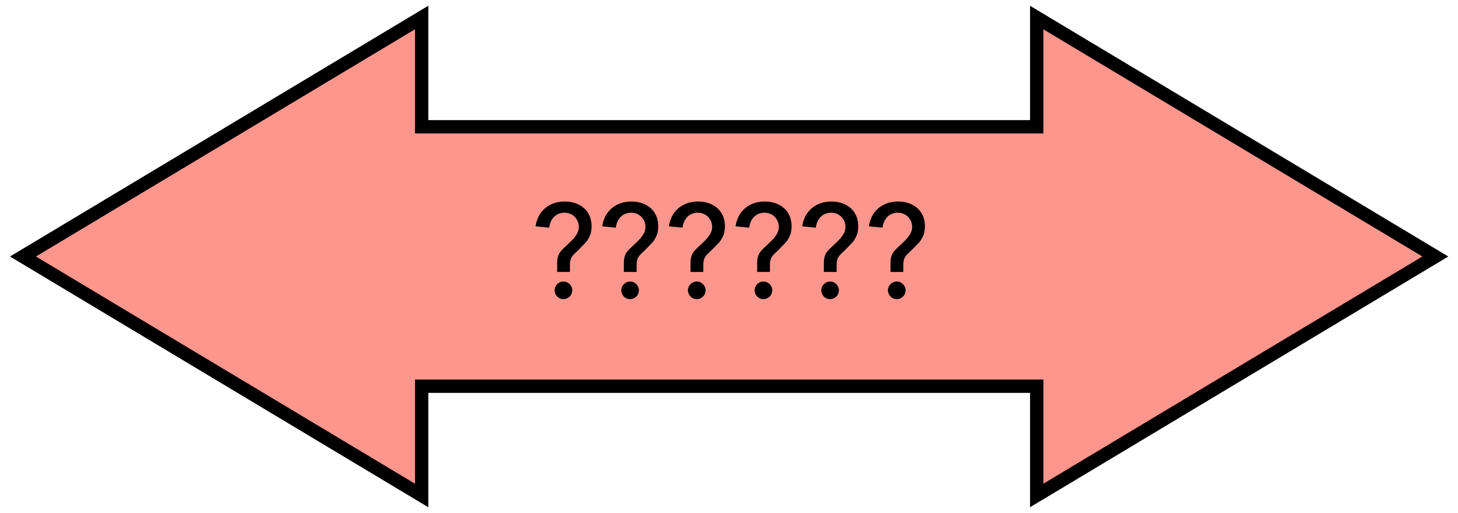
Aniruddha Kembhavi

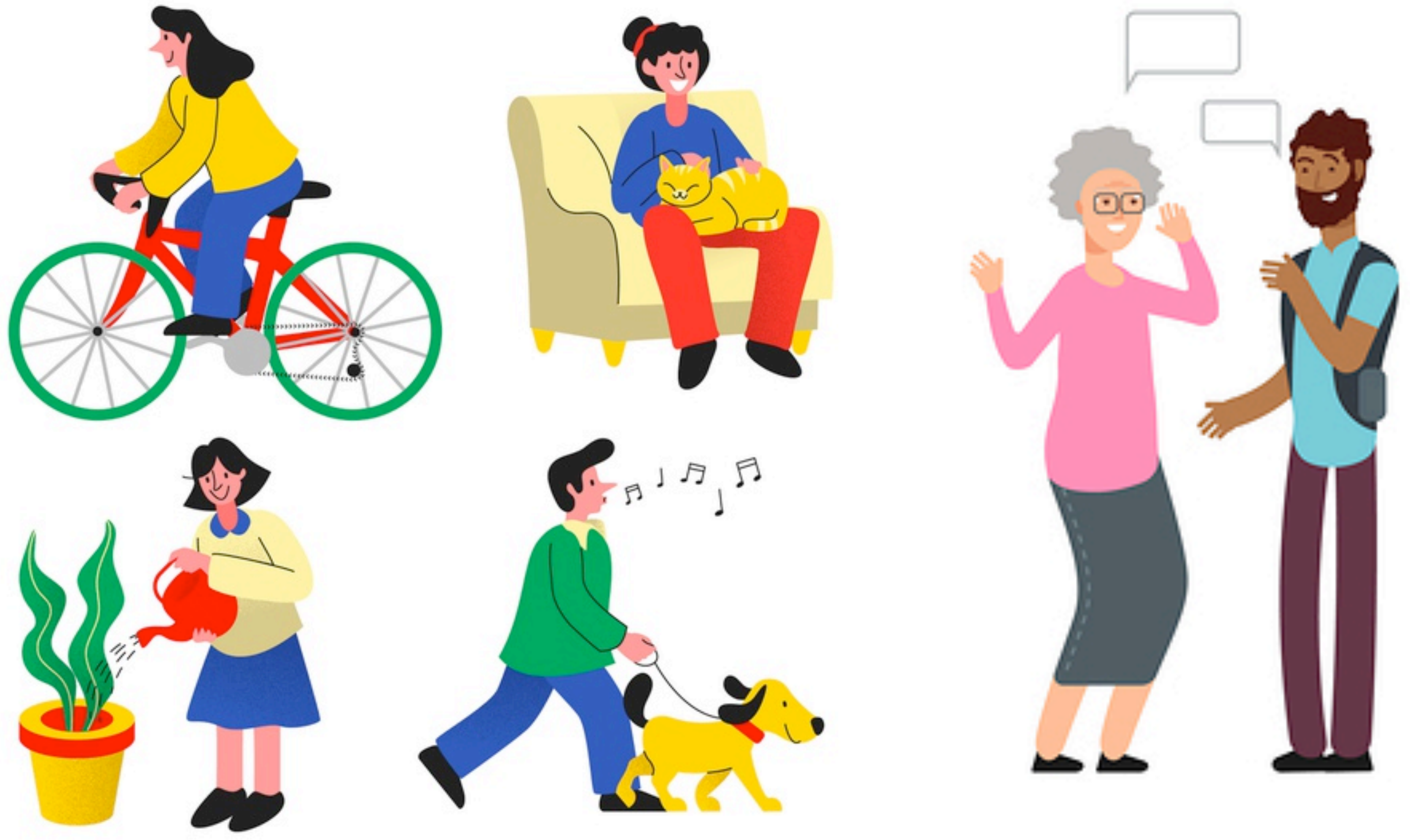
Ali Farhadi

Me

Problem: a gap between ***language form*** and ***commonsense grounded meaning***





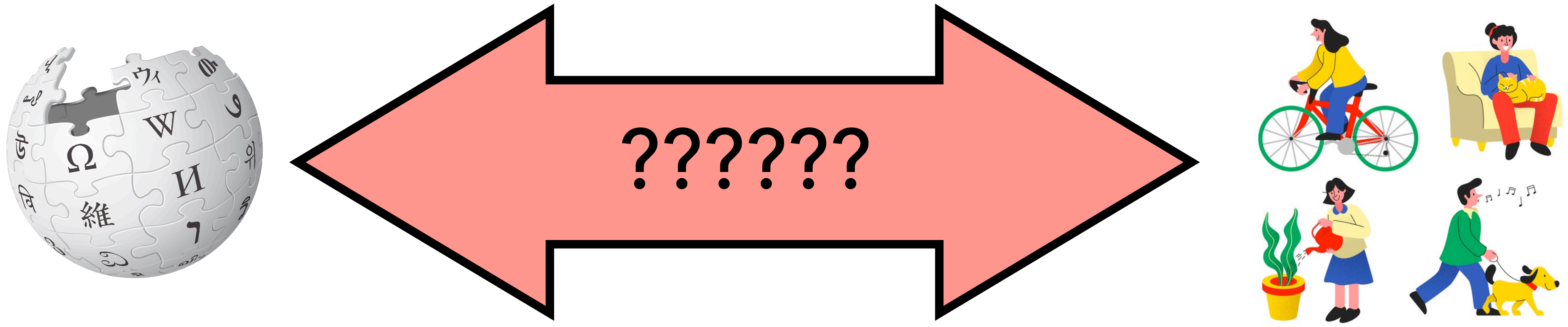


Written language
(*symbols*)

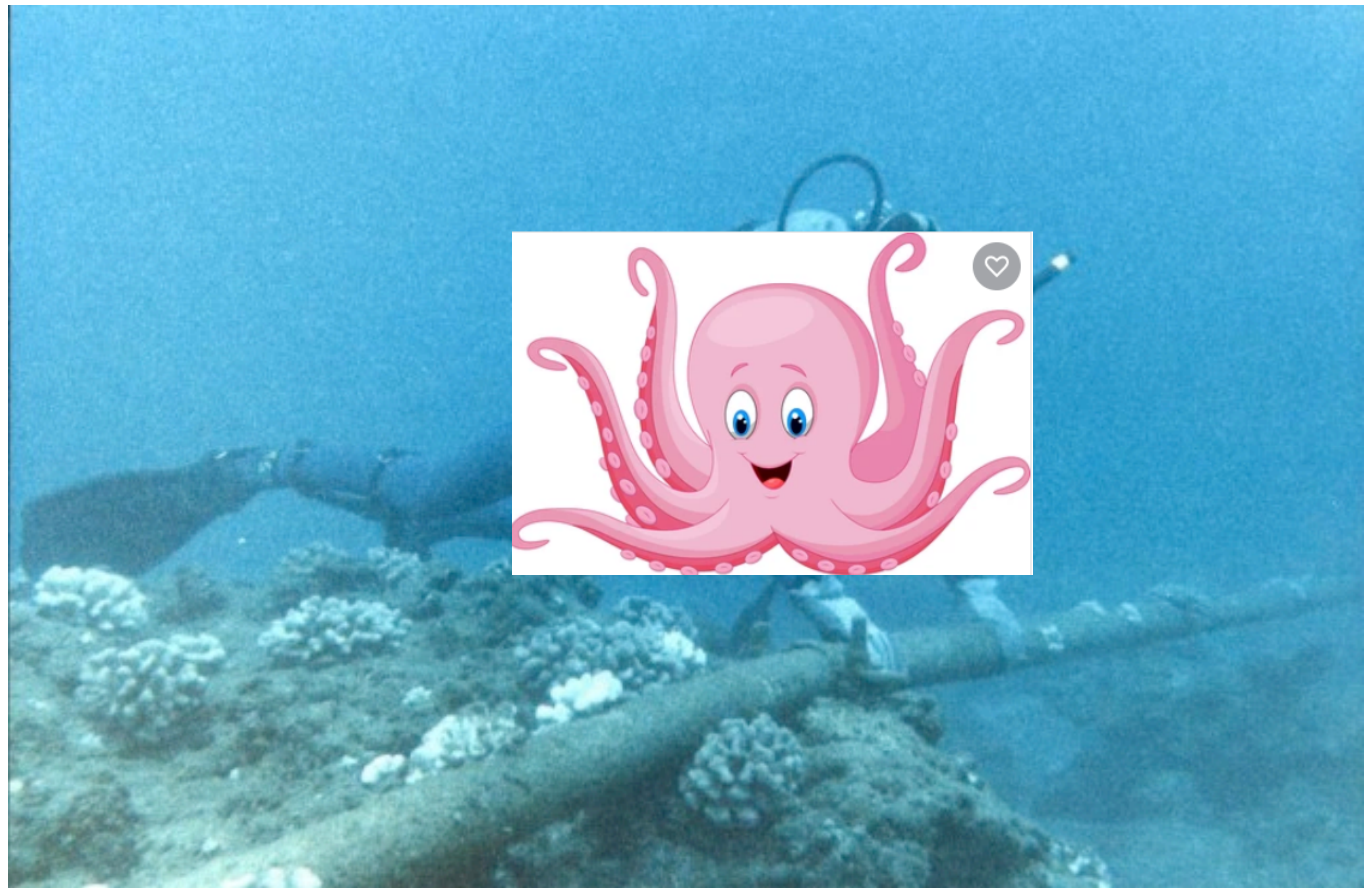
The world
(*continuous, subjective experience*)

Harnad 1992, inter alia

Problem: a gap between ***language form*** and ***commonsense grounded meaning***



Harnad 1992, inter alia

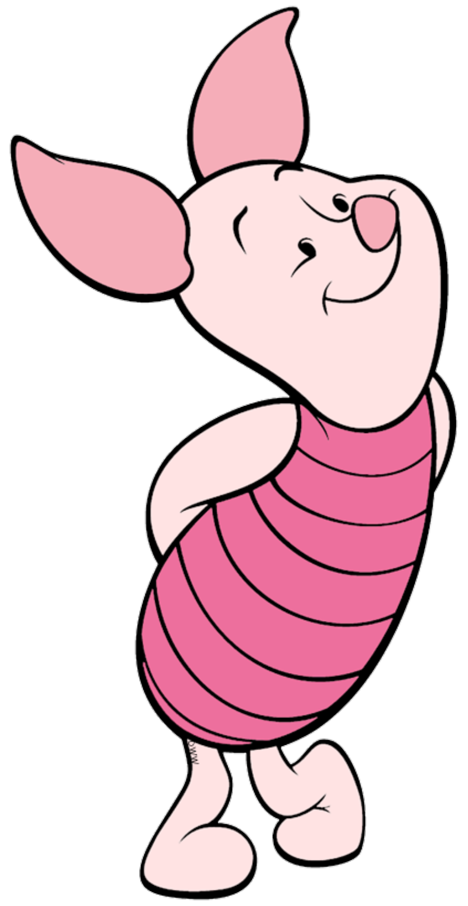


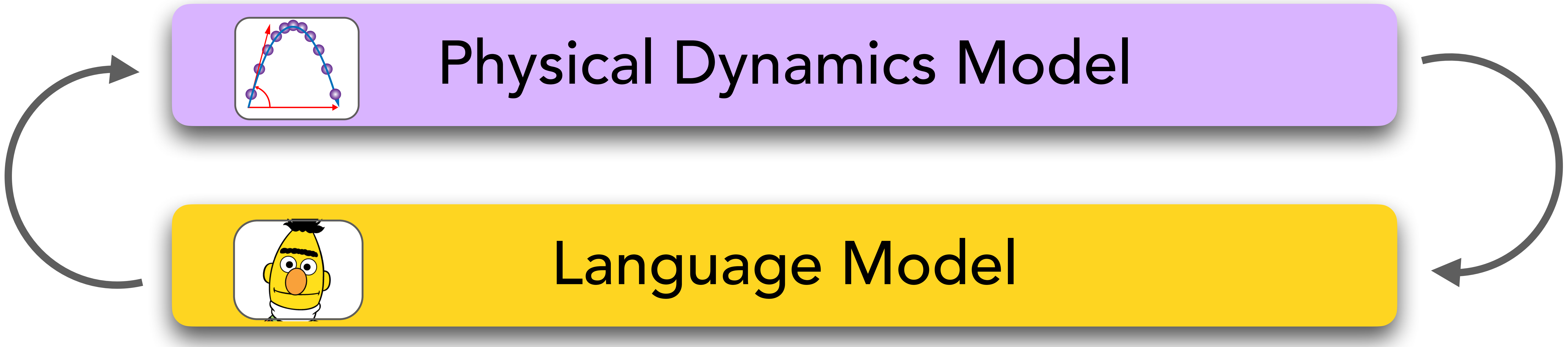
Bender and Koller 2020, inter alia

Proposal: ground language via a functional world representation, learned in simulation

PIGLeT: **P**hysical **I**nteractions as **G**rounding for **L**anguage **T**ransformers

Key idea: learn ***TWO*** model components for "how the world works" and "how to communicate it"



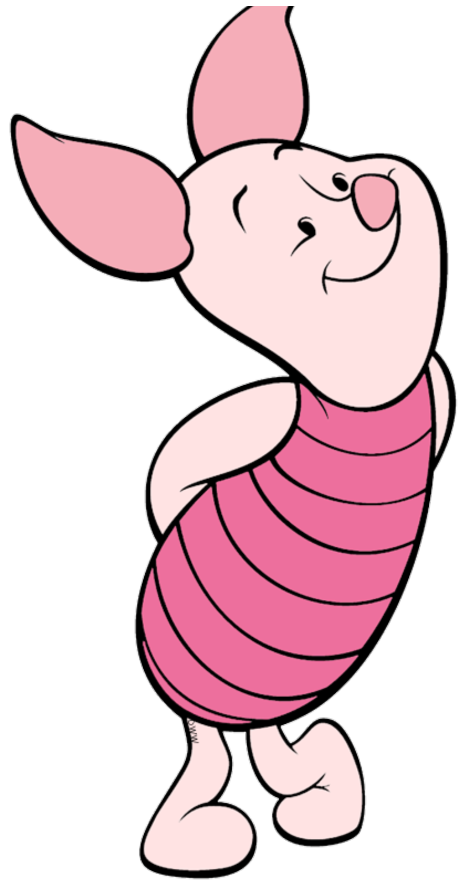


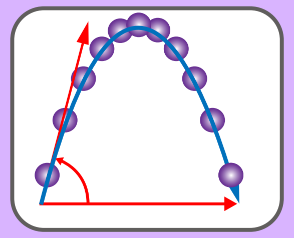
Learning “How the World Works”

Name:	Egg
Temperature:	RoomTemp
isCooked:	False
isBroken:	True

...

<heatUp, Pan>





Physical Dynamics Model

Language Model

Learning "How the World Works"

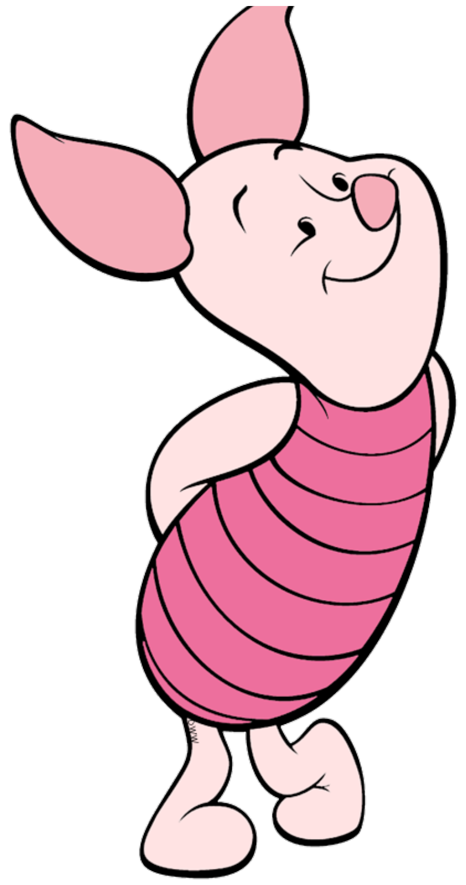
Name:	Egg
Temperature:	RoomTemp
isCooked:	False
isBroken:	True

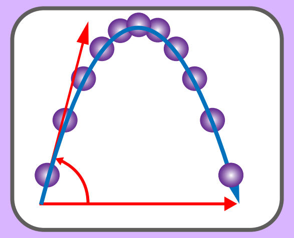
…

<heatUp, Pan>

Name:	Egg
Temperature:	***Hot***
isCooked:	***True***
isBroken:	True

…





Physical Dynamics Model

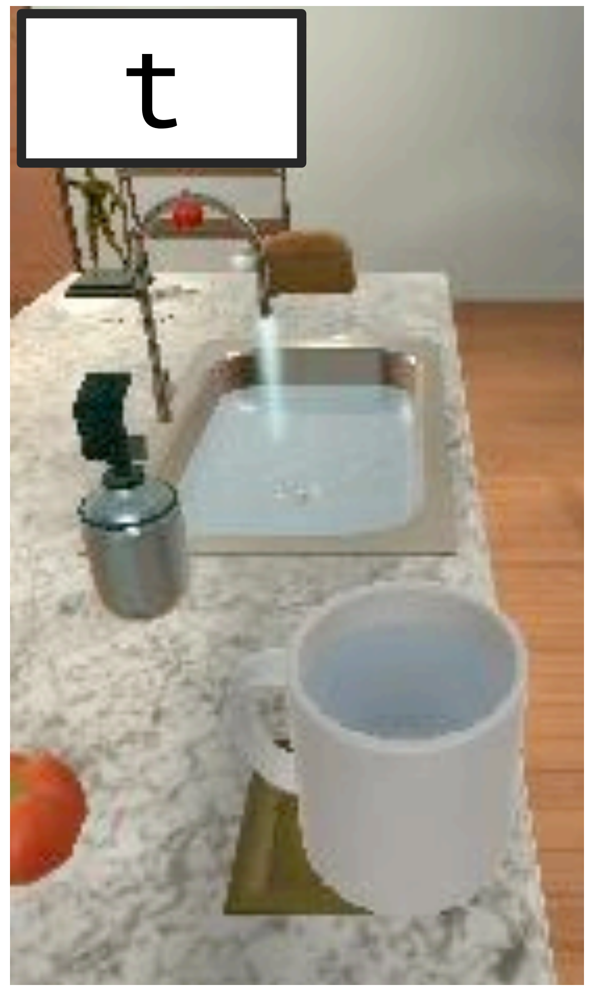
Language Model


Physical Dynamics Model
Name: Egg
Temperature: RoomTem
isCooked: False
isBroken: True
...
Object Encoder
<heatUp, Pan>
Action Encoder
Action Apply
Object Decoder
Name: Egg
Temperature: Hot
isCooked: True
isBroken: True
...


Physical Dynamics Model
Name: Egg
Temperature: RoomTem
isCooked: False
isBroken: True
...
Object Encoder
Action Apply
Object Decoder
Name: Egg
Temperature: Hot
isCooked: True
isBroken: True
...
<heatUp, Pan>
Action Encoder
The robot heats up the pan.
Language Model
Language Model
The pan becomes hot, and the egg gets cooked.

Results (accuracy of getting all attributes right)

Qualitative Example



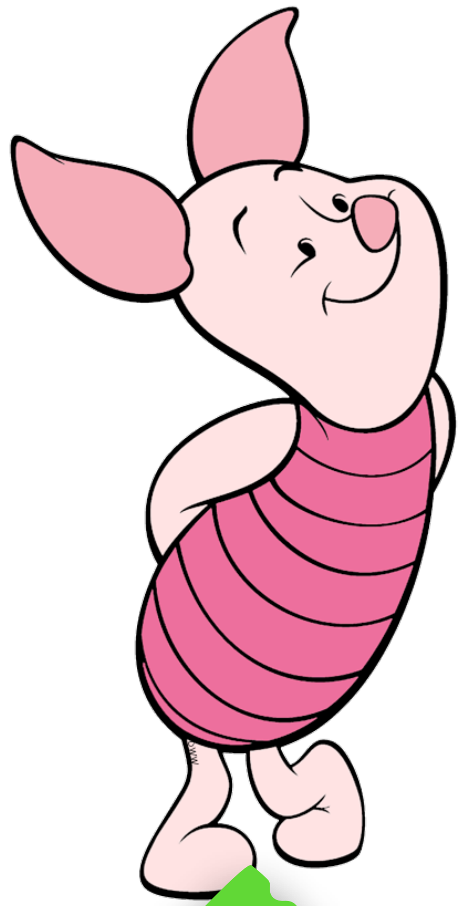
Name:	Sink
filledWithLiquid	True

Name:	Mug
filledWithLiquid	True
isPickedUp	True

The robot empties the mug.

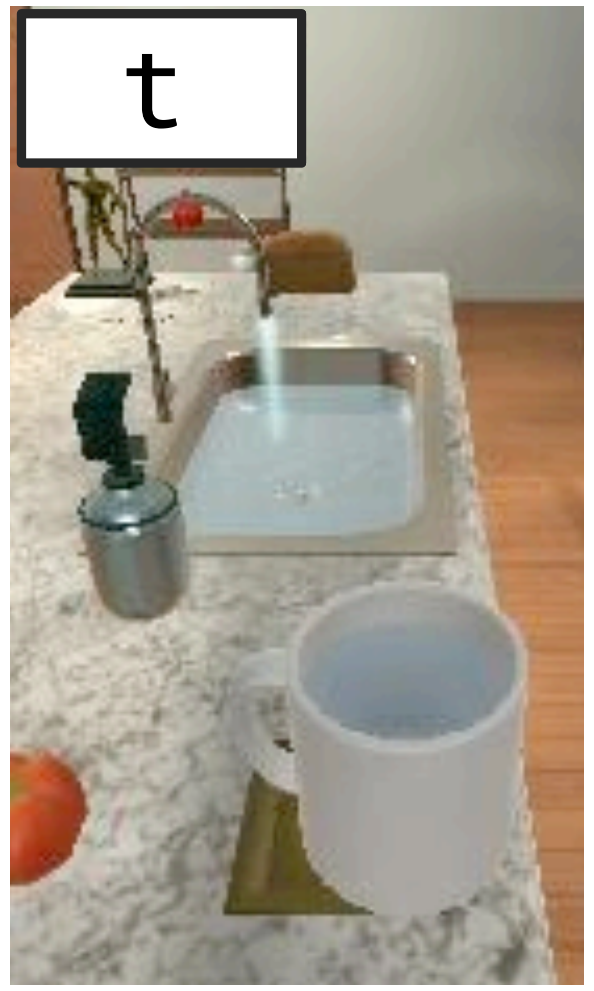
Name:	Sink
filledWithLiquid	True

Name:	Mug
filledWithLiquid	***False***
isPickedUp	True



Qualitative Example

T5



Name:	Sink
filledWithLiquid	True

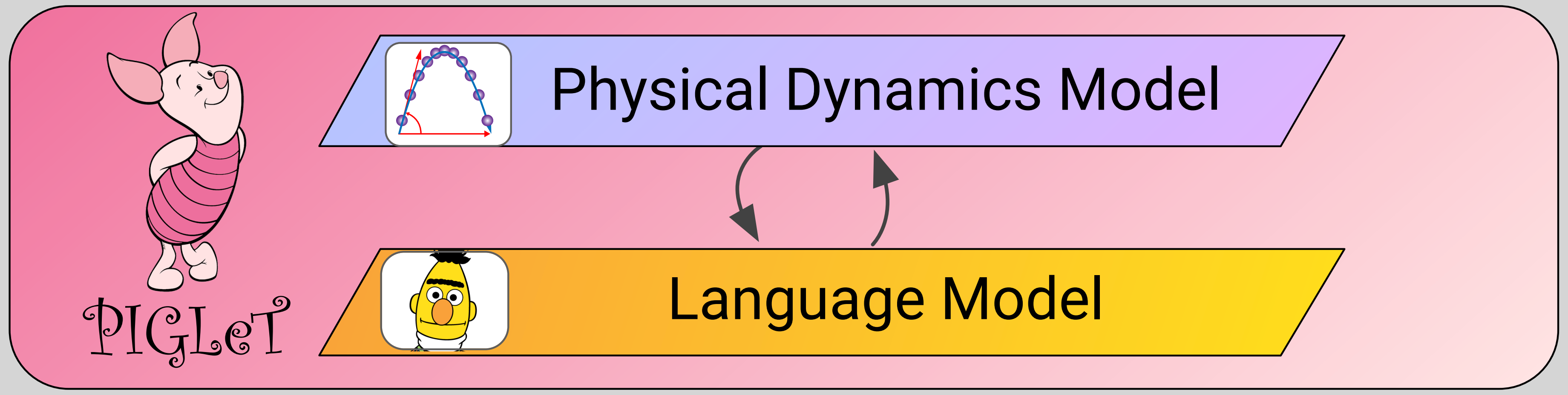
Name:	Mug
filledWithLiquid	True
isPickedUp	True

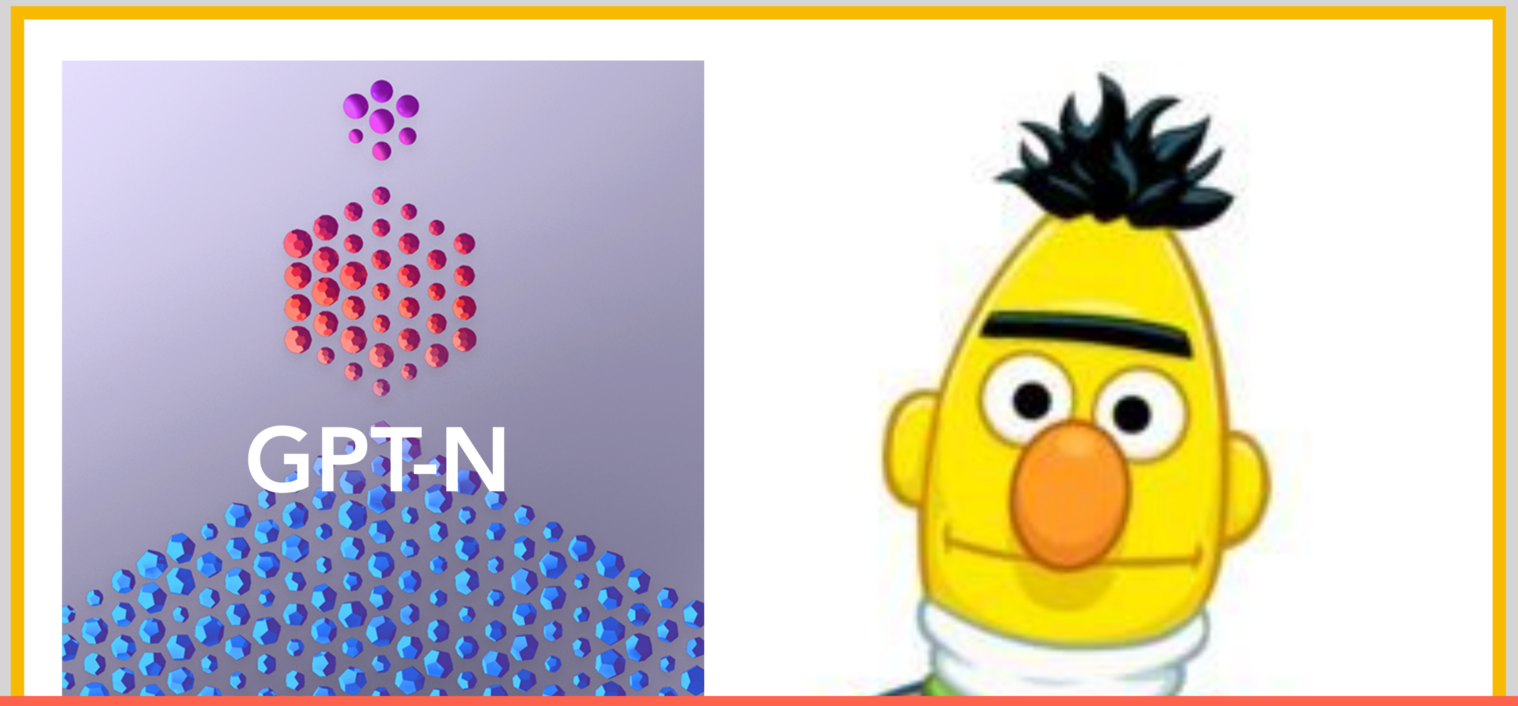
The robot empties the mug.

Name:	Sink
filledWithLiquid	False

T5, through text, learns "emptying liquid from an object" makes all objects in the room empty

PIGLeT: **P**hysical **I**nteractions as **G**rounding for **L**anguage **T**ransformers





Learning physical commonsense through interactions
=> higher performance with 100x smaller models

Learn a lightweight factorized world model for predicting *what might happen next*	**A single, heavyweight, entangled model**
Can generalize to new concepts without words	**Limited generalization to new concepts**

Harnad's Symbol Grounding Problem

- Grounding with 3D

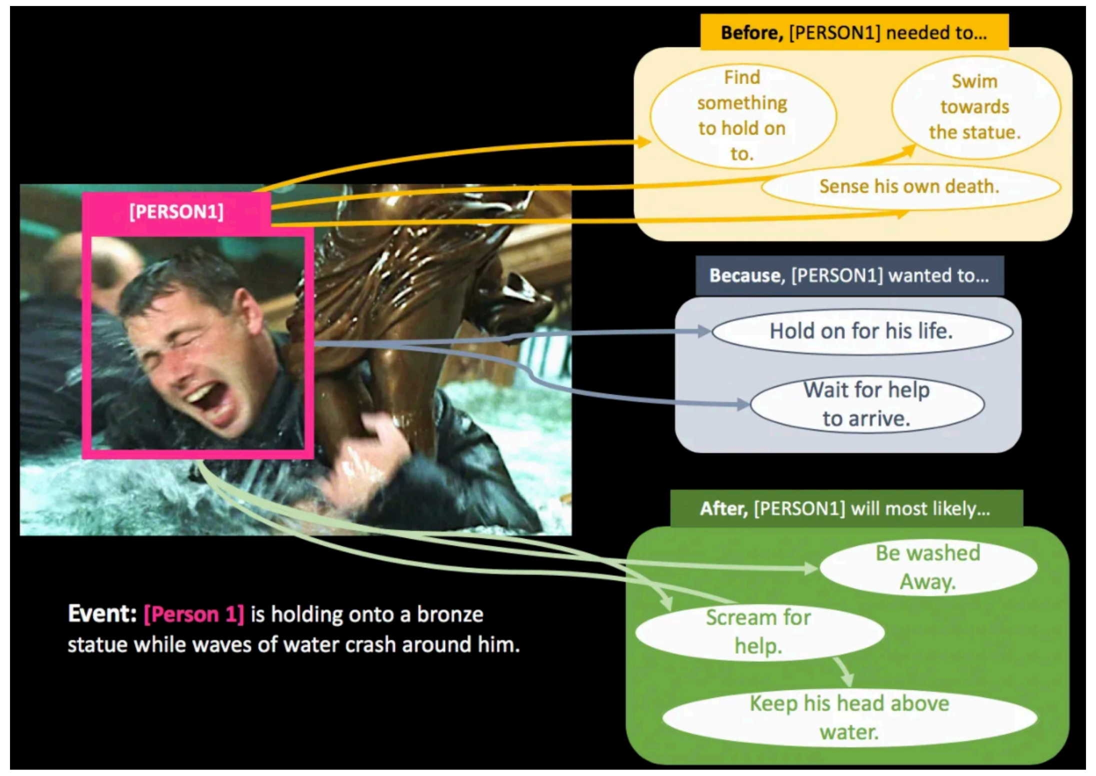
PIGLeT

- Grounding with 2D + Time

MERLOT

- Grounding with 2D + KG

Visual Comet



MERLOT: Multimodal Neural Script Knowledge Models

In Preparation

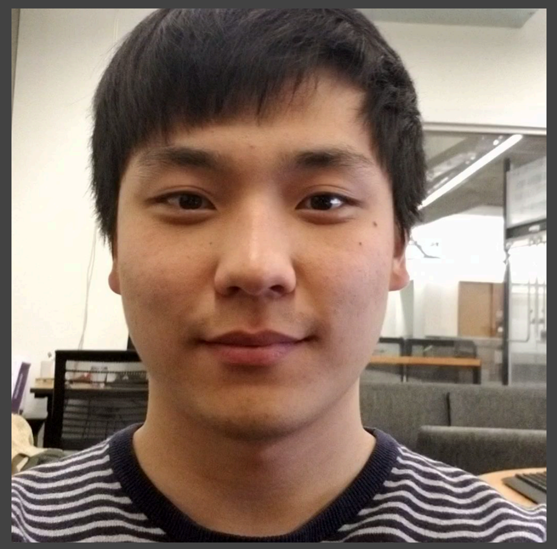
Rowan Zellers

Ximing Lu

Jack Hessel

Youngjae Yu

Jae Sung (James) Park



Jize Cao

Ali Farhadi

Me

VCR

Visual Commonsense Reasoning

Previously on VCR (cvpr 2019)

VCR Visual Commonsense Reasoning

Why is he pointing?


VCR
Visual Commonsense Reasoning
Why is he pointing?
<object: syrup bottle>
scene: a diner
<someone holding food>

Multimodal Script Knowledge

- Commonsense knowledge about events, including…
- What do people do at restaurants, and why?
- What might happen next in this event?

Script Knowledge

- (vanilla) script knowledge theory dates back to the early days of AI

SCRIPTS, PLANS, AND KNOWLEDGE

Roger C. Schank and Robert P. Abelson†

Yale University
New Haven, Connecticut USA

(1977)

"Of what a strange nature is knowledge! It clings to the mind, when it has once seized on it, like a lichen on the rock."

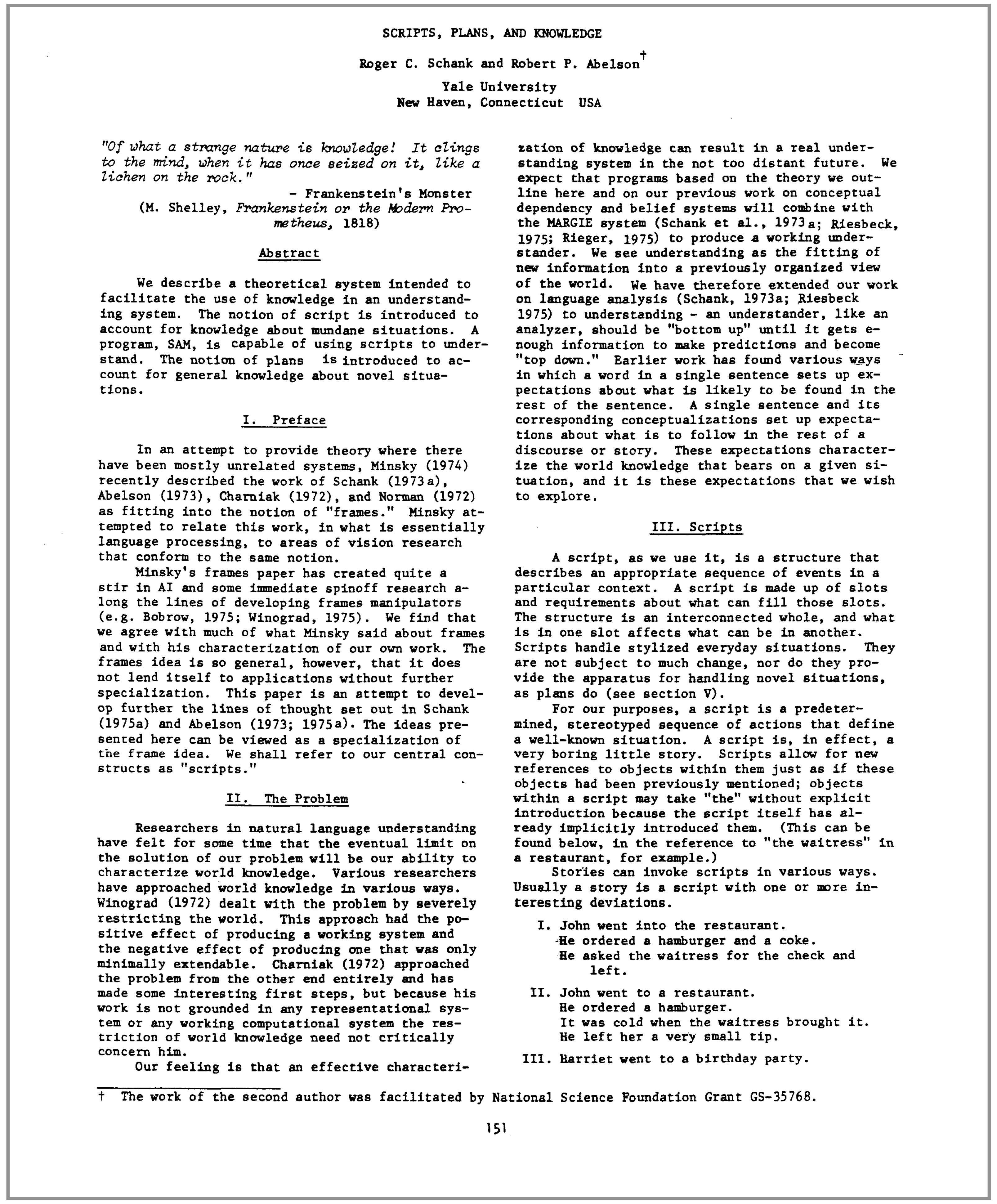
- Frankenstein's Monster
(M. Shelley, *Frankenstein or the Modern Prometheus*, 1818)

Abstract

We describe a theoretical system intended to facilitate the use of knowledge in an understanding system. The notion of script is introduced to

zation of knowledge can result in a real understanding system in the not too distant future. We expect that programs based on the theory we outline here and on our previous work on conceptual dependency and belief systems will combine with the MARGIE system (Schank et al., 1973a; Riesbeck, 1975; Rieger, 1975) to produce a working understander. We see understanding as the fitting of new information into a previously organized view of the world. We have therefore extended our work on language analysis (Schank, 1973a; Riesbeck 1975) to understanding - an understander, like an

Script Knowledge



SCRIPTS, PLANS, AND KNOWLEDGE

Roger C. Schank and Robert P. Abelson†

Yale University
New Haven, Connecticut USA

"Of what a strange nature is knowledge! It clings to the mind, when it has once seized on it, like a lichen on the rock."

- Frankenstein's Monster
(M. Shelley, *Frankenstein or the Modern Prometheus*, 1818)

Abstract

We describe a theoretical system intended to facilitate the use of knowledge in an understanding system. The notion of script is introduced to account for knowledge about mundane situations. A program, SAM, is capable of using scripts to understand. The notion of plans is introduced to account for general knowledge about novel situations.

I. Preface

In an attempt to provide theory where there have been mostly unrelated systems, Minsky (1974) recently described the work of Schank (1973a), Abelson (1973), Charniak (1972), and Norman (1972) as fitting into the notion of "frames." Minsky attempted to relate this work, in what is essentially language processing, to areas of vision research that conform to the same notion.

Minsky's frames paper has created quite a stir in AI and some immediate spinoff research along the lines of developing frames manipulators (e.g. Bobrow, 1975; Winograd, 1975). We find that we agree with much of what Minsky said about frames and with his characterization of our own work. The frames idea is so general, however, that it does not lend itself to applications without further specialization. This paper is an attempt to develop further the lines of thought set out in Schank (1975a) and Abelson (1973; 1975a). The ideas presented here can be viewed as a specialization of the frame idea. We shall refer to our central constructs as "scripts."

II. The Problem

Researchers in natural language understanding have felt for some time that the eventual limit on the solution of our problem will be our ability to characterize world knowledge. Various researchers have approached world knowledge in various ways. Winograd (1972) dealt with the problem by severely restricting the world. This approach had the positive effect of producing a working system and the negative effect of producing one that was only minimally extendable. Charniak (1972) approached the problem from the other end entirely and has made some interesting first steps, but because his work is not grounded in any representational system or any working computational system the restriction of world knowledge need not critically concern him.

Our feeling is that an effective characterization of knowledge can result in a real understanding system in the not too distant future. We expect that programs based on the theory we outline here and on our previous work on conceptual dependency and belief systems will combine with the MARGIE system (Schank et al., 1973a; Riesbeck, 1975; Rieger, 1975) to produce a working understander. We see understanding as the fitting of new information into a previously organized view of the world. We have therefore extended our work on language analysis (Schank, 1973a; Riesbeck 1975) to understanding - an understander, like an analyzer, should be "bottom up" until it gets enough information to make predictions and become "top down." Earlier work has found various ways in which a word in a single sentence sets up expectations about what is likely to be found in the rest of the sentence. A single sentence and its corresponding conceptualizations set up expectations about what is to follow in the rest of a discourse or story. These expectations characterize the world knowledge that bears on a given situation, and it is these expectations that we wish to explore.

III. Scripts

A script, as we use it, is a structure that describes an appropriate sequence of events in a particular context. A script is made up of slots and requirements about what can fill those slots. The structure is an interconnected whole, and what is in one slot affects what can be in another. Scripts handle stylized everyday situations. They are not subject to much change, nor do they provide the apparatus for handling novel situations, as plans do (see section V).

For our purposes, a script is a predetermined, stereotyped sequence of actions that define a well-known situation. A script is, in effect, a very boring little story. Scripts allow for new references to objects within them just as if these objects had been previously mentioned; objects within a script may take "the" without explicit introduction because the script itself has already implicitly introduced them. (This can be found below, in the reference to "the waitress" in a restaurant, for example.)

Stories can invoke scripts in various ways. Usually a story is a script with one or more interesting deviations.

I. John went into the restaurant.
He ordered a hamburger and a coke.
He asked the waitress for the check and left.

II. John went to a restaurant.
He ordered a hamburger.
It was cold when the waitress brought it.
He left her a very small tip.

III. Harriet went to a birthday party.

† The work of the second author was facilitated by National Science Foundation Grant GS-35768.

script: restaurant

roles: customer, waiter, chef, cashier

Scene 1: entering

PTRANS self into restaurant

ATTEND eyes to where empty tables are

MBUILD where to sit

PTRANS self to table

MOVE sit down

Scene 2: ordering

…

Multimodal Script Knowledge

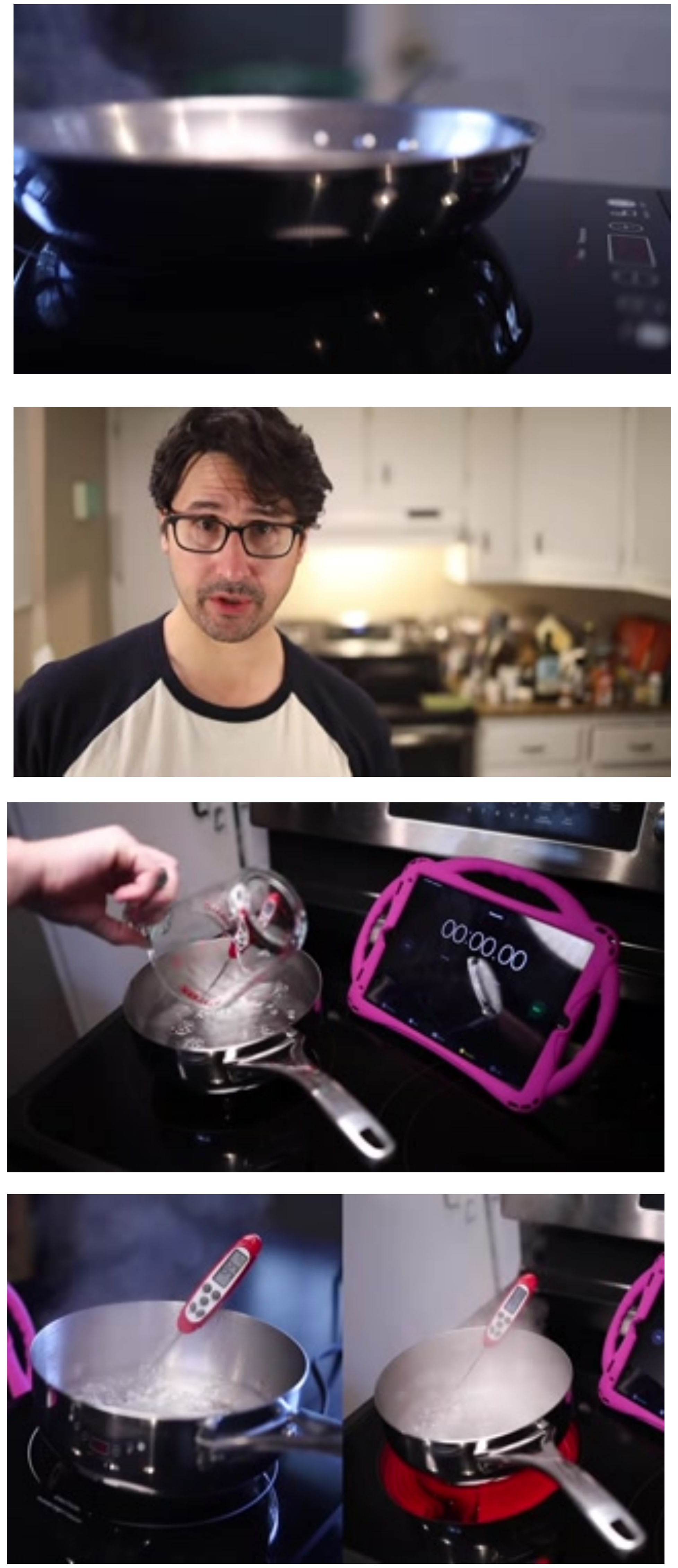
(Neural)

Multimodal Script Knowledge

(Neural)

From 6M youtube videos, we'll learn:

From 6M youtube videos, we’ll learn:



Recognition-level Knowledge

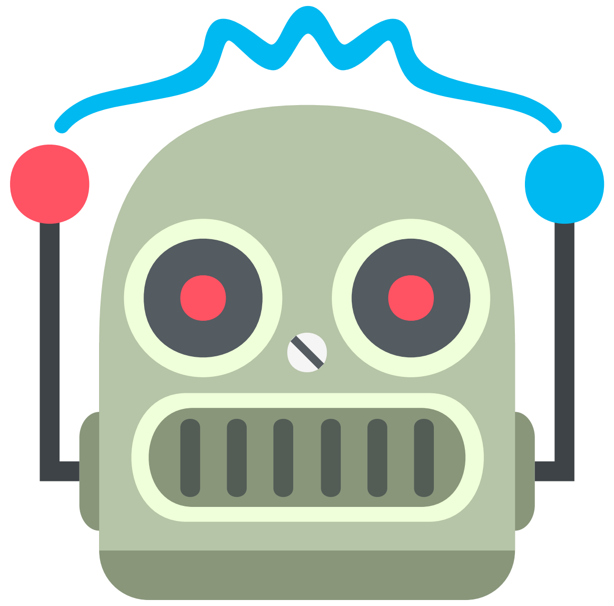
person

pan Burner

stopwatch
water pitcher

thermometer

Multimodal Script Knowledge

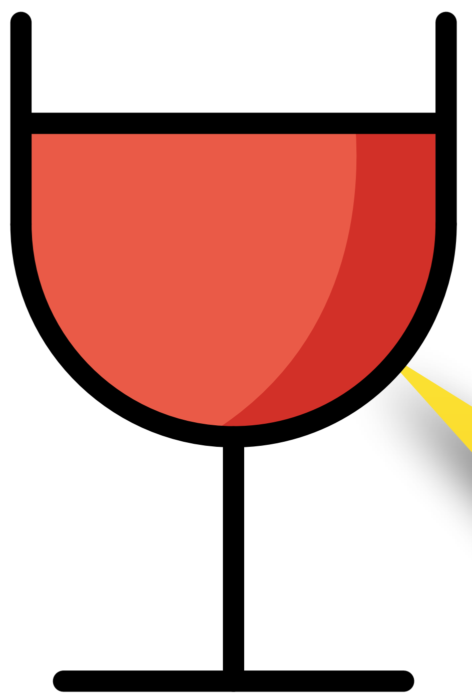


This person might be measuring how fast the water boils

From 6M youtube videos, we'll learn:

Recognition-level Knowledge

Multimodal Script Knowledge



Multimodal **E**vent **R**epresentation **L**earning **O**ver **T**ime

The result:

- Trained fully from scratch, we get...
- zero-shot temporal commonsense,
- Fine-tuned SOTA on 13 tasks

Setup: Videos and Transcripts

"I'm going to compare electric and induction stoves..."

"I'll use a stopwatch to time how fast my electric stove boils water...."

"In goes the cold water..."

"It took 4 and a half minutes to reach full boil..."

Time

Recognition-level learning

Image Encoder

OK, but..

Text Encoder

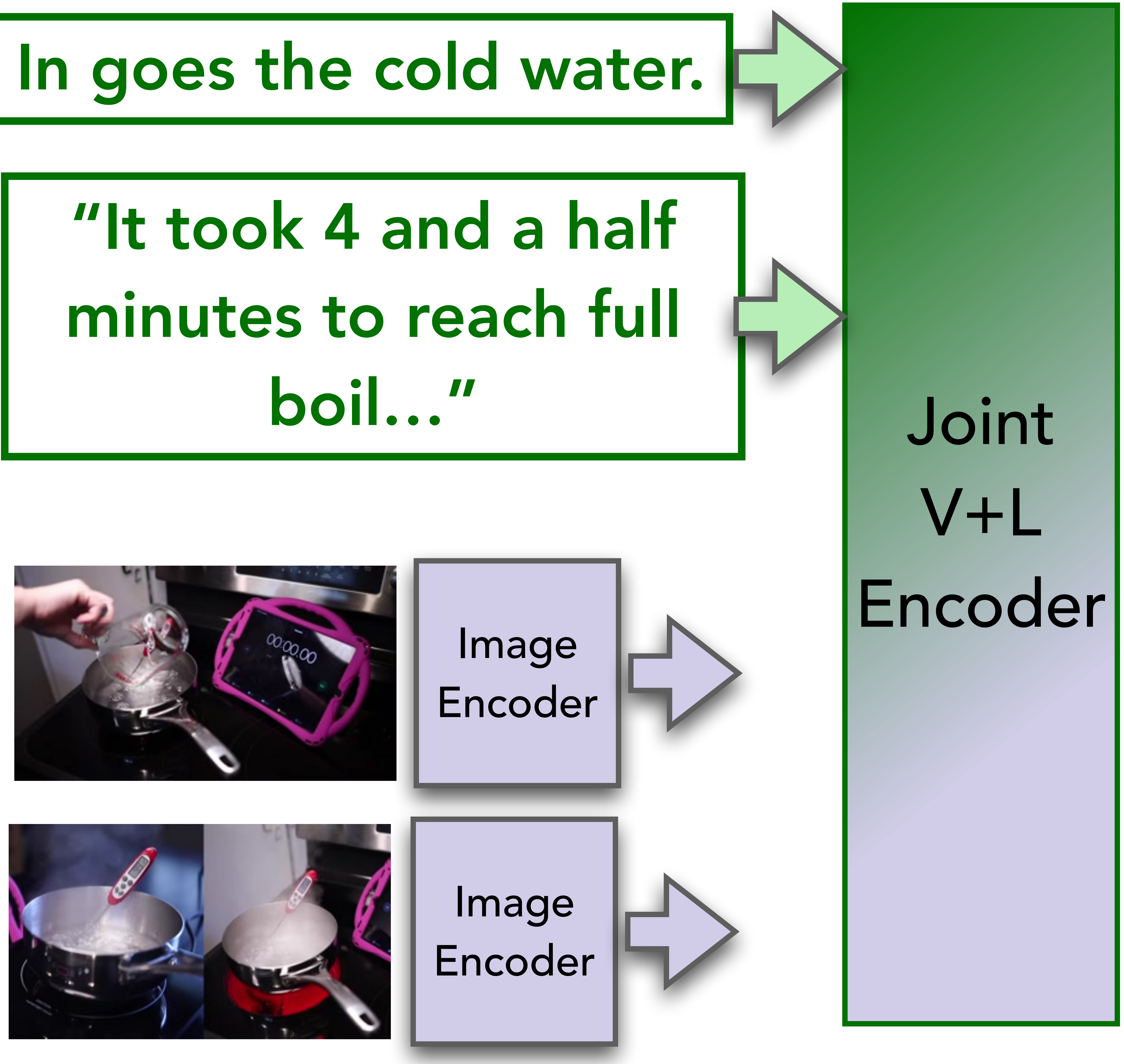
In goes the cold water.

(ConVIRT; Zhang et al 2020, CLIP; Radford et al 2021)

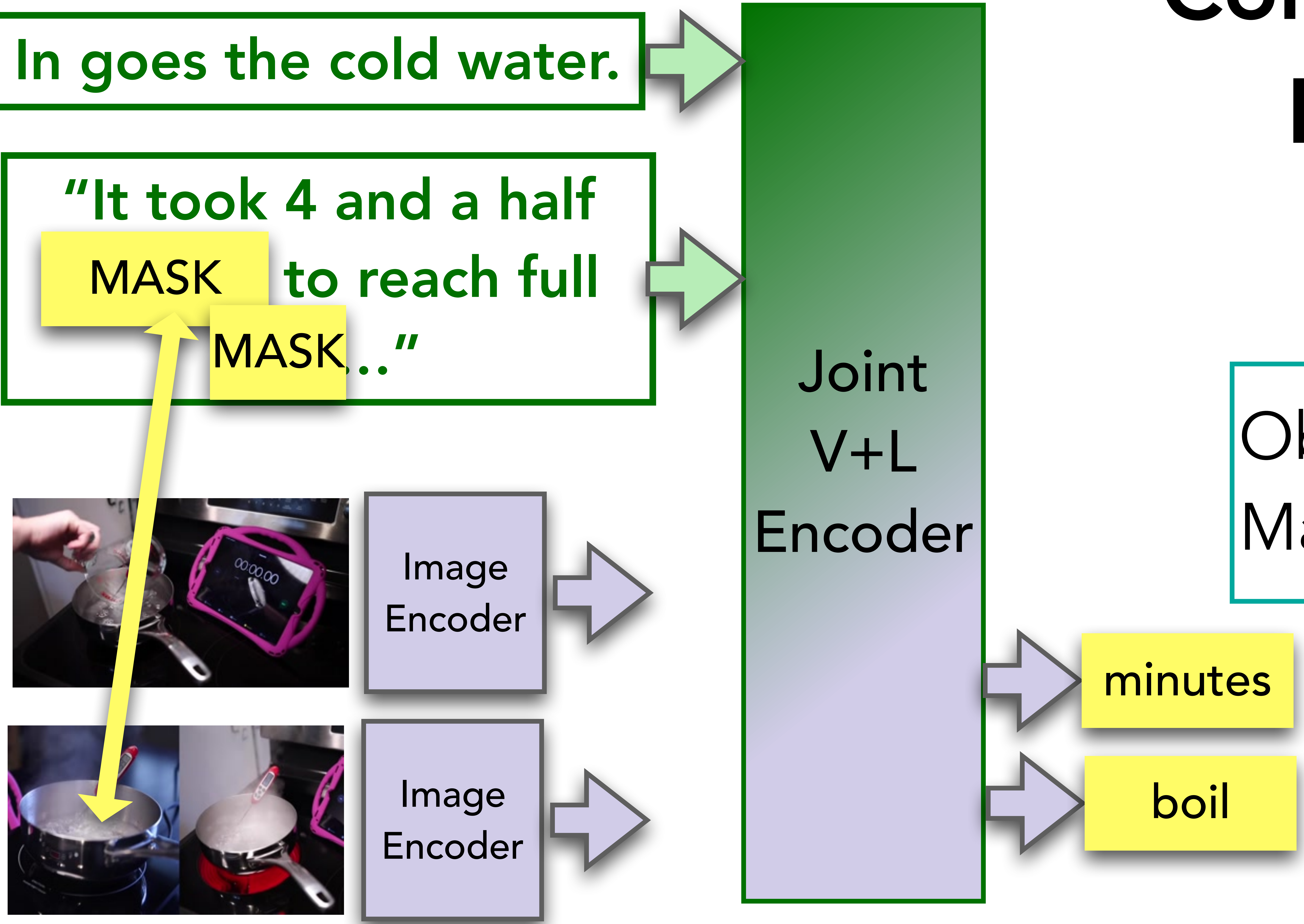
Recognition-level learning

Recognition-level learning

Commonsense Learning



Commonsense Learning



Objective 2:
Mask LM

Commonsense Learning

In goes the cold water.

"It took 4 and a half minutes to reach full boil..."

Image Encoder

Image Encoder

Joint V+L Encoder

Objective 3: Unshuffle frames

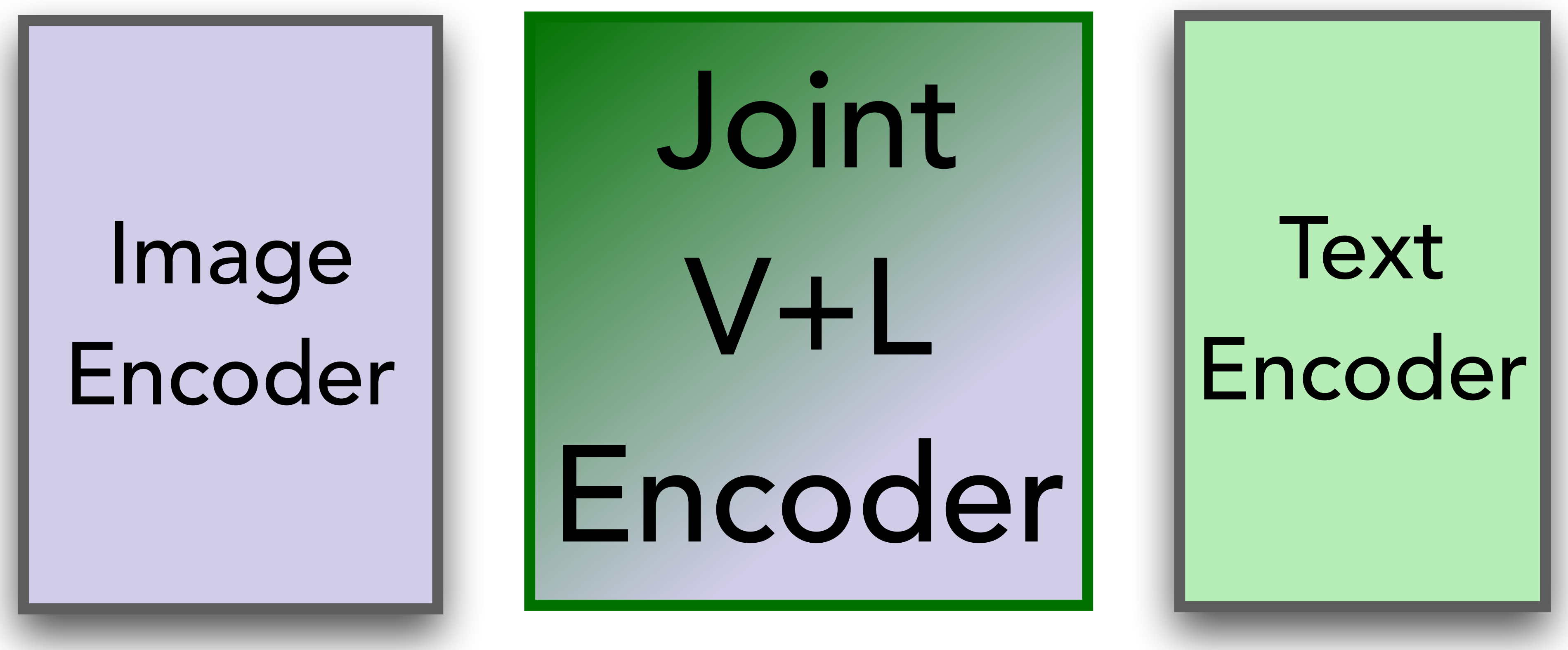
Frame 2 comes first

Objective1: Contextual Frame-Text Matching

Objective 2: Mask LM

Objective 3: Unshuffle frames

Using a 12-layer 'base' Transformer, train end-to-end on 6M videos



Evaluation

Evaluation 1: Zero-Shot Unscrambling Visual Stories

Task: Given the text of a visual story, match images to text to tell a narrative

(SIND; Huang et al 2016, Agrawal et al 2016)

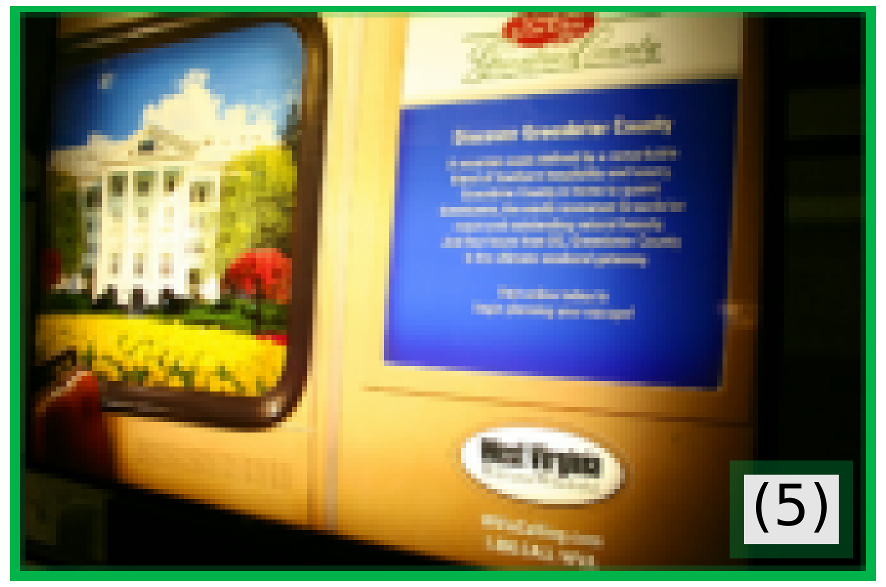
The old man was riding the escalator. → He was almost to the top. → His kids were already at the top. → At the top was a train station. → They then got on the train.

Task: Given the text of a visual story, match images to text to tell a narrative

The old man was riding the escalator. → He was almost to the top. → His kids were already at the top. → At the top was a train station. → They then got on the train.

Task: Given the text of a visual story, match images to text to tell a narrative

The old man was riding the escalator. → He was almost to the top. → His kids were already at the top. → At the top was a train station. → They then got on the train.



Our model gets this right ***without finetuning,***
using the unscrambling objective

Task: Given the text of a visual story, match images to text to tell a narrative

The old man was riding the escalator.

He was almost to the top.

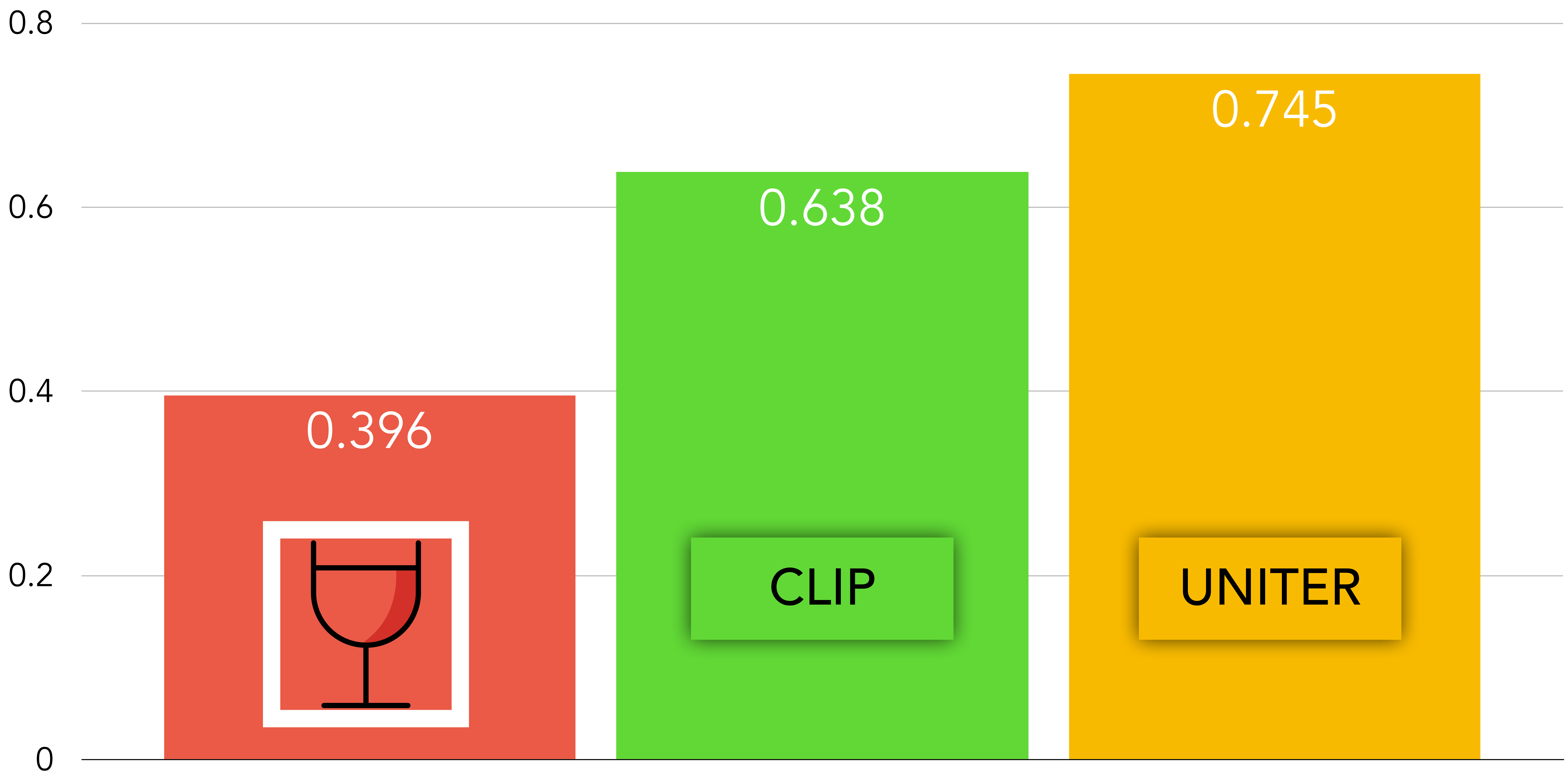
His kids were already at the top.

At the top was a train station.

They then got on the train.


The old man was riding the escalator.
He was almost to the top.
His kids were already at the top.
At the top was a train station.
They then got on the train.
(1)
(2)
(3)
(4)
(5)
CLIP
(Radford et al 2021)
(1)
(3)
(2)
(4)
(5)

Distance away from sorted order (lower is better, 5.0 is max)



(Chen et al 2019)

Even when our model is "wrong" it's kinda cool

I went to the fair with my kids last weekend. → There were a lot of people there. → They also had a barn. → We got to see a lot of animals. → We can't wait to go back later.

Even when our model is “wrong” it’s kinda cool

Evaluation 2: Fine-tuned Video QA

Evaluation 3: Visual Commonsense Reasoning (Q->AR)

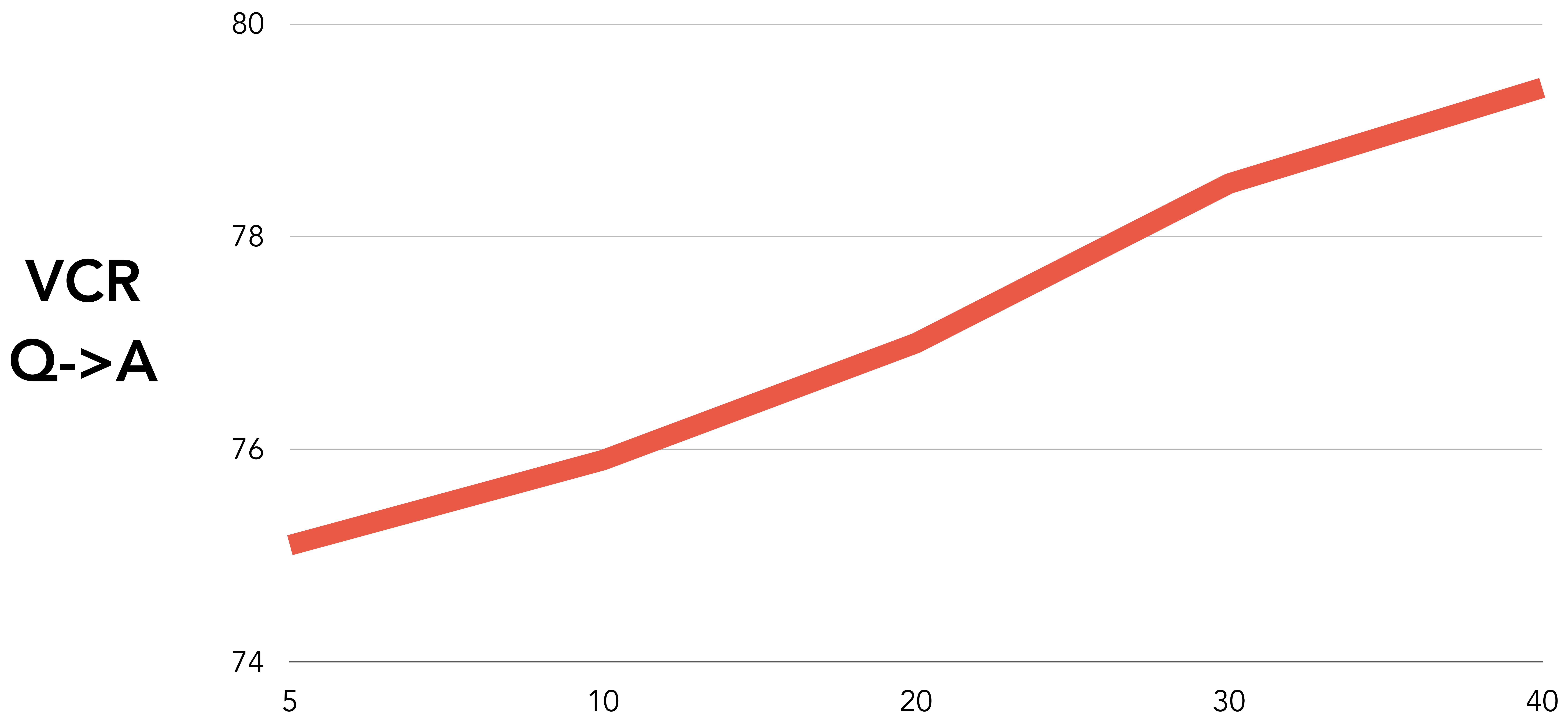
Despite no supervised object detector, and never seeing still images before

Analysis (on TVQA+)

Analysis (on TVQA+)

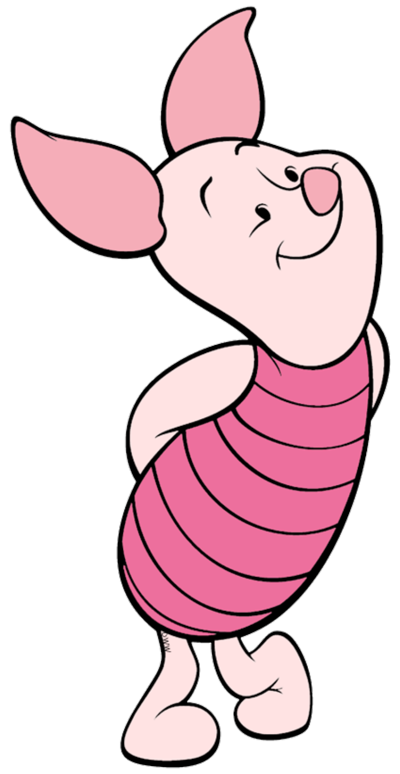
Analysis (on TVQA+)

Performance increases with # epochs



Harnad's Symbol Grounding Problem

- Grounding with 3D



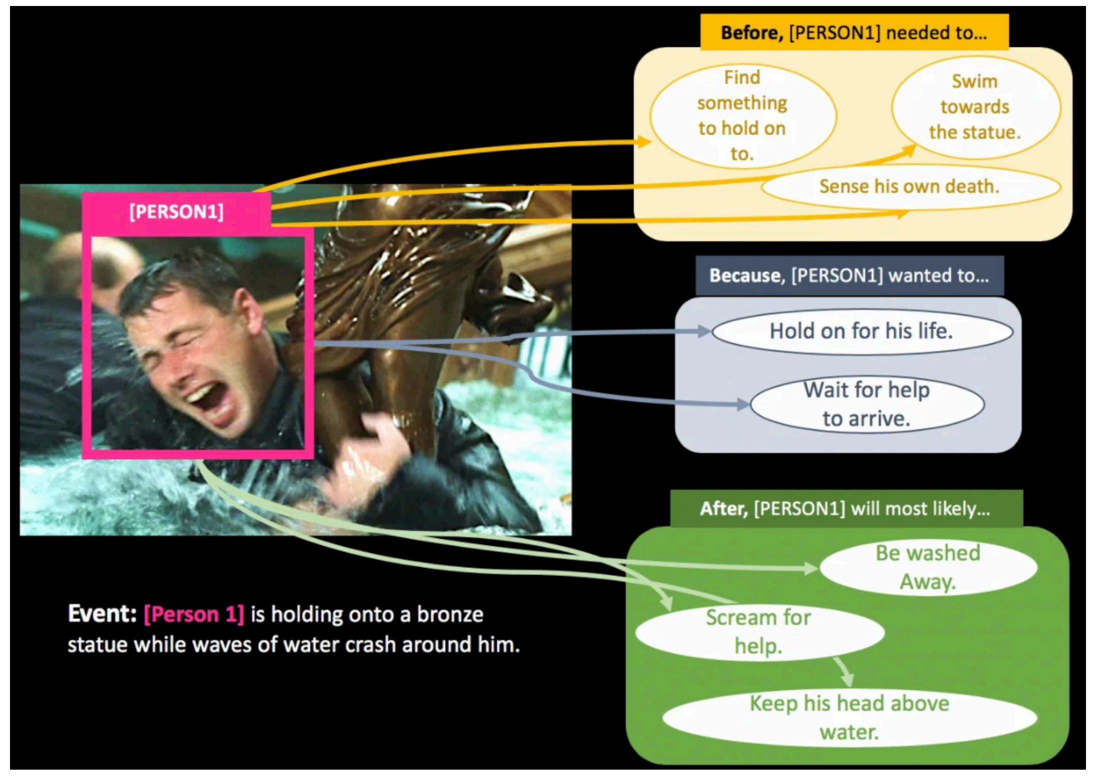
PIGLeT

- Grounding with 2D + Time

MERLOT

- Grounding with 2D + KG

Visual Comet

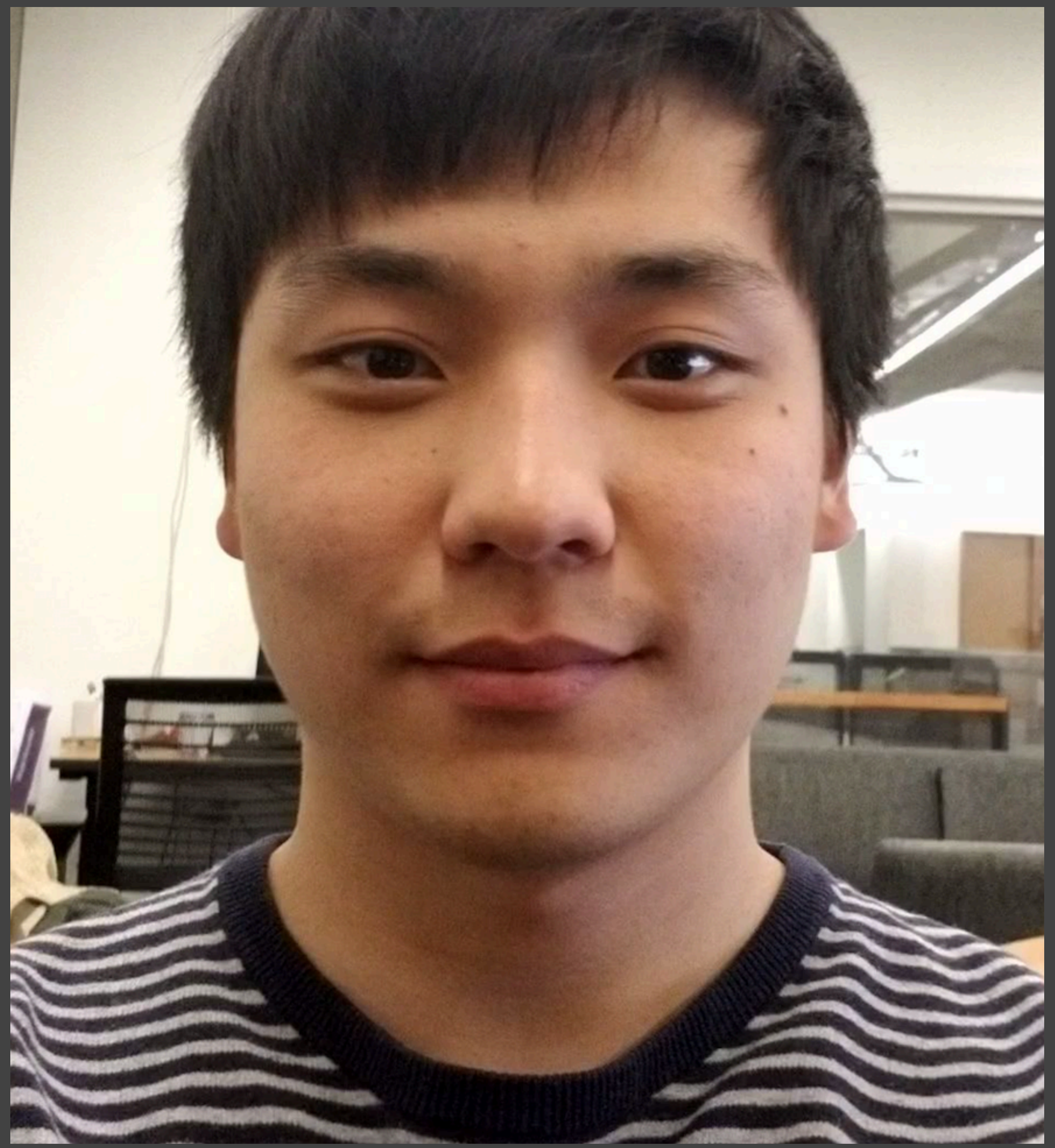


Visual COMET:

Reasoning about the *Dynamic* Context of a *Still* Image

ECCV 2020

Jae Sung (James) Park



Chandra Bhagavatula

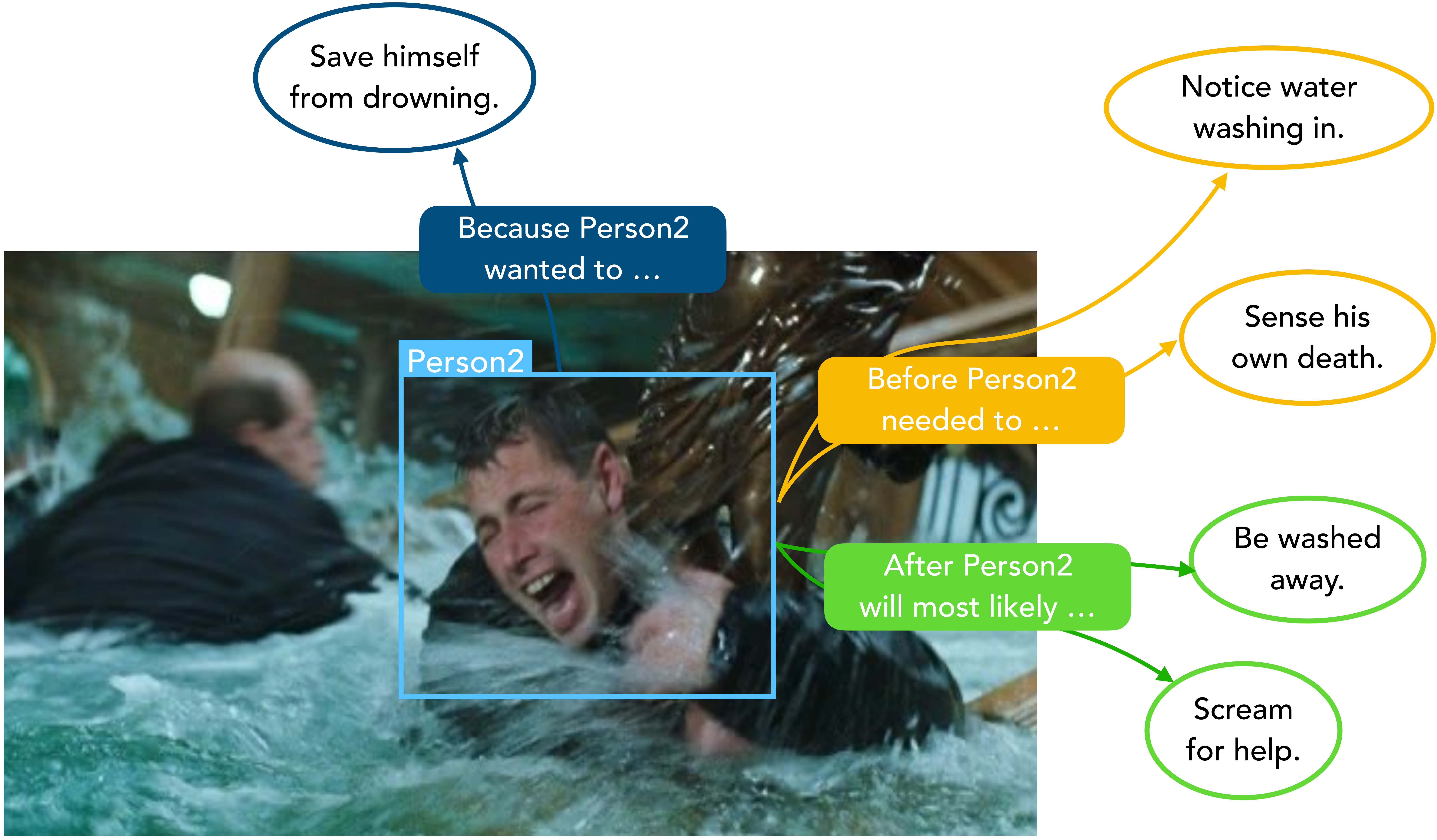
Roozbeh Mottaghi

Ali Farhadi

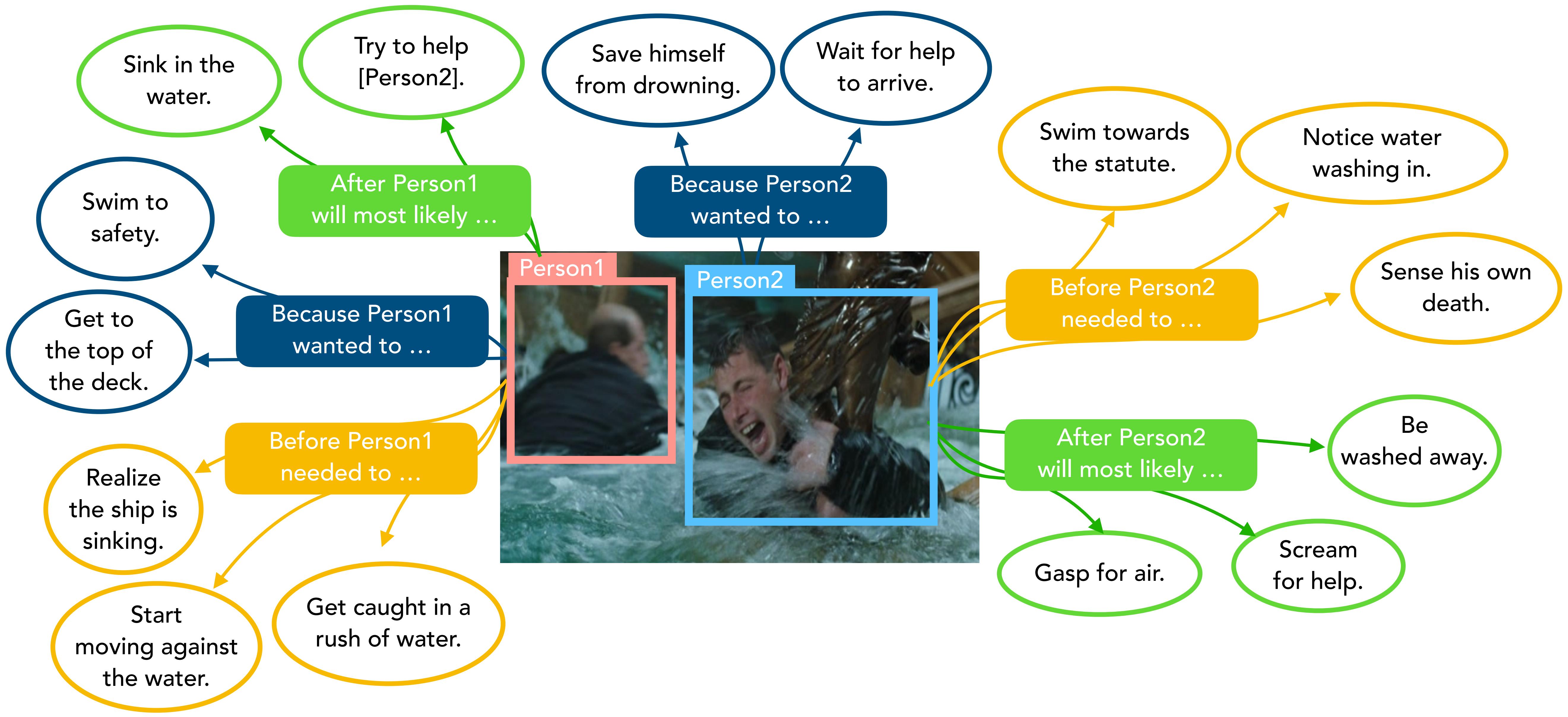
Yejin Choi

Visual Commonsense Graphs:

Reasoning about the *Dynamic* Context of a *Still* Image



Visual Commonsense Graphs:
Reasoning about the *Dynamic* Context of a *Still* Image



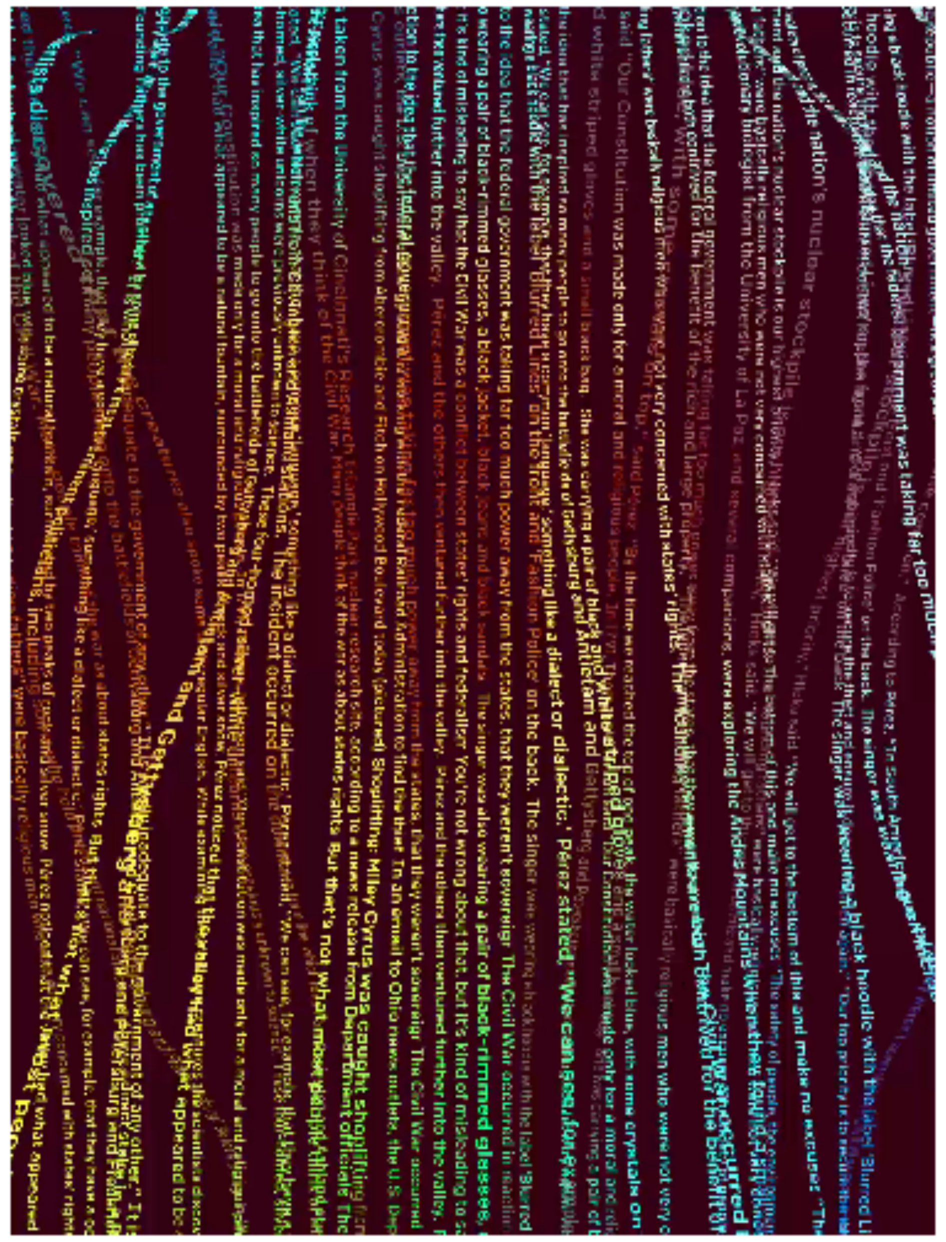
Sink in the water.
Try to help [Person2].
Save himself from drowning.
Wait for help to arrive.
Swim towards the statute.
Notice water washing in.
After Person1 will most likely …
Because Person2 wanted to …
Swim to safety.
Sense his own death.
Before Person2 needed to …
Person1
Person2
Get to the top of the deck.
Because Person1 wanted to …
Before Person1 needed to …
After Person2 will most likely …
Be washed away.
Realize the ship is sinking.
Gasp for air.
Scream for help.
Start moving against the water.
Get caught in a rush of water.

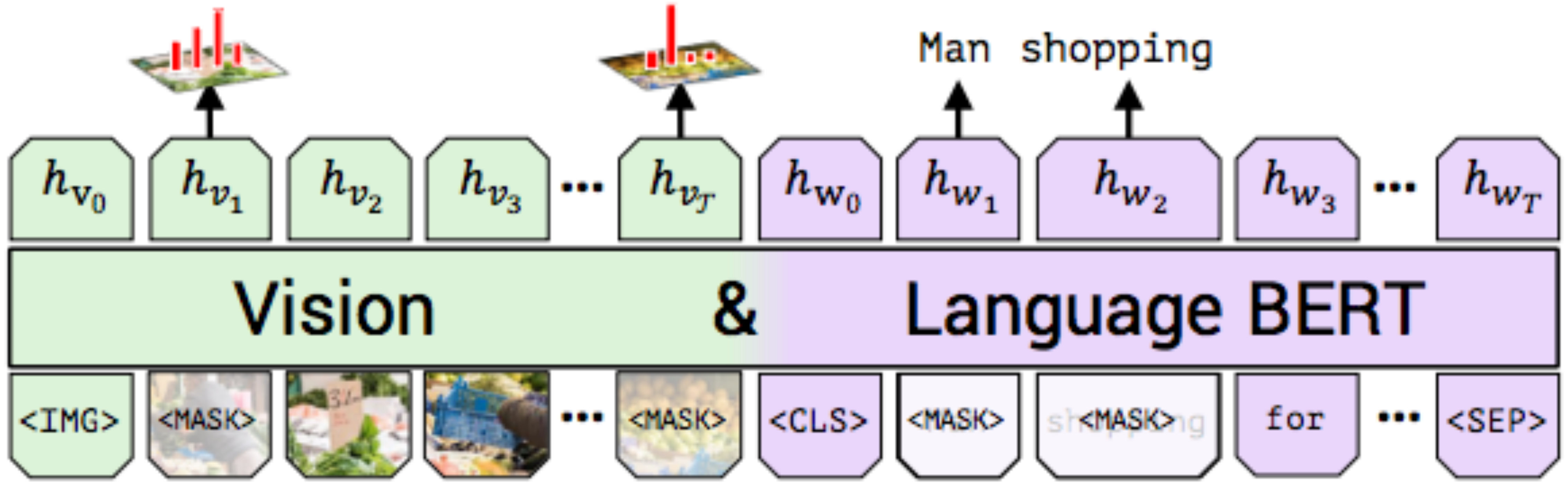
Lift [Person2] up.
Comfort [Person2].
Take [Person2] somewhere safer.
Because Person1 wanted to ...
Person1
Person2
After Person1 will most likely ...
Announce passengers to relax.
Get back to her seat.
Survey the damage.
Before Person2 needed to ...
Find something to hold onto.
Unfasten his seatbelt.
Fall to the floor.
Purchase a ticket.
Get prepared for his departure.
Before Person1 needed to ...
Get to the top of the deck.
Person1
Because Person1 wanted to ...
Say goodbye to his friend.
Express his excitement.
Person1
Person2
Because Person1 wanted to ...
Try to help [Person2].
After Person1 will most likely ...
Wait for help to arrive.
Get them out of the car.
Before Person2 needed to ...
Get knocked unconscious.
Crash the car.
Turn the steering wheel.
Sink in the water.
Save himself from drowning.
After Person1 will most likely ...
Because Person2 wanted to ...
Swim to safety.
Because Person1 wanted to ...
Person1
Person2
Before Person2 needed to ...
Swim towards the statute.
Notice water washing in.
Sense his own death.
Before Person1 needed to ...
Realize the ship is sinking.
Start moving against the water.
Get caught in a rush of water.
After Person2 will most likely ...
Be washed away.
Gasp for air.
Scream for help.
Get away from something.
Drive the car as quickly as he can.
Because Person1 wanted to ...
Person1
Turn to look out the windshield.
After Person1 will most likely ...
Step on the break.
Before Person1 needed to ...
Drive a car into the situation.
Feel terrified.
Try to get free.
Stay as hostage.
Get his hand off her mouth.
After Person1 will most likely ...
Because Person1 wanted to ...
Person2
Person1
Before Person1 needed to ...
Make noise.
Get Captured.
Adhere to police instruction.
Person1
Applaud for the speech.
After Person1 will most likely ...
Participate in the wave.
Start jumping in joy.
Person1
Because Person1 wanted to ...
Show approval.

1.4 million inferences over 60K images with https://visualcomet.xyz

Task: Generating Commonsense Inferences in Language

Our Model Builds on Pre-Trained Language Models





GPT-2 **for Conditional Generation**

Vision-Language Transformer Architecture

(Radford et. al., 2019)

(Lu et. al., 2020; Su et. al, 2020; Tan et. al, 2020)

Our Approach

<startimg>   <endimg>

Visual Context

<startevent> <Person2> Is Holding Onto ··· <endevent>

Text Event

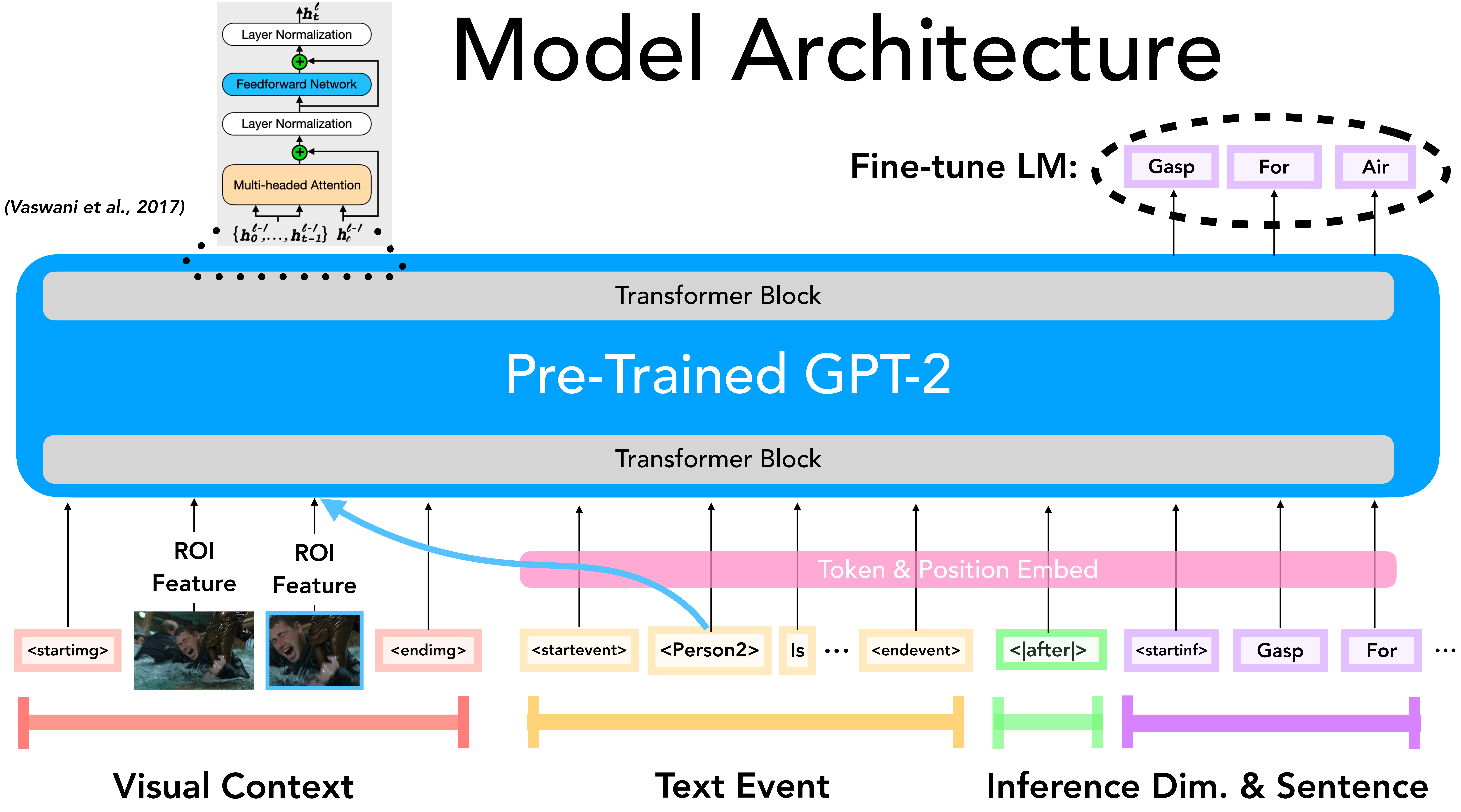
<|after|>

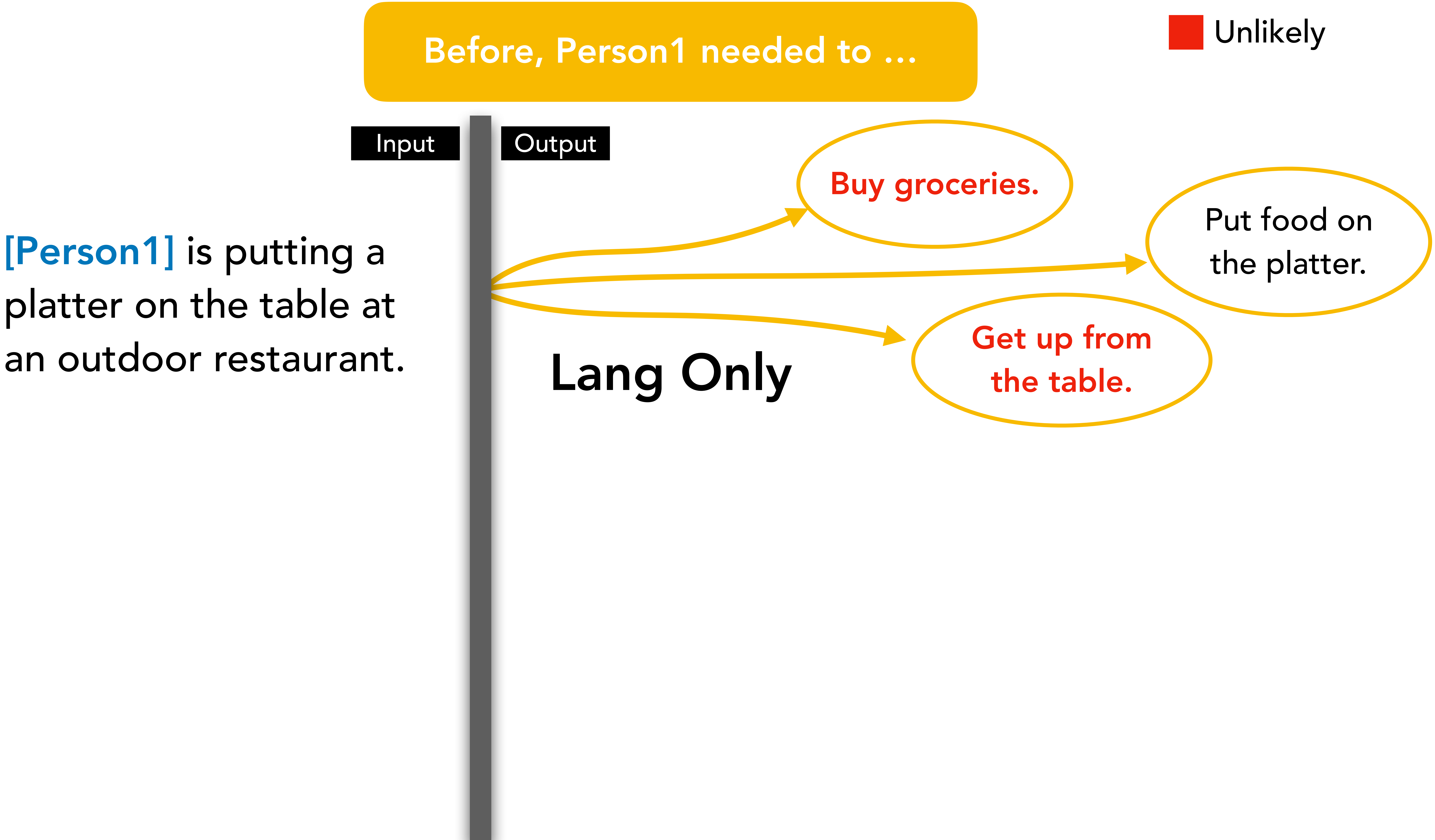
Inference Dimension

<startinf> Gasp For Air <endinf>

Inference Sentence

Model Architecture



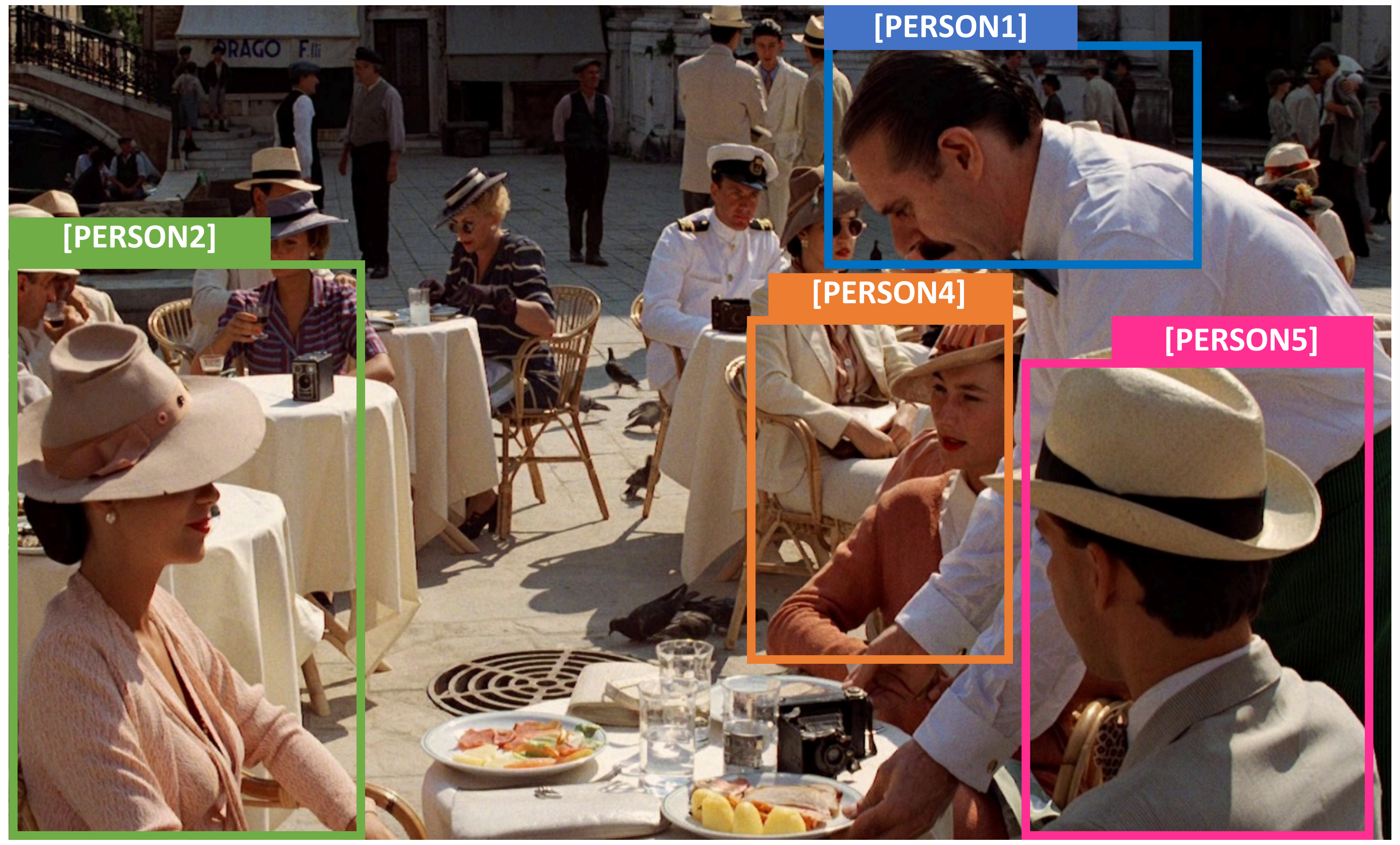

Before, Person1 needed to ...
Unlikely
Input
Output
[Person1] is putting a platter on the table at an outdoor restaurant.
Buy groceries.
Put food on the platter.
Get up from the table.
Lang Only

Before, Person1 needed to ...

Unlikely

Input

[Person1] is putting a platter on the table at an outdoor restaurant.



Output

Lang Only

- Buy groceries.
- Put food on the platter.
- Get up from the table.

Vision + Lang

- Wait for everyone to sit down.
- Be hired as a waiter.
- Receive an order for platter.

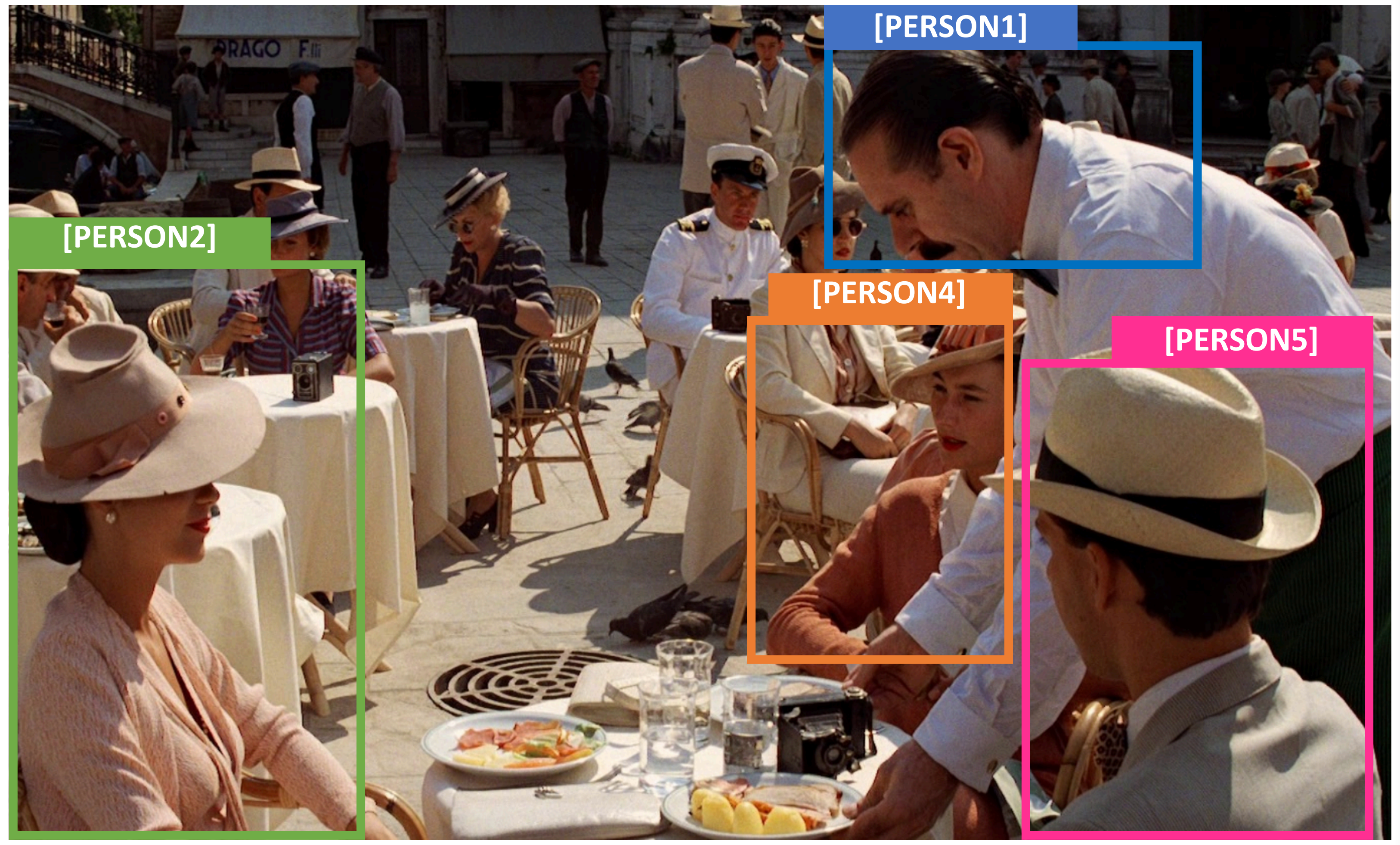
Because, Person1 wanted to …

Unlikely

Input

Output

[Person1] is putting a platter on the table at an outdoor restaurant.



Have dessert

Tend to the patrons.

Ensure the food is taken care of.

Lang Only

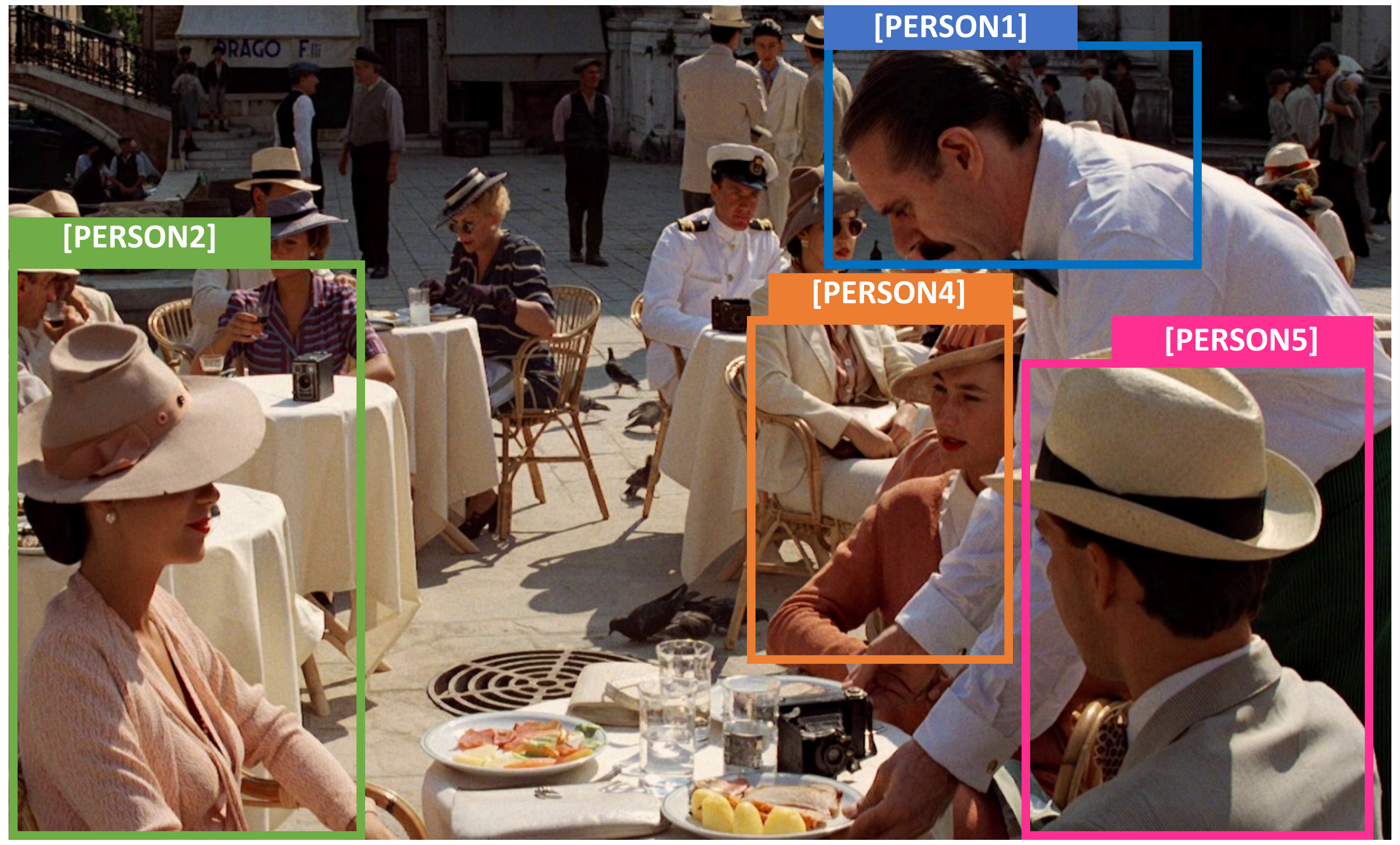
Because, Person1 wanted to ...
Unlikely
Input
Output
[Person1] is putting a platter on the table at an outdoor restaurant.
[PERSON1]
[PERSON2]
[PERSON4]
[PERSON5]
Have dessert
Tend to the patrons.
Ensure the food is taken care of.
Lang Only
Greet [P2], [P4], and [P5].
Serve [P2], [P4], and [P5].
Have [P2], [P4], and [P5] to eat.
Vision + Lang

After, Person1 will most likely ...

Unlikely

Input

[Person1] is putting a platter on the table at an outdoor restaurant.



Output

Sip the water.

Ask [P2] for a menu.

Get up and walk over to his table.

Lang Only

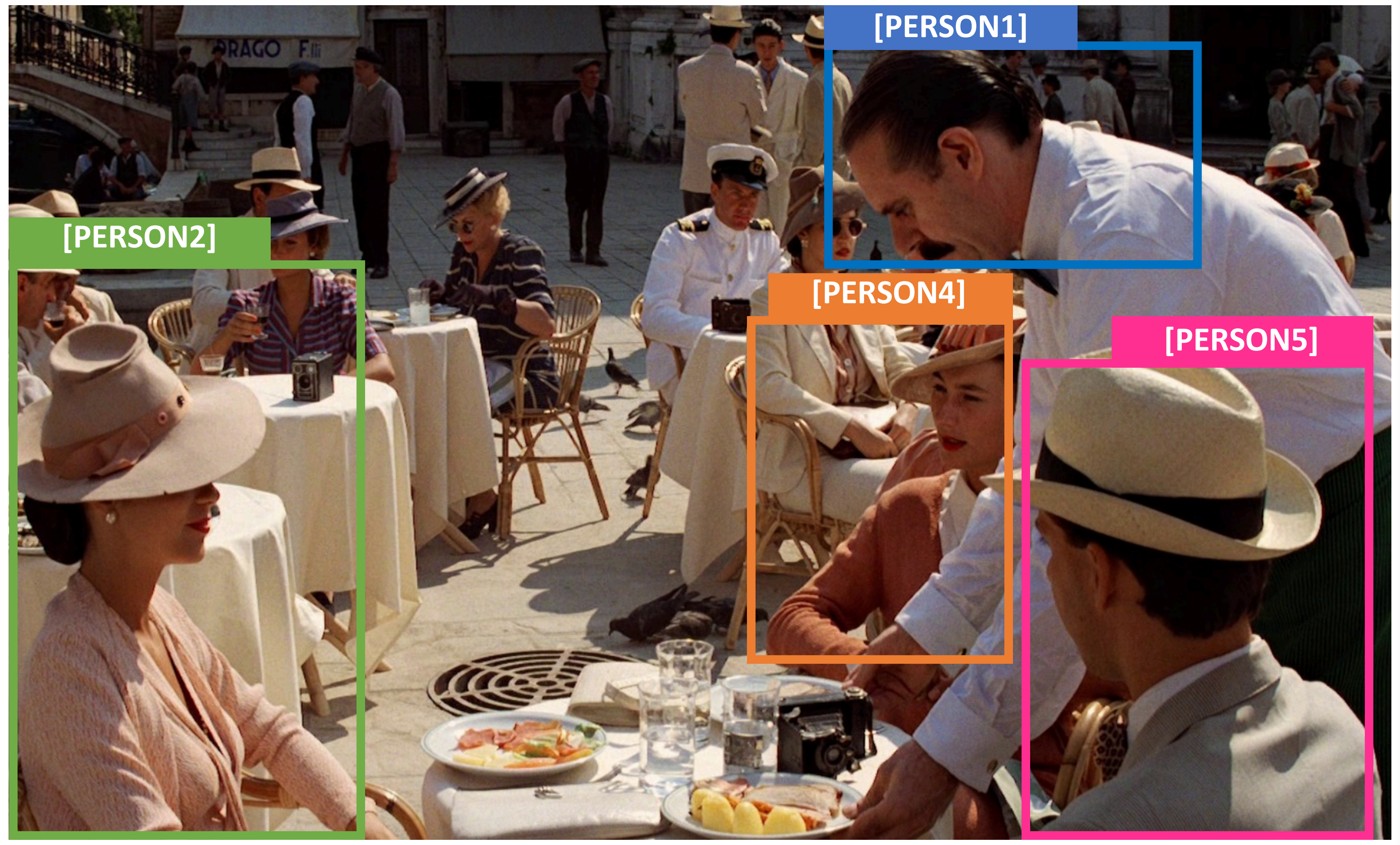
After, Person1 will most likely …

Unlikely

Input

Output

[Person1] is putting a platter on the table at an outdoor restaurant.



Lang Only

Sip the water.

Ask [P2] for a menu.

Get up and walk over to his table.

Vision + Lang

Get back to his work duties.

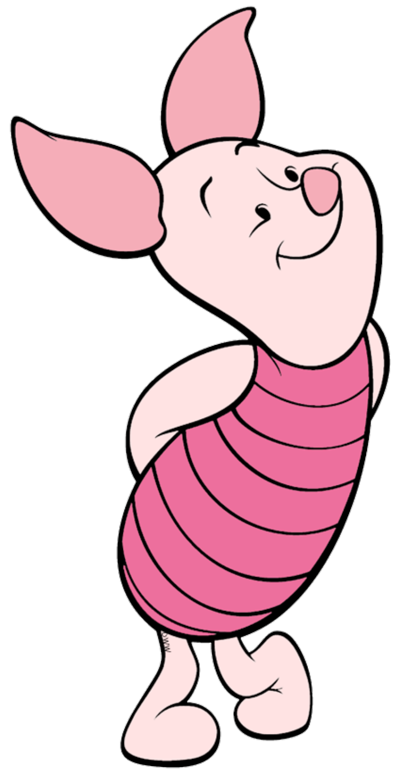
Take drinks.

Get back to the **kitchen** to get more food.

Harnad's Symbol Grounding Problem

- **Grounding with 3D**

Interactions at the cost of concept coverage



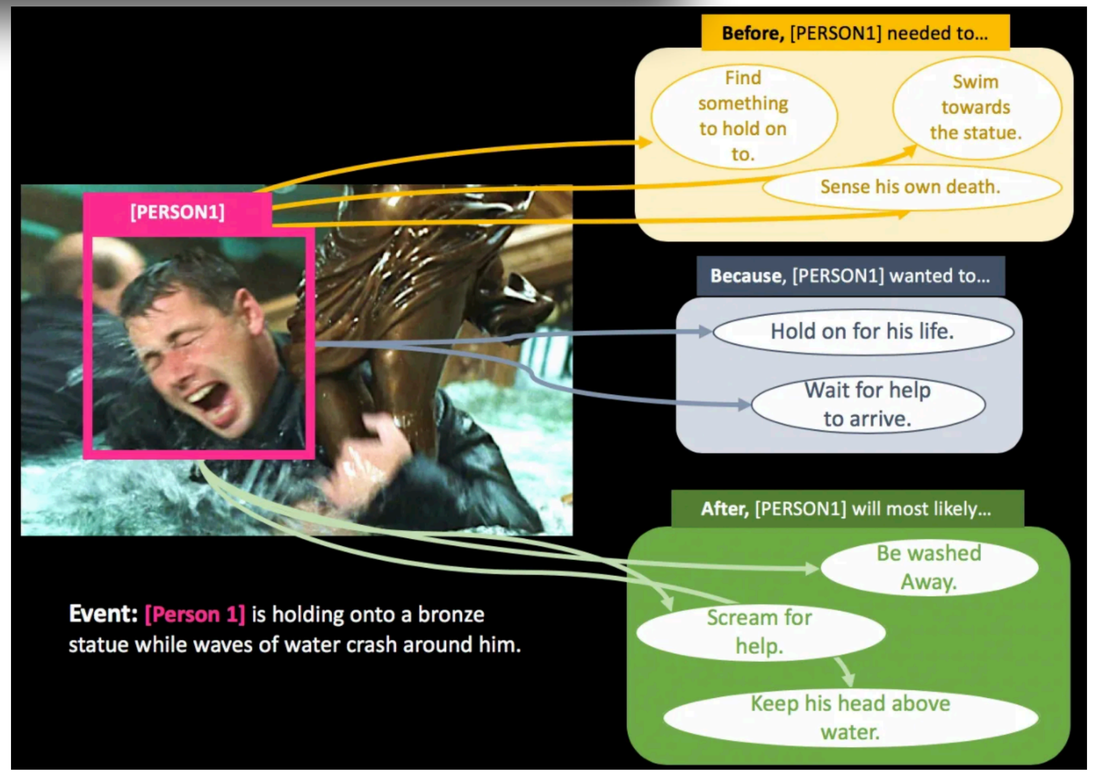
PIGLeT

- **Grounding with 2D + Time**

Far richer concepts (causal / temporal interactions) at the cost of direction interactions with the world

- **Grounding with 2D + KG**

Learning only from raw data vs from rich declarative knowledge about the world



Thanks! Questions?