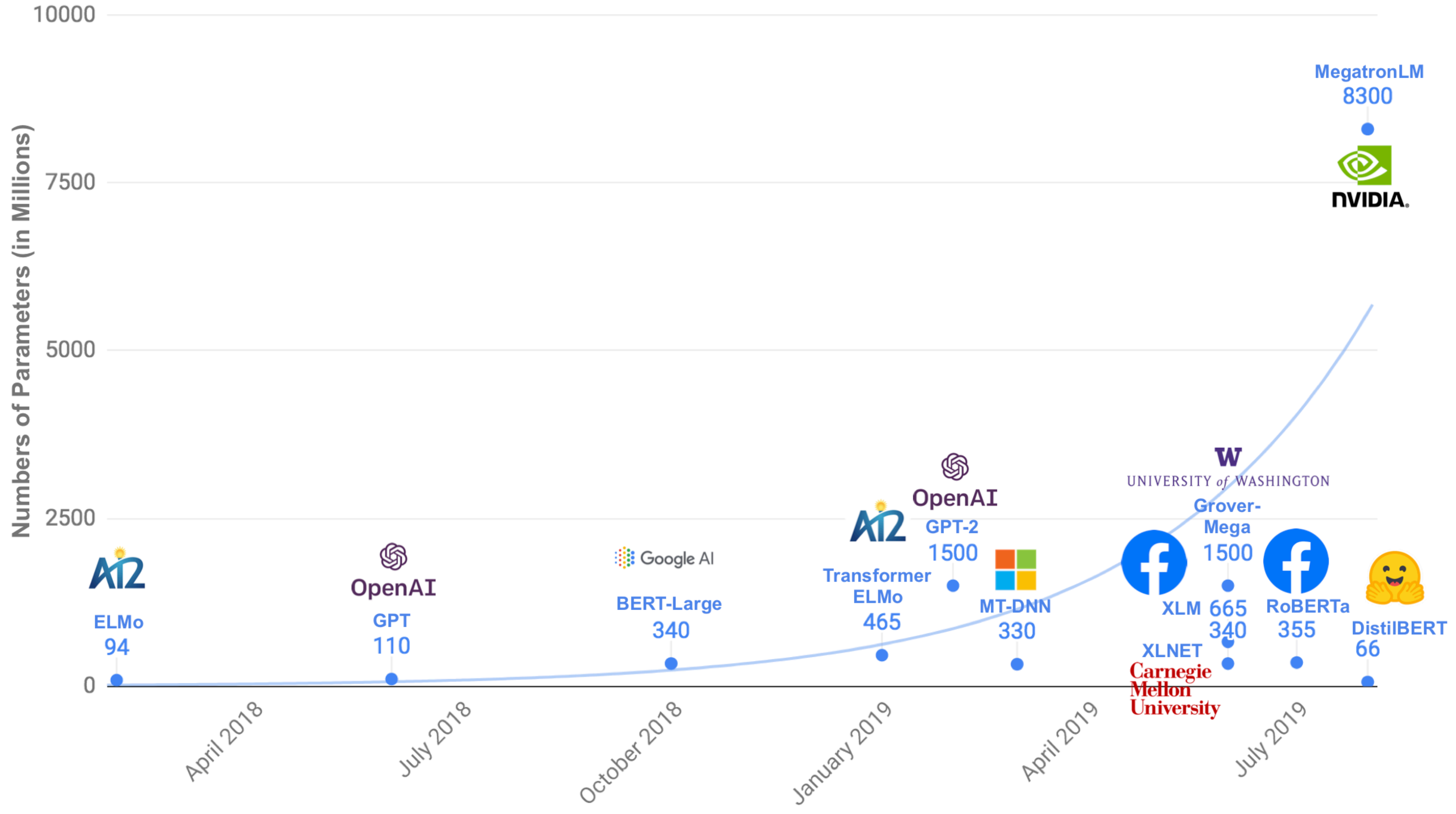
Learning Commonsense Understanding through Language and Vision

Rowan Zellers

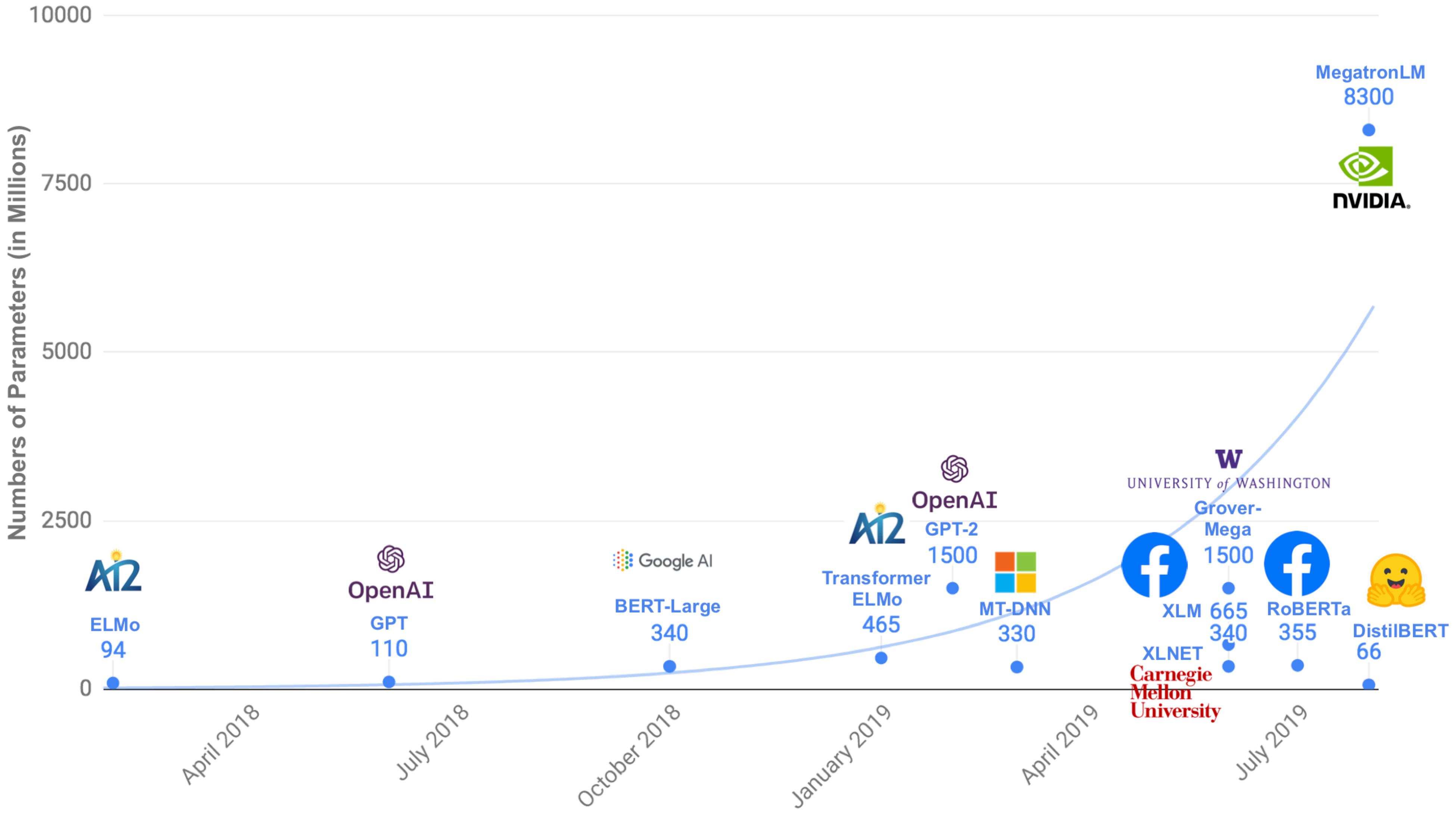
Paul G. Allen School of Computer Science & Engineering
University of Washington &
Allen Institute for Artificial Intelligence

AI progress... vs humans

exponential increase in model scale



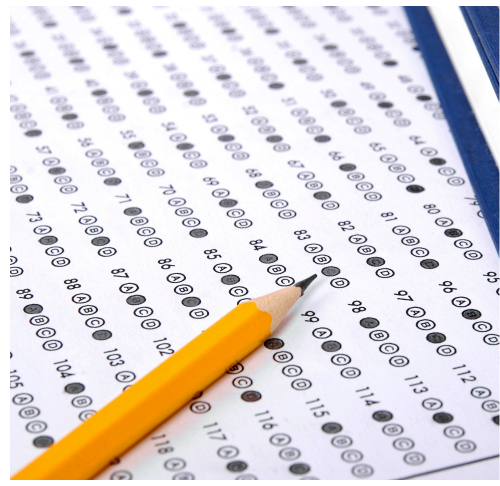
exponential increase in model scale



Strong performance on...

Text: multiple-choice QA

Raffel et al 2019, Brown et al 2020, inter alia



Vision: webly supervised classification, detection

Radford et al 2021, Kamath et al 2021, inter alia

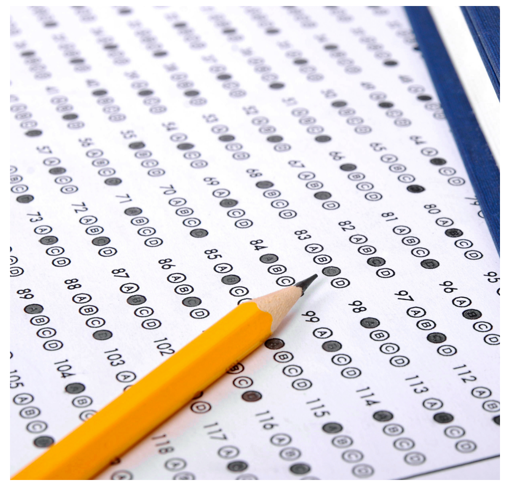
Panda

Vision + Language: learning from captions

Chen et al 2019, Zhang et al 2021, inter alia

A train on the tracks

Text: multiple-choice QA



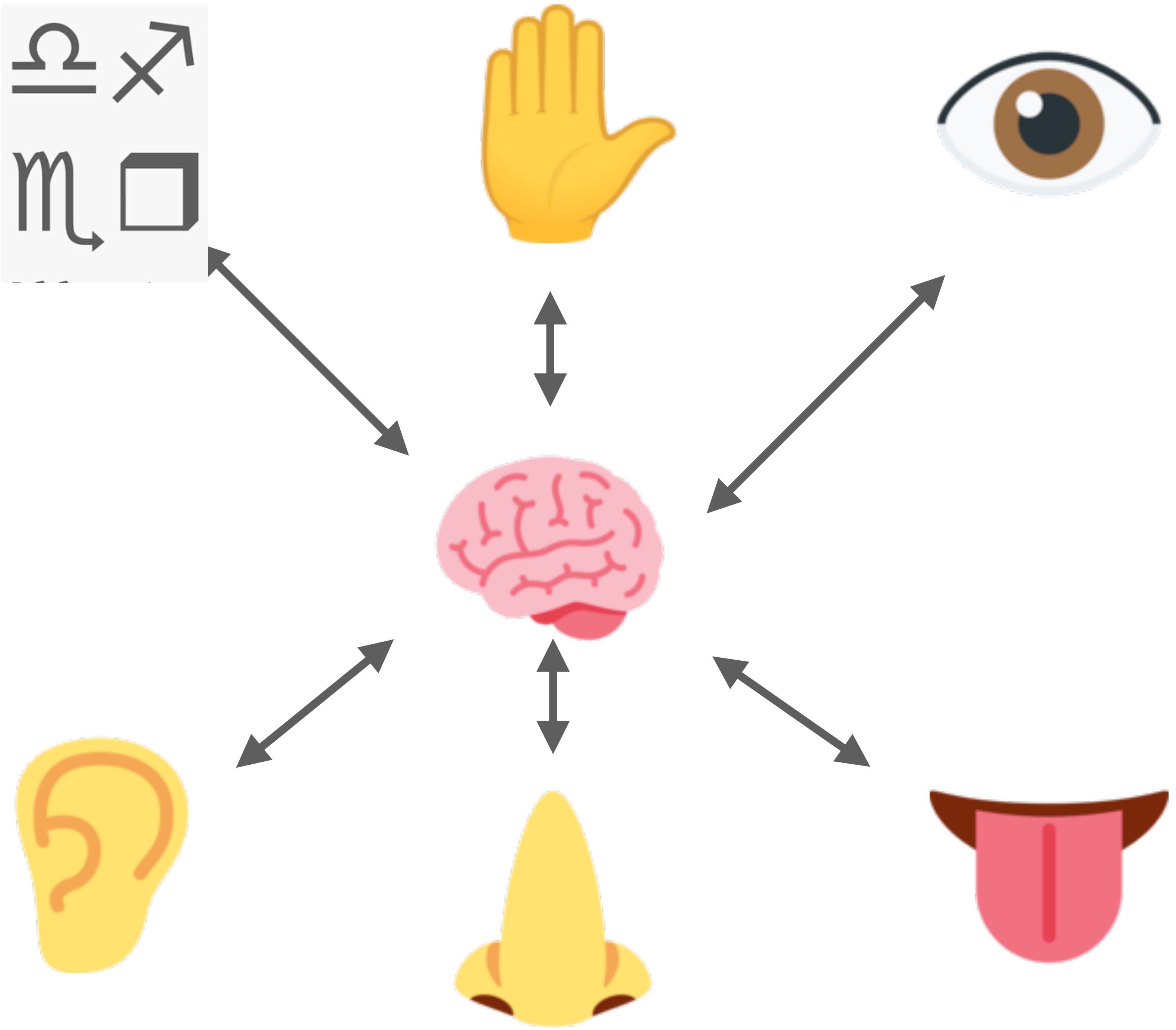
Vision: webly supervised classification, detection

Panda

Vision + Language: learning from captions

A train on the tracks

Humans....

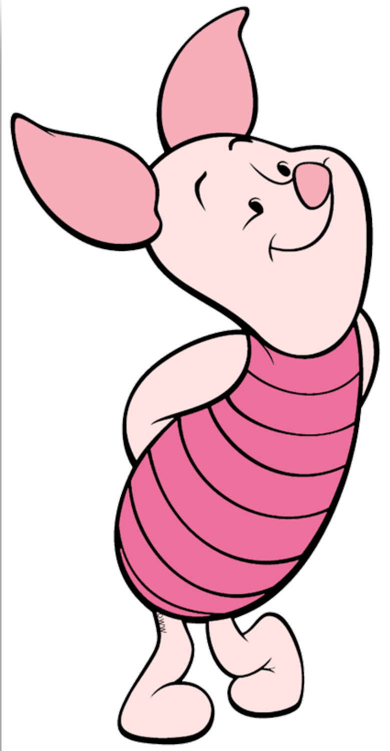


- *Integration of many modalities, learned from interaction*
- *Grounded in events, and daily life*

Today's talk

- *Integration of many modalities, learned from interaction*
- *Grounded in events, and daily life*

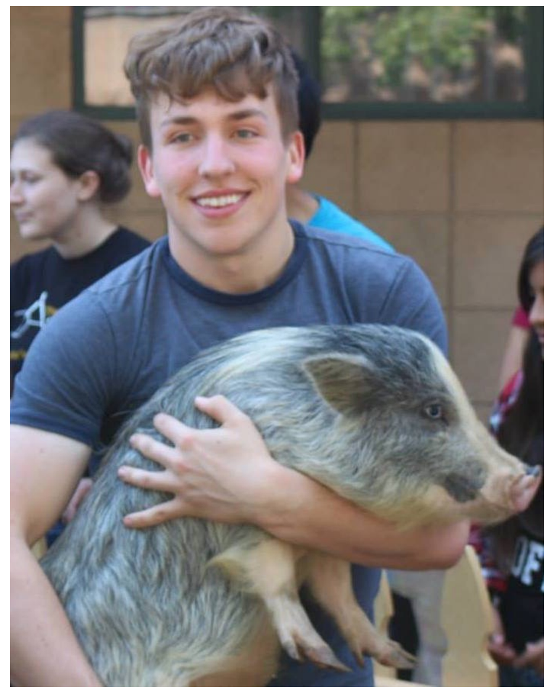
- *Integration of many modalities, learned from interaction*



PIGLeT

Language Grounding Through Neuro-Symbolic Interaction in a 3D World
(ACL 2021)

Me



Ari Holtzman

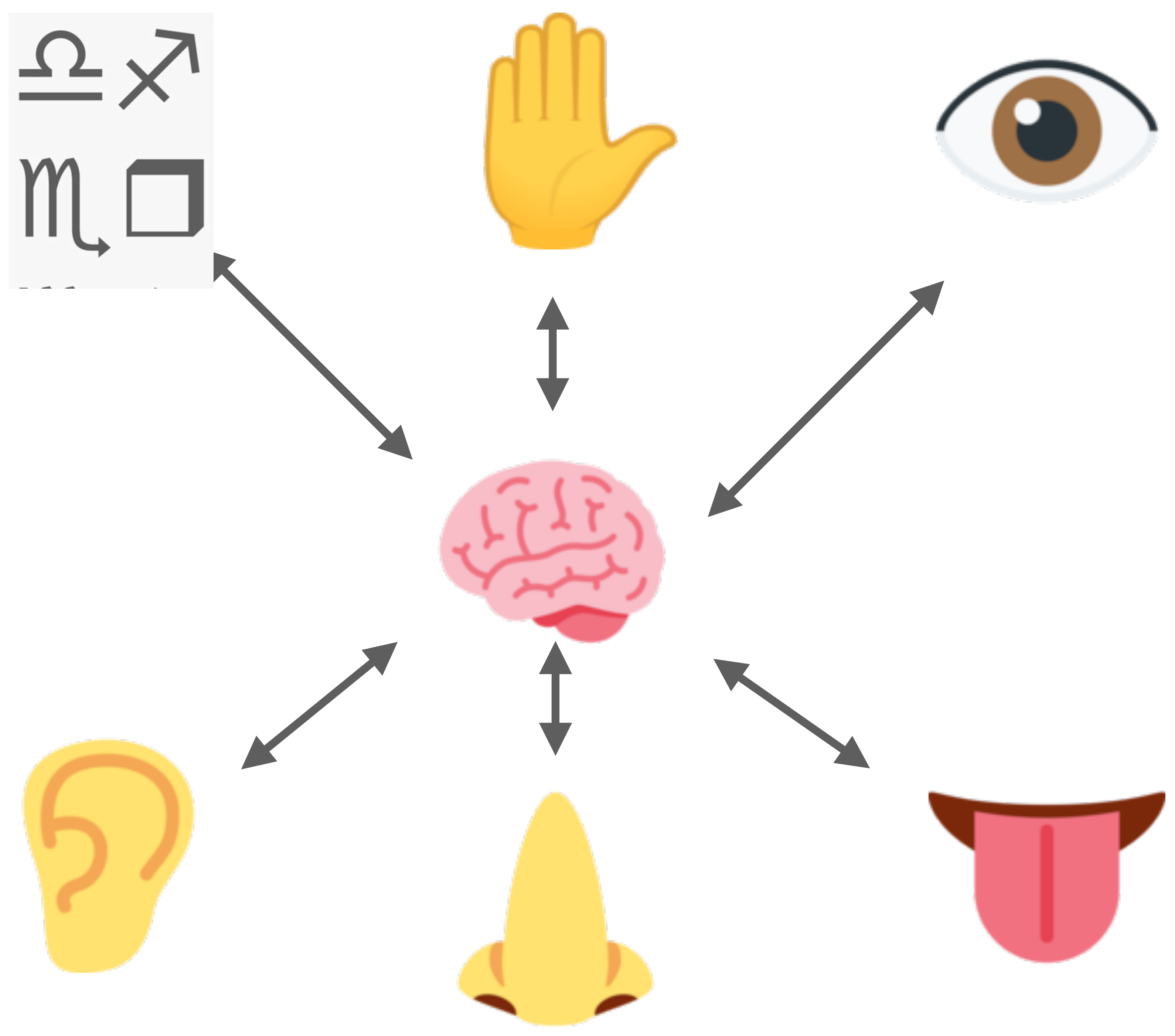
Matthew Peters

Roozbeh Mottaghi

Aniruddha Kembhavi

Ali Farhadi

Yejin Choi



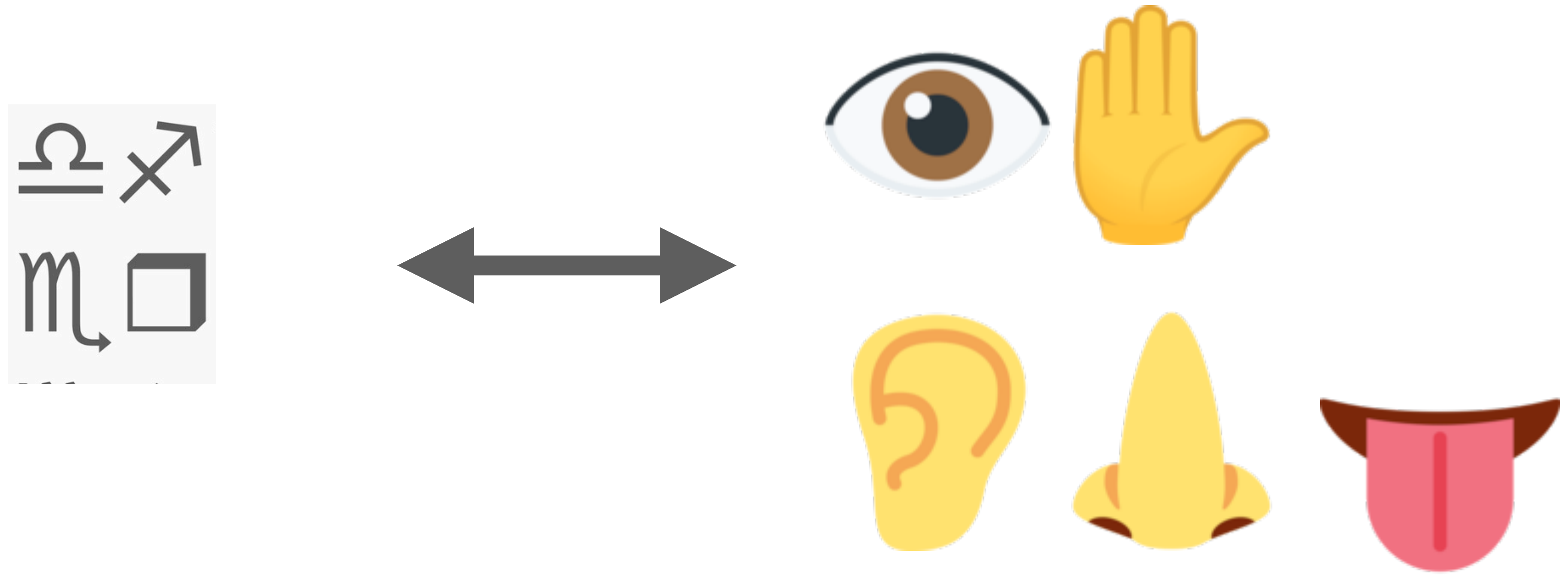
Problem: a gap between ***language form*** and ***commonsense grounded meaning***

Written language
(*symbols*)

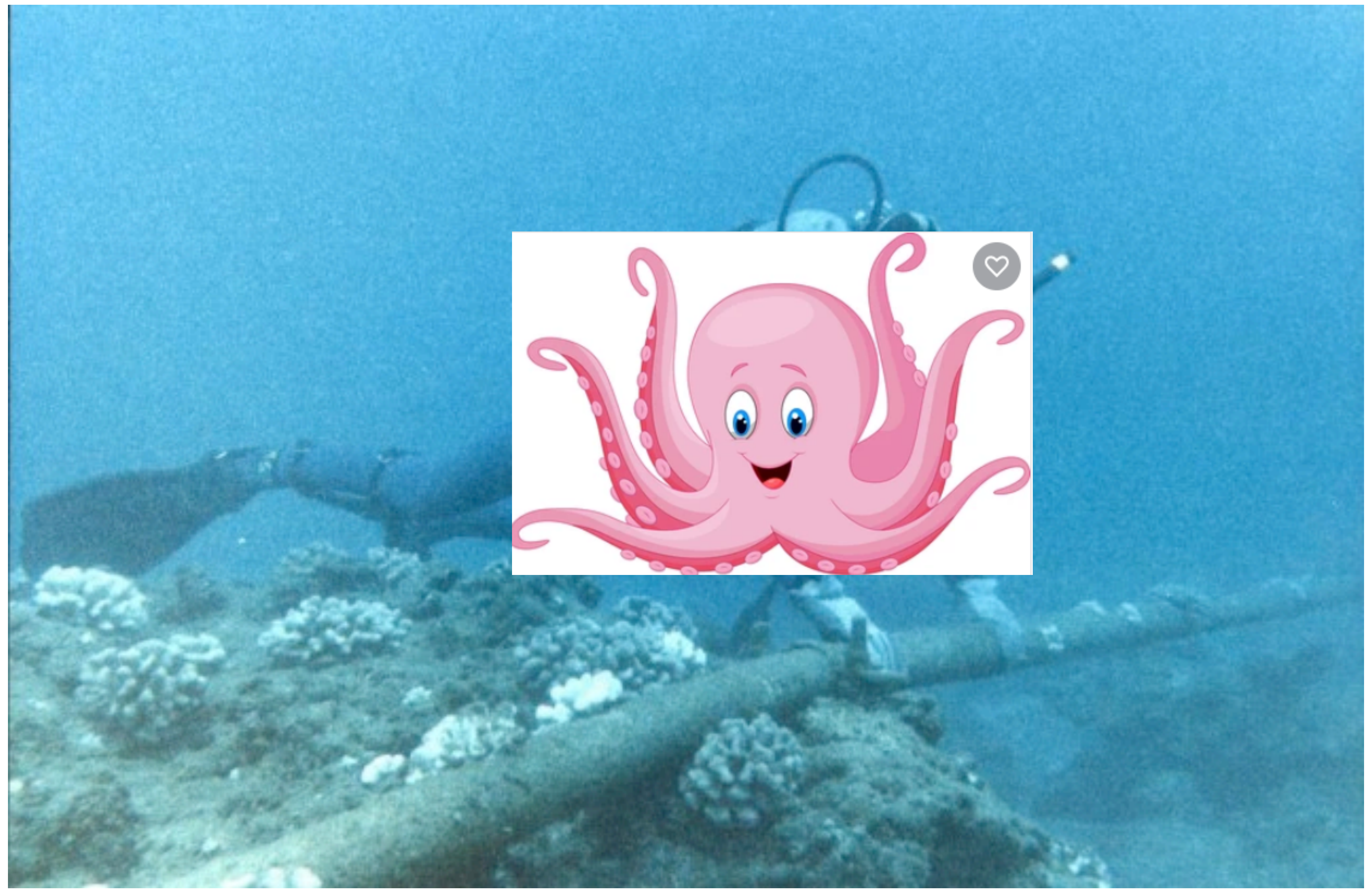
The world
(*continuous, subjective experience*)

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

Name:	Mug
Temperature:	RoomTemp
isBreakable	True
isFilledWithLiquid	True
…	

grounds

"I accidentally dropped the mug and it broke"

"I filled up my mug with coffee"

"I'm holding that mug with my hand"

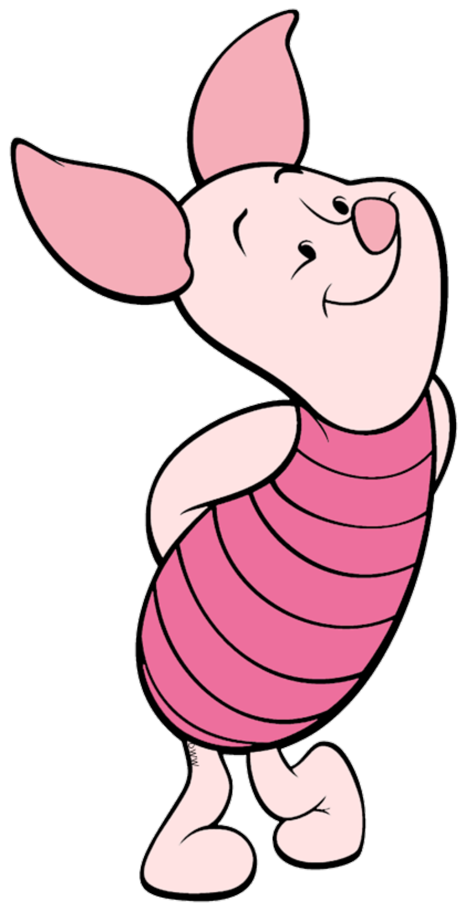
"Careful touching that mug, it's hot"

Learning from THOR

- THOR: An interactive 3D environment with 20 actions, 125 object types
- Actions are contextual
- Objects have a state (expressed by 42 attributes)

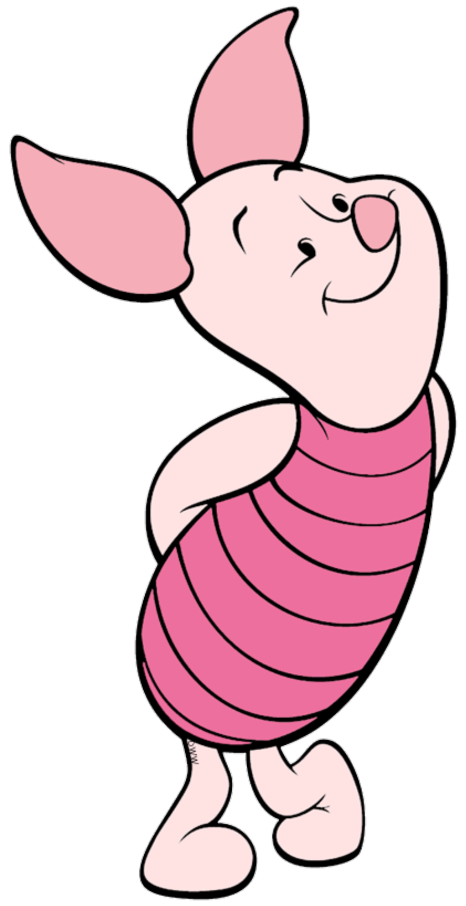
Model: PIGLeT

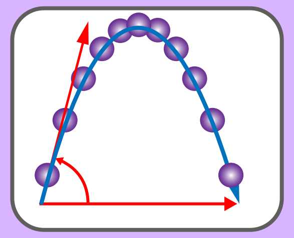
- We'll predict *explicitly* "what happens next" to an object given an event written out in English
- Or, write an English sentence summarizing the state change.



PIGLeT: Physical Interactions as Grounding for Language Transformers

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



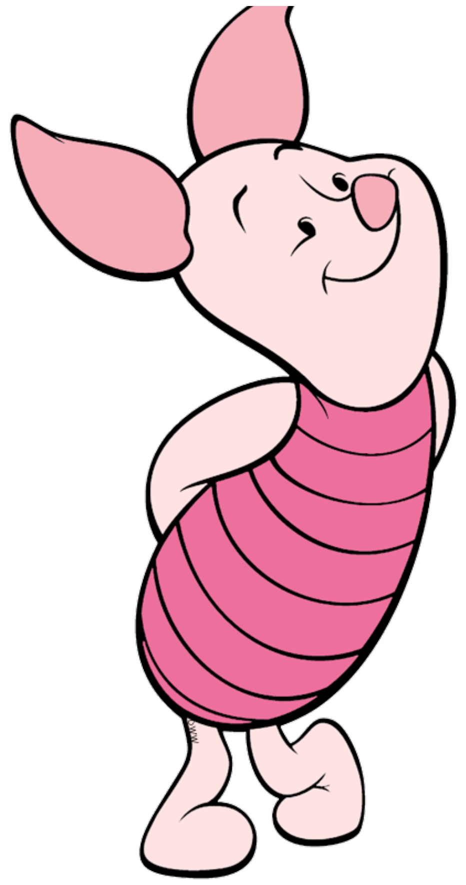
 Physical Dynamics Model

 Language Model

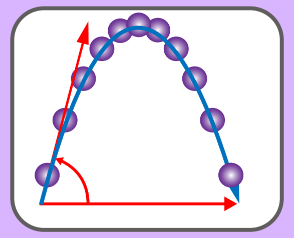
Learning “How the World Works”

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

...



<heatUp, Pan>



Physical Dynamics Model

Language Model

Learning "How the World Works"

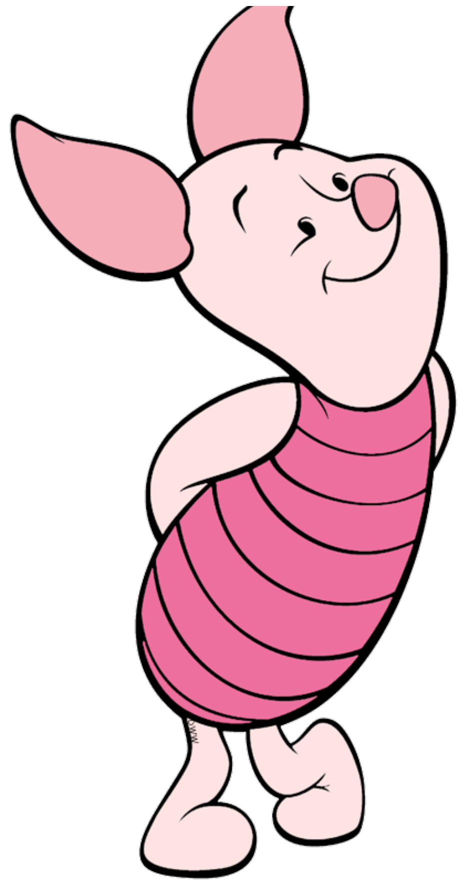
Name:	Egg
Temperature:	RoomTemp
isCooked:	False
isBroken:	True

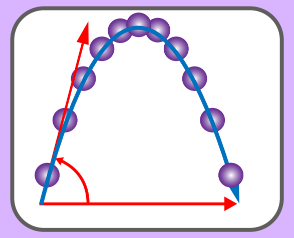
…

<heatUp, Pan>

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

…





Physical Dynamics Model

Language Model


Physical Dynamics Model
Name: Egg
Temperature RoomTemp
isCooked: False
isBroken: True
...
Object Encoder (Transformer)
<heatUp, Pan>
Action Encoder (MLP)
Action Apply MLP
Object Decoder (Transformer)
Name: Egg
Temperature: Hot
isCooked: True
isBroken: True
...


Physical Dynamics Model
Name: Egg
Temperature RoomTemp
isCooked: False
isBroken: True
...
Object Encoder
<heatUp, Pan>
Action Encoder
Action Apply
Object Decoder
Name: Egg
Temperature: Hot
isCooked: True
isBroken: True
...
Pretrain the physical dynamics model with a cross-entropy loss to predict "what happens next" over 280k situations

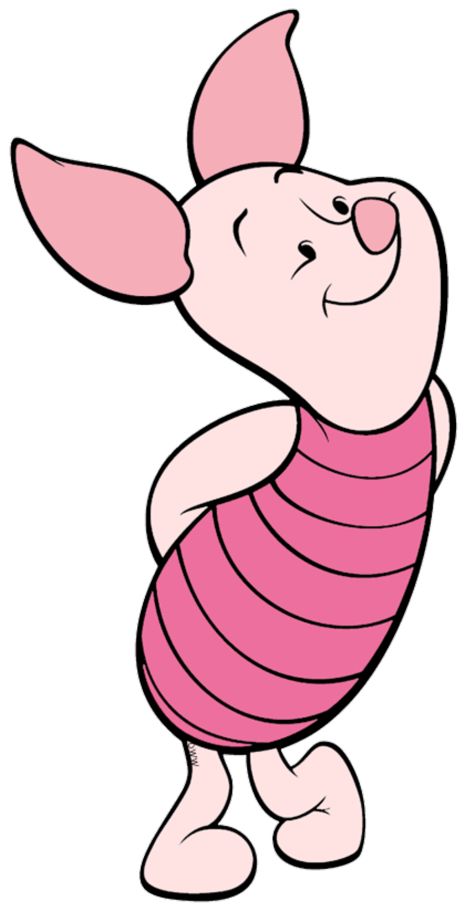
Physical Dynamics Model
Name: Egg
Temperature RoomTemp
isCooked: False
isBroken: True
...
Object Encoder
<heatUp, Pan>
Action Encoder
Action Apply
Object Decoder
Name: Egg
Temperature: Hot
isCooked: True
isBroken: True
...
The robot heats up the pan.
Language Model
Language Model
The pan becomes hot,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

Physical Dynamics Model
Name: Egg
Temperature RoomTemp
isCooked: False
isBroken: True
…
Object Encoder
<heatUp, Pan>
Action Encoder
Action Apply
Name: Egg
• We assume 500 paired (language, situation) examples, which we use to finetune the combined model.
• Both sub-models are pretrained separately, and the total model is BERT-Base sized.
The robot heats up the pan.
Language Model
Language Model
The pan becomes hot, and the egg gets cooked.

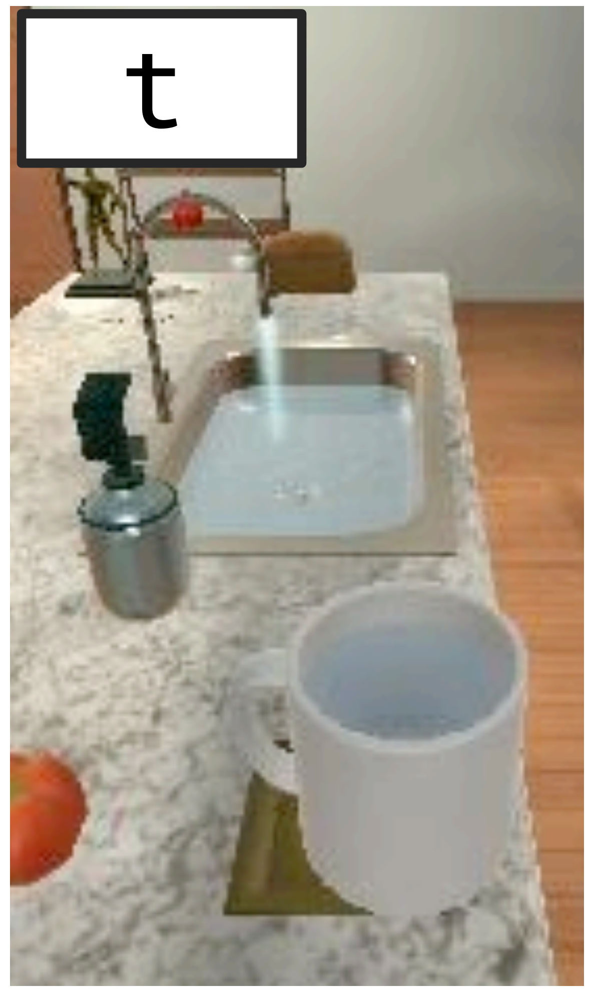
Model: PIGLeT

- We'll predict ***explicitly*** "what happens next" to an object given an event written out in English

- Or, write an English sentence summarizing the state change.



- predict "what happens next" to an object given an event written out in English



Name:	Sink
filledWith Liquid	True

Name:	Mug
filledWith Liquid	True
isPickedUp	True

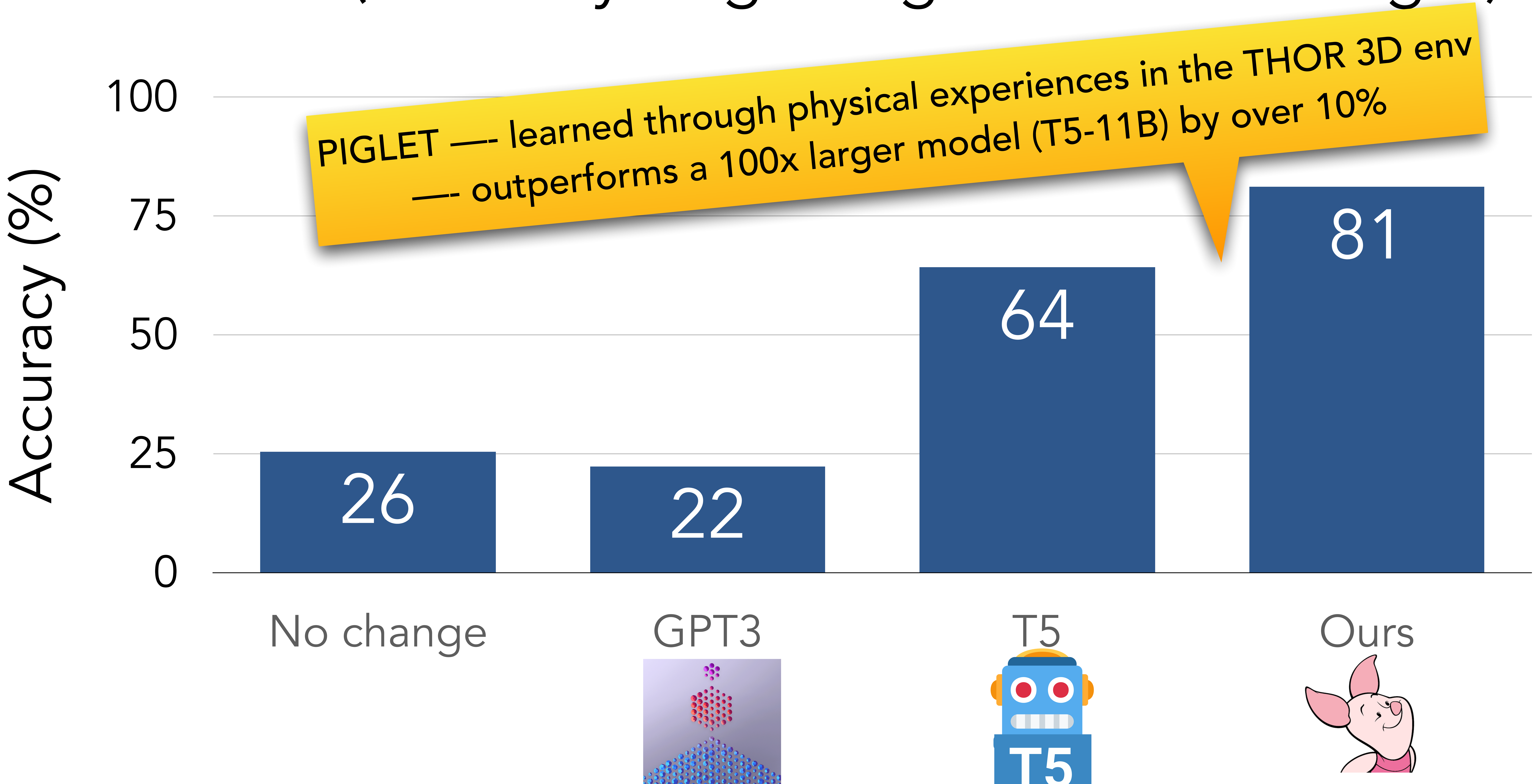
The robot empties the mug.

Name:	Sink
filledWith Liquid	True

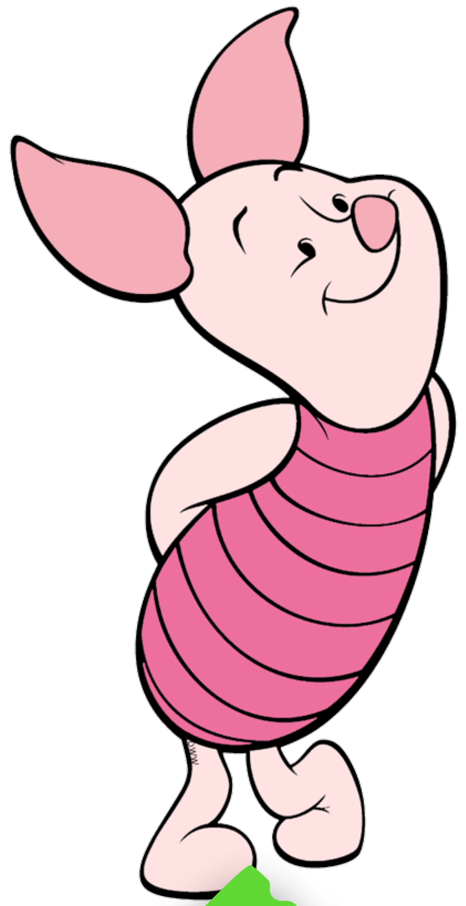
Name:	Mug
filledWith Liquid	***False***
isPickedUp	True

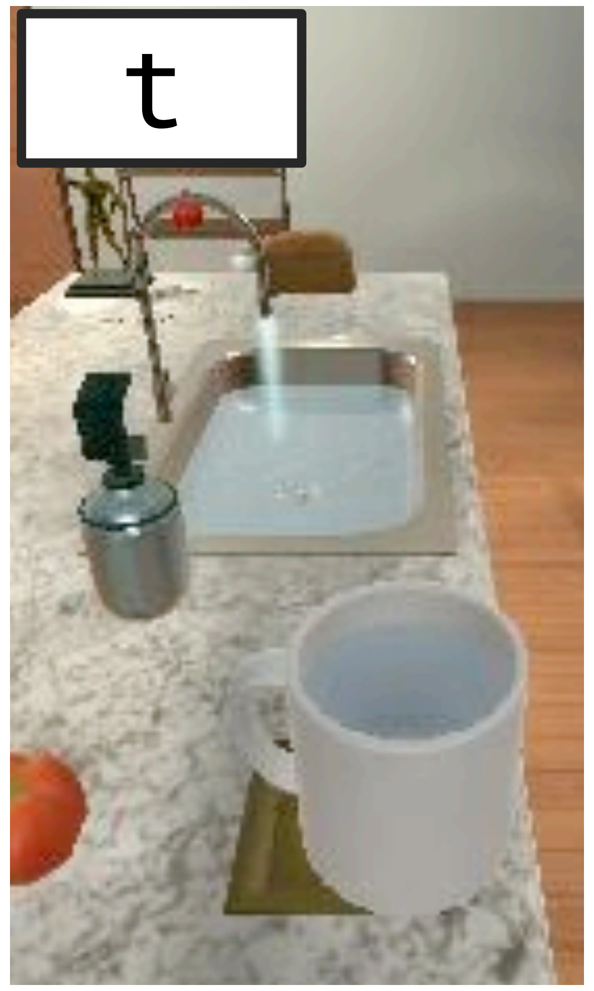
Evaluation: Accuracy (of getting all attributes right)

Results (accuracy of getting all attributes right)



Qualitative Example





Name:	Sink
filledWith Liquid	True

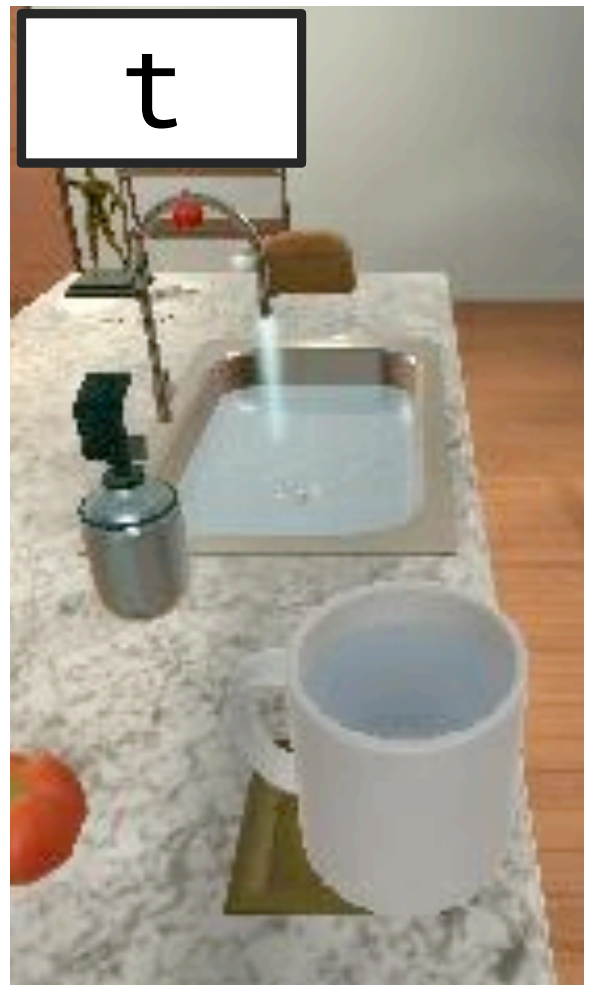
Name:	Mug
filledWith Liquid	True
isPickedUp	True

The robot empties the mug.

Name:	Sink
filledWith Liquid	True

Name:	Mug
filledWith Liquid	***False***
isPickedUp	True

Qualitative Example

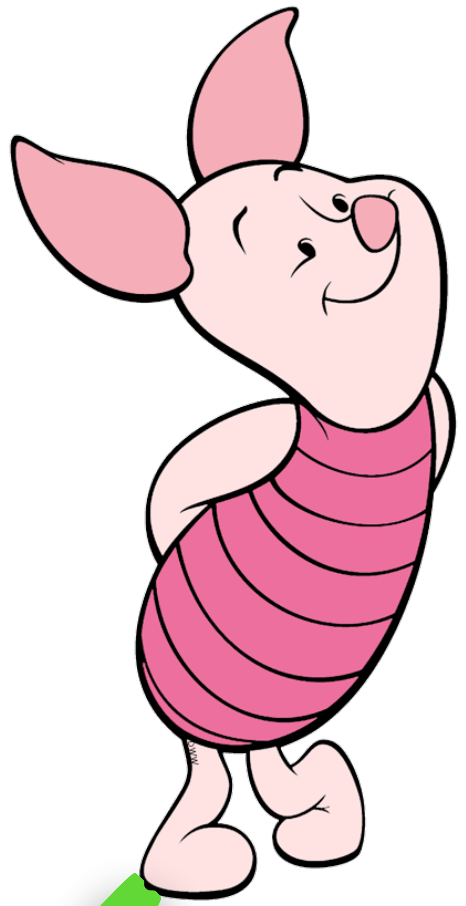


Name:	Sink
filledWith Liquid	True

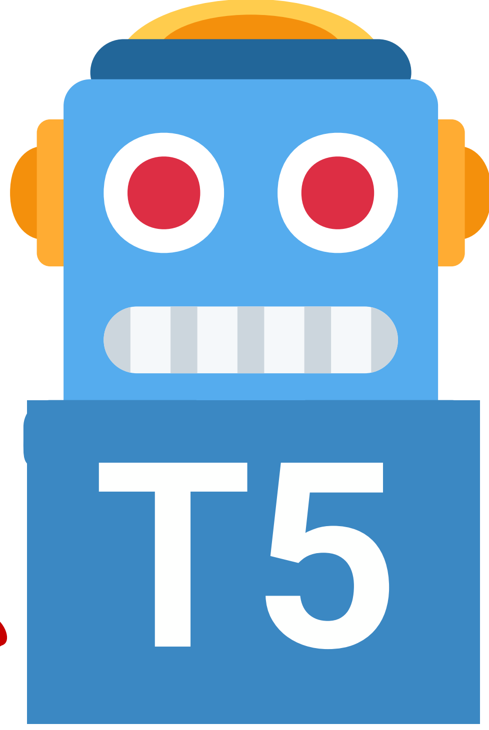
Name:	Mug
filledWith Liquid	True
isPickedUp	True

The robot empties the mug.

Name:	Sink
filledWith Liquid	True



Name:	Sink
filledWith Liquid	False

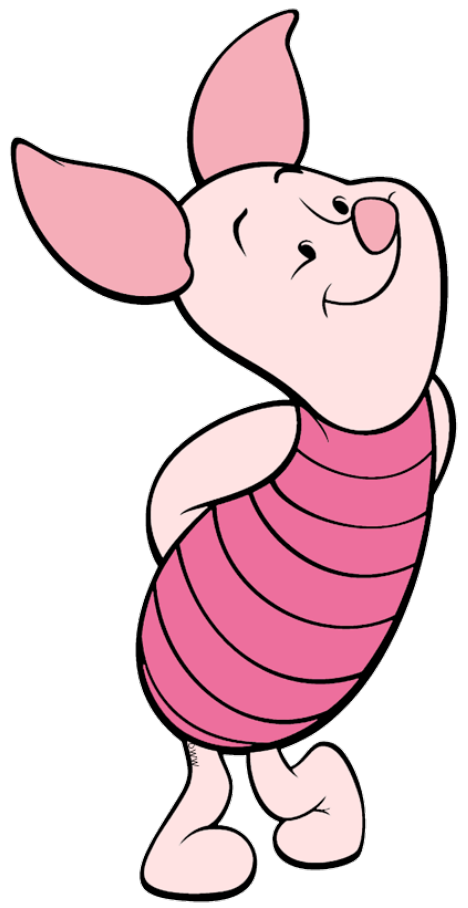


T5-11B, through text, learns “emptying liquid from an object” makes all objects in the room empty

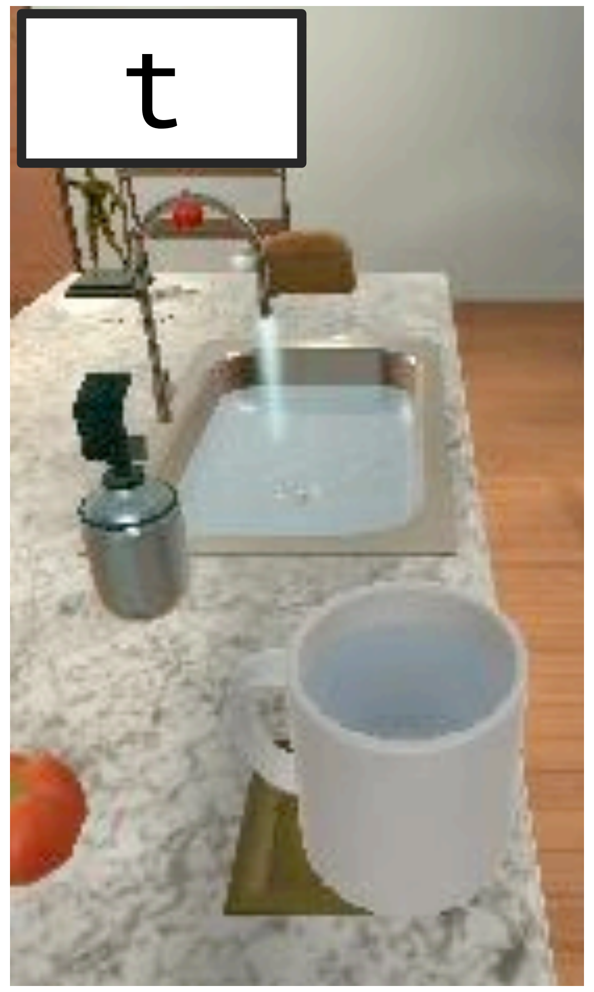
Model: PIGLeT

- We'll predict *explicitly* "what happens next" to an object given an event written out in English

- Or, write an English sentence summarizing the state change.



- and summarize this prediction in English



Name:	Sink
filledWith Liquid	True

Name:	Mug
filledWith Liquid	True
isPickedUp	True

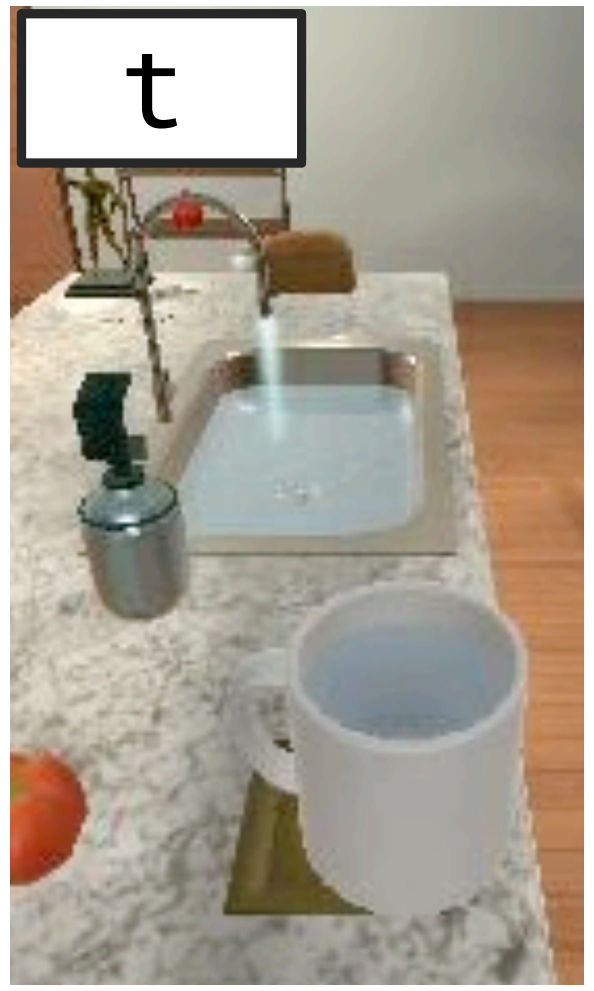
<empty, Mug>

The mug is no longer filled with water.

Evaluation: human, BLEU, BERTScore

Human eval [-1, 1]
1.0
0.5
0.0
-0.5
-1.0
0.82
0.92
0.15
0.22
T5
T5
Fluency
Faithfulness

PIGLeT's generations

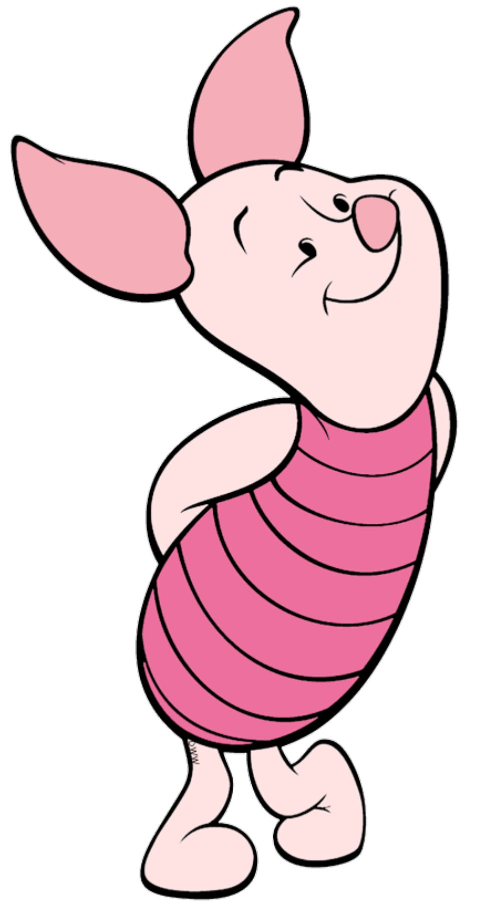


Name:	Sink
filledWith Liquid	True

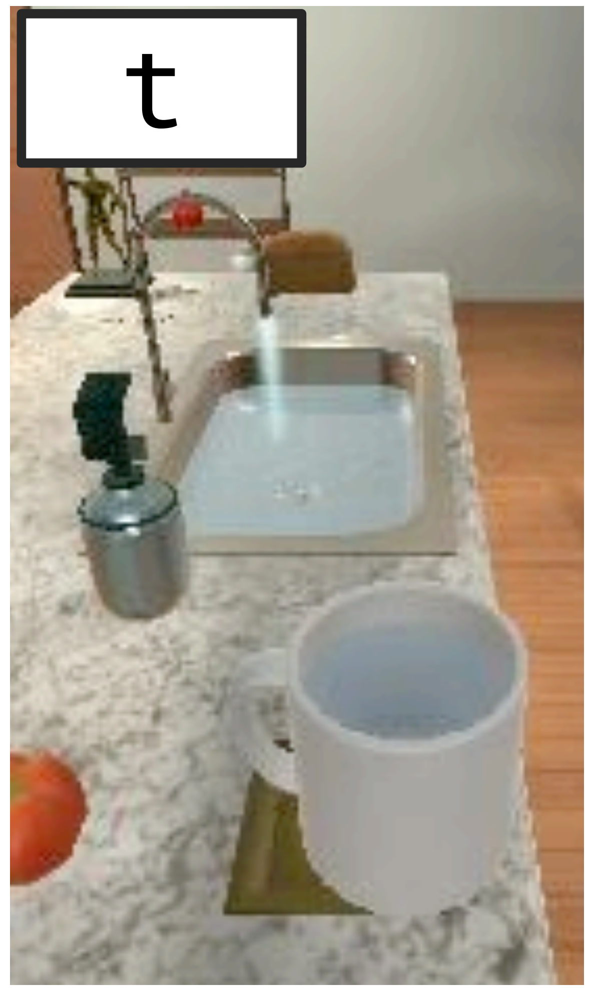
Name:	Mug
filledWith Liquid	True
isPickedUp	True

<empty, Mug>

The mug is now empty.



PIGLeT's generations

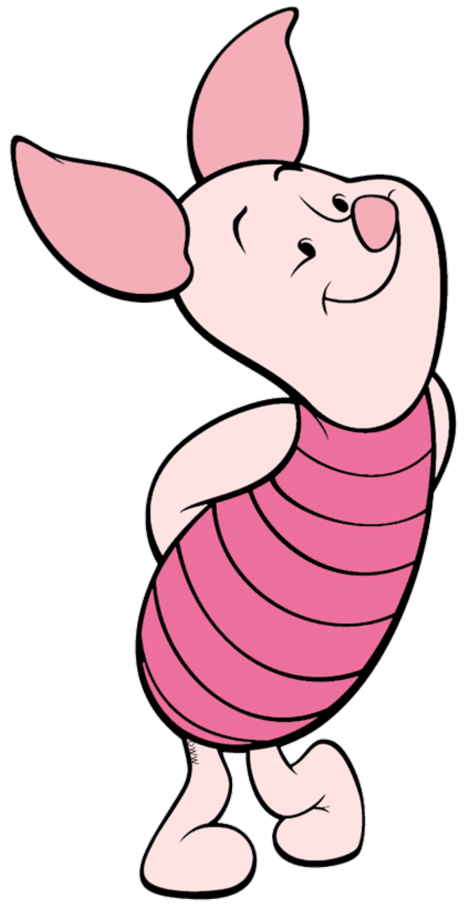


Name:	Sink
filledWith Liquid	True

Name:	Mug
filledWith Liquid	True
isPickedUp	True

<empty, Mug>

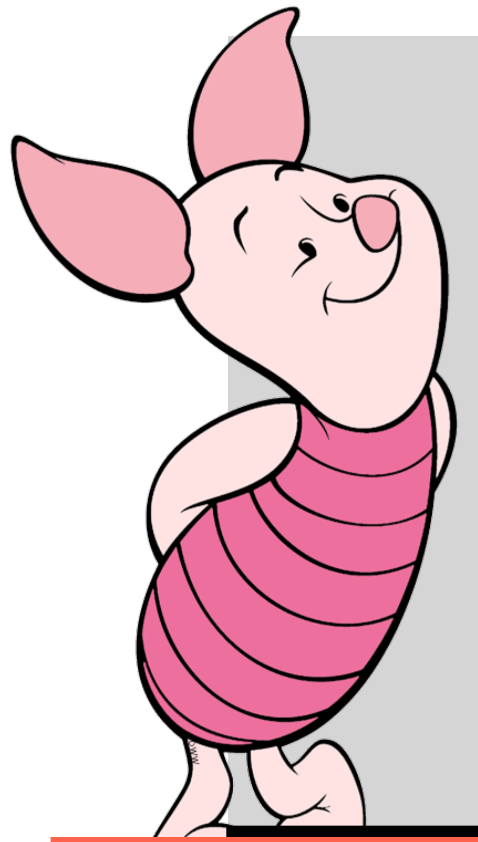
The mug is now empty.



The sink is now empty.

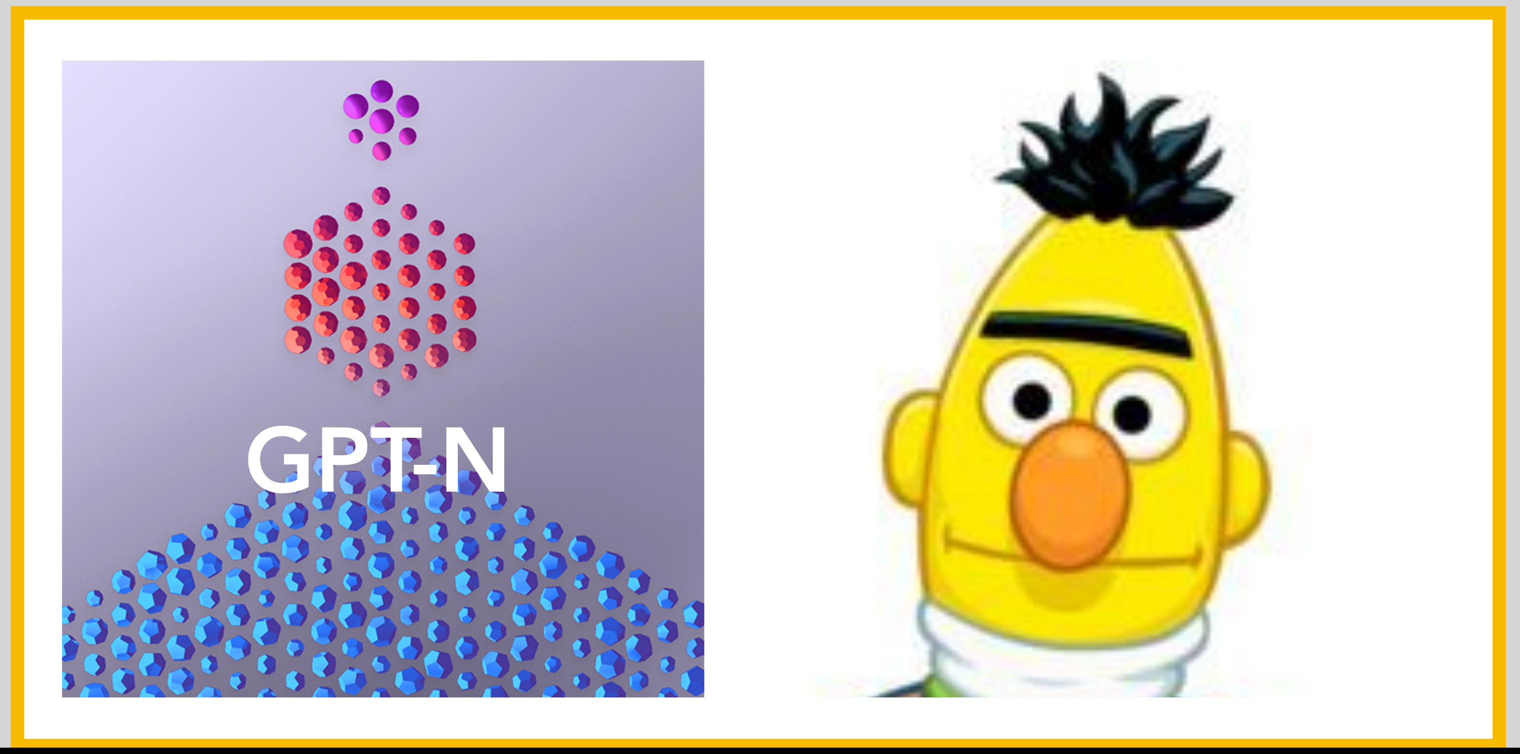
T5

PIGLET vs text-only learning



Physical Dynamics Model

Language Model



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
Paper-only bonus!! Can generalize to new concepts like "Dax" without words	Limited generalization to new concepts

Today's talk

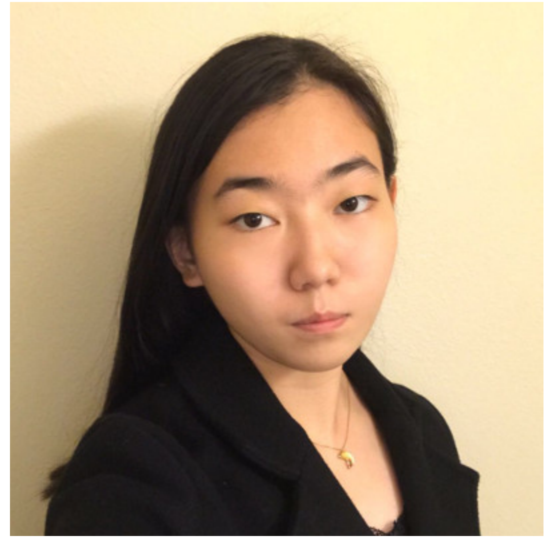
- *Integration of many modalities, learned from interaction*
- *Grounded in events, and daily life*

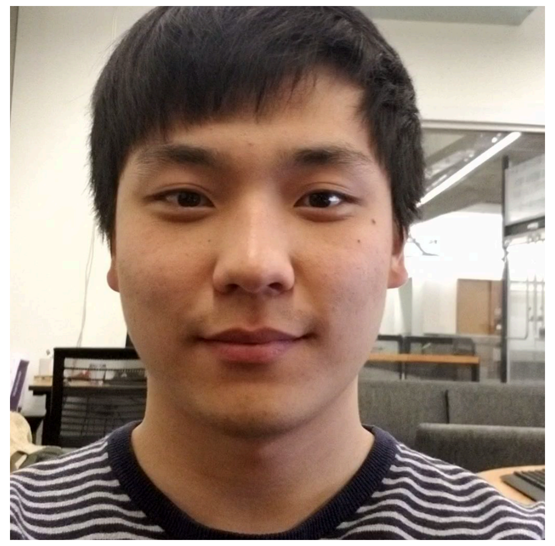
● *Grounded in events, and daily life*

MERLOT: Multimodal Neural Script Knowledge Models

arxiv 2021

Rowan Zellers* | Ximing Lu* | Jack Hessel* | Youngjae Yu | Jae Sung (James) Park | Jize Cao | Ali Farhadi | Yejin Choi


VCR

Visual Commonsense Reasoning

Previously on my slide deck...

Why is he pointing?

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

(1977)

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."

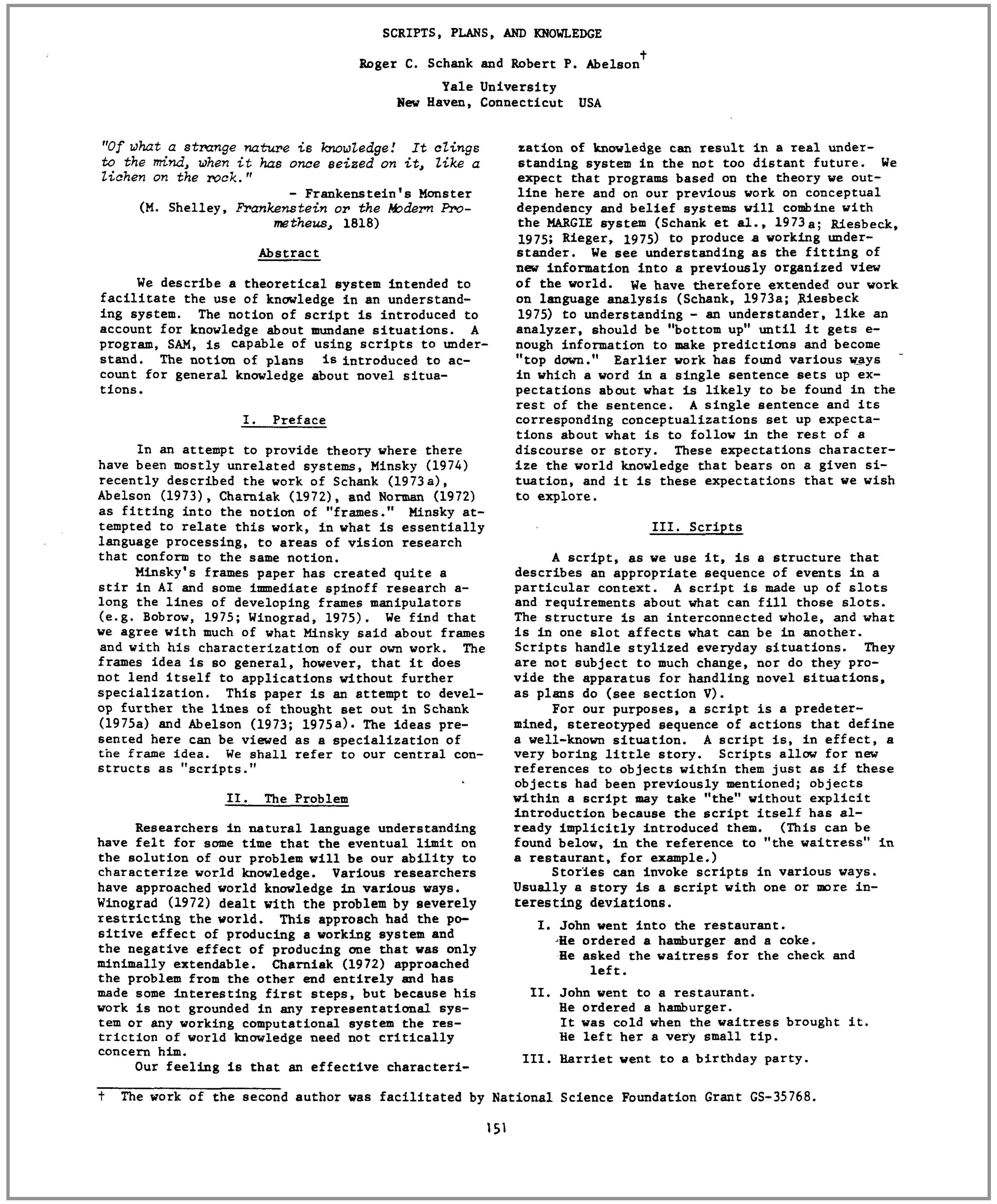
- Frankenstein's Monster
(M. Shelley, *Frankenstein or the Modern Prometheus*, 1818)

Abstract

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

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.

151

```
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

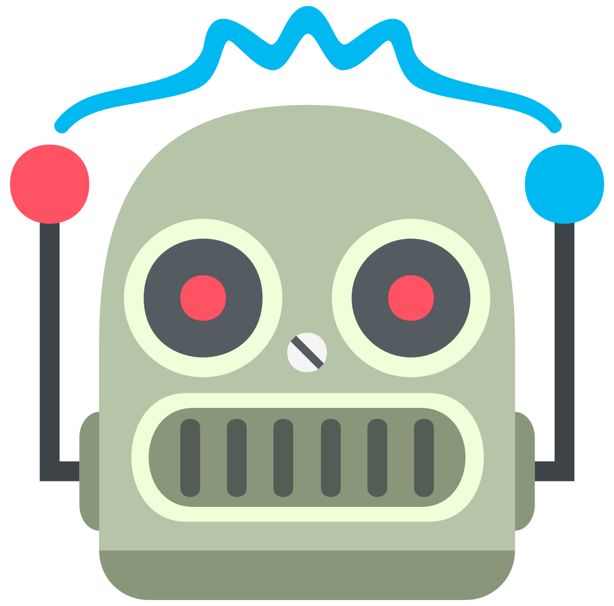
Burner

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:

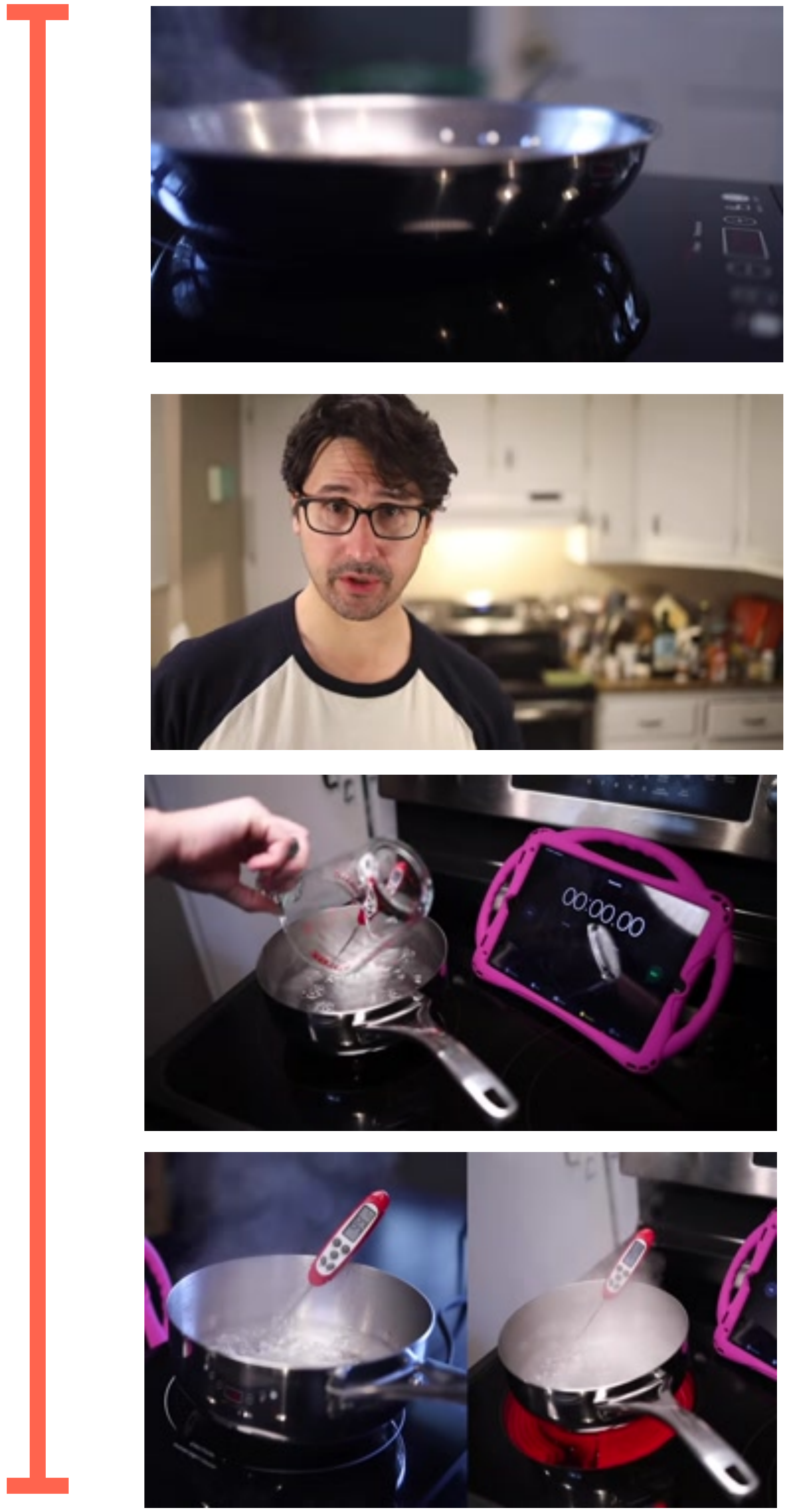
- Trained fully from scratch, we get…
- zero-shot temporal commonsense,
- Fine-tuned SOTA on 13 tasks

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

- Pretraining Strategy + Objectives

- Evaluation

Setup: Videos and Transcripts



Setup: Videos and Transcripts

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

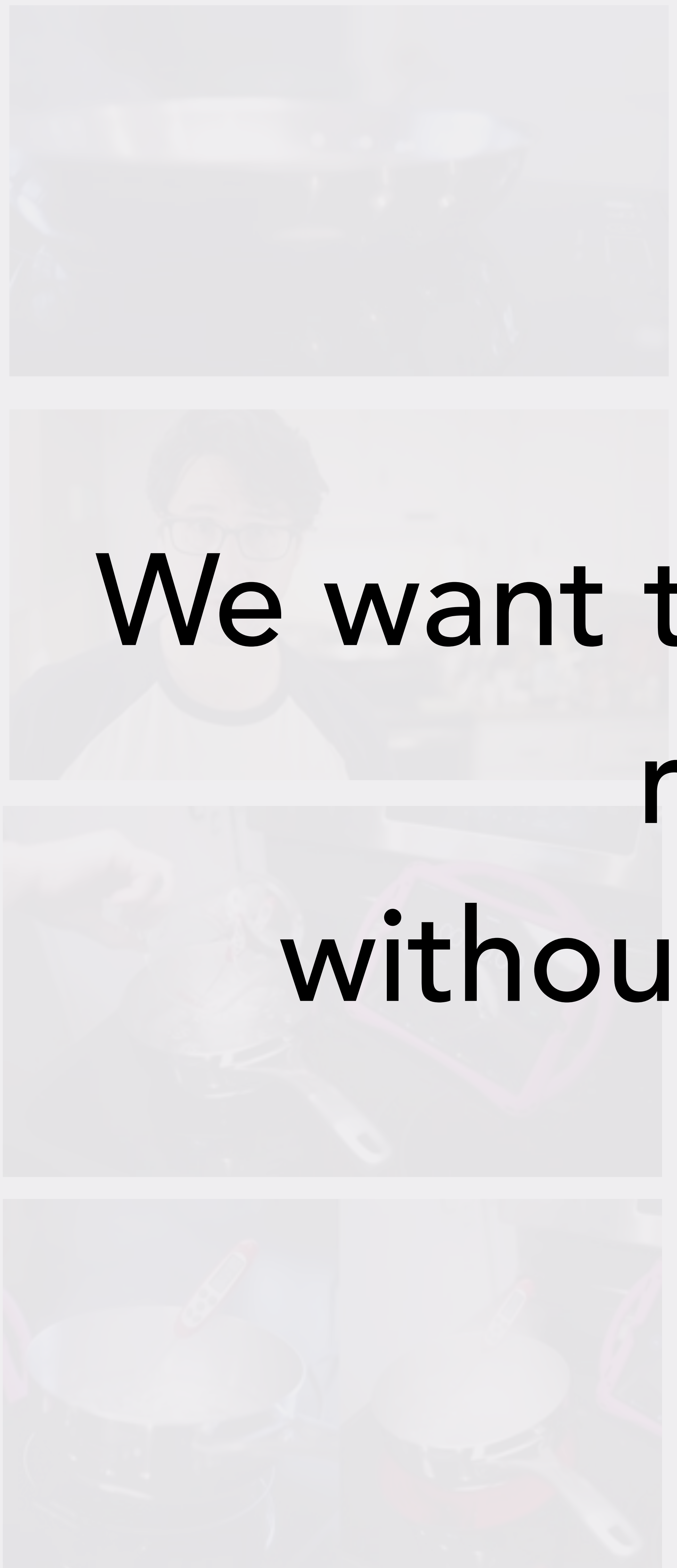
"I'll use a stopwatch to time how fast my electric stove boils water…."

"In goes the cold water…"

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

Time

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...."

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

Time

We want to use this (dynamic) data to first learn recognition-level reasoning...
without training on manually labeled data

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

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

Image Encoder

OK, but..

Text Encoder

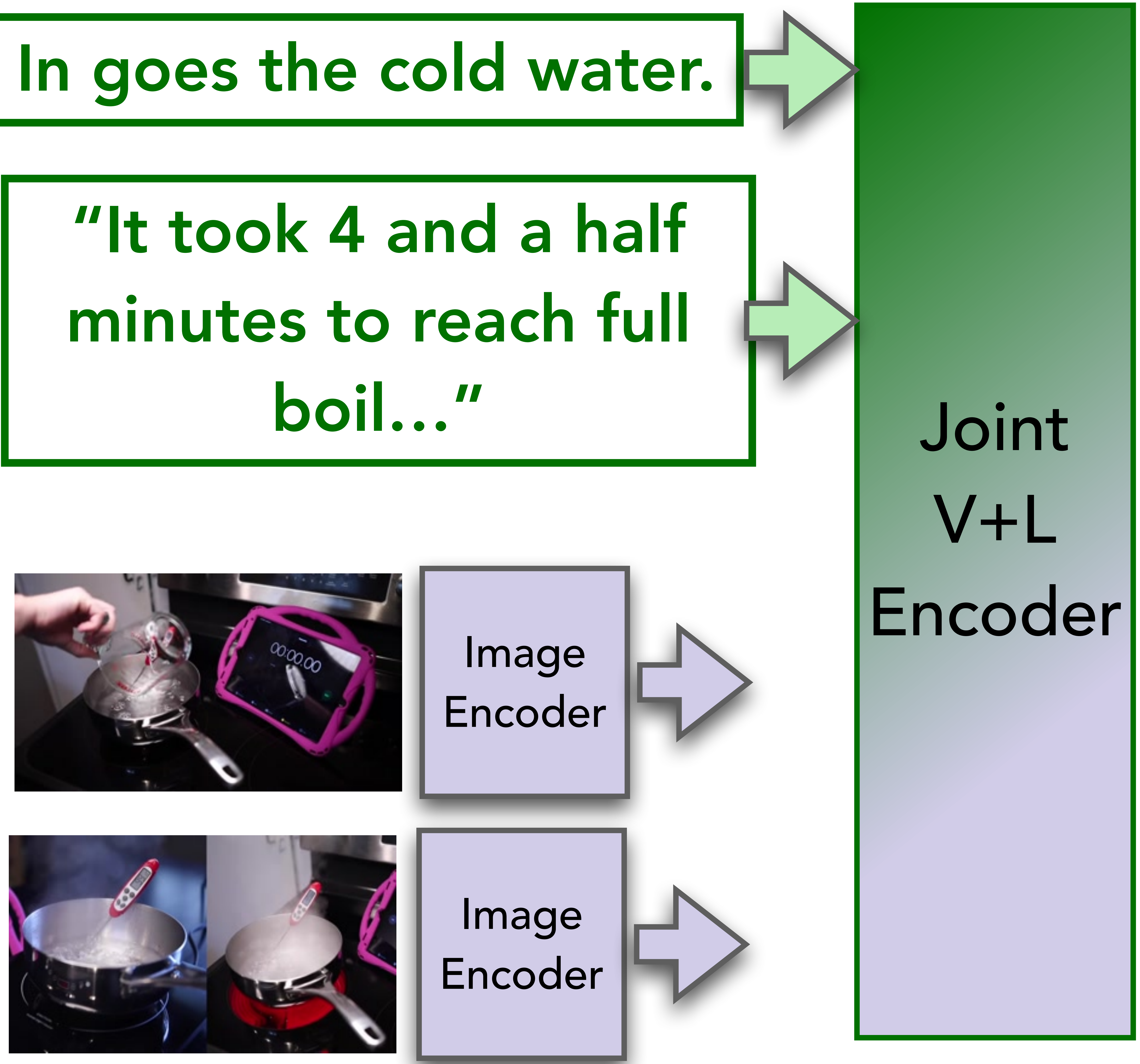
In goes the cold water.

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

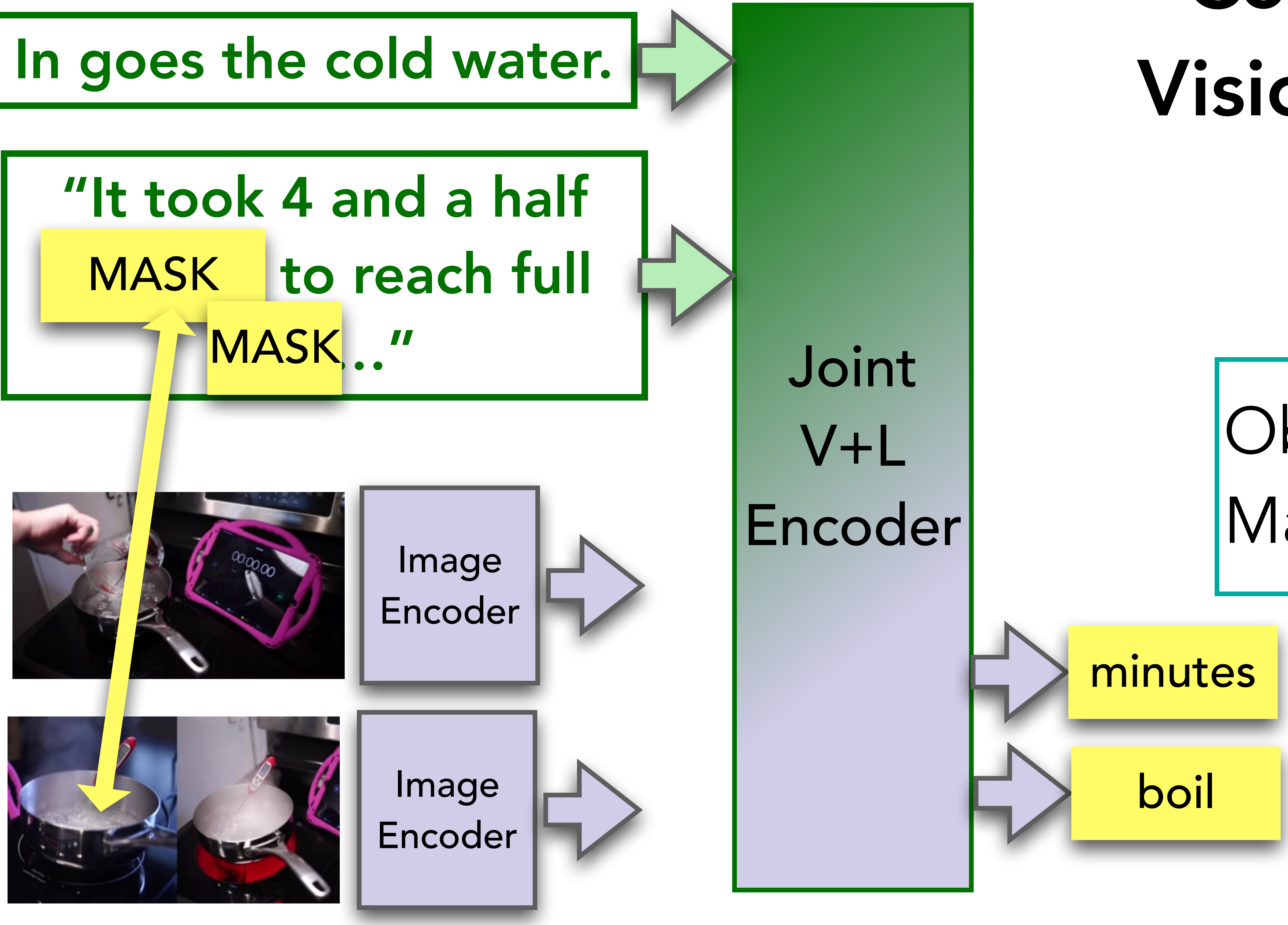
Recognition-level learning

Recognition-level learning

Commonsense Vision+Language Learning



Commonsense Vision+Language Learning



Objective 2:
Mask LM


In goes the cold water.
"It took 4 and a half
MASK
to reach full
MASK
..."
Image Encoder
Image Encoder
Joint V+L Encoder
minutes
boil
Objective 2: Mask LM
with careful selection of words for masking


In goes the cold water.
"It took 4 and a half
MASK to reach full
MASK..."
"Um, okay MASK that
took MASK minutes, so
now we'll..."
Joint
V+L
Encoder
minutes
boil
Objective 2:
Mask LM
with careful
selection of words
for masking


In goes the cold water.
"It took 4 and a half
MASK
to reach full
MASK
..."
Image Encoder
Image Encoder
Joint V+L Encoder
minutes
boil
Objective 2: Mask LM
with careful selection of words for masking

Commonsense Vision+Language 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

Commonsense Vision+Language Learning

[t=1] In goes the cold water.

[t=2] "It took 4 and a half minutes to reach full boil..."

[t=idk1]

Image Encoder

[t=idk2]

Image Encoder

Joint V+L Encoder

Objective 3: Unshuffle frames

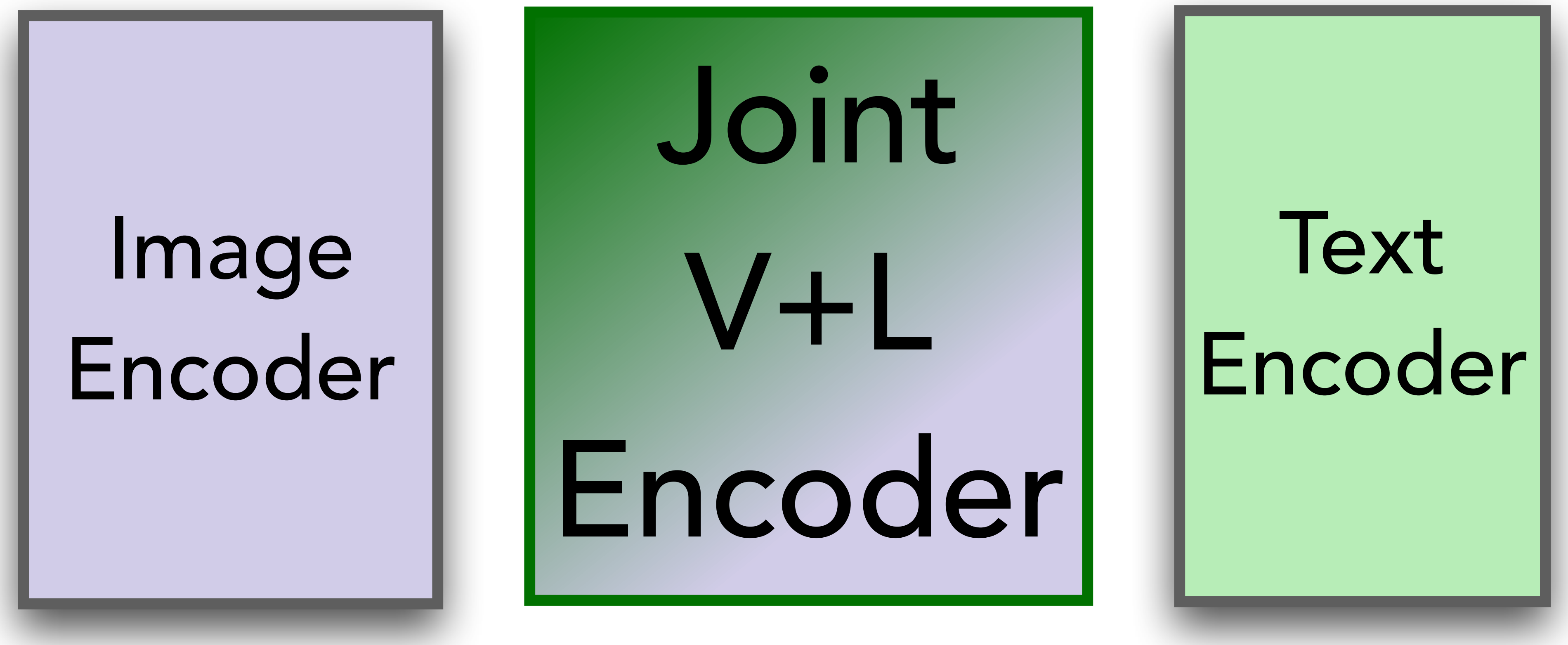
Frame "idk2" comes first

Objective1: Contextual Frame-Text Matching

Objective 2: Mask LM

Objective 3: Unshuffle frames

Using a 12-layer 'base' Transformer, train everything E2E on 6M videos



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

- Pretraining Strategy + Objectives

- 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)

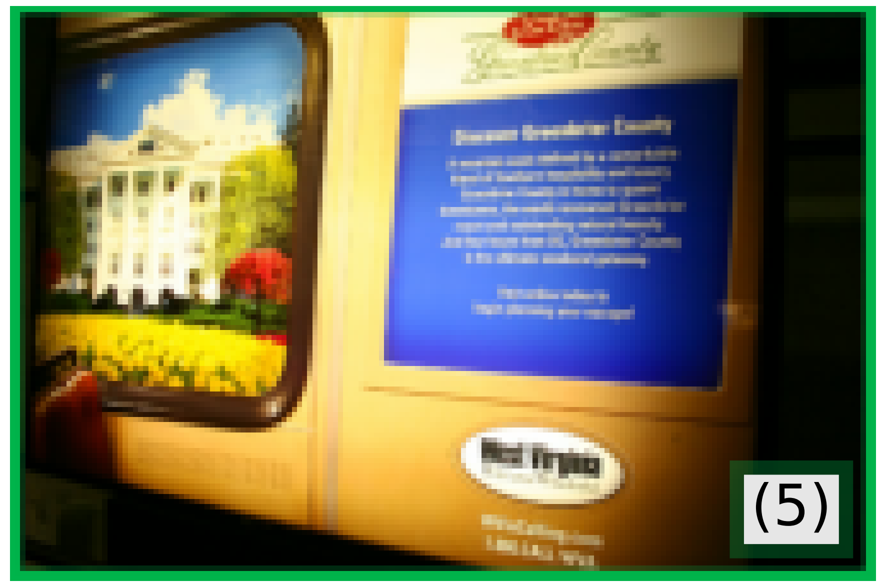
The old man was riding the escalator. → He was almost to the top. → His kids were already at the top. → At the top was a train station. → They then got on the train.

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.

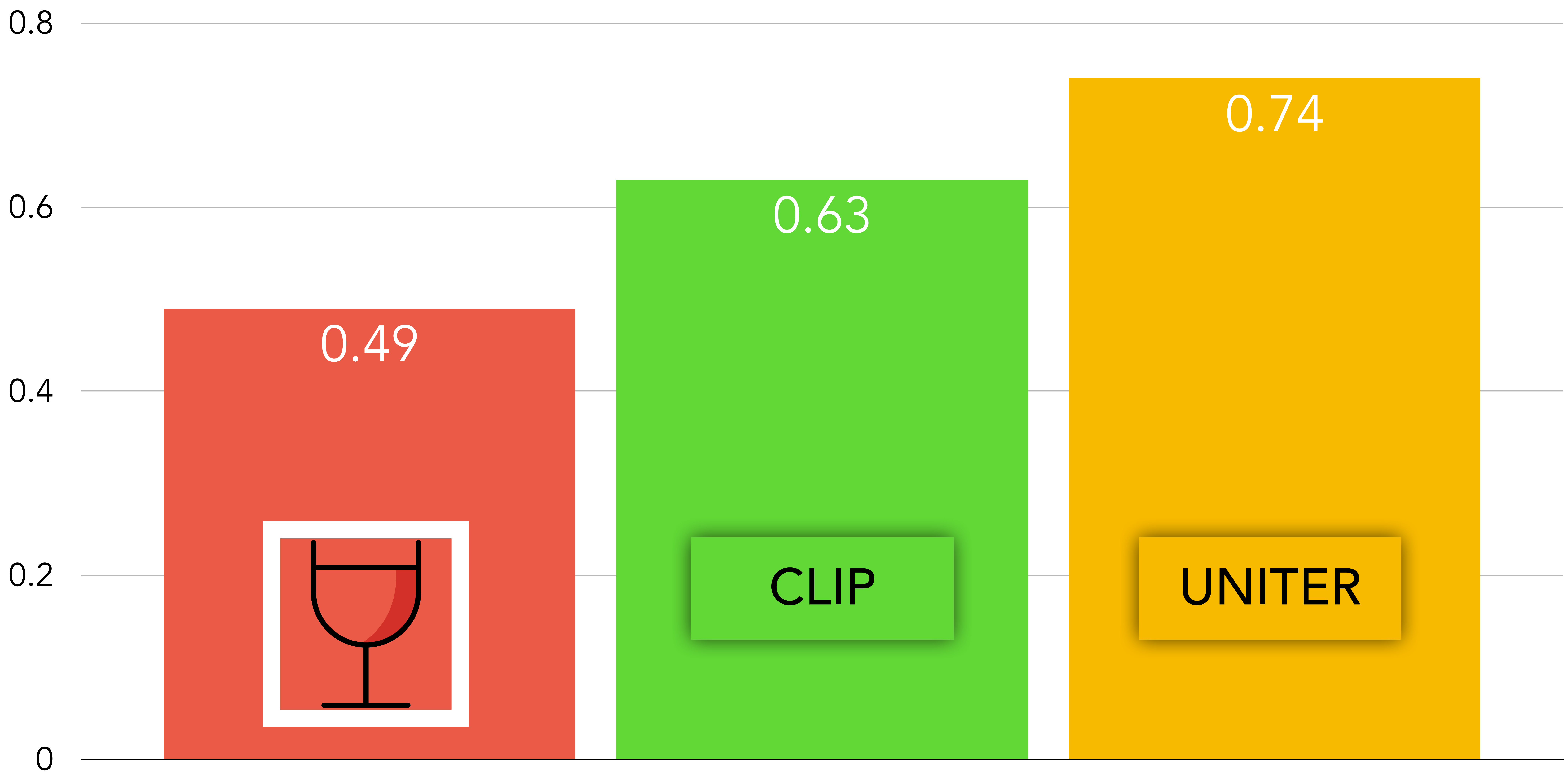
At the top was a train station.

They then got on the train.

Visual Coref over time!


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

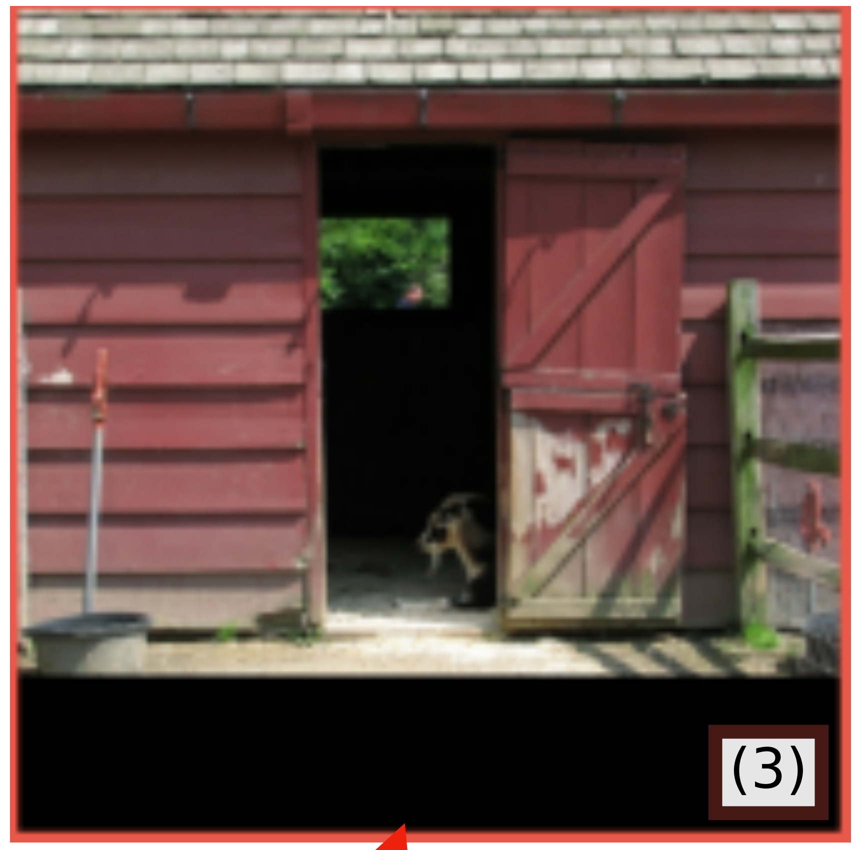
I went to the fair with my kids last weekend.

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.



MERLOT: people stay on the Merry-Go-Round for a while

Evaluation 2: Fine-tuned Video QA

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

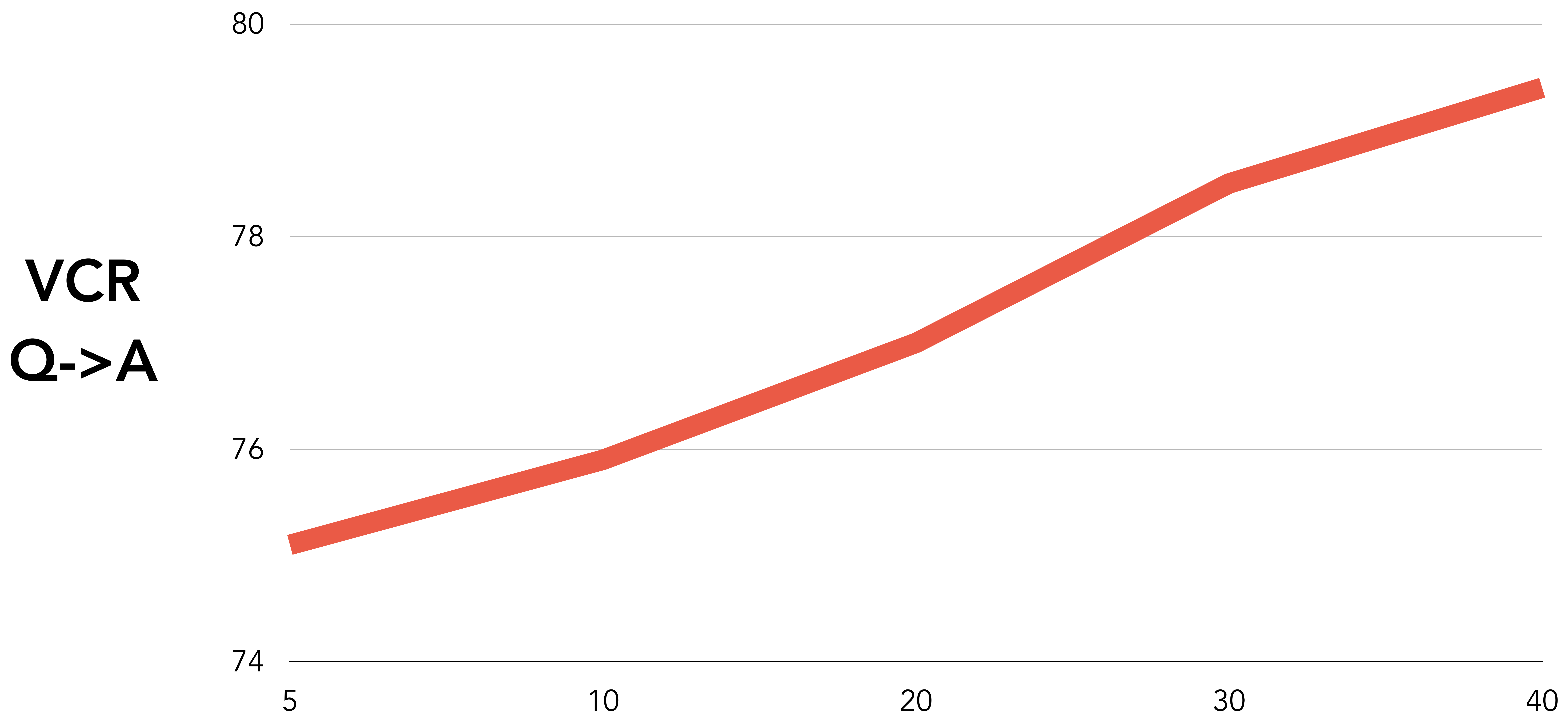
Despite no supervised object detector, and never seeing still images before

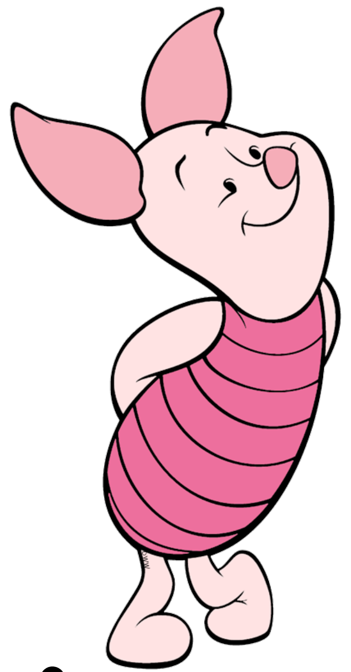
Analysis (on TVQA+)

Analysis (on TVQA+)

Analysis (on TVQA+)

Performance increases with # epochs





Discussion

- Simulation pros:
  - Learning to *act*, not just see/write
  - Future work: Models guiding the training loop, maybe based on curiosity
- Cons:
  - Limited vocabulary in simulation
  - Hard to learn human behavior

- Web video pros:
  - Super wide vocabulary
  - Learning human norms, behavior, events
- Cons:
  - Can't participate in the video
  - Privacy

Privacy (and other negative societal implications of training on multimodal Web Data)

- Things we did for MERLOT
    - data curation focused on big channels, not randos
        - on a public platform that people expect is public (Kang et al 2015)
        - … at a scale so that people are “in public without being public” (Marwick and boyd 2011)
    - distributing links, not the videos, for the “right to be forgotten”
    - Encouraging future work into these foundation models — not advocating for product use right now

Privacy (and other negative societal implications of training on multimodal Web Data)

- data curation focused on big channels, not randos
  - on a public platform that people expect is public (Kang et al 2015)
  - … at a scale so that people are "in public without being public" (Marwick and boyd 2011)
- distributing links, not the videos, for the "right to be forgotten"
- Encouraging future work into these foundation models — not advocating for product use right now

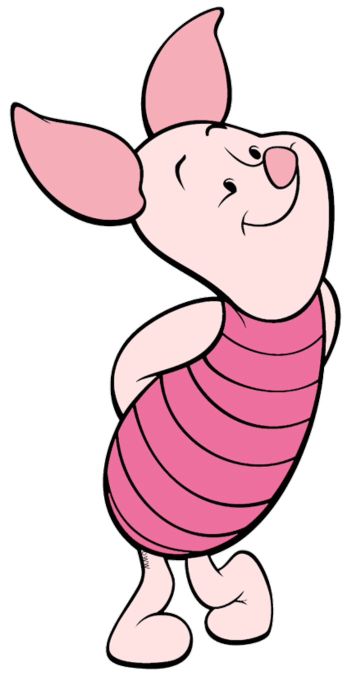
Lots of local news… which has bias issues (Gilliam Jr et al 1996)

Inherent bias with training on data that encodes a "view from nowhere" (Haraway et al 1988, Waseem et al 2021)

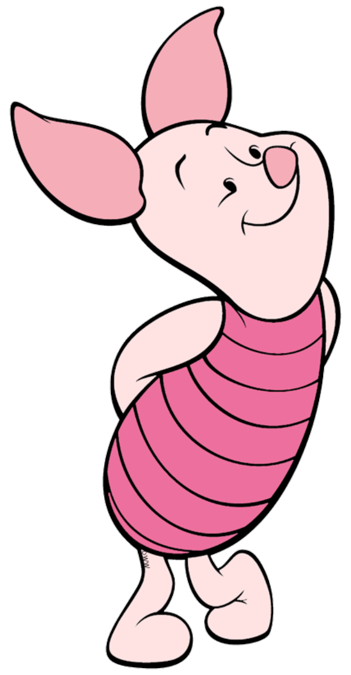
… bias that is amplified by culture and the "YouTube Algorithm" (Strangelove et al 2020)

Privacy (and other negative societal implications of training on multimodal Web Data)

- Future work: studying privacy, bias, and dual use,
- … exploring possibly a mix of technical and non-technical fixes here
- Hopefully the beginning, not the end, of this key conversation



Questions?



Thanks!!