Carnegie Mellon University
Language Technologies Institute

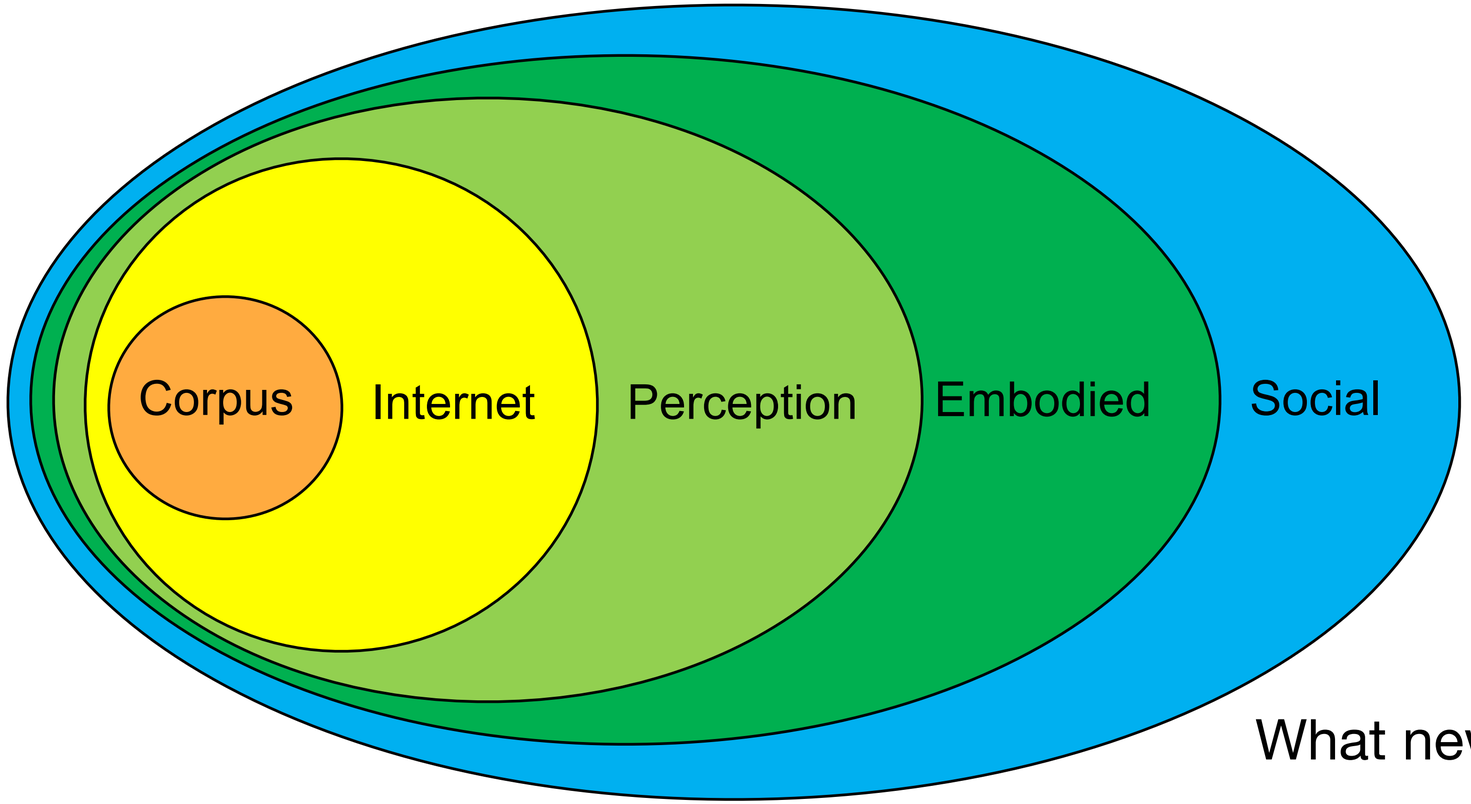
Language Should be Embodied

Wait, huh, what? Please explain…

Yonatan Bisk @ Stanford NLP

Nov 19, 2020

World Scopes



How does reporting bias change?

What new knowledge do models have access to?

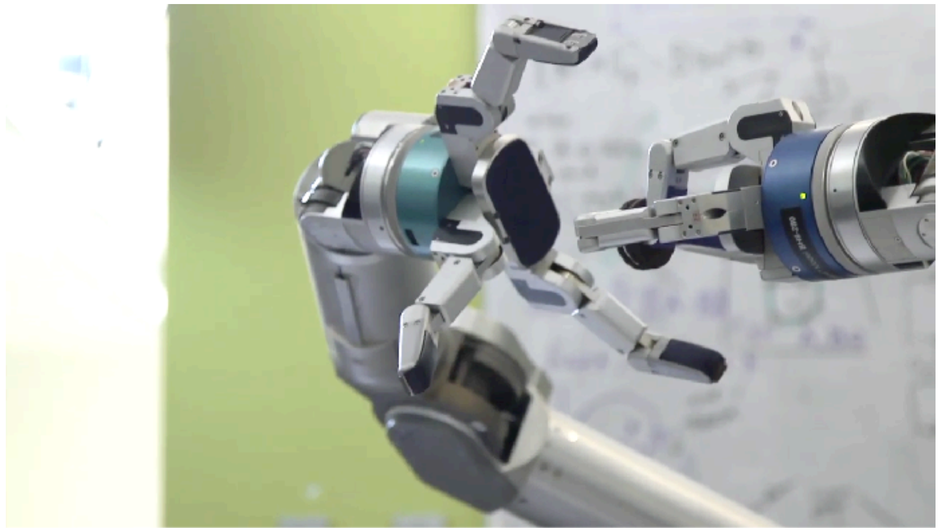
What advances in modeling and fusion are necessary?

Yonatan Bisk, Ari Holtzman, Jesse Thomason
Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, Joseph Turian
EMNLP 2020

Why?

Language that affects the world

Help! I like the cream separated from the oreo



HERB (Siddhartha Srinivasa)

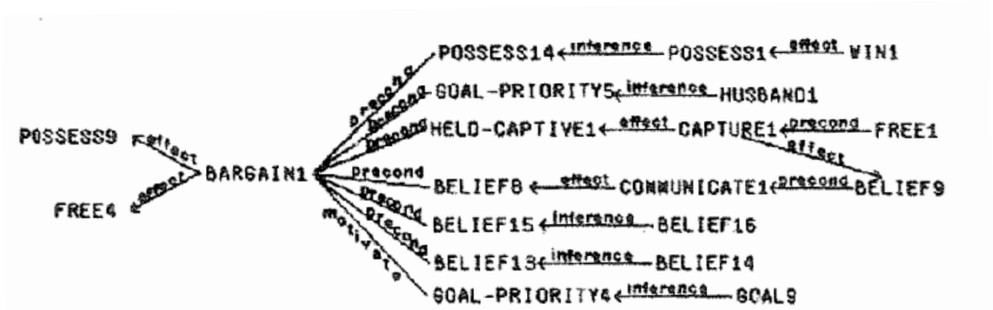
Access to Broader Semantics

What's it like to drive a bus?

How many hours of watching to achieve same level of performance as 30m of practice?

https://www.youtube.com/watch?v=SqGdH8lWvbA

Hierarchy, Abstraction, & Scripts



POSSESS9	Bob has $75,000.
FREE4	Alice is free.
BARGAIN1	Bob makes a bargain with Ted in which Bob releases Alice and Ted gives $75,000 to Bob.
POSSESS14	Ted has $75,000.
POSSESS1	Ted has $100,000.
WIN1	Ted wins $100,000 in the lottery.
GOAL-PRIORITY5	Ted wants Alice free more than he wants to have $75,000.
HUSBAND1	Ted is Alice's husband.
HELD-CAPTIVE1	Bob is holding Alice captive.
CAPTURE1	Bob captures Alice.
FREE1	Alice is free.
BELIEF8	Ted believes Bob is holding Alice captive.
COMMUNICATE1	Bob contacts Ted and tells him that he is holding Alice captive.
BELIEF9	Bob believes he is holding Alice captive.
BELIEF15	Bob believes Ted has $75,000.
BELIEF16	Bob believes Ted has $100,000.
BELIEF13	Bob believes Ted wants Alice to be free more than he wants to have $75,000.
BELIEF14	Bob believes Ted is Alice's husband.
GOAL-PRIORITY4	Bob wants to have $75,000 more than he wants to hold Alice captive.
GOAL9	Bob wants to have $75,000.

Infinite variations

1. How to generalize
2. How to instantiate

Mooney and DeJong — *Learning Schemata for Natural Language Processing* — IJCAI 1985

Hierarchy, Abstraction, & Scripts

“Strain the pasta”

Put the **strainer** in the **sink**

Once the **pot** with the pasta is cool enough, grab it by the **handles**

Pour the pasta and water into the **strainer** in the **sink**

…

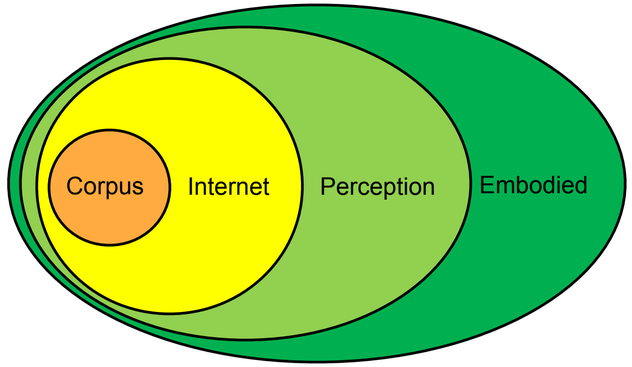
$\vec{\theta}_i$

$\vec{\theta}_j$

$\vec{\theta}_k$

Bisk et al — *Benchmarking Hierarchical Script Knowledge* — NAACL 2019
Multimodal: Fried et al — Learning to Segment Actions from Observation and Narration — ACL 2020

Lexical Semantics



“Strain” = *Burden, stress, colander, …*

=

= $\vec{\theta}_j$

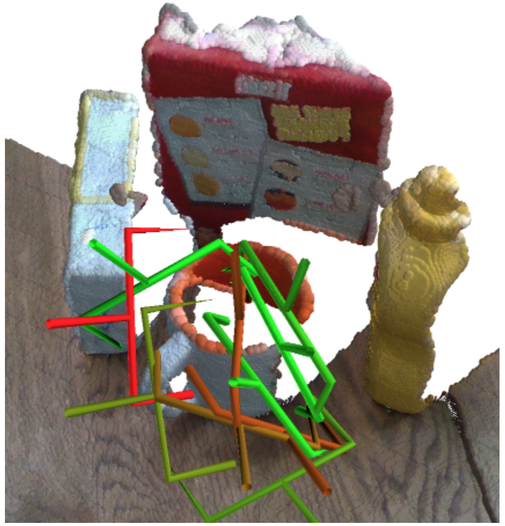
Pick-up

What's hidden in that?

Does "pick up" mean the same thing for all of these?

Does "pick up" correspond to a specific action sequence?

Pick-up isn't an action, it's a post-condition



Mousavian et al. 6-DOF GraspNet: Variational Grasp Generation for Object Manipulation — ICCV 2019

Why is embodiment hard?

Action Space Semantics

- Predicates
- Pixels
- End-effectors
- Torques
- …

Implicit Plans

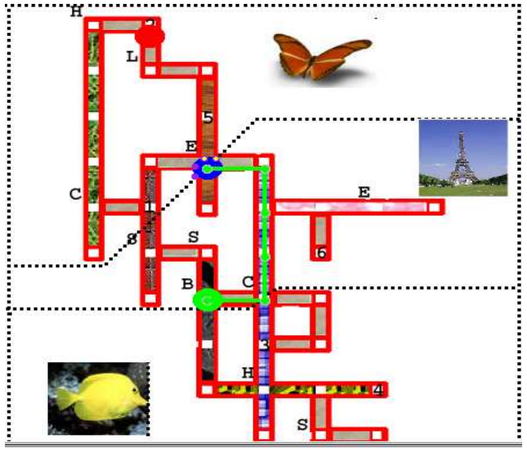
- Scripts
- Pre-conditions
- Post-conditions
- Success
- …

🚨Bold unsubstantiated claim that I can't deliver on🚨

None of these should be symbolic

The Gap

Navigation in simulation



MacMahon, Stankiewicz, and Kuipers AAAI 2006

Chen and Mooney AAAI 2011

+Visual Complexity

+Language Complexity

Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

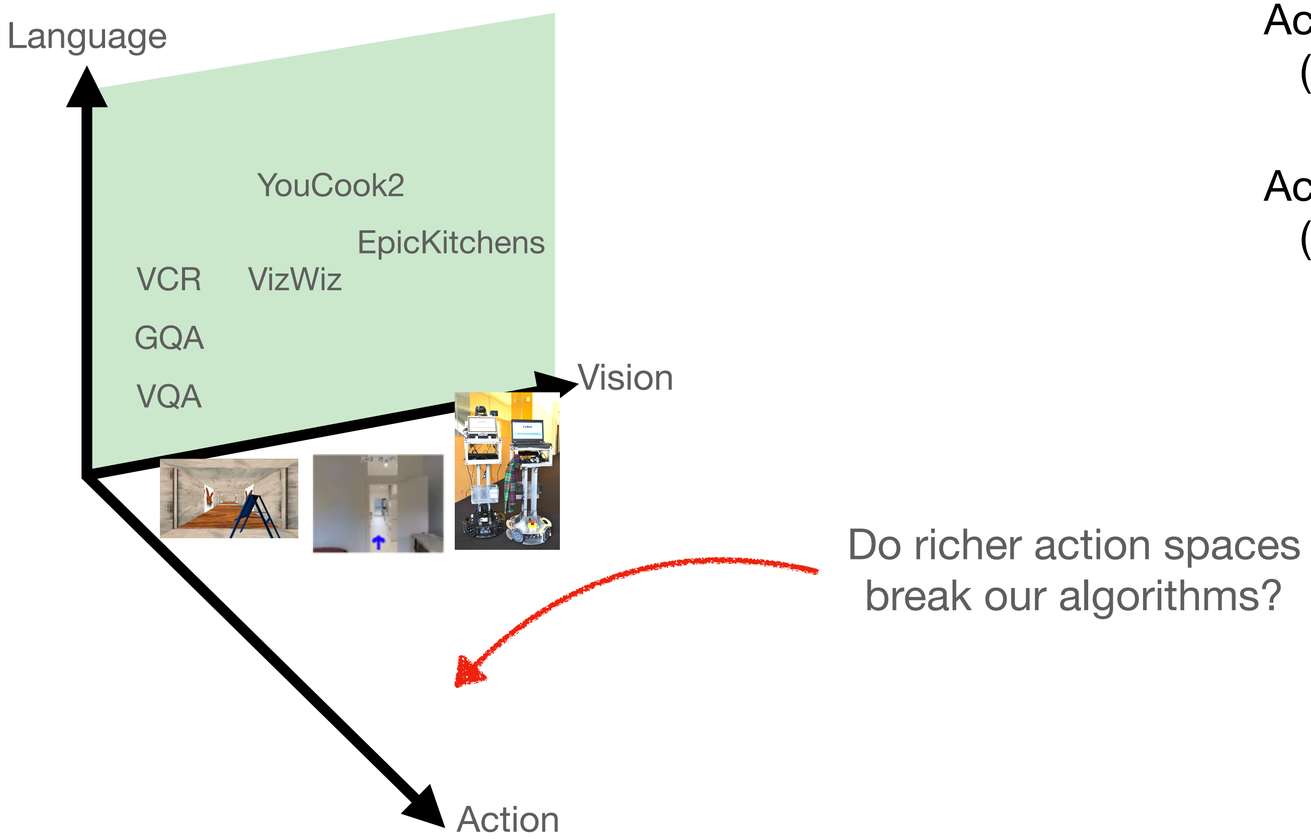
Anderson et al. CVPR 2018

- Simplified Action space
- Static Environment

Navigation in the world

Rosenthal and Veloso AAAI 2012

Grounding Spectrum



Action sequence length
(e.g. Memory, Alignment, pre-/post-conditions)

Action output space
(e.g. end-effector position? Quaternions? Torques?)

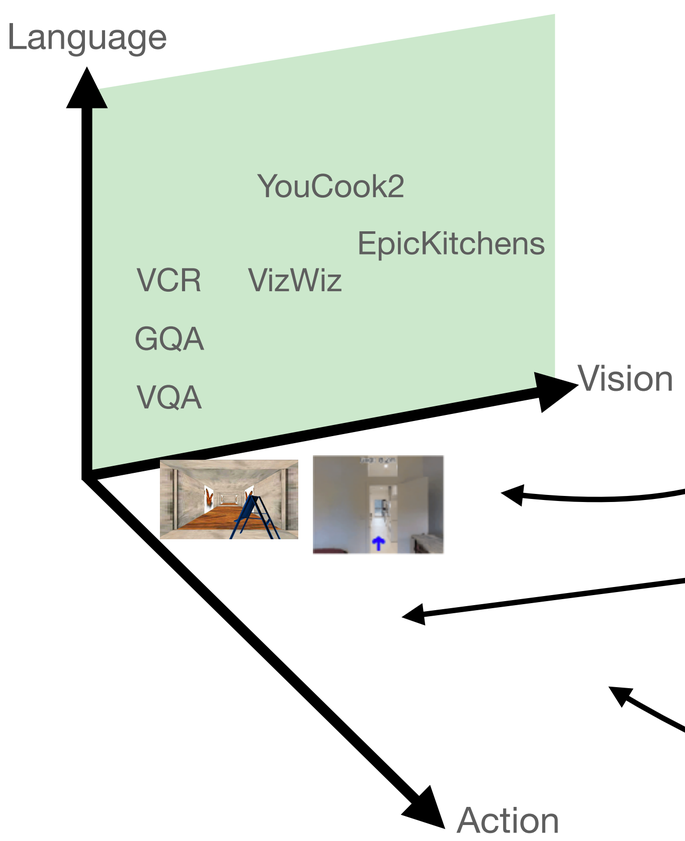
Do these change or inform language?

Grounding Spectrum

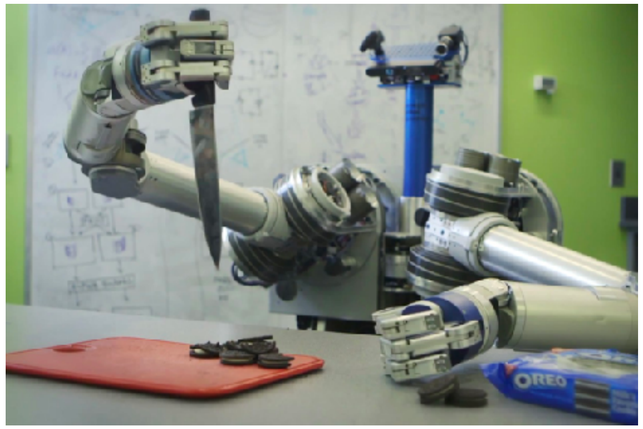
Action sequence length
(e.g. Memory, Alignment, pre-/post-conditions)

Action output space
(e.g. end-effector position? Quaternions? Torques?)

Do these change or inform language?



Sharma, Zhang, Kroemer 2019



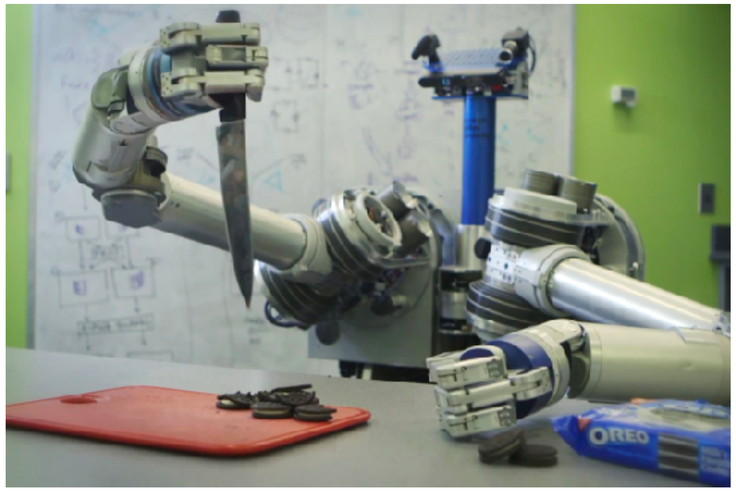
HERB

The Gap

+ Longer horizon plans
+ State changes
+ Language Complexity
+ Underspecified Language

• • • •▶

Manipulation in the world



HERB

- Masks for object interaction
- Discrete actions (no torques)

ALFRED

A Benchmark for Interpreting Grounded Instructions for Everyday Tasks

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk,
Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, Dieter Fox
CVPR 2020

https://askforalfred.com/

ALFRED

Action Learning From Realistic Environments and Directives

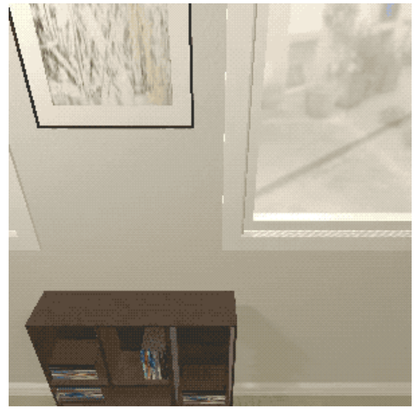
Seven High-level Tasks

Paths are generated by planner

Pick & Place

Double Place

Stack



Examine

Heat

Cool

Rinse

Data collection

Tuple

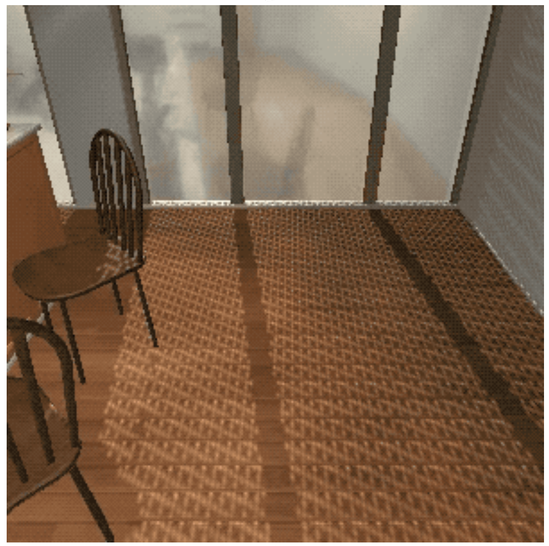
(Stack, Fork, Cup, CounterTop, Kitchen3)

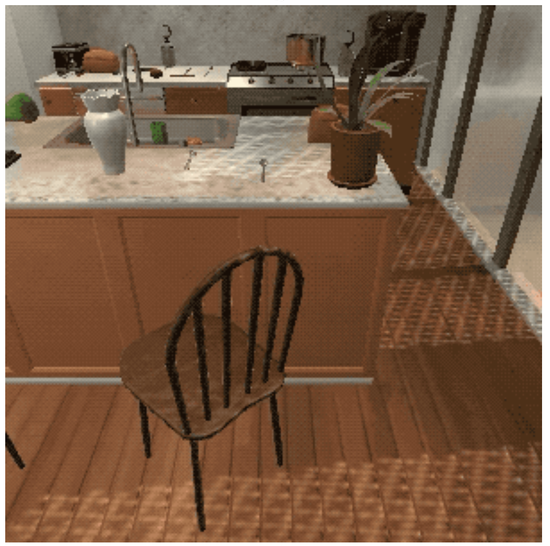
Planner

(x,y,z) | is_fork(x) ^ is_cup(y) ^ on(x, y) ^ is_counter(z) ^ on(y, z)

Sample

Execute





Annotate

Language
Language
Language

Language
Language
Language

Language
Language
Language

Promises

+ Longer horizon plans
+ Language Complexity
+ State changes
+ Underspecified Language
• Masks for object interaction
- Discrete actions (no torques)

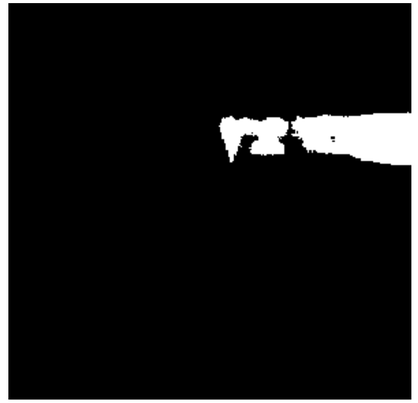
Place a hot bread slice on the counter

Promises

Wash the cup

“semantics” semantics

Put In



Toggle

+ Longer horizon plans
+ Language Complexity
+ State changes
+ Underspecified Language
- Masks for object interaction
- Discrete actions (no torques)

Hierarchy, Abstraction, & Scripts

"Place a heated apple slice on the large table"

create_slice(apple) heat(apple_slice) place(apple_slice, table)

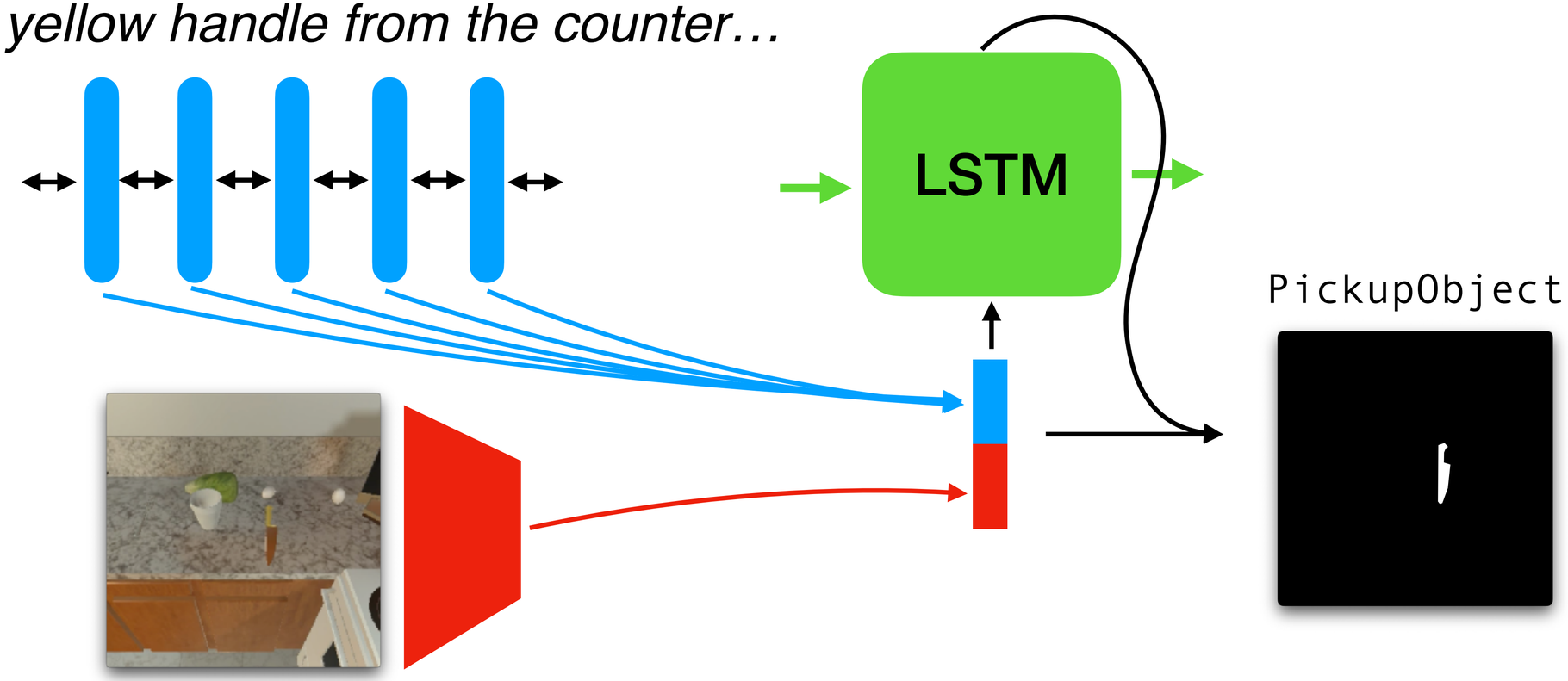
collect(knife) locate(apple) slice(apple)



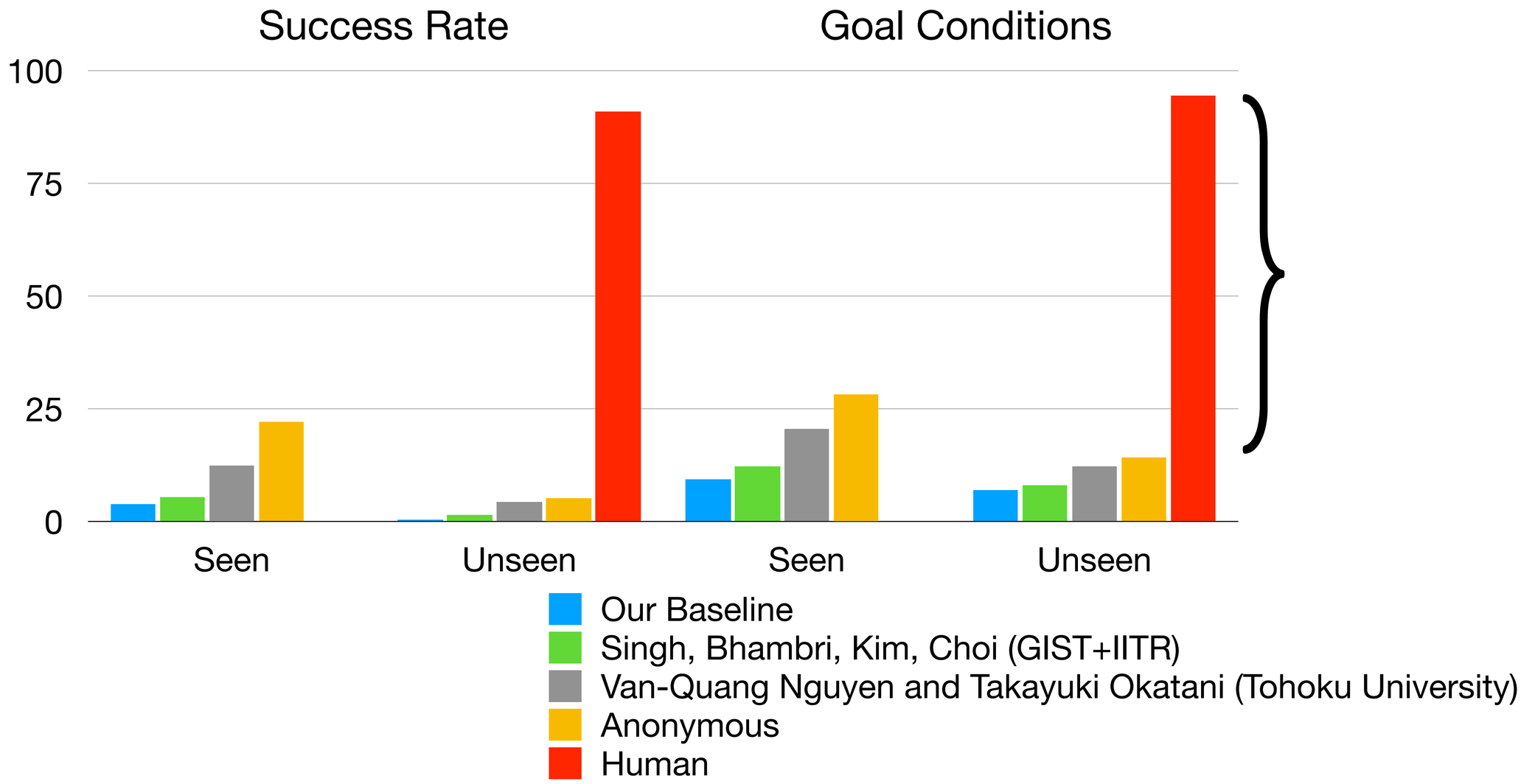

...

Deep Learning, End-to-End, Tabula Rasa, “magic”

Turn around and move to the stove, then turn left to face the counter to the left of the stove. Pick up the sharp knife with the yellow handle from the counter…



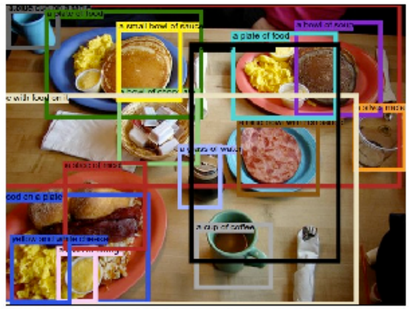
It’s hard


Success Rate
Goal Conditions
100
75
50
25
0
Seen
Unseen
Seen
Unseen
Our Baseline
Singh, Bhambri, Kim, Choi (GIST+IITR)
Van-Quang Nguyen and Takayuki Okatani (Tohoku University)
Anonymous
Human

Why is it hard?

Grounding Language + Masks

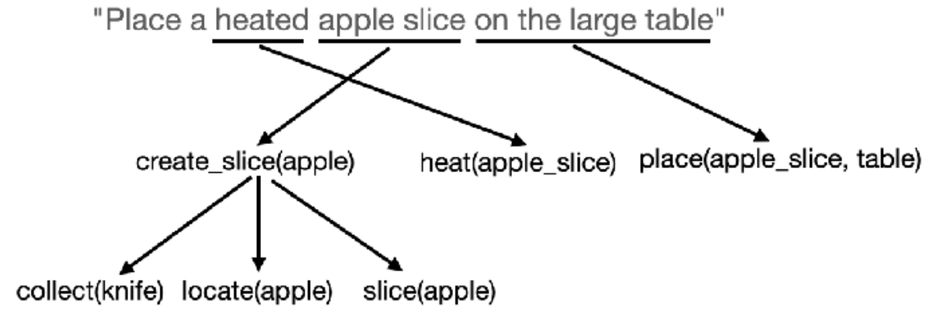
Dense Captioning



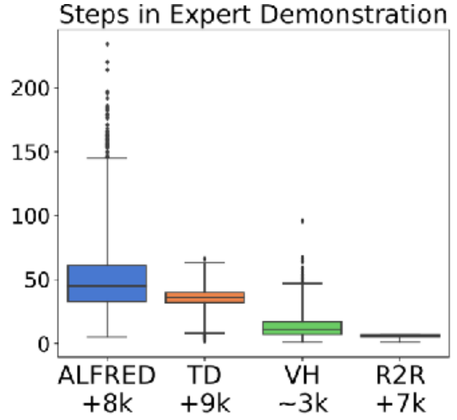
a plate of food. food on a plate. a blue cup on a table. a plate of food. a blue bowl with red sauce. a bowl of soup. a cup of coffee. a bowl of chocolate. a glass of water. a plate of food. a silver metal container. a small bowl of sauce. table with food on it. a slice of orange. a table with food on it. a slice of meat. yellow and white cheese.

Johnson et. al

Scripts / Program Induction?



Long Trajectory State Tracking?



… Everything?

ALFWorld

Aligning Text and Embodied Environments for Interactive Learning

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, Matthew Hausknecht

https://alfworld.github.io/

Procedure Learning via Exploration

Instruct

Explore (RL/IL/…)

Reward

"Place a heated apple slice on the large table"

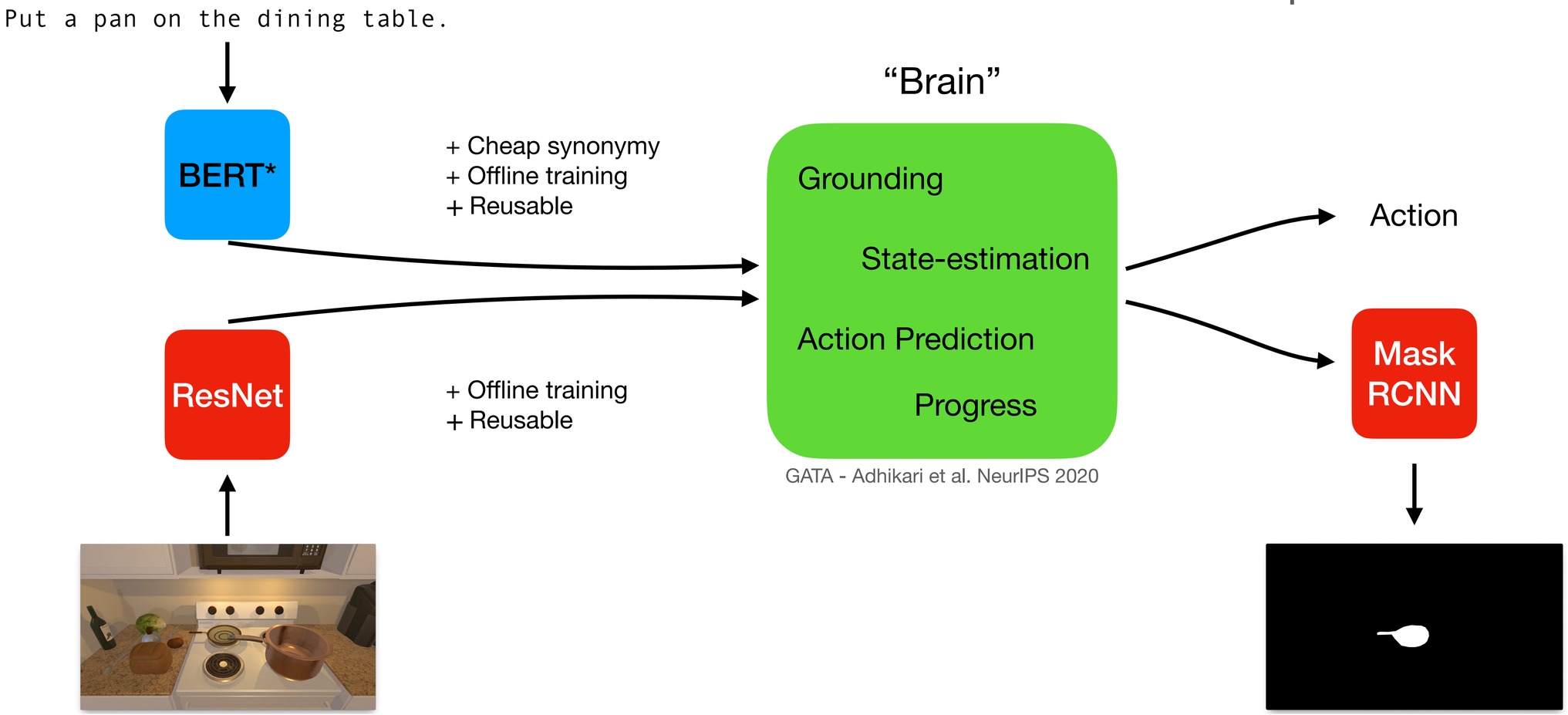
No 😔

🤫 You might be smarter than us though

$$\mathcal{R} \in \{-1, 1\}$$

Cheap Exploration with Generalization

Can the brain explore on its own?



GATA - Adhikari et al. NeurIPS 2020

*BERT lexical embeddings

Separating Exploration from Action

Welcome!
You are in the middle of the room. Looking around you, you see a dining table, a stove, a microwave, and a cabinet.

Your task is to:
Put a pan on the dining table.

> goto the cabinet

You arrive at the cabinet.
The cabinet is closed.

> open the cabinet

The cabinet is empty.

> goto the stove

You arrive at the stove. Near the stove, you see a pan, a pot, a bread loaf, a lettuce, and a wine bottle.

> take the pan from the stove

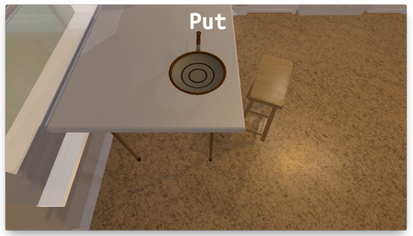
You take the pan from the stove.

> goto the dining table

You arrive at the dining table

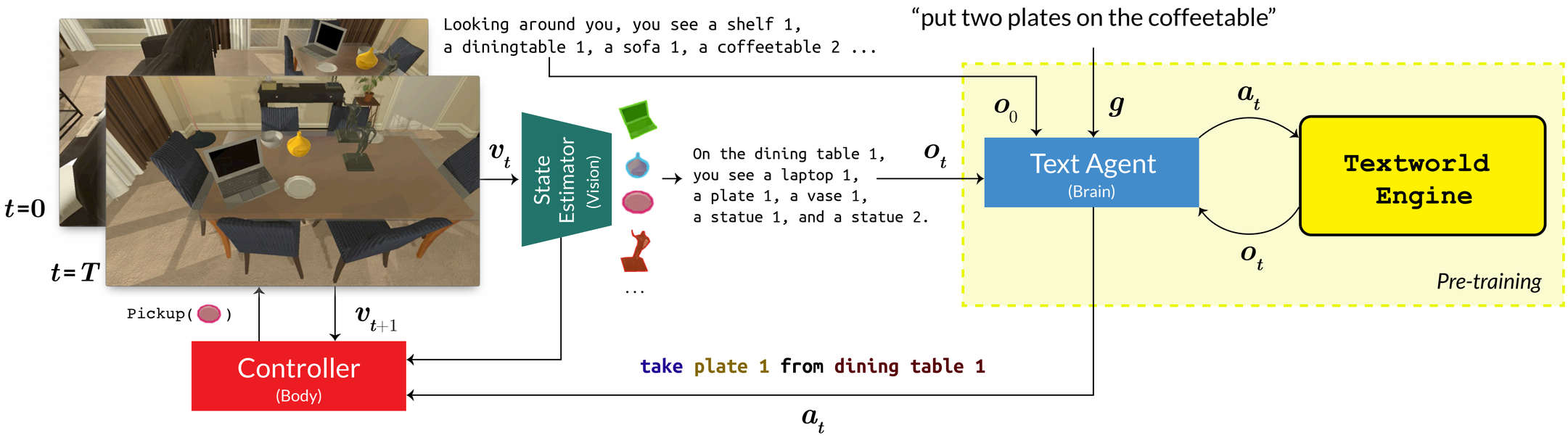
> put the pan on the dining table

You put the pan on the dining table.



AI2THOR
An Interactive Embodied AI Platform
AI2 Allen Institute for AI

Procedural Pre-training



BUTLER::Brain (Text Agent)

$$o_0, o_t, g \rightarrow a_t$$

BUTLER::Vision (State Estimation)

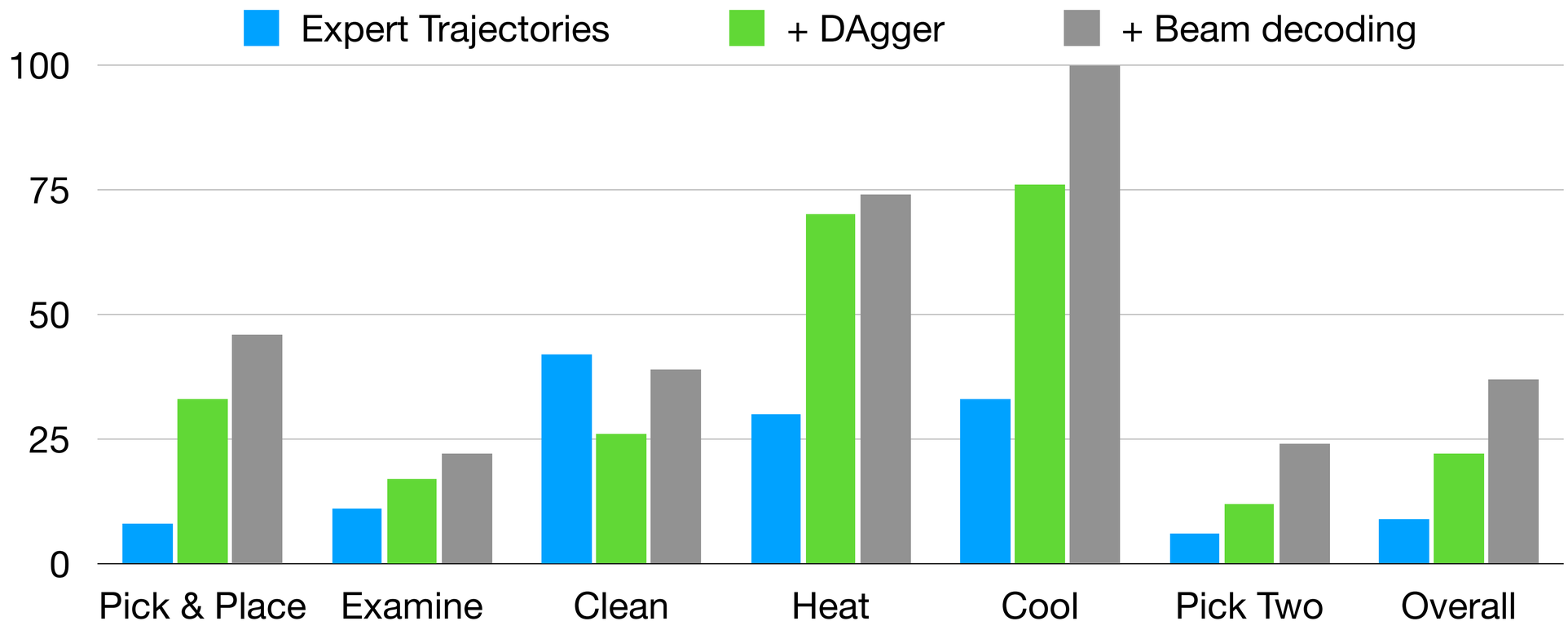
$$v_t \rightarrow o_t$$

BUTLER::Body (Controller)

$$v_t, a_t \rightarrow \{\hat{a}_1, \hat{a}_2, ..., \hat{a}_L\}$$

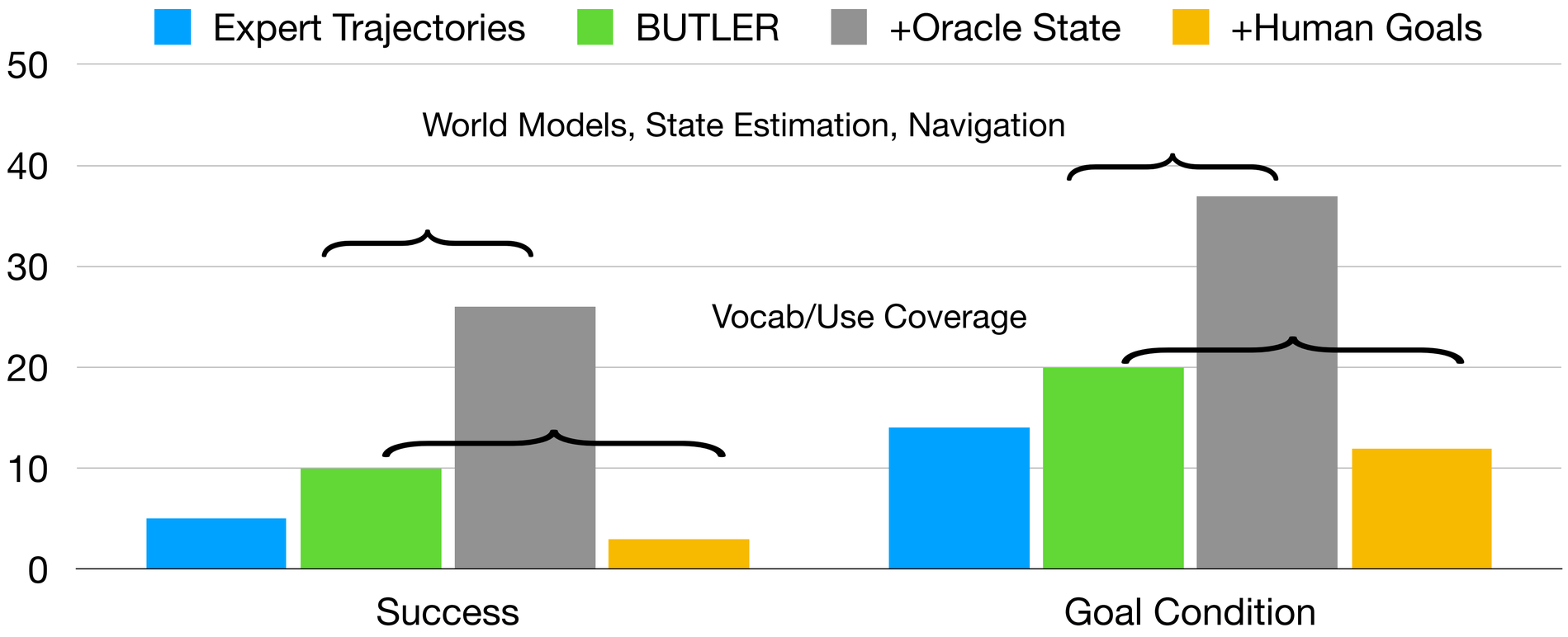
Environment Generalization within TextWorld

Unseen Environments



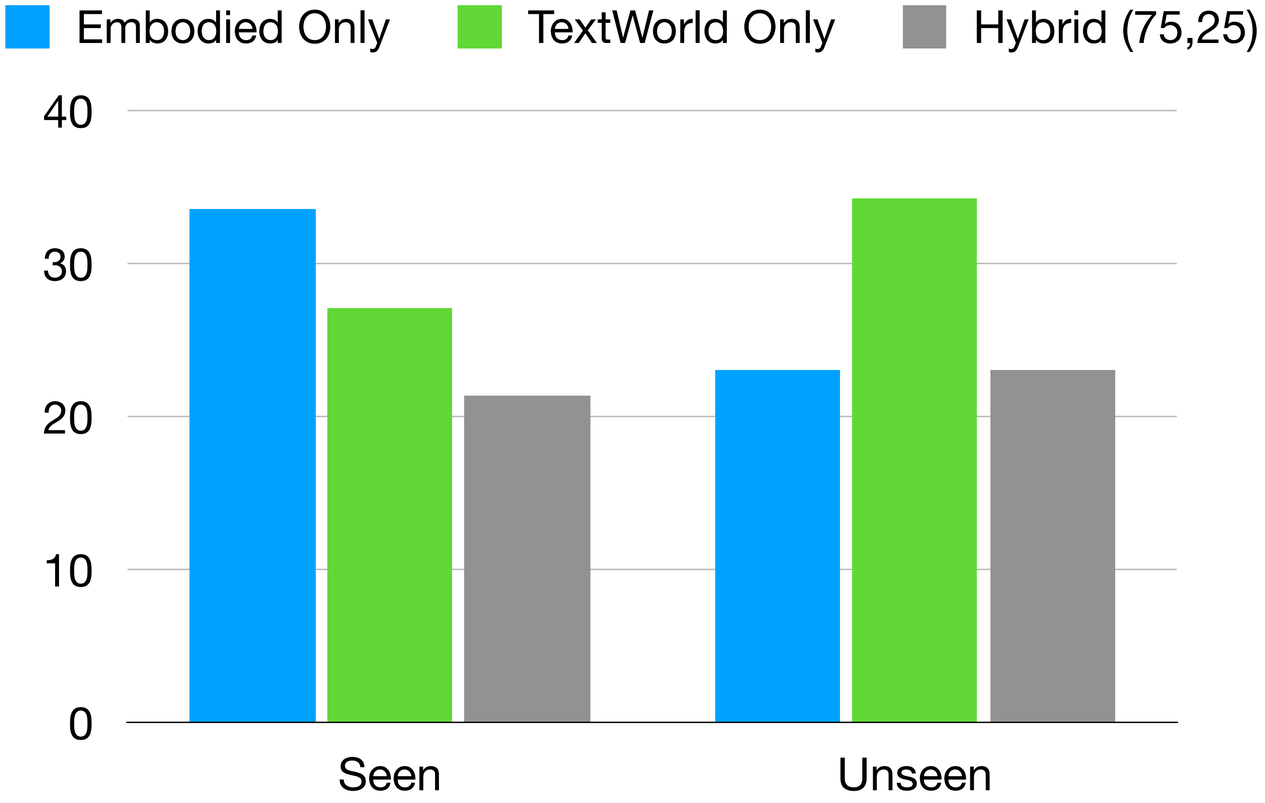
From Text to Embodied

Unseen Environments

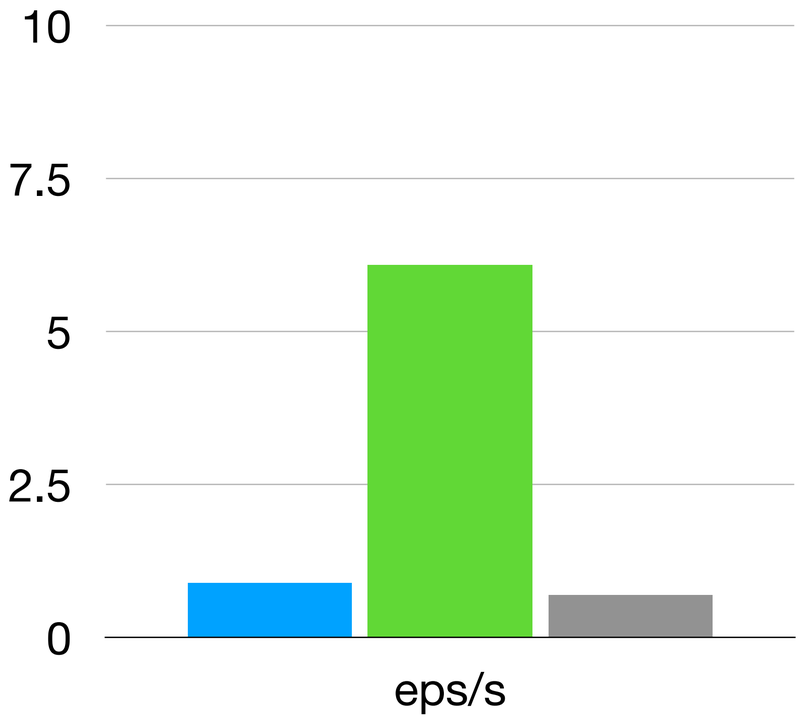


Is it worth it?

Should we train directly in the embodied world? [oracle detection + teleportation]



Fewer distractions?



The Gap(s)

Isolating Complexity

Goal Progress
Procedures
Memory
…

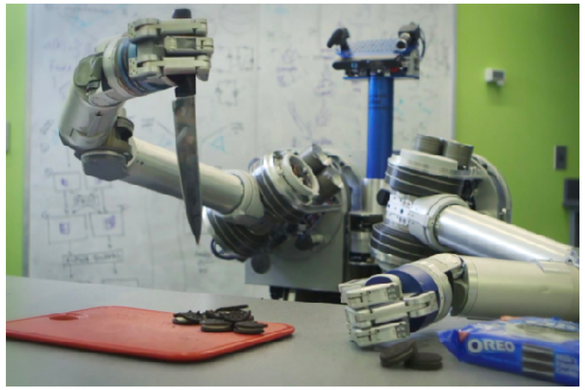
+ Physics
X can't fit in Y

+ Visual Synonymy
Apple? Tomato?

+ Interaction Masks

+ Navigation

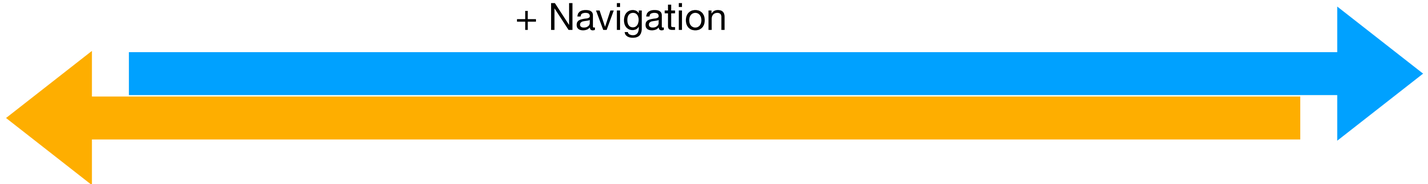
The goal



+ Sensor Noise

+ Manipulation

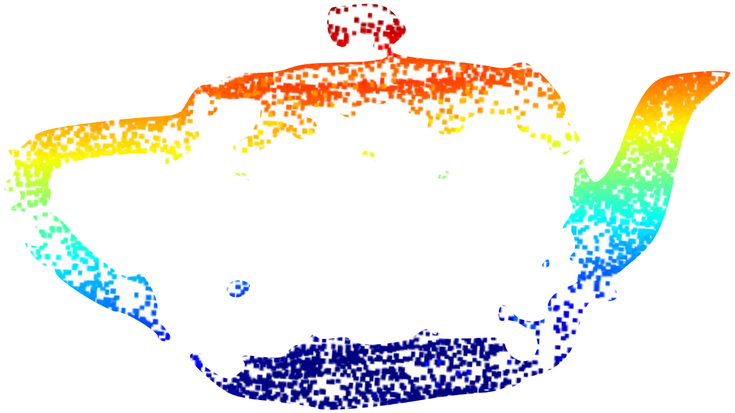
+ Motors



Ideally … WIP…

Tea Kettle

~



~

The Explainer

Automatic Instruction Generation for Egocentric Skill Learning

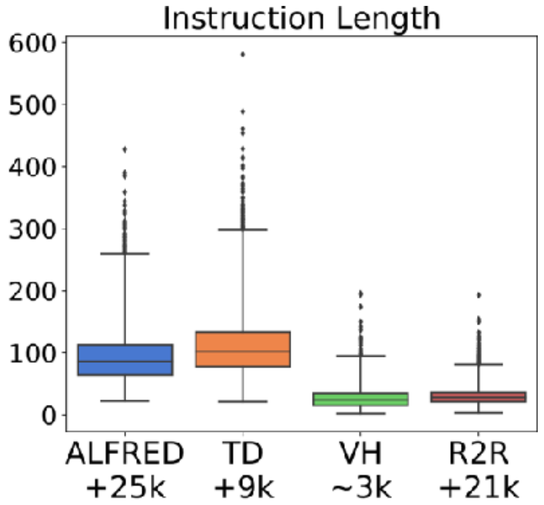
Legg Yeung, Yonatan Bisk, Alex Polozov

What about NLG?

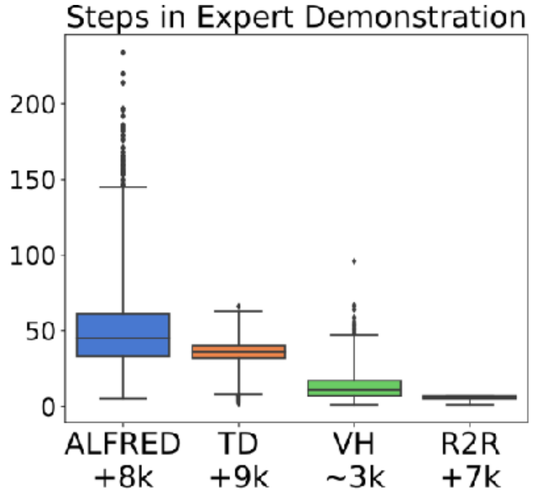
A larger search space

Speaker-Follower — Fried et al. 2018

Partial Observability

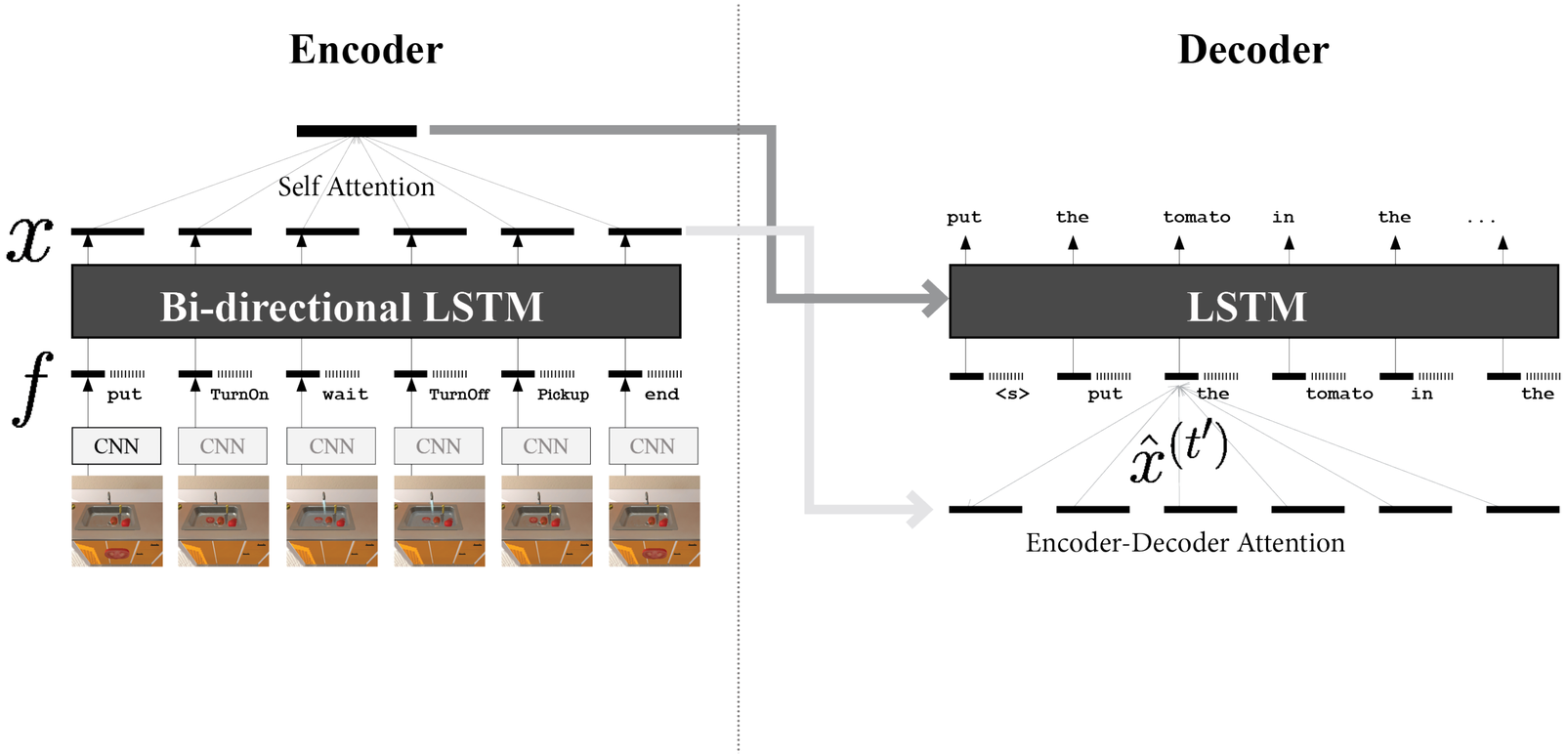


State Changes



Long Horizon

Simulator State based Aux Losses



Visibility (End of subgoal)

Sink 1

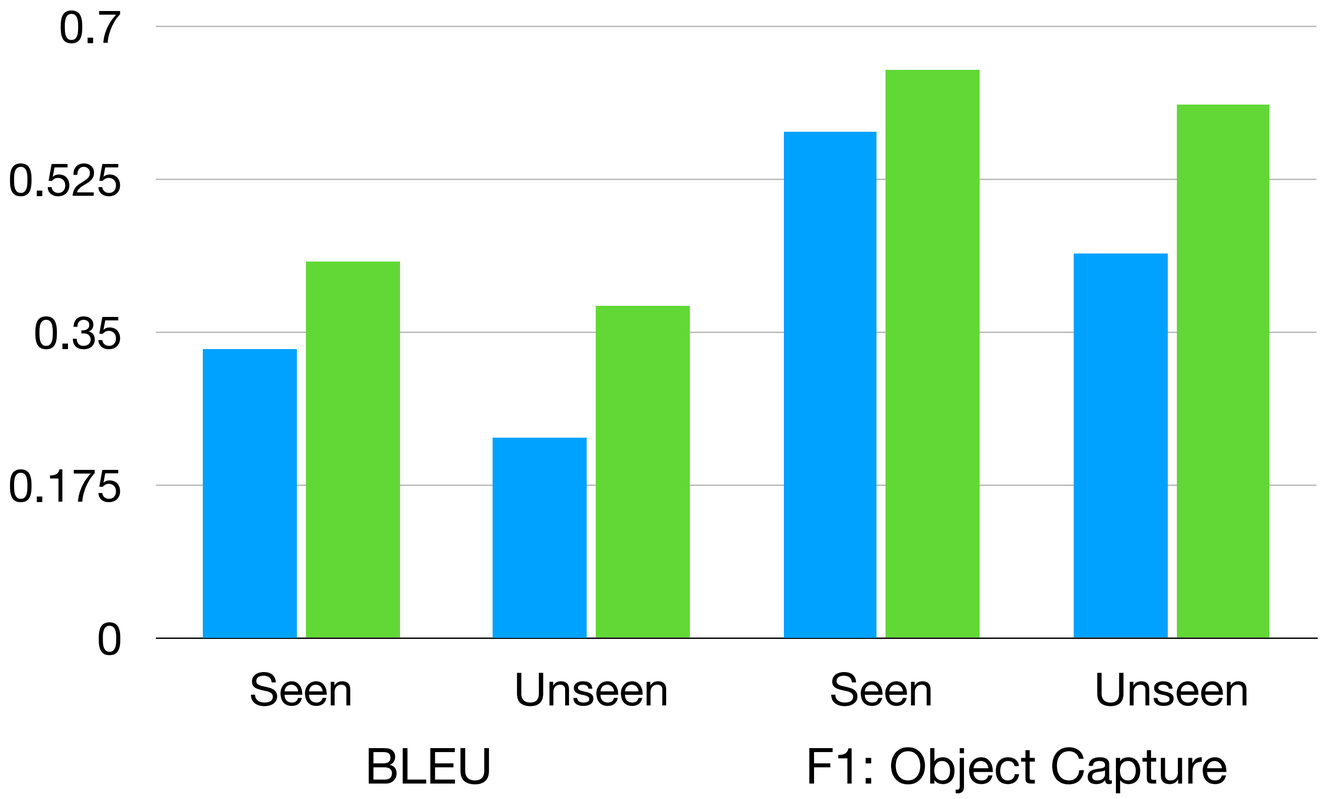
Tomato slice 1

Changed (End of subgoal)

Sink 0

Tomato slice 1

Very Hard to Focus

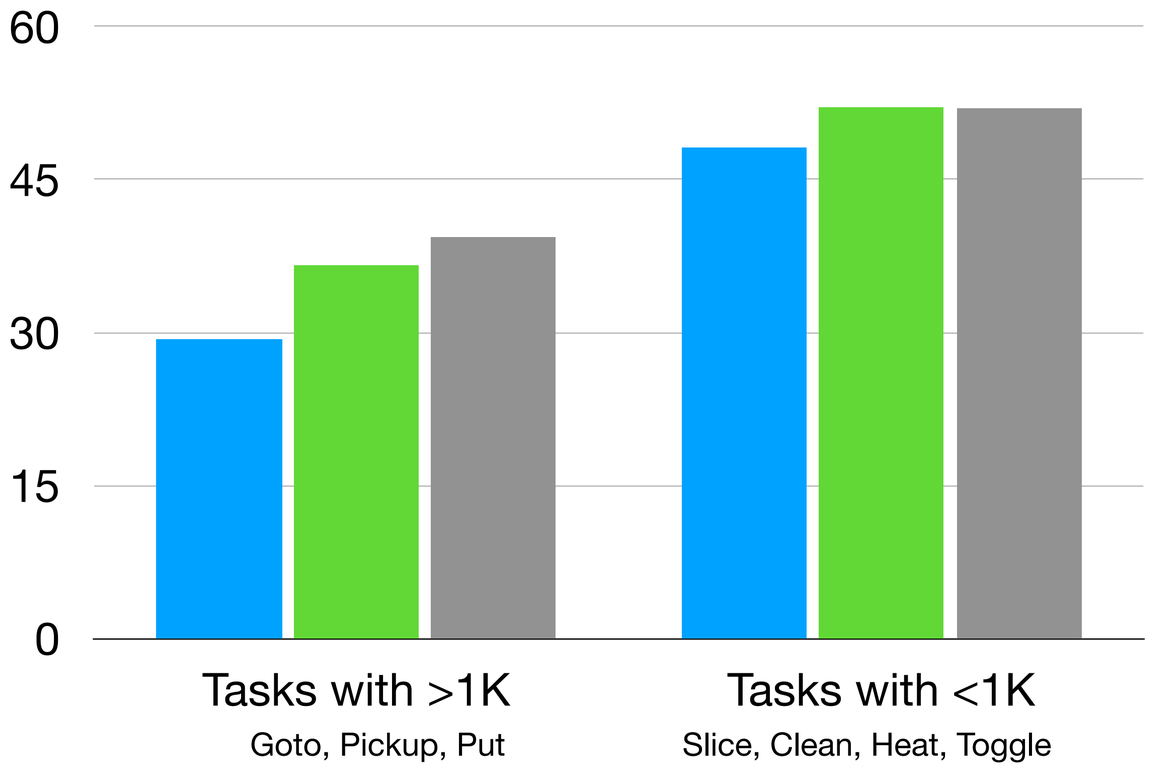


Object Capture:
Are correct objects mentioned per sub-goal

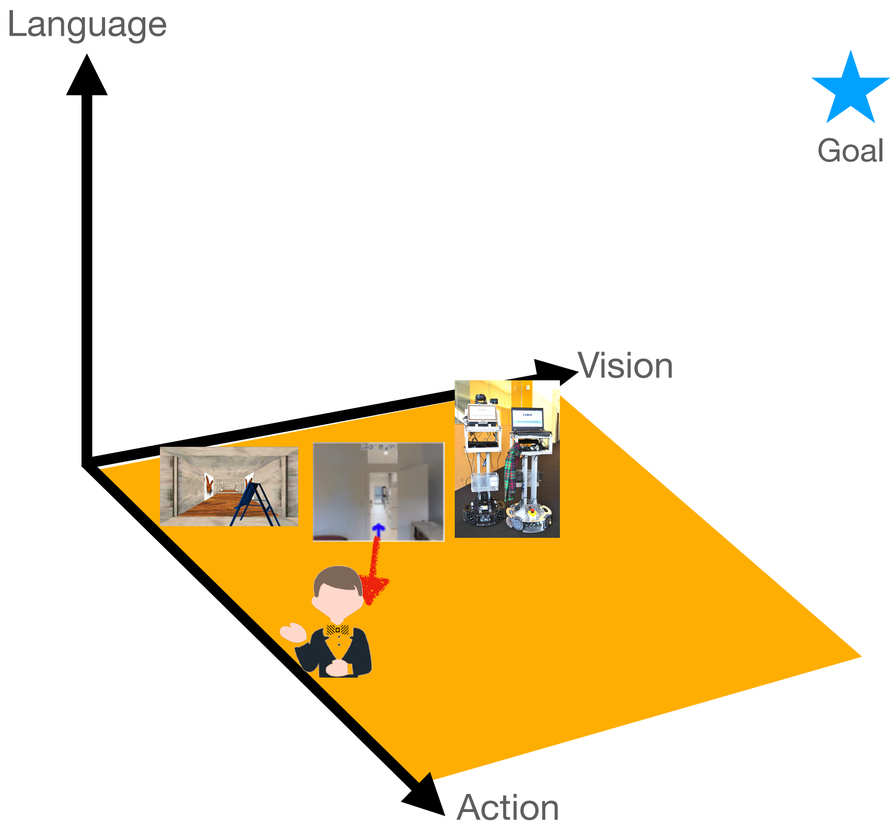
Ideally… A temporal scene-graph
and a SPICE style metric

Data Augmentation

Training with Failures



A Small Step



ALFRED **ALFWorld**

Mohit Shridhar

Explainer

Legg Yeung

Embodiment

- Lots of noise and spurious signals
- What are semantic primitives?
- Planning, Scripts, and Abstraction
- Language is woefully underspecified

Thanks!

All of these are the “same” verb