Carnegie Mellon University
Language Technologies Institute

Connecting Language to Actions

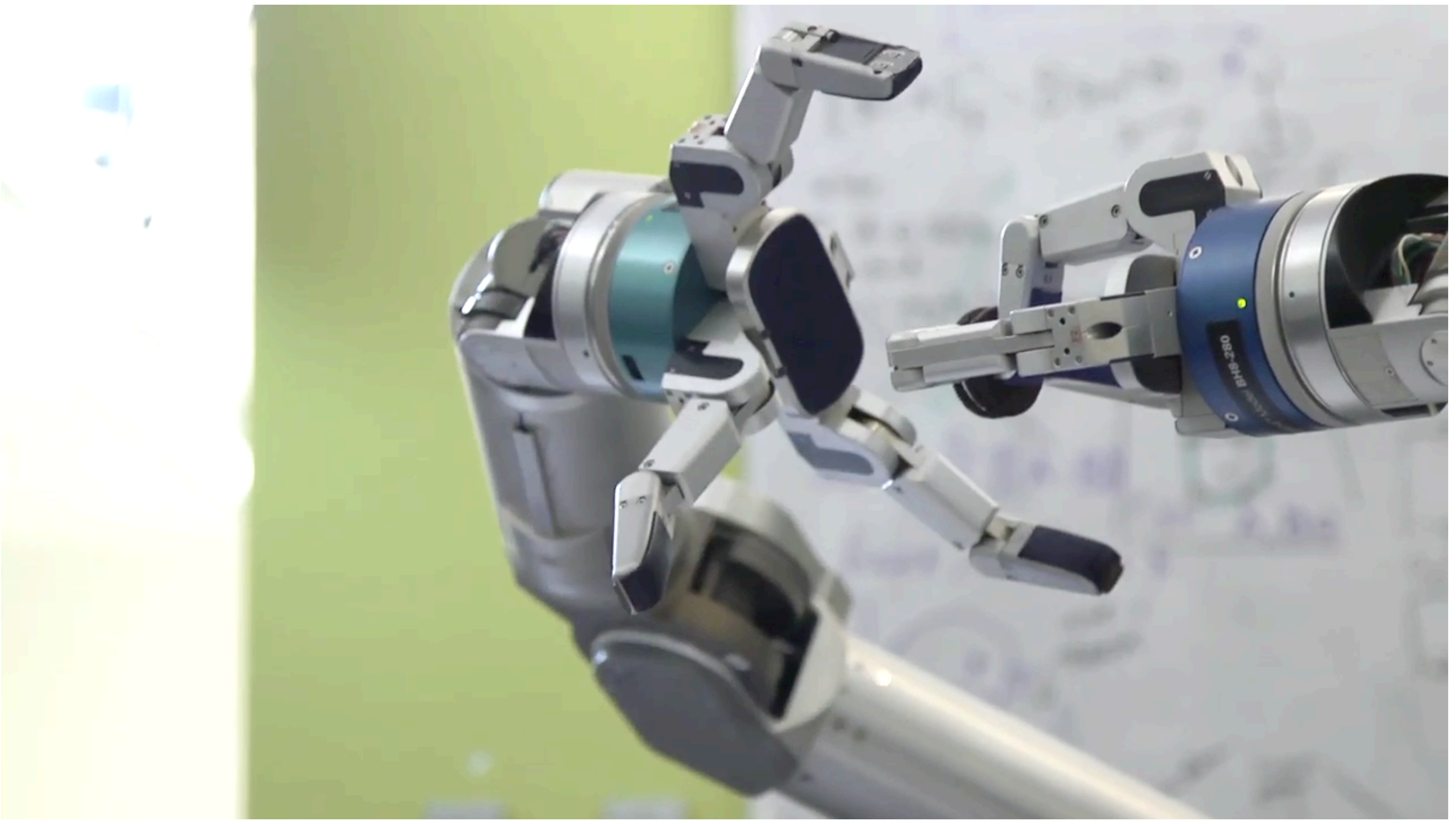
An MMML choose your own adventure game

Yonatan Bisk — Nov 17, 2020

Why?

Language that affects the world

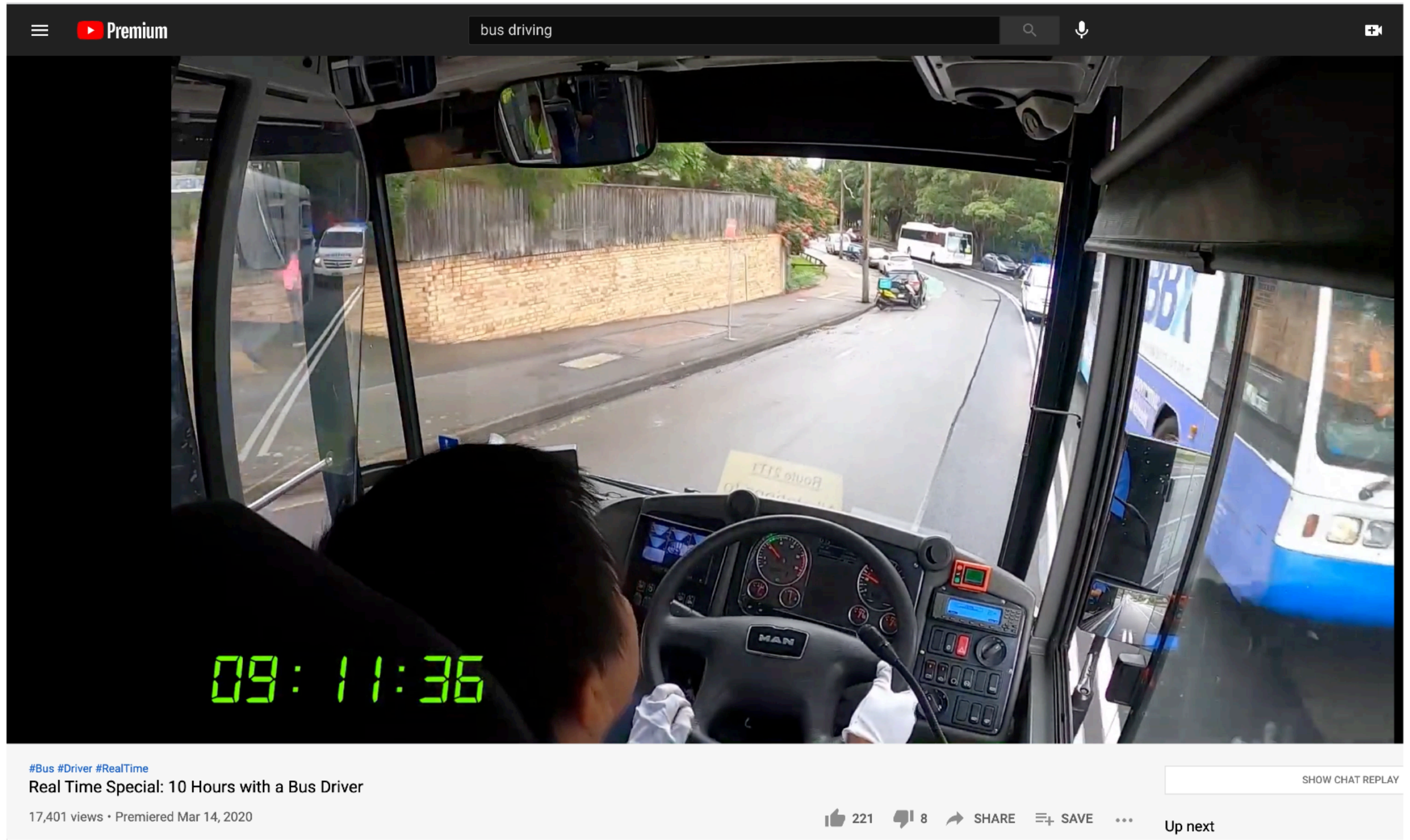
Remove the cream from the middle of the Oreo…



HERB (Siddhartha Srinivasa)

Access to Broader Semantics

What's it like to drive a bus?



#Bus #Driver #RealTime

Real Time Special: 10 Hours with a Bus Driver

17,401 views • Premiered Mar 14, 2020

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

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

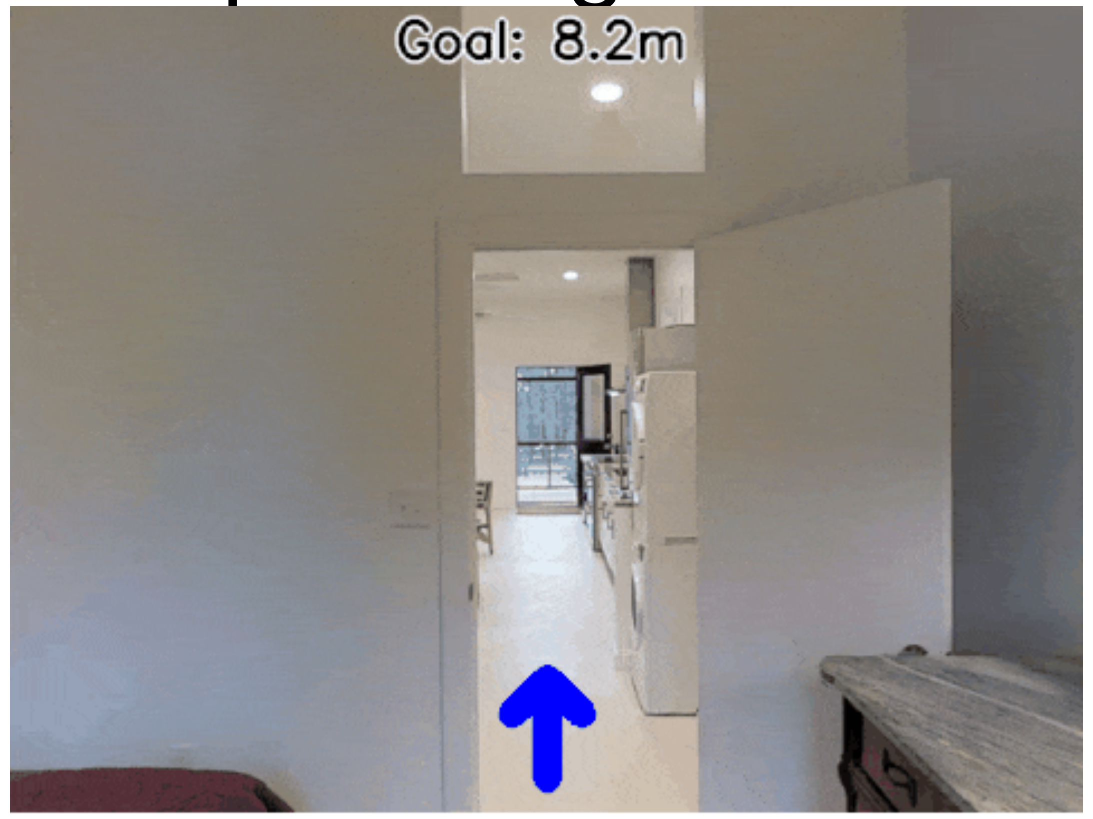
What does interaction mean?

Grid World?

0
1
2
3
4
ACTIONS
UP
DOWN
LEFT
RIGHT
STATE
CURRENT LOCATION
[1 , 2]
PREVIOUS LOCATION
[1 , 3]
VISUAL SCOPE
[WHOLE ROOM]

Reinforcement Learning: Crash Course AI#9
https://www.youtube.com/watch?v=nIgIv4IfJ6s

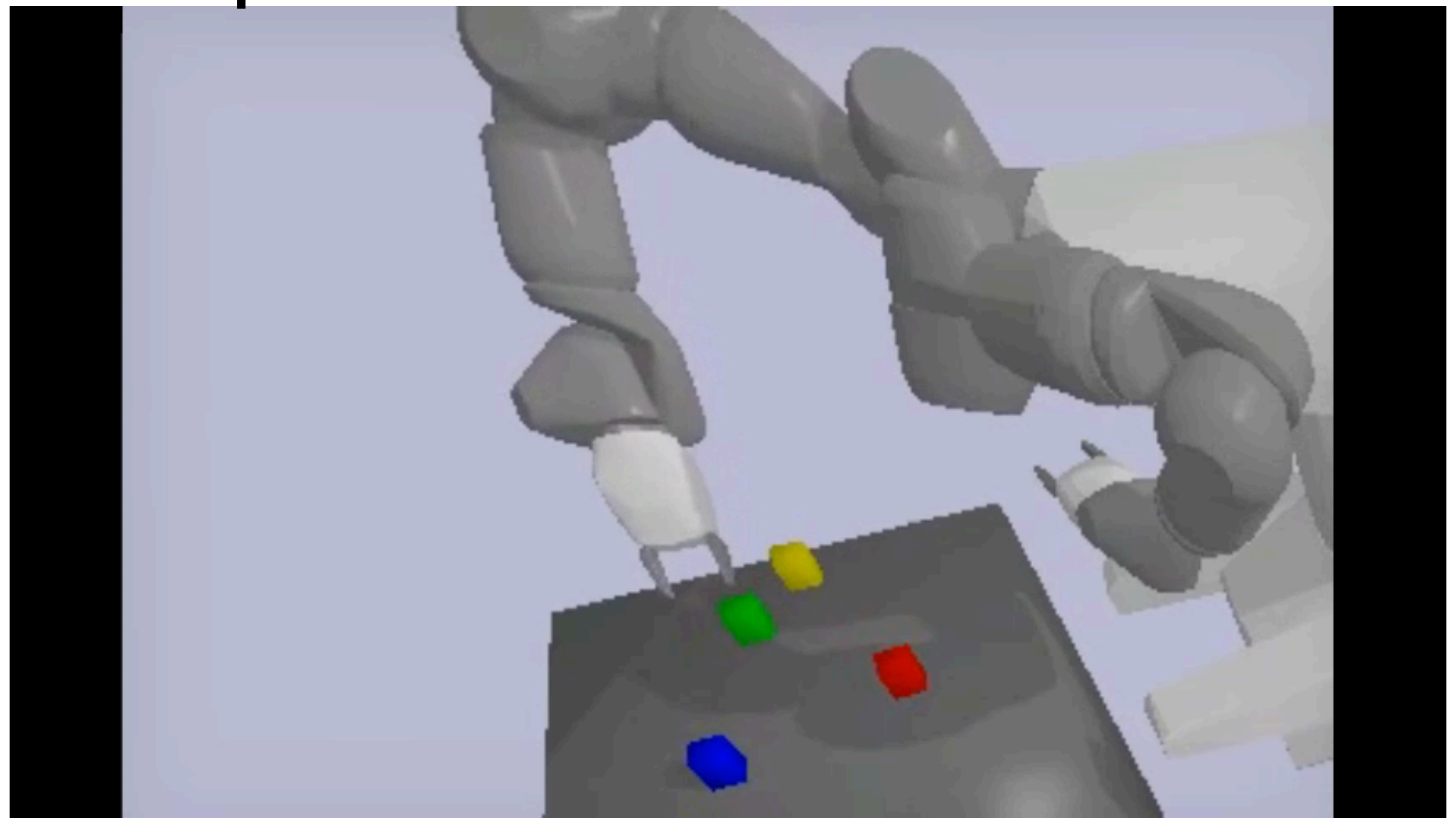
Graph Navigation?



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

Anderson 2018

Manipulation?

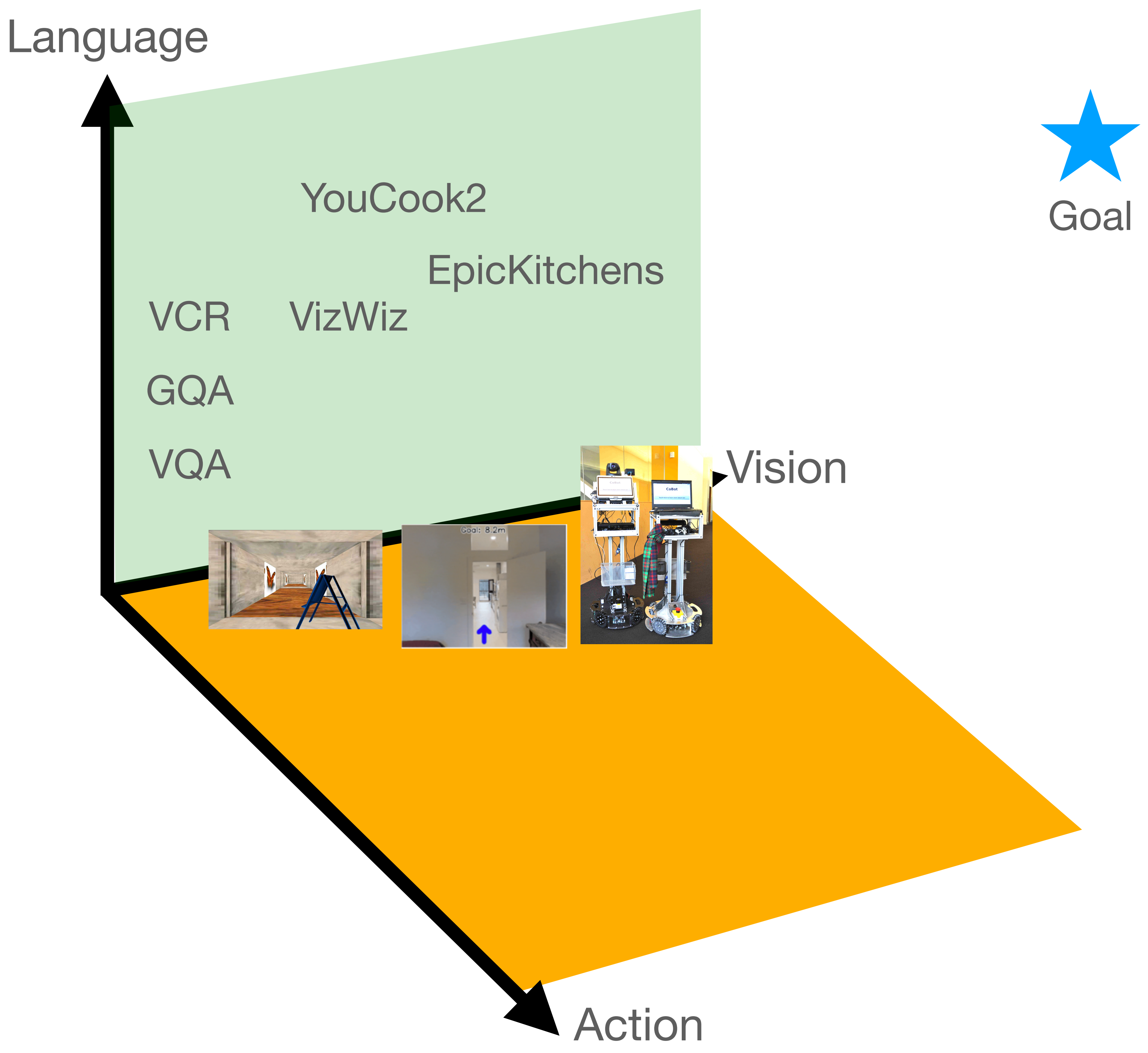


Paxton 2019

1. How does the agent move?
2. How many arms or legs does it have?
3. How many fingers (if any) do the grippers have?
4. How many joints do the limbs have?
5. What about physics? Real motor noise?

...

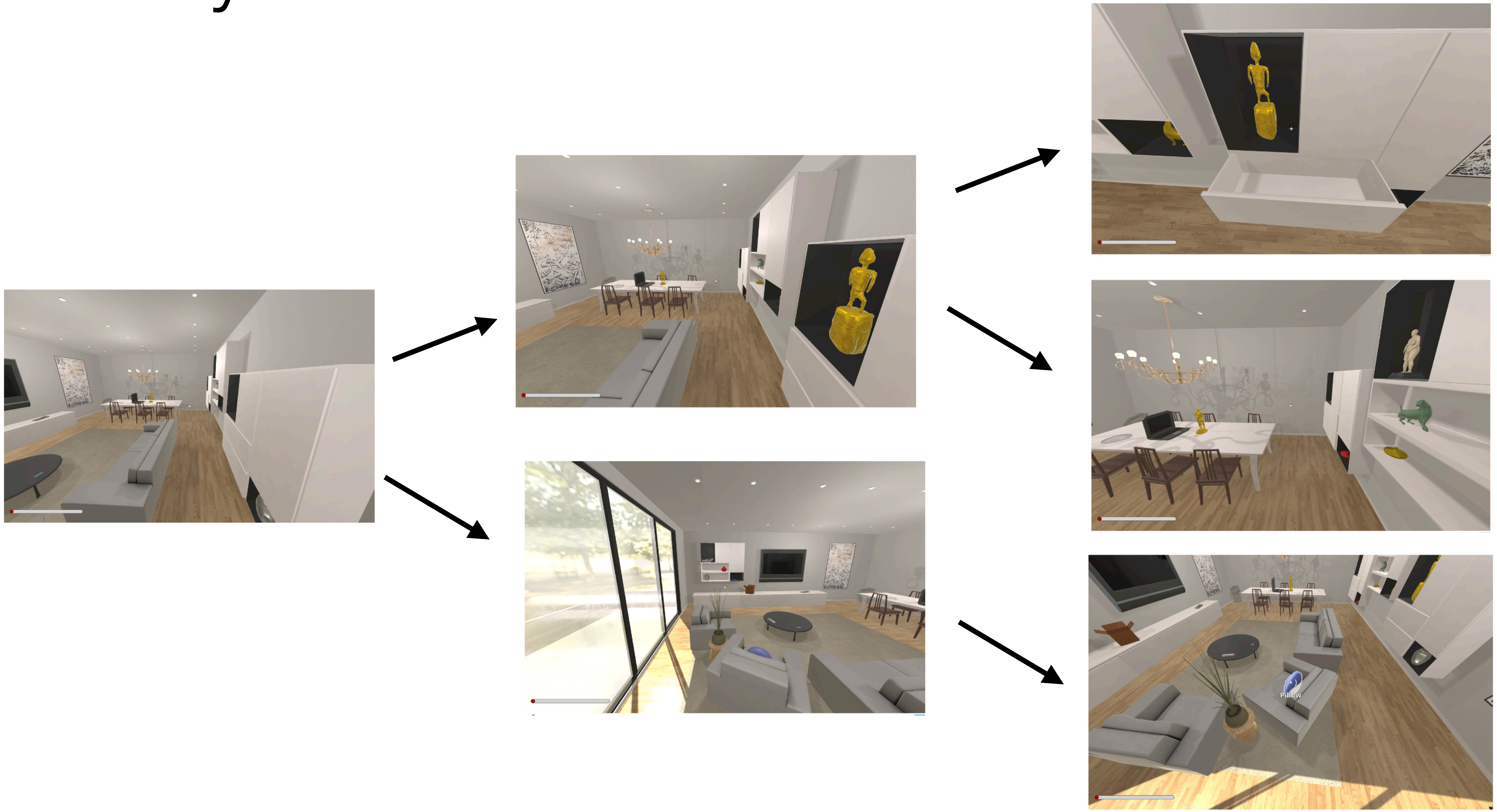
Every Dimension Interacts



1. How rich or abstract is the language?
2. How complex is the visual field?
3. Is the vision 2D, 3D, Lidar, … ?
4. What kind of supervision do you have?

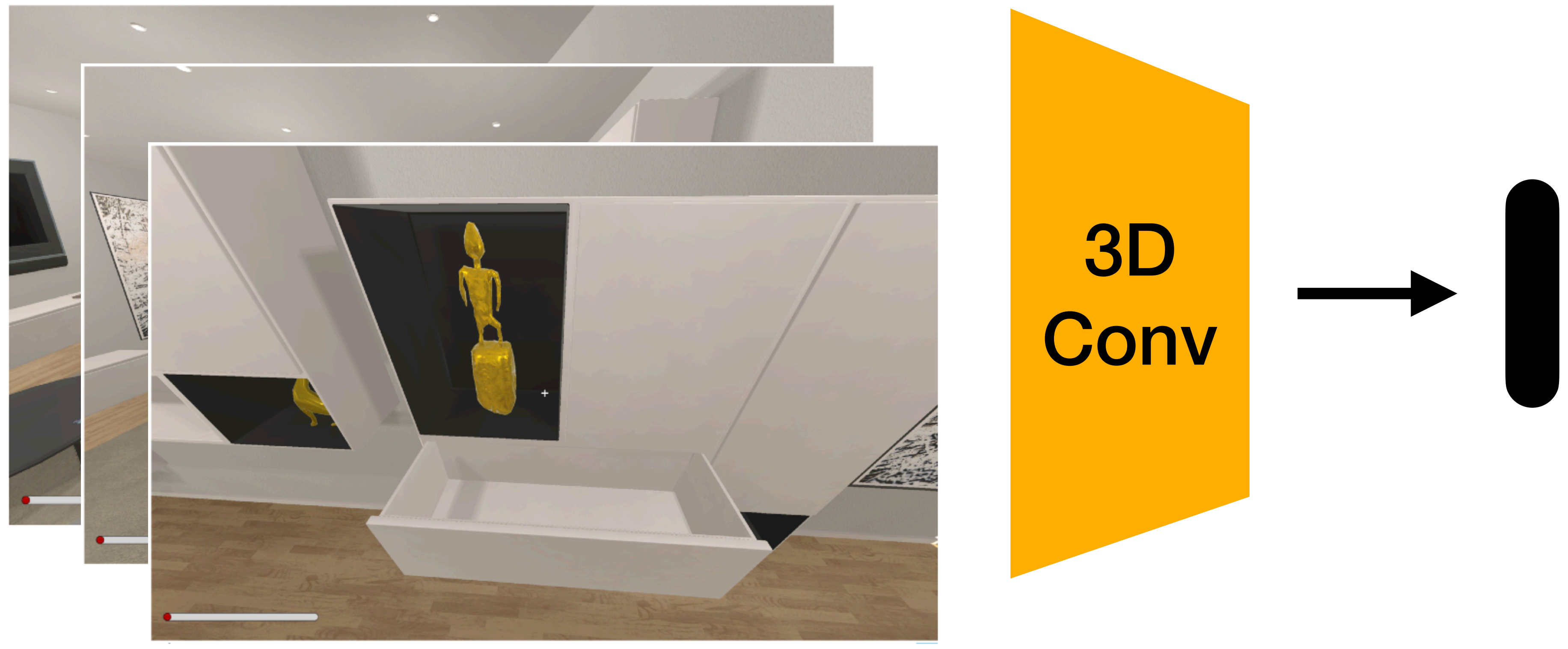
…

Choose your own adventure



Sequential and Online Modeling

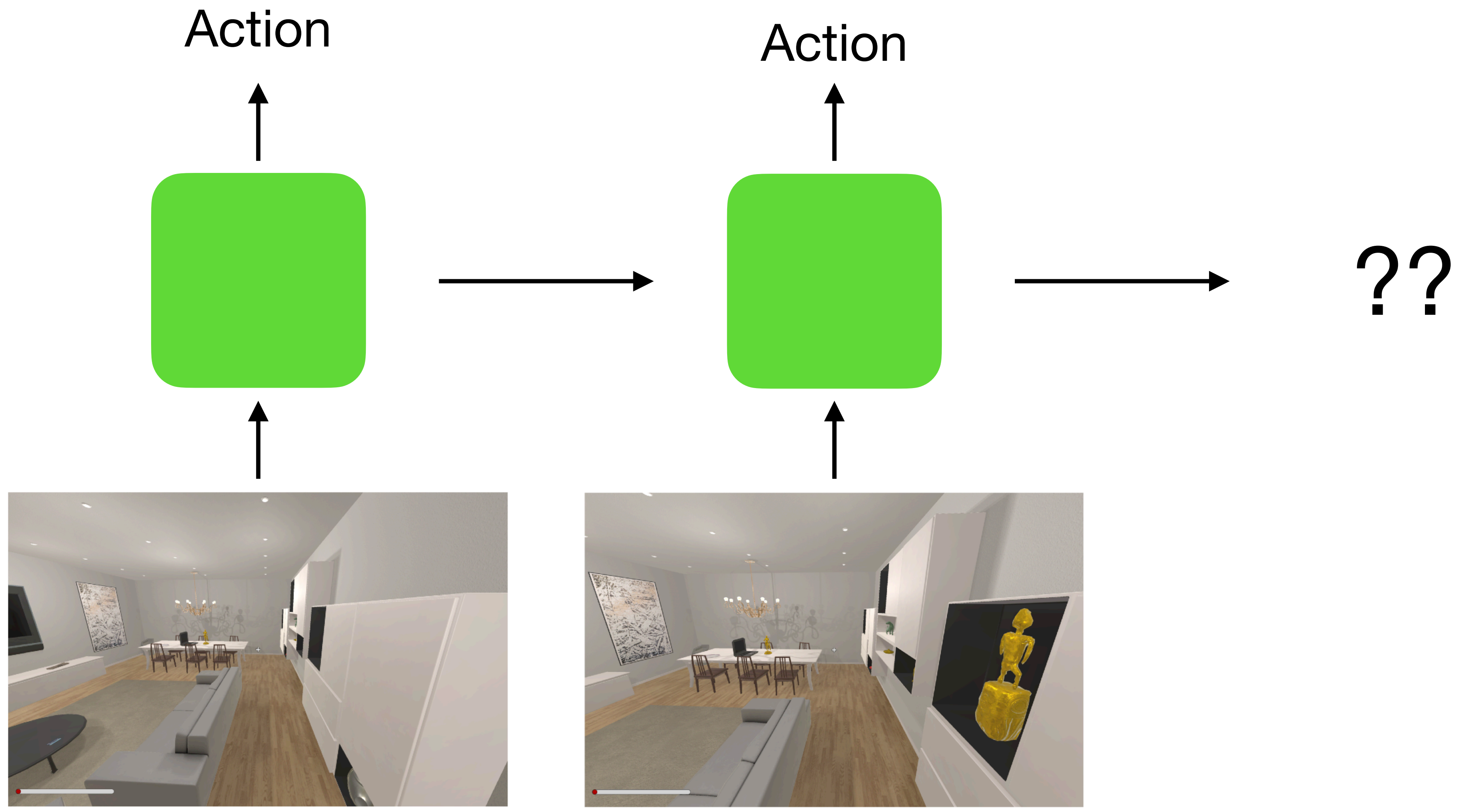
Action Recognition



“Action Summary”

$$p(\text{Action}|v_0, ..., v_t)$$

Embodied



$$p(v_t|v_0, ..., \text{Action})$$

Requirement: Have a goal

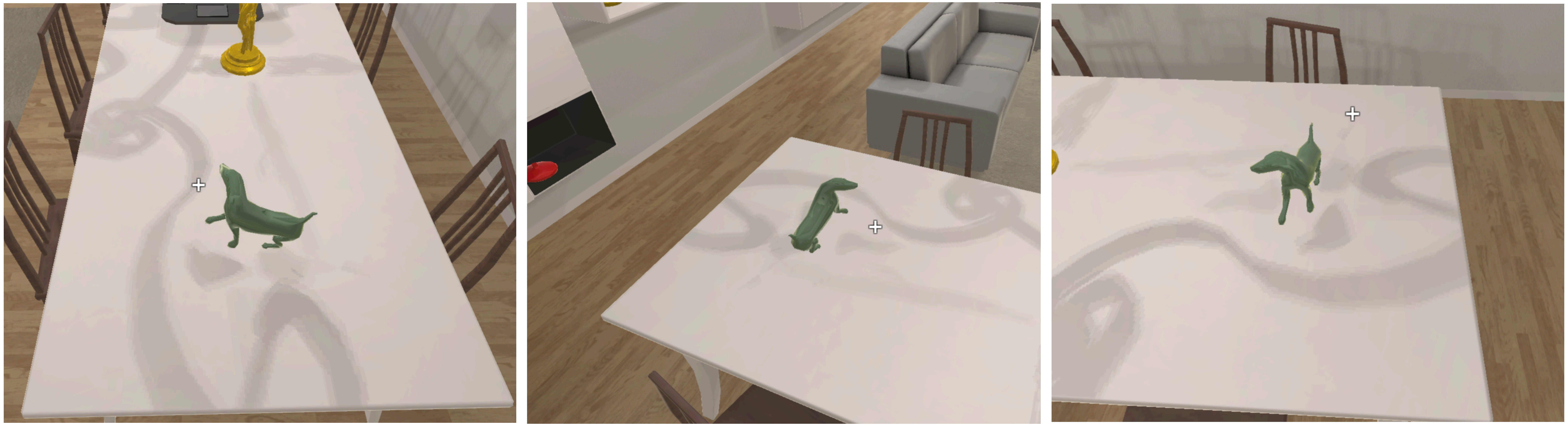
What is a “goal”?

“Put the green dog on the table”

$p(v_t|v_0, ..., \text{Action})$

$p(v_t|v_0, ..., v_{t-1}, a_0, ..., a_t)$

$v_t =$



Planning

Pre- and Post-Conditions

Task 4: Must locate object,
to move to object

Task 3: Must move to object,
to hold object

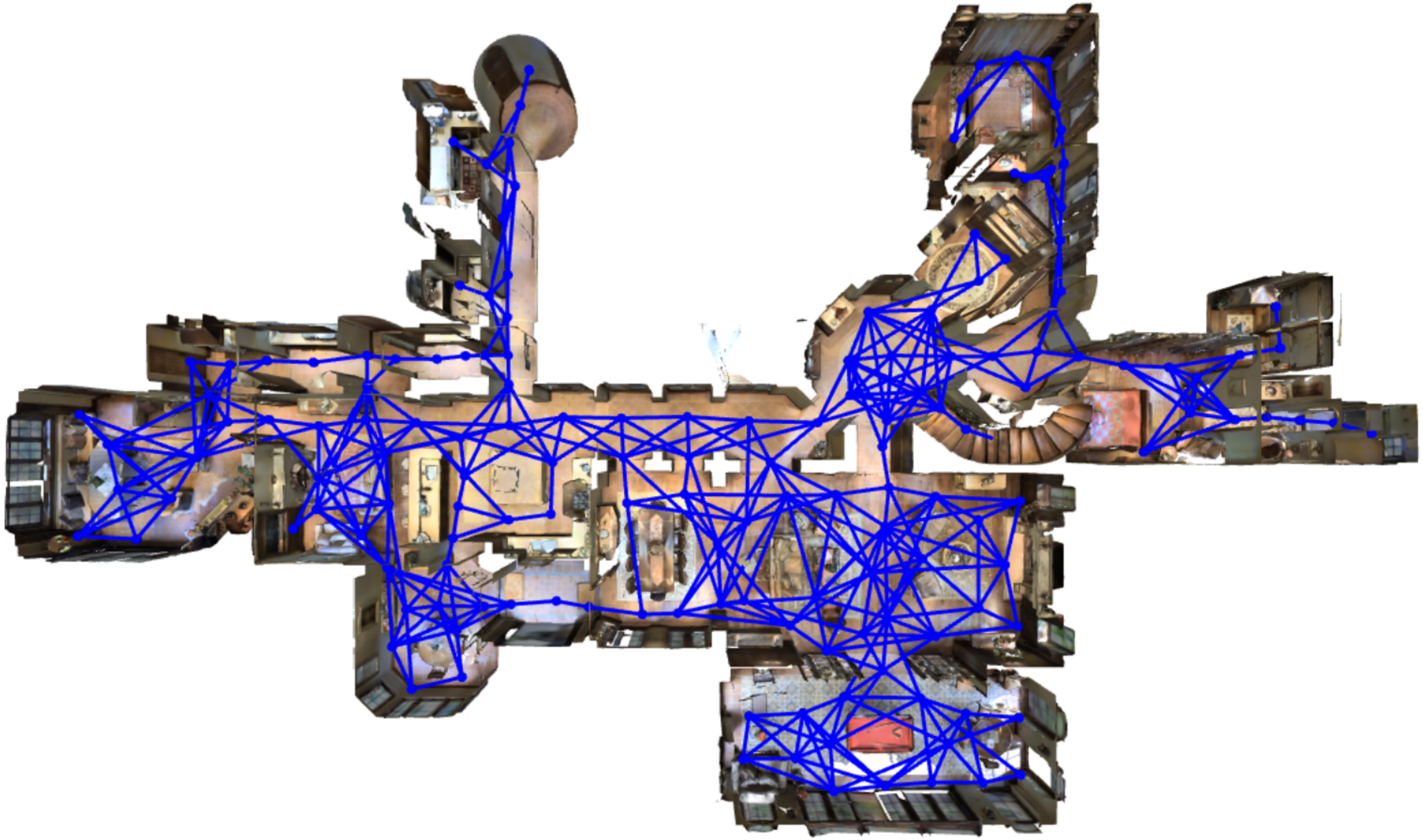
Task 2: Must hold object,
to place object

Task 1: Recognize Success

Instances of “green dog sculpture on table”

So what are we actually optimizing? What’s our actual goal?

Let’s Start Simple



Instruction Following

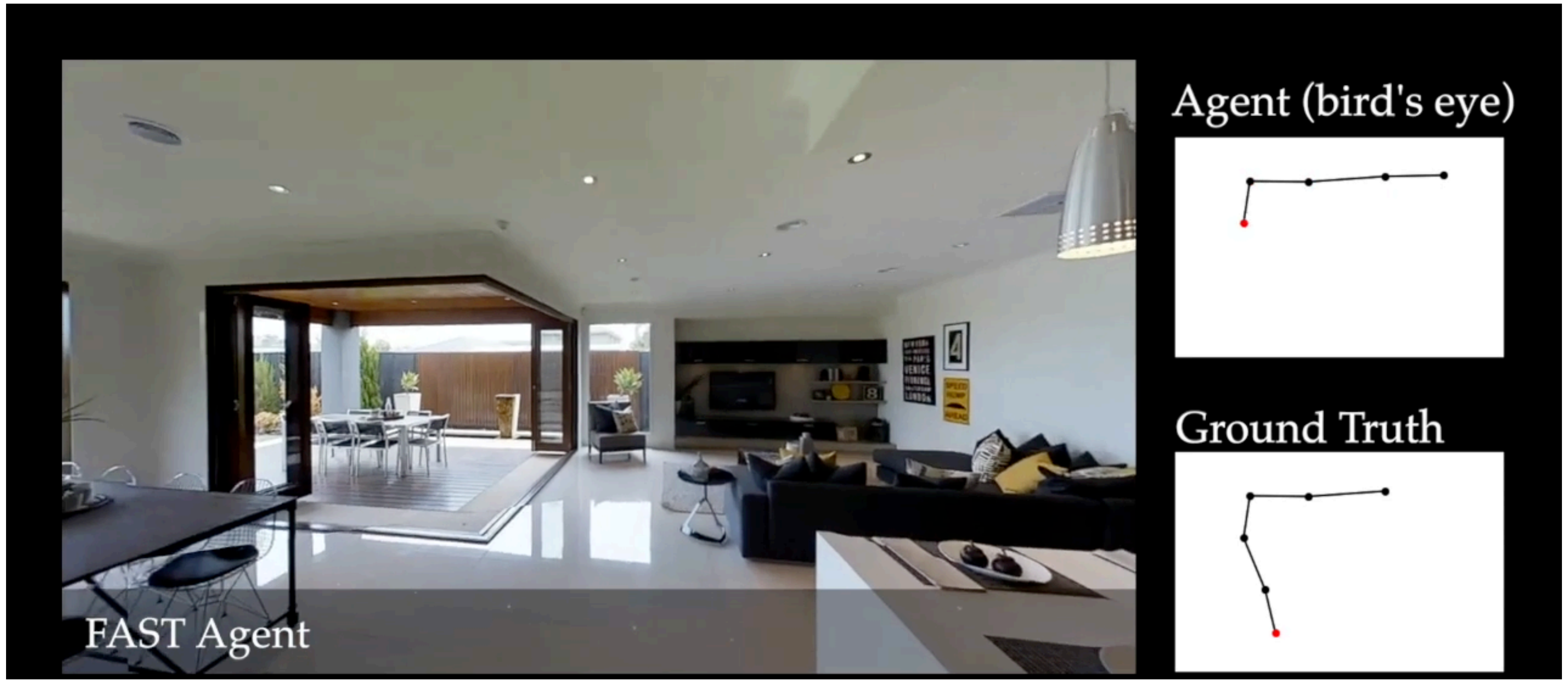
Explicit Action Supervision

Walk out of the bedroom through the open door into the hallway

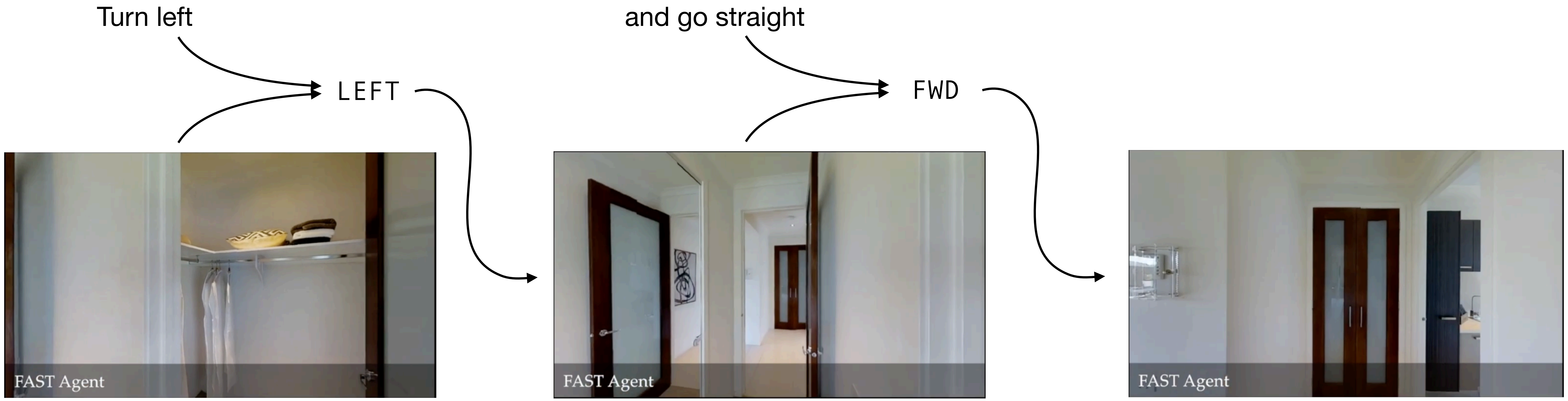
Turn the corner and walk into the dining area.

Pass the dining table and walk into the living room area towards the television.

Stop near the chair and open sliding doors to outside



V+L -> A



Turn left

LEFT

and go straight

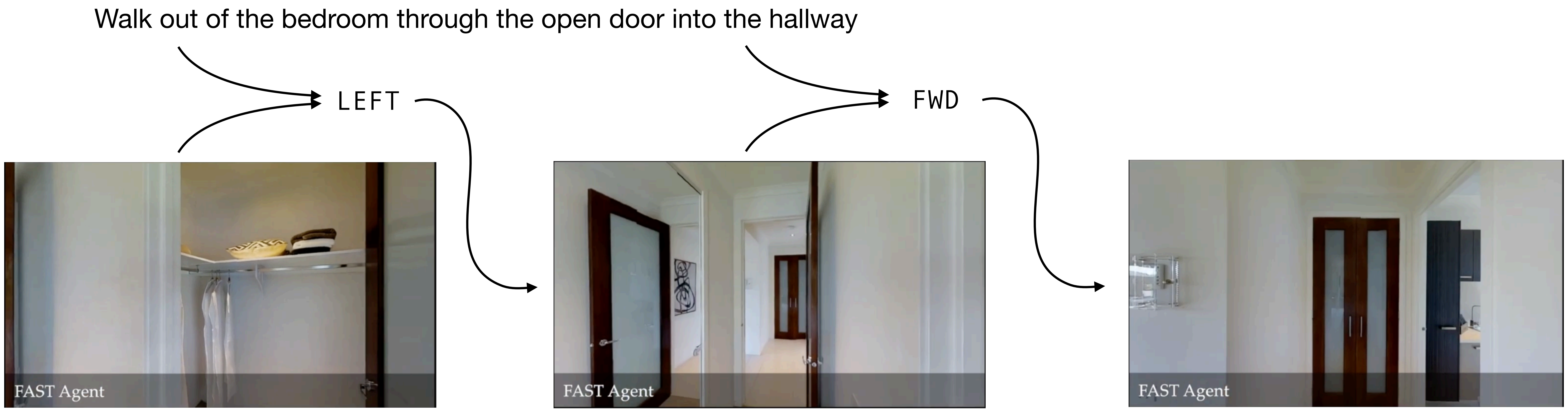
FWD

Does this actually need vision?

Does this understand plans?

No, this is ~Semantic Parsing

V+L -> A



Walk out of the bedroom through the open door into the hallway

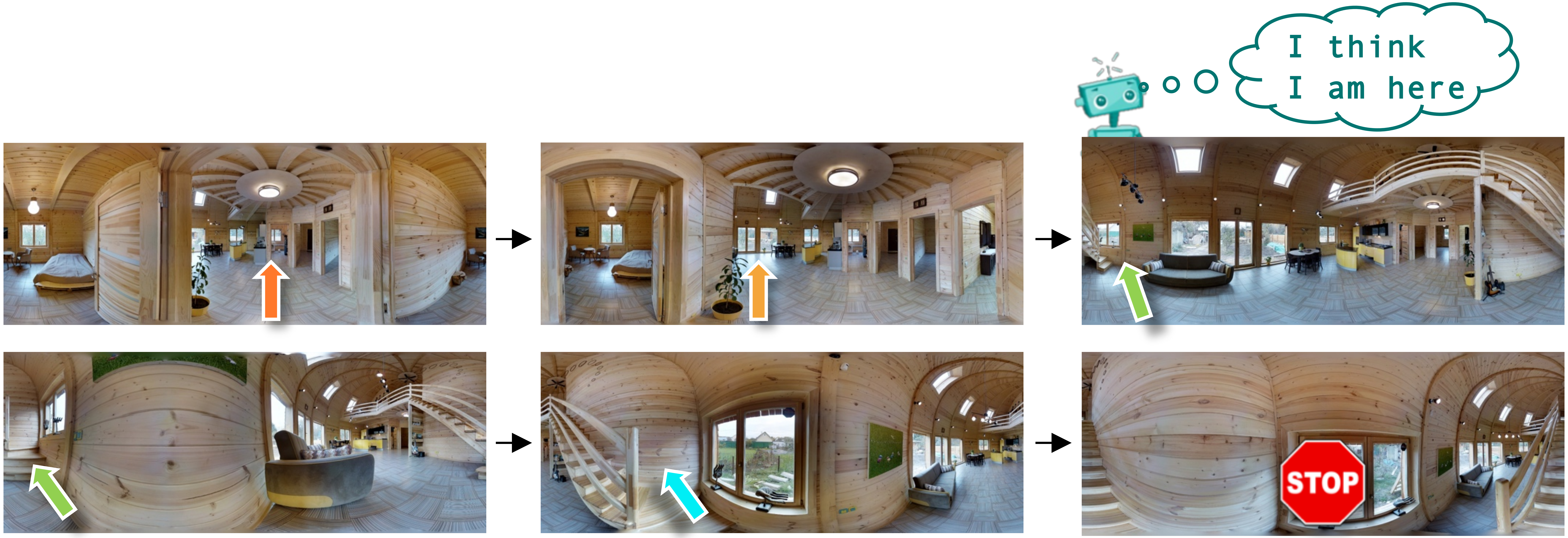
LEFT

FWD

Does this actually need vision?	Yes
Does this understand plans?	Maybe, probably not

First Major Question: Alignment

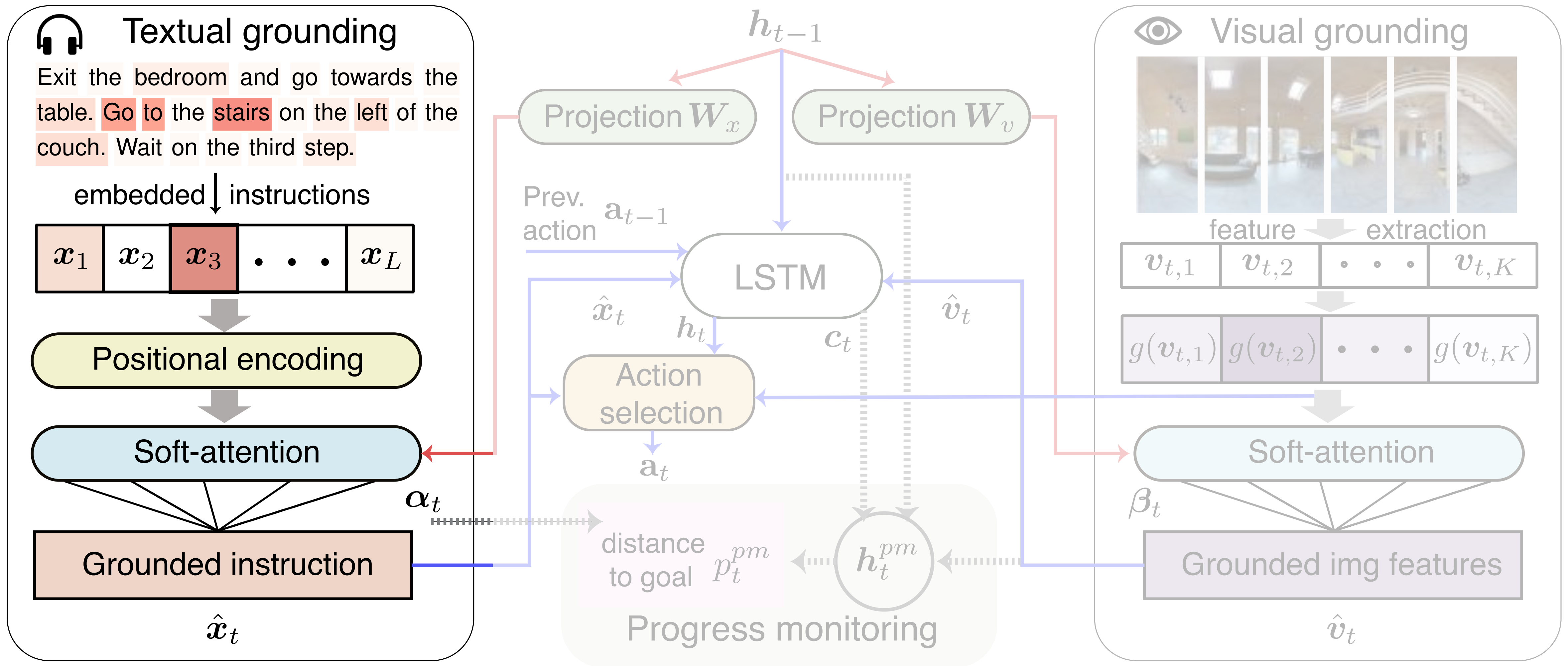
Exit the bedroom and go towards the table. Go to the stairs on the left of the couch. Wait on the third step.



Ma et al, "Self-Monitoring Navigation Agent via Auxiliary Progress Estimation" ICLR 2019

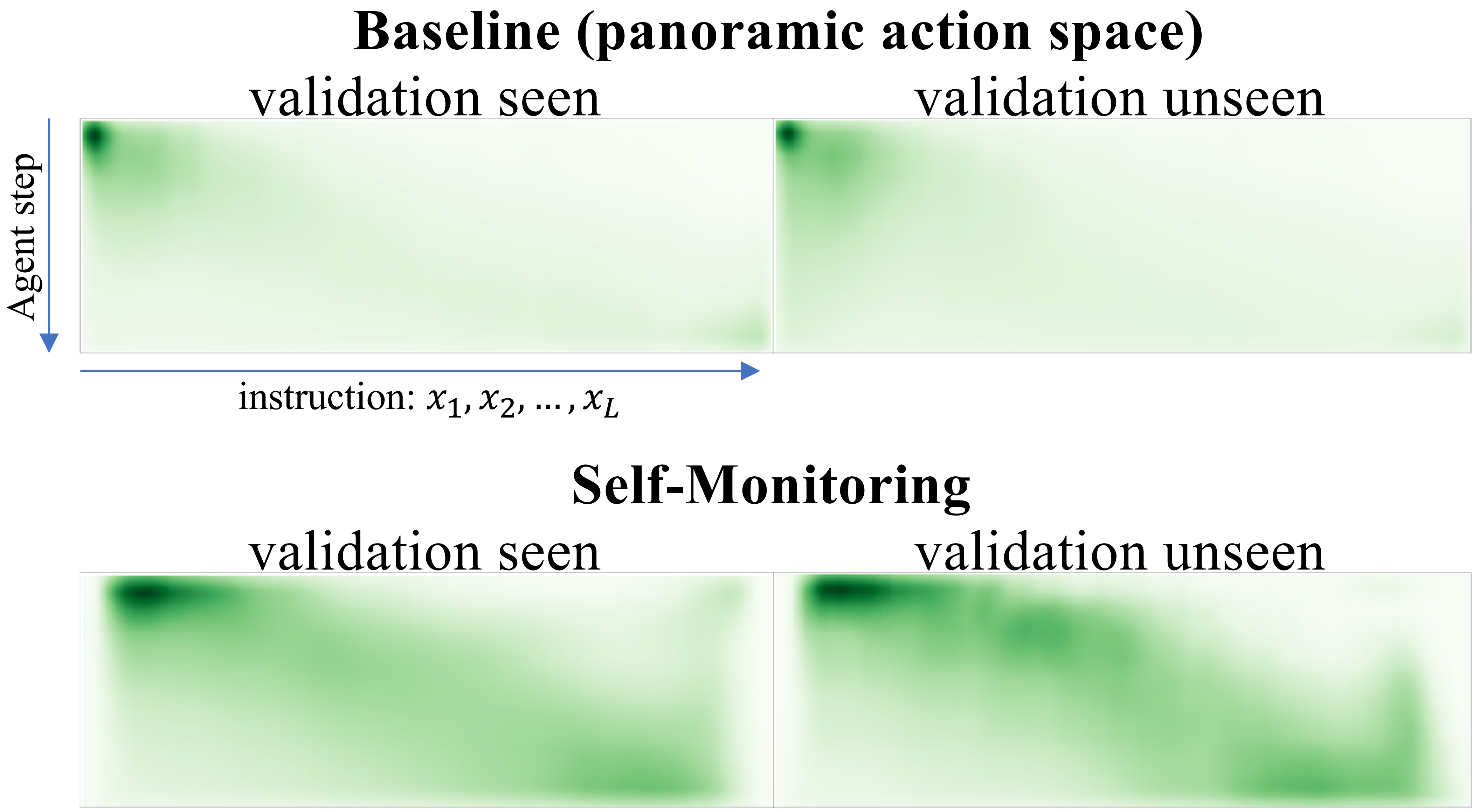
Alignment

~~Exit the bedroom~~ and ~~go towards the table~~. Go to the stairs on the left of the couch. Wait on the third step.



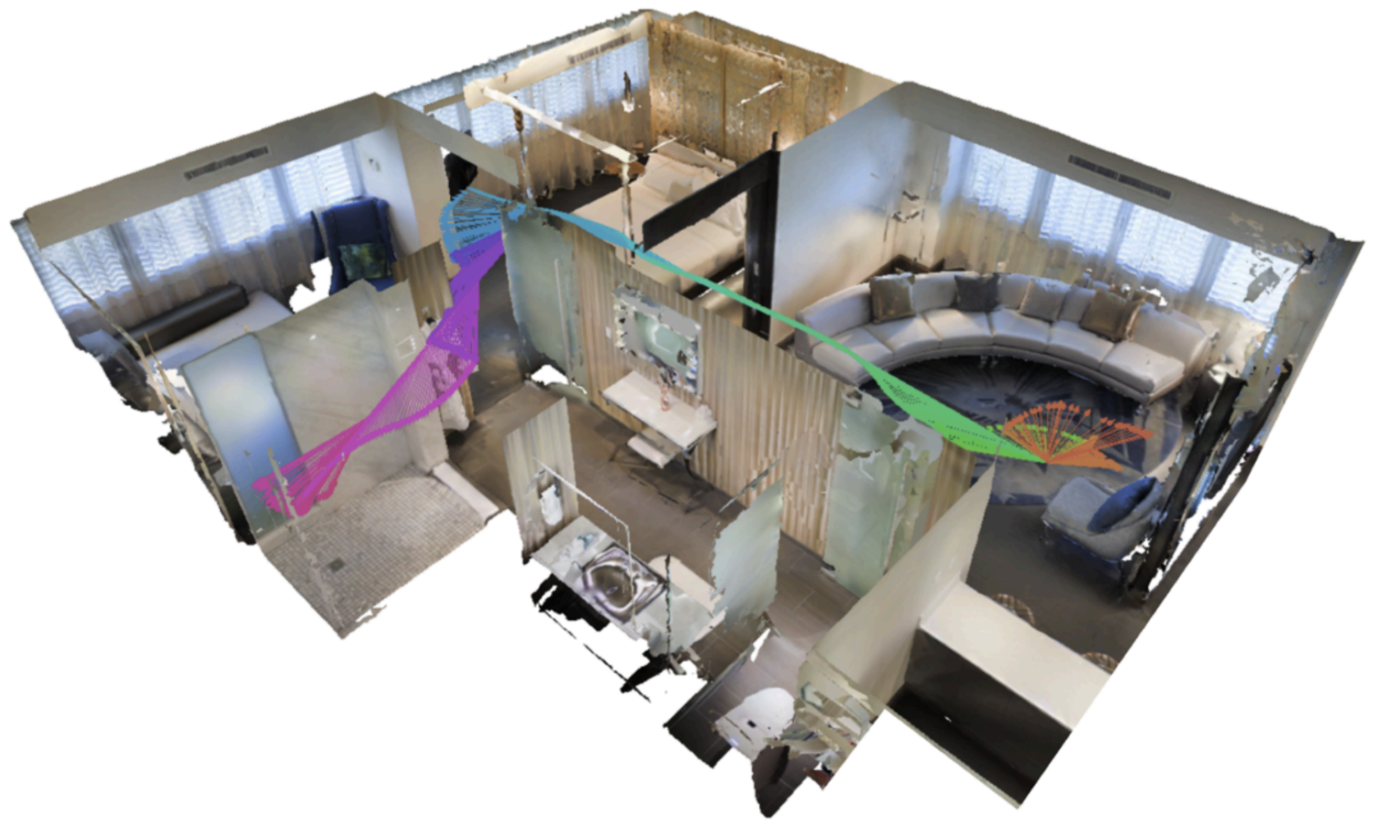
Baseline (panoramic action space)

Self-Monitoring



Lots of Data

Lots and lots of aligned data?



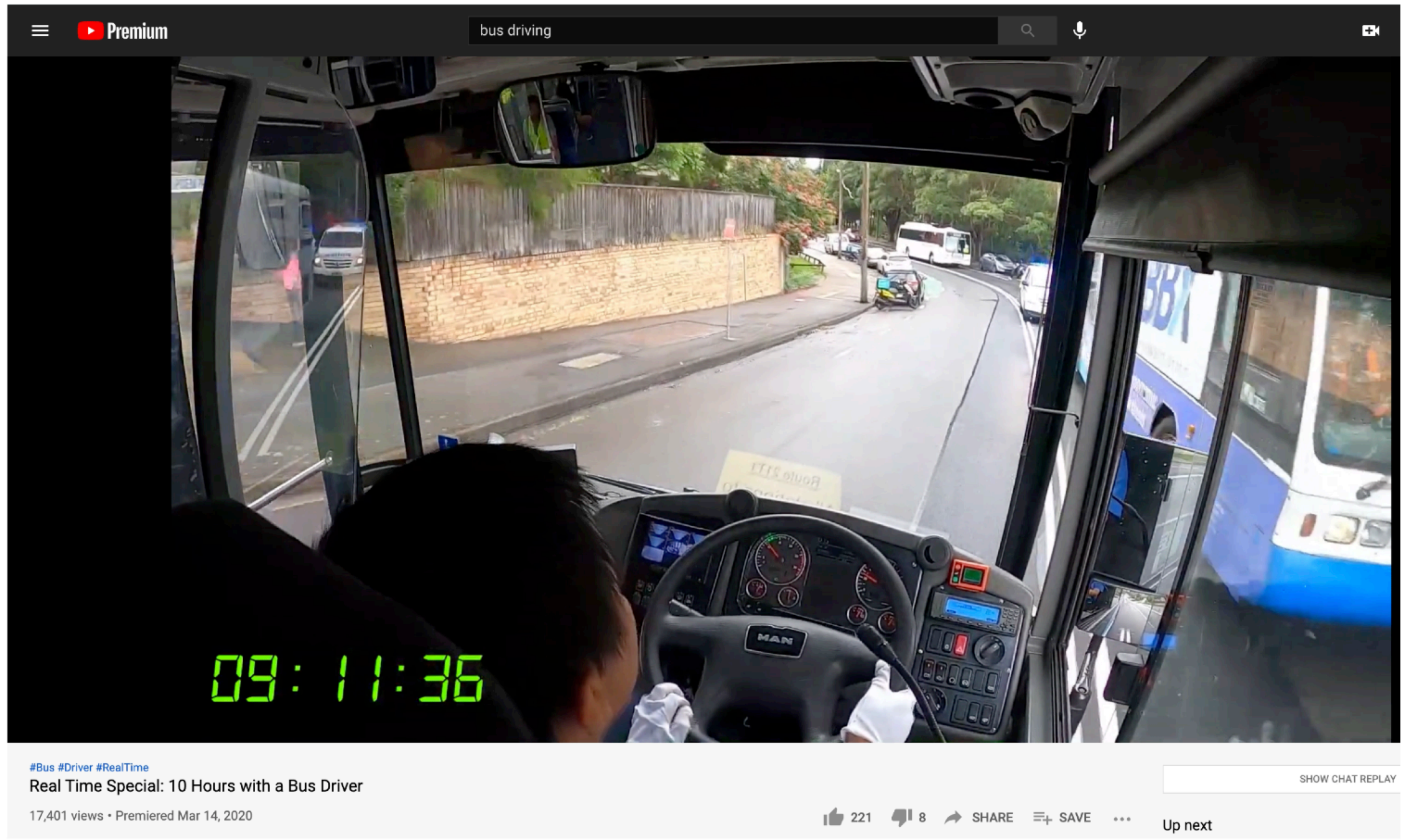
Our starting point is in a living room, we're facing towards a long beige sofa, and in front of the sofa there are three glass coffee tables, turn around and exit through the doorway that's in front of you, walk pass the bed that's on your right and then turn left, we're now facing towards another living room, and on the left there's an open door, walk towards that open door enter the bathroom that's in front of you, turn towards the right into the shower area, and that's your destination.

	Number of:				Includes:		
	Lang	Instruct	Words	Paths	Text	Ground	Demos
CVDN	1	2K†	167K	7K	✓		
R2R	1	22K	625K	7K	✓		
Touchdown	1	9K	1.0M	9K	✓	✓‡	
REVERIE	1	22K	388K	7K	✓	✓‡	
RxR	3	126K	9.8M	16.5K	✓	✓	✓

†The number of dialogues. ‡Grounding limited to one object per instruction.

Ku et al. Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding — EMNLP 2020

Wait, remember the bus driver question?

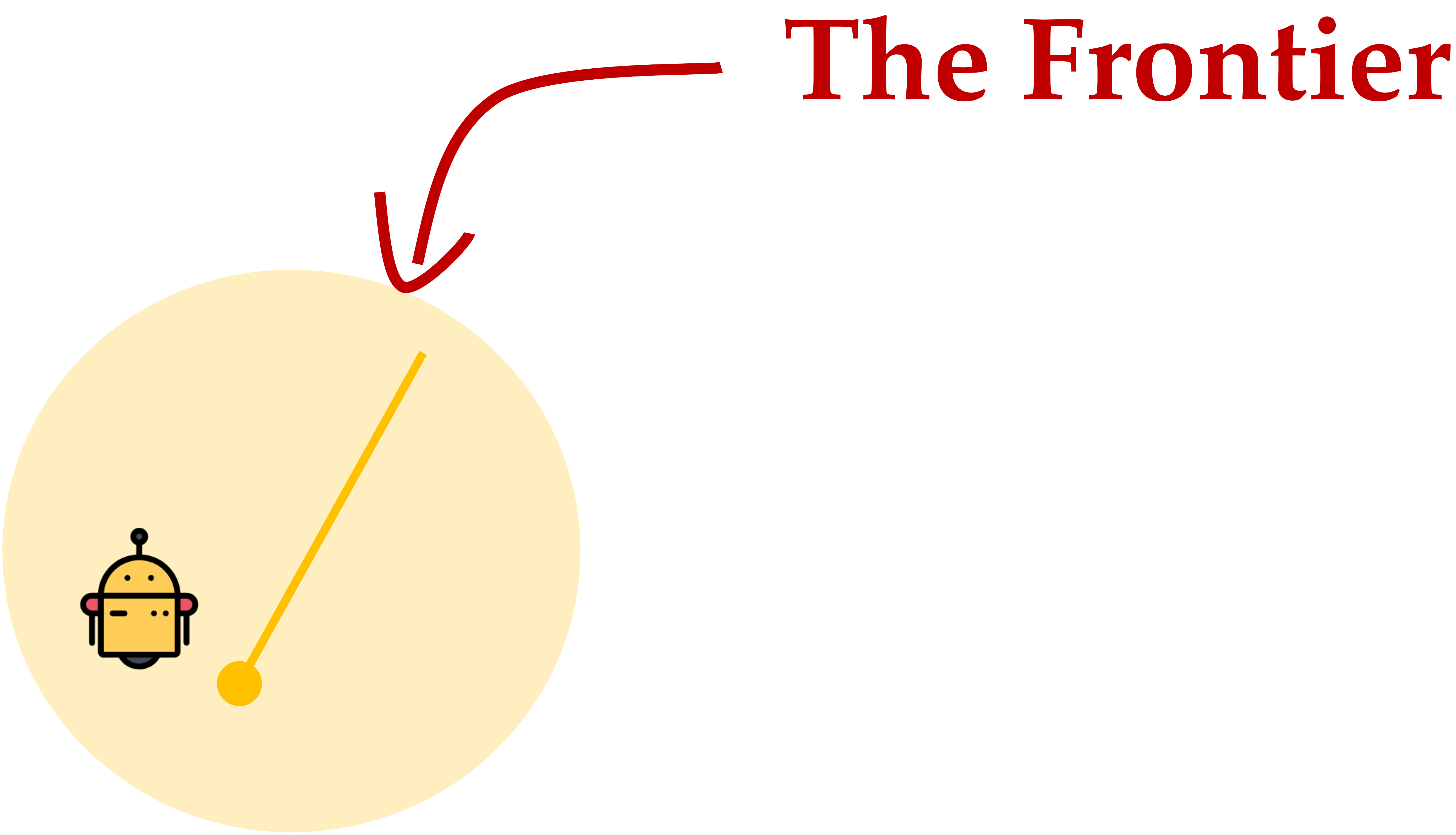


#Bus #Driver #RealTime

Real Time Special: 10 Hours with a Bus Driver

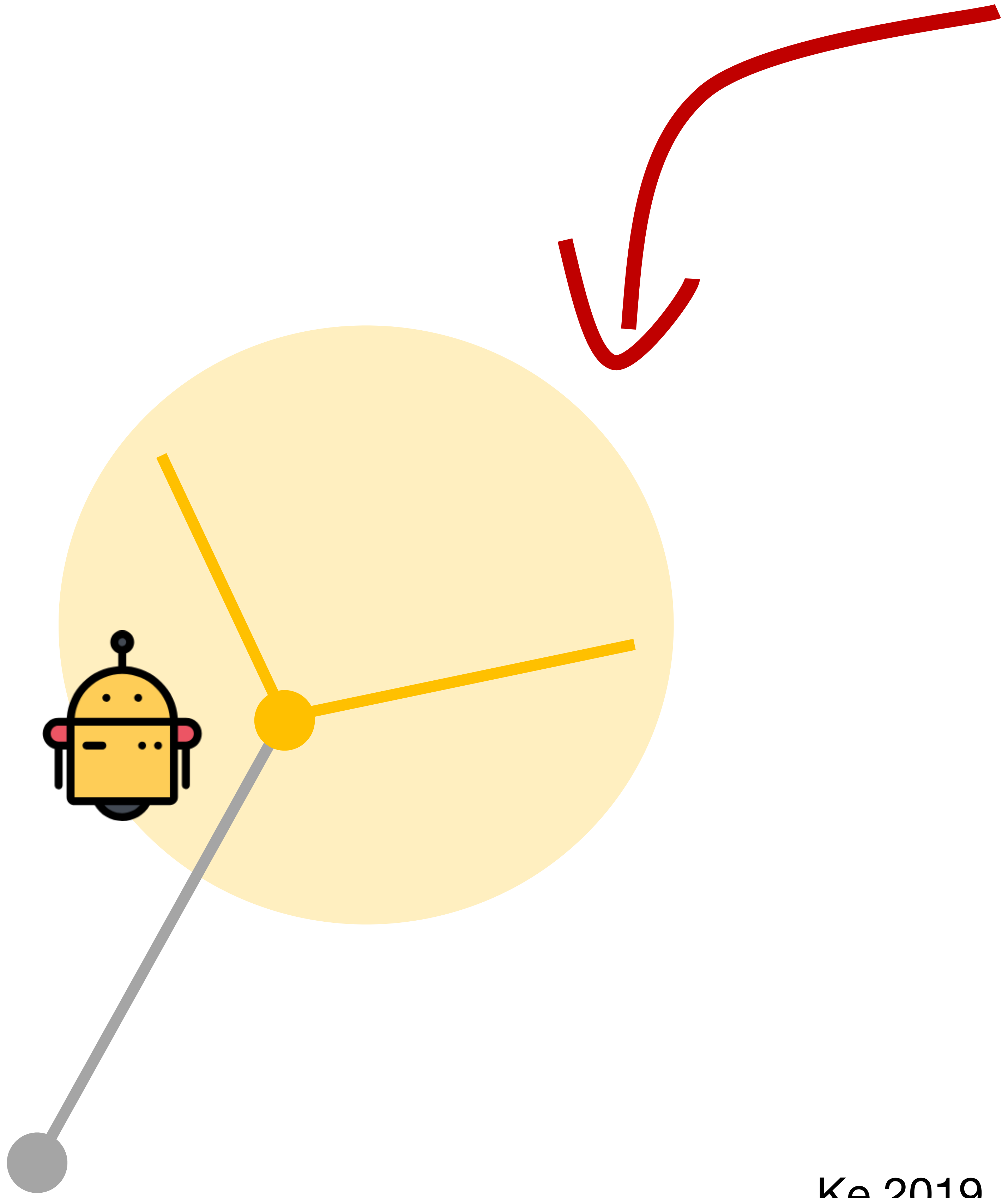
17,401 views • Premiered Mar 14, 2020

What if you make a mistake?



Ke 2019, Tactical Rewind: Self-Correction via Backtracking in Vision-and-Language Navigation - CVPR 2019

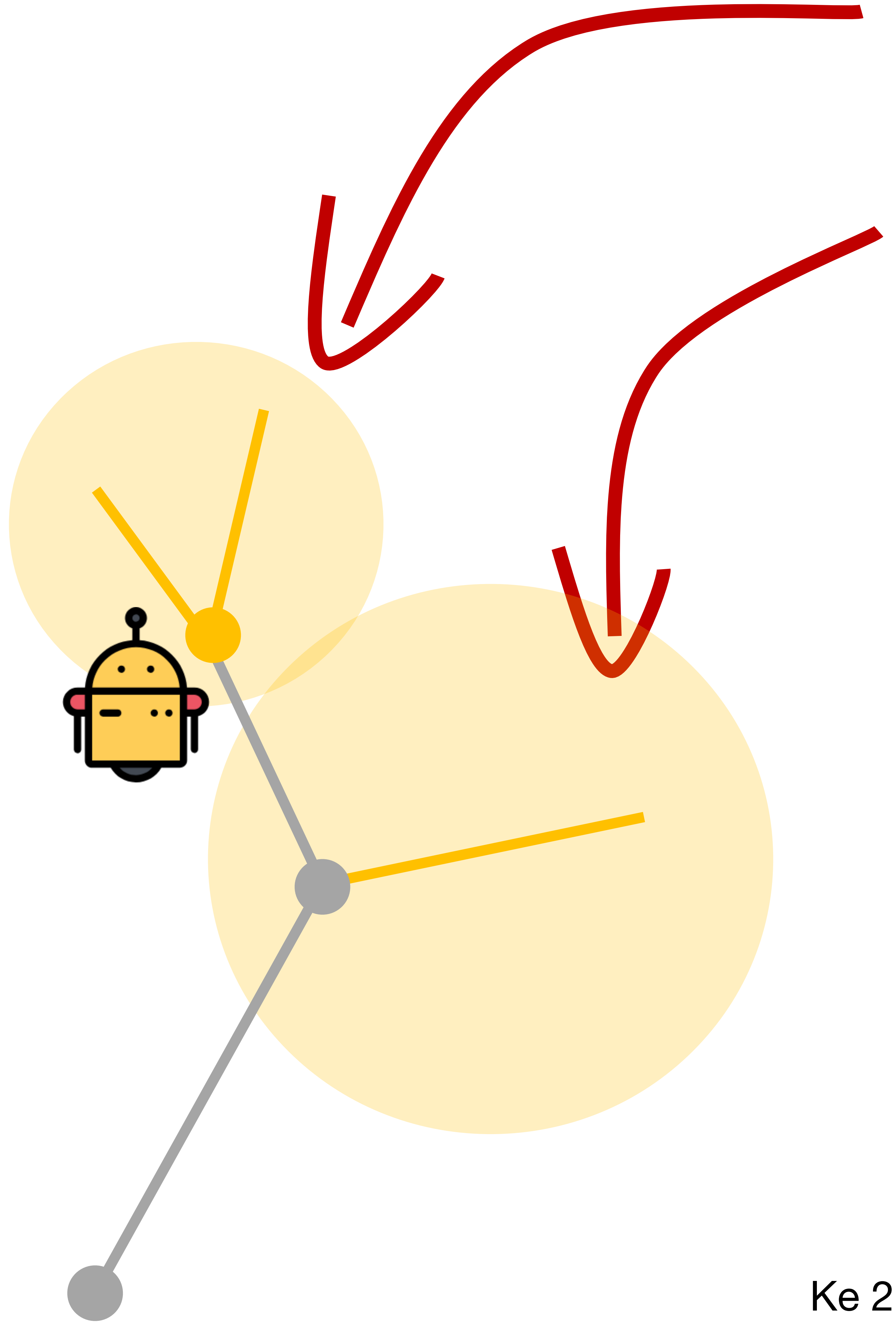
What if you make a mistake?



The New Frontier

Ke 2019, Tactical Rewind: Self-Correction via Backtracking in Vision-and-Language Navigation - CVPR 2019

What if you make a mistake?

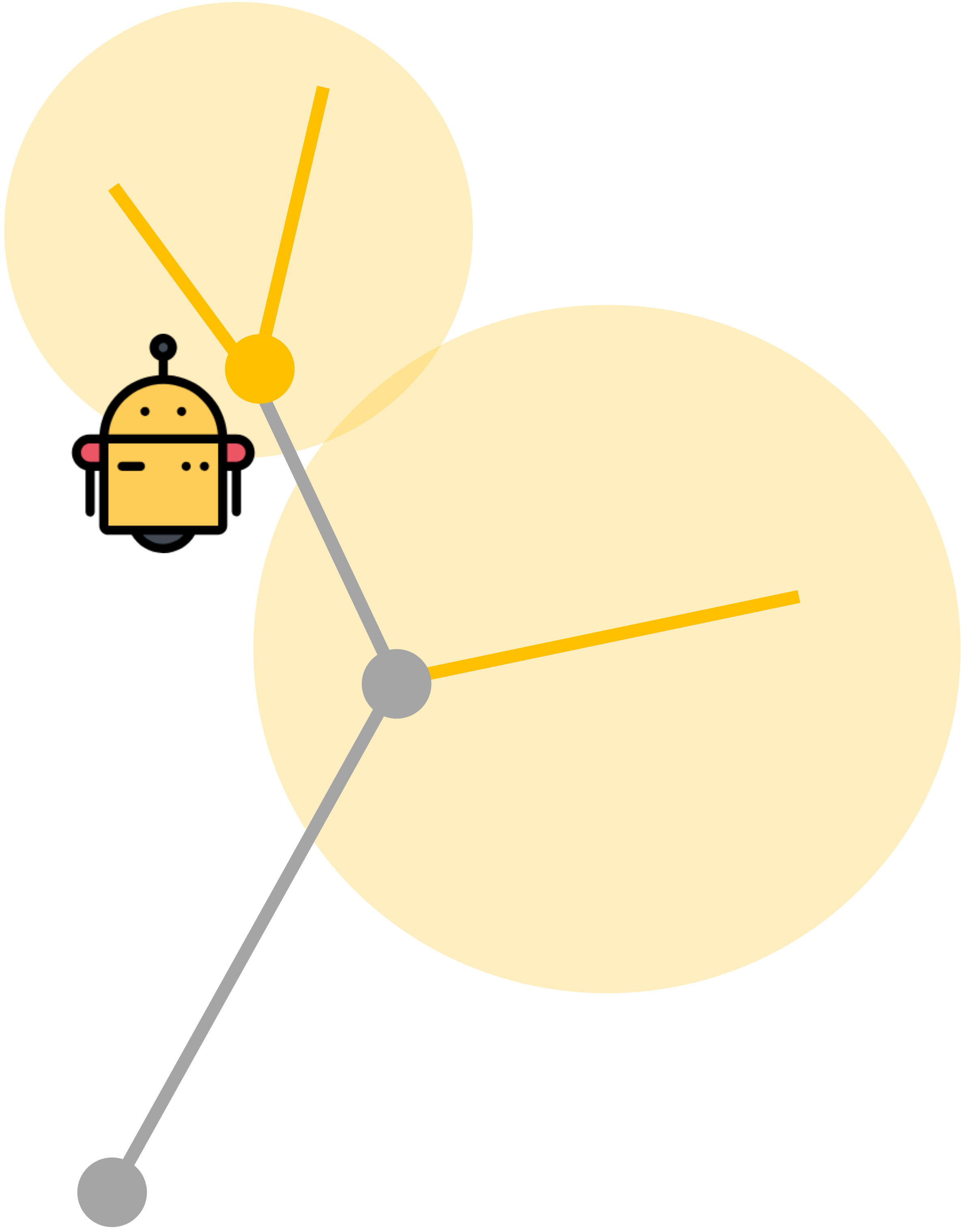


The Frontier

Ke 2019, Tactical Rewind: Self-Correction via Backtracking in Vision-and-Language Navigation - CVPR 2019

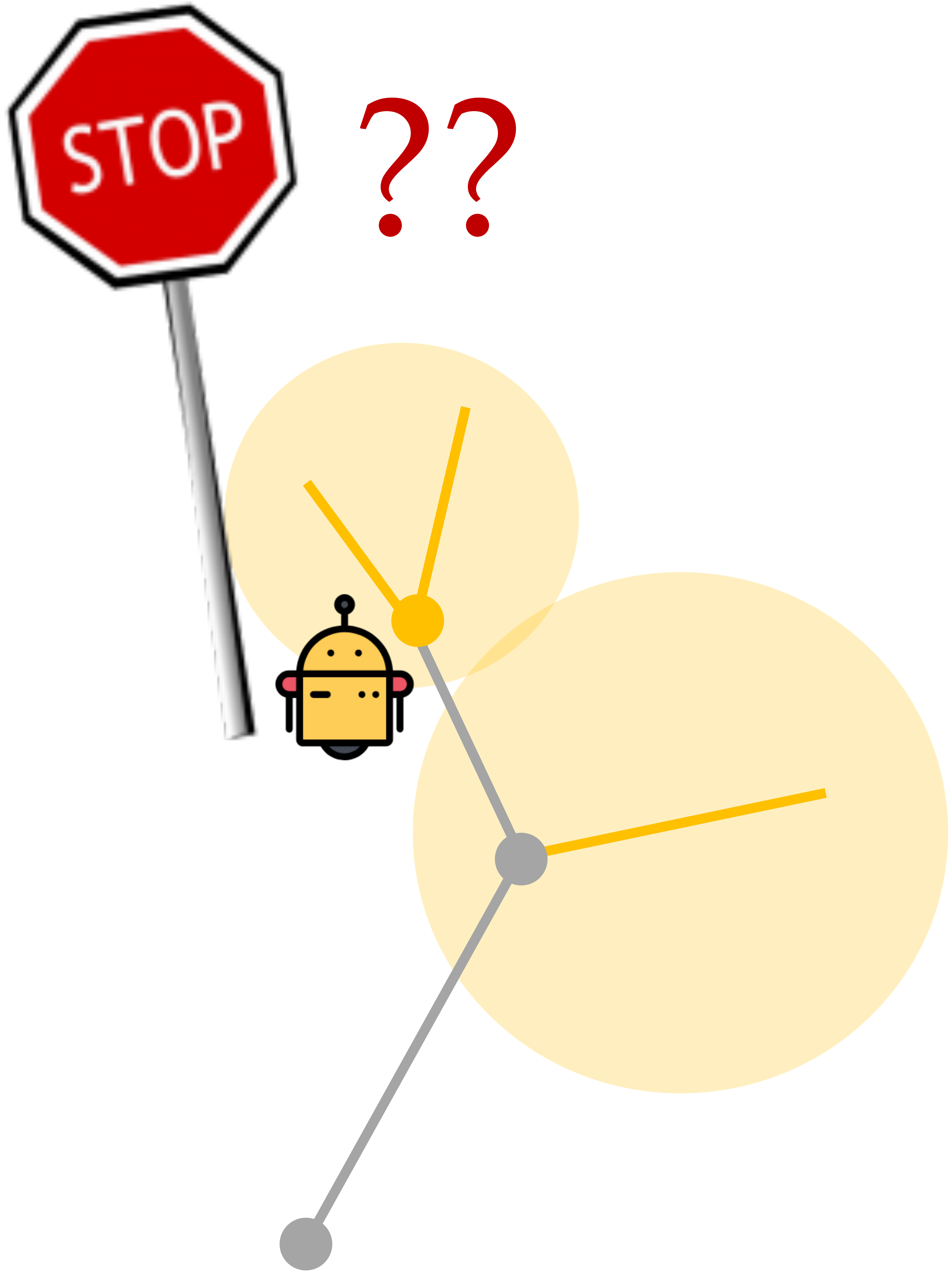
What if you make a mistake?

Eventually …



Ke 2019, Tactical Rewind: Self-Correction via Backtracking in Vision-and-Language Navigation - CVPR 2019

What if you make a mistake?

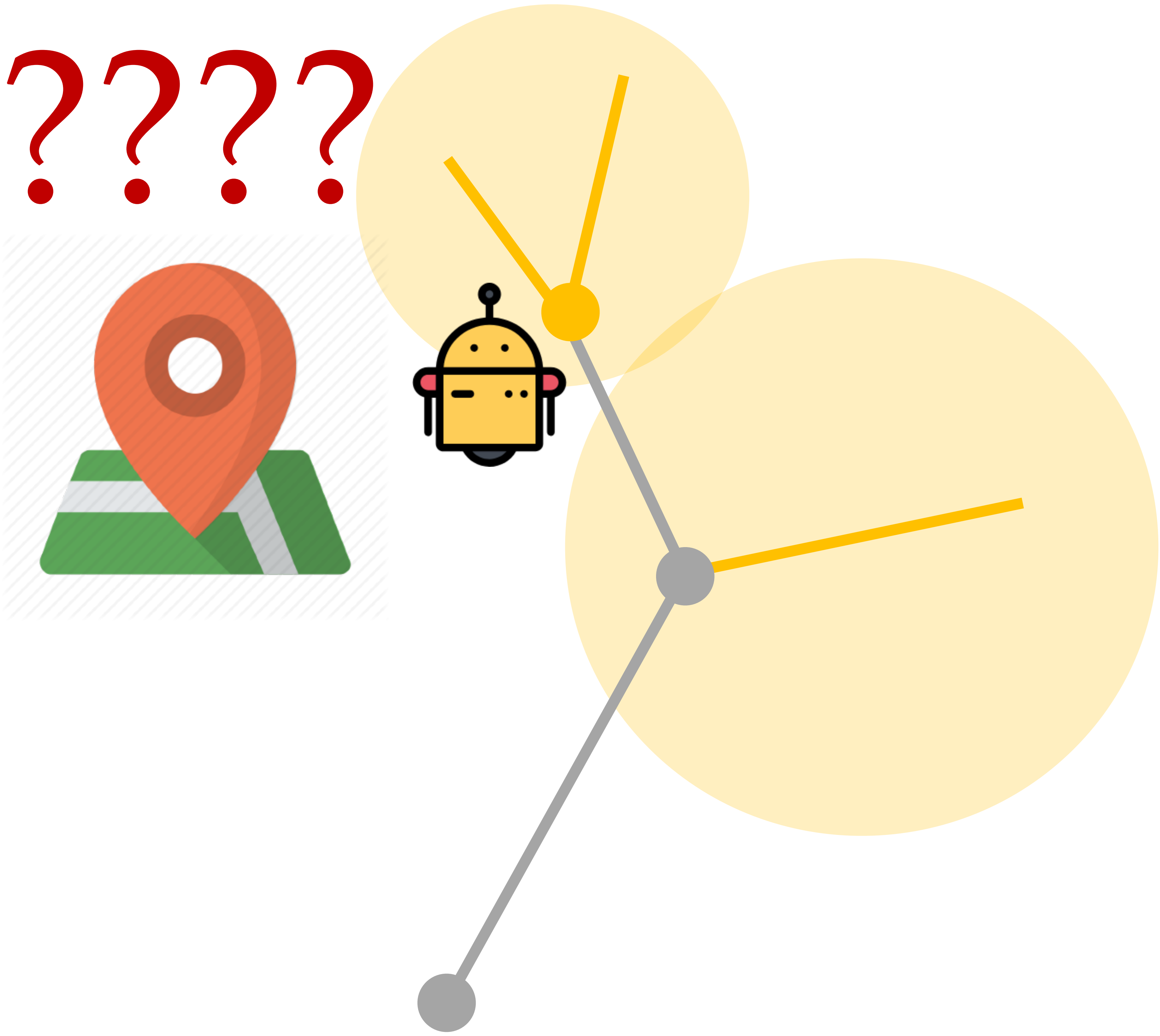


1. **Did I reach the target?**

Ke 2019, Tactical Rewind: Self-Correction via Backtracking in Vision-and-Language Navigation - CVPR 2019

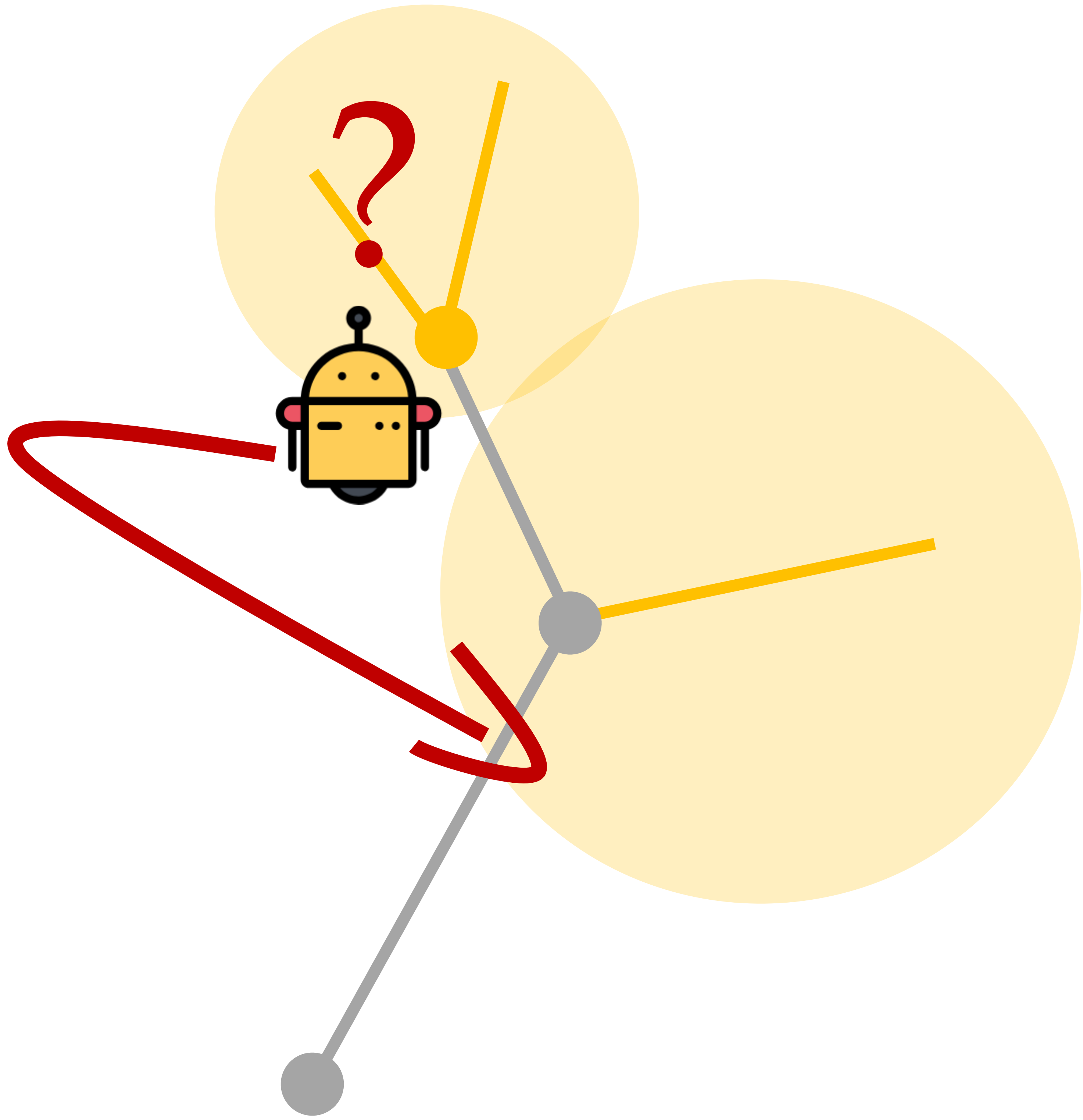
What if you make a mistake?

1. Did I reach the target?
2. Am I lost?



Ke 2019, Tactical Rewind: Self-Correction via Backtracking in Vision-and-Language Navigation - CVPR 2019

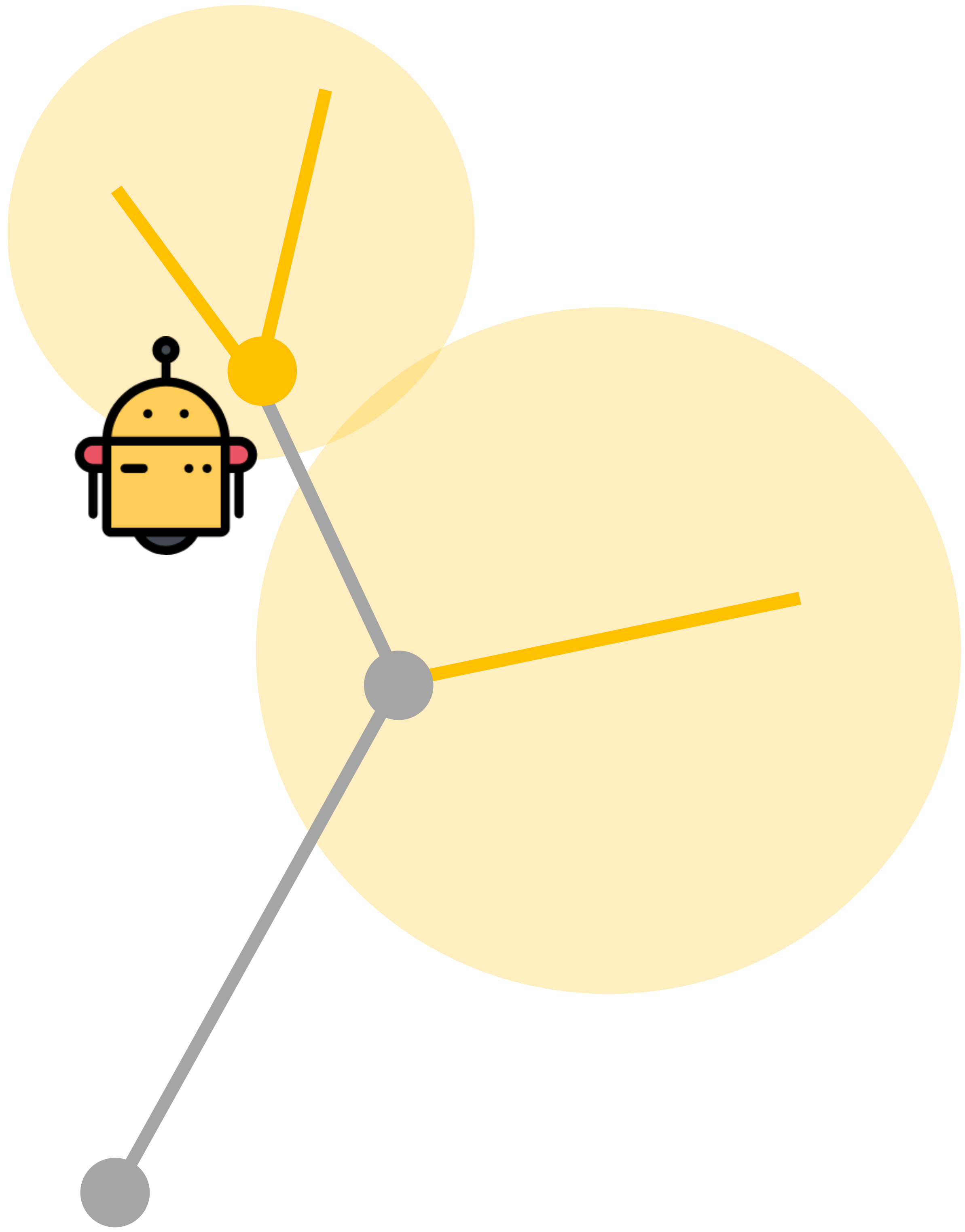
What if you make a mistake?



1. Did I reach the target?
2. Am I lost?
3. **Should I backtrack?**

Ke 2019, Tactical Rewind: Self-Correction via Backtracking in Vision-and-Language Navigation - CVPR 2019

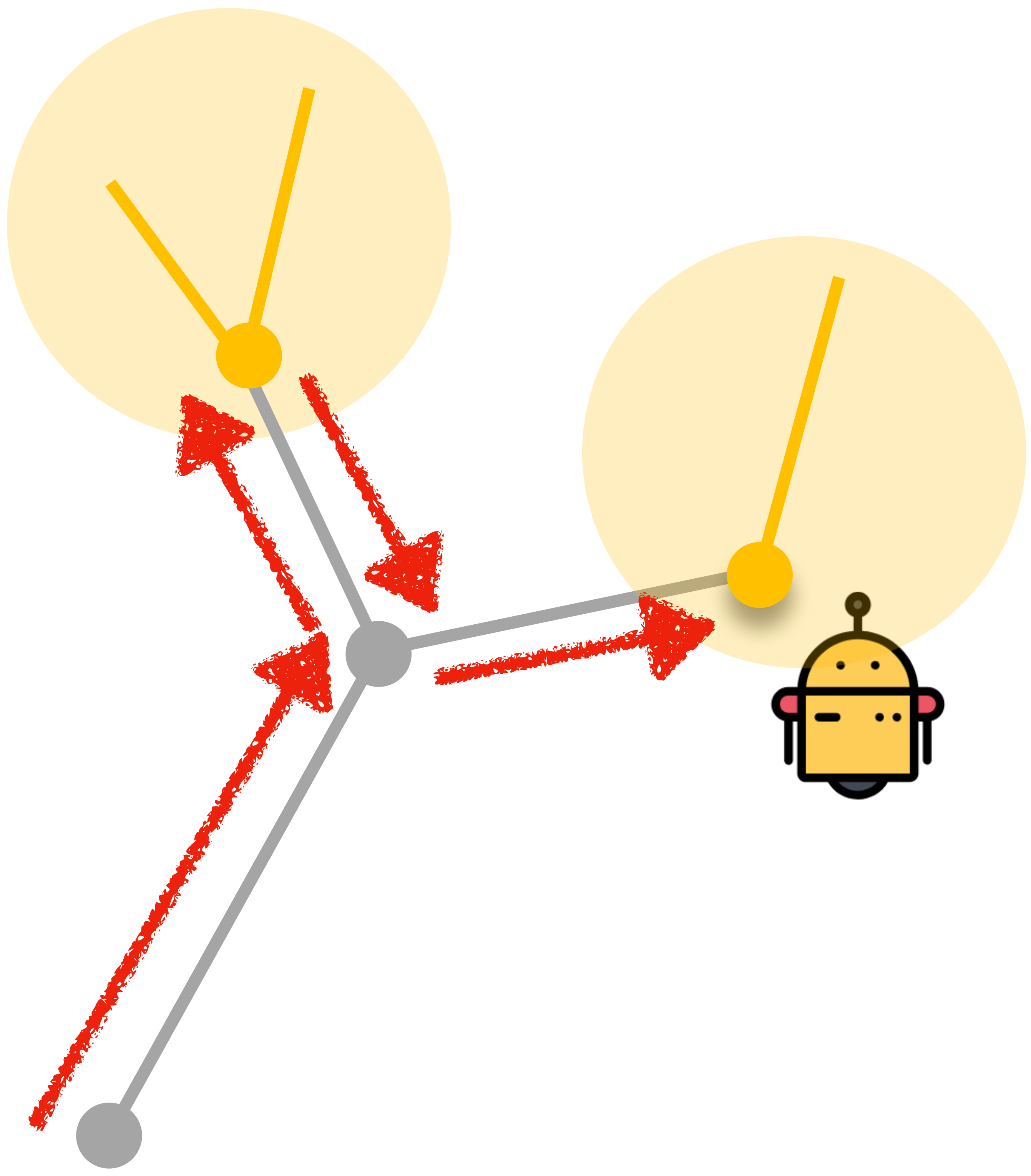
What if you make a mistake?



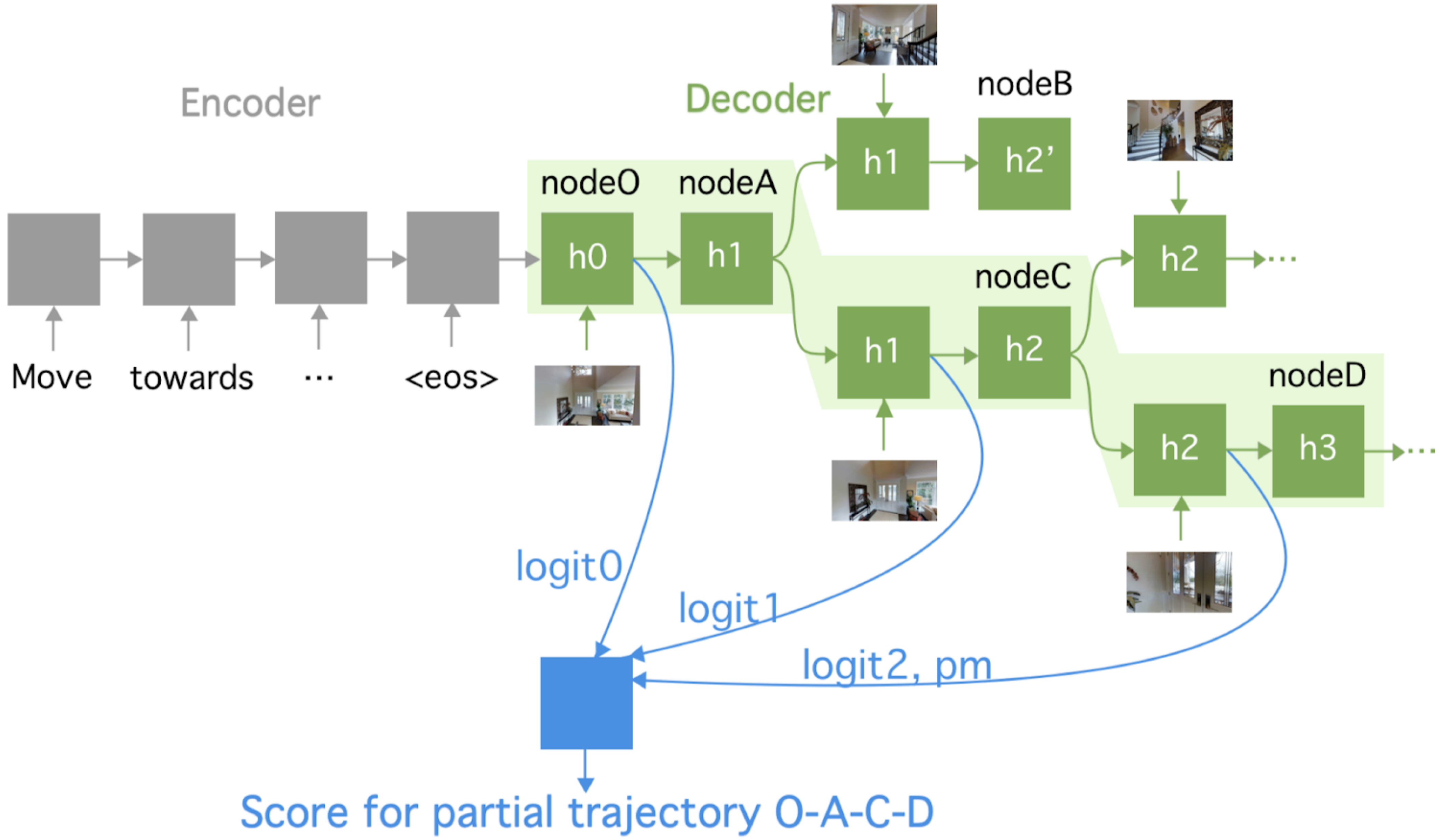
1. Did I reach the target?
2. Am I lost?
3. Should I backtrack?
4. **Where to backtrack to?**

Ke 2019, Tactical Rewind: Self-Correction via Backtracking in Vision-and-Language Navigation - CVPR 2019

What if you make a mistake?

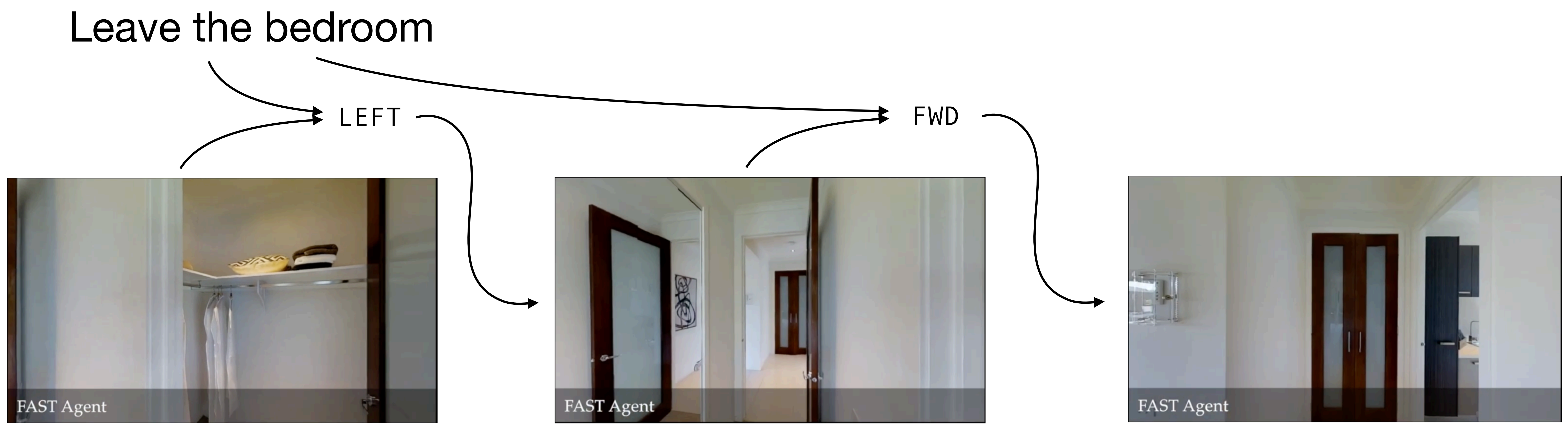


A lot of the visual observations and actions have no correspondence to the language



Ke 2019, Tactical Rewind: Self-Correction via Backtracking in Vision-and-Language Navigation - CVPR 2019

Underspecification



Leave the bedroom

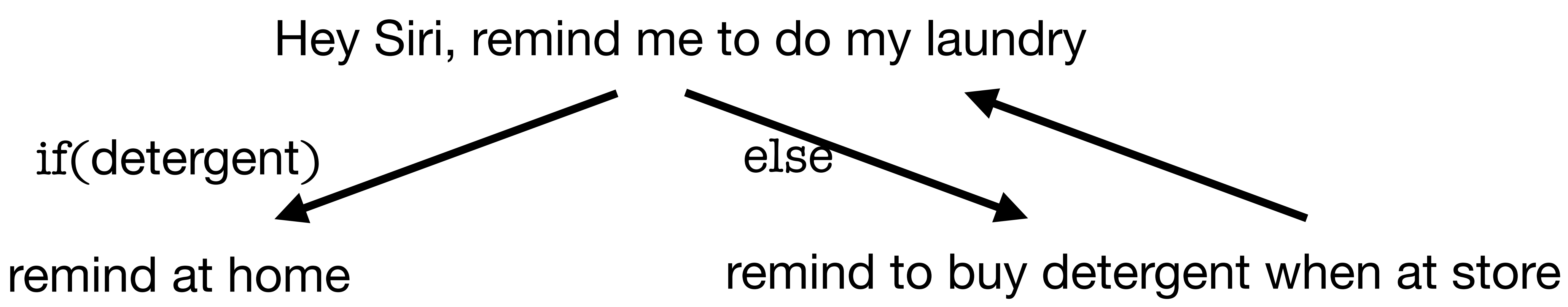
LEFT

FWD

Does this actually need vision?	Yes
Does this understand plans?	Maybe?

Why does this question matter?

Because in general, we can’t supervise everything



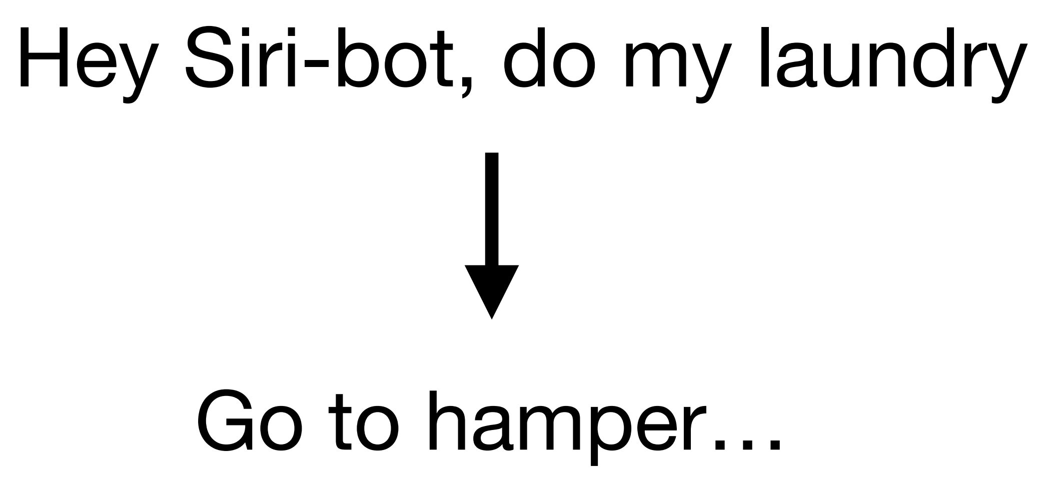
Hey Siri, remind me to do my laundry

if(detergent)

remind at home

else

remind to buy detergent when at store



Hey Siri-bot, do my laundry

Go to hamper…

ALFRED

Action Learning From Realistic Environments and Directives

Goal: "Place a warm slice of bread on the counter"

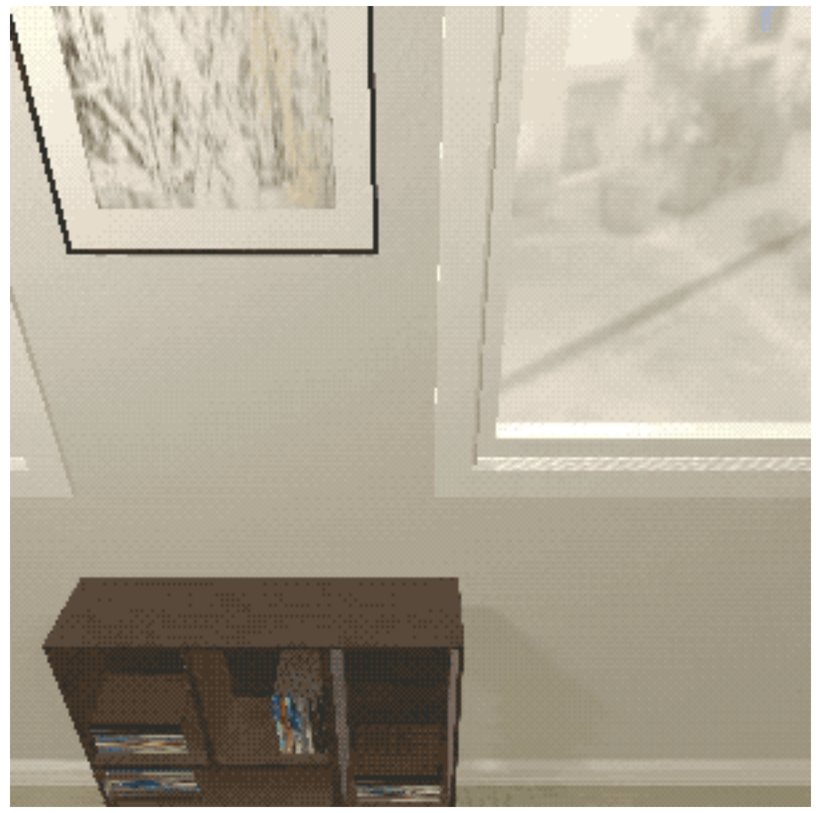
Seven High-level Tasks

Paths are generated by planner

Pick & Place

Double Place

Stack



Examine

Heat

Cool

Rinse

Data collection

Tuple

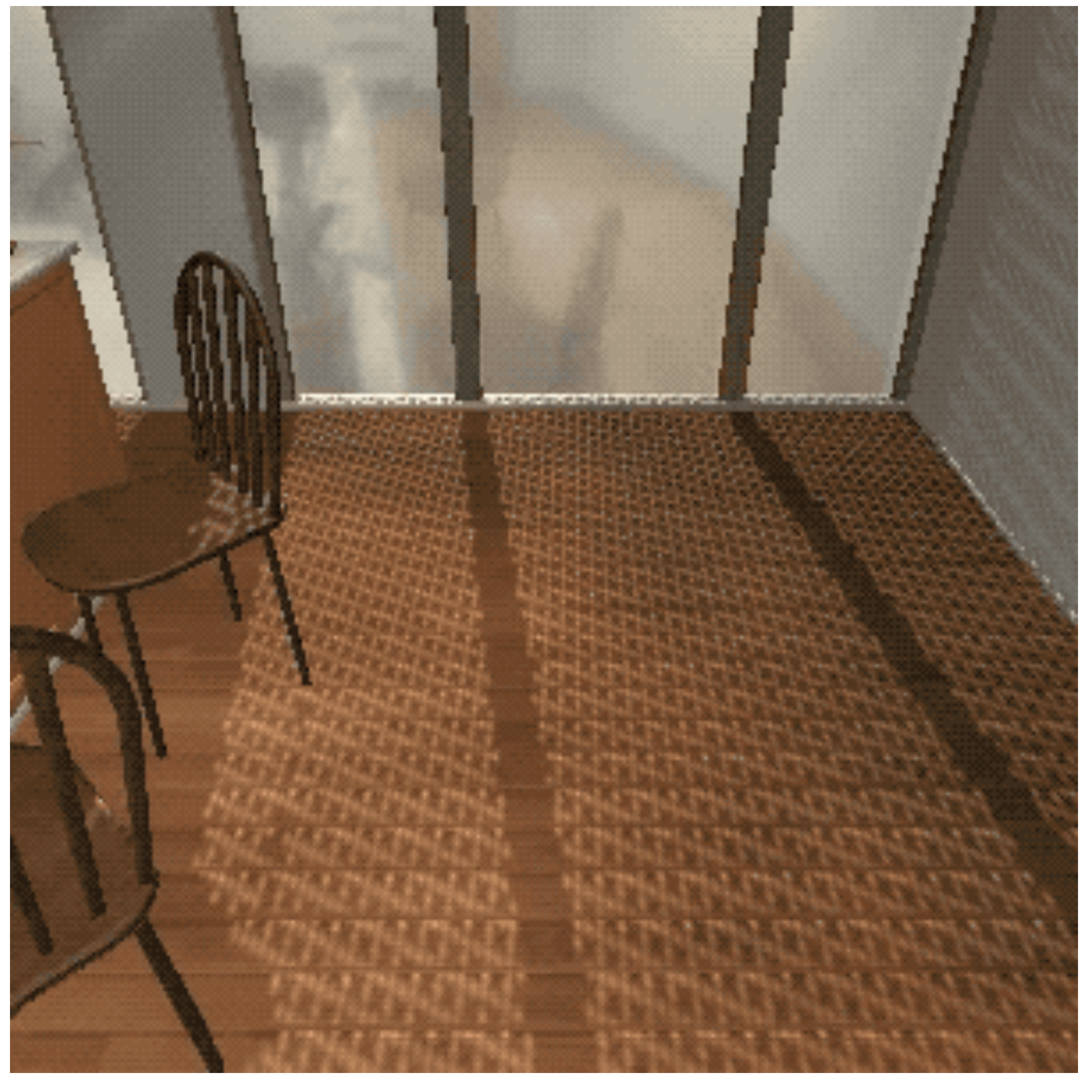
(Stack, Fork, Cup, CounterTop, Kitchen3)

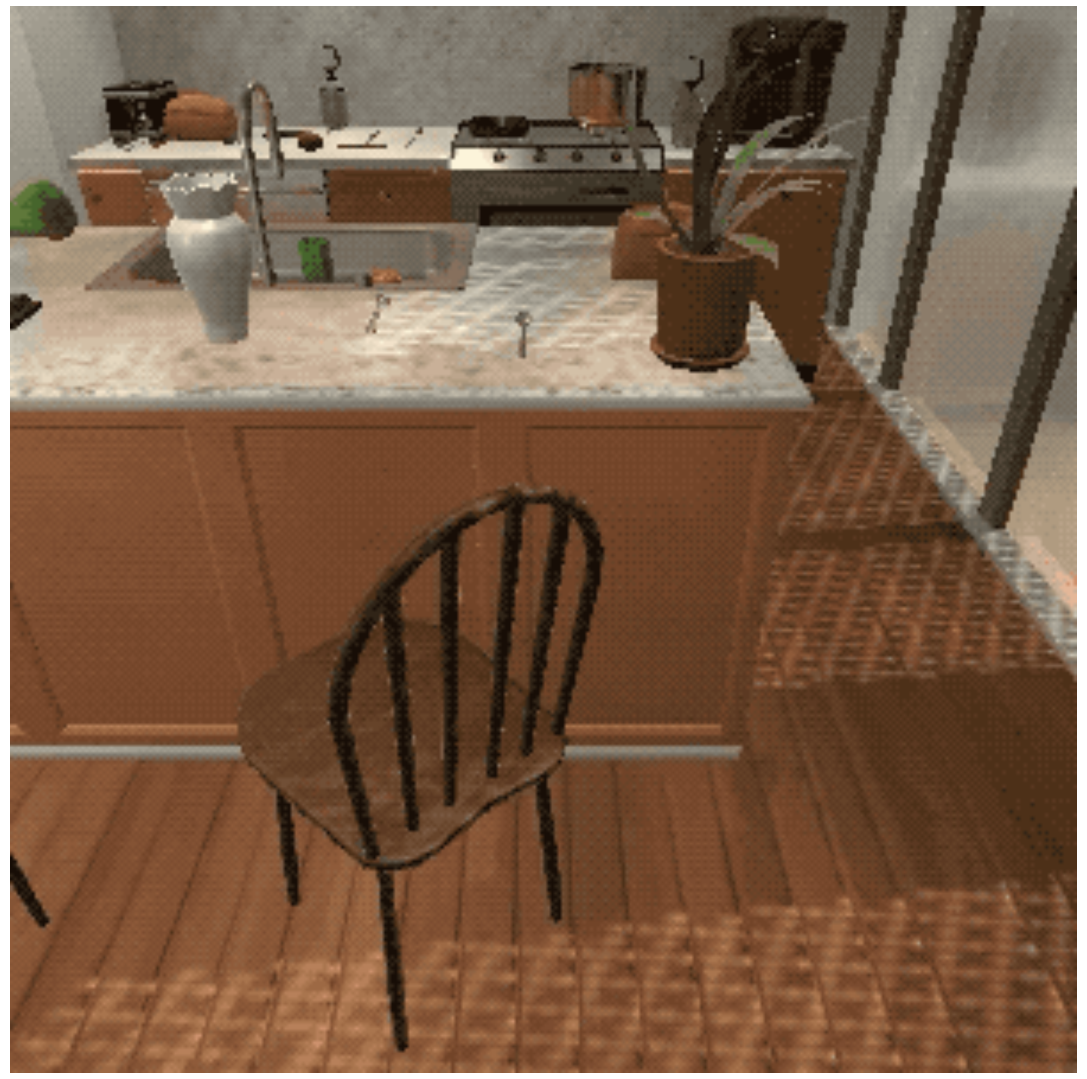
Planner

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

Sample

Execute





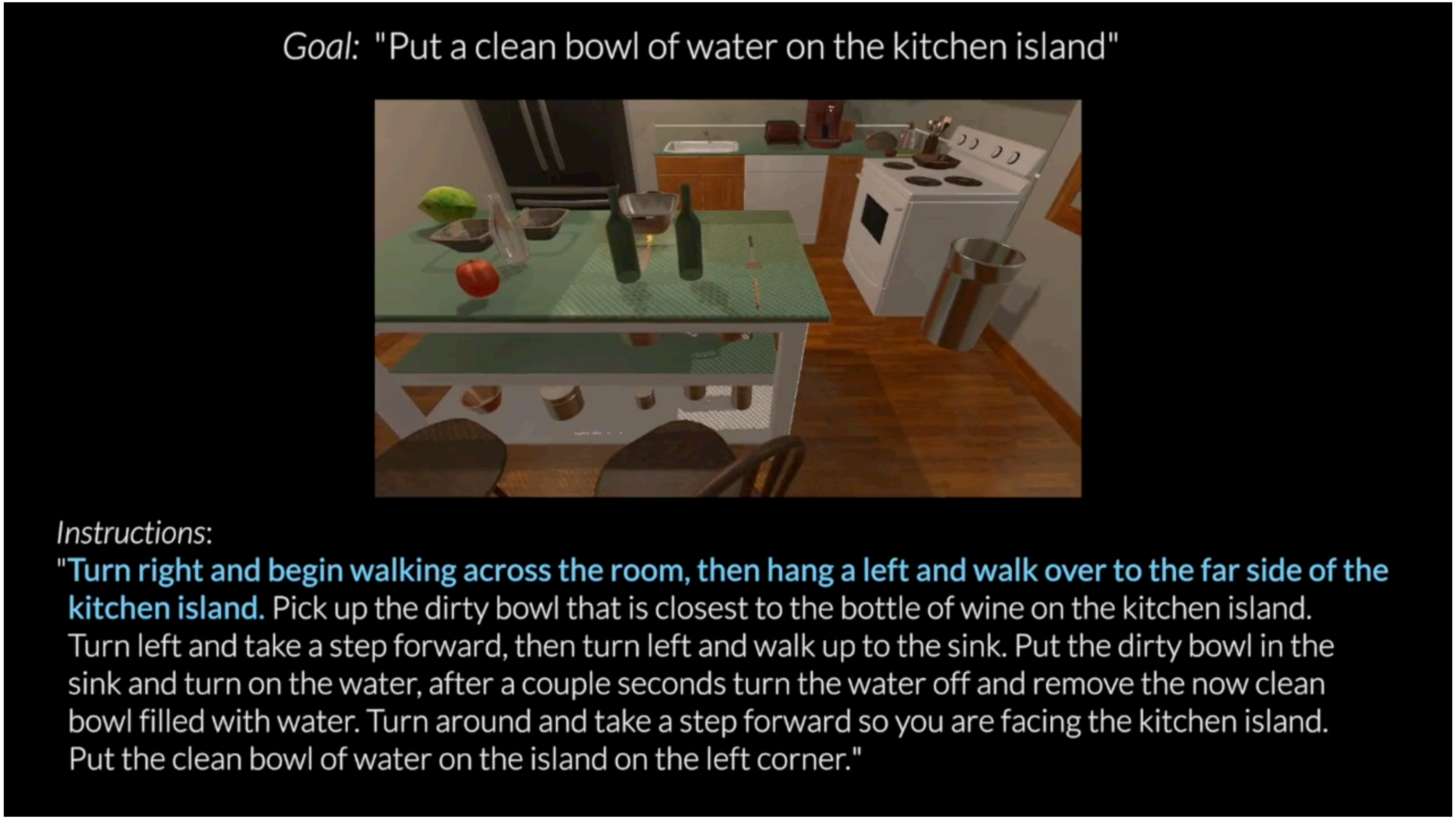
Annotate

Language Language Language

Language Language Language

Language Language Language

Example Language



Goal: "Put a clean bowl of water on the kitchen island"

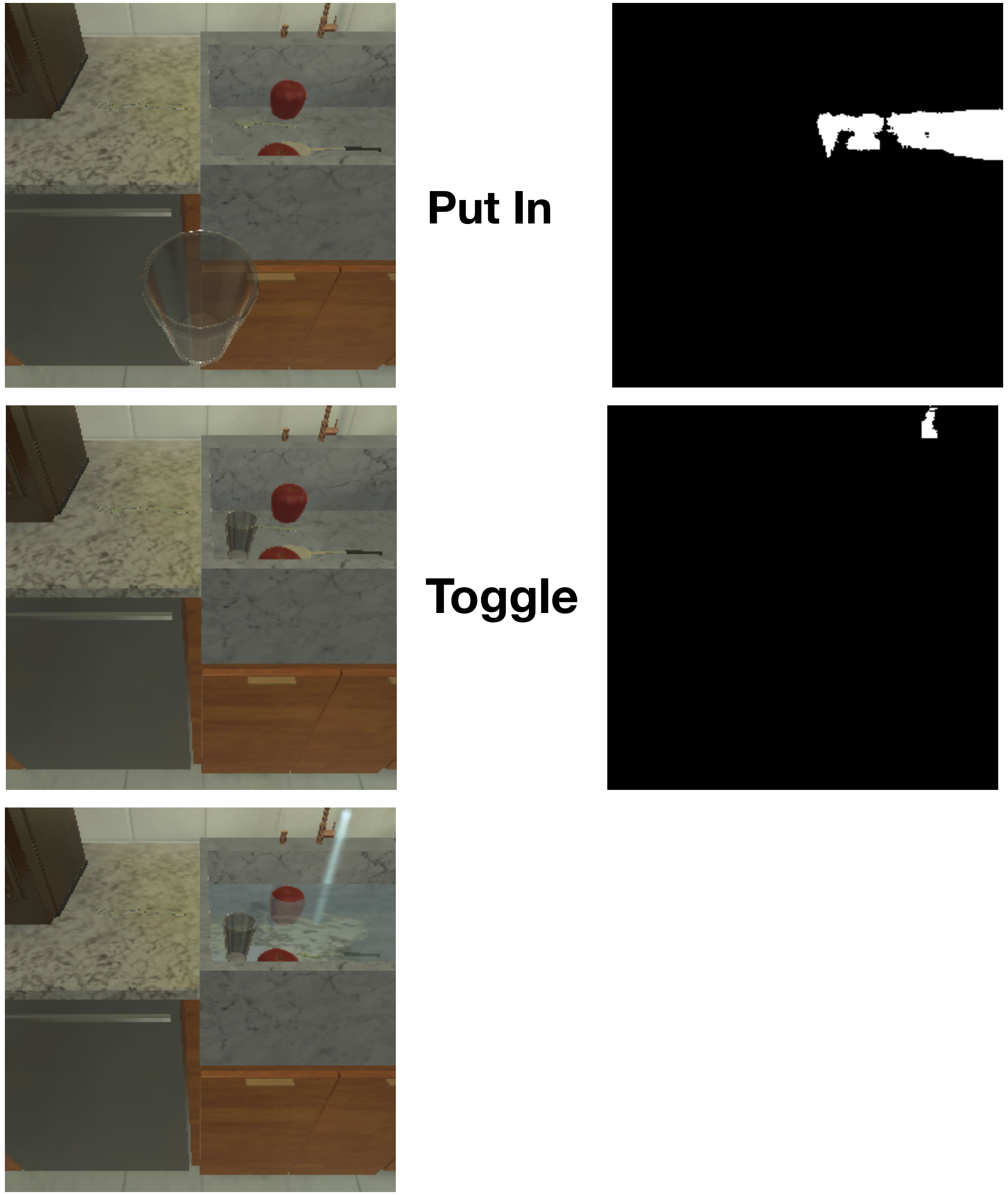
Instructions:

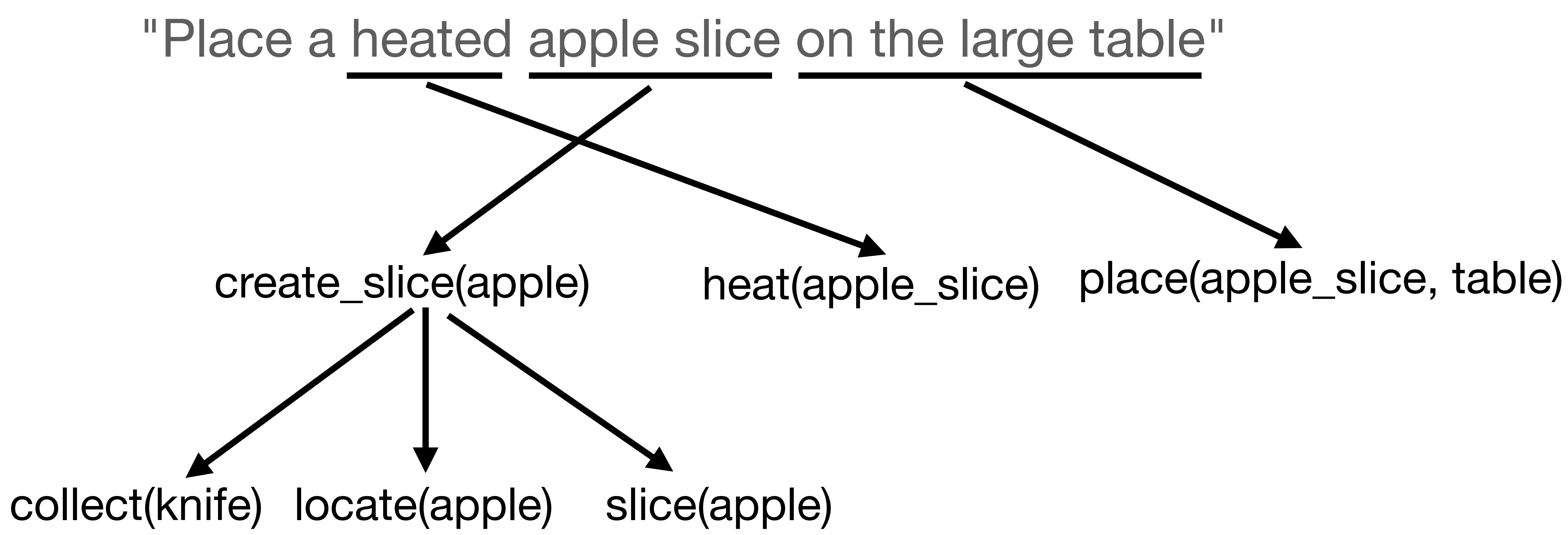
"**Turn right and begin walking across the room, then hang a left and walk over to the far side of the kitchen island.** Pick up the dirty bowl that is closest to the bottle of wine on the kitchen island. Turn left and take a step forward, then turn left and walk up to the sink. Put the dirty bowl in the sink and turn on the water, after a couple seconds turn the water off and remove the now clean bowl filled with water. Turn around and take a step forward so you are facing the kitchen island. Put the clean bowl of water on the island on the left corner."

Action Space

Wash the cup

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





"Place a heated apple slice on the large table"

create_slice(apple)

heat(apple_slice)

place(apple_slice, table)

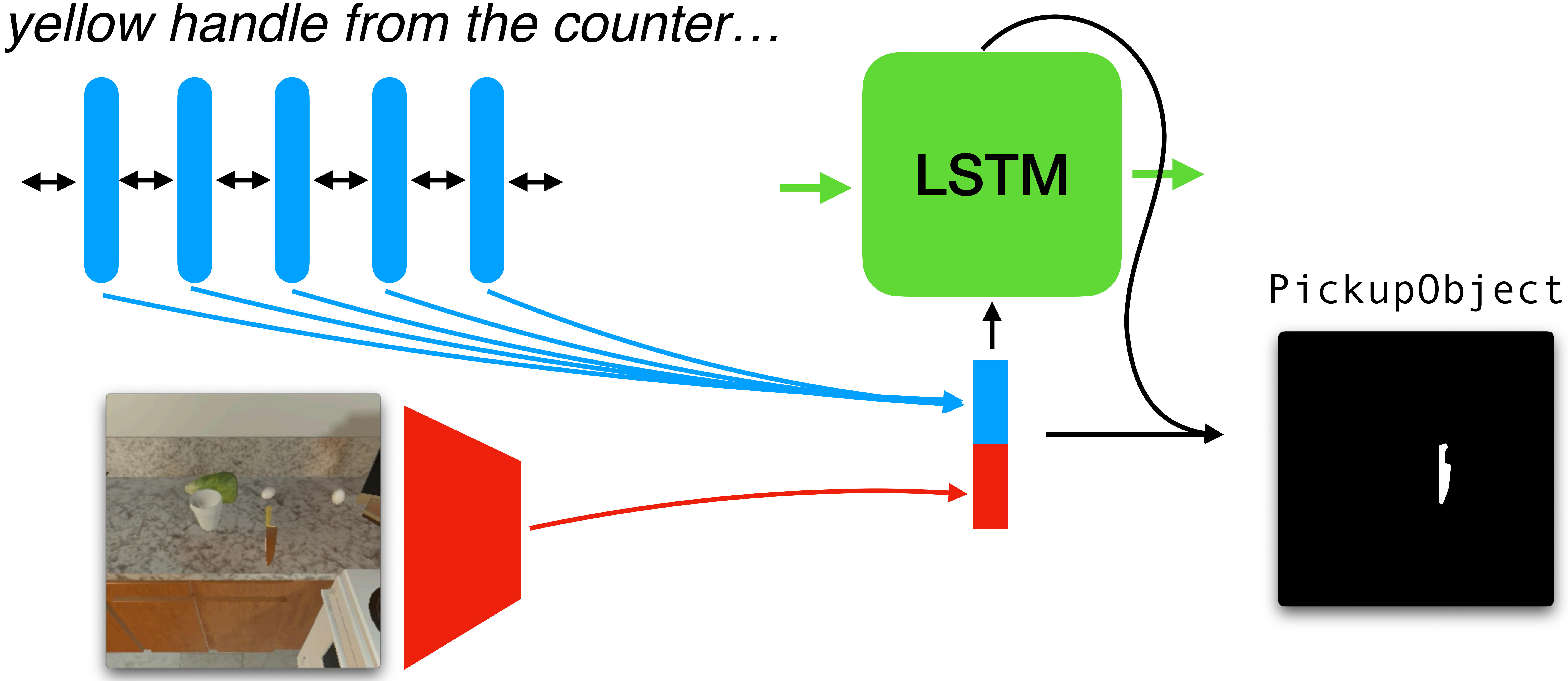
collect(knife)

locate(apple)

slice(apple)

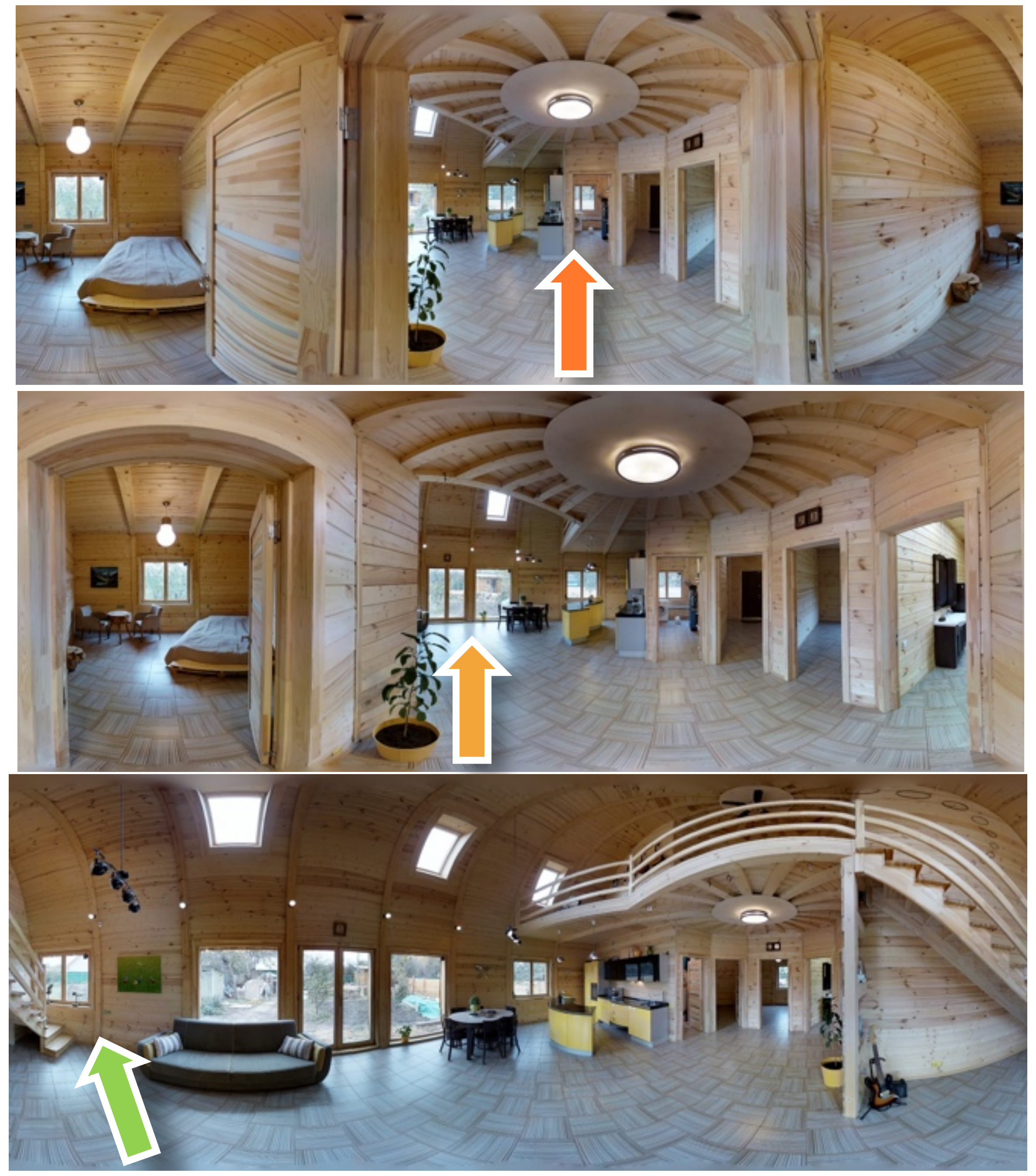
End-to-End Models

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…



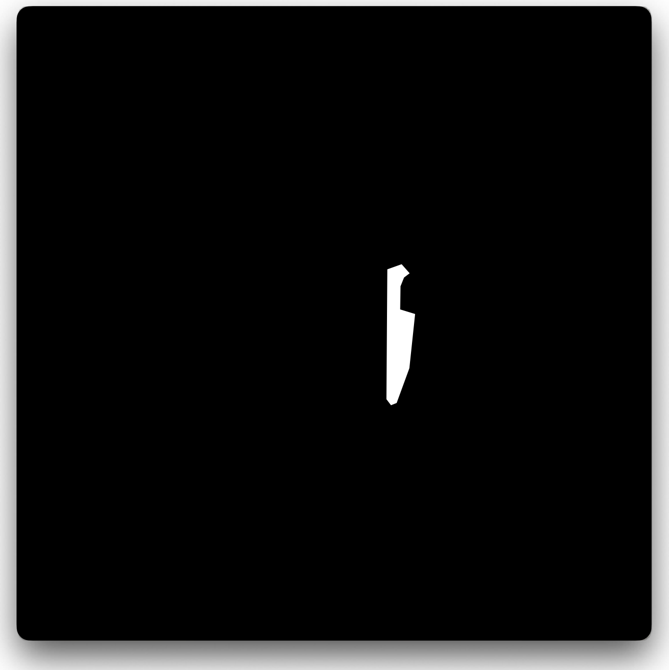
Action Spaces

Choose a view

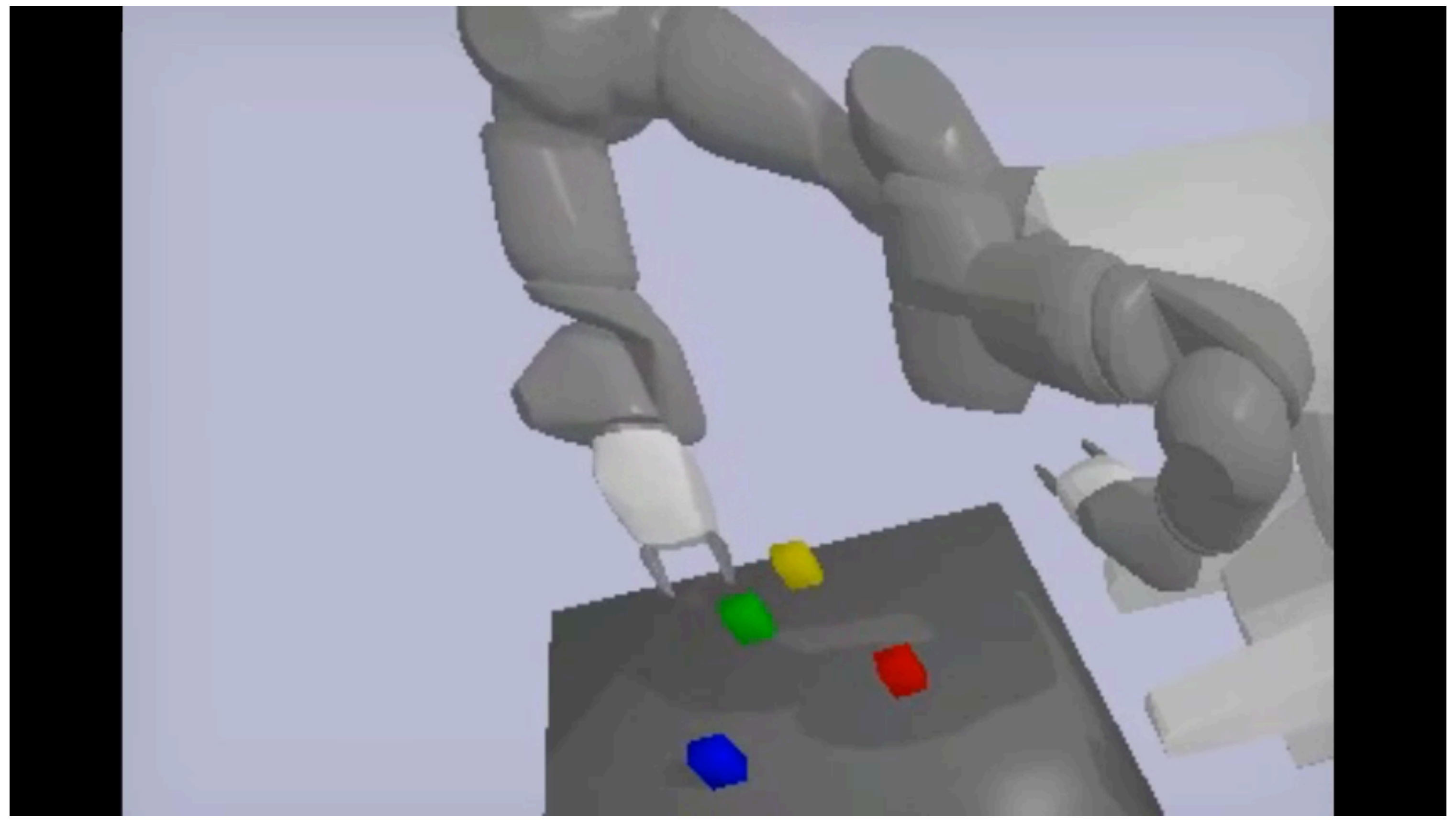


Outline an Object

PickupObject



Grasp an Object



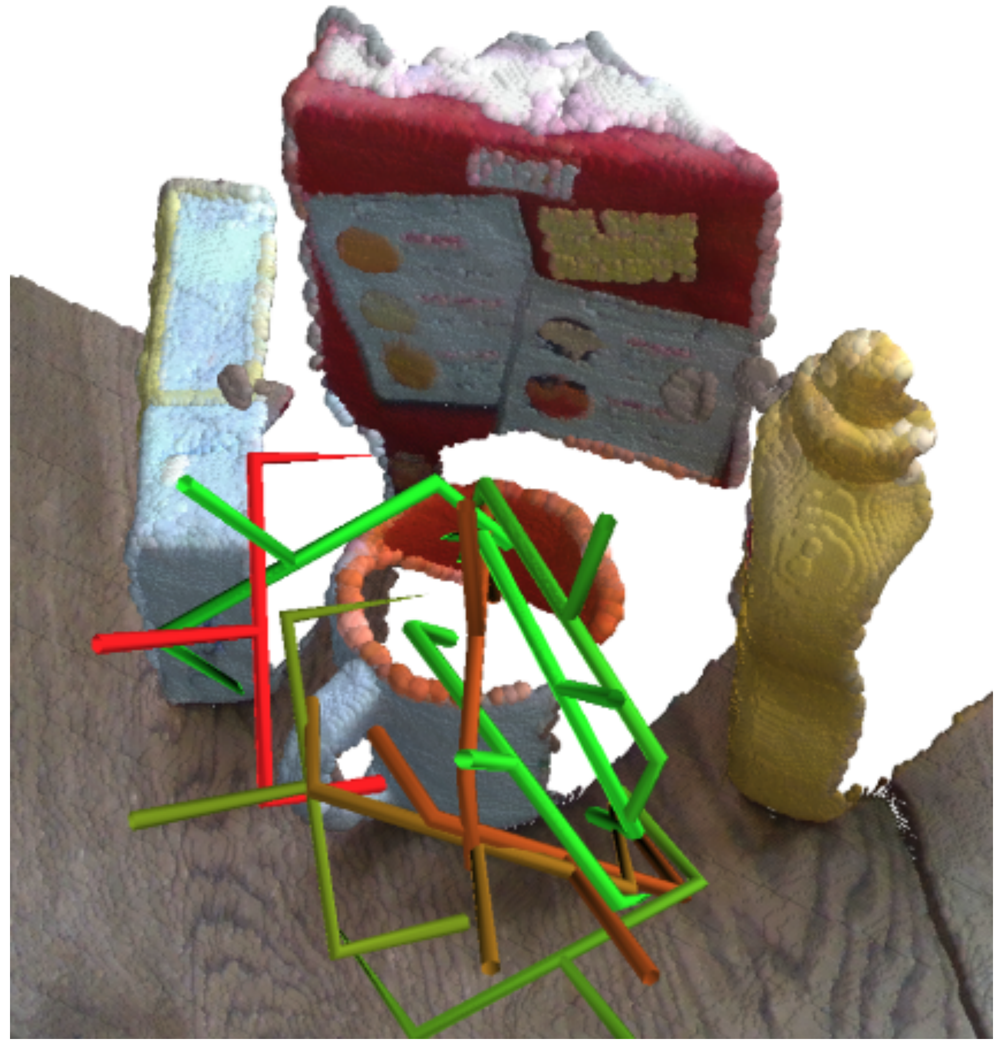
Pick-up

What's hidden in that?

If I gave you one of these and labeled it, could you abstract to the others?

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

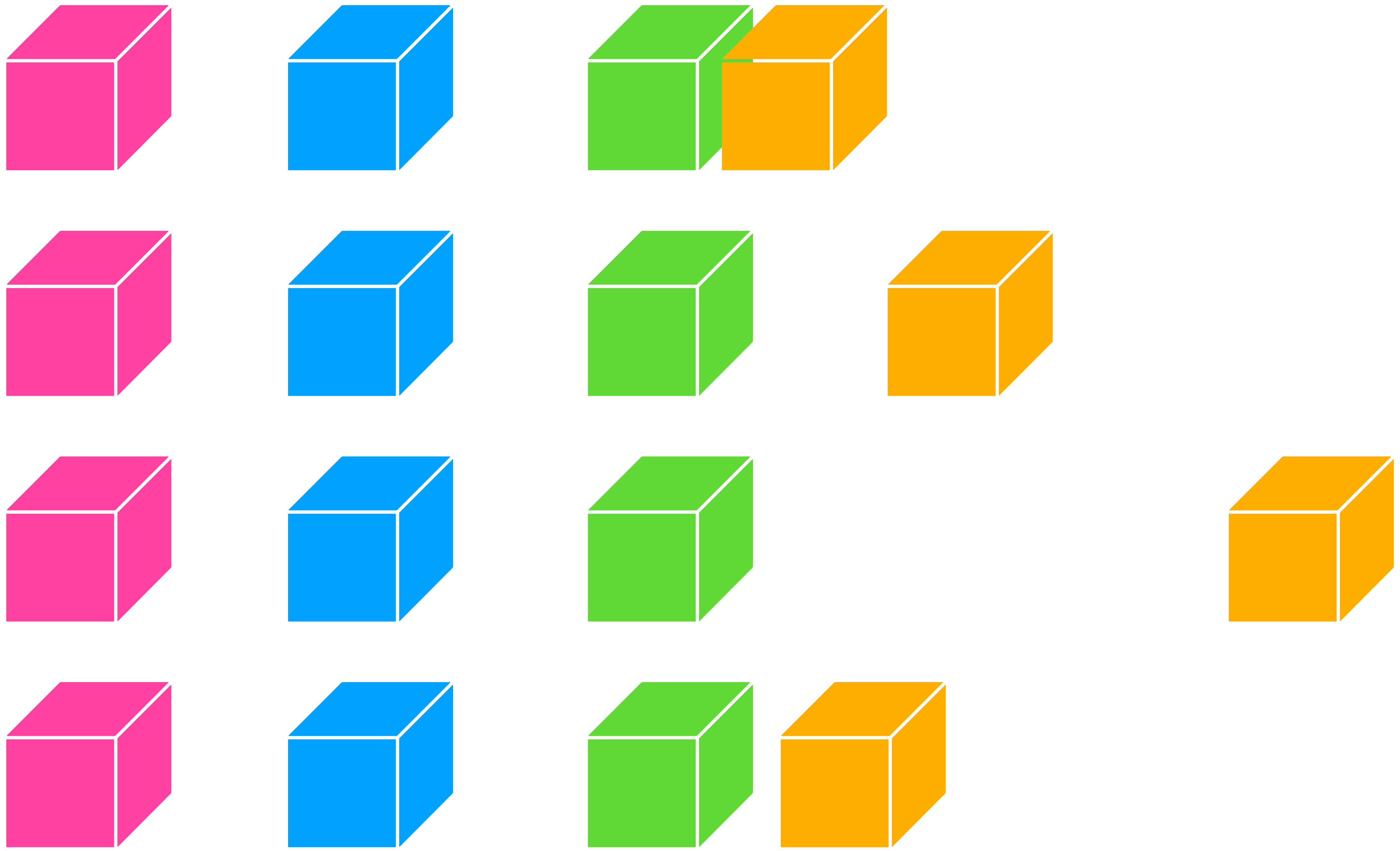
Does "pick up" correspond to a specific action sequence?



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

Simplify with Blocks and Coordinates

Put the orange block to the right of the green block



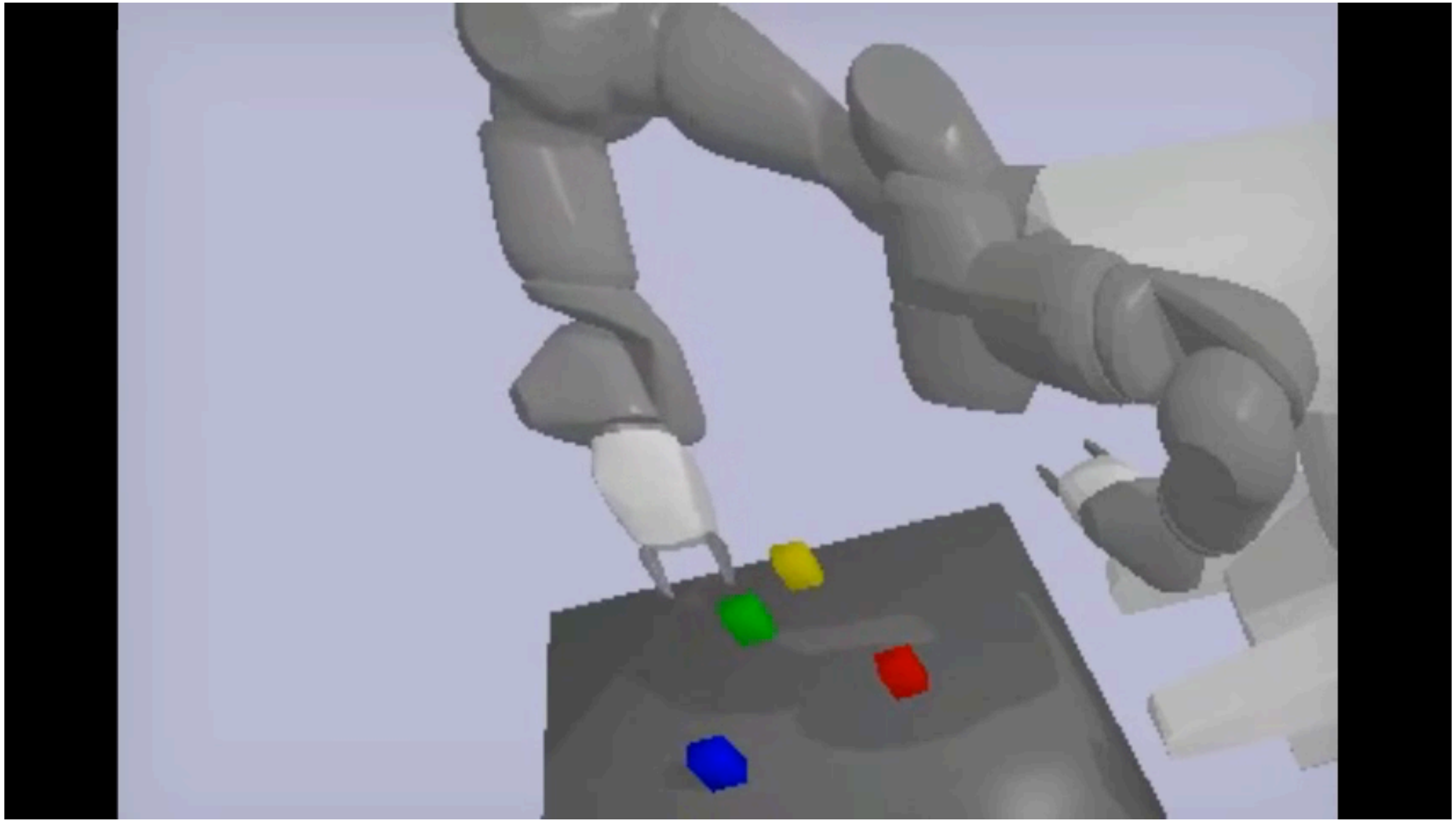
Why?

Is this a useful training datum?

("Put the orange block to right of the green block", 0.35)

We no longer have a discrete grounding

Simple Blocks

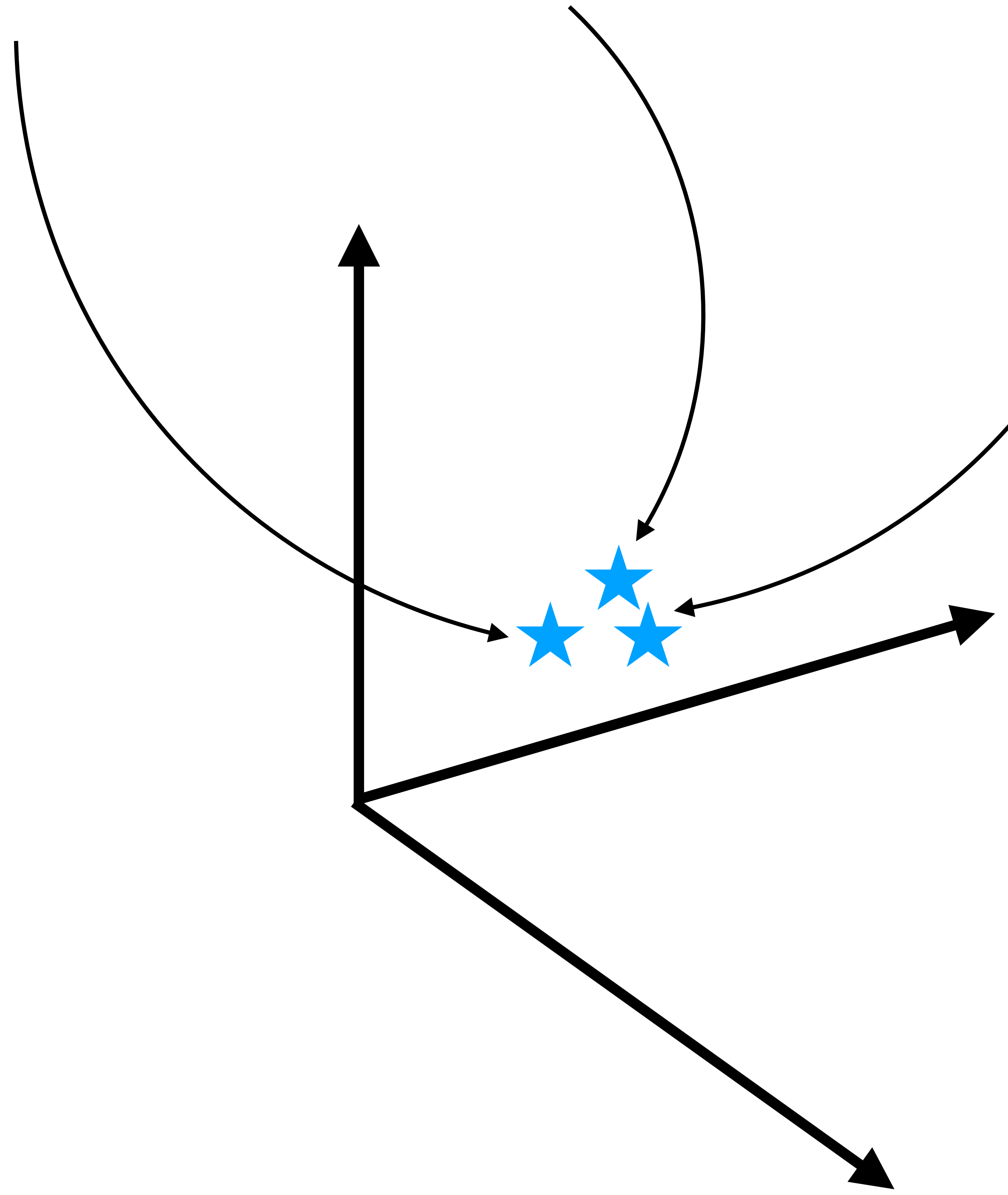


A Shared Semantic Space

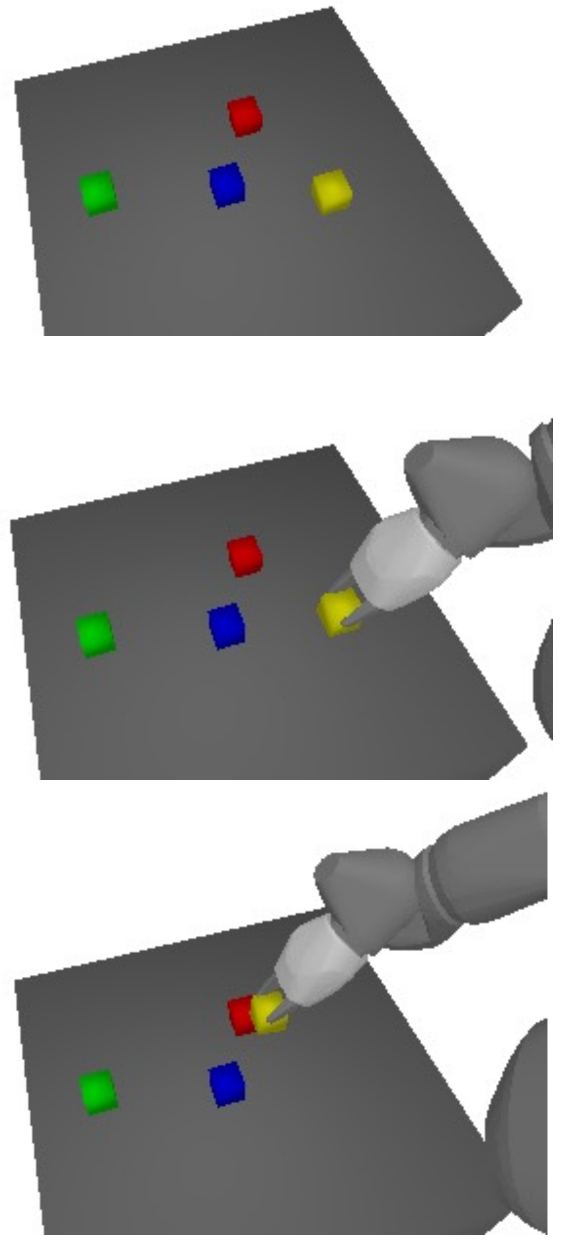
Language

“take the yellow object from the table and place it on top of the red object”

`move_to(yellow)` `grasp(yellow)` … `release(yellow)`



Observations



Paxton et al. Prospection: Interpretable Plans From Language By Predicting the Future ICRA 2019

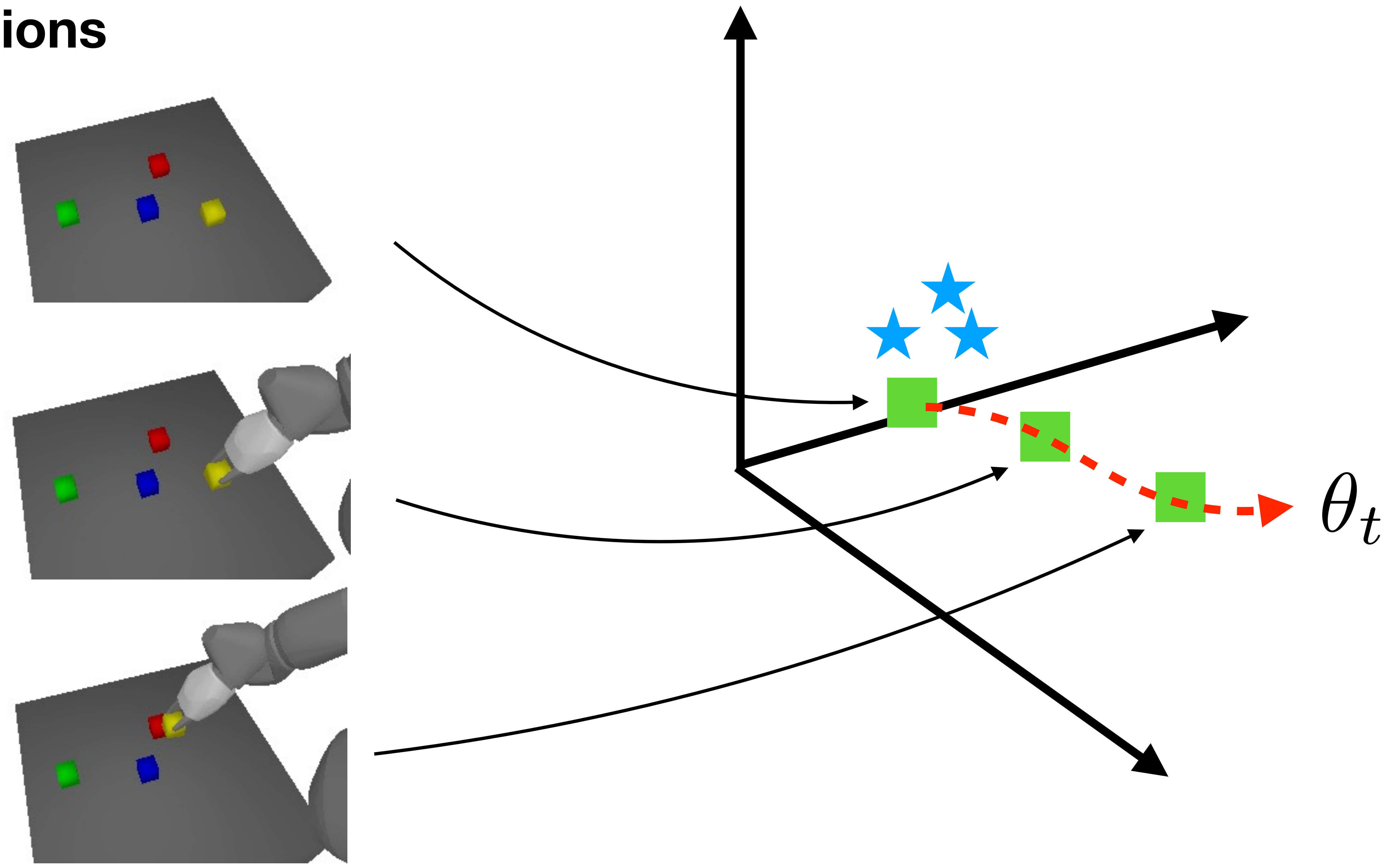
A Shared Semantic Space

Language

“take the yellow object from the table and place it on top of the red object”

```
move_to(yellow)  grasp(yellow)  … release(yellow)
```

Observations



7 DOF Joint Position

Paxton et al. Prospection: Interpretable Plans From Language By Predicting the Future ICRA 2019

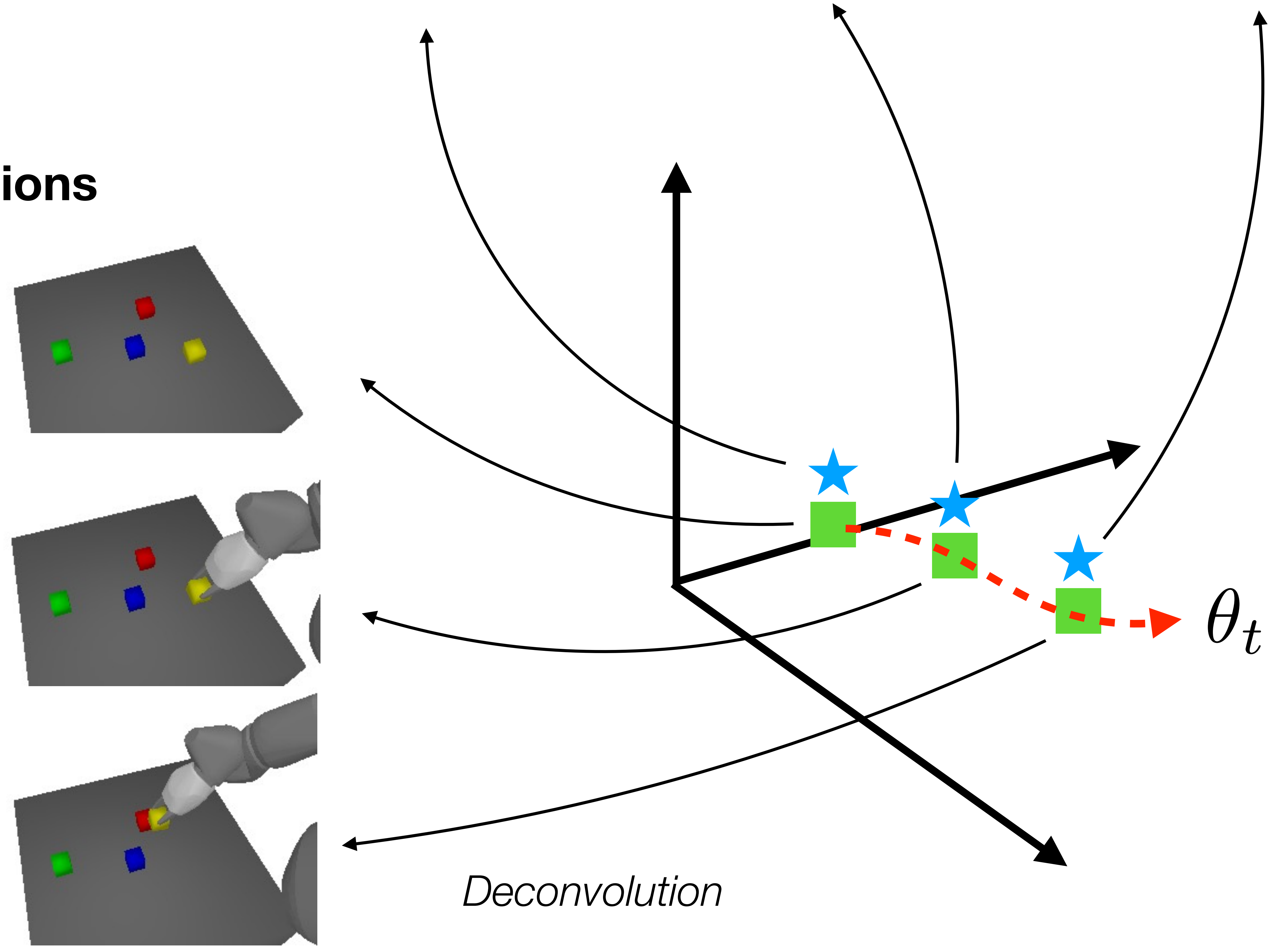
A Shared Semantic Space

Language

“take the yellow object from the table and place it on top of the red object”

`move_to(yellow)  grasp(yellow)  … release(yellow)`

Observations



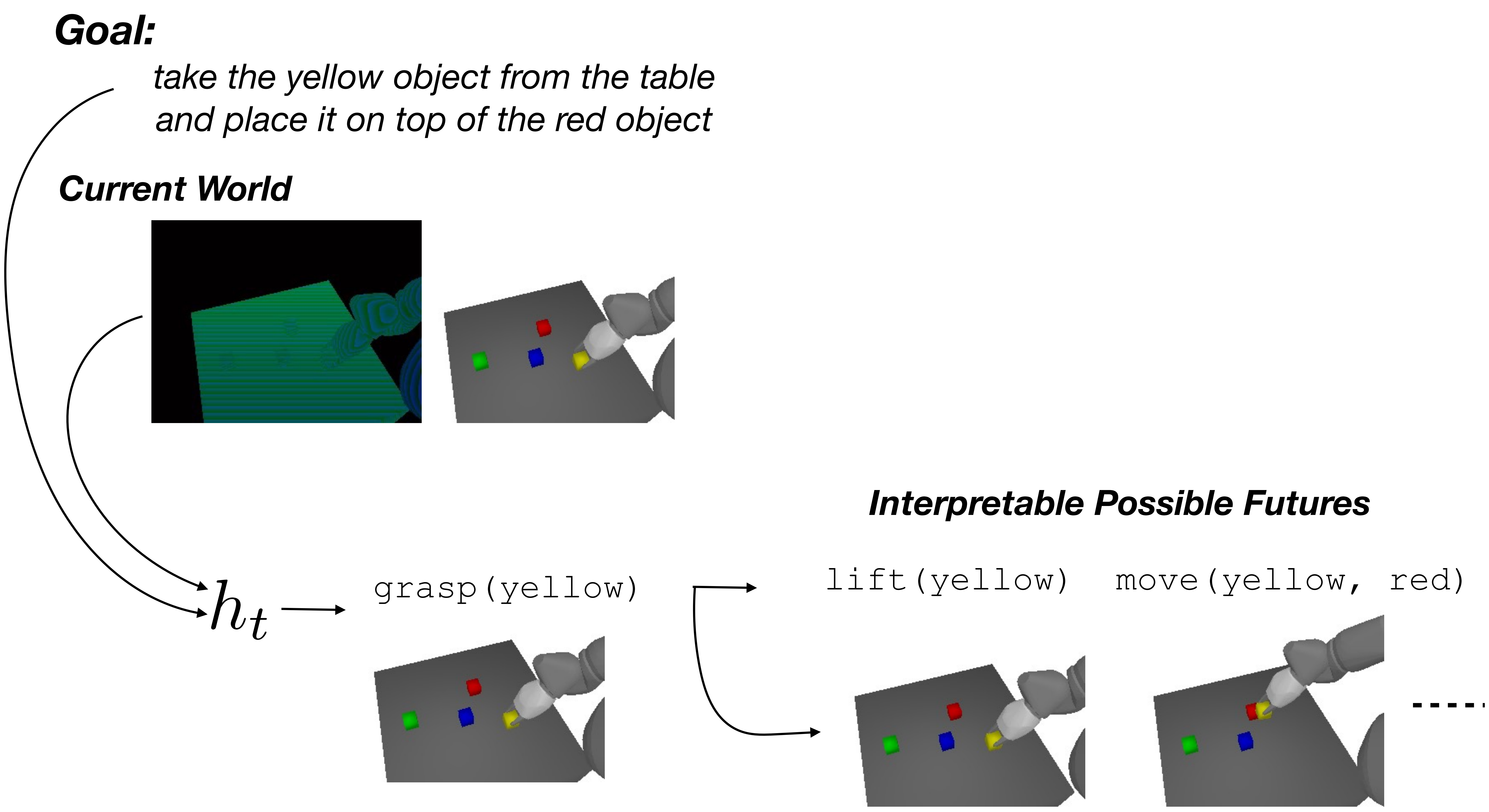
Generation

θ_t 7 DOF Joint Position

Deconvolution

Paxton et al. Prospection: Interpretable Plans From Language By Predicting the Future ICRA 2019

Predicting the Future



Goal:

take the yellow object from the table and place it on top of the red object

Current World

h_t

`grasp(yellow)`

Interpretable Possible Futures

`lift(yellow)` `move(yellow, red)`

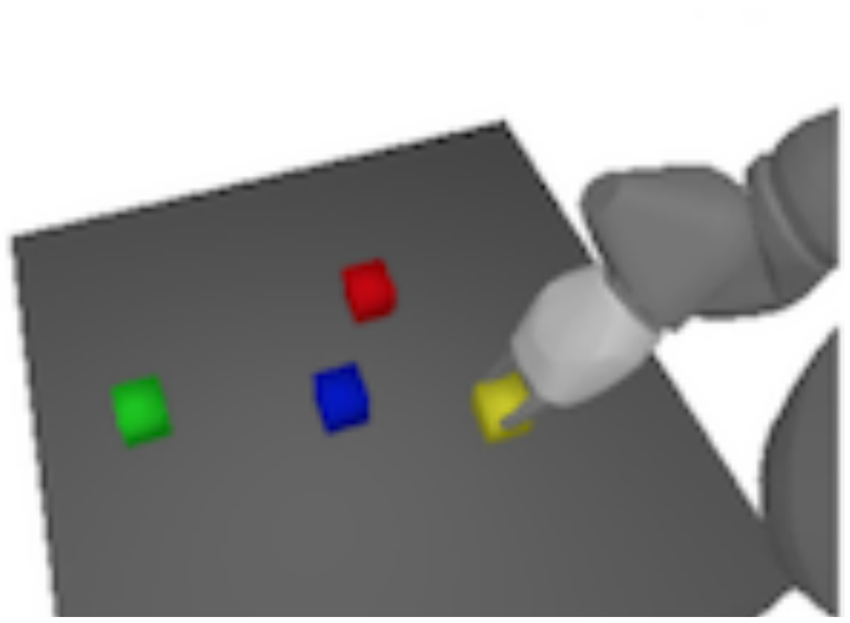
Paxton et al. Prospection: Interpretable Plans From Language By Predicting the Future ICRA 2019

Objectives

Latent Space Z_t

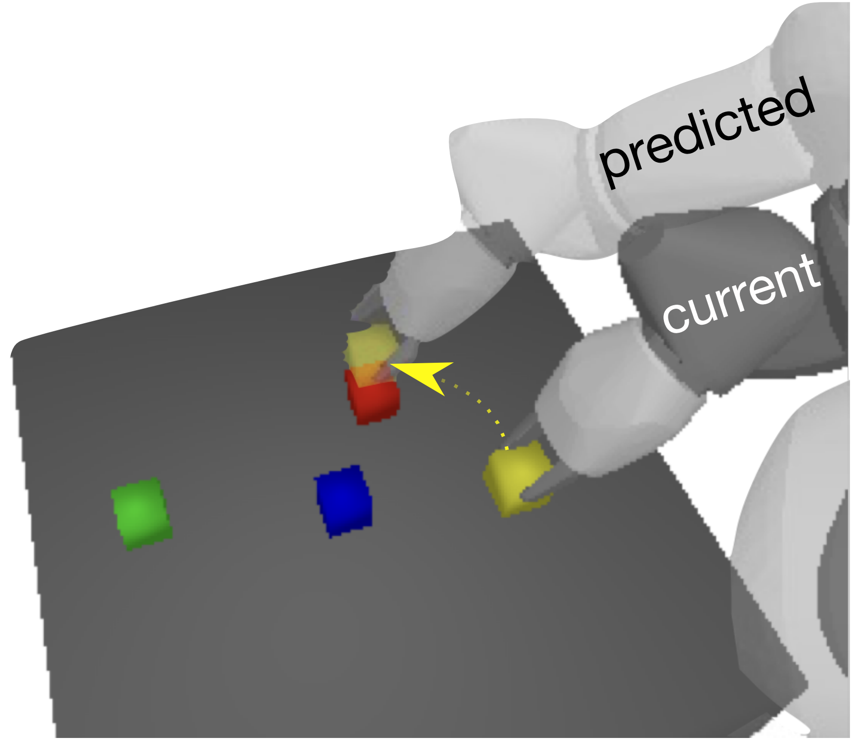
Reconstruction

$$||\hat{W}_t - W_t||_2^2$$



Pose

$$C_{actor}(\hat{\theta}_t, \theta_t)$$



SubGoal

$$C_G(\hat{G}_t, G_t)$$

`move(yellow, red)`

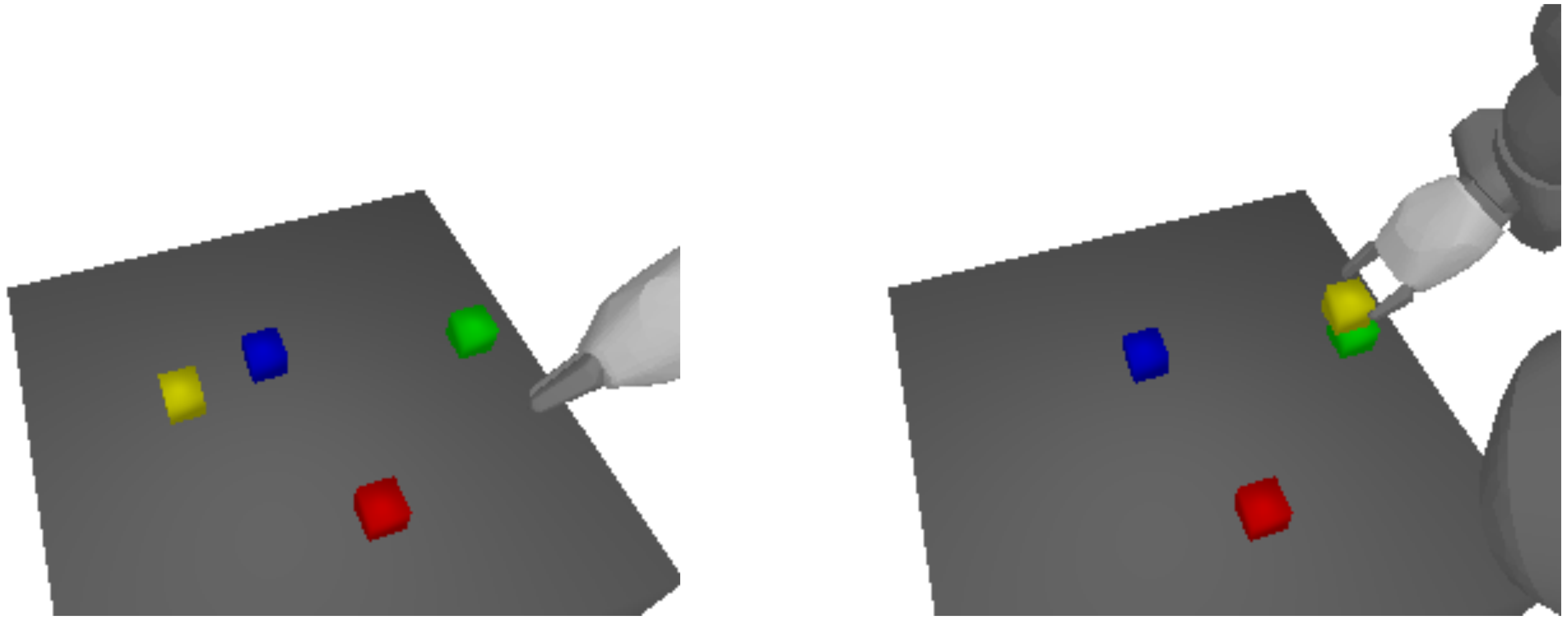
Block pos

$$C_{obj}(z_t)$$

x #steps in horizon

Paxton et al. Prospection: Interpretable Plans From Language By Predicting the Future ICRA 2019

Long Tails



Templates:

put the yellow one on the green block

Humans:

move the yellow cube to the right until it is on top of the green cube with the front half of the yellow cube touching the far half of the top of the green cube

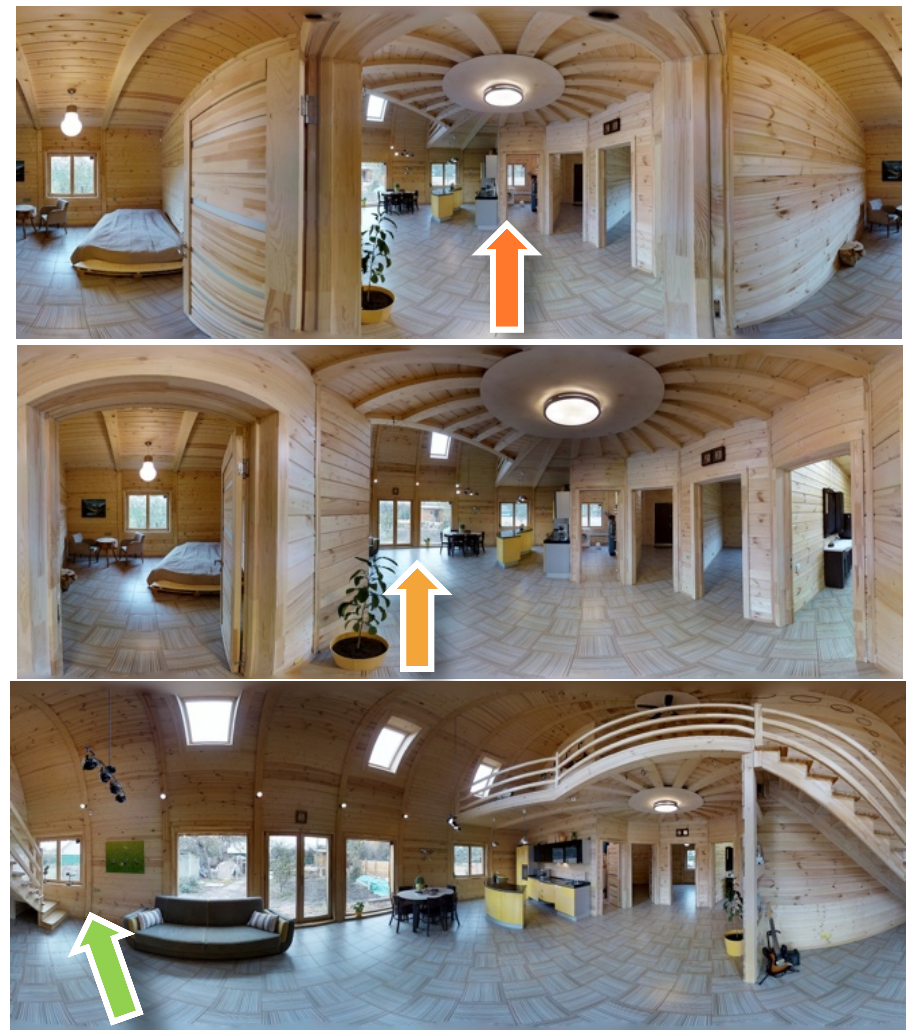
Paxton et al. Prospection: Interpretable Plans From Language By Predicting the Future ICRA 2019

Where does semantics come from?

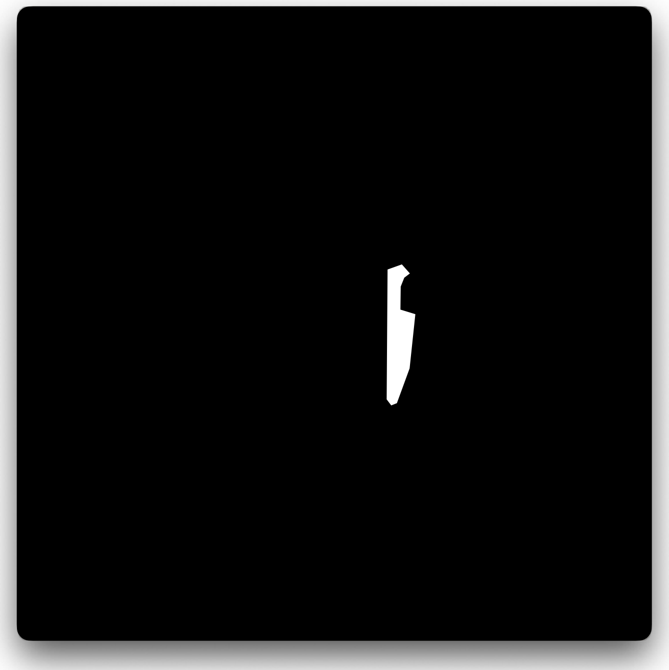
Someone labeled it?

$$p(a|v_0, ..., v_t)$$

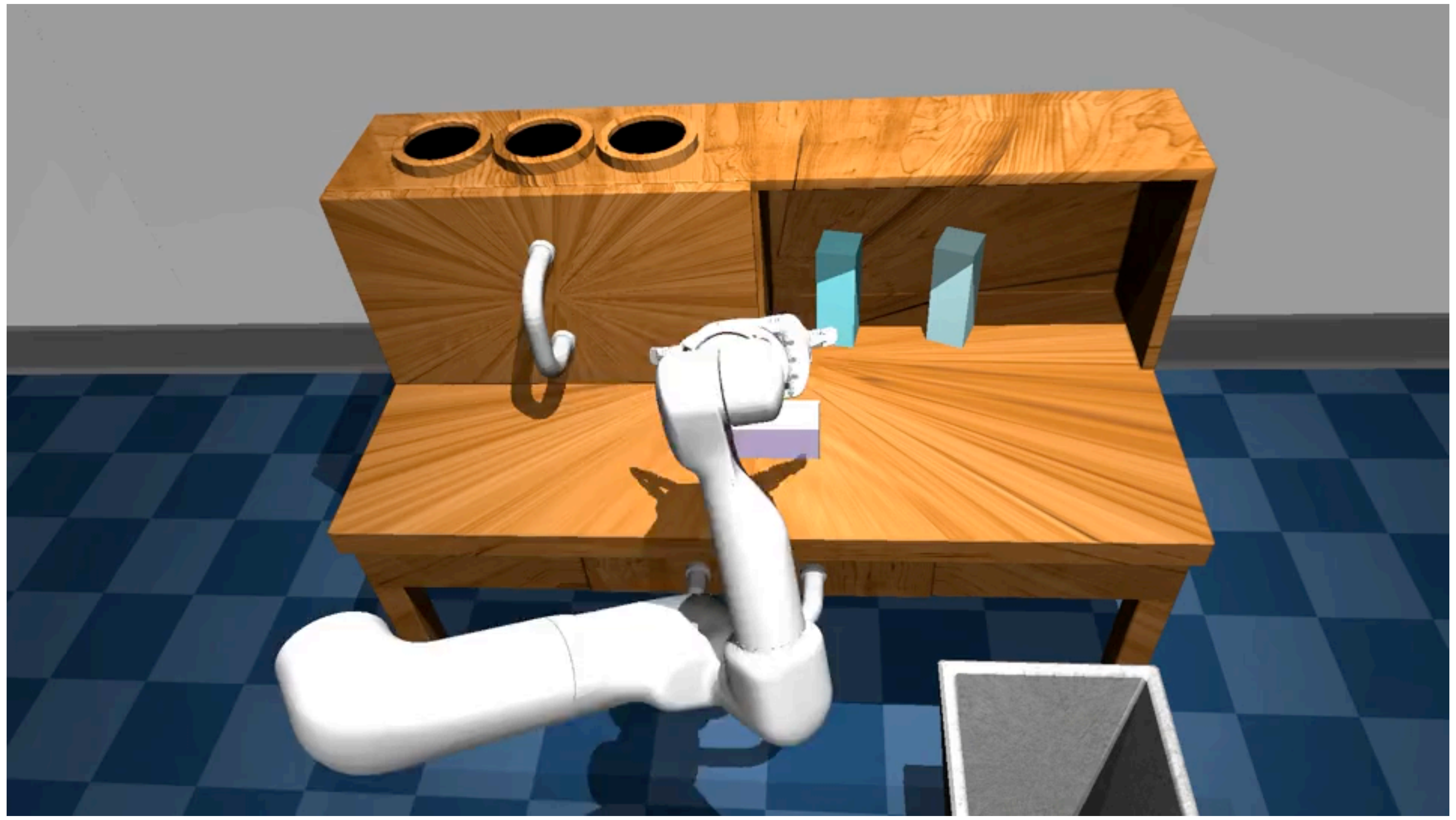
Simulator Definitions?



PickupObject



Self-Play and Physical Affordances?



Lynch et al. — Learning Latent Plans from Play — CoRL 2019

Embodiment

- Choose your own adventure — Lots of noise
- What does it mean to succeed?
- Where do concepts come from?
- What’s the role of exploration?
- Language is woefully underspecified

All of these are the “same” verb