

CMPT 983

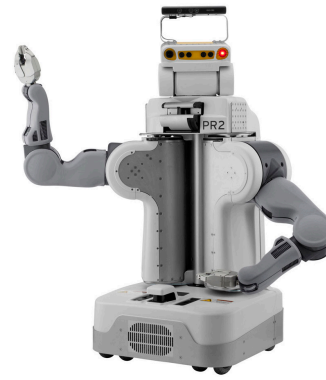
Grounded Natural Language Understanding

March 25, 2021

Instruction Following - RoboNLP

Robobarista

- PR2 robot make coffee?



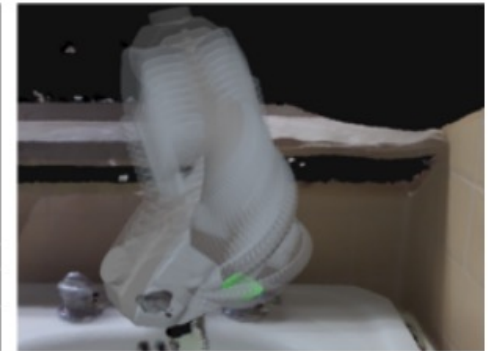
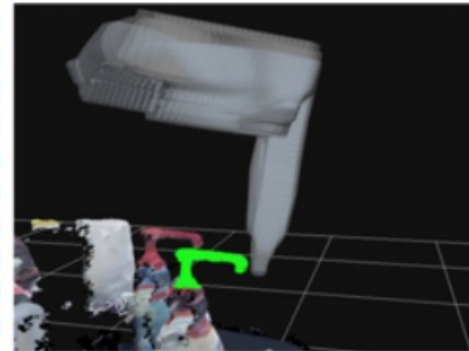
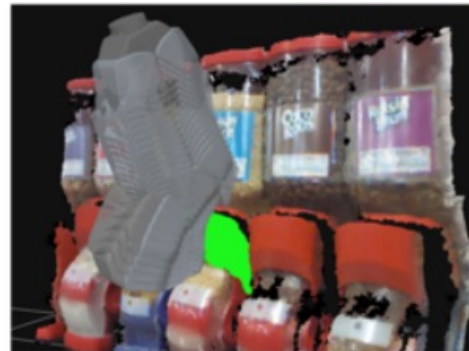
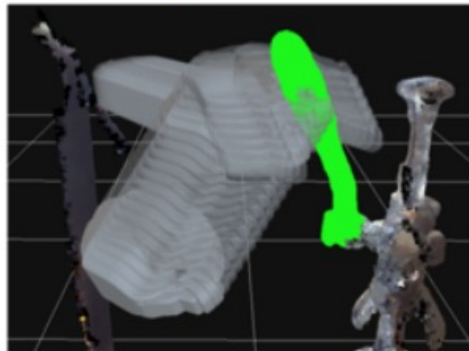
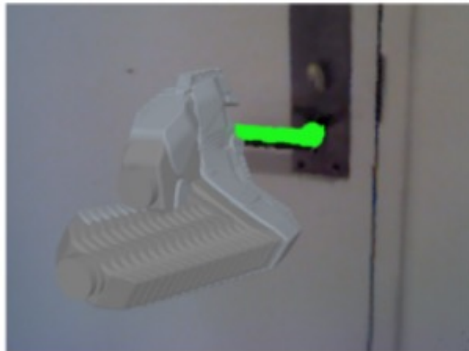
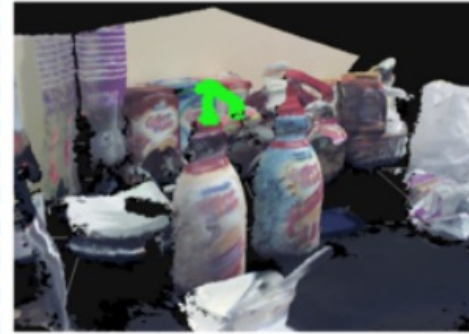
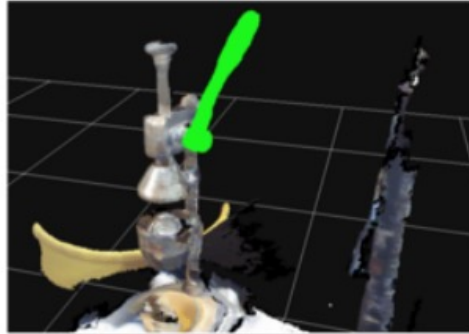
“Pull the handle down and then towards you to open the door.”

“Pull the handle to squeeze the juice from the fruit.”

“Pull the Crispy Rice handle to dispense.”

“Press down on the hazelnut syrup pump to dispense.”

“Turn handle counterclockwise to turn on cold water.”

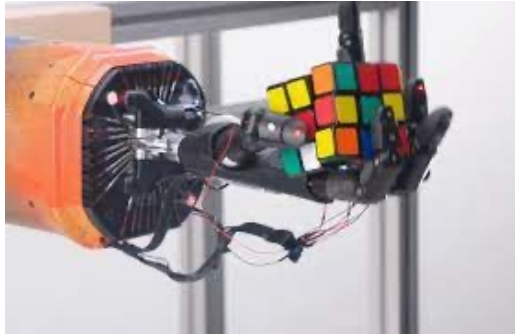


Deep Multimodal Embedding: Manipulating Novel Objects with Point-clouds, Language and Trajectories, Sung et al, ICRA 2017

RoboNLP

Robot following instructions

- Navigation
- Interaction
- Manipulation



Last week: Instruction-guided Visual Navigation



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

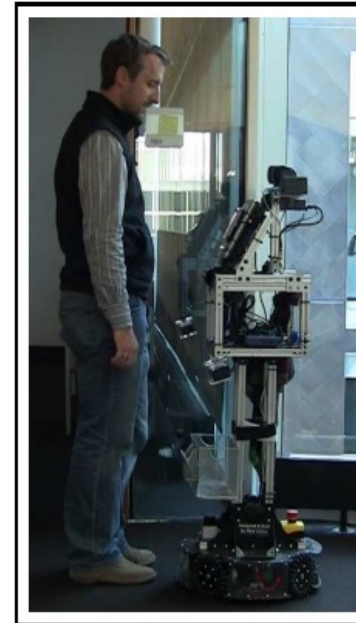
Navigation

Pioneer AT



“Go to the break room and report the location of the blue box.”
(Dzifcak et al, ICRA 2009)

CoBot



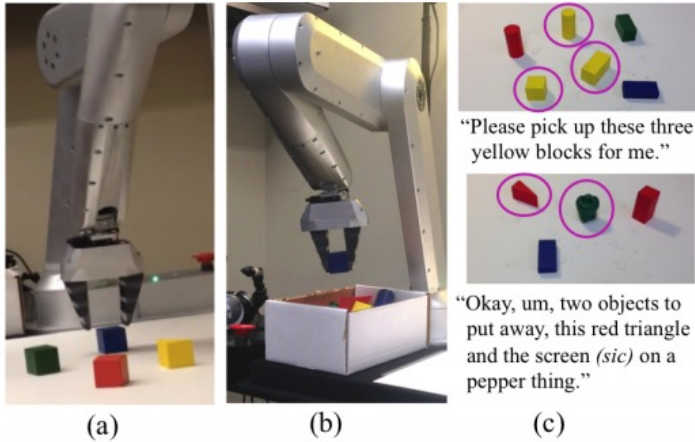
“Take me to the meeting room.”
(Kollar et al, ICRA 2013)

Commands

- Go to the bridge.
 - Go to the lab.
 - Bring me to the elevator.
 - Go to Christina’s office.
 - Take me to the meeting room.
-

Manipulation

Gambit manipulator

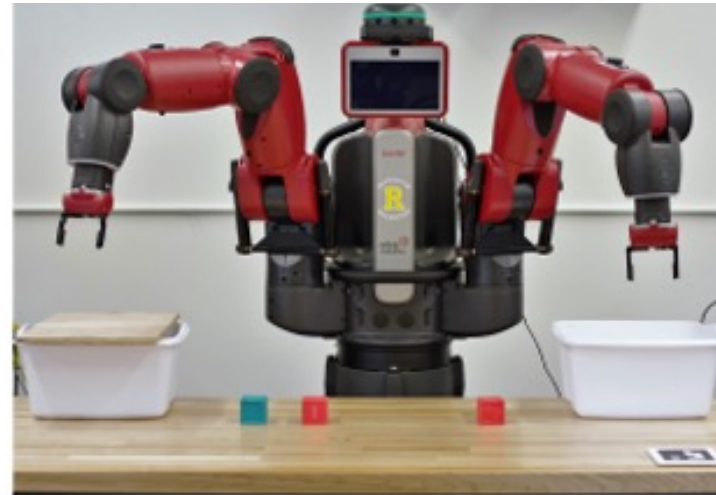


“Please pick up these three yellow blocks for me.”

“Okay, um, two objects to put away, this red triangle and the screen (*sic*) on a pepper thing.”

Pick and place
(Matuszek et al, AAI 2014)

Baxter



“Pick up the blue block and drop it in the box on your right. Pick up the red block and drop it in the box on your left”
(Boteanu et al, IROS 2016)

Baxter

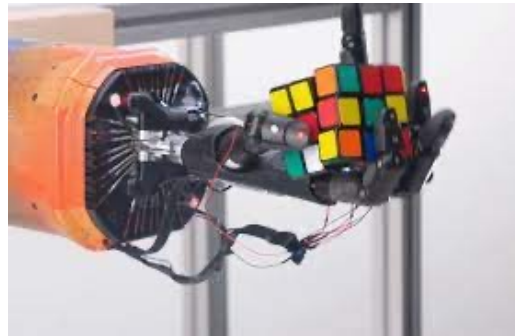


Learn from dialog and demonstrations
(Chai et al, IJCAI 2018)

RoboNLP

Robot following instructions

- Navigation
- Interaction
- Manipulation



Communication in embodied environments

- Human providing information
- Robot providing information
 - Question answering
- Robot asking for help
- Dialogue

Other tasks with manipulation

Jaco arm



Identify objects using attributes:
“silver, round, and empty”
(Thomason et al, IJCAI 2016)



Ask for help
(Tellex et al, RSS 2014)

TUM-Rosie



Make pancakes using recipes from
wikipantries.com (Nyga and Beetz,
IROS 2012)

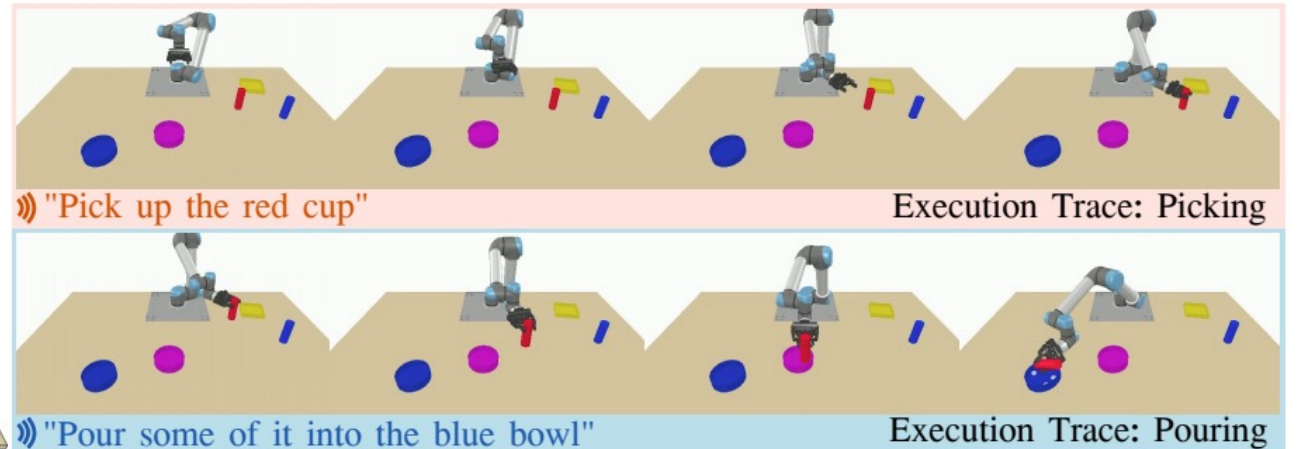
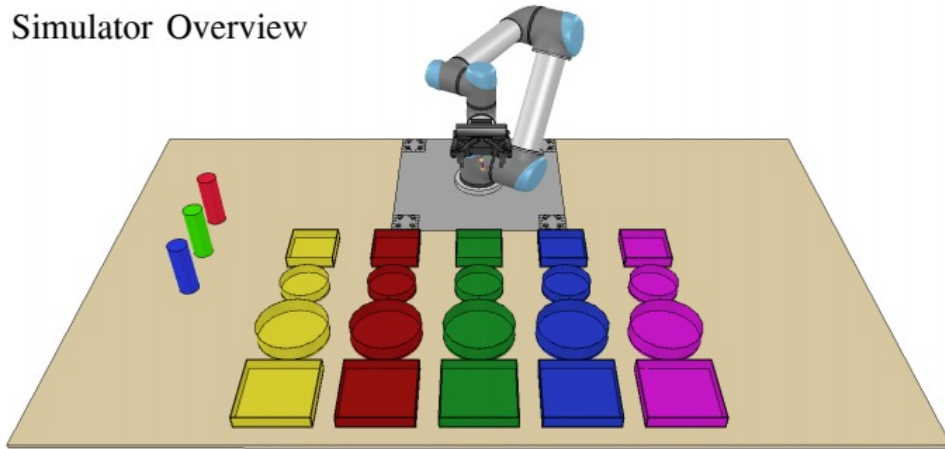
Most of these are very limited and constrained:
very small vocabulary, very specific scenario (pre-deep learning)

Examples from “Robots that use language: A survey”, Tellex et al, 2020

<https://h2r1.cs.brown.edu/wp-content/uploads/tellex20.pdf>

Manipulation

Simulator Overview



Language-Conditioned Imitation Learning for Robot Manipulation Tasks,
<https://arxiv.org/pdf/2010.12083.pdf>, Stepputtis et al, 2020

Navigation + Interaction/Manipulation

Why challenging?

- Need data
- Large task space
- Real robots are challenging (slow and tricky to work with)
- Accurate simulations involving manipulations is also challenging (modeling physics is not easy)



(a) Robotic forklift

Commands from the corpus

- Go to the first crate on the left and pick it up.
 - Pick up the pallet of boxes in the middle and place them on the trailer to the left.
 - Go forward and drop the pallets to the right of the first set of tires.
 - Pick up the tire pallet off the truck and set it down
-

(b) Sample commands

Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation, Tellex et al, AAAI 2011

Re-arrangement <https://arxiv.org/pdf/2011.01975.pdf>

Meta-benchmark

Goal state can be specified using **language**

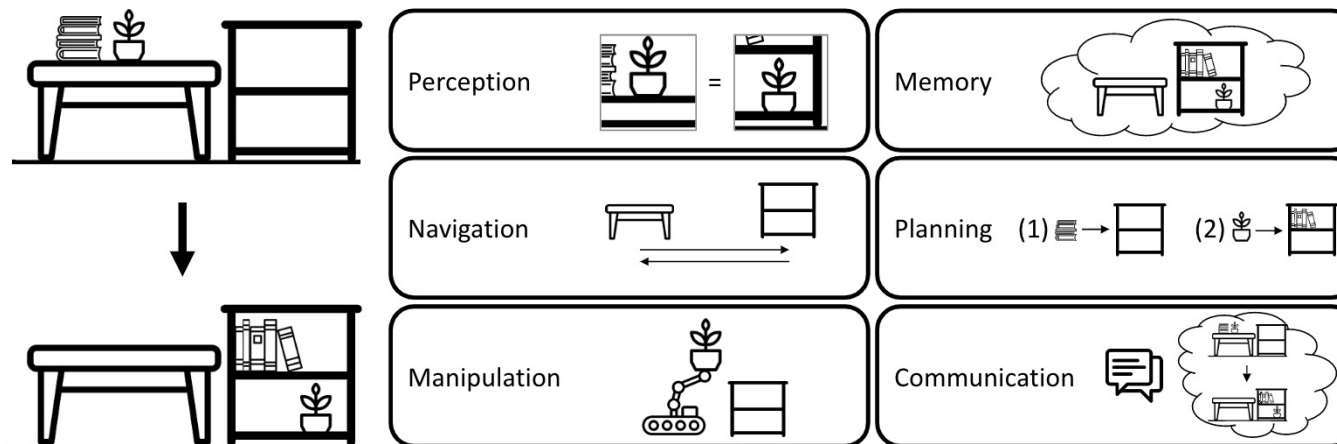


Whitepaper with

Intel, FAIR, AI2, Google, Berkeley, Princeton, GaTech, Imperial, UCSD

<https://arxiv.org/pdf/2011.01975.pdf>

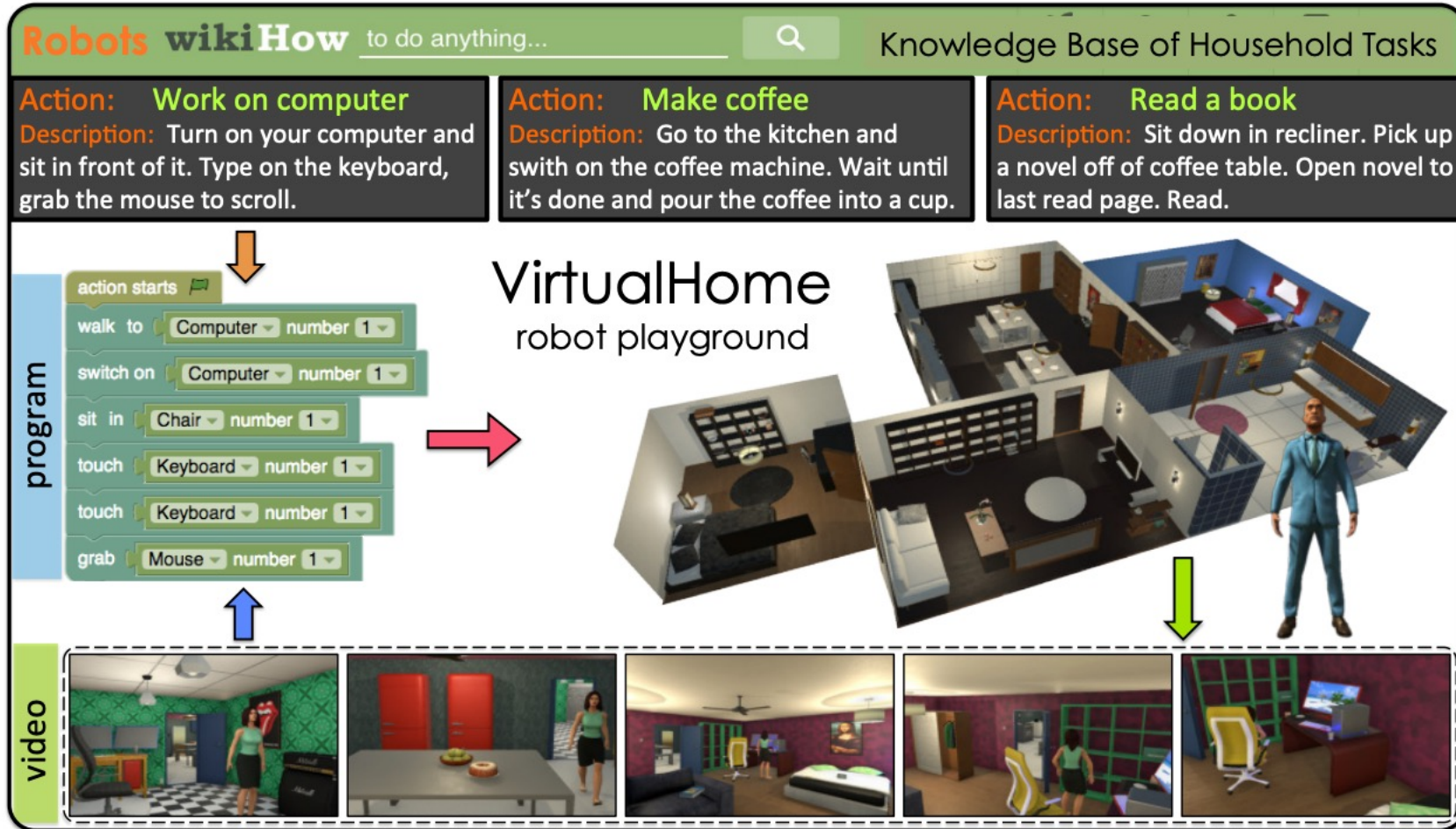
Encourage research across sub-areas



Interactive Habitat

Existing language +
interaction work

VirtualHome <http://virtual-home.org/>



- Collect common activities for 8 scene types
- Collect descriptions, and programs
- Use programs to generate videos

VirtualHome: Simulating Household Activities via Programs

<https://arxiv.org/pdf/1806.07011.pdf>

Puig et al, CVPR 2018

VirtualHome: Data collection

“Visual interface” to allow crowdworkers to piece together program

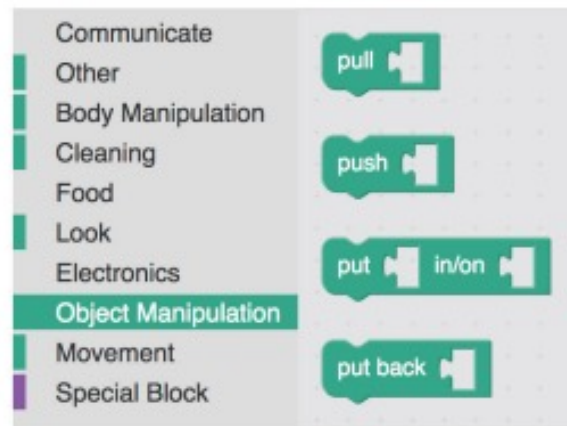
Action name:

Throw away newspaper

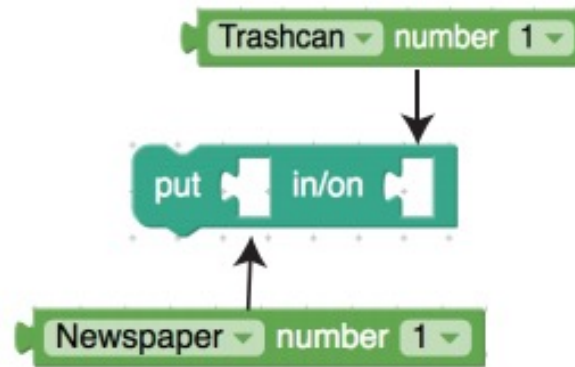
Description:

Take the newspaper on the living room table and toss it.

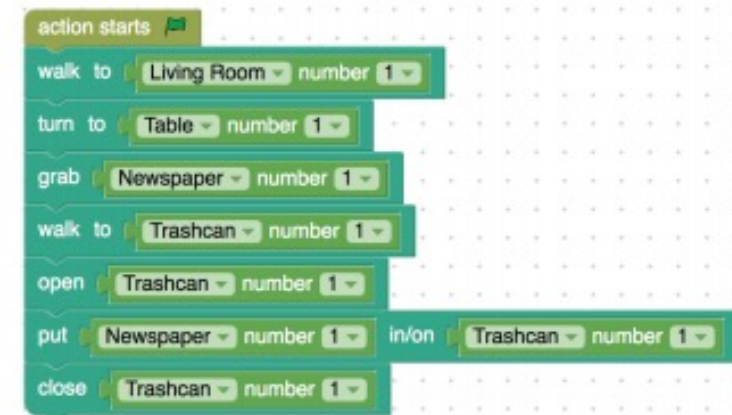
a)



b)



c)

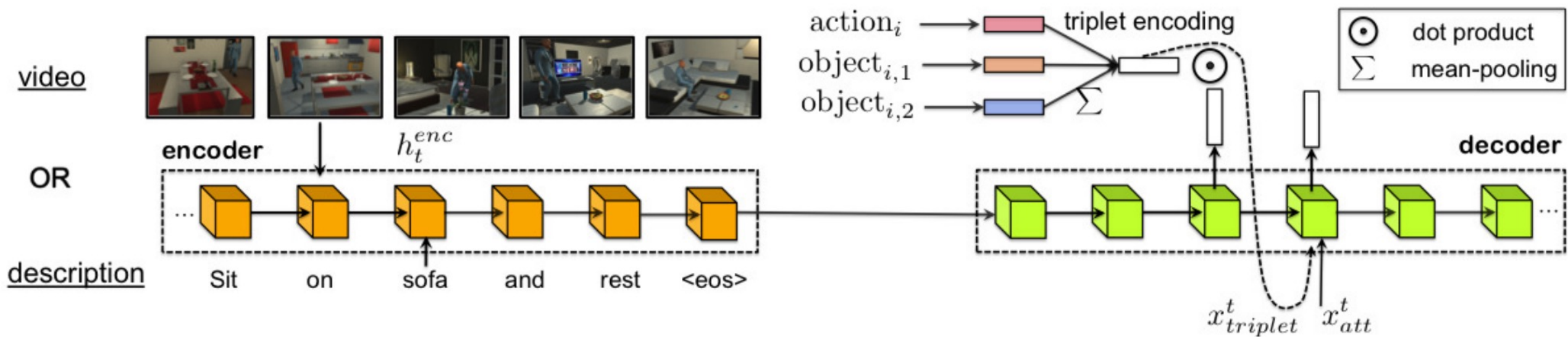


d)

Two datasets:

- ActivityPrograms: 1814 descriptions, 2821 programs for 75 actions, 308 objects
- VirtualHome Activity: synthesized 5193 programs (with human provided description), 12 most common actions

VirtualHome: Generate program from language or video



VirtualHome: Generated programs



Description: Get an empty glass. Take milk from refrigerator and open it. Pour milk into glass.



Description: Go watch TV on the couch. Turn the TV off and grab the coffee pot. Put the coffee pot on the table and go turn the light on.

Alfred

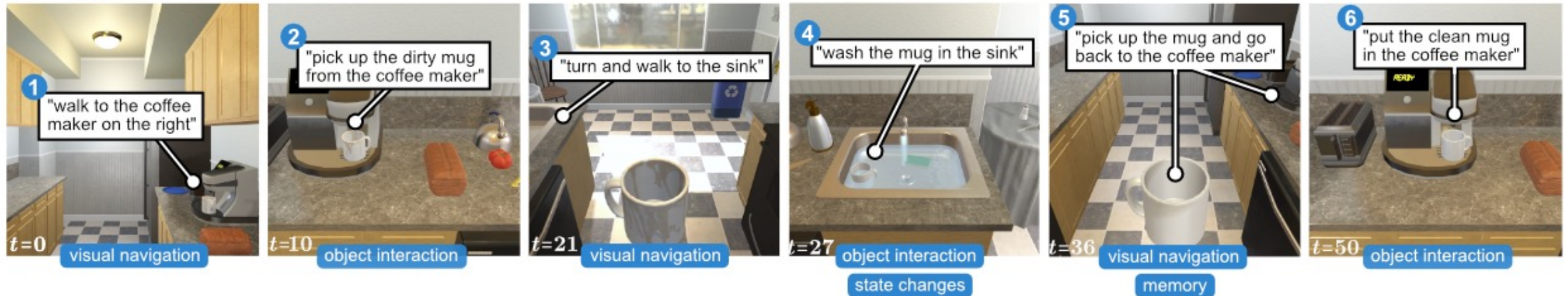
<https://askforalfred.com/>

25K English instructions

8K expert demonstrations (average 50 steps)

430K image-action pairs

Goal: "Rinse off a mug and place it in the coffee maker"



A Benchmark for Interpreting Grounded Instructions for Everyday Tasks

<https://arxiv.org/pdf/1912.01734.pdf>

Shridhar et al, CVPR 2020



ALFRED

A Benchmark for Interpreting
Grounded Instructions for Everyday Tasks




Alfred dataset

- 7 task types
- 84 object classes in 120 scenes

Episodes with Navigation + Interaction


- Generate expert demonstrations using planner and specifying task-specific planning rules with start and end position
- Interaction using object mask
- Collect instructions from AMT

	Pick & Place	Stack & Place	Pick Two & Place	Clean & Place	Heat & Place	Cool & Place	Examine in Light
item(s)	Book	Fork (in) Cup	Spray Bottle	Dish Sponge	Potato Slice	Egg	Credit Card
receptacle	Desk	Counter Top	Toilet Tank	Cart	Counter Top	Side Table	Desk Lamp
scene #	Bedroom 14	Kitchen 10	Bathroom 2	Bathroom 1	Kitchen 8	Kitchen 21	Bedroom 24
expert demonstration							

	Annotation # 1	Annotation # 2	Annotation # 3
Goals	Put a clean sponge on a metal rack.	Place a clean sponge on the drying rack	Put a rinsed out sponge on the drying rack
Instructions	Go to the left and face the faucet side of the bath tub. Pick up left most green sponge from the bath tub. Turn around and go to the sink. Put the sponge in the sink. Turn on then turn off the water. Take the sponge from the sink. Go to the metal bar rack to the left. Put the sponge on the top rack to the left of the lotion bottle.	Turn around and walk over to the bathtub on the left. Grab the sponge out of the bathtub. Turn around and walk to the sink ahead. Rinse the sponge out in the sink. Move to the left a bit and face the drying rack in the corner of the room. Place the sponge on the drying rack.	Walk forwards a bit and turn left to face the bathtub. Grab a sponge out of the bathtub. Turn around and walk forwards to the sink. Rinse the sponge out in the sink and pick it up again. Turn left to walk a bit, then face the drying rack. Put the sponge on the drying rack.

Other datasets

	— Language —		— Virtual Environment —			— Inference —		
	# Human Annotations	Granularity	Visual Quality	Movable Objects	State Changes	Vis. Obs.	Navigation	Interaction
TACoS [43]	17k+	High&Low	Photos	✗	✗	–	–	–
R2R [3]; Touchdown [14]	21k+; 9.3k+	Low	Photos	✗	✗	Ego	Graph	✗
EQA [15]	✗	High	Low	✗	✗	Ego	Discrete	✗
Matterport EQA [55]	✗	High	Photos	✗	✗	Ego	Discrete	✗
IQA [20]	✗	High	High	✗	✓	Ego	Discrete	Discrete
VirtualHome [42]	2.7k+	High&Low	High	✓	✓	3 rd Person	✗	Discrete
VSP [58]	✗	High	High	✓	✓	Ego	✗	Discrete
ALFRED 	25k+	High&Low	High	✓	✓	Ego	Discrete	Discrete + Mask

AI2Thor <https://ai2thor.allenai.org/>

- Unity game engine with designed actions and object states

Object Rearrangement

AI2THOR

AI2-THOR: “Actionable Properties”

<https://ai2thor.allenai.org/ithor/documentation/objects/actionable-properties/>

Properties on **126 objects** indicating **actions** that can be performed

- Openable (OpenObject/CloseObject)
- Pickupable (PickupObject/PutObject/DropHandObject/Throw/Pull/Push)
- Moveable (Push/Pull - too large to be Pickupable)
- Toggleable (ToggleObjectOn/ToggleObjectOff)
- Receptable (other objects can be placed on/in these objects)
- Fillable (FillObjectWithLiquid)
- Sliceable (SliceObject, one-way state change)
- Cookable (CookObject, one-way state change)
- Breakable (BreakObject, one-way state change)
- Dirty (DirtyObject toggles the state)
- UsedUp (UseUp, one-way state change)

AI2-THOR

- Material Properties <https://ai2thor.allenai.org/ithor/documentation/objects/material-properties/>
 - Temperature (Hot/Cold/RoomTemp)
 - All objects have Temperature
 - Some objects can change the temperature of other objects (e.g. Refrigerator, Stove Burner)
 - Mass
 - Salient Materials (Metal, Wood, Plastic, Glass, Ceramic, Stone, Fabric, Rubber, Food, Paper, Wax, Soap, Sponge, Organic)
 - Individual objects will have different salient materials (material types)
- Contextual interactions <https://ai2thor.allenai.org/ithor/documentation/objects/contextual-interactions/>
 - Rules that specify state change to objects under certain conditions
 - Example: BreadSliced, becomes Cooked if:
 - PutObject is used to place BreadSliced into Toaster objects that is on
 - Moved over Stove Burner that is on

Agent model

- Seq2seq model (CNN vision, LSTM language)
- Predicts **action + binary mask** of object from **concatenated input**
 - 13 actions (5 navigation + 7 interaction + stop)

Navigation

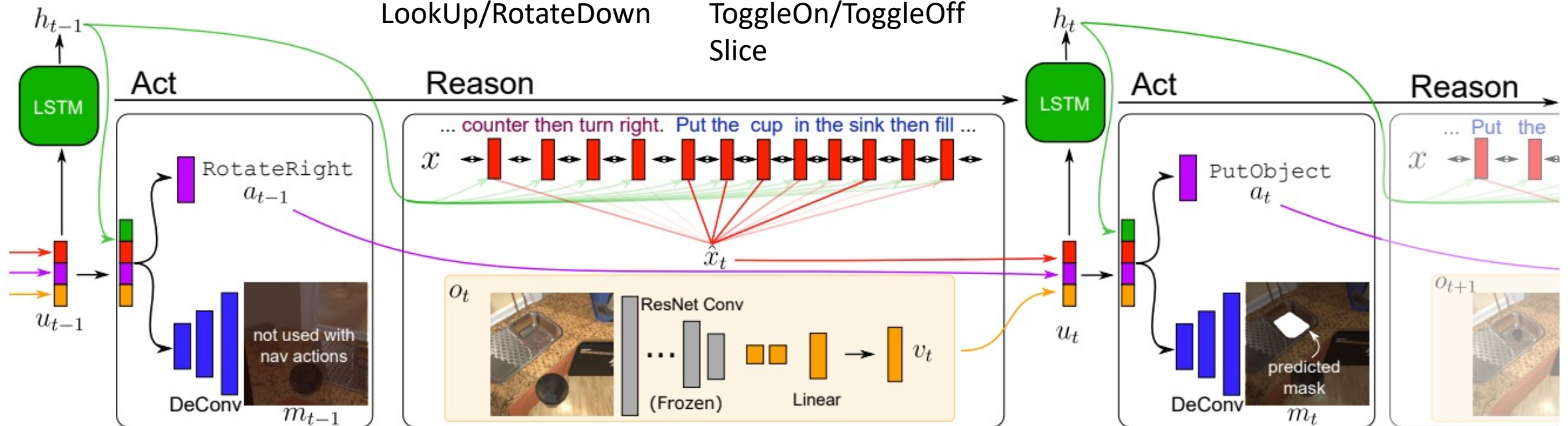
MoveAhead
RotateLeft/RotateRight
LookUp/RotateDown

Interaction

Pickup/Put
Open/Close
ToggleOn/ToggleOff
Slice

Concatenated input

Vision
Language
Last action



Alfred Training

- Train with imitation learning / teacher forcing
 - Dagger / student-forcing challenging
- Variations:
 - Progress Monitors
 - Estimate of progress toward goal
 - Helps to learn utility of each state
 - Two progress monitors (both trained with L2 loss)
 - Predict overall progress based on time t/T
 - Predict (normalized) number of subgoals accomplished
 - Uses as input LSTM hidden state + concatenated input (of vision + language + last action)

$$p_t = \sigma (W_p [h_t; u_t])$$

$$c_t = \sigma (W_c [h_t; u_t]).$$

t = current time step

T = total length of expert demonstration

Alfred evaluation

- Task Success:
 - Does the final object position and state match goal? (1 or 0)
- Goal-Condition Success
 - Ratio of goals completed (1 → task success)
- Path weighted Task and Goal-Condition Success

$$p_s = s \times \frac{L^*}{\max(L^*, \hat{L})}$$

Like SPL for navigation, but for action path

Alfred results

- Compare against simple no language and no vision baselines
- Random is at 0%, PM adds ~1% for seen environments
- Very poor performance!

Model	Validation				Test			
	<i>Seen</i>		<i>Unseen</i>		<i>Seen</i>		<i>Unseen</i>	
	Task	Goal-Cond	Task	Goal-Cond	Task	Goal-Cond	Task	Goal-Cond
NO LANGUAGE	0.0 (0.0)	5.9 (3.4)	0.0 (0.0)	6.5 (4.7)	0.2 (0.0)	5.0 (3.2)	0.2 (0.0)	6.6 (4.0)
NO VISION	0.0 (0.0)	5.7 (4.7)	0.0 (0.0)	6.8 (6.0)	0.0 (0.0)	3.9 (3.2)	0.2 (0.1)	6.6 (4.6)
GOAL-ONLY	0.1 (0.0)	6.5 (4.3)	0.0 (0.0)	6.8 (5.0)	0.1 (0.1)	5.0 (3.7)	0.2 (0.0)	6.9 (4.4)
INSTRUCTIONS-ONLY	2.3 (1.1)	9.4 (6.1)	0.0 (0.0)	7.0 (4.9)	2.7 (1.4)	8.2 (5.5)	0.5 (0.2)	7.2 (4.6)
SEQ2SEQ	2.4 (1.1)	9.4 (5.7)	0.1 (0.0)	6.8 (4.7)	2.1 (1.0)	7.4 (4.7)	0.5 (0.2)	7.1 (4.5)
+ PM PROGRESS-ONLY	2.1 (1.1)	8.7 (5.6)	0.0 (0.0)	6.9 (5.0)	3.0 (1.7)	8.0 (5.5)	0.3 (0.1)	7.3 (4.5)
+ PM SUBGOAL-ONLY	2.1 (1.2)	9.6 (5.5)	0.0 (0.0)	6.6 (4.6)	3.8 (1.7)	8.9 (5.6)	0.5 (0.2)	7.1 (4.5)
+ PM Both	3.7 (2.1)	10.0 (7.0)	0.0 (0.0)	6.9 (5.1)	4.0 (2.0)	9.4 (6.3)	0.4 (0.1)	7.0 (4.3)
HUMAN	-	-	-	-	-	-	91.0 (85.8)	94.5 (87.6)

Success metrics (with path weighted metrics in parenthesis)

Alfred results: Ablations

- Sub-Goal Evaluation
 - How well can the agent accomplish each sub-goal (assuming perfect performance up to that point)?

	Model	<i>Goto</i>	<i>Pickup</i>	<i>Put</i>	<i>Cool</i>	<i>Heat</i>	<i>Clean</i>	<i>Slice</i>	<i>Toggle</i>	Avg.
<i>Seen</i>	No Lang	28	22	71	89	87	64	19	90	59
	S2S	49	32	80	87	85	82	23	97	67
	S2S + PM	51	32	81	88	85	81	25	100	68
<i>Unseen</i>	No Lang	17	9	31	75	86	13	8	4	30
	S2S	21	20	51	94	88	21	14	54	45
	S2S + PM	22	21	46	92	89	57	12	32	46

Alfred progress (leaderboard)

Rank	Submission	Created	Unseen Success Rate	Seen Success Rate	Seen PLWSR	Unseen PLWSR	Seen GC	Unseen GC	Seen PLW GC Success Rate	Unseen PLW GC Success Rate
1	LWIT <i>Anonymous</i>	01/04/2021	0.0942	0.3092	0.2590	0.0560	0.4053	0.2091	0.3676	0.1634
2	E.T. <i>Anonymous</i>	02/22/2021	0.0857	0.3842	0.2778	0.0410	0.4544	0.1856	0.3493	0.1146
3	A test <i>A test</i>	03/09/2021	0.0530	0.2146	0.1472	0.0272	0.2771	0.1425	0.2171	0.0999
3	A new method <i>Anonymous</i>	10/26/2020	0.0530	0.2205	0.1510	0.0272	0.2829	0.1428	0.2205	0.0999
5	LWIT <i>Anonymous</i>	08/01/2020	0.0445	0.1239	0.0820	0.0224	0.2068	0.1234	0.1879	0.0944
6	Baseline Seq2Seq + Progress M... <i>Singh, Bhambri, Kim, Choi (Gl...</i>	07/22/2020	0.0150	0.0541	0.0251	0.0070	0.1232	0.0808	0.0827	0.0520
10	Baseline Seq2Seq+PM (both) <i>Shridhar et. al (UW)</i>	03/28/2020	0.0039	0.0398	0.0202	0.0008	0.0942	0.0703	0.0627	0.0426

ALFWorld <https://alfworld.github.io/>

- Aligned tasks and scenarios in TextWorld and AI2Thor (Alfred)
- Is it possible to train an agent in a text-only world and transfer to embodied setting?

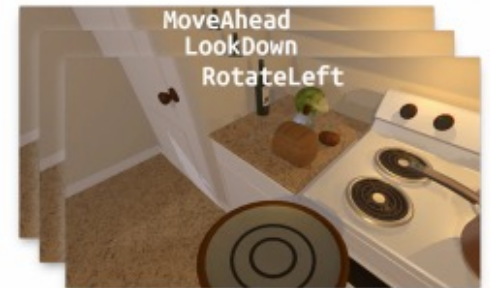
ALFWorld Aligning Text and Embodied Environments for Interactive Learning,
<https://arxiv.org/pdf/2010.03768.pdf>
Shridhar et al, ICLR 2021

TextWorld



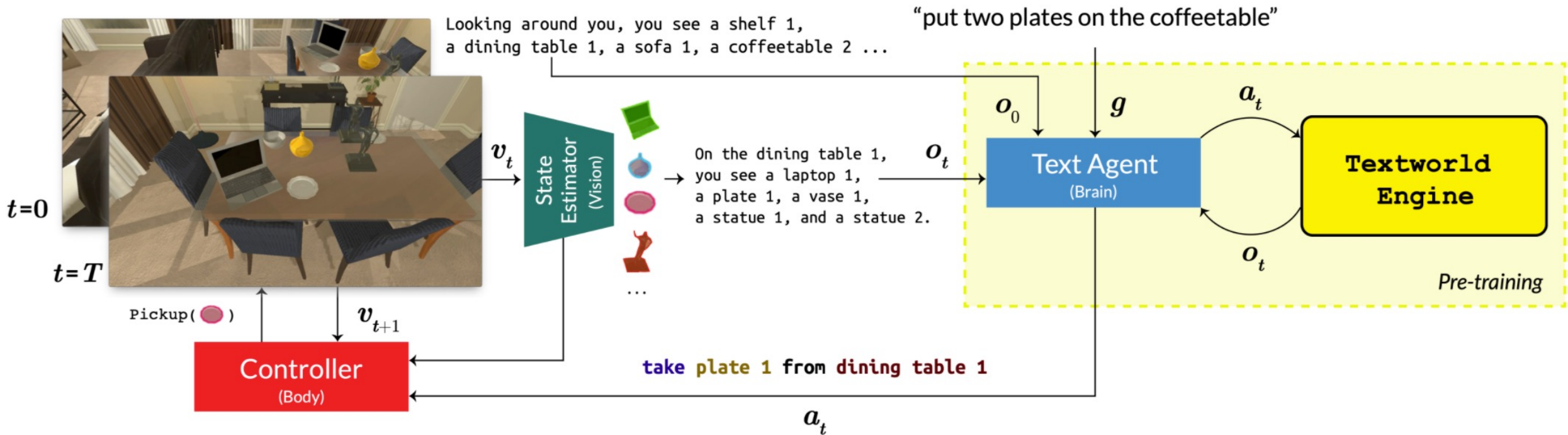
```
Welcome!  
You are in the middle of the room.  
Looking around you, you see  
a diningtable, a stove,  
a microwave, and a cabinet.  
  
Your task is to:  
Put a pan on the diningtable.  
  
> goto the cabinet  
You arrive at the cabinet.  
The cabinet is closed.  
  
> open the cabinet  
The cabinet 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 winebottle.  
  
> take the pan from the stove  
You take the pan from the stove.  
  
> goto the diningtable  
You arrive at the diningtable.  
  
> put the pan on the diningtable  
You put the pan on the  
diningtable.
```

Embodied



Transferring from text only domain

- State estimator translates from vision to text description



ALFWorld Results

Synthetic (generated) instructions

task-type	TextWorld		Seq2Seq		BUTLER		BUTLER-ORACLE		Human Goals	
	seen	unseen	seen	unseen	seen	unseen	seen	unseen	seen	unseen
Pick & Place	69	50	28 (28)	17 (17)	30 (30)	24 (24)	53 (53)	31 (31)	20 (20)	10 (10)
Examine in Light	69	39	5 (13)	0 (6)	10 (26)	0 (15)	22 (41)	12 (37)	2 (9)	0 (8)
Clean & Place	67	74	32 (41)	12 (31)	32 (46)	22 (39)	44 (57)	41 (56)	18 (31)	22 (39)
Heat & Place	88	83	10 (29)	12 (33)	17 (38)	16 (39)	60 (66)	60 (72)	8 (29)	5 (30)
Cool & Place	76	91	2 (19)	21 (34)	5 (21)	19 (33)	41 (49)	27 (44)	7 (26)	17 (34)
Pick Two & Place	54	65	12 (23)	0 (26)	15 (33)	8 (30)	32 (42)	29 (44)	6 (16)	0 (6)
All Tasks	40	35	6 (15)	5 (14)	19 (31)	10 (20)	37 (46)	26 (37)	8 (17)	3 (12)

Task success percentage (averaged across 3 evaluation runs)
with goal conditioned successes in parenthesis

Training Strategy	train (succ %)	seen (succ %)	unseen (succ %)	train speed (eps/s)
EMBODIED-ONLY	21.6	33.6	23.1	0.9
TW-ONLY	23.1	27.1	34.3	6.1
HYBRID	11.9	21.4	23.1	0.7

TW-Only (train in TW, zero-shot transfer to embodied),
Hybrid (75% TW, 25% embodied)

BUTLER

Building Understanding in Textworld via Language
for Embodied Reasoning

Next time

- Paper presentations (3/29)
 - RMM: A Recursive Mental Model for Dialogue Navigation (Ke)
 - Language-Conditioned Imitation Learning for Robot Manipulation Tasks (Carolina)
 - Few-shot Object Grounding and Mapping for Natural Language Robot Instruction Following (Discussion)
- Thursday (4/1): Interactive language grounding