Housekeep: Tidying Virtual Households using Commonsense Reasoning

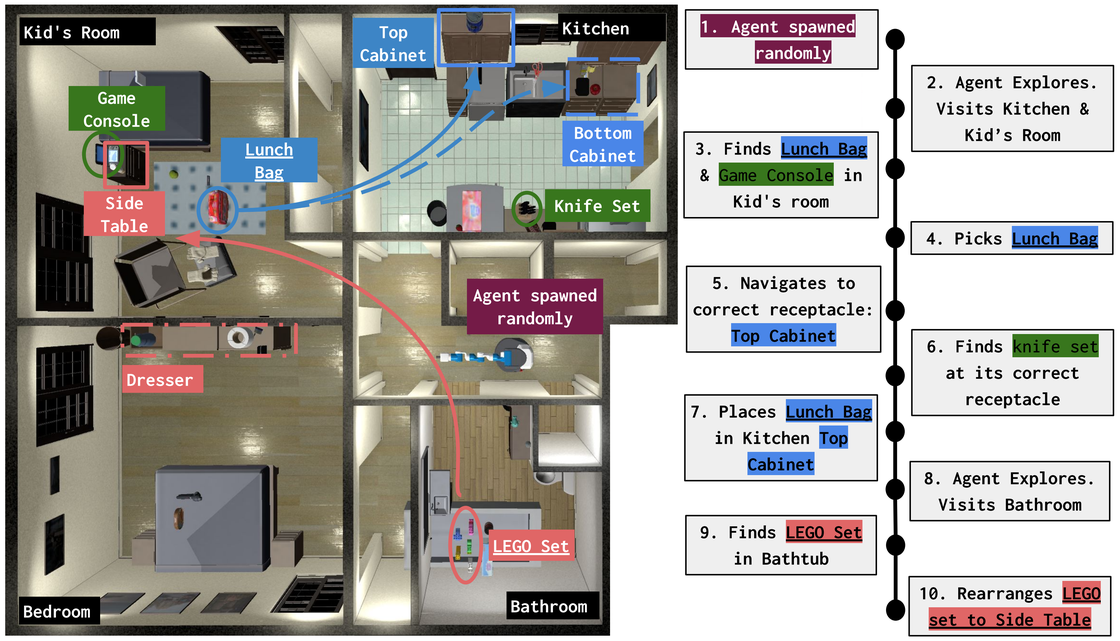
Yash Kant[1,2], Arun Ramachandran[2], Sriram Yenamandra[2], Igor Gilitschenski[1], Dhruv Batra[2,3], Andrew Szot*[2], Harsh Agrawal*[2]
Twitter: @yash2kant, @_arun_r, @yvsriram, @igilitschenski, @DhruvBatraDB, @andrew_szot, @harsh_092
[1]University of Toronto, [2]Georgia Tech, [3]Meta AI

https://yashkant.github.io/housekeep/

Overview

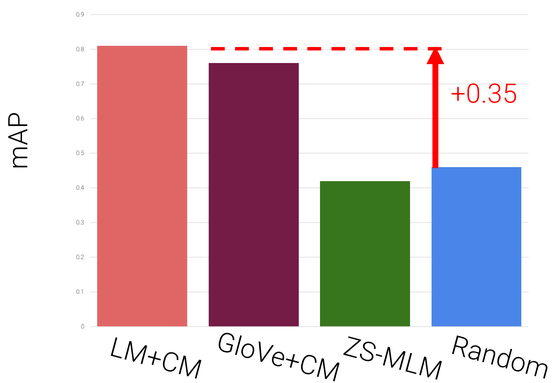
Can embodied agents do tasks without detailed human instructions?

- Housekeep, a new benchmark for common sense reasoning in embodied AI
- The agent must **rearrange objects without explicit instructions**
- We collect a **dataset of human object placement preferences**
- We build a **modular baseline that uses a language model (LM) for planning**



Language Models for Embodied Commonsense

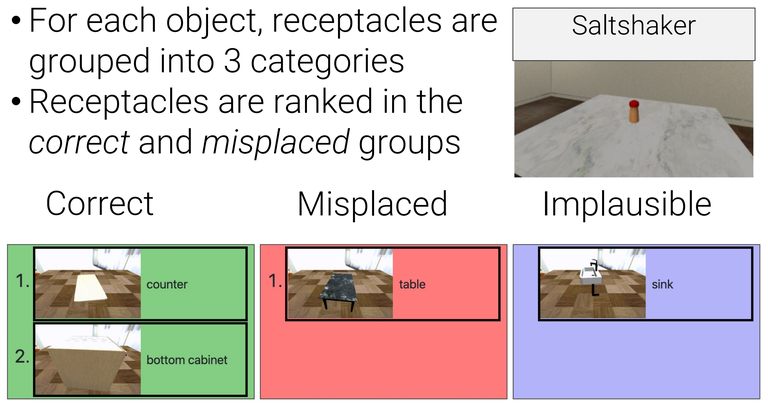
- Language models capture commonsense reasoning by ranking the compatibility of object / receptacle pairs.
- Right: comparing ranking strategies. Fine-tuned language model performs best



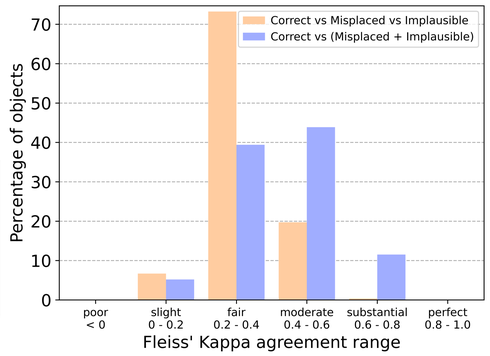
Human Preferences Dataset

- Over 45k annotations of human preferences for object placements
- 1799 objects, 268 object categories, 585 placements, and 105 rooms

- For each object, receptacles are grouped into 3 categories
- Receptacles are ranked in the *correct* and *misplaced* groups

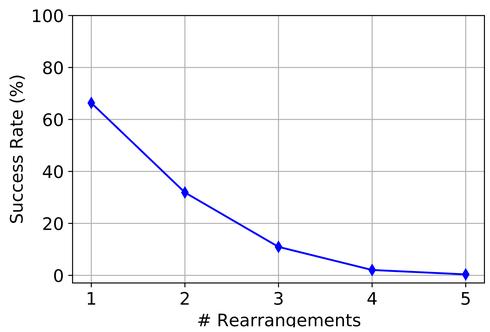


High agreement on object placements in dataset



Baseline and Ablations on Housekeep

- Modular Approach: Planning, exploration, navigation
- Right: Performance on Housekeep for different ranking and exploration modules



Episode success is low due to compounding errors between rearranging multiple objects

	#	Modules: Rank	Modules: Explore	Rearrange: ES ↑	Rearrange: OS ↑	Explore: MC ↑	Explore: OC ↑	Efficiency: PPE ↑
t-seen	1	OR	OR	1.00	1.00	–	1.00	1.00
	2	OR	FTR	0.35	0.64	73	0.73	1.00
	3	LM	OR	0.04	0.44	–	1.00	0.57
	4	LM	FTR	0.01	0.30	77	0.76	0.41
	5	GLV	FTR	0.01	0.29	71	0.73	0.39
t-unseen	6	OR	OR	1.00	1.00	–	1.00	1.00
	7	OR	FTR	0.35	0.65	74	0.74	1.00
	8	LM	OR	0.02	0.32	–	1.00	0.42
	9	LM	FTR	0.01	0.23	73	0.74	0.35
	10	GLV	FTR	0.00	0.23	72	0.74	0.26

- Success metrics computed using human preferences dataset
- Metrics for efficiency, soft alignment with human preferences, and exploration in paper
- Finding misplaced objects and rearranging them according to human preferences are difficult challenges of Housekeep