Housekeep: Tidying Virtual Households using Commonsense Reasoning

Yash Kant
University of Toronto, Georgia Tech
ykant6@gatech.edu

Arun Ramachandran
Georgia Tech
arunram@gatech.edu

Sriram Yenamandra
Georgia Tech
sriramy@gatech.edu

Igor Gilitschenski
University of Toronto
gilitschenski@cs.toronto.edu

Dhruv Batra
Georgia Tech, Meta AI
dbatra@gatech.edu

Andrew Szot*
Georgia Tech
aszot3@gatech.edu

Harsh Agrawal*
Georgia Tech
harsh.agrawal@gatech.edu

Abstract

*We introduce **Housekeep**, a benchmark to evaluate commonsense reasoning in the home for embodied AI. In Housekeep, an embodied agent must tidy a house by rearranging misplaced objects* without explicit instructions specifying which objects need to be rearranged. *Instead, the agent must learn from and is evaluated against human preferences of which objects* belong *where in a tidy house. Specifically, we collect a dataset of where humans typically place objects in tidy and untidy houses constituting 1799 objects, 268 object categories, 585 placements, and 105 rooms. Next, we propose a modular baseline approach for Housekeep that integrates planning, exploration, and navigation. It leverages a fine-tuned large language model (LLM) trained on an internet text corpus for effective planning. We show that our baseline agent generalizes to rearranging unseen objects in unknown environments.*

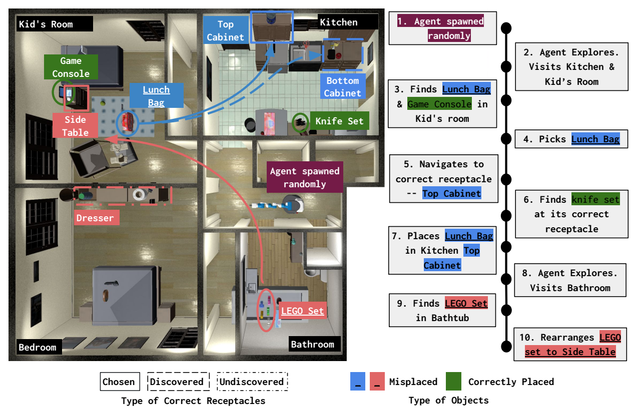

Figure 1. In Housekeep, an agent is spawned in an untidy environment and tasked with rearranging objects to suitable locations without explicit instructions. The agent explores the scene and discovers misplaced objects, correctly placed objects, and receptacles where objects belong. The agent rearranges a misplaced object (like a lunch box on the floor in the kid's room) to a better receptacle like the top cabinet in the kitchen.

1. Introduction

Imagine your house after a big party: there are dirty dishes on the dining table, cups left on the couch, and maybe a board game lying on the coffee table. Wouldn't it be nice for a household robot to clean up the house *without explicit instructions specifying which objects rearrange*?

Building AI reasoning systems that can perform such housekeeping tasks is an important scientific goal that has seen a lot of recent interest from the embodied AI community. The community has recently tackled various problems such as navigation [2, 4, 13, 16, 20, 28], interaction and manipulation [12, 27], instruction following [3, 26], and embodied question answering [11, 14, 29]. Each of these tasks defines a goal, e.g. navigating to a given location, moving objects to correct locations, or answering a question correctly. However, defining a goal for tidying a messy house is more tedious – one will have to write down a rule for where every object can or cannot be kept. Previous works in semantic reasoning frameworks for physical and relational commonsense [1, 5, 6, 10, 17, 18] are often limited to specific settings (*e.g*. evaluating multi-relational embeddings) without instantiating these tasks in a physically plausible scenario, or by not capturing the full context of a complete household (*e.g*. table-top organization). We believe the time is right to bridge the gap between these lines of research.

We introduce the Housekeep task to benchmark the ability of embodied AI agents to use physical commonsense reasoning and infer rearrangement goals that mimic human-preferred placements of objects. Figure 1 illustrates our task, where the Fetch robot is randomly spawned in an unknown house that contains unseen objects. Without explicit instructions, the agent must then discover objects placed in the house, classify the misplaced ones (LEGO set and lunch bag in Figure 1), and finally rearrange them to one of

* Equal Contribution

many suitable receptacles (matching color-coded square). We collect a dataset of human preferences of object placements in tidy and untidy homes and use this dataset for: a) generating semantically meaningful initializations of unclean houses, and b) defining evaluation criteria for what constitutes a clean house. This dataset contains rearrangement preferences for 1799 objects, in 585 placements, in 105 rooms from iGibson dataset scenes [25], constituting 1500+ hours of effort from 372 total annotators with 268 object categories curated from the Amazon-Berkeley [15], YCB objects [31], Google Scanned Objects [23], and iGibson [25] datasets. Housekeep evaluates how agents can rearrange novel objects not seen during training.

We propose a modular baseline and demonstrate that embodied (physical) commonsense extracted from large language models (LLMs) [7, 19] serves as an effective planner for Housekeep. Specifically, we find that finetuning LLM embeddings on a subset of human preferences generalizes well, and helps to reason about correct rearrangements for novel objects never seen during training. We integrate this LLM-based planning module into a hierarchical policy that coordinates navigation, exploration, and planning as a baseline approach to Housekeep. Our hierarchical approach also generalizes to unseen objects and scenes in Housekeep achieving an object success rate of 0.23 for unseen (versus 0.30 on seen objects).

2. Housekeep: Task and Dataset

The central challenge of Housekeep is understanding how humans prefer to put everyday household objects in an organized and disorganized house. We want to capture where objects are typically found in an unorganized house (before tidying the house), and in a tidy house where objects are kept in their correct position (after the person has tidied the house). To this end, we run a study on Amazon MTurk [9, 24] with 372 participants. Each participant is shown an object (*e.g.* salt-shaker), a room (*e.g.* kitchen) for context, and asked to classify all the receptacles present in the room as either *misplaced*, *correctly placed*, or at an *implausible placement*.

We then use this dataset of human annotations to generate episodes where some objects start misplaced. We also use the dataset to compute evaluation metrics for the agent's rearrangements. For example "Episode Success" is measured by if the majority of reviewers agree if the final object placements are correct at the episode end. Likewise, we define "Object Success" as the fraction of objects correctly placed. "Soft Object Success" is the ratio of reviewers that agree the object is placed correctly. Finally, "Rearrange Quality" is a normalized value in $[0, 1]$ (via mean reciprocal rank [8]) given to each object-receptacle based on the ranking collected from human preferences, with 0 given if misplaced.

Rank	Explore	Episode Success ↑	Object Success ↑	Soft Object Success ↑	Rearrange Quality ↑
OR	OR	1.00 $\pm$ 0.00	1.00 $\pm$ 0.00	0.64 $\pm$ 0.00	0.61 $\pm$ 0.00
OR	FTR	0.35 $\pm$ 0.02	0.65 $\pm$ 0.01	0.49 $\pm$ 0.01	0.40 $\pm$ 0.01
LM	OR	0.02 $\pm$ 0.00	0.32 $\pm$ 0.01	0.42 $\pm$ 0.00	0.20 $\pm$ 0.01
LM	FTR	0.01 $\pm$ 0.00	0.23 $\pm$ 0.01	0.36 $\pm$ 0.00	0.14 $\pm$ 0.01

Table 1. Results using our modular baseline on the Housekeep TEST-SEEN and TEST-UNSEEN splits. OR : Oracle, LM : LLM-based ranking, FTR : Frontier exploration.

Our baseline breaks the multi-stage rearrangement into three natural components: a) exploration and mapping, b) planning, and c) navigation and rearrangement. The planning module communicates with all the other modules and determines what the agent does (explore or rearrange). The exploration module uses frontier-based exploration [30]. While navigating, the agent stores a list of locations of discovered objects and receptacles. From this list, it produces a list of object-receptacle pairs indicating the order of rearrangements to perform via a submodule which ranks potential object-receptacle pairings by modeling the joint distribution $\mathbb{P}(\text{receptacle}, \text{room}|\text{object})$. To learn the probability scores for the ranking, we start by extracting word embeddings from a pretrained RoBERTa LLM [19] of all objects and receptacles. We experiment with various contextual prompts [21, 22] for extracting embeddings of paired (*e.g.* "(object) in (room)") combinations. We use RoBERTa as our base LLM and finetuning by Contrastive Matching on the collected dataset. We utilize the LLM as a scoring function within the RANKER module to continuously rerank (thus replan) newly discovered rooms and receptacles while exploring Housekeep episodes.

In Table 1, we show results on the test split with unseen object types, scenes, and rearrangement problems. The first row with oracle ranking and exploration denote the upper bounds achievable across all metrics. Compared to the oracle ranker, the language model impacts object success (OS) by -68%, and episode success by -98%. The dramatic drop is expected as Housekeep is a multi-step problem with compounding errors between rearrangements. Using Frontier exploration, object success drops by 65%. This drop in performance signifies the importance of task-driven exploration to better find misplaced objects or correct receptacles. Finally, we evaluate the fully non-oracle baseline (last row) which achieves a 23% object success rate. This performance on unseen objects supports our claim that LLMs can indeed serve as a generalizable planning module aligned with human preferences.

In this work we presented the Housekeep benchmark to evaluate commonsense reasoning in the home for embodied AI. Future work will explore a learned exploration and reasoning module to learn semantic priors in exploration and increase accuracy of identifying misplaced objects.

References

[1] Nichola Abdo, C. Stachniss, Luciano Spinello, and Wolfram Burgard. Robot, organize my shelves! tidying up objects by predicting user preferences. *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015. 1

[2] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. 1

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian D. Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018. 1

[4] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020. 1

[5] Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 2020. 1

[6] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. 1

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[8] Nick Craswell. Mean reciprocal rank. In *Encyclopedia of Database Systems*, 2009. 2

[9] Kevin Crowston. Amazon mechanical turk: A research tool for organizations and information systems scholars. In *Shaping the future of ict research. methods and approaches*. 2012. 2

[10] Angel Daruna, Weiyu Liu, Zsolt Kira, and Sonia Chernova. Robocse: Robot common sense embedding, 2019. 1

[11] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018. 1

[12] Kiana Ehsani, Winson Han, Alvaro Herrasti, Eli VanderBilt, Luca Weihs, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. ManipulaTHOR: A framework for visual object manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 1

[13] Chuang Gan, Jeremy I Schwartz, Seth Alter, Martin Schrimpf, James Traer, Júlian Letícia de Freitas, J. Kubilius, Abhishek Bhandwaldar, N. Haber, M. Sano, Kuno Kim, Elias Wang, Damian Mrowca, Michael Lingelbach, Aidan Curtis, Kevin T. Feigelis, Daniel M. Bear, Dan Gutfreund, David Cox, James J. DiCarlo, J. McDermott, J. Tenenbaum, and Daniel L. K. Yamins. Threedworld: A platform for interactive multi-modal physical simulation. *NeurIPS*, abs/2007.04954, 2020. 1

[14] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. IQA: visual question answering in interactive environments. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018. 1

[15] Collins Jasmine, Goel Shubham, Luthra Achleshwar, Xu Leon, Deng Kenan, Zhang Xi, Yago Vicente Tomas F, Arora Himanshu, Dideriksen Thomas, Guillaumin Matthieu, and Malik Jitendra. Abo: Dataset and benchmarks for real-world 3d object understanding. *arXiv preprint arXiv:2110.06199*, 2021. 2

[16] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*, 2017. 1

[17] Weiyu Liu, Dhruva Bansal, Angel Daruna, and Sonia Chernova. Learning Instance-Level N-Ary Semantic Knowledge At Scale For Robots Operating in Everyday Environments. In *Proceedings of Robotics: Science and Systems*, 2021. 1

[18] Weiyu Liu, Chris Paxton, Tucker Hermans, and Dieter Fox. Structformer: Learning spatial structure for language-guided semantic rearrangement of novel objects. *arXiv preprint arXiv:2110.10189*, 2021. 1

[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*, 2019. 2

[20] Medhini Narasimhan, Erik Wijmans, Xinlei Chen, Trevor Darrell, Dhruv Batra, Devi Parikh, and Amanpreet Singh. Seeing the un-scene: Learning amodal semantic maps for room navigation. *CoRR*, abs/2007.09841, 2020. 1

[21] Fabio Petroni, Patrick Lewis, Aleksandra Piktus, Tim Rocktäschel, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. How context affects language models' factual predictions. In *Automated Knowledge Base Construction*, 2020. 2

[22] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. 2

[23] Google Research. Google Scanned Objects. https://app.ignitionrobotics.org/GoogleResearch/fuel/collections/Google%20Scanned%20Objects, 2020. [Online; accessed Feb-2022]. 2

[24] Matthew J. Salganik. *Bit by Bit: Social Research in the Digital Age*. Open review edition edition, 2017. 2

[25] Bokui Shen, Fei Xia, Chengshu Li, Roberto Martín-Martín, Linxi Fan, Guanzhi Wang, Shyamal Buch, Claudia D'Arpino, Sanjana Srivastava, Lyne P Tchapmi, et al. igibson, a simulation environment for interactive tasks in large realistic scenes. *arXiv preprint arXiv:2012.02924*, 2020. 2

[26] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020. 1

[27] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[28] Saim Wani, Shivansh Patel, Unnat Jain, Angel X. Chang, and Manolis Savva. Multion: Benchmarking semantic map memory using multi-object navigation. In *NeurIPS*, 2020. 1

[29] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied question answering in photorealistic environments with point cloud perception. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019. 1

[30] Brian Yamauchi. A frontier-based approach for autonomous exploration. In *cira*, volume 97, 1997. 2

[31] B. Çalli, A. Singh, Aaron Walsman, S. Srinivasa, P. Abbeel, and A. Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. *2015 International Conference on Advanced Robotics (ICAR)*, 2015. 2