BEHAVIOR in Habitat 2.0: Simulator-Independent Logical Task Description for Benchmarking Embodied AI Agents

Ziang Liu[1], Roberto Martín-Martín[1], Fei Xia[2], Jiajun Wu[1], Li Fei-Fei[1]
[1]Stanford University [2]Google Research
{ziangliu, robertom, feixia, jiajunwu, feifeili}@cs.stanford.edu

1. Introduction

Robots excel in performing repetitive and precision-sensitive tasks in controlled environments such as warehouses and factories [2]. However, this excellence has not been yet extended to embodied AI agents providing assistance in uncontrolled environments, i.e. assisting humans in everyday tasks at home. Inspired by the catalyzing effect that benchmarks have played in the AI fields such as computer vision [4, 8] and natural language processing [12], the community is looking for new benchmarks for embodied AI, often running on simulation to leverage their safety, reproducibility and speed. Different to the very clear and uniformly adopted definitions of success in CV and NLP benchmarks, in embodied AI each benchmark defines tasks using a different formalism, often specific to one environment, simulator or domain, making it hard to develop general and comparable solutions.

The most common ways to define embodied AI tasks include geometric, image, language, experience, and predicate [1]. The most adopted is geometry: manually defining regions to place objects (rearrangement) or the robot (navigation) to claim success [5, 6, 11, 14, 15]. While providing an exact guidance to the agent, this formulation requires explicit knowledge of the object/robot valid goal poses for each scene, involving a manual process that does not generalize to other scenes. Experience-based goal allow agents to collect observations in the goal environment [1, 13]. However, despite its simplicity, providing a goal environment in the real world is challenging. Language goal describe configurations with natural language [9]. This formulation is the closest to defining in the logic domain and more interpretable to human, but is less concise and adds the challenge of language understanding. To alleviate the aforementioned limitations, we note that using logic predicates to define tasks provides more generalizability to different scenes and simulators, and is closer to real world task definitions.

BEHAVIOR [10] is a set of 100 household activities for evaluating embodied AI agents defined in BEHAVIOR Domain Definition Language (BDDL) with synsets and logic predicates instead of grounded object instances and representations that depend on simulator features, providing a level of abstraction that can be adapted to any simulator and scene assets while allowing a flexible configuration space similar to how humans define tasks in the real world. Although BEHAVIOR is simulator-agnostic, so far it has only been integrated with iGibson 2.0 (iG 2.0) [7]. Recent release of Habitat 2.0 (H2.0) [11] shows a promising test bed for BEHAVIOR, as they provide a significantly higher simulation speed and thus allowing more experiences in the same time period.

In this work, we bring 45 out of the 100 BEHAVIOR activities which involve only kinematic states into H2.0 to benefit from its fast simulation speed as a first step towards demonstrating the ease of adapting activities defined in the logic space into different simulators, in the process equip H2.0 with a even richer set of iG 2.0 interactive scenes and assets.

2. BEHAVIOR in H2.0

Fully supporting BEHAVIOR activities in a new simulator imposes five requirements, as stated in Section 5 in the BEHAVIOR paper [10]: 1) object-centric representation, 2) simulate physics and sensor signals, 3) non-kinematic states, 4) instance sampling from BDDL conditions, 5) state predicate checking. H2.0 naturally satisfies requirement 2) through its variety of sensor signals and the physics simulation through Bullet [3]. In this work, we extend H2.0 for the requirements 1), 4) and 5). For 1), we extend the H2.0 simulator to keep track of the additional object-centric state information needed for evaluating activity progress with BDDL. For 4), we enable H2.0 to use iG 2.0 assets and leverage the sampled instances from iG 2.0. For 5), we implement the pipeline to evaluate each kinematic state predicate. The missing requirement 3) is a current limitation of our effort: we have only kinematic states. This restricts our effort to support 45 out of 100 BEHAVIOR activities.

Loading BEHAVIOR Instances in H2.0. Many objects in daily household activities require interaction with ob-

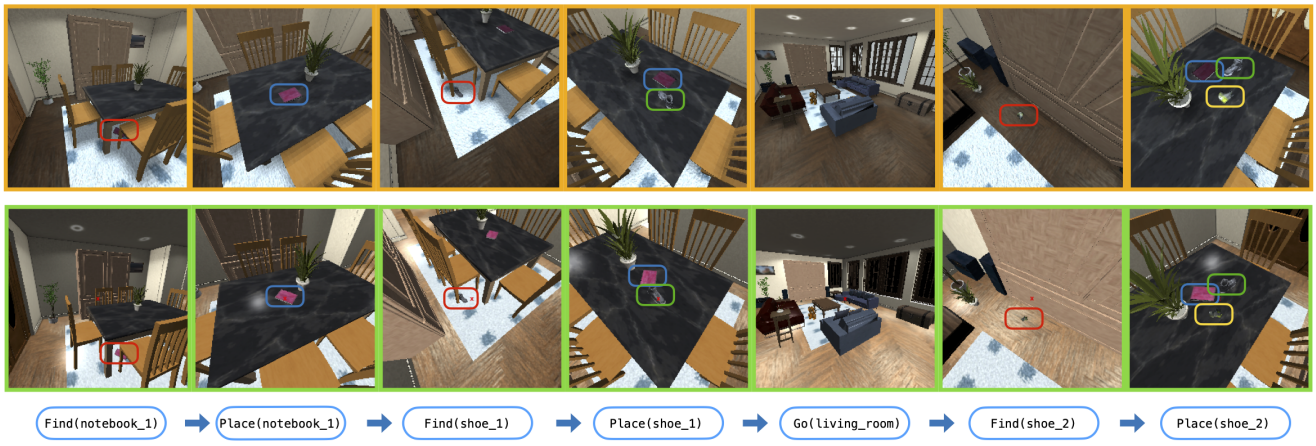

Figure 1. Performing one episode of BEHAVIOR activity to *collect_misplaced_item* through teleoperation. Top row: observation key-frames from iG 2.0. Bottom row: observation key-frames from H2.0.

jects' articulation mechanisms, from loading dishes into a dish washer to opening doors and beyond. For simulators to support scenes that closely resemble real world scenarios, having more articulated objects that represents their real world counterparts in various scene layouts is highly desirable. H2.0's ReplicaCAD dataset lacks abundance in object categories, articulated objects, and scene layouts, despite the richness in carefully designed room configurations, as shown in Table 1. Adding iG 2.0 scenes and assets allows H2.0 users to train and evaluate their AI agents with far more diverse environments and object set, and in particular more articulated objects to interact with.

Checking Predicates for BEHAVIOR Activities in H2.0. BEHAVIOR requires seven kinematic states (NextTo, OnTop, etc.) and fourteen non-kinematic states (Burnt, Sliced, etc.). In this work we focus on implementing kinematic states in H2.0 that are essential for many BEHAVIOR activities. We provide a BDDL backend for H2.0 that supports predicate checking for *NextTo* , *Inside*, *OnFloor*, *OnTop*, *Touching*, and *Under*. For validating task completion progress and task success, we leverage the logic evaluation mechanism from BDDL. Overall, our effort facilitates training and evaluating on 45 out of 100 BEHAVIOR activities.

3. Experimental Validation

To demonstrate and validate our implementation, we perform an episode of the *collect_misplaced_item* activity in the *Wainscott_0_int* apartment in both iG 2.0 and H2.0 through teleoperation.

The captured key-frames in Figure 1 correspond to observations when performing the activity. Benefited from BEHAVIOR's logic domain specification, we are able to implement the same activities in two different simulators without altering the activity definition in any way. Note that the differences in object appearances are due to lighting setup and using non-pbr rendering in H2.0.

Asset	Apt.	Rm.	Cat.	Obj.	A.O.
BEHAVIOR	**15**	**100**	**391**	**1217**	**339**
ReplicaCAD	1	1	41	1201	8

Table 1. BEHAVIOR (iG 2.0) and ReplicaCAD (H2.0) assets comparison, based on the number of apartments, rooms, layouts, object categories, objects, and articulated objects.

Performance Comparison of iG 2.0 vs. H2.0. One of our goals in bringing BEHAVIOR to H2.0 is to gain performance benefit. Our effort enables a fair performance comparison of iG 2.0 and H2.0 with the same assets.

From our evaluation, H2.0 provided 10.4x speed up in an iG 2.0 scene with 64 processes on 8 GPUs. However, as the number of objects increase, the performance benefit of H2.0 over iG 2.0 decreases to 1.5x with 16 processes on 1 GPU and 1.25x with 64 processes on 8 GPUs, and 0.94x on a single process.

4. Next Steps

In this work, we ported BEHAVIOR household activities into H2.0, demonstrating that defining tasks in the high-level logic domain allows simple implementation of the tasks in different simulators. To further demonstrate the behavior of agents trained on the same task in different simulators, we plan to provide simple baseline training results in both iG 2.0 and H2.0, and release the code publicly to facilitate research in this direction. As a main limitation, our work currently only enabled activities with kinematic states; a natural extension is to implement the relevant extended object states and predicate checking mechanisms for the non-kinematic states to support even more BEHAVIOR activities.

References

[1] Dhruv Batra, Angel X. Chang, Sonia Chernova, Andrew J. Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, Manolis Savva, and Hao Su. Rearrangement: A challenge for embodied AI. *CoRR*, abs/2011.01975, 2020. 1

[2] US Census Bureau. Annual survey of manufactures industrial robotic equipment: 2018, Dec 2021. 1

[3] Erwin Coumans. Bullet physics simulation. In *ACM SIGGRAPH 2015 Courses*, SIGGRAPH '15, New York, NY, USA, 2015. Association for Computing Machinery. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1

[5] Chuang Gan, Jeremy Schwartz, Seth Alter, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, Kuno Kim, Elias Wang, Damian Mrowca, Michael Lingelbach, Aidan Curtis, Kevin T. Feigelis, Daniel M. Bear, Dan Gutfreund, David D. Cox, James J. DiCarlo, Josh H. McDermott, Joshua B. Tenenbaum, and Daniel L. K. Yamins. Threedworld: A platform for interactive multi-modal physical simulation. *CoRR*, abs/2007.04954, 2020. 1

[6] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. Rlbench: The robot learning benchmark & learning environment. *CoRR*, abs/1909.12271, 2019. 1

[7] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Elliott Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, Andrey Kurenkov, Karen Liu, Hyowon Gweon, Jiajun Wu, Li Fei-Fei, and Silvio Savarese. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In *5th Annual Conference on Robot Learning*, 2021. 1

[8] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *arXiv preprint arXiv:2109.13410*, 2021. 1

[9] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[10] Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, C. Karen Liu, Silvio Savarese, Hyowon Gweon, Jiajun Wu, and Li Fei-Fei. BEHAVIOR: benchmark for everyday household activities in virtual, interactive, and ecological environments. *CoRR*, abs/2108.03332, 2021. 1

[11] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[12] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461, 2018. 1

[13] Luca Weihs, Matt Deitke, Aniruddha Kembhavi, and Roozbeh Mottaghi. Visual room rearrangement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 1

[14] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. *CoRR*, abs/1910.10897, 2019. 1

[15] Yuke Zhu, Josiah Wong, Ajay Mandlekar, and Roberto Martín-Martín. robosuite: A modular simulation framework and benchmark for robot learning. *CoRR*, abs/2009.12293, 2020. 1