BEHAVIOR-1K: A Benchmark for Embodied AI with 1,000 Everyday Activities and Realistic Simulation

Chengshu Li*[1], Ruohan Zhang*[1], Josiah Wong*[1], Cem Gokmen*[1], Sanjana Srivastava*[1], Roberto Martín-Martín*[3,4], Chen Wang*[1], Gabrael Levine*[1], Michael Lingelbach[1], Jiankai Sun[1], Mona Anvari[1], Minjune Hwang[1], Manasi Sharma[1], Arman Aydin[1], Dhruva Bansal[1], Samuel Hunter[1], Kyu-Young Kim[1], Alan Lou[1], Caleb R Matthews[1], Ivan Villa-Renteria[1], Jerry Huayang Tang[1], Claire Tang[1], Fei Xia[1], Silvio Savarese[1,2,4], Hyowon Gweon[1,2], C. Karen Liu[1,2], Jiajun Wu[1,2], Li Fei-Fei[1,2]

Stanford University[1], Stanford Institute for Human-Centered AI (HAI)[2], The University of Texas at Austin[3], Salesforce Research[4]

Abstract: We present BEHAVIOR-1K, a comprehensive simulation benchmark for human-centered robotics. BEHAVIOR-1K includes two components, guided and motivated by the results of an extensive survey on '*what do you want robots to do for you?*'. The first is the definition of 1,000 everyday activities, grounded in 50 scenes (houses, gardens, restaurants, offices, etc.) with more than 5,000 objects annotated with rich physical and semantic properties. The second is OMNIGIBSON, a novel simulation environment that supports these activities via realistic physics simulation and rendering of rigid bodies, deformable bodies, and liquids. Our experiments indicate that the activities in BEHAVIOR-1K are long-horizon and dependent on complex manipulation skills, both of which remain a challenge for even state-of-the-art robot learning solutions. To calibrate the simulation-to-reality gap of BEHAVIOR-1K, we provide an initial study on transferring solutions learned with a mobile manipulator in a simulated apartment to its real-world counterpart. We hope that BEHAVIOR-1K's human-grounded nature, diversity, and realism make it valuable for embodied AI and robot learning research. Project website: https://behavior.stanford.edu/.

Keywords: Embodied AI Benchmark, Everyday Activities, Mobile Manipulation

1 Introduction

Inspired by the progress that benchmarking brought to computer vision [1–11] and natural language processing [12–16], the robotics community has developed several benchmarks in simulation [17–30]. The broader goal of these benchmarks is to fuel the development of general, effective robots that bring major benefits to people's daily lives – human-centered AI that "serves human needs, goals, and values" [31–34]. Inspiring as they are, the tasks and activities in those benchmarks are designed by researchers; it remains unclear if they are addressing the actual needs of humans.

We observe that a human-centered robotic benchmark should not only be designed *for* human needs, but also originated *from* human needs: *what everyday activities do humans want robots to do for them?* To this end, we conduct an extensive survey with 1,412 participants (see Sec. 2) to rank a wide range of daily activities based on participants' desire to delegate these activities to robots. We also ask layperson annotators to provide definitions of those activities. The survey reveals systematicity in what activities people want robots to do, but more importantly, highlights two key factors that we should prioritize when designing robotic benchmarks: **diversity** in the type of scenes, objects, and activities, and **realism** of the underlying simulation environments.

* indicates equal contribution
correspondence to {chengshu,zharu}@stanford.edu

6th Conference on Robot Learning (CoRL 2022), Auckland, New Zealand.

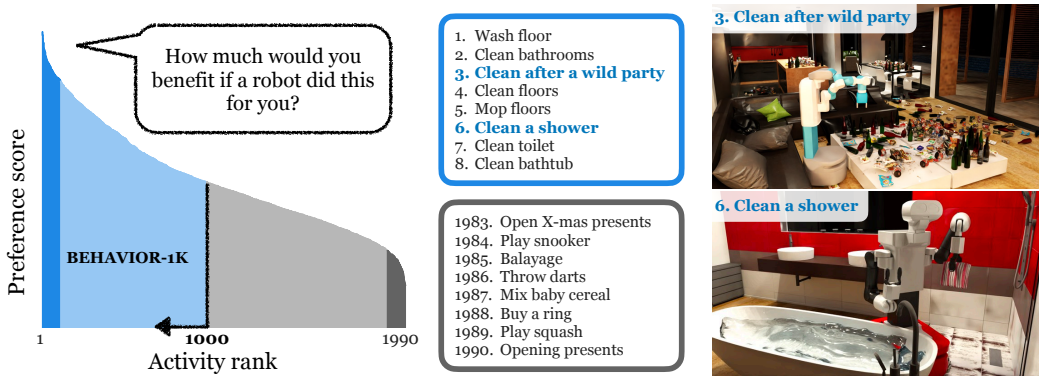

Figure 1: **Developing a Human-Centered Benchmark for Embodied AI.** Left: human preference score over 1990 activities, ranked based on a survey on 1412 participants. The distribution indicates the high **diversity** of needs and preferences of humans that should be reflected in a comprehensive benchmark. Middle: Example activities. Laborious activities are ranked the highest, while pleasurable ones are ranked the lowest. Right: visualization of two of the top 8 activities generated by our **realistic** OMNIGIBSON simulation environment.

The most needed activities indicated by the survey range from 'cleaning your house after a wild party' to 'scrubbing the bathroom floor.' Clearly, the diversity of these activities is far beyond what real-world robotics challenges may offer [35–42]. Developing simulation environments is a natural alternative: one can train and test robotic agents in many activities with diverse scenes, objects, and conditions efficiently and safely. However, for this paradigm to work, the activities have to be simulated realistically, reproducing accurately the circumstances that a robot may encounter in the real world. While significant progress in realism has been made in specific domains [43–45], achieving realism for a diverse set of activities remains a tremendous challenge, due to the effort required to provide realistic models and simulation features.

In this work, we present **BEHAVIOR-1K**, a Benchmark of 1,000 Everyday Household Activities in Virtual, Interactive, and Ecological Environments – the next generation of BEHAVIOR-100 [27]. BEHAVIOR-1K includes two novel components to address the demands for diversity and realism: the diverse **BEHAVIOR-1K DATASET** and the realistic **OMNIGIBSON** simulation environment. The BEHAVIOR-1K DATASET is a large-scale dataset comprising 1) a common-sense knowledge base for 1,000 activities with definitions in predicate logic (initial and goal conditions), as well as the objects involved, their properties, and their state transitions, and 2) high-quality 3D assets including 50 scenes and 5,000+ object models with rich physical and semantic annotations.

All activities in the BEHAVIOR-1K DATASET are instantiated in a novel simulation environment, OMNIGIBSON, which we build on top of Nvidia's Omniverse and PhysX 5 [46] to provide realistic physics simulation and rendering of rigid bodies, deformable bodies, and fluids. OMNIGIBSON expands beyond Omniverse's capabilities with a set of extended object states like temperature, toggled, soaked, and dirtiness. It also includes capabilities to generate valid initial activity configurations and discriminate valid goal solutions based on activity definitions. With all these realistic simulation features, OMNIGIBSON supports the 1,000 diverse activities in the BEHAVIOR-1K DATASET.

We evaluate state-of-the-art reinforcement learning algorithms [47, 48] in several activities of BEHAVIOR-1K, both with visuomotor control in the original action space, and with action primitives that leverage sampling-based motion planning [49]. Our analysis indicates that even a single activity in BEHAVIOR-1K is extremely challenging for current AI algorithms, and the baselines can only solve it with a significant injection of domain knowledge. Concretely, the difficulties derive in part from the length of BEHAVIOR-1K's activities and the complexity of the physical manipulation required. To calibrate the simulation-to-real gap of BEHAVIOR-1K, we provide an initial study on transferring solutions learned with a mobile manipulator in a simulated apartment to its real-world counterpart. We hope that the BEHAVIOR-1K benchmark, our survey, and our analysis will serve to support and guide the development of future embodied AI agents and robots.

2 Creating a Benchmark Grounded in Human Needs: A Survey Study

A significant amount of robotics research aspires to satisfy human needs, but those needs are typically assumed or speculated. Human-centric development requires direct information about what humans

		BEHAVIOR-1K	BEHAVIOR-100	AI2THOR Vis. Room Rearr.	TDW Transport	Rearr. T5 (Habitat)	ManipulaTHOR ArmPointNav	Interactive Gibson Benchmark	VirtualHome	ALFRED	Habitat 2.0 HAB	SAPIEN ManiSkill	Watch-And-Help	RFUniverse	Rearr. T2 (OCRTOC)	IKEA Furniture Assembly	RLBench	Metaworld	Robosuite	SoftGym	DeepMind Control Suite	OpenAIGym	Habitat 1.0	Gibson
		Mobile manipulation													Static manipulation								Navigation	
	Activities from human preference survey	✔	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘
Diversity	Activities	**1000**	100	1	1	1	1	2	549	7	3	4	5	5	5	100	50	1	5	10	28	8	2	3
Diversity	Scene types	**8**	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Diversity	Scenes / rooms	**50/306**	15/100	-/120	15/105	55 static/-	-/30	10/-	6/24	-/120	1/6	1/-	7/29	Gibson	1/-	1/-	1/-	1/-	1/-	1/-	1/-	1/-	HM3D	572 static
Diversity	Object categories	**1265**	391	118	50	YCB	150	5	308	84	41+YCB	4	117	UNK	12+YCB	73+	28	7	10	4	4	4	Mtpt.	N/A
Diversity	Object models	**5215**	1217	118	112	YCB	150	152	UNK	84	92+YCB	162	UNK	UNK	101+YCB	73+	28	80	10	4	4	4	N/A	N/A
Diversity	Objs. per activity	**3-47**	3-34	5	7-9	2-5	2-3	10	1-24	2	5	1	2-8	1-6	5-10	1-2	1-2	1	1-3	1-3	1-3	1	0-1	N/A
Diversity	Diff. state changes required per activity	**2-11**	2-8	4	4	4	2	1-3	1-7	2-3	1-2	1	1-3	1	1	1-3	1-4	4	1	1-3	1-2	1-2	1	1
Diversity	Infinite scene-agnostic instantiation	✔	✔	✘	✘	✘	✘	✘	✘	✔	✔	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	N/A	N/A
Realism	Visual quality score	**3.20**	1.69	1.73	1.65	-	1.73	-	-	1.73	1.74	-	-	-	-	-	-	-	-	-	-	-	-	-
Realism	Kinematics, dynamics	✔	✔	✔	✔	✔	✔	✔	✘	✔	✔	✔	✘	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Realism	Continuous extended states (temp., wetness)	✔	✔	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘
Realism	Flexible materials	✔	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✔	✘	✘	✘	✘	✘	✔	✘	✘	✘	✘
Realism	Deformable bodies	✔	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✔	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘
Realism	Realistic Fluid	✔	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✔	✘	✘	✘	✘	✘	✔	✘	✘	✘	✘
Realism	Thermal effects (steam/fire)	✔	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘	✔	✘	✘	✘	✘	✘	✘	✘	✘	✘	✘
Realism	Realistic action execution	✔	✔	✘	✔	✘	✔	✔	✘	✘	✔	✔	✘	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
	Benchmark focus: Task-Planning and/or Control	TP+C	TP+C	TP	TP+C	TP+C	TP+C	C	TP	TP	TP+C	C	TP	TP+C	TP+C	C	TP+C	C	C	C	C	C	C	C

Table 1: **Comparison of Embodied AI Benchmarks:** BEHAVIOR-1K contains 1,000 diverse activities that are grounded by human needs. It achieves a new level of diversity in scenes, objects, and state changes involved. OMNIGIBSON provides realistic simulation of these 1,000 activities, including some of the most advanced simulation and rendering features such as fluid and deformable bodies. This table is extended from [27].

want from autonomous agents [31]. To create a benchmark that reflects these needs, we conduct a survey targeting the general U.S. population that asks: *what do you want robots to do for you?* The survey sources around 2,000 activities from time-use surveys [50–52], which record how people spend their time, and from WikiHow articles [53]. We conduct the survey on Amazon Mechanical Turk with a total of 1,412 respondents (demographics in Appendix A.3) and fifty 10-point Likert scale responses per activity.

Survey results are summarized in Fig. 1 (left), in which we rank the activities based on their human preference score. The full list of ranked activities can be found on our website. The distribution shows large statistical dispersion (Gini index=0.158): humans want robots to perform a wide range of activities, from cleaning chores to cooking large feasts. Tedious tasks like "scrubbing the bathroom floor" score the highest, while recreational activities like game-play score the lowest. There are over 270 cleaning activities and over 310 cooking activities, among many other categories.

BEHAVIOR-1K activities include the 909 activities with the highest human preference scores and 91 activities from BEHAVIOR-100 [27], altogether the top-ranked 1000 activities. BEHAVIOR-1K sets itself apart from other embodied AI benchmarks by sourcing from time-use surveys and using survey data to prioritize activities considered most important and useful by humans, and by including a tremendously diverse set of activities.

3 Related Work: Embodied AI Benchmarks

We provide an extensive comparison between BEHAVIOR-1K, and other embodied AI benchmarks in simulation [17–26] in Table 1. We include a number of factors that contribute to diversity and realism and observe a significant step forward with BEHAVIOR-1K. First, no other benchmark grounds their activity set on the needs of lay people. Other benchmarks often target a relatively restricted set of activities, and their simulators are realistic only in the relevant aspects for those tasks. In fact, we often observe a diversity-realism tradeoff. For instance, instruction-following benchmarks such as VirtualHome [20] and ALFRED [20, 21] are diverse in the number of scenes, objects, and state changes, but offer a limited low-level physical realism. On the other hand, household rearrangement benchmarks such as Habitat 2.0 HAB [26], TDW Transport [19], and SAPIEN ManiSkill [54, 55] support realistic action execution and accurate physics simulation for rigid bodies, but only include a handful of tasks. Similarly, SoftGym [45] and RFUniverse [56] have the closest simulation features and hence realism to OMNIGIBSON, but they also lack the task diversity needed to support the development of human-centered general robots.

The most similar benchmark to us is the previous generation BEHAVIOR-100 [27]. BEHAVIOR-100 brought forward several beneficial design choices that we inherit in BEHAVIOR-1K such as the activity sources (ATUS [50]), activity definition logic language, and evaluation metrics. However, it fell short in the diversity and realism necessary to support a human-centered embodied AI benchmark in simulation, dimensions where BEHAVIOR-1K achieves unmatched levels. While BEHAVIOR-100 comprises 100 activities selected by researchers, our BEHAVIOR-1K increases diversity by one order of magnitude, to 1,000 activities, that are grounded in human needs thanks to our unique survey.

Figure 2: **Elements of BEHAVIOR-1K.** Our benchmark comprises two elements: BEHAVIOR-1K DATASET and OMNIGIBSON. Left: BEHAVIOR-1K DATASET includes 1,000 BDDL activity definitions (top left), 50 realistic and diverse scenes (top right), and 5,000+ objects with properties annotated in the knowledge base (bottom). Right: OMNIGIBSON provides the necessary functionalities to realistically simulate the 1000 activities, including thermal effects such as fire/steam/smoke (top left), fluid dynamics (bottom left), functional machines for transition rules (top center), deformable bodies/cloths (bottom center), realistic lighting and reflections (top right), and transparency rendering (bottom right). Together, they constitute a concrete, realistic instantiation of an everyday activity like `CookingDinner` in simulation.

Furthermore, BEHAVIOR-100 includes only 15 scenes (all houses) and 300+ object categories, while BEHAVIOR-1K increases to 50 scenes (houses, stores, restaurants, offices, etc.) and 1,200+ object categories. In terms of realism, BEHAVIOR-1K extends the simulatable physical states and processes with OMNIGIBSON: fluids, flexible materials, mixing substances, etc. The realism achieved in rendering by OMNIGIBSON for BEHAVIOR-1K is also significantly higher than what was possible in BEHAVIOR-100 and other benchmarks (see Fig. 3).

4 BEHAVIOR-1K DATASET

Once activities have been sourced to reflect human needs, they need to be concretely defined and instantiated the way they would occur in the real world. We build the BEHAVIOR-1K DATASET, which includes a knowledge base of crowdsourced activity definitions with relevant objects and object states, and a large-scale repository of high-quality, interactive 3D models.

We crowdsource concrete definitions of activities in the form of BEHAVIOR Domain Definition Language (BDDL) [27]. BDDL is based on predicate logic and designed to be accessible for laypeople to describe concrete initial and goal conditions for a given activity. Unlike geometric, image/video, or experience goal specifications [17, 18], BDDL definitions are in terms of objects and object states, allowing annotators to define at an intuitive semantic level. The semantic symbols also capture the fact that multiple physical states might be valid initializations and solutions to an activity. See Listing 1, 2 and 3 in Appendix for example definitions.

The object and object state spaces that activity definitions are built upon are annotated to be ecologically plausible. The object spaces are derived from 4,500 WikiHow articles for the 1,000 activities and mapped to 1,484 WordNet [57] synsets. Through crowdworkers, students, and GPT-3 [58], we also associate each object with our fully simulatable object states: for example, `apple` is associated with `cooked` and `sliced`, but not `toggledOn`. Many object-property pairs are augmented with parameters, e.g. "cooked temperature for apples", taking advantage of OMNIGIBSON's continuous extended states to make activities especially realistic. Finally, annotators and researchers also create transition rules, e.g. turning tomatoes and salt into sauces, or requiring sandpaper to remove rust. The result is a knowledge base of tens of thousands of elements underlying 1000 ecologically plausible activity definitions. We ensure annotation quality by having five experienced machine learning annotators verify a subset of all types of annotations and receive extremely high approval rates (>96.8%). See Appendix B for more details about the knowledge base.

The diversity of these activity definitions requires diverse object and scene models. On top of the 15 house scenes from BEHAVIOR-100 [27], we acquire 35 fully interactive scenes across diverse scene types, such as gardens, offices, restaurants, and stores, that are essential for everyday activities. This is unprecedented compared to other benchmarks (see Table 1). We also acquire 5,000+ object instances across 1,200+ categories required by the activities, and annotate rich physical (e.g., friction,

OMNIGIBSON	Habitat 2.0	AI2-THOR	iGibson 2.0	ThreeDWorld
$\mathbf{3.20 \pm 1.23}$	1.74 ± 1.33	1.73 ± 1.37	1.69 ± 1.24	1.65 ± 1.23

Figure 3: **Comparison of Visual Realism:** We evaluate OMNIGIBSON's visual realism against other simulation environments by running a survey with 60 human subjects. We ask them to rank the realism of sampled images from each environment with a score of 5 (most realistic) to 1 (least realistic). We report the mean and standard deviation and show a sampled image from the study. We observe that the participants consider OMNIGIBSON to be significantly more visually realistic than all other environments. See Appendix E.2 for more info.

mass, articulation) and semantic properties (e.g., category) for each object. Representative scene and object models can be seen in Fig. 2. More details of 3D models can be found in Appendix D.

5 OMNIGIBSON: Instantiating BEHAVIOR-1K with Realistic Simulation

BEHAVIOR-100 is implemented in iGibson 2.0 [59]; however, realistic simulation of the diverse activities in BEHAVIOR-1K is beyond the capability of iGibson 2.0. We present a novel simulation environment, OMNIGIBSON, that provides the necessary functionalities to support and instantiate BEHAVIOR-1K. OMNIGIBSON is built on top of Nvidia Omniverse and PhysX 5, providing the simulation of not only rigid bodies, but also deformable objects, fluids, and flexible materials (see Fig. 4), while generating highly realistic ray-traced or path-traced virtual images (see Fig. 3). These features significantly boost the realism of BEHAVIOR-1K compared to other benchmarks.

Similar to BEHAVIOR-100, OMNIGIBSON also simulates additional, non-kinematic extended object states (e.g. temperature, soaked level) based on heuristics (e.g. temperature increases when being next to a heat source that is toggled on). OMNIGIBSON also implements the functionalities to generate infinite valid physical configurations that satisfy the activities' initial conditions as logical predicates (e.g. `food` is `frozen`), and to evaluate their goal conditions (e.g. `food` is `cooked` and `onTop` of a `plate`, the `cloth` is `folded`) based on the object's physical states (pose and joint configuration) and extended states. OMNIGIBSON natively supports randomization during scene initialization, and can sample amongst object models and their poses/states. The full details of extended object states and logical predicates that OMNIGIBSON supports can be found in Appendix E.1.

Many everyday tasks are difficult to simulate because they require modeling complex physical processes, such as folding a towel or pouring a glass of water. OMNIGIBSON unlocks them by supporting realistic simulation of fluids, deformable bodies, and cloths (see Fig. 2). Indeed, without these features, over half of BEHAVIOR-1K activities would not be simulatable, highlighting how crucial these features are for capturing everyday activities. OMNIGIBSON also captures multiple physical processes that are not natively simulatable by Omniverse, such as baking pies or pureeing vegetables. Aside from the extended states mentioned above, we also design a modular *Transition Machine*, which specifies custom transitions between groups of objects when specified conditions are met. For example, a dough placed inside an oven that reaches a certain temperature threshold will turn into a pie. This further expands OMNIGIBSON's capacity to simulate complex, realistic activities that would otherwise be intractable to fully simulate physically.

6 Experiments: Evaluating Embodied AI Solutions in BEHAVIOR-1K

In our experiments, we aim to answer three questions: How do existing vision-based robot learning algorithms perform in BEHAVIOR-1K, and what assumptions have to be made to improve their success? What elements of the activities are the most problematic for current AI? What are the main sources of the sim-real gap in BEHAVIOR-1K/OMNIGIBSON? Our goal is to indicate promising research directions to improve AI's performance in BEHAVIOR-1K activities in simulation and, ultimately, in the real world.

6.1 Evaluating BEHAVIOR-1K Solutions in OMNIGIBSON

Experimental Setups. We selected three paradigmatic activities for our experiments: `CollectTrash`, where the agent gathers empty bottles and cups, and throws them into a trash bin (rigid body manipulation); `StoreDecoration`, where the agent stores items into a drawer (articulated object

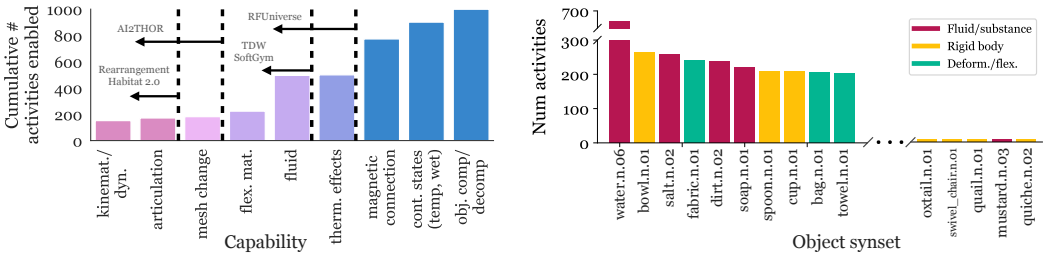

Figure 4: **Objects and States in Activity Definitions:** Left: the number of activities unlocked by each simulation capability that OMNIGIBSON has. None of the other simulation environments are sufficient to fully support BEHAVIOR-1K, e.g. Habitat 2.0 can support only 23% of the activities. Right: the number of activities that require each object synset (category). Several top-10 object synsets are fluids and flexible materials, necessitating the development of OMNIGIBSON. As we expect, the object synsets also follow a long-tail distribution: most objects are involved in only a few activities.

Method	Policy Features		Task success rate		
	Primitives	History	StoreDecoration	CollectTrash	CleanTable
RL-VMC	✗	✗	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
RL-Prim.	✓	✗	0.48 ± 0.06	0.42 ± 0.02	0.77 ± 0.08
RL-Prim.Hist.	✓	✓	0.55 ± 0.05	0.63 ± 0.03	0.88 ± 0.02

Table 2: Task success rates across three baseline methods. RL-VMC with end-to-end visuomotor control completely fails to solve any of the activities, whereas RL-Prim. and RL-Prim.Hist. with action primitives are able achieve decent performance. Memory of observations helps in longer horizon activities (e.g. CollectTrash).

manipulation); and CleanTable, where the agent wipes a dirty table with a soaked piece of cloth (manipulation of flexible materials and fluids). We evaluate three different baselines based on state-of-the-art reinforcement learning algorithms (RL) [60]:

- RL-VMC, a visuomotor control (from image to low-level joint commands) RL solution based on Soft Actor-Critic (SAC) [48];
- RL-Prim., a RL solution based on PPO [47] that leverages a set of action primitives based on a sampling-based motion planner [61, 62, 49] (pick, place, push, navigate, dip and wipe). The policy outputs a discrete selection of a primitive applied on an object;
- RL-Prim.Hist., a variant of RL-Prim. that takes in the history observations (3 steps) as additional inputs to help disentangle similar-looking states.

All agents are trained with a sparse task success reward without any reward engineering. Following the metrics proposed in BEHAVIOR-100 [27], we report the success rate and efficiency metrics (distance traveled, time invested, and disarrangement caused) in Table 2 and 3, and the success score Q in Table A.13 in Appendix.

Grasping is a challenging research topic on its own. To facilitate our experiments, we adopt an assistive pick primitive that creates a rigid connection between the object and the gripper if grasping is requested when all fingers are in contact with the object, a stricter form of *StickyMitten* used in prior works [26, 63, 64]. Furthermore, to accelerate training, the action primitives check only the feasibility (e.g., reachability, collisions) of the final configuration, e.g. the grasping pose for pick or the desired location for navigate. If kinematically feasible, the action primitives will directly set the robot state to the final configuration, and continue to simulate from then on. We include an ablation analysis of the effect of these assumptions and simplifications in our evaluation (see Table 4). Further details about our training and evaluation setup can be found in Appendix F.

Results: Task Completion. Table 2 contains task success rates across our baseline methods. The extreme long-horizon in our activities causes the visuor-motor control (RL-VMC) policy to fail in all three activities, potentially due to problems such as credit assignment [65], deep exploration [66, 67], and vanishing gradients [68] as reported by prior works. Our baselines with time-extended action primitives (RL-Prim. and RL-Prim.Hist.) obtain better success, achieving over 40% success rates across all three activities. We observe that longer-horizon activities are more challenging: while CleanTable can be accomplished by executing the optimal sequence of 6 primitive steps,

Method	Metrics in CollectTrash		
	Dist. Nav. [m]	Sim. Time [s]	Kin. Dis. [m]
RL-VMC	27.58 ± 5.95	16.67 ± 0.00	0.00 ± 0.00
RL-Prim.	17.98 ± 2.35	13.95 ± 5.14	12.34 ± 5.01
RL-Prim.Hist.	15.33 ± 2.70	12.48 ± 3.68	10.82 ± 3.90

Table 3: Efficiency metrics across three baseline methods. RL-VMC has low spatial and temporal efficiency because it fails to learn, whereas history information helps remove redundant actions and improve efficiency.

Phys. Realism		Task success rate		
Grasping	Full Motion	StoreDecoration	CollectTrash	CleanTable
✔	✔	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
✖	✔	0.46 ± 0.04	0.36 ± 0.08	0.73 ± 0.03
✖	✖	0.48 ± 0.06	0.42 ± 0.02	0.77 ± 0.08

Table 4: Ablation study of RL-Prim. on the impact of removing the simplifying assumptions of grasping and motion execution during evaluation. We observe a large drop in performance when enabling fully physics-based grasping, but not when enabling full trajectory motion execution.

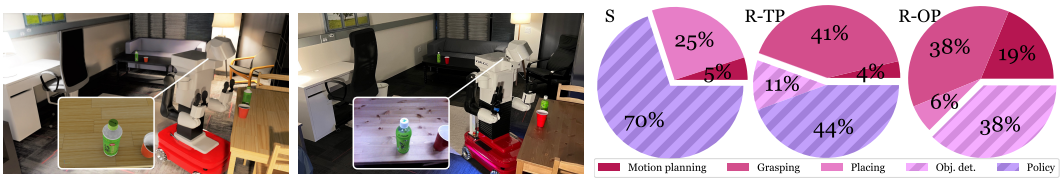

Figure 5: **Characterizing the Sim-Real Gap:** Left: a side-by-side comparison of the simulated and the real scene, including virtual and real images obtained by the robot. While the high resolution images are extremely similar, the mismatch in wooden texture and camera properties causes a sizable gap in the visual input to the agent. Right: source of failure in Simulation (S, left) and in Real-world with a Trained Policy in OMNIGIBSON (R-TP, middle) and an Optimal Policy (R-OP, right) due to actuation (solid color) or perception (striped). In simulation, without full simulation of grasping (see Sec. 6.1), policy failures (i.e., selecting the wrong action primitive) dominate. On the real robot, grasping is one of the main causes of error, as well as perception issues (policy errors with the trained visual policy, object detection errors with the optimal policy).

CollectTrash requires at least 16. This supports the idea that some form of action-space abstraction must be necessary to solve long-horizon activities of BEHAVIOR-1K, as others reported [27, 26, 69]. When analyzing the role of memory, we observe a sizable performance gain from RL-Prim. to RL-Prim.Hist., especially in long-horizon activities with aliased observations such as CollectTrash. In this task, when the robot is looking at the trash bin, it needs additional information to know what location has been cleaned already in order to proceed to other locations. Our results indicate that memory will play a critical role for embodied AI in long-horizon BEHAVIOR-1K activities.

Results: Efficiency. In addition to success, efficiency is also critical in the evaluation of embodied AI: a successful policy in simulation may be infeasible in the real world if it takes a long time or wastes too much energy. In Table 3, we report the results with three efficiency metrics proposed by Srivastava et al. [27]. We observe that the use of memory (RL-Prim.Hist.) improves efficiency across all metrics: distance navigated (Dist. Nav.), simulated time (Sim. Time), and kinematic object disarrangement (Kin. Dis.), i.e., amount of object displacement due to robot motion.

We also evaluate to what extent the simplifications we introduce in physics and actuation (grasping, motion execution) during training impact the performance of RL-Prim. during evaluation when these simplifications are removed. We report the results in Table 4. We observe a radical performance drop after enabling fully physics-based grasping during evaluation. Grasping is thus a critical component of any embodied AI task and researchers should be careful when simplifying its execution during training. While OMNIGIBSON supports fully physics-based grasping, designing a pick action primitive for arbitrary objects that leverages fully physics-based grasping is by itself an open research problem that we leave for future work. In contrast, there is much less performance drop after enabling full trajectory motion execution during evaluation. This result supports our hypothesis that it is reasonable to accelerate the training process by assuming that motion planning is likely to provide viable paths in free space during evaluation.

6.2 Evaluating BEHAVIOR-1K Solutions on a Real Robot

We performed a series of experiments with a real robot to answer the question: *what are the main sources of discrepancy between our realistic simulation and the real world?* To that end, we used a real-world counterpart of the simulated scene of a mockup apartment for the CollectTrash activity. We scanned the apartment and converted it into a virtual, interactive scene. We use a real bi-manual mobile manipulator Tiago, and leverage the RGB-D images from its onboard sensors and a YOLOv3 object detector [70, 71] to localize the objects in 3D space for manipulation. For navigation, the robot localizes with a particle filter [72] based on two LiDAR sensors and a map of the apartment. The

action primitives are implemented with the same sampling-based motion planning algorithm as in simulation [62, 49] with additional tuning. We evaluate two strategies for selecting action primitives in the real world: an optimal policy based on human input, and a vision-based policy (RL-Prim.) trained in OMNIGIBSON. To facilitate sim-to-real transfer, during training we additionally applied image-based data augmentation to the observations based on a prior work [73] (see Appendix G for further details). With the optimal policy, we evaluate the gap in actuation between the simulated and the real robot; with the learned policy, we also evaluate the gap in visual perception. We achieve different success rates in simulation (50 runs, $\sim$40% success) and in real world with optimal (27 runs, $\sim$22%) and trained policies (26 runs, 0%), hinting on a sim-real gap that we analyze below.

The failure cases are depicted in Fig. 5 (right). We observe that the majority of failures in simulation are due to the visual policy (perception), while others are caused by stochasticity in the `place` primitive and the sampling-based motion planner. The reason why none of the failures are due to grasping is because in simulation we evaluate with the assistive `pick` primitive. Grasping is fairly difficult in the real world, contributing to around 40% of the failures for both the trained and the optimal policy. For the learned policy, 44% of the errors come from the visual policy selecting the wrong action primitive due to the differences between the simulated and the real images. The visual discrepancy results from unmodeled effects such as the real camera's poor dynamic range (see Fig. 5, left and middle) and imperfect object modeling (e.g. the exact wooden texture and the surface reflectivity of the tables), which can be alleviated by more targeted domain randomization. Interestingly, several manipulation failures on the real robot are caused by unfavorable robot base placement resulting from navigation inaccuracies in the previous timestep. This compounding source of error is not present in simulation because we assume perfect localization and execution. We believe this analysis provides relevant information about the main sources and severity of the sim-real gap in BEHAVIOR-1K in OMNIGIBSON, and provide insights for future research avenues. Our plan is to use some of these insights to create novel sim-real solutions that make progress on BEHAVIOR-1K.

7 Discussion and Limitations

We presented BEHAVIOR-1K, a benchmark for embodied AI and robotics research with realistic simulation of 1000 diverse activities grounded in human needs. BEHAVIOR-1K comprises two elements: BEHAVIOR-1K DATASET, a semantic knowledge base of everyday activities, and a large-scale 3D model library; OMNIGIBSON, a simulation environment that provides realistic rendering and physics for rigid/deformable objects, flexible materials and fluids. In our evaluation, we observed that BEHAVIOR-1K is an extremely challenging benchmark: solving these 1,000 activities autonomously is beyond the capability of current state-of-the-art AI algorithms. We studied and attempted to solve a handful of the activities with action primitives in order to gain insights into the most challenging components, providing a starting point for other researchers to work on our benchmark. Similarly, we explored the sources of the sim-real gap by creating a digital twin of a real-world mock apartment, and by performing rigorous evaluation and analysis of policies in both simulation and the real world with a simulated and real mobile manipulator.

Limitations: We inherit several limitations from our underlying physics and rendering engine, Nvidia's Omniverse. In OMNIGIBSON, we trade off rendering speed for visual realism (ray-traced), reaching around 60 fps for a house scene of around 60 objects (v.s. around 100 fps in iGibson 2.0 [59]). We are actively working on performance optimization. Another limitation is that we only include activities that do not require interactions with humans. Realistic simulation of humans (behavior, motion, appearance) is extremely challenging and an open research area. We plan to include simulated humans when the technology becomes more mature. Finally, there is still room for improvement in OMNIGIBSON to further facilitate sim2real transfer, such as incorporating noise models of perception and actuation.

Acknowledgments

This work was done in part when Chengshu Li, Josiah Wong, and Michael Lingelbach were interns at Nvidia Research. The work is in part supported by the Stanford Institute for Human-Centered AI (HAI), the Toyota Research Institute (TRI), NSF CCRI #2120095, NSF RI #2211258, NSF NRI #2024247, ONR MURI N00014-22-1-2740, ONR MURI N00014-21-1-2801, Amazon, Bosch, Salesforce, and Samsung. Ruohan Zhang is partially supported by Wu Tsai Human Performance Alliance Fellowship. Sanjana Srivastava is partially supported National Science Foundation Graduate Research Fellowship Program (NSF GRFP).

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[3] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.

[4] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[5] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012.

[6] R. Goyal, S. Ebrahimi Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5842–5850, 2017.

[7] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari. Actor and observer: Joint modeling of first and third-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7396–7404, 2018.

[8] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[9] R. Martín-Martín, M. Patel, H. Rezatofighi, A. Shenoi, J. Gwak, E. Frankel, A. Sadeghian, and S. Savarese. Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[10] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[11] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018.

[12] M. A. Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Using Large Corpora*, page 273, 1994.

[13] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[14] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[15] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, 2013.

[16] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[17] D. Batra, A. X. Chang, S. Chernova, A. J. Davison, J. Deng, V. Koltun, S. Levine, J. Malik, I. Mordatch, R. Mottaghi, M. Savva, and H. Su. Rearrangement: A challenge for embodied ai. *arXiv preprint arXiv:2011.01975*, 2020.

[18] L. Weihs, M. Deitke, A. Kembhavi, and R. Mottaghi. Visual room rearrangement. *arXiv preprint arXiv:2103.16544*, 2021.

[19] C. Gan, S. Zhou, J. Schwartz, S. Alter, A. Bhandwaldar, D. Gutfreund, D. L. Yamins, J. J. DiCarlo, J. McDermott, A. Torralba, et al. The threedworld transport challenge: A visually guided task-and-motion planning benchmark for physically realistic embodied ai. *arXiv preprint arXiv:2103.14025*, 2021.

[20] X. Puig et al. Virtualhome: Simulating household activities via programs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[21] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 10740–10749, 2020.

[22] F. Xia, W. B. Shen, C. Li, P. Kasimbeg, M. E. Tchapmi, A. Toshev, R. Martín-Martín, and S. Savarese. Interactive gibson benchmark: A benchmark for interactive navigation in cluttered environments. *IEEE Robotics and Automation Letters*, 5(2):713–720, 2020.

[23] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, pages 1094–1100. PMLR, 2020.

[24] S. James, Z. Ma, D. Rovick Arrojo, and A. J. Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 2020.

[25] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019.

[26] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. S. Chaplot, O. Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[27] S. Srivastava, C. Li, M. Lingelbach, R. Martín-Martín, F. Xia, K. E. Vainio, Z. Lian, C. Gokmen, S. Buch, K. Liu, et al. Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In *Conference on Robot Learning*, pages 477–490. PMLR, 2022.

[28] Y. Zhu, J. Wong, A. Mandlekar, and R. Martín-Martín. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.

[29] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. d. L. Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

[30] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[31] M. L. Littman, I. Ajunwa, G. Berger, C. Boutilier, M. Currie, F. Doshi-Velez, G. Hadfield, M. C. Horowitz, C. Isbell, H. Kitano, et al. Gathering strength, gathering storms: The one hundred year study on artificial intelligence (AI100) 2021 study panel report. Technical report, Stanford University, 2021.

[32] M. O. Riedl. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1):33–36, 2019.

[33] W. Xu. Toward human-centered ai: a perspective from human-computer interaction. *Interactions*, 26(4):42–46, 2019.

[34] B. Shneiderman. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4):1–31, 2020.

[35] H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda, E. Osawa, and H. Matsubara. Robocup: A challenge problem for ai. *AI magazine*, 18(1):73–73, 1997.

[36] T. Wisspeintner, T. Van Der Zant, L. Iocchi, and S. Schiffer. Robocup@home: Scientific competition and benchmarking for domestic service robots. *Interaction Studies*, 10(3):392–426, 2009.

[37] L. Iocchi, D. Holz, J. Ruiz-del Solar, K. Sugiura, and T. Van Der Zant. Robocup@ home: Analysis and results of evolving competitions for domestic and service robots. *Artificial Intelligence*, 229:258–281, 2015.

[38] M. Buehler, K. Iagnemma, and S. Singh. *The DARPA Urban Challenge: Autonomous Vehicles in City Traffic*, volume 56. springer, 2009.

[39] E. Krotkov, D. Hackett, L. Jackel, M. Perschbacher, J. Pippine, J. Strauss, G. Pratt, and C. Orlowski. The darpa robotics challenge finals: Results and perspectives. *Journal of Field Robotics*, 34(2):229–240, 2017.

[40] N. Correll, K. E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J. M. Romano, and P. R. Wurman. Analysis and observations from the first amazon picking challenge. *IEEE Transactions on Automation Science and Engineering*, 15(1):172–188, 2016.

[41] C. Eppner, S. Höfer, R. Jonschkowski, R. Martín-Martín, A. Sieverling, V. Wall, and O. Brock. Lessons from the amazon picking challenge: four aspects of building robotic systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4831–4835, 2017.

[42] M. A. Roa, M. Dogar, C. Vivas, A. Morales, N. Correll, M. Gorner, J. Rosell, S. Foix, R. Memmesheimer, F. Ferro, et al. Mobile manipulation hackathon: Moving into real world applications. *IEEE Robotics & Automation Magazine*, pages 2–14, 2021.

[43] E. Heiden, M. Macklin, Y. Narang, D. Fox, A. Garg, and F. Ramos. Disect: A differentiable simulation engine for autonomous robotic cutting. *arXiv preprint arXiv:2105.12244*, 2021.

[44] Y. Urakami, A. Hodgkinson, C. Carlin, R. Leu, L. Rigazio, and P. Abbeel. Doorgym: A scalable door opening environment and baseline agent. *arXiv preprint arXiv:1908.01887*, 2019.

[45] X. Lin, Y. Wang, J. Olkin, and D. Held. Softgym: Benchmarking deep reinforcement learning for deformable object manipulation. In *Conference on Robot Learning*, 2020.

[46] Nvidia, Corp. Physx. https://developer.nvidia.com/physx-sdk, 2022. Accessed: 2022-06-10.

[47] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[48] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870. PMLR, 2018.

[49] M. Jordan and A. Perez. Optimal bidirectional rapidly-exploring random trees. Technical Report MIT-CSAIL-TR-2013-021, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 2013.

[50] U.S. Bureau of Labor Statistics. American Time Use Survey. https://www.bls.gov/tus/, 2019.

[51] European Commission. Harmonised european time use surveys. https://ec.europa.eu/eurostat/web/time-use-surveys, 2010.

[52] J. Gershuny, M. Vega-Rapun, and J. Lamote. Multinational time use study. https://www.timeuse.org/mtus, 2020.

[53] wikiHow, Inc. wikihow. https://www.wikihow.com, 2021. Accessed: 2021-06-16.

[54] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, et al. SAPIEN: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11097–11107, 2020.

[55] T. Mu, Z. Ling, F. Xiang, D. Yang, X. Li, S. Tao, Z. Huang, Z. Jia, and H. Su. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. *arXiv preprint arXiv:2107.14483*, 2021.

[56] H. Fu, W. Xu, H. Xue, H. Yang, R. Ye, Y. Huang, Z. Xue, Y. Wang, and C. Lu. Rfuniverse: A physics-based action-centric interactive environment for everyday household tasks. *arXiv preprint arXiv:2202.00199*, 2022.

[57] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

[58] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[59] C. Li, F. Xia, R. Martín-Martín, M. Lingelbach, S. Srivastava, B. Shen, K. E. Vainio, C. Gokmen, G. Dharan, T. Jain, A. Kurenkov, K. Liu, H. Gweon, J. Wu, L. Fei-Fei, and S. Savarese. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In *Annual Conference on Robot Learning*, 2021.

[60] A. G. Barto and S. Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems*, 13(1):41–77, 2003.

[61] S. M. LaValle. *Planning Algorithms*. Cambridge University Press, 2006.

[62] J. J. Kuffner and S. M. LaValle. Rrt-connect: An efficient approach to single-query path planning. In *Proceedings IEEE International Conference on Robotics and Automation*, volume 2, pages 995–1001. IEEE, 2000.

[63] K. Ehsani, W. Han, A. Herrasti, E. VanderBilt, L. Weihs, E. Kolve, A. Kembhavi, and R. Mottaghi. Manipulathor: A framework for visual object manipulation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 4497–4506, 2021.

[64] C. Li, F. Xia, R. Martín-Martín, and S. Savarese. Hrl4in: Hierarchical reinforcement learning for interactive navigation with mobile manipulators. In *Conference on Robot Learning*, pages 603–616. PMLR, 2020.

[65] V. Alipov, R. Simmons-Edler, N. Putintsev, P. Kalinin, and D. Vetrov. Towards practical credit assignment for deep reinforcement learning. *arXiv preprint arXiv:2106.04499*, 2021.

[66] T. Yang, H. Tang, C. Bai, J. Liu, J. Hao, Z. Meng, and P. Liu. Exploration in deep reinforcement learning: a comprehensive survey. *arXiv preprint arXiv:2109.06668*, 2021.

[67] I. Osband, B. V. Roy, D. J. Russo, and Z. Wen. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019.

[68] S. Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 6(2):107–116, 1998.

[69] F. Xia, C. Li, R. Martín-Martín, O. Litany, A. Toshev, and S. Savarese. ReLMoGen: Leveraging motion generation in reinforcement learning for mobile manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020.

[70] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[71] M. Bjelonic. YOLO ROS: Real-time object detection for ROS. https://github.com/leggedrobotics/darknet_ros, 2016–2018.

[72] S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. MIT Press, 2005.

[73] L. Fan, G. Wang, D.-A. Huang, Z. Yu, L. Fei-Fei, Y. Zhu, and A. Anandkumar. Secant: Self-expert cloning for zero-shot generalization of visual policies. *arXiv preprint arXiv:2106.09678*, 2021.