ReALFRED: An Embodied Instruction Following Benchmark in Photo-Realistic Environments

Taewoong Kim[1,*], Cheolhong Min[1,*], Byeonghwi Kim[1], Jinyeon Kim[1,2], Wonje Jeung[1], and Jonghyun Choi[1,†]

[1] Seoul National University
[2] Yonsei University
{twoongg.kim, cheolhong.min, byeonghwikim}@snu.ac.kr, jinyeonkim@yonsei.ac.kr, {specific0924, jonghyunchoi}@snu.ac.kr

Abstract. Simulated virtual environments have been widely used to learn robotic agents that perform daily household tasks. These environments encourage research progress by far, but often provide limited object interactability, visual appearance different from real-world environments, or relatively smaller environment sizes. This prevents the learned models in the virtual scenes from being readily deployable. To bridge the gap between these learning environments and deploying (*i.e.*, real) environments, we propose the ReALFRED benchmark that employs real-world scenes, objects, and room layouts to learn agents to complete household tasks by understanding free-form language instructions and interacting with objects in large, multi-room and 3D-captured scenes. Specifically, we extend the ALFRED benchmark with updates for larger environmental spaces with smaller visual domain gaps. With ReALFRED, we analyze previously crafted methods for the ALFRED benchmark and observe that they consistently yield lower performance in all metrics, encouraging the community to develop methods in more realistic environments. Our code and data are publicly available[3].

Keywords: Interactive Scanned Environments · Instruction Following · Embodied AI · Reality Gap · Dataset and Benchmark

1 Introduction

Building autonomous robotic assistants that can perform everyday household tasks has been an elusive aspiration within the research community for decades. To let them learn these intricate tasks, we may provide them with interactive environments where agents can learn task completion skills with numerous interactions with environments. A straightforward approach to train such agents that can carry out real-world activities is to directly deploy robots in real-world environments and let them learn to complete desired tasks. However, this often faces several practical challenges, including cost, time, or safety concerns [3,14,39,68].

[3] Homepage: https://github.com/snumprlab/realfred
[*] Equal contribution. [†] Corresponding author.

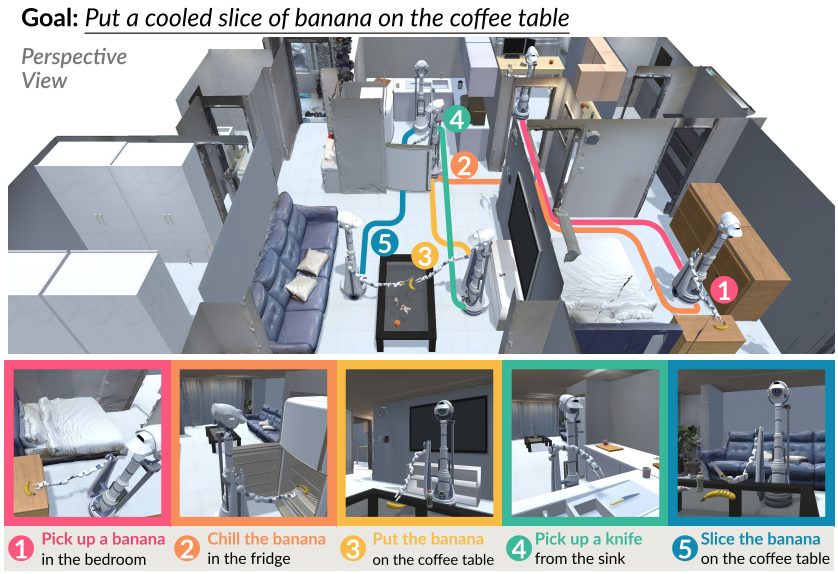


Fig. 1: Proposed ReALFRED benchmark. The top image provides a perspective view of one of our scenes. The images below represent third-person views at each time step, along with their corresponding descriptions, for better understanding. The agent is required to understand instructions in natural language and then complete the desired tasks by navigating large 3D-captured environments and interacting with objects.

As an alternative, several simulated environments have been introduced [59, 74] which leverage extensive 3D-captured environments obtained from real-world scenes [7, 56]. Compared to real-world deployment, these environments offer agents a faster process of taking actions and observing consequences, and the convenience of resetting the environment and trying again in case of failure, which enables agents to learn the skills to complete desired tasks. Adopting such simulated environments has produced remarkable advancements in various sub-tasks for embodied AI agents, including visual navigation [9,47,50,55,73], vision and language navigation [1, 37], and remote object grounding [54]. Due to the inherently static nature of these 3D-captured environments where objects (*e.g.*, books, chairs, *etc.*) remain non-interactive, however, current benchmarks [1, 9] for these tasks have less focused on object interaction, which might hinder deployability for more complex tasks that require object interaction. Recent studies [44, 57] insert liftable objects in scanned environments for object interaction, but they support limited object interaction such as picking up objects, which might not provide enough deployability for more challenging real-world scenarios such as heating objects using a microwave or cooling objects using a refrigerator.

Meanwhile, virtual game engines such as Unity have been exploited to build object-interactable environments with graphically crafted assets, including walls, floors, ceilings, and objects. These object-interactable environments have led to

notable progress in the handling of more intricate object-centric tasks, including rearrangement [67,72] and manipulation [25,26], beyond navigation-centric tasks. In particular, we have observed significant progress in the execution of more complex tasks by natural language instructions [5, 33, 46, 51, 61, 62]. However, object-interactable environments for training and evaluation of such agents often pose several issues, such as visual domain gaps [69] and relatively smaller room sizes compared to their counterparts in 3D-captured environments [7, 56, 66].

To bridge the gap between limited object interactability and environmental sizes with visual domain discrepancy, we propose the Real-World ALFRED (ReALFRED) benchmark that requires agents to complete long-horizon tasks by understanding free-form language instructions and interacting with objects in large 3D-captured environments, following a similar task setup to the widely used embodied instruction following benchmark, ALFRED [61]. The 3D-captured environments used in the ReALFRED benchmark encompass multiple rooms, providing ample space for agents to engage in multiple rooms in a single episode, which adds a sense of realism to the tasks. This resembles real-world scenarios in which agents navigate seamlessly between multiple rooms. Unlike prior benchmarks focusing primarily on single-room activities [61] or 3D-captured environments with limited object interaction [59], ReALFRED provides task evaluation in wide, realistic, and interactive environments to mirror the natural expectations of human-robot interactions. Fig. 1. illustrates the household task in one of the scenes in ours.

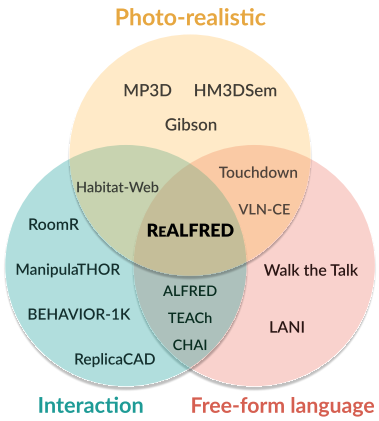


Fig. 2: While other benchmarks [7, 12, 19, 37, 41, 43, 48, 49, 56, 57, 61, 67, 72, 74] provide one or two aspects, our proposed ReALFRED benchmark addresses all of these aspects.

In our experiments, we observe that the models [5, 33, 46, 61–63] proposed in synthetic environments [61] do not perform well in our ReALFRED benchmark, implying that models developed for synthetic environments may not easily adapt to realistic environments.

We summarize our contributions as follows:

- We propose ReALFRED, a benchmark for embodied instruction following with 3D-captured multi-room environments and objects.
- We collect 3D-captured scenes and objects to reduce the simulation-reality gap and free-form language instructions to support task completion based on agents' language understanding.
- We provide analyses on the recent state-of-the-art models in the literature and the relevant Sim2Real transfer to empirically validate the necessity of our ReALFRED benchmark.

Table 1: Comparison of ReALFRED and other Embodied AI benchmarks. 'Language' column denotes the number of human annotated language directives. 'Environment' column compares spatial characteristics and whether it supports interactivity. 'Inference' column denotes whether their action space includes interactive capability with objects. The ReALFRED benchmark is the first benchmark that provides a 3D-captured and interactable environment to solve household tasks that require navigation and interaction at the same time directed by natural language commands. †We count the number of dialogue sessions. ‡We count the number of annotations in English. *Though [16] environment supports interaction, interaction is not required.

	Language	Environment				Inference
	Human Annotations	Visual Quality	Multi-Room Navigation	Movable Objects	State Changes	Object Interactability
IQA [24]	-	Synthetic	✗	✗	✓	✓
ManipulaTHOR [19]; RoomR [72]	-	Synthetic	✗	✓	✓	✓
RoboTHOR [16]	-	Synthetic	✓	✓	✓	✗*
ProcTHOR [18]	-	Synthetic	✓	✓	✓	✓
ReplicaCAD [67]	-	Synthetic	✓	✓	✓	✓
BEHAVIOR-1K [41]; HSSD-200 [31]	-	Synthetic	✓	✓	✓	✓
OpenRooms [42]	-	Photo	✗	✗	✗	✗
MP3D [7]; HM3D [56]; Gibson [74]	-	Photo	✓	✗	✗	✗
Habitat-Web [57]	-	Photo	✓	✓	✓	✓
ALFRED [61]; TEACh [49]; CHAI [48]	25k+; 2.0k+†;12k+	Synthetic	✗	✓	✓	✓
LANI [48] ; Walk the Talk [43]	28k+; 0.7k+	Synthetic	✓	✗	✗	✗
Virtualhome [52]	2.7k+	Synthetic	✓	✗	✗	✓
R2R [1]; RxR [38]; VLN-CE [37]	21k+; 42k+‡; 21k+	Photo	✓	✗	✗	✗
ReALFRED	**30k+**	**Photo**	✓	✓	✓	✓

2 Related Work

Fig. 2 illustrates the comparison of our ReALFRED with other benchmarks in three selected aspects. We first review 3D-captured environments and benchmarks that provide visual aesthetics similar to real-world environments. Then, we review simulation environments and benchmarks that support object interaction. Table 1 compares our ReALFRED with other indoor datasets.

Datasets and benchmarks with 3D-captured environments. The collection of advanced 3D scans has played a key role in enhancing our understanding of 3D objects [15, 64, 76] and their perception [2, 30, 45]. While these datasets offer valuable insights for a deeper understanding of 3D environments, they lack object interaction for learning interactive embodied agents. To further promote research on embodied agents for real-world applications, training and evaluation of such agents with physical spaces from scanned data [7, 66, 74] have been proposed. They provide a rich source of data for researchers to explore the capabilities of agents operating within real-world-inspired scenarios.

In these environments, notable progress has been achieved, primarily in the realms of navigation and exploration. Extensive research has been conducted on agents capable of navigating complex 3D environments, as evidenced by work such as navigating to a specified object [9, 47, 55], navigating to a certain point [50, 73], and navigating to the shown image [36]. Similarly, exploring novel environments has also yielded valuable insights [8, 13].

Additionally, there is a work to integrate multi-modal sensory information for developing an agent that can deeply understand environments using inputs such as vision and language [1,38] and audio and vision [10,11]. These advances enable agents to utilize a broader range of sensory data, enhancing their perception and interaction capabilities. Further research explores additional capabilities of embodied agents, such as object localization alongside navigation [54].

However, a fundamental limitation of these scan-based environments remains their static nature, which restricts the potential for object interaction. This limitation has resulted in the majority of prior research primarily focusing on navigation tasks, leaving the broader scope of agent adaptability to more complex, interactive challenges largely unaddressed. Consequently, this limitation hampers the adaptability of agents to more complex tasks.

Object-interactable simulated environments. Scan-based environments [56, 76] are typically composed of static scene representations, focusing on the visual aspects of real-world spaces. However, they often fail to adequately support interactivity with objects, representing a discrepancy between the dynamic and interactive characteristics of the real world, necessitating the development of environments better aligned with the demands of interactive agent research.

To address this gap, various simulators, such as AI2-THOR [35], ManipulaTHOR [19], RoboTHOR [16], VirtualHome [52], TDW [20], iGibson [40], and OmniGibson [41], have emerged as promising solutions. These environments are engineered with a primary focus on enabling interaction between agents and objects. They are built upon game engines, which provide a solid foundation for ensuring realistic interactivity within the virtual environment. On these simulators, researchers publish benchmarks [16, 21, 41, 49, 53, 61, 65, 75] that support interactions. These benchmarks have been promoting research in the field, leading to the development of robotic assistants capable of handling complex tasks.

However, the dataset that provide interactive environments are limited in size of a room because creating a large high-fidelity space is challenging [56]. Furthermore, despite their near-photo realism, agents would face a visual domain gap when deployed in real world environments [69, 71, 77]. To this end, HabitatWeb [57] proposes an interactive pick-and-place task based on the Habitat simulator [59,67]. Their template-based language instructions, however, might not be sufficient to express the complex nuances of human expression. Innovative methods for acquiring 3D scans from phone-captured layouts have been proposed for learning enviroment-specific policies [17]. Nevertheless, its synthetic assets may lead to visual domain gaps when deployed in the real world. On the contrary, our REALFRED supports photorealism, high interactivity with objects, and free-form language annotations. These features can offer a framework for developing language-driven agents with visual perception for complex household chores.

3 The REALFRED Benchmark

To develop agents capable of performing household tasks, substantial progress has been achieved in various domains, including navigation [1, 38], rearrange-

Fig. 3: Top-down view of 3D-captured environments. We provide two examples from our scanned indoor environments. White circles denote where scanners are deployed. By scanning scenes at diverse points, we can prevent including blind spots.

ment [72], and manipulation tasks [25,26]. In particular, [61] recently introduced the ALFRED benchmark that requires agents to complete long-horizon household tasks by jointly understanding egocentric visual observations and natural language instructions in household environments.

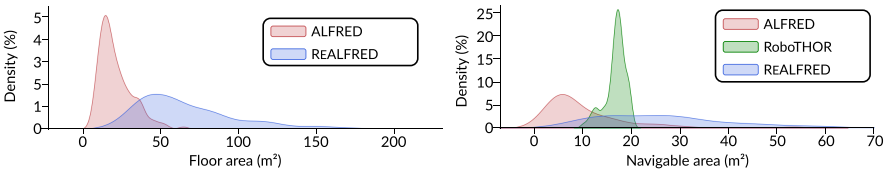
However, these environments are restricted to a single room size compared to previously proposed 3D-captured environments [7, 56] consisting of multiple rooms, which could potentially restrict the deployability of agents to larger environments. Furthermore, the environments used in the ALFRED benchmark [61] are built with synthetic CAD assets and therefore could potentially yield visual aesthetics different from those obtained from real-world environments [69], which could eventually cause performance degradation due to visual domain gaps.

To address these issues, we extend the ALFRED benchmark [61] and propose a challenging benchmark, named the REALFRED benchmark, which requires agents to perform household tasks in large indoor environments captured in 3D with object interaction. For training and evaluation, we follow the same protocol as [61] to collect expert demonstrations in the captured large environments.

3.1 Object-Interactable 3D-Captured Scenes

To reduce the visual domain gap, a straightforward approach is to use 3D scans of real world environments. However, the captured 3D scans (*i.e.*, meshes) remain *static* and thus, agents cannot interact with objects in the captured scenes. For object interaction, we manually replace object parts with 3D object assets to support object interaction. For photorealism comparison with previous environments using FID [27] and KID [4] metrics, please refer to Sec A.3.

We detail the process of collecting object-interactable 3D-captured environments and highlight key differences from previously proposed benchmarks below.
Data acquisition process. To collect 3D scans of real-world environments, we visit residential properties and employ scanners. Inspired by recent work [2], we collect scans outside of the US to add the diversity with public scanned environments. We utilize the same 3D scanner as [7], equipped with three RGB cameras



(a) Distribution of floor areas by scene. **(b)** Distribution of navigable areas by scene.

Fig. 4: Distribution of navigable and floor areas in interactive benchmarks by scenes. 'Floor area' denotes the overall size of the scene. 'Navigable area' denotes the size of the space in which the agent can actually navigate. For both metrics, the ReALFRED benchmark poses a more even distribution and provides larger areas. For (a), we exclude RoboTHOR since it consists of single-sized floors.

along with a depth sensor, and capture images from three distinct perspectives: front-facing and slightly vertical above and below. Panoramic images are acquired through six consecutive captures with horizontal rotation from a fixed viewpoint. We scan each house with 2.5-meter intervals and address blind spots due to furniture with additional scans from different viewpoints. Fig. 3 shows houses where scanners were deployed while capturing scans. It can be discerned that captures were densely concentrated in areas with a high presence of objects.

Object-interactable environments. To construct an environment with numerous interactable objects, separating each objects should be preceded. In other words, objects in the scans are initially merged into the background and therefore they remain as the background. Constructing an environment where objects can be interacted with requires the separation of objects from the background and from each other. Therefore, we manually separate the 3D scans into background elements and interactive objects. Furthermore, each object can exist in various states. To visualize changes in an object's state, we add state-relevant textures on objects. For example, we add a stain texture to a clean object when it becomes dirty. Finally, we reconstruct these individual object meshes within the Unity editor, making them compatible with the AI2-THOR simulator [35].

Comparison with previous benchmarks. To investigate the spatial characteristics of scenes in the ReALFRED benchmark, we compare ours with other benchmarks that support interaction with objects [16, 61, 67]. We observe enhancements in our dataset in both: 1) spatial sizes and 2) spatial complexity.

Spatial size. We compare the ReALFRED benchmark with other benchmarks in terms of spatial size [56] by measuring 'Floor area' and 'Navigable area' and provide the result in Fig. 4. 'Floor area' represents the total spatial size (m^2) of a scene, defined by the floor projection. 'Navigable area' measures the spatial size (m^2) of the space in which an agent can actually navigate. 'Navigable area' is smaller than or equal to 'Floor area' since 'Navigable area' excludes areas where the agent collides with any components in the scene from 'Floor area.'

We observe that the ReALFRED benchmark yields a diverse distribution of floor areas, compared to previous work [61], with a larger average per scene. Furthermore, the ReALFRED benchmark provides a broader range of navigable

Table 2: Component sizes of interactive embodied AI benchmarks. Each 'Total floor area' and 'Total navigable area' (Total nav. area) denotes the sum of all floor areas and navigation areas. 'Nav. complex.' denotes navigational complexity. 'Scene clutter' measures the amount of clutter in the scene. We do not compare ReplicaCAD's Navigable area, Navigation complexity, and Scene clutter as agent sizes differ across simulators. The highest value for each metric is shown in **bold**. †We count objects used as target objects in the ObjectNav task. ‡We count objects used in any of the tasks.

Simulator	Habitat2.0	AI2THOR		
Dataset	ReplicaCAD [67]	RoboTHOR [16]	ALFRED [61]	**ReALFRED**
# Scenes	111	75	120	**150**
Total floor area (m^2)	8,824.5	2,574	2,555	**10,060**
Total nav. area (m^2)	–	1,258	1,356	**4,251**
Nav. complex.	–	2.036	2.549	**3.020**
Scene clutter	–	**8.095**	5.119	8.072
# object class	92	14†	82‡	**112**‡

areas, with larger areas on average. This implies that an agent needs the ability to navigate effectively in spaces of varying sizes, generally wider on average.

We also investigate the spatial characteristics of each benchmark [16, 61, 67] and summarize the result in Table 2. 'Total floor area' denotes the sum of all floor areas and 'Total navigable area' denotes the sum of all navigable areas in the dataset. We observe that the ReALFRED benchmark has the highest total floor area and navigable area value, implying that ReALFRED provides more navigation space for an agent than previous work [16, 61, 67].

Spatial complexity. We now investigate the complexity of spatial structures in the ReALFRED benchmark using several metrics. To ensure a fair comparison with other datasets [16, 61, 67], we employ the navigation complexity introduced by [74] and the scene clutter measurement from [56]. A higher navigation complexity indicates an increased difficulty in navigating through the space, while a higher scene clutter implies the presence of more obstacles in the environment. By utilizing these metrics, we conduct a comparative analysis with object-interactable benchmarks [16, 61] and provide the result in Table 2.

We observe that the ReALFRED benchmark provides the environment with higher navigation complexity and scene clutter compared to other benchmarks [61, 67]. The high navigation complexity in our scenes stems from their multi-room composition. This setup requires the agent to execute more intricate navigation when moving from one room to another, in comparison to scenarios where the agent operates within a single room. We observe that ours poses the second-highest scene clutter with a marginal gap from [16]. This is because the spaces in [16] are relatively confined with a high density of furniture. We observe that ours has a similar value to [16], meaning that our scenes have a large amount of obstacles with a similar portion of [16] since our scenes' average size is larger than that of [16]. This implies that the ReALFRED benchmark provides a complex and challenging space for the agent to explore the environment.

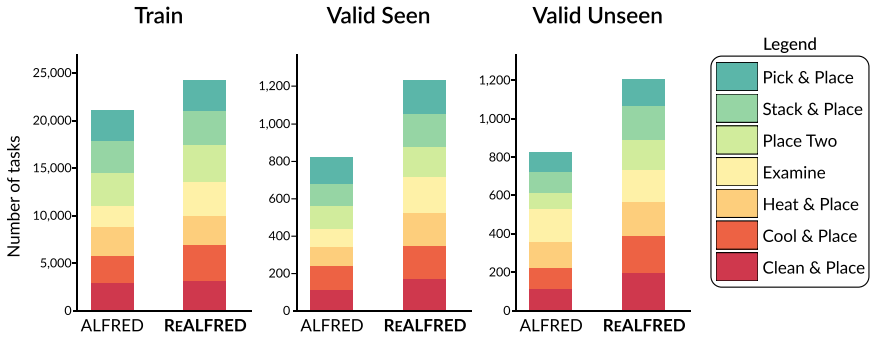


Fig. 5: The seven types of tasks' distribution in ReALFRED. We provide 37.6% more tasks in valid sets and 19.3% in total compared to previous benchmark [61].

3.2 Expert Demonstration Generation

Each expert demonstration includes a set of an egocentric RGB view and action information with an interaction mask if exists at each time step. Expert demonstrations for each task are generated by a planner [29] with encoded state spaces into Planning Domain Definition Language (PDDL) rules [22]. To generate household tasks, we utilize seven task types introduced in [61].

Data splits. We split the generated demonstrations into train, validation, and test folds. Specifically, we designate 135 scenes for *seen* and 15 scenes for *unseen* fold. Comparison of the amount of each task and the distribution across training and validation folds with previous work [61] is shown in Fig. 5.

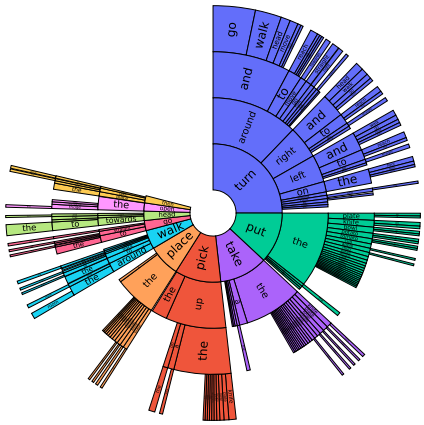


Fig. 6: Word distribution. The words are arranged starting from the center and extending outwards. The arc length corresponds to the word frequency in the instructions.

Curating free-form language instructions. Detailed language instructions describe a task that involves a sequence of actions, a point of interest in robotics. The ReALFRED benchmark offers 30,696 language directives, each comprising a human-annotated high-level goal and a set of step-by-step instructions. These directives are collected from 93 Amazon Mechanical Turk workers with a 'Master' qualification, ensuring high-quality. Collected annotations are validated through an additional voting survey, and invalid instructions are replaced with newly collected instructions. The distribution of language instructions by their first four words is presented in Fig. 6. We provide a detailed annotation process and examples of an expert demonstration with the instruction in Sec. A.1 and A.4 in the supplementary material.

Multi-room embodied instruction following. For household robots to be able to effectively assist humans, it would be more practical to deploy them in complex and diverse environments with various room types, rather than confining them to a single room. Cross-room navigation poses significant challenges, requiring agents to understand room references, plan efficient paths, and overcome obstacles proficiently. This agent is further tasked to interpret visual cues, demonstrate a keen sense of spatial awareness, and process natural language instructions.

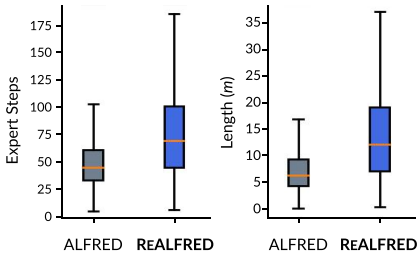


Fig. 7: Comparison of expert demonstrations (ALFRED *vs.* REALFRED). Due to expanded environments, longer steps and planning horizons are required in REALFRED.

The distribution of the number of steps and the length covered by the expert demonstrations is shown in Fig. 7. We observe that longer steps and trajectories are required to complete our tasks compared to the single-room constraint benchmark [61], meaning that our benchmark provides longer-horizon tasks.

To broaden a range of tasks, we provide an extensive number of object class types in the REALFRED benchmark. More object classes are added to the REALFRED benchmark, making it a superset of [61] with 86 pickupable and 26 receptacle objects. This has resulted in a more diverse set of tasks, with the number of unique tasks being 84.3% more than the ALFRED benchmark [61], and the proportion of unique tasks being higher in the REALFRED benchmark. We provide detailed information in Sec. A in the supplementary material.

4 Experiments

Metrics. We follow the same evaluation protocol of the ALFRED benchmark [61]. The primary metric is 'Success Rate (SR)' which measures the percentage of completed tasks. 'Goal-Condition Success Rate (GC)' measures the percentage of achieved goal conditions. We also use path-length-weighted metrics to measure how efficiently an agent completes tasks. For more details, kindly refer to [61].
Baselines. We evaluate several recent state-of-the-art methods [5,33,34,46,61–63] with competitive results in [61]. We provide more details in Sec. B.

4.1 Comparison of the State of the Arts

We evaluate the baselines in the proposed REALFRED benchmark over multiple runs and present the average result in Table 3. We report extended results with path-length-weighted metrics in Sec. C in supplementary.

For a fair comparison, we separate the baseline into two groups based on the use of additional depth supervision for semantic map reconstruction: 'Imitation Learning' where agents learn direct mapping from visual observations and language instructions to action sequences and 'Spatial Map Reconst.' where agents plan action sequences based on reconstructed semantic spatial representations.

Table 3: Task and Goal-Condition Success Rate. We train and evaluate recent state-of-the-art methods on our ReALFRED benchmark. For a fair comparison, we group these methods into two based on the usage of extra depth supervision: models learned by imitation learning without depth supervision ('Imitation Learning') and ones that maintain semantic spatial representations constructed by predicted depth maps ('Spatial Map Reconst.'). Path-length-weighted (PLW) metrics are reported in Sec. C for each value. †Authors' implementation as the code is not publicly available.

Learning	Model	Validation				Test			
		Seen		*Unseen*		*Seen*		*Unseen*	
		SR	GC	SR	GC	SR	GC	SR	GC
Imitation Learning	Seq2Seq [61]	0.77 ± 0.06	6.93 ± 0.06	0.00 ± 0.00	4.03 ± 0.00	1.10 ± 0.00	6.60 ± 0.00	0.00 ± 0.00	3.50 ± 0.00
	MOCA [62]	12.64 ± 0.12	20.95 ± 0.18	1.44 ± 0.05	6.76 ± 0.04	14.11 ± 0.03	22.84 ± 0.04	0.62 ± 0.08	5.14 ± 0.08
	ABP† [33]	24.71 ± 0.05	33.80 ± 0.14	4.22 ± 0.05	11.71 ± 0.27	27.44 ± 0.40	35.81 ± 0.23	3.54 ± 0.23	10.57 ± 0.22
Spatial Map Reconst.	HLSM [5]	4.23 ± 0.08	9.14 ± 0.09	1.08 ± 0.14	6.12 ± 0.23	6.27 ± 0.04	10.44 ± 0.13	0.49 ± 0.16	4.28 ± 0.13
	FILM [46]	7.08 ± 0.28	11.93 ± 0.23	4.44 ± 0.17	9.26 ± 0.13	8.79 ± 0.07	13.03 ± 0.08	2.15 ± 0.18	6.56 ± 0.15
	LLM-Planner† [63]	5.80 ± 0.19	11.69 ± 0.35	3.33 ± 0.22	8.29 ± 0.19	8.16 ± 0.20	13.20 ± 0.13	1.90 ± 0.13	6.33 ± 0.02
	CAPEAM† [34]	13.45 ± 0.05	18.16 ± 0.27	4.92 ± 0.22	9.47 ± 0.23	15.61 ± 0.15	20.22 ± 0.11	2.87 ± 0.13	7.36 ± 0.07
	Human	-	-	-	-	-	-	85.00 ± 3.54	91.30 ± 2.94

We observe that all these baselines, 'Imitation Learning' and 'Spatial Map Reconst.,' consistently achieve lower performance values for all metrics in both *seen* and *unseen* splits compared to performances achieved in [61], implying that our proposed ReALFRED provide more challenges compared to [61]. While in [61], the 'Spatial Map Reconst.' baselines [5,34,46] outperform learning-based approaches [33,61,62] by exploiting semantic spatial representation with deterministic algorithms (*e.g.*, obstacle-free navigation path planning [60]), we observe a contrasting result that the 'Imitation Learning' baselines outperforming the 'Spatial Map Reconst.' baselines in our ReALFRED benchmark.

We qualitatively observe that such a confined spatial map reconstruction is due to a limited field of view and map reconstruction methods. The agent's limited field of view often leads to failure in recognizing room corners with doors or narrow aisles, resulting in (single-room-sized) limited map reconstruction. In addition, [34, 46] perceives obstacles larger than they actually are for better obstacle-free path planning by sacrificing navigable area, but this can be quite critical for narrow doors and aisles as they have a small amount of navigable space. This may hinder navigation to other rooms and thus, fail at tasks. We provide a more detailed discussion, supported by figures, on the agents' difficulty in recognizing narrow passages and the resulting impact on the agents, who struggle with spatial map reconstruction in Sec. D in the supplementary material.

4.2 Comparison to the agents with sim-to-real adaptation

We investigate the transfer from simulation training to real-world scan evaluation (sim-to-real) and from real-world scan training to real-world scan evaluation (real-to-real) with [33]. We train a sim-to-real agent with synthetic visual data and a real-to-real agent with real scanned visual data, respectively. During inference, both agents predict an action and an object mask based on the scanned visual input frame and the given language instruction at every time step.

Training and evaluation data selection. For training the sim-to-real agent, we use the training dataset in [61], encompassing $21K$ language annotations. For a fair comparison, we train the real-to-real agent with tasks from the REALFRED benchmark's training fold, specifically involving the manipulation of the same objects that were used in training the sim-to-real agent with $19K$ language annotations. For evaluation, we select tasks from the valid unseen splits of our REALFRED benchmark. We evaluate the agent performance with the tasks that 1) do not require multi-room navigation 2) and those that do. This selection specifically includes tasks that feature objects used in the training phase.

Real-to-sim domain adaptation. The use of Generative Adversarial Networks (GANs) [23, 79] for domain adaptation has recently been examined in the literature on robotics [6, 28, 32, 58, 77]. These models are employed to adapt input images from real-world domains (*i.e.* Real $\rightarrow$ Sim) before they are passed to the agent policy. We train a CycleGan [79] and its off-the-shelf variant [70] with unpaired images collected from the REALFRED (real domain) and ALFRED benchmarks [61] (simulated domain). Among the trained generators, the one performing real to simulation conversion, referred to as the *real-to-sim goggle*, learns the mapping from the real domain to the simulation domain $G_S : \mathcal{R} \rightarrow \mathcal{S}$, where $\mathcal{R}$ denotes the real domain and $\mathcal{S}$ denotes the simulated domain.

Results. We present the results of the sim-to-real experiments in Table 4. We report the performance of agents when tasks 1) require agents to navigate multi-rooms, denoted as 'Multi + Single,' and 2) are solvable within a single room, denoted as 'Single only.' Firstly, we observe that the sim-to-real agent significantly underperforms its real-to-real counterpart in all metrics ($\#(a)$ *vs.* $\#(d)$). We then compare the results with *goggled* agents. By comparing agents evaluated in the 'Multi + Single' tasks, we do not observe improvements in the main metric, success rate (SR). We observe a slight increase in the PLWSR metric for the *goggled* sim-to-real agent compared to the vanilla sim-to-real agent ($\#(a)$ *vs.* $\#(b, c)$).

Table 4: Comparison to the agents with sim-to-real adaptation. Each 'Sim2Real' and 'Real2Real' denotes an agent trained in simulated environments and 3D-captured scenes. 'Goggle' denotes *real-to-sim* methods.

			Multi + Single		Single only	
#	Setting	Goggle	SR	PLWSR	SR	PLWSR
(*a*)	Sim2Real	*None*	0.115	0.012	0.0	0.0
(*b*)	Sim2Real	CycleGan [79]	0.115	0.016	0.327	0.065
(*c*)	Sim2Real	UVCGAN-v2 [70]	0.115	0.046	0.327	0.187
(*d*)	Real2Real	*None*	2.405	0.785	2.614	0.762

To isolate the impacts of the visual domain gap, we then compare agents evaluated on 'Single only' tasks. We observe that both *goggled* agents show improvements over the vanilla sim-to-real agent which results in zero success rate. However, a noticeable performance gap exists with the real-to-real agent, implying the need for learning in real scanned environments ($\#(d)$ *vs.* $\#(a, b, c)$).

4.3 Challenges in REALFRED

We propose two hypotheses for the low performance observed with state-of-the-art methods on the REALFRED benchmark: (1) navigating within larger scenes and (2) overcoming narrow pathways between rooms. These elements represent challenges of completing tasks within multi-room, household-scale environments.

Fig. 8: Distribution of collision spots. The heatmap illustrates collision frequency on each spot and the black outline depicts a navigable area. We illustrate collision points of [33] in *valid unseen* environments to represent the distribution. Specifically, we choose failure episodes from the two largest scenes in the *valid unseen* split. We observe a remarkable concentration of collisions at the entrance to the rooms. The corresponding egocentric views are presented on the right side of the figure.

We analyze failure cases with the best-performing model, ABP [33], on *valid unseen* split, since it seems the most promising on the ReALFRED benchmark.

Difficulty in large environments. Understanding the surrounding environment, including the location of objects and the positions of obstacles, can be beneficial for task completion. However, the agent's visual range is bounded, requiring it to expend more steps for exploration in order to perceive larger spaces. This intensifies the challenge of navigating larger spaces with limited steps.

We conduct an analysis to see the relationship between the success rate (SR) and the size of the space. We follow the interquartile range (IQR) method to set a threshold at $30.44m^2$ that covers the most navigable space sizes in the ALFRED [61]. This threshold marks the upper fence, defining spaces above it as outliers. We classify spaces in the ReALFRED benchmark smaller than threshold as smaller scenes, and those larger as larger scenes. Results indicate that the agent [33] with the highest SR in the *valid unseen* fold showed an average SR of 5.46% in the smaller scenes and only 1.77% in the larger scenes. This implies that solving tasks becomes more challenging as the space size increases.

In addition, we compare the difficulty of navigation between the previous [61] and ReALFRED benchmarks with different average spatial sizes. To quantify, we first define milestones for each task, which are spots to be reached to interact with target objects (*e.g.*, Apple, Knife, *etc.*). We consider navigation to be a success only if *all* milestones are visited, regardless of whether the actual interaction (*e.g.*, slicing an apple) is performed or not. Consequently, the navigation success rate in the ReALFRED benchmark is merely 59.18% for the *valid unseen* split while the agent's [33] navigation success rate is 84.82% in ALFRED [61]. The ReALFRED benchmark offering an average space size more than three times larger than the previous work [61] implies the need for model development capable of overcoming the challenges associated with this increased scale.

Navigation through narrow doorways. The ReALFRED benchmark supports environments with multiroom composition, unlike the previous dataset [61]

providing a single room scale. Specifically, it contains narrow doorways across rooms, which may hinder the agent from navigating through. Here, we hypothesize that the agent would frequently collide with the walls near the door.

We investigate collision spots and showcase two examples on top of the scene's layout in *valid unseen* fold in Fig. 8. The number of collisions in failed cases is accumulated and normalized to the maximum number of collisions, respectively, to indicate the collision frequency at each point. The black outline represents the agent's navigable area, as detailed in Sec. 3.1. Each value denotes the normalized collision frequency. In both scenes, we observe that collisions often occur on walls near the door, implying that our spatial characteristic (*i.e.*, including narrow spaces) may hinder the agent from properly navigating without collision.

Human evaluation. Following [61], we randomly select 100 directives from the *test unseen* fold and have them evaluated by humans. Five participants are given 20 tasks each, which they complete using a keyboard-and-mouse interface. Before starting, they are given the opportunity to become familiar with the interface.

Participants achieve a comparable high success rate of 85%, along with a goal-condition success rate of 91.30%, in average. We agree that human performance may appear slightly lower compared to the results presented in previous work [57, 61,78]. This is partly because human participants encountered several difficulties when controlling the agent, particularly in avoiding collisions within narrow corridors. Furthermore, navigating large spaces with a limited egocentric field of view introduced additional challenges, leading to task failures.

5 Conclusion

We present REALFRED, a new dataset and benchmark for embodied instruction following task on 3D-captured environments. We capture 150 indoor houses in 3D with interactable objects to enable complex household tasks. The reconstructed indoor scenes provide a larger spatial area and complex multi-room environments that are close to the real-world scenario and challenging for an agent to successfully complete a task. Expert demonstrations are also provided along with free-form human-language instructions.

In our empirical evaluations, we show that state-of-the-art methods struggle in large multi-room environments, provide analyses of our newly proposed benchmark, and perform Sim2Real transfer experiments. We have released our Embodied AI research data and code for reproducibility. We expect that the REALFRED benchmark will encourage further research on developing robotic agents that execute household tasks by language instructions in the real world.

Limitation and future work. Although we support a large number of interactable objects, the types of tasks to be completed are rather limited, considering more complex real-world scenarios. In addition, we currently address natural language in English but users may come from different regions with different languages. We can think of two future research avenues as follows. (1) adding additional complicated types of task that require both hands to complete. (2) supporting a multi-lingual interface for users from different regions.

Acknowledgements

This work was partly supported by the NRF grant (No.2022R1A2C400230012, 5%) and IITP grants (No.RS-2022-II220077 (5%), No.RS-2022-II220113 (5%), No.RS-2022-II220959 (5%), No.RS-2022-II220871 (15%), No.RS-2020-II201361 (5%, Yonsei AI), No.RS-2021-II211343 (5%, SNU AI), No.RS-2021-II212068 (5%, AI Innov. Hub), No.RS-2022-II220951(50%)) funded by the Korea government(MSIT).

References

1. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., van den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: CVPR (2018)
2. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., Shulman, E.: ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In: NeurIPS Datasets and Benchmarks Track (2021)
3. Berseth, G., Xie, C., Cernek, P., van de Panne, M.: Progressive reinforcement learning with distillation for multi-skilled motion control. In: ICLR (2018)
4. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. In: ICLR (2018)
5. Blukis, V., Paxton, C., Fox, D., Garg, A., Artzi, Y.: A persistent spatial semantic representation for high-level natural language instruction execution. In: CoRL (2021)
6. Bousmalis, K., Irpan, A., Wohlhart, P., Bai, Y., Kelcey, M., Kalakrishnan, M., Downs, L., Ibarz, J., Pastor, P., Konolige, K., et al.: Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In: ICRA (2018)
7. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. arXiv:1709.06158 (2017)
8. Chaplot, D.S., Gandhi, D., Gupta, S., Gupta, A., Salakhutdinov, R.: Learning to explore using active neural slam. In: ICLR (2020)
9. Chaplot, D.S., Gandhi, D.P., Gupta, A., Salakhutdinov, R.R.: Object goal navigation using goal-oriented semantic exploration. In: NeurIPS (2020)
10. Chen, C., Jain, U., Schissler, C., Gari, S.V.A., Al-Halah, Z., Ithapu, V.K., Robinson, P., Grauman, K.: Soundspaces: Audio-visual navigation in 3d environments. In: ECCV (2020)
11. Chen, C., Schissler, C., Garg, S., Kobernik, P., Clegg, A., Calamia, P., Batra, D., Robinson, P.W., Grauman, K.: Soundspaces 2.0: A simulation platform for visual-acoustic learning. In: NeurIPS Datasets and Benchmarks Track (2022)
12. Chen, H., Suhr, A., Misra, D., Snavely, N., Artzi, Y.: Touchdown: Natural language navigation and spatial reasoning in visual street environments. In: CVPR (2019)
13. Chen, T., Gupta, S., Gupta, A.: Learning exploration policies for navigation. In: ICLR (2019)
14. Choi, S., Ji, G., Park, J., Kim, H., Mun, J., Lee, J.H., Hwangbo, J.: Learning quadrupedal locomotion on deformable terrain. Science Robotics (2023)
15. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017)

16. Deitke, M., Han, W., Herrasti, A., Kembhavi, A., Kolve, E., Mottaghi, R., Salvador, J., Schwenk, D., VanderBilt, E., Wallingford, M., et al.: Robothor: An open simulation-to-real embodied ai platform. In: CVPR (2020)
17. Deitke, M., Hendrix, R., Farhadi, A., Ehsani, K., Kembhavi, A.: Phone2proc: Bringing robust robots into our chaotic world. In: CVPR (2023)
18. Deitke, M., VanderBilt, E., Herrasti, A., Weihs, L., Ehsani, K., Salvador, J., Han, W., Kolve, E., Kembhavi, A., Mottaghi, R.: Procthor: Large-scale embodied ai using procedural generation. In: NeurIPS (2022)
19. Ehsani, K., Han, W., Herrasti, A., VanderBilt, E., Weihs, L., Kolve, E., Kembhavi, A., Mottaghi, R.: Manipulathor: A framework for visual object manipulation. In: CVPR (2021)
20. Gan, C., Schwartz, J., Alter, S., Mrowca, D., Schrimpf, M., Traer, J., De Freitas, J., Kubilius, J., Bhandwaldar, A., Haber, N., et al.: Threedworld: A platform for interactive multi-modal physical simulation. In: NeurIPS Datasets and Benchmarks Track (2021)
21. Gao, X., Gao, Q., Gong, R., Lin, K., Thattai, G., Sukhatme, G.S.: Dialfred: Dialogue-enabled agents for embodied instruction following. RA-L (2022)
22. Ghallab, M., Howe, A., Knoblock, C., McDermott, D., Ram, A., Veloso, M., Weld, D., Wilkins, D.: Pddl the planning domain definition language (1998)
23. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
24. Gordon, D., Kembhavi, A., Rastegari, M., Redmon, J., Fox, D., Farhadi, A.: Iqa: Visual question answering in interactive environments. In: CVPR (2018)
25. Gu, J., Xiang, F., Li, X., Ling, Z., Liu, X., Mu, T., Tang, Y., Tao, S., Wei, X., Yao, Y., et al.: Maniskill2: A unified benchmark for generalizable manipulation skills. In: ICLR (2023)
26. Heo, M., Lee, Y., Lee, D., Lim, J.J.: Furniturebench: Reproducible real-world benchmark for long-horizon complex manipulation. In: RSS (2023)
27. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS (2017)
28. Ho, D., Rao, K., Xu, Z., Jang, E., Khansari, M., Bai, Y.: Retinagan: An object-aware approach to sim-to-real transfer. In: ICRA (2021)
29. Hoffmann, J., Nebel, B.: The ff planning system: Fast plan generation through heuristic search. In: JAIR (2001)
30. Hua, B.S., Pham, Q.H., Nguyen, D.T., Tran, M.K., Yu, L.F., Yeung, S.K.: Scenenn: A scene meshes dataset with annotations. In: 3DV (2016)
31. Khanna, M., Mao, Y., Jiang, H., Haresh, S., Shacklett, B., Batra, D., Clegg, A., Undersander, E., Chang, A.X., Savva, M.: Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In: CVPR (2024)
32. Khansari, M., Ho, D., Du, Y., Fuentes, A., Bennice, M., Sievers, N., Kirmani, S., Bai, Y., Jang, E.: Practical visual deep imitation learning via task-level domain consistency. In: ICRA (2023)
33. Kim, B., Bhambri, S., Singh, K.P., Mottaghi, R., Choi, J.: Agent with the big picture: Perceiving surroundings for interactive instruction following. In: Embodied AI Workshop @ CVPR (2021)
34. Kim, B., Kim, J., Kim, Y., Min, C., Choi, J.: Context-aware planning and environment-aware memory for instruction following embodied agents. In: ICCV (2023)

35. Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Gordon, D., Zhu, Y., Gupta, A., Farhadi, A.: Ai2-thor: An interactive 3d environment for visual ai. arXiv:1712.05474 (2017)
36. Krantz, J., Gervet, T., Yadav, K., Wang, A., Paxton, C., Mottaghi, R., Batra, D., Malik, J., Lee, S., Chaplot, D.S.: Navigating to objects specified by images. In: ICCV (2023)
37. Krantz, J., Wijmans, E., Majumdar, A., Batra, D., Lee, S.: Beyond the nav-graph: Vision-and-language navigation in continuous environments. In: ECCV (2020)
38. Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J.: Room-Across-Room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In: EMNLP (2020)
39. Kumar, A., Fu, Z., Pathak, D., Malik, J.: Rma: Rapid motor adaptation for legged robots. In: RSS (2021)
40. Li, C., Xia, F., Martín-Martín, R., Lingelbach, M., Srivastava, S., Shen, B., Vainio, K.E., Gokmen, C., Dharan, G., Jain, T., Kurenkov, A., Liu, K., Gweon, H., Wu, J., Fei-Fei, L., Savarese, S.: igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In: CoRL (2021)
41. Li, C., Zhang, R., Wong, J., Gokmen, C., Srivastava, S., Martín-Martín, R., Wang, C., Levine, G., Lingelbach, M., Sun, J., et al.: Behavior-1k: A benchmark for embodied ai with 1,000 everyday activities and realistic simulation. In: CoRL (2022)
42. Li, Z., Yu, T.W., Sang, S., Wang, S., Song, M., Liu, Y., Yeh, Y.Y., Zhu, R., Gundavarapu, N., Shi, J., Bi, S., Yu, H.X., Xu, Z., Sunkavalli, K., Hasan, M., Ramamoorthi, R., Chandraker, M.: Openrooms: An open framework for photorealistic indoor scene datasets. In: CVPR (2021)
43. MacMahon, M., Stankiewicz, B., Kuipers, B.: Walk the talk: Connecting language, knowledge, and action in route instructions. In: AAAI (2006)
44. Majumdar, et al.: Findthis: Language-driven object disambiguation in indoor environments. In: CoRL (2023)
45. Mao, Y., Zhang, Y., Jiang, H., Chang, A., Savva, M.: Multiscan: Scalable rgbd scanning for 3d environments with articulated objects. In: NeurIPS (2022)
46. Min, S.Y., Chaplot, D.S., Ravikumar, P., Bisk, Y., Salakhutdinov, R.: Film: Following instructions in language with modular methods. In: ICLR (2022)
47. Min, S.Y., Tsai, Y.H.H., Ding, W., Farhadi, A., Salakhutdinov, R., Bisk, Y., Zhang, J.: Object goal navigation with end-to-end self-supervision. In: IROS (2023)
48. Misra, D., Bennett, A., Blukis, V., Niklasson, E., Shatkhin, M., Artzi, Y.: Mapping instructions to actions in 3d environments with visual goal prediction. In: EMNLP (2018)
49. Padmakumar, A., Thomason, J., Shrivastava, A., Lange, P., Narayan-Chen, A., Gella, S., Piramuthu, R., Tur, G., Hakkani-Tur, D.: Teach: Task-driven embodied agents that chat. In: AAAI (2022)
50. Partsey, R., Wijmans, E., Yokoyama, N., Dobosevych, O., Batra, D., Maksymets, O.: Is mapping necessary for realistic pointgoal navigation? In: CVPR (2022)
51. Pashevich, A., Schmid, C., Sun, Chen: Episodic transformer for vision-and-language navigation. In: ICCV (2021)
52. Puig, X., Ra, K., Boben, M., Li, J., Wang, T., Fidler, S., Torralba, A.: Virtualhome: Simulating household activities via programs. In: CVPR (2018)
53. Puig, X., Undersander, E., Szot, A., Cote, M.D., Yang, T.Y., Partsey, R., Desai, R., Clegg, A.W., Hlavac, M., Min, S.Y., et al.: Habitat 3.0: A co-habitat for humans, avatars and robots. In: ICLR (2024)

54. Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W.Y., Shen, C., Hengel, A.v.d.: Reverie: Remote embodied visual referring expression in real indoor environments. In: CVPR (2020)
55. Ramakrishnan, S.K., Chaplot, D.S., Al-Halah, Z., Malik, J., Grauman, K.: Poni: Potential functions for objectgoal navigation with interaction-free learning. In: CVPR (2022)
56. Ramakrishnan, S.K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J.M., Undersander, E., Galuba, W., Westbury, A., Chang, A.X., Savva, M., Zhao, Y., Batra, D.: Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In: NeurIPS Datasets and Benchmarks Track (2021)
57. Ramrakhya, R., Undersander, E., Batra, D., Das, A.: Habitat-web: Learning embodied object-search strategies from human demonstrations at scale. In: CVPR (2022)
58. Rao, K., Harris, C., Irpan, A., Levine, S., Ibarz, J., Khansari, M.: Rl-cyclegan: Reinforcement learning aware simulation-to-real. In: CVPR (2020)
59. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., Parikh, D., Batra, D.: Habitat: A Platform for Embodied AI Research. In: ICCV (2019)
60. Sethian, J.A.: A fast marching level set method for monotonically advancing fronts. In: PNAS (1996)
61. Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., Fox, D.: Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In: CVPR (2020)
62. Singh, K.P., Bhambri, S., Kim, B., Mottaghi, R., Choi, J.: Factorizing perception and policy for interactive instruction following. In: ICCV (2021)
63. Song, C.H., Wu, J., Washington, C., Sadler, B.M., Chao, W.L., Su, Y.: Llm-planner: Few-shot grounded planning for embodied agents with large language models. In: ICCV (2023)
64. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: CVPR (2015)
65. Srivastava, S., Li, C., Lingelbach, M., Martín-Martín, R., Xia, F., Vainio, K.E., Lian, Z., Gokmen, C., Buch, S., Liu, K., et al.: Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In: CoRL (2021)
66. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv:1906.05797 (2019)
67. Szot, A., Clegg, A., Undersander, E., Wijmans, E., Zhao, Y., Turner, J., Maestre, N., Mukadam, M., Chaplot, D., Maksymets, O., Gokaslan, A., Vondrus, V., Dharur, S., Meier, F., Galuba, W., Chang, A., Kira, Z., Koltun, V., Malik, J., Savva, M., Batra, D.: Habitat 2.0: Training home assistants to rearrange their habitat. In: NeurIPS (2021)
68. Tan, J., Zhang, T., Coumans, E., Iscen, A., Bai, Y., Hafner, D., Bohez, S., Vanhoucke, V.: Sim-to-real: Learning agile locomotion for quadruped robots. In: RSS (2018)
69. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., Abbeel, P.: Domain randomization for transferring deep neural networks from simulation to the real world. In: IROS (2017)

70. Torbunov, D., Huang, Y., Tseng, H.H., Yu, H., Huang, J., Yoo, S., Lin, M., Viren, B., Ren, Y.: Uvcgan v2: An improved cycle-consistent gan for unpaired image-to-image translation. arXiv:2303.16280 (2023)
71. Truong, J., Chernova, S., Batra, D.: Bi-directional domain adaptation for sim2real transfer of embodied navigation agents. RA-L (2021)
72. Weihs, L., Deitke, M., Kembhavi, A., Mottaghi, R.: Visual room rearrangement. In: CVPR (2021)
73. Wijmans, E., Kadian, A., Morcos, A., Lee, S., Essa, I., Parikh, D., Savva, M., Batra, D.: Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. In: ICLR (2020)
74. Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson env: Real-world perception for embodied agents. In: CVPR (2018)
75. Yenamandra, S., Ramachandran, A., Yadav, K., Wang, A., Khanna, M., Gervet, T., Yang, T.Y., Jain, V., Clegg, A.W., Turner, J., et al.: Homerobot: Open-vocabulary mobile manipulation. In: CoRL (2023)
76. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: ICCV (2023)
77. Zhang, J., Tai, L., Yun, P., Xiong, Y., Liu, M., Boedecker, J., Burgard, W.: Vr-goggles for robots: Real-to-sim domain adaptation for visual control. RA-L (2019)
78. Zhu, H., Kapoor, R., Min, S.Y., Han, W., Li, J., Geng, K., Neubig, G., Bisk, Y., Kembhavi, A., Weihs, L.: Excalibur: Encouraging and evaluating embodied exploration. In: CVPR (2023)
79. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)