EXCALIBUR
Encouraging and Evaluating Embodied Exploration

Hao Zhu[♛], Raghav Kapoor[♛], So Yeon Min[♛],
Winson Han[♞], Jiatai Li[♛], Kaiwen Geng[♛],
Graham Neubig[♛], Yonatan Bisk[♛], Aniruddha Kembhavi[♞], Luca Weihs[♞]
[♛]Carnegie Mellon University, [♞]Allen Institute for Artificial Intelligence
zhuhao@cmu.edu

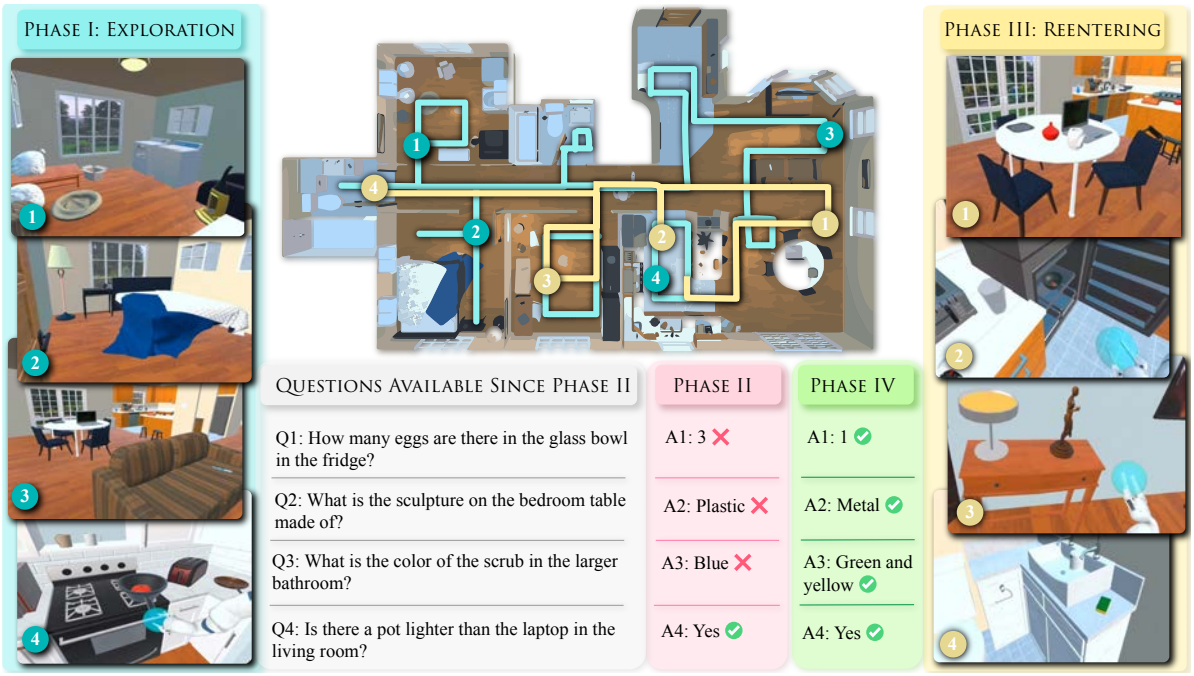

Figure 1. Episode in EXCALIBUR played by a human annotator. An episode is divided into four sequential phases: in Phase I, the agent explores the house for 2,500 steps (each action takes a step); in Phase II the agent needs to answer 20 questions (5 shown) about the explored environment; in Phase III the agent is given a second chance to reenter the house, now with knowledge of the questions; in Phase IV the agent answers the questions again. Performance is evaluated with the answer accuracy in Phases II&IV and the time spent in Phase III. The observation space is egocentric (see left and right panels). The action space includes navigation and manipulation actions (Fig. 2).

Abstract

Experience precedes understanding. Humans constantly explore and learn about their environment out of curiosity, gather information, and update their models of the world. On the other hand, machines are either trained to learn passively from static and fixed datasets, or taught to complete specific goal-conditioned tasks. To encourage the development of exploratory interactive agents, we present the EXCALIBUR *benchmark.* EXCALIBUR *allows agents to explore their environment for long durations and then query their understanding of the physical world via inquiries like: "is the small heavy red bowl made from glass?" or "is there a silver spoon heavier than the egg?". This design encourages agents to perform free-form home exploration without myopia induced by goal conditioning. Once the agents have answered a series of questions, they can renter the scene to refine their knowledge, update their beliefs, and improve their performance on the questions. Our experiments demonstrate the challenges posed by this dataset for the present-day state-of-the-art embodied systems and*

the headroom afforded to develop new innovative methods. Finally, we present a virtual reality interface that enables humans to seamlessly interact within the simulated world and use it to gather human performance measures. EXCALIBUR *affords unique challenges in comparison to present-day benchmarks and represents the next frontier for embodied AI research.*

1. Introduction

Humans are *active* learners, acquiring knowledge of the physical world through intentional experiments with their bodies and senses. Children as young as a few months old learn about objects and their environment through observation and interaction [6, 24]. This sensorimotor experience, as pointed out by Piaget [47], is critical in forming a fundamental understanding of reality. This is the cognitive motivation for the creation of EXCALIBUR.

In contrast, machine learning models typically obtain knowledge by passively observing web-crawled, encyclopedic, or crowd-sourced static datasets [67]. This *passive* approach has clear limitations. For instance, grounding physical concepts like *heavy*, *large*, and *long* requires moving beyond passive observation. To weigh an object, humans will often try to use different forces to move it. To compare the sizes of objects, they move around and perceive the objects from different angles and distances. Although large pre-trained models have made progress in aligning with the grounded world [41, 45], they still lack an embodied understanding of physical concepts [59].

Todays popular active, embodied-learning benchmarks in the Embodied AI community focus on directed task completion. These include navigating to specified GPS coordinates [3], locating an object of a specified category [7], translating commands into low-level actions [5, 56], and inspecting a scene to answer a question about the presence or count of an object category [15, 25]. A more recent benchmark, Room Rearrangement [62] requires agents to explore the scene, but the focus there is on navigation, observation, and memorization. Progress on these benchmarks has been promising. We can now train agents that can comprehend goal instructions reasonably well and complete simple tasks, particularly navigation heavy tasks. None of these benchmarks, however, explicitly probe how these models have learned to represent their environments, nor do they encourage the type of free-form, undirected, experimental, exploration performed by humans.

To encourage and evaluate the capacity of embodied agents to openly explore their environment and interact with objects within it, we present the EXCALIBUR[1] benchmark. EXCALIBUR is built using large procedurally generated houses via ProcTHOR [18]. Each episode in EXCALIBUR consists of four phases as shown in Fig. 1. Phase I Exploration – The agent must navigate to and interact with objects in the environment. Importantly, the agent isn't seeded with a goal and must instead perform open-ended exploration. Interacting with objects takes place via physics-enabled arm manipulation. Phase II Question Answering – We probe the agent's understanding of the physical world through natural language inquiries. Our questions go beyond simple primitive queries, *e.g.* regarding object existence, and include physical attributes (*e.g.* masses and materials) and visual attributes (*e.g.* colors and shapes). Phase III Reentering – This is a goal-directed phase, since the agent must interact with the environment to refine its understanding of the world in response to questions asked in the previous stage. Phase IV Refined Question Answering – This phase repeats the inquiries made in Phase II to query if the agent was able to successfully acquire the required knowledge about its world after being provided the goal question set.

Our use of question-answering in this benchmark which focuses on interaction and exploration has several benefits. Natural language inquiries allow us to probe the agent's understanding of the world. They also provide a clear and objective metric for EXCALIBUR. Further, they can serve as supervisory signals to encourage agents to interact with objects and explore the world. Finally, the introduction of language opens the door to using pre-trained language models in future work, given the recent rise of their use for planning for embodied agents [2].

EXCALIBUR is the first benchmark that offers the following new avenues and challenges for Embodied AI research: (1) It encourages open-ended exploration. (2) Agents in EXCALIBUR have access to a rich interactive action space that covers navigation, arm-based manipulation, and grasping with different degrees of force. (3) The questions in this benchmark move beyond existence and counting. They probe the agent on its abilities to learn physical and visual attributes of the world. (4) Our task requires long-horizon planning and reasoning. Most embodied benchmarks today have maximum episode lengths of up to 250 steps. Our task has four phases that include an exploration phase of 2500 steps. (5) Our task also evaluates the ability of an agent to refine and improve the existing knowledge of its environment. This is an ability that humans commonly showcase in their everyday experiences.

We present baselines using state-of-the-art Embodied AI neural models and learning methods. We also design a Virtual Reality interface to enable humans to navigate and interact with objects in ProcTHOR scenes in an immersive way. This allows for a more accurate human baseline measurement, which demonstrates that there remains substantial room for model improvement. Finally, in Sec. 5, we show that the failure patterns of models are distinct from

[1] **Ex**ploratory **C**urious **A**gents with **L**anguage **I**nduced Em**b**odied World Unde**r**standing

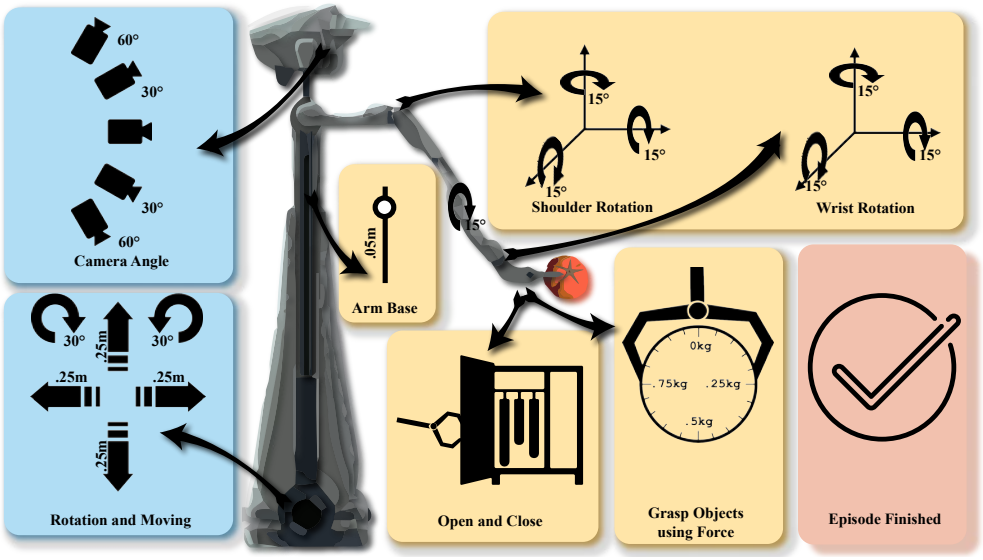

Figure 2. The action space of EXCALIBUR. The whole action space consists of two sets of actions: Navigation (left) and Manipulation (right). Navigation actions are used to move the agent (bottom left) and look at different angles (top left). Manipulation actions are used to move the arm (top right), grasp with force and open and close closets, drawers and fridges (which are implemented as high action which can be triggered when the gripper is close to the handles), and signal finishing the task (bottom right). All of the actions are discretized: angler motion are discretized into 15 degrees, linear motion are discretized into 0.05 meter for joints and 0.25 meter for base and force is discretized into 0.05 kilogram-force.

those of humans. Humans are great at exploration, but fall short at memorization, while agents tend to succeed at answering questions that depend on memory but are poor explorers – even when trained with popular exploration rewards. Altogether, we find that EXCALIBUR serves as a powerful and flexible framework and environment for evaluating and building **Ex**ploratory **C**urious **A**gents with **L**anguage **I**nduced Em**b**odied World **Under**standing.

2. EXCALIBUR

Consider the example depicted in Fig. 1: the embodied agent is spawned in the bedroom of a random house at a random position. It traverses the bedroom, living room, kitchen, and bathroom, opens closets, fridges, and drawers, and picks up various objects. After 2,500 steps, the agent is asked 20 questions and answers some questions correctly and some incorrectly, e.g. "How many silver objects are heavier than the white egg in the kitchen?". The agent then returns to the house and explores the scene again. This time it starts lifting silver objects in the room to estimate their weight. As the example reveals, EXCALIBUR encourages agents to openly explore their world in the first phase but also evaluates their ability to perform goal-directed exploration once the questions become known. Natural language inquiries are used to ascertain what the agent has learned about its environment. We now present details about the EXCALIBUR task, and contrast it to previous Embodied AI benchmarks in Sec. 6 and Tab. 3.

2.1. Task.

An EXCALIBUR task is defined as a triple $\langle \mathcal{H}, \mathcal{Q}, \mathcal{P} \rangle$, where a $\mathcal{H}$ouse consists of a floor plan and objects in it, a $\mathcal{Q}$uestion set is a list of English question-answer pairs, and a $\mathcal{P}$osition is a 2D location on the floor of $\mathcal{H}$ that is empty (*i.e.* at which the agent can be placed) along with an initial agent camera orientation. Each object in the house is defined by its type, colors, materials (full list of object types, colors, and materials is in Appendix B), meshes (3D shape of the objects), location, size (width, depth, height), and weight (under 5kg). At the start of each episode, the floor plan, objects (including colors, materials, sizes, and weights), questions, and agent spawn position are randomly sampled with the distribution specified in Appendix C.

Phases. The EXCALIBUR task consists of four phases: (I) exploration, (II) question answering, (III) reentering, and (IV) refined question answering. In both (I) and (III), the agent may navigate throughout the house and manipulate objects. One difference between (I) and (III) is that the time steps in (I) are limited to 2,500, while the steps in (III) T_3 are unlimited but used to discount the accuracy improvement in Eq. 1. In (II) the agent is asked 20 questions. This brings up another notable difference between (I) and (III). In (I), an agent must perform open-ended exploration, learning about objects and their relationships. In (III), its exploration is conditioned on its experience in (I), the goal questions and its own answers in (II), and it attempts to improve its answers in (IV). We denote the accuracy in Phase (II) and

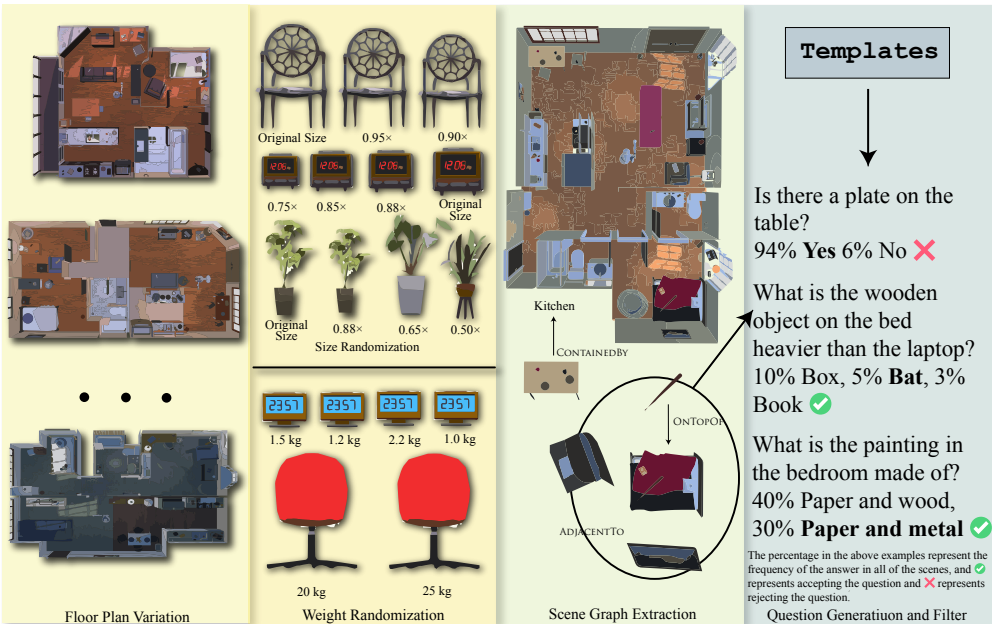

Figure 3. Dataset contruction procedure. We generate the dataset in four steps (each in a pane). (1) We consider the procedurally generated floor plans and houses generated with PROCTHOR. (2) We then randomize the sizes and weights of objects in the scene. (3) We then extract the scene graphs of objects and relations in the scenes. (4) Based on hand-crafted templates, we generate questions and filter out questions that can be answered without exploring the scenes.

Phase (IV) as Acc_{exp} and Acc_{ref}.

Agents. The breadth of embodied experience results from the versatility of human bodies. With this in mind, the agent used in EXCALIBUR is the MANIPULATHOR arm agent of Ehsani *et al*. [21]. This agent has a dexterous 6 DOF Kinova-inspired robotic arm, see Fig. 2. We extend their design by adding a force argument to grasping action.[2] This is one step further towards more realistic manipulation and also empowers the agents to "feel" the weights of objects through interaction. Fig. 2 shows the available actions of the armed agent in Phase (I) and (III). The "Done" action signals that the agent wishes to end Phase (III), the number of time steps spent before which are counted as T_3. At every timestep, the agent acts given egocentric RGB images (of size 800×600) as its observation.

Evaluation. We wish to evaluate two facets of exploration: (1) "how many questions can be answered with the knowledge acquired in Phase (I)?", and (2) "how efficient is the agent in refining its answers in Phase (IV)?". To define a unified metric measuring both facets, we propose the following exploration score (ExQA):

$$\text{ExQA} \triangleq \text{Acc}_{\text{exp}} + (\text{Acc}_{\text{ref}} - \text{Acc}_{\text{exp}}) \exp(-kT_3), \quad (1)$$

where we call $k > 0$ the *energy coefficient*. ExQA reduces to Acc_{ref} when $k = 0$ and reduces to Acc_{exp} as $k \to \infty$. Our choice of k thus determines how we prioritize accuracy after exploration versus after answer refinement. We choose a value for k that maximizes human performance, biasing models to uncover strategies of similar efficiency and efficacy as we see in human demonstrations.[3]

2.2. Dataset Construction

The EXCALIBUR dataset is built upon PROCTHOR-10k, a dataset of 10,000 procedurally generated home environments, each containing between 1-10 rooms [18]. For each PROCTHOR-10k home, we apply a variety of scene augmentations (*e.g.* randomizing object weight and sizes) and generate sets of challenging questions. We break our dataset generation process into four stages: randomization, scene graph generation, question generation, and filtering. We detail each stage below, see Fig. 3 for a visual overview.

Randomization. The diversity across PROCTHOR-10k houses is very large: objects placements, floor plans, materials, ||are all randomized while respecting sensible constraints common across real homes. Despite this diversity, we found that, without applying additional scene augmentations, many questions of interest become either trivial or answerable via commonsense. For instance, the weights of many objects in AI2-THOR (and thus in PROCTHOR-10k) are set uniformly across object categories. This means that a question such as "is the cup in the kitchen heavier than the bowl?", may have a constant answer across all cups and bowls. Thus, without applying weight randomization,

[2]Force feedback mechanisms are common in physical manipulators.

[3]Empirically, we find k which maximizes ExQA likelihood under a gaussian prior. The procedure for choosing an optimal value of k is described in App. F.

the agent may answer accurately without any exploration or object interaction. In EXCALIBUR, we apply two types of supplemental randomization to PROCTHOR-10k: object weight and size randomization. In particular, within each house, we uniformly sample the weights of *pickupable* (*i.e.* excluding large objects that cannot be held by the agent, *e.g.* a fridge) objects to be between $0.5\times$ and $1.5\times$ their starting values. Similarly, the size of pickupable objects (*i.e.* their scale) is randomized to be with $0.8\times$ and $1.0\times$ of their starting values. Note that we only downscale objects as this prevents potential collisions between nearby objects.

Scene Graph. Before moving to question generation, we first preprocess each house to produce a scene graph representation of the environment. This scene graph provides a compact summary of the objects in the house along with their relationships and attributes. In our formulation, rooms, objects, and agent are represented as nodes with edges between nodes representing their relationships. These relationships include, for example, CONTAINEDBY, ADJACENTTO, ONTOPOF. A full listing of object relationships and node attributes can be found in the appendix.

Question Generation. To generate our question sets, we follow the process used to generate the single-image visual question answering (VQA) dataset CLEVR [31]. In particular, we represent questions using functional programs whose answer values can be found by evaluating these programs upon the above described scene graph. As for CLEVR, we design a collection of (11) question families, which can be composed and chained to generate questions. This question generation process may produce degenerate or tautological questions, we prune these using the depth-first approach employed when constructing CLEVR. Details of question generation procedure is in App. E.

	Type	%
Question	Yes-no	78.8
	Count	12.3
	Query	8.9
Relation	Color	26.7
	Material	66.2
	CONTAINEDBY	8.2
	ADJACENTTO	39.5
	ONTOPOF	0.8
	HEAVIERTHAN[4]	30.6
	LARGERTHAN	18.9

Table 1. Dataset Distribution.

Filtering. In order to create questions that are challenging and whose answers are not overly biased to certain answers, we use extensive question filtering to remove easy questions. In particular, for each candidate question q, we compute the answer of q across all scenes, which produces a distribution over answers. We filter questions to only include those whose answer distribution is sufficiently balanced. For more details see, App. D.

The result of such a process is an underlying dataset with a range of difficult questions of 3 different types and 7 kinds of physical properties and relations (Fig. 1) Different types of questions are evaluated in slightly different ways: Yes-no questions are evaluated by exact matching, count questions are answered correctly when the prediction is only different than the standard answer by 5%, and query questions match prediction and the standard answer order-agnostically. In this way, we use accuracy as an umbrella metric for all of the questions. There are four splits in EXCALIBUR: (1) a training set with 10k PROCTHOR scenes, (2) a validation and a test set with 1k PROCTHOR scenes each, and (3) another test set with 9 hand-crafted ARCHITECTHOR scenes[5] for comparison between agents and humans.

3. Human Baseline with VR Interface

One challenge of comparing human performance fairly with that of our agents is that our agents are extensively trained on houses from our dataset while human annotators, on the other hand, are only exposed to a small handful of training episodes. It is therefore important to create a realistic environment where real-life experience and knowledge can be easily transferred to the simulated environment. For this, we create a VR interface to EXCALIBUR and ask human annotators to complete tasks while virtually embodied as the agent. In our experiments, human participants used the Meta Quest 2 VR headset[6] and were evaluated using the same metric as our agents. Concretely, to make the experience interactive and immersive, we ensured that our VR experience satisfied the following requirements.

- **Flexible Head Movement**: The head movement of the human annotators is smoothly reflected as camera movement in the VR environment, so that the information-seeking behavior of the human annotators can be easily transferred to the simulated environment.
- **Intuitive Arm Movement**: Human annotators should be able to intuitively manipulate the robotic, 6 DOF Kinova-like, the arm of the MANIPULATHOR agent used in EXCALIBUR. As the robotic arm has greater degrees of freedom than a human arm (ignoring human fingers) this means that special attention must be paid to ensure that humans need not worry about the rotation of joints of the arm, but only the position and orientation of the gripper.
- **Gripping With Force**: We leveraged the pressure on the grip button of the Meta Quest 2 controller to map it to the grasp force in the environment so annotators can use different magnitudes of forces to grip objects.
- **Open/Close**: We also facilitated the user to open and close various objects in the VR environment, to make the experience more immersive and allow the user to explore the house in greater depth.

For more details on VR interface, training and evaluating

[4] HEAVIERTHAN includes LIGHTERTHAN, and LARGERTHAN includes SMALLERTHAN, LONGERTHAN, and SHORTERTHAN

[5] One ARCHITECTHOR scene is used for training human annotators.

[6] https://www.meta.com/quest/products/quest-2/

human annotators, see App. A.

4. Reinforcement Learning Baselines

EXCALIBUR requires a model to actively plan, explore the houses, manipulate objects, memorize its history, and answer questions. In this work, and as is common across modern embodied benchmarks, we train reinforcement learning models as our baselines. Recurrent neural networks (RNNs) are frequently used as generic models for encoding language instructions, historical observations, and actions, into belief states for embodied agents [18, 21, 32, 62–64]. Following this prior work, we use a GRU [13] to encode the history of observations seen and actions taken by the agent to produce, at every time step $t \geq 0$, a vector *belief state* b_t corresponding to the output of the RNN at that timestep. We extend this practice by feeding the belief states as input to an actor-critic policy head as well as to a question answering module. To understand whether questions answering serves as a good stimulation for encourging exploration, we consider three training signals: a (1) coverage-based reward, (2) QA reward, and (3) QA cross-entropy loss. Our goal in the following experiments is to show that modern Embodied AI models and training techniques can achieve some level of success on EXCALIBUR with the goal of inspiring future work to build upon these results.

Actor-critic policy The belief state is fed into an MLP with one hidden layer, which we call the *actor-head*, and decoded into logits, one logit for each discrete action available to the agent (recall Sec. 2.1). By passing these logits through a softmax we produce the agent's policy (*i.e.* a distribution over agent actions). To enable training with PPO [52, 63], we also must produce an estimate of the value of the agent's current state. To do this, we feed the belief state through another similar MLP, the *critic-head*, which returns a 1-dimensional output.

Question answering To make full use of existing large, pretrained, language models, we follow [60] and propose to convert belief states into continuous *prefix* tokens using a prefix generator MLP with two hidden layers f_θ^{prefix}. These prefix tokens are preprepended to with the question tokens and fed into the encoder of pre-trained T5 [49]. We then use the, pretrained, T5 decoder module to produce a (distribution over) natural-language answers to the given question. Note that the T5 model has its parameters frozen and so is not trained in our experiments.

Featurizing agent observations We experiment with two different visual feature extractors for the agent's egocentric RGB observations: (1) a pre-trained CLIP ResNet50 model [32, 48] and (2) a MaskRCNN [29] model finetuned on our training scenes. Visual features and an embedding of the agent's last action are concatenated and passed as input to the above RNN. After Phase II, the agent additionally conditions the question embeddings from the T5 encoder as input to the RNN, which is also concatenated to observation and question embeddings.

	ProcTHOR Test Set				ArchitecTHOR Test Set			
	Acc_{exp}	Acc_{ref}	T_3	ExQA	Acc_{exp}	Acc_{ref}	T_3	ExQA
Random	41.7	41.7	-	41.7	39.1	39.1	-	39.1
Language	53.5	53.5	-	53.5	49.2	49.2	-	49.2
QA	58.5	60.2	131.2	60.0	52.4	56.0	159.1	55.7
Novelty	54.2	56.5	99.6	56.4	49.9	54.5	125.7	54.1
Novelty+QA	**58.7**	**63.1**	203.2	**62.4**	**53.5**	**56.3**	211.7	**55.9**
Human w/o replay	-	-	-	-	63.6	87.1	759.4	79.4
Human w/ replay	-	-	-	-	81.3	94.3	782.1	90.1

Table 2. Human and baseline performance across two test sets. We **bold** best metric values among AI systems.

Training Our training loss equals the unweightd sum of the standard PPO RL loss [52] and $\mathcal{L}_{\text{QA}}$, a cross-entropy loss for question answering defined as

$$\mathcal{L}_{\text{QA}} = \sum_{t=1}^{T} \sum_{(q,a)\in\mathcal{Q}} -\log p_{\text{T5}}(a \mid [f_\theta^{\text{prefix}}(h_t), f^{\text{emb}}(q)]), \tag{2}$$

where p_{T5} is the probability of answer a produced by a T5 encoder-decoder, and f^{emb} is the embedding layer of the T5 encoder, and $\mathcal{Q}$ is the set of question-answer pairs associated with an episode.

Rewards We consider two kinds of rewards in this paper: (1) a QA reward and (2) a novelty-based reward. The QA reward is calculated by comparing the answers generated through beam search from T5 and the ground truth answers:

$$r_t^{\text{QA}} = \frac{1}{|\mathcal{Q}|} \sum_{(q,a)\in\mathcal{Q}} \Big(\mathbb{I}(a = \text{T5}_t(q)) - \mathbb{I}(a = \text{T5}_{t-1}(q))\Big), \tag{3}$$

where $\text{T5}_t(q) = \text{T5}(f_\theta^{\text{decoder}}(h_t), f^{\text{emb}}(q))$ denotes the output of the T5 model when using beam search decoding. Note that r_t^{QA} can only be non-zero when the agent's answer to a question changes between time steps $t-1$ and t. Our novelty reward encourages the agent to exhaustively navigate and observe novel objects, in particular, we let

$$r_t^{\text{novelty}} = \frac{O_t^{\text{seen}} - O_{t-1}^{\text{seen}}}{O_{\text{all}}} + \frac{A_t - A_{t-1}}{A_{\text{reachable}}}, \tag{4}$$

where O_t^{seen} denotes the number of objects seen till time step t, O_{all} denotes number of objects in $\mathcal{H}$, A_t denotes the area covered by time step t, and $A_{\text{reachable}}$ denotes the total reachable area in $\mathcal{H}$.

5. EXCALIBUR Human and Agent Evaluation

To gain insight into the gap between humans' and state-of-the-art embodied AI models' performance on EXCALIBUR we first must train such embodied models. To this

end, we train several variants of the reinforcement learning baseline described in Sec. 4 on the training split of EXCALIBUR. In particular, we train three variants denoted *QA*, *Novelty*, and *Novelty+QA*; as suggested by their names, the *QA* agent is only given the QA reward signal, the *Novelty* agent has access to the novelty reward, and the *Novelty+QA* is given the sum of both rewards at every timestep. For all of these agents, cross entropy loss is used for optimizing the prefix generator. Beyond these RL baselines, we also include non-interactive *Random* and *Language* baselines; the *Random* baseline simply chooses answers at random from among plausible answers when conditioned on the question type while the *Language* model is trained to answer questions given only question text, which helps indentifying artifects in question generation.

To make cross-model and human-agent comparisons we evaluate our embodied models on two test sets: (1) the procedurally generally PROCTHOR-10k testing scenes and (2) the set of, human-designed, ARCHITECTHOR test houses [18]. We evaluate humans only in the ARCHITECTHOR houses as the ARCHITECTHOR test houses were meticulously crafted to closely imitate real-world houses and represent a smaller domain shift for human participants.

The results of these evaluations can be found in Table 2. Among AI systems, we see that the *Novelty+QA* agent performs best across the Acc_{exp}, Acc_{ref}, and ExQA metrics with the *QA* model close behind. This suggests that the novelty reward may provide only marginal benefits and, indeed, the *Novelty* agent obtains results only slightly above those of the *Langauge* model which, at best, simply reproduces the biases in our question-answer pairs.

For our human evaluations, we consider two experimental conditions *Human w/o replay* and *Human w/ replay*. In the *Human w/ replay* trials, unlike in *Human w/o replay*, humans are allowed to view a video of their behavior in Phase I and Phase III when answering questions in Phase II and IV, respectively. Hence participants in the *Human w/ replay* trials are relieved of the burden of needing to remember all of the details of their exploration. While humans outperform the AI systems in both experimental conditions, the gap between AI and human performance is far narrower (gap of $+10.1$ Acc_{exp} for *Human w/o replay v.s.* a gap of $+27.8$ for Acc_{exp} *Human w/o replay*). This suggests that memorization is a significant bottleneck for humans. Note that, in the *Human w/ replay* condition, humans achieve an extremely high Acc_{ref} value (94.3) showing clearly that EXCALIBUR is, in principle, solvable by intelligent systems.

Further analysis of our results as well as descriptive metrics of agent exploration behavior can be found in App. H.

6. Related Work

The domain of embodied AI has seen an explosion of attention in recent years [17, 19]. Here, we review three sub-areas of this community most relevant to this work.

6.1. Exploration, Execution and Manipulation

Tab. 3 summarizes recent embodied AI benchmarks and evaluation frameworks comparing our EXCALIBUR benchmark with those including those designed for question answering, instruction following, rearrangement, and visual navigation. We say that an embodied benchmark or framework requires: open-ended **exploration** if the agent must act *before* being given fully specified goal information, goal-driven **execution** if the agent must act *after* being given the task definition, and **manipulation** if the agent must directly interact with objects, either with a physically simulated *arm* (*e.g.*, [21,36,58]) to complete its goal or with a higher-level abstraction (*e.g.*, in [62], the agent picks up objects by specifying their semantic category). We can see that most benchmarks emphasize either exploration or execution and manipulation. Most similar to EXCALIBUR are the BEHAVIOR [57] and AI2-THOR Rearrangement [62] benchmarks. BEHAVIOR requires agents to complete activities, defined using predicate logic, using rich interaction and object manipulation but, unlike EXCALIBUR, does not emphasize open-ended exploration and experimentation. AI2-THOR rearrangement, on the other hand, includes an exploration component but this exploration requires only memorizing object states, unlike EXCALIBUR which rewards agents who directly interact with objects. In total, EXCALIBUR is the first benchmark that explicitly evaluates agents' understanding of the physical world after agents explore, and manipulate objects within, virtual homes. As argued previously, EXCALIBUR requires that agents understand scenes with their body, form a representation that can be used to answer symbolic questions, and apply the knowledge acquired from exploration to execution.

6.2. Visual Exploration

The task of visual exploration in embodied and robotics contexts has a long history of study with a rich diversity in perspectives. This diversity exists, in part, as the meaning of "exploration" is ambiguous: is an agent successful in exploration if it visits many locations, if it interacts with many objects, or something else entirely? The excellent survey of Ramakrishnan, *et al.* [51] divides space of existing exploration strategies into four groups: curiosity (seeking unexpected states), novelty (seeking unseen states), coverage (looking to visual reveal large areas), and reconstruction (seeking states that aid in predicting other unseen states). Some recent works that have touched on these areas include, curiosity [38,42,46,54], novelty [8,9,20,43], coverage [11, 12, 64], and reconstruction [30, 33, 50]. Of course not all work falls cleanly into these categories, for instance Eysenbach *et al.* perform skill discovery (*i.e.* exploration) by maximizing information theoretic quantities [22] and

	Work	Exploration	Execution	Manipulation	Human Perf.	Language
QA	EQA [15]	No	Yes	No	No	QA
	IQA [25]	No	Yes	Abstract	Keyboard	QA
	QA Probing [14]	Yes	No	No	No	QA
	EMQA [16]	No	No	No	No	QA
Instr.	RxR Habitat [34]	No	Yes	No	Keyboard	Instruction
	ALFRED [56]	No	Yes	Abstract	Keyboard	Instruction
	TEACh [44]	No	Yes	Abstract	Keyboard	Dialog
Rear.	AI2THOR [62]	Yes	Yes	Abstract	No	No
	Habitat [58]	Yes	No	Arm	No	No
Nav.	PointNav [3]	No	Yes	No	No	No
	ObjectNav [3]	No	Yes	No	No	No
	ArmPointNav [21]	No	Yes	Arm	No	No
	BEHAVIOR [57]	No	Yes	Arm	Immersive	Descriptive
	EXCALIBUR	Yes	Yes	Arm	Immersive	QA

Table 3. Comparison between Embodied AI agents and human evaluation frameworks.

Chaplot *et al*. perform a type of heurstic semantic-goal-guided exploration using learned priors [10].

We argue that question answering rewards act as highly versatile and symbolic training signal for embodied agents. While clearly a non-traditional exploration training signal, our work can be be seen as a type of reconstruction-based exploration. While existing reconstruction-based exploration generally uses a pixel-based objective (*e.g*. ability to predict how an environment would look from an unseen camera location), our natural language queries require the agent the agent to "reconstruct" a general semantic understanding of the environment.

6.3. Question Answering for Vision

The work of Agrawal *et al*. [1] introduced the task of large-scale free-form open-ended *Visual Question Answering* (VQA) where, given a static image and natural language question about the image, a model is expected to return a natural language answer to this question. This seminal work began a new subdomain of computer vision with hundreds of publications and dozens of related datasets, see [55] for a recent review. These VQA benchmarks probe model's ability to reason about, for example, common sense [68], spatial relationships [31], potential agent actions [37], and diverse world knowledge [53]. Fundamentally, VQA focuses on single-image-understanding while our work requires interaction-driven agent exploration of an entire environment; for instance, questions about an object's weight in our dataset are unanswerable without interaction.

More recently, several video question answering datasets have been introduced, *e.g*. [23, 26, 28, 35, 65, 66]. Among these datasets, perhaps most related to our work, as it requires answering questions from an egocentric perspective, is the *episodic memory* task from the Ego4D benchmark suite [26]; in this a task a model must answer natural language questions about a video by returning the segment of the video including the question's answer. While moving from single-images to videos requires utilizing long-term memory and building a holistic representation of the environment, the lack of agent-driven interaction in these tasks means that agent learning is constrained to the prefixed trajectories taken when filming the videos. This makes it challenging to train agents who run their own experiments and are able to flexibly correct their mistakes.

The vision and language research community has prodiced a vast array of models for VQA ranging from the earliest vanilla architectures [1], to using explicit object detectors [4], to pre-training with transformers [39] to general purpose unified architetcures [27, 40, 61]. In this work we use a T5 language decoder to answer questions that conditions on the belief state of the agent which forms a representation of its current and past observations.

7. Conclusion and Future Directions

In this paper, we present a novel benchmark EXCALIBUR for encouraging and evaluating the exploration ability of embodied agents. We build strong baseline models trained with both question answering reward and novelty based exploration-encouraging reward. We compared them with human performance in immersive environment. Human's great exploration ability not only leads to much higher accuracy after exploration but also leads to faster answer finding in reentering phase. Putting it all together, EXCALIBUR is still a challenging task and improvements on this benchmark could be make in various directions, including but not limit to fine-grained semantic map building, adversarial question generation, better features from observation, and leveraging large language models. A successful agent on EXCALIBUR will lead to embodied agents with the ability to actively understanding environments in a symbolic way and being able to ground and articulate embodied concepts in language.

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. VQA: visual question answering - www.visualqa.org. *Int. J. Comput. Vis.*, 123(1):4–31, 2017. 8

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. Do as i can and not as i say: Grounding language in robotic affordances. In *arXiv preprint arXiv:2204.01691*, 2022. 2

[3] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. 2, 8

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 8

[5] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018. 2

[6] Renée Baillargeon. Infants' physical world. *Current Directions in Psychological Science*, 13:89 – 94, 2004. 2

[7] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited: On evaluation of embodied agents navigating to objects. *ArXiv*, abs/2006.13171, 2020. 2

[8] Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1471–1479, 2016. 7

[9] Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. Exploration by random network distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 7

[10] Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 8

[11] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural SLAM. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 7

[12] Tao Chen, Saurabh Gupta, and Abhinav Gupta. Learning exploration policies for navigation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 7

[13] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014. 6, 16

[14] Abhishek Das, Federico Carnevale, Hamza Merzic, Laura Rimell, Rosalia Schneider, Josh Abramson, Alden Hung, Arun Ahuja, Stephen Clark, Greg Wayne, and Felix Hill. Probing emergent semantics in predictive agents via question answering. In *ICML*, 2020. 8

[15] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *CVPR*, June 2018. 2, 8

[16] Samyak Datta, Sameer Dharur, Vincent Cartillier, Ruta Desai, Mukul Khanna, Dhruv Batra, and Devi Parikh. Episodic memory question answering. In *CVPR*, 2022. 8

[17] Matt Deitke, Dhruv Batra, Yonatan Bisk, Tommaso Campari, Angel X. Chang, Devendra Singh Chaplot, Changan Chen, Claudia Pérez D'Arpino, Kiana Ehsani, Ali Farhadi, Li Fei-Fei, Anthony Francis, Chuang Gan, Kristen Grauman, David Hall, Winson Han, Unnat Jain, Aniruddha Kembhavi, Jacob Krantz, Stefan Lee, Chengshu Li, Sagnik Majumder, Oleksandr Maksymets, Roberto Martín-Martín, Roozbeh Mottaghi, Sonia Raychaudhuri, Mike Roberts, Silvio Savarese, Manolis Savva, Mohit Shridhar, Niko Sünderhauf, Andrew Szot, Ben Talbot, Joshua B. Tenenbaum, Jesse Thomason, Alexander Toshev, Joanne Truong, Luca Weihs, and Jiajun Wu. Retrospectives on the Embodied AI Workshop. *ArXiv*, 2022. 7

[18] Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Jordi Salvador, Kiana Ehsani, Winson Han, Eric Kolve, Ali Farhadi, Aniruddha Kembhavi, et al. Procthor: Large-scale embodied ai using procedural generation. *arXiv preprint arXiv:2206.06994*, 2022. 2, 4, 6, 7

[19] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied AI: from simulators to research tasks. *IEEE Trans. Emerg. Top. Comput. Intell.*, 6(2):230–244, 2022. 7

[20] Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. Go-explore: a new approach for hard-exploration problems. *CoRR*, abs/1901.10995, 2019. 7

[21] Kiana Ehsani, Winson Han, Alvaro Herrasti, Eli VanderBilt, Luca Weihs, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Manipulathor: A framework for visual object manipulation. In *CVPR*, 2021. 4, 6, 7, 8

[22] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 7

[23] Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. Knowit VQA: answering knowledge-based questions about videos. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 10826–10834. AAAI Press, 2020. 8

[24] Alison Gopnik, Andrew N. Meltzoff, and Patricia K. Kuhl. The scientist in the crib : minds, brains, and how children learn. 1999. 2

[25] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *CVPR*, June 2018. 2, 8

[26] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3, 000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18973–18990. IEEE, 2022. 8

[27] Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. Towards general purpose vision systems. *ArXiv*, abs/2104.00743, 2021. 8

[28] Vivek Gupta, Badri N. Patro, Hemant Parihar, and Vinay P. Namboodiri. Vquad: Video question answering diagnostic dataset. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV - Workshops, Waikoloa, HI, USA, January 4-8, 2022*, pages 282–291. IEEE, 2022. 8

[29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988. IEEE Computer Society, 2017. 6

[30] Dinesh Jayaraman and Kristen Grauman. Learning to look around: Intelligently exploring unseen environments for unknown tasks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1238–1247. Computer Vision Foundation / IEEE Computer Society, 2018. 7

[31] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society, 2017. 5, 8

[32] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: CLIP embeddings for embodied AI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14809–14818. IEEE, 2022. 6

[33] Klemen Kotar and Roozbeh Mottaghi. Interactron: Embodied adaptive object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14840–14849. IEEE, 2022. 7

[34] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4392–4412, 2020. 8

[35] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. TVQA+: spatio-temporal grounding for video question answering. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8211–8225. Association for Computational Linguistics, 2020. 8

[36] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Elliott Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, Andrey Kurenkov, C. Karen Liu, Hyowon Gweon, Jiajun Wu, Li Fei-Fei, and Silvio Savarese. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Conference on Robot Learning, 8-11 November 2021, London, UK*, volume 164 of *Proceedings of Machine Learning Research*, pages 455–465. PMLR, 2021. 7

[37] Rengang Li, Cong Xu, Zhenhua Guo, Baoyu Fan, Runze Zhang, Wei Liu, Yaqian Zhao, Weifeng Gong, and Endong Wang. AI-VQA: visual question answering based on

agent interaction with interpretability. In João Magalhães, Alberto Del Bimbo, Shin'ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni, editors, *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 5274–5282. ACM, 2022. 8

[38] Manuel Lopes, Tobias Lang, Marc Toussaint, and Pierre-Yves Oudeyer. Exploration in model-based reinforcement learning by empirically estimating learning progress. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 206–214, 2012. 7

[39] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 8

[40] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *ArXiv*, abs/2206.08916, 2022. 8

[41] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. *arXiv preprint arXiv:2209.15162*, 2022. 2

[42] Yaniv Oren, Matthijs T. J. Spaan, and Wendelin Böhmer. Planning with uncertainty: Deep exploration in model-based reinforcement learning. *CoRR*, abs/2210.13455, 2022. 7

[43] Georg Ostrovski, Marc G. Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 2721–2730. PMLR, 2017. 7

[44] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *AAAI*, 2022. 8

[45] Roma Patel and Ellie Pavlick. Mapping language models to grounded conceptual spaces. In *ICLR*, 2022. 2

[46] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 488–489. IEEE Computer Society, 2017. 7

[47] Jean Inhelder Brbel Piaget. *The construction of reality in the child*. 1954. 2

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 6, 17

[49] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020. 6, 17

[50] Santhosh K. Ramakrishnan, Dinesh Jayaraman, and Kristen Grauman. Emergence of exploratory look-around behaviors through active observation completion. *Sci. Robotics*, 4(30), 2019. 7

[51] Santhosh K. Ramakrishnan, Dinesh Jayaraman, and Kristen Grauman. An exploration of embodied visual exploration. *International Journal of Computer Vision*, 2021. 7

[52] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 6

[53] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. *CoRR*, abs/2206.01718, 2022. 8

[54] Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8583–8592. PMLR, 2020. 7

[55] Himanshu Sharma and Anand Singh Jalal. A survey of methods, datasets and evaluation metrics for visual question answering. *Image Vis. Comput.*, 116:104327, 2021. 8

[56] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *CVPR*, 2020. 2, 8

[57] Sanjana Srivastava, Chengshu Li, Michael Lingelbach, Roberto Martín-Martín, Fei Xia, Kent Elliott Vainio, Zheng Lian, Cem Gokmen, Shyamal Buch, Karen 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. 7, 8

[58] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in Neural Information Processing Systems*, 34:251–266, 2021. 7, 8

[59] Jesse Thomason, Mohit Shridhar, Yonatan Bisk, Chris Paxton, and Luke Zettlemoyer. Language grounding with 3d objects. In *Conference on Robot Learning*, pages 1691–1701. PMLR, 2022. 2

[60] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 6, 17

[61] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022. 8

[62] Luca Weihs, Matt Deitke, Aniruddha Kembhavi, and Roozbeh Mottaghi. Visual room rearrangement. In *CVPR*, 2021. 2, 6, 7, 8

[63] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. DD-PPO: learning near-perfect pointgoal navigators from 2.5 billion frames. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 6

[64] Joel Ye, Dhruv Batra, Abhishek Das, and Erik Wijmans. Auxiliary tasks and exploration enable objectgoal navigation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 16097–16106. IEEE, 2021. 6, 7

[65] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: collision events for video representation and reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 8

[66] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8807–8817. Computer Vision Foundation / IEEE, 2019. 8

[67] Anthony Zador, Blake Richards, Bence Ölveczky, Sean Escola, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, et al. Toward next-generation artificial intelligence: Catalyzing the neuroai revolution. *arXiv preprint arXiv:2210.08340*, 2022. 2

[68] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE, 2019. 8