DISCO: Embodied Navigation and Interaction via Differentiable Scene Semantics and Dual-level Control

Xinyu Xu[1] Shengcheng Luo[1] Yanchao Yang[2] Yong-Lu Li[1†] Cewu Lu[1†]

[1] Shanghai Jiao Tong University
{xuxinyu2000,yonglu_li,lucewu}@sjtu.edu.cn woodroof1998@gmail.com
[2] The University of Hong Kong
yanchaoy@hku.hk

Abstract. Building a general-purpose intelligent home-assistant agent skilled in diverse tasks by human commands is a long-term blueprint of embodied AI research, which poses requirements on task planning, environment modeling, and object interaction. In this work, we study primitive mobile manipulations for embodied agents, *i.e.* how to navigate and interact based on an instructed verb-noun pair. We propose **DISCO**, which features non-trivial advancements in contextualized scene modeling and efficient controls. In particular, DISCO incorporates differentiable scene representations of rich semantics in object and affordance, which is dynamically learned on the fly and facilitates navigation planning. Besides, we propose dual-level coarse-to-fine action controls leveraging both global and local cues to accomplish mobile manipulation tasks efficiently. DISCO easily integrates into embodied tasks such as embodied instruction following. To validate our approach, we take the ALFRED benchmark of large-scale long-horizon vision-language navigation and interaction tasks as a test bed. In extensive experiments, we make comprehensive evaluations and demonstrate that DISCO outperforms the art by a sizable +8.6% success rate margin in unseen scenes even without step-by-step instructions. Our code is publicly released at https://github.com/AllenXuuu/DISCO.

Keywords: Differentiable scene semantics · Dual-level control · Embodied instruction following

1 Introduction

Recent years have witnessed a huge effort [3, 28, 44, 48] in developing embodied agents to perform everyday household tasks in indoor environments. However, accomplishing long-horizon household tasks in the unstructured real world by human commands remains challenging for modern robots. Numerous fundamental capabilities are essential for such a comprehensive robotic application, encompassing multimodal understanding, decomposing long-term objectives, perceiving the environment, and taking actions. Fig. 1 gives an intuitive case. The

† Corresponding authors.

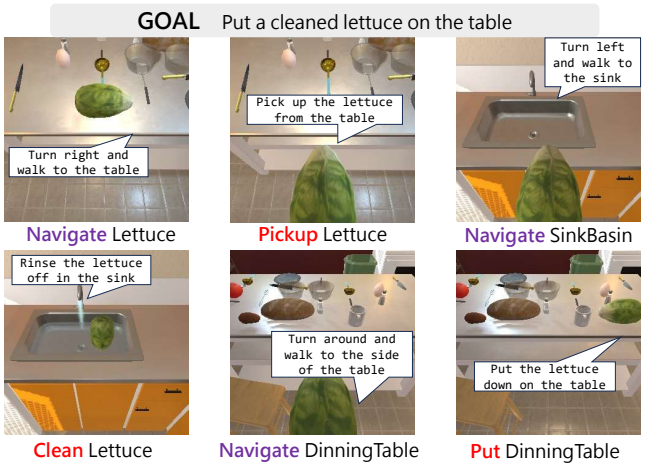

Fig. 1: An example of vision-language navigation and interaction task in ALFRED [44]. An agent is given a goal directive and step-by-step instructions to perform mobile manipulation of multiple subgoals. Our work can omit step-by-step instructions.

agent parses human directives to make corresponding plans. Then, it perceives the surroundings to localize semantic entities, navigates to desired waypoints, and interacts with objects.

We center on primitive tasks of mobile manipulations in this work, which necessitates fundamental capability to navigate and interact based on an instructed *verb-noun pair*, *e.g. Pickup Lettuce*. Besides, leveraging the impressive power of language models [2,13,14,22,40] in high-level task planning, our method leads to a comprehensive embodied application.

Existing works for mobile manipulation include neural policies [7, 41, 44, 47] and map-based planning [5, 24, 36]. The former requires numerous training trajectories and annotations of high costs and suffers from the conflict of long-horizon nature and memory-less perception. The latter lacks flexibility in execution and hardly self-adapts in running time. To this end, we present DISCO (DIfferentiable Scene Semantics and Dual-level COntrol). It learns dynamic scene representations of objects and affordances on the fly, which facilitates map-based coarse navigation planning. A neural policy is deployed next to perform fine controls and boost object interaction.

Learning a dependable spatial representation of the scene is crucial for robotic applications, which can be 2D top-down view [36] or 3D voxels [5] in practice. An ideal scene representation should embody the following attributes: (1) Rich semantics in objects and affordances. It encapsulates objects and potential actions in the spatial space. (2) On-the-fly update. The scene is always dynamic and subject to changes upon interaction. (3) Easy accessibility for queries. It can be queried to facilitate downstream tasks like map-based trajectory planning. (4) Generalizability. The representation can be learned in unseen scenes.

To the best of our knowledge, previous research in mobile robotics has hardly developed representations encompassing all these attributes. In contrast, we build differentiable scene representations with all these attributes and demonstrate its prowess in the realm of interactive navigation.

Recent large models [6,7] equipped with scaled datasets and model capacity, have demonstrated successes in generating end-to-end neural actions. However, when faced with limited data, neural policies [41,47] struggle on academic benchmarks [44], primarily due to the data-hungry issue. Besides, mobile manipulation tasks often involve lengthy trajectories, yet semantic objectives within egocentric observations are rare. This scarcity presents significant challenges to neural policies. To address this, we direct actions based on global and local spatial cues and formulate dual-level coarse-to-fine controls. First, we design analytical controls on the map to drive the agent coarsely toward the object, with global cues from the scene representations. Subsequently, a fine-grained short-horizon neural control tailored to local egocentric observation is designed to fine-tune the pose and manipulate the object efficiently. Prior works use manually-defined rules [36] or human-in-the-loop feedback [37] to perform some sort of adjustments. But they are hard to formulate and scale up. Our dual-level control paradigm reduces the need for lengthy action trajectories and enhances overall efficiency.

DISCO can easily integrate into embodied applications such as embodied instruction following, where an embodied agent takes multimodal inputs to accomplish mobile manipulation tasks. We deploy DISCO on the widely-used ALFRED [44] benchmark consisting of long-horizon vision-language navigation and interaction tasks simulated in AI2THOR [27] environment as a testbed. ALFRED incorporates high-level human language directives to define the ultimate goal of each task, coupled with low-level step-by-step instructions for agent planning, as depicted in Fig. 1. In this work, we also challenge the planning ability in long-horizon tasks under a setting that omits low-level instructions. In extensive experiments, we achieve a substantial 11.0% gain of success rate in unseen scenes without step-by-step instructions, which even outperforms the state-of-the-art method that uses step-by-step instructions by 8.6%. We have also made thorough analyses and qualitative studies for a comprehensive evaluation.

Our contributions include: (1) We develop differentiable representations enriched with object and affordance semantics. It is dynamic, easy to query, and can be easily deployed in unseen environments. (2) We propose a dual-level approach that integrates both global and local cues for coarse-to-fine controls, enabling efficient mobile manipulation within limited imitation data. (3) In extensive experiments, we have evaluated our agent on ALFRED [44] benchmark and achieved new *state-of-the-art* performance with sizeable improvements.

2 Related Works

Embodied Navigation and Interaction. Mobile robotics requires fundamental ability in navigation and interaction. In past years, many simulators, scene assets, and benchmarks [3,9–11,15,19,27–29,43,44,48,50,53,56], including both

indoor and outdoor scenes, have been developed to facilitate algorithm researches in embodied AI. Early works only require navigation in static environments, such as PointNav [1] and ObjectNav [4]. However, subsequent studies [29, 44, 50] have expanded to interactive navigation, with semantic changes in dynamic scenes and object manipulation. The body of related works can be diversified into different categories based on modalities. Some employ natural language as instructions, as seen in [3, 28, 44]. In contrast, other works like SoundSpace [9] utilize audio instructions, and DialFRED [17] incorporates dialogue into navigation. Besides, embodied mobile manipulation agents should possess the ability to perform complex reasoning tasks, such as room rearrangement [50] and question answering [11, 19, 45] as well. The long-term target of indoor navigation and interaction research is to build agents to accomplish everyday household activities like humans [48].

Works on ALFRED. As a widely-used interactive vision-language navigation benchmark, ALFRED [44] attracts much interest from the research community. Existing works on ALFRED can be divided into model-free [25, 39, 41, 44, 46, 47, 54] and model-based [5, 24, 33, 36, 37] schools. The former school deploys an end-to-end neural policy to generate actions, necessitating extensive and costly training trajectories and instructional annotations. Early works [41, 44] utilized LSTM or Transformer to encode visual observation, language instructions, and action history for next-step prediction. They are trained to mimic expert behavior but exhibit subpar performance in novel environments. Some other techniques like panoramic observation [25] and instruction alignment [47] have been proposed for improvement. In contrast, the latter school constructs a model of the scene to facilitate action planning. The modeled scene can be a representation of 3D voxels [5] or 2D top-down views [24, 36, 37]. Object search modules [24, 36] were designed to help agents find objects. In-context planning and memory were explored in [26]. In our work, the novel dual-level control utilizes the advantages of both sides.

A□ordance. Affordance [18] reveals the potential interactions in the physical world. It is a multidisciplinary concept of vision, cognition, and robotics. Affordance can be learned from 2D [31] and 3D [12, 52] visual contents, language model reasoning [23], experienced interactions [38] and reinforced value estimations [2]. This concept finds a variety of robotic applications, such as scene exploration [38], optimal view selection [30], and mobile manipulation [7, 23, 49]. In our study, we utilize the ground-truth knowledge from the embodied simulator and learn affordances through supervised learning.

LLMs for Mobile Manipulation. The recent advancements in Large Language Models (LLMs) suggest their substantial potential in scaling robotic mobile manipulations. LLMs contribute significantly to mobile agents by facilitating scene understanding [22], task planning [14, 51], affordance grounding [2] and decision making [6, 7]. They can be integrated into the robotic navigation framework via prompted query [55], multi-expert discussion [34], or fine-tuning [51], paving for more comprehensive robots.

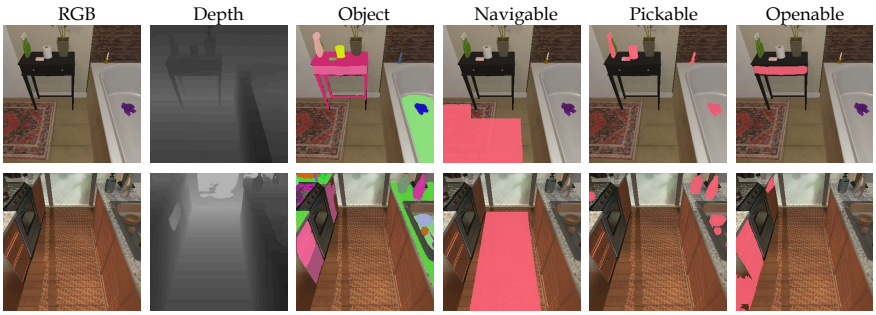

Fig. 2: The perception foundation. (i) 1st column: egocentric RGB frames as the initial observation. **(ii) 2nd column:** depth estimations. **(iii) 3rd column:** object instance segmentations. **(iv) 4th-6th columns:** affordance masks predictions: the navigable mask and two interactable masks (namely pickable and openable) as references.

3 Approach

We introduce how DISCO works in this section, including the perception system in Sec. 3.1, the scene representation in Sec. 3.2, the dual-level coarse-to-fine controls in Sec. 3.3, and its application to Embodied Instruction Following in Sec. 3.4.

3.1 Perception

Our agent perceives the surroundings from an egocentric RGB frame. We use three neural nets to estimate finer-grained spatial information: depth, instance segmentation, and affordances. All three generate pixel-wise information about the frame. Fig. 2 gives an intuitive example. We deploy a Mask R-CNN [20] to detect objects, and two U-Nets [42] to estimate depth and affordances respectively. The architecture settings of segmentation and depth estimation modules are the same as baselines [5, 24, 36].

Training reliable perception modules requires a substantial amount of high-quality curated data. They are of significant cost in the real world, but in our study, we collect data via querying a simulation oracle in AI2THOR [27]. The ground-truth depth is from the depth sensor. The ground-truth object segmentation of 85 classes is from the projection of the simulated objects in AI2THOR. The collected affordance includes one navigation class and seven interaction classes. For navigability, we discretize the scene into grids of $25cm \quad 25cm$, aligning with the agent moving step size. Then we teleport our agent to traverse over all grids and determine navigable ones. Pixels in the egocentric view that are localized in the navigable grids form the navigability mask. For interactivity, we directly query actionable properties of objects (*e.g.* *Openable*) from AI2THOR and merge all actionable objects in the view into the affordance mask of a certain interaction class.

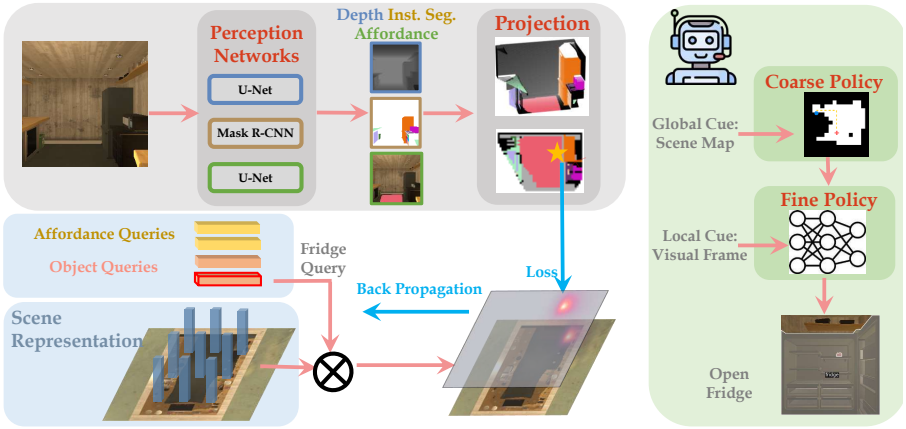

Fig. 3: An overview of the DISCO framework. Starting from the egocentric RGB frame, our perception system predicts pixel-wise depth, instance segmentation, and affordance frames. They are converted into semantic point clouds via projection and localized in the scene. We build differentiable scene representations with semantic queries to model the scene. They are optimized using gradient descent to match localized point cloud semantics. We apply dual-level coarse-to-fine controls. The coarse control depends on the global semantic map to approach the localized target. The fine control leverages a neural policy based on the local visual frame to interact.

We collect all frames from training trajectories to train perception networks, while unseen scenes are strictly prohibited. The Mask R-CNN [20] network for instance segmentation is initialized from a COCO [32] pre-trained checkpoint and then is finetuned by AdamW for 15 epochs with base learning rate 2e-4 and batch size 60. Default Mask R-CNN losses are adopted. A linear warm-up [35] of the learning rate is used in the first 1,000 steps. The U-Net [42] for depth estimation is optimized by AdamW for 15 epochs with base learning rate 1e-3 and batch size 80. The depth of each pixel is discretized into 50 bins of $10cm$ each and trained via a cross-entropy loss. The U-Net [42] for affordance estimation is optimized by AdamW for 25 epochs with base learning rate 1e-3 and batch size 80. We use binary cross-entropy loss to supervise all classes.

3.2 Learning Scene Representations

Prior works [5,24,36] mainly utilize cell-based representations to model the scene. However, discrete cells suffer from imperfect perception issues, *e.g.* hand-crafted rules are required to fix an object miss in one frame. We leverage continuous features to learn more robust scene representations. It softly models the scene map with a trade-off between historical and current observation. Different from the matching mechanism in previous continuous representation work [16], we use gradients to update the scene.

We model the scene as a $20m \quad 20m$ room and discretize it into $25cm \quad 25cm$ squares, which leads to $M \quad M(M = 80)$ grids in total. This configuration can cover all scenes in AI2THOR [27] and aligns the moving step size. Each grid is allocated a C dimensional($C = 256$ in our implementation) embedding. Additionally, we initialize $N^o + N^a$ semantic queries of C dimensions each, where N^o is the number of object classes and N^a is the number of affordance classes.

Our scene representation facilitates straightforward querying and yields a semantic map. Let $s_i(i = 1, 2 \quad , M^2)$ be the scene representation of the i-th grid, $q_j(j = 1, 2 \quad , N^o + N^a)$ be the j-th semantic query vector. We obtain the probability $p_{i;j}$ of the j-th class at i-th grid by a composition function f, such as $p_{i;j} = f(s_i, q_j)$. For system efficiency, we employ a minimal query mechanism via inner-dot followed by sigmoid, *i.e.* $f(s_i, q_j) = \sigma(s_i^T \ q_j)$, where σ is the sigmoid function. We apply zero-value initialization for s_i while random initialization from a normal distribution for q_j, thus it predicts $p_{i;j} = 0.5$ of no certainty at the beginning of the episode.

A scene is usually dynamically changed subject to embodied interactions. Therefore, a key factor of our scene representation is on-the-fly optimization at each step. We illustrate the learning framework in Fig. 3 and describe the more detailed process below. Starting from the egocentric RGB frame, our perception system predicts extra depth, segmentation, and affordance frames of rich geometric and semantic information. Following the semantic mapping procedure in [8, 36], the egocentric frame is converted to point clouds with semantics via camera projection. Notably, AI2THOR uses a discrete action space thus we estimate camera pose by the accumulation of historic actions. Top-down projection is followed to generate an allocentric point map of the current observation. The localized semantic-aware points are used to supervise the scene optimization. Let c_i be localized points in i-th grid and c_i^j be the points of j-th semantics among them. Then c_i^j / c_i is the proportion of semantic points. We normalize the proportion to get the soft grid-level semantic label y_i^j:

$$y_i^j = \frac{c_i^j / c_i}{\max_{1 \le k \le M^2} f c_k^j / c_k g}, \tag{1}$$

where the positive label 1 is assigned for the largest semantic proportion and other labels decrease. We optimize representations of visible grids at each step while leaving other grids unchanged. We threshold localized point cloud density to determine visibility, *i.e.* the visibility of i-th grid $v_i = \mathbf{1}[c_i > \rho]$ where $\mathbf{1}[\,]$ is the binary indicator function and ρ ($\rho = 500$ in our implementation) is the threshold. With all the estimated utilities above, we optimize the scene using back-propagation. For visible grids with $v_i = 1$, we compute the cross-entropy loss between y_i^j and $f(s_i, q_j)$ to update the feature. The learning step is formu-

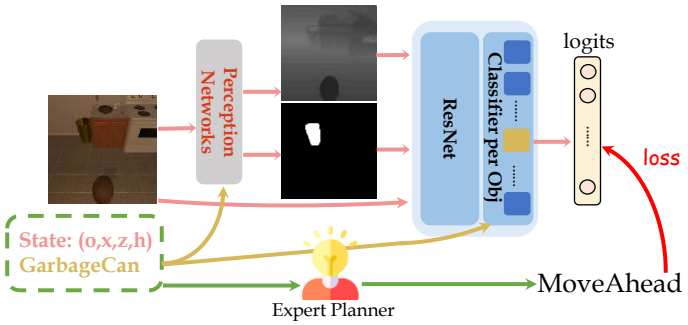

Fig. 4: The design of fine action control. DISCO employs a neural policy network to predict fine action steps. RGB, depth, and the object mask are sent to the network to derive a feature, followed by an object-specific classifier to predict the action. The policy is trained by mimicking expert actions.

lated as:

$$L(y_i^j, f(s_i, q_j)) = \quad (1 \quad y_i^j)(1 \quad f(s_i, q_j)) \quad y_i^j f(s_i, q_j), \tag{2}$$

$$s_i \quad s_i \quad \alpha \sum_j v_i \quad \frac{\partial}{\partial s_i} L(y_i^j, f(s_i, q_j)), \tag{3}$$

$$q_j \quad q_j \quad \alpha \sum_i v_i \quad \frac{\partial}{\partial q_j} L(y_i^j, f(s_i, q_j)). \tag{4}$$

For each step, we update s_i and q_j for 10 learning iterations with learning rate α = 0.01.

Till now, we have built scene representations that can be optimized by semantic scene differentiation. It omits some hand-crafted scene-updating strategies in prior work [36] and proves to be more generalizable.

3.3 Coarse-to-Fine Action Control

In this section, we introduce how DISCO acts to accomplish a primitive task represented as a verb-noun pair, *e.g. Pickup Lettuce*. It first starts with random walks until the target object is observed. Then we execute coarse-to-fine controls to perform mobile manipulation, where the coarse control navigates to approach near the object and fine control refines the agent pose for interaction.

Random Walk. Our agent begins with a random walk if the object has never been detected by the detector. We use the navigable query (one of the affordance classes) to obtain the navigability map of the scene. Then, we randomly select a navigable waypoint and apply the Breadth First Search (BFS) algorithm to plan a trajectory to the destination.

Coarse Control. The random walk terminates once the object is spotted by the detector. Then, we apply map-based coarse control to navigate the agent

close to the localized target. The coarse control is based on global cues, *i.e.* semantic scene maps. We query the scene representations to obtain probabilistic maps of the semantic objects and affordances of all grids, where affordances play a supplement role in localizing a desired interaction. Next, we select the grid of the largest object-affordance union probability as the target. For instance, when the agent is asked to *Pickup Lettuce*, it queries the object *Lettuce* distribution in the scene as well as the affordance *Pickupable* distribution, then the grid of maximal multiplied union probability is targeted. Our coarse control is designed to approach the object. We expand the target location and its neighboring grids within *1m* distance to be the destination of coarse navigation. BFS algorithm based on the navigability map is applied to plan trajectories.

Fine Control. Though map-based coarse actions are efficient in planning lengthy navigation trajectories. It can't be self-adapted to manipulate objects. The local state, such as view direction and distance to the object, greatly affects whether an interaction can be successful. Some refinement based on local cues, *i.e.* egocentric frames, are essential for better manipulations. To this end, we propose fine action controls by a neural network, illustrated in Fig. 4. In the fine action step, we formulate each state as (o, x, z, h), where o is the target object going to be manipulated, (x, z) is the agent location and h is the camera horizon. We adjust the agent direction to the target object by referring to the localized yaw first. This makes the target object visible. We use the concatenation of RGB, estimated depth, and the mask of the target object to be the input of the policy. The input includes geometric distance and identifies the target object. We use a ResNet50 [21] to encode the feature, followed by object class-specific classifiers to generate actions. The independent classifier design is motivated by the widely used object detector head design [20]. We find some insights into neural policy learning. First, we adjust agent rotation to keep the object in view, which reduces the ambiguity of control. Second, our neural policy is applied exclusively in short-horizon refinement stages, simplifying the learning process for lengthy trajectories. Notably, existing works [24,36] use hand-crafted rules or tests to deal with the openness of openable receptacles. Our affordance module (Fig. 2) automatically detects openable property and helps DISCO to act.

We build an expert planner with full knowledge of all scenes to help policy learning by imitation. The expert building process follows the generative pipeline of ALFRED [44]. It can label all interactable states and plan transiting actions by BFS search. Since coarse control navigates the agent close to the target within $1m$, the fine control next only requires short-horizon refining steps, which eases the difficulty of training. We collect data on states within 4 expert steps to interactions. We iterate over all objects and all short-horizon states to save frames, generating a training set of 316,935 frames. As default training trajectories of ALFRED [44] contain 1,051,308 frames, we find training short-horizon policy is more data-efficient. We train the policy by imitating planned actions from the expert, also known as behavior cloning. It is supervised by expert actions using a cross-entropy loss. We optimize the policy using an AdamW optimizer with a constant learning rate 5e-5 and train it for 40 epochs with a batch size of 100.

3.4 Application: Embodied Instruction Following

DISCO performs primitives of mobile manipulation tasks commanded by verb-noun pairs and can be easily applied in diverse embodied tasks. We take the embodied instruction following tasks from ALFRED [44] as a test bed, where the agent is commanded by natural language instructions to accomplish long-horizon tasks of many mobile manipulation subgoals. More introduction about the benchmark can be found in the supplementary material. However, though we take vision-language navigation and interaction to conduct main experiments, we believe DISCO has the potential for other modality instructions like sound [9] or dialog [17] with different instruction processing modules.

In the detailed implementation, for a fair comparison with the baseline, we inherit the instruction processing procedure and some building components from FILM [36]. Notably, though ALFRED provides low-level step-by-step instructions, our methods can also run without these annotations. High-level goal directives are only used by default to generate subgoal plans. To generate plans, we adopt fine-tuned BERTs [13] from [36] to parse natural language instructions into ALFRED internal parameters. Next, leveraging the patterned task nature of ALFRED, templates are used to convert parameters into multiple verb-noun manipulation subgoals. For instance, we take the case in Fig. 1 as an example. Language models recognize it as a *pick_clean_then_put* task with object argument *Lettuce* and receptacle argument *DinningTable*. It is converted to subgoal series: (*Pick*, *Lettuce*), (*Clean*, *Lettuce*), (*Put*, *DinningTable*) using the template. We give more details about the instruction processing and the holistic application in the supplementary material.

4 Experiments

4.1 Evaluation Protocols

We evaluate our method on ALFRED [44], consisting of large-scale long-horizon vision-language navigation and interaction tasks. The dataset is divided into *train*, *validation*, and *test* splits, containing 21,023/1,641/3,062 episodes respectively. We use *test* split to compare with competitive baselines, but *valid* split to make analyses. Both *valid* and *test* splits are divided into *seen* and *unseen* scenes. The *valid*/*test* splits have 820/1,533 episodes in seen scenes while 821/1528 episodes in unseen scenes.

Four metrics are used for evaluation. (1) Success Rate (SR). Rate of accomplished tasks with success. (2) Goal Condition (GC). Rate of achieved conditions for goals. (3) Path Length Weighted SR (PLWSR). Weighing SR by the agent trajectory length against expert trajectory length. (4) Path Length Weighted GC (PLWGC). Applying the same weight on GC. SR and GC mainly reflect the effectiveness of agents while PLW metrics reflect the efficiency of running steps. All metrics are the higher the better.

Table 1: Results in the test splits of ALFRED [44].

	step-by-step instructions	Test Seen				Test Unseen			
		SR	GC	PLWSR	PLWGC	SR	GC	PLWSR	PLWGC
Seq2Seq [44]	✔	4.0	9.4	2.0	6.3	3.9	7.0	0.1	4.3
MOCA [46]	✔	26.8	33.2	19.5	26.8	7.7	15.7	4.2	11.0
E.T. [41]	✔	38.4	45.4	27.9	34.9	8.6	18.6	4.1	11.5
LWIT [39]	✔	29.2	38.8	24.7	34.9	8.4	19.1	5.1	14.8
HiTUT [54]	✔	21.3	30.0	11.1	17.4	13.9	20.3	5.9	11.5
ABP [25]	✔	44.6	51.1	3.9	4.9	15.4	24.8	1.1	2.2
FILM [36]	✔	27.7	38.5	11.2	15.1	26.5	36.4	10.6	14.3
M-Track [47]	✔	24.8	33.3	13.9	19.5	16.3	22.6	7.7	13.2
LGS-RPA [37]	✔	40.1	48.7	21.3	29.0	35.4	45.2	15.7	22.8
Prompter [24]	✔	53.2	63.4	25.8	30.7	45.7	58.8	20.8	26.2
CAPEAM [26]	✔	51.8	60.5	21.6	25.9	46.1	57.3	19.5	24.1
DISCO (Ours)	✔	**59.5**	**66.1**	**40.6**	**47.4**	**56.5**	**66.8**	**36.5**	**44.5**
HiTUT [54]	✗	13.6	21.1	5.6	11.0	11.1	17.9	4.5	9.8
HLSM [5]	✗	25.1	35.8	6.7	11.5	16.3	27.2	4.3	8.5
FILM [36]	✗	25.8	36.2	10.4	14.2	24.5	34.8	9.7	13.1
LGS-RPA [37]	✗	33.0	41.7	16.7	24.5	27.8	38.6	12.9	20.0
EPA [33]	✗	40.0	44.1	2.6	3.5	36.1	39.6	2.9	3.9
Prompter [24]	✗	49.4	55.9	23.5	29.1	42.6	59.6	19.5	25.0
CAPEAM [26]	✗	47.4	54.4	19.0	23.8	43.7	54.6	17.6	22.8
DISCO (Ours)	✗	**58.0**	**64.9**	**39.6**	**46.5**	**54.7**	**65.5**	**35.5**	**43.6**

4.2 Baselines

We adopt competitive baselines reported on ALFRED to compare with our method. They include Seq2Seq [44], MOCA [46], E.T. [41], LWIT [39], ABP [25], M-Track [47], HiTUT [54], HLSM [5], FILM [36], LGS-RPA [37], EPA [33], prompter [24], and CAPEAM [26]. All methods use RGB vision and language instructions in test time. While step-by-step instructions can be omitted by some methods as well as DISCO, we separate the comparison for fairness. By default, DISCO only uses a high-level description of the goal.

4.3 Quantitative Comparisons

We report the quantitative results of DISCO and competitive baselines in test splits of ALFRED in Tab. 1 for fair comparisons. We find our method outperforms the art by sizable improvements in all slots. Equipped with step-by-step instructions, our method achieves 59.5% and 56.5% success rates in seen and unseen scenes, with substantial gains of 6.3% and 10.4% over the state-of-the-art methods, showcasing the effectiveness of DISCO. The superiority is consistent with other metrics. Integrating the path length into metrics, we achieve almost 1.57x and 1.75x performances than Prompter [24] on PLWSR metrics in seen

Table 2: Ablation study.

	Valid Seen		Valid Unseen	
	SR	GC	SR	GC
DISCO	57.3	63.9	55.0	65.5
+ step-by-step instr.	65.1	70.8	59.1	68.6
+ gt. lang.	70.5	75.5	64.1	71.9
+ gt. percep.	67.2	73.4	66.8	73.9
+ gt. percep. lang.	79.9	84.2	79.6	83.0
w.o. differentiable	47.4	54.0	42.7	52.3
w.o. interactive aff.	52.5	57.9	51.3	56.6
w.o. navigation aff.	48.1	56.1	46.0	54.8
w.o. coarse control	13.5	16.2	9.1	10.3
w.o. fine control	53.0	58.3	51.7	57.3

and unseen scenes, which means our agent accomplishes tasks using fewer steps in execution. This strongly validates the efficiency. Besides, in terms of goal condition metrics, we also have superior performances over baselines, namely 66.1% and 66.8% in seen and unseen scenes respectively. Next, we omit step-by-step instructions and challenge DISCO only using high-level goals. Under this setting, our method achieves 58.0% and 54.7% success rates in seen and unseen scenes, outperforming baselines by 8.6% and 11.0%. A bigger surprise is that DISCO without step-by-step instructions outperforms the art with step-by-step instructions, where we achieve success rate gains of 4.8% and 8.6% in seen and unseen scenes respectively. It proves we surpass all existing methods with non-trivial advancement in both efficiency and effectiveness. Overall, our method establishes new *state-of-the-art* performances on ALFRED.

4.4 Ablation Study

We conduct comprehensive ablation studies on the valid splits of ALFRED to explore the effects of components of DISCO. Results are reported in Tab. 2.

Increments of stronger multimodal inputs. We augment DISCO with stronger instruction understanding and perception modules. Equipped with low-level step-by-step instructions, DISCO achieves 65.1% and 59.1% success rates in seen and unseen scenes. Ground-truth language parsing pushes the results to 70.5% and 64.1%. This uncovers the potential of stronger language models in boosting embodied planning. Ground-truth perceptions further enhance performances to 79.9% and 79.6% in seen and unseen scenes.

Differentiable representations. We replace differentiable representations with widely-used cell representations [36]. Noticeable success rate drops of -9.9% and -12.3% are observed in seen and unseen scenes. This is because our differentiable representations provide soft semantics such as Eq. (1) and benefit acting. In contrast, cell-based representations are binary and require hand-crafted rules to

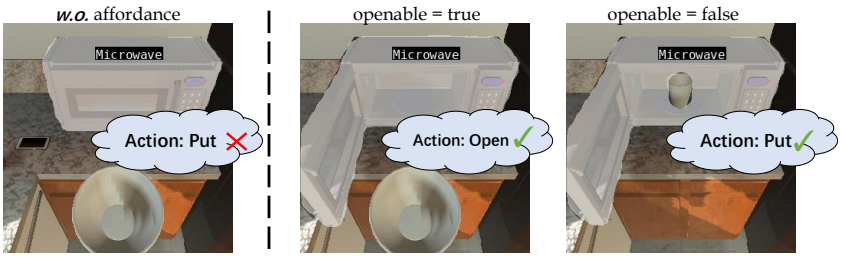

Fig. 5: Qualitative case of affordance. Left: The agent fails to put the bowl into the microwave without openable knowledge. Right: DISCO is aware of the openable affordance property in microwave interaction.

fix inconsistent perceptions. Our differentiable representations balance historical and current observations to alleviate the issue.

A□ordance. We first remove interactive affordances to validate their effects. In this test, the openable property and affordance-augmented localization are not used. We achieve 52.5% and 51.3% SR in seen and unseen scenes, with drops of -5.1% and -3.7%. Next, we replace the navigability affordance with point-obstacle-based navigation methods. The performance drops become larger, namely -9.3% and -9.0% SR in seen and unseen scenes respectively. These experiments validate affordance knowledge is crucial for embodied applications.

Dual-level controls. We remove the dual-level controls and apply coarse or fine control individually. Removing coarse control leads to a crash of our system, with extremely low success rates. Besides, removing fine control leads to SR performance drops of -4.3% and -3.3% in seen and unseen scenes respectively. The finding suggests that coarse control plays its role in most lengthy moving steps of irreplaceable significance. However, the fine control makes some local refinement to facilitate interactions.

4.5 Qualitative Results

We provide qualitative cases to reveal some running facts about DISCO.

A□ordance. We demonstrate the use of affordance in Fig. 5. Prior works use manually defined rules or tests to help interact with openable receptacles. Our affordance module can automatically detect the opening state and facilitate acting. Without the affordance knowledge, the agent cannot place the bowl into the microwave properly.

Dual-level control. Our control policy is illustrated in Fig. 6. DISCO executes map-based coarse actions to approach the object first. However, the object target is usually not interactable after the coarse stage. They may be attributed to state change expansion, view direction, position offset, and more diversified reasons. To this end, we apply neural fine action controls for short-horizon self-adjustment to flexibly address the trouble. We provide more qualitative results in the supplementary material.

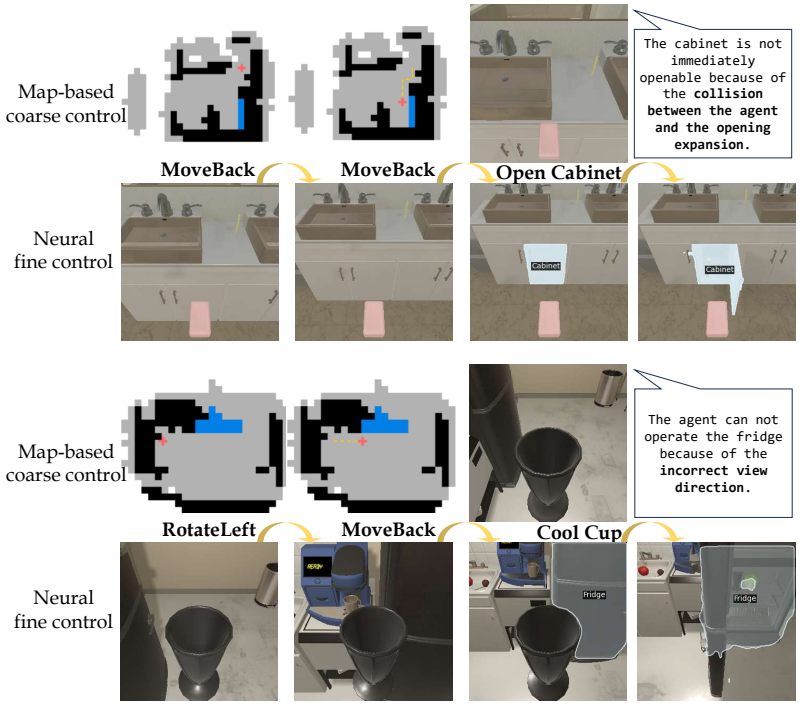

Fig. 6: Qualitative running cases. DISCO applies map-based coarse actions followed by neural fine actions. The agent may suffer from the opening collision (upper) or incorrect view direction (bottom) after the coarse navigation. Fine actions perform self-adjustment to facilitate interactions. More cases are in the supplementary material.

5 Conclusion

We introduce DISCO to perform mobile manipulation tasks in this work. It learns differentiable scene representations of rich semantics in objects and affordances in online exploration. The scene representations retrieve target semantics and facilitate map-based navigation planning. We propose dual-level coarse-to-fine action controls, which leverage a global scene map to coarsely approach the navigation target followed by neural fine actions to boost object interaction. We leverage language models to plan primitive tasks and integrate DISCO into an Embodied Instruction Following application. In extensive experiments, DISCO achieves new *state-of-the-art* results on ALFRED.

Acknowledgments

This work was supported by the National Key Research and Development Project of China (No. 2022ZD0160102), National Key Research and Development Project of China (No. 2021ZD0110704), Shanghai Artificial Intelligence Laboratory, and XPLORER PRIZE grants.

References

1. Abhishek Kadian*, Joanne Truong*, Gokaslan, A., Clegg, A., Wijmans, E., Lee, S., Savva, M., Chernova, S., Batra, D.: Are We Making Real Progress in Simulated Environments? Measuring the Sim2Real Gap in Embodied Visual Navigation. In: arXiv:1912.06321 (2019) 4
2. Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Ho, D., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jang, E., Ruano, R.J., Jeffrey, K., Jesmonth, S., Joshi, N.J., Julian, R., Kalashnikov, D., Kuang, Y., Lee, K.H., Levine, S., Lu, Y., Luu, L., Parada, C., Pastor, P., Quiambao, J., Rao, K., Rettinghouse, J., Reyes, D., Sermanet, P., Sievers, N., Tan, C., Toshev, A., Vanhoucke, V., Xia, F., Xiao, T., Xu, P., Xu, S., Yan, M., Zeng, A.: Do as i can, not as i say: Grounding language in robotic affordances (2022) 2, 4
3. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 1, 3, 4
4. Batra, D., Gokaslan, A., Kembhavi, A., Maksymets, O., Mottaghi, R., Savva, M., Toshev, A., Wijmans, E.: ObjectNav Revisited: On Evaluation of Embodied Agents Navigating to Objects. In: arXiv:2006.13171 (2020) 4
5. Blukis, V., Paxton, C., Fox, D., Garg, A., Artzi, Y.: A persistent spatial semantic representation for high-level natural language instruction execution. In: Conference on Robot Learning. pp. 706–717. PMLR (2022) 2, 4, 5, 6, 11
6. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., Florence, P., Fu, C., Arenas, M.G., Gopalakrishnan, K., Han, K., Hausman, K., Herzog, A., Hsu, J., Ichter, B., Irpan, A., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, L., Lee, T.W.E., Levine, S., Lu, Y., Michalewski, H., Mordatch, I., Pertsch, K., Rao, K., Reymann, K., Ryoo, M., Salazar, G., Sanketi, P., Sermanet, P., Singh, J., Singh, A., Soricut, R., Tran, H., Vanhoucke, V., Vuong, Q., Wahid, A., Welker, S., Wohlhart, P., Wu, J., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., Zitkovich, B.: Rt-2: Vision-language-action models transfer web knowledge to robotic control (2023) 3, 4
7. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jackson, T., Jesmonth, S., Joshi, N.J., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, K.H., Levine, S., Lu, Y., Malla, U., Manjunath, D., Mordatch, I., Nachum, O., Parada, C., Peralta, J., Perez, E., Pertsch, K., Quiambao, J., Rao, K., Ryoo, M., Salazar, G., Sanketi, P., Sayed, K., Singh, J., Sontakke, S., Stone, A., Tan, C., Tran, H., Vanhoucke, V., Vega, S., Vuong, Q., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., Zitkovich, B.: Rt-1: Robotics transformer for real-world control at scale (2023) 2, 3, 4
8. Chaplot, D.S., Gandhi, D., Gupta, S., Gupta, A., Salakhutdinov, R.: Learning to explore using active neural slam. In: International Conference on Learning Representations (ICLR) (2020) 7
9. 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) 3, 4, 10

10. Chen, H., Suhr, A., Misra, D., Snavely, N., Artzi, Y.: Touchdown: Natural language navigation and spatial reasoning in visual street environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12538–12547 (2019) 3
11. Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., Batra, D.: Embodied Question Answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 3, 4
12. Deng, S., Xu, X., Wu, C., Chen, K., Jia, K.: 3d affordancenet: A benchmark for visual object affordance understanding (2021) 4
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 2, 10
14. Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., Florence, P.: Palm-e: An embodied multimodal language model. In: arXiv preprint arXiv:2303.03378 (2023) 2, 4
15. Ehsani, K., Gupta, T., Hendrix, R., Salvador, J., Weihs, L., Zeng, K.H., Singh, K.P., Kim, Y., Han, W., Herrasti, A., et al.: Imitating shortest paths in simulation enables effective navigation and manipulation in the real world. arXiv preprint arXiv:2312.02976 (2023) 3
16. Gadre, S., Ehsani, K., Song, S., Mottaghi, R.: Continuous scene representations for embodied ai. CVPR (2022) 6
17. Gao, X., Gao, Q., Gong, R., Lin, K., Thattai, G., Sukhatme, G.S.: Dialfred: Dialogue-enabled agents for embodied instruction following. IEEE Robotics and Automation Letters **7**(4), 10049–10056 (oct 2022). https://doi.org/10.1109/lra.2022.3193254 4, 10
18. Gibson, J.J.: The ecological approach to the visual perception of pictures. Leonardo **11**(3), 227–235 (1978) 4
19. Gordon, D., Kembhavi, A., Rastegari, M., Redmon, J., Fox, D., Farhadi, A.: Iqa: Visual question answering in interactive environments. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4089–4098 (2018) 3, 4
20. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017) 5, 6, 9
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016) 9
22. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models. arXiv (2023) 2, 4
23. Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., Fei-Fei, L.: Voxposer: Composable 3d value maps for robotic manipulation with language models (2023) 4
24. Inoue, Y., Ohashi, H.: Prompter: Utilizing large language model prompting for a data efficient embodied instruction following (2022). https://doi.org/10.48550/ARXIV.2211.03267, https://arxiv.org/abs/2211.03267 2, 4, 5, 6, 9, 11
25. Kim, B., Bhambri, S., Singh, K.P.: Agent with the big picture: Perceiving surroundings for interactive instruction following (2021) 4, 11
26. Kim, B., Kim, J., Kim, Y., Min, C., Choi, J.: Context-aware planning and environment-aware memory for instruction following embodied agents. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10936–10946 (2023) 4, 11

27. 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 preprint arXiv:1712.05474 (2017) 3, 5, 7
28. Ku, A., Anderson, P., Patel, R., Ie, E., Baldridge, J.: Room-Across-Room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In: Conference on Empirical Methods for Natural Language Processing (EMNLP) (2020) 1, 3, 4
29. Li, C., Xia, F., Martín-Martín, R., Lingelbach, M., Srivastava, S., Shen, B., Vainio, K., Gokmen, C., Dharan, G., Jain, T., et al.: igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. arXiv preprint arXiv:2108.03272 (2021) 3, 4
30. Li, X., Wang, Y., Shen, Y., Iaroslav, P., Lu, H., Wang, Q., An, B., Liu, J., Dong, H.: Imagemanip: Image-based robotic manipulation with affordance-guided next view selection (2023) 4
31. Li, Y.L., Xu, Y., Xu, X., Mao, X., Yao, Y., Liu, S., Lu, C.: Beyond object recognition: A new benchmark towards object concept learning (2023) 4
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 6
33. Liu, X., Palacios, H., Muise, C.: A planning based neural-symbolic approach for embodied instruction following. In: CVPR Embodied AI Workshop (2022) 4, 11
34. Long, Y., Li, X., Cai, W., Dong, H.: Discuss before moving: Visual language navigation via multi-expert discussions (2023) 4
35. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts (2017) 6
36. Min, S.Y., Chaplot, D.S., Ravikumar, P., Bisk, Y., Salakhutdinov, R.: Film: Following instructions in language with modular methods. arXiv preprint arXiv:2110.07342 (2021) 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
37. Murray, M., Cakmak, M.: Following natural language instructions for household tasks with landmark guided search and reinforced pose adjustment. IEEE Robotics and Automation Letters **7**(3), 6870–6877 (2022) 3, 4, 11
38. Nagarajan, T., Grauman, K.: Learning affordance landscapes for interaction exploration in 3d environments (2020) 4
39. Nguyen, V.Q., Suganuma, M., Okatani, T.: Look wide and interpret twice: Improving performance on interactive instruction-following tasks. arXiv preprint arXiv:2106.00596 (2021) 4, 11
40. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R.: Training language models to follow instructions with human feedback (2022) 2
41. Pashevich, A., Schmid, C., Sun, C.: Episodic Transformer for Vision-and-Language Navigation. In: ICCV (2021) 2, 3, 4, 11
42. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015) 5, 6
43. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 3

44. 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: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10740–10749 (2020) 1, 2, 3, 4, 9, 10, 11
45. Shridhar, M., Yuan, X., Côté, M.A., Bisk, Y., Trischler, A., Hausknecht, M.: ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In: Proceedings of the International Conference on Learning Representations (ICLR) (2021), https://arxiv.org/abs/2010.03768 4
46. Singh, K.P., Bhambri, S., Kim, B., Mottaghi, R., Choi, J.: Factorizing perception and policy for interactive instruction following. arXiv preprint arXiv:2012.03208 (2020) 4, 11
47. Song, C.H., Kil, J., Pan, T.Y., Sadler, B.M., Chao, W.L., Su, Y.: One step at a time: Long-horizon vision-and-language navigation with milestones. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15482–15491 (2022) 2, 3, 4, 11
48. Srivastava, S., Li, C., Lingelbach, M., Martín-Martín, R., Xia, F., Vainio, K., Lian, Z., Gokmen, C., Buch, S., Liu, K., Savarese, S., Gweon, H., Wu, J., Fei-Fei, L.: Behavior: Benchmark for everyday household activities in virtual, interactive, and ecological environments. In: Conference in Robot Learning (CoRL). p. accepted (2021) 1, 3, 4
49. Wang, Y., Wu, R., Mo, K., Ke, J., Fan, Q., Guibas, L., Dong, H.: AdaAfford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions (2022) 4
50. Weihs, L., Deitke, M., Kembhavi, A., Mottaghi, R.: Visual room rearrangement. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2021) 3, 4
51. Wu, Z., Wang, Z., Xu, X., Lu, J., Yan, H.: Embodied task planning with large language models (2023) 4
52. Xu, C., Chen, Y., Wang, H., Zhu, S.C., Zhu, Y., Huang, S.: Partafford: Part-level affordance discovery from 3d objects (2022) 4
53. Yenamandra, S., Ramachandran, A., Yadav, K., Wang, A., Khanna, M., Gervet, T., Yang, T.Y., Jain, V., Clegg, A.W., Turner, J., Kira, Z., Savva, M., Chang, A., Chaplot, D.S., Batra, D., Mottaghi, R., Bisk, Y., Paxton, C.: Homerobot: Open vocabulary mobile manipulation (2023) 3
54. Zhang, Y., Chai, J.: Hierarchical task learning from language instructions with unified transformers and self-monitoring. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 4202–4213. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.findings-acl.368, https://aclanthology.org/2021.findings-acl.368 4, 11
55. Zhou, G., Hong, Y., Wu, Q.: Navgpt: Explicit reasoning in vision-and-language navigation with large language models (2023) 4
56. Zhu, Y., Gordon, D., Kolve, E., Fox, D., Fei-Fei, L., Gupta, A., Mottaghi, R., Farhadi, A.: Visual semantic planning using deep successor representations. In: Proceedings of the IEEE international conference on computer vision. pp. 483–492 (2017) 3