2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

# Embodied Referring Expression for Manipulation Question Answering in Interactive Environment

Qie Sima, Sinan Tan, Huaping Liu†, Fuchun Sun, Weifeng Xu and Ling Fu

***Abstract*—Embodied agents are expected to perform more complicated tasks in an interactive environment, with the progress of Embodied AI in recent years. Existing embodied tasks including Embodied Referring Expression (ERE) and other QA-form tasks mainly focuses on interaction in term of linguistic instruction. Therefore, enabling the agent to manipulate objects in the environment for exploration actively has become a challenging problem for the community. To solve this problem, We introduce a new embodied task: Remote Embodied Manipulation Question Answering (REMQA) to combine ERE with manipulation tasks. In REMQA task, the agent needs to navigate to a remote position and perform manipulation with the target object to answer the question. We build a benchmark dataset for the REMQA task in AI2-THOR simulator. To this end, a framework with 3D semantic reconstruction and modular network paradigms is proposed. The evaluation of the proposed framework on REMQA dataset is presented to validate its effectiveness.**

***Index Terms*—Embodied AI, Referring Expression,Visual Semantics, Question Answering**

## I. Introduction

Recently, the AI community has witnessed the prosperity of Embodied AI where agents are required to perform tasks in various forms with egocentric vision. The success of embodied AI brings up the interest of researchers in the robot community to transfer methods in off-shelf Embodied AI tasks to robot platforms.

Currently, most of works in Embodied AI have revolved around the task of navigation – including position-goal, object-goal, and area-goal [1]. However, the ability to actively manipulate objects and physically interact with the environment becomes crucial in the embodied robot task, where agents need to perform complex tasks in the real world. As the studies on embodied tasks have surged in recent years, a wide variety of embodied tasks has been proposed. However, very few works have looked into a general framework for embodied tasks that involve most modular models of robot task in the real world: Visual reception, Language comprehension, Active navigation and Manipulation. In an embodied robot task, how to localize the target object precisely and effectively has always been a challenge. Since many objects in the real scene are similar in shape and appearance (e.g., books on shelf, cabinets in the kitchen).

Referring Expression (RE) is a widely studied cross-modal task in both computer vision and natural language processing fields as a vision and language task. In a RE task, the agent needs to localize a specific target object in the image in response to a given natural language referring expression. Most of current studies in referring expression focus on passive image datasets (e.g. RefCOCO, RefCOCO+ [2], RefCOCOg [3]) where samples will not change with agent's decision. Recently, referring expression tasks in embodied scenarios has emerged. In an Embodied Referring Expression (ERE) task, the agent is required to navigate to the position mentioned in the given expression in a 3D environment and complete the RE task on the final scene. However, the process of navigating to the target object scene in most of above tasks merely consists of spatial movements without interaction with surrounding environments, such as opening closed objects or moving occlusion.

Therefore, we introduce a novel embodied task **Remote Embodied Manipulation Question Answering (REMQA)** where the agent is required to navigate to a remote position and manipulate the target object, which can be precisely localized by referring expression comprehension. Then, the agent infers the answer to the question from the post-manipulation layout of objects. As we illustrate in Fig. 1, the input question consists of a referring phrase explicitly referring to the target object (drawer). After navigating to the goal position (toaster), the agent needs to localize it by distinguishing it from other drawers with referring expression comprehension and conducting manipulation action (open drawer) to get the answer.

In this work, we focus on the referring expression comprehension problems for **Manipulation Question Answering (MQA)** task in a physically interactive environment. The main contributions of this work are listed below:

- **Problem.** a novel embodied robot task consists of vision perception, language comprehension and manipulation in an interactive environment, Remote Embodied Manipulation Question Answering.
- **Dataset.** a benchmark dataset of proposed task with a set of indoor object arrangements of different rooms in an interactive environment and questions within referring expression about the objects in the environment.
- **Method.** a framework to handle the proposed task in which Language Attention Network and 3D semantic

Q. Sima, S. Tan, H. Liu and F. Sun are with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. W. Xu and L. Fu are with Siemens Ltd., China. This work was supported in part by the National Natural Science Fund for Distinguished Young Scholars under Grant 62025304, and in part by the Seed Fund of Tsinghua University (Department of Computer Science and Technology)-Siemens Ltd., China Joint Research Center for Industrial Intelligence and Internet of Things.

†Corresponding author (hpliu@tsinghua.edu.cn)

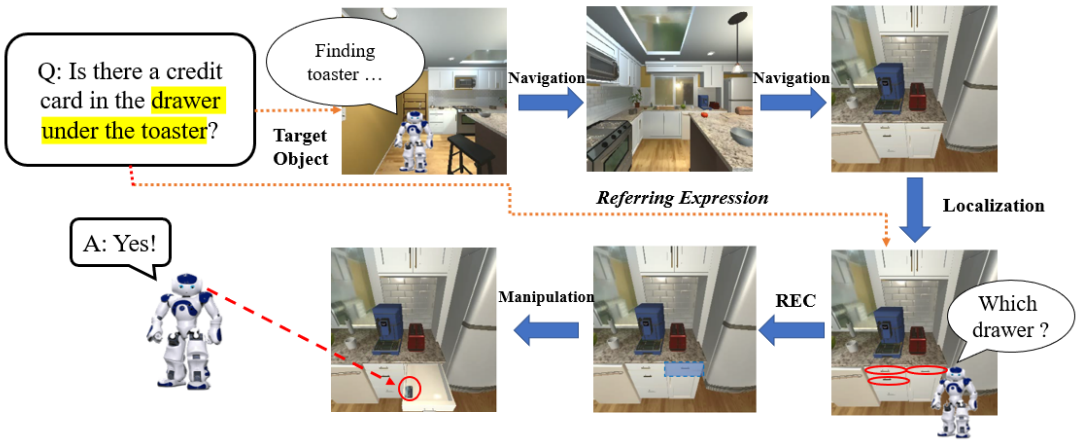



Fig. 1: A demonstration of the Remote Embodied Manipulation Question Answering task. The agent needs to navigate to the goal position, localize the target object and perform manipulation to answer the question.

memory prior-ed navigation are implemented. Experimental validation of the proposed model has been conducted in an interactive environment with the physical engine.

In the rest of the paper, Section II presents a review of related works. Summary of data for pretraining and proposed benchmark dataset is introduced in Section III. Section IV details our proposed model for the Remote Embodied Manipulation Question Answering task in an interactive environment. Section V presents the experimental results and Section VI concludes this work.

## II. Related Work

### A. Referring Expression

**Referring Expression on Static Dataset:** Most of works in referring expression comprehension focus on comprehension tasks in datasets built on classical static visual datasets (COCO, Flick et al.). Specifically, RE tasks can be categorized to two kinds with aspect to labels used for localization: 1) Referring Expression Comprehension (REC): a bounding box 2) Referring Expression Segmentation (RES): a segmentation mask. For REC task, Mao et al. [4]introduce the first CNN-LSTM method:MMI as a general solution to REC task. Yu et al. propose a visual comparative method (Visdif) to distinguish the target object from the surrounding objects rather than extracting features by CNN. Furthermore, Yu et al. [5] raise MAttNet: Modular Attention Network to decompose referring expressions into different modular channels for accurate matching. Besides CNN-LSTM methods, some works [6], [7]present models of the relationship between images and expressions and some others [8] utilize the pre-trained vision and language models for REC task. For RES task, Li et al. [9] propose a multi-modal LSTM for vision and linguistic fusion. To obtain more accurate results for long referring expressions, Shi et al. [10] employ an attention mechanism in raised keyword-aware network. Luo et al. [11] introduce Multi-task Collaborative Network (MCN) as a joint learning framework of RES.

**Embodied Referring Expression:** Due to the absence of interaction in conventional referring expression tasks, researchers have recently tried to transplant referring expression tasks to embodied scenarios. Several ERE tasks and datasets have been released in recent years. Most proposed ERE tasks can be classified into two main categories with aspect to platform: 1) ERE task in manipulator scenario: INGRESS [12] 2) ERE task in mobile navigation scenario: REVIERE [13], Touchdown-SDR [14], REVE-CE [15], ALFRED [16] The community havs developed several methods that enable agents to tackle embodied tasks that require active interaction with the environment. Wu et al. [13] propose a Navigator-Pointer model as a baseline for REVIERE dataset. Gao et al. employ room object-aware attention mechanism and transformer architecture in REVIERE. Lin et al. [17] pre-train agent with cross-modal alignment sub-tasks for ERE task.

### B. Embodied Robot Task

As an intersection of robotics, computer vision and natural language processing, the study of embodied robot tasks has gained much attention from all the above fields. A wide variety of embodied tasks has been formulated in recent years. The off-shelf embodied robot tasks can be categorized into two main types: Visual Navigation and Question Answering.

**Visual Navigation Task:** Visual Language Navigation(VLN) [18], Visual Semantic Navigation (VSN) [19] requires the agent to actively navigate to the goal position following linguistic information: language instructions for VLN semantic labels for VSN. Anderson et al. [18], [20] introduce the seq-to-seq framework and evaluation metrics for VLN task.

**Question Answering Task:** Antol et al. [19] firstly formulate Visual Question Answering (VQA). The agent needs to infer the answer from the image passively, which only relies on understanding questions and images. Many works [21], [22] on VQA have been proposed in the past decade.

In an EQA task [23], the agent is randomly spawned in a 3D environment and should explore the scenario with egocentric vision and answer the question with the final scene. Most of the methods proposed for EQA are based on Reinforcement Learning (RL) [23], [24]. Yu et al. [25] extend EQA to multi-target scenario. Tan et al. [26] employ a multi-agent system for EQA. Interactive Question Answering is an extension of EQA raised by Gordon et al. [27]. In IQA, the agent needs to do some simple standard virtual interactions (**e.g.** open the fridge) with the environment. Deng et al. [28] introduce Manipulation Question Answering (MQA), where a fixed-base manipulator is required to manipulate objects in the cluttered scene to render more information about objects initially unseen and answer the question better.

In following Table I, we compare the difference of several robot embodied tasks mentioned above. As we can see, only our proposed task has taken all 4 mentioned modules into consideration.

TABLE I: Comparison of different embodied robot tasks

| | VSN/VLN | VQA | EQA | IQA | MQA | Ours |
|---|---|---|---|---|---|---|
| Language | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Navigation | ✓ | - | ✓ | ✓ | - | ✓ |
| Interaction | - | - | - | ✓ | ✓ | ✓ |
| Manipulation | - | - | - | - | ✓ | ✓ |

## III. Dataset

### A. Embodied Environment

Training and evaluating an interactive agent in a real environment is temporarily uneconomic considering costs, time and generalizability. Therefore, we adopt AI2-THOR, a photo-realistic 3D environment simulator designed for embodied AI research [29], as our learning framework. The AI2Thor simulator consists of 120 different room layouts of 4 categories (Bedroom, Living room, Kitchen, and Bathroom), with 30 layouts for each category. We choose an extension of the AI2-THOR simulator: ManipulaTHOR, which has the same scenes as AI2-THOR and a realistic Kinova 6-DOF arm added to the agent [30]. ManipulaTHOR allows agents to interact physically with objects at a low control level via arm manipulators. Besides, several sensors including RGB-D frame, agent location, and arm configuration at the arm-joint level are also available, enabling us to render the metadata of agent to design embodied tasks.

### B. Pretraining Data

We build our datasets for pretraining our instance segmentation and referring expression comprehension modules. The agent samples more than 10000 images from all scenes in AI2-THOR with the BFS strategy. Annotations including semantic labels, positions in the scene and ground-truth segmentation masks are automatically generated from the metadata of the simulator.

We build the corresponding semantic scene graph incrementally from metadata during the sampling for referring expressions used for REC module pretraining. As shown in Fig. 2, scene frames and metadata of objects seen in the frame are sampled during navigation. By updating frames during sampling, we add new objects as nodes and spatial relationships between objects as edges into scene graph. Two metrics assign the spatial relationship between two objects: distance between central points $l$ and 3D IoU(intersection over union ) of ground-truth bounding boxes $S_{IoU}$. Then, we build our referring expression set with node-edge-node pairs in scene graph in form of template.$RE$ : *the* $\{$*OBJ1*$\}$ $\{$*RELATION*$\}$ *the* $\{$*OBJ2*$\}$ , where $\{$*OBJ1*$\}$,$\{$*OBJ2*$\}$ represent objects and $\{$*RELATION*$\}$ represents spatial relationship between them.

### C. Summary of REMQA dataset

We select 60 types of most frequently seen objects out of objects in AI2-THOR to build our REMQA dataset. Similar to the construction of dataset for REC module pretraining, we generate our questions using node-edge-node pairs from the scene graph. We designed three kinds of questions in our REMQA dataset: COUNTING, EXISTENCE, and SPATIAL questions. The templates of 3 kinds of questions are shown in Table. II where ***RE****(OBJ2)* represents the referring expression consists of object *OBJ2*

TABLE II: Question Templates

| | |
|---|---|
| EXISTENCE | *Is there a* $\{$*OBJ1*$\}$ *in the* $\{$***RE****(OBJ2)*$\}$*?* |
| COUNTING | *How many* $\{$*OBJ1*$\}$ *are there in the* $\{$***RE****(OBJ2)*$\}$*?* |
| SPATIAL | *Is there a* $\{$*OBJ1*$\}$ $\{$*RELATION*$\}$ *the* $\{$*OBJ2*$\}$*?* |

As the statistics of our proposed REMQA dataset shown in following Table. III , REMQA dataset is composed of 120 scenes. Among the dataset, we use 100 scene series for training and 20 for testing. There are altogether 4072 questions of three kinds.

TABLE III: REMQA dataset split

| | Scenes | EXISTENCE | COUNTING | SPATIAL |
|---|---|---|---|---|
| train | 100 | 1653 | 647 | 1083 |
| test | 20 | 324 | 143 | 222 |
| all | 120 | 1977 | 790 | 1305 |
| avg. Length | - | 9.5 | 11.8 | 7.1 |

Similar to off-shelf EQA datasets [23], [27], each scene is associated with multiple scene configurations that result in different answers to the same question. For every task sample in the dataset, a question with a referring phrase, ground-truth answer, scene configuration, the initial scene of the target object and final scene after manipulation are included.

## IV. Proposed Approach

As shown in Fig. 3,we have proposed a general framework to for REMQA task which consists of three main parts: 1) Navigation module 2) Referring expression comprehension module 3) Manipulation Question Answering module.

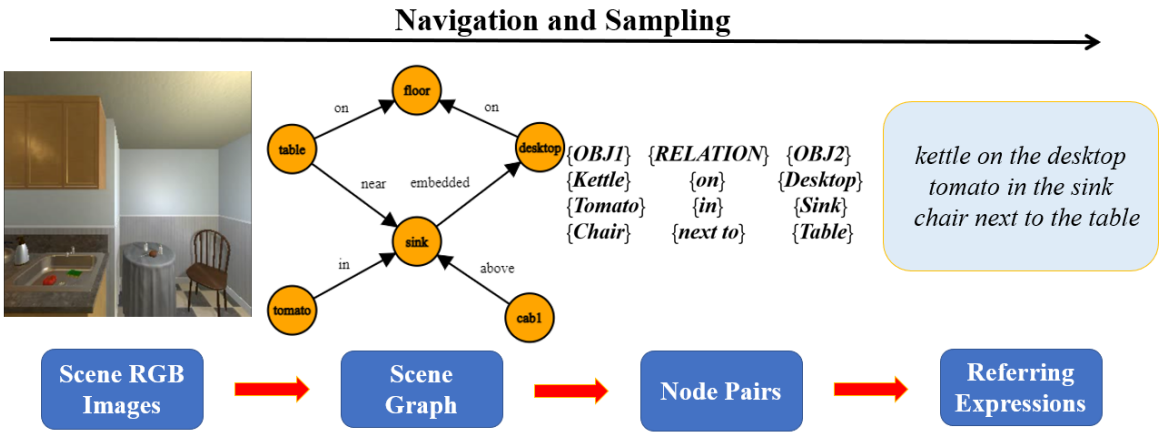



Fig. 2: An overview of proposed referring expression data generation pipeline

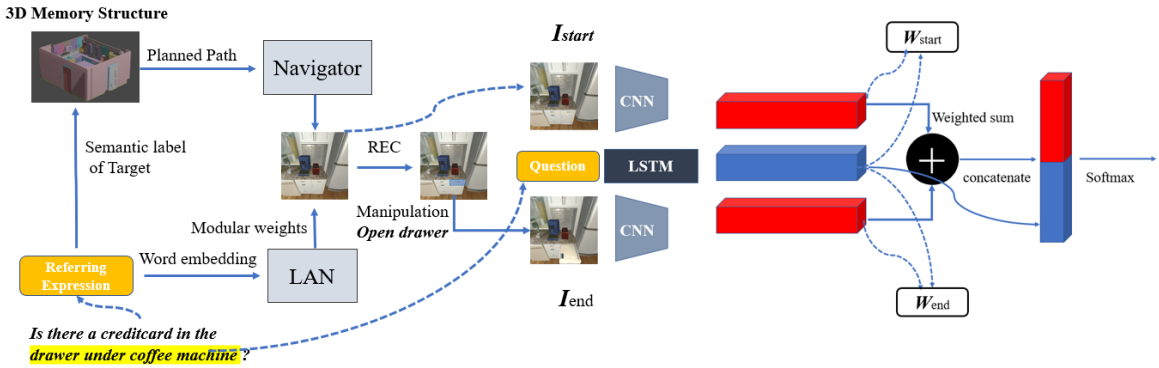



Fig. 3: The overall architecture of proposed framework: The referring expression in input question passes through LAN for REC. Then the scene frames before and after manipulations $\boldsymbol{I}_{start}, \boldsymbol{I}_{end}$ along with embedded question are extracted by CNN-LSTM network. Then extracted features are further fused in QA part and classified by a softmax classifier.

### A. Navigation

To enhance the performance of the navigation module with information of task scenarios, we build a knowledge-prior visual semantic navigation model based on a scene graph and semantic map of a given indoor environment. The construction of the scene graph is mentioned in Section III-B. To construct the semantic map, the agent first navigates in AI2THOR scenes with pre-designed sampling paths to incrementally build the 3D semantic memory structure of every room layout from metadata. The illustration of constructed structure is presented in Fig.4a. Every voxel and its color in the structure represents a tiny cubic space occupied by an object and the corresponding object type. Meanwhile, the 3D memory with semantic labels is dynamically transformed into a 2D semantic map by dimensional reduction during the sampling. As shown in Fig.4b, the agent samples the RGB frames and metadata of environment when navigating along the given path in each room of AI2THOR. The semantic map of the task scene is updated at every time step.

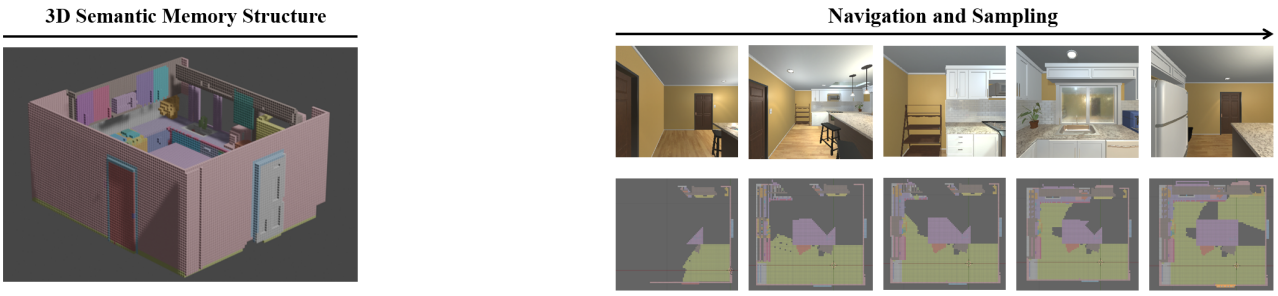
Employing generated 2D semantic maps as prior knowledge, the agent first locates the target object's position by searching its semantic label in the map. Then the shortest path between the initial agent position and the target object is planned by the Floyd algorithm in the semantic map for the navigation task.

### B. Referring Expression Comprehension

We build out referring expression comprehension module by adopting Language Attention Network (LAN) a modular design from MAttNet [5]. The LAN decomposes referring expressions in the form of word embedding into three modular components: subject attributes, location and spatial relationship to other objects. For every module, a phrase embedding is provided to calculate the matching scores of corresponding area in the given image without affecting each other. The overall matching scores weighted by modular weight are calculated to match objects with expressions.

### C. Manipulation Question Answering

We implement a Manipulation Question Answering (QA) module based on LSTM network. The upstream REC module




(a) (b)

Fig. 4: (a). The illustration of 3D semantic memory structure consist of semantic voxels (b). The RGB frames of the agent and incremental semantic map during sampling. The room is a kitchen scenario provided in AI2THOR simulator

pass the semantic label of target object which is usually an occlusion (e.g. Fridge, Cabinet) to a classifier. The classifier will choose the manipulation type from the action set: $\mathcal{A} = \{Open, Move, Pickup\}$ according to the type of target object (e.g. Open Fridge, Move chair,Pick up book). Then the agent moves its arm until the arm end reaches the vicinity of the target object by rendering the position from 3D semantic memory. Due to the incompleteness of the dynamic simulation of Manipulathor, the manipulation will be automatically conducted to objects within end sphere. Hence, dynamic planning is not included in our proposed manipulation module.

We denote the initial RGB frame before manipulations as $I_{start}$ and the frame after manipulations as $I_{end}$. The answering module encodes $I_{start}, I_{end}$ frames with CNN and the input question with a 2-layer LSTM network. Then, attention weights based on image-question similarity are computed to fuse the features of two images and image features with an encoded question. A softmax classifier outputs predictions from the space of possible answers with the fused feature passed through.

TABLE IV: Comparison of VSN methods

| Methods | Max steps=25 | | Max steps=50 | |
|---|---|---|---|---|
| | Success | SPL | Success | SPL |
| Random | 0.013 | 0.006 | 0.035 | 0.013 |
| A3C | 0.210 | 0.129 | 0.241 | 0.152 |
| SAVN | 0.283 | 0.121 | 0.396 | 0.178 |
| Scene Priors | 0.264 | 0.117 | 0.376 | 0.164 |
| Ours | **0.566** | **0.461** | **0.729** | **0.595** |

TABLE V: Comparison of REC methods

| Methods | Pred@0.5 | |
|---|---|---|
| | Val | Test |
| MCN | 71.04 | 73.16 |
| CGAN | 73.18 | 76.94 |
| BiLSTM [6] + detectron2 | 66.57 | 70.45 |
| LAN+MaskRCNN | 70.27 | 74.12 |
| LAN+detectron2(Ours) | **75.20** | **77.52** |

## V. Experiments

In this section, we first introduce the experimental settings and evaluation metrics. Then, we present the analysis of the quantitative results of our proposed method and several baseline variants. Finally, an MQA task sample is selected to illustrate how the agent navigates, comprehends and answers the question as qualitative results.

### A. Embodied Referring Expression

**Experimental settings:** To comprehensively illustrate the performance of our model on embodied referring expression task, we analyze the ERE results in two stages: the performance on VSN and the performance on REC toward ground-truth final scenes assuming that agent navigates successfully in the first stage. Two kinds of metrics are raised for evaluation of VSN task: the success rate of navigation $s$ and Success rate weighted by Path Length (SPL) [20]. The SPL calculates the success rate of the VSN task weighted by the ratio of shortest path distance from the starting position to the goal by the path distance the agent actually takes. For REC performance, we adopt prec@X [11] measures the percentage of test images with an intersection over union (IoU) score higher than the threshold $X$ and here we set $X = 0.5$.

In the VSN stage, we compared our proposed semantic map priored navigation model with Random agent, traditional RL methods: A3C and other embodied navigation methods that use metadata of scene in AI2THOR as priors for navigation: SAVN [31], Scene Priors [32]. All baselines are evaluated on seen objects and known objects. During the evaluations, the agent can move 0.25m or turn 90 degrees at each step. When the number of steps reaches the max steps, the task is terminated and marked as failed. Results are summarized in following Table IV.Taking advantage of the 2D semantic map obtained from the 3D memory structural, our proposed navigation model outperforms all baseline methods by a large margin in terms of both success rate and SPL metrics. It is noted that the 2D semantic map only represents information of scenes seen before. Our proposed method can be improved in the generalization of novel scenes.

In the REC stage, a set of off-shelf models proposed for referring expression tasks are selected as baselines: MCN [11],

TABLE VI: Evaluations of QA models on REMQA dataset

| Methods | EXISTENCE | | | COUNTING | | | SPATIAL | | |
|---|---|---|---|---|---|---|---|---|---|
| | $S_N$ | $S_L$ | $S_{QA}$ | $S_N$ | $S_L$ | $S_{QA}$ | $S_N$ | $S_L$ | $S_{QA}$ |
| EQA | 0.509 | 0.389 | 0.330 | 0.455 | 0.350 | 0.217 | 0.468 | 0.396 | 0.131 |
| IQA | 0.642 | 0.596 | 0.488 | **0.769** | **0.692** | 0.510 | 0.653 | 0.550 | 0.216 |
| Ours | **0.738** | **0.620** | **0.540** | 0.671 | 0.608 | **0.559** | **0.752** | **0.595** | **0.284** |

**Question: Is there a cup in the cabinet above the coffee machine? Groudtruth Answer: Yes**

Fig. 5: A qualitative example with the question. The trajectory of agent is presented on the left side and some rendered images of the agent egocentric view and third party camera on the right

CGAN [33] and combinations of language parser and object detector with other backbones. The validation and test REC datasets are split from referring phrase set in Section III-B. Results are summarized in following Table V. Our proposed REC model outperforms all baseline models on both validation and test set. It is worth noting that there is no significant margin between our model and SoTA baseline models. However, the performance gap between two recombined baselines is much larger than SoTAs. We infer the reason that SoTAs and our model all adopt the modular module to process the referring expressions into multi-modal channels. The result can also validate the effectiveness of the modular module that BiLSTM + detectron2 baseline performs much lower than LAN+MaskRCNN since Bi-LSTM can only encode the referring expressions without modular extraction.

### *B. MQA on REMQA dataset*

To validate the effectiveness of our proposed model, we separately compare our navigation and comprehension results with the state-of-the-art methods (SoTAs) on our proposed benchmark dataset. The results are presented in Table. VI. The $S_N, S_L$ denotes the ratio of samples successfully navigated to the goal position and successfully localized at the goal position with REC. The $S_{QA}$ represents the rate of correctly answered questions out of all samples in the REMQA dataset.

The results show that our framework for the REMQA task outperforms EQA and IQA models in most metrics except $S_N$ and $S_L$ in COUNTING questions. Most of the objects mentioned in COUNTING questions are receptacles (e.g.cabinets, drawers) that are many individuals in the same scene. IQA model is trained on a larger dataset (IQUAD v1 with 75000 questions) and may better distinguish instances from the same type. We also can notice that EQA model performs far below IQA and our model due to the absence of the ability to interact with the environment. Therefore, the effectiveness of manipulations for the REMQA task can be validated.

### *C. Qualitative Results*

To illustrate how our agent navigates and manipulates in proposed Embodied Environments, we select a task sample from proposed REMQA dataset as a qualitative example. As shown in following Fig. 5, the agent actively explores in kitchen scenarios to find the coffee machine mentioned in the input question. When successfully navigated, agent localizes the target cabinet with REC and moves its arm to open it to find whether there is a cup for final answer.

## VI. Conclusion

In this work, we have proposed a brand new embodied robot task Remote Embodied Manipulation Question Answering (REMQA) in a physically interactive environment. We build a benchmark dataset for the task by scene graph generation. To solve this problem, we propose a general framework consisting of a VSN module with scene semantic map as priors, a LAN for referring expression comprehension and manipulation decision, and a CNN-LSTM network for question answering. The experimental results on the new benchmark dataset validate the effectiveness of our proposed model. This task still challenges to our proposed method in more complicated question and multi-stage task. For the future study, a validation experiment in real robot platforms and physical environment is expected to validate the ability of to conduct more complex actions.

## REFERENCES


[1] J. Duan, S. Yu, H. L. Tan, H. Zhu, and C. Tan, "A survey of embodied ai: From simulators to research tasks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2022.

[2] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "Referitgame: Referring to objects in photographs of natural scenes," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[3] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *European Conference on Computer Vision*. Springer, 2016, pp. 69–85.

[4] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *IEEE conference on computer vision and pattern recognition*, 2016.

[5] L. Yu, Z. Lin, X. Shen, J. Yang, X. Lu, M. Bansal, and T. L. Berg, "Mattnet: Modular attention network for referring expression comprehension," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[6] R. Hu, M. Rohrbach, J. Andreas, T. Darrell, and K. Saenko, "Modeling relationships in referential expressions with compositional modular networks," in *IEEE conference on computer vision and pattern recognition*, 2017.

[7] S. Yang, G. Li, and Y. Yu, "Cross-modal relationship inference for grounding referring expressions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[8] M. Li and L. Sigal, "Referring transformer: A one-step approach to multi-task visual grounding," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[9] R. Li, K. Li, M. Shu, X. Qi, X. Shen, and J. Jia, "Referring image segmentation via recurrent refinement networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[10] H. Shi, H. Li, F. Meng, and Q. Wu, "Key-word-aware network for referring expression image segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[11] G. Luo, Y. Zhou, X. Sun, L. Cao, C. Wu, C. Deng, and R. Ji, "Multi-task collaborative network for joint referring expression comprehension and segmentation," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[12] M. Shridhar, D. Mittal, and D. Hsu, "Ingress: Interactive visual grounding of referring expressions," *The International Journal of Robotics Research*, vol. 39, no. 2-3, 2020.

[13] Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. v. d. Hengel, "Reverie: Remote embodied visual referring expression in real indoor environments," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[14] H. Chen, A. Suhr, D. Misra, N. Snavely, and Y. Artzi, "Touchdown: Natural language navigation and spatial reasoning in visual street environments," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[15] X. Li, D. Guo, H. Liu, and F. Sun, "Reve-ce: Remote embodied visual referring expression in continuous environment," *IEEE Robotics and Automation Letters*, 2022.

[16] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox, "Alfred: A benchmark for interpreting grounded instructions for everyday tasks," in *Computer Vision and Pattern Recognition*, 2020.

[17] X. Lin, G. Li, and Y. Yu, "Scene-intuitive agent for remote embodied visual grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[18] P. Anderson, Q. Wu, D. Teney, J. Bruce, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[19] L. F. Posada, F. Hoffmann, and T. Bertram, "Visual semantic robot navigation in indoor environments," *VDE*, 2014.

[20] P. Anderson, A. Chang, D. S. Chaplot, S. Gupta, V. Koltun, J. Kosecka, J. Malik, R. Mottaghi, M. Savva *et al.*, "On evaluation of embodied navigation agents," *arXiv preprint arXiv:1807.06757*, 2018.

[21] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, "Visual dialog," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[22] T. M. Le, V. Le, S. Venkatesh, and T. Tran, "Hierarchical conditional relation networks for video question answering," in *IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[23] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied question answering," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[24] E. Wijmans, S. Datta, O. Maksymets, A. Das, G. Gkioxari, S. Lee, I. Essa, D. Parikh, and D. Batra, "Embodied question answering in photorealistic environments with point cloud perception," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[25] L. Yu, X. Chen, G. Gkioxari, M. Bansal, T. L. Berg, and D. Batra, "Multi-target embodied question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[26] S. Tan, W. Xiang, H. Liu, D. Guo, and F. Sun, "Multi-agent embodied question answering in interactive environments," in *Computer Vision - ECCV 2020 - 16th European Conference*. Springer, 2020.

[27] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi, "Iqa: Visual question answering in interactive environments," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[28] Y. Deng, D. Guo, X. Guo, N. Zhang, H. Liu, and F. Sun, "MQA: answering the question via robotic manipulation," in *Robotics: Science and Systems XVII, 2021*, D. A. Shell, M. Toussaint, and M. A. Hsieh, Eds., 2021.

[29] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi, "Ai2-thor: An interactive 3d environment for visual ai," *arXiv preprint arXiv:1712.05474*, 2017.

[30] K. Ehsani, W. Han, A. Herrasti, E. VanderBilt, L. Weihs, E. Kolve, A. Kembhavi, and R. Mottaghi, "Manipulathor: A framework for visual object manipulation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[31] M. Wortsman, K. Ehsani, M. Rastegari, A. Farhadi, and R. Mottaghi, "Learning to learn how to learn: Self-adaptive visual navigation using meta-learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[32] W. Yang, X. Wang, A. Farhadi, A. Gupta, and R. Mottaghi, "Visual semantic navigation using scene priors," *arXiv preprint arXiv:1810.06543*, 2018.

[33] G. Luo, Y. Zhou, R. Ji, X. Sun, J. Su, C.-W. Lin, and Q. Tian, "Cascade grouped attention network for referring expression segmentation," in *28th ACM International Conference on Multimedia*, 2020.