

Learning-To-Rank Approach for Identifying Everyday Objects Using a Physical-World Search Engine

Kanta Kaneda, Shunya Nagashima, Ryosuke Korekata, Motonari Kambara and Komei Sugiura

Abstract—Domestic service robots offer a solution to the increasing demand for daily care and support. A human-in-the-loop approach that combines automation and operator intervention is considered to be a realistic approach to their use in society. Therefore, we focus on the task of retrieving target objects from open-vocabulary user instructions in a human-in-the-loop setting, which we define as the learning-to-rank physical objects (LTRPO) task. For example, given the instruction “Please go to the dining room which has a round table. Pick up the bottle on it,” the model is required to output a ranked list of target objects that the operator/user can select. In this paper, we propose MultiRankIt, which is a novel approach for the LTRPO task. MultiRankIt introduces the Crossmodal Noun Phrase Encoder to model the relationship between phrases that contain referring expressions and the target bounding box, and the Crossmodal Region Feature Encoder to model the relationship between the target object and multiple images of its surrounding contextual environment. Additionally, we built a new dataset for the LTRPO task that consists of instructions with complex referring expressions accompanied by real indoor environmental images that feature various target objects. We validated our model on the dataset and it outperformed the baseline method in terms of the mean reciprocal rank and recall@k. Furthermore, we conducted physical experiments in a setting where a domestic service robot retrieved everyday objects in a standardized domestic environment, based on users’ instruction in a human-in-the-loop setting. The experimental results demonstrate that the success rate for object retrieval achieved 80%. Our code is available at this URL¹.

Index Terms—Deep Learning Methods, Learning Categories and Concepts, Deep Learning for Visual Perception

I. INTRODUCTION

THE aging of the population has led to an increased demand for daily care and support. The shortage of home care workers has become a pressing societal problem, and the use of domestic service robots (DSRs) to assist individuals with disabilities is considered to be a possible solution. A realistic approach to using DSRs for societal applications is a human-in-the-loop setting that combines automation and intervention by their operator. Therefore, in this study, we consider that a DSR that is neither fully automated nor fully

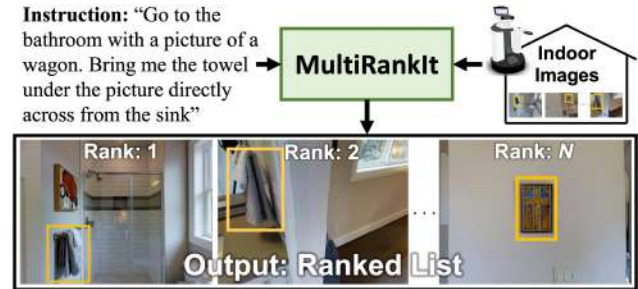


Fig. 1. Typical scene of the LTRPO task. The input is an instruction and output is the ranked list of target objects.

teleoperated, but rather serves as a cybernetic avatar [1] for the user. In this setting, a DSR that could provide relevant options to the operator would be highly beneficial.

In this study, we focus on the task of retrieving target objects from open-vocabulary user instructions, for example, manipulating or moving objects in the domestic environment, in a human-in-the-loop setting. We define this task as a learning-to-rank physical objects (LTRPO) task. Fig. 1 shows a typical scene for this task. Given the instruction “Go to the bathroom with a picture of a wagon. Bring me the towel under the picture directly across from the sink,” the model is required to output a ranked list of target objects. In the human-in-the-loop setting, it is important to display and rank a number of target objects that are easily selectable by the user/operator because human attention is a limited resource. Additionally, displaying a limited number of objects helps to minimize the cognitive load on the user/operator.

The LTRPO task is challenging because of its difficulty in accurately retrieving the specified target object from complex instructions. Specifically, the success rate for humans in a referring expression comprehension (REC) task with a simple referring expression is 90.76%, whereas it is only 48.98% for a state-of-the-art model [2].

The LTRPO task differs from REC tasks in that its objective is to retrieve images of the target object from complex instructions, often spanning more than two sentences and incorporating multiple referring expressions (see Fig.1). Note that the average expression length of a common REC task dataset is 8.43 words [3], while the average sentence length of the LTRPO dataset is 18.8 words, highlighting the significantly more complex nature of sentences in the LTRPO task. Additionally, while the REC task focuses on localizing the target object from a single image, the LTRPO task aims to

Manuscript received: August 4, 2023; Revised December 21, 2023; Accepted December 14, 2023. This paper was recommended for publication by Editor Cesar Cadena upon evaluation of the Associate Editor and Reviewers’ comments. This work was partially supported by JSPS KAKENHI Grant Number 20H04269, JST Moonshot, and NEDO.

Authors are with Keio University, 3-14-1 Hiyoshi, Kohoku, Yokohama, Kanagawa 223-8522, Japan. {k.kaneda, ng_sh, rkorekata, motonari.k714, komei.sugiura}@keio.jp

Digital Object Identifier (DOI): see top of this page.

¹<https://github.com/keio-smilab23/MultiRankIt>

identify the target object from hundreds of images within the environment. The human-in-the-loop setting changes the goal of the solution, allowing the user/operator to select the desired target object from the output ranked image list. Consequently, the methods proposed for the REC task are impractical for the LTRPO task.

Recent advancements in multimodal representation learning, such as CLIP [4], have resulted in the improved performance of crossmodal retrieval. However, as explained above, the performance of the crossmodal retrieval using texts that contain complex referring expressions remains insufficient. This is because existing methods often consider tasks that use only simple expressions (e.g., “A photo of {label}” [4]).

In this paper, we propose MultiRankIt, a method to retrieve target objects from open-vocabulary user instructions in a human-in-the-loop setting. In MultiRankIt, we introduce the Crossmodal Noun Phrase Encoder (CNPE), which handles noun phrases and prepositional phrases extracted from the instruction to model the relationship between the extracted phrases and the target bounding box. By introducing this module, it is expected that the model can retrieve the appropriate target object from an instruction with complex referring expressions. We also introduce the Crossmodal Region Feature Encoder (CRFE), which handles multiple images of the surrounding context. By handling context images that capture a broader range, rather than only those that are in close proximity to the target bounding box, it is expected that the model can effectively model the relationship between the target object and its surrounding contextual environment.

The difference between our method and existing methods is that our method considers the task as a learning-to-rank task [5] in which a DSR generates a ranked list of candidate target objects that are presented to the user. The proposed approach aims to strike a balance between automation and human intervention, where not everything is performed automatically and conversely, not everything is performed manually.

Our main contributions are summarized as follows:

- We propose MultiRankIt, which is a novel approach for identifying target objects from user instructions in a human-in-the-loop setting. While existing methods that address object localization as an image retrieval task (e.g., [6]) are limited to a closed-vocabulary setting, MultiRankIt handles open-vocabulary user instructions. In addition, to the best of our knowledge, this is the first attempt to propose this setting for mobile manipulation.
- We introduce the CNPE to model the relationship between phrases that contain referring expressions and target bounding boxes. While existing methods (e.g., [7]–[9]) are limited to handling simple noun phrases referring solely to the target object, the proposed method can accommodate various and hierarchical referring expressions within the instruction. This allows for the effective modeling of referring expressions related to both the target object and its surrounding environment.
- We introduce the CRFE to model the relationship between the target object and multiple images of its surrounding contextual environment. While existing methods (e.g., [10], [11]) are unable to consider relationships

when dealing with referring expressions related to objects beyond the image frame, the proposed method can adequately model the relationships with surrounding objects and landmarks across a broader viewpoint.

II. RELATED WORK

Numerous studies have been conducted in the field of vision and language [10]–[29]. In [12], the authors provide a comprehensive overview of various tasks, and the latest methods for vision and language tasks. In the following, we discuss related research in the field of vision and language, and focus on crossmodal retrieval, vision and language for robotics field and REC.

Crossmodal retrieval involves the retrieval of samples from one modality based on a query expressed in another modality. In [22], the authors present an image retrieval method that uses an input query in the form of an image and an accompanying text description of desired modifications. They introduce Text Image Residual Gating (TIRG), which combines the reference image and text query using a gating function and residual connection. DCNet [23] extends methods such as TIRG by introducing the Correction Network, which incorporates the difference between the target and reference image to learn robust multimodal representations.

Several standard datasets exist that handle the crossmodal retrieval task (e.g., [30], [31]). The CIRR dataset [31] consists of over 36,000 pairs of open-domain images paired with human-generated modifying text. It comprises real-life images that are not restricted to a specific domain, and places greater emphasis on distinguishing between visually similar images. The Fashion200k dataset [30] consists of over 200,000 images of fashion products. The dataset provides human-generated captions that distinguish similar pairs of fashion images, along with side information that consists of real-world product descriptions.

Additionally, studies related to vision and language for the robotics field include navigation [28], [29] and object manipulation [24]–[27]. LM-Nav [28] introduces a robotic navigation system that effectively utilizes ViNG [32], CLIP [4] and GPT-3 [33] to achieve successful long-horizon navigation in real outdoor environments based on natural language instructions. CLIPort [24] combines semantic understanding from CLIP with spatial precision from Transporter [34], allowing successful execution of tabletop tasks without explicit representations of object poses, memory, or syntactic structures. In [26], authors address the multimodal language understanding task for fetching instructions, in which the goal is to predict the target object as instructed by the user from an unconstrained sentence. HLMS-MAT [27] addresses the vision and language task in the execution of household tasks based on textual instructions using the Moment-based Adversarial Training algorithm, which uses two types of moments for perturbation updates.

The REC task shares some similarities with the LTRPO task. The REC task requires an agent to localize an object in an image given a natural language expression. RefTeacher [35] employs a semi-supervised teacher-student learning paradigm with Attention-based Imitation Learning and Adaptive Pseudo-label Weighting, achieving near fully supervised results using

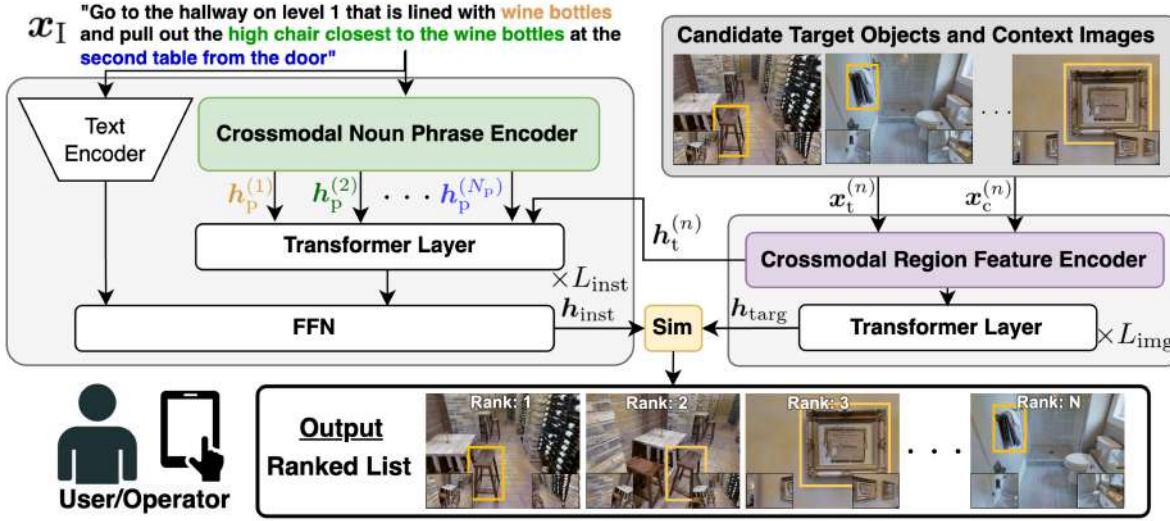


Fig. 2. Proposed method framework. MultiRankIt consists of Crossmodal Noun Phrase Encoder and Crossmodal Region Feature Encoder. “FFN” and “Sim” denote the feedforward network and cosine similarity, respectively.

only 10% labeled data. Target-Dependent UNITER [11] uses the transformer attention mechanism based on the UNITER [36] framework to model the relationship between objects and instructions for multimodal understanding for fetching instructions. However, as explained in Section I, methods proposed for the REC task are not directly applicable to the LTRPO task.

As mentioned above, our method differs from existing methods (e.g., [22], [23]) in that we consider the task as a learning-to-rank task [5] in which a DSR generates a ranked list of candidate target objects that are presented to the user/operator. In addition, while existing methods typically involve a robot operating in an environment where it has no prior information about its surroundings, the LTRPO task is specifically designed for situations where the robot already has access to images of the environment. This particular setup is highly practical for domestic service robots that are designed to operate within the same environment over extended periods. Our method is also different from [10] in that it retrieves the target object from multiple images showing the domestic environment, whereas in [10], the method identifies the target object from a single image with a fixed viewpoint. Additionally, our method introduces a module to model the relationship between the target object and multiple images of its surrounding contextual environment. Furthermore, our method differs from existing methods that handle noun phrases (e.g., [7]–[9]) by introducing a module to model the relationship between phrases that contain referring expressions and target bounding box.

III. PROBLEM STATEMENT

In this paper, we define the LTRPO task as the task of retrieving images of the target object taken by DSRs according to open-vocabulary user instructions. The terminology used in this paper is defined as follows:

- **Instruction x_I :** an open-vocabulary instruction to DSRs to perform household tasks.
- **Target object:** the object that is the target of the instruction.

- **Target bounding box:** the bounding box of the target object. Note that there are N_{targ} bounding boxes in an environment to be ranked by relevance. ($T = \{x_t^{(n)} | n = 1, 2, \dots, N_{\text{targ}}\}$)
- **Context images:** images of the target object and its surroundings. For each $x_t^{(n)}$, there exists a corresponding $x_c^{(n)}$, resulting in a total of N_{targ} surrounding images in the environment. Note that duplication of images might be included in this set. ($C = \{x_c^{(n)} | n = 1, 2, \dots, N_{\text{targ}}\}$)

The input and output of this task are elements $x = (x_I, T, C)$ and ranked list of T , respectively. Note that the details for the input and output are explained in Section IV. We assume that the coordinates of the target objects in the images are given.

IV. PROPOSED METHOD

In this paper, we propose MultiRankIt, which is a novel approach for identifying target objects from open-vocabulary user instructions, for example, manipulating or moving objects in the domestic environment, in a human-in-the-loop setting. The novelty of this study, that is, the introduction of a learning-to-rank approach to multimodal object retrieval in a human-in-the-loop setting, could be applied to other tasks, such as multimodal image retrieval (e.g., [37]).

Fig. 2 shows the structure of our method. The proposed method consists of two main modules: CNPE and CRFE. The details of the modules are explained below.

Input x to our method is defined as follows:

$$\begin{aligned} x &= (x_I, T, C), \\ T &= \{x_t^{(n)} | n = 1, 2, \dots, N_{\text{targ}}\}, \\ C &= \{x_c^{(n)} | n = 1, 2, \dots, N_{\text{targ}}\}, \end{aligned}$$

where $x_I \in \mathbb{R}^{V \times L}$, $x_t^{(n)} \in \mathbb{R}^{3 \times W \times H}$ and $x_c^{(n)} \in \mathbb{R}^{N_c \times 3 \times 256 \times 256}$ denote an instruction tokenized as a one-hot vector with vocabulary size V and maximum token length L , the bounding box of the target object of width W and height

H , and an image obtained by concatenating N_c surrounding images of $\mathbf{x}_t^{(n)}$, respectively.

First, we construct a query $(\mathbf{x}_I, \mathbf{x}_t^{(n)}, \mathbf{x}_c^{(n)})$ to compute the similarity between \mathbf{x}_I and $(\mathbf{x}_t^{(n)}, \mathbf{x}_c^{(n)})$. As mentioned in Section I, many of the instructions in the LTRPO task contain complex referring expressions related to, for example, target objects, landmarks, and rooms. Therefore, we introduce the CNPE, which handles noun phrases and prepositional phrases extracted from \mathbf{x}_I to model the relationship between the extracted phrases and \mathbf{x}_t .

The CNPE first extracts N_p phrases from \mathbf{x}_I through the following steps: First, it extracts noun phrases and prepositional phrases using the Stanford Parser [38]. Then, we obtain N_p phrases $\mathbf{x}_p^{(k)}$ ($k = 1, 2, \dots, N_p$) by grouping adjacent noun phrases and prepositional phrases. We then obtain text features $\mathbf{h}_I \in \mathbb{R}^{768}$ and $\mathbf{h}_p^{(k)} \in \mathbb{R}^{768}$ from \mathbf{x}_I and $\mathbf{x}_p^{(k)}$, respectively, using a pretrained CLIP text encoder [4].

Finally, we obtain the embedding for instruction \mathbf{h}_{inst} as follows:

$$\mathbf{h}_{\text{inst}} = f_{\text{FFN}} \left(\left[f_{\text{trm}} \left(\left[\mathbf{h}_p^{(1)}; \dots; \mathbf{h}_p^{(N_p)} \right] \right); \mathbf{h}_I; \mathbf{h}_t^{(n)} \right] \right)$$

where f_{FFN} , f_{trm} and $\mathbf{h}_t^{(n)}$ denote a feedforward network, L_{inst} transformer layers, and visual features extracted from $\mathbf{x}_t^{(n)}$ (details are explained later), respectively. Note that each transformer layer consists of multi-head self attention layers [39] and feedforward networks (FFNs), and a FFN consists of a linear layer, normalization layer and activation function.

The CRFE handles multiple images of the surrounding context. By handling context images that capture a broader range rather than just those that are in close proximity to the target bounding box, the proposed method can effectively model the relationship between the target object and its surrounding contextual environment.

First, we obtain visual features $\mathbf{h}_t^{(n)} \in \mathbb{R}^{768}$ and $\mathbf{h}_c^{(n)} \in \mathbb{R}^{N_c \times 768}$ from $\mathbf{x}_t^{(n)}$ and $\mathbf{x}_c^{(n)}$, respectively, using the pretrained CLIP image encoder (ViT/L-14). Then we obtain the embedding for image features \mathbf{h}_{targ} using L_{img} transformer layers. Finally, we define the similarity between \mathbf{x}_I and $(\mathbf{x}_t^{(n)}, \mathbf{x}_c^{(n)})$ using cosine similarity as follows:

$$s(\mathbf{x}_I, \mathbf{x}_t^{(n)}, \mathbf{x}_c^{(n)}) = \frac{\mathbf{h}_{\text{inst}} \cdot \mathbf{h}_{\text{targ}}}{\|\mathbf{h}_{\text{inst}}\| \|\mathbf{h}_{\text{targ}}\|}.$$

The output of our model is the ranked list of T based on $s(\mathbf{x}_I, \mathbf{x}_t^{(n)}, \mathbf{x}_c^{(n)})$.

The loss for each batch is defined as follows:

$$L_{\mathcal{B}} = -\frac{1}{|\mathcal{B}|} \sum_{\mathbf{x}_t^{(n)} \in \mathcal{B}} \log \frac{\exp(s(\mathbf{x}_I, \mathbf{x}_t^{(n)}, \mathbf{x}_c^{(n)}))}{\sum_{j=1}^{|\mathcal{B}|} \exp(s(\mathbf{x}_I, \mathbf{x}_t^{(j)}, \mathbf{x}_c^{(j)}))},$$

where $|\mathcal{B}|$ denotes the batch size.

V. DATASET AND EXPERIMENTAL SETUP

A. Dataset

To the best of our knowledge, no standard dataset exists for the LTRPO task. For example, the REVERIE dataset [40], which is the standard dataset for object-goal navigation tasks, contains the important elements for the object localization task in that it deals with real indoor environmental images

and object manipulation instructions. However, the REVERIE dataset alone is unsuitable for the LTRTO task. Therefore, we built the Learning to Rank in Real Indoor Environments (LTRRIE) dataset, which is a new dataset for LTRPO tasks.

We built the LTRRIE dataset by the following steps: First, we collected instructions from the REVERIE dataset and panoramic images from the Matterport3D Simulator [41]. Then, we obtained the target bounding boxes by cropping the panoramic images using the coordinates of the target objects provided in the REVERIE dataset. Note that we excluded samples with target bounding boxes at the edges of the former image because they may not have included the entire target object when cropped.

The LTRRIE dataset consists of 58 environments, 5501 instructions, and 4352 target bounding boxes. Note that the LTRRIE dataset consists of both complex (e.g., Fig. 1) and simple instructions (e.g., ‘‘Bring me the white plant,’’ ‘‘Pick up the brown towel’’). It has a vocabulary size of 53118 words, a total of 103118 words, and an average sentence length of 18.78 words. The LTRRIE dataset includes 4210, 397 and 501 samples in the training, validation and test sets, respectively. Each set contains 50, 4 and 4 environments, respectively, with no duplication of environments. Therefore, the samples included in the validation set can be regarded as unseen objects. Compared to existing datasets (e.g., PFN-PIC [10]), the LTRRIE dataset is a comprehensive and diverse dataset that closely aligns with the real-world scenarios typically encountered by domestic service robots. Note that we obtained the instructions for the training set from the REVERIE train set, and obtained the instructions for the validation and test sets by dividing the val_unseen set of the REVERIE dataset. We used the training set to update the parameters of our model and the validation set to tune the hyperparameters. We evaluated our model on the test set.

B. Experimental Setup

We adopted the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) for training with learning rate of 2.0×10^{-5} and batch size of 128. In the transformer layer, we set $L_{\text{inst}} = 4$, $L_{\text{img}} = 4$, $\#A = 4$, $\#H = 768$, where $\#H$ and $\#A$ denote the hidden size and number of attention heads in the f_{inst} and f_{targ} transformer layers, respectively.

Our model had 47M trainable parameters and 24.49G multiply-add operations. The proposed model was trained on an RTX 2080Ti with 11GB of GPU memory and an Intel Core i9 processor. It took approximately an hour to train our model. The inference time was approximately 20 ms/sample. We evaluated our model on the validation sets every epoch. The final performance on the test set was achieved by the model with maximum recall@10 for the validation set.

VI. EXPERIMENTAL RESULTS

A. Results

To the best of our knowledge, limited studies exist in which the LTRPO task has been addressed. Thus, we developed a baseline method that extends CLIP [4] to handle the LTRPO task. The baseline method takes \mathbf{x}_I , \mathbf{x}_t , and $\mathbf{x}_{c'}$ as input. Here, $\mathbf{x}_{c'}$ is obtained by concatenating N_c surrounding images of

TABLE I
QUANTITATIVE COMPARISON FOR THE VALIDATION SET. THE BEST SCORES ARE IN BOLD.

Model	Condition		Validation Set				
	w/ CNPE	w/ $x_c^{(n)}$	MRR	R@1[%]	R@5[%]	R@10[%]	R@20[%]
(a) Baseline		✓	0.433 ± 0.010	13.7 ± 0.6	47.1 ± 0.9	66.6 ± 1.4	82.0 ± 1.2
(b)	✓		0.365 ± 0.022	11.5 ± 1.0	38.0 ± 3.2	54.9 ± 2.9	71.7 ± 1.4
(c) Proposed		✓	0.472 ± 0.007	16.8 ± 0.4	53.9 ± 1.2	73.2 ± 0.1	85.1 ± 0.7
(d)	✓	✓	0.481 ± 0.014	17.6 ± 1.4	50.8 ± 1.3	68.1 ± 1.3	83.0 ± 1.3

TABLE II
QUANTITATIVE COMPARISON FOR THE TEST SET. THE BEST SCORES ARE IN BOLD.

Model	Condition		Test Set				
	w/ CNPE	w/ $x_c^{(n)}$	MRR	R@1[%]	R@5[%]	R@10[%]	R@20[%]
(a) Baseline		✓	0.415 ± 0.009	14.0 ± 1.0	45.3 ± 1.7	63.8 ± 2.5	80.8 ± 2.0
(b)	✓		0.373 ± 0.015	12.1 ± 0.5	39.6 ± 1.4	56.1 ± 1.1	70.2 ± 0.7
(c) Proposed		✓	0.426 ± 0.004	14.6 ± 0.4	45.3 ± 0.5	66.1 ± 1.7	80.6 ± 0.9
(d)	✓	✓	0.501 ± 0.008	18.3 ± 1.0	52.2 ± 1.4	69.8 ± 1.5	83.8 ± 0.6

$x_t^{(n)}$ horizontally and resizing the result to 256×256 pixels. x_t and $x_{c'}$ are input into the CLIP image encoder, and a simple MLP structure is used to obtain image features from these embeddings. Additionally, x_I is input into the CLIP text encoder to obtain language features. Then we used these features to calculate the similarity using the same method explained in Section IV. The output is a ranked list of T , obtained based on this similarity.

Tables I and II show the quantitative results of the baseline method, proposed method and ablation studies for the validation and test set, respectively. The values in the tables are the average and standard deviation over five trials. We explain the details of the ablation studies in Section VI-B. We evaluated the model using the MRR and recall@k ($k = 1, 5, 10, 20$). We used the MRR as the primary metric. MRR is defined as follows:

$$\text{MRR} = \frac{1}{N_{\text{inst}}} \sum_{j=1}^{N_{\text{inst}}} \frac{1}{r_1^{(j)}},$$

where N_{inst} and $r_1^{(j)}$ denote the number of instructions and the rank of the first relevant sample in the list of retrieved samples, respectively. Recall@k is defined as follows:

$$\text{Recall@k} = \frac{1}{N_{\text{inst}}} \sum_{j=1}^{N_{\text{inst}}} \frac{|A_j \cap B_j|}{|A_j|},$$

where A_j and B_j denote the set of relevant samples to be searched and the set of the top- k retrieved samples, respectively. We used the MRR and recall@k because they are standard metrics in learning-to-rank settings [5].

Table II shows that the MRR for the baseline method (a) and the proposed method (d) were 0.415 and 0.501, respectively. Therefore, the proposed method (d) outperformed the baseline method (a) on the test set by 0.086 points in terms of the MRR. Similarly, the proposed method also outperformed the baseline method in terms of recall@k ($k \leq 20$). For all the above results except for recall@20, there were statistically significant differences in performance ($p < 0.05$) between our method and the baseline method.

Fig. 3 shows the qualitative results. Fig. 3 (a) presents the top-3 samples in the output ranked list for x_I “Go to the

bathroom with a picture of a wagon and bring me the towel directly across from the sink.” For this sample, the MRR was 1 and 0.1 for the proposed method and the baseline method, respectively. It is worth noting that the “sink”, which was mentioned in x_I , was only presented in edge of $x_c^{(n)}$. Therefore, this result suggests that the introduction of the CRFE, which handles context images that capture a broader range, was beneficial to performance.

Fig. 3(b) shows the top-3 samples in the output ranked list for x_I “Go to the hallway on level 1 that is lined with wine bottles and pull out the high chair closest to the wine bottles at the second table from the door.” For this sample, MRR was 1 and 0.091 for the proposed method and baseline method, respectively. The proposed model successfully retrieved the appropriate x_t as top-1 result for the complex referring expression: “the high chair closest to the wine bottles at the second table from the door.” Therefore, this result suggests that the introduction of the CNPE module to model the relationship between noun phrases and target objects was beneficial to performance.

Fig. 4 illustrates a sample for which the model failed to retrieve the desired image as specified in x_I , “Proceed to the hallway on level 2 with the basketball and level painting above the open book.” The instruction required the model to retrieve an image of the “painting above the open book,” but the model was unsuccessful in retrieving images of paintings that were not positioned near an open book. We hypothesize that this failure was because of the lack of a “book” in $x_c^{(n)}$, which made it challenging for the model to accurately search using a referring expression.

B. Ablation Study

We set the following two conditions for the ablation studies to demonstrate the effectiveness of our proposed method. Tables I and II present the quantitative results for the ablation studies.

$x_c^{(n)}$ ablation: We removed $x_c^{(n)}$ to investigate the performance when only $x = (x_I, T)$ was used as input. Table II shows that the MRR decreased by 0.128 points under condition (b) on the test set. This result indicates that introducing



Fig. 3. Qualitative Results. The topmost images show the ground truth target image and the others show the top-3 ranked images. The yellow-bordered area and the green dotted-line area represent \mathbf{x}_t and contextual objects other than \mathbf{x}_t , respectively.



Fig. 4. A failed sample. The leftmost image shows the ground truth target image and the others show the top-3 ranked images. The model was unsuccessful in retrieving images of painting that were positioned near an open book.

TABLE III
RESULTS FOR ERROR ANALYSIS.

Error	#Samples
Referring Expression Comprehension Error	6
Target Noun Phrase Selection Error	5
Object Grounding Error	3
Annotation Error	2
Ambiguous Instruction	2
Others	2

CRFE to handle $x_c^{(n)}$ was most beneficial to performance. We believe that the introduction of CRFE was effective due to its capability to handle contextual images that capture a broader range, enabling the transformer layer subsequent to CRFE to effectively model the relationship between the target object and its surrounding environment.

CNPE ablation: We removed the CNPE to investigate the influence on performance. Table II shows that the MRR decreased by 0.075 points under condition (c) on the test set. This results indicates that the introduction of the CNPE, to handle noun phrases extracted from complex instructions, was also beneficial to performance. We believe that the introduction of CNPE proved effective because it enabled the model to effectively model the relationship between the image of the target object and the phrases associated with the object within the environment.

C. Error Analysis

We analyzed samples with an MRR less than 0.1, which we defined as failed cases. We identified 171 failed cases, and conducted a detailed analysis of the top-10 retrieved results for 20 of these failed cases. Table III categorizes the failed cases. We classified the causes of failure into the following six groups: (i) REC Error (RECE): This refers to the case in which the top retrieved results did not comply with the referring expression. (ii) Target Noun Phrase Selection Error: This refers to the case in which the model did not retrieve the target object but retrieved one of the landmark objects referred to in the sentence. (iii) Object Grounding Error: This refers to the case in which the model has achieved inadequate performance for object grounding. (iv) Annotation Error: This refers to the case in which the annotation for the target bounding box or the instruction was incorrect. (v) Ambiguous Instruction: This refers to the case in which the x_I was ambiguous and the top search results could not be definitively deemed incorrect. (vi) Others: This refers to errors that we could not classify under the previously mentioned categories. Table III shows that the main bottleneck was RECE. As a possible solution, we could effectively address this error by introducing a module that handles map information.

VII. PHYSICAL EXPERIMENTS

We used the Human Support Robot (HSR) [42], developed by the Toyota Motor Corporation (shown in Fig. 5(a)). The physical experiments involved YCB objects [43] (Fig. 5(b)). Fig. 5(a) shows the environment used in the physical experiments, which was based on the standardized environment of the World Robot Summit 2020 Partner Robot Challenge/Real Space [44], an international robotics competition that evaluated

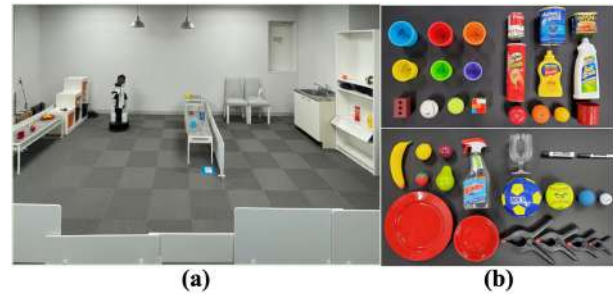


Fig. 5. (a) DSR platform in the experiment environment and (b) YCB objects used in the physical experiments.

tidying tasks in domestic settings. The environment's dimensions measured $6.0 \times 4.0\text{m}^2$, comprising six specified types of furniture pieces arranged as depicted in Fig. 5(a).

The procedure for the physical experiments are as follows: Firstly, we generated 5 distinct object placement patterns, ensuring all objects were placed on furniture pieces. Subsequently, we provided the DSR with a total of 50 English instructions. The DSR's behavior was designed as follows: It began by collecting environment images while navigating through waypoints from the initial position, employing standard path planning and navigation methods based on a pre-provided map. Each waypoint allowed the DSR to face each furniture piece and capture the target object from multiple viewpoints using an Asus Xtion Pro camera mounted on its head. We performed zero-shot transfer using our model trained on the LTRRIE dataset. Note that this indicates that the results of the physical experiments can be viewed as demonstrating the generalization capability of the proposed model. For grasping actions, the grasping point was determined based on the depth image and the bounding box of the target object. Using the intrinsic camera parameters, a point cloud was obtained from the bounding box of the depth image and transformed into the camera coordinate system. The median position in each coordinate axis was then identified as the grasping point. It took approximately 4 seconds to execute the process in real-time, including preprocessing, data transfer and inference, with our extended version.

Table IV shows the quantitative results of the physical experiments. We used MRR, recall@10 and task success rate as the evaluation metrics. To enable the user/operator to easily select the target object in standard UI setting, we only considered ranks within the top-10 for calculating the MRR. Table IV shows that the MRR and recall@10 were 0.45 and 88, respectively. Also, the task success rate for object retrieval and manipulation was 80%.

Fig. 6 shows qualitative results of the physical experiment, where the red rectangles represent candidate object regions. The top-3 samples in the ranked list of outputs are displayed, along with surrounding images for each candidate object area. In Fig. 6(a), the target object, a spam can, was appropriately retrieved as the top-1 result, and the user selected the appropriate object region. The robot successfully grasped the spam can and delivered it to the user. Similarly, in Fig. 6(b), the target object, a yellow cup, was appropriately retrieved as the top-1 result, and the yellow cup selected by the user was successfully delivered.

TABLE IV
 TASK SUCCESS RATES IN THE PHYSICAL EXPERIMENTS.

MRR	Recall@10 [%]	SR [%]
0.45	88	80 (40 / 50)



Fig. 6. Qualitative results of the physical experiments. The leftmost image shows the ground truth target image and the others show the top-2 ranked images. For (a) and (b), the DSR successfully retrieved images of target objects.

VIII. CONCLUSIONS

In this paper we focused on the LTRPO task, which is a task to retrieve images of a target object taken by DSRs according to given instructions in a human-in-the-loop setting. Our method outperformed the baseline method for all metrics on the LTRRIE dataset. Also in physical experiments, our method achieved task success rate of 80%, which indicates that language comprehension and action execution can be integrated in a physical robot. In future work, we plan to extend our model by introducing a new module to handle map information.

REFERENCES

[1] H. Ishiguro, “The Realisation of an Avatar-Symbiotic Society where Everyone can Perform Active Roles without Constraint,” *Advanced Robotics*, vol. 35, no. 11, pp. 650–656, 2021.

[2] L. Yu *et al.*, “MattNet: Modular Attention Network for Referring Expression Comprehension,” in *CVPR*, 2018, pp. 1307–1315.

[3] J. Mao, J. Huang, *et al.*, “Generation and Comprehension of Unambiguous Object Descriptions,” in *CVPR*, 2016, pp. 11–20.

[4] A. Radford, J. Kim, *et al.*, “Learning Transferable Visual Models from Natural Language Supervision,” in *ICML*, 2021, pp. 8748–8763.

[5] T. Liu, “Learning to Rank for Information Retrieval,” *Foundations and Trends in Information Retrieval*, vol. 3, no. 3, pp. 225–331, 2009.

[6] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell, “Natural Language Object Retrieval,” in *CVPR*, 2016, pp. 4555–4564.

[7] S. Subramanian, W. Merrill, T. Darrell, M. Gardner, S. Singh, and A. Rohrbach, “ReCLIP: A Strong Zero-shot Baseline for Referring Expression Comprehension,” in *ACL*, 2022, pp. 5198–5215.

[8] A. Baldrati, D. Morelli, G. Cartella, M. Cornia, M. Bertini, R. Cucchiara, *et al.*, “Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing,” in *ICCV*, 2023.

[9] G. Cartella, A. Baldrati, D. Morelli, M. Cornia, *et al.*, “OpenFashionCLIP: Vision-and-Language Contrastive Learning with Open-Source Fashion Data,” in *ICIAP*, 2023, pp. 245–256.

[10] J. Hatori, Y. Kikuchi, S. Kobayashi, K. Takahashi, Y. Tsuboi, *et al.*, “Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions,” in *ICRA*, 2018, pp. 3774–3781.

[11] S. Ishikawa and K. Sugiura, “Target-Dependent UNITER: A Transformer-based Multimodal Language Comprehension Model for Domestic Service Robots,” *IEEE RAL*, vol. 6, no. 4, pp. 8401–8408, 2021.

[12] S. Uppal, S. Bhagat, D. Hazarika, N. Majumder, *et al.*, “Multimodal Research in Vision and Language: A Review of Current and Emerging Trends,” *Information Fusion*, vol. 77, pp. 149–171, 2022.

[13] Z. Yu, J. Yu, Y. Cui, *et al.*, “Deep Modular Co-attention Networks for Visual Question Answering,” in *CVPR*, 2019, pp. 6281–6290.

[14] L. Li *et al.*, “Relation-Aware Graph Attention Network for Visual Question Answering,” in *ICCV*, 2019, pp. 10313–10322.

[15] Z. Yang, N. Garcia, *et al.*, “BERT Representations for Video Question Answering,” in *WACV*, 2020, pp. 1556–1565.

[16] Y. Luo, J. Ji, *et al.*, “Dual-Level Collaborative Transformer for Image Captioning,” in *AAAI*, vol. 35, no. 3, 2021, pp. 2286–2293.

[17] M. Cornia *et al.*, “Meshed-Memory Transformer for Image Captioning,” in *CVPR*, 2020, pp. 10578–10587.

[18] P. Anderson, A. Shrivastava, *et al.*, “Sim-to-Real Transfer for Vision-and-Language Navigation,” in *CoRL*, 2021, pp. 671–681.

[19] H. Wang, W. Liang, J. Shen, *et al.*, “Counterfactual Cycle-Consistent Learning for Instruction Following and Generation in Vision-Language Navigation,” in *CVPR*, 2022, pp. 15471–15481.

[20] J. Chen *et al.*, “Reinforced Structured State-Evolution for Vision-Language Navigation,” in *CVPR*, 2022, pp. 15450–15459.

[21] T. Weyand, A. Araujo, B. Cao, and J. Sim, “Google Landmarks Dataset v2-A Large-Scale Benchmark for Instance-Level Recognition and Retrieval,” in *CVPR*, 2020, pp. 2575–2584.

[22] N. Vo *et al.*, “Composing Text and Image for Image Retrieval-An Empirical Odyssey,” in *CVPR*, 2019, pp. 6439–6448.

[23] J. Kim, Y. Yu, *et al.*, “Dual Compositional Learning in Interactive Image Retrieval,” in *AAAI*, vol. 35, no. 2, 2021, pp. 1771–1779.

[24] M. Shridhar, L. Manuelli, and D. Fox, “CliPort: What and Where Pathways for Robotic Manipulation,” in *CoRL*, 2022, pp. 894–906.

[25] Y. Zhou, S. Sonawani, M. Phielipp, *et al.*, “Modularity through Attention: Efficient Training and Transfer of Language-Conditioned Policies for Robot Manipulation,” in *CoRL*, 2023, pp. 1684–1695.

[26] A. Magassouba *et al.*, “Understanding Natural Language Instructions for Fetching Daily Objects Using GAN-based Multimodal Target-Source Classification,” *IEEE RAL*, vol. 4, no. 4, pp. 3884–3891, 2019.

[27] S. Ishikawa and K. Sugiura, “Moment-based Adversarial Training for Embodied Language Comprehension,” in *IEEE ICPR*, 2022, pp. 4139–4145.

[28] D. Shah, B. Osiński, S. Levine, *et al.*, “LM-Nav: Robotic Navigation with Large Pre-trained Models of Language, Vision, and Action,” in *CoRL*, 2023, pp. 492–504.

[29] C. Huang, O. Mees, A. Zeng, and W. Burgard, “Visual Language Maps for Robot Navigation,” in *ICRA*, 2023, pp. 10608–10615.

[30] X. Han, Z. Wu, P. Huang, *et al.*, “Automatic Spatially-Aware Fashion Concept Discovery,” in *ICCV*, 2017, pp. 1463–1471.

[31] Z. Liu *et al.*, “Image Retrieval on Real-Life Images with Pre-trained Vision-and-Language Models,” in *CVPR*, 2021, pp. 2125–2134.

[32] D. Shah, B. Eysenbach, G. Kahn, N. Rhinehart, and S. Levine, “ViNG: Learning Open-World Navigation with Visual Goals,” in *ICRA*, 2021, pp. 13215–13222.

[33] T. Brown, B. Mann, N. Ryder, M. Subbiah, *et al.*, “Language Models are Few-shot Learners,” in *NeurIPS*, vol. 33, 2020, pp. 1877–1901.

[34] A. Zeng *et al.*, “Transporter Networks: Rearranging the Visual World for Robotic Manipulation,” in *CoRL*, 2021, pp. 726–747.

[35] J. Sun, G. Luo, Y. Zhou, X. Sun, G. Jiang, Z. Wang, and R. Ji, “RefTeacher: A Strong Baseline for Semi-Supervised Referring Expression Comprehension,” in *CVPR*, 2023, pp. 19144–19154.

[36] Y. Chen, L. Li, *et al.*, “UNITER: UNiversal Image-TExt Representation Learning,” in *ECCV*, 2020, pp. 104–120.

[37] H. Wu, Y. Gao, X. Guo, Z. Ai, S. Rennie, K. Grauman, and R. Feris, “Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback,” in *CVPR*, 2021, pp. 11307–11317.

[38] S. Schuster and C. Manning, “Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks,” in *LREC*, 2016, pp. 2371–2378.

[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, *et al.*, “Attention is All You Need,” in *NIPS*, vol. 30, 2017, pp. 5998–6008.

[40] Y. Qi, Q. Wu, P. Anderson, X. Wang, Y. Wang, C. Shen, and A. Hengel, “REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments,” in *CVPR*, 2020, pp. 9982–9991.

[41] A. Chang *et al.*, “Matterport3D: Learning from RGB-D Data in Indoor Environments,” in *IEEE 3DV*, 2018, pp. 667–676.

[42] T. Yamamoto, K. Terada, A. Ochiai, *et al.*, “Development of Human Support Robot as the Research Platform of a Domestic Mobile Manipulator,” *ROBOMECH Journal*, vol. 6, no. 1, pp. 1–15, 2019.

[43] B. Calli, A. Walsman, *et al.*, “Benchmarking in Manipulation Research: Using the Yale-CMU-Berkeley Object and Model Set,” *IEEE RAM*, vol. 22, no. 3, pp. 36–52, 2015.

[44] “World Robot Summit 2020 Partner Robot Challenge Real Space Rules & Regulations,” 2020.