IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Open-Vocabulary Mobile Manipulation Based on Double Relaxed Contrastive Learning with Dense Labeling

Daichi Yashima, Ryosuke Korekata, Komei Sugiura

***Abstract*—Growing labor shortages are increasing the demand for domestic service robots (DSRs) to assist in various settings. In this study, we develop a DSR that transports everyday objects to specified pieces of furniture based on open-vocabulary instructions. Our approach focuses on retrieving images of target objects and receptacles from pre-collected images of indoor environments. For example, given an instruction "Please get the right red towel hanging on the metal towel rack and put it in the white washing machine on the left," the DSR is expected to carry the red towel to the washing machine based on the retrieved images. This is challenging because the correct images should be retrieved from thousands of collected images, which may include many images of similar towels and appliances. To address this, we propose RelaX-Former, which learns diverse and robust representations from among positive, unlabeled positive, and negative samples. We evaluated RelaX-Former on a dataset containing real-world indoor images and human annotated instructions including complex referring expressions. The experimental results demonstrate that RelaX-Former outperformed existing baseline models across standard image retrieval metrics. Moreover, we performed physical experiments using a DSR to evaluate the performance of our approach in a zero-shot transfer setting. The experiments involved the DSR to carry objects to specific receptacles based on open-vocabulary instructions, achieving an overall success rate of 75%.**

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

I. Introduction

SERVICE robots capable of transporting objects and working alongside humans are becoming increasingly important in various situations such as restaurants, hospitals, and warehouses, especially with rising labor shortages and insufficient workforces in society. To enhance their functionality, these robots should incorporate natural language understanding capabilities. This capability is particularly valuable for domestic service robots (DSRs), which can assist elderly people by performing daily tasks. However, DSRs face significant challenges in identifying the target object or the receptacle from numerous similar objects based on open-vocabulary instructions that include complex referring expressions.

In this study, we focus on developing a DSR system that transports everyday objects to specific pieces of furniture.

Manuscript received: August 16, 2024; Revised November 12, 2024; Accepted December 14, 2024. This paper was recommended for publication by Editor Aleksandra Faust upon evaluation of the Associate Editor and Reviewers' comments. This work was partially supported by JSPS KAKENHI Grant Number 23K03478, JST Moonshot, and NEDO.

Author is with Keio University, 3-14-1 Hiyoshi, Kohoku, Yokohama, Kanagawa 223-8522, Japan. {ydaichi1207, rkorekata, komei.sugiura}@keio.jp

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

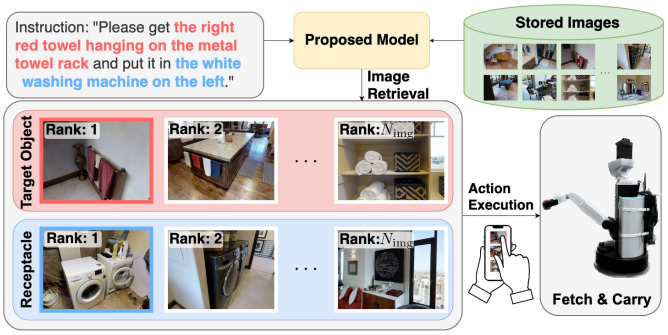

Fig. 1. Overview of our task. First, the DSR collects images of the environment through pre-exploration. Given an open-vocabulary instruction, it is required to retrieve the red and blue framed images as the target object image and receptacle image, respectively, from the collected images. Subsequently, the DSR carries the target object to the receptacle, based on the user-selected images.

Open-vocabulary instructions are used to retrieve images of target objects and receptacles from a collection of environmental images. Fig. 1 shows a typical scene for this task. First, the DSR collects images of indoor environments during pre-exploration. Next, given an instruction such as "Please get the right red towel hanging on the metal towel rack and put it in the white washing machine on the left," the proposed method outputs two ranked lists: one for the red towel hanging on the metal towel rack and the other for the white washing machine on the left. The DSR is then tasked with transporting the target object to the receptacle, based on the images selected by users from the two ranked lists.

The primary challenge in this task is to retrieve images from open-vocabulary instructions, which often include complex referring expressions that involve both target objects and receptacles. Moreover, it is challenging to retrieve the ground-truth image when the DSR has collected many similar images in the environment. While recent multimodal representation models such as CLIP [1] have improved cross-modal retrieval performance, they still have limitations for this task, as discussed in Section V-C. Most models use contrastive loss functions such as InfoNCE [2], where a single phrase is paired with a single positive image, and all other pairs within the batch are considered negative (e.g., [1], [3], [4]). This approach can be disadvantageous when the batch contains unlabeled positive samples. These unlabeled positives are typically treated as negatives, potentially resulting in undesired anti-correlation between samples that should be positively correlated. The lack of positive annotations often occurs due to the labor-intensive

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

and time-consuming nature of providing detailed labels for all similar instances[1].

In this paper, we propose RelaX-Former, a novel method that leverages unlabeled positive labels and introduces a double relaxed contrastive learning approach to handle unlabeled positive and negative samples, thereby improving the alignment between images and text. Most conventional methods follow the standard approach of using contrastive loss functions such as InfoNCE, often ignoring the presence of unlabeled positives [1], [5]. In contrast, the proposed method incorporates unlabeled positives by assigning unlabeled positive labels to images similar to the ground-truth image through the Dense Labeler. The Double Relaxed Contrastive (DRC) loss is then used to handle the relations among positive, unlabeled positive, and negative samples. By distinguishing between unlabeled positive and negative samples with the Dense Labeler and applying the proposed DRC loss, our approach enables a more diverse and effective learning representation. Our code is available at this URL[2].

The main contributions of this study are as follows:

- We propose the Spatial Overlay Grounding (SOG) module, which employs markers on segmented images to extract features from a multimodal large language model (MLLM). By generating dense captions for multiple image regions, it provides fine-grained, linguistically grounded features to enrich visual representations.
- We present the Dense Representation Learning (DRL) module, which employs the Dense Labeler to estimate unlabeled positives to images similar to the positive image and uses the DRC loss to effectively handle the relationship among positive, unlabeled positive, and negative samples.

II. Related Work

A. Language-Guided Mobile Manipulation

There has been widespread research in mobile manipulation tasks guided by natural language instructions [6]–[8]. For instance, [9]–[11] have conducted language-guided mobile manipulation tasks using DSRs within standardized real-world environments, demonstrating the application of natural language instructions in practical scenarios. While these tasks share similar objectives to ours, the task considered in the present study diverges by employing free-form, open-vocabulary instructions accompanied by referring expressions, as opposed to the use of template-based instructions. Although [11] utilizes open-vocabulary instructions, it differs from our study in that it does not incorporate fully free-form instructions.

Various studies have attempted to identify target objects from among similar objects in the environment based on manipulation instructions [3], [12], [13]. Specifically, [3] and [13] involve multimodal image retrieval in indoor environments, requiring the output of a ranked list of multiple images that contain the target object described in the instruction. Our task differs from these in that it handles both the target object and the receptacle within the instruction in the image retrieval setting. NLMap [7] employs a CLIP-based approach with open-vocabulary instructions and pre-explored images for target object and receptacle image retrieval. While it has a fully automated image retrieval process, our system enhances this by allowing user selection from the top-ranked retrieved images, improving the adaptability to complex scenarios.

B. Contrastive Learning

There have been numerous studies in the field of contrastive learning and self-supervised representation learning [14]. Contrastive learning maps similar (positive) samples close together and dissimilar (negative) samples farther apart in the embedding space. Representative studies address data augmentation (e.g., [15]–[17]), clustering (e.g., [18]), and loss function designs (e.g., [16], [19]).

InfoNCE [2] is widely used as a contrastive loss function (e.g., [1], [4], [20], [21]) but its primary limitation is the imbalance between positive and negative samples during training. This restricts the diversity and adaptability of the representations by consistently distancing all unpaired samples from the query [14]. To address this limitation, softened loss functions such as [5], [16], [22] have been proposed as more balanced alternatives to InfoNCE. ReCo [5] relaxes the strict contrast applied to negative samples within the embedding space. Although this alleviates the asymmetric relationship between positive and negative embeddings, it does not explicitly consider unlabeled positives. We address this issue by introducing the novel DRC loss, which can handle unlabeled positives and relaxes the spaces of unlabeled positive and negative embeddings, resulting in more refined representations.

III. Problem Statement

In this study, we focus on the Image Retrieval-based Open-Vocabulary Fetch-and-Carry (IROV-FC) task [4]. Fig. 1 shows a typical scene associated with this task. The task procedure is defined as follows:

1) The DSR collects images of an indoor environment during pre-exploration.
2) Image retrieval: given an instruction such as "Please get the right red towel hanging on the metal towel rack and put it in the white washing machine on the left," two ranked lists of images should be retrieved: one for the target object and the other for the receptacle. Images of the target object and the receptacle should be ranked highly in these lists.
3) The correct target object and receptacle are selected by the user from among the top-K retrieved images.
4) Action execution: the DSR moves to the location at which the target image was collected, grasps the target object, carries it to the receptacle, and places it there.

The terminology used in this paper is defined as follows:

- **Target object:** an everyday object identified as the target in the instruction.
- **Receptacle:** a piece of furniture identified as the designated placement area in the instruction.

[1]Comprehensively annotating all possible unlabeled positive samples in our training set of 5,814 would require approximately 188,000 h of human labor, assuming 10 s per annotation and checking each instruction against all images.

[2]https://github.com/keio-smilab24/RelaX-Former

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

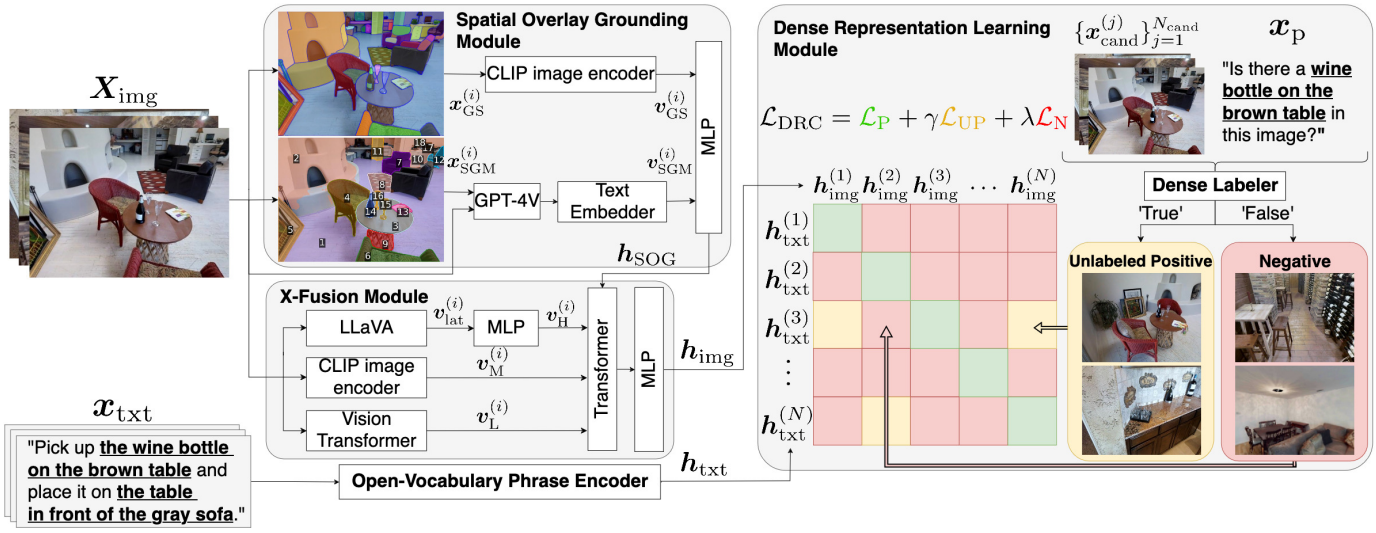

Fig. 2. Architecture of RelaX-Former. The proposed architecture consists of four modules: Spatial Overlay Grounding (SOG) module, X-Fusion (XF) module, Dense Representation Learning (DRL) module, and Open-Vocabulary Phrase (OVP) encoder. Here, N denotes the batch size.

We assume that the images of the indoor environment have already been collected through pre-exploration. This is reasonable because DSRs are typically used in the same indoor environment for long periods of time [7], [13]. It is also assumed that trajectory generation regarding navigation, object grasping, and object placement is based on heuristic methods.

IV. Proposed Method

Fig. 2 shows the structure of our proposed method, RelaX-Former. It consists of four modules: Spatial Overlay Grounding (SOG) module, X-Fusion (XF) module, Dense Representation Learning (DRL) module, and Open-Vocabulary Phrase (OVP) encoder.

The input $\boldsymbol{x}$ to RelaX-Former is defined as follows:

$$\boldsymbol{x} = \{X_{\text{img}}, \boldsymbol{x}_{\text{txt}}, m\},\ X_{\text{img}} = \{\boldsymbol{x}_{\text{img}}^{(i)} |\ i = 1, \cdots, N_{\text{img}}\},$$

where $\boldsymbol{x}_{\text{txt}} \in \{0,1\}^{V \times L}$, $m \in \{\langle\text{target}\rangle, \langle\text{receptacle}\rangle\}$ and $\boldsymbol{x}_{\text{img}} \in \mathbb{R}^{3\times W\times H}$ denote a tokenized instruction, an image, and a mode token that indicates the basis for the ranking, respectively. Here, V, L, i, N_{img}, W, and H denote the vocabulary size, maximum token length, index of each image in the set of collected images to be ranked, number of collected images to be ranked, image width, and image height, respectively.

A. Spatial Overlay Grounding Module

In the SOG module, visual features are obtained using two parallel streams. One stream applies a multimodal encoder with segmentation masks, while the other employs a MLLM with region-marked images, both leveraging mask information to enhance regional features. Typical methods obtain visual features either globally or by focusing on a single object or part of an object, which can lead to misinterpretations of the visual context, as shown in Section V-E. On the other hand, our approach uses foundation models for segmentation (e.g., [23], [24]) to isolate objects. This provides auxiliary information such as contours, shape, and the relative positions between objects, which helps reduce visual errors.

This module takes X_{img} as the input. We use SAM [23] and SEEM [24] to generate segmentation masks for $\boldsymbol{x}_{\text{img}}^{(i)}$. The masks generated by SAM and SEEM are overlaid on $\boldsymbol{x}_{\text{img}}^{(i)}$ to obtain $\boldsymbol{x}_{\text{GS}}^{(i)} \in \mathbb{R}^{3\times W\times H}$ and $\boldsymbol{x}_{\text{SS}}^{(i)} \in \mathbb{R}^{3\times W\times H}$, respectively. Next, visual features $\boldsymbol{v}_{\text{GS}}^{(i)} \in \mathbb{R}^{d_{\text{GS}}}$ are obtained using a pre-trained CLIP image encoder (ViT-L/14) [1] on $\boldsymbol{x}_{\text{GS}}^{(i)}$. Subsequently, as shown in Fig. 2, we obtain $\boldsymbol{x}_{\text{SGM}}^{(i)} \in \mathbb{R}^{3\times W\times H}$ by assigning numerical marks to each segmentation mask in $\boldsymbol{x}_{\text{SS}}^{(i)}$. This process consists of two steps: determining the mark location and assigning numerical labels. For mark location determination, we process masks from the smallest to largest area, determining suitable positions while excluding regions covered by previously processed masks. The final numerical labels are then assigned to these locations in descending order following [25].

An image description is then obtained using an MLLM based on three inputs: $\boldsymbol{x}_{\text{img}}^{(i)}$, $\boldsymbol{x}_{\text{SGM}}^{(i)}$, and a text prompt that instructs the description of the structured relationship of the room. The image description is embedded with text-embedding-large-3, and the resulting embedding is processed through a multi-layer perceptron (MLP) to obtain $\boldsymbol{v}_{\text{SGM}}^{(i)} \in \mathbb{R}^{d_{\text{SGM}}}$. To enhance the generation of image descriptions by MLLMs, we adopt a dual-input strategy that uses both $\boldsymbol{x}_{\text{img}}^{(i)}$ and $\boldsymbol{x}_{\text{SGM}}^{(i)}$ for inputs to GPT-4V. This addresses the issue of the marks overlapping and potentially hiding essential image details. While also preventing misinterpretation of the masked colors as those from the actual scene by providing the raw image for correct color recognition. Finally, the output $\boldsymbol{h}_{\text{SOG}} \in \mathbb{R}^{d_{\text{SOG}}}$ is obtained by concatenating $\boldsymbol{v}_{\text{GS}}^{(i)}$ and $\boldsymbol{v}_{\text{SGM}}^{(i)}$ and feeding it into a MLP.

B. X-Fusion Module

In the XF module, we comprehensively obtained three types of visual embeddings from pre-trained visual encoders(e.g.,

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

[26]), multimodal encoders(e.g., [1]), and MLLMs that provide latent features (e.g., [27]). Visual encoders capture edges, textures, and shapes, but lack semantic and spatial understanding. Multimodal encoders provide semantically aligned embeddings that connect visual and textual data, though they struggle with spatial detail. In contrast, MLLMs that provide latent features can obtain structural features that directly represent spatial expressions and complex referring relations through embeddings. Using these latent features are advantageous over text-based outputs because they eliminate the need to pass through a tokenizer, allowing for straightforward utilization of the features. Thus, we use these three types of embeddings in parallel to leverage their complementary strengths. Furthermore, these three specific feature types were chosen based on the results of preliminary experiments.

The XF module takes X_{img} as input. First, we derive the visual features $\boldsymbol{v}_{\text{L}}^{(i)} \in \mathbb{R}^{d_{\text{L}}}$ from a pre-trained visual encoder (ViT). Next, $\boldsymbol{v}_{\text{M}}^{(i)} \in \mathbb{R}^{d_{\text{M}}}$ is obtained by the pre-trained CLIP image encoder. The latent features $\boldsymbol{v}_{\text{lat}}^{(i)} \in \mathbb{R}^{d_{\text{lat}}}$ are acquired using an MLLM (LLaVA-v1.6-mistral-7b) from $\boldsymbol{x}_{\text{img}}^{(i)}$ and a prompt that instructs the MLLM to describe $\boldsymbol{x}_{\text{img}}^{(i)}$. We feed $\boldsymbol{v}_{\text{lat}}^{(i)}$ into a MLP to obtain $\boldsymbol{v}_{\text{H}}^{(i)} \in \mathbb{R}^{d_{\text{H}}}$. Finally, the comprehensive visual features $\boldsymbol{h}_{\text{img}} \in \mathbb{R}^{d_{\text{img}}}$ are obtained as follows:

$$\boldsymbol{h}_{\text{img}} = \text{MLP}\left(\text{Transformer}\left([\boldsymbol{v}_{\text{L}}; \boldsymbol{v}_{\text{M}}; \boldsymbol{v}_{\text{H}}; \boldsymbol{h}_{\text{SOG}}]\right)\right),$$

where Transformer$(\cdot)$ denote the transformer encoder.

C. Open-Vocabulary Phrase Encoder

The OVP encoder, following [4], efficiently processes open-vocabulary instructions for both target and receptacle modes within a single model framework. It takes $\boldsymbol{x}_{\text{txt}}$ and m as the input. An LLM identifies the phrases of either the target object or the receptacle based on the mode. It then generates standardized instructions, while a parser extracts noun phrases. Note that a standardized instruction in this case is a paraphrased instruction with a format for a typical fetch-and-carry task such as "Carry A to B."The encoder combines CLIP-processed text features from the original instruction, standardized instruction, and mode-specific phrases. These features, along with transformer-encoded outputs of the extracted phrases, are processed through a multi-layer perceptron to obtain the final text representation $\boldsymbol{h}_{\text{txt}} \in \mathbb{R}^{d_{\text{txt}}}$.

D. Dense Representation Learning Module

The DRL module leverages the novel DRC loss to optimize among positive, unlabeled positive, and negative samples through a relaxed contrastive approach. Recent multimodal pre-training methods (e.g., [1]) primarily use contrastive loss functions such as InfoNCE [2]. With InfoNCE, the model is optimized to maximize the similarity between the positive samples and minimize the similarity between the negative samples, targeting similarity scores closer to 1 and -1, respectively. However, in scenarios where the dataset contains similar images, this method results in those similar images being treated as negative samples when they should actually be considered positive or partially positive. The situation where only one annotation per ground truth is provided is due to various constraints, such as the labor-intensive and time-consuming costs associated with providing annotations as discussed in Section I. Therefore, the application of annotating unlabeled positives and a loss function that can appropriately handle these cases is crucial.

To address this issue, we propose the following DRC loss: $\mathcal{L}_{\text{DRC}} = \mathcal{L}_{\text{P}} + \gamma\mathcal{L}_{\text{UP}} + \lambda\mathcal{L}_{\text{N}}$, where γ and λ are hyperparameters that control the weights of the unlabeled positive sample loss and the negative sample loss, respectively. The components constituting $\mathcal{L}_{\text{DRC}}$ represent the losses for positive, unlabeled positive, and negative samples, respectively, which are defined as follows:

$$\mathcal{L}_{\text{P}} = \sum_{i} \left(1 - \text{sim}\left(\boldsymbol{h}_{\text{txt}}^{(i)}, \boldsymbol{h}_{\text{img}}^{(i)}\right)\right)^2,$$

$$\mathcal{L}_{\text{UP}} = \sum_{(i,j)\in\mathcal{S}} \max\left(\alpha - \text{sim}\left(\boldsymbol{h}_{\text{txt}}^{(i)}, \boldsymbol{h}_{\text{img}}^{(j)}\right), 0\right)^2,$$

$$\mathcal{L}_{\text{N}} = \sum_{(i,j)\notin\mathcal{S}} \max\left(\text{sim}\left(\boldsymbol{h}_{\text{txt}}^{(i)}, \boldsymbol{h}_{\text{img}}^{(j)}\right), 0\right)^2,$$

where $\text{sim}(\cdot,\cdot)$, $\mathcal{S}$, and α denote the similarity score, the set of indices corresponding to unlabeled positive samples, and a margin parameter that sets a threshold for the similarities of unlabeled positive samples, respectively. We use cosine similarity for $\text{sim}(\boldsymbol{h}_{\text{txt}}^{(i)}, \boldsymbol{h}_{\text{img}}^{(j)})$.

We introduce the Dense Labeler, which uses an MLLM (LLaVA-v1.6-mistral-7b) to label the unlabeled positives. The input is a text prompt and a set of candidate images. The text prompt describes whether the target object or receptacle can be seen in the image, depending on the mode. First, all scores $-1 \leq \text{sim}(\boldsymbol{h}_{\text{txt}}^{(i)}, \boldsymbol{h}_{\text{img}}^{(j)}) \leq 1$ are obtained using existing pre-trained models that have been successfully applied in image retrieval tasks (e.g., [1], [3]). After obtaining the score for $\boldsymbol{x}_{\text{txt}}^{(i)}$, the top N_{cand} images are selected as candidates and input into the MLLM. If the output text includes 'True', the image is classified as an unlabeled positive and the indices (i, j) are included in $\mathcal{S}$.

With the obtained $\mathcal{S}$, $\mathcal{L}_{\text{SP}}$ penalizes the model for samples $(i, j) \in \mathcal{S}$ where $\text{sim}(\boldsymbol{h}_{\text{txt}}^{(i)}, \boldsymbol{h}_{\text{img}}^{(j)})$ is less than α. The max function allows only those samples with a similarity score of less than α to contribute to the loss, thereby relaxing the contrastiveness by disregarding samples that have similarity scores greater than α. The parameters α and γ represent the leniency of the unlabeled positive labels and the reliance on the Dense Labeler, respectively. The same approach is applied to $\mathcal{L}_{\text{N}}$, where the model is penalized for negative samples $(i, j) \notin \mathcal{S}$ that have similarity scores higher than 0 by summing the squared similarities. By incorporating these components, the DRC loss balances the contribution of positive, unlabeled positive, and negative samples, leading to a more diverse and effective training process.

The output of RelaX-Former during inference is the ranked list of X_{img} arranged in descending order based on the similarity score $\text{sim}(\boldsymbol{h}_{\text{txt}}, \boldsymbol{h}_{\text{img}}^{(i)})$. Two ranked image lists, $\hat{Y}_{\text{targ}}$ and $\hat{Y}_{\text{rec}}$, for the target object and receptacle, respectively, are

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I
QUANTITATIVE COMPARISON BETWEEN RELAX-FORMER AND BASELINE METHODS ON THE TEST SETS. THE BEST SCORE FOR EACH METRIC IS PRESENTED IN **BOLD**. RECALL SCORES WERE CALCULATED INDIVIDUALLY FOR EACH TEST ENVIRONMENT AND THE AVERAGE OF THESE SCORES WAS REPORTED.

	Method	HM3D-FC (unseen)			MP3D-FC (unseen)		
		R@5↑ [%]	R@10↑ [%]	R@20↑ [%]	R@5↑ [%]	R@10↑ [%]	R@20↑ [%]
(i)	NLMap [7] (rep.)[3]	14.7	27.9	53.2	12.2	27.1	63.8
(ii)	MultiRankIt [3]	28.7 ± 3.4	48.3 ± 3.4	73.3 ± 2.6	35.7 ± 9.9	51.7 ± 8.9	72.7 ± 3.3
(iii)	DM^{2}RM [4]	47.8 ± 1.2	67.1 ± 2.4	87.0 ± 1.1	49.6 ± 0.7	64.1 ± 3.6	78.5 ± 0.5
(iv)	RelaX-Former (ours)	**55.4** ± 0.5	**76.3** ± 0.9	**91.6** ± 0.9	**57.0** ± 1.1	**72.4** ± 0.7	**82.5** ± 0.8

obtained through a total of two inferences by changing the mode.

V. EXPERIMENTS

A. Dataset

We used the LTRRIE-FC dataset [4] for the IROV-FC task. The LTRRIE-FC dataset was constructed from images collected from HM3D [28], [29] and MP3D [30], [31], featuring various everyday environments, with natural language instructions annotated by humans. Each instruction includes a target object and a receptacle, detailing the task of transporting the target object to the receptacle (e.g., "Pick up the green vase on the wash basin and put it on the counter-top table in the dining room.").

HM3D and MP3D are standard datasets in research involving everyday environments, such as navigation and scene understanding [13], [32]. To the best of our knowledge, there is no standard dataset that includes human-annotated instructions for fetch-and-carry tasks in both the HM3D and MP3D environments. Therefore, we used the LTRRIE-FC dataset. The dataset consists of 6,581 instructions and 7,148 images collected from 774 environments. The vocabulary size, total number of words, and average sentence length were 2,491, 103,263, and 15.69, respectively. In the LTRRIE-FC dataset, the training, validation, and test sets contained 5,814, 354, and 413 samples, respectively. These sets covered 690, 42, and 42 environments, respectively, with no duplication of environments. We used the train set to train the baseline and proposed models, the validation set to tune their hyperparameters, and the test sets to evaluate the models.

B. Parameter Settings

The AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.98$) was adopted for training with a learning rate of 1.0×10^{-4} and, batch size of $N = 128$. We used $\alpha = 0.7$ across all the experiments in this study. The number of candidate images selected for input into the MLLM, denoted as N_{cand}, was set to 20.

Our model had approximately 201M trainable parameters and a total of 329G multiply-add operations. We trained our model using a GeForce RTX4090 with 24 GB of GPU memory and an Intel Core i9-13900KF with 64 GB of RAM. The training process, which consisted of 20 epochs, took approximately 3 h. The inference time for computing the similarity between a single instruction and 100 images was approximately 79 ms. All visual features required for training and inference, including those from ViT, CLIP image encoder, LLaVA, GPT-4V, SAM, and SEEM, were obtained offline prior to training and the inference process, as these models remain frozen and do not need to be recomputed for each training or inference pass. The evaluation on the test sets was based on the model achieving the maximum value of recall@10 on the validation set.

[3]The results are based on our reproduction of NLMap, as the original code is not publicly available.

C. Quantitative Results

Table I presents the quantitative results for the performance of the baseline and proposed methods on the HM3D-FC and MP3D-FC test sets. The average and standard deviation values from five trials are included. We used NLMap [7], MultiRankIt [3], and DM^{2}RM [4] as the baseline methods. NLMap was chosen for its ability to retrieve images from a pre-explored set using a method based on CLIP [1], which is a representative method for zero-shot image retrieval tasks. Additionally, we selected MultiRankIt and DM^{2}RM following their successful application to tasks closely related to the IROV-FC task. Selecting models that do not use similarity scores for their outputs (e.g., [12]) would fail to effectively handle image retrieval based on instructions. The scores presented for NLMap were based on a single trial because a pre-trained frozen model provides consistent results across multiple trials. Furthermore, we trained MultiRankIt on target objects and receptacles separately and measured the average scores of the two trained models, as this approach cannot handle both modes within a single model. We used recall@K $(K = 5, 10, 20)$ as evaluation metrics, with recall@10 as the primary metric because it is a standard metric for image retrieval tasks [33]. Recall scores were calculated individually for each test environment, and the average of these scores was reported.

Table I indicates that the proposed method (iv) achieved the highest recall@10 scores of 76.3% on HM3D-FC and 72.4% on MP3D-FC. This represents improvements of 48.4 points and 45.3 points over (i), 28.0 points and 20.7 points over (ii), and 9.2 points and 8.3 points over (iii), respectively. Moreover, the proposed method (iv) outperformed the baseline methods across other recall metrics. The performance differences were statistically significant ($p < 0.01$). We believe this was because our proposed method handled unlabeled positives effectively.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

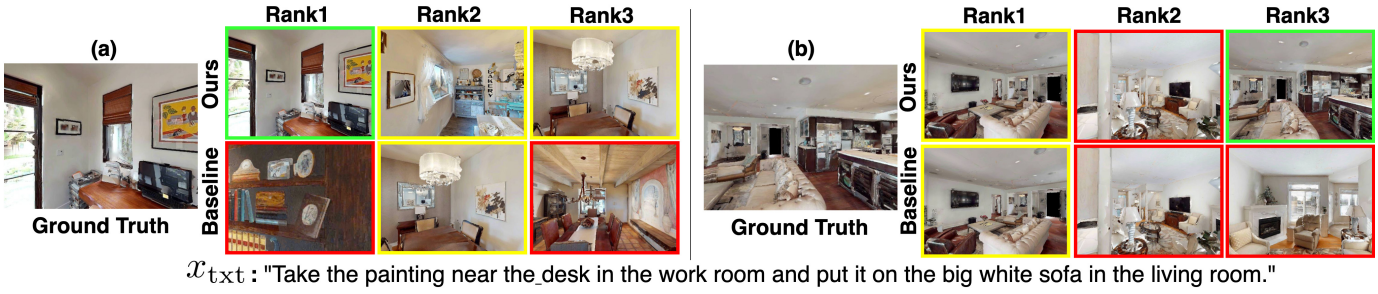

Fig. 3. Qualitative results of RelaX-Former and the most competitive baseline method [4] from the HM3D-FC test for $\boldsymbol{x}_{\text{txt}}$. The ground-truth image and the top-3 retrieved images are shown for each mode. (a) Target mode and (b) receptacle mode. Positive, unlabeled positive, and negative labels are colored in green, yellow, and red, respectively.

D. Qualitative Results & Discussions

Fig. 3 shows successful examples from the LTRRIE-FC dataset using the proposed method and compared them with the most competitive baseline method [4]. For each panel, a ground-truth image and the top-3 retrieved images are shown for each mode. Each retrieved image has a border color indicating its label: green for positive, yellow for unlabeled positive, and red for negative.

Figs. 3 (a) and (b) show an example from the HM3D-FC test set where $\boldsymbol{x}_{\text{txt}}$ was "Take the painting near the desk in the work room and put it on the big white sofa in the living room." In (a), the baseline method failed to retrieve the ground-truth image, while the proposed method successfully retrieved the ground-truth image as the top result, with the second and third ranks also including unlabeled positives. In (b), the baseline method did not include the ground-truth image in the top-3 retrieved images, instead ranking it 9th. Conversely, the proposed method retrieved the correct image in rank 3 and also retrieved an image with a similar receptacle at rank 1. This demonstrates that the DRC loss effectively improved the model's ability to retrieve positive and unlabeled positive images.

We further performed detailed error analysis to better understand the limitations. We defined samples with a rank less than 10 as failure samples. There were a total of 72 failure samples with 15 and 57 from HM3D-FC and MP3D-FC, respectively. We performed error analysis on 20 randomly sampled failed samples, 10 from each test set. Each sample was analyzed with the top-5 images for each mode. The main error for our model occurred in scenarios where similar images were among the top-ranked images, while the ground-truth image was not ranked within the top-20. Therefore, in future studies, we plan to develop a Dense Labeler that can assign accurate positive labels to similar images and incorporate a reranking model during inference.

E. Ablation Studies

Table II presents the results of our ablation studies. The ablation conditions were as follows:

SOG ablation: We conducted experiments by selectively removing streams within the SOG module to investigate their respective effect on performance. The results show a decrease in the recall@10 for models (b) and (c) compared with model (a) on the HM3D-FC and MP3D-FC test sets. These results imply that both streams within the SOG module substantially contribute to overall model performance, with the results indicating that segmenting images to isolate specific features enhances the image grounding capabilities of MLLMs and image encoders.

XF ablation: Similarly, while keeping the Transformer and MLP components, we removed the three image features (ViT, CLIP image encoder, and LLaVA) that were obtained and input into the Transformer within the XF module to assess its contribution. The results show a decrease in the recall@10 for models (d), (e), and (f) compared with model (a) on the HM3D-FC and MP3D-FC test sets. These results suggest that the presence of the XF module which integrates latent features from the MLLM alongside standard visual and multimodal encoders, enhances the model's ability to capture complex multimodal relationships.

Comparison with classic contrastive learning variants: We compared our proposed loss function with four classic contrastive learning approaches: using InfoNCE [2] (g), ReCo [5] (h), treating all unlabeled positive samples as positive (i), and setting α to a soft target value (j). In all cases, we observed decreased performance for all evaluation metrics on both the HM3D-FC and MP3D-FC test sets compared with model (a). These results demonstrate that our method's selective approach to managing unlabeled positives and the incorporation of the DRL module contribute significantly to improved retrieval performance.

Baseline comparisons with different loss functions: We conducted further experiments on using different loss functions across baseline methods. Specifically, we compared our proposed DRC loss with the InfoNCE loss across two baselines: MultiRankIt and DM^{2}RM. Note that we did not include NLMap in the experiment because it is a CLIP-based training free-method. Table III shows the results of our experiment. In both cases, we observed increased performnace in all evaluation metrics on the HM3D-FC and MP3D-FC test sets compared to the original baseline methods. These consistent improvements across both architectures demonstrate that our proposed DRC loss effectively addresses the limitations of InfoNCE by better handling unlabeled positives in the training data.

VI. Physical Experiments

We also validated the proposed method using a DSR in zero-shot, real-world experiments. The DSR executed fetch-

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE II
ABLATION STUDIES ON THE TEST SETS. THE HIGHEST VALUES FOR EACH METRIC ARE HIGHLIGHTED IN **BOLD**. RECALL SCORES WERE CALCULATED INDIVIDUALLY FOR EACH TEST ENVIRONMENT, AND THE AVERAGE OF THESE SCORES WAS REPORTED.

	Model	HM3D-FC (unseen)			MP3D-FC (unseen)		
		R@5↑ [%]	R@10↑ [%]	R@20↑ [%]	R@5↑ [%]	R@10↑ [%]	R@20↑ [%]
(a)	RelaX-Former (full)	**55.4** ± 0.5	**76.3** ± 0.9	91.6 ± 0.9	**57.0** ± 1.1	**72.4** ± 0.7	**82.5** ± 0.8
(b)	w/o $\boldsymbol{v}_{\mathrm{GS}}^{(i)}$	51.6 ± 1.2	73.5 ± 1.1	**91.8** ± 0.6	56.6 ± 2.3	70.8 ± 0.9	80.6 ± 0.6
(c)	w/o $\boldsymbol{v}_{\mathrm{SGM}}^{(i)}$	51.0 ± 1.7	73.8 ± 1.6	91.5 ± 1.0	55.8 ± 1.7	70.5 ± 0.9	81.0 ± 1.1
(d)	w/o $\boldsymbol{v}_{\mathrm{L}}^{(i)}$	54.2 ± 2.1	74.8 ± 0.7	89.9 ± 0.8	55.9 ± 1.7	71.1 ± 0.7	81.6 ± 1.2
(e)	w/o $\boldsymbol{v}_{\mathrm{M}}^{(i)}$	50.7 ± 1.6	72.3 ± 1.4	90.9 ± 0.4	54.3 ± 1.1	69.9 ± 1.3	82.0 ± 1.6
(f)	w/o $\boldsymbol{v}_{\mathrm{H}}^{(i)}$	51.2 ± 2.0	72.9 ± 3.0	90.2 ± 0.8	55.8 ± 1.4	70.6 ± 1.1	81.6 ± 1.3
(g)	w/ InfoNCE [2]	48.8 ± 0.9	70.9 ± 0.5	91.5 ± 0.5	54.8 ± 0.8	69.5 ± 0.8	81.8 ± 1.1
(h)	w/ ReCo [5]	52.5 ± 1.4	73.5 ± 1.1	91.4 ± 0.7	55.4 ± 0.7	69.1 ± 1.3	80.9 ± 1.3
(i)	unlabeled positives as positive	52.7 ± 1.6	73.7 ± 0.9	91.0 ± 1.0	55.7 ± 1.4	70.4 ± 0.9	82.3 ± 0.6
(j)	DRC, $\alpha = 0.9$	51.9 ± 0.1	73.6 ± 0.6	90.8 ± 0.8	55.2 ± 1.3	70.6 ± 0.2	81.1 ± 0.4

TABLE III
RECALL@10 SCORES ON THE TEST SETS ACROSS DIFFERENT MODELS AND LOSS FUNCTIONS. THE HIGHEST SCORES BETWEEN THE TWO LOSS FUNCTIONS FOR EACH METRIC ARE HIGHLIGHTED IN **BOLD**.

[%]	Model	Loss	HM3D-FC	MP3D-FC
(i)	MultiRankIt [3]	InfoNCE	48.3 ± 3.4	51.7 ± 8.9
(ii)		DRC	**57.6** ± 0.9	**58.3** ± 1.9
(iii)	DM^{2}RM [4]	InfoNCE	67.1 ± 2.4	64.1 ± 1.7
(iv)		DRC	**69.0** ± 2.1	**66.8** ± 1.1

and-carry tasks based on open-vocabulary instructions given by the users. These experiments involved objects that had not been seen during the training phase.

A. Settings

The experimental setup closely follows that of DM^{2}RM [4]. The test area measured 4.0×6.0 m^2 and contained nine pieces of furniture arranged in a specific layout. In the experiments, we used Toyota Motor Corporation's Human Support Robot, which has been the standard platform for RoboCup@Home since 2017 [9]. We used 30 everyday objects from the YCB object set [34]. The experiments consisted of 100 episodes across 20 different environmental setups, with five episodes per setup. In each environmental setup, 15 to 20 objects were randomly selected and placed at arbitrary locations on various pieces of furniture. There were approximately 400 different images throughout the physical experiment.

B. Implementation

The DSR collected RGB-D images of the environment using predetermined waypoints. The users then randomly selected one target object and one receptacle and provided a unique open-vocabulary fetch-and-carry instruction for each episode from the given environmental setup. The language instructions were provided by three human annotators from different countries with different cultural backgrounds. Instructions typically contained referring expressions and involved tasks such as picking up and carrying objects to specific pieces of furniture. After receiving instructions from the users, the DSR retrieved images from the pre-collected images using the proposed model in the zero-shot setting. The users then selected an appropriate image for the target object and the receptacle from the candidates displayed on a web interface. If there was no appropriate image of the target object among the top-K (=10) candidates, the task was considered a failure, and the DSR did not perform any further steps in the episode.

Subsequently, the DSR navigated to the location from which the selected image was collected and grasped the target object. The grasp was counted as successful only if the DSR succeeded in grasping the target object. Finally, the DSR carried the target object to the receptacle, completing the fetch-and-carry task. This final step was only attempted if the users had selected appropriate images for both the target object and the receptacle, and the DSR had successfully grasped the target object. The task was considered a success if the DSR successfully carried the target object and placed it on the receptacle. We did not utilize any learning-based methods in generating trajectories for object grasping and placing, as this aspect was beyond the scope of our research.

C. Results

Table IV presents the quantitative results of the physical experiments for the baseline methods [3], [4], [7] (as described in Section V-C) and the proposed methods. The evaluation metric for the physical experiments was the task success rate (SR), which is a standard evaluation metric for robotic manipulation tasks. SR is defined as SR $= N_{\mathrm{s}}/N_{\mathrm{a}}$, where N_{s} and N_{a} denote the numbers of successes and attempts, respectively. The grasping attempts were limited to instances in which the image of the target object was within the top-10 results retrieved in the target mode. Likewise, placing actions were only attempted when the following two conditions were met: the image of the receptacle ranked among the top-10 in receptacle mode, and the preceding grasping action had been successful. Table IV indicates that the proposed method

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE IV
QUANTITATIVE COMPARISON BETWEEN RELAX-FORMER AND BASELINE METHODS FOR PHYSICAL EXPERIMENTS. THE NUMBERS IN PARENTHESES SHOW THE SR (N_S/N_A). THE BEST SCORE IS IN **BOLD**.

	Method	SR↑ [%]
(i)	NLMap [7] (rep.)	64 (64 / 100)
(ii)	MultiRankIt [3]	53 (53 / 100)
(iii)	DM^{2}RM [4]	68 (68 / 100)
(iv)	RelaX-Former (ours)	**75** (75 / 100)

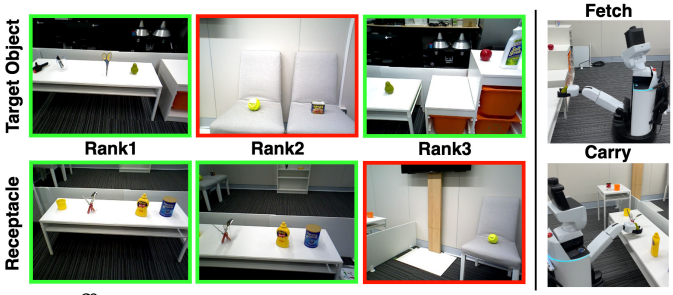

Fig. 4. Qualitative results of the physical experiments with a given instruction x_{txt}. The top-3 retrieved images for each mode are shown, alongside the scenes of fetching and carrying actions. The images that were considered correct or incorrect for each mode are framed in green or red, respectively.

(iv) achieved an overall SR of 75%, 7 points higher than the best baseline method (iii). The reason that method (i) was better than method (ii) in the real-world experiment, but worse in the simulated experiments, is because its training-free nature and dependence on CLIP limited its ability to handle the complex referring expressions frequently present in the instructions within the dataset.

Fig. 4 shows a successful example of the physical experiments. The x_{txt} was "Pick up the pear on the table and place it on the table next to the mustard." In target mode, the proposed method was able to retrieve the correct image in first place. For receptacle mode, the top-2 images retrieved by the proposed method were correct. These two images can both be considered correct because they contain the same scene captured from different angles. Subsequently, the DSR grasped the pear and placed it on the correct table.

VII. CONCLUSIONS

In this study, we focused on the IROV-FC task [4]. In this task, the DSR retrieved images of the target object and the receptacle based on an open-vocabulary instruction and subsequently transported the target object to the receptacle. We proposed RelaX-Former, a method that leveraged unlabeled positive samples and introduced a double relaxed contrastive learning approach to handle the relations among positive, unlabeled positive, and negative samples. RelaX-Former outperformed the baseline methods in terms of standard metrics on the LTRRIE-FC dataset [4]. Furthermore, in physical experiments using a DSR, RelaX-Former demonstrated robust performance in a zero-shot transfer setting, achieving an overall success rate of 75%.

REFERENCES

[1] A. Radford and W. Kim, "Learning Transferable Visual Models From Natural Language Supervision," in *ICML*, 2021, pp. 8748–8763.
[2] A. Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," *arXiv preprint arXiv:1807.03748*, 2018.
[3] K. Kaneda, S. Nagashima, R. Korekata, *et al.*, "Learning-To-Rank Approach for Identifying Everyday Objects Using a Physical-World Search Engine," *IEEE RA-L*, vol. 9, no. 3, pp. 2088–2095, 2024.
[4] R. Korekata, K. Kaneda, *et al.*, "DM2RM: Dual-Mode Multimodal Ranking for Target Objects and Receptacles Based on Open-Vocabulary Instructions," *arXiv preprint arXiv:2408.07910*, 2024.
[5] Z. Lin, E. Bas, Y. Singh, *et al.*, "Relaxing Contrastiveness in Multimodal Representation Learning," in *WACV*, 2023, pp. 2226–2235.
[6] B. Ichter, A. Brohan, *et al.*, "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances," in *CoRL*, 2022, pp. 287–318.
[7] B. Chen, F. Xia, *et al.*, "Open-vocabulary Queryable Scene Representations for Real World Planning," in *ICRA*, 2023, pp. 11 509–11 522.
[8] D. Driess, F. Xia, M. Sajjadi, C. Lynch, *et al.*, "PaLM-E: An Embodied Multimodal Language Model," in *ICML*, 2023, pp. 8469–8488.
[9] L. Iocchi, D. Holz, J. Solar, K. Sugiura, and T. Zant, "RoboCup@Home: Analysis and results of evolving competitions for domestic and service robots," *AIJ*, vol. 229, pp. 258–281, 2015.
[10] H. Okada, T. Inamura, and K. Wada, "What competitions were conducted in the service categories of the World Robot Summit?" *AR*, vol. 33, no. 17, pp. 900–910, 2019.
[11] S. Yenamandra, A. Ramachandran, K. Yadav, A. Wang, *et al.*, "HomeRobot: Open-Vocabulary Mobile Manipulation," in *CoRL*, 2023.
[12] R. Korekata, M. Kambara, Y. Yoshida, S. Ishikawa, *et al.*, "Switching Head-Tail Funnel UNITER for Dual Referring Expression Comprehension with Fetch-and-Carry Tasks," in *IROS*, 2023, pp. 3865–3872.
[13] G. Sigurdsson, J. Thomason, *et al.*, "RREx-BoT: Remote Referring Expressions with a Bag of Tricks," in *IROS*, 2023, pp. 5203–5210.
[14] P. Le-Khac *et al.*, "Contrastive Representation Learning: A Framework and Review," *IEEE Access*, vol. 8, pp. 193 907–193 934, 2020.
[15] T. Chen, S. Kornblith, *et al.*, "A Simple Framework for Contrastive Learning of Visual Representations," in *ICML*, 2020, pp. 1597–1607.
[16] C. Feng and I. Patras, "Adaptive Soft Contrastive Learning," in *ICPR*, 2022, pp. 2721–2727.
[17] J. Denize *et al.*, "Similarity Contrastive Estimation for Self-Supervised Soft Contrastive Learning," in *WACV*, 2023, pp. 2705–2715.
[18] M. Zheng, F. Wang, S. You, C. Qian, *et al.*, "Weakly Supervised Contrastive Learning," in *ICCV*, 2021, pp. 10 022–10 031.
[19] Y. Gao *et al.*, "PyramidCLIP: Hierarchical Feature Alignment for Vision-language Model Pretraining," in *NeurIPS*, 2022, pp. 35 959–35 970.
[20] Y. Gao, J. Liu, Z. Xu, T. Wu, *et al.*, "SoftCLIP: Softer Cross-Modal Alignment Makes CLIP Stronger," in *AAAI*, 2024, pp. 1860–1868.
[21] B. Wu, "Data Efficient Language-Supervised Zero-Shot Recognition with Optimal Transport Distillation," in *ICLR*, 2022.
[22] C. Ge, J. Wang, Z. Tong, *et al.*, "Soft Neighbors are Positive Supporters in Contrastive Visual Representation Learning," in *ICLR*, 2023.
[23] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, *et al.*, "Segment Anything," in *ICCV*, 2023, pp. 4015–4026.
[24] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, *et al.*, "Segment Everything Everywhere All at Once," in *NeurIPS*, 2023, pp. 19 769–19 782.
[25] J. Yang *et al.*, "Set-of-Mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V," *arXiv preprint arXiv:2310.11441*, 2023.
[26] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *ICLR*, 2021, pp. 12 888–12 900.
[27] H. Liu, C. Li, Q. Wu, and J. Lee, "Visual Instruction Tuning," in *NeurIPS*, 2023, pp. 34 892–34 916.
[28] S. Ramakrishnan *et al.*, "Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI," in *NeurIPS*, 2021.
[29] K. Yadav, R. Ramrakhya, K. Ramakrishnan, T. Gervet, *et al.*, "Habitat-Matterport 3D Semantics Dataset," in *CVPR*, 2023, pp. 4927–4936.
[30] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, *et al.*, "Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments," in *CVPR*, 2018, pp. 3674–3683.
[31] A. Chang, A. Dai, T. Funkhouser, *et al.*, "Matterport3D: Learning from RGB-D Data in Indoor Environments," in *3DV*, 2017, pp. 667–676.
[32] Y. Qi *et al.*, "REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments," in *CVPR*, 2020, pp. 9979–9988.
[33] M. Cao, S. Li, J. Li, L. Nie, *et al.*, "Image-text Retrieval: A Survey on Recent Research and Development," in *IJCAI*, 2022, pp. 5410–5417.
[34] B. Calli, A. Walsman, A. Singh, S. Srinivasa, *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.