Manufacturing Domain Instruction Comprehension using Synthetic Data

Kritika Johari[1⋆], Christopher Tay Zi Tong[1], Rishabh Bhardwaj[1], Vigneshwaran Subbaraju[2], Jung-Jae Kim[2], and U-Xuan Tan[1]

[1] Singapore University of Technology and Design, Singapore
kritika_johari@mymail.sutd.edu.sg, ctzy818@gmail.com,
rishabh_bhardwaj@mymail.sutd.edu.sg, uxuan_tan@sutd.edu.sg
[2] Agency for Science Technology and Research (A*STAR), Singapore
vigneshwaran_subbaraju@ihpc.a-star.edu.sg,
jjkim@i2r.a-star.edu.sg

Abstract. Referring Expression Comprehension (REC) system solves a task to localize objects in a given image, based on natural language expression. We propose a novel approach to adapting the pre-trained REC model for the manufacturing domain.

Introduction: Despite significant advances in REC research, current REC datasets fail to recognize objects from specific yet important domains such as manufacturing due to the absence of domain-specific samples during training. Thus, we introduce a synthetic data-based domain adaptation approach for REC.

Methods: To adapt a REC model to the manufacturing domain, we generated a synthetic REC dataset RefMD that consists of two sub-datasets: 1) dataset for manufacturing object classification, and 2) dataset for REC adaptation to manufacturing. Each dataset serves as one step toward the REC adaptation. Adaptation of the object classification network (visual backbone) is carried out by training ResNet50 on domain-specific labeled data, while the REC adaptation completes with the adaptation of modules in RealGIN altogether. The manufacturing domain-adapted model is further enhanced with the capability to handle ambiguous referring expressions through human-in-the-loop (HITL) interaction.

Results: The experiments on 3D printed manufacturing objects demonstrate that the interactive REC model can accurately comprehend human instructions with an 82% accuracy.

Conclusion: This paper introduces an approach to facilitate domain adaptation using solely synthetic data for the specific case of the manufacturing domain. However, the proposed adaptation methodology can be applied to any other domain by following the same synthetic data generation process.

Key words: Domain Adaptation, Referring Expressions comprehension, Human-in-the-loop, Human-Robot Collaboration.

⋆ corresponding author

1 Introduction

Pick-and-place robots are ubiquitous as they enhance the optimal functioning of a wide range of industries such as manufacturing, medical, and electronics. Manufacturing industries employ such robots in a variety of applications including assembling parts of a product, its inspection, packaging, and up to the final stage, i.e., product delivery. For instance, during the assembly process, a pick-and-place robot can fetch parts from different locations and assemble them into a well-functioning product [1]. However, to improve their performance in real-world situations, robots must be able to interpret natural language instructions, which are more intuitive than programming in most scenarios [2]. This requirement poses significant challenges, as spoken language instructions in everyday life are often ambiguous and not predefined. Humans do not always put effort into making their instructions clear, which further complicates the problem [3].

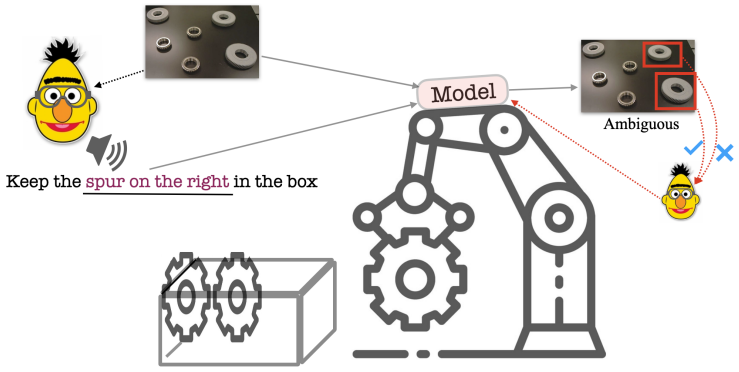

Fig. 1. Human-Robot Interaction in manufacturing domain: Human instructs a pick and place robot to pick an object (spur) and place it at some location (box) using natural language. The robot in return highlights candidate objects in AR glasses worn by the human and asks for confirmation to resolve ambiguity.

Natural language communication between humans and robots has gained attention due to its intuitive nature [4]. Figure 1 represents a scenario where a human gives a robot a verbal order to pick an object. Thus, it is desirable for the robot to have the capability to understand unconstrained natural language to identify the object to be picked. A manifestation of this capability is the task of referring expression comprehension (REC). REC aims to detect objects in an image specifically referenced by the query text [5]. The task is formulated as follows: Given an image $\mathcal{I}$ and a query text $\mathcal{Q}$, identifying the object bounding box $\mathcal{B}$ in $\mathcal{I}$ referred by $\mathcal{Q}$.

The existing benchmark REC datasets contain referring expressions for a limited number of object categories. For example, ReferIt [6] consists of referring expressions for 238 object categories including animals, people, buildings, etc. RefCOCO [7] and RefCOCOg [8] comprise referring expressions for 80 object categories from the MS-COCO dataset [9]. Thus, a REC model trained on these benchmark datasets might expe-

rience a performance drop in comprehension of referring expressions for unseen object categories. This work aims to address this challenge by facilitating a robot's comprehension of unconstrained human instructions for a specific domain without the need to collect any real-world data.

We hypothesize that a system that is efficient in comprehending generic referring expressions is capable of adapting to a specialized domain with the help of synthetic data. This obviates the need to collect real-world data samples of specific domains. As an illustration, we adapted the RealGIN REC model to the manufacturing domain. We also equip the manufacturing-adapted RealGIN with the capability to comprehend ambiguous natural language with the help of a HITL disambiguation methodology. We have also tried to show the performance of the domain-adapted model's instruction comprehension on 3D printed manufacturing objects (distinct from ReferIt's object categories). Our contributions can be summarised as follows :

1. We propose an approach to enable REC for a specific domain without the need to collect domain-specific data using **RefMD** (**Ref**erring Expressions for objects in **M**anufacturing **D**omain), a synthetic dataset for image classification and referring expression comprehension task in the manufacturing domain.
2. HITL disambiguation methodology to resolve ambiguous referring expressions in one-stage REC model. We established a bi-directional communication between the Epson AR glasses worn by the human and the manufacturing domain adapted RealGIN to accomplish the HITL interaction.
3. Demonstrate and evaluate the extension of the model trained on synthetic data to manufacturing objects and examination of its generalization capability to novel objects.

2 Related Work

2.1 Comprehension of Natural Language Expression

Referring expression comprehension (REC) requires an understanding of complex language semantics and various types of visual information, such as objects, attributes, and regions [5]. Existing approaches to supervised REC can be classified into two categories: two-stage and one-stage. Two-stage REC possesses an explicit object detection network (such as Faster RCNN [10]) to generate proposals, followed by a network to rank the identified objects with the grounding (representations) of textual query [11], [12]. The object detection network can only create the candidate regions corresponding to a limited number of trained categories [13]. An alternative approach overcomes this limitation, known as the one-stage or single-stage REC system. The single-stage network combines the proposal generation and a grounding network and formulates the task as a bounding box regression problem instead of a region-query ranking problem [14], [15]. It is faster and computationally efficient than the two-stage network and therefore suitable for practical applications like video surveillance and text-to-image retrieval [16].

2.2 Handling Ambiguous Instructions

Ambiguity is one of the main limitations of using natural language with robots. A natural language instruction referring to an object can be ambiguous due to such reasons as the following: 1) visual - object similarity, object proximity, and perspective [17], and 2) linguistic - despite clear visibility of objects and an uncluttered environment, humans often fail to give precise and clear instructions [18]. [2] proposes an approach to allow the robot to ask fixed generic questions, such as "which one?" and display candidate objects. The operator is then allowed to use unconstrained language expressions as feedback to direct the robot to the requested item.

In the case of INGRESS [19], an interactive process is employed for final object disambiguation. For each candidate object, a specific question is presented to the human, such as "Do you mean E?" The E is derived from the grounding process within INGRESS that involves a two-stage generation process. The initial stage utilizes a neural network to generate visual descriptions of objects and then compares these descriptions with the input language expression to identify a group of candidate objects. The second stage employs another neural network to analyze all potential pairwise relationships among the candidates and deduce the most probable referred object. Following the question, the user has the option to respond with "yes" to confirm the referred object, "no" to explore other possible objects or provide a specific response.

An extension of this approach is INGRESS-POMDP [19], which employs a mathematical framework to reason about the entire interaction history. It maintains a belief, which is a probability distribution over the object being referred to. The comprehension model then explores a tree that represents possible sequences of future interactions and selects the most appropriate question based on this belief. After the user answers the question, the belief is updated accordingly.

2.3 Synthetic Data for REC

Researchers have shifted toward synthetic data to overcome challenges such as cost, bias, and privacy associated with data collection [20]. The use of synthetic data is becoming a promising alternative to real-world data collection for the REC task [21]. Synthetic REC datasets such as CLEVR-Ref+ [22] and SynthRef [23] are based on an existing benchmark dataset CLEVR [24] and YouTube-VIS [25] respectively. CLEVR-Ref+ [22] transforms the questions and answers of [24] to referring expressions and referred objects. SynthRef [23] utilizes the classes and bounding box of the target object from a video frame of [25] and heuristically generates a referring expression. However, these datasets comprise generic object categories. Inspired by [22] and [23], we utilize a benchmark dataset of natural language corpus for object reference in manipulation scenario to create a manufacturing domain-specific synthetic REC dataset, which to the best of our knowledge is the first synthetically generated manufacturing domain REC dataset.

3 Methodology

Consider a database from manufacturing domain $\mathcal{X}^M$=$\{(im_1, txt_1), \ldots, (im_n, txt_n)\}$ where tuple (im_k, txt_k) $(k \in [1, n])$ represents an image-text pair with text referring to

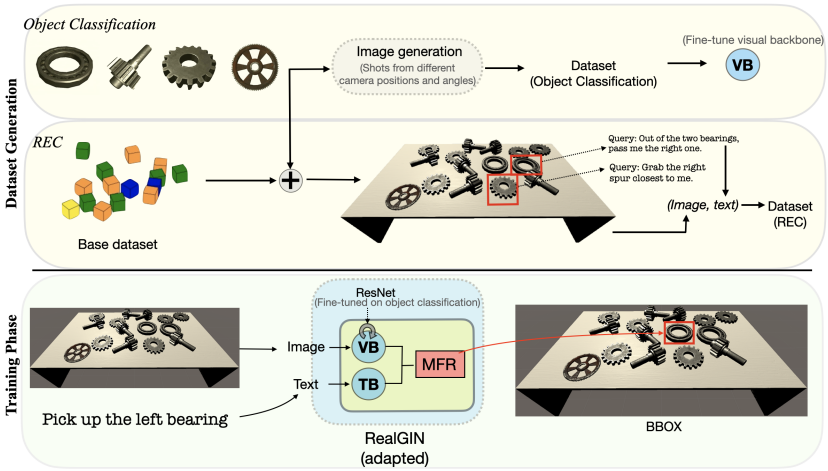

Fig. 2. An overview of the REC adaptation process. **(1) Dataset Generation**: In the top box, the object classification (OC) dataset is produced by capturing 3D mechanical objects from various camera angles and positions. RealGIN's visual backbone is fine-tuned with this OC dataset. In the middle box, 3D objects are combined with the cube arrangement of the instruction corpus to create a simulation environment from which image-text pairs are obtained. **(2) Training phase**: The image and text pairs generated from the simulation environment are used to train the RealGIN model with the ResNet fine-tuned on the OC dataset.

one or multiple objects inside the corresponding image[1]. The task is to train a learnable system that, given an unseen sample tuple of query image im_q and a text txt_q, can localize the manufacturing objects referred to by the text. We assume that the learnable system possesses enough capacity to encode the multimodal entangled (fusion) features important for learning to solve the REC task. On the other hand, learning a model of sufficient capacity warrants the need for a large number of labeled training samples, which is difficult to obtain, time-consuming, and cost ineffective for a niche domain such as manufacturing. Thus, instead of learning a domain-specific REC system, we exploit a widely used pre-trained model, real-time global inference network (RealGIN) [14], and perform domain adaptation [26]:

REC—RealGIN is a one-stage REC system that extracts multimodal features (denoted as MFR) from the image-text pair input using a ResNet-based visual backbone (VB) and a bi-GRU-based textual backbone (TB) as shown in Figure 2. VB takes the image as input and, after a sequence of convolutional operations, outputs a visual feature map $\mathbf{F}_{\mathrm{v}}$. TB is a bi-GRU (Bi-directional Gate Recurrent Unit) network [27] that maps the input discrete text to a vector space of a fixed dimension $\mathbf{F}_{\mathrm{t}}$. RealGIN is pre-trained on ReferIt, i.e., a phrase grounding benchmark dataset [6]. It is a subset of ImageClef [28] that consists of 20k images along with 85k query phrases. Natural language expressions are gathered from the images of different sports, actions, people, animals,

[1] We consider that multiple objects in the image can satisfy the referring text condition. We resolve the ambiguity later by bringing human in-loop.

cities, landscapes, and many other aspects of contemporary life in the two-player game. This dataset, with its vast size and diversity, inherently captures a range of referring expressions.. Therefore, we posit that RealGIN is an appropriate pre-trained model that encodes critical information required for a domain-generic language grounding.

To perform manufacturing domain adaptation of the underlying pre-trained REC model we propose a two-stage process, stage-1) domain adaptation of VB for manufacturing object classification, and stage-2) domain adaptation of the complete REC system for manufacturing object localization. The manufacturing domain often involves complexities such as clutter, spatial relationships between objects, orientations, and interactions within a confined space [29]. Therefore, before performing domain adaptation and as an important contribution to this work, we create a manufacturing domain-specific dataset. The dataset is a combination of two sub-datasets dedicated to each stage of the adaptation:

3.1 Datasets for REC in Manufacturing

Dataset for manufacturing object classification To mitigate the challenge of obtaining labeled data for the supervised classification task, we construct a dataset artificially. We utilize `Unity`, a game engine that can be used to create a virtual environment with objects mapped in three-dimensional space. We obtained four common 3D mechanical objects {bearing, gear, spur, and wheel} from CoolWorks [2]. To construct a large dataset from the four 3D mechanical objects for classifier tuning, we perform perspective transforms. As shown in Figure 3, we place a mechanical object at the center (origin) and change the location of the camera (perspective). At the start, we place the camera at a distance r from the center of the object in the same horizontal place as the object. With respect to the vertical axis and always facing the object, the camera revolves around the object at an angular speed of 9° per second and captures object perceptions (camera shots) at a rate of 60 images per second. After each revolution, we change the height of the camera by 9° to rotate around a new circle and repeat the process until it reaches the top. Following this process, all four objects produced labeled 8,648 images in total. A sample of thus obtained images can be seen in Figure 2 at the top [3].

Dataset for REC adaptation to manufacturing Using the four manufacturing objects, we create multiple scenes for object localization in a given image. A frugal yet sustainable alternative to collecting labeled data is to synthesize it using computer graphics [30]. The graphics allow us to create rich synthetic labeled data with easy control over variations and diversity without having to worry about collection and annotation noises that are usually a huge limitation of real data curation [31]. Thus, we created a synthetic referring expression dataset for the manufacturing domain, **RefMD** (**Ref**erring Expressions for objects in **M**anufacturing **D**omain) available in the supplementary material link. The main components of RefMD are–synthetic images, referring expressions, and ground-truth bounding boxes.

[2] The CoolWorks asset package is a set of various PBR (Physically Based Rendering) ready gears and accessories published by CoolWorks Studio designer www.coolworks.cz

[3] For better illustration in Figure 2, we crop extra white space around the image.

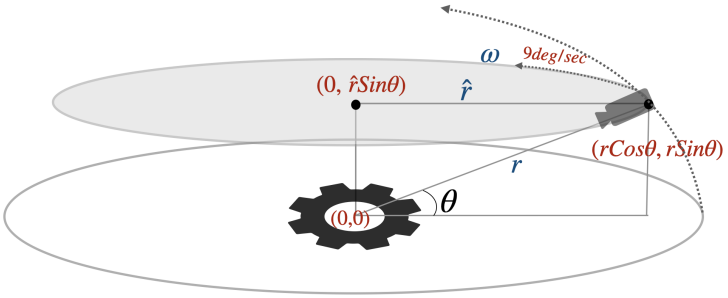

Fig. 3. Camera revolving around a circle of radius $\hat{r}$ with an angular velocity ω of 9° per second and capturing images of an object lying at the center of a hemisphere of radius r. The camera is at an angle θ with the x-axis which increases with a step size of 9° until it reaches the top.

We employed a distinct methodology to generate RefMD. Initially, we transformed a pre-existing synthetic dataset into a manufacturing domain-specific dataset. To determine the most appropriate dataset to exploit, we conducted a comparative analysis between existing synthetic REC datasets, as illustrated Table 1. Three distinct datasets were analyzed in this study. Firstly, the CLEVR-Ref+ (Compositional Language and Elementary Visual Reasoning for Referring Expression) [22] dataset, which was designed to mitigate dataset bias by employing modular programs to provide detailed reasoning ground truth without relying on human annotators. Secondly, the SynthRef dataset [23], which aimed to create the first large-scale repository of synthetic referring expressions tailored for video object segmentation. Lastly, the Natural Language Corpus [17], which was compiled to gather a comprehensive collection of domain-specific natural language for a robot manipulation task, thereby establishing a benchmark of image-instruction pairs to reveal the inherent challenges in tabletop object specification.

Table 1. Comparison analysis between various synthetic REC datasets

Dataset/ Attributes	CLEVR-Ref+	SynthRef	Natural Language Corpus
Type of Objects	Three object shapes, two absolute sizes, two materials, and eight colors.	40 object categories including everyday objects such as persons, animals, and vehicles.	Orange, yellow, green, and blue blocks.
Ref Exp Source	Functional programs.	Algorithmically generated.	Human written instructions.
Sample Size	100k images, 3 to 10 objects in an image.	3774 object types and 15,798 referring expressions.	15 objects, 28 images, and 1582 referring expressions.

Among the options discussed in Table 1, SynthRef [23] proves to be the least suitable choice for serving as the foundational dataset of a domain-specific synthetic dataset. This is because it was originally designed for video object segmentation and the objects it features are significantly different from those typically found in the manufacturing context.

On the other hand, the referring expressions in CLEVR-Ref+ [22] are derived from the same functional programs used for the CLEVR [24] VQA (visual question answer-

ing) dataset's questions. These functional programs are constructed using a small set of basic functions such as relate, query, filter, etc. Additionally, the dataset relies on predetermined definitions of spatial relationships like left versus right and behind versus front. As a result, CLEVR-Ref+'s referring expressions lack the linguistic breadth of natural human language. For instance, people might use spatial relationships like "left", "right", etc. differently than how they are used in CLEVR questions.

Therefore, due to the greater relevance to tabletop scenarios during assembly within the manufacturing domain and human written instructions, the natural language corpus is the most suitable choice for the base dataset. Also, the limited scale of the corpus helps to illustrate the approach of generating domain-specific synthetic REC data easily. The scenarios with 15 blocks in our dataset create a higher complexity for generating and understanding spatial references than most of the other datasets. While the dataset is specifically designed for the manufacturing domain, it is worth noting that the generation methodology can be readily adapted to generate for any other domain too.

For RefMD construction, we utilize a corpus [17] consisting of 1,582 natural language instructions that refer to objects of interest (or target objects) in a collaborative tabletop manipulation setting. Thus, the corpus is a collection of image-natural language instruction pairs referring to a target object. The Figure 4 shows two out of fourteen configurations with the text instruction for the right being more complex than the left configuration. Given the comprehensiveness of the baseline dataset encompassing simple and complex referring expressions, it becomes the base for the construction of RefMD.

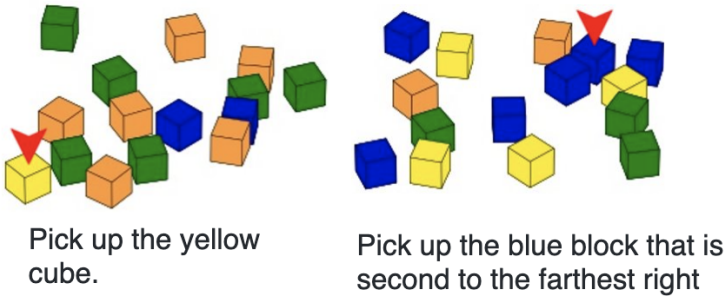

Fig. 4. Examples of two image-instruction pairs. The images are typical configurations with 15 randomly-spaced, randomly-colored blocks (blue, green, orange, or yellow) arranged on a table and instruction is a description of the unique location of the target object (block marked by a red arrow)

For our use case, we recreate object arrangements from the base dataset in `Unity` and replace the cubes with manufacturing objects' models [29] as shown in Figure 2. We adopted a color-based rule to replace the cubes in the base dataset images with various mechanical objects. The mechanical object image in Figure 2 illustrates this rule by substituting spurs for all the green cubes, gears for the orange cubes, bearings for the

blue cubes, and a wheel for the yellow cubes in the base dataset image. Additionally, we substituted the mechanical object type of the target color of each data instance with each of the other three object types. For instance, in the query "Out of the two bearings, pass me the right one" in Figure 2, the blue cubes are first changed to bearings, and the other cubes are substituted as mentioned above. Then, we substituted the blue cubes with, for example, spurs and the green cubes with bearings while substituting the orange and yellow cubes with gears and wheels, respectively, as mentioned above. Thus, each image of the cube configuration produced four images of the mechanical objects' configuration. While we recognize the removal of color references, it's important to note that such examples are already present in the ReferIt dataset, from which our model draws its pre-training.

We also substitute the corresponding box references in the text with the name of the mapped manufacturing object. To obtain ground truth bounding boxes, we utilize the `Unity` perception[4]. Using this, we obtain bounding boxes corresponding to each object in the image followed by manually selecting the correct bounding box (ground truth). This integration of programmatic techniques significantly reduced the manual burden associated with annotation, making the dataset generation process more efficient. One of the key advantages of our methodology is its scalability across various domains. By replacing 3D models programmatically, our approach becomes adaptable to different object types, facilitating the creation of datasets tailored to specific domains without the need for extensive manual re-annotation..

Thus, we obtained a labeled synthetic referring expression dataset RefMD with input set $\mathcal{X}^M$ and ground truth bounding boxes $\mathcal{Y}^M$. Input is an image-text tuple $(im_k, txt_k) \in \mathcal{X}^M$ and corresponding output is a set of candidate ground truth bounding boxes y_k=$\{b_1, \ldots, b_n\} \in \mathcal{Y}^M$ where n can be as high as the number of objects in the underlying image.

3.2 Domain Adaptation of REC

Adapting Visual Backbone. The first stage of our two-stage domain adaptation approach is the adaptation of VB for the manufacturing domain. The VB is a RetinaNet [32] with ResNet50 [33] network. We use the ResNet50 as image encoder, which is pre-trained with a widely used image classification dataset ImageNet [34]. Thus, to let the ResNet50 network learn about the new domain-specific classes, we further train it with the obtained images during dataset construction of manufacturing object classification (Section 3.1-(a)). We replace the last 1000 class classification layer of ResNet50 with a four-class classification layer where each class index corresponds to a unique object from the manufacturing domain, i.e., one of bearing, spur, gear, and wheel. The adaptation of ResNet50 architecture obviated the need to learn a neural architecture from scratch which would also warrant a much larger domain-specific dataset to prevent overfitting [35].

[4] An easy-to-use and highly customizable toolset which accelerates the process of generating synthetic datasets for computer vision tasks [20]

Adapting RealGIN. Once we obtain a domain-specific image classification backbone, we perform an adaptation of the complete REC system by performing RealGIN fine-tuning. As shown in Figure 2, during the training phase, RealGIN takes the image-text tuple as the input. The image is passed through VB to obtain a visual feature map. The text is passed through the bi-GRU to obtain textual features and the concatenation of forward and backward context. Both visual feature maps and textual features are concatenated to obtain a multimodal feature vector that is used to assign scores to the anchor boxes recognized by RealGIN.

We perform manufacturing domain-specific adaptation of RealGIN in two different ways: The first method is to fine-tune the pre-trained RealGIN model on the RefMD dataset (the training, test, and validation split is mentioned in 4.1), called (pre-training + fine-tuning). In the second method called (training from scratch), the RefMD dataset was combined with $\mathcal{X}^R$ (ReferIt) data and trained with the adapted ResNet50. The RealGIN pre-trained on the ReferIt [6] serves as the baseline for the adapted models from the two approaches. Also, out of the two approaches, fine-tuning is the widely used approach for adapting a model on a new dataset and therefore, is considered the baseline adaptation approach. The training and validation accuracy and loss are shown over 10 epochs for both approaches shown in Figure 6.

3.3 Inference-time Disambiguation through Human-in-the-loop

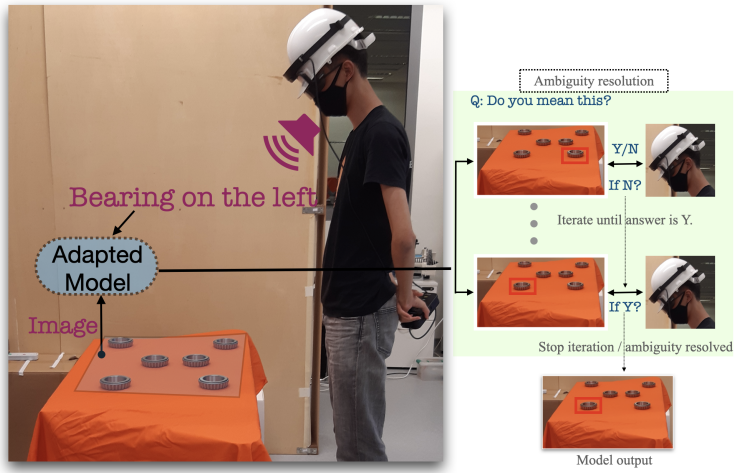

Fig. 5. Human-in-the-loop disambiguation approach: The participant describes an object from an object arrangement of 3D printed bearings while wearing the AR glasses. Images captured from the glasses along with the natural language object description are inputted into the adapted model. The adapted model presents candidate bounding boxes (shown in dotted red boxes whose confidence scores by RealGIN are above a threshold) to the human operator for confirmation.

During inference, owing to the limited domain-specific data as well as its characteristics away from realistic scenarios, the REC model faces difficulty in comprehending referring expressions when there is a comprehension ambiguity. Therefore, we propose a confidence threshold-based disambiguation approach for the adapted model to comprehend ambiguous referring expressions during inference as illustrated in Figure 5. The disambiguation requires human confirmation on the candidate bounding boxes output by the model. The confirmation is obtained by establishing bi-directional communication between the REC model and AR glasses worn by the human agent as illustrated in Figure 5.

$$\arg\min_{\theta} E_{q,o}[\max\{0, m - f_\theta(q,o) + f_\theta(q,o')\} + \max\{0, m - f_\theta(q',o) + f_\theta(q,o)\}] \tag{1}$$

The disambiguation approaches mentioned in the related work section [2] and [19] do not apply to our REC models, which is a one-stage REC. The REC model [2] essentially scores a pair of a sentence and an object. The training procedure involves minimizing a max-margin loss between scores of positive and negative object–sentence pairs as shown in Equation (1) where q and q' are correct and incorrect sentences, and o and o' are correct and incorrect objects. In other words, this training aims to guarantee that every correct pair of a sentence and an object has scored higher by a margin m than any other pair with a wrong sentence or object. If multiple objects have scores within this threshold, the model treats all such objects as potential options. This allows the model to ask for more details while highlighting relevant objects.

Algorithm 1 illustrates the proposed disambiguation methodology which depends on the confidence scores assigned by the REC model to the various bounding boxes in an image based on the natural language instruction. In this approach, we select the top 100 scoring bounding boxes from $\mathcal{C}^M$ (a set of pairs of the bounding box and confidence scores). Non-maximum separation technique is applied to $\mathcal{C}^M$ to discard overlapping bounding boxes. From the filtered bounding boxes C, the threshold confidence score C_{th} is calculated as 80% of the maximum confidence score of C. The bounding boxes scoring higher than C_{th} are the potential targets P^t and are presented for human confirmation.

The proposed disambiguation approach heavily depends on a manually chosen threshold confidence score obtained from the final bounding box predictions by RealGIN. In contrast, [2] and [19] approaches filter the candidate objects for human confirmation from the intermediate object proposals from the object detection and self-referential networks, respectively. It is relatively more challenging to handle ambiguity in a one-stage architecture due to the lack of intermediate object proposals. This is primarily because a one-stage REC fuses visual and textual features from the image and referring expression to determine the referred object.

Our proposed disambiguation algorithm consists of three primary steps: Non-maximum suppression, Generating a potential target set (referred to as P^t), and Threshold filtering. Step 1 and Step 3 have a time complexity of $O(1)$ as the nested loops iterate over a fixed number of candidates (100 and t respectively). Moreover, the operation within the loops involves removing candidates based on the Intersection over Union (IoU) and threshold, which also has a constant time complexity. However, Step 2, which involves

Algorithm 1: Disambiguation algorithm

Data: $\mathcal{C}^M$=$\{(b_1, s_1), \ldots, (b_n, s_n)\}$ where b_k and s_k denote bounding box and its confidence score, respectively.
Result: $P^t = [b_1, \ldots, b_t]$ where b_k represents a candidate bounding box
Select top 100 scoring candidate regions in $\mathcal{C}^M$;
$C \leftarrow \{(b_1, s_1), \ldots, (b_{100}, s_{100})\}$;
for $i = 0$ *to* 100 **do**
  for $j = i + 1$ *to* 100 **do**
    if $IoU(C[i][0], C[j][0]) \geq 0.5$ **then**
      remove (b_j, s_j) from C ; `/* Non-maximum Separation */`
    else
      continue;
    end
  end
end
$P^t \leftarrow \{(b_1, s_1), \ldots, (b_t, s_t)\}$;
$C_{th} \leftarrow 0.8 * \max(s_1, \ldots, s_t)$;
for $i = 1$ *to* t **do**
  if $P^t[i][1] \geq C_{th}$ **then**
    continue ; `/* Target Filtering */`
  else
    remove (b_i, s_i) from P^t;
  end
end

creating the final potential target set, requires copying a maximum of t candidates, resulting in a time complexity of $O(t)$. When we combine these steps, the algorithm's overall time complexity is determined by the step with the highest complexity. In this case, it's the creation of the final potential target set with a time complexity of $O(t)$. Thus, the asymptotic time complexity of our algorithm is $O(t)$, where t refers to the number of candidates in the final potential target set.

In contrast, [2] propose a method that employs a scoring mechanism to evaluate a pair comprised of a sentence and an object. This pair is represented as $f_\theta(q, o)$ in Equation (1). The computational complexity of evaluating the scoring function for a single pair is $O(g)$, meaning that the inner max function and the scoring function dominate the overall complexity of the optimization problem. As a result, the asymptotic complexity of the optimization problem will likely be O(g). The impact of the expectation term on the overall complexity of the problem will depend on its implementation.

In the initial phase of INGRESS [19], a neural network is utilized to generate visual descriptions, while the subsequent phase involves the use of another neural network to analyze pairwise relationships. The intricacy of these operations depends on several factors, including the network's architecture and parameters, which are typically denoted as $O(f(n))$, where n represents the input size. These operations are usually evaluated based on training time or inference time. During the interactive disambiguation phase, a new complexity factor is introduced, which is dependent on the number of candidate objects, represented as m. If the number of candidates remains constant, the complexity is expressed as $O(1)$. However, if the number of candidates increases with the input, the complexity may rise to $O(m)$. According to INGRSS-POMDP [19], updating the probabilities of a referred object involves maintaining a belief or probability distribution through user responses. The update process depends on the size of the belief space, which is represented as $O(g(m))$, where m is the number of candidates. To explore a tree that predicts future interaction sequences, one must navigate its structure, which is often denoted as $O(h(n))$ or $O(h(m))$, where h is a function that reflects the tree's growth with the input.

In conclusion, the efficiency of disambiguation methods, such as those described in [2] and [19], is influenced by various factors, such as the scoring function, neural network operations, the number of candidate objects, belief update process, and tree exploration. To gain a better understanding of the specific complexities involved, a detailed analysis of the neural network architectures, belief update algorithm, and tree is required. In contrast, our disambiguation method streamlines the process to $O(t)$, in addition to the inference time of the one-stage REC, where t represents the number of candidates. This makes our approach faster than the techniques outlined in [2] and [19].

4 Results

For real-time inference and robust evaluations of the proposed approach, i.e., adaptation of REC on manufacturing data, we utilize 3D printed versions of the objects designed in `Unity` (Figure 5-image input to the adapted model). We also evaluate the visual backbone's performance after it has passed through the image classification training and RealGIN-based domain adaptation. The following sections elaborate on the target

dataset, evaluation metric, and comparison results of two domain adaption methods– 1) training RealGIN from scratch on domain-specific data as well as pre-training data, i.e., ReferIt, and 2) fine-tuning RealGIN on domain-specific data. We also analyze the performance of the manufacturing domain synthetic REC data generated from a subset of CLEVR-Ref+ [22] to determine the impact of the base dataset choice on the adapted model's performance.

4.1 Evaluation Metric and Target Dataset Information

A REC model's performance is evaluated using the IoU performance metric [6], [12], [7]. IoU is the area of overlap divided by the area of union between the top predicted bounding box and the ground truth box. The prediction is correct if IoU is greater than 0.5. Otherwise, the prediction is considered incorrect. This is also known as the Precision@1 score. The Precision@1 scores are averaged over all referring expressions [36]. A total of 2,740 referring expression data samples were generated using Unity software. This dataset is divided into training (1,749), test (545), and validation (446).

Table 2. Dataset information and accuracy of different adaptation methods on target distribution

Method		**Pre-training + Fine-tuning**	**Training from scratch**
Dataset Information	Train	1,749	$\mathcal{X}_{tr}^{R}$ + 1,749
	Val	446	$\mathcal{X}_{val}^{R}$ + 446
	Test	545	$\mathcal{X}_{te}^{R}$ + 545
Accuracy on target data	Val	0.508	0.542
	Test	0.510	0.557
Base dataset from CLEVR-Ref+			
Accuracy on target data	Val	0.48	0.53
	Test	0.51	0.55

We carried out a simple experiment using a smaller portion (2740 referring expressions) of the CLEVR-Ref+ and generated a synthetic dataset similar to RefMD. This was done to ensure that the sample size was in line with the natural language corpus. During the experiment, we adapted the REC model by utilizing synthetic data generated from this subset of the CLEVR-Ref+ dataset. Our objective was to examine how the quality of the initial data affected the model's performance. The results indicated that the performance of the model adapted to the subset of CLEVR-Ref+ was similar to that of the model adapted to RefMD as shown in Table 2. Moreover, both adaptation techniques share a similar pattern to RealGIN's adaptation on RefMD. Training from scratch with the adapted visual backbone outperforms fine-tuning. However, conducting a user study with the REC model adapted using the subset of CLEVR-Ref+ is not

feasible due to time limitations. We utilized the input images and referring expressions from the previous user study. The model adapted to the subset of CLEVR-Ref+ had three more failure cases.

4.2 Experimental Settings and Results

Table 2 contains the dataset information and precision@1 score for the two adaptation methods described in Subsection b) of 3.2. $\mathcal{X}^R$ represents samples from the ReferIt [6] dataset. Both fine-tuning and training from scratch utilized the same underlying dataset, but the adaptation methodologies differed. Fine-tuning employed samples exclusively from RefMD for constructing training, validation, and test sets, as the pre-trained model already had samples from ReferIt. On the other hand, training from scratch involved using a combined dataset derived from ReferIt and RefMD for the same purpose.. We ran both methods (fine-tuning and training from scratch with adapted ResNet50) for 10 epochs and at a learning rate of 0.0001. We used a batch size of 16 and Adam optimizer [37]. We didn't modify any other hyperparameters from the original RealGIN setting. We emphasize that these specific hyperparameters yielded the most favorable outcomes across our comprehensive explorations shown in Table 3. For model training, the model uses the focal loss as described in [32] to learn to differentiate between foreground and background. A smooth-L1 loss is used for bounding box regression [10].

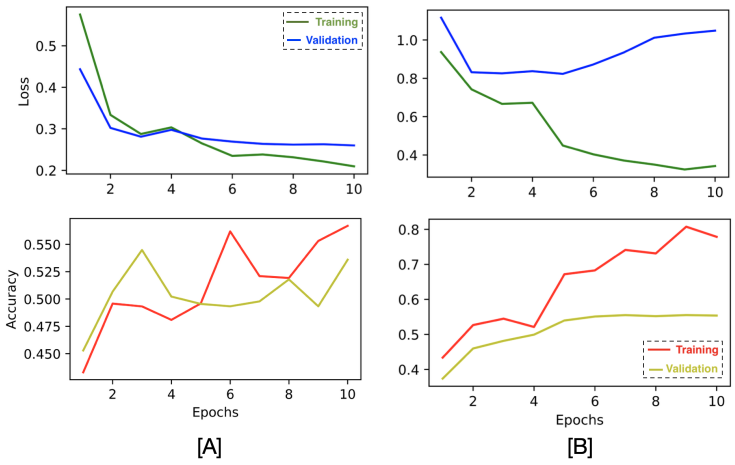

Fig. 6. Comparison of the two domain adaptation methods on the basis of training and validation loss and accuracy. The subplots on the right (B) represent the training from scratch using the adapted ResNet50 method while the subplots on the left (A) belong to the fine-tuning method.

Training RealGIN from scratch with adapted ResNet50 outperforms the fine-tuning method with a validation accuracy of 54.2% and test accuracy of 55.7% on the target dataset (validation = 446 and test = 545 data points) as mentioned in Table 2. We performed transfer learning [38, 39] on the pre-trained ResNet50 for five epochs using the

manufacturing object classification dataset and achieved 99.95% training, 100% validation, and 98.34% test accuracy.

The RefMD dataset has different referring expressions from ReferIt, causing challenges for the RealGIN model in understanding the nuances of referring expressions in a broader context. Both adaptation approaches face this issue, with the only difference in training-from-scratch being the use of an adapted visual backbone. Therefore, we demonstrate that adapting the visual backbone can improve performance when using unconstrained natural language to refer to mechanical objects in uncontrolled scenarios.

Table 3. Validation accuracy of the adapted REC model.

	Fine-tuning			**Training from scratch**		
LR/Epoch	**5**	**10**	**20**	**5**	**10**	**20**
0.001	0.4821	0.5187	0.5194	0.5143	0.5291	0.5304
0.0001	0.4955	0.5359	0.5273	0.5398	0.5538	0.5431
0.00001	0.4856	0.5238	0.5247	0.5267	0.5386	0.5411

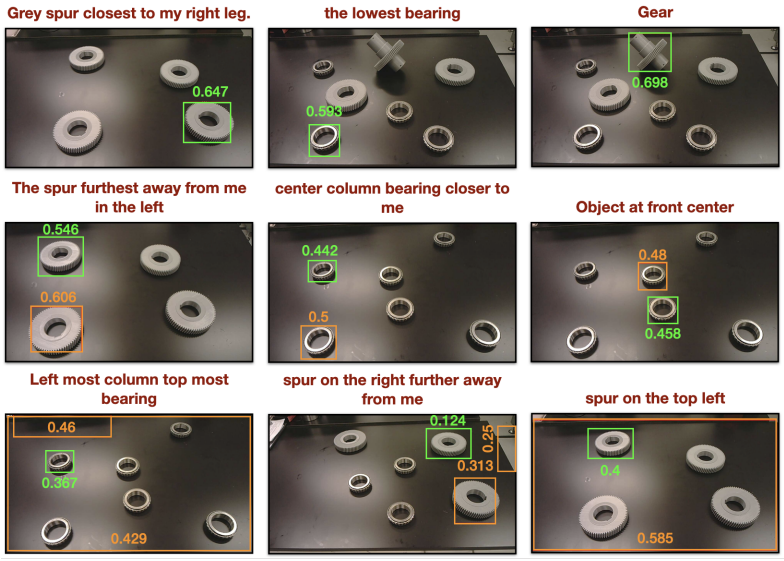

Fig. 7. Visualizations of the user-study results: In the top row are scenarios in which the candidate output by the adapted RealGIN is the referred object. In the middle row, multiple candidate boxes are output by the model for human confirmation. The green boxes are correct, while the orange ones are incorrect. Failure cases where none of the candidate boxes are referred objects are shown in the bottom row.

4.3 User-study Results with Real Objects

Since our training data only consisted of synthetic data for both visual backbone adaptation and domain adaptation of the complete REC system, we evaluated the potential of the adapted model to solve real-world instruction comprehension with 3D printed objects. We evaluated our grounding model's performance in a realistic human–machine interaction for the manufacturing task of pick and place, as shown in Figure 1. The objective was to examine its capacity to identify the referred object in the presence of diverse language expressions and visual input containing noise such as extra background, blurred image, etc.

We conducted a user study with 3D-printed objects for five human subjects (4 male, and one female) from the university community to give natural language instruction to the domain-adapted REC model to pick and place mechanical objects (a video illustrating a few scenarios are available in the supplementary link). All participants were competent in spoken English. Each participant responded to five different scenarios with 3D printed models of three mechanical parts: bearing, gear assembly, and spur. At the start of each configuration, the experimenter instructs the participant, "You have to describe this object" by touching the object. Each participant specified two objects in each object configuration.

Experiment Setup and Procedure The human operator interacts with the domain-adapted REC model with voice input through Epson Smart Glasses (Moverio Pro BT-2000), as shown in Figure 5 and in the user-study video available in the supplementary link. The adapted model (which runs on a PC workstation with an Intel i9-10900K CPU and an Nvidia RTX Turbo 2080Ti GPU) captures as input an RGB image (camera view of the smart glasses, 5 MP (Megapixel) and image resolution 1280x720 pixels) and a textual referring expression (using an in-house automatic speech recognition system developed using several large English database), and outputs a bounding box containing the referred object in the image as shown in Figure 5. The model relies on fixed generic question templates "Do you mean this?" to clarify ambiguous referring expressions interactively [2]. The user can respond as "Yes" to confirm the referred object and "No" to deny it.

This choice of interaction device is driven by the increasing presence of AR glasses in manufacturing and logistics settings [40], [41], where tasks like assembly, inspection, and machinery operation benefit from hands-free interaction. The AR glasses provide an integrated workflow and system, allowing the warehouse staff or forklift drivers to perform their tasks seamlessly, for a more efficient picking process [42]. While the impact of the interaction device might not be significant in controlled laboratory studies where objects are centralized, it becomes crucial in industrial settings. For instance, when the objects need to be gathered from different locations, the AR glasses are more suitable than a static robotic arm due to the ability to incorporate supplementary information, such as indoor navigation [43]. It also facilitates the overlay of information onto the screen [44], enhancing usability [45], [46]. The headset's lightweight nature eliminated the need to carry bulky manuals, especially valuable in sectors like industrial repair, maintenance, and heavy-duty industries such as automotive and aerospace engineering.

Data Collection We collected 50 (5 participants x 2 objects x 5 layouts) data samples to test the adapted model. We familiarised participants with object names but left them free to refer to objects in any way. Therefore, participants used multiple strategies to refer to the target object, such as using directions like the top, left, bottom, right, front, back, etc., and the position of the target object relative to nearby things.

Evaluation Criteria and Results The interactive disambiguation model depends on the user feedback to ground the natural language instruction. The grounding is accurate when the predicted bounding box has an IoU score above 0.5 with the ground truth box [6], [12], [7], which can occur with (multiple candidates above threshold) and without (only one candidate above threshold) participant confirmation. We observed that the model grounded 41 out of the 50 natural language descriptions correctly with the help of user feedback. Figure 7 illustrates nine samples of the user-study results. While the bottom row of Figure 7 displays three failure cases where none of the candidate boxes is the referred object, the top, and middle row represents the single (no confirmation required) and multiple candidate boxes (human confirmation required) scenarios. The system fails to comprehend the referring expression examples shown in the bottom row of Figure 7 because the confidence score assigned to the referred bounding box falls below the threshold confidence score. Thus, the referred bounding box doesn't make it to the human confirmation. The average grounding accuracy of the interactive model is 82%. Along with the high accuracy, the industrial pick and place system needs to have a faster inference speed. So we calculated the grounding response time of the one-stage REC model. Grounding response time is the time taken by the model from receiving the speech and image input to predicting the bounding boxes. On average, our model infers an image and text input in 2.25 seconds.

The absence of the interaction method results in lower performance compared to its inclusion, primarily attributed to the model's limited accuracy in object identification. This limitation arises from the REC's one-stage architecture lacking a dedicated object detection network. Therefore, our disambiguation helps to reduce the error along with handling ambiguous referring expressions. During the study, the participants utilized a total of 50 referring expressions. However, five of them were found to be ambiguous. Out of the 41 correctly comprehended referring expressions, a majority of 30 required no further confirmation or clarification. On the other hand, 11 of the expressions were referring to multiple things, with only three of them being genuinely ambiguous, while the remaining eight were unambiguous. Based on these findings, there seems to be a need for improvement in the current model.

Subjects Feedback Subjects expressed satisfaction with the response time of the system, finding it suitable for real-time interactions within the manufacturing workflow. Participants found it beneficial to provide clearer feedback about the referred object by highlighting it with a red bounding box, especially in situations that were complex or ambiguous. The headset used for interaction was user-friendly, with minimal disruption, enabling subjects to concentrate on their tasks while scarcely noticing the device's presence. Furthermore, the feedback mechanism involving "yes" or "no" responses felt intuitive and comfortable for the participants.

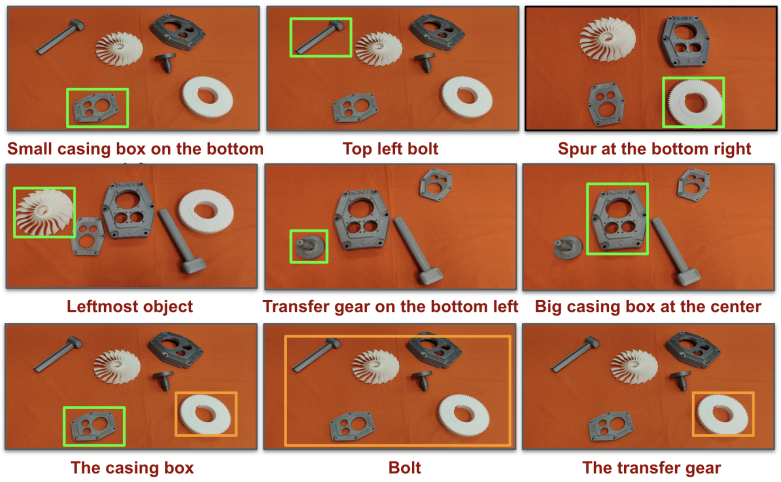

Fig. 8. Generalization of the adapted model: Assessment of the adapted model's comprehension of referring expressions for gearbox assembly objects, which are not included in ReferIt and RefMD datasets.

Generalization Capability Even though the user study employed 3D-printed objects that align in shape with the synthetic dataset, our approach capitalizes on a well-established vision-and-language alignment. As a result, our one-stage REC model theoretically can conduct fine-grained detection for unseen categories, even in scenarios where referents haven't appeared during pretraining or training. This implies that our method's generalization capability extends beyond the confines of shape-matching objects, encompassing cases where objects exhibit slight variations from the shapes encountered by the network. To further validate this, we conducted tests on objects from the gearbox assembly, which were absent in the ReferIt dataset used for RealGIN's pretraining as shown in Figure 8. This demonstrates the model's ability to generalize effectively to novel objects [47].

Our model has demonstrated impressive proficiency in recognizing object categories with spatial references that were previously unseen. However, it finds it difficult to assess its ability to identify objects in cluttered environments without clear spatial references. For example, in the left image of the bottom row in Figure 8, we introduced an ambiguous reference to a "casing box at the back." While the first candidate was correctly identified, the second candidate was not. This illustrates a potential limitation of our model when faced with ambiguous references. In other instances, such as identifying bolt and transfer gear object categories, the model made incorrect associations based on the background and spur objects, respectively.

5 DISCUSSION AND FUTURE WORK

The proposed adaptation approach applies to any domain (including medical, manufacturing, and electronics) and the REC model. We adapted the RealGIN model to the

manufacturing domain. In comparison to ReferIt [6] and other benchmark datasets, RealGIN's performance on RefMD has declined for several reasons. These mechanical objects are not present in the benchmark datasets. Furthermore, RefMD derives its referring expressions from the natural language instruction corpus which distinguishes the target object from many visually similar objects or distractions. Consequently, the combination of novel categories of objects and complex referring expressions results in decreased performance on RefMD. Improvement in disambiguation and REC performance can be achieved by comparing and exploring different and more sophisticated disambiguation approaches, such as attribute-guided disambiguation [48], to improve the accuracy of grounding as well as by incorporating gesture [49] and gaze [50] information. While the adapted model comprehends the natural language object descriptions with 82% accuracy in the user study, the domain gap between the synthetic and real-world application can further be reduced by incorporating more variation and randomization [29] in the RefMD dataset.

6 CONCLUSIONS

We propose an approach for adapting REC models using synthetic domain data. To illustrate this approach for the manufacturing domain, we generated a synthetic REC dataset RefMD from 3D models of manufacturing objects. It consists of two sub-datasets: 1) dataset for object classification, and 2) dataset for REC adaptation. In order to make the REC model's visual backbone (pre-trained on ImageNet [34]) identify manufacturing objects, it is fine-tuned on the first dataset. The second dataset is used to fine-tune the REC model (pre-trained on ReferIt [6]) and produces the adapted model. To handle the ambiguity in referring expressions, we equipped the adapted model with a confidence score-based HITL interaction method. Our user study with 3D printed manufacturing objects demonstrates the extensibility of the model adapted on synthetic data to real objects.

Compliance with Ethical Standards

In this paper, we have developed an approach whereby we generate synthetic data, thus skipping the process of data collection for machine learning. The objective of performing the experiment with humans is to demonstrate the feasibility. As the tasks are simple (minimal risks to participants) and no personal identifier is collected, we have obtained IRB exemption approval.

Data availability

Object Classification Dataset: The synthetic images of manufacturing objects used to adapt the visual backbone were generated using the data generation methodology demonstrated in Figure 3. These images are publicly available in the supplementary material link[5].

[5] `https://anonymous.4open.science/r/identify-referred-object-in-image-5917/Manufacturing_Domain/data/readMe`

RefMD Dataset: Additionally, the synthetic REC dataset used in this study contains synthetic images of manufacturing objects arranged in various layouts, as depicted in Figure 2 and Figure 4. The bounding box coordinates obtained from the Unity engine and natural language expressions from the base dataset are also available in the supplementary material link.

User-Study Video: Lastly, the video illustrating the experimental setup and some of the scenarios used in the user-study can be found at[6].

Conflict of Interest

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project # A18A2b0046). All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

1. Du, K.L., Huang, X., Wang, M., Hu, J.: Assembly robotics research: a survey. International Journal of Robotics and Automation **14** (1999) 171–183
2. Hatori, J., Kikuchi, Y., Kobayashi, S., Takahashi, K., Tsuboi, Y., Unno, Y., Ko, W., Tan, J.: Interactively picking real-world objects with unconstrained spoken language instructions. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE (2018) 3774–3781
3. Pramanick, P., Sarkar, C., Paul, S., dev Roychoudhury, R., Bhowmick, B.: Doro: Disambiguation of referred object for embodied agents. IEEE Robotics and Automation Letters **7** (2022) 10826–10833
4. Thomason, J., Padmakumar, A., Sinapov, J., Walker, N., Jiang, Y., Yedidsion, H., Hart, J., Stone, P., Mooney, R.J.: Improving grounded natural language understanding through human-robot dialog. In: 2019 International Conference on Robotics and Automation (ICRA), IEEE (2019) 6934–6941
5. Qiao, Y., Deng, C., Wu, Q.: Referring expression comprehension: A survey of methods and datasets. IEEE Transactions on Multimedia (2020)
6. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). (2014) 787–798
7. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European Conference on Computer Vision, Springer (2016) 69–85
8. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 11–20
9. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, Springer (2014) 740–755

[6] `https://anonymous.4open.science/r/identify-referred-object-in-image-5917/Manufacturing_Domain/video`

10. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497 (2015)
11. Deng, C., Wu, Q., Wu, Q., Hu, F., Lyu, F., Tan, M.: Visual grounding via accumulated attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2018) 7746–7755
12. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 1307–1315
13. Sadhu, A., Chen, K., Nevatia, R.: Zero-shot grounding of objects from natural language queries. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 4694–4703
14. Zhou, Y., Ji, R., Luo, G., Sun, X., Su, J., Ding, X., Lin, C.w., Tian, Q.: A real-time global inference network for one-stage referring expression comprehension. arXiv preprint arXiv:1912.03478 (2019)
15. Chen, X., Ma, L., Chen, J., Jie, Z., Liu, W., Luo, J.: Real-time referring expression comprehension by single-stage grounding network. arXiv preprint arXiv:1812.03426 (2018)
16. Liao, Y., Liu, S., Li, G., Wang, F., Chen, Y., Qian, C., Li, B.: A real-time cross-modality correlation filtering method for referring expression comprehension. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 10880–10889
17. Scalise, R., Li, S., Admoni, H., Rosenthal, S., Srinivasa, S.S.: Natural language instructions for human–robot collaborative manipulation. The International Journal of Robotics Research **37** (2018) 558–565
18. Shridhar, M., Hsu, D.: Interactive visual grounding of referring expressions for human-robot interaction. arXiv preprint arXiv:1806.03831 (2018)
19. Shridhar, M., Mittal, D., Hsu, D.: Ingress: Interactive visual grounding of referring expressions. The International Journal of Robotics Research **39** (2020) 217–232
20. Borkman, S., Crespi, A., Dhakad, S., Ganguly, S., Hogins, J., Jhang, Y.C., Kamalzadeh, M., Li, B., Leal, S., Parisi, P., et al.: Unity perception: Generate synthetic data for computer vision. arXiv preprint arXiv:2107.04259 (2021)
21. Gorniak, P., Roy, D.: Grounded semantic composition for visual scenes. Journal of Artificial Intelligence Research **21** (2004) 429–470
22. Liu, R., Liu, C., Bai, Y., Yuille, A.L.: Clevr-ref+: Diagnosing visual reasoning with referring expressions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 4185–4194
23. Kazakos, I., Ventura, C., Bellver, M., Silberer, C., Giró-i Nieto, X.: Synthref: Generation of synthetic referring expressions for object segmentation. arXiv preprint arXiv:2106.04403 (2021)
24. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2017) 2901–2910
25. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 5188–5197
26. Kong, X., Xia, S., Liu, N., Wei, M.: Gada-segnet: gated attentive domain adaptation network for semantic segmentation of lidar point clouds. The Visual Computer (2023) 1–11
27. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
28. Escalante, H.J., Hernández, C.A., Gonzalez, J.A., López-López, A., Montes, M., Morales, E.F., Sucar, L.E., Villasenor, L., Grubinger, M.: The segmented and annotated iapr tc-12 benchmark. Computer vision and image understanding **114** (2010) 419–428

29. Tang, P., Guo, Y., Zheng, G., Zheng, L., Pu, J., Wang, J., Chen, Z.: Two-stage filtering method to improve the performance of object detection trained by synthetic dataset in heavily cluttered industry scenes. The Visual Computer (2023) 1–20
30. Wood, E., Baltrusaitis, T., Zhang, X., Sugano, Y., Robinson, P., Bulling, A.: Rendering of eyes for eye-shape registration and gaze estimation. In: Proceedings of the IEEE international conference on computer vision. (2015) 3756–3764
31. Wood, E., Baltrušaitis, T., Hewitt, C., Dziadzio, S., Cashman, T.J., Shotton, J.: Fake it till you make it: face analysis in the wild using synthetic data alone. In: Proceedings of the IEEE/CVF international conference on computer vision. (2021) 3681–3691
32. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. (2017) 2980–2988
33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 770–778
34. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee (2009) 248–255
35. Bhardwaj, R., Saha, A., Hoi, S.C.: Vector-quantized input-contextualized soft prompts for natural language understanding. arXiv preprint arXiv:2205.11024 (2022)
36. Nagaraja, V.K., Morariu, V.I., Davis, L.S.: Modeling context between objects for referring expression understanding. In: European Conference on Computer Vision, Springer (2016) 792–807
37. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
38. Ahmad, N., Asghar, S., Gillani, S.A.: Transfer learning-assisted multi-resolution breast cancer histopathological images classification. The Visual Computer **38** (2022) 2751–2770
39. Prakash, A.J., Prakasam, P.: An intelligent fruits classification in precision agriculture using bilinear pooling convolutional neural networks. The Visual Computer **39** (2023) 1765–1781
40. Reif, R., Walch, D.: Augmented & virtual reality applications in the field of logistics. The Visual Computer **24** (2008) 987–994
41. Yu, K., Ahn, J., Lee, J., Kim, M., Han, J.: Collaborative slam and ar-guided navigation for floor layout inspection. The Visual Computer **36** (2020) 2051–2063
42. Latif, U.K., Shin, S.Y.: Op-mr: the implementation of order picking based on mixed reality in a smart warehouse. The Visual Computer **36** (2020) 1491–1500
43. Qin, Y., Chi, X., Sheng, B., Lau, R.W.: Guiderender: large-scale scene navigation based on multi-modal view frustum movement prediction. The Visual Computer (2023) 1–11
44. Xiang, N., Liang, H.N., Yu, L., Yang, X., Zhang, J.J.: A mixed reality framework for microsurgery simulation with visual-tactile perception. The Visual Computer (2023) 1–13
45. Ayadi, M., Scuturici, M., Ben Amar, C., Miguet, S.: A skyline-based approach for mobile augmented reality. The Visual Computer **37** (2021) 789–804
46. Jurado, D., Jurado, J.M., Ortega, L., Feito, F.R.: Geuinf: real-time visualization of indoor facilities using mixed reality. Sensors **21** (2021) 1123
47. Bhagat, P., Choudhary, P., Singh, K.M.: A study on zero-shot learning from semantic viewpoint. The Visual Computer **39** (2023) 2149–2163
48. Yang, Y., Lou, X., Choi, C.: Interactive robotic grasping with attribute-guided disambiguation. In: 2022 IEEE International Conference on Robotics and Automation (ICRA), IEEE (2022)
49. Liang, H., Yuan, J., Thalmann, D., Thalmann, N.M.: Ar in hand: Egocentric palm pose tracking and gesture recognition for augmented reality applications. In: Proceedings of the 23rd ACM international conference on Multimedia. (2015) 743–744

50. Johari, K., Tong, C.T.Z., Subbaraju, V., Kim, J.J., Tan, U., et al.: Gaze assisted visual grounding. In: International Conference on Social Robotics, Springer (2021) 191–202