2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

PGA: Personalizing Grasping Agents with Single Human-Robot Interaction

Junghyun Kim[12] Gi-Cheon Kang[12*] Jaein Kim[13*] Seoyun Yang[4] Minjoon Jung[12] Byoung-Tak Zhang[123]

***Abstract*— Language-Conditioned Robotic Grasping (LCRG) aims to develop robots that comprehend and grasp objects based on natural language instructions. While the ability to understand personal objects like *my wallet* facilitates more natural interaction with human users, current LCRG systems only allow generic language instructions, *e.g., the black-colored wallet next to the laptop*. To this end, we introduce a task scenario GRASPMINE alongside a novel dataset aimed at pinpointing and grasping personal objects given personal indicators via learning from a single human-robot interaction, rather than a large labeled dataset. Our proposed method, Personalized Grasping Agent (PGA), addresses GRASPMINE by leveraging the unlabeled image data of the user's environment, called *Reminiscence*. Specifically, PGA acquires personal object information by a user presenting a personal object with its associated indicator, followed by PGA inspecting the object by rotating it. Based on the acquired information, PGA pseudo-labels objects in the *Reminiscence* by our proposed label propagation algorithm. Harnessing the information acquired from the interactions and the pseudo-labeled objects in the *Reminiscence*, PGA adapts the object grounding model to grasp personal objects. This results in significant efficiency while previous LCRG systems rely on resource-intensive human annotations—necessitating hundreds of labeled data to learn *my wallet*. Moreover, PGA outperforms baseline methods across all metrics and even shows comparable performance compared to the fully-supervised method, which learns from 9k annotated data samples. We further validate PGA's real-world applicability by employing a physical robot to execute GRSAPMINE. Code and data are publicly available at https://github.com/JHKim-snu/PGA.**

I. INTRODUCTION

Empowering robots with the ability to comprehend human natural language presents a formidable yet vital challenge within the realms of AI and Robotics [1]. This capability, which involves understanding and executing human language instructions, allows for more intuitive human-robot interactions. The researchers have studied such capability in the context of Language-Conditioned Robotic Grasping (LCRG) [2]–[11], which focuses on robotic systems that ground and grasp objects based on language instructions.

Current approaches in LCRG [2]–[11] predominantly rely on generic language expressions when describing objects for manipulation, leading to less intuitive human-robot communication. For instance, a person might instinctively instruct, "*Get my wallet*". However, with current LCRG systems that operate on generic instructions, one might need to instruct, "*Get the black-colored wallet next to the laptop*". This forces users to craft their instructions to suit the robot's limited knowledge base, what can be termed as *robot-centric* directives. However, many non-expert users find this *robot-centric* directives both unfamiliar and cumbersome. To truly facilitate more intuitive *user-centric* interactions, robots should have more shared personal knowledge with their users beyond the generic, understanding personal objects.

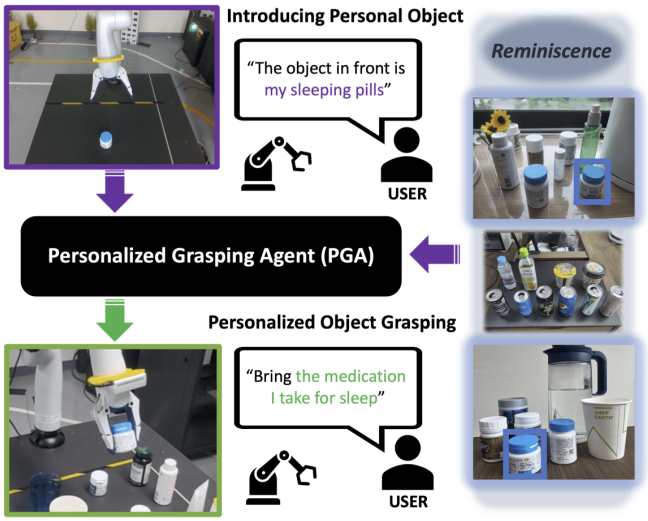

Fig. 1. **Personalizing grasping agents with single human-robot interaction.** Upon the user's introduction of a personal object, it retrieves the identical objects from its visual reminiscence. Leveraging the retrieved objects, the robot subsequently engages in an integrated learning process of the personal object. The Personalized Grasping Agent (PGA) can finally comprehend and grasp the personal object.

To address this problem, we introduce a novel personalized task scenario of LCRG called GRASPMINE with a benchmark dataset. GRASPMINE aims to locate and grasp personal objects given a personal indicator, *e.g., "my sleeping pills"*, that robotic systems with generic knowledge may not handle properly. According to the field of Personalization [12], [13], personalizing with a few examples is a common scenario since collecting a lot of personalized data for each user is infeasible in real-world applications. Considering this aspect, GRASPMINE requires robots to learn personal objects via minimal human-robot interaction, *i.e.*, a one-time verbal introduction of a personal object, as depicted in Fig. 1. In GRASPMINE dataset, we consider learning of 96 personal objects, containing five distinct test splits (Heterogeneous, Homogeneous, Paraphrased, and Cluttered), each sample including an image containing multiple objects, a personal indicator, and the associated object's coordinates.

*Authors have equal contributions
[1]AI Institute, Seoul National University
[2]Interdisciplinary Program in AI, Seoul National University
[3]Interdisciplinary Program in Neuroscience, Seoul National University
[4]Division of Engineering Science, University of Toronto

This research was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) (2021-0-02068-AIHub/10%, 2021-0-01343-GSAI/30%, 2022-0-00951-LBA/20%, 2022-0-00953-PICA/40%) grant funded by the Korean government.

However, with the prevailing learning strategies of current LCRG systems [3]–[11], supervised learning that requires vast amounts of data, it is challenging to tackle GRASPMINE with only a single sample. To address GRASPMINE, we propose **P**ersonalized **G**rasping **A**gent (PGA), that learns personal objects by propagating the information acquired from a user through a *Reminiscence*. Specifically, PGA first constructs the *Reminiscence*—a collection of raw images from the user's environment. Then PGA acquires personal object information in two successive steps: *human-robot interaction* and *robot-object interaction*. In *human-robot interaction*, a user introduces a personal item to the robot with a personal indicator, *e.g.,* "*my sleeping pills*". In *robot-object interaction*, the robot inspects the object from multiple perspectives, obtaining diverse images of personal objects. Leveraging the acquired information of the personal object, PGA pseudo-labels objects in the *Reminiscence* by propagating personal indicators based on the visual features of the objects, inspired from the label propagation algorithm [14]. We term this *Propagation through Reminiscence*. Harnessing the data obtained from the interactions and the pseudo-labeled objects in *Reminiscence*, PGA adapts the object grounding model, and with the personalized model, it adeptly grasps the user-specified personal objects.

In the experiment, we provide baselines for GRASPMINE, including Direct and Supervised methods. Direct employs supervised learning with only one sample per object acquired from *human-robot interaction*, while Supervised represents an upper-bound model trained with nearly 9k annotated samples. In offline experiments focusing on object grounding, PGA outperforms the Direct model by an average absolute increment of approximately 30%, and even demonstrates comparable results to the Supervised method. Notably, we observe performance improvements in offline experiments as the number of images in *Reminiscence* increases. Transitioning to the online setting, we validate PGA's real-world applicability by employing a physical robot to execute GRASPMINE. To qualitatively conclude our analysis, we present visual examples from different phases of PGA, providing insights into its robust performance.

To summarize, our contributions are mainly three-fold:

- We introduce a novel personalized task scenario in LCRG called GRASPMINE, aimed at grounding and grasping personal objects based on personal indicators. This scenario fills a gap in current LCRG systems, which primarily rely on generic expressions, thus enhancing the intuitive nature of human-robot interactions.
- We propose Personalized Grasping Agent (PGA) as a strong baseline for GRASPMINE. PGA learns to ground personal objects by leveraging information acquired from a single *human-robot interaction, robot-object interaction*, and *Propagation through Reminiscence*.
- We provide comprehensive experimental validation of PGA's performance against baselines, showcasing its effectiveness in grounding personal objects. Additionally, we demonstrate the PGA's practical applicability through deployment on a real-world physical robot.

II. RELATED WORK

A. Language-Conditioned Robotic Grasping (LCRG)

Language-Conditioned Robotic Grasping (LCRG) focuses on the robot's capability to grasp objects based on human instructions given in natural language. Numerous studies [3]–[10] have employed fully supervised training approaches to teach robots how to locate objects based on natural language instructions, making use of public datasets [15]–[17]. However, these supervised learning methods reach their limits when robots have to recognize or interact with personal objects unseen during the training. A study [11] curated its own dataset comprising unique objects. Still, it depended on supervised learning, which necessitates exhaustive annotations of language instructions. GVCCI [2] introduced a paradigm shift by unveiling an unsupervised approach to adapt to the user's personal objects. While GVCCI enabled robots to automatically learn the visual nuances of personal objects, it remained constrained. The learning was heavily reliant on generic object categories [18] and attributes inferred by pretrained classifier [19]. This forces non-expert users to use *robot-centric* instructions, a pain point discussed in Sec. I. Some studies [20], [21] delved deeper, augmenting robots' knowledge with semantic or intended meanings of objects, typified by phrases like "*I am thirsty*". While these studies enable robots to grasp user intentions, they also rely on fully-supervised learning that can not be applied to GRASPMINE, necessitating extensive annotations. Unlike previous works, PGA can efficiently learn the user's personal objects and their corresponding language descriptors with just a single user-robot interaction.

B. Personalization

Personalization has become an important factor within various domains of Machine Learning [22]–[29]. Personalization has also been studied in the field of Robotics, such as dressing assistance [30]–[32] and tidy-up task [33]. While these works of Personalization focus on the user's personal preferences, our work focuses on transferring the user's personal knowledge, standing as a pioneering work in the realm of personalized LCRG. Two studies [12], [13] align closely with our research objectives, focusing on learning visual and linguistic representations of personal objects. The work [13] utilizes personal videos paired with transcripts to learn personal objects and their corresponding names, without the need for explicit human labeling. Yet, their focus lies on temporal grounding in videos, overlooking the crucial aspect of spatial grounding. This limitation renders their approach unsuitable for LCRG, where pinpointing the precise object location is imperative. The work [12] broadens this focus to include spatial grounding. However, they assume that fine-grained annotations (*e.g.*, segmentation masks) and personal object names are readily available. In contrast, PGA only requires a single human-robot interaction without any further annotations. So, non-expert users can easily be involved in collecting the training data due to the intuitive and streamlined interface.

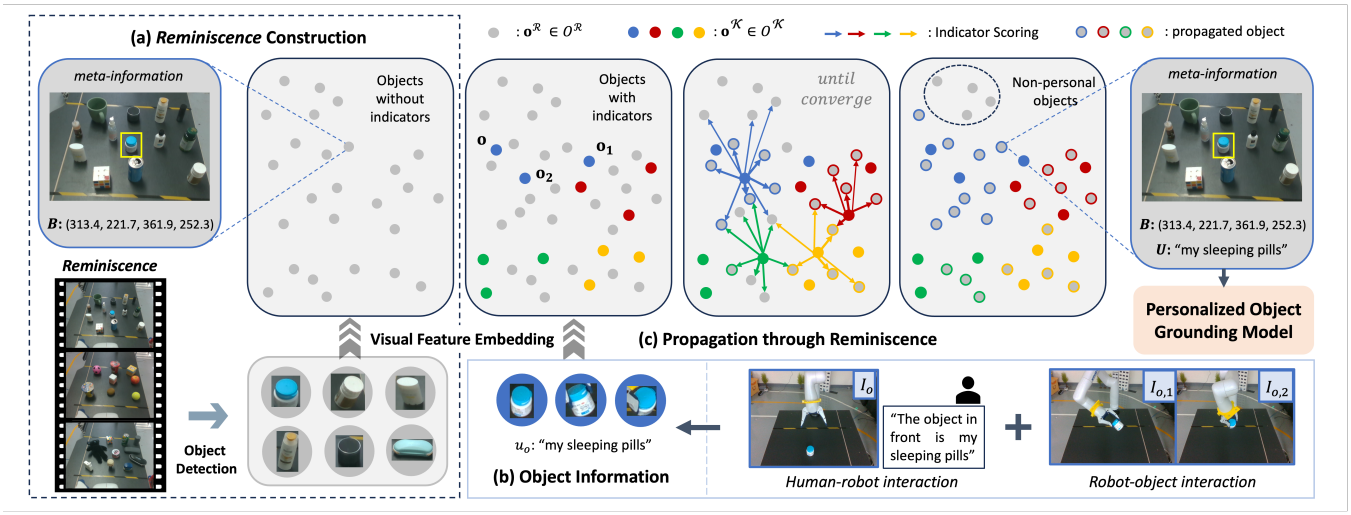

Fig. 2. **Overview of Personalized Grasping Agent (PGA).** (a) Initially, PGA gathers a collection of raw images, termed as *Reminiscence*, from the user's environment. With the personal indicator provided by the user from (b), unlabeled objects in the *Reminiscence* are pseudo-labeled via (c) Propagation through *Reminiscence*. It's vital to note that certain objects, particularly those not introduced by the user (*e.g.*, non-personal objects), remain unlabeled by our algorithm. Ultimately, PGA employs all the object nodes with labels (colored nodes) to train the Personalized Object Grounding Model.

III. METHOD

A. Reminiscence Construction

Personalized Grasping Agent (PGA) initiates by constructing the *Reminiscence*, as illustrated in Fig. 2-(a). PGA first collects a set of N images from the user's environment, represented as $\mathcal{R} = \{I_n^{\mathcal{R}}\}_{n=1}^{N}$, which we term as *Reminiscence*. PGA detects all the objects $\{o_m^{\mathcal{R}}\}_{m=1}^{M}$ from $\mathcal{R}$, via an off-the-shelf object detector [19], and constructs a node for each detected objects, embedding them with the pretrained visual encoder $f(\cdot)$, DINO [34]. These object nodes, represented as a set of vector embeddings $O^{\mathcal{R}} = \{\mathbf{o}_m^{\mathcal{R}}\}_{m=1}^{M} = \{f(o_m^{\mathcal{R}})\}_{m=1}^{M}$ where $M \gg N$, are tagged with the meta-information – an image from $\mathcal{R}$ and the bounding box (bbox) coordinates of the detected object – which will be leveraged in subsequent processes. Note that nodes in $O^{\mathcal{R}}$ are *without* the object's personal indicators.

B. Object Information Acquisition

To acquire the information of personal objects, PGA goes through two successive steps as shown in Fig. 2-(b): *human-robot interaction* and *robot-object interaction*. In *human-robot interaction*, the user introduces their personal object o by displaying the object and verbally describing it. The description consists of two indicators: a general indicator u_o^G, *e.g.*, "*the object in front*", and a personal indicator u_o, *e.g.*, "*my sleeping pills*". Using its visual perception I_o and u_o^G, PGA predicts the bbox coordinates b_o of the object via GVCCI [2]. PGA then constructs an object node $\mathbf{o} = f(o)$ using pretrained visual encoder $f(\cdot)$ [34]. This node is tagged with the meta-information including I_o, b_o, and u_o.

In *robot-object interaction*, PGA subsequently examines the personal object o from multiple views, aiming to capture its distinctive visual appearance. By leveraging b_o, PGA grasps and rotates the object capturing A images $\{I_{o,a}\}_{a=1}^{A}$ for each object, each image offering a distinct view of the object (see the bottom right of Fig. 2). PGA extracts a set of objects $\{o_a\}_{a=1}^{A}$ and their bounding box coordinates $\{b_{o,a}\}_{a=1}^{A}$ from the captured images $\{I_{o,a}\}_{a=1}^{A}$. PGA constructs a set of object nodes $O^{view} = \{\mathbf{o}_a\}_{a=1}^{A} = \{f(o_a)\}_{a=1}^{A}$, where each node $\mathbf{o}_a$ is tagged with the meta-information $I_{o,a}$, $b_{o,a}$, and u_o. Finally, each personal object is associated with $\mathbf{o}$ and O^{view}, *i.e.*, $A+1$ nodes that are tagged with personal indicators. Across all personal objects, the aggregation of nodes with personal indicators forms $O^{\mathcal{K}}$, while collective personal indicators form U, as described in lines 1-6 of Algorithm 1.

C. Propagation through Reminiscence

Propagation through Reminiscence draws inspiration from the semi-supervised method of label propagation [14], [35], as detailed in lines 7-20 of Algorithm 1 and Fig. 2-(c). The goal is to pseudo-label (tag) the missing indicators in $O^{\mathcal{R}}$ by propagating personal indicators from $O^{\mathcal{K}}$. For every node $\mathbf{o}^{\mathcal{R}} \in O^{\mathcal{R}}$, PGA computes an affinity score, $S(\mathbf{o}^{\mathcal{R}}, u)$, for each possible personal indicator $u \in U$. This score is the average cosine similarity ϕ between node $\mathbf{o}^{\mathcal{R}}$ and every known object node $\mathbf{o}^{\mathcal{K}}$ tagged with indicator u;

$$\phi(\mathbf{o}^{\mathcal{K}}, \mathbf{o}^{\mathcal{R}}) = \frac{\mathbf{o}^{\mathcal{K}} \cdot \mathbf{o}^{\mathcal{R}}}{\|\mathbf{o}^{\mathcal{K}}\|_2 \times \|\mathbf{o}^{\mathcal{R}}\|_2} = \frac{f(o^{\mathcal{K}}) \cdot f(o^{\mathcal{R}})}{\|f(o^{\mathcal{K}})\|_2 \times \|f(o^{\mathcal{R}})\|_2}; \quad (1)$$

$$S(\mathbf{o}^{\mathcal{R}}, u) = \frac{1}{N(u)} \sum_{\mathbf{o}^{\mathcal{K}} \in O^{\mathcal{K}}} 1(u_{\mathbf{o}^{\mathcal{K}}} = u) \cdot \phi(\mathbf{o}^{\mathcal{R}}, \mathbf{o}^{\mathcal{K}}); \quad (2)$$

where $N(u)$ denotes the number of $\mathbf{o}^{\mathcal{K}}$s tagged with indicator u and $1(\cdot)$ is the indicator function that takes the value of 1 if the condition is true, and 0 otherwise. Note that $u_{\mathbf{o}^{\mathcal{K}}}, u \in U$. Based on these scores, PGA assigns the indicator u with the highest affinity score to $\mathbf{o}^{\mathcal{R}}$; $u_{o^{\mathcal{R}}} = \text{argmax}\{S(\mathbf{o}^{\mathcal{R}}, u) \mid u \in U\}$; if the highest score is above the certain threshold. After pseudo-labeling $\mathbf{o}^{\mathcal{R}}$ with the most likely personal indicator, $O^{\mathcal{K}}$ is updated to incorporate the propagated node $\mathbf{o}^{\mathcal{R}}$, and label propagation repeats until the percentage of relabeled

nodes is less than 10%. *Propagation through Reminiscence* allows PGA to automatically gather pseudo-labeled examples, alleviating the scarcity of labeled data in GraspMine.

Algorithm 1 Personalized Grasping Agent (PGA)

Require: The set of user's personalized objects O
Require: Visual Encoder $f(\cdot)$ [34]
Require: Nodes $O^{\mathcal{R}} \leftarrow \{\mathbf{o}_m^{\mathcal{R}}\}_{m=1}^{M}$ from $\mathcal{R}$ ▷ (Sec. III-A)
Require: Set of personal indicators $U = \emptyset$
Require: Nodes with indicators $O^{\mathcal{K}} = \emptyset$
for $o \leftarrow O$ **do** ▷ (Sec. III-B)
  $\mathbf{o} \leftarrow f(o)$ and u_o from *human-robot interaction*
  Get $\{o_a\}_{a=1}^{A}$ from *robot-object interaction*
  $O^{view} \leftarrow \{f(o_a)\}_{a=1}^{A}$ tagged with u_o
  $O^{\mathcal{K}} \leftarrow O^{\mathcal{K}} \cup \{\mathbf{o}\} \cup O^{view}$ and $U \leftarrow U \cup \{u_o\}$
end for
repeat ▷ (Sec. III-C)
  for $\mathbf{o}^{\mathcal{R}} \leftarrow O^{\mathcal{R}}$ **do**
    $\forall u \in U$, set $S_u = 0$, $N_u = 0$
    for $\mathbf{o}^{\mathcal{K}} \leftarrow O^{\mathcal{K}}$ **do**
      $u \leftarrow u_{\mathbf{o}^{\mathcal{K}}}$
      $S_u \leftarrow S_u + \phi(\mathbf{o}^{\mathcal{K}}, \mathbf{o}^{\mathcal{R}})$ and $N_u \leftarrow N_u + 1$
    end for
    $\forall u \in U$, $S(\mathbf{o}^{\mathcal{R}}, u) \triangleq S_u / N_u$
    if max $\{S(\mathbf{o}^{\mathcal{R}}, u) \mid u \in U\} >$ threshold **then**
      $u_{o^{\mathcal{R}}} \leftarrow \text{argmax}\{S(\mathbf{o}^{\mathcal{R}}, u) \mid u \in U\}$
      $O^{\mathcal{K}} \leftarrow O^{\mathcal{K}} \cup \{\mathbf{o}^{\mathcal{R}}\}$
    end if
  end for
until Ratio of changed indicator of node < 10%

D. Personalized Object Grounding Model

The *Personalized Object Grounding Model* is a Transformer [36] based Vision-Language model that takes an image and a natural language indicator to infer the bounding box coordinates of the personal object. The model's optimization occurs by minimizing

$$\mathcal{L} = -\sum_{\mathbf{o}^{\mathcal{K}} \in O^{\mathcal{K}}} log P_\theta(b_{\mathbf{o}^{\mathcal{K}}} | I_{\mathbf{o}^{\mathcal{K}}}, u_{\mathbf{o}^{\mathcal{K}}}), \tag{3}$$

the negative log-likelihood of the bounding box for every object node $\mathbf{o}^{\mathcal{K}} \in O^{\mathcal{K}}$. Note that $O^{\mathcal{K}}$ is not from Sec. III-B but after *Propagation through Reminiscence* (Sec. III-C).

E. Personalized Object Grasping

The process of *Personalized Object Grasping* is initiated when a user instructs the robot to grasp a personal object as seen in Fig. 1. To execute this, PGA first infers the 2D coordinates of the queried object using the *Personalized Object Grounding Model*. Upon obtaining these 2D bounding box coordinates, PGA calculates the 3D segmented object coordinates by leveraging point cloud data and the RANSAC [37] algorithm. Specifically, it translates the 2D bounding box into a 3D spatial configuration using the point cloud, followed by segmenting the points of the object within the 3D bounding box via RANSAC. Lastly, PGA grasps the object by computing a trajectory of the robot arm [38].

IV. GraspMine

Our proposed task scenario, GraspMine, consists of a curated dataset comprising a Training set, *Reminiscence*, and Test set, featuring 96 personal objects along with 100+ everyday objects, totaling around 200 individual objects.

Training set consists of 96 pairs of images I_o and their respective personal indicators u_o, alongside general indicators u_o^G, collected through *human-robot interaction* described in Sec. III-B and illustrated in Fig. 2-(b).

Reminiscence, $\mathcal{R} = \{I_n^{\mathcal{R}}\}_{n=1}^{N}$, involves $N = 400$ raw images, each containing multiple objects. These unlabeled images can be optionally provided to aid the learning process. However, we annotated objects in the *Reminiscence* for analyzing the propagation ability of models and for training the Supervised method in Sec. V-B. Notably, these annotations were not utilized in any aspect of training PGA.

Test set. Examples of a Test set is presented in Fig.4-A~E. Given an image and a personal indicator, the agent is tasked with inferring the correct location (depicted by the black box). For the deeper analysis, Test set is categorized into five distinct splits as follows.

- **Heterogeneous** split incorporates scenes with randomly selected objects as exemplified in Fig.4-A. It consists of 60 images with 120 personal indicators and bboxes.
- **Homogeneous** split incorporates scenes with similar-looking objects with the same category, making discrimination more challenging as exemplified in Fig.4-B,C. It consists of 60 images with 120 indicators and bboxes.
- **Cluttered** split contains 106 images containing highly cluttered objects as exemplified in Fig.4-D,E, and a single personal indicator and bbox per image. Images of this split originate from the IM-Dial dataset [21].
- **Paraphrased** split comprises all Heterogeneous, Homogeneous, and Cluttered split with each personal indicator paraphrased by the annotators. For instance, the personal indicator "*my sleeping pills*" is queried as "*the medication I take for sleep*" as depicted in Fig. 1.
- **Generic** split is sourced from the *VGPI* dataset presented in GVCCI [2]. This dataset features images comprising generic objects, each annotated with indicators based on their basic-level object categories. Within the *VGPI* dataset, we specifically utilized the Test-E split, which aligns with our environmental settings.

V. EXPERIMENTS

A. Compared Methods

OFA [39] is a state-of-the-art visual grounding model pretrained using a public dataset [15]. This model has never been exposed to the user's personal objects, positioning it as the most basic compared method. It primarily relies on generic cues (*e.g.*, "*tennis ball*") present in personal indicators (*e.g.*, "*my Djokovic tennis ball*") to make predictions.

GVCCI [2] is a robotic lifelong learning framework designed to autonomously learn generic expressions of objects.

TABLE I

Results of offline experiments. Results show the personal object grounding score in GRASPMINE dataset. Bold scores represent the highest performance, while underlined scores the second-best results. The 'interaction' column represents the average number of interactions per object that users had with the robot to teach personal objects. 'annotated' indicates the total human annotations used in each method, while 'utilized' specifies the number of triplets - image, object bounding box, and personal indicator - employed for training the models. It's noteworthy that subtracting the 'annotated' count from the 'utilized' count reveals the amount of triplet automatically generated by each method without any human labor.

				Heterogeneous		Homogeneous		Paraphrased		Cluttered	Generic
Method	interaction	annotated	utilized	$IoU_{>0.5}$	$IoU_{>0.8}$	$IoU_{>0.5}$	$IoU_{>0.8}$	$IoU_{>0.5}$	$IoU_{>0.8}$	$IoU_{>0.8}$	$IoU_{>0.5}$
OFA [39]	-	-	-	49.2	47.5	23.7	20.3	35.8	34.1	34.6	65.7
GVCCI [2]	-	-	-	59.3	55.1	30.5	23.8	42.3	38.9	44.3	<u>79.1</u>
Direct	1.1	96	96	60.2	55.1	37.3	27.1	48.3	42.5	46.4	-
PassivePGA	1.1	96	4,828	89.0	76.3	68.6	53.4	73.6	62.7	62.4	-
PGA (ours)	1.1	96	6,492	<u>91.5</u>	<u>81.4</u>	<u>70.3</u>	<u>61.9</u>	<u>74.7</u>	<u>68.2</u>	<u>64.5</u>	<u>79.1</u>
Supervised	91.3	8,763	8,763	**97.5**	**90.7**	**92.4**	**83.1**	**84.8**	**79.0**	**71.4**	-

By leveraging images from both the training set and the *Reminiscence*, it automatically generates these generic expressions and utilizes the generated data for training.

Direct model is trained directly from the *human-robot interaction* (Sec. III-B). This model does not employ *robot-object interaction* (Sec. III-B) and the *Propagation through Reminiscence* (Sec. III-C), resulting in training the model with a single sample per personal object. This model serves as a representation of traditional LCRG systems that typically learn in a supervised manner when confronted with the task scenario of GRASPMINE. Consequently, the Direct model can only utilize 96 annotated data obtained from an average of 1.1 *human-robot interaction* per object (101 interactions for 96 objects due to five visual grounding failures).

PassivePGA model is an ablative model of PGA, excluding *robot-object interaction* in Sec. III-B. In other words, *Propagation through Reminiscence* (Sec. III-C) is conducted only with the nodes from *human-robot interaction* in Sec. III-B. Consequently, this model utilizes 96 annotated data points along with 4.7k pseudo-labeled data obtained from the *Propagation through Reminiscence*, resulting in a total of 4.8k utilized data points.

Supervised model is trained with a fully-supervised approach that utilizes ground-truth annotations (*i.e.*, object location coordinates and personal indicators) in *Reminiscence*. Since this model has access to all 8.7k ground truth annotations in the *Reminiscence*, whereas GRASPMINE assumes access of only one annotation per object (96 annotations in total), the Supervised model serves as the upper bound performance benchmark for GRASPMINE in the following experiments. Essentially, training this model requires approximately 100 times more manual effort compared to PGA.

B. Offline Experiment

Evaluation Protocol. In the offline experiment, we study the PGA's proficiency in grounding the target object given a personal indicator. By following standard practice [2], [39], we assess the grounding accuracy using the Intersection over Union (IoU) score, a key visual grounding metric, which calculates the overlap between the predicted and ground truth bounding boxes. While we present the percentage of predictions surpassing an IoU of 0.5, we also emphasize IoU above 0.8 to ensure a stricter alignment between the predicted and actual regions, acknowledging the precision needed for successful object grasping. For the Cluttered split, we only reported scores with an IoU exceeding 0.8 since boxes with an IoU above 0.5 frequently included multiple objects, rendering the score less meaningful.

Comparison with baselines. The offline results, presented in Tab. I, demonstrate that the Direct model, which relies solely on object information acquired from *human-robot interaction* without leveraging *Reminiscence*, exhibited only marginal improvement compared to methods reliant on generic object knowledge, *i.e.*, OFA and GVCCI. This indicates that naively training existing LCRG models on a small amount of labeled data does not make significant performance gains in GRASPMINE. However, PGA, utilizing a large amount of unlabeled data, demonstrated a significant improvement compared to the Direct model, achieving approximately 30% enhancement. Remarkably, even when compared to the Supervised method, which requires about 100 times more annotations than PGA, PGA exhibited comparable results. We also assess PGA in grounding objects when queried with generic instructions, using the Generic split, to examine how well PGA retains knowledge about the generic instructions after training the personal indicators. Even after acquiring personal knowledge, PGA's ability to ground objects through generic instructions remains impeccably intact. In short, by utilizing an *Reminiscence* and the robot's manipulative power, PGA proficiently grounds personal objects with learning from a single human-robot interaction, outperforming the baseline methods and showing comparable performance to the Supervised approach, all while retaining its knowledge of generic expressions.

Impact of *robot-object interaction*. Tab. I shows a notable enhancement of up to 8.5% in PGA over the PassivePGA model. This underscores the impact of *robot-object interaction* in improving performance by capturing objects from different views. Notably, when comparing the $IoU_{>0.5}$ and $IoU_{>0.8}$ across various splits in Tab. I, we found that the improvement in $IoU_{>0.8}$ (+5-8%) exceeded that of $IoU_{>0.5}$ (+1-2%). This suggests that leveraging *robot-object interaction*

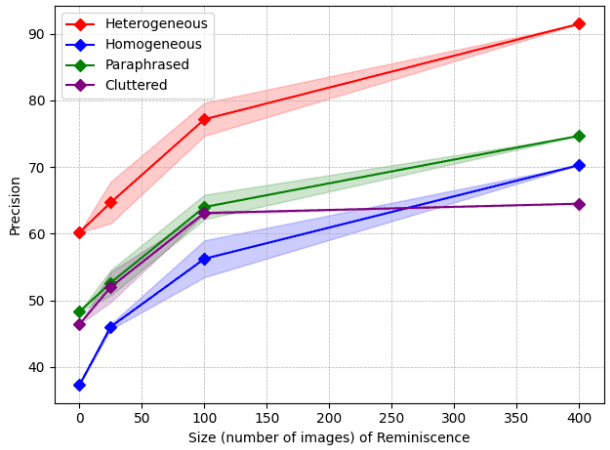

Fig. 3. **Impact of *Reminiscence* size.** PGA's offline scores based on the number of raw images utilized from *Reminiscence*, from '0' to '400'.

leads to more precise localization of objects, particularly beneficial for robotic grasping. To further understand these findings, we compare the propagation accuracy of PassivePGA and PGA. The accuracy, measured as the percentage of correctly pseudo-labeled objects among total nodes, revealed PGA achieving 79.4%, while PassivePGA lagged behind at 64.5%. We further investigate incorrectly labeled objects and observe that most of the propagation error comes from the noisiness of the off-the-shelf object detector [19]. Noisy objects were categorized into two groups: 1) ambiguous boxes, *i.e.*, boxes that only include a part of an object (*e.g.*, Fig. 4-(a)) or boxes that surround multiple objects, and 2) invalid boxes, *i.e.*, non-objects such as line stickers on desks or a robot arm (*e.g.*, Fig. 4-B). We found 393 ambiguous boxes and 870 invalid boxes in the *Reminiscence*. Regarding propagation to ambiguous boxes, PGA showed a reduced rate of 41.2% compared with that of 50.4% from the PassivePGA model. Moreover, PassivePGA model had a high tendency to propagate towards invalid boxes, at 83.7%, while PGA showed a nearly negligible rate at 0.07%. In summary, the impact of *robot-object interaction* on the inaccurate and noisy *Propagation through Reminiscence* appears to have influenced both the overall grounding accuracy and the precision of localization.

Impact of *Reminiscence* size. As shown in Fig. 3, our exploration delves into the relationship between PGA's performance and the size of the *Reminiscence* — the number of raw images utilized. A size of '0' is equivalent to the Direct model. We experimented on a log scale, examining sizes of 25, 100, and 400, with each configuration undergoing three separate trials. Our findings suggest a clear pattern: as PGA is exposed to an increasing number of scenes from the user's environment, its performance is incrementally enhanced. This indicates that, as the robot collects more raw images from the environment, its performance is expected to be improved without necessitating further human supervision.

TABLE II

Results of Real-world Online Experiment (LCRG) in GRASPMINE.

	Heterogeneous		Homogeneous	
Method	grounding	grasping	grounding	grasping
Direct	63.3	53.3	26.7	20.0
PGA (ours)	**90.0**	**76.7**	**66.7**	**53.3**

C. Online Experiment

Robotic Platform. Our robotic platform is equipped with the 6-DoF Kinova Gen3 Lite[1], complemented by the Intel Realsense Depth Camera D435. While all learning and inferences of PGA are managed on the remote server, manipulation planning is computed on the local platform.

Evaluation Protocol. From the Heterogeneous and Homogeneous splits, we randomly selected 30 images each and closely reproduced the depicted scenes. We assessed LCRG based on two human-evaluated metrics: object grounding success rate and grasping accuracy, *i.e.*, LCRG.

Results. Results in Tab. II reveal that PGA significantly surpassed the Direct model in grasping, along with enhanced grounding accuracy. The Direct model, which solely relies on object information acquired from *human-robot interaction* without leveraging raw images from *Reminiscence*, lagged behind PGA with a performance gap of 23.4% in the Heterogeneous split and 33.3% in the Homogeneous split. This result underscores the effectiveness of PGA, which achieves superior performance without requiring additional human labor for annotations compared to the Direct model.

D. Qualitative Analysis

Through qualitative analysis, shown in Fig. 4, we visualize examples from the various phases including *Object Information Acquisition*, *Propagation through Reminiscence*, and *Personalized Object Grounding*. For a comprehensive understanding, we compare the performance of PGA against the PassivePGA model.

One striking observation is PGA's adeptness at inspecting objects from diverse angles as seen in '*robot-object interaction*'. This allows PGA to recognize variations in object orientation, *e.g.*, views from the rear or the base. In instances such as those represented in (c), (d), (e), and (f), the PassivePGA model often overlooks objects in the *Reminiscence* that differ from their appearances in the *human-robot interaction* phase. We conjecture that the omitted samples, *e.g.*, (d), (e), and (f), in the training phase affect some wrong prediction results as in C and D. For example, consider the case of "the charger for my brand new MacBook". With the aid of *robot-object interaction*, PGA successfully propagates to samples viewed from different angles. However, PassivePGA fails to capture such samples as shown in (d) and (e). In sample D, it is evident that PassivePGA makes an incorrect inference, while PGA's inference is accurate. Moreover, as detailed in Sec. V-B, the PassivePGA model

[1]https://www.kinovarobotics.com/product/gen3-lite-robots

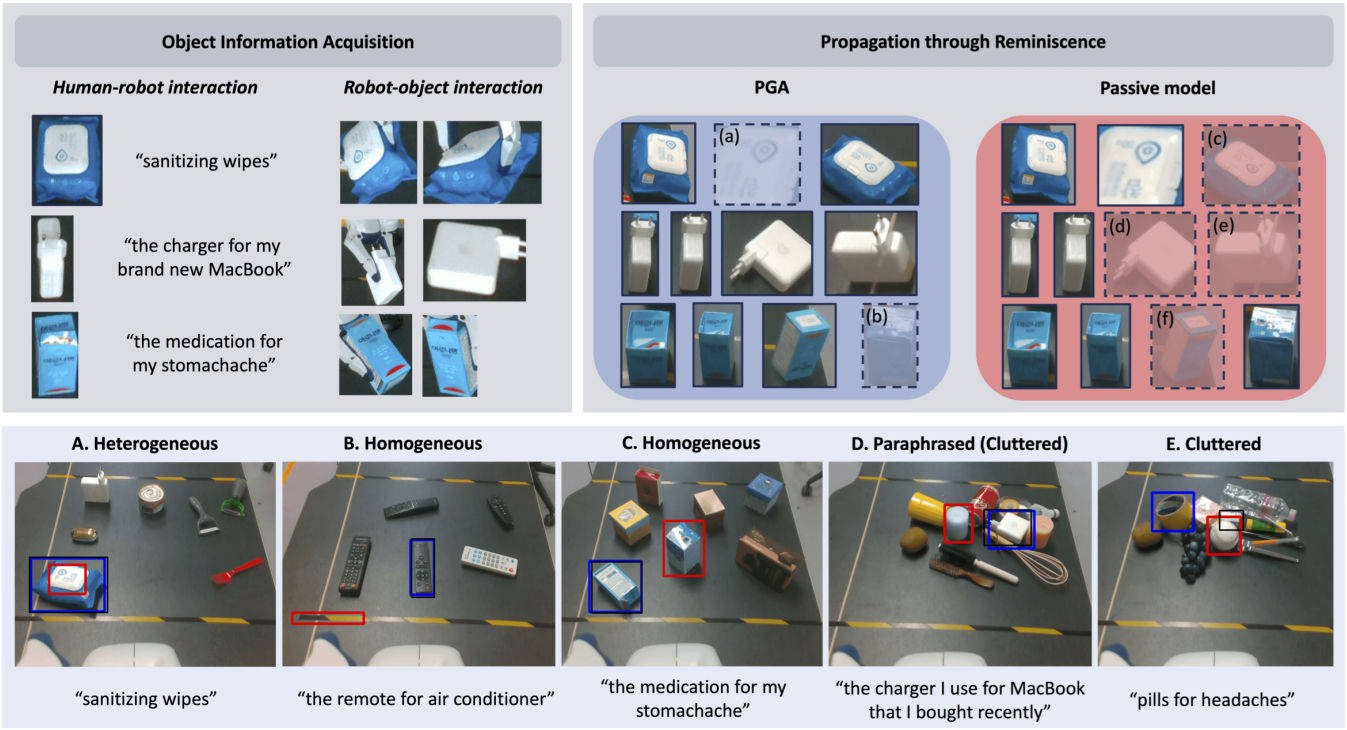

Fig. 4. **Qualitative Analysis.** The top row showcases examples of objects from each phase. In the bottom row, personalized object grounding results by PGA are depicted in blue boxes and those by PassivePGA model in red, set alongside the ground truth in black boxes. Within the *Propagation through Reminiscence* on the top row, solid lined boxes are the pseudo-labeled objects from the *Reminiscence* and dotted boxes denote objects that were NOT pseudo-labeled according to indicators in the respective models.

occasionally tends to propagate towards ambiguous boxes in the *Reminiscence*, as depicted in (a). Such behavior can lead to inaccurate grounding, as observed in the red box in sample A. The tendency of propagating towards incorrect boxes might have affected in some results shown in panel B.

Despite PGA's robust grounding abilities as seen through A-D, challenges persist in some complex scenarios, such as in a scenario like E, where objects are situated in highly cluttered environments—the difficulty reflected in the lower scores of Cluttered split in Tab. I. Nevertheless, our qualitative analysis reveals PGA's successful *Propagation through Reminiscence*, enhanced by the *robot-object interaction*, yielding promising results in Personalized Object Grounding.

VI. DISCUSSION

The Personalized Grasping Agent (PGA) has shown promising performance in the GraspMine dataset. However, it is essential to discuss several assumptions and limitations in our experimental setup to guide future studies for better improvement in dealing with GraspMine.

Due to the use of an arm robot without mobility necessitates two primary assumptions. Firstly, we confine the execution area to a 80*60 *cm* table where all objects are at least partially visible within this space. This assumption limits manipulation to flat and open surfaces, excluding scenarios such as retrieving objects from handbags or shelves. Overcoming this limitation would require developing deeper and dynamically adaptive reasoning in motion planning, a direction we leave for future exploration. Secondly, we assume a full access to raw images of *Reminiscence*. While obtaining these raw images from a mobile robot through random navigation in the user's environment is feasible, we assumed that these images were pre-acquired. Extending this work with a mobile robot capable of autonomously acquiring raw images from the user's environment would be a crucial topic for further studies.

During the online experiment with the physical robot (See Tab. II), a performance drop from grounding accuracy to grasping accuracy (around 13%) was observed. This decline can be attributed to limitations in motion planning, particularly in the manipulation strategy employed, which first aligns the object and the end effector's horizontal coordinates before making vertical adjustments. However, this approach may prove ineffective when attempting to grasp objects like wine bottles. Enhancing motion planning to tailor grasping strategies for diverse objects is crucial for future studies.

VII. CONCLUSION

This paper introduces a challenging task scenario GraspMine, along with a novel robotic framework PGA to equip robots with the capacity to learn a user's personalized knowledge through a single human-robot interaction. By expanding the mutual knowledge between a robot and its user beyond the generic, our approach promotes a shift from *robot-centric* to *user-centric* interactions. We anticipate that our method will foster more natural, intuitive interaction and significantly enhance the non-expert user's experience.

Acknowledgements. We extend our gratitude to Jungmin Lee for her invaluable contribution to video editing, and all the reviewers for their insightful comments and feedback.

References

[1] S. Tellex, N. Gopalan, H. Kress-Gazit, and C. Matuszek, "Robots that use language," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, pp. 25–55, 2020.

[2] J. Kim, G.-C. Kang, J. Kim, S. Shin, and B.-T. Zhang, "Gvcci: Lifelong learning of visual grounding for language-guided robotic manipulation," *arXiv preprint arXiv:2307.05963*, 2023.

[3] S. G. Venkatesh, A. Biswas, R. Upadrashta, V. Srinivasan, P. Talukdar, and B. Amrutur, "Spatial reasoning from natural language instructions for robot manipulation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 196–11 202.

[4] H. Zhang, Y. Lu, C. Yu, D. Hsu, X. La, and N. Zheng, "Invigorate: Interactive visual grounding and grasping in clutter," *arXiv preprint arXiv:2108.11092*, 2021.

[5] M. Shridhar, D. Mittal, and D. Hsu, "Ingress: Interactive visual grounding of referring expressions," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 217–232, 2020.

[6] O. Mees, A. Emek, J. Vertens, and W. Burgard, "Learning object placements for relational instructions by hallucinating scene representations," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 94–100.

[7] O. Mees and W. Burgard, "Composing pick-and-place tasks by grounding language," in *Experimental Robotics: The 17th International Symposium*. Springer, 2021, pp. 491–501.

[8] Y. Wang, K. Wang, Y. Wang, D. Guo, H. Liu, and F. Sun, "Audio-visual grounding referring expression for robotic manipulation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 9258–9264.

[9] M. Shridhar and D. Hsu, "Interactive visual grounding of referring expressions for human-robot interaction," *arXiv preprint arXiv:1806.03831*, 2018.

[10] D. Whitney, E. Rosen, J. MacGlashan, L. L. Wong, and S. Tellex, "Reducing errors in object-fetching interactions through social feedback," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1006–1013.

[11] J. Hatori, Y. Kikuchi, S. Kobayashi, K. Takahashi, Y. Tsuboi, Y. Unno, W. Ko, and J. Tan, "Interactively picking real-world objects with unconstrained spoken language instructions," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3774–3781.

[12] N. Cohen, R. Gal, E. A. Meirom, G. Chechik, and Y. Atzmon, ""this is my unicorn, fluffy": Personalizing frozen vision-language representations," in *European Conference on Computer Vision*. Springer, 2022, pp. 558–577.

[13] C.-H. Yeh, B. Russell, J. Sivic, F. C. Heilbron, and S. Jenni, "Meta-personalizing vision-language models to find named instances in video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 123–19 132.

[14] X. Zhu and Z. Ghahramani, "Learning from labeled and unlabeled data with label propagation," *ProQuest Number: INFORMATION TO ALL USERS*, 2002.

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

[16] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 11–20.

[17] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[18] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem, "Basic objects in natural categories," *Cognitive psychology*, vol. 8, no. 3, pp. 382–439, 1976.

[19] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[20] T. Nguyen, N. Gopalan, R. Patel, M. Corsaro, E. Pavlick, and S. Tellex, "Robot object retrieval with contextual natural language queries," *arXiv preprint arXiv:2006.13253*, 2020.

[21] G.-C. Kang, J. Kim, J. Kim, and B.-T. Zhang, "Prograsp: Pragmatic human-robot communication for object grasping," *arXiv preprint arXiv:2309.07759*, 2023.

[22] C. Chunseong Park, B. Kim, and G. Kim, "Attend to you: Personalized image captioning with context sequence memory networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 895–903.

[23] E. Denton, J. Weston, M. Paluri, L. Bourdev, and R. Fergus, "User conditional hashtag prediction for images," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 1731–1740.

[24] C. Long, X. Yang, and C. Xu, "Cross-domain personalized image captioning," *Multimedia Tools and Applications*, vol. 79, pp. 33 333–33 348, 2020.

[25] X. Jia, H. Zhao, Z. Lin, A. Kale, and V. Kumar, "Personalized image retrieval with sparse graph representation learning," in *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020, pp. 2735–2743.

[26] W. Zhuo, Z. He, M. Zheng, B. Hu, and R. Wang, "Research on personalized image retrieval technology of video stream big data management model," *Multimedia Tools and Applications*, pp. 1–18, 2021.

[27] Y. Zhang, C.-B. Zhang, P.-T. Jiang, M.-M. Cheng, and F. Mao, "Personalized image semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 549–10 559.

[28] F. Amat, A. Chandrashekar, T. Jebara, and J. Basilico, "Artwork personalization at netflix," in *Proceedings of the 12th ACM conference on recommender systems*, 2018, pp. 487–488.

[29] A. Yan, Z. He, J. Li, T. Zhang, and J. McAuley, "Personalized showcases: Generating multi-modal explanations for recommendations," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 2251–2255.

[30] A. Jevtić, A. F. Valle, G. Alenyà, G. Chance, P. Caleb-Solly, S. Dogramadzi, and C. Torras, "Personalized robot assistant for support in dressing," *IEEE transactions on cognitive and developmental systems*, vol. 11, no. 3, pp. 363–374, 2018.

[31] Y. Gao, H. J. Chang, and Y. Demiris, "User modelling for personalised dressing assistance by humanoid robots," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 1840–1845.

[32] S. D. Klee, B. Q. Ferreira, R. Silva, J. P. Costeira, F. S. Melo, and M. Veloso, "Personalized assistance for dressing users," in *Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7*. Springer, 2015, pp. 359–369.

[33] J. Wu, R. Antonova, A. Kan, M. Lepert, A. Zeng, S. Song, J. Bohg, S. Rusinkiewicz, and T. Funkhouser, "Tidybot: Personalized robot assistance with large language models," *arXiv preprint arXiv:2305.05658*, 2023.

[34] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[35] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Label propagation for deep semi-supervised learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5070–5079.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[37] R. Wahl, M. Guthe, and R. Klein, "Identifying planes in point-clouds for efficient hybrid rendering," in *The 13th Pacific Conference on Computer Graphics and Applications*, vol. 3, 2005.

[38] D. Coleman, I. Sucan, S. Chitta, and N. Correll, "Reducing the barrier to entry of complex robotic software: a moveit! case study," *arXiv preprint arXiv:1404.3785*, 2014.

[39] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *International Conference on Machine Learning*. PMLR, 2022, pp. 23 318–23 340.