Vision Beyond Boundaries: An Initial Design Space of Domain-specific Large Vision Models in Human-robot Interaction

Yuchong Zhang
yuchongz@kth.se
KTH Royal Institute of Technology
Stockholm, Sweden

Yong Ma
University of Bergen
Bergen, Norway
Yong.Ma@uib.no

Danica Kragic
KTH Royal Institute of Technology
Stockholm, Sweden

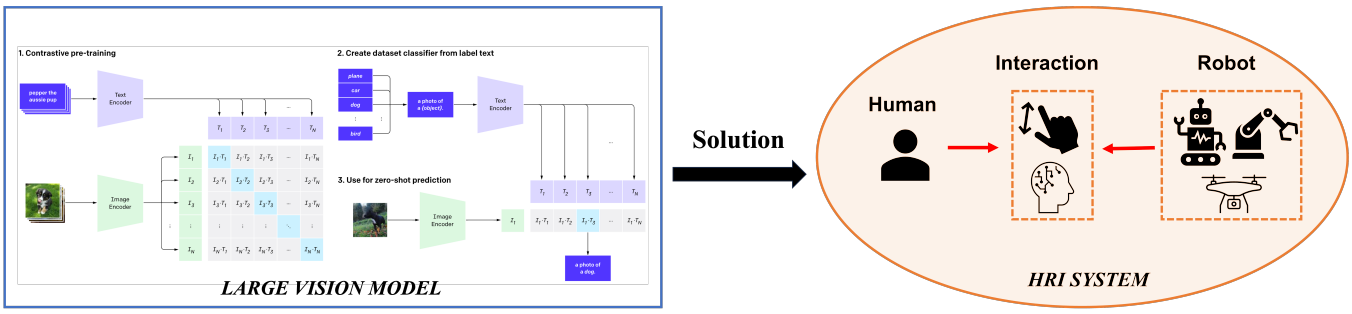

Figure 1: Large vision models (the example is OpenAI CLIP model [38]) for human-robot interaction systems. It is foreseen that the superiority of LVMs will enhance the robustness and performance of HRI systems, paving the way for more efficient and intuitive interactions between humans and robots compared to conventional computer vision models.

ABSTRACT

The emergence of Large Vision Models (LVMs) is following in the footsteps of the recent prosperity of Large Language Models (LLMs) in following years. However, there's a noticeable gap in structured research applying LVMs to Human-Robot Interaction (HRI), despite extensive evidence supporting the efficacy of vision models in enhancing interactions between humans and robots. Recognizing the vast and anticipated potential, we introduce an initial design space that incorporates domain-specific LVMs, chosen for their superior performance over normal models. We delve into three primary dimensions: HRI contexts, vision-based tasks, and specific domains. The empirical validation was implemented among 15 experts across six evaluated metrics, showcasing the primary efficacy in relevant decision-making scenarios. We explore the process of ideation and potential application scenarios, envisioning this design space as a foundational guideline for future HRI system design, emphasizing accurate domain alignment and model selection.

CCS CONCEPTS

• **Computing methodologies** → **Vision for robotics**; **Computer vision**; • **Human-centered computing** → **HCI theory, concepts and models**.

Conference acronym 'XX, June 03–05, 2024, Woodstock, NY

ACM ISBN XXX-X-XXXX-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

KEYWORDS

domain-specific, large vision models, human-robot interaction, empirical study

ACM Reference Format:
Yuchong Zhang, Yong Ma, and Danica Kragic. 2024. Vision Beyond Boundaries: An Initial Design Space of Domain-specific Large Vision Models in Human-robot Interaction. In *Woodstock '24: ACM Symposium on Neural Gaze Detection, June 03–05, 2024, Woodstock, NY*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

With the astonishing development of robotics and human-computer interaction (HCI), an extraordinary progress has been shown in the field of human–robot interaction (HRI) during past years, which is now benefiting the everyday life. From robots helping with household tasks to humans and robots working together in production lines, it has a significant impact on human society. The key to research in this area is understanding the nature of interactivity and social behavior between robots and humans [6], which aims at endowing robots with the ability to estimate the intent of humans and to complete tasks more effectively that meet the requirements of users. Drawing from existing literature [25, 34, 44], the conceptualization of HRI has undergone iterative refinement and predominantly encompasses several core aspects: the supervisory control of robots by humans, the cooperative and collaborative engagement between humans and robots in executing domain-specific tasks, and the facilitation of social objectives through human-robot interactions, etc. Among them, visual data is an imperative component which has been widely used in the design of intuitive HRI contexts, such as gestural interaction [10], object segmentation [23], and video tracking [32].

Over the past decade, the field of computer vision has undergone a remarkable evolution, paralleling the rapid advancements in deep learning technologies. Leveraging the satisfying performance of deep neural networks, vision models have achieved significant breakthroughs in a plethora of context-specific tasks [51]. These tasks range from object detection and image segmentation to the more complex realm of scene reconstruction. In more recent years, the sphere of computer vision has further expanded, witnessing notable progress in areas such as visual tracking, video captioning, pose estimation, and innovative content generation. Particularly noteworthy is the emergence and application of generative artificial intelligence (GenAI) across various domains, where it has demonstrated impressive results. At the forefront of GenAI technology are large language models (LLMs), exemplified by OpenAI's ChatGPT, which has garnered considerable attention for its proficiency in handling text-based data. Following the trail blazed by LLMs, large vision models (LVMs) have gradually emerged as a parallel strand of large-scale foundational models [39]. These models are increasingly gaining traction, playing a pivotal role in shaping the future of vision-based analysis and interpretation. LVMs represent a significant leap forward in the field, offering novel perspectives and capabilities in the processing and understanding of visual data, thereby setting new benchmarks in the landscape of AI and machine learning.

The emergence of LVMs has opened the potential to address long-standing challenges in HRI through their advanced capabilities in visual perception and interpretation [51]. Initially developed for image-based tasks such as object recognition and scene understanding, these models are now being applied in robotics. Equipped with advanced visual processing, they are capable of interpreting complex visual scenes and making informed decisions, as well as processing multimodal input alongside visual data. The techniques applied in developing these models are essential for further progress, and may improving robustness and efficiency. Similar to the transformative impact of LLMs on text-oriented applications [55], LVMs are spearheading a revolution in the field of visual data processing. However, a notable distinction arises in their applicability: while LLMs trained on internet text are generally effective across a wide range of scenarios where the language used keeps identical, the imagery used by most formal companies significantly differs from the typical visuals found on social platforms. This discrepancy underscores the necessity for LVMs that are tailored to specific domains. Customized domain-specific LVMs have shown superior efficacy in comprehending the unique and nuanced visual content relevant to particular contexts, excelling in tasks such as visual classification, object detection, and segmentation which are pervasively utilized in recent HRI architectures.

The landscape of AI, and more specifically computer vision, has undergone significant transformation in recent years, largely due to the advent of vision transformers and large, general-purpose vision models. Several LVMs have paved the first step of visual data processing revolution, such as OpenAI's CLIP (Contrastive Language-Image Pretraining) [38], Landing AI's LandingLens [1], Google's ViT [8], and SWIN Transformer [28]. For domain-specific LVMs that are particularly focused on specialized enterprise applications, the GPT-4V [2] developed by OpenAI and SAM (Segment Anything Model) [24] by Meta have pioneered the progression of training extensive datasets of images, empowering businesses to customize them for a variety of applications within their specific domains. Beyond improved performance, domain-specific LVMs offer several advantages compared to normal LVMs and commonly-used vision models [3]: reduced training costs attributable to smaller databases, enhanced accuracy within specific domains, and the unprecedented scale via extensive parameters for fine-tuning.

According to Landing AI [4], the LVM revolution is trailing LLM by two or three years, while domain-specific LVMs outperform others. Thus, we foresee that new advances brought by those models are being made regularly, leading to the development of more intelligent HRI systems and allowing robots to be more seamlessly integrated into human society (Figure 1). To our knowledge, this is the first work to explore the potential of LVMs in human–robot interaction, and our contributions can be summarized as follows: Developing an initial design space with empirical evaluation that helps to explore the possible usage of domain-specific LVMs in future HRI systems.

2 RELATED WORK

Our work is established upon the intersection between the design space theory of HCI and the current status of LVMs in HRI.

2.1 Design Space with HRI

In HCI, a design space serves as a tool that highlights the diverse possibilities in crafting a specific type of artifact, thereby enriching and facilitating the design process. It also evolves iteratively, incorporating new design potentials into its existing dimensions based on insights gained from each design iteration [29]. Many research on the exploratory design spaces of various components in HRI have been accomplished during the past decades. In 2002, Dautenhahn investigated the design space of social robots that can be believed by humans, underpinned by finding the niche spaces [5]. Woods et al. developed a design space in identifying children's perspectives on distinguishing robots, pointing out that the two dimensions labelled 'Behavioural Intention' and 'Emotional Expression' included are the decisive factors [52]. An original empirical framework was proposed by Walters to probe social robots' appearance and behaviour, which suggested that humans favour the cases when the behaviour is consistent with the robot's appearance [50]. Kalegina et al. conducted a survey study and characterized a design space specifying the preferences of human perception regarding multifaceted robot faces based on different contexts. They indicated that the presence or absence of certain facial features can significantly impact how users perceive the robot's capabilities and trustworthiness, thereby influencing its appropriateness for specific tasks or environments [21].

[1] https://landing.ai/platform/

[2] https://openai.com/research/gpt-4v-system-card

[3] https://www.solulab.com/large-vision-models/#What_are_Large_Vision_Models

[4] https://landing.ai/

2.2 LVMs in HRI: status

A considerable volume of research has been successfully conducted by integrating advanced vision models into HRI systems for specialized tasks [7, 22, 27, 31, 36, 41, 45, 53, 57]. A prevalent task is object detection [20, 26, 30, 42, 46], where vision models are tailored to enhance performance, enabling robots to interact more effectively with humans through improved recognition capabilities. Azagra et al. proposed a pipeline including interaction type recognition and target object detection learned by robots based on a series of vision models [1]. Moreover, Fang et al. developed a robust handheld object detection framework by using YOLO [40] and KCF [18] algorithms on RGB-D video data, for better understanding human-object interaction with a humanoid robot [12]. Yu et al. exploited the renowned LSTM model [19] for video object detection, where they tested the proposition in a space robot-human scenario with gestural interactions [54]. Another commonly-designed task in recent HRI research is visual segmentation that partitions images and videos into desired fragments [4, 13, 15]. Ückermann et al. presented a real-time 3D segmentation algorithm that robustly separates objects from unstructured and cluttered environment. The evaluation was implemented in a human-interfered robot grasping mission with a Shadow Robot Hand [48]. Fan et al. proposed a gesture segmentation (together with recognition) architecture based on deep convolutional neural networks with high accuracy, realizing it via an HRI control module with a mechanical arm crawler [10]. Similarly, Kim et al. designed a multimodal HRI framework involving online object segmentation and gesture recognition with visual features extraction, achieving effcient and interactive object learning [23]. Besides, visual tracking is another task that is frequently employed in vision-related HRI research. My et al. [35] and Putro et al. [37] formulated real-time face tracking systems based on camera data for seamless interaction with moving robots. Distinctively, Song et al. developed a face tracking method but by using image data within a human-mobile robot interactive environment [47]. Nonetheless, other tasks such as pose estimation, content generation, visual recognition, etc are also engaged in many HRI related studies. Yet, there is a notable scarcity of research integrating well-defined LVMs into human-robot interaction contexts, with a particular absence of domain-specific LVMs. Given the proven effectiveness of LLMs in text-based information processing, we posit that domain-specific LVMs have the potential to transform vision-based contextual analysis across various industries, leveraging their specialized domain expertise.

3 METHODOLOGY

In this section, we present the formulation of the proposed design space encapsulating three main dimensions with included components in each dimension.

3.1 HRI Contexts

HRI has emerged as a critical area intersecting robotics and HCI during past decades. As robots increasingly coexist with humans in various work environments, the creation of systems that are both safe and user-friendly has become paramount. The challenge lies in the close physical interaction between humans and robots, necessitating the integration of human behaviors into the robot's decision-making framework. As AI technology advances, research in HRI is dual-focused: not only on ensuring the safest possible physical interactions but also on fostering socially appropriate interactions that are sensitive to cultural nuances. The objective is to develop robots capable of intuitive communication with humans, leveraging speech, gestures, and facial expressions for seamless and natural exchanges. The scope of extends to multiple domains where robotic technologies are employed, encompassing industrial automation, medical assistance, and personal companionship, among other diverse applications.

Various studies in the literature have categorized different contexts of interactions between humans and robots, yielding similar outcomes yet grounded in distinct underlying principles [3, 43, 44]. In our paper, we distill the interaction contexts of HRI, a critical dimension within the design space we propose, into three primary categories based on the literature examined (Figure 2):

- ***Human-initiated:*** In this context, human supervisory control over robots is initiated for executing routine tasks. It is anticipated that humans will make decisions about the manner and timing of actions, subsequently directing, instructing, or commanding the robots to carry out these actions and provide relevant feedback. Under this framework, robots are considered passive entities, springing into action only upon receiving explicit control signals from human operators.
- ***Robot-proactive:*** In this context, robots proactively determine the appropriate ways and timing of actions, and also guide humans on their following behaviours. The robots autonomously identify the necessary actions to progress to the subsequent steps required for routine tasks. Concurrently, they independently monitor the situation and observe human actions. Should the predefined goal not be achieved within a set time frame, the robots will then commit an action to ensure the accomplishment of the objective.
- ***Neutral:*** In this context, both humans and robots proactively initiate actions, collaboratively deciding on the how and when of task execution to achieve mutual goals. Actions required for progressing to the next sub-task are determined simultaneously by both parties. Robots select actions as they would in a robot-proactive context, but they also inform humans about their anticipated contributions, based on the robot's task plan. A key application of this approach is in social interactions between humans and robots, where robotic devices are designed to entertain, educate, comfort, and assist specific human users.

3.2 Vision-based tasks

The second dimension identified in our design space pertains to vision-based tasks, developed by synthesizing insights from published works on visual engagement in HRI and current categorizations of computer vision tasks [49]. As illustrated in Figure 2, this dimension comprises nine distinct tasks:

- ***Visual Detection:*** At present, this is the most prevalent task designed in contemporary HRI research employing vision models, such as object detection [20, 46] and face detection [27, 37]. In the majority of cases, the visual detection tasks

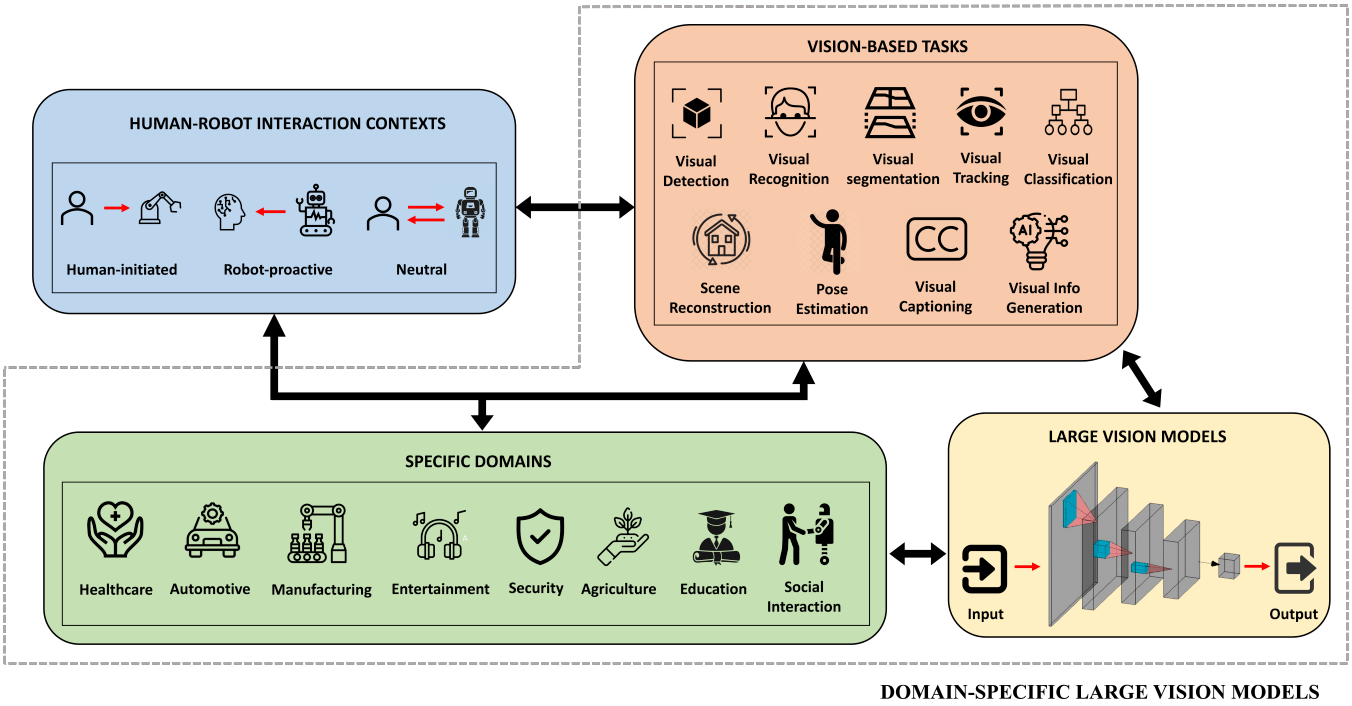

Figure 2: Our proposed design space. The HRI contexts interplay with vision-based tasks tailored to specific domains. Corresponding LVMs deliver performance that meets the unique requirements of each case.

have significantly enhanced the robustness of interaction with robots.

- ***Visual Recognition:*** This task is widely-adopted particularly in gesture-based interactive frameworks [4, 14], where the gestural recognition ensures the smooth control of the robots.
- ***Visual Segmentation:*** Similarly, segmenting a specific set of areas of interest of gestures is currently another important focus used for instructing the robots [23, 48].
- ***Visual Tracking:*** In this task, human motion tracking, human behaviour tracking, and human body tracking are the most used instances for interacting mainly with mobile robots and humanoid robots [32, 33].
- ***Visual Classification:*** As a fundamental task that has evolved alongside deep learning, this area is anticipated to bring significant advantages to HRI systems, particularly in instances like human/robot intention classification [36] and object classification in robot teaching [7].
- ***Scene Reconstruction:*** The sophisticated task of reconstruction has seen rapid development in recent years. Several studies have already leveraged it for decision-making and proactive collaboration within the human-robot loop [11, 56].
- ***Pose Estimation:*** This task is often integrated with visual tracking, enabling robots to proficiently perform 3D estimation of either humans or objects [53].
- ***Visual Captioning:*** As an emerging vision task, it is expected to bring more convenience where robots provide education, or social companionship [22].
- ***Visual Info Generation:*** With the swift advancements in GenAI, this task has demonstrated its high capability in the field of robotics in recent years. In HRI, we foresee its more utilization, building on the current successes in areas like motion generation [57] and location generation [42].

3.3 Specific domains

The primary objective of our paper is to highlight the significant potential of domain-specific LVMs in enhancing the efficiency of HRI systems. In comparison to generic LVMs, these domain-specific models require only about 10% to 30% of the amount of labeled data, while producing considerably fewer errors. This represents a substantial advancement in reducing computational demands and achieving improved performance. Consequently, the specific domains of LVMs form another critical dimension in our design space (as illustrated in Figure 2). We have pinpointed eight domains that are expected to encompass most application areas of HRI, both current and prospective, based on insights from published works and research trends:

- ***Healthcare:*** HRI is revolutionizing health such as robotic-assisted surgery, where precision and reduced invasiveness are critical. Robots like exoskeletons are also being deployed for rehabilitation, providing physical therapy with adaptive routines. The superiority of LVMs in detecting anomalies in medication-related imaging and forecasting disease progression will enhancing the decision-making.

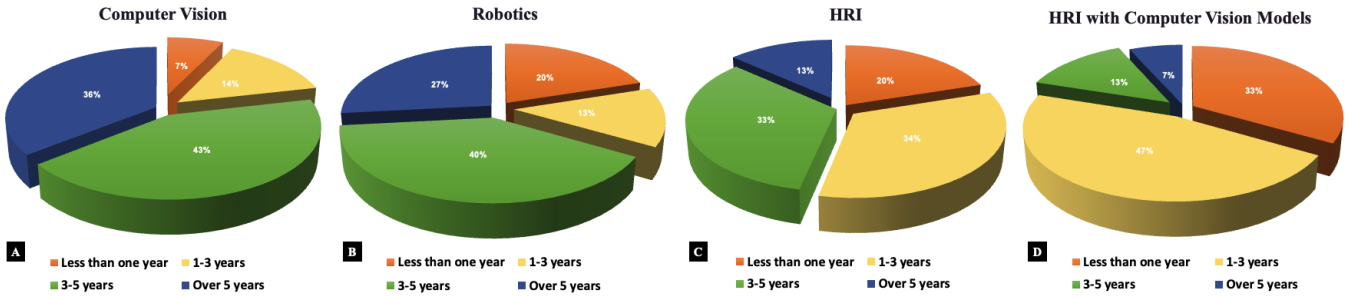

Figure 3: Demographics: academic background information of all participants. A: computer vision; B: robotics; C: HRI; D: HRI with computer vision models.

- ***Automotive:*** LVMs are central in advancing the autonomous vehicles. The real-time object and pedestrian detection, will enable safer navigation and interaction with the environment. Furthermore, LVMs enhance the design of ergonomic vehicle interiors, thereby elevating user experience and safety in driver-assist systems.
- ***Manufacturing:*** LVMs enable robots to execute precise quality inspections, detect defects, and ensure accurate assembly, collaborating with human operators. In addition, the process monitoring and maintenance significantly improve the efficiency and workplace safety.
- ***Entertainment:*** LVMs in HRI are anticipated to vastly improve entertainment experiences through their prowess in generating visual content, including enhanced video and gaming experiences. Moreover, the capacity for analyzing visual data, like predicting human intentions, is set to further enrich the quality of interactive experiences between humans and robots.
- ***Security:*** In critical situations, robots are enabled to perform continuous surveillance, recognize suspicious activities, and identify potential threats with desirable accuracy. Also, utilizing advanced facial recognition and behavior analysis from LVMs to ensure human safety and security publicly.
- ***Agriculture:*** LVMs enable agricultural robots (such as drones) to accurately identify and classify crops, assess plant health, predict incoming yields, and detect pests or diseases, optimizing the cultivating and harvesting.
- ***Education:*** This domain has been ubiquitously probed in HRI. With the aid of LVMs, educational robots are to precisely recognize and respond to manifold gestures and expressions from humans, facilitating more engaging and personalized teaching performance. Some vision-related techniques such as virtual reality/augmented reality are able to make learning more immersive and accessible.
- ***Social Interaction:*** Social robotics is a crucial facet of HRI. In social settings, LVMs enable robots to interpret and respond to human emotions and social cues, fostering more natural and empathetic communication and engagement. A wide-used case is eldercare, where the social robots can provide companionship and emotional support, tailored to individual needs.

4 USER VALIDATION

Exploring the design space's dimensions offers valuable insights to HRI researchers and practitioners about the diversity within the domains of LVMs. However, this exploration does not necessarily reveal how the possible variations may trigger different responses from individuals. To bridge this gap, we carried out an expert evaluation aimed at gathering empirical evidence to guide the development of HRI systems that can evoke specific reactions to LVMs. We engaged a total of 15 participants aged between 24 and 42 (7 male, 6 female, 2 prefer not to say, M = 30.33, SD = 4.66) through email invitations and personal outreach. Among these participants, we have identified their prior experience in computer vision, robotics, HRI, and the experience in proximity with HRI and computer vision models (refer to Figure 3). We found that the majority of participants have several years of relevant experience, which strengthens this empirical validation with enhanced reliability and robustness. We presented them of the design space with a brief background introduction and formulated a questionnaire to delve into the experts' perspectives on the three individual dimensions and the design space itself.

4.1 Questionnaire Design

This study employed a questionnaire-based evaluation alongside quantitative statistical methods to assess the utility of the design space by measuring several perceived metrics motivated from existing literature [2, 16, 17, 21]. In addition to collecting basic demographic information, the questionnaire introduced to participants included an overview of the design space (as shown in Figure 2) and posed a series of targeted questions. These questions aimed to gauge participants' perceptions on several key aspects: perceived likeability, trustworthiness, usefulness, intent to use, comprehensiveness, and system usability of the dimensions incorporated in the design space, utilizing a 7-point Likert scale. Following the questionnaire, participants underwent a system usability scale (SUS) assessment eventually, which yields six metrics in total. The questions posed were as follows:

- How much do you like this dimension/design space? (Very dislike – Very like)
- How trustworthy do you consider this dimension/design space to be? (Very untrustworthy – Very trustworthy)

- How useful do you find this dimension/design space in guiding or inspiring future research in related areas? (Very useless – Very useful)
- How likely are you to utilize this dimension/design space in designing future HRI systems that incorporate LVMs? (Very unlikely – Very likely)
- How comprehensive do you find the dimension/design space in covering the aspects relevant to the future research in related areas? (Very incomprehensive – Very comprehensive)

4.2 Data Analysis

On average, participants completed each trial within approximately 15 minutes, which included a brief interview session at the end, where they were request to share their overall impressions of the design space. Each participant successfully finished the study and received a small gift as appreciation. Prior to conducting statistical analysis, we carried out normality tests and confirmed that all the metrics adhered to a normal distribution ($p < 0.001$) except the SUS score. One-way ANOVA with repeated measurements ($p < 0.05$) for the five metrics and the Friedman test for the SUS score were implemented to determine the statistical significance of the results, which are comprehensively depicted in following subsections and Figure 4. Overall, each dimension and the design space itself garnered satisfactory ratings across all metrics and the SUS score. Notably, the HRI context dimension achieved the highest ratings in every metric, whereas the vision-based tasks dimension registered the lowest.

4.2.1 Likeability. Participants showed an overall preference favoring the HRI contexts dimension as the most likable, closely followed by the design space itself and then the specific domains dimension. However, the differences in likeability ratings did not reach statistical significance.

4.2.2 Trustworthiness. In terms of trustworthiness, perceptions varied slightly from likeability. Here, the specific domains dimension was considered second only to the HRI contexts dimension, though the gap was marginal. Again, these differences were not statistically significant.

4.2.3 Usefulness. The vision-based tasks dimension was seen as the least motivational for future research, distinctly lagging behind the other dimensions. This assessment was supported by a significant statistical difference, as indicated by a repeated measures ANOVA (($F(3, 42) = 4.145, p < 0.05$)), highlighting the variance in perceived usefulness across dimensions.

4.2.4 Intent to Use. The variance in the intention for potential adoption was the smallest across all dimensions with the design space itself when compared to other five evaluated metrics. The vision-based tasks dimension was rated only slightly falling behind other dimensions in this metric, although no significance was detected.

4.2.5 Comprehensiveness. Within the whole study, the HRI contexts dimension received the highest rating for comprehensiveness, nonetheless, the vision-based tasks dimension obtained the lowest rating. Statistical significance was affirmed by the repeated measures ANOVA on the comprehensiveness scales with ($F(3, 42) = 5.119, p < 0.05$).

4.2.6 SUS Score. Consistent with the majority of metrics, the HRI contexts dimension was rated obviously the highest in terms of the SUS score, followed by the design space itself, and then the specific domains dimension. Significant differences were revealed in usability scores, highlighting the distinct user experiences across these dimensions, $\chi^2(3) = 43.591, p < 0.001$.

4.2.7 Summary of Qualitative Feedback. During our brief interview sessions, we collected some valuable insights. The majority of participants expressed positivity of the design space and its three dimensions. One participant, whose research interests align closely with the proposed topic, mentioned, *"The design space you've introduced seems like it will be a useful tool and guide for my future research endeavors to a significant extent, as it clearly lists the essential components."* In terms of the three dimensions, another participant appreciated the strategy of outlining specific domains of focus before diving into practical research challenges brought by LVMs: *"This approach of summarizing key domains before tackling practical research problems can streamline the process of identifying the precise area of interest for applying the vision model. It not only saves time but also leads to enhanced outcomes."* Moreover, a participant with years of experience in computer vision and robotics commented on the utility of the HRI contextual categorization, stating, *"This categorization of contexts will aid in laying down the essential structure for the interactive system I plan to develop in the future, especially since my prior experience in this area is limited."*

5 DISCUSSION

We believe that our design space offers a pioneering framework for designing future HRI architectures, utilizing LVMs across specific domains. In this section, we delve into various facets of the proposed design space.

It serves as a preliminary tool for designers and researchers. In this study, we formulated an initial design space, featuring three dimensions tailored for users to develop future HRI systems using domain-specific LVMs. Our validation process included an empirical testing by experts and the demonstration of its foundational utility across six distinct metrics. Notably, the dimension focusing on HRI contexts received the highest appreciation, whereas the vision-based tasks dimension was evaluated as the least effective according to our metrics. This discrepancy might stem from experts' perception that, despite our efforts to be exhaustive, the list of vision tasks might not fully encapsulate the breadth of existing research in this area. Given the fast-paced advancements in computer vision, it's plausible that more innovative and effective vision tasks will emerge. Nevertheless, we believe that the interaction contexts between humans and robots can be effectively categorized within the three classes we proposed, based on the current state of knowledge. The overall design space received satisfactory appraisal in all metrics, implying that its fundamental utility was primarily endorsed by experts.

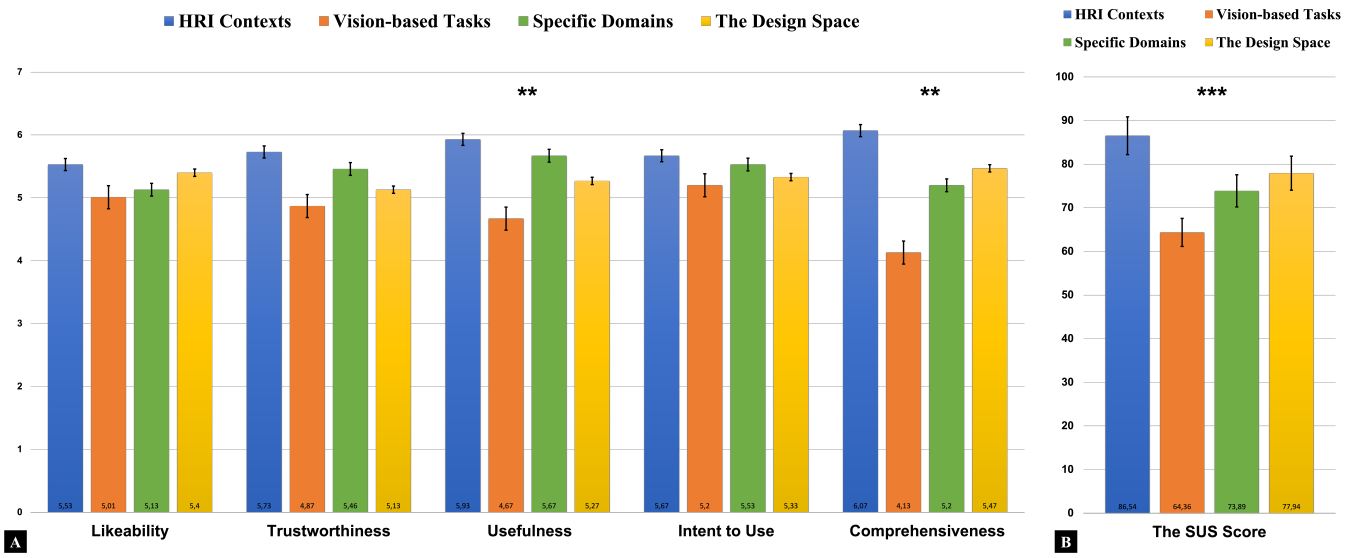

Figure 4: The outcomes of the six evaluated metrics, which assessed the three dimensions and the overall design space. A: Perceived metrics; B: SUS scores. **: $p < 0.05$; ***: $p < 0.001$.

The ideation is based on the existing work and emerging trends. While numerous studies have focused on integrating various vision models in HRI, primarily for enhanced visual detection and robot control, there is a lack of structured approaches in utilizing LVMs for efficient interaction between humans and robots. Given the current surge in LLMs, we anticipate a similar revolutionary trajectory for LVMs in following years. The concept of domain-specific LVMs, first promoted and advocated by Landing AI [5], has shown superior performance over traditional models. This underscores the need for a foundational guideline that assists in designing future HRI systems, delineating specific contexts, vision tasks, and domains.

The proposed approach offers several advantages over current models. Utilizing domain-specific unlabeled data substantially reduces the reliance on expensive labeled data, making the development of models more economical. This cost reduction could lead to enhanced interactions with robots. Additionally, the need for less data is capable to yield faster training periods, enabling the rapid execution and implementation of LVMs in real-world HRI applications especially in wearable and social robotics. Furthermore, by selecting datasets that are specialized for specific domains, the models are expected to achieve significantly higher accuracy due to a deeper understanding of the intricacies unique to different areas.

Using the design space. Our proposed design space serves as an effective tool for preliminarily identifying key elements in human-robot collaborative environments. For instance, the development of a robot arm controlled by human gestures to grasp objects includes: a human-initiated context is established, focusing on tasks such as gesture recognition (visual recognition) and object detection (visual detection). The selection and application of LVMs then follow, tailored to the domains (such as education or entertainment), enhancing visual task performance and facilitating smoother control and interaction. Similarly, in a social companion robot scenario, interaction occurs within a neutral context, employing LVMs specialized for social interaction domain. These models can be utilized for recognizing human emotions and behaviors (visual recognition) and providing empathetic responses (visual captioning). Through our expert evaluation, we are expecting that our design space will aid in informed decision-making in relevant cases.

General challenges. Despite their considerable promise, several challenges need addressing for the broader adoption and practical application. A primary issue is accurately determining the appropriate domain before utilizing specific LVMs to guarantee the precision of visual data, as many HRI studies often overlook domain specification. Data availability may be constrained in some fields due to the difficulty of accessing substantial amounts of domain-specific data, particularly in certain industries. Furthermore, establishing effective interaction methods with robots remains an ongoing challenge in the broader field of HRI research.

Ethical considerations. Another significant factor impacting the successful deployment of the design space is ethic concerns. LVMs require a vast number of database, which can raise a significant factor – data bias and fairness, due to the bias inheriting in the training data. This can lead to inequitable or unethical outcomes, especially in sensitive areas like facial recognition in security and healthcare, where privacy concerns are paramount. Additionally, the interpretability and explainability of these large models are crucial; a lack of understanding of how these models function could compromise transparency. The substantial computational resources required for LVMs pose another significant gap, potentially impeding the successful deployment of an effective HRI system. Finally, ethical issues from HRI's perspective, such as robots being safe

[5] https://landing.ai/

and trustworthy rather than threatening data privacy and human well-being, are expected to be addressed [9].

Limitations of the initial design space. While our design space is formulated on existing HRI research and the evolving trends of large foundational models, certain limitations are inherent. We have identified nine vision-based tasks and eight domains aimed at encompassing the most relevant scenarios. However, the continuous advancement of vision models may give rise to new visual tasks in the future, and HRI could potentially expand into currently not-well-explored domains such as pharmacy and electronics. Moreover, the absence of user studies or expert evaluations in our approach means a lack of practical feedback and insights from professionals in the field, which is crucial for comprehensive validation and enhancement of the design space.

6 CONCLUSION

In this paper, we present an initial design space focused on the future development of HRI systems, leveraging domain-specific LVMs. Mirroring the recent success of LLMs, we foresee LVMs as transformative agents in vision-based information processing within HRI endeavors. We carried out an empirical validation with 15 expert participants, and all perceived metrics and SUS evaluation pointed towards the promising employment of our proposed methodology. Our advocacy for domain-specific models instead of the normal ones stems from their potential to drive significant enhancements and progress in targeted scenarios, enabling more fluid interactions with robots. We believe that this design space will serve as an accelerator for innovative HRI designs across a variety of domains.

ACKNOWLEDGMENTS

To Robert, for the bagels and explaining CMYK and color spaces.

REFERENCES

[1] Pablo Azagra, Florian Golemo, Yoan Mollard, Manuel Lopes, Javier Civera, and Ana C Murillo. 2017. A multimodal dataset for object model learning from natural human-robot interaction. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 6134–6141.

[2] Jeanine MD Baartmans, Francisca JA van Steensel, Lynn Mobach, Tessa AM Lansu, Geraly Bijsterbosch, Iris Verpaalen, Ronald M Rapee, Natasha Magson, Susan M Bögels, Mike Rinck, et al. 2020. Social anxiety and perceptions of likeability by peers in children. *British Journal of Developmental Psychology* 38, 2 (2020), 319–336.

[3] Judith Bütepage and Danica Kragic. 2017. Human-robot collaboration: from psychology to social robotics. *arXiv preprint arXiv:1705.10146* (2017).

[4] Harold Vasquez Chavarria, Hugo Jair Escalante, and L Enrique Sucar. 2013. Simultaneous segmentation and recognition of hand gestures for human-robot interaction. In *2013 16th International Conference on Advanced Robotics (ICAR)*. IEEE, 1–6.

[5] Kerstin Dautenhahn. 2002. Design spaces and niche spaces of believable social robots. In *Proceedings. 11th ieee international workshop on robot and human interactive communication*. IEEE, 192–197.

[6] Kerstin Dautenhahn. 2007. Socially intelligent robots: dimensions of human–robot interaction. *Philosophical transactions of the royal society B: Biological sciences* 362, 1480 (2007), 679–704.

[7] Masood Dehghan, Zichen Zhang, Mennatullah Siam, Jun Jin, Laura Petrich, and Martin Jagersand. 2019. Online object and task learning via human robot interaction. In *2019 international conference on robotics and automation (ICRA)*. IEEE, 2132–2138.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[9] Reza Etemad-Sajadi, Antonin Soussan, and Théo Schöpfer. 2022. How ethical issues raised by human–robot interaction can impact the intention to use the robot? *International journal of social robotics* 14, 4 (2022), 1103–1115.

[10] Jinlong Fan, Yang Yue, Yu Wang, Bei Wan, Xudong Li, and Gengpai Hua. 2022. A continuous gesture segmentation and recognition method for human-robot interaction. In *Journal of Physics: Conference Series*, Vol. 2213. IOP Publishing, 012039.

[11] Junming Fan, Pai Zheng, and Shufei Li. 2022. Vision-based holistic scene understanding towards proactive human–robot collaboration. *Robotics and Computer-Integrated Manufacturing* 75 (2022), 102304.

[12] Zhiwen Fang, Junsong Yuan, and Nadia Magnenat-Thalmann. 2018. Understanding human-object interaction in RGB-D videos for human robot interaction. In *Proceedings of Computer Graphics International 2018*. 163–167.

[13] Haolin Fei, Ziwei Wang, Darren Williams, and Andrew Kennedy. 2023. Hybrid Approach for Efficient and Accurate Category-Agnostic Object Detection and Localization with Image Queries in Human-Robot Interaction. In *IECON 2023-49th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 1–6.

[14] Juan M Gandarias, Jesús M Gómez-de Gabriel, and Alfonso J García-Cerezo. 2018. Enhancing perception with tactile object recognition in adaptive grippers for human–robot interaction. *Sensors* 18, 3 (2018), 692.

[15] Saeed Shiry Ghidary, Yasushi Nakata, Hiroshi Saito, Motofumi Hattori, and Toshi Takamori. 2001. Multi-modal human robot interaction for map generation. In *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No. 01CH37180)*, Vol. 4. IEEE, 2246–2251.

[16] Matthias Gries. 2004. Methods for evaluating and covering the design space during early design development. *Integration* 38, 2 (2004), 131–183.

[17] Jeannie L Haggerty, Marie-Dominique Beaulieu, Raynald Pineault, Frederick Burge, Jean-Frédéric Lévesque, Darcy A Santor, Fatima Bouharaoui, and Christine Beaulieu. 2011. Comprehensiveness of care from the patient perspective: comparison of primary healthcare evaluation instruments. *Healthcare policy* 7, Spec Issue (2011), 154.

[18] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. 2014. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence* 37, 3 (2014), 583–596.

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[20] Shih-Chung Hsu, Yu-Wen Wang, and Chung-Lin Huang. 2018. Human object identification for human-robot interaction by using fast R-CNN. In *2018 Second IEEE International Conference on Robotic Computing (IRC)*. IEEE, 201–204.

[21] Alisa Kalegina, Grace Schroeder, Aidan Allchin, Keara Berlin, and Maya Cakmak. 2018. Characterizing the design space of rendered robot faces. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 96–104.

[22] Soo-Han Kang and Ji-Hyeong Han. 2023. Video captioning based on both egocentric and exocentric views of robot vision for human-robot interaction. *International Journal of Social Robotics* 15, 4 (2023), 631–641.

[23] Soohwan Kim, Dong Hwan Kim, and Sung-Kee Park. 2010. On-line object segmentation through human-robot interaction. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1734–1739.

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643* (2023).

[25] Kazuhiro Kosuge and Yasuhisa Hirata. 2004. Human-robot interaction. In *2004 IEEE International Conference on Robotics and Biomimetics*. IEEE, 8–11.

[26] Arsalan Latif, Aimen Mughal, Muhammad Hasan Danish Khan, and Muhammad D Khan. 2023. Human robot Interaction–Object Detection and Distance Measurement Using Kinect V2. In *2023 International Conference on IT and Industrial Technologies (ICIT)*. IEEE, 1–5.

[27] Sang-Seol Lee, Sung-Joon Jang, Jungho Kim, and Byeongho Choi. 2016. A hardware architecture of face detection for human-robot interaction and its implementation. In *2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*. IEEE, 1–2.

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.

[29] Allan MacLean, Richard Young, Victoria Bellotti, and Thomas Moran. 1991. Design space analysis: Bridging from theory to practice via design rationale. *Proceedings of Esprit* 91, 720-730 (1991), 2.

[30] Elisa Maiettini, Vadim Tikhanoff, and Lorenzo Natale. 2021. Weakly-supervised object detection learning through human-robot interaction. In *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*. IEEE, 392–399.

[31] Aleix M Martínez and Jordi Vitria. 2001. Clustering in image space for place recognition and visual annotations for human-robot interaction. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 31, 5 (2001), 669–682.

[32] P Menezes, L Brethes, F Lerasle, P Danes, and J Dias. 2003. Visual tracking of silhouettes for human-robot interaction. In *Proceedings of The 11th International Conference on Advanced Robotics (ICAR 2003), Coimbra, Portugal*. 971–986.

[33] Luis Molina-Tanco, JP Bandera, Rebeca Marfil, and F Sandoval. 2005. Real-time human motion analysis for human-robot interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1402–1407.

[34] Robin R Murphy, Tatsuya Nomura, Aude Billard, and Jennifer L Burke. 2010. Human–robot interaction. *IEEE robotics & automation magazine* 17, 2 (2010), 85–89.

[35] Vo Duc My and Andreas Zell. 2013. Real time face tracking and pose estimation using an adaptive correlation filter for human-robot interaction. In *2013 European Conference on Mobile Robots*. IEEE, 119–124.

[36] Kai Siang Ong, Yuan Han Hsu, and Li Chen Fu. 2012. Sensor fusion based human detection and tracking system for human-robot interaction. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 4835–4840.

[37] Muhamad Dwisnanto Putro and Kang-Hyun Jo. 2018. Real-time face tracking for human-robot interaction. In *2018 International Conference on Information and Communication Technology Robotics (ICT-ROBOT)*. IEEE, 1–4.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[39] Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922* (2023).

[40] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[41] Shoichiro Sakurai, Eri Sato, and Toru Yamaguchi. 2007. Recognizing pointing behavior using image processing for human-robot interaction. In *2007 IEEE/ASME international conference on advanced intelligent mechatronics*. IEEE, 1–6.

[42] Azhar Aulia Saputra, Chin Wei Hong, and Naoyuki Kubota. 2019. Real-time grasp affordance detection of unknown object for robot-human interaction. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. IEEE, 3093–3098.

[43] Ruth Schulz, Philipp Kratzer, and Marc Toussaint. 2018. Preferred interaction styles for human-robot collaboration vary over tasks with different action types. *Frontiers in neurorobotics* 12 (2018), 36.

[44] Thomas B Sheridan. 2016. Human–robot interaction: status and challenges. *Human factors* 58, 4 (2016), 525–532.

[45] PS Febin Sheron, KP Sridhar, S Baskar, and P Mohamed Shakeel. 2021. Projection-dependent input processing for 3D object recognition in human robot interaction systems. *Image and Vision Computing* 106 (2021), 104089.

[46] Ming-Yuan Shieh, Yen-Hao Chen, Jeng-Han Li, Neng-Sheng Pai, and Juing-Shian Chiou. 2013. Fast object detection for human-robot interaction control. In *Proceedings of the 2013 IEEE/SICE International Symposium on System Integration*. IEEE, 616–619.

[47] Kai-Tai Song and Wen-Jun Chen. 2004. Face recognition and tracking for human-robot interaction. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, Vol. 3. IEEE, 2877–2882.

[48] Andre Ückermann, Robert Haschke, and Helge Ritter. 2013. Realtime 3D segmentation for human-robot interaction. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2136–2143.

[49] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, et al. 2018. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience* 2018 (2018).

[50] Michael L Walters. 2008. *The design space for robot appearance and behaviour for social robot companions*. Ph. D. Dissertation.

[51] Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, Songyao Zhang, et al. 2023. Review of large vision models and visual prompt engineering. *Meta-Radiology* (2023), 100047.

[52] Sarah Woods. 2006. Exploring the design space of robots: Children's perspectives. *Interacting with Computers* 18, 6 (2006), 1390–1418.

[53] Chengjun Xu, Xinyi Yu, Zhengan Wang, and Linlin Ou. 2020. Multi-view human pose estimation in human-robot interaction. In *IECON 2020 The 46th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 4769–4775.

[54] Jiahui Yu, Hongwei Gao, Yongquan Chen, Dalin Zhou, Jinguo Liu, and Zhaojie Ju. 2022. Deep object detector with attentional spatiotemporal LSTM for space human–robot interaction. *IEEE Transactions on human-machine systems* 52, 4 (2022), 784–793.

[55] Ceng Zhang, Junxin Chen, Jiatong Li, Yanhong Peng, and Zebing Mao. 2023. Large language models for human-robot interaction: A review. *Biomimetic Intelligence and Robotics* (2023), 100131.

[56] Zeyu Zhang, Lexing Zhang, Zaijin Wang, Ziyuan Jiao, Muzhi Han, Yixin Zhu, Song-Chun Zhu, and Hangxin Liu. 2023. Part-level scene reconstruction affords robot interaction. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 11178–11185.

[57] Xuan Zhao, Sakmongkon Chumkamon, Shuanda Duan, Juan Rojas, and Jia Pan. 2018. Collaborative human-robot motion generation using LSTM-RNN. In *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*. IEEE, 1–9.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009