Multimodal Human-Urban Robot Interaction Interface Using Large Language Models, Computer Vision and Virtual Reality

Akash Reddy Mallepally[1], Hongrui Yu, Ph.D. [2*]

[1] M.S. Student, Bradley Dept. of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, 1185 Perry St., Blacksburg, VA 24061-0002; e-mail: mallepally@vt.edu
[2*] Assistant Professor, The Charles E. Via, Jr. Dept. of Civil & Environmental Engineering, 750 Drillfield Dr., Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0002; e-mail: hryu42@vt.edu

ABSTRACT

Robotic systems are increasingly being developed to enhance intelligence and user interaction within the built environment, supporting low physical effort lifestyles. However, pre-programmed robots often lack the situational awareness to adapt to unstructured domestic environments and result in task errors. Verbal corrections typically lag, often after mistakes have already caused workpiece damage. To address these challenges, we present a real-time user monitoring and proactive plan adaptation system designed to improve task fluency and user engagement. The system integrates an LLM-based commander model, a computer vision (CV)-based engagement recognition module, and a Virtual Reality (VR) interface for low-workload communication. The LLM acts as a central controller, interpreting inputs, initiating interactions, and translating prompts into robot instructions. Meanwhile, the VR-based emotion monitoring system detects negative user emotions and dynamically adjusts work plans to preemptively avoid errors. Physical embodiment with a desktop setup was conducted to validate and evaluate the system's efficiency.

INTRODUCTION

Robots are increasingly recognized as valuable tools in urban environments, including industrial workplaces, healthcare facilities, and homes for elderly care and individuals with disabilities (Park et al. 2023; Silvera-Tawil 2024). Their adoption streamlines workflows and assists with hazardous and physically demanding tasks (Park et al. 2023; Yu et al. 2023b). However, many robotic systems lack the "intelligence" to perceive user needs and emotions, limiting adaptability. Awareness and proactive adaptation are essential for improving efficiency and user satisfaction (Burden et al. 2022, Yu et al. 2023a). When interacting with civil robots, often through Virtual Reality (VR) or a Digital Twin to enhance usability (Burden et al. 2022, Park et al. 2023), users frequently must reason through commands and instructions. To alleviate this cognitive burden, we propose a human-centric multimodal interaction framework aimed at improving Human-Robot Interaction (HRI). Human-centric design focuses on systems that naturally interpret diverse human needs, offering intuitive experiences that complement user skills (Crnokic et al. 2024; Panagou et al. 2023). Motivated by the synergy among different interfaces and methods, as described below, we aim to address the limitations by developing an intelligent collaboration framework integrating:

Virtual Reality (VR): VR is a key medium in Human-Computer and Human-Robot interaction, enhancing user interaction through intuitive interfaces (Sun et al. 2024). As part of Extended Reality (XR), VR is chosen for its convenience, portability, and safety, compared to direct physical robot interactions. It serves as a virtual user-robot communication channel, facilitating the integration of various components essential for inducing intelligence. VR's potential as an intuitive interface for robot communication remains underexplored in the context of real-time engagement and proactive interaction.

Computer Vision (CV): CV is vital as Robotic Vision for basic robot tasks which necessitate image segmentation or object detection (Crnokic et al. 2024). However, CV can also serve a critical role in detecting human engagement, gaze, and emotions, which is studied for applications in mental health, education, and robotics (Savchenko et al. 2022; Alameda-Pineda et al. 2024, Zixiang et al. 2020). Despite its potential, CV-based engagement recognition has not been fully integrated with real-time decision-making frameworks that dynamically adjust robot behavior based on user interaction. In this study, the authors use this to conduct additional analysis of the scene to gain general awareness of the changes occurring over time. Emotion and engagement recognition are also added to the scene analysis to provide comprehensive human state understanding. We also incorporate these methods as part of system "intelligence."

Natural Language Interaction: LLM-based Natural Language Interaction embedded with robots is also being researched lately as means for effective communication (Alameda-Pineda et al. 2024, Crnokic et al. 2024). Other than its utilitarian role in enabling conversation for easy communication, it also imparts a social nature onto the robot which is key to developing a social-companionship robot (Silvera-Tawil 2024). However, most existing robotics research focuses on static, text-based dialogues, and few studies explore their integration within a multimodal interaction framework involving VR and CV. We use this combination to give the user a better perception of the robot as an engaged and social collaborator.

While previous studies have highlighted the impact of these individual components in enhancing HRI, no existing work, to the authors' knowledge, unifies VR, CV-based engagement recognition, and LLM-driven interaction into a single HRI framework. This study aims to bridge the gap between passive command-based robots and truly interactive, user-aware robots.

METHODOLOGY

We designed a multimodal VR scene featuring an interface that records and interprets user instructions. The VR environment replicates a common domestic environment with furniture that the user attempts to arrange. Domestic environments in virtual interfaces are useful for preliminary testing, particularly with elderly assistive living research (Budzevski et al. 2023). The overview of the User Interface (UI) is shown in Figure 1. The user is presented with a VR interface to speak with the system through OpenAI's Whisper speech-to-text encoding model (Radford 2022). The OpenAI-Unity plugin (Srcnalt 2024) was used to integrate Whisper and provide the primary UI for experiment. The text is then sent to the central robot command model, which leverages GPT-4o-mini API (OpenAI 2025) to give an appropriate response to the user. In the background, however, the system is periodically fed with the images of the scene and the user's facial

expression data for analysis. The backend CV models of the Engagement Recognition (ER) and Scene Analysis form the basis of the proposed system intelligence. The models then pass the extracted scene changes and user emotion state as aggregated descriptive text inputs to the central robot command model to proactively decide when to communicate with the user through natural language and determine an action for the assistive robot. If the user chooses to respond or command the system to help, the response is parsed by the prompt-engineered GPT model into a chain of robot-executable commands. The system continues to monitor changes in scene and user emotions after task execution to proactively determine further action.

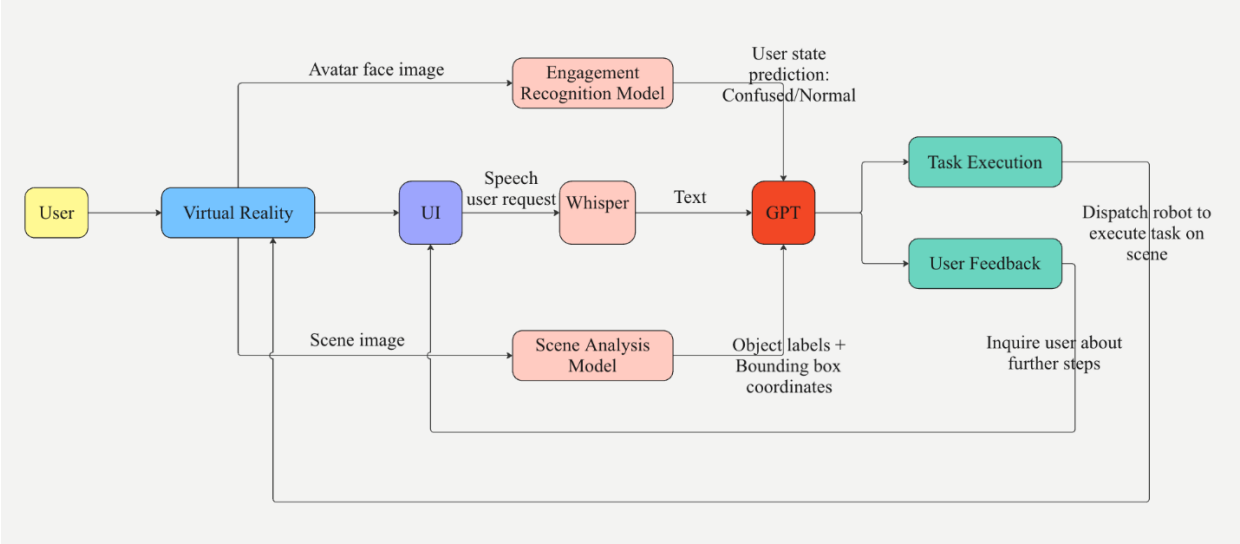

Figure 1. Architecture of the interaction framework.

Engagement Recognition. Recent studies have used neural networks trained to recognize emotions to determine the engagement level of users (Savchenko et al. 2022). The spectrum of emotions and their corresponding trained weights from the neural network have been shown to overlay with the user's focus and engagement levels in a VR setting (Savchenko et al. 2022). For example, negative emotions (sadness, anger, etc.) indicate disengagement or dissatisfaction.

EfficientNets are particularly adept at capturing and predicting facial expressions (Savchenko et al. 2022). For this experiment, we have fine-tuned the enet_b0_8_best_afew model to identify facial expressions of the user's avatar. The experiment uses the Meta Quest Pro Headset that can track facial expressions represented via an avatar (Chen 2021). As directly capturing and processing the user's facial information can be computationally intensive, we have utilized Meta's natural expression and avatar feature as a reference for emotion capturing.

Rather than making the model predict each individual emotion, the use of a binary classifier with "Confused" and "Normal" labels fits this use-case in determining user engagement. As the emotions -contempt, anger, fear, disgust, sad, neutral, and happy (Figure 2) learnt by the pre-trained model fall in the range of negative-neutral-positive emotions, the classifier and the final neural layer were fine-tuned to predict the binary classes.

Figure 2. 'Negative', 'Borderline', Neutral', and 'Positive' samples.

To fine-tune the pre-trained EfficientNet model to the VR avatar and the new labels, we created 83 samples of the avatar's full range of emotions. We chose a smaller dataset and froze all layers except the classifier and the final neural layer, as the model has already learnt to classify emotions.

The fine-tuning experiment took place in 3 stages, as shown in Table 1, to ensure model generalization over new samples. New test samples were added for every successive experiment to prevent training bias. To counter the imbalance of samples in the "Normal" and "Confused" categories (2 Normal vs. 5 Confused emotions), the authors introduced class weights and weighted cross-entropy loss. Image augmentation was also incorporated to improve diversity in a small dataset. In Experiment 3, the test set included avatar samples captured from skewed angles from the real-time experiment to observe the generalizability of the fine-tuned model.

Table 1. Finetuning Experiments for Engagement Recognition Model.

	Number of Samples	Validation Set	Test Set	Accuracy
Experiment 1	33 – Confused 22 – Normal	80-20 split	10 additional	Pre-train: ~45.45% Validation: ~72.73% Post-train: ~28.57%
Experiment 2	38 – Confused 27 - Normal	80-20 split	18 additional	Pre-train: ~46.15% Validation: ~84.62% Post-train: ~16.67
Experiment 2b	47 – Confused 36 - Normal	No validation set	80-20 split	Pre-train: ~58.2% Validation: - Post-train: ~70%
Experiment 3	47 – Confused 36 - Normal	80-20 split	24 additional	Pre-train: ~58.82% Validation: ~70.59% Post-train: ~45.83%

Scene Analysis. The second component of system intelligence comes from its scene analysis capability. Just like how a collaborator would be expected to be aware of task updates, the system is scene aware with the State-of-The-Art (SOTA) YOLO 11 object detection model (Jocher 2024). The model predicts objects and their bounding boxes in the scene. The trajectory is calculated based on the changes in the bounding box values. This information is passed onto the central robot

command model where the labels of objects identified and changes in bounding box values are observed over time. This is the second approach to infer how actively the user works on the task.

Language Interaction. The GPT-4o-mini base model is used as the key point of culmination for inputs from the user and outputs from engagement and scene analysis. As the focus of this work is on establishing the interaction framework rather than creating an application specific LLM, the base model was used without any fine-tuning. Based on this correspondence and situation data, the model decides the occurrence of task execution, i.e., parses user communication (Figure 5) into executable task information (Figure 6) or further communicates with the user to understand their needs better. The feedback process engages the user with the intelligent system by proactively asking the user for next steps in task progression or inquiring about their current satisfaction level. This gives the user a perception of a human and situation-aware collaborative entity.

The model is prompt-engineered to act as a classifier that decides between task intervention and silent observation of the situation. The intervention has two approaches: initiating a conversation with the user or providing commands to the user to assist them. The analysis continues to take place while the robot is on the scene to observe the task progress and user satisfaction with the robot's actions. Sample prompts are provided in the Appendix.

RESULTS AND DISCUSSION

The above models are implemented in a heavy furniture moving case study. The user will provide a habitual vague instruction of "moving forward". The virtual robot will monitor the users' facial expressions and proactively determine when and what questions to ask. The results are as follows:

Engagement Recognition. The model demonstrated a notable ability to learn the relationship between emotions and engagement labels, successfully predicting outcomes when the avatar faced the camera as in training samples. However, validation revealed issues with overfitting, particularly in detecting nuanced emotions like "boredom" and "annoyance," which can be subtle for a 7 or 8-label classifier. These hidden categories are crucial for engagement analysis, as users often do not exhibit direct or extreme emotions during tasks.

The authors implemented a classification threshold (Figure 3a) in experiment 2b to soften the classifier's decision boundary to more appropriately determine which side the emotion leans towards. This was also a necessary step as there was a natural imbalance in samples owing to the higher number of original labels for the "Confused" class (negative emotions - Anger, disgust, contempt, sadness) than for the "Normal" class (positive/neutral emotions - Happy, Neutral). Using a classification threshold during prediction helped prevent the misclassification of ambivalent emotions and made the model generalize better on the borderline range. The accuracy improved from 70% to 76% in the final experiment after thresholding.

However, as seen in experiment 3, the introduction of real-time skewed images into the test set deteriorated the model's performance. In Figure 3b, the confusion matrix shows that the model was unable to identify confused user states in skewed images.

Scene Analysis. As seen in Figure 4, the YOLO v11 model successfully identifies the furniture and their bounding boxes in the Figure which the user wants to move. This gives the central robot command model regular information about the scene being witnessed by the user.

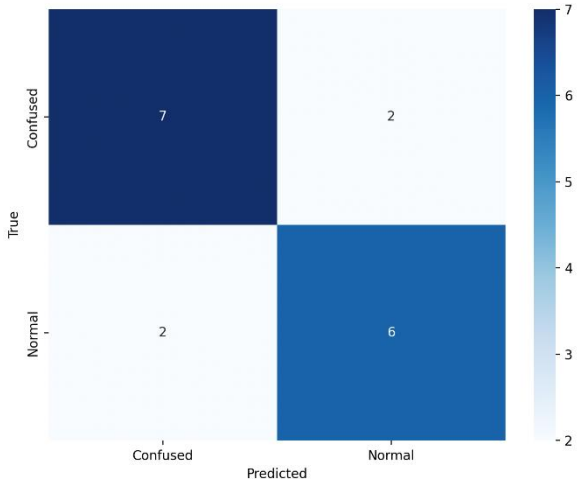

Figure 3a. Experiment 2b Confusion matrix.

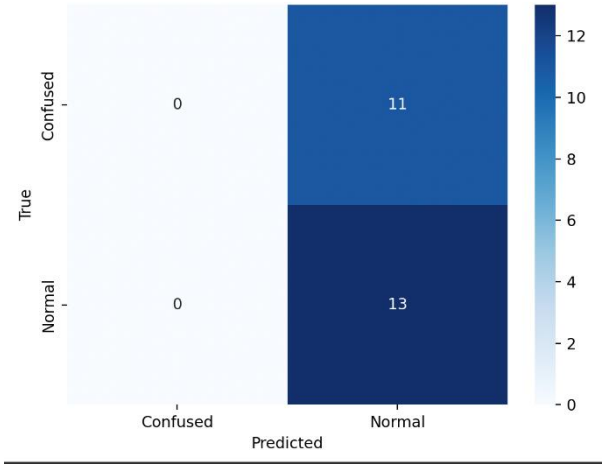

Figure 3b. Experiment 3.

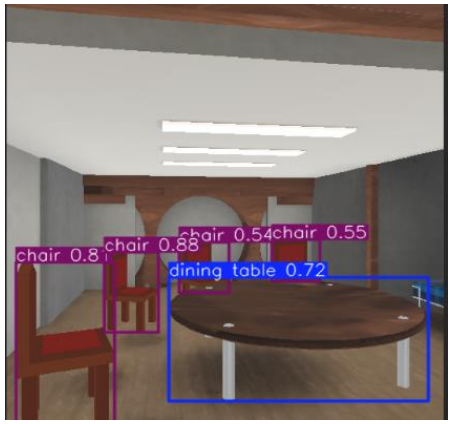

Figure 4. YOLO model predicting objects and bounding boxes.

Language Interaction. We have found that the central robot command model (powered by the GPT-4o-mini interface) has always given a suitable response to the user and has appropriately asked to help them if it detected a lack of change in the scene or confused user state.

As shown in Figures 7 a, b, c, d, and e, the user first requests the robot to move the chair closer to the table (User Prompt: "Help me move the chair closer to the table"). After the robot performs the requested task, the user informs the system of their dissatisfaction with robot task execution (User Prompt: "You moved the chair too little. Can you please move it more.") The robot proceeds to alter the task progression accordingly by moving the chair more. We also validated this system with a physical robot arm of Reactor RX200, as shown in Figure 8.

DISCUSSIONS

This work established a multimodal intelligent-communication pipeline between a human and a robot. Intelligence induction is achieved through engagement recognition, scene analysis, and natural language interaction. The proposed pipeline is robot-neutral, meaning that it is flexible

enough to be overlayed over any available robot without the need for embedding additional programs into the robotic system. Only the pipeline can be modified according to the use-case.

However, the authors acknowledge several limitations that exist in this study. Firstly, the current study validated the feasibility and accuracy of enabling multimodal communication between humans and robots. The model needs to be embodied with a physical robot and a physical urban setting. In the future, the authors plan to extend the suggested framework onto a Mixed Reality system overlaid on a physical environment with a physical robot via Robot Operating System and localization of real-world objects in the virtual environment.

Secondly, both scene and engagement awareness can be expanded to create a comprehensive body of knowledge regarding the scene. The engagement model needs to be further fine-tuned to detect more relevant and clear-cut emotion labels for engagement analysis. As mentioned before, the relation between the "borderline" category of ambiguous emotions and user engagement needs to be studied in detail to deduce appropriate class labels and training techniques. A larger number of diverse avatar samples is needed to generalize across scenarios. The image capture must also be calibrated to reduce skew while capturing the avatar face.

Finally, the effect of applying a humanoid robot using this framework on collaboration, intelligence, and companionship perceptions will be studied to better understand the multimodal framework's broader application capabilities in Human-Robot Interaction.

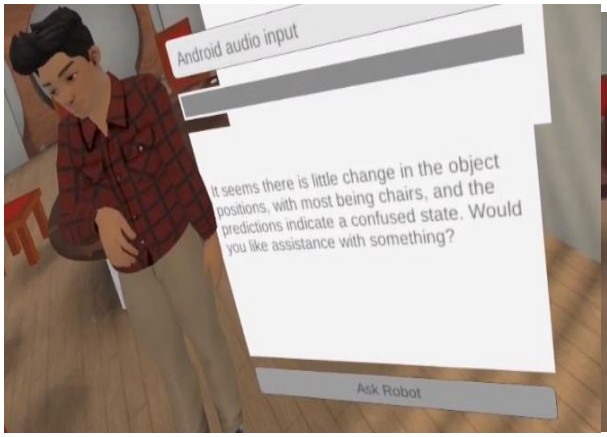

Figure 5. GPT prompting user.

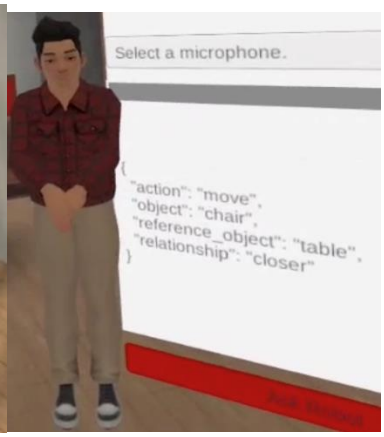

Figure 6. GPT parsing user commands to robot format.

Figure 7a. Task execution.

Figure 7b. Unsatisfactory user feedback detected.

Figure 7c. Task re-execution.

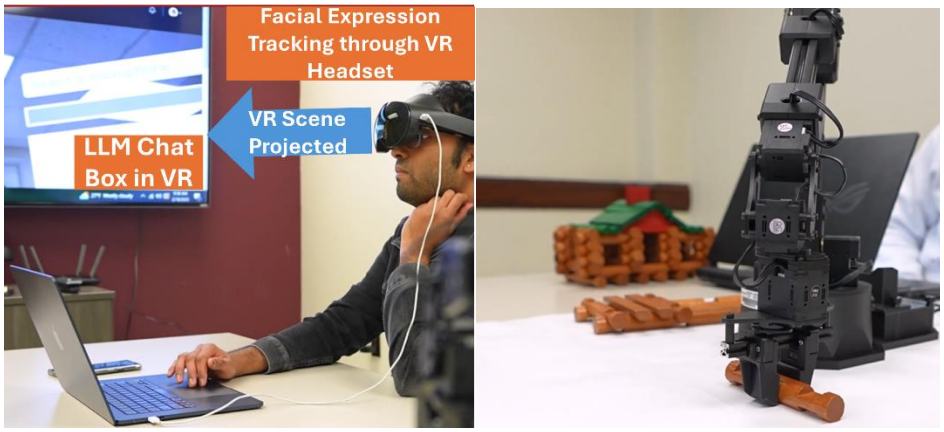

Figure 8. Validation of the proposed multimodal interaction system with a robot arm.

CONCLUSION

By integrating key Human-Computer Interaction methods in a virtual environment, our proposed framework is a step towards a comprehensive awareness-instilled robot communication system. We highlight the primary components of human-centricity that are necessary to facilitate and impart the perception of intelligent-collaboration. In future studies, transferring to a Mixed Reality modality is crucial to study the full impact of this framework. Digital Twinning of the robot and ghosts of relevant objects in the MR environment can be explored further in their capacity to enrich user experience while facilitating better scene analysis for robot decision making. The notion and components of "collaboration" must also be studied to make robotic systems more adept at recognizing the user's need for different degrees of collaboration, thereby personalizing the entity. Owing to the flexibility and freedom offered by VR, and Extended Reality, any new advances in the formulation of human-centricity can be incorporated and studied within our framework to further contribute to the field of Human-Robot Interaction.

REFERENCES

Alameda-Pineda X, A. Addlesee, D. Hernández García, C. Reinke, S. Arias, F. Arrigoni, A. Auternaud, L. Blavette, C. Beyan, L. Gomez Camara, O. Cohen, A. Conti, S. Dacunha, C. Dondrup, Y. Ellinson, F. Ferro, S. Gannot, F. Gras, N. Gunson, R. Horaud, M. D'Inca, I. Kimouche, S. Lemaignan, O. Lemon, C. Liotard, L. Marchionni, M. Moradi, T. Pajdla, M. Pino, M. Polic, M. Py, A. Rado, B. Ren, E. Ricci, A-S. Rigaud, P. Rota, M. Romeo, N. Sebe, W. Sieińska, P. Tandeitnik, F. Tonini, N. Turro, T. Wintz , and Y. Yu. 2024. "Socially Pertinent Robots in Gerontological Healthcare." *arXiv*. doi:10.48550/arXiv.2404.07560. Available from: https://arxiv.org/abs/2404.07560

Budzevski, L., N. Surana, T. Bulbul, and R. Zhang. 2024. "Augmented Reality and Wearable Technology-Supported Biophilic Design of Senior Housing for Improving Quality of Life in Older Adults." *In Computing in civil engineering*, Corvallis, OR: ASCE, 280–287.

Burden, A., G. Caldwell, and M. Guertler. 2022. "Towards human–robot collaboration in construction: current cobot trends and forecasts." *Construction Robotics*. 6, 209–220.

Chen, L., C. Cao, F. De la Torre, J. Saragih, C. Xu, and Y. Sheikh. 2021. "High-fidelity face tracking for ar/vr via deep lighting adaptation." *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Virtual: CVF, 13059–13069.

Crnokic, B., I. Peko, J. Gotlih. 2024. "The Development of Assistive Robotics: A Comprehensive Analysis Integrating Machine Learning, Robotic Vision, and Collaborative Human Assistive Robots." In *Digital Transformation in Education and Artificial Intelligence Application*, Mostar, Bosnia and Herzegovina: Springer, Cham, vol 2124. https://doi.org/10.1007/978-3-031-62058-4_12

Javaid M., A. Haleem, R.P. Singh, R. Suman. 2021. "Substantial capabilities of robotics in enhancing industry 4.0 implementation." *Cognitive Robotics*, 1, 58–75

Jocher, Glenn, and Qiu, Jing. (2024). "Ultralytics YOLO11, Version 11.0.0." Software. Available at: https://github.com/ultralytics/ultralytics. Licensed under AGPL-3.0. ORCID: 0000-0001-5950-6979, 0000-0002-7603-6750, 0000-0003-3783-7069.

OpenAI. 2025. *GPT-4o-mini*. Retrieved from https://openai.com.

Panagou, S., W. P. Neumann, and F. Fruggiero. 2023. "A scoping review of human robot interaction research towards Industry 5.0 human-centric workplaces." *International Journal of Production Research*, *62*(3), 974–990. https://doi.org/10.1080/00207543.2023.2172473

Park, S., H. Yu, C. C. Menassa, and V. R. Kamat. 2023. "A comprehensive evaluation of factors influencing acceptance of robotic assistants in field construction work." *Journal of Management in Engineering*, *39*(3), 04023010.

Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. 2022. "Robust Speech Recognition via Large-Scale Weak Supervision." *arXiv*. doi: 10.48550/ARXIV.2212.04356. Available from: https://arxiv.org/abs/2212.04356

Savchenko, A. V., Savchenko, L. V., & Makarov, I. 2022. "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network." *IEEE Transactions on Affective Computing, 13(4), 2132–2143*.

Sawik, B., S. Tobis, E. Baum, A. Suwalska, S. Kropińska, K. Stachnik, E. Pérez-Bernabeu, M. Cildoz, A. Agustin, and K. Wieczorowska-Tobis. 2023. "Robots for Elderly Care: Review, Multi-Criteria Optimization Model and Qualitative Case Study." *Healthcare*, *11*(9), 1286. https://doi.org/10.3390/healthcare11091286

Silvera-Tawil, D. 2024. "Robotics in Healthcare: A Survey." *SN COMPUT. SCI.* **5**, 189. https://doi.org/10.1007/s42979-023-02551-0.

Sun, T., Y. Yu, Q. Zheng, Z.Wang, C. Zheng. 2024. "Extended reality: Enhancing human-centered capabilities for human-cyber-physical systems (HCPS)." *Procedia CIRP*, 130, 368–373.

Wang X., Shen L., and L. H. Lee. 2024. "Towards Massive Interaction with Generalist Robotics: A Systematic Review of XR-enabled Remote Human-Robot Interaction Systems." *arxiv*. Available from: https://arxiv.org/abs/2403.11384

Yu, H., V. R. Kamat, C. C. Menassa, W. McGee, Y. Guo, and H. Lee. 2023a. "Mutual physical state-aware object handover in full-contact collaborative human-robot construction work." *Automation in Construction*, 150, 104829.

Yu, H., V. R. Kamat, C. C. Menassa, W. McGee, Y. Guo, and H. Lee. 2023b. "Grip state recognition for enabling safe human-robot object handover in physically collaborative construction work." *In Computing in civil engineering*, Corvallis, OR: ASCE, 787–795.

Zixiang, F., E. Yang, D. D. U. Li, S. Butler, W. Ijomah, X. Li, H. Zhou. 2020. “Deep convolution network based emotion analysis towards mental health care.” *Neurocomputing*, 388, 212–227.

APPENDIX

Prompt 1: “You are a personal assistant for an elderly adult, synced with a robot which you can dispatch when necessary. Your tasks include: 1) Interacting with the user when they speak. 2) Receiving the user engagement (Figure 2) and scene analysis (Figure 4) inputs from a virtual environment scene every few seconds. 3) Analyzing changes in these figures over time.
If there is little to no change in object positions or if the predicted class indicates a confused state, decide whether to ask the user if they need help. When the user requests help with a task, respond with a strictly JSON-formatted response containing action, object, reference object, the secondary object, and relationship (The spatial or contextual relationship).

If the user still looks confused after you perform the task, inquire if they would like you to proceed further with the task. Do not prompt the user every time you receive input from engagement and scene analysis. Use your analysis to decide if interaction is necessary. Always respond thoughtfully, balancing the user’s needs and the task urgency.

Prompt 2: “Analyze this image information containing object names and their respective coordinates carefully while picking up the nuance. The information also contains the user state prediction (confused/normal). Track the changes in information over time and ask if the user needs help whenever you feel is necessary and/or when not much progress seems to be made and/or if the user looks confused. If they requested you to perform a task, but they still look confused after execution, ask them if they would like you to proceed in a different way.”