

Intelligent Virtual Sonographer (IVS): Enhancing Physician-Robot-Patient Communication

Tianyu Song^{1,2*}, Feng Li^{1,2*}, Yuan Bi^{1,2}, Angelos Karlas³, Amir Yousefi⁴, Daniela Branzan³, Zhongliang Jiang^{1,2}, Ulrich Eck¹, and Nassir Navab^{1,2}

¹ Chair for Computer-Aided Medical Procedures and Augmented Reality, Technical University of Munich, Munich, Germany

² Munich Center for Machine Learning, Munich, Germany

³ Department for Vascular and Endovascular Surgery, Rechts der Isar University Hospital, Technical University of Munich, Munich, Germany

⁴ Clinic for Vascular Surgery, Helios Klinikum München West, Munich, Germany
zl.jiang@tum.de

Abstract. The advancement and maturity of large language models (LLMs) and robotics have unlocked vast potential for human-computer interaction, particularly in the field of robotic ultrasound. While existing research primarily focuses on either patient-robot or physician-robot interaction, the role of an intelligent virtual sonographer (IVS) bridging physician-robot-patient communication remains underexplored. This work introduces a conversational virtual agent in Extended Reality (XR) that facilitates real-time interaction between physicians, a robotic ultrasound system (RUS), and patients. The IVS agent communicates with physicians in a professional manner while offering empathetic explanations and reassurance to patients. Furthermore, it actively controls the RUS by executing physician commands and transparently relays these actions to the patient. By integrating LLM-powered dialogue with speech-to-text, text-to-speech, and robotic control, our system enhances the efficiency, clarity, and accessibility of robotic ultrasound acquisition. This work constitutes a first step toward understanding how IVS can bridge communication gaps in physician-robot-patient interaction, providing more control and therefore trust into physician-robot interaction while improving patient experience and acceptance of robotic ultrasound. The code is available at <https://github.com/stytim/IVS>

Keywords: Extended reality · Robotic ultrasound · Intelligent agent

1 Introduction

Advancements in robotic ultrasound systems (RUSs) [17, 3, 15] and telemedicine [23, 9] are transforming medical diagnostics by enabling automated and remote procedures. These technologies enhance accessibility to ultrasound imaging, particularly in rural or underserved areas, where trained sonographers may be unavailable. However, the integration of robotic ultrasound systems into clinical

* These authors contributed equally to this work.

practice presents new challenges in communication, decision-making, and user acceptance [10]. A crucial aspect of this integration is the interaction between physicians, robotic systems, and patients—a communication dynamic that is often overlooked in existing research.

While prior studies have explored virtual agents for either patient engagement [5, 27, 24] or physician assistance [19, 18], there remains a significant research gap in designing intelligent agents introduced in this work that facilitate communication among all three entities: the physician, the RUS, and the patient. This gap is critical, as effective coordination among these stakeholders is essential for diagnostic efficiency, physician trust, and patient acceptance. Without a structured communication framework, physicians may struggle to effectively convey instructions to the robotic system, patients may have difficulty understanding the physician’s intent, and physicians may misinterpret patient concerns. This lack of seamless multidirectional communication can lead to errors in robotic control, patient uncertainty, and decreased trust in the procedure.

Recent research in Embodied Conversational Agents (ECAs) [6, 7] has shown that virtual avatars with human-like behaviors, facial expressions, and speech synchronization enhance trust, engagement, and communication effectiveness, particularly in healthcare applications. ECAs have been used to guide patients through medical procedures [4, 26] and support physician decision-making [21], demonstrating their potential in clinical interactions. However, existing ECAs typically focus on single-user interactions, either with the patient or the physician, rather than facilitating real-time multidirectional communication between multiple stakeholders and a robotic system.

Building on this foundation, we introduce an Intelligent Virtual Sonographer (IVS) designed to bridge the interaction gap between the physician, the RUS, and the patient. This agent operates in an Extended Reality (XR) environment, offering assistance and procedural control for the physician, and providing empathetic guidance and reassurance for patients. The agent facilitates bidirectional information exchange, ensuring that physicians receive real-time updates on patient conditions while patients remain informed about the robotic ultrasound procedure. Additionally, the agent enhances transparency and user trust by explicitly communicating the robotic system’s actions to both parties.

Our approach leverages recent advancements in large language models (LLMs) to enable real-time, natural language interaction. The IVS facilitates multidirectional communication by translating physician instructions into robotic ultrasound commands and relaying system updates back to the physician. It allows patients to request pauses or pressure adjustments while keeping them informed about the procedure. Additionally, it ensures effective physician-patient communication by relaying patient concerns to the physician and conveying the physician’s intent to the patient. In this paper, we present the design, implementation, and evaluation of IVS. We assess its impact in bridging physician-robot-patient interactions and analyze its usability, communication effectiveness, perceived intelligence, and overall user satisfaction. Additionally, our system runs entirely locally to address medical privacy concerns, utilizing a relatively

small LLM that operates efficiently without requiring a data center, yet still achieves strong performance. Our findings contribute to the broader field of LLM-driven AI agents in medical applications, particularly in telemedicine and robotic ultrasound-assisted diagnostics.

2 Method

Our system implements an XR-driven robotic ultrasound workflow, as illustrated in Fig. 1, allowing both the physician and patient to interact with virtual sonographers. The RUS autonomously plans and performs ultrasound scanning by extracting intent from natural conversations. The following sections provide a detailed overview of the RUS architecture and the IVS.

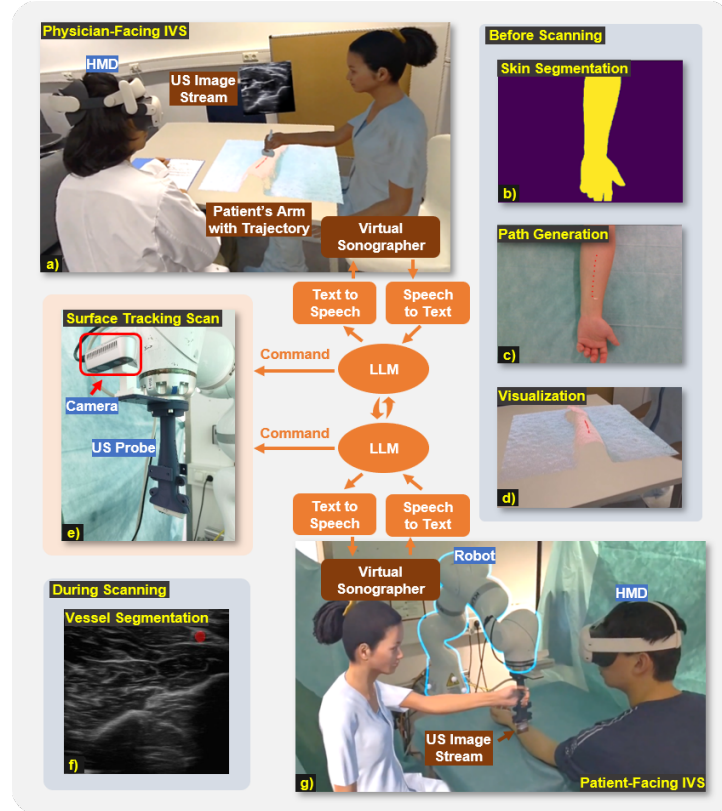


Fig. 1: System overview. (a) IVS interacting with the physician in XR. (b)-(d) Automatic trajectory generation visualized in XR. (e) RUS hardware setup with tracking camera, ultrasound probe, and robot. (f) Real-time vessel segmentation on the ultrasound image. (g) IVS interacting with the patient. The straight arrow represents the data transmission flow, the curvy arrow shows the interaction.

2.1 Robotic Ultrasound System

To enable automatic scanning, the robotic ultrasound system is designed with a KUKA LBR iiwa 14 R820 robotic arm and a Siemens ACUSON Juniper ultrasound machine. The Siemens 12L3 ultrasound probe is rigidly attached to the robot’s end-effector using a 3D-printed probe holder. Additionally, an Intel Realsense D435i depth camera is mounted between the probe and the robot end-effector (Fig. 1(e)) to capture the 3D point cloud from the patient side, which is then reprojected for the physician’s visualization. The hand-eye calibration [13] is used to transform the camera pose into the robot end-effector pose.

Before scanning, the physician can instruct the IVS system to display the patient’s arm and the planned scanning trajectory. During this process, one RGB image along with ten consecutive depth images are acquired. To generate the scanning path, we employ the MediaPipe [20] framework to extract specific body features, including skin segmentation and hand landmark detection. First, the arm mask is obtained using skin segmentation with a trained model, SelfieMulti-class (Fig. 1(b)). To better determine the scanning direction, HandLandmarker model is applied to identify the wrist point pose. Additionally, the arm mask’s skeleton is computed, starting from the wrist and extending to the widest part of the forearm. While these steps rely on the RGB image, generating the full scanning path requires 3D point information, which is extracted from the depth images. We capture ten depth images, extract the depth values at each selected path point, and apply a median filter to refine the depth estimation. Once the origins of each scanning point (t_x, t_y, t_z) are determined, 16 neighboring points for each selected point along the path are acquired and a plane is fitted to calculate the normal vector. This step ensures that the robot’s end effector maintains the smallest possible angle with respect to the normal direction of each path point, allowing for smooth movement and precise surface tracking [16]. As a result, high-quality ultrasound images can be obtained. Finally, we compute the full 6D pose $(t_x, t_y, t_z, r_x, r_y, r_z)$ in the robot base coordinate system for each path point (Fig. 1(c)) and visualize the trajectory for the physician (Fig. 1(d)).

During scanning, the RUS employs impedance control to optimize the contact between the ultrasound probe and the skin. The system can dynamically adjust its behavior—pausing, increasing pressure, or reducing pressure—to accommodate the physician’s specific needs. Additionally, the UNet model [22] trained on 3000 US images (Dice=0.954 $\pm$ 0.012, precision=0.942 $\pm$ 0.021) is used to segment the blood vessels in real-time (Fig. 1(f)), and the blood vessels’ position is relayed to the physician-facing IVS to provide prompts for the physician. After completing the scan, the RUS signals the IVS to conclude the process.

2.2 Intelligent Virtual Sonographer

Multidirectional Communication Framework. Effective communication between the physician, RUS, and patient is essential for ensuring accurate diagnostics, efficient workflow, and patient comfort. The IVS facilitates this interaction through three key communication pathways, shown in Fig. 2:

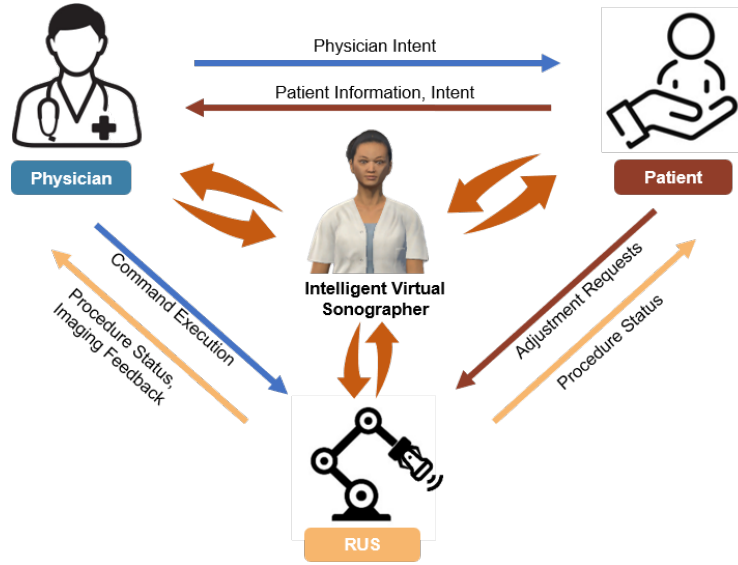


Fig. 2: The IVS facilitates seamless interaction between the physician, robotic ultrasound system, and patient.

Physician ↔ Robot: The physician issues verbal commands to control the RUS, ensuring accurate execution of scanning maneuvers. The robotic system provides real-time feedback on probe positioning and detected anatomical structures, such as vessel locations, helping the physician make informed decisions.

Physician ↔ Patient: The IVS enhances physician-patient communication by conveying the physician’s intent and procedural instructions in an accessible manner. Simultaneously, it ensures the physician receives patient-specific information, including medical history, current condition, and patient preferences, supporting personalized care.

Patient ↔ Robot: Patients can interact with the RUS by expressing discomfort or requesting adjustments, such as reducing probe pressure or pausing the scan. The robotic system, in turn, provides real-time updates on its status, reassuring the patient about the ongoing procedure.

To enable multidirectional communication, the IVS must support simultaneous interactions between the physician, patient, and RUS. Physician-patient communication is inherently asynchronous—the physician and patient may request or provide information at different times, requiring the IVS to handle simultaneous interactions without blocking one conversation for the other. Running two independent LLM instances allows the IVS to manage both interactions in parallel, ensuring that the patient-facing conversation does not delay or interfere with the physician’s workflow and vice versa. To achieve this, the IVS utilizes two separate instances of a LLM, each running Meta Llama 3.1 8B model [8] without fine-tuning. These LLM instances were deployed locally on

an NVIDIA IGX Orin system equipped with an RTX 6000 Ada GPU. To facilitate dual-role interaction, each LLM instance was prompted differently: one was designated for professional communication with the physician, while the other engaged in empathetic conversation with the patient, ensuring clear and supportive explanations. When the physician requested patient-specific information, the physician-facing LLM relayed the query to the patient-facing LLM if it lacked the required data. The patient-facing LLM then asked the patient for a response, after which it relayed the information back to the physician-facing LLM and notified the physician accordingly. Beyond conversation, the IVS also interpreted verbal physician instructions to execute robotic ultrasound actions. Using an approach similar to Xu *et al.* [28], the physician-facing LLM converted spoken commands into robotic control API calls, enabling the IVS to execute the scanning, adjust ultrasound probe pressure, pause or resume the procedure. The IVS also transparently communicated these actions to the patient, reinforcing trust and clarity in robotic-assisted diagnostics.

XR Visualization and Embodied Interaction. The visual representation of the IVS was developed in Unity 2022.3.55, utilizing avatars from the Microsoft Rocketbox library [11]. Predefined animations, such as sitting with breathing and blinking, were baked into the system, while dynamic animations—including head turning during conversations and procedural grasping of the ultrasound probe—were generated in real time using Final IK [1]. Facial lip-syncing animations were driven by SALSA [2], enabling mouth movements synchronized with speech output. The Unity application was executed on a Windows PC equipped with an NVIDIA RTX 4060 GPU and visualized on a Meta Quest 3 headset using Quest Link mode. For speech processing, the system employed the OpenAI Whisper base model for speech-to-text (STT), running on the same PC to transcribe user speech for the LLM. The LLM-generated responses were then synthesized into speech using Kokoro [12], an open-source text-to-speech (TTS) model, which was deployed on the same server as the LLM. This pipeline enabled real-time spoken communication between the IVS and users.

3 Experiment and Results

3.1 User Study

To evaluate the IVS, we conducted a user study assessing its communication accuracy, perceived intelligence, interaction quality, usability, and overall satisfaction. The study aimed to measure both objective communication effectiveness and subjective user experience through structured participant feedback.

While prior work by Song *et al.* [24] has explored the effects of a virtual agent designed for patient interaction, our study primarily focuses on the physician-facing IVS and its role in facilitating physician-robot-patient communication. To this end, we recruited 14 participants (4 female, 10 male), ranging in age from 24 to 42 years (mean age: 31.6 years). Among them, 7 were medical doctors

with an average of 6.7 years of professional experience, while the remaining were biomedical engineers. The study was conducted in accordance with the Declaration of Helsinki. Ethical approval was obtained from the ethics committee of the Technical University of Munich. To ensure a controlled and consistent evaluation, participants took on the role of physicians, while the authors of this paper acted as patients, adopting pre-defined patient personas with different names, ages, and medical histories for each physician participant.

In the user study, the physician was required to retrieve the patient-specific information, including name, age, and medical history, which the IVS relayed during the interaction. To assess communication accuracy, the information provided by the patient was compared to what the physician received through the IVS. Beyond information retrieval, the physician needed to initiate the scanning process verbally and interact with the IVS to control the RUS. Through natural conversation, the physician could instruct the IVS to adjust probe pressure, pause the movement to examine the ultrasound image more carefully, and resume scanning as needed. Additionally, the patient could also interact with the IVS by requesting adjustments to the probe pressure or pausing the procedure, simulating situations where they might feel discomfort during the examination. These tasks evaluated whether the IVS effectively facilitated multidirectional intent communication, ensuring that the physician’s instructions were accurately executed by the robot, that the patient’s concerns were properly relayed and addressed, and that both the physician and patient correctly understood each other’s intent during the interaction.

3.2 Quantitative Evaluation

For the IVS to support natural and fluid conversation, system responsiveness is critical. In our setup, the IVS achieved an average frame rate of 72 FPS at $2,064 \times 2,208$ pixels per eye. On average, the latency for STT was 80 ms, while LLM response generation was 860 ms. The TTS synthesis introduced an additional 150 ms delay, resulting in a total conversational latency of 1.09 seconds per turn.

In terms of communication effectiveness, the IVS achieved an accuracy of 90.48% in relaying patient-specific information to the physician. Additionally, the IVS accurately executed 85.71% of patient-requested actions, such as adjusting probe pressure or pausing the procedure. For physician-requested actions, including scanning maneuvers and robotic adjustments, the system achieved an accuracy of 92.86%. Despite these high accuracy rates, some failure cases were observed during the user study. Occasionally, the LLM hallucinated patient-specific information, generating responses from example patients in the prompt instead of accurately retrieving the correct participant information.

3.3 Subjective Ratings

The subjective ratings are categorized into four aspects: perceived intelligence, interaction quality, system usability, and overall satisfaction, in a 5-point Likert scale, where higher scores indicate better performance. The results, as shown in

Fig. 3, are presented separately for novice and physician participants (mean $\pm$ standard deviation). Perceived intelligence was rated 4.11 ± 0.45 (median: 4), and 4.17 ± 0.64 (median: 4.2) by physicians. Interaction quality received 4.33 ± 0.43 (median: 4.33) from novices, and 4.14 ± 0.54 (median: 4) from doctors. System usability was rated 4.21 ± 0.51 (median: 4.25) by novices but lower by physicians at 3.82 ± 0.59 (median: 3.86). Overall satisfaction was 4.28 ± 0.75 (median: 4) for novices, and 4.00 ± 0.81 (median: 4) for physicians. While both groups provided high ratings, the lower interaction, usability and satisfaction score from physicians suggested potential areas for further improvement.

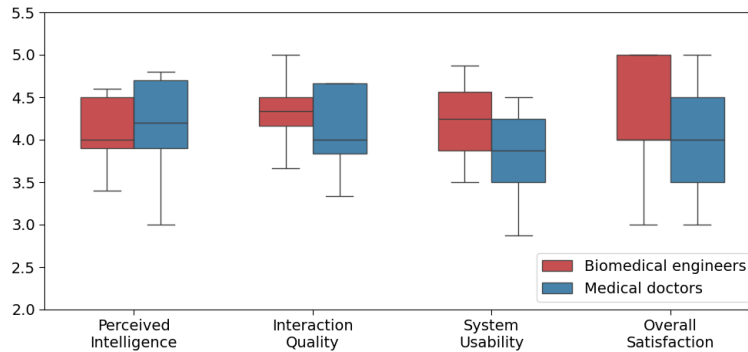


Fig. 3: Subjective ratings of biomedical engineers and medical doctors.

4 Discussion and Conclusion

In this work, we introduced IVS, a dual-LLM-driven embodied conversational agent, designed to enhance physician-robot-patient communication by facilitating natural and interactive dialogue in robotic ultrasound. Even with a moderate-sized LLM without fine-tuning, IVS provides initial evidence of facilitating communication, improving the efficiency and transparency of robotic ultrasound. However, challenges remain in latency and LLM reliability. While the system latency exceeds typical human response times [25], it remains within an acceptable range for interactive medical applications [14]. Another critical aspect is LLM hallucination, which risks generating incorrect clinical information. Our dual-instance LLM architecture mitigates this to some extent, but integrating medical knowledge bases and physician validation mechanisms could further enhance reliability. Despite these challenges, IVS received positive ratings, demonstrating its potential as a scalable, locally deployed AI agent in medical robotics. While current LLMs have limitations in complex medical contexts, rapid advancements in contextual understanding and reliability are underway, and our framework is positioned to benefit from these improvements. As IVS intelligence evolves, the

interaction paradigm could also shift from low-level robot control to higher-level clinical communication, allowing physicians to request anatomical views rather than specifying probe manipulations—aligning with existing clinical workflows. Additionally, future integration of sensing modalities such as eye-tracking, heart rate, facial expression, or voice analysis may enable IVS to infer patient emotional states and relay them to physicians, enriching remote diagnostic interactions. This novel approach paves the way for future advancements, encouraging the community to further develop AI-driven virtual agents that enhance trust, usability, and real-world impact in robotic-assisted diagnostics.

Acknowledgments. The authors wish to thank NVIDIA for their in-kind hardware and advisory support. This work was partly supported by the state of Bavaria through Bayerische Forschungsförderung (BFS) under Grant AZ-1592-23-ForNeRo.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Final IK. <https://assetstore.unity.com/packages/tools/animation/final-ik-14290>, accessed: 2025-02-26
2. SALSA LipSync Suite. <https://assetstore.unity.com/packages/tools/animation/salsa-lipsync-suite-148442>, accessed: 2025-02-26
3. Bi, Y., Jiang, Z., Duelmer, F., Huang, D., Navab, N.: Machine learning in robotic ultrasound imaging: Challenges and perspectives. *Annual Review of Control, Robotics, and Autonomous Systems* **7** (2024)
4. Bickmore, T., Gruber, A., Picard, R.: Establishing the computer–patient working alliance in automated health behavior change interventions. *Patient education and counseling* **59**(1), 21–30 (2005)
5. Bickmore, T.W., Pfeifer, L.M., Jack, B.W.: Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In: *Proceedings of the SIGCHI conference on human factors in computing systems*. pp. 1265–1274 (2009)
6. Cassell, J.: Embodied conversational interface agents. *Communications of the ACM* **43**(4), 70–78 (2000)
7. Cassell, J.: Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine* **22**(4), 67–67 (2001)
8. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024)
9. Eck, U., Wechner, M., Pankratz, F., Yu, K., Lazarovici, M., Navab, N.: Real-time 3d reconstruction pipeline for room-scale, immersive, medical teleconsultation. *Applied Sciences* **13**(18), 10199 (2023)
10. Eilers, C., van Kemenade, R., Busam, B., Navab, N.: On the importance of patient acceptance for medical robotic imaging. *International Journal of Computer Assisted Radiology and Surgery* **18**(7), 1261–1267 (2023)

11. Gonzalez-Franco, M., Ofek, E., Pan, Y., Antley, A., Steed, A., Spanlang, B., Maselli, A., Banakou, D., Pelechano, N., Orts-Escolano, S., et al.: The rocket-box library and the utility of freely available rigged avatars. *Frontiers in virtual reality* **1**, 561558 (2020)
12. Hexgrad: Kokoro-82m (2025). <https://doi.org/10.57967/hf/4329>, <https://huggingface.co/hexgrad/Kokoro-82M>
13. Horaud, R., Dornaika, F.: Hand-eye calibration. *The international journal of robotics research* **14**(3), 195–210 (1995)
14. Huq, S.M., Maskeliūnas, R., Damaševičius, R.: Dialogue agents for artificial intelligence-based conversational systems for cognitively disabled: A systematic review. *Disability and Rehabilitation: Assistive Technology* **19**(3), 1059–1078 (2024)
15. Jiang, Z., Bi, Y., Zhou, M., Hu, Y., Burke, M., Navab, N.: Intelligent robotic sonographer: Mutual information-based disentangled reward learning from few demonstrations. *The International Journal of Robotics Research* **43**(7), 981–1002 (2024)
16. Jiang, Z., Grimm, M., Zhou, M., Hu, Y., Esteban, J., Navab, N.: Automatic force-based probe positioning for precise robotic ultrasound acquisition. *IEEE Transactions on Industrial Electronics* **68**(11), 11200–11211 (2020)
17. Jiang, Z., Salcudean, S.E., Navab, N.: Robotic ultrasound imaging: State-of-the-art and future perspectives. *Medical image analysis* p. 102878 (2023)
18. Killeen, B.D., Chaudhary, S., Osgood, G., Unberath, M.: Take a shot! natural language control of intelligent robotic x-ray systems in surgery. *International journal of computer assisted radiology and surgery* **19**(6), 1165–1173 (2024)
19. Killeen, B.D., Suresh, A., Gomez, C., Inigo, B., Bailey, C., Unberath, M.: Intelligent control of robotic x-ray devices using a language-promptable digital twin. *arXiv preprint arXiv:2412.08020* (2024)
20. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019)
21. Martínez-Miranda, J.: Embodied conversational agents for the detection and prevention of suicidal behaviour: current applications and open challenges. *Journal of medical systems* **41**(9), 135 (2017)
22. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18. pp. 234–241. Springer (2015)
23. Roth, D., Yu, K., Pankratz, F., Gorbachev, G., Keller, A., Lazarovici, M., Wilhelm, D., Weidert, S., Navab, N., Eck, U.: Real-time mixed reality teleconsultation for intensive care units in pandemic situations. In: *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. pp. 693–694. IEEE (2021)
24. Song, T., Pabst, F., Eck, U., Navab, N.: Enhancing patient acceptance of robotic ultrasound through conversational virtual agent and immersive visualizations. *IEEE Transactions on Visualization and Computer Graphics* **31**(5), 2901–2911 (2025). <https://doi.org/10.1109/TVCG.2025.3549181>
25. Stivers, T., Enfield, N.J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G., Rossano, F., De Ruiter, J.P., Yoon, K.E., et al.: Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences* **106**(26), 10587–10592 (2009)
26. Ter Stal, S., Kramer, L.L., Tabak, M., op den Akker, H., Hermens, H.: Design features of embodied conversational agents in ehealth: a literature review. *International Journal of Human-Computer Studies* **138**, 102409 (2020)

27. Walker, T., Christensen, H., Mirheidari, B., Swainston, T., Rutten, C., Mayer, I., Blackburn, D., Reuber, M.: Developing an intelligent virtual agent to stratify people with cognitive complaints: a comparison of human–patient and intelligent virtual agent–patient interaction. *Dementia* **19**(4), 1173–1188 (2020)
28. Xu, H., Wu, J., Cao, G., Chen, Z., Lei, Z., Liu, H.: Transforming surgical interventions with embodied intelligence for ultrasound robotics. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 703–713. Springer (2024)