Published in final edited form as:
Int J Comput Assist Radiol Surg. 2024 June ; 19(6): 1165–1173. doi:10.1007/s11548-024-03120-3.

Take a shot! Natural language control of intelligent robotic X-ray systems in surgery

Benjamin D. Killeen[1], **Shreayan Chaudhary**[1], **Greg Osgood**[2], **Mathias Unberath**[1]

[1]Laboratory for Computational Sensing and Robotics, Johns Hopkins University, Baltimore, MD 21218, USA

[2]Department of Orthopaedic Surgery, Johns Hopkins University, Baltimore, MD 212187, USA

Abstract

Purpose—The expanding capabilities of surgical systems bring with them increasing complexity in the interfaces that humans use to control them. Robotic C-arm X-ray imaging systems, for instance, often require manipulation of independent axes via joysticks, while higher-level control options hide inside device-specific menus. The complexity of these interfaces hinder “ready-to-hand” use of high-level functions. Natural language offers a flexible, familiar interface for surgeons to express their desired outcome rather than remembering the steps necessary to achieve it, enabling direct access to task-aware, patient-specific C-arm functionality.

Methods—We present an English language voice interface for controlling a robotic X-ray imaging system with task-aware functions for pelvic trauma surgery. Our fully integrated system uses a large language model (LLM) to convert natural spoken commands into machine-readable instructions, enabling low-level commands like “Tilt back a bit,” to increase the angular tilt or patient-specific directions like, “Go to the obturator oblique view of the right ramus,” based on automated image analysis.

Results—We evaluate our system with 212 prompts provided by an attending physician, in which the system performed satisfactory actions 97% of the time. To test the fully integrated system, we conduct a real-time study in which an attending physician placed orthopedic hardware along desired trajectories through an anthropomorphic phantom, interacting solely with an X-ray system via voice.

Conclusion—Voice interfaces offer a convenient, flexible way for surgeons to manipulate C-arms based on desired outcomes rather than device-specific processes. As LLMs grow increasingly capable, so too will their applications in supporting higher-level interactions with surgical assistance systems.

✉Benjamin D. Killeen, killeen@jhu.edu.

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

Informed consent This article does not contain patient data collected by any of the authors.

Keywords

Image-guided surgery; Autonomous imaging; Large language models; Speech-to-text; Machine learning

Introduction

As surgical assistance systems become more capable, their user interfaces likewise grow more complicated. High-level functions may exist, but if the actual steps to use them are not obvious, then their availability is merely theoretical, or “present-at-hand.” A hammer, by contrast, is “ready-to-hand” in that its function and use are both available at a glance. Making complex surgical assistance systems ready-to-hand may encourage adoption and use of potentially time- and risk-reducing capabilities. C-arm X-ray imaging devices, for instance, provide guidance to many minimally invasive procedures in orthopedics, interventional radiology, and angiology. Fully robotic C-arms are capable of precisely orienting themselves to achieve desired views, but they often require manipulation of independent axes via joysticks. Higher-level functions, such as patient-specific imaging [1, 2], may reduce radiation or operating time, but they remain hidden to all but dedicated power users of specific systems.

Natural language offers a flexible, familiar interface for surgeons to manipulate X-ray systems by expressing their desired outcome rather than remembering the device-specific steps necessary to achieve it [3]. Unlike graphical user interfaces, which demand users’ attention, and foot pedal controls, which can only trigger a small number of functions, natural language interfaces enable human–robot collaboration between surgeons and imaging systems, so that surgeons can manipulate imaging devices themselves without interrupting the procedure [4], as they have been shown to decrease cognitive load for human collaborators [5]. One major challenge in the development of such systems is the conversion of domain-specific instructions, such as “More inlet” or “Go to the obturator oblique view,”—for which there is no large dataset available—into machine-readable instructions. Recently, the advent of large language models (LLMs) has accelerated the development of natural language interfaces for controlling robotic manipulation systems on a range of real-world tasks, such as rearranging blocks, fetching household objects, folding laundry, and more [6]. As LLMs become even more capable, they have been shown to be effective few-shot learners, able to learn new behaviors without re-training on large datasets [7]. This key capability opens the door to rapid development of a natural language interface for controlling X-ray systems, as in Fig. 1.

Here, we present a fully integrated English language interface for controlling a robotic X-ray system with spoken commands, with additional support for patient-specific imaging in X-ray-guided pelvic surgery. Based on observations of X-ray-guided surgery, we generate a limited set of episodes indicating how to adjust a robotic X-ray system given certain commands. We additionally provide support for patient-specific imaging, which chooses the optimal pose to achieve a desired view based on automated image analysis. When provided with these episodes and a short set of instructions for standardizing communication, an LLM is able to generate machine-readable commands for positioning a Brainlab Loop-X device, a fully robotic X-ray system with six degrees of freedom. In our integrated system, we utilize an open-source transcription model [8] to convert spoken commands into text prompts, which are relayed to an LLM [9]. The resulting commands are either relayed directly to the Loop-X, in the case of a low-level movement, or referred to an image analysis engine for determining the patient-specific view. We evaluate the system's ability to interpret natural commands and to achieve standard views of the pelvis. Finally, we conduct a real-time study with an attending physician, in which he successfully achieved desired alignment of a Kirschner wire (K-wire) with the S2 corridor under image guidance, using only voice commands to control with the system.

Related work

The promise of natural language interfaces has always been to distill complex systems into ready-at-hand tools for humans to use. As such, much work has explored language control for general-purpose robotics [6, 10–13]. In [10, 12], for example, the user expresses high-level goals like, "Move the red block left of the blue shoe," which is embedded using a language model to be combined with RGB-D observations. These latent-space embeddings are then converted into machine-readable instructions in the form of pick-and-place instructions in the field of view. More recently, LLMs have enabled even more powerful language interfaces for real-world tasks in everyday life. PaLM-E [6], for instance, is a multi-modal, embodied model capable of interpreting images and natural language, breaking down complex goals into achievable sub-goals, corresponding to policies in [14]. In medicine, similar models may be capable of assisting physicians interfacing with electronic medical records [15] and planning treatments [16], although significant challenges remain to validate such models given the increased risk [17, 18]. As of yet, however, the potential for LLMs to interpret the domain-specific language necessary to control medical robots has yet to be explored.

As image models have enabled complex tasks in general robotics, they have also yielded promising advances toward patient-specific and task-aware interventional X-ray imaging systems [19]. Notable successes in this area demonstrate the effectiveness of deep neural networks (DNNs) at recognizing anatomical landmarks [20–22], localizing surgical instruments [22, 23], anticipating complications [24], analyzing surgical workflows [25, 26], and planning X-ray system movements [1, 2, 27]. [2, 27] in particular train a DNN to regress the pose change necessary for a mobile C-arm system to achieve standard views of the pelvis and spine, given the current X-ray image, while [1] plan optimal image acquisitions for assisting percutaneous pelvic fracture fixation. Here, we implement a similar system for achieving standard views, relying on DNNs to triangulate anatomical landmarks as surgery progresses.

Interpreting domain-specific language in X-ray-guided surgery

Our approach consists of a minimal protocol enabling an LLM to control a robotic X-ray system, namely the Brainlab Loop-X device. Although our protocol is specialized to control the degrees of freedom in this system, it can be easily adapted to similar systems due to the brevity of the instructions, which are less than 250 lines including examples. Messages consist of single-line strings delimited by semicolons, with the first element indicating the topic or `message_type`. After transcribing a spoken command from the surgeon, our system obtains the patient orientation and current X-ray system pose to provide additional context along with the user input. These are sent to the LLM in an initial `command` message, to which the LLM can respond with an `action or a question`, in case the LLM is unsure how to proceed. Actions may consist of an axis movement (`pose`), patient-specific view (`view`), image acquisition (`shot`), or no-op action (`none`). In the case of a question, the user may respond with a clarification that is stored for future reference. Messages and instructions are concise in order to minimize token usage, a common bottleneck in existing LLMs.

In our case, the Loop-X X-ray system consists of six independent axes, as shown in Fig. 2a. Unlike conventional C-arm systems, the Loop-X is an O-arm-like device with independently moving X-ray source and detector arms (`source_angle and detector_angle`). It moves on the floor in any direction, in a coordinate system specified by the lateral (`x`) and longitudinal (`y`) position. It can also rotate freely (`yaw`). Finally, the gantry tilts from −30° to 60° about `x`. Communication with the Loop-X is achieved via a Remote Function Call Protocol (RFCP) that enables queries and commands based on these axes. The RFCP also provides functionality to achieve precise views specified by an orientation vector in coordinate frame of the ring f_{ring}, with which to align the principle ray $\hat{\mathbf{r}}$, and acquire navigated 2D images semi-automatically. For safety reasons, the Loop-X requires the user to confirm these actions with a physical button, but at no time in our experiments did the user initiate actions from the Brainlab user interface itself. This physical confirmation is a requirement for the Loop-X device, but future systems may avoid additional confirmation by building sufficient trust with the surgeon.

After the full text of the instructions specifying this protocol (see Appendix), we append 35 example interactions. These consist of the initial command, followed by the anticipated response from the LLM, and finally a summary of the interaction. For example, the following episode details the “Push in” command, which should move the Loop-X laterally toward the surgeon. (The `patient_side` provided at start-time here, but could be inferred from imaging or external camera sources.)

```
command; True; supine; patient_right; 180.0;0.0;0.0;0.0;0.0;0.0; Push in.
action; pose; 180.0;0.0;0.0;10.0;0.0;0.0
summary;“Push in ” means moving along X toward the surgeon . Since the
surgeon is on patient right, head_first = True, and patient_pose = supine,
you should increase X by 10.
```

Determining patient-specific views

To determine patient-specific views, we align 2D images with a 3D statistical shape model, in which standard views are known. First, we identify a suitable set of recent acquisitions from which to triangulate points. For percutaneous pelvic fracture fixation, surgeons generally alternate between two nearly orthogonal views, making it convenient to obtain suitable acquisitions in a passive manner. Given a set of suitable images, then, we obtain 2D positions for anatomical landmarks present $\left\{\mathbf{u}_i, l \in \mathbb{R}^2 | l \in \mathscr{L}_i\right\}$, where $\mathcal{L}_i$ is the set of landmarks present in image *i*. These anatomical landmarks are well defined in the literature [20, 22, 25, 28], consisting of identifiable points on the surface of the pelvis, such as bony protrusions. This is accomplished using automated anatomical landmark detection with a U-Net like architecture from [25]. For each landmark present in at least two images, we estimate its 3D position using triangulation. That is, we find the 3D point $\mathbf{x}_{i,l} \in \mathbb{R}^3$ that minimize the reprojection error back onto each image,

$$\mathbf{x}_l = \underset{x}{\operatorname{argmin}} \sum_i \|\mathbf{P}_i\tilde{\mathbf{x}} - \tilde{\mathbf{u}}\|^2 \tag{1}$$

where $\mathbf{P}_i$ is the projection matrix of image *i* relative to an optical marker body fixed to the patient, and ~ indicates the point in homogeneous coordinates. The transformation from the marker body coordinates to f_{ring} is provided by the RFCP. Once anatomical landmarks have been obtained in 3D, we fit an SSM of the pelvis to the triangulated points using deformable iterative closest point [29]. This is essential to ensure that the anterior superior iliac spine (ASIS) landmarks can be estimated, from which the anterior pelvic plane (APP) frame is determined. Following [25], the principle ray for each standard view, including the AP, lateral, inlet, outlet, and obturator oblique views, is defined in this coordinate frame. Figure 2b shows the image projections, detected landmarks, and standard views for three acquisitions sampled randomly, such as might be acquired in the course of fluoro-hunting. Once these views have been defined, they can be achieved through the inverse kinematic solver of a robotic X-ray system. In our experiments, the Brainlab Loop-X is able to achieve all the defined views given a reasonable table height and patient pose. It avoids any potential gimbal lock by approaching trajectories from a standard upright starting position.

Results

We evaluate our system in three ways. First, we evaluate the language interface by showing that satisfactory actions are chosen based on 212 prompts commonly occurring in pelvic trauma surgery. Second, for actions which requested a patient specific view, we evaluate the system's ability to attain this view based on 1, 2, and 3 random prior images, using a rating system of 1 (wrong view) to 5 (no adjustment needed), as rated by an attending physician. Finally, we evaluate the fully integrated system in a phantom study with spoken commands carried out by a commercial robotic X-ray device.

To generate prompts in our test set, we begin with the base prompts listed in Table 1. These are paraphrased from conversations with an attending orthopedic surgeon and professional C-arm technologists. Additional variations on the base prompts are then generated by asking a LLM for more ways to say the same phrase, being sure to inject alternate phrases and synonyms. For example, the simplest command to achieve more inlet tilt is "More inlet," based on which we introduce variations such as "Provide a downward tilt away from the patient’s head,” and “Adjust the C-Arm for a lower view, away from the head.” We then cull the resulting list of 100 commands to ensure they still align with sensible outputs for controlling the Loop-X system, resulting in 53 prompts not seen during the example episodes. For each prompt, we test the LLM supplied with our instructions on four randomly sampled Loop-X poses as the input pose. We evaluate a response as “correct” if it chooses the intended action according to an attending physician. The system performs quite well, choosing the correct response for (206 / 212) 97% commands. In general, failures resulted from extreme Loop-X pose inputs dissimilar to those supplied in the example episode.

We also evaluate real images obtained with no feedback from a physician, relying solely on the patient-specific imaging system described here. This assesses the ability of our system to align with physicians’ expectations in terms of the quality of the performed command. After two random acquisitions 30° apart, we prompt the system to acquire patient-specific views of a radiopaque anthropomorphic pelvic phantom by saying, for example, “Go to an AP view.” An attending physician then rated each image in terms of its alignment with the standard view as evaluated on a 10-point scale, where 10 is a perfect acquisition and 1 indicates the wrong view (see Table 2). The AP, lateral, and inlet views were successfully or very nearly achieved based on voice commands, with an average rating of 8.8 / 10. The outlet and obturator oblique views received a rating of 6.7, which indicates that some fine adjustment is needed to achieve the ideal view. This is in part due to the fact that the Loop-X is limited to a 30° tilt in the outlet direction, although the outlet view is typically closer to 40° away from AP.

Phantom study with an attending physician

Combining all the components of our system, we conduct a study emulating percutaneous fracture fixation in the pelvis. A pair of Shokz OpenRun wireless headphones streamed audio to a 2019 MacBook Pro laptop, which transcribed commands in real time using the Whisper ASR model [8]. For this study, the investigators maintained silence in the OR except for the surgeon’s commands, although this could be avoided by the use of a wake word or task-aware wake-up [30]. The laptop also relayed commands to the LLM, in this case GPT-4 [9]. A Linux server with an RTX 3090 GPU was responsible for automated landmark detection and triangulation, as well as communication with the Loop-X via the RFCP. Images acquired on the Loop-X were relayed in DICOM format, including navigation information, to the server directly. To sum-up, the power-on steps for our integrated system were (1) turn on the Loop-X and set-up the patient, (2) start the Linux server providing AI-based patient-specific imaging and communication with the Loop-X, (3) connect the microphone to the MacBook laptop, and (4) start the laptop server, which runs the Whisper ASR model and communicates with GPT-4. From there the system is ready to relay commands to the Loop-X.

During the study, an attending orthopedic surgeon interfaced with the system solely via voice. While interfacing with the Loop-X, the surgeon aimed to align a surgical pointer with the S2 corridor, a narrow bony corridor often used for stabilize sacroiliac fractures. Alternating between inlet and outlet views, he positioned the pointer, which was held in place with a passive positioning arm to allow for trajectory verification. In total, 33 X-ray images were acquired during the study, and successful placement along the S2 corridor was evaluated with a CT scan by the Loop-X.

Figure 3 shows our system in action, moving from an inlet view to a patient-specific AP view. Note that because the Loop-X source and detector move independently, it is possible to take non-isocentric images that are post-processed to correct for skew, accounting for the angled detector. In this study, the system requested clarifications for commands it did not understand, which could be avoided by specifying the amount of movement desired for fine-grained adjustments. The LLM also responded with `question` messages in the case of a transcription error, such as confusing “Inlet” for “In that.” Once the user provided a clarification, the LLM was able to compensate for similar errors.

Discussion

Despite their complexity, higher-level capabilities of surgical assistance systems promise real benefits that are worth using. They can streamline procedures, reduce risk, and improve patient outcomes. Patient-specific imaging, for example, can reduce the number of images acquired in the process of fluoro-hunting [2, 27], reducing the radiation exposure for patients and staff as well as the time under general anesthesia. Currently, high-level functions can only be accessed through graphical user interfaces (GUIs), which are complicated and differ among manufacturers. Although further investment in GUI development can improve these interfaces, they can never be simpler than the functions they control, so that in practice, device-specific power users arise who are familiar with each device’s quirks and higher-level capabilities. In effect, *these users already provide a natural language interface to higher-level functions*, albeit an indirect one through regular conversation, which LLMs have begun to emulate. With the recent advances in language models’ capabilities, natural language interfaces are being proposed at a rapid pace for human–robot collaboration in complex tasks [4, 5], because they enable individuals to access sophisticated capabilities quickly and while focusing on performing tasks themselves.

There are immediate opportunities for improving the specific system presented here. For instance, although we have demonstrated that few-shot learning is capable of supporting a usable system, fine-tuning an LLM to adhere to our protocol would avoid long message exchanges. In our current system, the full instruction set including examples is included in every exchange with the LLM, increasing cost and limiting memory. A fine-tuned LLM would support exchanges that persist throughout the surgery and would enable more sophisticated protocols with longer instruction sets, improving the robustness of low-level control by relegating pose changes to an underlying, verifiable function. At the same time, a specialized language model would not need the same computational resources. The resources required to run general-purpose LLMs, such as GPT-4, translate to significant environmental impact where carbon-free energy sources are unavailable [31]. Utilizing smaller models with augmented capabilities, such as access to code environments or real-world information, may be sufficient for many specialized tasks, reducing their energy consumption [32].

At a higher level, there are significant challenges remaining before an LLM-based interface can be incorporated into the healthcare setting [18]. The protection of patient privacy is paramount, and steps must be taken to ensure that recorded conversations in the OR, which very often include patient data, are processed in a secure, transparent manner. Furthermore, open questions persist regarding the quality and ethical use of data used to train state-of-the-art LLMs, which tend to ingest as much of the internet as possible [31]. The entire Wikipedia corpus, for example, arose because of the committed effort of volunteers and charitable donations, but it has proven to be an invaluable resource for preventing hallucination in many domains [33]. How to honor the efforts of millions of content creators whose efforts have enabled LLMs to exist at all has yet to be determined, but it is no less relevant in this context than in any other.

Another challenge is the classification of commands from cluttered audio streams. Like voice interfaces on smart-phones and speakers, an X-ray language interface must be able to distinguish between intentional commands and unrelated speech, including essential communication among the surgical team. The idea of an OR which is silent except for the surgeon’s commands to an automated system is not practical nor desirable. Wake words like “Alexa” confront this problem by making the distinction explicit, but current standard practice in the OR, in which surgeons do interface with X-ray technologists, residents, and other support staff via natural language, suggests that more seamless solutions are possible. Technologists regularly distinguish commands which are intended for the X-ray system without being addressed by name, and recent work shows LLMs are capable of the same task-aware extraction [30].

Incorporating advances in human–robot interaction (HRI) can ensure that such an intelligent assistance system integrates smoothly into the operational workflow. For instance, biases in available training data can lead to inaccurate transcriptions for non-native English speakers or other under-represented populations [34], and it is well known that LLMs are prone to hallucination, which in this context may produce content that adheres to protocol but results in incorrect or unsafe actions [31, 35]. Intelligent systems that acknowledge mistakes and double-check safety-critical operations can maintain trust with their human collaborators [5, 36, 37]. In the future OR, an X-ray language interface may be one component of a broader human–robot collaborative system.

Conclusion

We have presented a natural language interface for controlling robotic X-ray systems, embodied in a fully integrated system. We include an interface to high-level functionality in the form of patient-specific imaging of the pelvis as well as enabling low-level control of a fully robotic X-ray system, the Brainlab Loop-X device. We rely on a commercially available LLM to perform few-shot interpretation of spoken commands according to a specified protocol. We demonstrate the effectiveness of our system for supporting pelvic trauma surgery in a user study with an attending physician, interacting with the system solely through spoken commands. As these technologies continue to develop, we envision natural language interfaces will serve as a unified entry point into complex capabilities for robotic assistance systems in the OR.

Funding

This work was supported by the Link Foundation Fellowship for Modeling, Training, and Simulation; the NIH under Grant No. R21EB028505, the NSF under Award No. 2239077, and Johns Hopkins University Internal Funds.

Appendix: Full instructions

The following instructions are provided to the LLM in every interaction.

```
Please only receive and respond to messages in the following single-line
format with no additional text. Do not explain your reasoning except in a
summary message when requested.
Messages are a single line with the following structure:
“‘
message_type;…
“‘
where message_type={ command | question | action | clarification | accepted
| summary }
The ‘command ‘ message type has this structure:
“‘
command; patient_head_first; patient_pose; surgeon_side; source_angle;
detector_angle;tilt;x;y;yaw;user_input
“‘
where patient_head_first={ True, False }, patient_pose={ supine | prone },
and surgeon_side={ left | right }. source_angle, detector_angle, tilt, x, y,
and yaw are all floats giving the loopx pose. user_input is a string.
You have two options for a response. Please always reply with one of the
following:
1. If you are confused, send a ‘ question ‘ message:
“‘
question; question_text
“‘
where question_text is a string.
2. If you are ready, reply with an ‘ action ‘ message:
“‘
action; action_type;…
“‘
where action_type={ pose | shot | view | none }
“‘
– A ‘pose ‘ action moves the loopx to the given pose:
“‘
action; pose; source_angle; detector_angle; tilt;x;y; yaw
“‘
– A ‘shot ‘ action is has no extra parameters and just takes a shot.
“‘
```

```
action; shot
“‘
– A ‘view ‘ action requests a specific view:
“‘
action; view; view_name; anatomy
“‘
where view_name={ ap | lateral | inlet | outlet | oblique_left |
oblique_right | teardrop }
and anatomy={ hip_left | hip_right | femur_left | femur_right | sacrum
| vertebra_L5 | pelvis | s1_left | s1_right | s1 | s2 | ramus_left |
ramus_right | teardrop_left | teardrop_right }
– ‘none ‘ performs no action.
The user will then send one of the following:
If a question was sent, the user will send a ‘ clarification ‘ message:
“‘
clarification; clarification_text
“‘
where clarification_text is a string with the explanation. You can then
reply with an ‘action ‘ message as defined above.
Otherwise, the user might send a ‘ finished ‘ message:
“‘
finished; accepted
“‘
where accepted={ True | False }. You should respond with a ‘summary
‘ message giving your reasoning
“‘
reasoning; reasoning_text
“‘
where reasoning_text is a string with your reasoning. Only send a ‘ summary
‘ message when the user sends a ‘ finished ‘ message. Otherwise, no
explanation is needed.
```

Each interaction also includes a set of 35 example episodes, such as:

```
command; True; supine; patient_right; 180.0;0.0;0.0;0.0;0.0;0.0; move the
detector clockwise 20 degrees
action; pose; 180.0; −20.0;0.0;0.0;0.0;0.0
summary; negative rotation corresponds to clockwise direction, so subtract
20 from the given detector angle
command; True; supine; patient_right; 180.0;0.0;0.0;0.0;0.0;0.0; go to an AP
view
action; view; ap; pelvis
summary; the surgeon requested a patient – specific AP view. Since they didn
‘t specify the anatomy, assume they want the pelvis.
```

```
command; True; supine; patient_right; 180.0;0.0;0.0;0.0;0.0;0.0; go to an AP
shot of the S2
action; view; ap; s2
summary; even though the surgeon said “shot”, they meant “view” in this
context
```

References

1. Killeen BD, Gao C, Oguine KJ, Darcy S, Armand M, Taylor RH, Osgood G, Unberath M (2023) An autonomous X-ray image acquisition and interpretation system for assisting percutaneous pelvic fracture fixation. Int J CARS 18(7):1201–1208. 10.1007/s11548-023-02941-y
2. Kausch L, Thomas S, Kunze H, Privalov M, Vetter S, Franke J, Mahnken AH, Maier-Hein L, Maier-Hein K (2020) Toward automatic C-arm positioning for standard projections in orthopedic surgery. Int J CARS 15(7):1095–1105. 10.1007/s11548-020-02204-0
3. Hendrix G (1982) Natural-language interface. Am J Comput Linguist 8(2):56–61
4. Zhang C, Chen J, Li J, Peng Y, Mao Z (2023) Large language models for human–robot interaction: a review. Biomim Intell Robot 3(4):100131. 10.1016/j.birob.2023.100131
5. Ye Y, You H, Du J (2023) Improved trust in human–robot collaboration with ChatGPT. IEEE Access 11:55748–55754. 10.1109/ACCESS.2023.3282111
6. Driess D, Xia F, Sajjadi MSM, Lynch C, Chowdhery A, Ichter B, Wahid A, Tompson J, Vuong Q, Yu T, Huang W, Chebotar Y, Sermanet P, Duckworth D, Levine S, Vanhoucke V, Hausman K, Toussaint M, Greff K, Zeng A, Mordatch I, Florence P (2023) PaLM-E: an embodied multimodal language model. arXiv. 10.48550/arXiv.2303.03378arXiv:2303.03378
7. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D (2020) Language models are few-shot learners. arXiv 10.48550/arXiv.2005.14165arXiv:2005.14165
8. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I (2022) Robust speech recognition via large-scale weak supervision. arXiv (2022). 10.48550/arXiv.2212.04356arXiv:2212.04356
9. OpenAI:GPT-4 Technical Report. arXiv (2023). 10.48550/arXiv.2303.08774arXiv:2303.08774
10. Shridhar M, Manuelli L, Fox D (2021) CLIPort: what and where pathways for robotic manipulation. arXiv. 10.48550/arXiv.2109.12098arXiv:2109.12098
11. Hundt A, Killeen B, Greene N, Wu H, Kwon H, Paxton C, Hager GD (2020) “Good Robot!”: efficient reinforcement learning for multi-step visual tasks with sim to real transfer. IEEE Robot Autom Lett 5(4):6724–6731. 10.1109/LRA.2020.3015448
12. Hundt A, Murali A, Hubli P, Liu R, Gopalan N, Gombolay M, Hager GD (2022) Good robot! Now watch this!”: repurposing reinforcement learning for task-to-task transfer. In: Conference on robot learning. PMLR, pp 1564–1574. https://proceedings.mlr.press/v164/hundt22a.html
13. Tellex S, Gopalan N, Kress-Gazit H, Matuszek C (2020) Robots that use language. Annu Rev Control Robot Autonom Syst 3(1):25–55. 10.1146/annurev-control-101119-071628
14. Lynch C, Wahid A, Tompson J, Ding T, Betker J, Baruch R, Armstrong T, Florence P (2023) Interactive language: talking to robots in real time. IEEE Robot Autom Lett 66:1–8. 10.1109/LRA.2023.3295255
15. Hazlehurst B, Sittig DF, Stevens VJ, Smith KS, Hollis JF, Vogt TM, Winickoff JP, Glasgow R, Palen TE, Rigotti NA (2005) Natural language processing in the electronic medical record: assessing clinician adherence to tobacco treatment guidelines. Am J Prev Med 29(5):434–439. 10.1016/j.amepre.2005.08.007 [PubMed: 16376707]
16. Huang H, Zheng O, Wang D, Yin J, Wang Z, Ding S, Yin H, Xu C, Yang R, Zheng Q, Shi B (2023) ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. Int J Oral Sci 15(29):1–13. 10.1038/s41368-023-00239-y [PubMed: 36593250]

17. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW (2023) Large language models in medicine. Nat Med 29(8):1930–1940. 10.1038/s41591-023-02448-8 [PubMed: 37460753]

18. Meskó B, Topol EJ (2023) The imperative for regulatory oversight of large language models (orgenerative AI) in healthcare. npj Digit Med 6(120):1–6. 10.1038/s41746-023-00873-0 [PubMed: 36596833]

19. Killeen BD, Cho SM, Armand M, Taylor RH, Unberath M (2023) In silico simulation: a key enabling technology for next-generation intelligent surgical systems. Prog Biomed Eng 5(3):032001. 10.1088/2516-1091/acd28b

20. Bier B, Unberath M, Zaech J-N, Fotouhi J, Armand M, Osgood G, Navab N, Maier A (2018) X-ray-transform invariant anatomical landmark detection for pelvic trauma surgery. In: Medical image computing and computer assisted intervention—MICCAI 2018. Springer, Cham, Switzerland, pp 55–63. 10.1007/978-3-030-00937-3_7

21. Liu W, Wang Y, Jiang T, Chi Y, Zhang L, Hua X-S (2020) Landmarks detection with anatomical constraints for total hip arthroplasty preoperative measurements. In: Medical image computing and computer assisted intervention—MICCAI 2020. Springer,Cham, Switzerland, pp 670–679. 10.1007/978-3-030-59719-1_65

22. Gao C, Killeen BD, Hu Y, Grupp RB, Taylor RH, Armand M, Unberath M (2023) Synthetic data accelerates the development of generalizable learning-based algorithms for X-ray image analysis. Nat Mach Intell 5(3):294–308. 10.1038/s42256-023-00629-1 [PubMed: 38523605]

23. Kügler D, Sehring J, Stefanov A, Stenin I, Kristin J, Klenzner T, Schipper J, Mukhopadhyay A (2020) i3PosNet: instrument pose estimation from X-ray in temporal bone surgery. Int J CARS 15(7):1137–1145. 10.1007/s11548-020-02157-4

24. Killeen BD, Chakraborty S, Osgood G, Unberath M (2022) Toward perception-based anticipation of cortical breach during K-wire fixation of the pelvis. In: Proceedings Volume 12031, medical imaging 2022: physics of medical imaging. SPIE, pp 410–415. 10.1117/12.2612989

25. Killeen BD, Zhang H, Mangulabnan J, Armand M, Taylor RH, Osgood G, Unberath M (2023) Pelphix: surgical phase recognition from X-ray images in percutaneous pelvic fixation. arXiv. 10.48550/arXiv.2304.09285arXiv:2304.09285

26. Arbogast N, Kurzendorfer T, Breininger K, Mountney P, Toth D, Narayan SA, Maier A (2019) Workflow phase detection in fluoroscopic images using convolutional neural networks. In: Bildverarbeitung Fr die Medizin 2019. Springer, Wiesbaden, Germany, pp 191–196. 10.1007/978-3-658-25326-4_41

27. Kausch L, Thomas S, Kunze H, Norajitra T, Klein A, El Barbari JS, Privalov M, Vetter S, Mahnken A, Maier-Hein L, Maier-Hein KH (2021) C-arm positioning for spinal standard projections in different intra-operative settings. In: Medical image computing and computer assisted intervention —MICCAI 2021. Springer, Cham, Switzerland, pp 352–362. 10.1007/978-3-030-87202-1_34

28. Grupp RB, Unberath M, Gao C, Hegeman RA, Murphy RJ, Alexander CP, Otake Y, McArthur BA, Armand M, Taylor RH (2020) Automatic annotation of hip anatomy in fluoroscopy for robust and efficient 2D/3D registration. Int J Comput Assist Radiol Surg 15(5):759–769. 10.1007/s11548-020-02162-7.arXiv:3233.3361 [PubMed: 32333361]

29. Seshamani S, Chintalapani G, Taylor R (2011) Iterative refinement of point correspondences for 3D statistical shape models. In: Medical image computing and computer-assisted intervention—MICCAI 2011. Springer, Berlin, Germany, pp 417–425. 10.1007/978-3-642-23629-7_51

30. Cámbara G, López F, Bonet D, Gómez P, Segura C, Farrús M, Luque J (2022) TASE: task-aware speech enhancement for wake-up word detection in voice assistants. Appl Sci 12(4):1974. 10.3390/app12041974

31. Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: Can language models be too big? xn–st9h. In: FAccT’21: proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, NY, USA, pp 610–623. 10.1145/3442188.3445922

32. Mialon G, Dessì R, Lomeli M, Nalmpantis C, Pasunuru R, Raileanu R, Rozière B, Schick T, Dwivedi-Yu J, Celikyilmaz A, Grave E, LeCun Y, Scialom T (2023) Augmented language models: a survey. arXiv. 10.48550/arXiv.2302.07842.arXiv:2302.07842

33. Semnani S, Yao V, Zhang H, Lam M (2023) WikiChat: stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia. ACL Anthol. 10.18653/v1/2023.findings-emnlp.157

34. Sloos M, Ariza García A, Andersson A, Neijmeijer M (2019) Accent-induced bias in linguistic transcriptions. Lang Sci 76:101176. 10.1016/j.langsci.2018.06.002

35. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Chen Q, Peng W, Feng X, Qin B, Liu T (2023) A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. arXiv. 10.48550/arXiv.2311.05232arXiv:2311.05232

36. Chen M, Nikolaidis S, Soh H, Hsu D, Srinivasa S (2020) Trust-aware decision making for human–robot collaboration: model learning and planning. J Hum–Robot Interact 9(2):1–23. 10.1145/3359616

37. Cuadra A, Li S, Lee H, Cho J, Ju W (2021) My bad! Repairing intelligent voice assistant errors improves interaction. Proc ACM Hum–Comput Interact 5(CSCW1):1–24. 10.1145/3449101 [PubMed: 36644216]

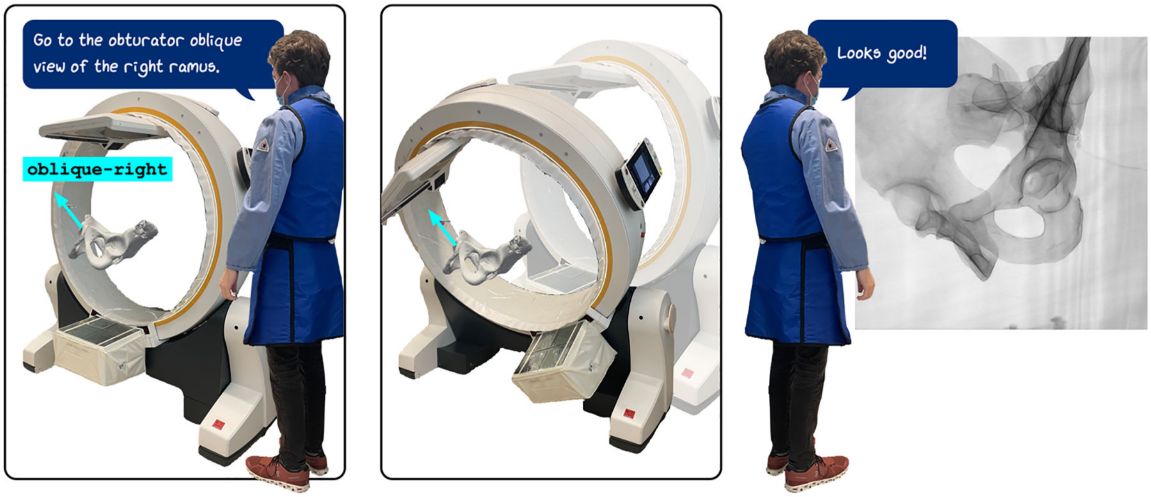

Fig. 1.

Our system allows a user to control independent axes of a robotic X-ray system, the Brainlab Loop-X, using spoken natural language. Patient-specific views, such as the obturator oblique, are supported via automated image analysis

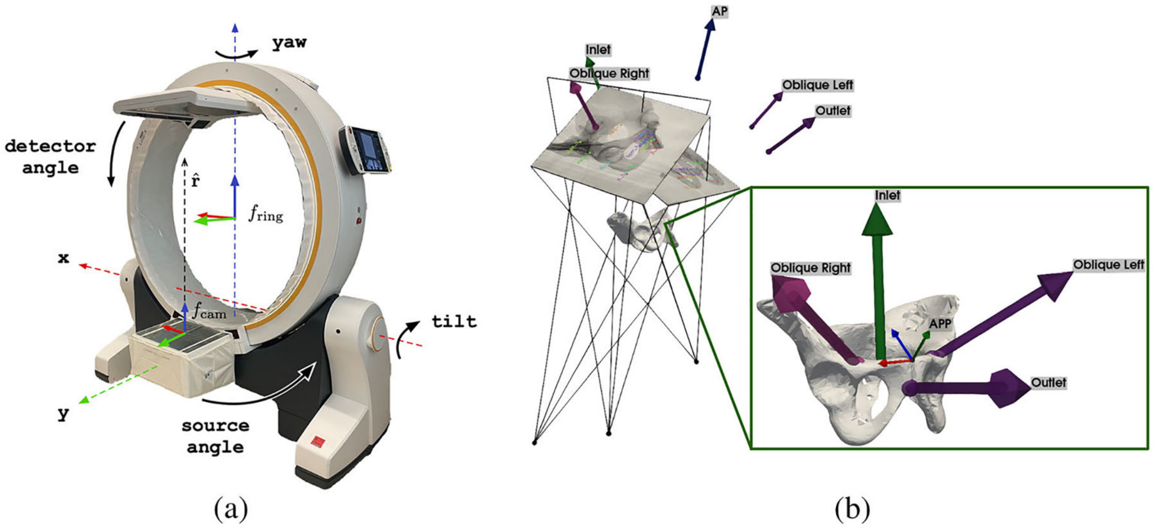

Fig. 2.

2a The Brainlab Loop-X X-ray system comprises six independent axes, with independent source and detector movement. It is capable of aligning itself to achieve precise views along a desired principle ray $\hat{\mathbf{r}}$ specified in f_{ring}. 2 To determine patient-specific views, we rely on automated landmark detection to determine the patient pose with respect to a statistical shape model (SSM). The standard view directions are defined with respect to the SSM following [25]

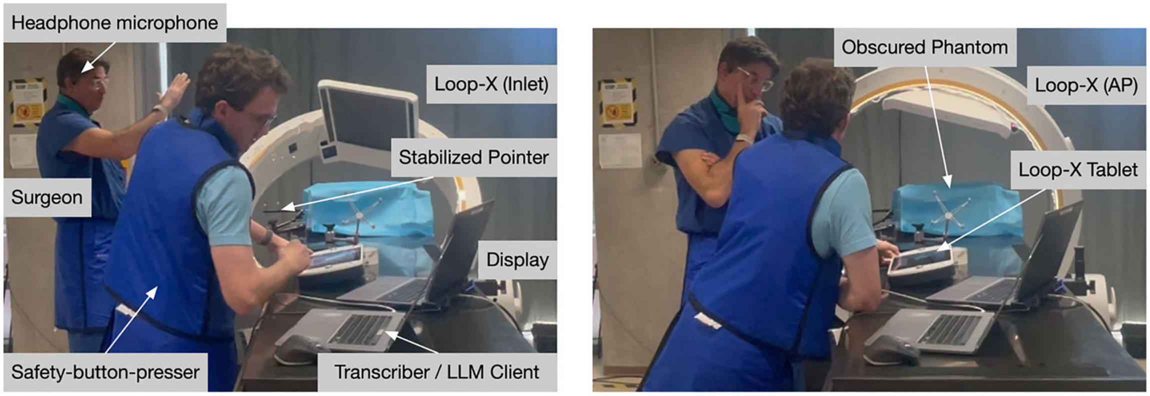

Fig. 3.

In a phantom study, an attending orthopedic surgeon was able to emulate X-ray-guided percutaneous fixation of the sacroiliac joint by interfacing with the Loop-X X-ray system solely through voice commands. For safety reasons, movements and acquisitions require physical confirmation on the control tablet, but at no point were any commands initiated from the built-in interface

Table 1

System actions given responses

Base prompt	Variations	Anticipated response	Correct (%)
“More outlet.”	20	Angular rotation toward superior	95
“More inlet.”	24	Angular rotation toward inferior	83.3
“Roll back.”	12	Orbital rotation toward surgeon	100
“Roll over.”	12	Orbital rotation away from surgeon	100
“Go north.”	44	Movement in the superior direction	100
“Go south.”	12	Movement in the inferior direction	91.7
“Push in.”	16	Lateral movement toward surgeon	100
“Push out.”	12	Lateral movement away from surgeon	100
“Rotate clockwise.”	8	Rotation on the floor	100
“Inlet view.”	8	Patient-specific view	100
“Take a shot!”	44	Acquire an image	100
Total	212		97.2

Table 2

Physician rating for patient-specific views

View	Avg. rating
Anteroposterior (AP)	7.7/10
Lateral	8.3/10
Inlet (R. Ramus)	9.3/10
Outlet	6.7/10
OO (L. Ramus)	6.7/10
OO (R. Ramus)	6.7/10