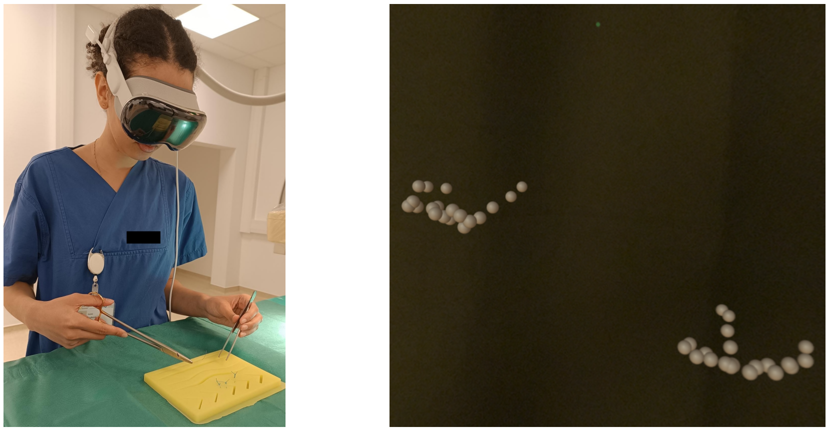
# Learning Surgical Skills by Imitation: A Work-in-Progress XR Method Using Expert Hand Trajectories

Pooja Salunke[*] Thore Menk[†] Agunda Chekhoeva[‡] Artur Lichtenberg[§]
Alina Huldtgren[¶] Hug Aubin[‖] Falko Schmid[**]

[*‡§‖**]Digital Health Lab Düsseldorf, University Hospital Düsseldorf, Germany

[†¶]University of Applied Sciences Düsseldorf, Düsseldorf, Germany


[redacted]

(a) XR-based suturing practice setup

(b) Rigged hand in the Apple Vision Pro

Figure 1: Work-in-progress mixed reality system for cardiac suturing training. (a) A trainee wearing the Apple Vision Pro practices suturing on a silicone phantom. (b) Headset view showing a rigged hand model driven by 21 tracked landmarks.

## Abstract

Critical cardiac suturing techniques are difficult to scale due to reliance on expert supervision and limited hands-on training. Without timely feedback, trainees may develop suboptimal motor patterns that compromise procedural quality. We present an Extended Reality (XR) pipeline to capture and visualize expert hand motions for autonomous training. Expert demonstrations are recorded via RGB-D camera, and 3D hand landmarks are reconstructed through pose estimation with depth integration. These trajectories are processed offline and integrated into a Unity-based environment deployed to the Apple Vision Pro (AVP). Preliminary results demonstrate successful capture and replay of suturing trajectories using a rigged hand model. Ongoing work includes system calibration, ethics approval, and user studies to evaluate guidance clarity. This work establishes the feasibility of AVP-based high-fidelity XR for surgical tasks, laying the foundation for scalable, supervision-light skill acquisition.

**Index Terms:** Mixed Reality, Extended Reality Surgical Training, Surgical Skill Assessment, Apple Vision Pro.

## 1 Introduction

In cardiac surgery, numerous operative sub-steps, particularly precise suturing techniques, are critical to procedural success. The acquisition of fine-motor skills typically requires guidance and feedback from experienced surgeons. Due to time constraints, staff shortages, and high surgical workload, continuous individual training is only partially feasible, in particular with highly specific suturing techniques, as e.g. found in cardiac surgery. Without immediate expert feedback, trainees may develop suboptimal techniques that are difficult to correct later and could potentially compromise patient safety. This creates a gap between the operative expertise of experienced surgeons and the ability to convey these skills to trainees in a precise, objective, and reproducible manner.

Immersive technologies such as XR offer a potential solution to this training gap. They can provide risk-free practice of complex surgical motions. In parallel, advances in Artificial Intelligence (AI)-based video analysis are increasingly capable of automatically capturing hand movements. Building on these advances, XR systems that model expert surgical hand motions as traceable paths could support guided practice and evaluation in the future.

## 2 Related Work

XR technologies, including Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR), have been increasingly explored for surgical training, enabling safe, repeatable practice and objective skill assessment [8]. XR devices support both technical (motor dexterity) and non-technical (decision-making) skills [4].

Recent reviews show that these technologies can measurably improve hand–eye coordination, spatial understanding, and operative performance in surgical training [13]. Hand-tracking and skeleton-based motion capture, using approaches such as MediaPipe, enable monitoring of hand posture and surgical instrument movements [1].

[*] e-mail: poojachandrakant.salunke@med.uni-duesseldorf.de
[†] e-mail: thore.menk@study.hs-duesseldorf.de
[‡] e-mail: agunda.chekhoeva@med.uni-duesseldorf.de
[§] e-mail: artur.lichtenberg@med.uni-duesseldorf.de
[¶] e-mail: alina.huldtgren@hs-duesseldorf.de
[‖] e-mail: hug.aubin@med.uni-duesseldorf.de
[**] e-mail: falko.schmid@med.uni-duesseldorf.de

However, most current XR systems provide only coarse cues, such as tool-tip visualizations, and lack fine-grained hand-level guidance required for complex surgical tasks. Trajectory-based skill assessment is well established for evaluation [6], but it typically focuses on instrument trajectories and is used only indirectly to assess motor skill acquisition.

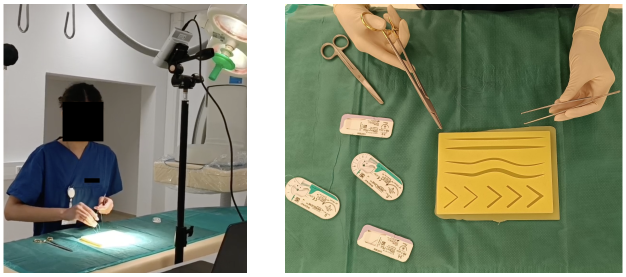
High-end spatial computing platforms, such as the AVP [2], which is a video-see-through (VST) XR device, provide high-fidelity immersive experiences while maintaining real-world visual perception [5]. Studies have explored the AVP for medical training and visualization, reporting realistic 3D representations and minimal user fatigue [5, 11, 12]. Javaheri et al. [9] conducted a controlled evaluation of the AVP for surgical suturing, demonstrating its feasibility and usability, though not as a training tool. Similarly, Microsoft HoloLens has been widely used in medical applications [7]; however, limitations remain in providing reusable expert motion trajectories and detailed hand-level guidance.

Overall, these studies highlight a gap in XR-based surgical training, particularly in providing expert-modeled hand motions that trainees can study and practice independently.

## 3 System Overview

### 3.1 Data Acquisition

The preliminary experiment was conducted at (prepared for blind review). Suturing demonstrations were performed by an experienced cardiac surgeon to capture expert hand motions for subsequent training purposes.



(a) Recording with RGB-D camera (b) Workspace with suture training phantom

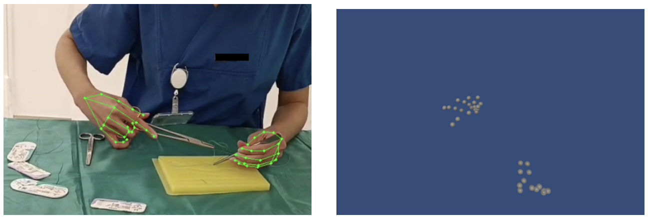
Figure 2: The recording environment showing RGB-D camera placement for expert hand capture and the workspace with the suturing phantom used to acquire hand motions.

Recordings were obtained using a single Orbbec Femto Mega depth camera positioned to capture hand motions from a frontal–oblique perspective (see Fig. 2a). The demonstrations were carried out in a simulation setting using a silicone suturing pad, surgical instruments, and clinical lighting. The recorded demonstrations include two suturing techniques: simple interrupted sutures (SIS), which involves basic hand motions, and continuous subcuticular sutures (CSS), which require comparatively more complex movements. Both techniques were performed using 3-0 polypropylene and standard surgical instruments, including needle holder, forceps, and suture scissors (see Fig. 2b). Recordings were performed both with and without gloves and under varying surgical lighting conditions to assess the robustness of hand tracking.

While Javaheri et al. [9] reported blurred vision and depth perception challenges with the AVP, we similarly observed that depth accuracy is highly sensitive to eye-to-object distance at close range; consequently, the AVP was not used for the preliminary expert data acquisition to ensure high-fidelity motion capture. In the context of trainee imitation, close-range viewing may further exacerbate these depth perception issues. To mitigate this effect in future iterations of the system, we plan to incorporate a digital magnification step within the XR application. This would allow trainees to inspect fine hand motions while maintaining a sufficient eye-to-object distance, thereby promoting a more stable viewing position and reducing depth-related visual distortions.

### 3.2 Hand Pose Extraction and Depth Integration

Expert demonstrations are processed offline to generate smooth 3D hand trajectories suitable for immersive training. Hand pose estimation is performed using MediaPipe Hands [14], which detects 21 hand landmarks (finger joints + palm) from a single RGB frame (see Fig. 3a), providing normalized 2D image coordinates (x, y) and a relative depth value (z) for each joint, where depth is expressed relative to the wrist landmark. The 2D pixel coordinates are clipped to the image frame to prevent out-of-bound values. The underlying training datasets include diverse lighting conditions and hand appearances to mitigate tracking bias and improve generalization across diverse skin tones.



(a) Hand landmarks detected using MediaPipe (b) Reconstructed 3D hand joints visualized in Unity

Figure 3: Detected hand landmarks are integrated with camera depth information and visualized in Unity after offline processing, providing motion data for the AVP application.

To reduce measurement noise and improve the camera's depth map reliability, a local median filter is applied, which removes spurious depth values. Metric 3D joint positions are subsequently computed by combining the filtered depth measurements with the 2D landmarks from MediaPipe. If a joint lacks a valid depth measurement, its 3D position is estimated by anchoring to the wrist depth from the RGB-D camera and scaling the joint's relative z-coordinate provided by MediaPipe. In cases of temporary occlusion or detection dropouts, trajectory gaps are interpolated using a constant velocity motion model [3], which assumes that the joint's velocity changes gradually between consecutive frames.

### 3.3 Replay Pipeline and Vision Pro Deployment

The processed 3D hand trajectories are serialized into Newline Delimited JSON (NDJSON) files, where each entry corresponds to the 21 hand landmarks of a single frame. This format facilitates sequential, low-latency data streaming into the Unity development environment (see Fig. 3b), which serves as the primary rendering engine for the AVP application. In Unity, a dedicated playback component parses the data to drive a rigged hand model. Additional smoothing filters are applied if necessary to reduce any residual jitter from capture or processing.

To maintain versatility across different recording conditions, the system dynamically maps and scales joint positions based on data availability. When RGB-D depth integration is successful, true-scale 3D coordinates are used, allowing the virtual hand to be rendered at a 1:1 scale relative to the real-world suturing pad in the AVP passthrough view. In scenarios where depth data is missing or invalid, the system reverts to MediaPipe's normalized landmarks. These values are uniformly scaled to match the anatomical proportions of the rigged model, ensuring a visually consistent representation even in the absence of absolute metric data. This pipeline ensures that the rigged hand reproduces the captured expert trajectory, preserving spatial and temporal consistency before deployment in the AVP headset.

## 4 Planned Evaluation

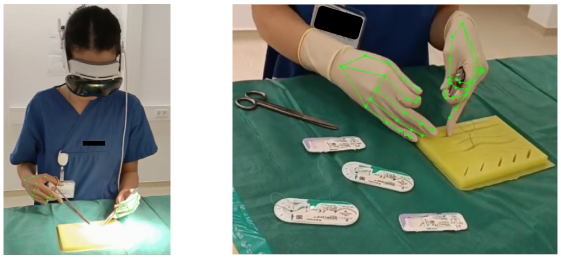
Preliminary expert suturing demonstrations have been recorded, and the resulting 3D hand trajectories have been successfully replayed using rigged hand models on the AVP. Initial testing confirms the technical feasibility of the data pipeline, from motion acquisition to spatial visualization within the headset.

Ongoing work involves system calibration and securing ethics approval for exploratory user studies with surgical trainees. These studies will focus on system feasibility, focusing on the perceived usefulness of 3D guidance and the clarity of the visual cues. This qualitative phase will ensure the system meets the practical needs of surgical education before proceeding to quantitative assessments of skill acquisition and training effectiveness.

## 5 Discussion and Limitations

This work shows that captured expert hand motions have the potential to serve as spatial guidance in a Vision Pro–based XR environment, enabling trainees to practice by imitation without continuous supervision. The prototype is intended for evaluation in a simulated training setting rather than intra-operative scenarios.

Current limitations include offline-only guidance and dependence on hand-pose estimation and depth alignment quality. During preliminary recordings, hand joints were reliably detected even when gloves were worn without markers (see Fig. 4b), although occasional tracking glitches occurred. In the operating room procedures are typically performed under intense surgical lighting to ensure clear visualization of fine anatomical structures and suture placement. To support realistic training, the XR environment can be combined with these lighting conditions (see Fig. 4a), as performing precise tasks under such illumination is essential for developing a better hand–eye coordination and procedural fidelity.



(a) Tracking under surgical lights

(b) Tracking with standard surgical gloves

Figure 4: Demonstration of hand tracking robustness showing stability under intense surgical glare and successful 21-joint landmark detection while wearing standard surgical gloves.

Prior work, such as by Karelin et al. [10], demonstrates that the Vision Pro can achieve sub-centimeter hand tracking accuracy suitable for immersive XR applications. This suggests that it could be used to track and record trainee motions for direct trajectory comparison against expert baselines. However, maintaining optimal headset positioning and physical lighting remains critical to minimizing depth errors and ensuring consistent alignment for objective assessment.

## 6 Conclusion and Future Work

This work explores hand-level XR guidance for cardiac surgical training, allowing repeated access to expert demonstrations and hands-on practice in a controlled, immersive environment. Next steps include enabling trainees to follow expert hand trajectories interactively in the AVP headset. The system will be extended to additional fine motor tasks. Planned user studies will evaluate skill development and training efficacy. Overall, this work indicates the feasibility of Vision Pro–based XR for repeatable, expert-informed training without continuous supervision. Future work may explore real-time guidance and adaptive feedback tailored to individual learner performance.

## Acknowledgments

We wish to thank the Susanne Bunnenberg Herzstiftung for its generous support, which made this research possible.

## References

[1] G. Amprimo, G. Masi, G. Pettiti, G. Olmo, L. Priano, and C. Ferraris. Hand tracking for clinical applications: validation of the Google MediaPipe hand framework. *arXiv preprint arXiv:2305.14133*, 2023. 1

[2] Apple Inc. Apple Vision Pro. https://www.apple.com/apple-vision-pro/, 2024. Accessed: 2026-02-19. 2

[3] N. L. Baisa. Derivation of a constant velocity motion model for visual tracking. *arXiv preprint arXiv:2005.00844*, 2020. 2

[4] T. R. Coles, D. Meglan, and N. W. John. The role of haptics in medical training simulators: A survey of the state of the art. *IEEE Transactions on Haptics*, 4(1):51–66, 2011. doi: 10.1109/TOH.2010.19 1

[5] J. Egger, C. Gsaxner, G. Luijten, J. Chen, X. Chen, J. Bian, J. Kleesiek, and B. Puladi. Is the Apple Vision Pro the ultimate display? A first perspective and survey on entering the wonderland of precision medicine. *JMIR Serious Games*, 12:e52785, 2024. doi: 10.2196/52785 2

[6] I. Funke, S. T. Mees, J. Weitz, and S. Speidel. Video-based surgical skill assessment using 3D convolutional neural networks. *arXiv preprint arXiv:1903.03043*, 2019. 2

[7] C. Gsaxner, J. Li, A. Pepe, Y. Jin, J. Kleesiek, D. Schmalstieg, and J. Egger. The HoloLens in medicine: A systematic review and taxonomy. *Medical Image Analysis*, 85:102757, 2023. doi: 10.1016/j.media.2023.102757 2

[8] S. Jain, S. Lee, S. R. Barber, and Y.-J. Son. Surgical proficiency assessment using Virtual Reality (VR)-based hybrid simulation for minimally invasive procedures. *Computers & Education: X Reality*, 8:100128, 2025. doi: 10.1016/j.cexr.2025.100128 1

[9] H. Javaheri, V. Fortes Rey, P. Lukowicz, G. Stavrou, J. Karolus, and O. Ghamarnejad. Assessing the feasibility of using Apple Vision Pro while performing medical precision tasks: Controlled user study. *JMIR XR Spatial Computing*, 2:e73574, 2025. doi: 10.2196/73574 2

[10] A. Karelin, D. Brazhenko, G. Kliukovkin, and Y. Chernenko. Real-time hand tracking and collision detection for immersive mixed-reality boxing training on Apple Vision Pro. *Sensors (Basel)*, 25(16):4943, 2025. doi: 10.3390/s25164943 3

[11] M. Masalkhi, E. Waisberg, J. Ong, et al. Apple Vision Pro for ophthalmology and medicine. *Annals of Biomedical Engineering*, 51:2643–2646, 2023. doi: 10.1007/s10439-023-03283-1 2

[12] J. Olexa, A. Trang, J. Cohen, et al. The Apple Vision Pro as a neurosurgical planning tool: A case report. *Cureus*, 16(2):e54205, 2024. doi: 10.7759/cureus.54205 2

[13] E. Toni, E. Toni, M. Fereidooni, and H. Ayatollahi. Acceptance and use of extended reality (XR) in surgical training: an umbrella review. *Systematic Reviews*, 13(1):299, 2024. doi: 10.1186/s13643-024-02723-w 1

[14] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann. MediaPipe hands: On-device real-time hand tracking. In *Proc. CVPR Workshop on Computer Vision for Augmented and Virtual Reality*, pp. 1–5, June 2020. 2