communications engineering

A Nature Portfolio journal

Article

https://doi.org/10.1038/s44172-026-00692-7

Real-time 3D ultrasound in augmented reality accelerates training and narrows novice–expert performance gaps

Jason F. Hou[1], Shrihari Viswanath[1], Cinay Dilibal[1,2], Bowen Wu[3], Tanisha Shende[1,4] & Canan Dagdeviren[1]

Ultrasound imaging requires users to infer three-dimensional anatomy from two-dimensional slices, imposing steep training demands that limit broader adoption. Here we present AR-VIU, a mixed-reality platform that streams real-time volumetric ultrasound as point-cloud renderings into an augmented-reality headset with true-scale spatial registration. To isolate the contributions of volumetric imaging and immersive display, we tested four conditions—two-dimensional imaging on a screen, two-dimensional imaging in augmented reality, three-dimensional imaging on a screen, and three-dimensional imaging in augmented reality—in a controlled study with 18 participants (9 novices, 9 experts). Participants performed object recognition and localization tasks. The augmented-reality volumetric system was associated with the highest accuracy, lowest variability, and near-elimination of the novice-expert performance gap. These results demonstrate technical feasibility for real-time three-dimensional ultrasound in mixed reality and establish an evaluation framework for perceptual and cognitive performance in clinically relevant scenarios, with near-term applications in training and education.

Ultrasound imaging is one of the most widely used medical imaging modalities due to its portability, safety, and cost-effectiveness. However, its utility is highly dependent on operator skill, as both image acquisition and interpretation require extensive training[1] and often produce varied assessments between operators[2]. Although advances in device miniaturization and portable systems have increased accessibility, the challenge of reducing training requirements remains[3–8]. This operator dependence creates barriers to broader adoption, particularly in settings with limited access to specialized sonographers or radiologists.

Two-dimensional (2D) ultrasound represents inherently three-dimensional (3D) anatomy as a series of planar slices, requiring users to mentally reconstruct spatial relationships from sequential images[9]. This process increases cognitive load, contributes to variability in interpretation, and extends the learning curve for new users[10]. 3D imaging can capture volumetric information more directly but when rendered on a conventional 2D screen, it remains an inherently lossy representation[11], similar to how volumetric MRI data must be sliced to be interpretable. Unlike MRI, ultrasound is used in dynamic and real time contexts where slicing is not feasible, which forces users to infer 3D structure from flattened views. As with all screen based displays, this also forces users to translate visual information back into physical space, which remains a persistent source of difficulty for both novices and experienced operators[12].

Volumetric ultrasound images allow for anatomy to be more completely captured without the limitations of a single viewing angle, granting access to coronal planes and complex spatial geometries that are physically impossible to visualize with conventional 2D planar imaging. 3D ultrasound images can either be captured directly via full-matrix[13] or 1.5D transducer array[14], or reconstructed by combining sequential slice captures[15]. The volume reconstruction process requires image co-registration either through sensors, motion capture, or deep-learning approaches[16–19]. With the use of volumetric capture, inter-user diagnostic variability has been shown to improve, as users miss diagnostic planes less often[20]. However, 3D ultrasound imaging systems inherently trade off imaging speed for volumetric reconstruction, requiring substantially greater computational resources for beamforming as volume dimensions increase. Additionally, matrix-array transducers carry substantially higher equipment costs—driven primarily by the transducer, cables, and connectors—and face image quality tradeoffs due to inherent beamforming limitations such as decreased through-plane resolution and lower SNR.

[1]Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Computer Science, Dartmouth College, Hanover, NH, USA. [3]Department of Electrical Engineering & Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Department of Computer Science, Oberlin College, Oberlin, OH, USA. e-mail: canand@media.mit.edu

Early efforts to provide in-situ ultrasound visualization relied on holographic screen overlays or half-silvered mirrors to optically align 2D projections with the patient[21,22]. While the transition to Head-Mounted Displays (HMDs) has improved clinical ergonomics, the majority of these systems remain constrained to projecting 2D planar slices[23–25]. This approach fails to resolve the cognitive reconstruction bottleneck, sometimes called mental tomography, as users are still required to mentally reconstruct 3D anatomy from planar views—a process hindered by the ambiguous depth cues inherent to floating 2D overlays[26]. Although volumetric rendering can mitigate this cognitive load, prior demonstrations of 3D-AR have been largely limited to static pre-operative CT models or offline 3D ultrasound reconstructions, sacrificing the live temporal resolution critical for guiding dynamic interventions[27,28]. Recent work advanced this by rendering real-time volumetric ultrasound as in situ volume projections through a HoloLens 2[29]. However, these systems typically rely on Optical See-Through (OST) displays and inside-out tracking for transducer co-registration, which suffer from intrinsic calibration drift greater than 6 mm, latency, and transparency issues that degrade volumetric contrast[26,30,31]. Finally, studies emphasize that while 2D overlays reduce head and eye shifts, they do not fully resolve the core challenge of cognitively reconstructing 3D anatomy from 2D planes[32]. These related contributions are summarized in Supplementary Table S1.

To address these gaps, we developed an augmented reality (AR) platform that integrates directly with a custom low-cost, real-time 3D ultrasound imaging system[33]. Our system, AR-VIU, is a system for Augmented reality, Real-time Volumetric Imaging in Ultrasound, which combines real-time volumetric acquisition with point-cloud rendering, a representation that preserves individual spatial samples rather than interpolating a continuous volume, displayed in a video see-through mixed-reality headset (Varjo XR-4). This architecture enables depth-consistent compositing with the physical environment and avoids the field-of-view and brightness limitations of optical see-through displays. Furthermore, AR-VIU integrates the volumetric imaging data stream into the UE's point cloud engine, preserving viewing angle stability during natural head movement and enabling dynamic user interactions with hand and controller tracking. AR-VIU integrates live volumetric ultrasound, point-cloud visualisation, and Video See-through (VST) mixed reality (Figs. 1A and 1B)—a combination not previously demonstrated—and is evaluated here in a controlled comparative study against 2D and 3D baselines with both novice and expert users ($N = 18$).

We further integrated this AR platform with a Verasonics 2D imaging system to create four experimental systems: 2D ultrasound displayed on a conventional monitor (2D/screen), 2D ultrasound visualized in augmented reality (2D/AR), 3D ultrasound displayed on a conventional monitor (3D/screen), and 3D ultrasound visualized in augmented reality (AR-VIU). This design allowed us to isolate the contributions of volumetric imaging and AR visualization independently as well as evaluate their combined impact. Using this framework, we conducted a controlled study with 18 participants (9 novices and 9 experts), who completed two tasks: 1) object recognition, tested perceptual understanding of complex structures, 2) object localization, assessed the ability to interpret scale and spatial positioning. While currently optimized for instructional environments, AR-VIU serves a specialized clinical function in workflows requiring intricate 3D spatial mapping.

We position AR-VIU primarily as a training and education platform, with complementary clinical roles in procedures that demand explicit three-dimensional spatial understanding such as vascular access, biopsy guidance, and echocardiographic assessment.

Results

Fast 3D image acquisition and beamforming

In this work, we use a chirped data acquisition system (cDAQ) developed in our earlier work on real time 3D ultrasound imaging[33]. The cDAQ employs a chirped excitation with programmable start and stop frequencies and is produced on an field-programmable gate array (FPGA) and transmitted through the array. Demodulation followed by low pass filtering converts the received signal into frequency shifts, which correspond to time delays and shifts the received signal to baseband, allowing each channel to be sampled at only 125 kS s-1 instead of several megahertz. The reduction in sampling rate not only decreases data throughput to the beamformer, but also enables lower latency data transfer, more efficient beamforming, and consequently real-time 3D imaging that can seamlessly support point cloud mixed reality visualization. Acquisition and beamforming occur in parallel. Depending on the chirp bandwidth and receive duration (depth), the acquisition time ranges from 0.4 to 1 s. Graphical processing unit (GPU)-accelerated beamforming using the delay-multiply and sum (DMAS) algorithm has a dependency on volumetric dimensions and ranges from 0.4 s for a 2cmx3cmx3cm box to 0.7 s for a 3cmx3cmx6cm box. Beamforming was run on a laptop GPU (NVIDIA GeForce RTX 4070) to maintain portability of the system, but more powerful desktop GPUs can enhance frame rates. System timing and latencies are shown in Supplementary Table S2. We achieve practical frame rates of 1.4 to 2 Hz, which allows dynamic imaging of moving objects, demonstrated by imaging a small metal ball moving through a fluid filled tube within a gelatin phantom (Fig. 1C), which is further detailed in the Supplementary Materials and Methods (Fig. S1 and Supplementary Movie S1).

Real-time data pipeline

Beamformed images are converted to volumetric point cloud representations and compressed via LZ4 compression, packetized, and asynchronously streamed across an ethernet connection by user-datagram Protocol (UDP) into the game engine at rates of ~3.09 Gbps. We were able to achieve compression rates of ~4x in an average of 0.81 ms with negligible transfer latency across UDP (0.02 ms), a decompression time of 0.5 ms, rendering time of 38.4 ms, and end-to-end performance supporting 2 frames per second (*fps*). A diagram for this data flow is shown in Fig. 1D.

As LZ4 packets arrive asynchronously, a custom Blueprint function in Unreal Engine performs a multi-threaded decompression of the conjoined packets and loads it into a native point cloud octree representation in Unreal Engine with additional color filters for contrast improvement, thresholding to remove noise, and transparency through dithering and material design. The supported latency allows for an average measured rendering rate of 77.9 million voxels per second.

The end-to-end data pipeline from acquisition to display operates as follows: chirp-based acquisition (0.4–1.0 s depending on depth), GPU-accelerated delay-multiply-and-sum beamforming on an NVIDIA RTX 4070 laptop GPU (0.4–0.7 s depending on volume dimensions), LZ4 compression (4x ratio, 0.81 ms), asynchronous UDP streaming (3.09 Gbps, 65000 Byte datagram size, 0.02 ms transfer latency), multi-threaded decompression (0.5 ms), and octree-based point cloud rendering in Unreal Engine 5.3 (38.4 ms). Decompressed frames are verified and a cyclic redundancy check (CRC) is performed to determine whether to partially render frames from dropped packets, which occur rarely due the short local-area network (LAN) connection. Acquisition and beamforming are pipelined in parallel, achieving end-to-end frame rates of 1.4–2.0 Hz with a rendering throughput of 77.9 million voxels per second at ~250 µm spatial resolution. Three copies of the point cloud are maintained in the rendering scene—one 4x scaled interactive copy and two true-to-scale copies refreshed interchangeably to minimize flickering. Foveated rendering, driven by the headset's eye tracking cameras, supports display refresh rates up to 60 Hz. More powerful desktop GPUs can further enhance beamforming throughput and frame rate.

Within the AR environment, the application also initializes and performs controller and hand tracking along with object detection via visual fiducial markers based on ArUco tags[34]. Therefore, both the interface elements of the user and the objects of interest in the scene can be simultaneously visualized, localized, and designated as rendering locations for point clouds (Fig. 1E)

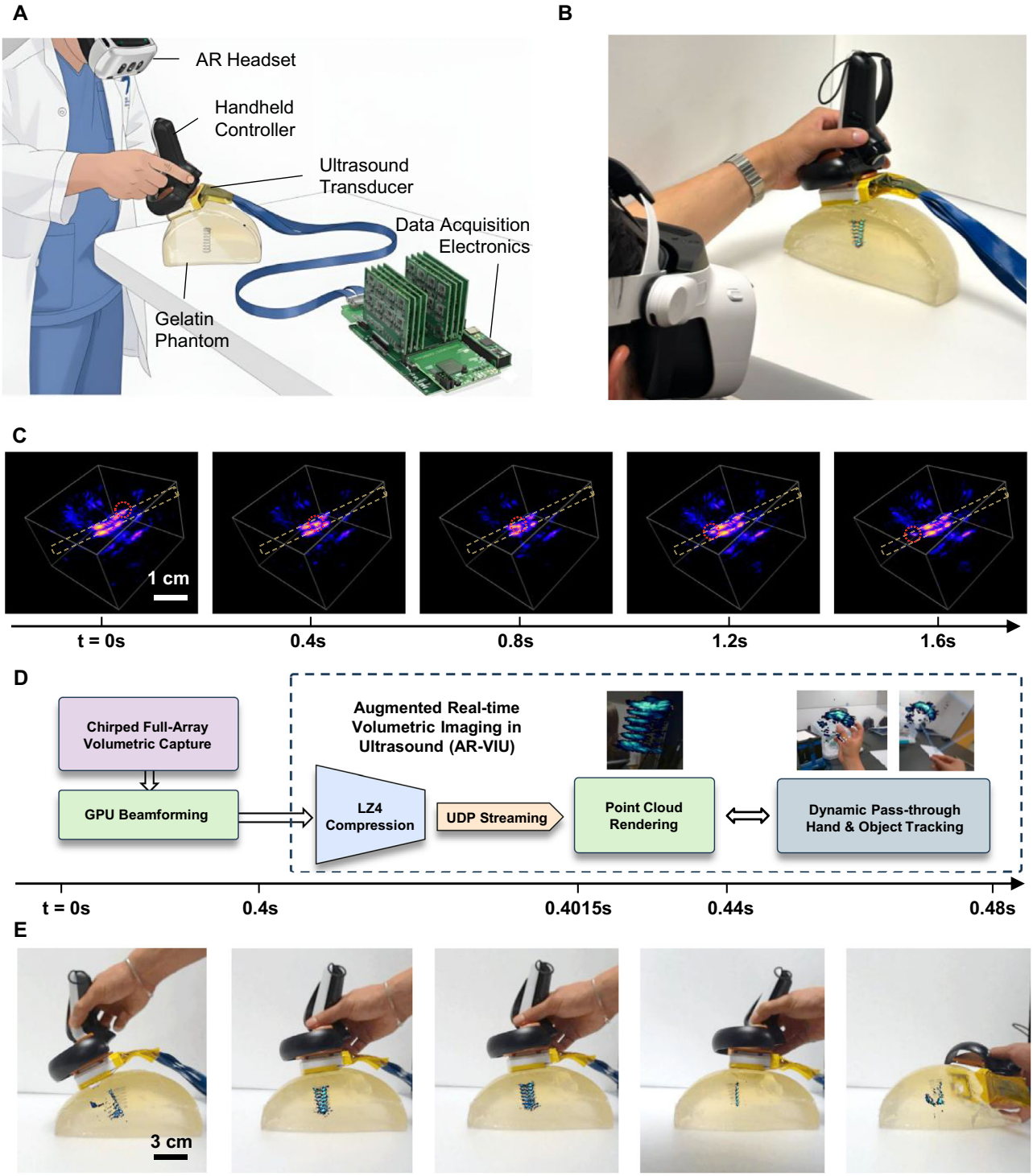

Fig. 1 | AR-VIU System Overview and Operation. A Schematic of AR-VIU system. Real-time volumetric images were captured by the ultrasound transducer and processed by custom data acquisition electronics. The handheld headset controller provides fast localization of the volumetric point cloud image which is overlaid over the physical object via augmented reality (AR) pass-through on the headset. **B** A volumetric point cloud image is shown overlaid on top of a metal spring embedded in a gelatin phantom. This view would typically be only visible to the user wearing the headset but is reconstructed in the image shown. **C** A time-series volume capture of a metal ball reflector (marked in red dashes) rolling through a tube (marked in yellow dashes) within a gelatin phantom depicting a dynamically changing scene (Box dimensions: 3 cm x 3 cm x 2 cm, Scale bar: 1 cm). **D** Schematic of the data pipeline with timing, which shows image acquisition with chirped full-array volumetric capture to graphics processing unit (GPU)-parallelized beamforming (0.4 s), subsequent LZ4 compression (0.81 ms), and asynchronous user datagram protocol (UDP) streaming (0.02 ms) into the Unreal Engine game environment where point clouds were rendered in relation to object and hand-tracked assets (38.9 ms). **E** Headset point clouds can be seen accurately overlaid on the metal spring suspended in transparent gelatin. Imaging from different angles revealed different apertures and details of the spring (Scale bar: 3 cm).

Multiple Systems comparing augmented reality versus 2D displays and 3D versus 2D imaging

The system architecture defines four distinct configurations: the 2D/screen setup uses the Verasonics system with a standard 2D linear-array transducer (Fig. S2), while the 2D/AR configuration pairs the 2D transducer with the Varjo XR-4 headset system (Fig. S3) The 3D/screen system employs our custom 3D imaging system electronics visualized via Napari 3D software on a monitor (Fig. S4), whereas the AR-VIU system combines these 3D imaging electronics with the Varjo XR-4 headset system (Figs. 2A, S5). The systems are characterized using a matrix that maps the system's dimensionality (2D vs. 3D) against its display method (On-Screen vs. AR). Figure 2B visually illustrates both the physical components and the resulting user interface. 2D/screen represents the conventional standard of ultrasound imaging in widespread medical practice.

Tracking and image overlay

We developed two object-localization systems and used both to assess the tracking accuracy and image overlay of the rendered ultrasound image. First, we used visual fiducial marker tracking, which uses the HMD's pass-through cameras and a single CPU core to track a 25 mm square target. We varied the viewing angle of the HMD relative to a 2D binary Varjo marker (Fig. 2C) and measured tracking error between the corners of the virtual bounding box and physical marker corners (Fig. 2D). The mean error was minimized when the HMD cameras faced the marker head-on at 0° to 1.53 ± 1.07 mm and increased as the view angle became more peripheral to 3.88 ± 0.627 mm at 30°, 4.2 ± 0.743 mm at 50°, 4.96 mm ± 0.396 mm at 60°, 7.28 ± 0.864 mm at 75° ($N = 8$, Figs. 2E, S6). Next, we designed a snap-fit attachment that connects a Varjo controller and the ultrasound imaging transducer housing allowing the controller handle to be used simultaneously as a handle for the transducer (Figs. S7, S8). We used a gelatin phantom cast in a cuvette-like container with a flat viewing face to test overlay accuracy at a 0° angle (head-on), incorporating both visual fiducial markers and the physical attachment of the hand-held controller to gather comparative data (Fig. S9). The results illustrate the head-on tracking comparison (0°), where the marker method shows a mean accuracy of 1.48 ± 0.544 mm and the controller method shows a mean accuracy of approximately 1.31 ± 0.589 mm ($N = 9$, Fig. 2F).

To test the controller tracking method across geometries, we performed overlay measurements on a curved phantom symmetrically from angles ranging from −40° to 40° (Figs. 2G, S10). A mix of outside-in inertial measurement unit (IMU) sensor tracking and inside-out external camera tracking ensures that the controller can track with low latency and even when the transducer or controller is partially occluded (Fig. 2H). Overlay measurement accuracy on the curved phantom on average remained within 2.3 mm on average across 40–0° ($N = 9$, Figs. 2I, S11, Supplementary Movie S2). This data indicates that the practical registration error of the overlay is well below 5 mm for the viewing angles used in the study. Visual fiducial marker tracking and virtual object localization are used in the 2D/AR system instead of controller tracking due to the ergonomic cost of mechanically attaching a controller to the 2D ultrasound transducer and their comparable tracking accuracy when set up for head-on tasks in the user study.

User study

We next recruited volunteers to participate in a study that tested the four systems [2D/screen (conventional), 2D/AR, 3D/screen, and 3D/AR (AR-VIU)] across two tasks focusing on object identification and object localization, respectively (Supplementary Movie S3 and S4). Furthermore, we recruited participants from two different inclusion groups: i) ultrasound experts, which include ultrasound researchers, medical practitioners, and sonographers with prior ultrasound experience and ii) novices who have no experience using ultrasound imaging systems. Out of the novices, 33% had AR experience, 89% had virtual reality (VR) experience, and the group consisted of 44% females and 56% males. Out of the experts, 33% had AR experience, 89% had VR experience, and the group also consisted of 44% females and 56% males. Within the ultrasound experts, 44% were researchers, 33% were practicing physicians, 11% were medical students, and 11% were sonographers. We administered the study to a total of 18 participants with an overall study protocol diagrammed in Figs. S13 and S14.

Object identification

Accurate identification of anatomy in imaged structures drives diagnosis and planning, with errors having the potential to cascade into mis-staging, delays, and procedural complications[35,36]. By restoring true-to-scale 3D context and spatial registration, AR-VIU aims to lower cognitive load and inter-user variability, improving accuracy, speed, and safety.

To assess a user's accuracy in object identification (Task 1), study participants were instructed to determine an object hidden in an opaque cup and suspended in gelatin (Figs. 3A, S15). The users were given the dimensions and pictures of the six objects (Fig. S16). Four of the six cups were selected at random and presented sequentially for subjects to image and determine the hidden object. We found that AR-VIU had the highest average accuracy of 91.7 % and on average 17.7% higher than other systems, especially compared to conventional 2D/screen ultrasound (Fig. 3B).

Task 1 accuracy was analysed using a binomial generalised estimating equations (GEE) model with exchangeable correlation structure and robust standard errors, treating each trial (correct/incorrect) as the unit of analysis and accounting for within-participant clustering ($N = 288$ trials, 18 participants). The model included visualisation system, expertise status, and their interaction as fixed effects (ref.2:D/screen, Novice).

A significant main effect of the 'System' was observed (Wald χ^2=21.84, $df = 3$, $p < 0.001$). AR-VIU was associated with 8.06-fold higher odds of correct identification relative to 2D/screen ($OR = 8.06$ [95% CI: 3.07–21.17], $p < 0.001$). Expertise was significant ($OR = 3.43$ [1.03–11.38], $p = 0.045$), and the 'System' by 'Expertise' interaction was not significant (χ^2=2.56, $df = 3$, $p = 0.464$), indicating that the system effect was consistent across experience levels. Predicted probabilities of correct identification were: 2D/screen 57.7% (novice) / 82.4% (expert); 2D/AR 69.4% / 78.4%; 3D/screen 72.8% / 93.8%; AR-VIU 91.7% / 94.0%.

Subgroup analyses confirmed the system effect among novices (Wald χ^2=21.84, $df = 3$, $p < 0.001$; AR-VIU OR = 8.06 [3.07–21.16] vs 2D/screen). Among experts, the system effect was also significant (Wald χ^2=9.86, $df = 3$, $p = 0.020$); 3D/screen showed OR = 3.22 [1.54–6.71] vs 2D/screen ($p = 0.002$), and AR-VIU showed $OR = 3.36$ [0.85–13.19] vs 2D/screen ($p = 0.083$). All statistics are reported in tabular form (Supplementary Table S3).

Experts performed better than novices across all systems, although the size of this difference varied across imaging configurations. Under conventional 2D/screen imaging, novices showed the lowest accuracy and the widest gap from expert performance, reflecting the greater challenge of inferring object structure from planar slices. Both 3D imaging systems improved accuracy for experts and novices, indicating that volumetric information benefits users regardless of prior experience. AR-VIU produced the highest accuracy for both groups and yielded the smallest difference between novice and expert performance. 3D imaging and AR-VIU in particular, help narrow experience-related disparities in object identification (Fig. 3C and D).

Error classification revealed that 88% of all misidentifications (52/59) were geometry errors confusing shape classes, rather than scale errors confusing sizes within the same class (7/59). This pattern was uniform across visualisation systems (χ^2=1.22, $p = 0.749$); however, AR-VIU reduced total errors by 70% compared to 2D/screen (Fisher exact $OR = 4.23$, $p = 0.004$). The most common confusions involved small balls and thick screws (10 and 7 errors in each direction) and thin screws misidentified as small springs (7 errors)—objects with similar acoustic cross-sections in single-plane slices or from reflection artifacts. AR-VIU virtually eliminated these geometry confusions (shape-class accuracy: 92% vs 76% for 2D/screen; error rate: 8% vs 28%), consistent with the hypothesis that

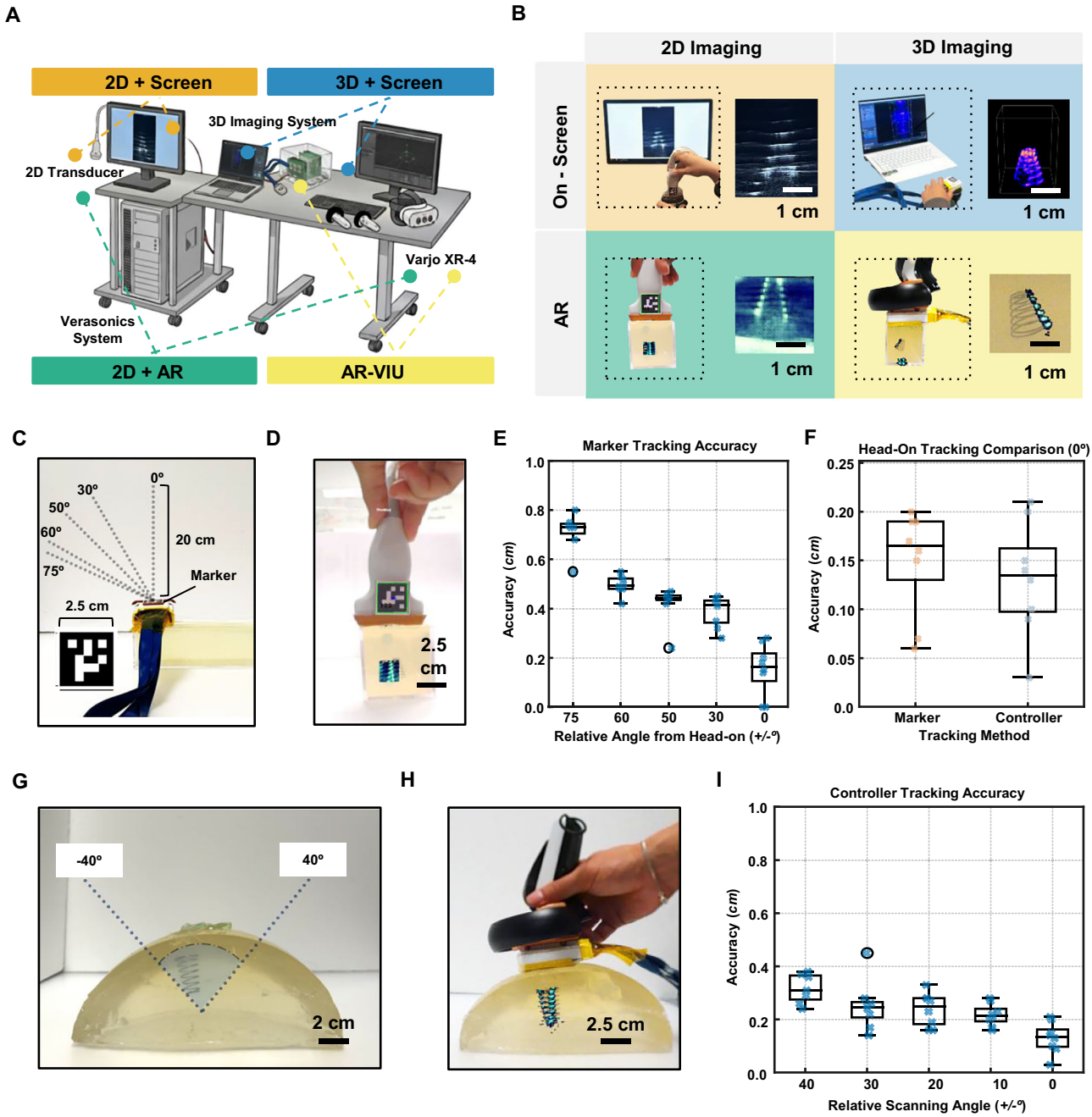

Fig. 2 | Cross-system overview, transducer tracking, and overlay accuracy. **A** Schematic of the hardware used for the four systems that test 2D vs. 3D ultrasound imaging and Screen vs. AR visualization, which includes the Verasonics research console, a L11-5V linear transducer, a custom 3D ultrasound imaging system, a laptop for rendering, and the Varjo XR-4 AR headset. **B** Close-up images of the user interface for each system as well as images of a small metal spring produced by each of the four systems: 2D/screen, 2D/AR, 3D/screen, 3D/AR (AR-VIU) (Scale bar: 1 cm). **C** Schematic of lateral view of the angles tested to determine visual marker tracking accuracy of a 2.5 cm × 2.5 cm fiducial from a distance of 20 cm. **D** A head-on view of a 2D ultrasound slice captured by a linear handheld array, showing the marker tracking overlaid by the game engine in a bounding green box (Scale bar: 2.5 cm). **E** Marker tracking accuracy across captured headset frames at different angles relative to the 'head-on' position at a viewing distance of 20 cm. Mean error at the head-on view at 0° was 1.53 ± 1.07 mm, 3.88 ± 0.627 mm at 30°, 4.2 ± 0.743 mm at 50°, 4.96 mm ± 0.396 mm at 60°, 7.28 ± 0.864 mm at 75°, reported as error ± SD ($N = 8$). **F** Head-on tracking comparison (0°) between the overlaid and physical object with a mean accuracy of 1.48 ± 0.544 mm with marker tracking (orange) and 1.31 ± 0.589 mm with controller tracking (blue), reported as error ± SD ($N = 9$). **G** Schematic of the range of angles tested of the imaging beam relative to direct positioning over the object (Scale bar: 2 cm). **H** View from the AR headset of the point cloud rendering and physical object at the 0° position (Scale bar: 2.5 cm). **I** Tracking accuracy of the point cloud overlay using controller tracking across different angles relative to the object. 1.31 ± 0.589 mm at 0°, 2.18 ± 0.427 mm at 10°, 2.38 ± 0.627 mm at 20°, 2.51 ± 0.930 mm at 20°, 3.13 ± 0.523 mm at 30°, reported as error ± SD ($N = 9$).

volumetric point-cloud rendering resolves shape features that remain ambiguous in two-dimensional slices. The full 6×6 confusion matrices are provided in Supplementary Fig. S18.

In the 2D imaging system, interpreting the true dimensions of a spherical target can be inherently challenging because the operator must mentally reconstruct the object based on isolated slices, while ensuring that the narrow fan-like beam intersects the target. 2D and 3D ultrasound image references for the six objects used for the object identification task are provided in Supplementary Fig. S19. Under AR-VIU, where the full volume is rendered simultaneously and to true-scale within the user's field-of-view,

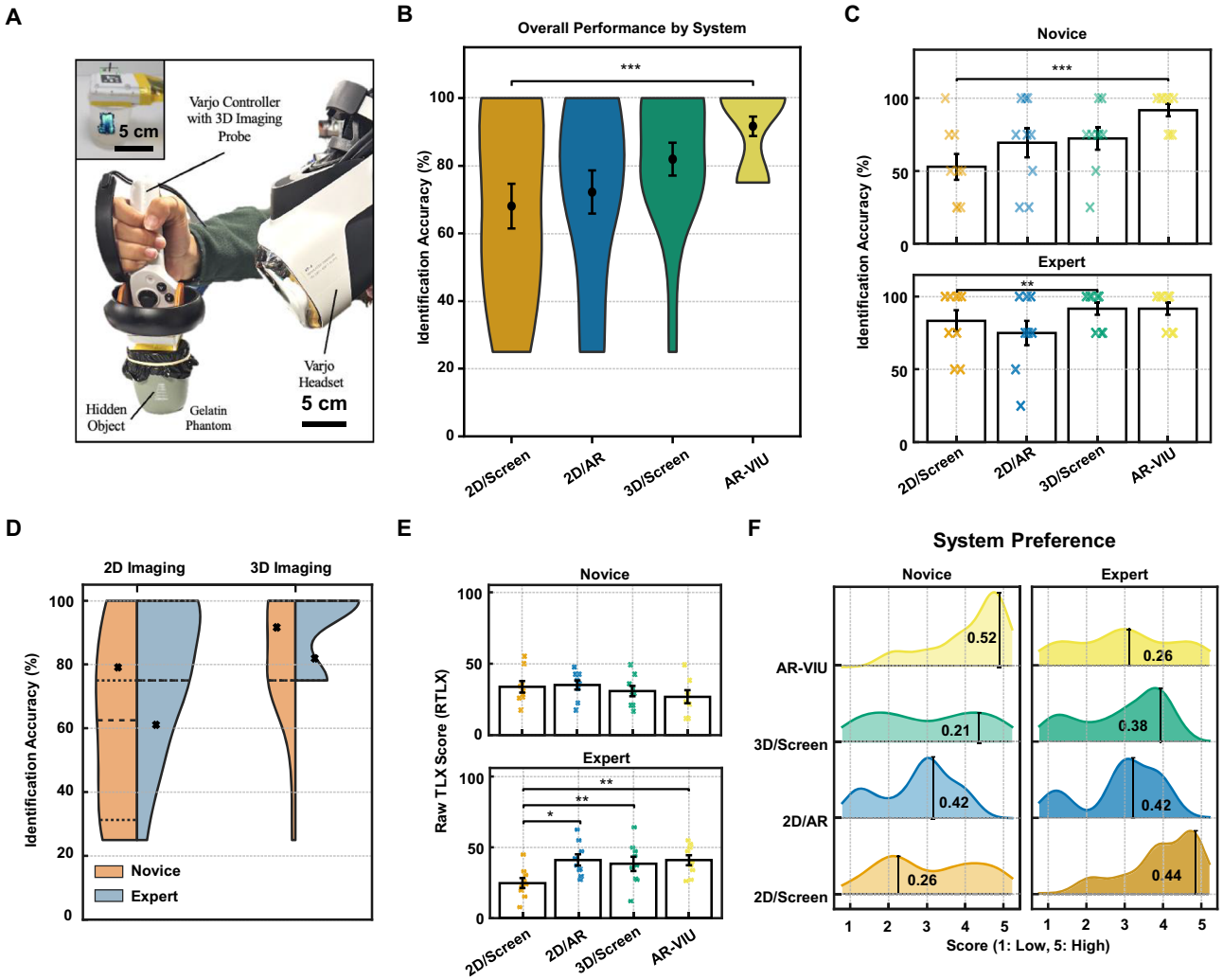

Fig. 3 | User study: object identification task results between novices and experts. **A** Schematic of a subject performing an object identification task with AR-VIU. The subject is able to see the hidden object with the AR point cloud visualization. (Scale bars: 5 cm). **B** Overall identification accuracy across visualisation systems for all participants ($N = 18$ participants, 4 trials each). Violin distributions show participant-level accuracy with mean ± standard error of the mean (SEM) overlay. A binomial GEE model confirmed a significant main effect of system (Wald $\chi^2(3) = 21.84$, $p < 0.001$); AR-VIU was associated with 8.06-fold higher odds of correct identification relative to 2D/Screen ($OR = 8.06$ [95% *CI*: 3.07–21.17], $p < 0.001$; Supplementary Table S3). **C** Identification accuracy stratified by expertise. Novices ($N = 9$): AR-VIU vs 2D/Screen $OR = 8.06$ [3.07–21.16], $p < 0.001$. Experts ($N = 9$): 3D/Screen vs 2D/Screen $OR = 3.22$ [1.54–6.71], $p = 0.010$. The 'System' by 'Expertise' interaction was not significant ($\chi^2(3) = 2.56$, $p = 0.464$), indicating a consistent system effect across experience levels. **D** Comparison of 2D imaging systems (2D/Screen and 2D/AR) versus 3D imaging systems (3D/Screen and AR-VIU), illustrating the accuracy gain conferred by volumetric information. **E** Raw Task Load Index (RTLX) workload scores (unweighted mean of six NASA-TLX subscales, 0–100) by system and expertise group. Expert participants showed a significant system effect (RM-ANOVA: $F(3,24) = 8.046$, $p < 0.001$, partial $\eta^2 = 0.50$), with 2D/Screen associated with the lowest workload (24.4 ± 10.6). Novice workload differences were not significant ($F(3,21) = 1.200$, $p = 0.334$). **F** System preference ratings (Likert scale, 1–5) for the object identification task, shown as ridgeline kernel density plots for novice and expert groups. Novices most frequently preferred AR-VIU; experts preferred 2D/Screen. The density scale bars show the peak density value for each panel.

these geometry confusions were lessened (shape-class accuracy 92% vs 76% for 2D/screen).

Raw composite NASA–TLX scores show clear differences in how novices and experts perceive the task demands across systems (Fig. 3E). A repeated-measures ANOVA on Raw Task Load Index (*RTLX*) scores revealed a significant main effect of the visualisation system for expert participants in Task 1 ($F(3,24) = 8.046$, $p < 0.001$, partial $\eta^2 = 0.50$). Post-hoc comparisons (Holm-corrected) confirmed that 2D/screen was associated with significantly lower workload ($RTLX = 24.4 \pm 10.6$) than 2D/AR (40.5 ± 11.7; $p = 0.036$, *Hedges' g* = 1.37), 3D/screen (37.8 ± 15.1; $p = 0.006$, $g = 0.98$), and AR-VIU (40.4 ± 10.6; $p = 0.005$, $g = 1.44$). For novice participants, AR-VIU was associated with the numerically lowest workload ($RTLX = 26.8 \pm 12.8$), but this did not reach significance ($F(3,21) = 1.200$, $p = 0.334$, partial $\eta^2 = 0.15$). These results are summarized in Supplementary Table S4.

Preference ratings tracked these workload patterns: novices most frequently preferred AR-VIU, followed by 3D/screen imaging, whereas 2D/AR received the lowest preference scores; experts most often preferred conventional 2D/screen imaging and expressed more variable preferences for the three non-conventional systems (Fig. 3F). Collectively, these findings indicate that AR-VIU minimized workload for novices while supporting the highest accuracy, whereas expert users reported the lowest workload on the familiar 2D/screen configuration and perceived all non-conventional systems as more demanding despite comparable or higher objective performance. Despite the higher perceived workload, experts were able to perform most accurately with AR-VIU.

Object localization

Precise localization is central to ultrasound-guided tasks such as central venous catheter placement[37,38], peripheral nerve blocks[39], and percutaneous biopsy[40,41]—procedures where spatial understanding of subsurface anatomy directly affects accuracy and safety. Localization remains operator-dependent: novices and even trained users face depth-perception and cognitive-load challenges when inferring spatial position from 2D views[11].

To assess a user's ability to localize objects within a phantom, we constructed a task where participants perform a simulated targeted biopsy test (Task 2). Instead of a catheter or needle, the participant used a pen to mark those locations accordingly. Surface marking was chosen over needle insertion because it provides (i) continuous millimetre-scale measurement without the binary hit-or-miss outcome of needle puncture, (ii) is accessible to novice participants who lack procedural needle-handling skills, and (iii) isolates the perceptual-cognitive components of target localization from the psychomotor skill and haptic feedback demands of needle guidance[42]. Needle-guided procedures can involve additional dexterity, tissue deformation, and real-time feedback demands not captured by this surface-marking paradigm. Participants scanned the CIRS Multi Purpose Multi Tissue phantom under each imaging configuration and drew perceived target positions on a front facing sheet affixed to the phantom (Fig. 4A and B). Localization error was computed as the Euclidean distance between each drawn target and the corresponding ground truth position from the manufacturer's schematic (Fig. 4C).

Localization error for Task 2 was scored by three independent raters blinded to participant identity, system condition, and experimental group. Participant drawings were digitally aligned to the reference template, anonymised, and assigned randomised codes. Raters marked paired points (participant mark and reference target) via a web-based scoring interface, from which Euclidean error distances were computed. Inter-rater reliability was excellent ($ICC(2,1) = 0.955$, 95% CI [0.93, 0.97]; mean absolute inter-rater difference: 1.22 ± 3.05 mm). All reported localization errors are three-scorer averages. Per-system ICC values and Bland-Altman plots are provided in Supplementary Fig. S20.

The error distributions shown in Fig. 4D demonstrate that augmented reality substantially reduced localization error for both experience groups compared with screen based visualization. A linear mixed model on log-transformed measurement-level distances ($N = 1{,}079$ three-scorer-averaged measurements, 18 participants) confirmed a significant main effect of System (Wald $\chi^2(3) = 165.41$, $p < 0.001$) and a significant System × Status interaction ($\chi^2(3) = 14.01$, $p = 0.003$; Supplementary Table S5). Back-transformed geometric means from estimated marginal means were lowest for 2D/AR (5.18 mm) and AR-VIU (5.46 mm), approximately half those of 3D/Screen (8.39 mm) and 2D/Screen (10.53 mm). 2D/Screen geometric mean localization error was approximately twice that of 2D/AR ($ratio = 2.03$, $p < 0.001$) and AR-VIU ($ratio = 1.93$, $p < 0.001$).

AR visualisation virtually eliminated the novice–expert accuracy gap: the LMM-derived geometric mean error for novices on AR-VIU ($EMM = 5.94$ mm) was comparable to that of experts ($EMM = 5.02$ mm), representing a 56% reduction from novice 2D/Screen error ($EMM = 13.44$ mm). Experts also benefited, with AR-VIU reducing their geometric mean error by 39% relative to 2D/Screen (5.02 vs 8.26 mm). The significant System × Status interaction ($\chi^2(3) = 14.01$, $p = 0.003$) confirms that the magnitude of system-related improvement differed by experience level, with novices showing the largest gains.

System specific performance is shown in Fig. 4E for novices and Fig. 4F for experts with per-subject distributions shown in Fig. S21. No speed–accuracy tradeoffs were detected in either task (all Spearman $|\rho| < 0.46$, all $p > 0.08$). At the trial level, incorrect identifications in Task 1 tended to take longer than correct ones under 2D/AR (incorrect: 42.0 s vs correct: 27.4 s; Mann–Whitney $p = 0.076$), suggesting that participants struggled longer before making errors rather than trading speed for accuracy (Figs. S22, S23).

Because the phantom is symmetric along its width, depth cues in the third axis were less informative, which limited the advantage of 3D imaging on screens. The addition of AR improved performance across both 2D and 3D imaging modes by spatially anchoring ultrasound information to the physical phantom. Across systems, experts annotated more wire targets than novices, suggesting that they were able to extract information more comprehensively even when overall error magnitudes differed.

We fit a linear mixed-effects model (LMM) to trial-level distances in millimeters with the natural log of distance as the response to accommodate for long-tailed distributions. Estimated marginal means (EMMs) were computed on the log scale and back-transformed to geometric means in millimeters for interpretation. Post hoc comparisons, corrected using the Holm method to control for multiple testing, showed that both AR systems produced statistically reliable reductions in geometric-mean localization error relative to the two-dimensional screen baseline (Table S5). In novices, the estimated error ratios for 2D/AR and AR-VIU were ~0.48 and ~0.45, respectively, with Holm-adjusted p-values $p < 0.001$, confirming that these improvements were not only large in magnitude but also robust after stringent multiplicity correction. Experts demonstrated a similar pattern with ratios of 0.66 for 2D/AR and 0.54 for AR-VIU (adjusted $p < 0.001$), again indicating strong and reliable performance gains. By contrast, the 3D/screen condition produced ratios of 0.82 in novices and 1.05 in experts, and neither comparison remained significant after Holm adjustment. AR-VIU enabled experts and novices to perform on average 50.5% better than with the conventional 2D/screen system (Fig. S24).

In Task 2, a significant effect of the visualisation system on Raw TLX workload was detected among novice participants (RM-ANOVA: $F(3,24) = 3.812$, $p = 0.023$, partial $\eta^2 = 0.32$). The 3D conditions were associated with lower workload (3D/screen: 30.8 ± 11.4; AR-VIU: 29.5 ± 12.2) than the 2D conditions (2D/screen: 41.7 ± 14.0; 2D/AR: 38.7 ± 14.4). After Holm correction, no individual pairwise comparison reached significance, though the 2D/screen vs AR-VIU contrast showed a large effect ($p\text{-}holm = 0.070$, Hedges' $g = 0.88$). No significant workload differences were observed among experts ($F(3,24) = 0.182$, $p = 0.908$) (Table S4 and Fig. S25).

System preference ratings mirrored these trends (Fig. 4H). Most novices preferred AR-VIU, followed by 3D/screen, and assigned the lowest scores to the 2D/AR system, consistent with its higher workload and error rates. Experts instead showed the strongest preference for 2D/screen and strong secondary preference for AR-VIU, demonstrating a marked shift in preference variance among experts compared to the first object identification task.

System preferences and performance comparison

Following the user study, participants were given a questionnaire to gauge system preferences for aspects of the experience such as comprehension, satisfaction, ease of learning, and clinical integration potential (Fig. 5A). Novice participants preferred AR-VIU most strongly in all categories but with particular preference in relation to the conventional 2D/screen system in its ease of learning. AR-VIU consistently received the highest novice ratings across categories. Expert participants preferred the conventional 2D/screen system most strongly in all categories; however, AR-VIU scored with the second highest preference score relative to the other non-conventional systems (2D/AR and 3D/screen) with a comparable score to the conventional system in the ease of learning criteria. For ease of learning in particular, expert ratings for AR-VIU and the conventional 2D/screen system cluster toward the upper end of the scale, whereas 2D/AR and 3D/screen show broader distributions that extend more toward intermediate scores. Finally, experts were asked to gauge the potential for clinical integration and use for each of the systems with the expectation that the gold standard (2D/screen) conventional case would have the highest score given its widespread clinical adoption. We observed that among the other non-standard systems (3D/screen, 2D/AR, and AR-VIU), AR-VIU had the highest preference for clinical integration by distribution weight at higher scores around 4 out of 5 (Fig. 5B).

In lieu of the Likert preferences selected by the participant, we sought to understand their subjective experience metrics in relation to their true task

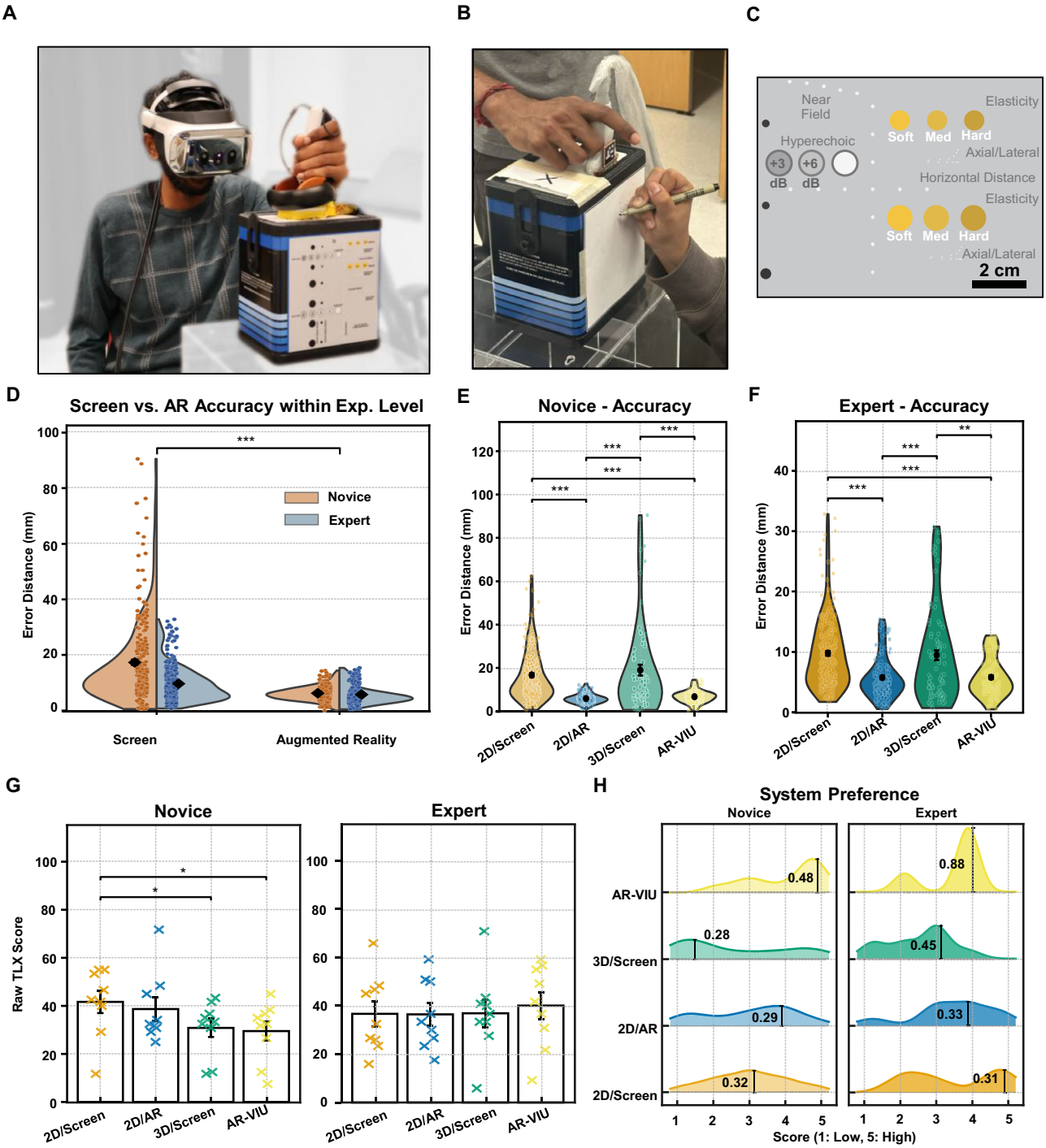

Fig. 4 | User study: object localization task results between novices and experts. **A** Participant performing the localization task with the AR-VIU system, using the 3D ultrasound probe while wearing the head-mounted display and scanning the CIRS Multi-Purpose Multi-Tissue phantom (scale bar, 4 cm). **B** Participant completing the localization task with the 2D imaging system while drawing the perceived target locations on a sheet of paper attached to the front face of the phantom (scale bar, 3 cm). **C** Digital recreation of the datasheet schematic of the CIRS Multi-Purpose Multi-Tissue phantom showing the arrangement of wire targets and echoic chambers that are able to be imaged by the participant due to the fixed imaging depth (6 cm) set for both 2D and 3D imaging systems during the user study; this layout served as the ground truth against which participant drawings were evaluated (scale bar, 2 cm). **D** Localization error distributions across all four visualization systems for the combined cohort ($N = 18$). Each measurement represents the three-scorer-averaged Euclidean distance (mm) between the participant's mark and the reference target. A linear mixed model on log-transformed distances confirmed a significant main effect of System (Wald $\chi^2(3) = 165.41$, $p < 0.001$) and a significant System × Status interaction ($\chi^2(3) = 14.01$, $p = 0.003$; Supplementary Table S5). Back-transformed geometric mean EMMs: 2D/Screen 10.53 mm, 3D/Screen 8.39 mm, 2D/AR 5.18 mm, AR-VIU 5.46 mm. **E** Localization error for novice participants ($N = 9$). AR-VIU reduced novice geometric mean error by 56% relative to 2D/Screen (*EMM*: 5.94 vs 13.44 mm, *ratio* = 0.44, $p < 0.001$). **(F)**Localization error for expert participants ($N = 9$). AR-VIU reduced expert geometric mean error by 39% relative to 2D/Screen (*EMM*: 5.02 vs 8.26 mm, *ratio* = 0.61, $p < 0.001$). **G** Raw Task Load Index (RTLX) workload scores by system and expertise group. A significant system effect was detected among novices (RM-ANOVA: $F(3,24) = 3.812$, $p = 0.023$, partial $\eta^2 = 0.32$), with 3D conditions associated with lower workload than 2D conditions. No significant differences were observed among experts ($F(3,24) = 0.182$, $p = 0.908$). **H** System preference ratings (Likert scale, 1–5) for the localization task, shown as kernel density ridgeline plots with scale bar representing peak density value for each panel. Novices most frequently preferred AR-VIU; experts preferred 2D/Screen with a strong secondary preference for AR-VIU.

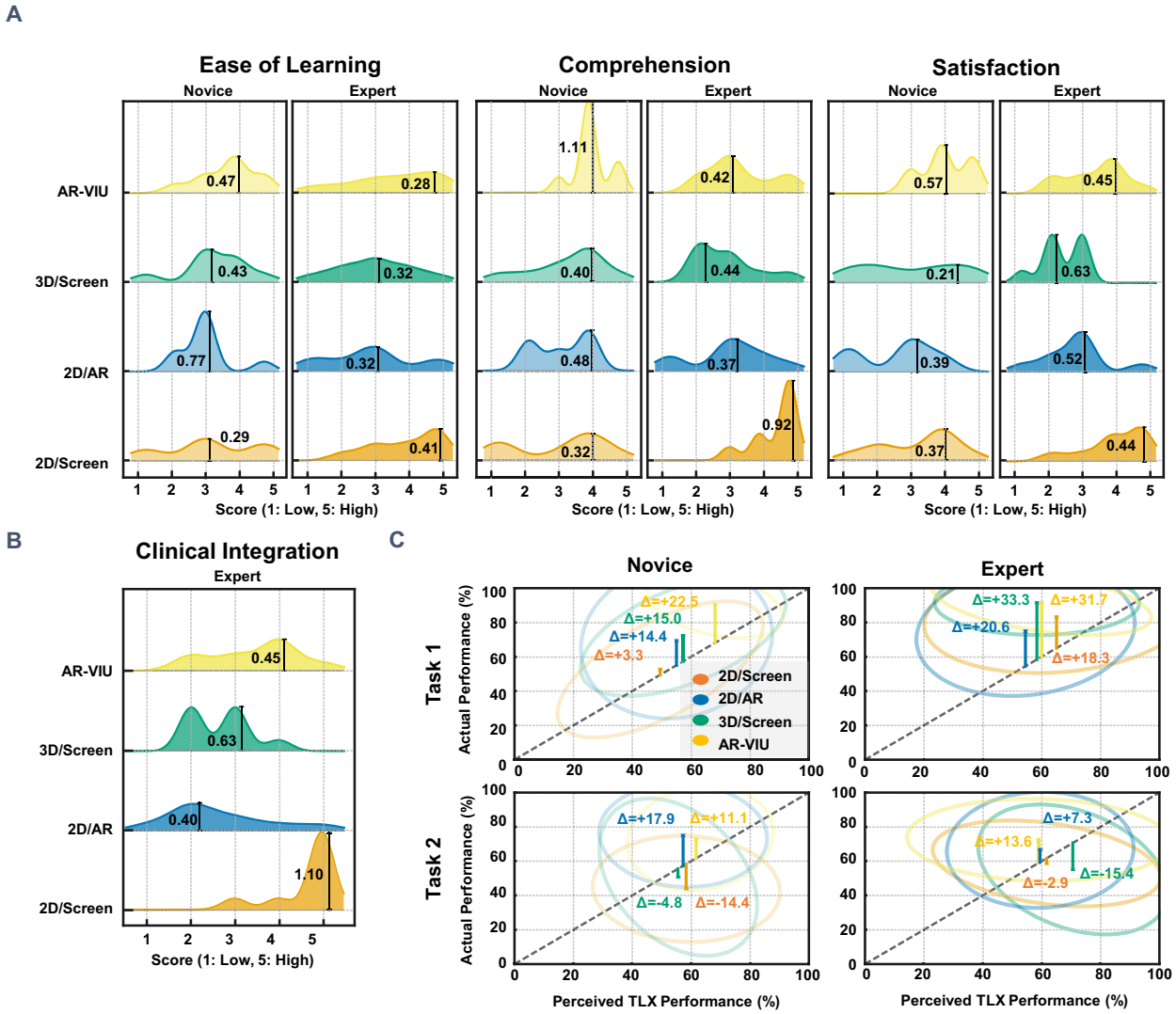

Fig. 5 | User Study: System Preferences and Perception-Performance Analysis. **A** Post-study questionnaire ratings across four subscales: Satisfaction, Ease of Learning, Comprehension, and Clinical Integration—shown as kernel density ridgeline plots (Likert scale, 1–5) for novice and expert groups. Clinical Integration was assessed for experts only. **B** Clinical Integration preference ratings for expert participants. The density scale bars show the peak density value for each panel. **C** Actual Performance' versus 'Perceived NASA-TLX Performance' across systems between experts and novices. Dotted line represents the axis where 'Actual Performance' matches 'Perceived Performance'. Ellipse centroid is marked, and its circumference represents a 68% *CI* with χ^2 scaling. A delta value showing Actual Performance deviation from the dotted line indicates relative overperformance or underperformance.

performance. Across both tasks, participants' objective performance and their TLX-based perceived performance were compared (Fig. 5C). We found that experts and novices alike performed better than they perceived when using AR-VIU. However, in Task 2, both groups underperformed relative to their perceived performance with screen-based systems and overperformed with AR-based systems. Notably, among experts, AR-VIU yielded the highest objective accuracy despite being rated among the lowest in perceived performance.

Post-study interview

Finally, we conducted semi-structured interviews with the participants to understand their learning experience, comprehension and confidence, suggestions for improvement, and thoughts on clinical integration. We also recorded the professional experience of expert users, demographics, and experience with AR or VR systems (Supplementary Table S6.). We analyzed their responses through inductive thematic coding, and these codes were then quantified to show how often each theme appeared among novices and experts (Fig. S26). Our findings corroborate and contextualize participants' quantitative performance and system preferences.

In line with their quantitative performance, participants' preferences diverged sharply by experience level. Among experts, 44% referenced the conventional system's familiarity and diagnostic reliability as a central reason for preferring it. For instance, one expert participant, an ultrasound researcher, noted that their training and regular exposure to 2D imaging in clinical and academic environments reinforce their confidence in interpreting 2D slices. This makes the conventional system feel more trustworthy and more immediately usable to her than the non-conventional systems. In contrast, 78% of novices described AR-VIU as aiding their spatial identification and intuition, and 33% explicitly stated a preference for 3D imaging due to its simplicity and learnability. Overall, experts trusted the familiar 2D workflow while novices favored the systems that provided the most direct and intuitive correspondence to the underlying object.

Despite this divide in preferences, 67% of experts still identified clinical use cases for AR-VIU. They noted that the system would be particularly valuable for tasks requiring fast and accurate spatial interpretation, such as trauma assessment, vein localization, biopsy guidance, and tumor identification. For example, an intensive care physician stated that 2D imaging suffices for stationary organs, but echocardiography would benefit from 3D

visualization of cardiac wall movement. Furthermore, two participants acknowledged that 3D images are more understandable than 2D slices and thus AR-VIU would be useful for anatomy instruction and communicating results to patients in a more accessible way.

Although experts identified meaningful clinical applications for AR-VIU, they also emphasized that its effectiveness in such settings would depend on system reliability and user comfort. Reliability concerns were one of the most frequently cited issues in interviews, with 67% of experts and 44% of novices reporting latency, frame-rate instability, and low resolution in the non-traditional systems. Ergonomic challenges were similarly widespread: 33% of experts and 67% of novices commented on headset weight, cable interference, and controller dependence. To make the system more user-friendly and practical for prolonged use, 33% of novices benefitted from but suggested more accurate hand-tracking, an expert stated that AR glasses would be less distracting and more lightweight than a headset.

In summary, the interviews highlight why novices performed well with AR-VIU and why experts tend to prefer more conventional workflows. While participants recognized meaningful clinical and educational applications, they also emphasized that adoption will depend on improved reliability and ergonomics. These findings contextualize the quantitative results and identify priorities for future iterations.

Discussion

In this work, we demonstrate that real time 3D ultrasound visualized in augmented reality can meaningfully alter how users perceive and interpret volumetric information. Across two clinically motivated tasks, AR-VIU was associated with reduced object identification errors, higher localization accuracy, and suppressed extreme outliers relative to conventional 2D/screen imaging. These benefits were observed in both novices and experts, with the largest relative gains in novice participants, indicating that mixed reality visualization can narrow experience related disparities in performance. Beyond the specific system implementation, our study provides a task based evaluation framework that enables quantifiable measurement of user performance, workload, and system preference in a clinical ultrasound context.

The stronger gains observed in novices highlight the potential of mixed reality visualization to lower the expertise barrier for high quality ultrasound use. When volumetric imaging and augmented reality visualization were combined in AR-VIU, novice accuracy approached expert levels, and localization error distributions overlapped more closely. These results suggest that immersive 3D visualization can enhance spatial understanding. This is particularly relevant for training programs and point of care users such as acute care physicians, anesthesiologists, or emergency clinicians who often rely on ultrasound but may not have access to prolonged supervised practice, unlike expert sonographers.

The localization task served as a controlled analogue of ultrasound guided targeting, where small spatial errors can detrimentally impact procedural outcomes. In clinical settings, similar demands arise during central venous catheter placement[36,37], regional nerve blocks[38], percutaneous biopsies[39,40], and drainage procedures, where the operator must position a device relative to nearby vessels, nerves, or other "no-go" structures. In our study, the augmented reality conditions yielded more stable performance with fewer large deviations, which is particularly relevant because rare but substantial targeting errors often drive complications. We acknowledge that needle-guided procedures can involve additional dexterity, tissue deformation, and real-time feedback demands not captured by this surface-marking paradigm. Future studies should evaluate AR-VIU performance with interventional needle tasks in tissue-mimicking phantoms. Although the phantom task simplifies the dynamics of in vivo procedures, these findings suggest that spatially registered mixed reality visualization could increase the safety margin for ultrasound guided interventions by providing more intuitive depth and position cues within realistic workflow constraints.

The mismatch between performance and perceived workload shows that familiarity and trust, rather than accuracy alone, drive how users judge and adopt new imaging systems. Novices reported the lowest workload and strongest preference for AR-VIU, consistent with its gains in accuracy and stability. Experts, however, favored the familiar 2D screen system, even when AR-VIU supported comparable or better performance. Post study interviews suggest that this reflects deeply ingrained workflows and reduced trust in newer displays that occasionally show lag or require recalibration. These insights indicate that improvements in tracking robustness, frame rate, and ergonomic comfort will be as important as advances in imaging quality, and that successful adoption will depend on predictable behavior and compatibility with established clinical practice. In the near term, mixed reality ultrasound may therefore find its most natural role in complementary settings such as simulation and skills training, supervision of less experienced operators or selected interventional procedures, where additional spatial context and shared visualization between team members provide a clear and defensible benefit over conventional imaging alone.

Expert preference for conventional 2D displays does not reflect resistance to innovation but rather well-calibrated clinical judgement: these practitioners have optimised perceptual strategies for 2D interpretation through years of deliberate practice. As one expert participant (cardiologist, >10 years of echocardiography) explained, "I preferred [2D] just because that's what I was used to, and I could just refer to it on a flat plane." Another expert participant (sonographer) highlighted that "the operator dependence is very heavy" in conventional ultrasound, underscoring that clinical interpretation is driven by practitioner skill and that the value proposition of AR-VIU differs by user population. While experts leveraged existing mental models to achieve high accuracy even under 2D conditions (ceiling effects in Task 1 for experts: Friedman $p = 0.248$), novices lacked these compensatory strategies and benefited disproportionately from volumetric rendering—a pattern evidenced by the complete elimination of the novice–expert accuracy gap under AR-VIU (Cohen's $d = 0.00$ vs $d = 0.88$ for 2D/screen).

These findings position AR-VIU primarily as a training and education tool in the near term, with complementary clinical roles in procedures requiring explicit 3D spatial understanding. Indeed, 67% of expert interviewees independently cited specific clinical applications, including vascular access, biopsy guidance, echocardiographic assessment, and trauma triage, where volumetric AR visualisation could augment rather than replace conventional imaging. Multiple experts endorsed the system's educational potential; one expert suggested that in training, "you'd be able to project what you see… a whole class seeing exactly what you're seeing," while another observed that "novices who've never seen this stuff… really like the VR and [it] make[s] sense." These responses suggest that AR-VIU may accelerate the steep learning curve associated with freehand ultrasound by providing spatial context as novices learn to construct mental mappings from ultrasound images.

While the proposed augmented reality system demonstrates a marked performance improvement, it is necessary to consider the fundamental physical constraints of acoustic imaging. Unlike computed tomography or magnetic resonance imaging, conventional ultrasound is traditionally characterized by extreme anisotropy where axial resolution vastly outperforms lateral and elevational dimensions. Although the custom system developed for this study is not perfectly isotropic, it was engineered to substantially narrow this resolution gap compared to standard clinical matrix arrays[33]. This more balanced resolution profile substantially mitigates the severe smearing typically observed when rotating standard acoustic volumes in virtual space.

Nevertheless, volumetric rendering still faces a persistent compromise between surface opacity and internal visualization. High opacity is required to define unambiguous surfaces for object identification, yet this inevitably obscures deeper structures. Indeed, experienced practitioners are adept at building robust mental models from planar slices alone where no anatomy is hidden. As one participant observed, experts become very good at making a mental math of things just by getting a lot of feedback. Additionally, the presence of acoustic shadowing and attenuation complicates volumetric interpretation because these artifacts create regions of missing data that a novice might misinterpret as anatomical voids. Finally, in near-eye HMD rendering, the vergence-accommodation conflict (VAC) arises because the

eyes converge on a virtual object at one depth while headset offset accommodates the focal plane of the optics causing visual fatigue, discussed further in Supplementary Note 3.

To address these visual challenges, our current implementation mitigated rendering ambiguity through intensity thresholding, dithered transparency, and a user accessible toggle between point cloud and real world views. Furthermore, the integration of motion parallax and proprioceptive feedback provides operators with additional spatial context to help analyze remaining acoustic artifacts. By reorienting the probe to look around shadows or optimize the acoustic window, the user can better manage the interpretability issues often associated with static volumetric renderings. This active spatial exploration leverages human stereopsis to ease the mental tomography bottleneck. The findings from the study suggest that immersive acoustic volumes convey actionable clinical information and can substantially narrow the performance gap between novices and experts. Future iterations should incorporate real time segmentation algorithms and adaptive transfer functions that map rendering parameters to clinically meaningful acoustic properties, thereby reducing perceptual ambiguity even further.

Several considerations frame the interpretation of this work. The current AR-VIU implementation operates at approximately two volumes per second with end to end latencies in the hundreds of milliseconds. Although sufficient for the controlled tasks performed here, these parameters represent an early stage of a rapidly evolving technology, and continued improvements in hardware and reconstruction pipelines are likely to narrow the gap with conventional ultrasound systems. Our evaluation was conducted in phantoms under well controlled conditions to enable reproducible comparisons across systems. Moving toward in vivo imaging will introduce additional complexity due to physiological motion, tissue deformation, and complex anatomical structures, providing important opportunities to further stress-test rendering capabilities, tracking pipelines and overlay stability. Most current HMDs also impose ergonomic constraints, as reflected in several participants' feedback regarding long-duration use. Further design considerations and ergonomics of HMDs are discussed in Supplementary Note 1 and 2.

Future work will extend both the technical and clinical dimensions of AR-VIU. Increasing volume rate and reducing latency through optimized beamforming, parallel processing, and dedicated hardware accelerators could enable temporally intensive procedures[43] and imaging moving organs[44]. Enhancing tracking robustness under occlusion, variable lighting, and continuous probe motion will improve reliability in realistic environments. On the translational side, studies in healthy volunteers and focused procedural contexts represent natural next steps for identifying where mixed reality ultrasound offers the greatest advantage. Integrating automated plane detection, quality feedback, or artificial intelligence based guidance could further amplify the benefits for less experienced users and create a more comprehensive training and decision support platform.

Taken together, these results indicate that real time 3D ultrasound in mixed reality can make structural information more accessible and reduce reliance on extensive operator experience. Specifically, AR-VIU eliminated the novice–expert accuracy gap in object identification (Cohen's $d = 0.00$ vs $d = 0.88$ for 2D/screen; Fig. 3C–D), reduced novice geometric mean localization error by 56% relative to 2D/screen (EMM: 5.94 vs 13.44 mm; Fig. 4E), and produced near-identical error distributions across experience levels (Fig. 4D). A binomial GEE model confirmed that AR-VIU was associated with 8.06-fold higher odds of correct identification compared to 2D/screen ($OR = 8.06$ [95% CI: 3.07–21.17], $p < 0.001$; Table S3). As the supporting hardware and tracking technologies continue to mature, systems like AR-VIU may provide a practical path toward training tools and procedural adjuncts that bring expert-level spatial understanding closer to the point of care. By lowering perceptual and cognitive barriers, mixed reality ultrasound has the potential to broaden the safe and effective use of ultrasound in settings where specialized sonographers are limited and where enhanced spatial context can meaningfully improve decision making.

Methods

Ethics approval

This study was reviewed and approved by the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects (COUHES) under protocol #2408001392.

Informed consent

All participants provided written informed consent prior to participation. Participants were informed of the study's purpose, procedures, potential risks, data handling practices, and their right to withdraw at any time without penalty.

3D ultrasound image acquisition

Real-time 3D ultrasound images were acquired using a custom chirped data acquisition (cDAQ) system[33] developed to enable real time volumetric imaging with a low channel count. The imaging probe incorporates 64 transmit and 64 receive elements arranged in a box geometry based on the convolutional optimally distributed array (CODA) sparse architecture[45]. The convolution of the orthogonal transmit and receive subarrays forms a fixed virtual aperture that achieves unity spatial sampling without electronic aperture steering. This compact configuration provides wide angle volumetric coverage while minimizing the total number of active channels.

The cDAQ employs a swept frequency continuous wave excitation rather than discrete pulses. A digitally generated chirp waveform with programmable start and stop frequencies and duration was produced on an FPGA and transmitted through the array. Because the chirp frequency increases linearly with time, echoes from deeper structures correspond to proportionally larger frequency shifts. On the receive path, in phase and quadrature demodulation followed by low pass filtering converts these frequency shifts into equivalent time delays. This analog frequency to delay mapping compresses the received signal bandwidth, allowing each channel to be sampled at only 125 kS s-1 instead of several megahertz. The reduction in sampling rate not only decreased data throughput but also enabled faster and lower latency data transfer, more efficient beamforming, and consequently real time three dimensional imaging that can seamlessly support 3D mixed reality visualization.

The hardware consists of a probe head containing the CODA array, low noise amplifiers, and transmit drivers connected via micro coaxial cables to a motherboard housing the demodulators, analog-to-digital converters (ADCs), an FPGA controller, and a Teensy 4.1 microcontroller. The FPGA coordinates transmit timing and data acquisition while Teensy manages high speed universal serial bus (USB) communication with the host workstation.

Data acquisition and reconstruction are implemented in Python using Numba and CuPy for parallel computation. The Napari library runs on the main thread for 3D visualization on the local workstation display, while acquisition and beamforming tasks execute asynchronously in background threads. Synthetic aperture reconstruction was performed using the baseband Delay Multiply and Sum algorithm ($p = 2.5$) accelerated on a laptop NVIDIA RTX 4070 GPU.

For mixed reality implementation, the main Napari visualization thread was replaced with a custom thread handling real time data transfer over UDP to the AR workstation. The AR workstation received the reconstructed volumetric frames and rendered them in the headset display. This architecture maintains continuous data flow from acquisition through beamforming to display, achieving low latency volumetric updates suitable for immersive mixed reality interaction.

The cDAQ frame rate is determined by the relationship between the chirp duration, sampling rate, bandwidth, imaging depth, and the number of transmit events required to synthesize the full aperture. Increasing the imaging depth or the chirp bandwidth directly increases the chirp duration, which lowers the achievable frame rate but improves depth penetration and spatial resolution. Conversely, higher ADC sampling rates shorten the chirp period and permit faster acquisitions at the expense of increased data throughput and power consumption. Under typical experimental settings in

this work, with a 125 kHz sampling rate, a 2–5 MHz chirp bandwidth, and an 6 cm imaging depth using 64 transmit events, the system achieved approximately 1.4 *fps* with 250 μm spatial resolution. Because these parameters are freely configurable, the same system can be adjusted to balance imaging depth, resolution, and acquisition speed according to experimental requirements, supporting both high fidelity volumetric imaging and real time mixed reality visualization.

2D ultrasound image acquisition

2D imaging was performed on a Verasonics Vantage system using an L11-5v linear array probe. The system was configured for standard B-mode acquisition with one transmit–receive cycle per scan line. The transmit waveform was set within a 2–6 MHz frequency range, matching the frequency band used in the cDAQ experiments. This range lies within the operating bandwidth of the L11-5v probe and was selected to ensure consistent imaging conditions and fair comparison across both systems.

Each transmit event excited one element in sequence, and received echoes were digitized through the Verasonics front end and stored in a continuously refreshed acquisition buffer. Real-time beamforming was performed using Verasonics' built-in delay-and-sum reconstruction, followed by envelope detection, log compression, and grayscale mapping to produce B-mode images. Depth-dependent time gain compensation was applied to balance signal intensity across the imaging field, and imaging parameters such as depth and contrast were tuned to match those of the 3D cDAQ experiments. In the screen based implementation, the system imaged at a rate of approximately 50 *fps*.

For the mixed-reality configuration, the same acquisition and reconstruction pipeline was used. An additional process captured each completed B-mode frame immediately after reconstruction and transmitted it over UDP to a dedicated mixed-reality workstation. The AR system received these frames and rendered them in real time on the headset display, enabling synchronized standard monitor visualization and immersive mixed-reality viewing without altering the Verasonics imaging sequence or processing pipeline. The 2D AR system operated on an end-to-end frame rate of approximately 10 *fps*.

Mixed reality environment

The mixed reality environment was built using the VarjoOpenXR framework within Unreal Engine 5.3. Using the built-in functions, Varjo marker detection, hand tracking, and controller tracking were exposed along with several gameplay modes that were mapped to keyboard commands. On each event loop, a PointCloud Manager blueprint object loaded incoming point cloud data from the exposed UDP port. A custom script compiled, decompressed, and loaded the volumetric point clouds into the LiDAR point cloud plugin structure, which contained native octree structure representation. Points were then rendered using dynamic level of detail scaling, which adjusts the resolution of the point cloud based on the viewing distance and other factors and provides substantial improvement in refresh rate and rendering speed compared to other game engines. Combined with the foveated rendering feature of the HMD, which uses eye tracking cameras to selectively sharpen display resolution at the user's visual focal points, the system supported refresh rates up to 60 Hz. Furthermore, we used the game engine's ability to render millions of points to maintain three copies of the point cloud image: a 4-times scaled interactive copy, and two overlapped, true-to-scale copies that are updated and refreshed interchangeably to reduce flickering during fast refresh rates. We applied thresholding to reduce noise and the number of rendered points. Point cloud objects were further color-mapped for pass-through contrast enhancement, scaled by size, and localized in the scene. The resulting interface consisted of one interactive point cloud in front of the user and another that was localized to the transducer via a visual marker or controller.

The mixed reality environment included several key features that helped the user mentally bridge objects in the visual and physical environment, including hand tracking, controller localization, and visual fiducial tracking. We also implemented tools to provide depth cues, including a pointer to localize a distinct point in 3D space and controller button toggles to hide and show the point clouds. We also provided hand tracking to manipulate, move, and rotate the scaled-up point cloud visualization in 3D space, which allowed users to manipulate the 3D point clouds naturally without having to hold a controller.

Controller attachment calibration

Spatial registration between the ultrasound volume and the mixed-reality coordinate frame was achieved using a snap-fit, 3D-printed computer-aided designed (CAD) mounting bracket (fused deposition modeling, 0.8 mm layer height). The bracket enforced a fixed geometric transform between the centroid of the transducers and the virtual origin of the Varjo XR-4 controller, consisting of a − 20 degree pitch, −3 cm translation along the z-axis, and +3 cm translation along the x-axis. This eliminated the need for runtime calibration. To quantify registration repeatability, the bracket was attached and detached $N = 5$ times, and four keypoints were annotated on overlay-ON and overlay-OFF frame pairs for each cycle. After centroid alignment to remove systematic translational offset, the mean residual registration error was 0.69 ± 0.33 mm across 20 keypoint pairs (Fig. S12, Supplementary Movie S5). This sub-millimetre residual was consistent across all five reattachment cycles (per-cycle means: 0.44–0.89 mm) and is substantially smaller than the 6–15 mm system-level differences observed in Task 2 localization accuracy, confirming that registration precision did not confound inter-system comparisons.

Pixel coordinates of the four vessel-corner keypoints were annotated in each overlay-ON and overlay-OFF frame. A scale factor s (mm px^{-1}) was derived from the known physical dimension of the imaging vessel face ($d_{face} = 50.75$mm) and its corresponding pixel extent (d_{ref} in px) measured in the same frame:

$$s = \frac{d_{face}}{d_{ref}} \tag{1}$$

Raw registration error for each keypoint was computed as

$$e_{raw} = s \cdot \left| |p_{on} - p_{off}| \right|, \tag{2}$$

where p_{on} and p_{off} are the pixel coordinates of the same anatomical landmark in the overlay-on and overlay-off frames. To isolate rotational and local deformation error from any global translational offset, the centroids of the four ON and four OFF keypoint sets were aligned by centroid subtraction, and per-keypoint residual errors were computed from the aligned coordinates.

This evaluation differs from the point reconstruction accuracy (PRA) metric from freehand ultrasound calibration work[46], which quantifies the residual between a reconstructed image point and its ground-truth location after transformation through the full tracking-to-image chain:

$$PRA = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left| |p^i{}_W - r^i{}_W| \right|^2} \tag{3}$$

$$p_W = {}^W T_C \cdot {}^C T_I \cdot p_I \tag{4}$$

where p_W is a reconstructed point's 3D position, r_W is its known 3D position, ${}^C T_I$ is the image-to-tracker calibration obtained via phantom-based optimization and ${}^W T_C$ is the controller pose reported by the tracking system. PRA is the appropriate metric when ${}^C T_I$ is estimated at runtime, because the calibration algorithm is the dominant error source. In AR-VIU, ${}^C T_I$ is mechanically fixed by a snap-fit CAD bracket rather than estimated, so the error budget reduces to

$$\epsilon^2{}_{total} = \epsilon^2{}_{bracket} + \epsilon^2{}_{tracking} + \epsilon^2{}_{imaging} \tag{5}$$

with the calibration-optimization term absent by construction. The tracking term $\epsilon_{tracking}$ is governed by the Varjo XR-4's Lighthouse tracking configuration, which has been independently characterised with submillimetre static accuracy and submillimetre dynamic accuracy against robotic ground truth in multi-base-station setups[47]. The localization stability of these methods are highlighted in Supplementary Movie S6. The imaging term $\epsilon_{imaging}$ reflects intrinsic beam geometry and is common to all freehand systems.

Our evaluation targets $\epsilon_{bracket}$ directly, since this is the only term introduced by our attachment method. The measured residual of 0.69 ± 0.33 mm is the mechanical analogue to the calibration residual in PRA and sits at the low end of the range reported for well-calibrated optimisation-based freehand systems[46], supporting mechanically constrained attachment as a viable alternative to phantom-based calibration.

Probe tracking

Accurate tracking of the imaging transducer is crucial for constructing an accurate pass-through overlay of the ultrasound image. In works that reconstruct and coregister 3D volumes from 2D slices, tracking accuracy determines the reconstruction fidelity. Previous works have explored a combination of inside-out and outside-in tracking methods that include electromagnetic, optoelectronic motion capture, and HMD-native camera methods and achieve tracking accuracies within millimeters[48,49]. These approaches require external or add-on hardware systems that can add considerable cost or clunkiness.

In our experiments, the ultrasound probe was tracked by the headset using two different techniques: visual fiducial markers and physical attachment of the hand-held controller. While visual fiducial markers are easy to attach and low-profile, they often add noticeable latency to the ultrasound image rendering due to the computational overhead to read the marker. They also can become easily occluded and deviate in accuracy when viewed at oblique angles. Although multiple markers at orthogonal angles can be used to improve tracking accuracy, this technique can add considerable computational overhead and interfere with the ergonomics of the handheld device.

Probe tracking with visual fiducial markers was integrated into the app with Varjo marker tracking. We used a single 25 mm marker, printed it on paper, cut it out, and attached it to a 3D printed frame with double-sided tape. Marker tracking functions identified the identification number of the marker (ID) and instantiated a Blueprint object that includes a green bounding box around where the system has last tracked the marker ID. The ultrasound image point cloud was then translated relative to the marker location for each different transducer probe to which it is attached.

We also demonstrated fast, occlusion-resistant tracking by integrating the headset's handheld controller with the ultrasound imaging transducer. We exploited the fact that controllers native to the HMD are optimized to support fast controller localization with millimeter accuracy using built-in inside-out and outside-in tracking techniques. A custom, 3D-printed snap-fit bracket attached the ultrasound transducer to the controller and allowed for quick and precise attachment and detachment. A CAD-designed mounting bracket enforced a fixed geometric transform between the centroid of the transducers and the virtual origin point of the Varjo XR-4 controller consisting of a -20 degree pitch, -3 cm shift in the z-axis, and $+3$ cm x-translation. It also served as an ergonomic grip for using the ultrasound transducer on the AR-VIU system. Because the controllers use a combination of outside-in tracking via the headset cameras and inside-out tracking via its internal IMU, the controllers were able to relay precise positional and pose information within milliseconds. The mixed reality app environment then used these position and pose measurements to reference the translation of the point clouds in real-time.

Image overlay accuracy assessment

To assess the overlay accuracy, we conducted experimental testing using several different gelatin phantoms embedded with conical springs.

To further understand the impact of viewing angles 0°, 30°, 50°, 60°, and 75° on overlay accuracy, we compared the accuracy of the visual tracking marker overlay (green bounding box) across different viewing angles measured by a protractor and at a fixed distance of 20 cm. A custom Python script was used to transform the image of the bounding box and marker when at an oblique viewing angle such that the marker and bounding box become squared (Fig. S6). Pixel distances were then measured and converted to corresponding pixel distances to determine visual marker tracking accuracy at different viewing angles relative to a head-on view.

Next, to compare the overlay accuracy between visual marker tracking and controller-based tracking from a head-on view, we fabricated a gelatin phantom with a suspended small conical spring (0.7×6.5–14×15 304 stainless steel, Cilky) in a cuvette container with a flat viewing face. This setup allowed us to compare the accuracy of key points of the image with corresponding key points of the physical object as viewed through the headset cameras. For the experimental testing, a toggle on/off functionality within the game environment allowed for the observation of the actual object and the overlay at the same frame and angle, which was used to calculate the distance between key points on the object.

To assess the overlay accuracy with various scanning angles with controller tracking, we fabricated a hemispherical gelatin phantom with a suspended large conical spring (0.8×12–20×30 304 stainless steel, Cilky). We then cut the phantom with a fishing line to create a flat viewing face. A heat-gun set at 200 °C was used to smooth out lines made from cutting the viewing face. This setup allowed for designated relative scanning angles of 0°, 10°, 20°, 30°, and 40° to quantify the precision of the AR overlay with respect to the physical object. For the experimental testing, a toggle on/off functionality within the game environment allowed for the observation of the actual object and the overlay at the same frame and angle, which was used to calculate the distance between key points on the object. In this specific test, we exclusively employed the hand-held controller technique for tracking. We measured distances by creating a custom Python script that co-registers the 'visualization on' image from the headset with the 'visualization off' image. Common keypoints in the two images are manually chosen and the pixel distances between keypoints are converted to physical distances and saved.

Image overlay

Beyond the accurate registration of the virtual and physical environments, the virtual image overlay was designed to improve clarity and contrast of the ultrasound point cloud while maintaining depth perception in pass-through. An additive and translucent material design were implemented and controlled on a toggle by preference. For effective and smooth rendering of transparency in a variety of lighting conditions, especially in bright hospital lighting, dithering was implemented. This technique created the perception of intermediate opacity levels by patterning fully opaque or fully culled-pixels, thus avoiding the computational expense of traditional alpha-blending and sorting. Furthermore, the system supported various color display modes, which provided further depth and passthrough contrast. To refine the quality of the point cloud data that formed the overlay, a thresholding function removed noise from the image data, culling occluding low-valued noise points rather than displaying them.

Thresholding and transparency parameters were fixed at preset values for the duration of the user study to ensure consistent imaging conditions across all participants and system comparisons. The toggle between point-cloud overlay and direct camera view was the only rendering control available to participants during trials.

Study design

We conducted a study approved under IRB #2408001392 by the MIT Committee on the Use of Humans as Experimental Subjects (COUHES). Informed consent was obtained after the nature and possible consequences of the studies were explained through a Consent Form approved by the MIT COUHES. The study was designed to systematically evaluate how imaging dimensionality (2D versus 3D) and visualization modality (screen-based

versus mixed reality) influence ultrasound interpretation. To isolate the independent and combined effects of these two factors, four imaging configurations were constructed: (1) 2D ultrasound displayed on a conventional monitor, (2) 2D ultrasound visualized in mixed reality, (3) 3D ultrasound displayed on a conventional monitor, and (4) 3D ultrasound visualized in mixed reality. Each configuration used identical probe positions, imaging depths, and acquisition parameters to ensure that only the visualization mode and imaging dimensionality varied between conditions.

A total of eighteen participants (9 novices and 9 experts) completed two structured tasks under each condition. We selected expert participants based on their experience and training with ultrasound imaging systems, which was required to exceed 3 months. Novice participants were selected based on no prior experience with using ultrasound imaging. Before the experimental trials, participants completed a familiarization session using a commercially available elasticity quality-assurance phantom (CIRS Model 049, Norfolk, VA, USA). This phantom contains calibrated inclusions of varying stiffness embedded at controlled depths, providing a standardized medium for participants to practice probe manipulation and image interpretation. The session was used solely for training and system familiarization and was not included in the experimental data analysis.

For the screen-based systems (2D/screen and 3D/screen), on-screen scale markers were displayed alongside the ultrasound image; the location and scale of reference markers were explained during the training phase (Fig. S16). For the AR-based systems (2D/AR and AR-VIU), a physical ruler was provided so that participants could measure the rendered objects and compare dimensions against the reference sheet (Fig. S17). Participants were free to manipulate both the ultrasound probe and the gelatin phantom cups throughout the trial, enabling inspection from multiple scanning planes and viewing angles.

The first experimental task (Task 1) was an object identification task designed to assess participants' ability to recognize structures from ultrasound images alone. Commonly identifiable objects were embedded within gelatin phantoms prepared in small containers and sealed with opaque film to eliminate visual cues (Fig. S15). Each participant was provided with a printed reference sheet showing all possible objects to scale and in the same orientation as they were positioned within the phantoms. Six phantoms were fabricated in total, and four were randomly selected for each trial to reduce bias from repeated exposure to specific objects. The order in which participants used the four imaging systems was also randomized to further minimize learning effects and ensure balanced comparisons across experimental conditions. Before each trial, ultrasound gel was applied to the phantom surface to ensure consistent acoustic coupling. Participants were then handed the prepared phantom and timed from the moment of receipt until they reached a final decision regarding the identified object. Violin plots showing the distribution of task completion times between novices and experts can be found in the Supplementary Materials and Methods (Fig. S21).

The second experimental task (Task 2) was an object localization task designed to evaluate participants' ability to interpret spatial relationships and scale from ultrasound images. This task used the CIRS Multi-Purpose, Multi-Tissue Ultrasound Phantom (Model 040GSE), which contains calibrated wire targets and echoic chambers positioned at known depths within a tissue-mimicking Zerdine background. The phantom has a symmetric internal structure with wire targets and circular chambers distributed uniformly along its width. For each trial, a sheet of paper was affixed to the front face of the phantom, corresponding to the orientation shown in the manufacturer's schematic[50]. Participants were instructed to draw the observed targets to scale and in spatial correspondence with their true positions in the phantom relative to the probe placement. Because the phantom is symmetric along its width, participants were asked to represent only the front-view cross section. A ruler was provided to assist in accurate scaling and distance estimation. After each trial, the edges of the phantom were marked on the paper to align the participant drawings with the ground truth geometry specified in the datasheet. This procedure enabled quantitative comparison between perceived and true target positions across all imaging conditions. Violin plots showing the distribution of task completion times between novices and experts can be found in the Supplementary Materials and Methods (Fig. S21).

After completing each imaging condition, participants completed a NASA Task Load Index (NASA-TLX) questionnaire to assess perceived workload, including mental and physical demand, effort, and overall task difficulty.

Raw data collected from each task can be found in the Supplementary Materials and Methods (Data S1 and S2).

Post-study questionnaire and interview analysis

Following the experimental trials, participants completed a structured questionnaire assessing their user experience and system preferences (Fig. S27). The questionnaire contained 5-point Likert-type items evaluating: (1) likelihood of using each ultrasound system for tasks similar to the object identification task and the object localization task, (2) overall satisfaction with each system, (3) ease of learning each system, and (4) perceived comprehension of each system's output. Experts answered an additional item on how likely they were to recommend each system for clinical use. Participants also ranked the four systems from most preferred to least preferred. We created ridgeline kernel density plots, separated by experience level, to visualize differences in preference, ease of learning, comprehension, satisfaction, and clinical integration ratings across systems (Fig. 4H, Fig. 5A, Fig. 5B).

After participants completed the questionnaire, we conducted semi-structured interviews to contextualize their quantitative responses and probe deeper into their user experience. Questions were organized into four categories: 1) learning experience, 2) comprehension and confidence, 3) suggestions for improvement, and 4) clinical integration (experts-only) (Fig. S27). We also asked participants to compare systems based on imaging dimensionality (2D versus 3D) and visualization modality (screen-based versus mixed reality), so we could isolate how each factor influenced their user experience. Ridgeline kernel density estimates of system preference ratings (1=low, 5=high). All panels within each figure use identical kernel density estimate bandwidth and density scaling to enable direct visual comparison across systems and experience groups. The density scale bars show the peak density value for each panel. The kernel density estimate is computed via *scipy.stats.gaussian_kde* with *bw_method* = 0.4 on the raw Likert scores (1–5). Interviews lasted approximately 15-20 minutes. All interviews were audio-recorded, transcribed, and analyzed using inductive thematic analysis. Codes were grouped into higher-level themes, and we quantified how frequently each theme appeared among novices and experts to compare recurring opinions across groups (Fig. S26).

Statistical analysis for workload

Subjective workload was measured using the NASA Task Load Index, scored as the Raw TLX (RTLX)—the unweighted arithmetic mean of all six subscale ratings (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, Frustration), each scored on a 0–100 scale. RTLX was chosen over the original weighted composite[51], as the weighting procedure adds respondent burden without meaningfully improving sensitivity. Each participant provided one RTLX score per system per task, yielding a within-subject repeated-measures design (4 systems × 1 score per task). The effect of the visualisation system on RTLX was tested using repeated-measures ANOVA (RM-ANOVA) within each task × expertise group (4 analyses: Task 1 Novice, Task 1 Expert, Task 2 Novice, Task 2 Expert). Sphericity was assessed via Mauchly's test, and Greenhouse–Geisser corrections were applied when violated. Significant omnibus tests were followed by pairwise comparisons with Holm correction for familywise error. Effect sizes are reported as partial η^2 for omnibus tests and Hedges' g for pairwise contrasts. As a robustness check, the nonparametric Friedman test was applied to each analysis; the two methods agreed in 7 of 8 group × task comparisons. The single discrepancy, Task 2 novice workload, was significant under RM-ANOVA ($p = 0.023$) but not Friedman ($p = 0.209$), consistent with the greater statistical power of the parametric test at $N = 9$. The directional

consistency between RTLX and the original weighted TLX analysis (22 of 24 pairwise comparisons agreed in direction; 23 of 24 in significance) supports the robustness of the overall pattern (Supplementary Table S4).

Statistical analysis for object identification

Task 1 accuracy (correct/incorrect per trial) was analysed using a binomial generalised estimating equations (GEE) model, which accommodates the correlated binary outcomes arising from repeated measurements within participants. An exchangeable working correlation structure was assumed, and robust (sandwich) standard errors were used to ensure valid inference regardless of the true within-subject correlation pattern. The model included visualisation system (4 levels; reference2:D/Screen), expertise status (Novice/Expert), and their interaction as fixed effects, with each trial as the unit of analysis ($N = 288$ trials across 18 participants). Effects were quantified as odds ratios (*OR*) with 95% Wald confidence intervals. The overall system effect was tested via a Type III Wald chi-square test. Subgroup GEE models stratified by expertise were fit to characterize system effects within each group separately. Predicted probabilities of correct identification were computed from the marginal model for each system by status cell. All Task 1 statistical results are reported in Supplementary Table S3.

Image analysis for object localization

The object recognition task evaluated participants' ability to identify and interpret structural features, while the object localization task assessed spatial understanding and depth perception, which are both essential skills needed for ultrasound use in practice. We directly compared the effects of image dimensionality and visualization mode, as well as their potential interaction, on user performance and perceptual accuracy.

We printed a bounding box (5.84" x 6.31") on a blank sheet of paper and taped the bounding box with rough alignment over the side of the wire phantom with its internal diagram. Participants then marked their assessment on top of the blank sheet covering the diagram. After the task trial, we carefully marked 3 alignment points of the underlying diagram appearing through the blank paper, which co-registers the bounding box on the paper to the phantom. Images were analyzed using a custom semi-automated image analysis software built in Python and using a PyQt6-based GUI. Localization error for was scored by three independent raters blinded to participant identity, system condition, and experimental group. Participant drawings were digitally aligned to the reference template, anonymised, and assigned randomised image codes via a blinding manifest. Raters marked paired points—the participant's drawn mark and the corresponding reference target—via a web-based scoring interface, from which Euclidean error distances were computed. Performance statistics from each individual participant show consistencies among distributions in score sheets (Figs. S28 and S29). Representative images of trials from a subject measured by this fashion are shown in the Supplementary Information (Fig. S30).

Statistical analysis for object localization

Localization error in Task 2 was defined as the Euclidean distance (mm) between each participant's marked target position and the corresponding ground-truth reference, averaged across three blinded independent raters. Because the resulting distance distribution was right-skewed, concentrated at small values with a minority of large errors, we applied a natural-log transformation to satisfy the normality and homoscedasticity assumptions of the linear model. One measurement with a recorded distance of zero was clipped to 0.1 mm prior to transformation to avoid undefined values. A linear mixed-effects model (LMM) was fit using restricted maximum likelihood (REML) with the log-transformed distance as the response. The fixed-effects structure included visualisation system (4 levels), expertise status (Novice/Expert), and their interaction. A random intercept for participants accounted for the repeated-measures design and the unequal number of target measurements per participant–system combination (range: 8–17 measurements per image). The model was fit to all 1,079 three-scorer-averaged measurement-level observations from 18 participants. The significance of fixed effects was evaluated using Type III Wald chi-square tests. Estimated marginal means (EMMs) were computed on the log scale and back-transformed via exponentiation to geometric means in millimetres for clinical interpretability. Post-hoc pairwise system contrasts were computed as log-ratios, back-transformed to multiplicative ratios and percent change, with 95% confidence intervals derived from the model standard errors. Familywise error across the six pairwise system comparisons was controlled using Holm's sequential correction. Subgroup analyses stratified by expertise level were conducted by fitting separate LMMs within each group, with contrasts referenced to the group-specific 2D/Screen baseline to avoid cross-group conflation. All localization statistics are reported in Supplementary Table S5.

Participant selection

Participants were recruited to take part in a controlled experimental study comparing performance across visualization and dimensionality conditions. All participants provided informed consent in accordance with institutional ethical guidelines. To be eligible, participants were required to be at least 18 years of age, in good physical health, and free from any medical or behavioral conditions that could interfere with task performance. Individuals with chronic or acute cardiovascular disease, epilepsy, motion sickness, or any condition impairing their ability to follow directions, complete study activities, or manipulate handheld objects were excluded. Participants were also required to understand and communicate effectively in English to ensure comprehension of instructions and the ability to provide valid consent. Participants were categorized into two groups based on prior experience with handheld ultrasound probes: experts, defined as having at least three months of training and/or experience using handheld ultrasound probes, and novices, defined as having less than three months of such training or experience. These criteria established a clear distinction in domain expertise while maintaining consistent general eligibility across all participants.

Phantom preparation

A variety of gelatin phantoms were prepared with buried metal reflector objects for representative ultrasound imaging along the surface of the gelatinized phantom. The metal objects include a small conical spring (0.7 × 6.5–14 × 15 304 stainless steel, Cilky) a large conical spring (0.8×12-20×30 304 stainless steel, Cilky), a small metal ball bearing (3/15" 304 stainless steel, Taiss), a 1.5 cm pipe screen mesh ball (steel, Nuanchu), a thin screw (M4 × 0.7 × 50 mm steel Phillips-drive), and a thick screw (3/8-16 in UNC steel hex-head screws, 1.75 in long). Boiling water is mixed with dry gelatin powder (Medley Hills Farm) at a ratio of 4% gelatin to water by weight to mimic the stiffness of human soft-tissue. In a separate vessel that is used to mold or contain the final phantom, metal objects such as springs, screws, and metallic balls are fixed to the vessel by a nylon fishing line (Zebco Omniflex 0.013 dia). The gelatin is mixed until dissolved completely and poured into the phantom vessel. Phantoms are transferred to a refrigerator to complete gelatinization.

Sample size and statistics

Eighteen participants (9 novices, 9 experts) completed all four system conditions (AR/Screen × 2D/3D) across two tasks, yielding extensive within-subject data. In Task 1, participants made accuracy judgments across four objects per condition; in Task 2, they produced dozens of point-level accuracy measures per participant. This rich, repeated-measures design, analyzed with mixed-effects models incorporating random effects for participants and items, maximizes statistical sensitivity while maintaining valid inference. With $N = 18$, the study achieves approximately 80% power to detect medium-to-large within-subject effects (paired $dz = 0.6–0.7$) and is well suited to identify meaningful Display Interface × Image Dimensionality × Experience Group interactions. Although smaller between-group effects (experts vs. novices) are less powered, the design provides robust estimation of effect sizes and confidence intervals, ensuring clear insights into how visualization modality and dimensionality influence performance across expertise levels.

Data availability

The datasets generated during and/or analysed during the current study are available in the public repository https://doi.org/10.6084/m9.figshare.32049102.

Code availability

Unreal Engine Game build and code are available at https://github.com/jasonf-hou/AR-VIU.

Received: 14 January 2026; Accepted: 7 May 2026;
Published online: 10 June 2026

References

1. Nielsen, A. B., Konge, M. J. & Tolsgaard, M. G. Assessment methods in medical ultrasound education. *Front. Med.* **9**, 871957 (2022).
2. Blehar, D. J., Barton, B. & Gaspari, R. J. Learning curves in emergency ultrasound education. *Acad. Emerg. Med.* **22**, 574–582 (2015).
3. Yoon, H. et al. Decoding tissue biomechanics using conformable electronic devices. *Nat. Rev. Mater.* **10**, 4–27 (2024).
4. Kim, J., Yoon, H., Viswanath, S. & Dagdeviren, C. Conformable piezoelectric devices and systems for advanced wearable and implantable biomedical applications. *Annu. Rev. Biomed. Eng.* **27**, 255–282 (2025).
5. Hou, J. F. et al. An implantable piezoelectric ultrasound stimulator (ImPULS) for deep brain activation. *Nat. Commun.* **15**, 4601 (2024).
6. Du, W. et al. Conformable ultrasound breast patch for deep tissue scanning and imaging. *Sci. Adv.* **9**, eadh5325 (2023).
7. Yu, C. et al. A conformable ultrasound patch for cavitation-enhanced transdermal cosmeceutical delivery. *Adv. Mater.* **35**, e2300066 (2023).
8. Ornellas, S. B. et al. Ultrasound in Women's Health: Mechanisms, applications, and emerging opportunities. Advanced Materials, e20454 (2026). https://doi.org/10.1002/adma.202520454
9. Varga, E., Pattynama, P. M. T. & Freudenthal, A. Manipulation of mental models of anatomy in interventional radiology and its consequences for design of human–computer interaction. *Cogn. Tech. Work* **15**, 457–473 (2013).
10. Breunig, M., Hanson, A. & Huckabee, M. Learning curves for point-of-care ultrasound image acquisition for novice learners in a longitudinal curriculum. *Ultrasound J.* **15**, 31 (2023).
11. Fenster, A., Parraga, G. & Bax, J. Three-dimensional ultrasound scanning. *Interface Focus* **1**, 503–519 (2011).
12. Klatzky, R. L., Wu, B. & Stetten, G. Spatial representations from perception and cognitive mediation. *Curr. Directions Psychol. Sci.* **17**, 359–364 (2008).
13. von Ramm, O. T. & Smith, S. W. Real time volumetric ultrasound imaging system. *J. Digit Imaging* **3**, 261–266 (1990).
14. Wildes, D. G. et al. Elevation performance of 1.25D and 1.5D transducer arrays. *IEEE Trans. Ultrason. Ferroelectr. Frequency Control* **44**, 1027–1037 (1997). Sept.
15. C. Herickhoff, J. Lin and J. Dahl, "Low-cost Sensor-enabled Freehand 3D Ultrasound. 2019 IEEE International Ultrasonics Symposium (IUS), 498-501 (IUS, 2019), https://doi.org/10.1109/ULTSYM.2019.8925917.
16. Johnson, A. N. et al. Ultrasound-guided needle technique accuracy: prospective comparison of passive magnetic tracking versus unassisted echogenic needle localization. *Regional Anesthesia Pain. Med.* **42**, 223–232 (2017b).
17. Kwitt, R., Vasconcelos, N., Razzaque, S. & Aylward, S. Localizing target structures in ultrasound video - a phantom study. *Med. Image Anal.* **17**, 712–722 (2013).
18. Mozaffari, M. H. & Lee, W. Freehand 3-D ultrasound imaging: a systematic review. *Ultrasound Med. Biol.* **43**, 2099–2124 (2017).
19. Rüger, C. et al. Ultrasound in augmented reality: a mixed-methods evaluation of head-mounted displays in image-guided interventions. *Int. J. Comput Assist Radio. Surg.* **15**, 1895–1905 (2020).
20. Krönke, M. et al. Tracked 3D ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. *PLoS One* **17**, e0268550 (2022).
21. Stetten, G. D. & Chib, V. S. Overlaying ultrasonographic images on direct vision. *J. Ultrasound Med.* **20**, 235–240 (2001).
22. Al-Nimer, S. et al. 3D holographic guidance and navigation for percutaneous ablation of solid tumor. *J. Vasc. Interventional Radiol.* **31**, 526–528 (2020).
23. Nguyen, T., Plishker, W., Matisoff, A., Sharma, K. & Shekhar, R. HoloUS: Augmented reality visualization of live ultrasound images using HoloLens for ultrasound-guided procedures. *Int. J. Comput Assist Radio. Surg.* **17**, 385–391 (2022).
24. von Haxthausen, F., Moreta-Martinez, R., Pose Díez de la Lastra, A., Pascau, J. & Ernst, F. UltrARsound: in situ visualization of live ultrasound images using HoloLens 2. *Int. J. Computer Assist. Radiol. Surg.* **17**, 2081–2091 (2022).
25. Montoro, R. et al. Mixed reality ultrasound-guided mini-ECIRS with apple vision Pro™ - first case report. *Int. Braz. J. Urol.* **51**, e20240610 (2025).
26. V. Nirmal, O. G. Ngwu, E. P. Sridhar, M. A. Habib, and M. Rahman, "Depth Perception and Task Performance in Robotic Teleoperation," in *Proceedings of the IISE Annual Conference*, 2024.
27. Ng, K. W. et al. HoloPOCUS: Portable Mixed-Reality 3D Ultrasound Tracking, Reconstruction and Overlay. In *Simplifying Medical Ultrasound* (eds Kainz, B. et al.) 111–121 (2023).
28. Maddali, D., Brun, H., Kiss, G., Hjelmervik, J. M. & Elle, O. J. Spatial orientation in cardiac ultrasound images using mixed reality: design and evaluation. *Front. Virtual Real.* **3**, 881338 (2022).
29. von Haxthausen, F., Rüger, C., Sieren, M. M., Kloeckner, R. & Ernst, F. Augmenting image-guided procedures through in situ visualization of 3D ultrasound via a head-mounted display. *Sensors* **23**, 2168 (2023).
30. Soares, I., B. Sousa, R., Petry, M. & Moreira, A. P. Accuracy and Repeatability Tests on HoloLens 2 and HTC Vive. *Multimodal Technol. Interact.* **5**, 47 (2021).
31. Cutolo, F., Fontana, U., Carbone, M., Ferrari, V. & Ferrari, M. Architecture of a hybrid video/optical see-through head-mounted display-based augmented reality surgical navigation platform. *IEEE Trans. Med. Robot. Bionics* **4**, 203–214 (2022).
32. Davrieux, C. F. et al. Mixed reality navigation system for ultrasound-guided percutaneous punctures: a pre-clinical evaluation. *Surg. Endosc.* **34**, 226–230 (2020). JanEpub 2019 Mar 25. PMID: 30911919.
33. C. Marcus., et al. Real-time 3D ultrasound imaging with an ultra-sparse, low power architecture. *Adv. Healthcare Mater*. e05310 (2026). https://doi.org/10.1002/adhm.202505310.
34. Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F. J. & Marín-Jiménez, M. J. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognit.* **47**, 2280–2292 (2014).
35. Masopa, D., Tityiwe, J. S., Ngwenya, S. & Sibanda, L. Error and discrepancy in ultrasound reporting by sonographers: inevitable or negligence. *Open J. Radiol.* **15**, 122–135 (2025).
36. Park, M. K. et al. Effectiveness of US surveillance of hepatocellular carcinoma in chronic hepatitis B: US LI-RADS visualization score. *Radiology* **307**, e222106 (2023).
37. McGee, D. C. & Gould, M. K. Preventing complications of central venous catheterization. *N. Engl. J. Med.* **348**, 1123–1133 (2003).
38. Zhang, G. et al. Augmented reality for point-of-care ultrasound-guided vascular access in pediatric patients using Microsoft HoloLens 2: a preliminary evaluation. *J. Med. Imaging* **11**, 062604 (2024).
39. Dodge, K. L., Lynch, C. A., Moore, C. L., Biroscak, B. J. & Evans, L. V. Use of ultrasound guidance improves central venous catheter insertion success rates among junior residents. *J. Ultrasound Med.* **31**, 1519–1526 (2012).

40. Marhofer, P., Greher, M. & Kapral, S. Ultrasound guidance in regional anaesthesia †. *Br. J. Anaesth.* **94**, 7–17 (2004).
41. Costa, N. et al. Augmented reality-assisted ultrasound breast biopsy. *Sensors* **23**, 1838 (2023).
42. Baum, Z. et al. Augmented reality training platform for neurosurgical burr hole localization. *J. Med. Robotics Res.* **4**, 1942001:1-1942001:13 (2019).
43. Lu, N., Foiret, J., Guo, Y., Yoon, B. C. & Ferrara, K. W. Improving real-time ultrasound spine imaging with a large-aperture array. *Sci. Adv.* **11**, eadw2601 (2025).
44. Perrin, D. P., Vasilyev, N. V., Marx, G. R. & del Nido, P. J. Temporal enhancement of 3D echocardiography by frame reordering. *JACC Cardiovasc Imaging* **5**, 300–304 (2012).
45. Cohen, R. & Eldar, Y. C. Sparse convolutional beamforming for ultrasound imaging. *IEEE Trans. Ultrason., Ferroelectr. Frequency Control* **65**, 2390–2406 (2018).
46. Cattari, N. et al. Wearable AR and 3D ultrasound: towards a novel way to guide surgical dissections. *IEEE Access* **9**, 156746–156757 (2021).
47. Carbajal, G., Lasso, A., Gómez, Á & Fichtinger, G. Improving N-wire phantom-based freehand ultrasound calibration. *Int. J. Computer Assist. Radiol. Surg.* **8**, 1063–1072 (2013).
48. Ameler, T. et al. A comparative evaluation of SteamVR tracking and the Optitrack system for medical device tracking. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* **2019**, 1465–1470 (2019).
49. Rosenthal, M. et al. Augmented reality guidance for needle biopsies: an initial randomized, controlled trial in phantoms. *Med. Image Anal.* **6**, 313–320 (2002).
50. Computerized Imaging Reference Systems, Inc. (2013). Multi-Purpose, Multi-Tissue Ultrasound Phantom Model 040GSE. https://www.cirsinc.com/wp-content/uploads/2021/09/040GSE-DS-093021.pdf
51. Hart, S. G. NASA-Task Load Index (NASA-TLX); 20 years later. *Proc. Hum. Factors Ergonomics Soc. Annu. Meet.* **50**, 904–908 (2006).

Acknowledgements

We sincerely thank all study participants for their invaluable contributions, time, and cooperation, without which this research could not have been completed. We also acknowledge the dedicated support of the staff, equipment, and the guidance provided by the MIT Center for Clinical and Translational Research (CCTR). We also would like to thank Talis Reks, Professor Brian Anthony, MIT.nano, and the Immersion Lab for providing guidance, training, and the high-end head-mounted displays that made this work possible.

Author contributions

J.F.H. conceived the project, designed the study, developed the mixed-reality platform and data pipeline, conducted all experiments, performed data analysis and visualization, and wrote the original manuscript. S.V. contributed to development of the cDAQ imaging system and integration with the mixed-reality environment, performed experiments, devised study methodologies and reviewed and edited the manuscript. C.Di. contributed to methodology, investigation, visualization, and manuscript writing. B.W. contributed to the mixed-reality system conceptualization, implementation, and methodology. T.S. contributed to experimental investigation and manuscript writing. C.D. supervised the project, acquired funding, contributed to visualization, and reviewed and edited the manuscript. C.D. = Canan Dagdeviren. C.Di. = Cinay Dilibal.

Funding

This work was supported by Media Lab Consortia funding, National Science Foundation Award Number 2524831, the MIT HEALS Graduate Fellowship (J.F.H.), and the MIT-Tata Graduate Fellowship (S.V.).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44172-026-00692-7.

Correspondence and requests for materials should be addressed to Canan Dagdeviren.

Peer review information *Communications Engineering* thanks Khoo Eng Tat, who co-reviewed with Kian Wei Ng, Ghazaleh Tanhaei, and the other anonymous reviewer(s) for their contribution to the peer review of this work. Primary Handling Editors: [Geng-Shi Jeng] and [Philip Coatsworth]. A peer review file is available.