Medical Devices: Evidence and Research

Dovepress
Taylor & Francis Group

Open Access Full Text Article

REVIEW

Evaluating Augmented Reality Head-Mounted Devices in Healthcare: A Review of Hardware, Software, and Usability Approaches

Peiming Zhang[1], Zihe Wang[1], Tao Wang[2], Tielong Liu[3], Jing Wang[3], Yimeng Gao[1], Weiqi Li[1]

[1]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, 200093, People's Republic of China; [2]Department of Electrical Engineering, Shandong Vocational College of Industry, Zibo, 256414, People's Republic of China; [3]Department of Orthopedics, Changzheng Hospital, Naval Medical University, Shanghai, 200003, People's Republic of China

Correspondence: Peiming Zhang, University of Shanghai for Science and Technology, No. 516, Jungong Road, Yangpu District, Shanghai, People's Republic of China, Email zpmking@163.com

Abstract: Augmented reality head-mounted devices (AR HMDs) are increasingly deployed in healthcare. Given the stringent safety and efficacy requirements of medical settings, proactive quantitative testing of key performance attributes prior to deployment is critical for risk assessment. A systematic performance evaluation framework is essential not only to support clinical adoption but also to secure regulatory approval. This review systematically summarizes hardware, software, and usability assessment methods for AR HMDs in healthcare, analyzes current research and experimental designs, and identifies challenges arising from device heterogeneity, limited coupling with real-world clinical scenarios, and subjective bias. To address these issues, we propose five design principles to guide the development of objective and practical evaluation methods: (1) identify key components based on core functions; (2) prioritize testing by functional contribution; (3) replicate authentic clinical and human-visual conditions; (4) objectify subjective perception; (5) test functionally linked components jointly.

Keywords: augmented reality, head-mounted display, performance evaluation, usability assessment, visual perception

Introduction

Augmented Reality (AR) is a technology that overlays computer-generated virtual information onto the real world to enhance the user's perception and interaction capabilities. A head-mounted display (HMD) is a device worn on the user's head that typically overlays virtual images onto the user's field of view through optical see-through or video see-through methods. It also incorporates built-in cameras and sensors to acquire spatial information, enabling precise registration with the real-world scene.[1] AR HMDs commonly integrate display, sensing, tracking, and computing functions in their hardware design. Depending on the level of integration and power configuration, current devices can be categorized into two main types: one type integrates all the functions and power into a single device; the other type offloads computing or power modules to an external mobile terminal, achieving overall functionality through a wired connection. Additionally, AR HMDs are often combined with visual markers or natural feature recognition to ensure that virtual images are displayed at their intended locations and updated in real-time.

With their display capabilities, simple structure, and ease of learning, AR HMDs are rapidly expanding in medicine. Some AR HMD systems have received FDA 510(k) clearance and are being used clinically,[2] including spine surgery,[3–5] dental treatment,[6–8] knee replacement surgery,[9,10] hip surgery,[11,12] tumor resection,[13–15] nasopharyngeal surgery,[16,17] and plastic surgery.[18] Furthermore, their value in rehabilitation is increasingly evident, with AR HMDs offering a cost-effective and user-friendly alternative for rehabilitation assessment and training.[19] In medical training, they can create immersive training environments for physicians and trainees,[20–22] enhancing training efficiency and reducing costs. Unlike consumer AR, medical AR participates in patient care and must meet stricter demands: millimeter spatial accuracy, high reliability and stability, patient safety assurance, sterile compatibility, and compliance with medical-device regulations and ergonomic standards.[23]

Received: 31 May 2025
Accepted: 2 August 2025
Published: 22 August 2025

Translating medical AR from concept to clinic demands deeper, context-specific optimization. Most systems remain anchored to consumer-grade hardware and software.[24] Under the FDA Medical Extended Reality Program,[25] current AR HMD performance testing in healthcare confronts the following challenges:[26,27]

1. Insufficient Hardware Performance Evaluation: The performance of key components of current AR HMDs, such as displays and cameras, has not been adequately clinically validated. For example, displays may suffer from veiling glare, which can impair the presentation of medical image details and reduce diagnostic accuracy.[28]
2. Insufficient Software Performance Evaluation: The software development platforms relied upon by AR HMDs exhibit significant heterogeneity, and there is currently no unified evaluation standard. For instance, color rendering deviations by virtual engines may lead to color distortions in medical images, affecting doctors' diagnoses and decision-making.[29]
3. Lack of Objective Usability Assessment Standards: Current usability assessment studies are mostly based on subjective evaluations, and there is a lack of objective measurement methods for individual differences in the Human Visual System (HVS).[30] Additionally, the usability metrics required for some clinical scenarios are not unified, making it difficult to form a common evaluation standard.

These challenges collectively reveal a deeper, systemic bottleneck: the absence of a unified, standardized performance-evaluation framework for medical AR HMDs, resulting in fragmented research, high translational barriers, and constrained industrial growth. This paper centers on evaluation methods for hardware, software, and usability, reviews existing evaluation approaches, and analyzes their challenges and optimization strategies in medical applications. Finally, the paper proposes five principles for the design of evaluation methods, as shown in Figure 1, to build a more scientific and comprehensive performance evaluation system and further promote the application and development of AR HMDs in the medical field.

Search Strategy

A systematic literature search was conducted exclusively within the Web of Science Core Collection, limited to English-language records published between 2015 and 2025. The Boolean query was formulated as follows: ("augmented reality" OR AR) AND ("head-mounted device*" OR "head-mounted display*" OR HMD*) AND (performance OR accuracy OR evaluation OR assessment OR testing OR validation OR regulation) AND (healthcare OR surgery OR rehabilitation OR training).

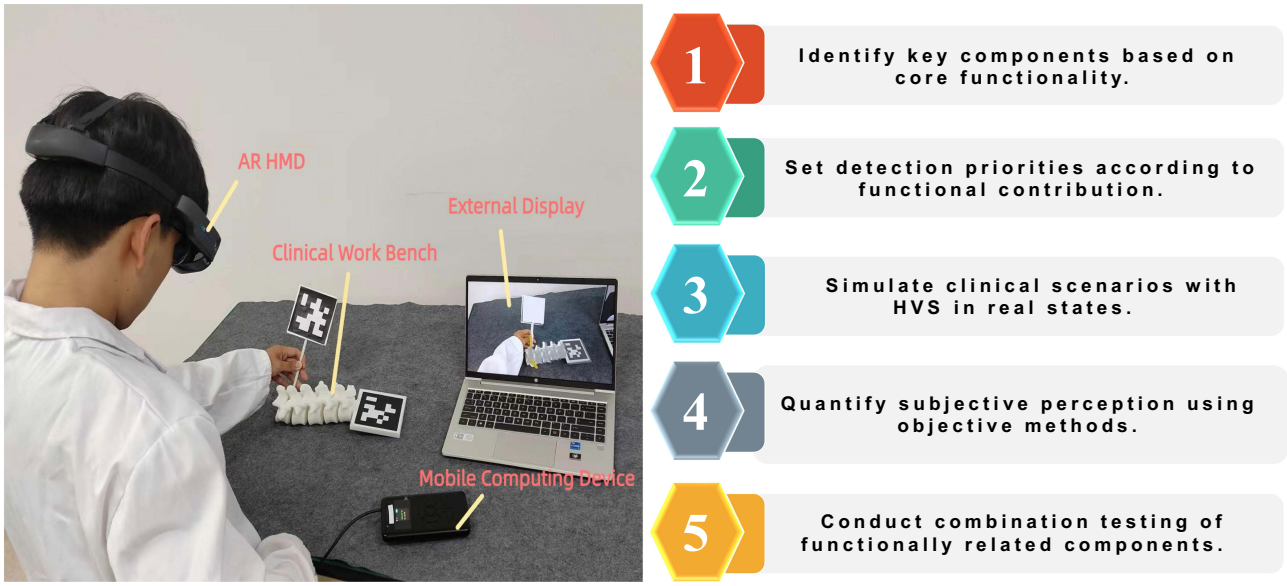

Figure 1 Five design principles for constructing a performance evaluation framework for AR HMDs in healthcare.

The search returned 677 unique citations. After automated deduplication and manual cross-checking, two reviewers independently screened titles and abstracts against predefined eligibility criteria (Cohen's $\kappa = 0.78$); disagreements were resolved through consensus, yielding 300 candidate articles. Full-text evaluation resulted in the inclusion of 70 studies that explicitly reported performance testing or validation protocols for AR-HMD systems in healthcare contexts. Despite the systematic approach, the inherent breadth of the query and the interpretive nature of screening may have introduced residual subjectivity.

Hardware Performance Evaluation

The hardware of AR HMDs is the main carrier for functional implementation. The performance of the display, camera, and processor is crucial and directly affects the safety and effectiveness of medical applications.[31] The display is used for near-eye presentation of AR images, and its image reproduction performance determines the visualization quality of medical images. The camera is responsible for spatial recognition and tracking, and insufficient accuracy and stability can lead to errors in operation guidance. The processor handles data processing and image rendering, and its computational power and response speed affect the overall efficiency of the system. Therefore, it is important to conduct objective and precise quantitative evaluation of the performance of these core hardware components to provide a basis for the safe and compliant use of these devices in clinical applications.[26]

Display Image Quality

The display is a core component of AR HMDs and directly affects the visualization quality of medical images,[26] which in turn can influence the diagnostic and therapeutic processes of physicians. There are many types of display technologies for AR HMDs,[31,32] including Half Mirror, Birdbath, Free-Form Prism, Waveguide, Retinal Scanning, Multiple LCDs, Microlens Arrays, Pinlight, and 3D Holographic Display.[33] Although the implementation technologies vary, the goal is to display medical images and information. Therefore, from the perspective of using the display as the evaluation terminal, it is important to prioritize the evaluation of key performance parameters that are common to HMD displays.[26,27,32] In addition to basic parameters such as display resolution, FOV, and display frame rate, there are also parameters that manufacturers may not specify but could potentially affect the display quality of medical images,[27] including color uniformity, veiling glare, lateral chromatic aberration, and spatiotemporal resolution.[31] Given the characteristics of these parameters, each needs to be detected separately, and the most suitable evaluation method is often related to specific hardware evaluation methods. Traditional evaluation methods for medical flat-panel displays do not consider spatial-related image artifacts caused by additional optical components such as HMD lenses,[27] so it is necessary to explore evaluation methods that are suitable for HMDs. Additionally, since these devices use near-eye display, it is important to fully consider and simulate the conditions under which the HVS observes during evaluation, including factors such as interpupillary distance (IPD), pupil entry position, convergence angle, and scanning field of view.

Existing evaluation methods for the above parameters mainly include spot measurement methods based on optical sensors, indirect measurement methods based on reflective light paths, and precision measurement methods based on optical platforms. Spot measurement methods typically place measurement tools directly behind the HMD display for evaluation. Tools are not limited to photometers,[33] colorimeters,[34] spectroradiometers,[35] and cameras[36] to measure parameters such as the color, brightness, and contrast of the HMD display, as shown in Figure 2(A). However, it is difficult to precisely control the pupil entry position during measurement, which affects the accuracy of the measurement. At the same time, the measurement space behind AR HMDs is narrow and some lenses are curved, making direct measurement difficult. Therefore, some studies have proposed indirect measurement methods based on reflective light paths. By changing the position of the measuring instrument, the difficulty of arranging the measuring instrument is simplified, thereby improving the evaluation accuracy. Some studies have placed a front-surface mirror at a 45° angle at the HMD observation position, reflecting the displayed content completely to the external measurement device,[37] improving measurement stability, as conceptually shown in Figure 2(B). However, this measurement process may introduce diffuse reflection errors. A method to avoid this problem is to use a conical measurement probe connected to an integrating sphere, placed behind the HMD to measure the parameters, reducing the interference of bright areas with the measurement signal.[36] Precision measurement methods based on optical platforms use high-precision optical platforms to fix the measuring

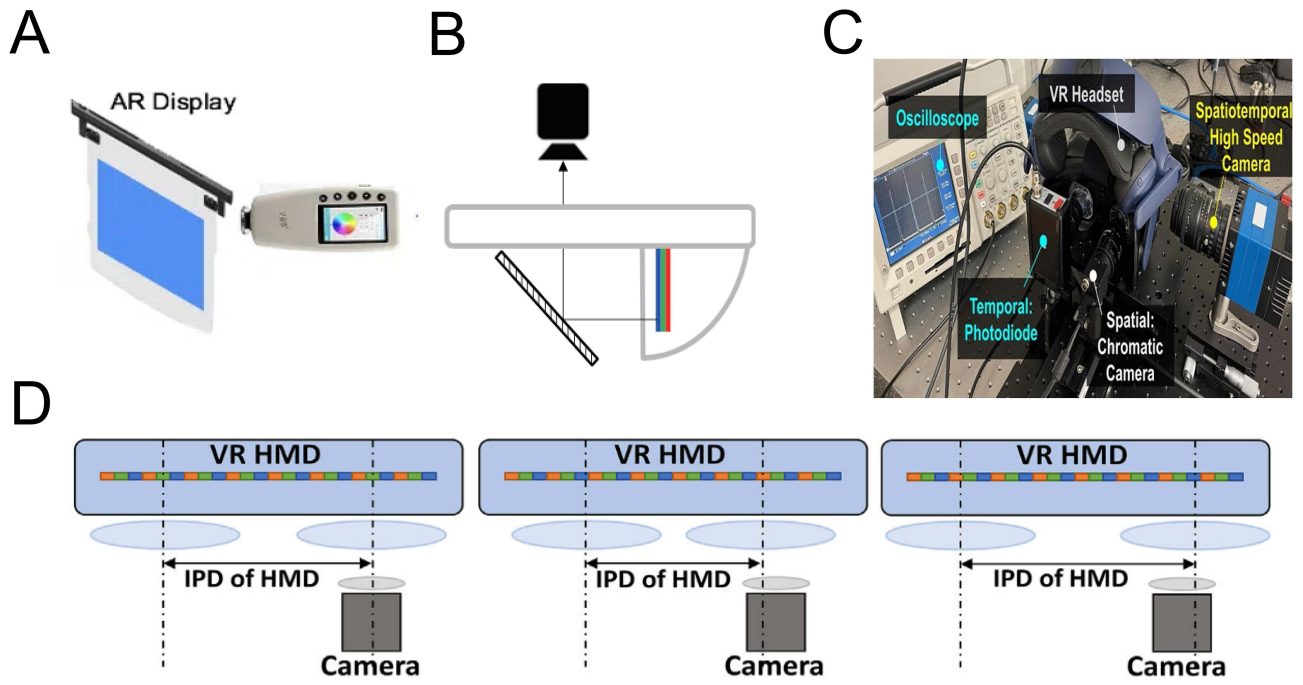

Figure 2 Methods for evaluating AR HMD display performance: (**A**) Instrument-based direct measurement. (**B**) Displacement-based measurement. (**C**) Test bench-based precise measurement; Reprinted from Zhao C, Kim AS, Beams R, Badano A. Spatiotemporal image quality of virtual reality head mounted displays. *Sci Rep*. 2022;12(1):20235. Creative Commons.[37] (**D**) Simulated IPD measurement; Reprinted from Zhao C, Beams R, Badano A. Radially variant contrast measurement in virtual reality headsets using circular concentric ring patterns. *J Soc Inf Disp*. 2023;31(5):387–397. This article is a U.S. Government work and is in the public domain in the USA.[38]

devices and combine them with precision adjustment devices such as electric linear translation stages and universal stages to precisely control the measurement angle and spatial position, ensuring precise and stable measurement of key parameters in the center and edge areas of the FOV, as conceptually shown in Figure 2(C).[28] In addition, when designing specific evaluation schemes, it is also necessary to consider the characteristics of the evaluation parameters and build dedicated functional evaluation platforms. For example, to detect the spatial and temporal characteristics (spatiotemporal resolution) of displaying moving images, some studies have combined cameras with silicon photodetectors to detect the spatial resolution, temporal response, and spatiotemporal performance of the display under motion. At the same time, the display content of the HMD can be driven by an external graphics card to eliminate the error of system delay on the evaluation results.[38] The main advantage of such methods is that they can significantly improve the reproducibility of experiments, reduce the impact of measurement device placement errors or environmental variable fluctuations, and thus improve the stability and reliability of evaluation data.

In addition, several evaluation studies have pointed out that attention should be paid to the settings of HVS and lighting conditions during the evaluation of display performance to better match the actual clinical use and reflect the performance status of the HMD under real use conditions. Key factors of the HVS include IPD, pupil entry position, and pupil size. For IPD adjustment, some devices can use the HMD's own parameter adjustment settings (common in VST devices), as shown in Figure 2(D).[39] For OST AR devices, the measurement device can be moved horizontally to achieve the same effect of IPD adjustment. Similarly, the influence of different pupil entry positions can be simulated by moving the evaluation device radially forward and backward. Different pupil sizes can be simulated by adjusting the aperture of the observation device's camera or installing an iris diaphragm, etc., to simulate the size of the human eye pupil under actual use conditions.[40]

The image quality evaluation of the display should also appropriately consider the combination with other related hardware. For example, in the operating room environment, OST AR HMDs often use colored sunshades to reduce environmental light interference and enhance the AR effect. However, some studies have found that they may reduce contrast perception in actual use.[33] Therefore, it is recommended to include functional implementation-related accessories in the display performance evaluation to improve the reliability of the evaluation experiment design.

Camera Tracking Accuracy

Another key function of AR HMDs in medical application scenarios is spatial positioning and tracking.[41] This function mainly relies on the built-in camera of the HMD combined with computer vision algorithms (based on or not based on visual markers) to be realized. The tracking accuracy has a direct impact on the quality of clinical applications, especially in high-risk scenarios such as intraoperative treatment. Therefore, it is crucial to conduct systematic evaluation of the camera tracking accuracy of AR HMDs.[42]

Currently, the evaluation of HMD camera tracking performance is mainly based on two types of methods. The first type of method assesses the accuracy by measuring the HMD's ability to track a known spatial location. That is, a target point with a known spatial location in the real world is selected as a reference point. At the same time, the tracking data of the HMD system for this point is recorded as the experimental point, and the spatial distance information between the two points is compared and analyzed for error[43] (as shown in Figure 3(A)). This type of method has strong flexibility and can be customized according to specific medical application scenarios. For example, in a simulated incision surgery scenario, subjects can be asked to wear the HMD and perform a pre-set tracking curve task on a tablet to test the accuracy of the tracking system. Finally, the tracking accuracy is assessed by comparing the deviation between the curve trajectory drawn under the AR-assisted state and the pre-set curve trajectory.[44] However, this method may be affected by the perceptual bias of the HVS, thereby introducing subjective judgment error.[45] To reduce the interference of subjective factors, cameras and other devices can be used to replace the human eye for measurement[46] (as shown in Figure 3(B)). Alternatively, tasks that do not require subjective judgment can be designed. For example, a movable grid plate with evenly distributed holes for installing tracking markers recognizable by the HMD can be used. By moving the grid plate at a fixed distance on a slide rail and adjusting the two-dimensional movement distance and longitudinal depth, a quantitative analysis of the three-dimensional tracking performance of the HMD camera can be achieved.[43] However, this method is mainly suitable for static tracking accuracy assessment and is difficult to fully reflect the demand for high-precision dynamic tracking ability in medical scenarios.

The design of the second type of evaluation method mainly involves introducing an external tracking system. By performing the same task and comparing the tracking data recorded by the HMD and the external system, the tracking performance of the HMD camera can be quantified, as shown in Figure 3(C). This method takes advantage of the non-interference characteristic between Inside-Out and Outside-In tracking methods and the stability of the external independent tracker hardware.[48] In specific experiments, recognition markers are usually attached to the HMD and the target object so that they can be precisely tracked in six degrees of freedom by the external tracking system.[47] Subsequently, by setting and performing standardized tracking tasks, the measurement data of the HMD system and the external tracking system are compared to quantify tracking accuracy and reliability.[49] To meet the demand for high-precision tracking in medical applications, higher-precision tracking systems can also be built or used in the evaluation. For example, using high-power infrared emitters as markers and equipping them with dual photodiode stereo sensors with infrared low-pass filters for tracking can achieve a sampling rate as high as 50kHz and a motion-to-pose latency as

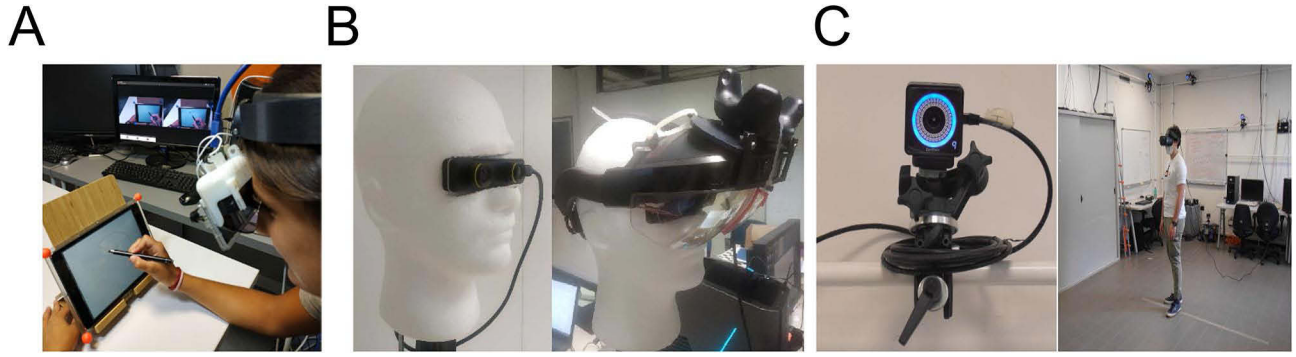

Figure 3 Methods for evaluating AR HMD camera tracking performance: (**A**) Comparison with known real-world locations under AR guidance; Reprinted from Cutolo F, Fida B, Cattari N, Ferrari V. Software framework for customized augmented reality headsets in medicine. *IEEE Access*. 2019;8:706–720. Creative Commons.[43] (**B**) Camera-based assessment; Reprinted from Ballestin G, Chessa M, Solari F. Assessment of optical see-through head mounted display calibration for interactive augmented reality. 2019. © Copyright 2019 IEEE.[45](**C**) Comparison using an external tracking system; Reprinted from Monica R, Aleotti J. Evaluation of the oculus rift s tracking system in room scale virtual reality. *Virtual Real*. 2022;26(4):1335–1345. Creative Commons.[47]

https://doi.org/10.2147/MDER.S541187

low as 28μs.[50] In addition, considering that the HMD may operate for a long time in medical applications, the stability of its camera performance also needs to be assessed. Some studies have run the HMD camera for a long time and statistically analyzed the ranging data of the ToF camera at different time intervals (such as 40 minutes after startup, 40–60 minutes, and 1 hour later) to evaluate its long-term operational reliability.[51]

To adapt to more complex medical application scenarios, such as surgical navigation involving soft tissues, markerless inside-out tracking technology is becoming a research focus.[52,53] This technology eliminates the dependence on external markers, thereby enhancing the system's applicability and clinical operational convenience to a certain extent. However, compared with marker-based tracking systems, the accuracy of markerless tracking technology is more susceptible to environmental lighting, target surface characteristics, and dynamic changes, and there are still relatively limited methods for detecting its accuracy separately. Therefore, further in-depth exploration of the accuracy evaluation of markerless tracking systems is needed to ensure their reliability and feasibility in clinical environments.

Performance Evaluation of Key Hardware Components

In the process of realizing the core functions of HMDs, in addition to the display and camera, other key components that significantly impact the device's performance also need to be tested accordingly. However, the current heterogeneity of HMD hardware architectures is quite large. Specific tests should be based on the actual composition of the hardware and prioritize evaluation according to their importance in the main function implementation. At the same time, the evaluation methods should match the characteristics of the components to ensure the scientific nature and accuracy of the evaluation. For example, in real-time intraoperative navigation application scenarios, the built-in processor of the HMD plays a key role in system latency and frame rate, which affects the user experience of doctors. Therefore, its performance needs to be tested through experiments. Some studies have used high-speed cameras to capture and analyze frame image changes to quantify performance.[54] Similarly, a evaluation platform based on microcontrollers, photodetectors, and microphones can be built to detect the asynchrony of audiovisual stimuli.[55] However, research on the functional contribution of other key hardware components of AR HMDs and their corresponding evaluation methods is still limited and requires further exploration and study.

Software Performance Evaluation

As the core component for hardware function scheduling, the software performance of AR HMDs has a decisive impact on medical applications. However, existing AR HMD software is mostly developed based on consumer-grade software development platforms, without fully considering the operational constraints and high-precision requirements of medical environments. Moreover, there is a lack of unified evaluation standards to verify their reliability in medical guidance. The focus of function implementation and performance evaluation lies in the precision of virtual-to-real registration, the fidelity of image rendering, and the medical adaptability of the software framework.

Registration Accuracy of Virtual Models

Virtual-to-real registration refers to the key process of precisely aligning and superimposing virtual models or images with real-world scenes in terms of spatial position, scale, and orientation, which is often jointly realized in combination with the camera hardware of the HMD. The precision of virtual-to-real registration of AR HMDs is one of the important links to ensure the accuracy of medical applications. The mainstream registration techniques currently include manual registration, computer-aided registration, and external registration.[48,56] Although these techniques differ in implementation methods, their common goal is to achieve precise alignment between virtual and real models, and evaluation can be based on this feature to establish a standardized evaluation scheme.

The key to registration precision evaluation methods can be understood as calculating the spatial error between the AR registration point and the actual physical reference point. Compared with the performance evaluation of hardware camera tracking, virtual-to-real registration precision evaluation focuses more on the spatial alignment error between the virtual model and the physical model after the software completes the registration. The commonly used method to achieve evaluation is as follows: First, select one or more spatial points to be measured in the real environment and record their coordinates. Then, establish a matching model in the virtual environment and perform registration. The subject wears the AR HMD and uses a tracking pointer to touch the feature points of the AR virtual image under AR guidance, records the coordinates, and

compares them with the physical reference points for error analysis.[57] Usually, this type of experiment relies on an external optical tracking system to provide a unified coordinate system for measurement. However, in medical scenarios with limited operating space, optical tracking markers may be affected by environmental constraints, limiting their applicability. Therefore, some studies have optimized the external tracking system according to actual evaluation needs, such as using electromagnetic tracking systems to enhance adaptability,[58] as shown in Figure 4(A). However, the measurement accuracy in this type of evaluation experiment is easily affected by individual physiological characteristics (such as IPD, VAC, etc)., which reduces the objectivity and consistency of the evaluation. Therefore, developing objective evaluation methods that can avoid subjective errors and improve measurement reliability is the development direction of this type of performance evaluation.

To enhance the objectivity of measurement, one improved approach is to install a small camera on the HMD to capture image data from the user's perspective, combined with precise displacement control of the HMD using a robotic arm to reduce human operational errors. This method can measure display pattern errors on a standardized evaluation board, providing a new idea for the quantitative assessment of the registration precision of AR HMDs.[59] In addition, another method is to record the AR image with the HMD, export the video, and analyze the spatial deviation between the AR image and the target image frame by frame,[60] as shown in Figure 4(B). Both types of evaluation methods can enhance the objectivity of evaluation, reduce errors brought by subjective assessment, and provide a reference for the experimental design of objective evaluation of the registration precision of AR HMDs.

Image Rendering Fidelity

In medical applications of AR HMDs, virtual images not only need to be accurately registered with the expected anatomical locations but also should maintain high fidelity in terms of color, shape, and proportion of the digital models to ensure the credibility of diagnostic and treatment decisions. Currently, the rendering of most three-dimensional medical digital models mainly relies on consumer-grade engines (such as Unity, Unreal Engine, etc).[22] Since there is no unified rendering standard, different engines may lead to significant differences in model fidelity, thereby affecting the accuracy and stability of intraoperative visualization.

To quantitatively analyze the performance differences of different rendering engines, existing studies have used ColorChecker color cards, TG18-QC test patterns, cell slice images, and grayscale medical images. By controlling a single variable in the rendering engine, the impact of different rendering parameters on color performance is detected, and the CIE 1976 color difference standard is used for quantification.[29] Similarly, other rendering parameters that affect the visual perception of the model, such as materials, textures, light sources, and perspective, also need to be included in

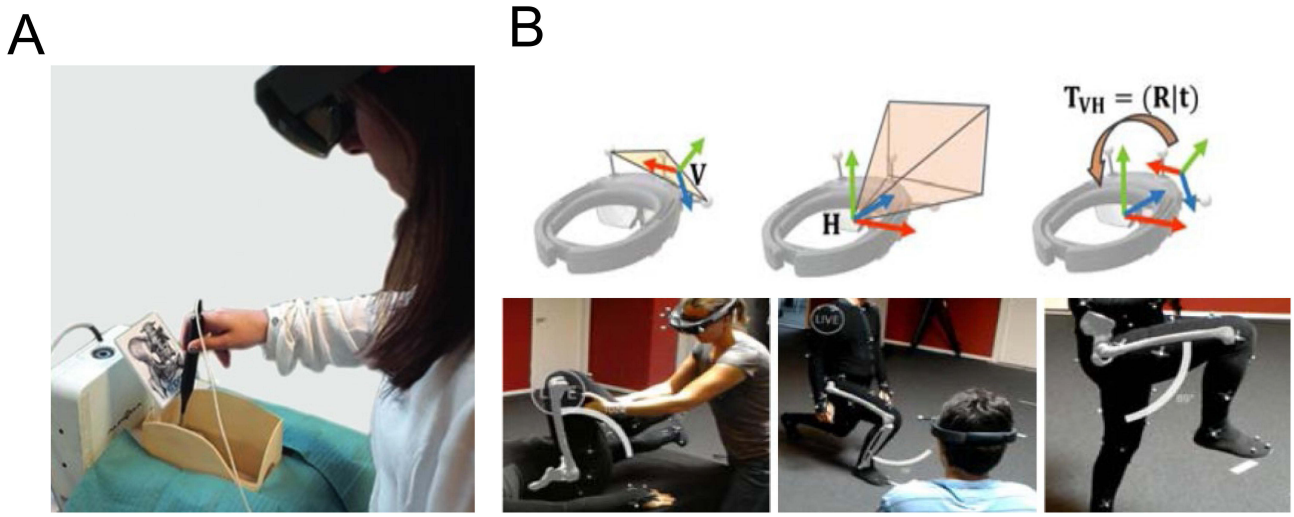

Figure 4 Methods for evaluating AR HMD registration precision: (**A**) Subjective evaluation method; Reprinted from Condino S, Turini G, Parchi PD, et al. How to build a patient-specific hybrid simulator for orthopaedic open surgery: benefits and limits of mixed-reality using the microsoft holoLens. *J Healthc Eng*. 2018;2018(1):5435097. Creative Commons.[56] (**B**) Video-analytical approach using HMD camera captures; Reprinted from Debarba HG, De Oliveira ME, Dermann LA, Chaqu A, S E, Charbonnier C. Tracking a consumer HMD with a third party motion capture system. *IEEE*. 2018:539–540. © Copyright 2018 IEEE.[58]

the evaluation to comprehensively evaluate the rendering quality of the virtual model.[61] However, the above evaluation methods are mostly aimed at the performance analysis of rendering parameters within a single platform and cannot better cope with the current challenge of significant differences in HMD development platforms.

Faced with this challenge, a feasible solution is to develop a cross-platform, easy-to-deploy standardized test toolkit to unify test content and indicators.[62] The toolkit can include benchmark test images or models with the same parameter characteristics and allow users to customize scenes to evaluate key indicators such as color mapping, grayscale response, display resolution, and spatiotemporal characteristics on different devices and software platforms. To meet the compatibility requirements of most medical AR HMDs, test data can be packaged in a universal format (such as JSON) and distributed over the network to achieve automated and standardized testing. Such methods can provide a more efficient and concise new approach for rendering performance testing of various platforms in medical scenarios.

Medical Adaptability of Software Frameworks

The adaptability of software frameworks refers to the ability of AR HMD software platforms to meet diverse medical needs, including accurate responses and compatibility with different clinical departments, hardware configurations, data security, and operational processes. Since medical scenarios have particularly strict requirements for safety, real-time performance, and scalability, medical AR HMDs lacking sufficient adaptability often struggle to cope with challenges brought by cross-departmental, cross-system, and even subsequent technological iterations. Therefore, before building an efficient and unified software framework, it is necessary to fully explore the adaptability standards that meet medical scenarios and their evaluation methods. Currently, the performance evaluation of the medical adaptability of AR HMD software frameworks mainly focuses on the following two ideas: The first is horizontal comparison of specialized software frameworks: by developing or selecting various software frameworks and applying them to similar medical scenarios for comparative testing of core functions (such as real-time tracking and computational efficiency). For example, in AR HMD-guided surgical scenarios, the fast processing capability of a tracking system based on the CUDA architecture within 0.007 seconds[44] can be regarded as an embodiment of high adaptability, capable of meeting the rapidly changing surgical needs. In addition, the selection of different architectures for software frameworks (such as distributed and monolithic) will also affect the adaptability of the framework in multi-department collaboration, telemedicine, or large-scale data processing environments.

The other is comprehensive scoring based on standardized development guidelines: according to the standardized guidelines for AR HMD applications, a series of evaluation indicators covering software design to application construction are established. The core lies in measuring the performance of software in dimensions such as graphical programming, application programming, analysis, and optimization, and assessing whether it supports various medical image formats, and whether it is easy to develop secondary to meet different clinical needs and other "adaptability" ky points.[63] Through these multidimensional scores, the generality and scalability of the software in a broader range of medical scenarios can be judged.

When assessing the adaptability of medical AR HMD software frameworks, it is necessary to closely integrate with the needs of the clinical front line. Inviting doctors or relevant medical teams to participate in the assessment helps to set standards for application scenarios and processes that are closer to clinical work, such as the precision requirements for medical image rendering and surgical navigation, the integration needs for multi-department general functions, and the expandability of secondary development.

Usability Assessment

The usability assessment of AR HMDs is crucial for ensuring the success of medical procedures, as it greatly affects the operational experience of medical staff and clinical outcomes.[64] When evaluating the usability of AR HMDs, the focus is on considering the real operational environment of medical professionals and the impact of the HVS.[26,31] The factors involved mainly include the human information processing system, lighting environment, color environment, human-computer interface design, and multiple other dimensions.[65] These factors directly or indirectly affect the terminal operation or judgment of medical users, thereby impacting the quality of medical care. However, the evaluation of these factors often involves subjective perception, with significant individual differences, making it difficult to achieve accurate and objective quantitative evaluation like hardware and software performance testing. Quantitative indicators for

measuring these parameters are challenging for assessing usability impact, but this does not prevent the continuous exploration of more precise and practical evaluation methods in the assessment.

Visual Information Processing System

In AR medical applications, the efficacy evaluation of the visual information processing system is a crucial aspect to ensure the usability of AR HMDs. Its core lies in quantifying the user's egocentric depth perception of the displayed information, a parameter that directly affects the user's accuracy in integrating virtual information with the real environment. Therefore, the precision assessment of this parameter is particularly critical for AR HMDs in high-precision medical scenarios.

Traditional evaluation paradigms mainly rely on subjective measurement methods, employing a mixed assessment framework of depth estimation tasks (Depth Estimation Task) and Likert scale questionnaires.[66] These tasks may include observing and estimating the depth of AR content display, spatial perception through finger movements, etc., as shown in Figure 5(A). However, these methods fall short in terms of precision and reproducibility of experimental results, failing to meet the high standards required in medical scenarios.[67]

To achieve higher precision in evaluation, some studies have attempted to transform human subjective perception results into quantifiable objective data. A typical approach is to use an improved optical platform, with plastic tubes installed on both sides of the platform: one side for fixing and supporting the framework of the physical model, and the other side equipped with a movable tube with a pointer. The subject moves the tube under the table to operate the pointer to point to the target position, thereby objectively quantifying the depth perception error of the AR HMD,[68] as shown in Figure 5(B). This device can measure errors in the millimeter range, approaching the 1 mm or smaller precision required for AR image-guided brain surgery.[69] In addition, when evaluating depth perception, multiple depth cues, including model material, lighting environment, shadows, and textures, should be comprehensively considered,[61] and the experimental conditions should be consistent with the attributes of the test objects.

Since depth perception assessment is inherently affected by subjective factors and individual differences cannot be completely eliminated, based on existing research, it is recommended to combine high-precision objective measurement platforms with corresponding subjective questionnaires when constructing evaluation methods to form a comprehensive

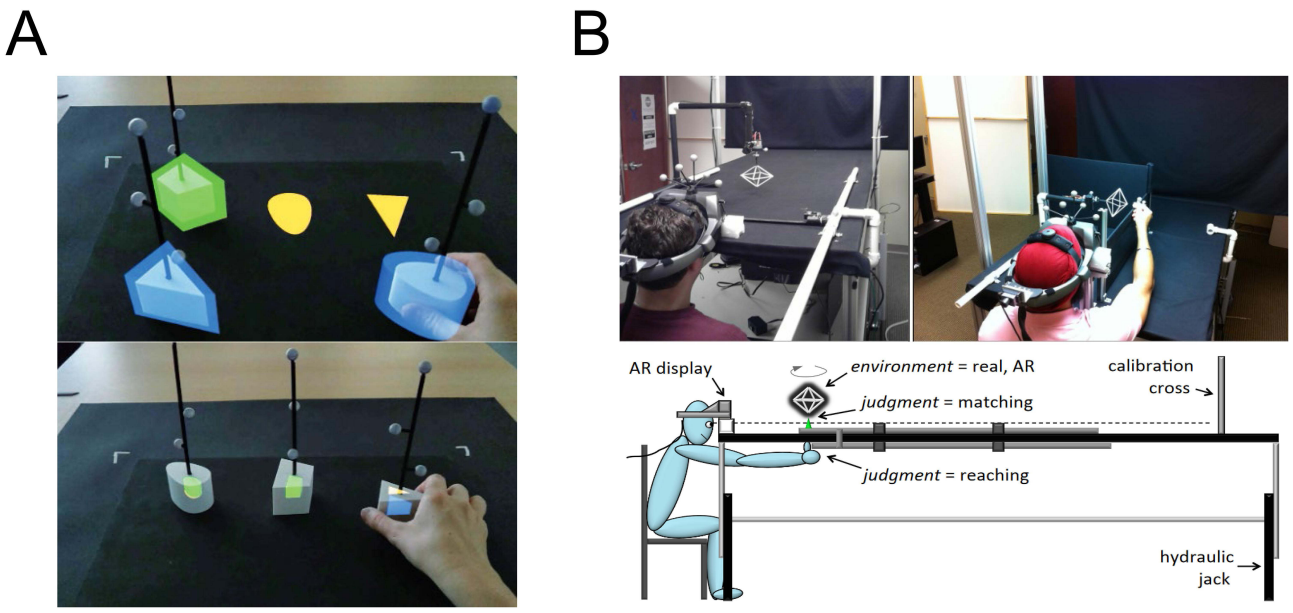

Figure 5 Methods for evaluating depth perception in AR HMDs: (**A**) Subjective assessment; Reprinted from Kahl D, Ruble M, Kr U, Ger A. The influence of environmental lighting on size variations in optical see-through tangible augmented reality. *IEEE*. 2022:121–129. © Copyright 2022 IEEE.[67] (**B**) Objective quantification of subjective judgments using experimental apparatus; Reprinted from Swan JE, Singh G, Ellis SR. Matching and reaching depth judgments with real and augmented reality targets. *IEEE Trans Vis Comput Graph*. 2015;21(11):1289–1298. © Copyright 2015 IEEE.[64]

evaluation scheme for AR HMD depth perception assessment,[54] which can improve quantification accuracy while also considering the subjective differences of the HVS.

Lighting Environment

The lighting environment is one of the important variables affecting the usability of AR HMDs. Changes in light intensity, direction, and nature can significantly affect the user's visual perception accuracy of virtual content, thereby directly affecting the effectiveness and safety of medical operations. In specific medical environments such as operating rooms, complex lighting conditions (such as strong light interference from shadowless lamps) can greatly affect the visibility and accuracy of AR display information.

To systematically detect the impact of lighting conditions on AR HMDs, some studies have set up experimental scenarios in different lighting environments and asked subjects wearing AR HMDs to manually adjust the physical model to match the fixed virtual model while using an external tracker to record the matching error between the physical and virtual positions.[70] On this basis, more practical elements of medical scenarios, such as zoning settings, personnel movement, and anchor placement, can be integrated,[71] and a more comprehensive assessment of lighting effects can be achieved by adjusting light intensity or type. Evaluation indicators can be selected in combination with the real needs of the operating room, such as recording the accuracy of hologram overlay, task completion time, number of registration rejections, real-time overlay delay of holograms, and number of operational failures. It is worth noting that the lighting environment may also indirectly affect device performance, such as the degree of interference with the gesture and voice recognition functions of AR HMDs by surgical gloves of different colors under different lighting conditions.[72] Therefore, in the assessment of the lighting environment, the dual impact of lighting on the main performance of AR HMD hardware and software and the overall workflow of the operating room should be comprehensively considered.

Clinical Usability Analysis

With the increasing research on the practical medical applications of AR HMDs, reaching a consensus on therapeutic effects has gradually become one of the important indicators for assessing the usability of AR HMDs. The evaluation criteria for clinical applicability need to be adjusted according to specific application targets, especially for surgical applications, which require higher strictness and comprehensiveness in the usability assessment of HMDs. Current clinical usability studies of AR HMDs mainly focus on two aspects: the results of clinical application performance and the postoperative reports from clinical doctors.

In surgical scenarios involving rigid anatomical structures, the clinical usability research of AR HMDs is relatively mature, and more unified methods and standards for usability evaluation have emerged in some specific surgeries. For example, in spinal pedicle screw surgery, researchers usually analyze the intraoperative or postoperative fluoroscopic images of screw placement and assess the position and angular deviations of pedicle screws or K-wires according to the Gertzbein-Robbins screw placement grading criteria. In addition, some studies also consider clinical indicators such as screw placement time, intraoperative blood loss, and related complications to comprehensively evaluate the application effects of AR HMDs in actual surgeries,[3,5,70–73] as shown in Table 1. In the field of dental implant surgery, the accuracy of AR HMD-guided implant placement is often assessed by comparing the deviations in the coronal, apical, and angular aspects between preoperative planning and postoperative implants,[7,8] as shown in Table 2. For AR HMD-assisted acetabular cup placement surgery, the position and orientation of the acetabular cup relative to the pelvis after surgery are often used as the evaluation baseline.[11] The usability evaluation indicators for clinical application scenarios of AR HMDs are becoming standardized, promoting their promotion and application.[12]

In contrast, non-rigid surgeries, due to factors such as tissue deformation and unstable tracking, still lack unified usability evaluation criteria. Current research is mostly conducted in simulated in vitro environments, comparing the registration accuracy of AR HMDs with traditional navigation technologies and measuring the rotational and translational errors of markerless tracking in clinical applications.[14] Due to the lack of stable marker recognition and predictable tissue deformation patterns, more clinical research data is needed to support these studies.

Beyond surgical scenarios, AR HMDs also show potential in rehabilitation and diagnostic fields. For patients with neurological disorders (such as stroke and Parkinson's disease), AR technology can enhance patients' understanding and interest in rehabilitation exercises through real-time visual feedback and immersive interaction.[92] Traditional motion

Table 1 Summary of Studies Highlighting Clinical Applications in Pedicle Screw Placement Surgery

Author & Year	Study Scale	Equipment	Clinical Evaluation Criteria
Butler, A.J. et al, 2023[3]	606 screws / 164 patients	Xvision Spine system (XVS) (Augmedics Ltd, USA)	Average time for screw placement, intra-operative screw revision rate
Molina, C.A. et al, 2020[73]	113 total implants / 5 cadavers	Xvision Spine system (XVS) (Augmedics Ltd, USA)	Gertzbein-Robbins scale, linear deviation, angular deviation
Liu, A. et al, 2021[71]	205 screws/28 patients	Xvision Spine system (XVS) (Augmedics Ltd, USA)	Gertzbein-Robbins scale
Felix, B. et al, 2022[5]	124 thoracolumbar pedicle screws/ 7 cadavers	VisAR (Novarad Corporation, USA)	Gertzbein-Robbins scale
Chang, C. et al, 2022[74]	24 screws/1 cadaveric	Caduceus (Taiwan)	Gertzbein-Robbins scale
M U Ller, F. et al, 2020[75]	20 K-wires/3 lumbar cadaver spines	HoloLens (Microsoft Corporation, USA)	Translational Error, Angular Error
Yahanda, A.T. et al, 2021[76]	63 screws / 9 patients	Xvision Spine system (XVS) (Augmedics Ltd, USA)	Gertzbein-Robbins scale
Charles, Y.P. et al, 2021[77]	80 pedicle screws / 20 patients	AlluraClarity Flexmove, (Philips. The Netherlands)	Gertzbein-Robbins scale
Farshad, M. et al, 2021[78]	4 pedicle screws /1 patient	HoloLens 2 (Microsoft Corporation, USA)	Angular Deviation, Translational Deviation
Frisk, H. et al, 2022[70]	48 pedicle screws / 4 spine phantoms	Magic Leap1(Magic Leap, USA)	Gertzbein-Robbins scale
Cao, B. et al, 2023[79]	82 pedicle screws / 5 spine cadavers	HoloLens, (Microsoft Corporation, USA)	Gertzbein-Robbins scale
Bhatt, F.R. et al, 2023[80]	222 pedicle screws / 32 patients	xvision Spine system (XVS) (Augmedics Ltd, USA)	Gertzbein-Robbins scale
Ghenbot, Y. et al, 2024[81]	64 pedicle screws / 3 cadavers	HoloLens 2 (Microsoft Corporation, USA)	Gertzbein-Robbins scale
Yanni, D.S. et al, 2021[82]	192 pedicle screws / 24 lumbar spine models	SpineAR (Surgical Theater, Inc).	Gertzbein-Robbins scale
Spirig, J.E.M. et al, 2021[83]	18 k-wires / 2 cadavers	HoloLens (Microsoft Corporation, USA)	3D Distance, Trajectory Angle

https://doi.org/10.2147/MDER.S541187

Table 2 Summary of Studies Highlighting Clinical Applications in Dental Implant Surgery

Author & Year	Study Scale	Equipment	Clinical Evaluation Criteria
Tao, B. et al, 2024[8]	121 dental implants/32 phantoms	HoloLens 2 (Microsoft Corporation, USA)	Global coronal, apical, and angular deviations
Fan, X. et al, 2023[7]	102 dental implants /30 patient phantoms	HoloLens 2 (Microsoft Corporation Corp).	Global coronal, apical, and angular deviations
Pellegrino, G. et al, 2019[84]	2 dental implants /2 patients	HoloLens (Microsoft Corporation, USA)	Entry point, apical point and angular deviations
Yotpibulwong, T. et al, 2022[85]	5 dental implants/5 patients	Moverio BT-300, (Seiko Epson Corporation, Japan)	Entry point, apical point, angular deviations
Kivovics, M.A.R. et al, 2022[86]	16 dental implants/4 models	Magic Leap 1(Magic Leap, USA)	Angular deviation, coronal, and apical global deviation.
Gonz A Lez-Rueda, J.O.N. et al, 2023[87]	20 dental implants/4 models	HoloLens 1 (Microsoft Corporation, USA)	Global coronal, apical, and angular deviations
Liu, L. et al, 2023[88]	25 dental implants/25 models	HoloLens (Microsoft Corporation, USA)	Entry point, middle point, apex point, and angular deviations
Elena et al, 2023[89]	207 dental implants/14 models	HoloLens 2 (Microsoft Corporation, USA)	Global coronal, apical, and angular deviations
Marhuenda Ramos et al, 2024[90]	20 dental implants/1 models	HoloLens 2 (Microsoft Corporation, USA)	Global coronal, apical, and angular deviations
Bochet, Q. et al, 2023[91]	2 dental implants/1 patient's models	HoloLens 2 (Microsoft Corporation, USA)	Entry point, apical point and angular deviations

analysis relies on optoelectronic systems or inertial measurement units (IMUs), while the built-in sensors and tracking algorithms of AR HMDs provide a low-cost, easy-to-deploy alternative for objectively recording patients' motion trajectories and assessing their rehabilitation progress.[93,94] For stroke patients, AR-assisted real-time monitoring and virtual guidance can help clinicians adjust rehabilitation strategies in a timely manner, bringing more convenient and efficient rehabilitation pathways for patients with impaired movement and cognition.[19]

Furthermore, AR HMDs have a broad prospect in medical training. Some studies have simulated in vitro training for spinal pedicle screw insertion, evaluating the usability of AR HMDs in medical training by calculating the distance between the virtual navigation screw and the edge of the pedicle bone.[95] Other studies have simulated central venous catheter insertion surgery, using an infrared reflection marker-based method to track the movement of non-rigid tissues.[96] In life support training, a study has used a questionnaire survey (covering six dimensions: user input, system output, usability, simulation fidelity, immersion, and preference) to evaluate the overall effectiveness of the AR system.[97] These studies indicate that the interactivity and immersion of AR HMDs provide a more intuitive and repeatable training platform for medical students and clinical staff. To accommodate the diversity of assessment methods, the complexity of the entire clinical workflow must be considered.[98] Tailored to the specific clinical context of AR HMDs, qualitative and quantitative analyses may be conducted via heuristic evaluation, cognitive walkthrough, task analysis, GOMS, usability testing, field studies, structured interviews, think-aloud protocols, or mixed-method approaches.

The clinical usability of AR HMDs needs to be comprehensively evaluated in a multidisciplinary and multi-scenario context. Research on the usability evaluation of rigid structure surgeries has accumulated a relatively systematic set of criteria, while soft tissue surgeries, due to technical limitations such as marker stability and registration accuracy, have fewer clinical studies and await the establishment of unified evaluation standards. The fields of rehabilitation and diagnostics show potential for objective assessment and real-time training, offering more rehabilitation options for those with impaired movement and cognitive functions. In the future, if their safety and effectiveness can be verified in large-scale clinical practice and training, and comprehensive evaluation methods and standards are established, AR HMDs are expected to play a more important role in clinical medicine.

Discussion

Hardware Performance Evaluation

Considering the hardware characteristics of AR HMDs, the following suggestions are made for constructing evaluation schemes: determining the components to be tested based on functional priority; employing standardized evaluation platforms with adjustable high-precision parameters to ensure measurement accuracy; and designing experimental parameters to simulate real clinical environments and human visual characteristics.

Specifically, the hardware evaluation of medical AR HMDs should first clarify the functional importance of key components such as displays, cameras, and processors, and determine the evaluation priority accordingly. For the performance evaluation of specific components, a dedicated experimental measurement platform should be constructed to precisely control the spatial position of the measuring instruments relative to the HMD and flexibly adjust the evaluation parameters. In addition, objective measurement tools should be used as much as possible to reduce errors introduced by human factors, such as replacing human observation with cameras and manual operations with robotic arms. At the same time, experimental conditions should simulate real application scenarios as closely as possible, including characteristics of the HVS such as IPD and pupil size, to better match the actual clinical usage environment.

Current research on hardware performance evaluation of AR HMDs mainly focuses on the two core components of displays and cameras, but there are still some methodological shortcomings. First, hardware performance evaluation methods have not yet formed a unified standard. For example, in the process of display performance evaluation, the spatial relationship between the measuring instruments and the HMD, the angle, the division of the measurement area, and the simulation methods of visual parameters (such as IPD and iris size) have not been unified. At present, only a few studies have attempted standardization through 3D printing or optical experimental platforms.[24]

Second, the current evaluation methods often fail to fully reflect the real clinical usage environment. Most experimental settings are overly idealized, ignoring actual scene factors such as physiological micro-movements of patients,

changes in environmental lighting, interference in marker recognition, and movements of personnel in the operating room. This may lead to discrepancies between the actual performance of the device in real applications and the experimental results.[68] Therefore, future research should focus on developing more objective evaluation methods that are closer to real clinical applications.

In addition, considering the diversity of medical applications and the high heterogeneity of AR HMD devices (such as display technologies, camera types and quantities), when conducting hardware evaluation, targeted evaluation scheme design should be carried out according to specific clinical needs. Other key hardware components, such as processors, microphones, speakers, and sunshades, should be selectively tested and evaluated based on their impact on device functionality. Currently, research on hardware evaluation of AR HMDs is still insufficient, and the reason is the diversity of device structures and the differences in manufacturers' design goals. In the future, research in these areas should be further strengthened to promote the systematic and objective construction of the hardware performance evaluation system for AR HMDs.

Future work must prioritize establishing standardized and systematic frameworks for evaluating AR HMD hardware performance. This requires unified protocols defining consistent measurement environments, device configurations, and parameter settings to ensure reproducibility and comparability across studies. Crucially, evaluation scenarios need higher ecological validity, achieved through testbeds that realistically simulate dynamic visuals, complex lighting, and human movement. Incorporating automated and intelligent assessment tools can significantly enhance objectivity and efficiency. Furthermore, the scope of evaluation should broaden beyond core displays to encompass critical supporting components like processors, microphones, and speakers, focusing on aspects such as multimodal interaction, noise handling, and thermal performance. Finally, fostering cross-institutional collaboration and developing open-access performance databases are essential for the continuous improvement and validation of evaluation methodologies.

Software Performance Evaluation

Software, as the core driving force of AR HMDs, plays an important role in the interaction between virtual and real scenes and the effectiveness of clinical applications, with a focus on registration accuracy, the fidelity of medical image rendering, and the smoothness of clinical applications. However, there are still some limitations in the evaluation methods for these important software performance aspects, mainly concentrated in the following areas: the objective evaluation of virtual-to-real registration accuracy of AR HMDs, where traditional evaluation methods rely on subjective human perception, and individual differences make it complex to establish standardized and objective evaluation methods; the consistency of rendering across heterogeneous platforms, where differences in rendering capabilities of different rendering engines may affect the quality of medical images; and the lack of specialized software frameworks tailored for surgical environments and corresponding evaluation standards. To address these challenges, the following points are suggested when constructing software performance evaluation methods for AR HMDs:

First, establish objective registration accuracy evaluation methods. Traditional approaches rely on user visual judgment, which is susceptible to individual differences. Subsequent HMD registration evaluation can combine machine vision to automatically recognize registration feature points, record spatial coordinates through robotic arm movement, or directly calculate through algorithms. This can replace or reduce the subjective intervention of the human eye and improve the quantification accuracy of registration errors.

Second, evaluate the consistency of rendering across heterogeneous platforms. Although current medical AR applications are mostly based on general consumer-grade rendering engines, these platforms have not yet standardized the rendering of medical images in terms of color, grayscale, and depth perception. It is recommended to develop a unified image test toolkit that includes standard color cards, medical images, and dedicated test patterns to quantify rendering differences across platforms in multiple dimensions, thereby addressing the issue of rendering heterogeneity across platforms. Additionally, construct specialized evaluation software frameworks tailored for medical scenarios. Current software evaluation methods lack a standardized evaluation system specifically designed for medical environments, making it difficult to ensure the effectiveness of software in actual medical workflows.

Future work must also establish specialized software evaluation frameworks for medical AR, addressing the current lack of standardized assessment criteria tailored to clinical settings which hinders validation of real-world effectiveness. Progress requires developing high-fidelity automated methods for assessing registration accuracy under clinically relevant conditions

using sensor fusion and computer vision. Concurrently, medical-oriented rendering quality benchmarks, aligned with diagnostic needs and supported by curated datasets and core fidelity metrics, are essential. Furthermore, context-specific validation frameworks—such as those for surgical navigation or rehabilitation—must incorporate real-world variables and undergo multicenter validation. Advancing standardization through universal protocols, open toolkits, and AI-assisted analytics is crucial across these efforts. Together, these developments will enhance objectivity, standardization, and critically, clinical relevance, securing a robust evaluative foundation for the reliability and safety of medical AR software.

Usability Assessment

Usability assessment aims to analyze the efficiency and accuracy of healthcare workers operating AR HMDs in actual diagnosis, treatment, or training processes, involving both subjective experience and objective operational performance dimensions. Unlike hardware and software performance evaluation, usability assessment places greater emphasis on user perception and human factors engineering. Current related research mainly focuses on several dimensions such as the visual information processing system, lighting environment, and clinical usability to evaluate the degree of AR HMD usability. In constructing evaluation methods, the following points should be considered for the common characteristics of usability:

Comprehensive assessment of multi-level perception factors, in addition to the routine subjective questionnaire surveys, objective measurement methods should be added, such as depth perception tests, completion time of operational tasks, and action error rates, to reduce the bias caused by pure subjective evaluation and ensure that the assessment results are more comprehensive and reliable.[99]

Establish comprehensive evaluation criteria closely linked to clinical endpoints. For different types of surgeries (such as spinal pedicle screw surgery, dental implant surgery, and soft tissue resection), differentiated clinical effectiveness indicators should be developed, such as the Gertzbein-Robbins screw placement grading criteria or dental implant angle error indicators, and compared and analyzed with traditional surgical methods to scientifically evaluate the added value and potential risks brought by AR HMDs.

In addition to the usability evaluation aspects discussed above, there are many other factors that need to be considered in the usability evaluation of AR HMDs. This part mainly involves factors such as human-machine systems, human-machine interface design, and mental workload in human factors engineering. For example, the impact of system latency in human-machine systems on usability can be comprehensively judged by combining subjective questionnaire surveys with objective indicators in specific operations, such as the total number of sutures and the number of incorrect sutures in simulated suture surgery tasks, to assess the actual impact of system latency on surgeons.[100] The comfort of using HMDs and the intraoperative load can be quantified in terms of their impact on usability. Studies can use surface electromyography technology to assess the burden on neck muscles or use inverse dynamics models to quantify the load on the cervical spine, to guide the optimization design of the device and comprehensive preclinical testing, and to reduce the fatigue risk and operation error rate of medical staff.[101] Other usability factors that need to be considered include, but are not limited to, the impact of the device's FOV on the senses,[102] ease of use,[99] privacy and security,[103] and human factors design of wearable systems.[104] Through comprehensive and systematic usability evaluation, the effectiveness and safety of AR HMDs in medical applications can be ensured, and the experience and satisfaction of medical staff and patients can be further improved.

Future research should establish a standardized, modular evaluation toolkit integrating core technical metrics, user experience, clinical efficacy, and non-technical factors like device-management workflows;Objective measures—including eye-tracking, task performance, and physiological signals—must enhance comprehensiveness and comparability across clinical contexts.[105] Longitudinal real-world usability studies examining sustained performance, cumulative fatigue, learning curves, and team collaboration under authentic clinical stressors remain critical. Deepening interdisciplinary collaboration among human factors, clinical medicine, computer science, and hospital administration is essential to develop holistic frameworks balancing technical performance, user experience, clinical benefit, safety, and organizational adoption, thereby fully addressing real-world clinical deployment complexity.

Conclusions

Conventional AR HMDs were engineered for industrial contexts and evaluated mainly through subjective impressions and rudimentary metrics. Their rapid penetration into healthcare has exposed the inadequacy of these methods under stringent medical-device regulations and clinical specificity, manifesting in three limitations:

1. heterogeneity barriers: absence of unified hardware/software standards.
2. clinical detachment: test conditions fail to mimic real medical environments.
3. subjective bias: perceptual metrics lack objective quantification.

To address these gaps, we propose five design principles for AR HMD performance testing:

1. identify key components by core function.
2. prioritize testing by functional contribution.
3. replicate authentic clinical and human-visual conditions.
4. objectify subjective perception.
5. test functionally linked components jointly.

The framework may not cover all edge cases and requires large-scale validation and refinement.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (grant no. 12302417), the Shanghai Pujiang Program (grant no. 23PJ1409200) and the Chenguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (grant no. 24CGA51).

Disclosure

All authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Navab N, Martin-Gomez A, Seibold M, et al. Medical augmented reality: definition, principle components, domain modeling, and design-development-validation process. *J Imaging*. 2022;9(1):4. doi:10.3390/jimaging9010004
2. Food and Drug Administration. Augmented reality and virtual reality medical devices. FDA website.Available from: https://www.fda.gov/medical-devices/digital-health-center-excellence/augmented-reality-and-virtual-reality-medical-devices. Accessed 10, Jul. 2025.
3. Butler AJ, Colman MW, Lynch J, Phillips FM. Augmented reality in minimally invasive spine surgery: early efficiency and complications of percutaneous pedicle screw instrumentation. *Spine J*. 2023;23(1):27–33. doi:10.1016/j.spinee.2022.09.008
4. Ghaednia H, Fourman MS, Lans A, et al. Augmented and virtual reality in spine surgery, current applications and future potentials. *Spine J*. 2021;21(10):1617–1625. doi:10.1016/j.spinee.2021.03.018
5. Felix B, Kalatar SB, Moatz B, et al. Augmented reality spine surgery navigation: increasing pedicle screw insertion accuracy for both open and minimally invasive spine surgeries. *Spine*. 2022;47(12):865–872. doi:10.1097/BRS.0000000000004338
6. Wang J, Suenaga H, Hoshi K, et al. Augmented reality navigation with automatic marker-free image registration using 3-D image overlay for dental surgery. *IEEE Trans Biomed Eng*. 2014;61(4):1295–1304. doi:10.1109/TBME.2014.2301191
7. Fan X, Tao B, Tu P, Shen Y, Wu Y, Chen X. A novel mixed reality-guided dental implant placement navigation system based on virtual-actual registration. *Comput Biol Med*. 2023;166:107560. doi:10.1016/j.compbiomed.2023.107560
8. Tao B, Fan X, Wang F, Chen X, Shen Y, Wu Y. Comparison of the accuracy of dental implant placement using dynamic and augmented reality-based dynamic navigation: an in vitro study. *J Dent Sci*. 2024;19(1):196–202. doi:10.1016/j.jds.2023.05.006
9. Hiraoka C, Kutsuna T, Kinoshita T, et al. Smartphone-based navigation system enables low-volume surgeons to perform accurate acetabular cup placement: a single-center retrospective study. *Journal of Joint Surgery and Research*. 2023;1(1):241–245. doi:10.1016/j.jjoisr.2023.11.001
10. Tsukada S, Ogawa H, Kurosaka K, Saito M, Nishino M, Hirasawa N. Augmented reality-aided unicompartmental knee arthroplasty. *J Exp Orthop*. 2022;9(1):88. doi:10.1186/s40634-022-00525-4
11. Hasegawa M, Naito Y, Tone S, Sudo A. Accuracy of augmented reality with computed tomography-based navigation in total Hip arthroplasty. *J Orthop Surg Res*. 2023;18(1):662. doi:10.1186/s13018-023-04155-z
12. Alexander C, Loeb AE, Fotouhi J, Navab N, Armand M, Khanuja HS. Augmented reality for acetabular component placement in direct anterior total Hip arthroplasty. *J Arthroplasty*. 2020;35(6):1636–1641. doi:10.1016/j.arth.2020.01.025
13. Ragnhildstveit A, Li C, Zimmerman MH, et al. Intra-operative applications of augmented reality in glioma surgery: a systematic review. *Front Surg*. 2023;10:1245851. doi:10.3389/fsurg.2023.1245851
14. Van Gestel F, Frantz T, Buyck F, et al. Neuro-oncological augmented reality planning for intracranial tumor resection. *Front Neurol*. 2023;14:1104571. doi:10.3389/fneur.2023.1104571
15. E BC, H BVCUC, Pugonja R, et al. Augmented reality integration in skull base neurosurgery: a systematic review. *Medicina*. 2024;60(2):335. doi:10.3390/medicina60020335

16. Thavarajasingam SG, Vardanyan R, Arjomandi Rad A, et al. The use of augmented reality in transsphenoidal surgery: a systematic review. *Br J Neurosurg*. 2022;36(4):457–471. doi:10.1080/02688697.2022.2057435
17. Campisi BM, Costanzo R, Gulino V, et al. The role of augmented reality neuronavigation in transsphenoidal surgery: a systematic review. *Brain Sci*. 2023;13(12):1695. doi:10.3390/brainsci13121695
18. Khoong YM, Luo S, Huang X, et al. The application of augmented reality in plastic surgery training and education: a narrative review. *J Plast Reconstruct Aesthetic Surg*. 2023;82:255–263. doi:10.1016/j.bjps.2023.04.033
19. Fregna G, Paoluzzi C, Baroni A, Cano-de-la-Cuerda R, Casile A, Straudi S. Head-mounted displays for upper limb stroke rehabilitation: a scoping review. *J Clin Med*. 2023;12(23):7444. doi:10.3390/jcm12237444
20. Shafarenko MS, Catapano J, Hofer SO, Murphy BD. The role of augmented reality in the next phase of surgical education. *Plastic and Reconstructive Surgery–Global Open*. 2022;10(11):e4656.
21. Hey G, Guyot M, Carter A, Lucke-Wold B. Augmented reality in neurosurgery: a new paradigm for training. *Medicina*. 2023;59(10):1721. doi:10.3390/medicina59101721
22. Lungu AJ, Swinkels W, Claesen L, Tu P, Egger J, Chen X. A review on the applications of virtual reality, augmented reality and mixed reality in surgical simulation: an extension to different kinds of surgery. *Expert Rev Med Devices*. 2021;18(1):47–62. doi:10.1080/17434440.2021.1860750
23. Baashar Y, Alkawsi G, Wan Ahmad WN, Alomari MA, Alhussian H, Tiong SK. Towards wearable augmented reality in healthcare: a comparative survey and analysis of head-mounted displays. *Int J Environ Res Public Health*. 2023;20(5):3940. doi:10.3390/ijerph20053940
24. Oun A, Hagerdorn N, Scheideger C, Cheng X. Mobile devices or head-mounted displays: a comparative review and analysis of augmented reality in healthcare. *IEEE Access*. 2024;12:21825–21839. doi:10.1109/ACCESS.2024.3361833
25. Food and Drug Administration. Medical extended reality program: research on medical extended reality-based medical devices. FDA website. Available from: https://www.fda.gov/medical-devices/medical-device-regulatory-science-research-programs-conducted-osel/medical-extended-reality-program-research-medical-extended-reality-based-medical-devices. Accessed 24, Mar. 2021.
26. Beams R, Brown E, Cheng W, et al. Evaluation challenges for the application of extended reality devices in medicine. *J Digit Imaging*. 2022;35(5):1409–1418. doi:10.1007/s10278-022-00622-x
27. Penczek J, Boynton PA, Beams R, Sriram RD. Measurement challenges for medical image display devices. *J Digit Imaging*. 2021;34(2):458–472. doi:10.1007/s10278-021-00438-1
28. Zhao C, Beams R, Dahal E, Badano A. Spatially dependent veiling glare degrades image quality in medical extended reality. *Virtual Real*. 2024;28(1):11. doi:10.1007/s10055-023-00893-x
29. Kim AS, Cheng W, Beams R, Badano A. Color rendering in medical extended-reality applications. *J Digit Imaging*. 2021;34(1):16–26. doi:10.1007/s10278-020-00392-4
30. Birlo M, Edwards PE, Clarkson M, Stoyanov D. Utility of optical see-through head mounted displays in augmented reality-assisted surgery: a systematic review. *Med Image Anal*. 2022;77:102361. doi:10.1016/j.media.2022.102361
31. Syberfeldt A, Danielsson O, Gustavsson P. Augmented reality smart glasses in the smart factory: product evaluation guidelines and review of available products. *IEEE Access*. 2017;5:9118–9130. doi:10.1109/ACCESS.2017.2703952
32. Erickson A, Kim K, Bruder G, Welch GF. Exploring the limitations of environment lighting on optical see-through head-mounted displays. 2020:1–8.
33. Gabbard JL, Smith M, Merenda C, Burnett G, Large DR. A perceptual color-matching method for examining color blending in augmented reality head-up display graphics. *IEEE Trans Vis Comput Graph*. 2020;28(8):2834–2851. doi:10.1109/TVCG.2020.3044715
34. Zhang L, Murdoch MJ, Bachy R. Color appearance shift in augmented reality metameric matching. *J Optical Soc Am A*. 2021;38(5):701–710. doi:10.1364/JOSAA.420395
35. Dahal E, Eby N, Lemaillet P, Beams R, Badano A. Method for characterizing small-spot luminance in medical virtual reality headsets. *Optics Continuum*. 2023;2(5):1180–1190. doi:10.1364/OPTCON.486520
36. Zhang L, Murdoch MJ. Color matching criteria in augmented reality. *J Perceptual Imaging*. 2018;1:1–8. doi:10.2352/J.Percept.Imaging.2018.1.1.010506
37. Zhao C, Kim AS, Beams R, Badano A. Spatiotemporal image quality of virtual reality head mounted displays. *Sci Rep*. 2022;12(1):20235. doi:10.1038/s41598-022-24345-9
38. Zhao C, Beams R, Badano A. Radially variant contrast measurement in virtual reality headsets using circular concentric ring patterns. *J Soc Inf Disp*. 2023;31(5):387–397. doi:10.1002/jsid.1208
39. Beams R, Kim AS, Badano A. Transverse chromatic aberration in virtual reality head-mounted displays. *Opt Express*. 2019;27(18):24877–24884. doi:10.1364/OE.27.024877
40. Ferrari V, Carbone M, Condino S, Cutolo F. Are augmented reality headsets in surgery a dead end? *Expert Rev Med Devices*. 2019;16(12):999–1001. doi:10.1080/17434440.2019.1693891
41. Barcali E, Iadanza E, Manetti L, Francia P, Nardi C, Bocchi L. Augmented reality in surgery: a scoping review. *Appl Sci*. 2022;12(14):6890. doi:10.3390/app12146890
42. Khadem R, Yeh CC, Sadeghi-Tehrani M, et al. Comparative tracking error analysis of five different optical tracking systems. *Comput Aided Surg*. 2000;5(2):98–107. doi:10.3109/10929080009148876
43. Cutolo F, Fida B, Cattari N, Ferrari V. Software framework for customized augmented reality headsets in medicine. *IEEE Access*. 2019;8:706–720. doi:10.1109/ACCESS.2019.2962122
44. Swan JE, Jones A, Kolstad E, Livingston MA, Smallman HS. Egocentric depth judgments in optical, see-through augmented reality. *IEEE Trans Vis Comput Graph*. 2007;13(3):429–442. doi:10.1109/TVCG.2007.1035
45. Ballestin G, Chessa M, Solari F. Assessment of optical see-through head mounted display calibration for interactive augmented reality. 2019.
46. Andrews CM, Henry AB, Soriano IM, Southworth MK, Silva JR. Registration techniques for clinical applications of three-dimensional augmented reality devices. *IEEE J Transl Eng Health Med*. 2020;9:1–14. doi:10.1109/JTEHM.2020.3045642
47. Monica R, Aleotti J. Evaluation of the oculus rift s tracking system in room scale virtual reality. *Virtual Real*. 2022;26(4):1335–1345. doi:10.1007/s10055-022-00637-3

48. Rudolph T, Ebert L, Kowal J. Comparison of three optical tracking systems in a complex navigation scenario. *Comput Aided Surg*. 2010;15(4–6):104–109. doi:10.3109/10929088.2010.530793
49. Blate A, Whitton M, Singh M, et al. Implementation and evaluation of a 50 kHz, 28μs motion-to-pose latency head tracking instrument. *IEEE Trans Vis Comput Graph*. 2019;25(5):1970–1980. doi:10.1109/TVCG.2019.2899233
50. Bner HU, Clintworth P, Liu K, Weinmann Q, Wursthorn M. Evaluation of HoloLens tracking and depth sensing for indoor mapping applications. *Sensors*. 2020;20(4):1021. doi:10.3390/s20041021
51. Hu X, FR YB, Cutolo F. Head-mounted augmented reality platform for markerless orthopaedic navigation. *IEEE J Biomed Health Inform*. 2021;26(2):910–921. doi:10.1109/JBHI.2021.3088442
52. Khang S, Park T, Lee J, Kim KW, Song H, Lee J. Computer-aided breast surgery framework using a markerless augmented reality method. *Diagnostics*. 2022;12(12):3123. doi:10.3390/diagnostics12123123
53. Qian L, Barthel A, Johnson A, et al. Comparison of optical see-through head-mounted displays for surgical interventions with object-anchored 2D-display. *Int J Comput Assist Radiol Surg*. 2017;12:901–910. doi:10.1007/s11548-017-1564-y
54. Eckhoff D, Schnupp J, Cassinelli A. Temporal precision and accuracy of audio-visual stimuli in mixed reality systems. *PLoS One*. 2024;19(1):e295817. doi:10.1371/journal.pone.0295817
55. Frantz T, Jansen B, Duerinck J, Vandemeulebroucke J. Augmenting Microsoft's HoloLens with vuforia tracking for neuronavigation. *Healthc Technol Lett*. 2018;5(5):221–225. doi:10.1049/htl.2018.5079
56. Condino S, Turini G, Parchi PD, et al. How to build a patient-specific hybrid simulator for orthopaedic open surgery: benefits and limits of mixed-reality using the microsoft holoLens. *J Healthc Eng*. 2018;2018(1):5435097. doi:10.1155/2018/5435097
57. Holloway RL. Registration error analysis for augmented reality. *Presence*. 1997;6(4):413–432. doi:10.1162/pres.1997.6.4.413
58. Debarba HG, De Oliveira ME, Dermann LA, Chaqu A, S E, Charbonnier C. Tracking a consumer HMD with a third party motion capture system. *IEEE*. 2018:539–540.
59. Diaz C, Walker M, Szafir DA, Szafir D. Designing for depth perceptions in augmented reality. *IEEE*. 2017:111–122.
60. Collins B, Beams R, Lago M. Open-source pattern creation tool for medical extended reality image quality assessment. *J Open Source Software*. 2024;9(93):6021. doi:10.21105/joss.06021
61. Ablyaev MR, Abliakimova AN, Seidametova ZS. Criteria of evaluating augmented reality applications. *Advanced Engineering Res*. 2020;20(4):414–421.
62. Behringer R, Christian J, Holzinger A, Wilkinson S. Some usability issues of augmented and mixed reality for e-health applications in the medical domain. *Springer*. 2007;255–266.
63. Schneider D, Biener V, Otte A, et al. Accuracy evaluation of touch tasks in commodity virtual and augmented reality head-mounted displays. 2021:1–11.
64. Swan JE, Singh G, Ellis SR. Matching and reaching depth judgments with real and augmented reality targets. *IEEE Trans Vis Comput Graph*. 2015;21(11):1289–1298. doi:10.1109/TVCG.2015.2459895
65. Omar K, Fakhouri H, Zraqou J, Marx GOM. Usability heuristics for metaverse. *Computers*. 2024;13(9):222. doi:10.3390/computers13090222
66. Edwards PJ, King AP, Maurer CR, et al. Design and evaluation of a system for microscope-assisted guided interventions (MAGI). *IEEE Trans Med Imaging*. 2000;19(11):1082–1093. doi:10.1109/42.896784
67. Kahl D, Ruble M, Kr U, Ger A. The influence of environmental lighting on size variations in optical see-through tangible augmented reality. *IEEE*. 2022:121–129.
68. Suter D, Hodel S, Liebmann F, Rnstahl FU, Farshad P. Factors affecting augmented reality head-mounted device performance in real OR. *Eur Spine J*. 2023;32(10):3425–3433. doi:10.1007/s00586-023-07826-x
69. Kubben PL, Sinlae RS. Feasibility of using a low-cost head-mounted augmented reality device in the operating room. *Surg Neurol Int*. 2019;10:26. doi:10.4103/sni.sni_228_18
70. Frisk H, Lindqvist E, Persson O, et al. Feasibility and accuracy of thoracolumbar pedicle screw placement using an augmented reality head mounted device. *Sensors*. 2022;22(2):522. doi:10.3390/s22020522
71. Liu A, Jin Y, Cottrill E, et al. Clinical accuracy and initial experience with augmented reality--assisted pedicle screw placement: the first 205 screws. *J Neurosurg Spine*. 2021;36(3):351–357. doi:10.3171/2021.2.SPINE202097
72. Chang RM, Kauffman RJ, Kwon Y. Understanding the paradigm shift to computational social science in the presence of big data. *Decis Support Syst*. 2014;63:67–80. doi:10.1016/j.dss.2013.08.008
73. Molina CA, Phillips FM, Colman MW, et al. A cadaveric precision and accuracy analysis of augmented reality--mediated percutaneous pedicle implant insertion: presented at the 2020 AANS/CNS joint section on disorders of the spine and peripheral nerves. *J Neurosurg Spine*. 2020;34(2):316–324. doi:10.3171/2020.6.SPINE20370
74. Chang C, Kuo C, Chang H, et al. Augmented reality-assisted percutaneous pedicle screw instrumentation: a cadaveric feasibility and accuracy study. *Appl Sci*. 2022;12(10):5261. doi:10.3390/app12105261
75. Müller F, Roner S, Liebmann F, Spirig JM, Fürnstahl P, Farshad M. Augmented reality navigation for spinal pedicle screw instrumentation using intraoperative 3D imaging. *Spine J*. 2020;20(4):621–628. doi:10.1016/j.spinee.2019.10.012
76. Yahanda AT, Moore E, Ray WZ, Pennicooke B, Jennings JW, Molina CA. First in-human report of the clinical accuracy of thoracolumbar percutaneous pedicle screw placement using augmented reality guidance. *Neurosurg Focus*. 2021;51(2):E10. doi:10.3171/2021.5.FOCUS21217
77. Charles YP, Cazzato RL, Nachabe R, Chatterjea A, Steib J, Gangi A. Minimally invasive transforaminal lumbar interbody fusion using augmented reality surgical navigation for percutaneous pedicle screw placement. *Clin Spine Surg*. 2021;34(7):E415–E424. doi:10.1097/BSD.0000000000001132
78. Farshad M, Rnstahl FU, P SJEM. First in man in-situ augmented reality pedicle screw navigation. *North Am Spine Soc J*. 2021;6:100065. doi:10.1016/j.xnsj.2021.100065
79. Cao B, Yuan B, Xu G, et al. A pilot human cadaveric study on accuracy of the augmented reality surgical navigation system for thoracolumbar pedicle screw insertion using a new intraoperative rapid registration method. *J Digit Imaging*. 2023;36(4):1919–1929. doi:10.1007/s10278-023-00840-x
80. Bhatt FR, Orosz LD, Tewari A, et al. Augmented reality-assisted spine surgery: an early experience demonstrating safety and accuracy with 218 screws. *Global Spine J*. 2023;13(7):2047–2052. doi:10.1177/21925682211069321

81. Ghenbot Y, Ahmad HS, Chauhan D, et al. Effects of augmented reality on thoracolumbar pedicle screw instrumentation across different levels of surgical experience. *World Neurosurg*. 2024;182:e284–e291. doi:10.1016/j.wneu.2023.11.100
82. Yanni DS, Ozgur BM, Louis RG, et al. Real-time navigation guidance with intraoperative CT imaging for pedicle screw placement using an augmented reality head-mounted display: a proof-of-concept study. *Neurosurg Focus*. 2021;51(2):E11. doi:10.3171/2021.5.FOCUS21209
83. Spirig JEM, Roner S, Liebmann F, Rnstahl FU, Farshad P. Augmented reality-navigated pedicle screw placement: a cadaveric pilot study. *Eur Spine J*. 2021;30:3731–3737. doi:10.1007/s00586-021-06950-w
84. Pellegrino G, Mangano C, Mangano R, Ferri A, Taraschi V, Marchetti C. Augmented reality for dental implantology: a pilot clinical report of two cases. *BMC Oral Health*. 2019;19:1–8. doi:10.1186/s12903-019-0853-y
85. Yotpibulwong T, Arunjaroensuk S, Pimkhaokham A. Application of augmented reality with navigation system and laboratory guide affected on the deviation of dental implant-a pilot clinical study. 2022:193–201.
86. Kivovics M, Takács A, Pénzes D, Németh O, Mijiritsky E. Accuracy of dental implant placement using augmented reality-based navigation, static computer assisted implant surgery, and the free-hand method: an in vitro study. *J Dent*. 2022;119:104070. doi:10.1016/j.jdent.2022.104070
87. Gonz A Lez-Rueda JON, Galparsoro-Catal A N AIN, de Paz-Hermoso VIC, et al. Accuracy of zygomatic dental implant placement using computer-aided static and dynamic navigation systems compared with a mixed reality appliance. An in vitro study. *J Clin Exp Dentistry*. 2023;15(12):e1035. doi:10.4317/jced.61097
88. Liu L, Wang X, Guan M, et al. A mixed reality-based navigation method for dental implant navigation method: a pilot study. *Comput Biol Med*. 2023;154:106568. doi:10.1016/j.compbiomed.2023.106568
89. Riad Deglow E, Zubizarreta-Macho Á, h GM, et al. Comparative analysis of two navigation techniques based on augmented reality technology for the orthodontic mini-implants placement. *BMC Oral Health*. 2023;23(1):542. doi:10.1186/s12903-023-03261-y
90. MIAT MR, Faus-Matoses I, Zubizarreta-Macho AL, et al. Influence of augmented reality technique on the accuracy of autotransplanted teeth in surgically created sockets. *BMC Oral Health*. 2024;24(1):415. doi:10.1186/s12903-024-04173-1
91. Bochet Q, GENE R, Lauwers L, Nicot R. Augmented reality in implantology: virtual surgical checklist and augmented implant placement. *J Stomatol Oral Maxillofacial Surg*. 2024;125(5):101813. doi:10.1016/j.jormas.2024.101813
92. De Pasquale P, Bonanno M, Mojdehdehbaher S, Quartarone A, Calabr O RS. The use of head-mounted display systems for upper limb kinematic analysis in post-stroke patients: a perspective review on benefits, challenges and other solutions. *Bioengineering*. 2024;11(6):538. doi:10.3390/bioengineering11060538
93. Casile A, Fregna G, Boarini V, et al. Quantitative comparison of hand kinematics measured with a markerless commercial head-mounted display and a marker-based motion capture system in stroke survivors. *Sensors*. 2023;23(18):7906. doi:10.3390/s23187906
94. Hussain N, Alt Murphy M, Sunnerhagen KS. Upper limb kinematics in stroke and healthy controls using target-to-target task in virtual reality. *Front Neurol*. 2018;9:300. doi:10.3389/fneur.2018.00300
95. Gibby JT, Swenson SA, Cvetko S, Rao R, Javan R. Head-mounted display augmented reality to guide pedicle screw placement utilizing computed tomography. *Int J Comput Assist Radiol Surg*. 2019;14:525–535. doi:10.1007/s11548-018-1814-7
96. Rochlen LR, Levine R, Tait AR. First-person point-of-view–augmented reality for central line insertion training: a usability and feasibility study. *Simulation in Healthcare*. 2017;12(1):57–62. doi:10.1097/SIH.0000000000000185
97. Ingrassia PL, Mormando G, Giudici E, et al. Augmented reality learning environment for basic life support and defibrillation training: usability study. *J Med Internet Res*. 2020;22(5):e14910. doi:10.2196/14910
98. Mohammadzadeh N, Katigari MR, Hosseini R, Pahlevanynejad S. Evaluation methods in clinical health technologies: a systematic review. *Iran J Public Health*. 2023;52(5):913. doi:10.18502/ijph.v52i5.12708
99. Moosburner S, Remde C, Tang P, et al. Real world usability analysis of two augmented reality headsets in visceral surgery. *Artif Organs*. 2019;43(7):694–698. doi:10.1111/aor.13396
100. Khan T, Zhu TS, Downes T, et al. Understanding effects of visual feedback delay in ar on fine motor surgical tasks. *IEEE Trans Vis Comput Graph*. 2023;29(11):4697–4707. doi:10.1109/TVCG.2023.3320214
101. Astrologo AN, Nano S, Klemm EM, Shefelbine SJ, Dennerlein JT. Determining the effects of AR/VR HMD design parameters (mass and inertia) on cervical spine joint torques. *Appl Ergon*. 2024;116:104183. doi:10.1016/j.apergo.2023.104183
102. Kim M, Choi SH, Park K, Lee JY. User interactions for augmented reality smart glasses: a comparative evaluation of visual contexts and interaction gestures. *Appl Sci*. 2019;9(15):3171. doi:10.3390/app9153171
103. Bhalla A, Sluganovic I, Krawiecka K, Martinovic I, Naranje P, Gupta AK. MoveAR: continuous biometric authentication for augmented reality headsets. *Curr Prob Diagnostic Radiol*. 2021;50:41–52. doi:10.1067/j.cpradiol.2019.07.009
104. Maisto M, Pacchierotti C, Chinello F, Salvietti G, De Luca A, Prattichizzo D. Evaluation of wearable haptic systems for the fingers in augmented reality applications. *IEEE Trans Haptics*. 2017;10(4):511–522. doi:10.1109/TOH.2017.2691328
105. IJUN BDN, Abdulazeem H, Vasanthan LT, et al. Barriers and facilitators to utilizing digital health technologies by healthcare professionals. *NPJ Digit Med*. 2023;6(1):161. doi:10.1038/s41746-023-00899-4