Artificial Intelligence In Medicine 170 (2025) 103270

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

journal homepage: www.elsevier.com/locate/artmed

Evaluating AI-driven characters in extended reality (XR) healthcare simulations: A systematic review

David Dasa [a],*, Michele Board [b], Ursula Rolfe [c], Tom Dolby [d], Wen Tang [a],*

[a] Department of Creative Technology, Bournemouth University, UK
[b] Department of Nursing Science, Bournemouth University, UK
[c] Department of Midwifery and Health Sciences, Bournemouth University, UK
[d] i3 Simulations, UK

ARTICLE INFO

Keywords:
Extended reality
Virtual reality
Artificial intelligence
Healthcare simulation
Medical education
Non-player characters

ABSTRACT

AI-driven characters in extended reality (XR) healthcare simulations are increasingly used for clinical training, yet their effectiveness, implementation, and quality assurance remain poorly understood.

We conducted a systematic review of 132 studies published between January 2015 and July 2025, including 11 randomized controlled trials (RCTs), sourced from biomedical, computing, and education databases and targeted proceedings. Most studies used virtual reality (62.1%) and focused on effectiveness (n = 71), with fewer examining implementation (n = 45) or quality assurance (n = 44). Meta-analysis of two RCTs found a large effect on knowledge and decision-making (Hedges' g = 1.31, 95% CI 0.08–2.54, I^2 = 85%), while one RCT reported faster task performance with AI-driven characters (g = -0.68, 95% CI -1.32 to -0.04). Certainty of evidence was low due to small samples and high heterogeneity. Implementation success was often associated with phased roll-outs and faculty training, but quality assurance practices (particularly bias audits and transparency measures) were rarely documented.

The review proposes the DASEX framework to address these gaps and guide future integration of AI-driven characters in XR training.

Contents

* Corresponding authors.
E-mail addresses: ddasa@bournemouth.ac.uk (D. Dasa), mboard@bournemouth.ac.uk (M. Board), urolfe@bournemouth.ac.uk (U. Rolfe), tom@i3simulations.com (T. Dolby), wtang@bournemouth.ac.uk (W. Tang).

https://doi.org/10.1016/j.artmed.2025.103270
Received 17 June 2025; Received in revised form 8 September 2025; Accepted 11 September 2025
Available online 24 September 2025

1. Introduction

1.1. Background and context

Extended reality (XR) encompassing virtual (VR), augmented (AR), and mixed reality (MR) is now commonplace in healthcare simulation, where it offers safe, repeatable, and scalable practice for technical and non-technical skills. Despite accelerating adoption, the evidence base for effectiveness, implementation feasibility, and quality assurance remains fragmented, and prior reviews rarely isolate the specific contribution of AI-driven characters in XR healthcare training.

1.2. The need for a new review

Three critical factors necessitate this review. First, since 2023, advances in large language models and XR hardware may confound comparisons with earlier studies. Second, outcome reporting is heterogeneous (different scales, designs, and endpoints), which has limited previous meta-analytic synthesis. Third, evaluation frameworks widely used in simulation (e.g., MLASE, CFIR, Kirkpatrick) only partially address AI-specific concerns (adaptivity, transparency, algorithmic safety), highlighting a gap between technical capabilities and existing educational governance frameworks. Finally, questions of equity and privacy are under-reported yet highly relevant for deployment.

1.3. Addressing prior limitations

We conducted a PRISMA-aligned search spanning 1 January 2015 to 31 July 2025 across biomedical, education, and computing sources, explicitly including the ACM Digital Library and targeted conference proceedings such as IEEE VR and key ACM/IEEE HCI venues to capture cutting-edge AI+XR research often published outside traditional journals. To manage the inherent methodological heterogeneity across studies without over-interpretation, we grouped outcomes into meaningful families such as task time, error rate, knowledge scores, OSCE scores, diagnostic accuracy, motion metrics, and confidence ratings, to enable meaningful synthesis. Where at least two comparable studies were available, we performed random-effects meta-analysis using the REML estimator, reporting τ^2, I^2, and 95% prediction intervals to better convey the uncertainty and generalizability of pooled effects. Sensitivity analyses were implemented to assess robustness in cases of small sample sizes or high risk of bias. Small-study effects were evaluated using funnel plots and Egger's test where the number of studies permitted ($k \geq 10$). Study quality was rigorously assessed using the RoB 2 tool for randomized trials and ROBINS-I for non-randomized designs, with overall certainty in evidence summarized using GRADE methodology and presented in structured Summary-of-Findings tables [1–3]. We also reference the SPIRIT-AI and CONSORT-AI reporting extensions where applicable [4,5].

1.4. Contributions

This review makes four key contributions: (1) First cross-disciplinary synthesis isolating AI-driven characters in XR healthcare training (2015–2025), (2) PRISMA-aligned methodology with transparent reporting, (3) Novel meta-analytic approach addressing heterogeneity through outcome families, and (4) Introduction of DASEX–an AI-specific evaluation framework for NPC-driven XR with prospective validation protocol.

1.5. Research questions

This review focuses on three core areas: the effectiveness of AI-driven characters in XR healthcare simulations, approaches for their implementation and scalability, and mechanisms ensuring their reliability, safety, and fairness. Accordingly:

Research questions.

- **RQ1** *In healthcare training, what is the effect of AI-driven characters in XR simulations compared to (a) traditional training methods, (b) non-AI XR simulations, or (c) no intervention, on learner performance*

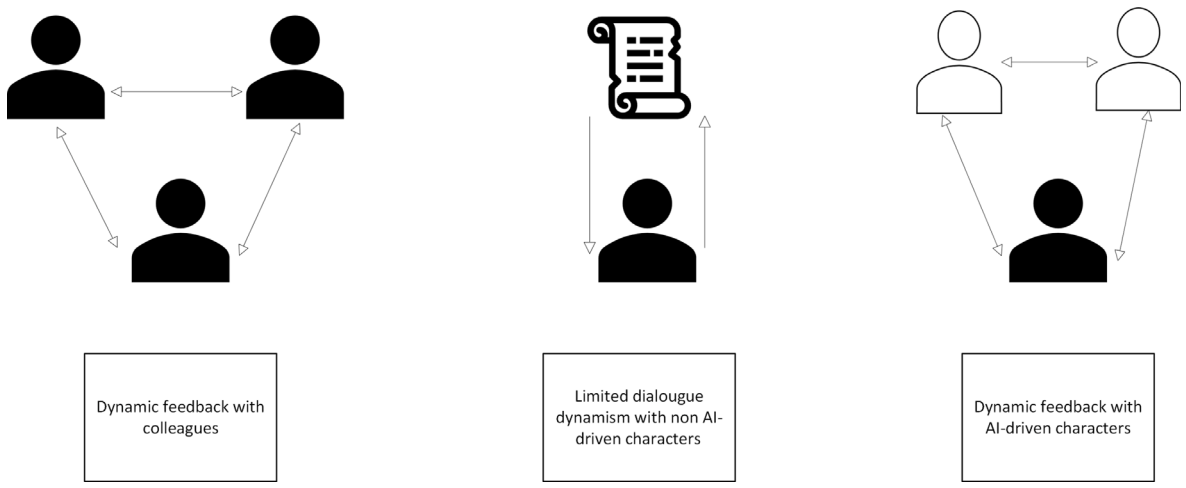

Fig. 1. Single and multi-participant training dynamics enabled by AI-driven characters in XR simulations.

outcomes including procedural task time, error rate, knowledge scores, OSCE scores, diagnostic accuracy, motion metrics, and self-reported confidence?

- **RQ2** *What implementation approaches, instructional designs, and technological configurations are used to integrate AI-driven characters into XR healthcare simulations, and what contextual factors (e.g., setting, resources, faculty training, device type, connectivity) influence successful adoption and scalability?*
- **RQ3** *What mechanisms are reported to ensure the reliability, safety, fairness, and explainability of AI-driven characters in XR healthcare simulations, and how do these align with or extend existing evaluation frameworks (e.g., MLASE, CFIR, RE-AIM)?*

2. Related work

This section reviews prior literature relevant to AI-driven characters in XR healthcare simulations, situating our study within the broader landscape of simulation-based education, artificial intelligence integration, and evaluation frameworks. We begin by tracing the evolution of simulation in healthcare, from early OSCE models to the adoption of XR technologies.

2.1. Background

Simulation in healthcare: From OSCE to XR

Simulation-based training was introduced into healthcare to develop clinical competencies in a manner that reflects actual operational conditions, with studies identifying its role in improving clinical outcomes and patient safety [6–10]. The concept of clinical training by simulating real-life conditions is not new; studies dating back to the year 2000 highlight the benefits of clinical simulation for skill acquisition and knowledge retention. Extended Reality (XR) was identified as a promising medium for delivering immersive clinical environments [11, 12].

Research into simulation validity followed, aiming to ensure alignment with clinical needs [13,14]. Traditional assessments like long-case exams, where a candidate reviewed a case independently and then presented their findings to an examiner, were eventually largely supplemented by standardized assessments [15]. Harden's Objective Structured Clinical Examination (OSCE), introduced in 1975 [16], was a turning point, enabling fairer evaluation via structured scenarios and checklists [17,18]. OSCEs have since become a gold standard [19], demonstrating the importance of standardization in clinical training and assessment tools.

The emergence of AI-driven characters

Extended Reality (XR), encompassing Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR), has transformed digital simulation-based education. VR offers fully immersive environments suited to procedural practice in areas such as sonography [20,21], psychiatry [22], and emergency care [23], with immersion linked to improved educational outcomes [24–26]. AR overlays digital information on real-world views to enable interactive, context-aware learning [27], while MR integrates both immersive and real-world elements to support team-based scenarios and flexible clinical training [28,29].

Recent advances have introduced AI-driven characters into these environments, capable of adapting dialogue, emotional tone, and clinical decision-making in real time. This advancement moves XR training beyond scripted, deterministic interactions, allowing simulations to respond dynamically to learner actions, replicate realistic patient behavior, and simulate collaborative team members or assessors, as illustrated in Fig. 1.

Artificial intelligence and conversational capabilities

Artificial Intelligence (AI) refers to the ability of machines to perform tasks, communicate, and reason in ways comparable to humans [30], reflecting Alan Turing's seminal question on machine intelligence [31]. The emergence of large language models (LLMs) based on transformer architectures [32] has enabled AI systems to generate nuanced, context-sensitive dialogue far beyond the capabilities of earlier rule-based or statistical approaches such as HMMs and n-gram models [33–35].

The 2022 release of ChatGPT accelerated both public and academic engagement with conversational AI [36]. Since then, clinical applications have proliferated [37–40], including safety-enhanced versions tailored for healthcare [41,42]. The conversational realism of these models enables AI-driven characters in XR simulations to convincingly mimic human interaction, with studies reporting that nearly half of participants could not distinguish AI from humans in blind tests [43]. Throughout, we use the term AI-driven characters to refer to AI-enabled agents used in XR as patients, teammates, instructors, or assessors; we use virtual patient only when the character specifically plays a patient role.

2.2. Prior reviews and gaps

What prior reviews have covered—and missed

Prior systematic reviews have explored the use of extended reality (XR) in healthcare education [44–47], yet few have isolated the

specific role of AI-driven characters from scripted avatars or human-controlled simulations [48–50]. Many existing syntheses focus narrowly on single clinical disciplines or particular skill domains, limiting their generalizability across specialties. Additionally, several reviews exclude conference proceedings, despite these venues being common outlets for cutting-edge AI+XR research in computing and human–computer interaction [28,48]. The heterogeneity in outcome reporting across studies has often precluded meta-analysis, resulting in purely qualitative summaries [46,50]. A critical gap is that many reviews fail to address AI-specific considerations, including bias auditing, or the fidelity of dynamic adaptivity, features that fundamentally distinguish AI-driven characters from static simulation tools [4,5,40,51–54]. These omissions highlight a gap in the current literature: a need for a review that not only identifies the presence of AI-driven characters but evaluates their unique contributions and risks within immersive training environments.

Rationale for this review

Given the rapid evolution of XR hardware, generative AI models, and adaptive pedagogical approaches, a new, comprehensive synthesis is warranted. This review addresses the limitations of prior work by drawing on cross-disciplinary evidence from healthcare, computer science, and educational research [47,55]. It organizes findings by outcome family to enable meta-analysis where methodological consistency allows, enhancing the interpretability of effect sizes and aligning with PRISMA 2020 reporting guidance [56]. Beyond effectiveness, the review evaluates implementation strategies and contextual factors such as institutional readiness, technical infrastructure, and faculty adoption that influence real-world deployment [57,58]. Crucially, it incorporates both established evaluation frameworks and novel, AI-specific criteria to assess quality, safety, and equity, ensuring that the unique challenges posed by autonomous or semi-autonomous agents are not overlooked in high-stakes clinical training contexts [4,5,37,39,40,51].

2.3. Existing evaluation frameworks

Existing evaluation frameworks including MLASE (Medical Learning, Analytics, and Simulation Evaluation), CFIR (Consolidated Framework for Implementation Research), and Kirkpatrick's model have been instrumental in guiding the design and assessment of simulation-based education [45,57–59]. However, these models were developed prior to the widespread integration of AI into immersive environments and therefore lack constructs to evaluate dynamic, non-deterministic AI-driven characters behaviors [48,60]. They do not account for real-time adaptivity to learner performance, nor do they provide guidance on assessing algorithmic transparency or the explainability of AI-generated responses; similarly, there is limited emphasis on formal protocols for detecting and mitigating bias in AI-driven interactions, or on ensuring the reliability of outputs when underlying models are probabilistic or context-sensitive [52,53]. Furthermore, these frameworks offer little support for evaluating the quality of immersive AI-driven characters interactions such as conversational realism, emotional responsiveness, or clinical plausibility of NPC decisions, features that are central to learner engagement and educational fidelity in AI-enhanced simulations [61,62].

2.4. Development and positioning of the DASEX framework

To address these shortcomings, this review introduces the DASEX, an AI-specific extension designed to evaluate XR simulations featuring intelligent, interactive characters. DASEX complements existing models by introducing specialized metrics for conversational fidelity, response quality, and the coherence of AI-generated clinical reasoning. It includes structured protocols to assess the degree and effectiveness of adaptivity in response to learner actions, as well as safety guardrails that ensure clinical accuracy and prevent harmful or misleading recommendations. The framework also incorporates dedicated modules for bias detection, enabling evaluators to scrutinize training data provenance, demographic representation, and fairness in feedback delivery. By integrating these dimensions, DASEX provides a more nuanced and technically grounded evaluation lens for AI-driven simulations [4,5,37,39,40,51]. Its validation methodology and implementation protocol are detailed in Section 9.3.1, illustrating how the framework operationalizes critical AI-specific concerns and supports the responsible development and deployment of AI-driven characters in healthcare education.

3. Methods

This section details the PRISMA-aligned methodology, following PRISMA 2020 guidelines [56]. It describes the search strategy, eligibility criteria, screening process, and data extraction methods, as well as the approaches used for quality assessment and synthesis. The methodological framework was designed to maximize transparency, reproducibility, and alignment with best practices for evidence synthesis in emerging technology domains.

3.1. Search strategy

We conducted a PRISMA-compliant search combining database queries, targeted conference proceedings, and citation snowballing. Five databases were searched: PubMed (biomedical literature), IEEE Xplore (engineering/computer science), Scopus (interdisciplinary sciences), Web of Science (multidisciplinary scholarship), and ACM Digital Library (computing research). Targeted conference proceedings included IEEE VR, ACM VRST, IEEE AIxVR, IEEE VIS, and ACM CHI (2015–2025).

The search timeframe spanned 1 January 2015 to 31 July 2025 with English-language restrictions. Boolean search strings (provided in Appendix A) combined three conceptual domains:

1. XR technologies: ("virtual reality" OR "augmented reality" OR "mixed reality" OR "extended reality" OR "immersive simulation")
2. AI-driven characters: ("artificial intelligence" OR "intelligent agent" OR NPC OR "virtual patient" OR "conversational agent" OR "large language model" OR "generative AI")
3. Healthcare context: ("clinical training" OR "medical education" OR "healthcare simulation" OR "team training")

To validate coverage, we performed citation snowballing on 14 keystone studies using Connected Papers (backward citation search) and Google Scholar (forward citation search). This yielded a 64% duplicate recovery rate (9/14 keystones already captured) and identified 5 additional studies, indicating excellent coverage of the citation backbone.

3.2. Eligibility criteria

Inclusion criteria

- **Population:** Healthcare professionals or students at any training level
- **Intervention:** XR simulations (VR, AR, MR) incorporating *AI-driven interactive agents*
- **Context:** Healthcare education or clinical training
- **Study Design:** Empirical investigations reporting quantitative or qualitative outcomes
- **Publication Type:** Peer-reviewed journal articles or full conference proceedings (2015–2025)

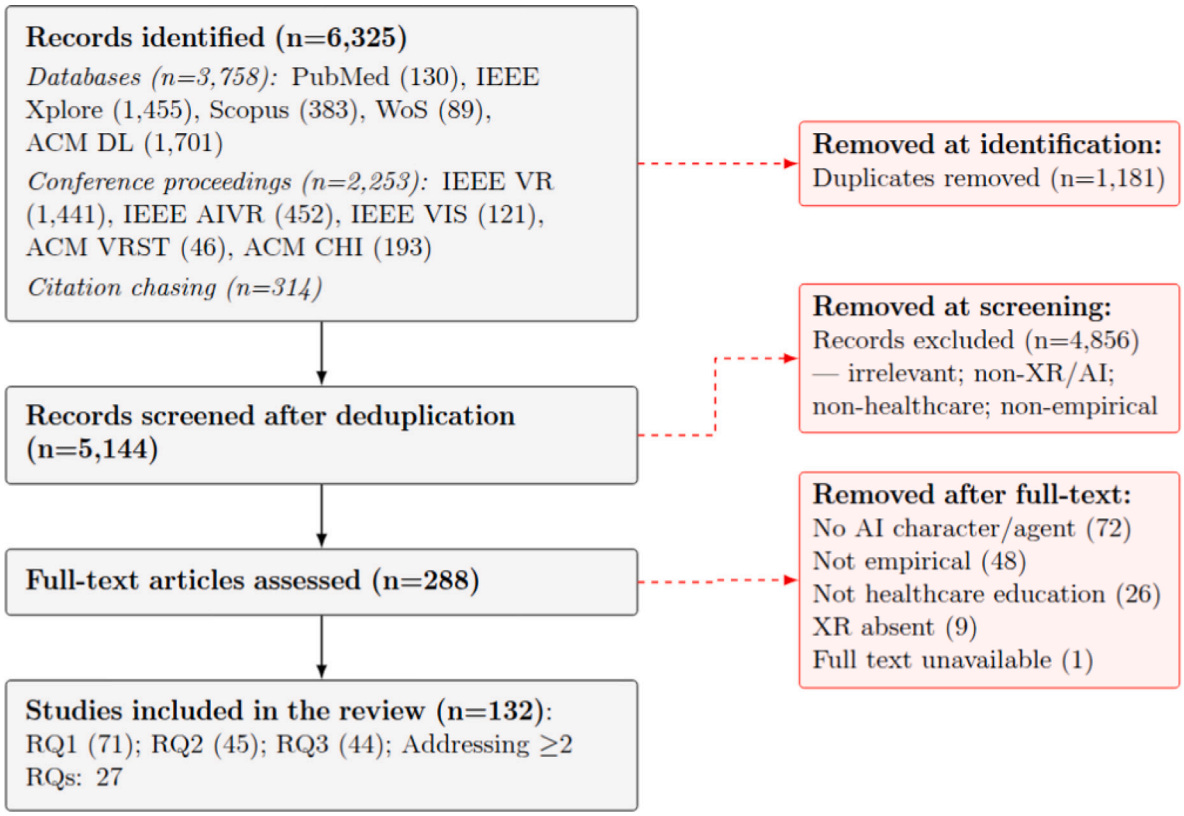

Fig. 2. PRISMA 2020 flow diagram of study selection.

Exclusion criteria

- Simulations without AI-driven characters
- XR applications outside healthcare education
- Non-empirical reports (e.g., conceptual papers, untested frameworks)
- Non-English publications
- Inaccessible full texts

3.3. Data extraction and quality assessment

The study selection process followed the PRISMA 2020 guidelines. A total of 6325 records were initially identified, comprising 3758 from database searches, 2253 from conference proceedings, and 314 from citation snowballing.

Study selection process

After removing 1181 duplicates, 5144 unique records remained for screening. Titles and abstracts were screened independently by two reviewers, resulting in the exclusion of 4856 records based on relevance and scope. The remaining 288 full-text articles were assessed for eligibility, of which 156 were excluded for the following reasons: absence of an AI-driven character (72), non-empirical design (48), non-healthcare context (26), lack of XR modality (9), and unavailability of full text (1). This process yielded 132 studies included in the final synthesis, with discrepancies resolved through consensus or adjudication by a third reviewer. The complete workflow is summarized in Fig. 2.

Data extraction

Data were extracted using a standardized framework designed to capture key dimensions of each study. This included study characteristics such as publication year, country of origin, research design, sample size, and participant demographics. Technological specifications were recorded in detail, encompassing the XR modality used (VR, AR, or MR), hardware platforms, AI architecture (e.g., rule-based, NLP, machine learning), and the functional roles of AI-driven characters within the simulation. Educational context was documented with attention to clinical specialty, learning objectives, and the method of integration into training curricula. Outcome data were systematically collected across performance metrics, including task time, error rates, and OSCE scores, as well as measures of knowledge gain and learner confidence. Finally, information on quality assurance practices was extracted, focusing on validation methods, reliability metrics, and protocols for human oversight to ensure educational and clinical fidelity.

Study characteristics synthesis

Across the included corpus ($n = 132$), randomized trials constituted a small minority (11; ≈8%), with the literature dominated by experimental and feasibility-style studies (60), alongside mixed-methods (7), pilot studies (5), surveys/cross-sectional designs (3), case/case-series (3), and reviews (18). Trials spanned team training and decision-making in VR with AI-driven characters [63–65], dental surgical decision-making with an intelligent tutor [66], and AI-guided bronchoscopy skill acquisition [67]. Experimental and feasibility work covered AI+AR guidance for robotic surgery and real-time VR/MR simulations with empathic AI-driven characters [68,69], while qualitative and project evaluations explored interaction modalities (menu vs. voice), student experience, and early service deployment [70,71]. Reviews synthesized biofeedback+AI in XR training, immersive reality in anesthesia, and metaverse opportunities and risks for medical education [46,49,72–74].

Sample sizes were modest. Among RCTs with a numeric N reported ($n = 7$), the median was 70 (IQR 54–120), with most trials conducted at single institutions and focused on short-term outcomes [63,64,67]. Across all studies with a numeric N ($n = 27$), the median was 41 (IQR 16–67). This scale, together with single-site implementation, should temper inferences about generalizability.

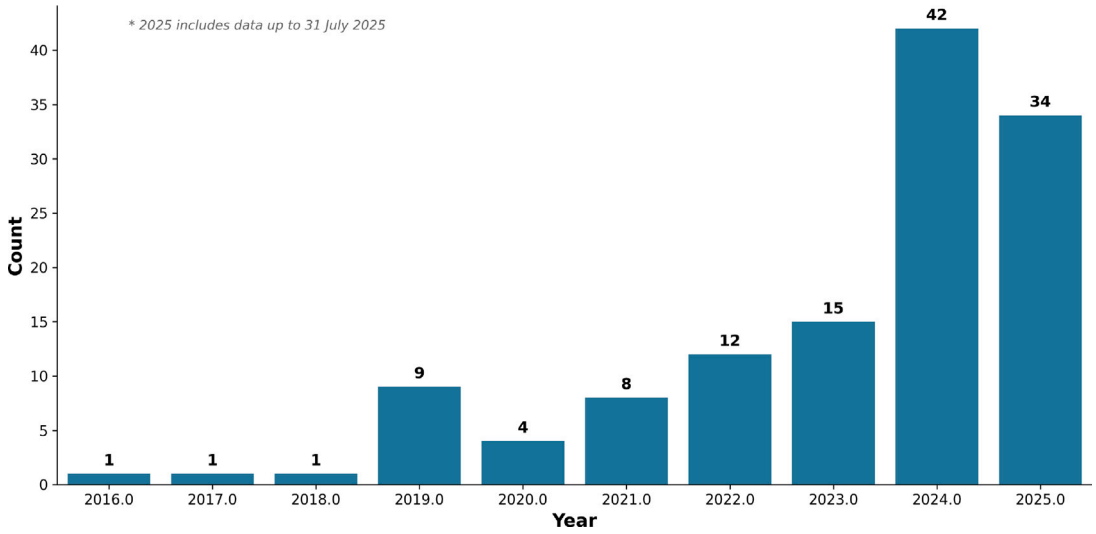

Fig. 3. Included studies per year (2015–2025).

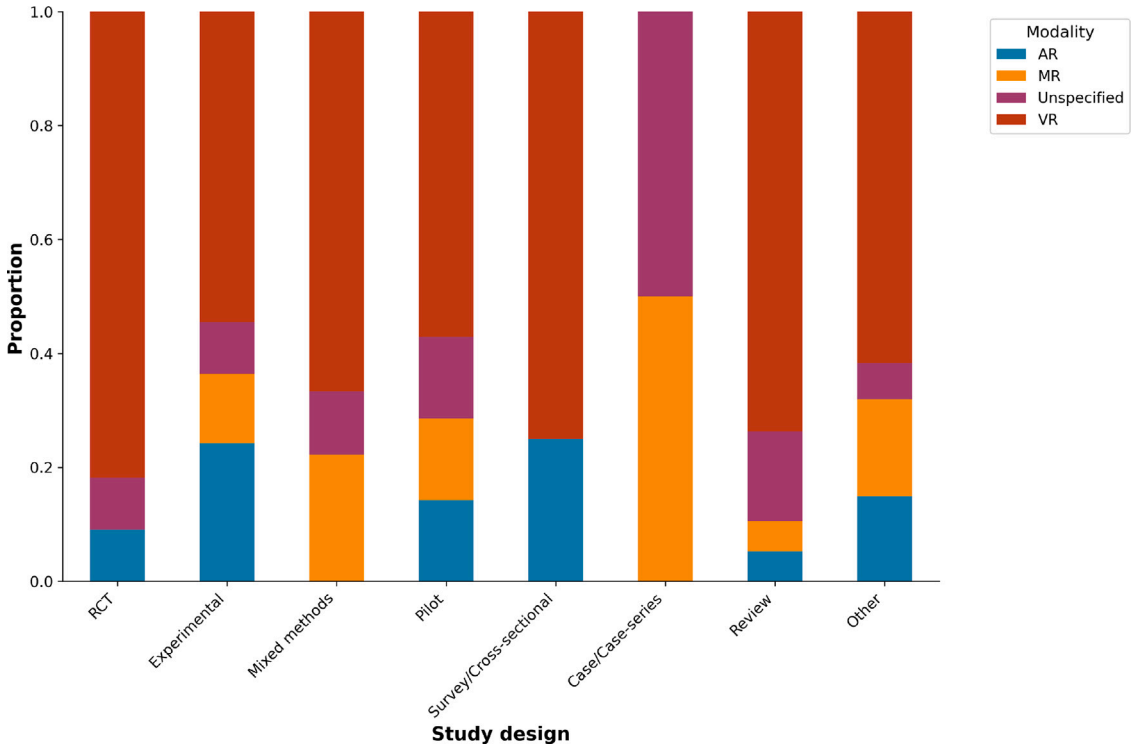

Fig. 4. Study design by XR modality.

Temporally, publication volume was strongly recent: 2019: 9; 2020: 4; 2021: 8; 2022: 12; 2023: 15; 2024: 42; 2025: 34. Approximately 68% of all studies were published from 2023–2025, underscoring that findings may reflect current toolchains and hardware capabilities more than stable effects over time [46,49,66,71].

By modality (coded from each record's *Technology Type* as shown in Fig. 4), VR was the most frequent label, followed by AR and MR. RCTs were predominantly VR-focused (e.g., sepsis team training in VR with AI-driven characters; [63–65]), with fewer AR trials (e.g., AI-guided bronchoscopy; [67]). Implementation studies and validation/QA work commonly integrated AR overlays, explainable-AI support, or automated scoring [75–77], reflecting a shift toward instrumentation, trust calibration, and scalable assessment.

Research question coverage was fragmented. Using the dataset's RQ flags, only-RQ1 (effectiveness) accounted for 47 studies; only-RQ2 (implementation) for 31; and only-RQ3 (quality assurance/ethics/validity) for 27. Overlaps were limited (RQ1+RQ2: 10; RQ1+RQ3: 13; RQ2+RQ3: 3), and just one study spanned all three. This pattern mirrors the field's separation between efficacy trials in controlled VR settings (e.g., [63,64]), implementation-focused deployments and frameworks [70,71], and QA/validity strands on explainability, trust, and automated assessment [75–77].

Together, these observations suggest: (i) limited but growing RCT evidence anchored in VR; (ii) small-to-mid samples with short-term outcomes; (iii) a recent surge in publications that may introduce temporal

Table 1
Distribution of included studies by research question (RQ) focus.

RQ Bucket	Count	Example studies (Title; Authors, Year)
Only RQ1	47	– Exploring the Application Capability of ChatGPT as an Instructor in Skills Education for Dental Medical Students: randomized Controlled Trial; Vannaprathip et al. 2025 – Effect of Feedback Modality on Simulated Surgical Skills Learning Using Automated Educational Systems: A Four-Arm randomized Control Trial; Yilmaz et al. 2023
Only RQ2	31	- Metaverse-based simulation: a scoping review of charting medical education over the last two decades in the lens of the 'marvelous medical education machine'; Ghaempanah et al. 2024 – Using Extended Reality (XR) for Medical Training and Real-Time Clinical Support during Deep Space Missions; Nam et al. 2022
Only RQ3	27	– First validated automated scoring system using the diagnostic arthroscopy skill score (DASS 2.0) for assessing proficiency in virtual reality arthroscopy; Anetzberger et al. 2025 – Enhancing Virtual Human Interactions by Designing a Real-Time Dialog Filter for Mitigating Nonsensical Responses; Harari et al. 2024
RQ1+RQ2	10	– Integrating Biofeedback and Artificial Intelligence into eXtended Reality Training Scenarios: A Systematic Literature Review; Blackmore et al. 2024 – How Managers Perceive AI-Assisted Conversational Training for Workplace Communication; Osborne et al. 2025
RQ1+RQ3	13	– Integrating large language model-based agents into a virtual patient chatbot for clinical anamnesis training; Huang et al. 2025 – SurgBox: Agent-Driven Operating Room Sandbox with Surgery Copilot; Agbontaen et al. 2024
RQ2+RQ3	3	– Immersive training of clinical decision making with AI-driven virtual patients – a new VR platform called medical tr.AI.ning; Liaw et al. 2023 – Human in the Loop for XR Training: Theory, Practice and Recommendations for Effective and Safe Training Environments; Teixeira et al. 2023
All three	1	– Artificial intelligence-enabled virtual reality simulation for clinical deterioration training: An effectiveness-implementation hybrid study; Liaw et al. 2023

effects; and (iv) sparse end-to-end studies uniting effectiveness, implementation feasibility, and robust QA. Bridging these strands, e.g., combining AI-adaptive VR trials [63,66] with implementation frameworks and validated, scalable outcome measurement [75–77], remains a key opportunity for the next wave of AI+XR medical education research.

4. Results

This section presents the findings from our systematic review of AI-driven characters in XR healthcare simulations. We first describe the study selection process and scope of included literature, followed by a breakdown of results according to the three research questions: RQ1 (effectiveness), RQ2 (implementation), and RQ3 (quality assurance). Summary statistics, illustrative examples, and patterns in study overlap are provided to contextualize the evidence base.

4.1. Study selection and scope

Across 2015–2025 we included 132 empirical studies that addressed at least one RQ. RQ coverage was RQ1 effectiveness (n = 71), RQ2 implementation (n = 45), and RQ3 quality assurance (n = 44).

Overlap was limited: 47 studies targeted only RQ1, 31 only RQ2, and 27 only RQ3; 10 addressed RQ1+RQ2, 13 RQ1+RQ3, 3 RQ2+RQ3, and 1 spanned all three RQs (Table 1). Publication output accelerated sharply after 2022, with 42 studies in 2024 and 33 by July 2025 (Fig. 3), indicating a field dominated by recent evidence and thus susceptible to temporal effects (e.g., LLM-era systems).

4.2. Designs, samples, and modalities

Study designs were heterogeneous, with 11 randomized controlled trials (RCTs) (~8%–9% of all studies), 33 experimental studies, 9 mixed-methods, 5 pilot, 3 surveys, 2 case/case-series, and 19 reviews. For studies reporting numeric samples, the median sample size was 41 (IQR 17–81), while RCTs with extractable Ns (n = 8) had a median 67 (IQR 56–120).

XR modality distribution (bucketed from "Technology Type") skewed toward VR (n = 82; 62.1%), followed by AR (n = 36; 27.3%); MR (n = 1; 0.8%) and XR-unspecified (n = 7; 5.3%) were rare (Fig. 5). When trials occurred, they were predominantly VR (RCTs by modality: VR = 10, AR = 2; Fig. 5). Modality aligned with RQ focus: VR dominated effectiveness (RQ1), while AR appeared more often in implementation/QA studies (RQ2–RQ3).

4.3. Effectiveness (RQ1)

Across healthcare education, the convergence of AI with XR is reshaping how learners acquire technical and non-technical skills. Evidence suggests AI can support adaptivity, reduce cognitive load, and personalize feedback, while preserving trainer oversight for safety [60,78].

Procedural/time outcomes: In an RCT of AI-guided bronchoscopy with AR overlays, clinicians completed tasks faster and more efficiently than expert-tutored controls (procedure time −77 s, P = .022; segments revisited −7, P = .019; SMD for post-only time ≈−0.68, 95% CI −1.32 to −0.04; negative = faster) [67]. A four-arm surgical VR RCT similarly showed that automated visual/visuospatial feedback improved benchmark attainment over practice-alone by the second/third repetition [64]

The chart provides an effect-direction overview across all included studies, grouped by outcome family and stratified by study design (see Fig. 6). Marker size is proportional to reported sample size (or CASP quality score when sample size was unavailable), with numbers indicating study counts per cell. The distribution highlights consistent advantages for AI/XR in procedural/time measures, more heterogeneous effects in knowledge/decision outcomes, mixed results for communication/teamwork, and limited but generally positive signals in errors/accuracy. Many studies reported unclear or within-group-only improvements, reflecting heterogeneity in design and reporting.

Knowledge/decision-making: A pooled analysis of two pre–post RCTs sepsis team training with an AI doctor [63] and a dental intelligent tutor [66] yielded a large standardized mean difference for knowledge/decision outcomes (g_{REML} = 1.31, 95% CI 0.08–2.54), but with substantial inconsistency ($\tau^2 = 0.67$; $I^2 = 85\%$), reflecting domain/task differences (sepsis cognition vs. endodontic decision-making).

Communication/teamwork: In the sepsis RCT, the AI-doctor group achieved higher knowledge but not superior communication performance versus human-controlled simulation; the human-controlled group reported higher communication self-efficacy [63]. Small experimental work suggests affective AI-driven characters can enhance realism

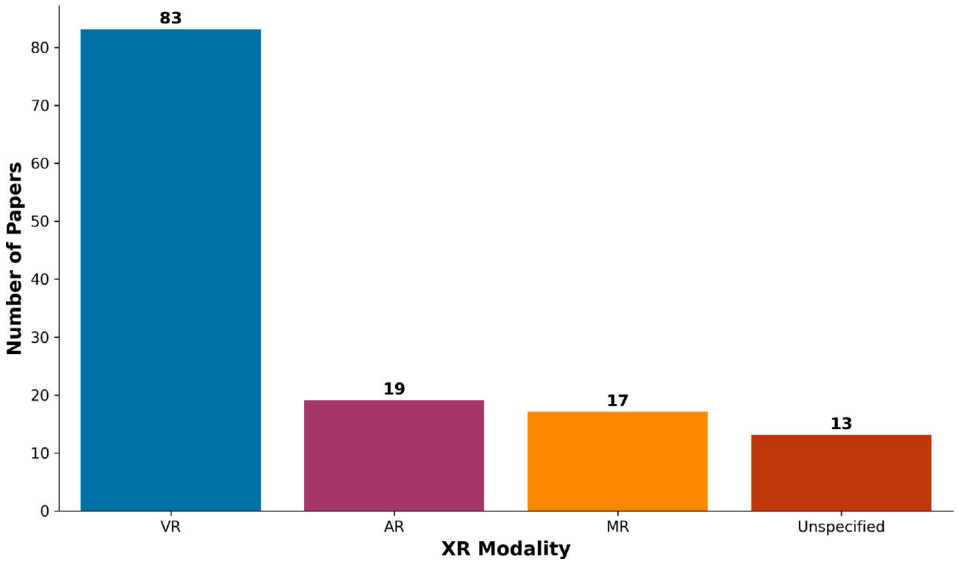

Fig. 5. XR modality distribution.

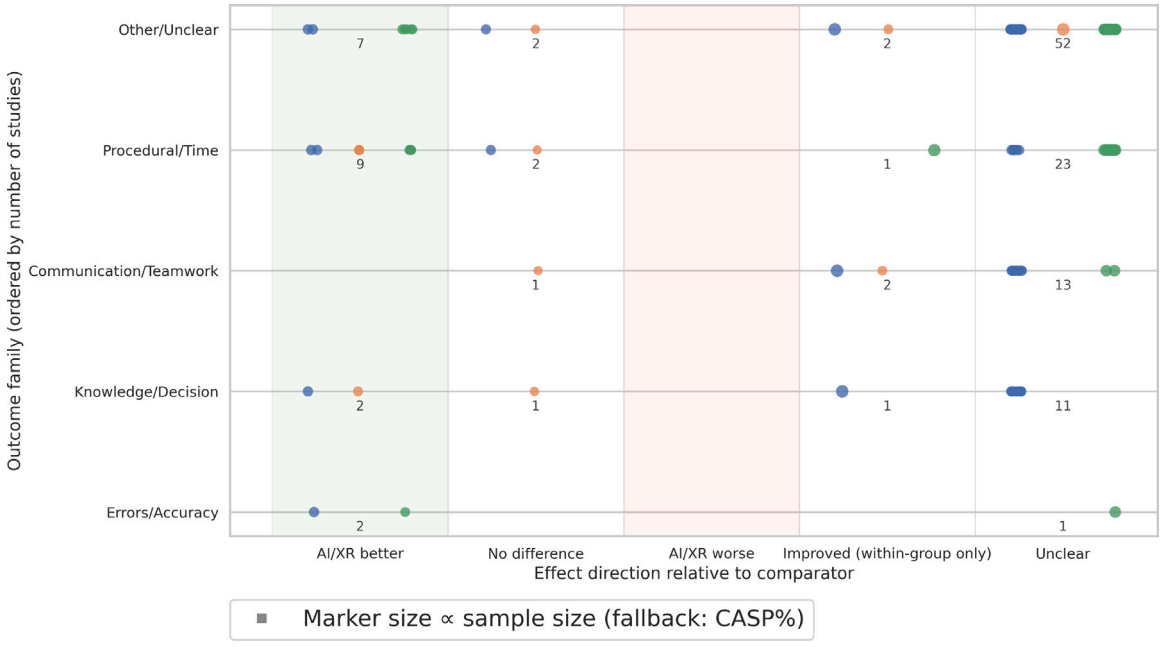

Fig. 6. Effect direction.

and communication skill practice, but samples are underpowered (n = 9) [69]. Controlled trials showed mixed effects on communication outcomes, in contrast to more consistent gains in procedural measures.

Measurement/assessment robustness: Automated assessment reliability supports scalable evaluation of technical performance AI scoring reached moderate-to-almost-perfect agreement with experts in laparoscopic tasks ($\kappa \approx 0.59$–0.86) and high ICC in automated arthroscopy scoring (ICC ≈ 0.89) [76,77], strengthening confidence in procedural outcome measurement; analogous validated instruments are less consistently reported for communication.

Across controlled trials in this review, procedural/time metrics consistently favored AI/XR (2/2 RCTs with significant improvements), knowledge/decision outcomes were favorable but heterogeneous (pooled $g \approx 1.31$ with wide PI), and communication performance effects were mixed (0/1 RCT superiority), despite signals from empathic AI-driven characters. Given small samples and domain variation, interpretation should emphasize outcome family and task context [63,64,66,67,69,70,76,77,79].

4.4. Implementation (RQ2)

Implementation reports commonly referenced phased rollouts (pilot → faculty development → broader integration) and compatibility considerations (headset availability, space, network constraints). Framework usage (e.g., CFIR, RE-AIM) was present but inconsistent; successful programmes emphasized leadership buy-in, faculty training, and low-friction tech stacks (e.g., cloud/offline modes, simple device provisioning). Barriers included cost, staff time, and data governance approvals.

Table 2
Summary of Findings (SOF) for AI-enhanced XR training in healthcare education.

Outcome	Absolute effect* (95% CI)	Relative effect (95% CI)	Participants (studies)	Certainty (GRADE)
Knowledge/Decision-making	AI/XR: 1.31 SD higher (0.08 to 2.54)	SMD 1.31 (0.08 to 2.54)	99 (2 RCTs)	•••• LOW[a]
Task performance time	AI/XR: 0.68 SD faster (1.32 to 0.04 faster)	SMD −0.68 (−1.32 to −0.04)	40 (1 RCT)	•••• LOW[b]
Cue recognition	AI/XR: 22% higher (7% to 40%)	RR 1.22 (1.07 to 1.40)	64 (1 RCT)	•••• LOW[c]

[a] Pooled effect from sepsis team training [63] and dental decision-making [66]; substantial heterogeneity ($I^2 = 85\%$).
[b] Single RCT of AI-guided bronchoscopy training [67]; negative SMD indicates faster AI performance.
[c] Single RCT of AR-virtual patient study [80].

4.5. Quality assurance and safety (RQ3)

Quality assurance practices included scenario versioning, expert review of AI behaviors, logging, and rule-based guardrails for unsafe actions. Some studies described algorithmic checks (e.g., bias or plausibility filters) and Human-in-the-Loop oversight for escalation. However, formal bias audits, privacy safeguards, and inter-rater reliability of assessment instruments were less frequently detailed, motivating the need for the AI-specific DASEX criteria introduced here and future empirical validation.

5. Meta-analysis

We prespecified four outcome families and fit random-effects models with REML where pooling was feasible, reporting τ^2, I^2, and 95% prediction intervals (PI). For knowledge/decision-making, two pre–post RCTs were eligible for change-score SMDs (Hedges g using the Morris method, $r = 0.50$): [63] ($n = 32/32$) and [66] ($n \approx 18/17$ per full text). Study effects were $g = 0.72$ (SE 0.26) and $g = 1.98$ (SE 0.42); the pooled effect favored AI/XR ($g_{\text{REML}} = 1.31$, 95% CI 0.08 to 2.54) with substantial heterogeneity ($\tau^2 = 0.67$, $I^2 = 85\%$) and a wide PI (−0.72 to 3.34), consistent with differing tasks (sepsis team cognition versus dental decision-making). Sensitivity analyses using $r = 0.30$ and $r = 0.70$ gave similar results. For task time, one RCT of AI-guided bronchoscopy [67] indicated faster performance with AI (post-only SMD $g = -0.68$, 95% CI −1.32 to −0.04; negative values denote faster AI). A four-arm RCT comparing feedback modalities [64] favored visual or visuospatial feedback over practice alone across multiple metrics, but group means and SDs were not extractable, so we summarize this trial narratively pending author data. For measurement validity, an AI assessor showed moderate to almost-perfect agreement with experts in laparoscopic tasks [77] ($\kappa = 0.59$ to 0.86) and high agreement for automated DASS 2.0 scoring [76] (ICC = 0.89), supporting scalable, blinded outcome scoring.

A four-arm RCT on feedback modalities [64] favored visual/visuospatial feedback over practice-alone on multiple performance metrics; figures/tables did not report group mean ± SDs in extractable form, so we report this study narratively pending author-level data.

Finally, to address measurement validity, an AI assessor showed moderate to almost-perfect agreement with experts in laparoscopic tasks [77] ($\kappa = 0.59$–0.86) and high agreement in automated assessment of DASS 2.0 scores [76] (ICC = 0.89), supporting the feasibility of scalable, blinded outcome scoring in this domain.

6. Summary of key quantitative findings

This section summarizes the main quantitative results for AI-enhanced extended reality (XR) training in healthcare education. Patient or population: healthcare learners (nursing students, dental students, critical care physicians). Setting: virtual and augmented reality simulation environments. Intervention: XR training systems incorporating AI-driven characters. Comparison: traditional training methods or human-controlled simulation (see Table 2).

GRADE certainty definitions: High: very confident the true effect is close to the estimate. Moderate: moderately confident; the true effect is likely close to the estimate, but may differ substantially. Low: limited confidence; the true effect may differ substantially from the estimate. Very low: very little confidence; the true effect is likely substantially different from the estimate.

Notes on evidence synthesis: Due to heterogeneity in study designs, outcome measures, and statistical reporting, only 4 studies (3.0%) met the eligibility criteria for quantitative synthesis. The remaining 128 studies (97.0%) were synthesized narratively due to methodological incompatibility or insufficient statistical data for pooling.

7. Critical appraisal using CASP

We evaluated study quality using an adaptation of the Critical Appraisal Skills Programme (CASP), retaining its ten core domains study aims, design alignment, recruitment transparency, outcome validity, analytic rigor, reflexivity, ethics, and practical value while implementing a quantitative rubric that assigns up to 20 points per study (converted to percentage scores). Weighting emphasized objective indicators (e.g., randomization, validated instruments, preregistration, explicit ethical approval). This enabled consistent evaluation across heterogeneous designs, from randomized trials to exploratory prototypes. The full adapted checklist is provided in Appendix.

In our corpus, randomized controlled trials achieved the highest quality scores, exemplified by an RCT integrating AI tutoring with VR-based skill assessment (95%; [79]). Systematic literature reviews similarly scored highly, reflecting clear methods and transparent synthesis (95%; [49]). Experimental studies with small samples also performed well when methods and measures were explicit (85%; [69]). Narrative field reviews showed wider variability, ranging from strong, well-scoped syntheses (90%; [46]) to more general overviews with limited protocol detail (70%; [72]). Early descriptive deployments and feasibility/technical prototypes tended to score lower due to limited recruitment detail, short timeframes, and minimal outcome validation (75% and 60%, respectively; [68,71]). Qualitative project evaluations with clear sampling and analysis could reach high scores (90%; [70]). Conceptual and theoretical contributions, while valuable for framing, typically scored mid-range given the absence of empirical methods (65%; [73,78]). Technology surveys landed in the upper mid-range when search strategies and coverage were explicit (80%; [81]). Overall, the adapted CASP rubric provided an evidence-informed, design-agnostic approach to compare methodological quality, highlighting where standards are strongest (RCTs, systematic reviews) and where further rigor is needed (feasibility prototypes, early narrative work) (see Fig. 7).

8. Discussion

This review synthesized evidence from 132 empirical studies investigating AI-driven characters in XR healthcare simulations, addressing three domains: effectiveness (RQ1), implementation frameworks (RQ2), and quality assurance (RQ3). The breadth of the included literature reflects growing interest in XR–AI clinical training, yet the heterogeneity of study designs, outcome measures, and reporting standards limits the ability to pool results quantitatively.

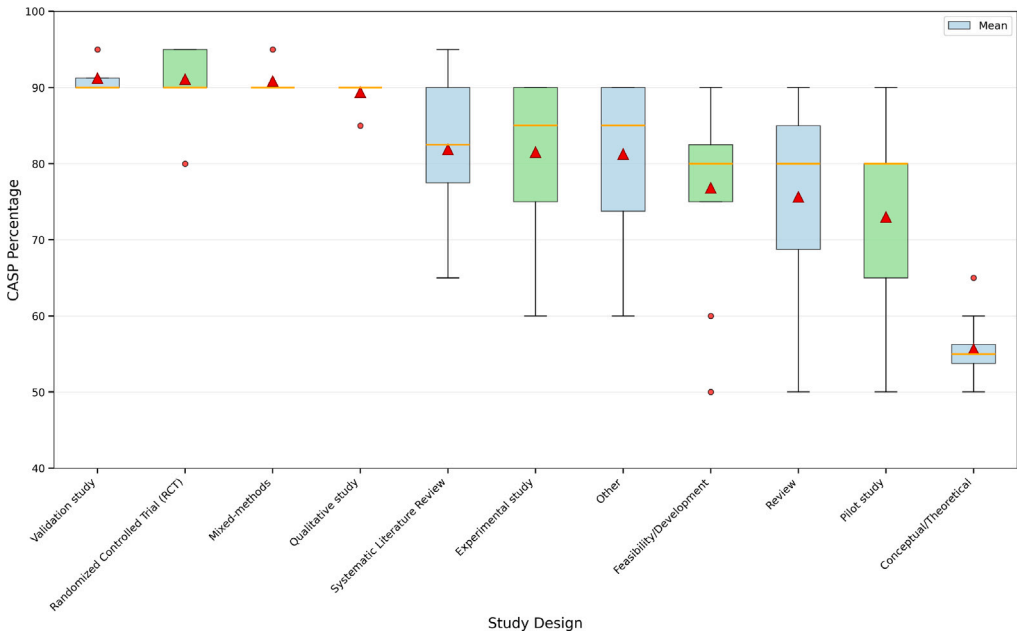

Fig. 7. Distribution of CASP scores by study design.

8.1. Effectiveness of AI-driven XR simulations (RQ1)

Across the 71 studies addressing RQ1, the evidence indicates that AI-enhanced healthcare simulations can improve knowledge acquisition, procedural skill performance, and decision-making speed compared to traditional training methods. These effects were most consistent in controlled studies evaluating discrete procedural skills (e.g., airway management, surgical suturing) or team-based crisis scenarios. Improvements were often accompanied by gains in learner confidence and engagement, although studies using standardized satisfaction scales (e.g., SUS, NASA-TLX) reported variable results. Notably, very few studies evaluated transfer to real-world clinical settings, and follow-up periods beyond immediate post-training were rare.

8.2. Implementation frameworks (RQ2)

The 45 studies mapped to RQ2 revealed a fragmented approach to implementation. While some adopted structured frameworks such as Kirkpatrick's model or Miller's pyramid to guide evaluation, most relied on bespoke protocols without explicit theoretical grounding. Hardware configurations were dominated by VR headsets (particularly Oculus/Meta Quest), with AR deployments often limited to pilot studies or procedural guidance applications. Integration of AI-driven NPCs varied from rule-based branching logic to advanced NLP and reinforcement learning, though few studies provided technical transparency on AI model architecture or training data provenance. Cross-disciplinary collaboration between clinicians, simulation designers, and AI developers emerged as a key facilitator of successful implementation, supporting previous observations on the importance of interdisciplinary design in complex simulation environments [82–85].

8.3. Quality assurance and safety (RQ3)

The 44 studies addressing RQ3 demonstrated a notable gap in formal quality assurance protocols. Safety features such as intervention plausibility checks, bias audits, or error interception were infrequently reported [50,84]. Only a minority of studies described logging AI reasoning processes or providing learners with transparent explanations for NPC actions. The lack of structured QA raises concerns regarding reproducibility, learner trust calibration, and risk management, particularly when simulations are deployed in high-stakes training contexts. Examples such as HealthBench's example-specific rubrics, validated across 48,000+ criteria [37], and the MLASE surgical checklist [45,59] illustrate how consensus scoring, stratified dimensions (e.g., accuracy, completeness, communication), and human-in-the-loop oversight could be adapted to XR-AI systems. Incorporating these elements together with algorithmic-bias mitigation, fidelity validation, and expert review would improve both rigor and flexibility, enabling reproducible yet context-sensitive quality control [86,87].

8.4. Experience level, cognitive load, and adaptive difficulty

These findings align with Cognitive Load Theory (CLT), which emphasizes balancing intrinsic (task complexity), extraneous (inefficient design or distractions), and germane (schema-building) load to optimize learning [88–90]. Communication-focused scenarios often demand emotionally rich, socially dynamic interactions such as empathy, negotiation, or conflict resolution that can increase extraneous load [91–94], and current AI-driven characters still struggle to manage this complexity consistently [95,96]. While novice learners benefit from scaffolded AI guidance [97–99], poorly tuned adaptivity can overwhelm experienced users, reducing performance and engagement [100,101]. One approach is to implement manual or semi-automated difficulty modes (e.g., scaffolded, standard, advanced) that adjust scenario complexity, AI prompt precision, and required autonomy to maintain optimal cognitive load [88,90] (see Table 3).

8.5. Geographic distribution and LMIC representation

Most studies originated from high-income countries, notably the United States, Canada, Western Europe, Japan, Republic of Korea, and Switzerland, with growing contributions from East Asia and the Middle East [102–105] (see Table 3). Bibliometric mapping confirms concentration in the United States, United Kingdom, and China, with collaborations clustered in North America, Western Europe, and East Asia [106]. LMIC and UMIC studies such as those from Egypt [107], Palestine [108], South Africa [109], Thailand [66], and China [110]—were fewer and tended to focus on feasibility, cost-conscious or hybrid delivery models, and local barriers (infrastructure, training, affordability) [107,109,111]. Large geographic gaps persist in Sub-Saharan Africa and much of Latin America and South/Southeast Asia outside major hubs [47,106].

Table 3
Geographic distribution of included studies.

Region	Count
North America	45
Europe	32
East Asia & Pacific	25
Middle East & North Africa	15
Unspecified	14
South Asia	11
Central Asia	2
Latin America & Caribbean	1
Sub-Saharan Africa	1

8.6. Ethics and equity

Ethical considerations reported include equitable access, inclusion of underserved learners, bias mitigation, transparency in AI-assisted feedback, informed consent in immersive studies, and cultural/linguistic adaptation [47,112,113]. Reviews highlight ethics as a key governance priority encompassing fairness, autonomy, explainability, and consent for multimodal data capture [47,112]. Equity is closely linked to geography: disparities in device cost, bandwidth, and digital literacy risk widening gaps unless systems adopt accessibility-first designs (mobile/low-bandwidth operation, localization, disability accommodations) and receive institutional support [102,111].

8.7. Cost and economic considerations

Several studies report potential cost savings compared with standardized patients or high-fidelity manikins, particularly with virtual patients, consumer-grade VR, and AI-assisted assessment [105,107,110,114]. Common cost-reduction levers include low-cost training boxes and scalable cloud models [110,115]. Reported cost challenges include hardware, content development, maintenance, and faculty training; high up-front and integration costs remain common, and few formal cost–utility analyses exist [47,102,104]. Suggested strategies include consumer-grade devices, open/reusable content, cloud-plus-edge deployment, shared regional hubs, and partnerships to offset licensing and support costs [107,114,115].

8.8. Data privacy and security

Privacy, security, and consent are frequently addressed through de-identification, IRB/REB approval, and informed consent procedures [47,103,112]. Risks include re-identification from multimodal XR data, insecure data pipelines, cross-platform sharing, and adversarial vulnerabilities in AI models [116,117]. Recommended safeguards include encryption, access controls, audit logging, data minimization/retention limits, differential privacy or synthetic data, federated learning, and governance aligned with GDPR/HIPAA analogues, alongside model transparency and XR/AI-specific threat modeling [74,102,116,117].

Implications:. To avoid exacerbating inequities, programs should integrate privacy-by-design and formal safety cases with accessibility-first deployments, and publish context-specific cost-effectiveness data from LMIC/UMIC pilots e.g., offline/low-bandwidth operation, on-device inference where privacy or connectivity is constrained, and local data stewardship [107,109,111,115].

Key gaps:. Few comparative multi-region evaluations, especially involving LMICs, report standardized outcomes for learning, cost, safety, or equity [106,108,109]. The corpus remains geographically uneven, ethically aware but inconsistently operationalized, economically promising yet under-evaluated, and privacy-conscious but with limited real-world validation of XR/AI-specific safeguards [47,102,114,116]. Key limitations include the underrepresentation of Sub-Saharan Africa and lower-income contexts, a lack of long-term outcomes, scarce head-to-head cost–utility studies, and limited testing of privacy/security measures in deployed systems [47,106,117].

Recommended reporting minimums:. Report country/income classification, device/bandwidth requirements, accessibility features, bias assessments, consent/anonymization procedures, data governance structures, security testing (including adversarial robustness), and a cost worksheet covering devices, licences, content, support, and refresh cycles [112,115,117].

Priority research needs:. Multicentre LMIC trials, standardized economic evaluations, cultural/linguistic adaptation studies, and red-team evaluations of AI-augmented XR training pipelines [107,109,114,116].

Practical implementation checklist:. Mobile-first/offline modes; consumer-grade hardware; localized content; bias audits; federated or synthetic data; encryption and access controls; audit and incident response; transparent assessment rubrics; faculty training; and community partnerships for equitable rollout [102,112,113,115].

8.9. Methodological limitations

Substantial heterogeneity in study designs (RCTs, quasi-experiments, pre–post single-group studies) and outcome measures limits comparability and meta-analysis.

Sample sizes varied widely, and many pilots were underpowered. Reporting on allocation concealment, blinding, and attrition was inconsistent, introducing potential bias. Most studies lacked cost-effectiveness analysis, limiting assessment of scalability across health system contexts.

8.10. Implementation barriers

Key barriers to routine adoption of AI-driven characters in XR curricula include high initial costs, technical complexity, faculty resistance, curriculum misalignment, and sustainability concerns [83,124,125]. Organizational support is a decisive adoption factor [126,127]; without it, resource constraints and unfamiliarity with XR-AI can stall integration [128,129]. Even with support, staff trust may dip if rollouts outpace governance [130]. About 34% of deployments required custom hardware or proprietary integrations [131]. Lightweight training systems that reduce local processing needs, e.g., through cloud rendering, can expand reach [132,133,133–136].

The Consolidated Framework for Implementation Research (CFIR) offers a structured approach to overcoming these challenges, addressing readiness, leadership commitment, resource allocation, sustainability, and stakeholder buy-in [57,58]. CFIR-based integration has doubled staff compliance and halved onboarding time in some deployments [137]. As AI-XR systems scale beyond pilots, frameworks like CFIR will be vital for aligning innovation with institutional capacity.

8.11. Ethics and equity (RQ2 and RQ3)

Ethics and equity remain unevenly addressed. Technical studies largely focus on performance, accuracy, and usability, with minimal attention to ethical risks or societal impacts [72,86,138]. Conceptual and review papers more often cover fairness, oversight, and inclusivity [85]. Key concerns include algorithmic bias from non-representative training data [139,140], automation bias and over-reliance on AI [47,141,142], data privacy risks from behavioral/biometric capture without robust governance [127,141], and implementation challenges in resource-limited settings [85,132,140,143]. Fewer than 10% of reviewed studies addressed these systematically, with limited LMIC representation exacerbating inequities [144,145].

Recommendations include mandatory Algorithmic Impact Assessments (AIA) and regular bias audits [144,146], embedding the FAIR and CARE principles [58], and ensuring human oversight mechanisms [87]. As Singaram et al. note, addressing equity requires not only mitigation but proactive inclusive design [145].

Table 4
Summary of cross-cutting challenges and proposed solutions in XR-AI healthcare simulation research.

Challenge	Why it is a problem	Suggested solution
Misalignment in assessment modalities between technical and communication skills	Limits comprehensive training effectiveness; objective metrics for communication are underdeveloped	Develop reproducible, validated evaluation frameworks for socio-emotional competencies; explore context-aware AI tools Section 4.2
Methodological heterogeneity	Hinders comparability and generalizability of results; many studies lack empirical rigor	Encourage use of RCTs, mixed-methods designs, and structured validation tools like CASP [118]; emphasize longitudinal/multi-site studies
Domain-specific variability in outcomes	Overstates utility in technical domains while underrepresenting challenges in transferrable skill training	Use domain-specific evaluation frameworks; apply Cognitive Load Theory (CLT) to structure XR scenarios more thoughtfully [88,89]
Differential impact based on learner experience	Uniform AI behavior can overwhelm novices or underserve experts	Implement adaptive modes (e.g., guided, standard, leadership) and allow manual calibration of difficulty [90,100]
Institutional and curricular barriers	Without structural integration, even good tools fail to scale	Apply CFIR to guide sustainable, phased adoption [119,120]; reduce hardware requirements through cloud-based delivery
Ethical and equity concerns	Risk of algorithmic harm, poor global applicability, and user distrust	Use Algorithmic Impact Assessments (AIA)[52,53], embed FAIR/CARE principles [121], ensure human oversight, and prioritize upstream inclusion
Lack of standardized validation frameworks	Undermines efforts for synthesis, benchmarking, and replication	Develop unified validation checklists (e.g., based on MLASE, HealthBench), with explicit fidelity, bias, and oversight components [35,122,123]

8.12. Summary of key challenges and solutions

To consolidate the themes outlined in this discussion, Table 4 summarizes the primary challenges identified, their implications, and the solutions proposed in each subsection.

9. Proposed frameworks

To address the challenges and gaps identified in this review, we propose an integrated framework, drawn from established models and modified for AI-driven XR simulation contexts and recent work on open-ended AI health evaluations such as HealthBench [37]. While HealthBench focuses on free-form LLM interactions, its methodology, particularly its rubric-based grading, physician consensus process, and performance stratification by behavioral axes, offers insights that can be adapted to the more structured, role-bound nature of simulation AI-driven characters.

We introduce the DASEX framework, a comprehensive, structured approach designed to ensure robust evaluation and optimal deployment of AI-driven healthcare simulations. DASEX aligns with contemporary evaluation and reporting guidance, including DECIDE-AI for early-stage clinical evaluation and TRIPOD+AI for transparent reporting of AI-enabled prediction models [40,51]; it also incorporates recent proposals for human evaluation of LLMs in healthcare [37]. It directly addresses the identified challenges across clinical accuracy, adaptive learning, risk management, teamwork, and transparency. A synthesis of the foundational frameworks to adopt during simulation development is summarized in Table 5. Each lens is aligned with a specific research aim and mapped to relevant research questions (RQs).

9.1. Pedagogical alignment

AI-driven XR simulations align well with Kolb's experiential learning cycle—from hands-on engagement to reflective abstraction. Adaptive feedback loops, guided resets, and scenario variation help support deliberate practice.

Integration of Cognitive Load Theory (CLT) is essential: studies show that emotional complexity in AI–NPC interactions can overload novice users [96,150]. It is important to align interaction modes with user cognitive profiles. *Future systems should incorporate cognitive-load telemetry to dynamically adjust difficulty and prevent disengagement* [147]. This pedagogical alignment supports more effective skill acquisition, addressing RQ1.

9.2. Implementation

Explicit application of CFIR constructs can systematically address implementation barriers as shown in Fig. 8, including organizational readiness, technology compatibility, sustainability, and stakeholder engagement.

Structured implementation guided by CFIR enhances practical feasibility, scalability, and stakeholder buy-in across diverse educational settings [57,58]. Utilizing RE-AIM complements this by tracking long-term adoption, implementation fidelity, and sustainability. This approach aligns with RQ2 and ensures AI-driven XR training solutions meet enduring educational and clinical-impact standards [57].

9.3. Positioning of DASEX

Existing implementation and evaluation frameworks, such as MLASE and CFIR, do not capture the dynamic, AI-specific behaviors of NPCs in XR simulations, for example e.g., real-time diagnostic reasoning, adaptive learner feedback, or bias mitigation. The DASEX framework addresses this gap through five domains: Diagnosis, Adaptation, Safety, Engagement, and Transparency. While full validation is pending, the planned protocol (provided in Appendix A) outlines a multi-coder reliability study and an iterative refinement process to ensure complementarity with existing models.

9.3.1. Proposed empirical validation of the DASEX framework

DASEX's criteria require live AI-XR simulations with interactive NPCs to assess features such as physiology-driven realism, safety intercepts, and adaptive communication, making retrospective application to published studies impractical. To address this, an Enhanced Pilot Validation Study will recruit healthcare trainees to interact with a functional AI-XR simulation, scored using the 25-item DASEX checklist alongside independent usability, workload, and self-efficacy measures. Structured debrief interviews will capture qualitative feedback for framework refinement, with inter-rater consistency analyzed for reliability. Performance thresholds for different training contexts (e.g., practice vs. high-stakes assessment) and follow-up measures at Kirkpatrick Level 3 will be explored. The current review thus provides the conceptual basis, and the forthcoming pilot represents DASEX's operationalization and empirical validation.

Table 5
Strategic lenses for AI–XR integration in healthcare training.

Lens	What to adopt	Aim	Research question
Pedagogy	Kolb-aligned learning cycles with adaptive feedback and CLT-based modulation [58,95,147–149]	Prevents overchallenge, reinforces deliberate practice.	RQ1: Effectiveness
Implementation	CFIR for phased roll-out, RE-AIM for life-cycle evaluation [57,58]	Aligns technology with institutional readiness & long-term impact.	RQ2: Implementation
Quality assurance	MLASE-inspired checklist + rubric scoring à la HealthBench + HITL safeguards [59,86,87]	Enables replicability, ethical robustness, and learner trust.	RQ3: Quality assurance
Equity & Ethics	FAIR + CARE data principles and mandatory AIAs with audit cycles [58,144,145]	Promotes inclusivity and transparency.	RQ3: Quality assurance

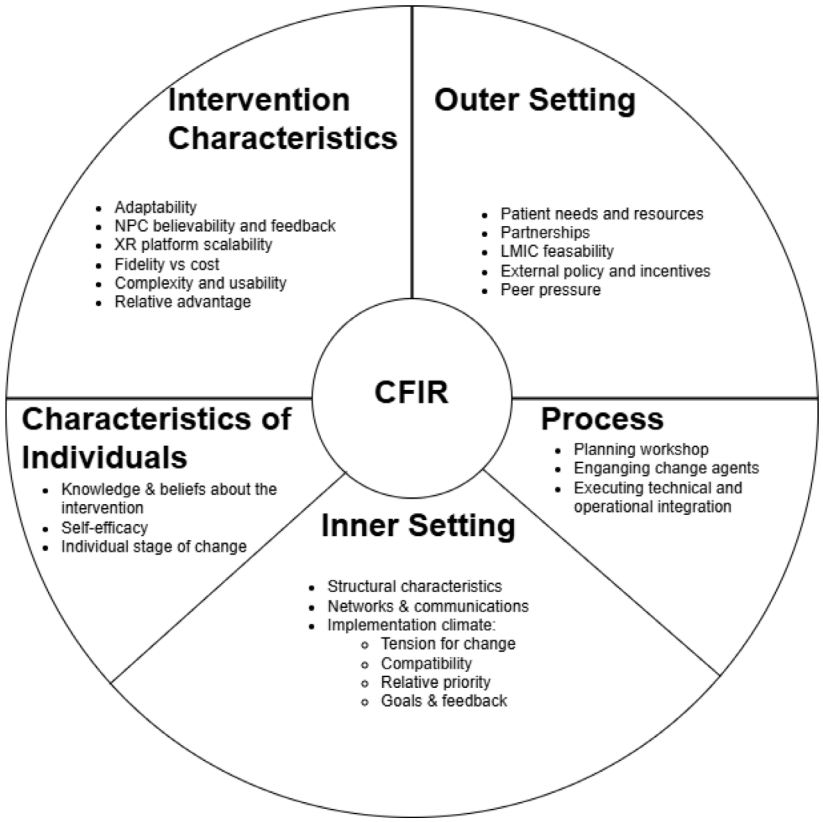

Fig. 8. Application of CFIR to structure implementation of AI-driven XR healthcare simulations.

9.4. DASEX framework explanation

The DASEX framework provides a structured approach to evaluating AI-driven healthcare simulations across five axes (Diagnosis, Adaptation, Safety, Engagement and Teamwork, and Explainability) ensuring clinical realism, learner adaptability, risk mitigation, collaborative training, and transparency.

Diagnosis (D): Clinical accuracy and reasoning robustness

Clinical realism and effective diagnostic training form the cornerstone of healthcare simulation. The Diagnosis axis emphasizes accurate initial diagnosis, ensuring AI-generated judgments align closely with presenting patient features. Active differential management systematically evaluates alternative diagnoses using evidence-based reasoning. Clinical-pathway fidelity ensures that AI-driven diagnostic actions strictly adhere to validated clinical standards and protocols. Dynamic diagnostic updates continually refine diagnoses with new clinical data, such as vital signs or laboratory results. Diagnostic robustness checks proactively monitor AI diagnoses, flagging significant deviations from realistic or expected clinical outcomes to enhance reliability and safety.

Adaptation (A): Personalization and scenario responsiveness

Personalized and adaptive learning experiences are essential for optimal educational outcomes. The Adaptation axis tailors scenario difficulty, feedback, and guidance specifically to individual learners' demonstrated expertise through learner-profile adaptation. Physiology-driven realism ensures patient physiological responses accurately reflect trainee interventions, such as vital sign changes after fluid administration. Emotional state calibration adjusts the intensity of AI interactions based on trainee stress indicators. Adaptive replay dynamically modifies scenarios to reinforce identified trainee skill gaps, promoting targeted skill improvement. Scenario logic consistency ensures AI actions remain logically coherent across adaptive scenario branches.

Safety (*S*): Harm prevention and AI risk management

Ensuring patient safety and managing AI-related risks within healthcare simulations is critical. The Safety axis includes intervention plausibility checks to prevent clinically implausible trainee or AI-generated actions. Strict protocol adherence and escalation mechanisms ensure compliance with established clinical guidelines, such as ACLS and NICE. Error interception actively prevents unsafe interventions, including medication overdoses or incorrect procedures. Algorithmic bias safeguards systematically audit AI decisions to detect and mitigate

Table 6
Checklist format of the DASEX framework for evaluating AI-driven XR healthcare simulations. Each criterion is scored from 1 (poor) to 5 (excellent).

Axis	Focus area	Criteria	Score (1–5)
D	Clinical accuracy and reasoning processes	1. **Accurate initial diagnosis** First-pass clinical judgment aligns with presenting features.	☐ ☐ ☐ ☐ ☐
		2. **Active differential management** Systematically rules in/out alternatives using evidence-based logic.	☐ ☐ ☐ ☐ ☐
		3. **Clinical-pathway fidelity** Diagnostic actions mirror real-world protocols.	☐ ☐ ☐ ☐ ☐
		4. **Dynamic diagnostic updates** Adjusts diagnosis in response to new data.	☐ ☐ ☐ ☐ ☐
		5. **Diagnostic robustness checks** Flags AI diagnoses that deviate from realistic clinical expectations.	☐ ☐ ☐ ☐ ☐
A	Personalization and scenario responsiveness	1. **Learner-profile adaptation** Adjusts difficulty and feedback to learner expertise.	☐ ☐ ☐ ☐ ☐
		2. **Physiology-driven realism** Vitals evolve plausibly after interventions.	☐ ☐ ☐ ☐ ☐
		3. **Emotional state calibration** Adjusts tone, urgency, or affective feedback based on learner stress level.	☐ ☐ ☐ ☐ ☐
		4. **Adaptive replay/reinforcement** Replay adapts to learner gaps.	☐ ☐ ☐ ☐ ☐
		5. **Scenario logic consistency** Ensures AI actions and patient responses remain coherent across branches.	☐ ☐ ☐ ☐ ☐
S	Harm prevention and risk management	1. **Intervention plausibility check** Prevents clinically implausible or dangerous actions.	☐ ☐ ☐ ☐ ☐
		2. **Protocol adherence & escalation** Follows escalation pathways.	☐ ☐ ☐ ☐ ☐
		3. **Error interception** Prevents unsafe actions.	☐ ☐ ☐ ☐ ☐
		4. **Algorithmic bias safeguards** Audits and mitigates AI biases to ensure equitable performance.	☐ ☐ ☐ ☐ ☐
		5. **Safety audit trail** Logs and reports events.	☐ ☐ ☐ ☐ ☐
E	Collaboration, communication, and user experience	1. **Turn-taking coordination logic** Maintains procedural order and prevents conflicting actions.	☐ ☐ ☐ ☐ ☐
		2. **Structured communication** Uses SBAR or equivalent.	☐ ☐ ☐ ☐ ☐
		3. **Coordination efficiency** Prevents duplication or missed actions.	☐ ☐ ☐ ☐ ☐
		4. **Trust monitoring & calibration** Measures and adapts to trainee trust in AI.	☐ ☐ ☐ ☐ ☐
		5. **Contextual debrief anchoring** Links simulation events to learning points for structured reflection.	☐ ☐ ☐ ☐ ☐
X	Transparency and evidence-based reasoning	1. **Transparent reasoning logs** AI decision processes are logged and reviewable.	☐ ☐ ☐ ☐ ☐
		2. **Evidence alignment** References clinical guidance.	☐ ☐ ☐ ☐ ☐
		3. **On-demand explanations** Explanations toggle for clarity.	☐ ☐ ☐ ☐ ☐
		4. **Persistent documentation** Logs actions and reasoning.	☐ ☐ ☐ ☐ ☐
		5. **Counterfactual rationale generation** Provides alternative outcome explanations for comparison.	☐ ☐ ☐ ☐ ☐

biases, promoting equitable and safe AI performance. Comprehensive safety audit trails transparently log safety-related AI decisions and interventions, facilitating thorough quality assurance and review processes.

Engagement and Teamwork (E): Collaboration and trust-aware user experience

Effective teamwork and clear communication are essential clinical competencies. The Engagement and Teamwork axis emphasizes turn-taking coordination logic to maintain procedural order and prevent action conflicts or duplications. Structured communication frameworks, such as SBAR and closed-loop communication, guide clear interactions. Workflow coordination efficiency ensures seamless AI–human teamwork, avoiding redundancies and oversights. Trust monitoring and calibration explicitly measures trainee trust in AI using defined metrics, dynamically adjusting AI interaction levels to maintain optimal trust levels. Contextual debrief anchoring links simulation events directly to teamwork and collaboration objectives, facilitating structured reflective learning.

Explainability (X): Transparency and evidence-based AI decision-making

Transparency and evidence-based reasoning underpin learner trust and educational credibility in AI-driven simulations. The Explainability axis includes transparent reasoning logs, explicitly documenting AI decision-making processes for learner review and audit. Authoritative evidence alignment ensures AI recommendations reference recognized clinical guidelines, such as NICE or WHO, reinforcing educational validity. On-demand explanation granularity allows learners to control the detail of AI explanations, matching cognitive and experiential needs. Persistent documentation systematically records comprehensive AI decision logs for validation and reflective learning. Counterfactual decision explanations clearly articulate why alternative clinical actions were not selected, deepening learner clinical reasoning and critical thinking (see Table 6). .

9.5. Implications for practice and research

For educators, these findings suggest that AI-driven extended reality (XR) can enhance learner engagement and skill acquisition; however, adoption should be guided by robust quality assurance processes and grounded in established educational theory to ensure pedagogical fidelity and safety. For researchers, key priorities include conducting multi-site, adequately powered randomized controlled trials using standardized outcome measures to improve comparability and generalizability. There is also a need to evaluate long-term knowledge retention and transfer of skills to real-world clinical practice, as current evidence is largely confined to immediate post-intervention outcomes. Geographically, the evidence base must expand to include low and middle income countries to address equity and scalability. Future studies should routinely incorporate cost and implementation feasibility analyses to inform policy and institutional investment. Finally, domain-specific evaluation frameworks, such as DASEX for AI-enhanced simulations, require empirical validation to support consistent, transparent, and safe deployment of AI-driven characters in healthcare training.

9.6. Limitations of the review

Our exclusion of non-English literature and most gray literature may have omitted innovative work, particularly from regions where English is not the primary language of publication. While targeted ACM

searches and citation snowballing partially addressed this, the risk of geographic bias remains, potentially excluding relevant developments from countries with significant XR and AI innovation where English is not the dominant language. The exclusion of broader gray literature, including industry whitepapers and non-peer-reviewed technical reports, may also have omitted emerging but unpublished work. Formal small-study effect tests (e.g., Egger's) were only feasible when $k \geq 10$; most outcomes had insufficient studies for reliable testing. Given the rapid pace of AI development, studies published after the search cut-off (July 2025) may reflect substantive technological advances not captured here.

CRediT authorship contribution statement

David Dasa: Writing – review & editing, Writing – original draft, Validation, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Michele Board:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Ursula Rolfe:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Tom Dolby:** Writing – review & editing, Funding acquisition. **Wen Tang:** Writing – review & editing, Supervision, Methodology, Data curation.

Disclosure

This work was supported in part by a Bournemouth University Ph.D. studentship match-funded by i3 Simulations. One co-author (Tom Dolby) is an employee of i3 Simulations. i3 Simulations had no role in the design, conduct, data collection, analysis, interpretation, or writing of the systematic review, and no influence over the decision to submit the manuscript for publication. All other authors declare no competing interests.

Ethics statement

This study did not involve human participants, human data, or human tissue. Therefore, ethical approval and informed consent were not required. All procedures complied with relevant institutional, national, and international guidelines.

Funding

This work was supported in part by a Bournemouth University Ph.D. studentship match-funded by i3 Simulations. No other funding was received for this study.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: David Dasa reports financial support was provided by i3 Simulations. Tom Dolby reports a relationship with i3 Simulations that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.artmed.2025.103270.

References

[1] Sterne Jonathan AC, Savović Jelena, Page Matthew J, et al. Rob 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898.
[2] Sterne Jonathan AC, Hernán Miguel A, Reeves Barnaby C, et al. Robins-i: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ 2016;355:i4919.
[3] Guyatt Gordon H, Oxman Andrew D, Schünemann Holger J, Tugwell Peter, Knottnerus Andre. Grade guidelines: A new series of articles in the journal of clinical epidemiology. J Clin Epidemiol 2011;64(4):380–2.
[4] Rivera Samantha Cruz, Liu Xiaoxuan, Chan An-Wen, Denniston Alastair K, Calvert Melanie J, Ashrafian Hutan, Beam Andrew L, Collins Gary S, Darzi Ara, Deeks Jonathan J, Khair ElZarrad M, Espinoza Cyrus, Esteva Andre, Faes Livia, Ruffano Lavinia Ferrante di, Fletcher John, Golub Robert, Harvey Hugh, Haug Charlotte, Holmes Christopher, Jonas Adrian, Keane Pearse A, Kelly Christopher J, Lee Aaron Y, Lee Cecilia S, Manna Elaine, Matcham James, McCradden Melissa, Moher David, Monteiro Joao, Mulrow Cynthia, Oakden-Rayner Luke, Paltoo Dina, Panico Maria Beatrice, Price Gary, Rowley Samuel, Savage Richard, Sarkar Rupa, Vollmer Sebastian J, Yau Christopher. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the spirit-ai extension. Lancet Digit Heal 2020;2(10):e549–60.
[5] Liu Xiaoxuan, Rivera Samantha Cruz, Moher David, Calvert Melanie J, Denniston Alastair K, CONSORT-AI, Group SPIRIT-AI Working. Reporting guidelines for clinical trials evaluating artificial intelligence interventions: the consort-ai extension. Nature Med 2020;26(9):1364–74.
[6] Salas E, Wilson KA, Burke CS, Priest HA. Using simulation-based training to improve patient safety: What does it take? Jt Comm J Qual Patient Saf 2005;31(7):363–71.
[7] Ayaz O, Ismail FW. Healthcare simulation: A key to the future of medical education – a review. Adv Med Educ Pr 2022;13:301–8.
[8] Elendu C, et al. The impact of simulation-based training in medical education: A review. Medicine 2024;103(27).
[9] Abildgren L, et al. The effectiveness of improving healthcare teams' human factor skills using simulation-based training: a systematic review. Adv Simul 2022;7(1):12.
[10] Bienstock J, Heuer A. A review on the evolution of simulation-based training to help build a safer future. Medicine 2022;101(25).
[11] Halamek LP, et al. Time for a new paradigm in pediatric medical education: Teaching neonatal resuscitation in a simulated delivery room environment. Pediatrics 2000;106(4):e45.
[12] Ziv SDS, Wolpe Paul Root, Amitai. Patient safety and simulation-based medical education. Med Teach 2000;22(5):489–95.
[13] Dias AAL, Souza RS, Eduardo AHA, Felix AMS, Figueiredo RM. Validation of two clinical scenarios for simulation-based learning for the prevention and control of healthcare-associated infections. Rev Eletr Enferm 2022;24:70072.
[14] Cioffi J. Clinical simulations: development and validation. Nurse Educ Today 2001;21(6):477–86.
[15] Wass V, Van Der Vleuten C. The long case. Med Educ 2004;38(11):1176–80.
[16] Patrício MF, Miguel J, Filipa F, Carneiro AV. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? Med Teach 2013;35(6):503–14.
[17] Norman G. The long case versus objective structured clinical examinations. BMJ 2002;324(7340):748.
[18] Hodges B. Validity and the osce. Med Teach 2003;25(3):250–4.
[19] Al-Hashimi K, Said UN, Khan TN. Formative objective structured clinical examinations (osces) as an assessment tool in UK undergraduate medical education: a review of its utility. Cureus 2023;15(5).
[20] Hsieh K-S, Yang T-H, Hsu Y-L, Hsu K-C. Application of virtual reality in heart ultrasound education. In: 2024 IEEE 7th Eurasian conference on educational innovation. ECEI, IEEE; 2024, p. 1–3.
[21] Dang Z, Yang Q, Deng Z, Han J, He Y, Wang S. Digital twin-based skill training with a hands-on user interaction device to assist in manual and robotic ultrasound scanning. IEEE J Radio Freq Identif 2022;6:787–93.
[22] Graf L, Gradl-Dietsch G, Masuch M. Depressed virtual agents: development of a playful vr application for the training of child and adolescent psychiatry students. In: Proceedings of the 23rd ACM international conference on intelligent virtual agents. 2023, p. 1–3.
[23] Lerner D, Mohr S, Schild J, Göring M, Luiz T. An immersive multi-user virtual reality for emergency simulation training: Usability study. JMIR Serious Games 2020;8(3):e18822.
[24] Bowman DA, McMahan RP. Virtual reality: How much immersion is enough? Comput (Long Beach Calif) 2007;40(7):36–43.
[25] Dang BK, Palicte JS, Valdez A, O'Leary-Kelley C. Assessing simulation, virtual reality, and television modalities in clinical training. Clin Simul Nurs 2018;19:30–7.
[26] Chen S-Y, Wathen C, Speciale M. Online clinical training in the virtual remote environment: Challenges, opportunities, and solutions. Prof Couns 2020;10(1):78–91.
[27] Xiong J, Hsiang E-L, He Z, Zhan T, Wu S-T. Augmented reality and virtual reality displays: emerging technologies and future perspectives. Light Sci Appl 2021;10(1):1–30.

[28] Palmas F, Klinker G. Defining extended reality training: A long-term definition for all industries. In: 2020 IEEE 20th international conference on advanced learning technologies. ICALT, IEEE; 2020, p. 322–4.
[29] Chheang V, et al. Collaborative virtual reality for laparoscopic liver surgery training. In: 2019 IEEE international conference on artificial intelligence and virtual reality. AIVR, 2019, p. 1–17.
[30] Gil de Zúñiga H, Goyanes M, Durotoye T. A scholarly definition of artificial intelligence (ai): advancing ai as a conceptual framework in communication research. Polit Commun 2024;41(2):317–34.
[31] Turing AM. Mind. Mind 1950;59(236):433–60.
[32] Vaswani A, et al. Attention is all you need. Adv Neural Inf Process Syst 2017;30.
[33] Brown PF, Della Pietra VJ, Desouza PV, Lai JC, Mercer RL. Class-based n-gram models of natural language. Comput Linguist 1992;18(4):467–80.
[34] Costa VG, Pedreira CE. Recent advances in decision trees: An updated survey. Artif Intell Rev 2023;56(5):4765–800.
[35] Rabiner LR. A tutorial on hidden markov models and selected applications in speech recognition. Proc IEEE 1989;77(2):257–86.
[36] Hager Paul, Jungmann Friederike, Holland Robbie, Bhagat Kunal, Hubrecht Inga, Knauer Manuel, Vielhauer Jakob, Makowski Marcus, Braren Rickmer, Kaissis Georgios, Rückert Daniel. Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nature Med 2024;30(9):2613–22.
[37] Tam Thomas Yu Chow, Sivarajkumar Sonish, Kapoor Sumit, Stolyar Alisa V, Polanska Katelyn, McCarthy Karleigh R, Osterhoudt Hunter, Wu Xizhi, Visweswaran Shyam, Fu Sunyang, Mathur Piyush, Cacciamani Giovanni E, Sun Cong, Peng Yifan, Wang Yanshan. A framework for human evaluation of large language models in healthcare derived from literature review. Npj Digit Med 2024;7(1):258.
[38] Singhal Karan, Azizi Shekoofeh, Tu Tao, Sara Mahdavi S, Wei Jason, Chung Hyung Won, Scales Nathan, Tanwani Ajay, Cole-Lewis Heather, Pfohl Stephen, Payne Perry, Seneviratne Martin, Gamble Paul, Kelly Chris, Babiker Abubakr, Schärli Nathanael, et al. Large language models encode clinical knowledge. Nat 2023;620(7972):172–80.
[39] Seo Junhyuk, Choi Dasol, Kim Taerim, Cha Won Chul, Kim Minha, Yoo Haanju, Oh Namkee, Yi YongJin, Lee Kye Hwa, Choi Edward. Evaluation framework of large language models in medical documentation: Development and usability study. J Med Internet Res 2024;26:e58329.
[40] Collins Gary S, Moons Karel GM, Dhiman Paula, Riley Richard D, Beam Andrew L, Calster Ben Van, et al. Tripod+ai: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 2024;385:e078378.
[41] Tu Tao, Schaekermann Mike, Palepu Anil, Saab Khaled, Freyberg Jan, Tanno Ryutaro, Wang Amy, Li Brenna, Amin Mohamed, Cheng Yong, Vedadi Elahe, Tomasev Nenad, Azizi Shekoofeh, Singhal Karan, Hou Le, Webson Albert, Kulkarni Kavita, Sara Mahdavi S, Semturs Christopher, Gottweis Juraj, Barral Joelle, Chou Katherine, Corrado Greg S, Matias Yossi, Karthikesalingam Alan, Natarajan Vivek. Towards conversational diagnostic artificial intelligence. Nat 2025;642(8067):442–50.
[42] Kleine Anne-Kathrin, Kokje Eesha, Hummelsberger Pia, Lermer Eva, Schaffernak Insa, Gaube Susanne. Ai-enabled clinical decision support tools for mental healthcare: A product review. Artif Intell Med 2025;160:103052.
[43] Jones C, Bergen B. Does GPT-4 pass the Turing test? In: Proceedings of the 2024 conference of the North American chapter of the association for computational linguistics: human language technologies (volume 1: long papers). 2024, p. 5183–210.
[44] Logeswaran A, Munsch C, Chong YJ, Ralph N, McCrossnan J. The role of extended reality technology in healthcare education: Towards a learner-centred approach. Futur Heal J 2021;8(1):e79–84.
[45] Lungu AJ, Wout S, Luc C, Puxun T, Jan E, 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.
[46] Fleet A, Kaustov L, Belfiore EB, Kapralos B, Matava C, Wiegelmann J, Giacobbe P, Alam F. Current clinical and educational uses of immersive reality in anesthesia: Narrative review. J Med Internet Res 2025;27(1):e62785.
[47] Ogundiya O, Rahman T, Valnarov-Boulter I, Young T. Looking back on digital medical education over the last 25 years and looking to the future: Narrative review. J Med Internet Res 2024;26:e60312.
[48] Hirzle T, Müller F, Draxler F, Schmitz M, Knierim P, Hornbæk K. When xr and ai meet-a scoping review on extended reality and artificial intelligence. In: Proceedings of the 2023 CHI conference on human factors in computing systems. 2023, p. 1–45.
[49] Blackmore Karen L, Smith Shamus P, Bailey Jacqueline D, Krynski Benjamin. Integrating biofeedback and artificial intelligence into extended reality training scenarios: A systematic literature review. Simul Gaming 2024.
[50] Harmon J, Pitt V, Summons P, Inder KJ. Use of artificial intelligence and virtual reality within clinical simulation for nursing pain education: A scoping review. Nurse Educ Today 2021;97:104700.
[51] Vasey Baptiste, Nagendran Myura, Campbell Bruce, et al. Decide-ai: Reporting guideline for early-stage clinical evaluation of ai decision support. Nature Med 2022;28(5):924–33.
[52] Stahl Bernd Carsten, Antoniou Joseph, Bhalla Neha, et al. A systematic review of artificial intelligence impact assessments. Artif Intell Rev 2023;56:12799–831.
[53] Kaminski Margot E, Malgieri Gianclaudio. Algorithmic impact assessments under the gdpr: producing multi-layered explanations. Int Data Priv Law 2021;11(2):125–44.
[54] Huo Bright, Collins Gary, Chartash David, Thirunavukarasu Arun, Flanagin Annette, Iorio Alfonso, Cacciamani Giovanni, Chen Xi, Liu Nan, Mathur Piyush, Chan An-Wen, Laine Christine, Pacella Daniela, Berkwits Michael, Antoniou Stavros A, Camaradou Jennifer C, Canfield Carolyn, Mittelman Michael, Feeney Timothy, Loder Elizabeth, Agha Riaz, Saha Ashirbani, Mayol Julio, Sunjaya Anthony, Harvey Hugh, Ng Jeremy Y, McKechnie Tyler, Lee Yung, Verma Nipun, Stiglic Gregor, McCradden Melissa, Ramji Karim, Boudreau Vanessa, Ortenzi Monica, Meerpohl Joerg, Vandvik Per Olav, Agoritsas Thomas, Samuel Diana, Frankish Helen, Anderson Michael, Yao Xiaomei, Loeb Stacy, Lokker Cynthia, Liu Xiaoxuan, Guallar Eliseo, Guyatt Gordon. Reporting guideline for chatbot health advice studies: The chart statement. Artif Intell Med 2025;168:103222.
[55] Bottrighi Alessio, Maconi Antonio, Nera Stefano, Piovesan Luca, Raina Erica, Terenziani Paolo. Ontology-based student testing through clinical guidelines: An ai approach. Artif Intell Med 2025;168:103226.
[56] Page Matthew J, McKenzie Joanne E, Bossuyt Patrick M, et al. The prisma 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71.
[57] Goh P-S, Sandars J. A vision of the use of technology in medical education after the COVID-19 pandemic. MedEdPublish 2020;9:49.
[58] Hamilton A. Artificial intelligence and healthcare simulation: the shifting landscape of medical education. Cureus 2024;16(5).
[59] Winkler-Schwartz A, et al. Artificial intelligence in medical education: Best practices using machine learning to assess surgical expertise in virtual reality simulation. J Surg Educ 2019;76(6):1681–90.
[60] Pretolesi Daniele, Zechner Olivia, Zafari Setareh, Tscheligi Manfred. Human in the loop for xr training: Theory, practice and recommendations for effective and safe training environments. In: 2023 IEEE international conference on metrology for eXtended reality, artificial intelligence and neural engineering. MetroXRAINE, 2023.
[61] Graf Linda, Sykownik Philipp, Gradl-Dietsch Gertraud, Masuch Maic. Towards believable and educational conversations with virtual patients. Front Virtual Real 2024;5.
[62] Woo J, Shidara K, Achard C, Tanaka H, Nakamura S, Pelachaud C. Adaptive virtual agent: Design and evaluation for real-time human-agent interaction. Int J Hum Comput Stud 2024;103321.
[63] Liaw Sok Ying, Tan Jian Zhi, Lim Siriwan, Zhou Wentao, Yap John, Ratan Rabindra, Ooi Sim Leng, Wong Shu Jing, Seah Betsy, Chua Wei Ling. Artificial intelligence in virtual reality simulation for interprofessional communication training: Mixed method study. Nurse Educ Today 2023.
[64] Recai Yilmaz MD, Ali M. Fazlollahi MSc, Alexander Winkler-Schwartz MD, PhD, Anna Wang BSc, Hafila Hassan Makhani BSc, Ahmad Alsayegh MBBS, MSc, Mohamad Bakhaidar MBBS, MSc, Dan Huy Tran MSc, Carlo Santaguida MD, Rolando F. Del Maestro MD, PhD. Effect of feedback modality on simulated surgical skills learning using automated educational systems: A four-arm randomized control trial. J Surg Educ 2023.
[65] Liaw Sok Ying, Ooi Sim Win, Rusli Khairul Dzakirin Bin, Lau Tang Ching, Tam Wilson Wai San, Chua Wei Ling. Nurse-physician communication team training in virtual reality versus live simulations: Randomized controlled trial on team communication and teamwork attitudes. J Med Internet Res 2020.
[66] Vannaprathip Narumol, Haddawy Peter, Schultheis Holger, Suebnukarn Siriwan. Sdmentor: A virtual reality-based intelligent tutoring system for surgical decision making in dentistry. Artif Intell Med 2025.
[67] Agbontaen Kaladerhan O, Cold Kristoffer M, Woods David, Grover Vimal, Aboumarie Hatem Soliman, Kaul Sundeep, Konge Lars, Singh Suveer. Artificial intelligence-guided bronchoscopy is superior to human expert instruction for the performance of critical-care physicians: A randomized controlled trial. Crit Care Med 2025.
[68] Long Yonghao, Cao Jianfeng, Deguet Anton, Taylor Russell H, Dou Qi. Integrating artificial intelligence and augmented reality in robotic surgery: An initial dvrk study using a surgical education scenario. In: 2022 international symposium on medical robotics. ISMR, 2022.
[69] Nam Hyeongil. Multimodal vr/mr simulation with empathic agents for nursing education. In: 2024 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops. VRW, 2024.
[70] Teixeira Luis, Mitchell Aby, Martinez Neus Carlos, Salim Behnam Jafari. Virtual reality with artificial intelligence-led scenarios in nursing education: a project evaluation. Br J Nurs 2024.
[71] Osborne Matthew, Truman Cheyne, Sheridan Sheridan. Perspectives: row a's nursing education is here today: experiences of artificial intelligence and virtual reality technologies. J Res Nurs 2024.
[72] Ghaempanah F, et al. Metaverse and its impact on medical education and health care system: A narrative review. Heal Sci Rep 2024;7(9):e70100.
[73] Mekki Yosra Magdi, Simon Leslie V, Freeman William D, Qadir Junaid. Medical education metaverses (meded metaverses): Opportunities, use case, and guidelines. COMPUTER 2025.

[74] Chaddad Ahmad, Jiang Yuchen. Integrating technologies in the metaverse for enhanced healthcare and medical education. IEEE Trans Learn Technol 2025.
[75] Harari Ryan, Shokoohi Hamid, Al-Taweel Abdullah, Ahram Tareq. Evaluating clinician trust and performance with explainable ai and augmented reality in transesophageal echocardiography (tee) imaging. In: 2024 IEEE international conference on artificial intelligence and eXtended and virtual reality. AIxVR, 2024.
[76] Anetzberger Hermann, Kugler Andreas, Mohr Michael, Haasters Florian, Reppenhagen Stephan, Becker Roland. First validated automated scoring system using the diagnostic arthroscopy skill score (dass 2.0) for assessing proficiency in virtual reality arthroscopy. Knee Surg Sport Traumatol Arthrosc 2025.
[77] Belmar Francisca, Gaete María Inés, Escalona Gabriel, Carnier Martín, Durán Valentina, Villagrán Ignacio, Asbun Domenech, Cortés Matías, Neyem Andrés, Crovari Fernando, Alseidi Adnan, Varas Julián. Artificial intelligence in laparoscopic simulation: a promising future for large-scale automated evaluations. Surg Endosc 2022.
[78] Stanney Kay M, Archer JoAnn, Skinner Anna, Horner Charis, Hughes Claire, Brawand Nicholas P, Martin Eric, Sanchez Stacey, Moralez Larry, Fidopiastis Cali M, Perez Ray S. Performance gains from adaptive extended reality training fueled by artificial intelligence. J Déf Model Simul: Appl Methodol Technol 2021.
[79] Huang Siyu, Wen Chang, Bai Xueying, Li Sihong, Wang Shuining, Wang Xiaoxuan, Yang Dong. Exploring the application capability of chatgpt as an instructor in skills education for dental medical students: Randomized controlled trial. J Med Internet Res 2025.
[80] Hernandez Olivia K, Sushereba Christen, Militello Laura, Miguel Christopher San, Wolf Steve, Allen Theodore T, Patterson Emily S. Strategies for case-based training with virtual patients: An experimental study of the impact of integrating mental model articulation and self-reflection. Appl Ergon 2024.
[81] Vuthea Chheang, Sharmin Shayla, Màrquez-Hernàndez Rommy, Patel Megha, Rajasekaran Danush, Caulfield Gavin, Kiafar Behdokht, Li Jicheng, Kullu Pinar, Barmaki Roghayeh Leila. Towards anatomy education with generative ai-based virtual assistants in immersive virtual reality environments. In: 2024 IEEE international conference on artificial intelligence and eXtended and virtual reality. AIxVR, 2024.
[82] Boal MWE, et al. Evaluation of objective tools and artificial intelligence in robotic surgery technical skills assessment: a systematic review. Br J Surg 2024;111(1).
[83] Chatha WA. From scalpel to simulation: Reviewing the future of cadaveric dissection in the upcoming era of virtual and augmented reality and artificial intelligence. Cureus 2024;16(10).
[84] Varas J, et al. Innovations in surgical training: exploring the role of artificial intelligence and large language models (llm). Rev Col Bras Cir 2023;50:e20233605.
[85] Liu W, et al. Self-guided dmt: Exploring a novel paradigm of dance movement therapy in mixed reality for children with ASD. IEEE Trans Vis Comput Graphics 2024.
[86] Ebina K, et al. Development of machine learning-based assessment system for laparoscopic surgical skills using motion-capture. In: 2024 IEEE/sICE international symposium on system integration. SII, IEEE; 2024, p. 1–6.
[87] Gupta I, Gupta I, Diwakar M, Arya C, Singh P, Pandey NK. A review of augmented reality (ar) and virtual reality (vr) applications in the digital healthcare training and education sector. In: 2024 second international conference computational and characterization techniques in engineering & sciences. IC3TES, IEEE; 2024, p. 1–7.
[88] Fraser KL, Ayres P, Sweller J. Cognitive load theory for the design of medical simulations. Simul Heal: J Soc Simul Heal 2015;10(5):295–307.
[89] Reedy Gabriel B. Using cognitive load theory to inform simulation design and practice. Clin Simul Nurs 2015;11(8):355–60.
[90] Haryana Muhammad Roy Aziz, Warsono Sony, Achjari Didi, Nahartyo Ertambang. Virtual reality learning media with innovative learning materials to enhance individual learning outcomes based on cognitive load theory. Int J Manag Educ 2022;20(3):100657.
[91] Plass JL, Kalyuga S. Four ways of considering emotion in cognitive load theory. Educ Psychol Rev 2019;31(2):339–59.
[92] Fraser K, Ma I, Teteris E, Baxter H, Wright B, McLaughlin K. Emotion, cognitive load and learning outcomes during simulation training. Med Educ 2012;46(11):1055–62.
[93] Skulmowski A, Xu KM. Understanding cognitive load in digital and online learning: a new perspective on extraneous cognitive load. Educ Psychol Rev 2022;34(1):171–96.
[94] Tzafilkou K, Perifanou M, Economides AA. Negative emotions, cognitive load, acceptance, and self-perceived learning outcome in emergency remote education during COVID-19. Educ Inf Technol (Dordr) 2021;26(6):7497–521.
[95] Hashim NC, Abd Majid NA, Arshad H, Hashim H, Alyasseri ZAA. Mobile augmented reality based on multimodal inputs for experiential learning. IEEE Access 2022;10:78953–69.
[96] Daher and S, et al. Matching vs. non-matching visuals and shape for embodied virtual healthcare agents. In: 2019 IEEE conference on virtual reality and 3D user interfaces. VR, IEEE; 2019, p. 886–7.
[97] Latour M, Alvarez I, Knackstedt M, Yim M. Virtually augmented surgical navigation in endoscopic sinus surgery simulation training: A prospective trial of repeated measures. Otolaryngology–Head Neck Surg 2024;171(6):1897–903.
[98] Mergen M, Junga A, Risse B, Valkov D, Graf N, Marschall B. Immersive training of clinical decision making with ai-driven virtual patients: a new vr platform called medical tr.ai.ning. GMS J Med Educ 2023;40(2):Doc19.
[99] Lee DK, Choi H, Jheon S, Jo YH, Im CW, Il SY. Development of an extended reality simulator for basic life support training. IEEE J Transl Eng Heal Med 2022;10:1–7.
[100] Skulmowski Alexander, Xu Kate Man. Understanding cognitive load in digital and online learning: a new perspective on extraneous cognitive load. Educ Psychol Rev 2022;34:171–96.
[101] Albus Patrick, Vogt Andrea, Seufert Tina. Signaling in virtual reality influences learning outcome and cognitive load. Comput Educ 2021;166:104154.
[102] Kim Kisoo, Yang Hyosill, Lee Jihun, Lee Won Gu. Metaverse wearables for immersive digital healthcare: A review. Adv Sci 2023.
[103] Masuda Mana, Hachiuma Ryo, Saito Hideo, Kajita Hiroki, Takatsume Yoshifumi. Os-nerf: Generalizable novel view synthesis for occluded open-surgical scenes. In: 2024 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops. VRW, 2024.
[104] Rubin John E, Pandian Balaji, Jotwani Rohan, Pryor Kane O, Rubin Lori A, Mack Patricia F. Leveraging spatial computing to improve crisis management training in anesthesiology. J Clin Anesth 2024.
[105] Gruenewald Armin, Schmidt Ricardo, Sayn Lukas, Gießer Christian, Eiler Tanja Joan, Schmuecker Vanessa, Braun Veit, Brueck Rainer. Virtual reality training application to prepare medical student's for their first operating room experience. In: 2021 IEEE international conference on artificial intelligence and virtual reality. AIVR, 2021.
[106] Maimaiti Zulipikaer, Li Zhuo, Li Zhiyuan, Fu Jun, Xu Chi, Chen Jiying, Chai Wei, Liu Liang. Ortho-digital dynamics: Exploration of advancing digital health technologies in musculoskeletal disease management. Digit Heal 2024.
[107] Sherif Meriam, Barakat Nahla, Hamdy Abeer. Adaptvr: An adaptive learning vr system for dental students training. In: 2024 international conference on computer and applications. ICCA, 2024.
[108] Jallad Samar Thabet, Alsaqer Khitam, Albadareen Baker Ishaq, Almaghaireh Duaa. Artificial intelligence tools utilized in nursing education: Incidence and associated factors. Nurse Educ Today 2024.
[109] RUMO Jabulani Tshepo, ISAFIADE Omowunmi Elizabeth, KOUMOU Kessel Okinga, EGIEYEH Samuel. Digidosear: Redefining the future of pharmacy education with augmented reality. In: IST-africa 2025 conference proceedings. 2025.
[110] Qi Yuxiao, Yang Yancheng, Lu Xiaofeng. Design and implementation of an intelligent simulation system for laparoscopic surgery standardized training. In: 2024 5th international symposium on artificial intelligence for medicine science. ISAIMS 2024, 2024.
[111] Ghaempanah Faezeh, Ghafari Bahar Moasses, Hesami Darya, Zadeh Reza Hossein, Noroozpoor Rashin, Ghalibaf AmirAli Moodi, Hasanabadi Parsa. Metaverse and its impact on medical education and health care system: A narrative review. Heal Sci Rep 2024.
[112] Schwind Valentin, Tadesse Netsanet Zelalem, Cunha Estefania Silva Da, Hamidi Yeganeh, Sultani Soltan Sanjar, Sehrt Jessica. A scoping review of informed consent practices in human–computer interaction research. ACM Trans Comput-Human Interact 2025.
[113] Elbasi Ersin, Nadeem Muhammad, Alzoubi Yehia Ibrahim, Topcu Ahmet E, Varghese Greeshma. Machine learning in education: Innovations, impacts, and ethical considerations. IEEE Access 2025.
[114] Bogar Peter Zoltan, Virag Mark, Bene Matyas, Hardi Peter, Matuz Andras, Schlegl Adam Tibor, Toth Luca, Molnar Ferenc, Nagy Balint, Rendeki Szilard, Berner-Juhos Krisztina, Ferencz Andrea, Fischer Krisztina, Maroti Peter. Validation of a novel, low-fidelity virtual reality simulator and an artificial intelligence assessment approach for peg transfer laparoscopic training. Sci Rep 2024.
[115] Cha Guoping. Cloud computing intelligence based framework for adaptive nursing education with artificial intelligence monitoring and blockchain certification. In: 2025 3rd international conference on data science and information system. ICDSIS, 2025.
[116] Otoum Yazan, Gottimukkala Navya, Kumar Neeraj, Nayak Amiya. Machine learning in metaverse security: Current solutions and future challenges. ACM Comput Surv 2024.
[117] Qayyum Adnan, Bilal Muhammad, Hadi Muhammad, Capik Pawe, Caputo Massimo, Vohra Hunaid, Al-Fuqaha Ala, Qadir Junaid. Can we revitalize interventional healthcare with ai-xr surgical metaverses? In: 2023 IEEE international conference on metaverse computing, networking and applications. MetaCom, 2023.
[118] Long HA, French DP, Brooks JM. Optimising the value of the critical appraisal skills programme (casp) tool for quality appraisal in qualitative evidence synthesis. Res Methods Med Heal Sci 2020;1(1):31–42.
[119] Chomutare Taridzo, Tejedor Marco, Svenning Tor Ole, Marco-Ruiz Luis, Tayefi Mohammad, Lind Kristin, Godtliebsen Fred, Moen Anne, Ismail Lamia, Makhlysheva Alena, Ngo Phuoc Duy. Artificial intelligence implementation in healthcare: A theory-based scoping review of barriers and facilitators. Int J Environ Res Public Heal 2022;19(23):16359.

[120] Finkelstein Joseph, Gabriel Ariella, Schmer Sam, et al. Identifying facilitators and barriers to implementation of ai-assisted clinical decision support in an electronic health record system. J Med Syst 2024;48:89.
[121] Carroll Stephanie R, Garba Ibrahim, Figueroa-Rodríguez Olga L, Holbrook Jarita, Lovett Ray, Materechera Solomon, Hudson Maui. The care principles for indigenous data governance. In: Open scholarship press curated volumes: policy. 2023, Retrieved from Open Scholarship Press.
[122] Gupta A, Nisar H. An improved framework to assess the evaluation of extended reality healthcare simulators using machine learning. In: 2022 IEEE/ACM conference on connected health: applications, systems and engineering technologies. CHASE, Arlington, VA, USA: IEEE; 2022, p. 188–92.
[123] Hu Guilin, Qiu Minghua. Machine learning-assisted structure annotation of natural products based on ms and nmr data. Nat Prod Rep 2023;40(11):1735–53.
[124] Meta quest 3 512 gb — meta product page (£469). 2024, [Accessed 12 August 2025].
[125] Microsoft hololens 2 enterprise price announced at $3,500. 2019, [Accessed 12 August 2025].
[126] Almarzouqi A, Aburayya A, Salloum SA. Prediction of user's intention to use metaverse system in medical education: A hybrid sem-ml learning approach. IEEE Access 2022;10:43421–34.
[127] Shardeo V, Sarkar BD, Mir UB, Kaushik P. Adoption of metaverse in healthcare sector: an empirical analysis of its enablers. IEEE Trans Eng Manag 2024.
[128] Morimoto T, H. H, M. U, N. F, T. S, M. S, T. K, M. T, T. Y, Y. T, Morimoto. Digital transformation will change medical education and rehabilitation in spine surgery. Medicina (B Aires) 2022.
[129] Tolentino R, Rodriguez C, Hersson-Edery F, Lane J, Abbasgholizadeh Rahimi S. Perspectives on virtual interviews and emerging technologies integration in family medicine residency programs: a cross-sectional survey study. BMC Med Educ 2024;24(1):975.
[130] Guckert M, et al. The disruption of trust in the digital transformation leading to health 4.0. Front Digit Heal 2022;4:815573.
[131] Lastrucci A, Giansanti D. Radiological crossroads: Navigating the intersection of virtual reality and digital radiology through a comprehensive narrative review of reviews. Robotics 2024;13(5):69.
[132] Nguyen H-S, Voznak M. A bibliometric analysis of technology in digital health: Exploring health metaverse and visualizing emerging healthcare management trends. IEEE Access 2024;12:23887–913.
[133] Chengoden R, et al. Metaverse for healthcare: a survey on potential applications, challenges and future directions. IEEE Access 2023;11:12765–95.
[134] Lewandrowski K-U, et al. The changing environment in postgraduate education in orthopedic surgery and neurosurgery and its impact on technology-driven targeted interventional and surgical pain management: Perspectives from Europe, Latin America, Asia, and the United States. J Pers Med 2023;13(5):852.
[135] Bhatia B, Joshi S. Applications of metaverse in the healthcare industry. In: 2023 international conference on innovative data communication technologies and application. ICIDCA, IEEE; 2023, p. 344–50.
[136] Alrashed and FA, et al. Incorporating technology adoption in medical education: a qualitative study of medical students' perspectives. Adv Med Educ Pr 2024;615–25.
[137] Hernandez OK, Sushereba C, Militello L, San Miguel C, Wolf S, Allen TT, Patterson ES. Strategies for case-based training with virtual patients: An experimental study of the impact of integrating mental model articulation and self-reflection. Appl Ergon 2024;118:104265.
[138] Bissonnette V, Mirchi N, Ledwos N, Alsidieri G, Winkler-Schwartz A, Del Maestro RF. Artificial intelligence distinguishes surgical training levels in a virtual reality spinal task. JBJS 2019;101(23):e127.
[139] Lu Haohui, Lin Ye, Li Zhidong, Yiu Man Lung, Gao Yu, Uddin Shahadat. Toward fair medical advice: Addressing and mitigating bias in large language model-based healthcare applications. Artif Intell Med 2025;168:103216.
[140] Raman R, Hughes L, Mandal S, Das P, Nedungadi P. Mapping metaverse research to the sustainable development goal of good health and well-being. IEEE Access 2024;12:180631–51.
[141] Gordon M, et al. A scoping review of artificial intelligence in medical education: Beme guide no. 84. Med Teach 2024;46(4):446–70.
[142] Nedbal C, et al. The role of 'artificial intelligence, machine learning, virtual reality, and radiomics' in pcnl: A review of publication trends over the last 30 years. Ther Adv Urol 2023;15:17562872231196676.
[143] Dicheva NK, Rehman IU, Anwar A, Nasralla MM, Husamaldin L, Aleshaiker S. Digital transformation in nursing education: A systematic review on computer-aided nursing education pedagogies, recent advancements and outlook on the post-covid-19 era. IEEE Access 2023;11:135659–95.
[144] Iqbal and J, et al. Reimagining healthcare: unleashing the power of artificial intelligence in medicine. Cureus 2023;15(9).
[145] Singaram VS, Pillay R, Mbobnda Kapche EL. Exploring the role of digital technology for feedback exchange in clinical training: a scoping review. Syst Rev 2024;13(1):1–28.
[146] Metcalf Jacob, Moss Emanuel, Watkins Elizabeth Anne, Singh Ranjit, Elish Madeleine Clare. Algorithmic impact assessments and accountability: The co-construction of impacts. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. FAccT '21, New York, NY, USA: Association for Computing Machinery; 2021, p. 735–46.
[147] Teixeira L, Mitchell A, Martinez NC, Salim BJ. Virtual reality with artificial intelligence-led scenarios in nursing education: a project evaluation. Br J Nurs 2024;33(17):812–20.
[148] Stanney KM, et al. Performance gains from adaptive extended reality training fueled by artificial intelligence. J Déf Model Simul 2022;19(2):195–218.
[149] Ahuja AS, Polascik BW, Doddapaneni D, Byrnes ES, Sridhar J. The digital metaverse: Applications in artificial intelligence, medical education, and integrative health. Integr Med Res 2023;12(1):100917.
[150] Votintseva A, Johnson R. Lemino–let me know how to negotiate: Virtual simulator for negotiation training. In: 2024 international conference on smart computing, IoT and machine learning. SIML, IEEE; 2024, p. 335–40.