

REVIEW



Augmented reality simulation for surgical skill retention and operative transfer: a systematic review

Ee Hng Ian Lim¹ · Aaleyah Adam¹ · Athanasios Hassoulas¹

Received: 14 August 2026 / Revised: 14 August 2026 / Accepted: 17 September 2026
© The Author(s) 2026

Abstract

Purpose Simulation-based training is embedded in surgical curricula. Augmented reality (AR) overlays digital guidance onto real-world training environments while preserving haptic feedback, and recent systematic reviews have established that AR improves technical performance during training. None of these reviews, however, has formally assessed whether AR-acquired skill is retained over time or transfers to intraoperative or high-fidelity ex-vivo performance. This review aimed to systematically synthesise and appraise the evidence on skill retention and transfer of AR-acquired surgical skills to intraoperative or high-fidelity ex-vivo performance.

Methods PubMed/MEDLINE, CINAHL, EMBASE, and Cochrane Library were searched from January 2000 to March 2026. Eligible studies used validated tools (OSATS, GOALS, GEARS, or equivalent) to assess AR skill retention beyond the immediate training session, or transfer to intraoperative or ex-vivo performance, in surgical trainees of any specialty. Two reviewers independently screened records and extracted data; quality and risk of bias were assessed using MERSQI and Cochrane RoB2, with narrative synthesis given the anticipated heterogeneity. This review followed PRISMA 2020 guidance and is registered on PROSPERO (CRD420261347095).

Results Of 304 records screened, 18 underwent full-text review and 9 were included (324 participants; laparoscopic, urological, neurosurgical, and orthopaedic specialties). Skill retention was maintained for up to 8 weeks, though gains were outcome-dependent. Five studies assessed high-fidelity ex-vivo transfer and favoured AR on validated instruments; the sole intraoperative transfer study found no significant difference on primary outcomes despite fewer trainer corrections in the AR group. MERSQI scores ranged from 13.5 to 16.5 (all high quality); RoB2 showed some concerns in 7 studies and high risk in 2, predominantly from unblinded outcome assessment.

Conclusions AR simulation shows technical benefits and short-term ex-vivo retention, but intraoperative transfer evidence rests on a single, constrained study, and no included study measured effect sizes or clinical outcomes. This small, heterogeneous evidence base, distinct from the technical-performance literature already synthesised elsewhere, warrants adequately powered, blinded intraoperative transfer trials before curricular implementation.

Keywords Augmented reality · Simulation training · Clinical competence · Surgical education · Skill retention · Skill transfer

Introduction

Surgical education has moved away from the “see one, do one, teach one” model toward structured, simulation-based curricula. The shift reflects restricted duty hours, patient safety concerns, and evidence that deliberate practice accelerates skill acquisition [1, 2]. Virtual reality (VR) and high-fidelity ex-vivo simulators have since been adopted across specialties, supported by strong evidence for early skills acquisition [3].

✉ Ee Hng Ian Lim
ian555ad@gmail.com

¹ Cardiff University School of Medicine, University Hospital of Wales (UHW), Main Building, Heath Park, Cardiff CF14 4XN, UK

Augmented reality (AR) differs from VR in one important respect: it overlays digital cues onto a trainee's real-world field of view rather than replacing that view, preserving haptic feedback throughout. Head-mounted displays and projector-based overlays have been investigated for their effects on spatial understanding, cognitive load, and early-stage skill acquisition [4], but evidence for AR's capacity to bridge early acquisition and clinical application remains fragmented across heterogeneous study designs, procedures, and outcome measures [5].

A central challenge for any simulation modality is distinguishing performance achieved during training from durable, transferable competence [6]. Retention, the persistence of a skill without continued practice, and transfer, whether an AR-acquired skill carries over into operative or high-fidelity ex-vivo settings, are the two outcomes that actually justify curricular investment [7]. Despite a growing number of AR training studies, this literature has not been synthesised with a specific focus on these two outcomes, and heterogeneity in follow-up intervals, outcome measures, and study design has made it difficult to draw firm conclusions [8].

That gap has, if anything, become more visible over the past year. El Ashry et al. [9] pooled eleven studies across seven surgical specialties and reported consistent AR advantages in technical performance, but none of their included studies measured retention or operative transfer. Xiong et al. [10], restricting their meta-analysis to basic laparoscopic skills, reached a similar endpoint and called explicitly for longer follow-up. Farooq et al. [11], synthesising AR telestration trials specifically, pooled task completion time and error rate but likewise flagged long-term retention as unmeasured. All three reviews, in other words, answer a version of the question "does AR improve performance during training" and leave the harder one, whether that performance survives time and the transition to the operating room, open. A related scoping review [7] mapped the broader AR implementation landscape, including retention and equity considerations in low-resource settings, and identified durability and clinical transferability as the area most in need of a quality-appraised, risk-of-bias-assessed synthesis.

This review addresses that gap directly. Rather than a further audit of whether AR improves training-session performance, a question the literature above has already answered with reasonable consistency, it restricts inclusion to studies reporting retention beyond the immediate training session, or transfer to live intraoperative or high-fidelity ex-vivo performance, and appraises each using PRISMA 2020 methodology, MERSQI, and formal risk-of-bias assessment.

Methods

Search strategy

This review followed PRISMA 2020 guidance [12] and was prospectively registered on PROSPERO (CRD420261347095). No funding was received and no conflicts of interest were declared.

Protocol amendments broadened the retention interval to any delayed post-training assessment, and redefined transfer to include both in vivo (operating theatre) and ex vivo (high-fidelity simulation or cadaveric) performance.

PubMed/MEDLINE, EMBASE, Cochrane Library, and CINAHL were searched from January 2000 to March 2026, combining MeSH headings and free-text terms across three concept blocks: augmented or mixed reality and AR simulation; surgical training and education; and skill retention, transfer, and validated outcome measures (OSATS, GOALS, GEARS). Reference lists of included studies and related reviews were screened manually for additional records.

Eligibility criteria

Eligible studies evaluated an AR-based simulation intervention and reported skill retention beyond the immediate training session, or transfer to live intraoperative or high-fidelity ex-vivo performance, in surgical trainees of any specialty, from medical students to fellows, using validated tools (OSATS, GEARS, GOALS, or equivalent). Randomised and non-randomised comparative studies, cohort studies, and single-arm pre-post studies were eligible. Studies of purely virtual reality systems without a physical component, non-surgical disciplines, subjective-only outcomes such as satisfaction or confidence, and grey literature or abstract-only records were excluded.

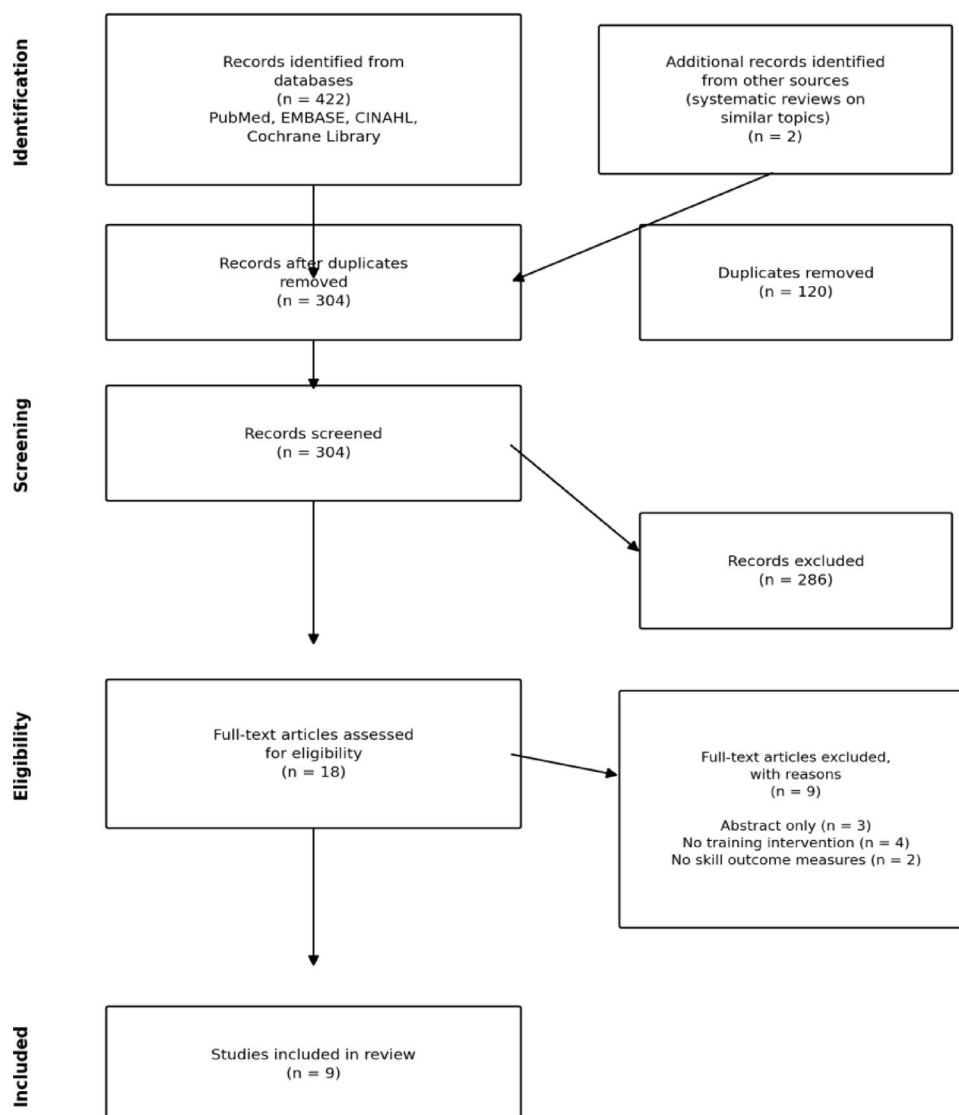
Screening and study selection

Records were exported in .ris format into Rayyan for deduplication. Two independent reviewers screened titles and abstracts, then full texts, against the eligibility criteria; disagreements were resolved by discussion, with a third reviewer available for arbitration (Fig. 1).

Data extraction and analysis

A piloted extraction form recorded study design, year, country, specialty, AR platform and fidelity, training protocol, comparator, assessment tool and validity evidence, follow-up interval, and outcome data. Outcomes were categorised as skill retention or intraoperative/ex-vivo transfer.

Fig. 1 PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram of study identification, screening, and inclusion. CINAHL = Cumulative Index to Nursing and Allied Health Literature



Quality and risk of bias assessment

Methodological quality was assessed using MERSQI, a validated 10-item, eight-domain instrument classifying studies as low (0–6), medium (7–12), or high (13–18) quality [13]. Risk of bias in randomised trials was assessed using Cochrane RoB2 across five domains, randomisation, deviations from intended interventions, missing data, outcome measurement, and selective reporting, each rated low risk, some concerns, or high risk, with an overall study-level judgement [14]. Domain 4 received particular attention, given how susceptible observer-rated performance assessments are to bias without formal blinding. Both instruments were applied independently by two reviewers, with disagreements resolved through discussion. Non-randomised studies were assessed using MERSQI alone, per protocol.

Use of artificial intelligence

This manuscript was written entirely by the authors. A large language model was used in an advisory capacity to suggest condensation and restructuring strategies during revision; no AI-generated text was incorporated into the manuscript, and all tables were formatted manually.

Results

Study selection and characteristics

The search identified 424 records; after removing 120 duplicates, 304 were screened, 18 underwent full-text review, and 9 studies were included.

Nine studies (2012–2025) enrolled 324 participants across four specialties: laparoscopic/visceral surgery ($n = 5$),

urology ($n = 2$), orthopaedics ($n = 1$), and neurosurgery ($n = 1$). All were prospective comparative designs, seven RCTs, one stratified three-arm RCT [15], and one multicentre two-phase RCT [16]. Sample sizes ranged from 9 to 60, with participants having no or limited prior procedural experience. AR platforms were heterogeneous: iSurgeon was used in three studies [17–19], with each remaining study using a distinct platform. Comparators were predominantly verbal or video-based instruction, with one study [20] including an additional VR arm (Table 1).

Skill retention outcomes

Four studies assessed performance beyond the immediate post-training session, with retention intervals ranging from 1 to 8 weeks (Table 2). Shepard et al. [23] reported the strongest evidence for durable retention: checklist scores at 8 weeks improved significantly from baseline at both the lower pole (45.5–66.0%; $p = 0.006$) and upper pole (39.6–55.2%; $p = 0.020$) access sites, with only modest, non-significant decline from post-test to retention at either site ($p = 0.057$ and $p = 0.150$). Brinkman et al. [21] found

Table 1 Characteristics of studies assessing skill retention or intraoperative/ex-vivo transfer in augmented reality simulation training

Author (year)	Specialty/procedure	<i>n</i>	Participants	Design	AR platform	Comparator	Primary outcome
Sugand et al. [20]	Orthopaedic surgery; DHS guidewire insertion	45	Medical students	Single-blinded RCT	FluoroSim	Minimal-exposure control; TraumaVision VR	Simulator performance (TAD, cut-out rate, time, fluoroscopic images, retries)
Felinska et al. [17]	Visceral/laparoscopic surgery; cholecystectomy	40	Medical students (laparoscopically naïve)	Randomized crossover trial	iSurgeon (RGB-D AR telestration)	Verbal instruction only (box trainer)	Ex vivo transfer performance (Global and task-specific OSATS; task duration)
Brinkman et al. [21]	Laparoscopic surgery (basic skills: clipping, grasping, cutting, object positioning, sharp dissection)	34	Medical interns (pre-residency; novice)	RCT	Haptica ProMIS (transfer assessment); Simbionix LAP Mentor II (VR, primary training)	Fixed-duration vs. criterion-based training endpoint	Simulator performance metrics (time, path length, economy of movement, accuracy); 1-week retention
Cizmic et al. [18]	General/hepatobiliary surgery (laparoscopic cholecystectomy)	40	Medical students (years 3–6; novice)	RCT	iSurgeon (real-time AR tutor hand-gesture overlay); ex vivo porcine liver box trainer	Verbal-only tutoring	Training performance across 10 sessions (OSATS, GOALS, operative time, complication rates, CVS)
Wild et al. [19]	Laparoscopic surgery (LC)	60	Medical students; laparoscopic novices	Randomized crossover RCT	iSurgeon (AR telestration)	Verbal instruction only	GOALS; OSATS (plus complication rate as secondary outcome)
Huetttl et al. [22]	Laparoscopic cholecystectomy (live surgery)	32	Surgical trainees (PGY-2–PGY-6) and fellows	Prospective single-centre RCT	HoloPointer (HoloLens AR intraoperative annotation)	Standard intraoperative teaching (verbal/gestural)	GOALS; CVS
Chowriappa et al. [16]	Robot-assisted urology (urethrovesical anastomosis)	52	Residents and fellows (low robotic experience)	Multicentre 2-phase RCT	HoST (AR-guided RoSS simulator training)	Video-based instruction	GEARS; procedure-specific UVA score
Shepard et al. [23]	Urology (ultrasound-guided PCNL)	12	Urology residents (PGY1–PGY6)	Randomised crossover trial	Vuzix M4000 smart glasses + HelpLightning MR remote guidance	In-person expert instruction	20-item checklist-based procedural score + procedure time; retention at 2 months
Dodier et al. [15]	Neurosurgery (aneurysm clipping simulation)	9	Neurosurgery residents (PGY1–PGY6)	Prospective stratified 3-arm RCT	HoloLens 1 + SOFA-based FEM patient-specific AR simulation	Video review; no additional training (control)	Clip selection accuracy; occlusion success (Raymond-Roy Class I); procedural efficiency (time, attempts); tremor-based dexterity

ARaugmented reality, DHSdynamic hip screw, TADtip–apex distance, VRvirtual reality, RGB-Dred-green-blue depth, OSATSObjective Structured Assessment of Technical Skills, RCTrandomized controlled trial, LClaparoscopic cholecystectomy, GOALSGlobal Operative Assessment of Laparoscopic Skills, CVScritical view of safety, PGYpostgraduate year, GEARSGlobal Evaluative Assessment of Robotic Skills, UVAurethrovesical anastomosis, PCNLpercutaneous nephrolithotomy, MRmixed reality, SOFASimulation Open Framework Architecture, FEMfinite element method

Table 2 Summary of outcome findings—skill retention and transfer of AR-acquired procedural skills

Study	Specialty/Setting	Interval/ Training Detail	Assessment Tool(s)	Group vs. Comparator	Summary of Finding
A. Skill Retention					
Shepard et al. [23]	Urology PCNL	8 weeks	Procedure checklist score	MR vs. in-person	Retention significantly above baseline; slight non-significant decline from post-test
Brinkman et al. [21]	Laparoscopy	1 week	Simulator performance composite	VR vs. AR training	Skills maintained at 1 week; no between-group differences
Dodier et al. [15]	Neurosurgery clipping	2 weeks	Procedural performance composite	AR vs. control	Retention improvement in AR group; process metrics unchanged
Sugand et al. [20]	Orthopaedics DHS fixation	1 week (washout)	Procedural performance composite	AR vs. control	AR group retained improved performance; controls did not
B. Intraoperative Transfer					
Huettl et al. [22]	Live LC; operating theatre	—	GOALS (5 domains); CVS achievement rate	AR (HoloPointer) vs. conventional intraoperative instruction	No significant between-group difference in operative performance (GOALS or CVS); AR group received fewer trainer corrections but trainer-rated assistance quality was significantly worse
C. Ex vivo/high-fidelity simulation transfer					
Felinska et al. [17]	Porcine cadaveric specimen; box trainer	—	OSATS (global + task-specific)	AR (iSurgeon) vs. verbal instruction only	Significantly higher OSATS scores in AR group on both global and task-specific subscales; no difference in task duration
Cizmic et al. [18]	Porcine liver; box trainer	10 sessions	GOALS; OSATS; CVS rate; LC completion rate; complication rate	AR (iSurgeon) vs. verbal-only guidance	AR group significantly superior on all technical performance and safety metrics; operative time did not differ; baseline randomisation imbalance acknowledged
Wild et al. [19]	Porcine liver; box trainer	—	GOALS; OSATS total; complication rate	AR (iSurgeon) vs. verbal instruction only	Significantly higher GOALS and OSATS scores and lower complication rate in AR group; operative time did not differ
Chowriappa et al. [16]	Inanimate UVA model; daVinci; simulation lab	—	GEARS (6 domains); UVA score; NASA-TLX	HoST AR robotic training vs. passive video observation	AR-trained group significantly outperformed controls on overall robotic skill and all UVA sub-domains; cognitive load significantly reduced; crossover arm confirmed findings
Dodier et al. [15]	Perfused aneurysm phantoms; ex vivo lab	—	Occlusion rate (Raymond-Roy); clip selection match; wrist tremor	AR holographic simulation vs. video-review-only/no-training control	Within-group improvement in AR group occlusion rate; between-group comparison limited by baseline randomisation imbalance; process metrics (tremor, simulation time) unchanged

AR augmented reality, CVS Critical View of Safety, DHS dynamic hip screw, GEARS Global Evaluative Assessment of Robotic Skills, GOALS Global Operative Assessment of Laparoscopic Skills, LC laparoscopic cholecystectomy, MR mixed reality, NASA-TLX NASA Task Load Index, OSATS Objective Structured Assessment of Technical Skills, PCNL percutaneous nephrolithotomy, UVA urethrovesical anastomosis, VR virtual reality.

significant training gains on all LAP Mentor II metrics in both time-based and criterion-based groups ($p < 0.05$), maintained without significant decline at one week and without between-group difference in retention magnitude. Dodier et al. [15] found aneurysm occlusion rate (Raymond-Roy Class I) improved significantly in the AR group at 14 days (76–93%; $p = 0.046$), while active simulation time ($p = 0.755$) and wrist tremor ($p > 0.999$) did not change; pooled arms showed only a non-significant trend toward higher occlusion versus control ($p = 0.096$), and baseline imbalances mean this should be read cautiously. Sugand et al. [20] used a one-week washout rather than a true post-training interval: within-group improvement across all five FluoroSim metrics was significant in the AR group ($p <$

0.05) but not in controls ($p = 0.230–0.961$), and between-group differences were significant for three of five metrics (procedural time, radiograph count, guidewire retries) but not tip-apex distance or predicted cut-out rate. This washout design does not constitute true retention evidence.

Transfer of AR-acquired skills to operative and ex-vivo performance

Six studies assessed skills transfer: one intraoperatively [22] and five to high-fidelity ex-vivo models [15–19], spanning laparoscopic cholecystectomy, aneurysm clipping, and robot-assisted urethrovesical anastomosis (Table 2), using GOALS/CVS [22], GOALS/OSATS/completion/

complication rates/operative time [17–19], GEARS/NASA-TLX [16], and a multimetric occlusion/tremor framework [15].

Intraoperative transfer

Huettl et al. [22] found no significant difference between AR-assisted (HoloPointer) and conventional intraoperative guidance in GOALS total score (18.9 ± 4.5 vs. 18.6 ± 4.8 ; $p = 0.94$), any GOALS sub-domain ($p = 0.39$ – 0.66), CVS score (3.8 ± 2.1 vs. 4.6 ± 1.5 ; $p = 0.38$), or operative time ($p = 0.31$). The AR group received fewer gestural (1.9 vs. 4.6; 59.4% reduction) and verbal corrections (11.4 vs. 17.8; 36.1% reduction), yet trainer-rated assistance quality was significantly lower (2.1 vs. 1.5 ; $p = 0.04$).

Ex vivo/high-fidelity simulation transfer

All three studies assessing transfer to porcine box-trainer cholecystectomy reported significant AR advantages [17–19]. Felinska et al. [17] found higher global and task-specific OSATS scores in the AR group ($p < 0.01$ for both), with no difference in task duration ($p = 0.98$). Wild et al. [19] reported higher GOALS and OSATS scores ($p = 0.007$, $p = 0.015$) and lower complication rates (60% vs. 86.7%; $p = 0.020$), again with no difference in operative time ($p = 0.458$). Cizmic et al. [18], across a ten-session programme, found the AR group superior on GOALS/OSATS ($p \leq 0.001$), CVS achievement (79.5% vs. 41.4%; $p \leq 0.001$), completion (53.6% vs. 27.3%; $p \leq 0.001$), and complication rates ($p \leq 0.001$), with no operative time difference ($p = 0.458$) despite an acknowledged baseline imbalance favouring the AR group ($p = 0.022$).

Operative time, notably, did not differ in any of these three studies, despite consistent performance advantages elsewhere.

Chowriappa et al. [16] found significantly higher GEARS scores in the AR-trained group (14.4 ± 1.2 vs. 11.9 ± 4.1 ; $p = 0.012$), advantages across all six GEARS domains and three UVA metrics, and significantly lower NASA-TLX time pressure (5.9 ± 3.4 vs. 9.3 ± 3.0 ; $p = 0.001$) and effort (5.8 ± 2.5 vs. 11.9 ± 3.7 ; $p = 0.035$); the crossover arm replicated these findings.

Dodier et al. [15] found significant within-group improvement in occlusion rate ($p = 0.046$), but the pooled between-group comparison did not reach significance ($p = 0.096$); simulation time ($p = 0.755$) and tremor ($p > 0.999$) were unchanged, and baseline imbalances favouring the AR group ($p = 0.001$, $p = 0.012$) constrain interpretation. Overall, four of five ex-vivo studies showed significant AR advantages on at least one primary outcome; the fifth,

Dodier et al. [15], showed within-group but not between-group significance, and no study reported effect sizes.

Quality assessment

MERSQI score

All nine studies were rated high quality (13.5–16.5 of 18), with eight scoring at least 15.0 (Supplementary Table S1). All studies scored maximally on Domains 1, 4, 6, and 7, and achieved response rates of at least 75% (Domain 3). Domain 5 (validity evidence) was near-maximal except for Brinkman et al. [21], which lacked internal structure validity data, and Shepard et al. [23], which reported content validity only via a Delphi-derived checklist. The most constrained domains were Domain 2 (sampling institutions), reflecting predominantly single-centre recruitment, and Domain 8 (outcome type), reflecting ex-vivo or simulator-only outcomes in every study but Huettl et al. [22]. No study measured patient- or healthcare-level outcomes.

Risk of bias (RoB2)

No study was judged overall low risk: seven had some concerns and two were high risk, both driven by Domain 4 (Supplementary Table S2). Domain 1 was rated some concerns in 8 of 9 studies, owing to incomplete reporting of sequence generation and allocation concealment; Wild et al. [19] was the sole exception, using an independent electronic randomiser with sealed, sequentially numbered envelopes. Domains 2 and 3 (deviations, missing data) were predominantly low risk, with some concerns in Cizmic et al. [18] (shared working space between groups), Wild et al. [19] (parallel rather than crossover design for its final task), Huettl et al. [22] (AR not used in 3 of 16 allocated cases, addressed via intention-to-treat and as-treated analyses), and Chowriappa et al. [16], due to unexplained differential dropout.

Domain 4 (outcome measurement) was the most consequential source of risk, rated high in two studies. In Felinska et al. [17], gaze data annotation was performed by an unblinded assessor, and OSATS rater blinding was not reported. In Wild et al. [19], the training tutor assessed outcomes in real time with no blinding or independent video review, and the AR overlay remained visible during assessment, which precluded blinding altogether. Cizmic et al. [18] was rated some concerns despite blinded video review, since secondary outcomes were assessed by an unblinded tutor in real time. Domain 5 (selective reporting) was rated some concerns in eight studies, owing to absent prospective registration or pre-specified analysis plans; the consistent reporting of non-significant alongside significant findings,

though, suggests this reflects a transparency limitation rather than active selective reporting.

Discussion

Overall interpretation of the evidence base

The nine included studies show a consistent directional signal favouring AR simulation, with no study reporting inferior AR performance on its primary endpoint. But methodological homogeneity (MERSQI 13.5–16.5; Supplementary Table S1) and the complete absence of reported effect sizes mean statistical significance cannot be read as educational or clinical meaningfulness [24]. What the evidence supports is narrower than a blanket endorsement: AR simulation looks like a useful early-stage skills adjunct, with confidence that does not yet extend beyond controlled simulation environments.

Position relative to existing evidence syntheses

This is a different question from the one recent AR reviews have been answering. El Ashry et al. [9], Xiong et al. [10], and Farooq et al. [11] each demonstrate, across overlapping but not identical study sets, that AR improves technical performance during training; four of the five laparoscopic studies included here, Felinska [17], Cizmic [18], Wild [19], and Huettl [22], also appear in one or more of those meta-analyses. None of the three, however, formally appraised retention or transfer as an outcome, and each names this as the gap that remains. The present review is not a further count of technical-performance trials; it is built around the subset of that same literature, plus additional studies these reviews did not capture, that actually reported a delayed or transfer outcome, appraised with MERSQI and RoB2 rather than pooled effect sizes alone. Two studies do the load-bearing work here: Cizmic et al. [18], the only ex-vivo study with a training programme long enough (ten sessions) to speak to something beyond a single exposure, and Huettl et al. [22], the only study of any AR platform assessing performance in a live operating theatre. Both are included in the immediate-performance meta-analyses above; neither has, until now, been analysed specifically for what it says about retention or transfer.

Simulation-to-clinical transfer gap

The most clinically significant tension in this review is the contrast between consistent AR advantages in ex-vivo transfer studies and the absence of AR superiority in the one study assessing true intraoperative transfer. Huettl et al.

[22] found no difference between AR and control groups in GOALS scores, CVS achievement, or operative time, directly contrasting the significant advantages reported by Felinska, Cizmic, and Wild et al. [17–19] under porcine box-trainer conditions. Several factors could explain the divergence. Ex-vivo settings remove contextual variables present in the operating theatre, unpredictable anatomy, time pressure, the consequences of error, that may otherwise favour AR guidance [25]. Huettl et al. [22] also evaluated a distinct AR modality, real-time intraoperative annotation, rather than the pre-operative simulation training used in every ex-vivo study. And AR-trained participants in Huettl et al. [22] received fewer gestural and verbal corrections yet were rated as receiving lower-quality assistance, which suggests AR changed the structure of the teaching interaction rather than the participants' underlying technical performance. Correction frequency, in other words, is not a safe proxy for skill acquisition, at least not on this evidence. A single, moderately powered null result cannot settle the question either way: until adequately powered, pre-registered, blinded intraoperative trials exist, the gap between ex-vivo and operative transfer remains the field's largest evidence deficit, and the one El Ashry [9], Xiong [10], and Farooq's [11] reviews could not close, because none of their included studies addressed it.

Nature of learning effects

Operative time showed no significant difference in any study that measured it ($n = 8$), regardless of specialty, platform, or comparator. This is probably not a null finding so much as a ceiling effect: in novice populations, duration is driven more by hesitation and error recovery than by movement quality, so AR may sharpen precision without shortening the procedure. That would fit the Fitts-Posner model, where qualitative gains precede speed and automaticity [26]. Chowriappa et al. [16] provided the review's only cognitive load evidence, with significantly lower NASA-TLX time-pressure and effort scores in the AR group, which is at least consistent with AR functioning as a cognitive scaffold, offloading spatial information from working memory. Retention findings were outcome-dependent: Dodier and Sugand [15, 20] showed differential retention across metrics, favouring precision-guided measures over efficiency ones, and Sugand's [20] washout design is mechanistically distinct from the true retention paradigms used by Shepard and Dodier [15, 23] and should not be read as confirmatory. With a maximum follow-up of eight weeks across the entire evidence base, any inference about consolidation timescales beyond that window is premature. These findings sit against real methodological limits. Unblinded assessors in Felinska and Wild et al. [17, 19] (RoB2 Domain 4: high risk) create a

plausible mechanism for performance inflation on observed instruments, and only Huettl et al. [22] reached Kirkpatrick Level 3. No study reported patient complication rates, patient-reported outcomes, or safety data, so what this evidence demonstrates is improved simulator performance, not established clinical benefit.

Implications for practice and future research

The evidence supports a fairly circumscribed set of conclusions. AR simulation produces measurable technical performance benefits in controlled ex-vivo and simulator-based settings, predominantly in novice learners undergoing laparoscopic or robotic training, which justifies cautious integration as a pre-operative or early-curriculum adjunct, most defensibly for laparoscopic cholecystectomy and robotic surgery, where the evidence is most internally consistent. What is not supported matters just as much: nothing here demonstrates benefit for skill consolidation beyond an eight-week horizon, procedural efficiency gains, or AR as a substitute for conventional intraoperative supervision.

The highest research priority, and one El Ashry, Xiong, and Farooq's reviews converge on from the opposite direction, is replication of intraoperative transfer studies at Kirkpatrick Level 3–4, adequately powered, with pre-registered outcomes and blinded assessment [27]. A coherent research agenda should also prioritise retention studies beyond three months, enrol senior trainees, mandate effect size reporting, and standardise platform characterisation so that future studies can actually be pooled. Field-wide consensus on these reporting standards, frankly, would do more for this literature than another single-platform trial. AR simulation occupies a genuinely promising but evidentially immature position, and it warrants the same methodological rigour demanded of any other training intervention before system-level adoption.

Limitations

This review's central limitation is size: nine studies is a small base on which to draw conclusions about retention and transfer, and it spans laparoscopic, robotic, urological, orthopaedic, and neurosurgical contexts with real heterogeneity in AR platform, training dose, retention interval, and assessment setting, which precludes meaningful quantitative synthesis. A second, related limitation is assessor blinding: randomisation sequence generation was inadequately reported in eight of nine studies, and RoB2 assessment surfaced further gaps in allocation concealment and outcome-assessor independence, which, combined with small sample sizes, raises the risk of both inflated effect estimates and type II error. Grey literature may have been

missed, English-language restriction may have introduced language bias, and the short retention intervals across the evidence base limit any conclusion about long-term durability, a genuinely important gap for curriculum design, but one that is more a property of the field than of this review's methodology.

Conclusion

This systematic review synthesised evidence from nine studies evaluating AR simulation for skill retention and transfer to high-fidelity ex-vivo and intraoperative performance, across laparoscopic, urological, neurosurgical, and orthopaedic training. AR-based simulators were associated with improved technical performance and short-term ex-vivo retention, but the evidence base remains small, methodologically heterogeneous, and variable in quality. Where recent reviews of this literature have established that AR improves performance during training, this review's contribution is narrower and, we think, more useful: it shows what that same literature does, and mostly does not, say about whether the improvement lasts and whether it survives the operating theatre.

AR simulation holds potential as an adjunct within structured, competency-based curricula, supplementing rather than replacing ex-vivo training or operative exposure. Until higher-quality, adequately powered trials with longer follow-up and blinded intraoperative assessment are available, AR should be treated as complementary rather than stand-alone, with implementation guided by specialty-specific evidence and institutional resources.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s44186-026-00585-5>.

Acknowledgements We thank Cha Won Jang for his contribution as an external reviewer during title and abstract screening.

Author contributions E.H. contributed to conception and design, screening/data extraction, quality appraisal, drafting of the manuscript, and critical revision. A.A. contributed to screening/data extraction, quality appraisal, and critical revision of the manuscript. A.H. contributed to supervision/conception and design/critical revision. All authors read and approved the final manuscript.

Funding No sources of funding were provided for this review; the corresponding author is not the recipient of any research scholarship.

Data availability All data extracted and analysed during this review are included in this published article and its supplementary information files (Tables S1–S2).

Declarations

Competing Interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Chung K, Kotsis S. Application of the see one, do one, teach one concept in surgical training. *Plast Reconstr Surg*. 2013;131(5):1194–201.
- Szent-Gyorgyi LE, Coblyn J, Turchin A, Loscalzo J, Kachalia A. Building a departmental quality program: a patient-based and provider-led approach. *Acad Med*. 2011;86(3):314–20.
- Dawe SR, Pena GN, Windsor JA, Broeders JA, Cregan PC, Hewett PJ, et al. Systematic review of skills transfer after surgical simulation-based training. *Br J Surg*. 2014;101(9):1063–76.
- Vávra P, Roman J, Zonča P, Ihnát P, Němec M, Kumar J, et al. Recent development of augmented reality in surgery: a review. *J Healthc Eng*. 2017;2017:4574172.
- Kovoor JG, Gupta AK, Gladman MA. Validity and effectiveness of augmented reality in surgical education: a systematic review. *Surgery*. 2021;170(1):88–98.
- McGaghie WC, Issenberg SB, Cohen ER, Barsuk JH, Wayne DB. Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. *Acad Med*. 2011;86(6):706–11.
- Lim EHI, Tan L-Z, Foo HM, Adrian SS. Augmented reality-enhanced simulation in surgical education: a scoping review of implementation, skill retention, and equity in low-resource settings. *Cureus*. 2026;18(1):e102079.
- Suresh D, Aydin A, James S, Ahmed K, Dasgupta P. The role of augmented reality in surgical training: a systematic review. *Surg Innov*. 2023;30(3):366–82.
- El Ashry M, El Ashry A, Khaliq H, Abdalle Y, Yeung T. Augmented reality in surgical training: systematic review of its impact on technical performance in surgical trainees. *JMIR Med Educ*. 2026;12:e71572.
- Xiong J, Dai X, Zhang Y, Liu X, Zhou X. Augmented reality for basic skills training in laparoscopic surgery: a systematic review and meta-analysis. *Surg Endosc*. 2025;39(1):307–18.
- Farooq M, Akram U, et al. Telestration with augmented reality in surgery: a comprehensive systematic review and meta-analysis. *J Surg Res*. 2025;314:690–700.
- Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
- Reed DA, Beckman TJ, Wright SM. An assessment of the methodologic quality of medical education research studies published in the American Journal of Surgery. *Am J Surg*. 2009;198(3):442–4.
- Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*. 2019;366:l4898.
- Dodier P, Civilla L, Mallouhi A, Haider L, Cho A, Lederer P, et al. An evaluation of physical and augmented patient-specific intracranial aneurysm simulators on microsurgical clipping performance and skills: a randomized controlled study. *Neurosurg Focus*. 2024;56(1):E9.
- Chowriappa A, Raza SJ, Fazili A, Field E, Malito C, Samarasakera D, et al. Augmented-reality-based skills training for robot-assisted urethrovaginal anastomosis: a multi-institutional randomised controlled trial. *BJU Int*. 2015;115(2):336–45.
- Felinska EA, Fuchs TE, Kogkas A, Chen ZW, Otto B, Kowalewski KF, et al. Telestration with augmented reality improves surgical performance through gaze guidance. *Surg Endosc*. 2023;37(5):3557–66.
- Cizmic A, Müller F, Wise PA, Häberle F, Gabel F, Kowalewski KF, et al. Telestration with augmented reality improves the performance of the first ten ex vivo porcine laparoscopic cholecystectomies: a randomized controlled study. *Surg Endosc*. 2023;37(10):7839–48.
- Wild C, Lang F, Gerhäuser AS, Schmidt MW, Kowalewski KF, Petersen J, et al. Telestration with augmented reality for visual presentation of intraoperative target structures in minimally invasive surgery: a randomized controlled study. *Surg Endosc*. 2022;36(10):7453–61.
- Sugand K, Wescott RA, Carrington R, Hart A, van Duren BH. Training and transfer effect of FluoroSim, an augmented reality fluoroscopic simulator for dynamic hip screw guidewire insertion: a single-blinded randomized controlled trial. *J Bone Joint Surg Am*. 2019;101(17):e88.
- Brinkman WM, Buzink SN, Alevizos L, de Hingh IH, Jakimowicz JJ. Criterion-based laparoscopic training reduces total training time. *Surg Endosc*. 2012;26(4):1095–101.
- Huettl F, Heinrich F, Boedecker C, Vradelis L, Ludt A, Kneist W, et al. Real-time augmented reality annotation for surgical education during laparoscopic surgery: results from a single-center randomized controlled trial and future aspects. *J Am Coll Surg*. 2023;237(2):292–300.
- Shepard L, Im C, Li O, Schuler N, Holler T, Saxton A, et al. Comparison of remote mixed reality versus in-person training of ultrasound-guided percutaneous nephrolithotomy with urological residents. *Urology*. 2025;199:230–5.
- Sullivan GM, Feinn R. Using effect size-or why the P value is not enough. *J Grad Med Educ*. 2012;4(3):279–82.
- Hamstra SJ, Brydges R, Hatala R, Zendejas B, Cook DA. Reconsidering fidelity in simulation-based training. *Acad Med*. 2014;89(3):387–92.
- Bugdadi A, Sawaya R, Olwi D, Al-Zhrani G, Azarnoush H, Sabagh AJ, et al. Automaticity of force application during simulated brain tumor resection: testing the Fitts and Posner model. *J Surg Educ*. 2018;75(1):104–15.
- Woodall WJ, Chang EH, Toy S, Lee DR, Sherman JH, Liu M, et al. Does extended reality simulation improve surgical/procedural learning and patient outcomes when compared with standard training methods? A systematic review. *Simul Healthc*. 2024;19(1s):S98–111.