Title	Empowering multimodal learning analytics using agentic AI: A comprehensive platform for simulation-based clinical training with Intelligent Assessment
Authors	Nguyen, Duc Hai;Salim, Kinza;Power, David;Connolly, Murray;Hamdy, Ahmed;Contreras, Maya;Mai, Tai Tan;Shorten, George;O'Sullivan, Barry;Nanjappan, Vijayakumar;Nguyen, Hoang D.
Publication date	2026-04-26
Original Citation	Nguyen, D H, Salim, K, Power, D, Connolly, M, Hamdy, A, Contreras, M, Mai, T T, Shorten, G, O'Sullivan, B, Nanjappan, V & Nguyen, H D 2026, Empowering multimodal learning analytics using agentic AI: A comprehensive platform for simulation-based clinical training with Intelligent Assessment. in 16th International Learning Analytics and Knowledge Conference, Bergen, Norway, 27 April - 1 May 2026. 16th International Learning Analytics and Knowledge Conference, LAK 2026, Association for Computing Machinery, Inc, pp. 674-684, 16th International Conference on Learning Analytics and Knowledge, LAK 2026, Bergen, Norway, 27/04/26. https://doi.org/10.1145/3785022.3785129;conference
Type of publication	Conference contribution (Peer reviewed)
Link to publisher's version	10.1145/3785022.3785129
Rights	cc_by
Download date	2026-08-29 00:44:08
Item downloaded from	https://hdl.handle.net/10468/19039

University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Empowering Multimodal Learning Analytics using Agentic AI: A Comprehensive Platform for Simulation-based Clinical Training with Intelligent Assessment

Duc-Hai Nguyen
University College Cork
Cork, Ireland
125109073@umail.ucc.ie

Kinza Salim
University College Cork
Cork, Ireland
124106946@umail.ucc.ie

David Power
University College Cork
Cork, Ireland
d.power@ucc.ie

Murray Connolly
Cork University Hospital
Cork, Ireland
murrayconnolly@gmail.com

Ahmed Hamdy
Cork University Hospital
Cork, Ireland
ahmed_elsaka888@yahoo.com

Maya Contreras
College of Anaesthesiologists of Ireland
Dublin, Ireland
mcontreras@coa.ie

Tai Tan Mai
Dublin City University
Dublin, Ireland
tai.tanmai@dcu.ie

George Shorten
University College Cork
Cork, Ireland
g.shorten@ucc.ie

Barry O'Sullivan
University College Cork
Cork, Ireland
b.osullivan@cs.ucc.ie

Vijayakumar Nanjappan
University College Cork
Cork, Ireland
vnanjappan@ucc.ie

Hoang D. Nguyen*
University College Cork
Cork, Ireland
hn@cs.ucc.ie

Abstract

Simulation-based clinical training generates rich multimodal data that remains underused due to fragmented modalities, annotation bottlenecks, weak provenance, and tools misaligned with educator workflows. We introduce **ClinVision**, an educator-in-the-loop platform that operationalizes end-to-end multimodal learning analytics: synchronized multi-camera review, ISBAR-aligned scoring (a validated clinical communication framework), and templated reports with jump-to-evidence provenance. Agentic AI - systems that act on behalf of users while preserving human authority - assists with phrasing under explicit control (accept/edit/reject) and visible provenance, supporting accountable use rather than prescriptive automation. An in-learning deployment with five educators revealed full ISBAR coverage and time-efficient workflows, though AI suggestions were used selectively. We surface three transferable design tensions (assistance vs. authority, structure vs. flexibility, evidence vs. overload) and demonstrate that workflow integration, temporal primitives, and background AI assistance may better support high-stakes assessment than analytics or automation alone.

*The corresponding author

LAK 2026, Bergen, Norway

ACM ISBN 979-8-4007-2066-6/26/04
https://doi.org/10.1145/3785022.3785129

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; *Visualization systems and tools*; • **Applied computing** → **Interactive learning environments**; **Health care information systems**.

Keywords

multimodal learning analytics, simulation-based training, human-AI collaboration, clinical education, agentic AI, trustworthy AI, video analytics, ISBAR assessment, human-centered design

ACM Reference Format:
Duc-Hai Nguyen, Kinza Salim, David Power, Murray Connolly, Ahmed Hamdy, Maya Contreras, Tai Tan Mai, George Shorten, Barry O'Sullivan, Vijayakumar Nanjappan, and Hoang D. Nguyen. 2026. Empowering Multimodal Learning Analytics using Agentic AI: A Comprehensive Platform for Simulation-based Clinical Training with Intelligent Assessment. In *LAK26: 16th International Learning Analytics and Knowledge Conference (LAK 2026), April 27–May 01, 2026, Bergen, Norway*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3785022.3785129

1 Introduction

Learning analytics offer a principled way to transform the dense traces of simulation-based clinical training into evidence that strengthens educator judgment and improves assessment quality. In competency-based medical education, robust supervisor evaluation is foundational: workplace-based assessment aggregates multiple observations to inform high-stakes decisions, and programmatic assessment combines diverse data over time to maximize validity and feedback for learning [29, 37, 41]. Within this landscape, *multimodal learning*

analytics (MMLA) has been recognized as a promising means to organize, analyze, and communicate rich evidence responsibly in health professions education [4, 30].

However, moving from simple digital records to MMLA exposes persistent obstacles that current workflows rarely solve. First, multimodality introduces heterogeneity of formats and sampling rates and demands precise temporal synchronization across sources [4, 30]. Second, producing high-quality, rubric-aligned, time-stamped annotations is labor-intensive and difficult to scale; manual annotation remains common in MMLA pipelines [28]. Third, gaps in standardization and interoperability limit reuse and comparison across cohorts and sites [30]. Finally, adoption in high-stakes settings depends on privacy, provenance, and human control [19, 25]. In practice, these issues manifest as loss of temporal fidelity, idiosyncratic phrasing undermining inter-rater consistency, and difficulty scaling assessment.

Existing practice also falls short in how video is used during debriefing. Although video-assisted debriefing is widely employed in technology-enhanced simulation, effectiveness depends on facilitation skill and context; syntheses report mixed or modest incremental benefits over verbal debriefing, underscoring the need for structured, evidence-anchored prompts [8, 47]. Routine use of video in real clinical contexts raises concerns among healthcare providers, highlighting adoption barriers [35]. At the same time, structured communication frameworks such as ISBAR (Identification, Situation, Background, Assessment, Recommendation) provide a validated mental model for making clinical feedback clear and actionable [15, 17]. Originally developed to improve patient handoffs, ISBAR has been widely adopted in simulation training. Validated instruments such as OSAD and electronic variants streamline data entry and support reliable scoring [1, 3].

These observations point to an emerging opportunity: platforms that do more than replay video. To enable MMLA in clinical education, systems should make the creation of time-locked annotations efficient and consistent, then convert those labels into structured, ISBAR-aligned feedback and reports, closing the loop from data collection to action.

This paper introduces *ClinVision*, a web-based platform designed with educators to address these problems within an end-to-end workflow (Fig. 1). ClinVision integrates multi-camera session review with time-linked annotations; provides ISBAR-aligned scoring; and uses *agentic AI* - AI systems that act on behalf of users while preserving human authority and control - to draft phrasing via accept/edit/reject mechanisms and provenance-linked jump-to-evidence. Rather than replacing professional judgment, this agentic approach positions AI as an assistant operating on educator-authored artifacts, suggesting refinements with visible provenance.

We therefore evaluate the platform on three reviewer-salient dimensions:

- **RQ1 - Evidence creation and consistency:** Does the platform help educators find, segment, and label salient behaviors across video and audio with higher coverage than current practice, and what features support or hinder consistency?
- **RQ2 - Effort and flow:** Does educator-in-the-loop, agentic assistance reduce the time and interaction cost of producing structured assessments while keeping educators in control of wording and emphasis?
- **RQ3 - Actionability and traceability:** Do provenance-linked timestamps/spans and ISBAR-aligned summaries make feedback clearer, easier to justify, and more useful for debriefing and supervisory decisions?

Our contributions are threefold:

(1) An end-to-end educator-in-the-loop MMLA workflow that unifies synchronized multi-camera review, time-linked labeling, ISBAR-aligned scoring, and provenance-preserving report generation, turning raw multimodal streams into decision-ready evidence.
(2) Empirically-grounded design insights for human-AI collaboration in multimodal assessment that address tensions between automation efficiency and professional authority through role-based AI scoping (error detection, post-hoc summarization) and mandatory provenance preservation, patterns potentially transferable to other MMLA contexts requiring auditability.
(3) Qualitative evidence from an in-learning pilot plus focus group revealing three design tensions (assistance vs. authority, structure vs. flexibility, evidence vs. overload) that constrain MMLA adoption in this high-stakes setting, along with usage patterns and educator feedback that informed concrete design refinements.

2 Background and Related Work

2.1 From Debrief to Evidence: Current Practice

Debriefing is the primary learning phase in simulation training, where participants reflect on actions and consolidate competencies [12, 40]. Video-assisted debriefing can enhance self-awareness by anchoring discussion in shared artifacts [2, 32], while annotation marks events for review and reduces reliance on memory [33]. However, curating clips is time-consuming, repeated viewing may induce stress, and practical frictions limit routine uptake [22, 35]. These constraints motivate learning-analytics approaches coupling educator judgment with scalable, structured evidence.

2.2 AI and Multimodal Analytics in Simulation

AI, multimodal learning analytics (MMLA), and sensing technologies enable capture and analysis of complex behaviors in simulation-based training [4, 16, 23]. Speech processing identifies communication patterns such as interruptions and silence gaps [18, 48], while computer vision tracks pose, gaze, and movement to infer engagement or hesitation [14, 43]. Interfaces translate analytics into trajectory displays, teamwork dashboards, and narrative tools [11, 13]. However, adoption constraints surface most clearly in *in-learning* deployments [7, 34], highlighting needs for temporal fidelity, alignment with structured frameworks, and educator control.

2.3 MMLA in Learning Analytics Research

The learning analytics community has advanced MMLA through three streams. *Frameworks for multimodal data orchestration* establish design principles emphasizing stakeholder needs and workflow

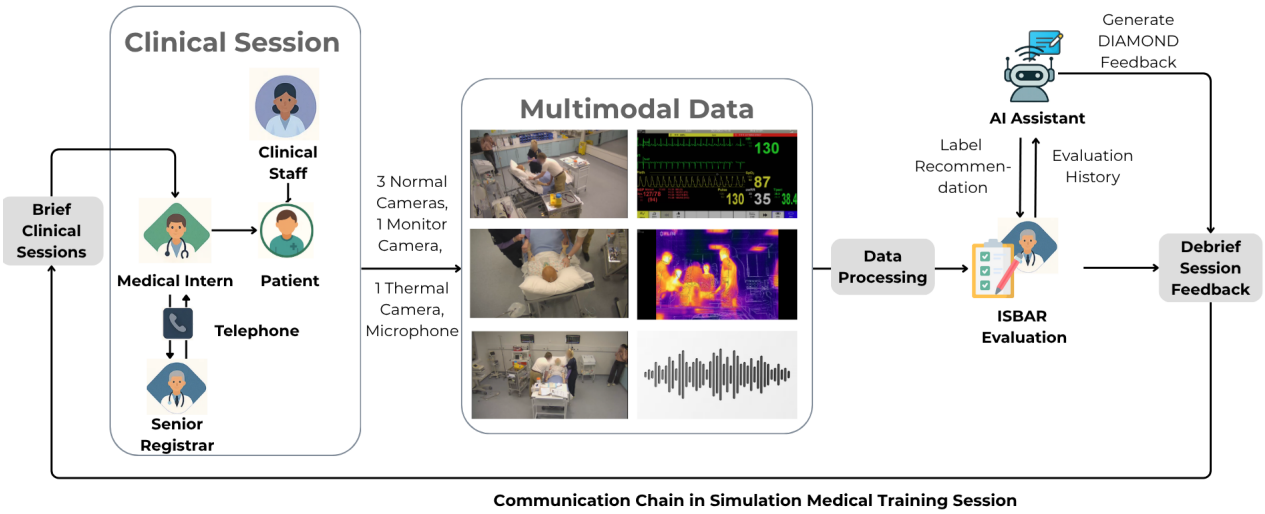

Figure 1: Overall workflow for simulation-based clinical training and ClinVision support. Multimodal data from synchronized cameras and audio feeds enable ISBAR-aligned evaluation with AI-assisted feedback generation. See Fig. 2 for data processing details.

integration [26, 31]. *Sensor-based skill tracking* reveals patterns invisible to observers, such as hesitation in gesture sequences or breakdown in collaboration rhythms [10, 44], yet automated inference struggles with intent, leaving manual annotation as a bottleneck. *Trust and transparency research* shows that opacity and misalignment with professional judgment constrain uptake [38, 42]. ClinVision extends this work by operationalizing end-to-end workflow integration and documenting design tensions emerging in consequential assessment practice.

2.4 ClinVision's Positioning

Existing MMLA research demonstrates multimodal capture feasibility [30] and promise of educator-facing dashboards [11]. Three gaps persist: **(1) Fragmented pipelines:** few operationalize *end-to-end* workflow from capture to auditable reporting [43]. **(2) Analytics without assessment standards:** dashboards rarely align with validated frameworks (ISBAR [17]), limiting uptake for high-stakes evaluation. **(3) Opaque AI:** work explores AI summaries [20] but seldom addresses trust requirements [19, 25]. This motivates *agentic AI*, systems preserving human authority through explicit control. Unlike fully automated assessment, agentic AI operates on educator-authored artifacts, providing accept/edit/reject suggestions with visible provenance. ClinVision addresses these gaps by integrating end-to-end pipeline, validated standards, and agentic AI within educator-controlled workflow.

3 MMLA Challenges in Simulation-Based Clinical Training

Simulation-based clinical training generates rich multimodal traces (multi-camera video, audio, interaction events, monitoring streams) that remain underused because three persistent challenges complicate the transition from data collection to actionable feedback [28, 30]. We articulate these challenges within surgical simulation training, showing how they manifest as workflow breakdowns.

3.1 Challenge 1: Temporal Heterogeneity Across Modalities

Multimodal capture involves streams with different temporal properties: video fps, continuous audio, discrete events (millisecond precision), and human annotations. Two problems arise: *synchronization* of streams from independent devices with clock drift [4], and *temporal fidelity during review* (maintaining precise timing relationships when analyzing recorded events)-educators debriefing from memory lose access to precise event timing, making it difficult to anchor feedback or measure response latencies [29].

In our surgical training center, sessions follow three phases: prebrief (10 min), simulated procedure (20–30 min), and debrief (30 min using frameworks like ISBAR). Educators rely on handwritten notes or memory; video playback is cumbersome-manually scrubbing multiple unsynchronized feeds-often abandoned under time pressure. This forces a choice between efficiency (memory-based) and precision (time-intensive video review).

3.2 Challenge 2: The Annotation Bottleneck

High-quality analytics depend on timestamped annotations linking behaviors to assessment criteria [28]. In simulation, relevant behaviors are brief, overlapping, and multimodal-e.g., delayed sepsis recognition inferred from speech hesitation (audio), extended gaze at monitors (video), and failure to escalate (temporal gap). Manual annotation is labor-intensive: educators in our partner center spend 30–45 minutes producing detailed, ISBAR-aligned feedback for a 25-minute scenario-a 2:1 ratio incompatible with routine use. Automated detection offers partial relief but cannot reliably infer intent; speech recognition misses pragmatic failures (e.g., omitting patient identifiers). Consequently, annotations remain sparse, inconsistent across raters, and disconnected from structured rubrics [41].

3.3 Challenge 3: Trust, Transparency, and Human Control

AI interest in automating assessment (feedback summaries, score suggestions, flagging failures [20]) faces trust requirements: transparency about what AI observed, how it reached conclusions, and where human override is possible [19, 25]. In competency-based medical education, assessments aggregate into high-stakes decisions about progression and certification [29]; educators cannot delegate to opaque systems yet need efficiency gains.

Our study context reveals skepticism toward "black-box" AI combined with frustration at manual inefficiency. Educators want assistance but insist on control over wording, emphasis, and scores. They want to see evidence (video timestamp, transcript) justifying AI claims and to edit or reject suggestions. Current tools rarely provide this: dashboards lack drill-down; summaries lack provenance links; accept-or-reject workflows don't support partial editing [19].

3.4 ClinVision's Response: Multimodal Data Infrastructure

ClinVision addresses these challenges through integrated infrastructure (Fig. 2): multi-camera video and audio ingestion, automated de-identification and temporal alignment, relational storage linking synchronized streams to timestamped annotations and ISBAR evaluations, and educator-facing tools for annotation, scoring, and templated report generation-with agentic AI assistance and role-based access control.

Five complementary modalities address the challenges (We refer to key challenges as C): **(1) Multi-camera video** (3–4 synchronized cameras) provides complementary perspectives while preserving temporal relationships (C1). **(2) Audio capture** (omnidirectional microphones) enables assessment of communication clarity and ISBAR structure, time-aligned with video. **(3) Annotation timelines** let educators mark key moments linked to ISBAR categories, reducing effort by enabling deliberate annotation rather than real-time capture (C2). **(4) Structured ISBAR evaluations** link scores to supporting annotations and video evidence, standardizing assessment language [17]. **(5) AI-assisted phrasing** suggests comments and generates draft reports; all AI content is visibly labeled, editable, and linked to source evidence via jump-to-timestamp buttons (C3).

A relational architecture centers on *Session* entities linking video files, timestamped comments, ISBAR evaluations, and survey responses, preserving temporal fidelity so every observation traces to concrete timeline evidence (C1). Secure anonymized identifiers protect participants; face blurring and masking of personally identifiable information are applied where appropriate. Once ingested, streams are synchronized server-side and exposed via a web interface supporting multi-camera playback, filters, search, and jump-to-evidence-operationalizing the full MMLA pipeline [26, 45]. These challenges informed the co-design process (Section 4), where educators translated them into concrete design requirements.

The platform is implemented as a web application with Python/Frappe backend, React/TypeScript frontend, and MariaDB relational database. HTTP range-request video streaming enables efficient playback of multi-camera sessions without requiring full file downloads, while RESTful APIs connect frontend components to backend services for session management, annotation persistence, and report generation.

4 Design of ClinVision

We adopted human-centered design (HCD) to iteratively develop *ClinVision* [5, 24, 46]. MMLA tools often fail when designers misunderstand educator workflows, trust requirements in high-stakes assessment, or the gap between analytics affordances and actionable feedback [7, 19, 25]. Pilot deployments in our surgical simulation center confirmed feasibility of multi-camera capture and synchronized playback, while revealing limitations mapping to the three challenges: unsynchronized streams (C1), slow annotation workflows (C2), and skepticism toward AI-generated text lacking provenance (C3).

4.1 Co-Design with Educators and Practitioners

We convened 7 medical educators, 5 AI researchers, and 2 learning-analytics experts for five months of iterative co-design (Table 1). Three stages unfolded: W1–W2 identified pain points; W3–W4 refined wireframes and calibrated AI scope; W5–W6 validated workflows. Thematic analysis [6] yielded design considerations (Section 4.2) and features (Section 4.3).

Three design tensions. Workshop discussions (W3–W4) surfaced three tensions. **(1) Assistance vs. authority:** educators wanted AI efficiency but rejected automation overriding judgment. Workshop W5 validated scoping AI to suggestion (not prescription). **(2) Structure vs. flexibility:** ISBAR standardization was valued, but rigid rubrics created "impossible scores" for edge cases where learner intent and environmental factors diverged. W6 piloted dual modes (brief vs. detailed) to resolve this. **(3) Evidence vs. overload:** multi-camera richness improved coverage but created review burden when all streams displayed simultaneously. These tensions inform six design considerations (D1–D6, Table 2) addressing MMLA challenges while operationalizing principles from learning analytics and human-AI interaction literature.

4.2 Design Instantiation: Features as Challenge Resolutions

ClinVision's four key capabilities instantiate D1–D6 to address the three MMLA challenges (Fig. 3 shows the end-to-end educator-controlled workflow from session selection through ISBAR scoring to AI-assisted report generation):

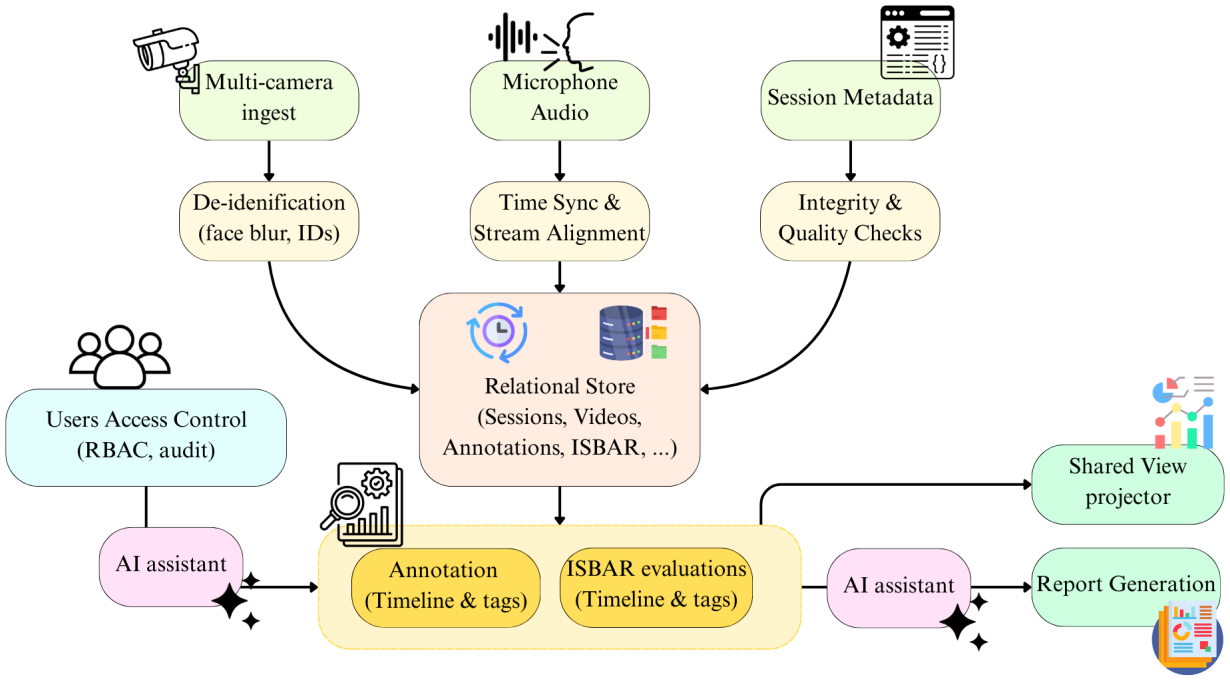

Figure 2: System pipeline: ingestion flows through de-identification and alignment into relational storage linking sessions, videos, annotations, ISBAR evaluations, and metadata. Educator workflows feed shared-screen views and templated reports. AI supports phrasing with human oversight; RBAC and audit logs govern access.

Table 1: Co-design process: five months of iterative development.

Stage	Participants	Methods	Key Outcomes
W1–W2	7 educators, 5 AI, 2 LA	Workflow mapping, pain-point analysis	Requirements; MMLA challenges; ISBAR alignment
W3–W4	Developers + 2 LA	Wireframes, critiques	Architecture validated; AI scope calibrated
W5	6 educators + dev + LA	Hands-on trials	ISBAR workflow validated; usability issues
W6	Full team	Build review, templates	Templates finalized; MVP agreed
Pilot	5 educators (in-learning)	Real sessions, surveys	Annotation value confirmed; provenance needs

Table 2: Design considerations for integrating multi-video review, structured evaluation, and AI assistance into debriefing.

ID	Consideration	Challenge Addressed & Rationale	Implementation in ClinVision
D1	AI assists, humans decide	**C3 (Trust):** AI acceleration valued but unvetted automation erodes authority.	All AI labeled, optional, editable; accept/edit/reject controls; visible provenance; audit logs.
D2	Multi-video without overload	**C1 (Temporal heterogeneity):** Multiple streams needed but create cognitive burden.	Synchronized timelines; linked play/pause; quick camera toggles; jump-to-moment navigation.
D3	Embed ISBAR in workflow	**C2 (Annotation):** Separate rubric mapping reduces consistency and increases effort.	Native ISBAR tags in timeline/forms; one-tap categorization; ISBAR filters; aligned reports.
D4	Efficient & flexible annotation	**C2 (Annotation):** Full automation unreliable; educators require control.	Lightweight bookmarks/tags/notes; AI-assisted phrasing (optional); reusable templates.
D5	Assistive AI across LA pipeline	**C2, C3:** Accelerate workflow without replacing educator judgment.	AI phrasing suggestions; ISBAR report drafts; pattern surfacing; educator-controlled filters.
D6	Trust, transparency, privacy by design	**C3 (Trust):** Opaque systems and privacy breaches block adoption.	Human-in-the-loop approval; visible provenance; de-identification; RBAC; audit trails.

(1) Multi-video playback with timeline annotations (C1, D2/D4): Synchronized multi-camera player with linked scrubbing; timeline controls (zoom, jump-to-annotation, color-coded ISBAR tags); timestamped comments categorized by ISBAR at annotation time (Fig. 4).

(2) ISBAR evaluation interface (C2, D3): Ordinal scoring (0–3) per category linked to annotations/timestamps; visual summaries (bar charts, radar plots); freeform fields for edge cases that rigid rubrics cannot capture [17] (Fig. 4).

(3) AI-assisted comment suggestions (C3, D1/D5/D6): A large language model (Google Gemini 2.5 Flash) generates context-aware phrasing suggestions for educator-authored annotations, using the annotation text and ISBAR category as input context. Educators retain full control via accept/edit/reject mechanisms; all AI-generated content displays visible provenance indicators ("AI-suggested" badges, jump-to-evidence links).

(4) Templated report generation (C3, D5/D6): ISBAR-aligned reports compiled from tagged comments/scores; each claim linked to source evidence (timestamp, annotation, scorer ID); educators review/refine before sharing [41].

5 In-learning Classroom Study

To evaluate whether ClinVision resolves the three MMLA challenges in authentic practice, we conducted an in-learning deployment. In-learning evaluation is critical because controlled studies cannot capture workflow integration challenges, competing cognitive demands, and trust dynamics that emerge when educators use analytics for consequential decisions [7, 9, 34]. Our design prioritizes ecological validity over experimental control.

The study took place at a university surgical simulation center. Educators used ClinVision during post-scenario debriefing to review performance, annotate key moments, and generate feedback reports. All received guidelines on system usage and a short meeting for tutorial to get familiar with the system. The study received ethics approval.

5.1 Participants and Study Context

Five medical educators (E1–E5, with years of simulation training experience in emergency medicine, surgery, or critical care) participated. None had MMLA or AI-assisted assessment experience. Each educator evaluated the same two pre-recorded simulation sessions: a sepsis scenario requiring escalation and ISBAR handoff, and a anaphylaxis scenario requiring deterioration recognition and multidisciplinary coordination. Both sessions were 20–25 minutes, captured by four synchronized cameras, and designed to elicit observable ISBAR behaviors where memory-based evaluation typically misses temporal details.

5.2 Data Collection and Analysis

Data sources. Four data types were captured: **(1) Annotation logs:** timestamped actions (start/stop, text, ISBAR categories, AI acceptance/editing/rejection). **(2) ISBAR assessment data:** ordinal scores (0–3) per category with timestamps and evidence spans. **(3) Post-session surveys:** workflow ease, efficiency, and traceability. **(4) Focus-group discussion:** 60-minute semi-structured discussion with all educators after both sessions, yielding 8,427 words transcribed for thematic analysis. All data were de-identified [39].

Analytical approach. We triangulated annotation logs, surveys, and focus-group transcripts [30]. For **RQ1**, we analyzed annotation density (comments per session, timestamps per category), ISBAR coverage (whether all five categories received at least one annotation), and AI acceptance rates; thematic coding [6, 27] identified consistency barriers. For **RQ2**, we analyzed educators' reported efficiency perceptions from surveys and thematic analysis of focus-group discourse on workflow friction points, complemented by timing patterns from annotation logs. For **RQ3**, we triangulated AI logs (acceptance/editing/rejection rates) with focus-group responses on trust and acceptable AI use cases.

6 Results and Discussion

We present qualitative findings organized by research question, integrating usage patterns from annotation logs with insights from surveys and focus-group discussion. Our findings suggest that ClinVision's design addresses Challenge 1 (temporal heterogeneity) in this context and partially addresses Challenge 2 (annotation bottleneck), but exposes gaps in Challenge 3 related to interaction sequencing, ISBAR temporal completeness, and AI role scoping. Each subsection links results to the MMLA challenges and design considerations from Sections 3 and 4.

6.1 RQ2: Effort, Flow, and the Interaction Sequence Mismatch

Time-efficient review enabled by synchronization. Educators reported completing evaluations more quickly than their typical practice and found the time investment manageable for routine use. Annotation logs showed that educators created multiple timestamped annotations per session, with most annotations created during initial review phases, suggesting front-loaded attention to critical moments followed by rapid ISBAR scoring. All educators noted that synchronized playback and timeline navigation enabled deliberate, evidence-anchored review within practical time constraints. One educator (E2) stated: "Before, I'd be frantically scribbling notes while the scenario was running, then trying to remember what happened when. Now I can pause, rewind, mark the exact moment, and move on." This shift from irreversible real-time capture to deliberate, evidence-anchored review suggests that ClinVision addresses Challenge 1 (temporal heterogeneity) by making synchronized multi-camera review practical within debrief time constraints.

Critical usability friction: forced interaction sequence. However, efficiency gains were constrained by a critical usability issue around forced sequencing of annotation actions. The system required entering comment text *before* the Start button became active, contradicting educators' natural mental model: observe event, mark start timestamp, articulate observation, assign priority, mark end timestamp. This mismatch, which mirrors Klein's Recognition-Primed Decision model [21] where experts first recognize patterns then articulate reasoning, induced repeated errors even after educators understood the constraint. Educators reported frequent attempts to click Start before entering text, requiring workarounds such as

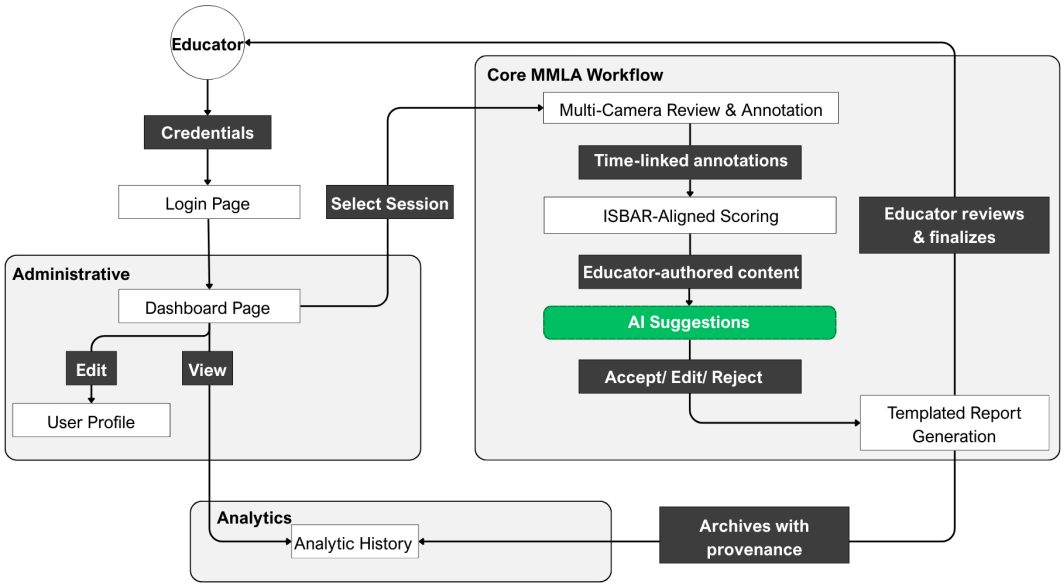

Figure 3: End-to-end educator-controlled MMLA workflow. Core sequence: multi-camera review and annotation; ISBAR-aligned scoring; AI suggestions (accept/edit/reject control); templated report generation with educator review. Reports archive with provenance to activity history. Interactive demonstration: https://duchai972.github.io/ClinVision_static

typing placeholder characters (periods, "x") to unlock timestamping functionality, a strategy they recognized as "incongruous" and "frustrating" (E3, E5).

Cognitive load from simultaneous interface elements. Beyond sequencing, educators reported divided attention across multiple simultaneous interface elements: video playback, audio monitoring, timestamp controls, comment composition, AI suggestion panels, color-label selection, and ISBAR scoring. This cognitive load was compounded during rapid scenario events, where educators needed to decide whether to capture an observation immediately or risk missing the temporal window. The pattern confirms that D2 (multi-video without overload) remains partially unresolved: progressive disclosure, surfacing controls in task-aligned sequence, is needed to reduce friction and accelerate the learning curve.

6.2 RQ1: Evidence Creation, Consistency, and ISBAR's Temporal Gaps

Full ISBAR coverage achieved. Analysis showed full ISBAR coverage across all educator reviews, with all five categories annotated in both sessions by all educators. Educators contrasted this with their traditional practice, where ISBAR application was often inconsistent or incomplete, suggesting that D3 (embed ISBAR in workflow) addresses Challenge 2 (annotation bottleneck). In the focus group, educators noted that native ISBAR tagging during annotation, rather than post-hoc mapping, reduced cognitive overhead: "I don't have to remember which parts are Identification versus Situation; I tag them as I see them" (E4). The ISBAR filter and aligned reports reinforced this structure throughout the workflow, making rubric-aligned evidence capture the path of least resistance rather than an additional step.

Temporal dimensions missing from current interface. However, educators identified a fundamental mismatch: the scoring interface conflates three temporal dimensions: (i) recognition that ISBAR is required (binary event + timestamp), (ii) latency from cue to initiation (interval in seconds), and (iii) execution quality (ordinal/checklist) into a single 0–3 scale. This creates "impossible scores" for edge cases. For example, one educator (E3) described a scenario where a learner correctly recognized the need for escalation but did not receive prompting from the senior clinician when expected: "Do I score that as zero because it didn't happen, or as two because they recognized it? The rubric doesn't let me capture both." Such cases force arbitrary choices that undermine validity. ClinVision captures the data (timestamped annotations marking recognition moments, intervals between cue and action) needed to compute these dimensions separately, but the interface does not expose them as first-class controls in the scoring panel.

Concerns about inter-rater consistency. Despite full ISBAR coverage, educators anticipated high inter-rater variance without explicit behavioral anchors. Multiple participants noted that scoring decisions required "intuition" rather than observable criteria, and expressed doubt that independent raters would converge on similar scores for the same video segment. Note that we did not measure actual inter-rater reliability in this study; these are educator perceptions rather than quantified agreement. Educators recommended: (i) support for both validated ISBAR instruments (5-point global rating for rapid formative assessment vs. 30-item checklist for summative evaluation [15]), and (ii) brief hands-on calibration training using shared clips before deployment. Challenge 2 is only partially resolved: ClinVision accelerates annotation and ensures full rubric

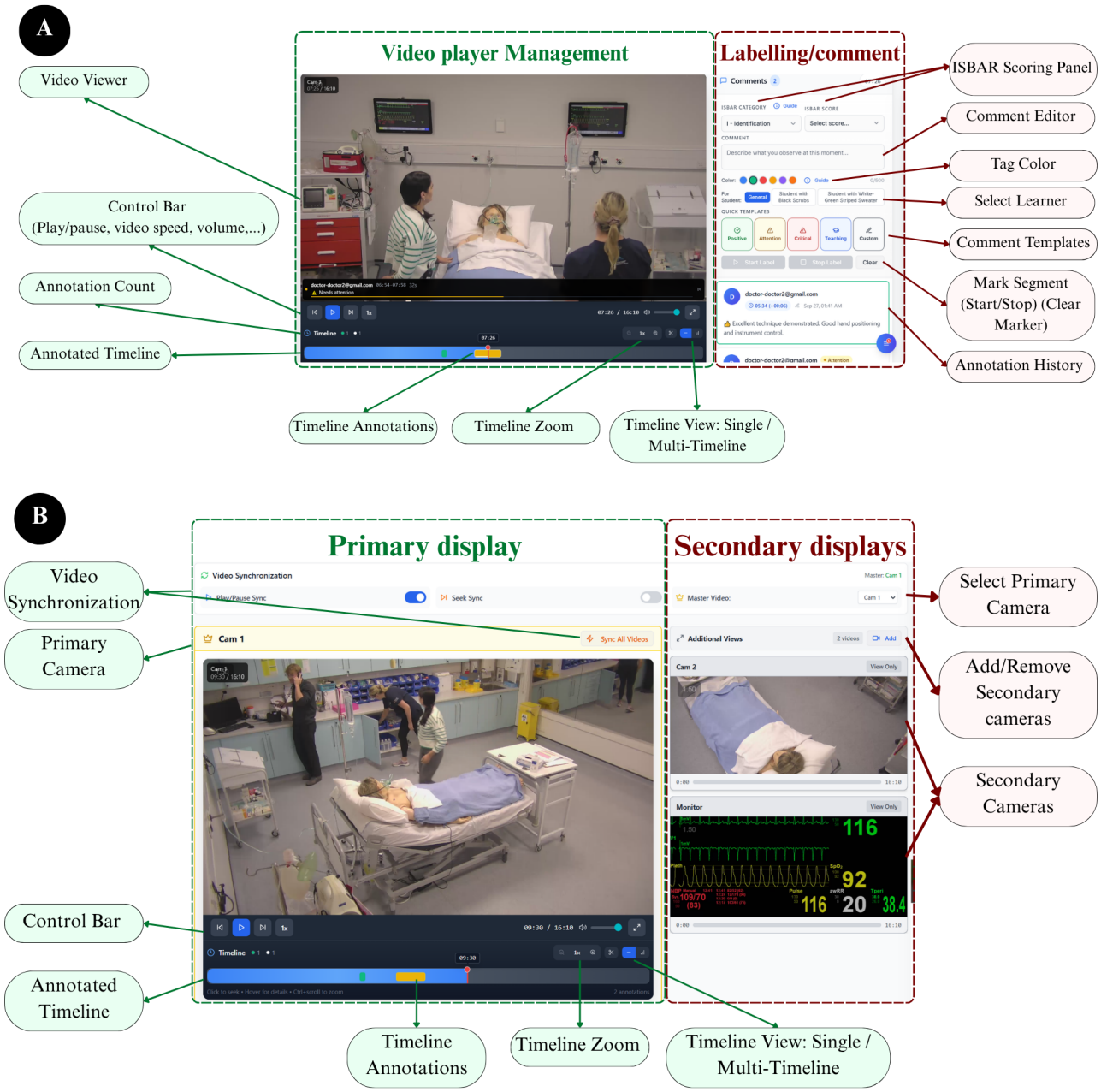

Figure 4: Session workspace: (A) Video player with annotated timeline, timeline zoom, multi-camera view; labelling panel with ISBAR scoring, comment editor, templates, segment marking. (B) Primary and secondary camera displays with synchronized timelines.

coverage, but would require temporal fields, explicit anchors, and empirical IRR validation for program-level assessment [41].

6.3 RQ3: Actionability, Traceability, and AI Assistance Limits

High AI rejection rates reveal trust barriers. AI suggestion logs and focus-group analysis revealed predominantly low acceptance of AI suggestions, with educators frequently editing or rejecting

AI-generated content. Three reasons emerged. First, generic prewritten comment templates lacked scenario-specific detail and often mischaracterized clinical context, for example suggesting procedural feedback when the issue was communication. Educators found themselves deleting template text rather than editing it, negating any efficiency gain. Second, real-time sentiment-classification pop-ups interrupted workflow; educators consistently dismissed these prompts as "distractions" (E2, E4) that added cognitive load during moments requiring focused attention on video content. Third, and most critically, any AI-generated text that replaced or obscured the educator's original wording was perceived as undermining professional judgment and "erasing my voice" (E1). One educator stated: "If I write something and the system changes it, that's no longer my feedback, it's the machine's version of what I meant."

Acceptable AI use cases: background, not prescription. Despite high rejection rates, educators did not reject AI assistance in principle. Instead, they proposed two acceptable use cases that preserve authority while accelerating workflow. The first is *error detection*: if an educator applies a negative color label to text that language models interpret as clearly positive (or vice versa), a non-modal alert could prompt review without blocking workflow. The second is *post-hoc summarization*: after educators complete timestamped annotations, AI-generated draft summaries, clearly labeled, fully editable, and linked to source annotations with jump-to-evidence buttons, could accelerate report preparation while maintaining human control over final wording and emphasis. This shift from real-time prescription to background augmentation aligns with trustworthy AI frameworks [19, 25], resolving the assistance-vs-authority tension by scoping AI to support rather than replace professional judgment.

Provenance valued but underutilized. Educators valued jump-to-evidence features (linking report claims to video timestamps) for supporting traceability and noted using them during report review. However, educators emphasized that provenance indicators were not prominent enough during AI suggestion review, making it difficult to assess *why* the AI made a particular suggestion. This gap constrains actionability: while the infrastructure for auditable feedback exists, the interface does not foreground provenance sufficiently to build trust in AI-assisted content. Addressing Challenge 3 (trust and transparency) requires not only technical provenance but also interface design that makes evidence pathways immediately visible.

6.4 Implications for MMLA Design and Practice

Our findings yield both theoretical implications for the MMLA research community and concrete design refinements for ClinVision's next iteration.

Four cross-cutting implications. **(1) Temporal fidelity requires temporal tools:** Educators distinguished recognition, latency, and execution, collapsed in traditional rubrics. MMLA platforms must expose temporal primitives (event detection, intervals, sequences) as first-class data types, not post-hoc annotations.

(2) The multimodal evidence paradox: Richer data improve evidence quality but worsen cognitive load. Progressive disclosure, surfacing controls in task-aligned sequence, resolves this better than simultaneous display. The forced text-before-timestamp sequencing illustrates how violating domain-specific task decomposition creates persistent friction.

(3) Trustworthy AI requires role clarity: Educators rejected AI that replaced judgment but valued AI that accelerated workflow. Design principle: agentic AI should operate on educator-authored artifacts, not generate assessments de novo. Real-time prescription erodes trust; background augmentation preserves authority.

(4) Standards enable and constrain: ISBAR ensured full coverage yet created "impossible scores" for edge cases. Reusable frameworks must include escape valves to preserve validity. The structure-vs-flexibility tension requires flexible instrumentation adapting to scenario diversity.

Five concrete refinements. Based on these insights, we identified five refinements that will be implemented in the next iteration: **(i)** resequenced annotation workflow supporting educators' natural observe-mark-articulate mental model by enabling Start button activation before text entry; **(ii)** scenario-aware ISBAR instrumentation providing behavioral anchors for consistent scoring plus dual modes (5-point global rating for formative use vs. 30-item checklist for summative evaluation); **(iii)** time-aware fields exposing temporal primitives (recognition timestamp, response latency, external prompting presence) as first-class scoring controls rather than collapsing them into ordinal scales; **(iv)** background AI assistance replacing real-time pop-ups with post-annotation summarization and error-detection alerts that preserve educators' original wording while accelerating report preparation; and **(v)** calibration training workshops using shared video clips to establish inter-rater reliability baselines before program-scale deployment, with periodic spot-checks to detect scoring drift.

6.5 Limitations and Future Work

Our single-site qualitative study (5 educators, 2 sessions) offers rich in-learning description but limited generalizability; findings should be interpreted as exploratory insights rather than confirmatory evidence. The study design prioritized ecological validity over experimental control, capturing authentic workflow integration challenges at the cost of statistical power. We did not measure inter-rater reliability quantitatively; educator concerns about consistency are perceptions rather than empirical agreement coefficients. Three research opportunities emerge: **(1) Multi-site validation:** Test whether temporal dimensions and AI acceptance patterns generalize across diverse scenarios, clinical specialties, and assessment frameworks. **(2) Longitudinal learning curves:** Determine if usability frictions persist with extended use or resolve as educators develop expertise with the interface. **(3) Inter-rater reliability studies:** Quantify whether time-locked annotations and calibration training improve scoring consistency compared to traditional methods, establishing psychometric foundations for program-scale deployment where aggregated assessments inform progression decisions.

6.6 Beyond Clinical Simulation: Transferability to Other MMLA Contexts

While our study focuses on clinical training, three findings appear transferable to high-stakes MMLA domains requiring professional judgment and accountability.

Design Tension 1 (Assistance vs. Authority) maps to contexts where expert interpretation cannot be delegated: teaching observation, sports coaching [44], workplace training [10]. The principle that AI should operate on human-authored artifacts, not generate assessments from scratch, likely generalizes wherever provenance matters [19, 25].

Design Tension 2 (Structure vs. Flexibility) parallels tensions in using rubrics for formative assessment across disciplines [36]. Our finding that rigid frameworks create "impossible scores" for edge cases echoes validity concerns: standardization improves consistency but may sacrifice nuance [41]. The underlying insight transfers: MMLA platforms that capture fine-grained temporal data must expose temporal primitives (event timestamps, intervals, sequences) as first-class interface elements, not collapse them into summary metrics. This design principle applies to any time-sensitive assessment domain: response latency in crisis management, turn-taking in collaborative learning [11], or cueing in embodied skill training [44].

Design Tension 3 (Evidence vs. Overload) is well-documented in dashboard design for learning analytics [38, 42]. Our specific insight, that progressive disclosure matching expert mental models reduces friction better than presenting all modalities simultaneously, extends prior work to video-heavy contexts. The usability issue we uncovered (forced text-before-timestamp sequencing contradicting educators' observe-then-mark mental model) illustrates a broader principle: MMLA interaction design should align with domain-specific task decomposition, not impose generic annotation-tool conventions. This likely transfers to other video analytics contexts where timing and sequencing matter [31].

Domain-specific elements such as clinical communication frameworks (ISBAR), scenario types, and regulatory requirements require local adaptation. However, the design tensions, resolution strategies, and underlying workflow principles appear robust across high-stakes MMLA contexts that demand auditability, professional authority, and temporal precision.

7 Conclusion

ClinVision operationalizes the end-to-end MMLA pipeline for simulation-based clinical training. Co-designed with seven educators and evaluated in-learning with five educators, the platform offers qualitative evidence suggesting that workflow integration, not analytics alone, may drive adoption in high-stakes settings. Triangulating usage logs, surveys, and focus-group analysis, we contribute MMLA insights along three dimensions: **(1) Addressing MMLA challenges in this context:** Synchronized playback enabled deliberate review; ISBAR interfaces ensured full rubric coverage but exposed needs for temporal fields and behavioral anchors; background AI assistance (not real-time prescription) preserved authority. **(2) Design tensions:** We surface assistance vs. authority, structure vs. flexibility, and evidence vs. overload, tensions that may transfer to other auditable MMLA contexts where professional judgment cannot be delegated. The design principles we identify (AI operating on human-authored artifacts, temporal primitives as first-class controls, progressive disclosure matching task decomposition) offer actionable guidance for MMLA researchers. **(3) Research agenda:** in-learning findings reveal interaction mismatches, temporal gaps in rubrics, and trust requirements that controlled studies rarely surface, framing multi-site, longitudinal, and inter-rater reliability studies needed to establish validity and impact. Looking forward, MMLA adoption depends not only on algorithmic sophistication but on human-centered design that preserves professional authority, makes evidence pathways transparent, and closes the loop from data collection to actionable feedback.

Acknowledgments

This publication has emanated from research supported in part by grants from Research Ireland under Grant [12-RC-2289-P2] and [18/CRT/6223] which is co-funded under the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[1] Sandra Abegglen, Alexander Krieg, Heidi Eigenmann, and Rudolf Greif. 2020. Objective Structured Assessment of Debriefing (OSAD) in simulation-based medical education: Translation and validation of the German version. *PLOS ONE* 15, 12 (2020), e0244816. doi:10.1371/journal.pone.0244816

[2] Abeer Alhaj ali and Elaine Miller. 2018. Effectiveness of Video-Assisted Debriefing in Health Education: An Integrative Review. *Journal of Nursing Education* 57 (01 2018), 14–20. doi:10.3928/01484834-20180102-04

[3] Sonal Arora, Maria Ahmed, John Paige, Debra Nestel, Jane Runnacles, Louise Hull, Ara Darzi, and Nick Sevdalis. 2012. Objective Structured Assessment of Debriefing: Bringing Science to the Art of Debriefing in Surgery. *Annals of Surgery* 256, 6 (2012), 982–988. doi:10.1097/SLA.0b013e3182610c91

[4] Paulo Blikstein. 2013. Multimodal Learning Analytics. In *Proceedings of the 3rd International Learning Analytics & Knowledge Conference (LAK '13)*. ACM, New York, NY, USA, 102–106. doi:10.1145/2460296.2460316

[5] Sara Bly and Elizabeth F. Churchill. 1999. Design through Matchmaking: Technology in Search of Users. *interactions* 6, 2 (1999), 23–31. doi:10.1145/296165.296174

[6] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. In *APA Handbook of Research Methods in Psychology, Vol. 2*. American Psychological Association, 57–71. doi:10.1037/13620-004

[7] Alan Chamberlain, Andy Crabtree, Tom Rodden, Matt Jones, and Yvonne Rogers. 2012. Research in the Wild: Understanding 'in the Wild' Approaches to Design and Development. In *Proceedings of the Designing Interactive Systems Conference (DIS '12)*. ACM, Newcastle, UK, 795–796. doi:10.1145/2317956.2318078

[8] Adam Cheng, Walter Eppich, Vincent Grant, Jonathan Sherbino, Benjamin Zendejas, and David A. Cook. 2014. Debriefing for technology-enhanced simulation: a systematic review and meta-analysis. *Medical Education* 48, 7 (2014), 657–666. doi:10.1111/medu.12432

[9] Andy Crabtree, Alan Chamberlain, Mark Davies, Kevin Glover, Steve Reeves, Tom Rodden, Peter Tolmie, and Matt Jones. 2013. Doing Innovation in the Wild. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI (CHItaly '13)*. ACM, Trento, Italy, 25:1–25:9. doi:10.1145/2499149.2499150

[10] Daniele Di Mitri, Jan Schneider, Marcus Specht, and Hendrik Drachsler. 2019. Detecting Mistakes in CPR Training with Multimodal Data and Neural Networks. *Sensors* 19, 14 (2019). doi:10.3390/s19143099

[11] Vanessa Echeverria, Roberto Martinez-Maldonado, and Simon Buckingham Shum. 2019. Towards Collaboration Translucence: Giving Meaning to Multimodal Group Data. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 1–16. doi:10.1145/3290605.3300269

[12] Ruth M. Fanning and David M. Gaba. 2007. The role of debriefing in simulation-based learning. *Simulation in Healthcare* 2, 2 (2007), 115–125. doi:10.1097/SIH.0b013e3180315539

[13] Gloria Fernandez-Nieto, Pengcheng An, Jian Zhao, Simon Buckingham Shum, and Roberto Martinez-Maldonado. 2022. Classroom Dandelions: Visualising Participant Position, Trajectory and Body Orientation Augments Teachers' Sensemaking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New York, NY, USA, Article 564, 17 pages. doi:10.1145/3491102.3517736

[14] Gloria Fernandez-Nieto, Roberto Martinez-Maldonado, Vanessa Echeverria, Kirsty Kitto, Pengcheng An, and Simon Buckingham Shum. 2021. What can analytics for teamwork proxemics reveal about positioning dynamics in clinical simulations? *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–24. doi:10.1145/3449284

[15] Cynthia L. Foronda, Susana Barroso, Vicky J.-H. Yeh, Karina Gattamorta, and Elizabeth Bauman. 2021. A Rubric to Measure Nurse-to-Physician Communication: A Pilot Study. *Clinical Simulation in Nursing* 50 (2021), 38–42. doi:10.1016/j.ecns.2020.09.005

[16] Michail Giannakos, Mutlu Cukurova, and Sofia Papavlasopoulou. 2022. Sensor-based analytics in education: Lessons learned from research in multimodal learning analytics. In *The Multimodal Learning Analytics Handbook*. Springer, Cham, 329–358. doi:10.1007/978-3-031-08076-0_13

[17] Kathleen M. Haig, Staci Sutton, and John Whittington. 2006. SBAR: A Shared Mental Model for Improving Communication Between Clinicians. *The Joint Commission Journal on Quality and Patient Safety* 32, 3 (2006), 167–175. doi:10.1016/S1553-7250(06)32022-3

[18] Swathi Jagannath, Neha Kamireddi, Katherine Ann Zellner, Randall S. Burd, Ivan Marsic, and Aleksandra Sarcevic. 2022. A Speech-Based Model for Tracking the Progression of Activities in Extreme Action Teamwork. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 6. 1–26. doi:10.1145/3512920

[19] Rogers Kaliisa, Ioana Jivet, and Paul Prinsloo. 2023. A checklist to guide the planning, designing, implementation, and evaluation of learning analytics dashboards. *International Journal of Educational Technology in Higher Education* 20, 28 (2023), 1–24. doi:10.1186/s41239-023-00394-6

[20] Jayden Khakurel and Kirsimarja Blomqvist. 2022. *Artificial Intelligence Augmenting Human Teams. A Systematic Literature Review on the Opportunities and Concerns*. 51–68. doi:10.1007/978-3-031-05643-7_4

[21] Gary Klein. 2008. Naturalistic decision making. *Human factors: : The Journal of the Human Factors and Ergonomics Society* 50 (06 2008), 456–60.

[22] Kristian Krogh, Margaret Bearman, and Debra Nestel. 2015. Expert Practice of Video-Assisted Debriefing: An Australian Qualitative Study. *Clinical Simulation in Nursing* 11, 3 (2015), 180–187. doi:10.1016/j.ecns.2015.01.003

[23] Lai Hoang Le, Hoang D. Nguyen, Martin Crane, and Tai Tan Mai. 2024. Multimedia learning analytics feedback in simulation-based training: A brief review. In *Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia* (Phuket, Thailand) *(AIQAM '24)*. Association for Computing Machinery, New York, NY, USA, 25–30. doi:10.1145/3643479.3662053

[24] Q Vera Liao and S Shyam Sundar. 2022. Designing for responsible trust in AI systems: A communication perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 1257–1268.

[25] Qinyi Liu and Mohammad Khalil. 2023. Understanding privacy and data protection issues in learning analytics using a systematic review. *British Journal of Educational Technology* 54, 6 (2023), 1715–1747. doi:10.1111/bjet.13388

[26] Roberto Martinez-Maldonado, Vanessa Echeverria, Gloria Fernandez-Nieto, Lixiang Yan, Linxuan Zhao, Riordan Alfredo, Xinyu Li, Samantha Dix, Hollie Jaggard, Rosie Wotherspoon, Abra Osborne, Simon Buckingham Shum, and Dragan Gašević. 2023. Lessons Learnt from a Multimodal Learning Analytics Deployment In-the-Wild. *ACM Transactions on Computer-Human Interaction* 31, 1 (2023), 8:1–8:41. doi:10.1145/3622784

[27] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23. doi:10.1145/3359174

[28] Shuai Mu, Qian Zhang, Yan Zhao, and et al. 2020. Multimodal Data Fusion in Learning Analytics. *Sensors* 20, 23 (2020), 6856. doi:10.3390/s20236856

[29] John Norcini and Vanessa Burch. 2007. Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Medical Teacher* 29, 9-10 (2007), 855–871. doi:10.1080/01421590701775453

[30] Xavier Ochoa. 2022. Multimodal Learning Analytics – Rationale, Process, Examples, and Direction. In *The Handbook of Learning Analytics (2nd ed.)*. SoLAR, 54–65. doi:10.18608/hla22.006

[31] Luis P. Prieto, Kshitij Sharma, Pierre Dillenbourg, and María Jesús. 2016. Teaching analytics: towards automatic extraction of orchestration graphs using wearable sensors. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (Edinburgh, United Kingdom) *(LAK '16)*. Association for Computing Machinery, New York, NY, USA, 148–157. doi:10.1145/2883851.2883927

[32] Shelly J. Reed, Claire M. Andrews, and Patricia Ravert. 2013. Debriefing Simulations: Comparison of Debriefing with Video and Debriefing Alone. *Clinical Simulation in Nursing* 9, 12 (2013), e585–e591. doi:10.1016/j.ecns.2013.05.007

[33] Peter Rich and Michael Hannafin. 2008. Video Annotation Tools. *Journal of Teacher Education - J TEACH EDUC* 60 (11 2008), 52–67. doi:10.1177/0022487108328486

[34] Yvonne Rogers, Kay Connelly, Lenore Tedesco, William Hazlewood, Andrew Kurtz, Robert E. Hall, Josh Hursey, and Tammy Toscos. 2007. Why It's Worth the Hassle: The Value of In-Situ Studies When Designing Ubicomp. In *UbiComp 2007: Ubiquitous Computing (Lecture Notes in Computer Science, Vol. 4717)*. Springer, Berlin, Heidelberg, 336–353. doi:10.1007/978-3-540-74853-3_20

[35] L.H. Rosvig, S. Lou, L. Hvidman, T. Manser, N. Uldbjerg, O. Kierkegaard, and L. Brogaard. 2023. Healthcare providers' perceptions and expectations of video-assisted debriefing of real-life obstetrical emergencies: a qualitative study from Denmark. *BMJ Open* 13, 3 (2023), e062950. doi:10.1136/bmjopen-2022-062950

[36] D. Royce Sadler. 2010. Beyond feedback: developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education* 35, 5 (2010), 535–550. doi:10.1080/02602930903541015

[37] Lambert W. T. Schuwirth and Cees P. M. van der Vleuten. 2011. Programmatic Assessment: From Assessment of Learning to Assessment for Learning. *Medical Teacher* 33, 6 (2011), 478–485. doi:10.3109/0142159X.2011.565828

[38] Beat A. Schwendimann, Maria Jesús Rodríguez-Triana, Andrii Vozniuk, Luis P. Prieto, Mina Shirvani Boroujeni, Adrian Holzer, Denis Gillet, and Pierre Dillenbourg. 2017. Perceiving Learning at a Glance: A Systematic Literature Review of Learning Dashboard Research. *IEEE Transactions on Learning Technologies* 10, 1 (2017), 30–41. doi:10.1109/TLT.2016.2599522

[39] Michael Seropian and Robert Lavey. 2010. Design Considerations for Healthcare Simulation Facilities. *Simulation in healthcare : journal of the Society for Simulation in Healthcare* 5 (12 2010), 338–45. doi:10.1097/SIH.0b013e3181ec8f60

[40] Scott I Tannenbaum and Christopher P Cerasoli. 2013. Do team and individual debriefs enhance performance? A meta-analysis. *Human Factors* 55, 1 (2013), 231–245.

[41] Cees P. M. van der Vleuten, Lambert W. T. Schuwirth, Erik W. Driessen, Martijn J. B. Govaerts, and Sylvia Heeneman. 2012. A Model for Programmatic Assessment Fit for Purpose. *Medical Teacher* 34, 3 (2012), 205–214. doi:10.3109/0142159X.2012.652239

[42] Katrien Verbert, Sten Govaerts, Erik Duval, José Luis Santos, Frans Van Assche, Gonzalo Parra, and Joris Klerkx. 2020. Learning dashboards: an overview and future research opportunities. *Personal and Ubiquitous Computing* 18, 6 (2020), 1499–1514. doi:10.1007/s00779-013-0751-2

[43] Kerrin E. Weiss, Michaela Kolbe, Andrina Nef, Bastian Grande, Bravin Kalirajan, Mirko Meboldt, and Quentin Lohmeyer. 2023. Data-driven resuscitation training using pose estimation. *Advances in Simulation* 8, 1 (2023), 12. doi:10.1186/s41077-023-00251-6

[44] Marcelo Bonilla Worsley. 2018. Multimodal Learning Analytics' Past, Present, and Potential Futures. In *CrossMMLA@LAK*. https://api.semanticscholar.org/CorpusID:52075421

[45] Lixiang Yan, Linxuan Zhao, Dragan Gašević, and Roberto Martinez-Maldonado. 2022. Scalability, Sustainability, and Ethicality of Multimodal Learning Analytics. In *Proceedings of the 12th International Learning Analytics and Knowledge Conference (LAK '22)*. ACM, Online, USA, 13–23. doi:10.1145/3506860.3506862

[46] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13.

[47] Hui Zhang, Evalotte Mörelius, Sam Hong Li Goh, and Wenru Wang. 2019. Effectiveness of Video-Assisted Debriefing in Simulation-Based Health Professions Education: A Systematic Review of Quantitative Evidence. *Nurse Educator* 44, 3 (2019), E1–E6. doi:10.1097/NNE.0000000000000562

[48] Linxuan Zhao, Lixiang Yan, Dragan Gašević, Samantha Dix, Hollie Jaggard, Rosie Wotherspoon, Riordan Alfredo, Xinyu Li, and Roberto Martinez-Maldonado. 2022. Modelling Co-Located Team Communication from Voice Detection and Positioning Data in Healthcare Simulation. In *Proceedings of the 12th International Learning Analytics and Knowledge Conference (LAK22)*. ACM, New York, NY, USA, 370–380. doi:10.1145/3506860.3506935