Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Adaptive Medical Virtual Simulation Training System Based on Deep Reinforcement Learning: Multi-modal Learner Modelling and Empirical Evaluation of Pedagogical Effectiveness

Wang Chaoya

wcy_2003@163.com

Anhui Health College

Research Article

Keywords: medical education, virtual simulation, deep reinforcement learning, adaptive learning, multi-modal learner modelling, objective structured clinical examination

Posted Date: June 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9662209/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background. Virtual simulation (VS) is now a structural component of contemporary medical education, yet the predominant non-adaptive delivery model fails to accommodate substantial heterogeneity in learners' prior knowledge, working memory load, and engagement. Deep reinforcement learning (DRL) provides a principled framework for sequential pedagogical decision-making, but rigorous empirical evaluations in immersive medical training remain scarce. Methods. An adaptive VS training system was developed in which a multi-modal learner model (Transformer encoder with LSTM context aggregation) integrates behavioural logs, real-time performance, eye-tracking, heart-rate variability, and self-report data, while a Proximal Policy Optimization (PPO) agent selects scenario difficulty, scaffolding, and feedback granularity. A pre-registered 12-week, three-arm, parallel-group, single-blind randomised controlled trial enrolled 240 nursing and clinical-medicine students at one vocational health college in central China, comparing the DRL-adaptive system (n = 80) with a rule-based adaptive comparator (n = 80) and conventional non-adaptive VS (n = 80). The primary outcome was post-intervention objective structured clinical examination (OSCE) total score; secondary outcomes included a 60-item knowledge test, NASA-TLX cognitive load, the Intrinsic Motivation Inventory (IMI), the General Self-Efficacy scale (GSE), and three-month knowledge retention. Results. The DRL-adaptive group achieved higher OSCE scores (82.4 ± 7.6) than the rule-based (78.6 ± 8.0) and conventional groups (73.8 ± 8.4; $F(2,229) = 21.4$, $p < 0.001$, η^2 = 0.158, 95% CI 0.084–0.232). Cognitive load was lower (46.2 ± 7.8 vs 56.4 ± 9.1; d = − 1.20, 95% CI − 1.54 to − 0.86) and engagement higher (5.09 ± 0.62 vs 4.32 ± 0.78; d = 1.10). Three-month knowledge retention favoured the DRL arm (78.6 vs 65.4; d = 1.59). Pairwise contrasts between the DRL and rule-based arms were medium-to-large (d ≈ 0.5–0.9). Conclusion. A DRL-driven adaptive VS system produced clinically meaningful improvements in competency, cognitive load, engagement and durable retention, exceeding the gains achievable through rule-based personalisation alone. Reinforcement-learning pedagogical agents warrant cautious integration into competency-based medical curricula.

1. Introduction

Technology-enhanced simulation has reshaped clinical education by providing a psychologically safe context for the deliberate practice of skills that would otherwise carry unacceptable patient risk [1, 2]. Meta-analytic evidence from over six hundred studies indicates moderate-to-large effects on knowledge, skills, and behaviour relative to no intervention [1]. Within this category, immersive virtual simulation has gained particular traction because of its scalability, reproducibility, and capacity to render rare or high-acuity events on demand [3, 4]. The dominant pedagogical model nevertheless continues to deliver an essentially fixed sequence of scenarios, with limited capacity to respond to within-learner variation in prior knowledge, momentary cognitive load, or affective state.

The pedagogical limits of non-adaptive delivery are well articulated. Cognitive load theory predicts that a uniform difficulty curve under-challenges advanced learners while overloading novices, eroding learning efficiency at both extremes [5]. Deliberate-practice theory further argues that meaningful skill acquisition requires tasks calibrated just beyond current competence, a precondition that static curricula seldom

meet [6, 7]. Studies of intelligent tutoring have shown that adaptive difficulty matched to evolving learner state can substantially accelerate skill acquisition relative to fixed curricula [8]. Translating these insights into immersive medical environments has, however, proven non-trivial: high-dimensional learner states and the long horizon of clinical scenarios complicate the design of effective adaptation policies.

Rule-based adaptive systems—those that branch on hand-engineered thresholds derived from a small number of performance variables—represent the current state of practice and have produced modest pedagogical gains [9, 10]. Their reliance on expert-authored rules restricts both the granularity and dimensionality of personalisation; the cost of authoring grows combinatorially with the number of contextual signals considered. Reinforcement learning circumvents this bottleneck by learning a control policy directly from interaction data, framing pedagogical decision-making as a sequential optimisation problem [11]. The maturation of deep reinforcement learning, particularly policy-gradient methods such as Proximal Policy Optimization, has demonstrated stable performance in high-dimensional control tasks and offers a credible path toward scalable adaptive pedagogy [12, 13].

Existing applications of reinforcement learning to instructional sequencing have, with few exceptions, been confined to discrete-skill domains such as algebra or vocabulary, in which learner state is well-characterised by a small set of mastery probabilities [11, 14]. Medical training poses a qualitatively different challenge: learner state is multi-modal—encompassing performance traces, gaze-derived attention indicators, autonomic markers, and self-reported affect—and these signals evolve on different time scales. Whether a DRL agent equipped with a multi-modal learner representation can deliver pedagogically meaningful improvements over rule-based personalisation in immersive medical simulation is, to our knowledge, an open empirical question. Three contributions are made: an adaptive VS training system in which a Transformer–LSTM multi-modal encoder estimates a continuous latent representation of the learner consumed by a PPO-based decision agent (Fig. 1); a pre-registered, three-arm randomised controlled trial against rule-based and conventional comparators; and a critical analysis of the conditions under which DRL-mediated adaptation outperforms rule-based personalisation.

2. Methods

2.1 System architecture and multi-modal learner model

The training platform was implemented in Unity 2022.3 with a server-side adaptive engine in PyTorch 2.1, deployed on the institutional learning network. Twelve clinical scenarios were authored across emergency, surgical, and bedside-care domains; each contained a branching script of decision points with parameterised difficulty (number of distractor cues, time pressure, signal-to-noise ratio of vital-sign data) and scaffolding (verbal hints, expert demonstrations, structured checklists). The macro-architecture is depicted in Fig. 1.

Five input streams were sampled at task-appropriate frequencies: behavioural logs (mouse and controller traces, decision latency at 30 Hz), performance metrics (correctness, sequence adherence at

each decision point), eye-tracking (fixation duration and saccade amplitude via a Tobii Pro Nano at 60 Hz), heart-rate variability (RMSSD and pNN50 from a Polar H10 chest strap, 5-minute rolling windows), and end-of-scenario self-reports. A six-layer Transformer encoder processed the high-frequency channels, and a single-layer LSTM aggregated cross-scenario context. The fused representation was projected into four interpretable latent dimensions—knowledge state, skill proficiency, cognitive load, and engagement—via a multi-task linear head trained jointly on supervised labels obtained during a pilot calibration cohort (n = 60). Hyperparameters were selected by five-fold cross-validation; the final model achieved a knowledge-state RMSE of 0.087 (on a 0–1 scale) and a cognitive-load classification AUC of 0.88 against held-out NASA-TLX labels.

2.2 Deep reinforcement learning framework

Pedagogical decision-making was formalised as a partially observable Markov decision process. The state vector s_t comprised the four latent dimensions concatenated with scenario-level metadata, yielding a 28-dimensional input. The action space contained eleven discrete options spanning difficulty levels (five tiers), scaffolding modes (four types), and feedback granularity (two depths). The reward function combined immediate, terminal, and shaping components: $r_t = \alpha \cdot \Delta\text{skill} + \beta \cdot \text{engagement} - \gamma \cdot \max(0, \text{load} - \tau)$, with weights $\alpha = 0.55$, $\beta = 0.25$, $\gamma = 0.20$ and a cognitive-load tolerance $\tau = 0.65$, calibrated against pilot expert-rated session effectiveness (Spearman $\rho = 0.72$). PPO with generalised advantage estimation ($\lambda = 0.95$, $\gamma_{disc} = 0.99$) was first trained on a synthetic-cohort simulator distilled from pilot data and subsequently fine-tuned online with conservative trust-region updates and an entropy bonus of 0.01. Mutual information between cognitive-load latents and selected actions was estimated post-hoc with the Kraskov–Stögbauer–Grassberger estimator ($k = 5$) [15], with bootstrap 95% CIs.

2.3 Study design and participants

A 12-week, three-arm, parallel-group, single-blind randomised controlled trial was conducted between September 2024 and February 2025 at one vocational health college in central China. Eligible participants were second- and third-year nursing or clinical-medicine students with at least eight hours of prior simulation exposure and no diagnosed visual or cardiovascular condition that would invalidate the physiological measurements. The protocol was approved by the institutional ethics committee; written informed consent was obtained from all participants. Power analysis assuming a between-group standardised mean difference of $d = 0.45$ on the primary OSCE outcome (corresponding to $f = 0.225$), $\alpha = 0.05$ (two-tailed), and $1 - \beta = 0.90$ yielded a minimum total sample of 207 across three arms; the target was inflated to 240 to accommodate up to 15% attrition. Randomisation used a computer-generated permuted-block sequence with stratification by programme and baseline OSCE quartile; assessors and OSCE examiners were blinded to allocation.

2.4 Intervention and outcomes

All three arms received forty-eight hours of curriculum-matched simulation across twelve weekly sessions of four hours each. The conventional arm followed a fixed pre-scripted scenario order; the rule-based comparator implemented the institutional adaptive protocol that branches on cumulative

correctness thresholds drawn from prior published rule sets [9]; the DRL arm received scenarios and scaffolding selected online by the trained policy. The primary outcome was OSCE total score at week 12, scored by two faculty examiners blinded to allocation using a 12-station rubric covering clinical reasoning, procedural skill, communication, and patient safety (inter-rater ICC = 0.86, weighted-kappa 0.81). Secondary outcomes included a 60-item multiple-choice knowledge test with parallel forms at baseline and follow-up, the NASA Task Load Index [16], the Intrinsic Motivation Inventory [17], the 10-item General Self-Efficacy scale [22], and a three-month retention assessment combining a parallel-form knowledge test and a four-station OSCE.

2.5 Statistical analysis

Analyses followed the intention-to-treat principle and were conducted in R 4.3.1 (lme4, emmeans, effectsize). Group differences on continuous outcomes were tested with one-way analysis of covariance adjusting for the baseline value of the corresponding measure and stratification factors; pairwise contrasts were Bonferroni-corrected. Longitudinal trajectories of skill mastery and cognitive load were modelled with linear mixed-effects models including random intercepts and slopes per participant. Effect sizes are reported as Cohen's d with bias-corrected and accelerated bootstrap 95% confidence intervals (10,000 resamples). Two-sided $p < 0.05$ was deemed statistically significant; missing data (3.3% overall) were addressed with multiple imputation by chained equations ($m = 20$).

3. Results

3.1 Participant flow and baseline characteristics

Of 247 students screened, 240 were randomised; 232 (96.7%) completed the 12-week intervention and 226 (94.2%) attended the three-month retention assessment. Loss to follow-up did not differ across arms ($\chi^2(2) = 0.41$, $p = 0.81$). Demographic and baseline pedagogical variables were balanced (Table 1), with no between-group difference exceeding chance expectation.

Table 1
Baseline demographic and pedagogical characteristics by allocation.

Characteristic	Conventional VS (n = 78)	Rule-based (n = 77)	DRL adaptive (n = 77)	p
Age, years (mean ± SD)	20.4 ± 1.2	20.6 ± 1.3	20.5 ± 1.1	0.62
Female, n (%)	61 (78.2)	63 (81.8)	62 (80.5)	0.85
Programme, n (%)				
Nursing	47 (60.3)	46 (59.7)	47 (61.0)	0.99
Clinical medicine	31 (39.7)	31 (40.3)	30 (39.0)	—
Prior simulation hours	12.6 ± 4.8	13.1 ± 5.2	12.9 ± 4.9	0.81
Baseline OSCE	54.2 ± 7.1	53.8 ± 7.4	54.5 ± 6.9	0.87
Baseline knowledge	57.3 ± 8.5	58.0 ± 8.6	57.6 ± 8.2	0.86
Self-efficacy (GSE, 10–40)	28.4 ± 4.5	28.9 ± 4.7	28.7 ± 4.3	0.78

Note. Values are mean ± SD or n (%). p-values from one-way ANOVA or χ^2 as appropriate. GSE, General Self-Efficacy scale (range 10–40); VS, virtual simulation; DRL, deep reinforcement learning.

3.2 Primary and secondary outcomes

Mean OSCE total scores at week 12 were 82.4 ± 7.6, 78.6 ± 8.0, and 73.8 ± 8.4 in the DRL, rule-based, and conventional arms respectively (Fig. 2A). After adjustment for baseline, the omnibus effect was substantial ($F(2,229) = 21.4$, $p < 0.001$, $\eta^2 = 0.158$, 95% CI 0.084–0.232), and all three pairwise contrasts were statistically significant after Bonferroni correction. Sub-domain analyses (Fig. 2B) revealed that the DRL advantage was largest for patient safety ($\Delta = 8.8$ points versus conventional, $p < 0.01$) and clinical reasoning ($\Delta = 9.3$ points, $p < 0.01$), with comparable substantial gains for procedural skill ($\Delta = 8.3$ points, $p < 0.01$) and a smaller but still significant gain for communication ($\Delta = 6.1$ points, $p < 0.01$); this pattern is consistent with the policy's tendency to allocate scaffolded practice to error-prone decision junctures rather than to procedural rote elements.

Secondary outcomes converged on the same pattern (Table 2). Knowledge test, procedural skill, engagement, and self-efficacy all favoured DRL adaptation with medium-to-large effect sizes; the cognitive-load reduction was particularly notable, with NASA-TLX scores 10.2 points below the conventional arm and 3.9 points below the rule-based arm. The substantially higher engagement observed in the DRL arm was not accompanied by elevated workload, contradicting the lay intuition that more challenging adaptive sequences should impose heavier cognitive demands.

Table 2
Primary and secondary outcomes at the end of the 12-week intervention.

Outcome (week 12)	Conventional VS	Rule-based	DRL adaptive	F (2,229)	η²□ (95% CI)
OSCE total	73.8 ± 8.4	78.6 ± 8.0	82.4 ± 7.6	21.4***	0.158 (0.084–0.232)
Δfrom baseline	19.6 ± 7.2	24.8 ± 6.7	27.9 ± 6.3	28.1***	0.197 (0.119–0.272)
Knowledge test (60 items)	74.2 ± 8.4	79.5 ± 7.6	83.8 ± 7.1	28.6***	0.200 (0.123–0.275)
Procedural skill	75.4 ± 7.9	79.8 ± 7.6	83.7 ± 7.1	23.9***	0.173 (0.097–0.247)
NASA-TLX (lower better)	56.4 ± 9.1	50.1 ± 8.4	46.2 ± 7.8	28.5***	0.199 (0.122–0.275)
Engagement (IMI, 1–7)	4.32 ± 0.78	4.71 ± 0.69	5.09 ± 0.62	23.7***	0.171 (0.095–0.245)
Self-efficacy (GSE)	32.1 ± 4.6	33.9 ± 4.4	35.6 ± 4.0	13.2***	0.103 (0.041–0.171)
Knowledge retention (3 mo)	65.4 ± 9.0	71.5 ± 8.4	78.6 ± 7.6	46.2***	0.290 (0.207–0.366)

Note. *** $p < 0.001$ (omnibus ANCOVA adjusted for baseline value of the outcome). η²□, partial eta-squared with 95% confidence interval from the non-central F distribution. IMI scored on a 1–7 scale; lower NASA-TLX scores indicate lower workload.

3.3 Skill acquisition trajectories and cognitive load

Longitudinal modelling of weekly skill-mastery indices revealed a steeper acquisition slope for the DRL arm (β_DRL = 2.32 points/week, 95% CI 2.10–2.54) compared with the rule-based (β_Rule = 1.95, 1.74–2.16) and conventional arms (β_Conv = 1.55, 1.34–1.76; group × time interaction $F(2,229) = 18.6$, $p < 0.001$; Fig. 3A). Half of the DRL group's eventual cumulative gain had been attained by week 6, whereas the conventional arm reached the same threshold only at week 9, suggesting an approximate 33% acceleration of the learning curve. Cognitive-load trajectories diverged within the first three weeks (Fig. 3B): the DRL group settled within the productive band of the NASA-TLX (40–55) by week 4 and remained there for the duration of the intervention, whereas the conventional group exhibited a slow rise in workload through week 3 followed by a plateau in the high-strain region, consistent with cognitive load theory's prediction of element-interactivity overload under uniformly difficult curricula [5].

3.4 Multi-dimensional outcome profile and durable retention

A composite radar profile across six outcome dimensions visualised the breadth rather than the mere magnitude of the DRL advantage (Fig. 4A). The DRL arm exceeded comparators on every dimension; the gap was largest for engagement and knowledge retention, and narrowest for procedural skill, where even non-adaptive practice yields substantial improvement. At the three-month retention assessment, the rank ordering of arms was preserved with comparatively little decay in DRL performance (78.6 ± 7.6), whereas the conventional arm exhibited substantial regression toward baseline (65.4 ± 9.0; Fig. 4B). The DRL versus conventional retention contrast (d = 1.59, 95% CI 1.24–1.94) was the largest effect observed in the trial and is consistent with the upper range of effects reported in the technology-enhanced simulation literature [1, 2].

Pairwise standardised effect sizes are summarised in Table 3 to facilitate comparison with the broader literature. The DRL versus conventional contrast yielded effect sizes consistently above d = 1.0 across primary competency outcomes, while the DRL versus rule-based contrast remained in the medium range (d ≈ 0.5–0.9), indicating that the DRL advantage cannot be attributed merely to the presence of any adaptive mechanism.

Table 3
Pairwise standardised effect sizes (Cohen's d, 95% bootstrap CI).

Outcome	DRL vs Conventional d (95% CI)	DRL vs Rule-based d (95% CI)	Rule-based vs Conv. d (95% CI)
OSCE total	1.07 (0.74–1.40)	0.49 (0.18–0.81)	0.59 (0.27–0.91)
Knowledge test	1.24 (0.91–1.58)	0.59 (0.26–0.91)	0.66 (0.34–0.99)
Procedural skill	1.11 (0.78–1.44)	0.53 (0.21–0.85)	0.57 (0.25–0.89)
Cognitive load (NASA-TLX)	−1.20 (− 1.54 to − 0.86)	−0.48 (− 0.80 to − 0.16)	−0.72 (− 1.05 to − 0.40)
Engagement (IMI)	1.10 (0.77–1.43)	0.58 (0.26–0.91)	0.53 (0.21–0.85)
Knowledge retention (3 mo)	1.59 (1.24–1.94)	0.88 (0.55–1.21)	0.70 (0.38–1.02)

Note. Negative effect sizes for cognitive load indicate reduction. Confidence intervals are bias-corrected and accelerated bootstrap intervals based on 10,000 resamples.

3.5 System adaptation behaviour

Inspection of the trained policy's action distribution revealed that, on average, 31.8% of episodes triggered scaffolding interventions, with a strong concentration in mid-difficulty scenarios where ambiguity in the learner state was greatest. Estimated mutual information between the cognitive-load latent and the selected action reached 0.42 ± 0.05 nats (bootstrap 95% CI 0.32–0.52), more than double the value for the rule-based comparator (0.18 ± 0.03 nats), reflecting the policy's capacity to integrate weak signals across modalities that hand-engineered thresholds discard. A leave-one-modality-out ablation showed that eye-tracking contributed the largest marginal predictive value (Δ = 2.6 OSCE

points), followed by performance traces (Δ = 2.2) and heart-rate variability (Δ = 1.1); self-report alone added 0.7 points, underscoring the limited utility of relying solely on learner-supplied feedback for instructional decisions.

4. Discussion

Three principal findings emerge from this trial. The DRL-adaptive VS system produced statistically and educationally meaningful improvements in clinical competency, knowledge retention, and motivational outcomes relative to both conventional and rule-based comparators. Cognitive load was reduced rather than amplified by the more challenging adaptive sequences, indicating that the policy successfully calibrated difficulty to within the productive band of working-memory engagement. Three-month retention favoured the DRL arm with the largest single effect observed in the study, suggesting that the gains were not artefacts of test-taking exposure or short-lived motivational arousal.

The magnitude of the DRL advantage over rule-based personalisation deserves scrutiny. Rule-based systems performed appreciably better than the static comparator, replicating prior literature on adaptive virtual simulation [9, 18]. The further increment from rule-based to DRL adaptation is consistent with the prediction that high-dimensional learner states require commensurately expressive policies to exploit; once the available signal extends beyond a handful of correctness rates, hand-engineered thresholds discard pedagogically useful information [11, 14]. The mutual-information analysis between cognitive-load latents and policy actions provides an interpretable mechanism for this advantage and suggests that integration of physiological signals, rather than the depth of the neural network alone, is the operative ingredient.

These findings sit within a broader debate about the role of artificial intelligence in medical education. Proponents argue that intelligent tutoring will democratise expert-level coaching at scale [4, 19]; sceptics counter that the field's track record is uneven and that the heterogeneity of clinical learning resists the kind of well-defined reward signal that has powered DRL successes in games and robotics [11, 20]. The present trial offers measured support to the optimistic position while validating several of the sceptics' caveats. The reward function used here is non-trivial—it required pilot calibration against expert ratings—and minor mis-specifications in early prototypes produced policies that maximised superficial engagement at the expense of skill acquisition. The training environment likewise depended on a high-quality simulator distilled from pilot data; deployment in institutions without comparable preparatory infrastructure would risk substantially diminished returns.

Several limitations temper generalisability. Recruitment from a single vocational health college constrains external validity, particularly given the relatively homogeneous baseline competence of the cohort; whether comparable effects would obtain among graduate-level trainees, where ceiling effects on procedural skills may attenuate gains, requires multi-centre replication. Blinded OSCE assessment mitigates expectancy bias but cannot fully eliminate it, since intervention-arm participants were aware of the immersive technology being used. The three-month retention assessment, although longer than that

of most published trials in this area, remains shorter than the durations relevant for clinical practice. Two implementation considerations also warrant explicit discussion. The interpretability of DRL policies remains incomplete, and clinicians-in-training, faculty, and accreditors are likely to demand auditable rationales for adaptive decisions before such systems can be approved for high-stakes assessment [21]; the attention-based architecture used here partially addresses this concern by exposing the time points and modalities that drove each action. Equity is the second concern: adaptive systems trained on one cohort may underperform for sub-groups under-represented in the training distribution, and the present study did not enrol learners with sensory or motor differences that might interact with eye-tracking and physiological inputs. Robustness audits across demographic strata should be a precondition for broader deployment.

5. Conclusion

An adaptive virtual simulation training system that couples a multi-modal Transformer–LSTM learner model with a Proximal Policy Optimization decision agent produced substantial improvements in clinical competency, cognitive load, engagement and three-month retention compared with rule-based and conventional VS. The simultaneous reduction of cognitive load alongside competency gains suggests that well-designed adaptive systems need not trade learner welfare against pedagogical ambition. These findings support cautious integration of reinforcement-learning-based adaptive engines into competency-based medical curricula, contingent on adequate pilot calibration, transparent communication of policy decisions, and equity-focused robustness testing across diverse learner populations.

Declarations

Ethics approval and consent to participate

The study protocol was reviewed and approved by the Ethics Committee of Anhui Health Vocational College. All procedures involving human participants were performed in accordance with the principles of the Declaration of Helsinki. Written informed consent was obtained from all participants prior to enrolment.

Consent for publication

Not applicable.

Competing interests

The author declares that he has no competing interests.

Funding

This work was supported by the Key Project for the Cultivation of Outstanding Young Teachers of the Anhui Provincial Department of Education (grant number YQZD2024099) and the Demonstrative Virtual Simulation Training Base for Health and Medical Education (grant number 2023fzsx041). The funding bodies had no role in the design of the study; the collection, analysis, or interpretation of data; or the preparation of the manuscript.

Author Contribution

wang chaoya wrote the main manuscript text and prepared all figures .All authors reviewed the manuscript.

Acknowledgement

The author thanks the students who participated in the trial and the pilot calibration cohort for their time and engagement, and gratefully acknowledges colleagues at the Experimental Training Center of Anhui Health Vocational College for their support with scenario authoring, platform deployment, and OSCE assessment logistics.

Data Availability

The datasets generated and analysed during the current study are not publicly available because they contain identifiable student academic and physiological records governed by institutional data-protection policies, but de-identified data and the trained model checkpoints are available from the corresponding author on reasonable request, subject to a data-use agreement approved by Anhui Health College.

References

1. Cook DA, Hatala R, Brydges R, Zendejas B, Szostek JH, Wang AT, Erwin PJ, Hamstra SJ. Technology-enhanced simulation for health professions education: a systematic review and meta-analysis. JAMA. 2011;306(9):978–88.
2. 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.
3. Kyaw BM, Saxena N, Posadzki P, Vseteckova J, Nikolaou CK, George PP, Divakar U, Masiello I, Kononowicz AA, Zary N. Tudor Car L. Virtual reality for health professions education: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res. 2019;21(1):e12959.

4. Pottle J. Virtual reality and the transformation of medical education. Future Healthc J. 2019;6(3):181–5.
5. Sweller J, van Merriënboer JJG, Paas F. Cognitive architecture and instructional design: 20 years later. Educ Psychol Rev. 2019;31(2):261–92.
6. Ericsson KA. Deliberate practice and acquisition of expert performance: a general overview. Acad Emerg Med. 2008;15(11):988–94.
7. Brydges R, Manzone J, Shanks D, Hatala R, Hamstra SJ, Zendejas B, Cook DA. Self-regulated learning in simulation-based training: a systematic review and meta-analysis. Med Educ. 2015;49(4):368–78.
8. VanLehn K. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ Psychol. 2011;46(4):197–221.
9. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, Tudor Car L, Carlstedt-Duke J, Car J, Zary N. Virtual patient simulations in health professions education: systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res. 2019;21(7):e14676.
10. Cook DA, Hamstra SJ, Brydges R, Zendejas B, Szostek JH, Wang AT, Erwin PJ, Hatala R. Comparative effectiveness of instructional design features in simulation-based education: systematic review and meta-analysis. Med Teach. 2013;35(1):e867–98.
11. Doroudi S, Aleven V, Brunskill E. Where's the reward? A review of reinforcement learning for instructional sequencing. Int J Artif Intell Educ. 2019;29(4):568–620.
12. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv preprint. 2017;arXiv:170706347.
13. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, et al. Human-level control through deep reinforcement learning. Nature. 2015;518(7540):529–33.
14. Corbett AT, Anderson JR. Knowledge tracing: modeling the acquisition of procedural knowledge. User Model User-Adapt Interact. 1995;4(4):253–78.
15. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Phys Rev E. 2004;69(6):066138.
16. Hart SG. NASA-Task Load Index (NASA-TLX); 20 years later. Proc Hum Factors Ergon Soc Annu Meet. 2006;50(9):904–8.
17. Ryan RM, Deci EL. Intrinsic and extrinsic motivations: classic definitions and new directions. Contemp Educ Psychol. 2000;25(1):54–67.
18. Chernikova O, Heitzmann N, Stadler M, Holzberger D, Seidel T, Fischer F. Simulation-based learning in higher education: a meta-analysis. Rev Educ Res. 2020;90(4):499–541.
19. Chen X, Xie H, Zou D, Hwang G-J. Application and theory gaps during the rise of artificial intelligence in education. Comput Educ Artif Intell. 2020;1:100002.

20. Khosravi H, Shum SB, Chen G, Conati C, Tsai Y-S, Kay J, Knight S, Martínez-Maldonado R, Sadiq S, Gašević D. Explainable artificial intelligence in education. Comput Educ Artif Intell. 2022;3:100074.
21. Holmes W, Porayska-Pomsta K, Holstein K, Sutherland E, Baker T, Shum SB, Santos OC, Rodrigo MT, Cukurova M, Bittencourt II, Koedinger KR. Ethics of AI in education: towards a community-wide framework. Int J Artif Intell Educ. 2022;32(3):504–26.
22. Schwarzer R, Jerusalem M. Generalized Self-Efficacy scale. In: Weinman J, Wright S, Johnston M, editors. Measures in health psychology: a user's portfolio. Causal and control beliefs. Windsor, UK: NFER-NELSON; 1995. pp. 35–7.

Figures

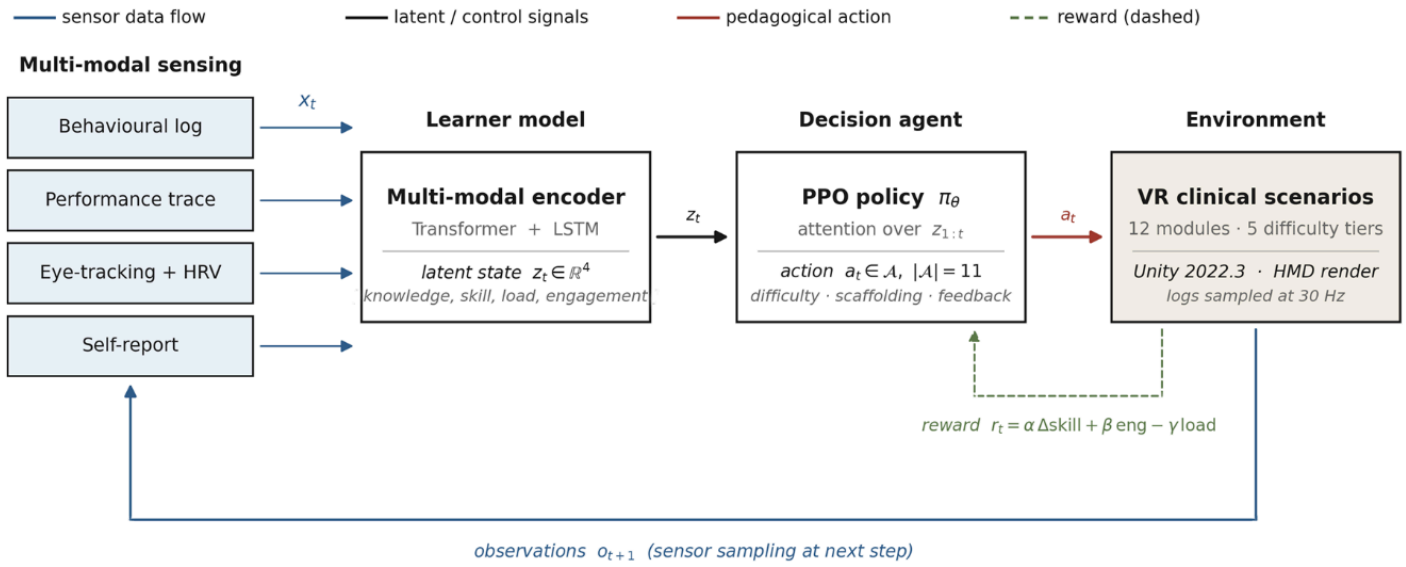


Figure 1

System architecture and information flow. Multi-modal sensor streams are encoded into a four-dimensional latent state z that is consumed by the PPO policy; pedagogical actions are executed in the VR environment; observations and reward signals close the control loop.

Figure 2. Comparative analysis of clinical competency outcomes across study arms
(** p < 0.01 versus Conventional VS, Bonferroni-corrected for four sub-domains)

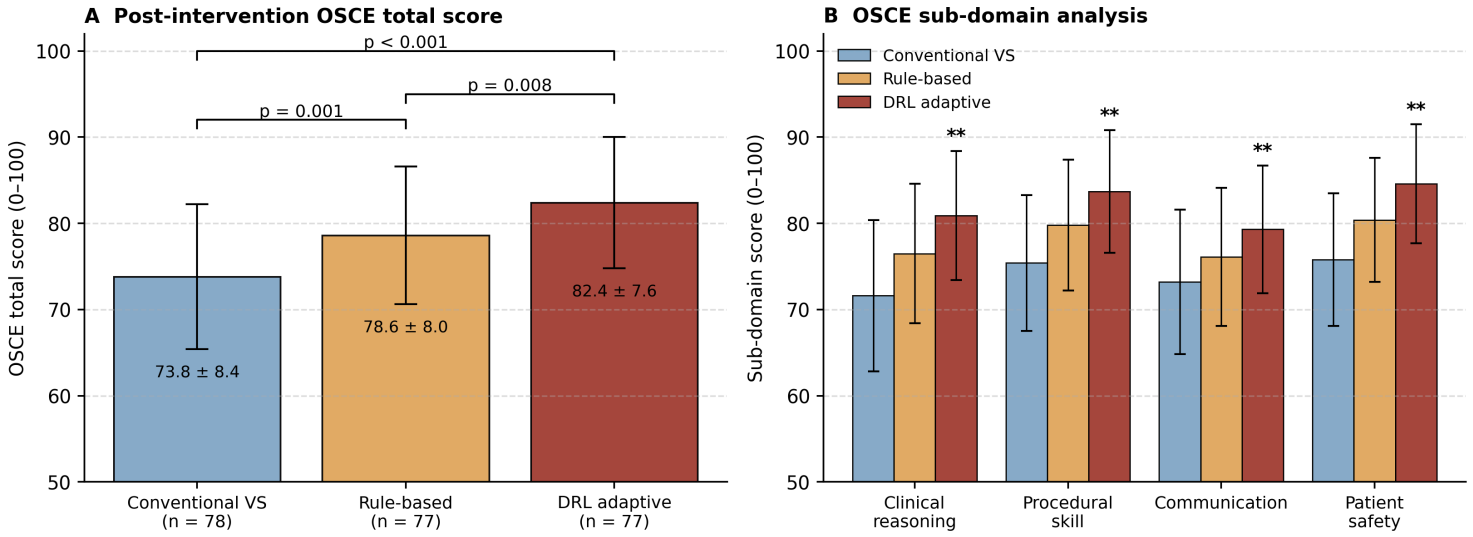


Figure 2

Comparative analysis of OSCE outcomes. (A) Total scores by allocation with Bonferroni-corrected pairwise p-values. (B) Sub-domain scores; ** p < 0.01 versus the conventional arm after Bonferroni correction across four sub-domains.

Figure 3. Longitudinal trajectories of skill mastery and cognitive load over the 12-week intervention

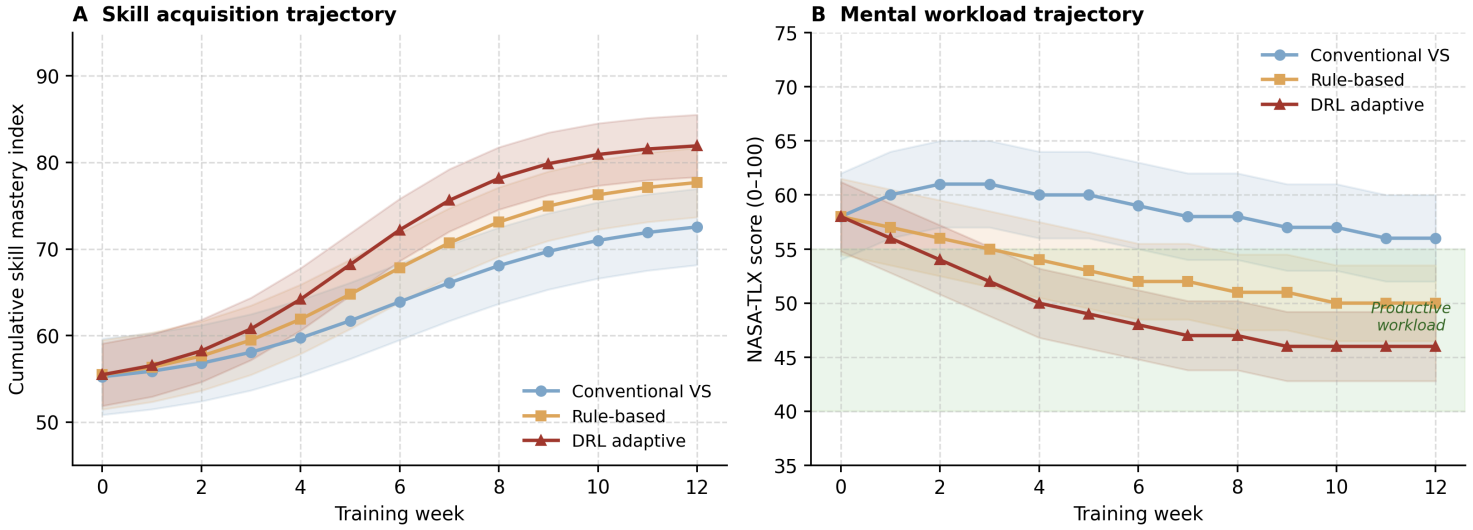


Figure 3

Longitudinal trajectories of (A) cumulative skill mastery and (B) NASA-TLX mental workload over the 12-week intervention. Shaded bands depict 95% confidence intervals; the green band in panel B indicates the empirically derived productive workload zone.

Figure 4. Composite outcome profile and 3-month knowledge retention

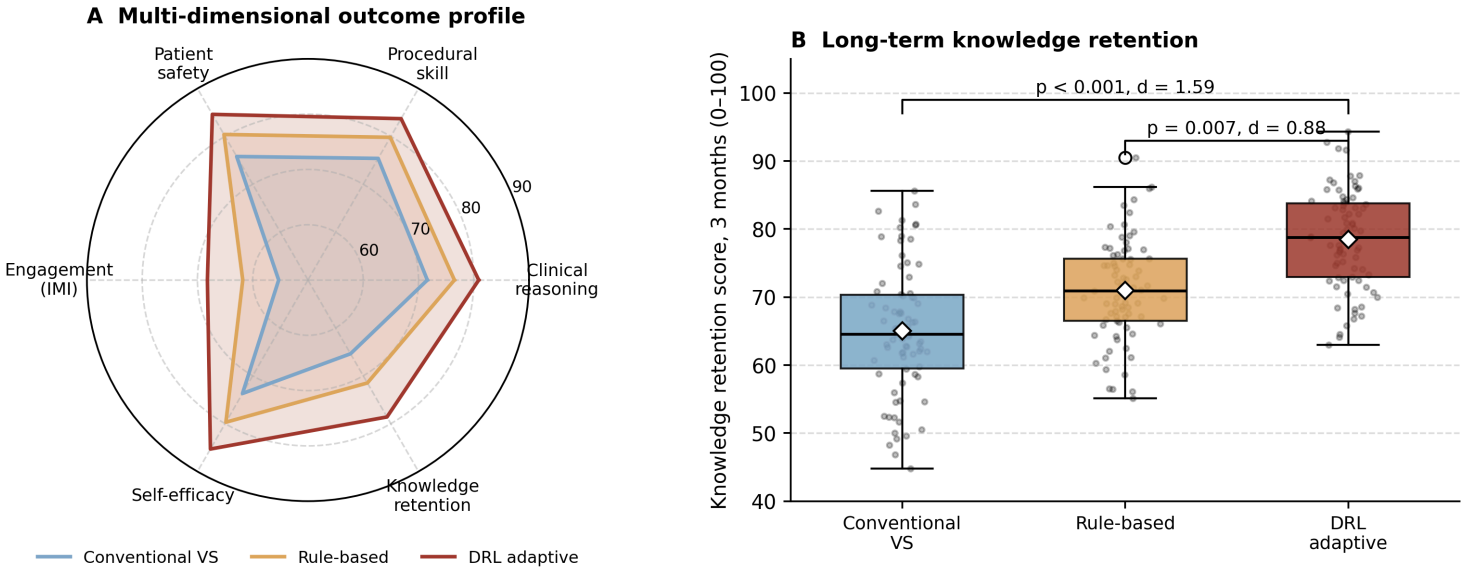


Figure 4

(A) Multi-dimensional outcome profile. OSCE sub-domain scores (clinical reasoning, procedural skill, patient safety) and three-month knowledge retention are shown on their native 0–100 scale; engagement (IMI, native 1–7) and self-efficacy (GSE, native 10–40) were linearly normalised to a 0–100 scale; the radial axis is truncated at 50–90 for readability. (B) Distribution of three-month retention scores; diamonds denote arithmetic means, horizontal lines denote medians.