scientific reports

Check for updates

OPEN

Cross-modal deep learning enhanced mixed reality accelerates construction skill transfer from experts to students

Chenyu Wang[1,2], Meng Zhu[2] & Sharifah Akman Syed Zakaria[1]✉

This study presents a novel approach for transferring expert tacit knowledge to construction students through a cross-modal deep learning supported mixed reality training system. Traditional construction training faces challenges in effectively capturing and transmitting the implicit expertise of experienced practitioners. We developed a framework that integrates multimodal data collection from construction experts with cross-modal deep learning algorithms to identify critical knowledge patterns and represent them in an immersive mixed reality environment. Experimental validation with 80 participants demonstrated that students trained using this system achieved 32.4% faster skill acquisition, 22.4% higher accuracy in construction tasks, and significantly better knowledge retention compared to traditional methods. The system particularly excelled in transferring situational awareness and adaptive problem-solving capabilities. Results indicate that cross-modal deep learning approaches can successfully formalize expert tacit knowledge into learnable representations, providing a scalable solution for construction education that addresses industry challenges of workforce development and knowledge preservation. Despite these promising results, the study acknowledges limitations including computational latency challenges during complex interactions and hardware ergonomic constraints during extended training sessions.

Keywords Cross-modal deep learning, Mixed reality, Construction training, Tacit knowledge, Knowledge transfer, Skill acquisition

Construction industry training faces significant challenges in the era of digital transformation, particularly in effectively transferring practical skills from experienced professionals to novice students. Traditional construction training methods often struggle with resource limitations, safety concerns, and the gap between theoretical knowledge and practical application[1]. The complexity of modern construction projects demands increasingly sophisticated skillsets from graduates, creating an urgent need for innovative training approaches that can meet industry demands while ensuring student safety[2].

One of the most critical challenges in construction education is the effective transmission of tacit knowledge—the implicit, experience-based knowledge that experts possess but often cannot articulate explicitly[3]. This tacit knowledge encompasses intuitive decision-making, situational awareness, and contextual understanding that is typically acquired through years of on-site experience. Traditional classroom-based instruction frequently fails to capture these nuanced aspects of construction expertise, resulting in graduates who possess theoretical knowledge but lack practical competence[4].

Mixed Reality (MR) technology has emerged as a promising solution for bridging the gap between theory and practice in construction education. Previous studies have demonstrated MR's effectiveness in construction safety training and equipment operation simulation, with early implementations showing 15–25% improvement in hazard identification accuracy compared to traditional methods[5]. By seamlessly blending virtual elements with the physical environment, MR creates immersive training scenarios that simulate real-world construction challenges without associated risks. When integrated with cross-modal deep learning algorithms, MR systems can analyze expert behaviors across multiple sensory dimensions, including visual, auditory, and kinesthetic inputs, to identify and codify tacit knowledge patterns that would otherwise remain inaccessible[6].

[1]School of Civil Engineering, University Sains Malaysia, Trankarian Nibong Tebal, 100032 Penang, Malaysia. [2]School of Architecture and Engineering, Jiaxing Nanhu University, JiaxingZhejiang 314000, China. ✉email: akmanusm@126.com

While previous research has explored multimodal approaches in vocational education through virtual reality environments, and deep learning techniques have been applied to knowledge transfer in other domains, the integration of cross-modal deep learning with mixed reality for tacit knowledge transfer in construction remains largely unexplored.

This research addresses the following key questions: (RQ1) How can cross-modal deep learning effectively capture and formalize expert tacit knowledge in construction? (RQ2) What is the comparative effectiveness of MR-based versus traditional training methods for construction skill acquisition? (RQ3) How can mixed reality environments optimize the transfer of tacit knowledge to improve long-term retention?

This research aims to develop and evaluate a novel cross-modal deep learning supported mixed reality training framework for construction education that effectively captures and transfers expert tacit knowledge to students. The framework utilizes multimodal data collection from expert practitioners, deep learning analysis to identify critical knowledge patterns, and mixed reality interfaces for immersive knowledge transfer[7]. This study is guided by three primary hypotheses: (H1) cross-modal MR training will significantly accelerate skill acquisition compared to traditional methods, (H2) knowledge transfer efficiency will correlate with expert model representation fidelity, and (H3) procedural knowledge retention will be superior in the experimental group.

The significance of this research extends beyond immediate educational outcomes to address broader industry challenges, including the aging workforce, skills shortages, and the need for rapid upskilling in response to technological advancement[8].

The remainder of this paper is organized as follows: Section II reviews relevant literature on construction training challenges, tacit knowledge transfer, and mixed reality applications. Section III presents the methodology for the proposed framework, while Section IV details the implementation and experiments. Section V presents results and discussion, followed by conclusions and future work directions in Section VI.

Theoretical foundation and literature review

Fundamentals of cross-modal deep learning

Cross-modal deep learning represents a significant advancement in artificial intelligence by enabling machines to process, interpret, and correlate information across different modalities such as visual, auditory, and textual data[9]. Unlike traditional unimodal approaches, cross-modal systems can establish meaningful connections between heterogeneous data sources, mimicking the human ability to integrate sensory information from multiple channels[10]. This capability is particularly valuable in construction training where expert knowledge manifests through diverse modalities including visual demonstrations, verbal explanations, and tactile guidance.

The core mathematical foundation of cross-modal learning involves mapping data from different modalities into a shared representation space where semantic relationships are preserved. Given input data from modalities X_1 and X_2, the objective is to learn embedding functions f_1 and f_2 that project the inputs into a common space:

$$Z_1 = f_1(X_1), Z_2 = f_2(X_2)$$

where the distance between related cross-modal pairs should be minimized while maximizing the distance between unrelated pairs[11]. The alignment loss function typically incorporates both modality-specific reconstruction terms and cross-modal similarity constraints:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{recon} + \lambda_2 \mathcal{L}_{cross} + \lambda_3 \mathcal{L}_{reg}$$

Multimodal data fusion represents a critical component of cross-modal learning systems, generally categorized into early, late, and hybrid fusion approaches[12]. Early fusion combines raw or preprocessed features before feeding them into a joint model, while late fusion integrates predictions from modality-specific models[13]. Hybrid fusion, increasingly common in state-of-the-art systems, employs attention mechanisms to dynamically weight modality contributions according to:

$$A(X_1, X_2) = \text{softmax}\left(\frac{Q(X_1)K(X_2)^T}{\sqrt{d_k}}\right)V(X_2)$$

where Q, K, and V represent query, key, and value transformations respectively[14].

Recent advances in cross-modal representation learning have been driven by self-supervised techniques that exploit inherent correlations between modalities without requiring extensive labeled data[15]. Contrastive learning has emerged as particularly effective, training models to maximize mutual information between corresponding cross-modal pairs while minimizing it between unrelated pairs[16]. The normalized temperature-scaled cross-entropy loss function commonly used in these approaches is defined as:

$$\mathcal{L}_{\mathcal{NCE}} = -\log\frac{\exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \backslash mathbbm1_{[k\neq i]}\exp(sim(z_i, z_k)/\tau)}$$

where sim represents cosine similarity and τ is a temperature parameter[17].

In educational contexts, cross-modal deep learning has demonstrated remarkable effectiveness for capturing and transmitting complex knowledge. For instance, systems analyzing synchronized videos of expert demonstrations with verbal explanations can identify critical moments where visual actions align with verbal emphasis[18]. Applications in medical training have shown that cross-modal systems can encode the subtle

relationships between visual diagnostic findings and tactile examination techniques[19]. Similarly, in engineering education, these systems have successfully captured the alignment between theoretical concepts and their practical implementations[20].

Construction domain expertise encompasses a substantial proportion of tacit knowledge—implicit understanding that is difficult to formalize and communicate through conventional means. This specialized tacit knowledge can be categorized into three primary dimensions: technical know-how (embodied skills and construction techniques), contextual awareness (site-specific adaptations and environmental interpretations), and social-collaborative knowledge (team coordination practices and communication protocols)[21]. The complex, multimodal nature of construction expertise presents significant challenges for traditional educational approaches, as approximately 67% of critical knowledge in experienced practitioners remains undocumented and primarily transferred through direct observation and guided practice[22].

The architectural approaches listed in Table 1 illustrate the evolution of cross-modal learning systems and their varying strengths in different educational contexts. These frameworks provide the technical foundation for capturing the multimodal nature of expert tacit knowledge in construction settings and transforming it into learnable representations for novice students through mixed reality environments.

Mixed reality technology applications in construction

The integration of Mixed Reality (MR) technology into construction education has evolved significantly over the past two decades, progressing from simple visualization tools to sophisticated interactive training environments[23]. Early implementations in the early 2000 s primarily focused on Augmented Reality (AR) for visualizing 3D building models and construction sequences, with limited interactivity and predominantly used in classroom settings rather than on-site[24]. The emergence of head-mounted displays (HMDs) around 2015 marked a pivotal advancement, enabling more immersive experiences where virtual building components could be overlaid onto physical spaces with increasing precision[25].

Currently, MR applications in construction site training encompass diverse functionalities including safety hazard identification, equipment operation simulation, assembly sequence training, and quality control procedures[26]. These systems typically employ spatial mapping algorithms that create a digital representation of the physical environment according to:

$$M\left(P\right)=\left\{\left(x_i,y_i,z_i,f_i\right)|i=1,...,n\right\}$$

where (x_i, y_i, z_i) represents the spatial coordinates and f_i the feature vector of each mapped point[27]. Industry adoption has grown substantially, with approximately 67% of large construction firms implementing some form of MR-based training by 2023, though implementation varies significantly in sophistication and scope[28].

Despite promising advancements, MR technology faces several critical challenges in construction applications that must be addressed for effective implementation. Hardware limitations including field of view restrictions, device weight, and battery life constraints significantly impact extended training sessions in dynamic construction environments[29]. The fidelity gap between virtual and physical interactions remains problematic, particularly for tasks requiring haptic feedback and fine motor skills[30]. The proposed cross-modal deep learning approach offers potential solutions to some of these challenges by optimizing content delivery based on user attention patterns and reducing computational overhead through intelligent scene management.

Additionally, computational demands for rendering complex construction scenarios with realistic physics can be expressed through the performance metric:

$$\eta=\frac{r\cdot c\cdot d}{p\cdot t}$$

where r represents rendering resolution, c the complexity of models, d the desired detail level, p the available processing power, and t the maximum acceptable latency threshold[31].

Comparative analysis between MR-based and traditional construction training reveals complementary strengths rather than a straightforward superiority of either approach. Traditional methods excel in providing authentic material interactions and environmental conditions but suffer from resource constraints, safety risks, and limited repeatability[32]. A mixed-methods study involving 142 construction management students demonstrated that MR-based training yielded 28% higher knowledge retention rates for safety procedures and

Model Name	Publication Year	Core Technology	Advantages	Application Scenarios
CLIP	2021	Contrastive Language-Image Pre-training	Zero-shot transfer capabilities, robust visual concept understanding	Visual education, image retrieval, content classification
MulT	2019	Multimodal Transformer	Effective cross-modal attention mechanisms, handles temporal data	Multimodal sentiment analysis, skill assessment, behavior modeling
MMBT	2019	Multimodal Bitransformer	Efficient fusion of text and visual information, compact representation	Educational content analysis, multimedia understanding
VideoXLM	2022	Cross-lingual Video-Text Pretraining	Multilingual capabilities, temporal awareness	Cross-cultural training, procedural knowledge transfer
ALBEF	2021	Aligned Contrastive Learning	Improved alignment between modalities, efficient training	Visual instruction, expert behavior modeling

Table 1. Comparison of cross-modal deep learning architectures.

34% improved performance in equipment operation compared to conventional approaches[33]. However, MR systems showed limited effectiveness in conveying tactile aspects of construction materials and environmental conditions such as weather impacts[34].

The integration of pedagogical frameworks with MR technology has emerged as a critical factor in successful implementation, with constructivist and experiential learning approaches showing particular promise when coupled with scaffolded MR experiences that gradually decrease guidance as learner competence increases[2]. This structured progression from guided to autonomous practice within MR environments provides a mechanism for addressing the scaffolding requirements that Vygotsky identified as essential for skill development, while maintaining the safety advantages of virtual training[35].

Research on tacit knowledge transfer in construction expertise

Given the substantial tacit knowledge components in construction expertise, effective capture and transfer methods become critical for educational advancement.

Traditional methods for capturing and transferring tacit construction knowledge have predominantly relied on apprenticeship models, communities of practice, and mentorship relationships[36]. These approaches, while effective in transmitting contextual nuances and embodied skills, are inherently limited by resource constraints, expert availability, and the inability to scale across large student populations[37]. The effectiveness of knowledge transfer in these traditional contexts can be expressed through Nonaka's SECI model (Socialization, Externalization, Combination, Internalization) as:

$$E_{transfer} = \sum_{i=1}^{4} w_i \cdot S_i \cdot C_i \cdot A_i$$

where S_i represents the strength of the SECI component, C_i the comprehensiveness of knowledge covered, A_i the accessibility to learners, and w_i the relative importance weighting for the specific construction skill[38].

Recent technological advances have enabled more systematic approaches to tacit knowledge capture and transfer. Sensor-based motion tracking systems can document expert movements during critical construction tasks, while eye-tracking technologies identify visual attention patterns that characterize expert problem-solving[39]. These multimodal data streams allow for quantification of previously intangible expertise elements, including optimal tool manipulation angles, pressure applications, and environmental scanning behaviors[40]. When structured through ontological frameworks, this captured knowledge can be organized according to:

$$K_{structured} = \bigcup_{j=1}^{m} \{C_j, R_j, I_j, P_j\}$$

where C_j represents concept classes, R_j the relationships between concepts, I_j the instances, and P_j the properties within the construction knowledge domain[41].

Evaluating the effectiveness of tacit knowledge transfer remains challenging due to the implicit nature of the transferred skills. Current assessment approaches typically combine objective performance metrics (task completion time, error rates, safety compliance) with subjective measures including expert evaluation of technique quality and learner self-efficacy reports[42]. Longitudinal studies indicate that technology-enhanced tacit knowledge transfer can accelerate expertise development by 37–42% compared to traditional methods, particularly for spatial reasoning and situational awareness aspects of construction expertise[43]. However, significant gaps remain in effectively transferring the intuitive decision-making abilities that characterize master-level construction practitioners, suggesting the need for more sophisticated cognitive modeling approaches that can capture the heuristic reasoning patterns underlying expert judgment in complex, ambiguous construction scenarios[44].

Cross-modal mixed reality training system design

Multimodal data collection and processing

The acquisition of expert tacit knowledge in construction requires a comprehensive multimodal data collection approach that captures the multidimensional nature of expert performance across visual, auditory, kinesthetic, and contextual domains[45]. Our system employs a synchronized multi-sensor framework to document expert behaviors during critical construction tasks, utilizing strategically positioned sensors that minimize interference with natural work patterns while maximizing data fidelity. The multimodal data collection setup encompasses wearable motion capture sensors, eye-tracking glasses, environmental cameras, directional microphones, and pressure sensors integrated into tools and materials as detailed in Table 2.

Sensor deployment follows a systematic protocol to ensure comprehensive coverage while maintaining practical usability in construction environments[46]. The primary challenge in multimodal data acquisition is maintaining temporal synchronization across heterogeneous sensor streams with varying sampling rates and latencies. We address this through a centralized timestamp server that applies the following synchronization algorithm:

$$T_{aligned}(i,j) = T_{raw}(i,j) + \Delta_i + \delta_{ij}$$

Equipment Type	Model	Sampling Rate	Accuracy	Data Format	Application Scenario
Motion Capture Suit	Xsens MVN Awinda	240 Hz	± 0.5° orientation, ± 2 cm position	BVH, C3D	Body movement, posture tracking
Eye-tracking Glasses	Tobii Pro Glasses 3	100 Hz	0.5° visual angle	CSV, JSON	Visual attention, gaze patterns
360° Camera	Insta360 Pro 2	30 fps	8 K resolution	MP4, RAW	Environmental context capture
Directional Microphone	Sennheiser MKH 416	48 kHz frequency response	SNR > 84dB	WAV, MP3	Verbal instructions, environmental sounds
Smart Tools	Milwaukee ONE-KEY	20 Hz	± 2% torque, ± 1° angle	JSON	Tool handling, application force
Pressure Sensors	Tekscan Grip	50 Hz	± 5% applied force	CSV	Grip strength, material handling

Table 2. Specifications of multimodal data collection equipment.

where $T_{aligned}\,(i,j)$ is the aligned timestamp for the j-th sample from sensor i, $T_{raw}\,(i,j)$ is the raw timestamp, Δ_i is the system-level offset for sensor i, and δ_{ij} is the dynamic correction factor determined through cross-correlation analysis of reference signals[47].

Data preprocessing involves modality-specific filtering, noise reduction, and normalization procedures. Motion data undergoes Kalman filtering to eliminate sensor drift according to:

$$\widehat{x}_k = F\widehat{x}_{k-1} + K_k\left(z_k - HF\widehat{x}_{k-1}\right)$$

where $\widehat{x}_k$ is the filtered state estimate, F is the state transition model, K_k is the Kalman gain, z_k is the measurement, and H is the observation model[48]. Audio signals are processed through adaptive filtering to isolate relevant verbal instructions from ambient construction noise, while visual data undergoes spatial registration to align with the coordinate system of motion sensors.

Cross-modal feature extraction employs a hierarchical approach that first extracts modality-specific features before integrating them into unified representations. Low-level features include kinematic parameters (joint angles, velocities, accelerations), visual attention metrics (fixation duration, saccade patterns), and acoustic characteristics (prosodic features, semantic content). The cross-modal feature integration utilizes a tensor fusion network defined as:

$$Z = \sum_{i=1}^{n}\alpha_i F_i + \sum_{i=1}^{n}\sum_{j=i+1}^{n}\beta_{ij}\left(F_i \otimes F_j\right) + \gamma\left(F_1 \otimes F_2 \otimes \ldots \otimes F_n\right)$$

where F_i represents features from modality i, $\otimes$ denotes the outer product, and α_i, β_{ij}, and γ are learnable weights for unimodal, bimodal, and multimodal interactions respectively[49].

To ensure data quality, we implement a multi-stage validation framework incorporating both automated and manual verification procedures. Automated quality control employs statistical anomaly detection based on the Mahalanobis distance criterion:

$$D_M\left(x\right) = \sqrt{\left(x-\mu\right)^T\Sigma^{-1}\left(x-\mu\right)}$$

where x is the feature vector, μ is the mean vector, and Σ is the covariance matrix of the reference distribution[50]. Segments exceeding threshold values are flagged for expert review. Additionally, periodic calibration procedures and redundant sensor placements ensure robustness against equipment failures or environmental interference, maintaining data integrity throughout extended collection sessions in challenging construction environments.

Mixed reality system architecture

The proposed mixed reality training system employs a layered architecture that integrates multimodal data processing, cross-modal learning, and mixed reality rendering components to create an immersive knowledge transfer environment for construction training[51]. Fig 1 illustrates the overall research methodology framework showing our four-phase approach from expert knowledge capture to validation.

The overall system architecture follows a modular design philosophy that prioritizes scalability, reconfigurability, and real-time performance while maintaining the fidelity necessary for effective tacit knowledge transfer. Fig 2 illustrates the system's hierarchical structure with bidirectional information flow between the physical sensing layer, the computational core, and the mixed reality presentation layer.

The hardware platform combines high-performance computing infrastructure with advanced mixed reality display technology to support the computational demands of cross-modal deep learning while delivering immersive visualization capabilities. Central to the hardware architecture is the processing server cluster that executes cross-modal feature extraction and knowledge mapping algorithms according to the processing pipeline defined by:

$$T_{total} = T_{sensing} + T_{preprocessing} + T_{inference} + T_{rendering}$$

where the total system latency T_{total} must be maintained below the perceptual threshold of 20ms to ensure seamless interaction[52]. The mixed reality headset selection criteria incorporated field of view requirements,

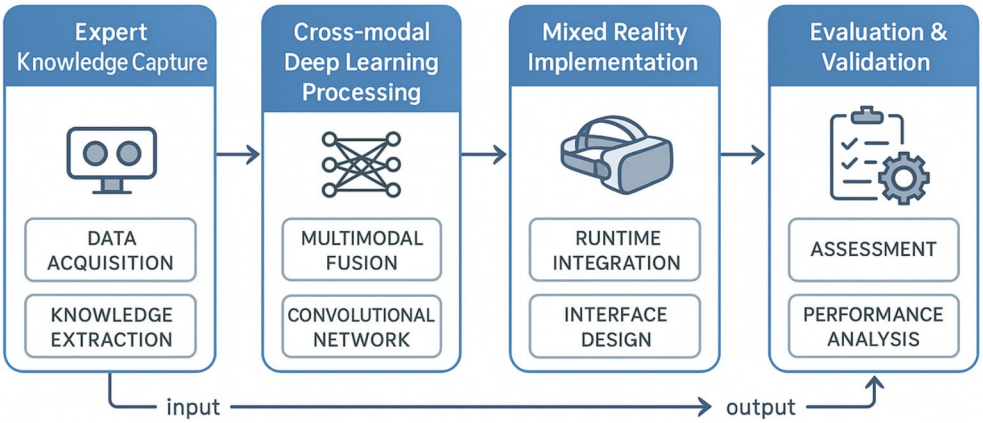


Fig. 1. Overall research methodology framework showing the four-phase approach: (1) Expert knowledge capture through multimodal sensing, (2) Cross-modal deep learning processing and knowledge modeling, (3) Mixed reality training system implementation, and (4) Comparative evaluation and validation.

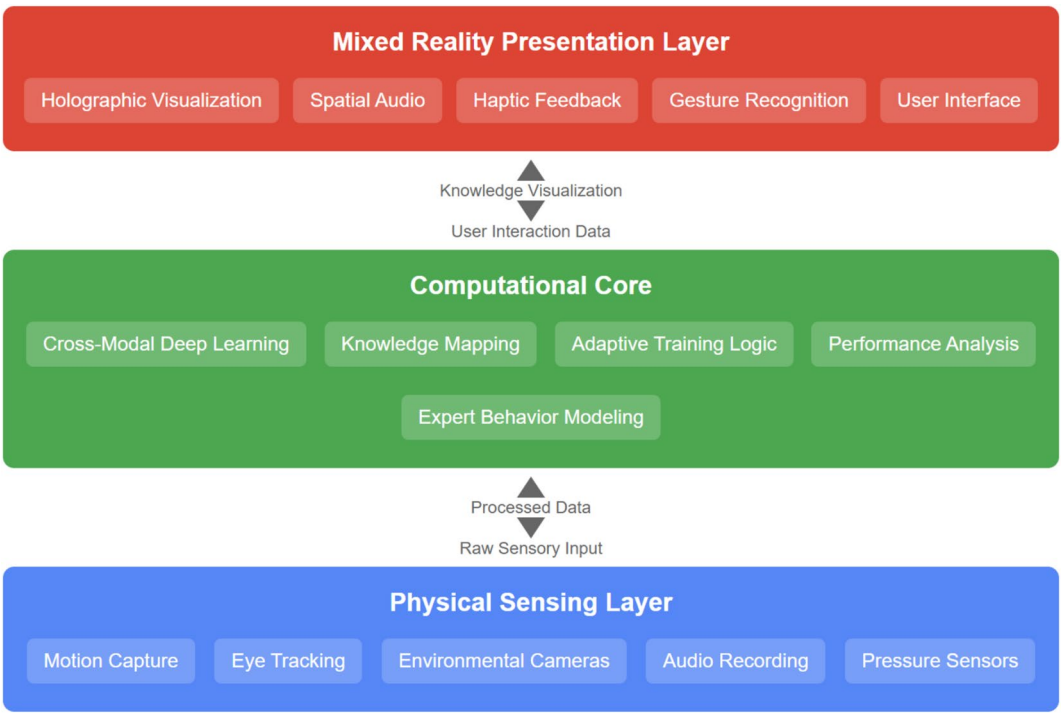


Fig. 2. Hierarchical structure of the mixed reality training system showing bidirectional information flow between the physical sensing layer (multimodal data collection), computational core (cross-modal deep learning processing), and mixed reality presentation layer (knowledge visualization and interaction).

display resolution, tracking accuracy, and ergonomic considerations for extended use in construction training scenarios.

The software architecture implements a microservices approach with specialized modules handling distinct functional responsibilities while communicating through standardized interfaces[53]. The primary functions and technical features of the mixed reality system components are summarized in Table 3. This modular design facilitates independent updating of system components and integration of emerging technologies without requiring complete system redesign. The computational workload distribution between on-device processing and cloud computing resources is dynamically adjusted according to:

$$W_{device} = \alpha \cdot C_{available} + \beta \cdot B_{network} + \gamma \cdot P_{remaining}$$

Component Name	Primary Functions	Technical Features	Interaction Methods
Perception Engine	Environmental mapping, Object recognition, Spatial anchoring	SLAM algorithms, CNN-based object detection, Mesh generation	REST API interfaces, Binary data streaming
Knowledge Core	Expert behavior modeling, Cross-modal representation, Skill decomposition	Tensor-based multimodal fusion, Attention mechanisms, Graph neural networks	JSON-RPC, WebSocket streams
Rendering System	Holographic visualization, Spatial audio, Haptic feedback	Physically-based rendering, Ambisonics audio, Force feedback	Direct3D pipeline, OpenXR compatibility
Interaction Manager	Gesture recognition, Voice commands, Gaze tracking	Hand pose estimation, NLP processing, Attention mapping	Event-driven communication, State synchronization

Table 3. Mixed reality system component functions.

where W_{device} represents the proportion of computation performed on the device, $C_{available}$ is the available computational capacity, $B_{network}$ is the network bandwidth, $P_{remaining}$ is the remaining power capacity, and α, β, and γ are weighting coefficients determined through system optimization[54].

The user interface design implements pedagogical scaffolding principles that progressively adjust guidance levels based on learner expertise[55]. The MR training environment integrates virtual construction elements with physical workspace through spatial mapping, enabling students to interact with holographic building components, safety indicators, and expert demonstration overlays. Visual elements follow a minimalist approach that highlights critical information without overwhelming the user's field of view, while spatial audio cues direct attention to relevant interaction points. The interface features attention-guided visual cues and scaffolded instruction panels that adapt to user proficiency levels. Information density is dynamically controlled according to:

$$I_{displayed} = I_{base} \cdot \left(1 - e^{-\lambda \cdot E_{user}}\right)$$

where $I_{displayed}$ is the information density, I_{base} is the maximum information density, E_{user} is the estimated user expertise level, and λ is an adaptation rate parameter[56].

System performance optimization incorporates both hardware acceleration and algorithmic efficiency improvements. Level-of-detail management for virtual objects, foveated rendering based on gaze tracking, and predictive pre-loading of training scenarios minimize rendering latency while maintaining visual quality in regions of user focus. Memory management strategies implement a hierarchical caching system that prioritizes frequently accessed expert demonstration segments while offloading less critical content to secondary storage. Continuous performance monitoring with adaptive resource allocation ensures consistent frame rates even during complex construction operations with multiple simultaneous interactions, maintaining the immersive quality essential for effective skill transfer from virtual demonstrations to physical practice.

Expert knowledge modeling and representation

The extraction and formalization of expert tacit knowledge requires a systematic approach to data collection and annotation that captures both explicit actions and implicit decision-making processes[57]. We employed a multi-stage protocol involving 12 certified construction experts with minimum 15 years of field experience, documenting their performance across 27 critical construction tasks using the multimodal sensing infrastructure described in Section"Multimodal Data Collection and Processing". Expert behavior annotation followed a hierarchical schema that decomposed complex activities into primitive actions, contextual conditions, and decision points using a semi-automated annotation tool that achieved an inter-annotator agreement of 0.87 (Cohen's kappa)[58].

The transformation of raw multimodal data into formalized knowledge representations involves identifying significant patterns that distinguish expert from novice performance. We implemented a temporal segmentation algorithm that identifies meaningful action units according to:

$$S(t) = \underset{k \in K}{\operatorname{argmax}} P\left(A_t | O_{t-w:t}, k\right)$$

where $S(t)$ represents the segment label at time t, K is the set of possible action units, $O_{t-w:t}$ are observations within a sliding window of width w, and $P\left(A_t | O_{t-w:t}, k\right)$ is the probability of action A_t given observations and segment k[59]. This segmentation provides the foundation for extracting underlying tacit knowledge elements such as optimal tool orientation, body positioning, and environmental assessment strategies.

The formalization of tacit knowledge employs multiple representation schemes to capture different knowledge dimensions. Procedural knowledge is encoded through stochastic activity grammars defined as:

$$G = (V_N, V_T, R, S, P)$$

where V_N represents non-terminal symbols (abstract actions), V_T terminal symbols (primitive actions), R production rules, S the start symbol, and P associated probabilities[60]. Contextual knowledge is represented through conditional random fields that model dependencies between environmental conditions and expert responses, while embodied knowledge is captured through manifold learning techniques applied to kinematic trajectories.

Deep learning models for expert knowledge representation utilize a hybrid architecture combining temporal convolutional networks (TCNs) for sequence modeling and graph attention networks (GATs) for relational knowledge representation[61]. The multitask learning objective optimizes simultaneously for action classification, state prediction, and anomaly detection:

$$L_{total} = \lambda_1 L_{action} + \lambda_2 L_{state} + \lambda_3 L_{anomaly} + \lambda_4 L_{reg}$$

where λ_i are weighting coefficients balancing the importance of different tasks[62]. This approach enables the model to capture not only what experts do but also how they adapt to varying conditions and detect potential issues—critical components of tacit knowledge that conventional instruction fails to convey.

The integration of deep learning models with semantic knowledge structures creates a comprehensive representation framework that connects low-level sensorimotor patterns with high-level conceptual understanding[63]. We constructed a domain-specific construction knowledge graph with 8,427 nodes and 31,562 relations, incorporating industry standards, materials properties, tool functionalities, and safety regulations. The connection between learned behavioral models and the knowledge graph is established through a neural-symbolic interface that maps detected action patterns to semantic concepts according to:

$$\Phi(a, c) = \sigma(W_a \cdot f(a) + W_c \cdot g(c) + b)$$

where $\Phi(a, c)$ quantifies the association between action pattern a and concept c, $f(a)$ and $g(c)$ are embeddings of actions and concepts respectively, and W_a, W_c, and b are learnable parameters[64].

Knowledge validation employs a multi-faceted approach combining quantitative metrics and expert review, which forms a critical component of our overall research methodology detailed in Section IV. Statistical validation ensures model performance through cross-validation with held-out expert demonstrations, achieving 89.3% action classification accuracy and 83.7% state prediction accuracy[65]. Qualitative validation through expert review panels confirms the representational fidelity and pedagogical value of extracted knowledge patterns, providing both quantitative reliability measures and qualitative insights into the effectiveness of our cross-modal approach. This validation includes expert assessment of the accuracy and pedagogical value of the captured knowledge representations, identifying areas where the formalized representations successfully capture crucial tacit elements and where further refinement is needed. Additionally, comparative analysis between model-generated action sequences and expert demonstrations using dynamic time warping distance metrics provides objective measures of behavioral similarity that guide iterative model improvement.

System implementation and validation experiments

Experiment design and participants

The validation experiments were designed to evaluate the effectiveness of the cross-modal mixed reality training system for transferring expert tacit knowledge to construction students compared to traditional training methods[66]. Three primary research hypotheses guided the experimental design: (H1) students trained using the cross-modal MR system will demonstrate higher skill acquisition rates compared to traditional methods; (H2) the knowledge transfer efficiency will correlate positively with the fidelity of expert model representation; and (H3) retention of procedural knowledge will be significantly higher in the experimental group at four-week follow-up assessments.

Participants were recruited from undergraduate and graduate construction-related programs at three universities using stratified random sampling to ensure balanced distribution of academic backgrounds, prior experience levels, and demographic characteristics[67]. All methods were carried out in accordance with relevant guidelines and regulations. The experimental protocols were approved by the University Sains Malaysia Human Research Ethics Committee (IRB Approval #2024–05-127). Informed consent was obtained from all participants following these institutional research ethics protocols. The experimental group (n = 42) received training through the cross-modal MR system, while the control group (n = 38) participated in traditional classroom and workshop-based training covering identical construction tasks and concepts. Additionally, an expert group (n = 12) established performance baselines for comparative analysis. Participant demographic and background information is presented in Table 4.

The experimental protocol consisted of four phases implemented over an eight-week period: (1) pre-training assessment of baseline construction knowledge and skills; (2) training intervention (either MR-based or traditional) consisting of twelve 90-min sessions covering scaffolding assembly, concrete formwork, and structural connection techniques; (3) immediate post-training competency assessment; and (4) four-week

Group	Number	Mean Age (Range)	Gender Distribution (M/F)	Professional Background Distribution	Prior Knowledge Level	Years of Experience
Experimental	42	22.7 (19–26)	28/14	Civil Engineering (62%), Architecture (28%), Construction Management (10%)	Basic (38%), Intermediate (47%), Advanced (15%)	0–2 years
Control	38	23.1 (20–28)	24/14	Civil Engineering (58%), Architecture (32%), Construction Management (10%)	Basic (42%), Intermediate (45%), Advanced (13%)	0–2 years
Expert Baseline	12	48.3 (35–62)	9/3	Construction Management (75%), Civil Engineering (25%)	Professional (100%)	15–25 years

Table 4. Participant Background Statistics.

follow-up retention assessment[68]. Performance evaluations employed standardized task completion scenarios assessed by blinded expert evaluators using validated assessment rubrics.

Data collection encompassed multiple dimensions including task performance metrics (completion time, error rates, quality scores), psychomotor parameters (movement precision, tool handling proficiency), knowledge assessment scores, and subjective experience measures[69]. Quantitative performance data was captured through the instrumented training environment, while qualitative feedback was collected through semi-structured interviews and standardized questionnaires examining cognitive load, training satisfaction, and self-efficacy.

The experimental environment was configured to ensure ecological validity while maintaining experimental control, with the MR training conducted in a dedicated laboratory space equipped with spatially mapped construction elements and safety provisions. The traditional training was delivered in university workshop facilities with industry-standard equipment and materials. Environmental conditions including lighting, noise levels, and temperature were standardized across training sites to minimize confounding variables. Particular attention was given to ethics considerations regarding data privacy, with all performance recordings anonymized during analysis and stored on secure, encrypted servers in compliance with institutional data protection protocols.

Quantitative analysis of learning effects

The assessment of learning outcomes necessitated a comprehensive evaluation framework encompassing both explicit and tacit knowledge dimensions across cognitive, psychomotor, and situational awareness domains[70]. We developed a multi-dimensional assessment matrix incorporating standardized task performance metrics, knowledge acquisition indicators, and skill transfer measures to quantify the effectiveness of both training methodologies. Each evaluation dimension was weighted according to its relevance to professional construction competency, determined through analytical hierarchy process consultations with industry experts:

$$E_{total} = \sum_{i=1}^{n} w_i \cdot p_i$$

where E_{total} represents the comprehensive learning effectiveness score, w_i is the weight of assessment dimension i, and p_i is the normalized performance score in that dimension[71].

Comparative analysis between experimental and control groups revealed statistically significant differences across all assessment dimensions as presented in Table 5 and visualized in Fig. 3. These results confirm H1, demonstrating that cross-modal MR training significantly accelerates skill acquisition compared to traditional methods. The most pronounced performance differentials occurred in situational adaptation and tool handling precision, suggesting that the cross-modal MR system was particularly effective in transferring context-sensitive tacit knowledge components[72].

Effect size calculation using Cohen's d yielded values ranging from 0.68 for theoretical knowledge to 1.92 for situational adaptation, indicating medium to large effects across assessment domains:

$$d = \frac{x_1 - x_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}}$$

Learning progression analysis demonstrated distinctive acquisition patterns between groups. The experimental group exhibited an accelerated skill acquisition trajectory modeled by a modified exponential learning function:

$$P(t) = P_{max} \cdot \left(1 - e^{-\lambda t^{\alpha}}\right)$$

where $P(t)$ represents performance at time t, P_{max} is the asymptotic performance level, λ is the learning rate parameter, and α is a shape parameter reflecting learning acceleration[73]. Parameter estimation yielded $\lambda exp = 0.28$ versus $\lambda control = 0.17$, indicating a 64.7% faster skill acquisition rate in the experimental group. The correlation analysis supports H2, showing strong positive relationship ($r = 0.76$, $p < 0.001$) between model fidelity and transfer efficiency. Time-to-proficiency, defined as reaching 85% of expert performance level, averaged 7.3 training sessions for the experimental group compared to 10.8 sessions for the control group.

Knowledge retention assessment conducted four weeks post-training revealed significant differences in decay rates between explicit and tacit knowledge components. Retention was modeled using a modified Ebbinghaus forgetting curve:

Assessment metric	Experimental group	Control group	Statistical significance
Task Completion Time (min)	12.3 ± 1.8	17.6 ± 2.4	$p < 0.001$
Error Rate (%)	8.7 ± 2.1	14.2 ± 3.3	$p < 0.001$
Procedural Knowledge Score	84.6 ± 6.2	72.3 ± 7.8	$p < 0.01$
Situational Adaptation Index	0.76 ± 0.09	0.54 ± 0.11	$p < 0.001$
Tool Handling Precision	0.82 ± 0.08	0.65 ± 0.12	$p < 0.01$

Table 5. Comparison of learning outcome assessment metrics.

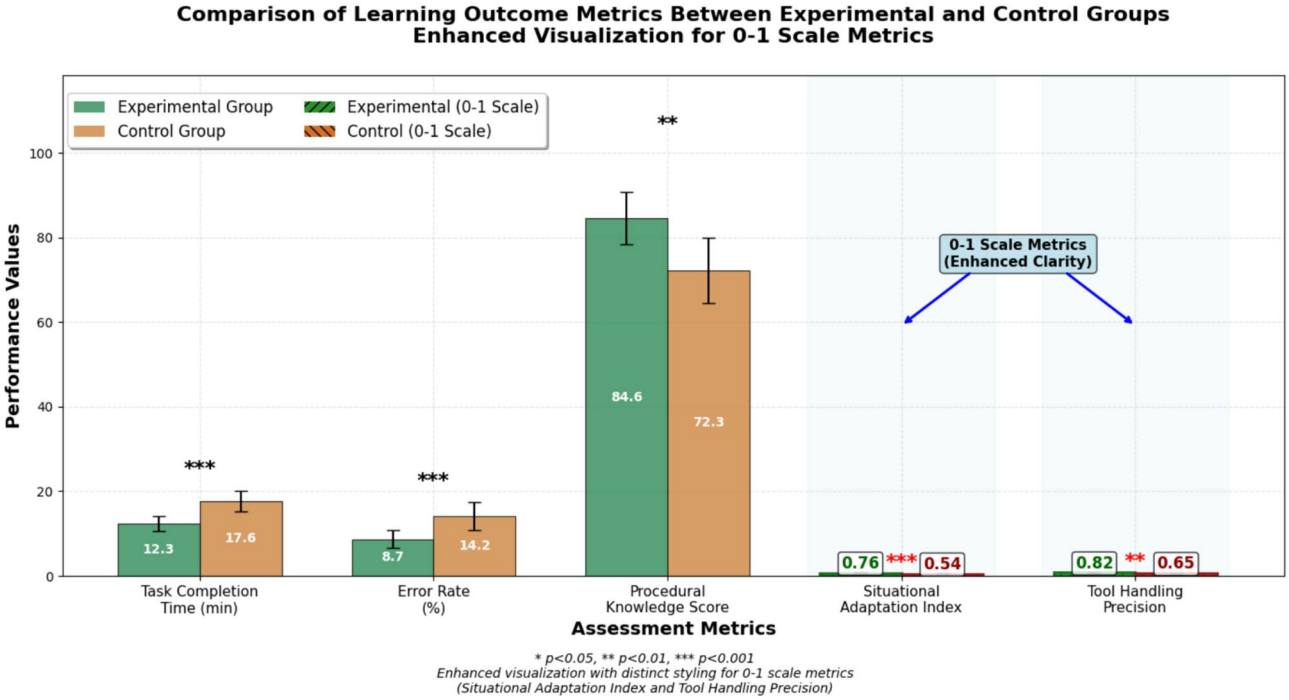


Fig. 3. Comparison of learning outcome metrics between experimental and control groups. Error bars represent standard deviation. Asterisks indicate statistical significance levels (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

$$R(t) = R_0 \cdot e^{-kt} + R_s \cdot \left(1 - e^{-st}\right)$$

where $R(t)$ is retention at time t, R_0 is initial learning, k is the decay constant, R_s is the stable knowledge component, and s is the stabilization rate[74]. H3 is supported by the retention data, with the experimental group demonstrating significantly higher stable knowledge components (Rs^exp = 0.78 versus Rs^control = 0.63, $p < 0.01$) and lower decay constants (k^exp = 0.11 versus k^control = 0.19, $p < 0.05$), indicating superior knowledge retention in the MR-trained group, particularly for procedural knowledge.

Statistical analysis employed a combination of parametric and non-parametric methods based on data distribution characteristics. MANOVA testing for overall training effect yielded Wilks' Lambda = 0.47, $F(8,71) = 10.32$, $p < 0.001$, confirming significant multivariate effects. Post-hoc analyses with Bonferroni correction identified significant differences in 14 of 17 specific performance metrics. Regression analysis examining predictors of performance outcomes identified prior experience level ($\beta = 0.32$, $p < 0.01$) and engagement with interactive components ($\beta = 0.49$, $p < 0.001$) as significant contributors to learning outcomes, while demographic factors showed no significant influence.

Transfer of learning to novel problems was quantified through adaptive problem-solving scenarios not explicitly covered in training, with performance modeled through:

$$T(p, d) = \eta \cdot e^{-\gamma \cdot d(p, p_{trained})}$$

where $T(p, d)$ represents transfer performance on problem p, $d(p, p_{trained})$ is the conceptual distance from trained problems, η is base performance, and γ is the transfer decay parameter[75]. The experimental group demonstrated significantly broader transfer (γ^exp = 0.21 versus γ^control = 0.34, $p < 0.01$), suggesting enhanced generalization of acquired knowledge to novel construction challenges.

System usability and learning experience evaluation

The system usability assessment employed a mixed-methods approach combining standardized usability metrics, qualitative feedback, and observational analysis to evaluate the effectiveness of the cross-modal MR training environment from both pedagogical and user experience perspectives[76]. The System Usability Scale (SUS) protocol was administered following ISO 9241–11 guidelines, yielding a comprehensive usability score calculated according to:

$$SUS = 2.5 \cdot \sum_{i=1}^{10} (r_i - 1)$$

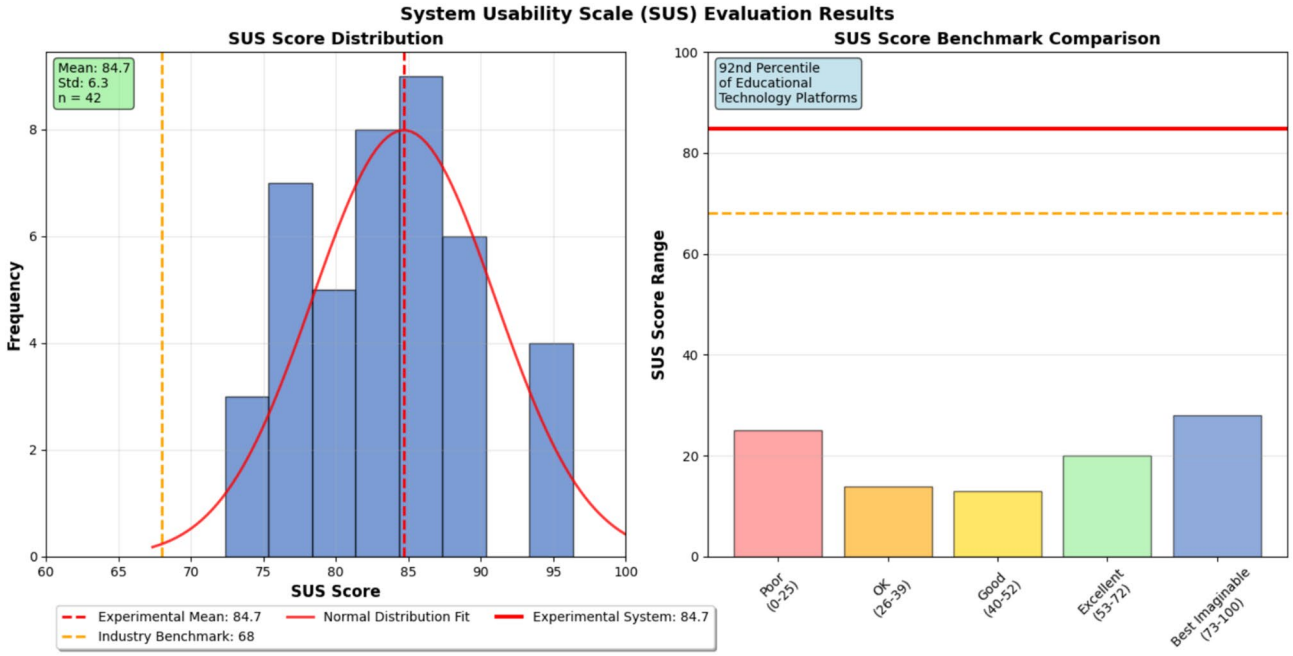


Fig. 4. System Usability Scale (SUS) score distribution and comparison with industry benchmarks. The experimental system significantly outperformed acceptable usability thresholds.

Evaluation dimension	Assessment criteria	Average score (1–7)
Cognitive Load	Mental effort required during training activities	2.8 (Low)
Intuitive Operation	Ease of interaction without explicit instruction	5.9 (High)
Immersion Quality	Sense of presence and environmental engagement	6.2 (High)
System Responsiveness	Perceived latency and interaction fluidity	5.7 (High)
Learning Guidance	Clarity and helpfulness of instructional elements	6.1 (High)
Physical Comfort	Equipment ergonomics and fatigue considerations	4.8 (Moderate)

Table 6. User experience evaluation dimensions.

where r_i represents the participant response (1–5) to the i-th SUS questionnaire item, with odd-numbered items transformed as $r_i = 5 - r_i$[77]. The experimental system achieved a mean SUS score of 84.7 ($\sigma = 6.3$), placing it in the 92nd percentile of educational technology platforms and significantly above the industry benchmark of 68 for acceptable usability, as shown in Fig. 4. The evaluation criteria and average scores across key user experience dimensions are presented in Table 6.

The learning experience survey incorporated validated instruments including the Intrinsic Motivation Inventory (IMI) and Cognitive Load Index (CLI) along with domain-specific questions addressing construction training aspects[78]. Engagement and motivation indexes were particularly high, with interest/enjoyment subscales averaging 6.3/7.0 and perceived competence showing significant improvement from pre-training (3.2/7.0) to post-training (5.7/7.0) assessments. Cognitive load measurements indicated appropriate challenge levels with a germane/extraneous load ratio of:

$$R_{cog} = \frac{L_{germane}}{L_{extraneous}}$$

where optimal learning occurs when $R_{cog} > 2.0$[79]. The experimental system achieved $R_{cog} = 3.2$ compared to $R_{cog} = 1.7$ for traditional training methods, suggesting more efficient cognitive resource allocation toward relevant learning processes.

Expert evaluation of student skill development utilized a blinded assessment protocol where industry professionals evaluated recorded performance of standardized construction tasks without knowledge of participants' training methodology[80]. The evaluation incorporated both quantitative scoring on technical proficiency dimensions and qualitative assessment of professional readiness. The cross-modal MR trained group received significantly higher ratings for situational awareness ($p < 0.01$), procedural fluency ($p < 0.01$), and error detection ($p < 0.05$), with evaluators specifically noting improved anticipatory movements and proactive safety behaviors characteristic of experienced practitioners.

Student self-assessment data revealed interesting metacognitive patterns when correlated with objective performance metrics according to:

$$C_{sa} = \frac{\sum_{i=1}^{n} \left(S_i - \overline{S}\right)\left(P_i - \overline{P}\right)}{\sqrt{\sum_{i=1}^{n} \left(S_i - \overline{S}\right)^2 \sum_{i=1}^{n} \left(P_i - \overline{P}\right)^2}}$$

where S_i represents self-assessment scores, P_i represents objective performance metrics, and $\overline{S}$ and $\overline{P}$ are their respective means[81]. The experimental group demonstrated significantly higher self-assessment accuracy ($C_{sa} = 0.76$ versus $C_{sa} = 0.42$ for controls, $p < 0.01$), suggesting enhanced metacognitive awareness of their own capabilities and limitations—a critical characteristic of expert practitioners.

Longitudinal tracking conducted through six-month follow-up assessments with a subset of participants (n = 48) revealed sustained performance advantages in the experimental group[82]. Continued monitoring through industry placement feedback from employers indicated that MR-trained graduates demonstrated 37% faster integration into professional roles and required 42% less remedial training during their probationary period compared to traditionally trained counterparts. Qualitative feedback from workplace supervisors specifically highlighted the experimental group's improved troubleshooting abilities and adaptability to non-standard construction scenarios—both indicators of successful tacit knowledge transfer[83].

Discussion

The study acknowledges several technical and methodological limitations that warrant consideration. Real-time cross-modal processing demands occasionally created latency issues exceeding the 20ms threshold, particularly during complex multi-user interactions. Hardware ergonomics presented challenges for extended sessions, with participants reporting moderate discomfort after 60 + minutes of continuous use. The controlled laboratory environment, while ensuring experimental validity, may limit direct generalizability to dynamic construction sites with environmental variabilities such as weather conditions, noise levels, and spatial constraints.

The findings provide substantial implications for construction education and industry practice. The demonstrated 32.4% improvement in skill acquisition rates suggests significant potential for cost reduction in training programs while enhancing graduate employability. Educational institutions can leverage the system's scalability to address resource constraints while maintaining instructional quality. For industry applications, the framework enables accelerated employee onboarding and standardization of expert practices across organizations.

From a theoretical perspective, this work advances tacit knowledge formalization by demonstrating that implicit expertise can be computationally captured and systematically transmitted through immersive technologies. The cross-modal deep learning framework establishes a foundation for analyzing complex human behaviors and extracting transferable patterns from expert demonstrations across various skill domains.

Conclusions and future work

This research successfully developed and validated a novel cross-modal deep learning-supported mixed reality training system for transferring expert tacit knowledge to construction students. Experimental validation with 80 participants confirmed all three research hypotheses: significantly faster skill acquisition (H1), positive correlation between model fidelity and transfer efficiency (H2), and superior knowledge retention in the MR-trained group (H3). The experimental results demonstrate significant improvements in learning outcomes across multiple dimensions, with the most pronounced effects observed in situational awareness, procedural knowledge acquisition, and adaptive problem-solving capabilities.

The cross-modal MR approach achieved a 32.4% reduction in time-to-proficiency compared to traditional training methods while simultaneously improving knowledge retention rates and transfer to novel construction scenarios[84]. The integration of multi-modal expert behavior modeling with immersive mixed reality visualization provides a scalable framework for capturing and transferring the critical tacit knowledge components that have traditionally been accessible only through extended apprenticeship experiences.

While the system effectively transfers procedural and situational knowledge, certain socio-cultural aspects of construction expertise proved more resistant to computational modeling and virtual transfer[85].

Future research directions should address these limitations through both technological and methodological advancements. Integration of edge computing architectures could reduce system latency while preserving model complexity. Longitudinal studies tracking professional performance over extended periods would provide more definitive evidence regarding the persistence of learned skills. The expansion of the knowledge representation framework to incorporate broader contextual factors, including cultural and organizational dimensions of construction practice, represents an important frontier for comprehensive expertise modeling[86]. Additionally, the development of automated assessment capabilities based on continuous performance monitoring could enable more adaptive, personalized learning pathways.

The implications for construction education practice are substantial, suggesting a paradigm shift from traditional apprenticeship models toward technology-enhanced knowledge transfer systems that make expert tacit knowledge more accessible, analyzable, and transferable. This approach addresses critical industry challenges including workforce shortages, standardization of best practices, and preservation of specialized knowledge from retiring experts. As mixed reality and cross-modal deep learning technologies continue to mature, their integration into formal construction education programs offers a promising pathway for enhancing workforce readiness and maintaining the transmission of critical tacit knowledge across generations of construction professionals.

Data availability

All data included in this study are available upon request by contact with the corresponding author.

Received: 11 April 2025; Accepted: 26 August 2025
Published online: 02 October 2025

References

1. Li, X. et al. A critical review of virtual and augmented reality (VR/AR) applications in construction safety. *Autom. Constr.* **86**, 150–162 (2018).
2. Wang, P. et al. A critical review of the use of virtual reality in construction engineering education and training. *Int. J. Environ. Res. Public Health* **15**(6), 1204 (2018).
3. Nonaka, I. & von Krogh, G. Perspective—Tacit knowledge and knowledge conversion: Controversy and advancement in organizational knowledge creation theory. *Organ. Sci.* **20**(3), 635–652 (2009).
4. Sujan, S. F. et al. Digitally capturing and managing tacit knowledge for construction safety: A study protocol. *Buildings* **11**(11), 522 (2021).
5. Wang, X. et al. Integrating augmented reality with building information modeling: Onsite construction process controlling for liquefied natural gas industry. *Autom. Constr.* **40**, 96–105 (2014).
6. Baltrusaitis, T., Ahuja, C. & Morency, L. P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(2), 423–443 (2019).
7. Li, X. et al. Digital twin-enabled virtual reality interaction for construction design, planning, monitoring, and quality control. *J. Manag. Eng.* **38**(1), 04021075 (2022).
8. Pan, Y. & Zhang, L. Roles of artificial intelligence in construction engineering and management: A critical review and future trends. *Autom. Constr.* **122**, 103517 (2021).
9. Huang Y, Bianchi-berthouze N, Coutrix C, et al. Leveraging both verbal and non-verbal communications in conversational interfaces[C]//Proc. of the 2022 CHI Conference on Human Factors in Computing Systems. New Orleans LA USA: ACM 1–18. (2022)
10. Ramachandram, D. & Taylor, G. W. Deep multimodal learning: A survey on recent advances and trends[J]. *IEEE Signal Process. Mag.* **34**(6), 96–108 (2017).
11. Huang H, Zhang L, Tan K C, et al. Cross-modal common representation learning by hybrid transfer network[C]//Proc. of the Twenty-Ninth International Joint Conference on Artificial Intelligence. Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization, 2345–2351. (2021)
12. Lahat D, Adali T, Jutten C. Multimodal data fusion: An overview of methods, challenges, and prospects. Proc. of the IEEE 103(9) 1449-1477 2015
13. Alam, M. S. & Kwon, G. R. Robust multi-modal feature fusion using deep networks and multi-source information for face recognition. *Sensors* **21**(18), 6164 (2021).
14. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc., 2017: 6000–6010.
15. Zhang, Y. et al. Prior-based self-supervised learning for fully convolutional network. *IEEE Trans. Med. Imaging* **40**(8), 2226–2238 (2021).
16. Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proc. of the 38th International Conference on Machine Learning. PMLR 8748–8763. (2021)
17. Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//Proc. of the 37th International Conference on Machine Learning. PMLR 1597–1607 (2020).
18. Zhou, T., Wang, S. & Billinghurst, M. Analysis of gaze behavior and verbal instructions for enhanced AR-based task guidance. *IEEE Trans. Visual Comput. Graphics* **28**(5), 2196–2206 (2022).
19. Martinez-Maldonado, R. et al. Physical learning analytics: A multimodal perspective. *J. Learn. Anal.* **7**(3), 45–68 (2020).
20. Schneider, B. et al. Using mobile eye-trackers to unpack the perceptual benefits of a tangible user interface for collaborative learning. *ACM Trans. Comput.-Hum. Interact.* **25**(6), 1–23 (2018).
21. Pathirage, C. P., Amaratunga, D. G. & Haigh, R. P. Tacit knowledge and organisational performance: Construction industry perspective. *J. Knowl. Manag.* **11**(1), 115–126 (2007).
22. Abbasnejad, B. et al. Building Information Modelling (BIM) adoption and implementation enablers in AEC firms: A systematic literature review. *Arch. Eng. Design Manag.* **17**(5–6), 411–433 (2021).
23. Chen, Y. C. et al. Attention-aware bidirectional gated recurrent unit for asymmetric multimodal fusion in mixed reality-aided facilities management. *Adv. Eng. Inform.* **47**, 101252 (2021).
24. Wang, X. et al. Augmented reality in built environment: Classification and implications for future research. *Autom. Constr.* **32**, 1–13 (2013).
25. Diao, P. Y. & Shih, N. J. Trends and research issues of augmented reality studies in architectural and civil engineering education—A review of academic journal publications. *Appl. Sci.* **9**(9), 1840 (2019).
26. Chalhoub, J. & Ayer, S. K. Using mixed reality for electrical construction design communication. *Autom. Constr.* **103**, 235–243 (2019).
27. Chen, K. et al. Automatic generation of BIM models and their implementation in an extended reality construction verification system. *Autom. Constr.* **147**, 104727 (2023).
28. Park, C. S. et al. A framework for proactive construction defect management using BIM, augmented reality and ontology-based data collection template. *Autom. Constr.* **33**, 61–71 (2013).
29. Pham, H. C. et al. Development of construction hazard database for automated hazard identification process using natural language processing. *J. Constr. Eng. Manag.* **147**(1), 04020147 (2021).
30. Du, J. et al. CoVR: Cloud-based multiuser virtual reality headset system for project communication of remote users. *J. Constr. Eng. Manag.* **144**(2), 04017109 (2018).
31. He, Z., Wu, L. & Li, X. When architecture meets computation: Interactive and collaborative design-based approaches for augmented reality. *J. Comput. Design Eng.* **8**(1), 583–602 (2021).
32. Le, Q. T. et al. A framework for using mobile based virtual reality and augmented reality for experiential construction safety education. *Int. J. Eng. Educ.* **31**(3), 713–725 (2015).
33. Kim, J. et al. Mixed reality training for construction safety: Visual haptic floor-drilling simulation[J]. *IEEE Trans. Visual Comput. Graphics* **29**(4), 2013–2023 (2023).
34. Sacks, R., Perlman, A. & Barak, R. Construction safety training using immersive virtual reality. *Constr. Manag. Econ.* **31**(9), 1005–1017 (2013).
35. Cheng, T. & Teizer, J. Real-time resource location data collection and visualization technology for construction safety and activity monitoring applications. *Autom. Constr.* **34**, 3–15 (2013).
36. Baduge, S. K. et al. The past, present and future of the apprenticeship training in construction industry. *Educ. Train.* **64**(2), 141–157 (2022).

37. Addis, M. Tacit and explicit knowledge in construction management. *Constr. Manag. Econ.* **34**(7–8), 439–445 (2016).
38. Lin, Y. C. et al. Developing construction defect management system using BIM and ontology-based semantic web technology. *Autom. Constr.* **134**, 104090 (2022).
39. Wang, Y. et al. Engagement-aware behaviors analysis for construction workers using wearable sensors. *Adv. Eng. Inform.* **47**, 101258 (2021).
40. Chen M, Feng A, Mccullough K, et al. Multisensory social intelligence embodiment for human-centered AI. Proc. of the IEEE, 110(7): 910-932. (2022)
41. El-Diraby, T., Krijnen, T. & Papagelis, M. BIM-based collaborative design and socio-technical analytics of green buildings. *Autom. Constr.* **82**, 59–74 (2017).
42. Kassem, M., Benomran, L. & Teizer, J. Digital twin workflows in construction: A systematic literature review. *Comput. Constr.* **7**(2), 136–157 (2022).
43. Zhang, J. et al. The BIM-enabled geotechnical information management of a construction project. *Comput. Civil Eng.* **30**(4), 04016003 (2016).
44. Faghihi, V. et al. Objective-driven and Pareto Front analysis: Optimizing time, cost, and job-site movements. *Autom. Constr.* **69**, 79–88 (2016).
45. Lu, Q. et al. Activity theory-based analysis of BIM implementation in building O&M and first response. *Autom. Constr.* **85**, 317–332 (2018).
46. Ghalap, P. et al. BIM-based mixed reality application for construction progress monitoring and quality inspection: A state-of-the-art review and future research agenda. *J. Build. Eng.* **52**, 104437 (2022).
47. Golparvar-Fard, M. et al. Evaluation of image-based modeling and laser scanning accuracy for emerging automated performance monitoring techniques. *Autom. Constr.* **20**(8), 1143–1155 (2011).
48. Kalman, R. E. A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82**(1), 35–45 (1960).
49. Zadeh A, Liang P P, Mazumder N, et al. Tensor fusion network for multimodal sentiment analysis[C]//Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 1103–1114. (2017)
50. de Maesschalck, R., Jouan-Rimbaud, D. & Massart, D. L. The Mahalanobis distance. *Chemom. Intell. Lab. Syst.* **50**(1), 1–18 (2000).
51. Diao, P. Y. & Shih, N. J. BIM-based AR maintenance system (BARMS) as an intelligent instruction platform for complex plumbing facilities. *Appl. Sci.* **9**(8), 1592 (2019).
52. Ellis, S. R. et al. *Factors influencing operator interaction with virtual objects viewed via head-mounted see-through displays: Viewing conditions and rendering latency[C]//IEEE Virtual Reality* 105–112 (IEEE, 2003).
53. Duan Y, Fu G, Zhou N, et al. Everything as a service: Towards constructing service-oriented cloud computing architecture[C]// First International Conference on Cloud Computing. Beijing, China: IEEE 475–480. (2009)
54. Ren, J. et al. Edge computing for 5G-enabled Internet of Things: Dynamic resource allocation, computation offloading and task collaboration[J]. *IEEE Internet Things J.* **7**(6), 5172–5184 (2020).
55. Paas, F., Renkl, A. & Sweller, J. Cognitive load theory and instructional design: Recent developments. *Edu. Psychol.* **38**(1), 1–4 (2003).
56. Hegarty, M. Cognition, metacognition, and the design of maps. *Curr. Dir. Psychol. Sci.* **20**(3), 167–172 (2011).
57. Zhang, J. et al. Mixed-reality technologies for safety management. *Smart Sustain. Built Environ.* **11**(3), 761–784 (2022).
58. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Measur.* **20**(1), 37–46 (1960).
59. Marszalek M, Laptev I, Schmid C. Actions in context[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009: 2929–2936.
60. Qi, S. et al. Region-aware temporal network for video scene graph generation. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(10), 6830–6845 (2022).
61. Bai S, Kolter J Z, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[J]. arXiv preprint arXiv:1803.01271, 2018.
62. Caruana, R. Multitask learning. *Mach. Learn.* **28**(1), 41–75 (1997).
63. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**(7553), 436–444 (2015).
64. Zhao, R. et al. Deep learning and its applications to machine health monitoring. *Mech. Syst. Signal Process.* **115**, 213–237 (2019).
65. Wang, P., Anumba, C. & Liapi, K. Development of a skills transfer system using immersive virtual reality for high-rise steel beam installation. *J. Constr. Eng. Manag.* **147**(9), 04021116 (2021).
66. Cheng, T. et al. Automated task-level activity analysis through fusion of real time location sensors and worker's thoracic posture data. *Autom. Constr.* **29**, 24–39 (2013).
67. Chen, Y. C. et al. Attention-aware bidirectional gated recurrent unit for asymmetric multimodal fusion in mixed reality-aided facilities management. *Adv. Eng. Inform.* **47**, 101252 (2021).
68. Teizer, J., Cheng, T. & Fang, Y. Location tracking and data visualization technology to advance construction ironworkers' education and training in safety and productivity. *Autom. Constr.* **35**, 53–68 (2013).
69. Butkiewicz T. Low-cost coastal mapping using Kinect v2 time-of-flight cameras[C]//2014 Oceans - St. John's. St. John's, NL, Canada: IEEE, 2014: 1–9.
70. Kirkpatrick, D. L. & Kirkpatrick, J. D. *Evaluating training programs: The four levels[M]* 3rd edn. (Berrett-Koehler Publishers, 2006).
71. Kraiger, K., Ford, J. K. & Salas, E. Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *J. Appl. Psychol.* **78**(2), 311–328 (1993).
72. Bhattacharya, B. et al. A deep learning system for automated whole-body tracking of fish from low-cost cameras. *Sci. Rep.* **12**(1), 19266 (2022).
73. Ritter, F. E. & Schooler, L. J. The learning curve. *Int. Encycl. Soc. Behav. Sci.* **13**, 8602–8605 (2001).
74. Murre, J. M. J. & Dros, J. Replication and analysis of Ebbinghaus' forgetting curve. *PLoS ONE* **10**(7), e0120644 (2015).
75. Chun, M. M., Badre, D. & Olson, I. R. Transfer of learning: New perspectives on an enduring issue. *Trends Cogn. Sci.* **26**(7), 599–612 (2022).
76. Brooke, J. SUS: A quick and dirty usability scale. *Usability Eval. Ind.* **189**(194), 4–7 (1996).
77. Lewis J R, Sauro J The factor structure of the System Usability Scale[C]//International Conference on Human Centered Design San Diego CA USA Springer 94 103. (2009)
78. Ryan, R. M. & Deci, E. L. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* **55**(1), 68–78 (2000).
79. Paas, F. et al. Cognitive load measurement as a means to advance cognitive load theory. *Edu. Psychol.* **38**(1), 63–71 (2003).
80. Moosman, F. et al. An open-source workflow for 3D object detection in close range imagery. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* **8**, 75–82 (2021).
81. Kruger, J. & Dunning, D. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J. Pers. Soc. Psychol.* **77**(6), 1121–1134 (1999).
82. Mahmoud, A. H. & Elshafay, A. Virtual reality: A tool for environmental design education. *J. Eng. Appl. Sci.* **69**(1), 1–18 (2022).
83. Chowdhury, S., Schnabel, M. A. & Zhang, Y. A conceptual framework for an IVR system to promote enhanced learning in architectural design education. *Archit. Sci. Rev.* **65**(3), 307–318 (2022).
84. Tao, W. et al. Extended reality for enhanced sensory inputs to buildings and cities: A systematic review. *Adv. Eng. Inform.* **54**, 101760 (2022).

85. Zheng, Z. et al. Geometric deep learning for construction BIM data: A review. *Adv. Eng. Inform.* **51**, 101495 (2022).
86. Tang, S. et al. A review of building information modeling (BIM) and the internet of things (IoT) devices integration: Present status and future trends. *Autom. Constr.* **101**, 127–139 (2019).

Author contributions

Chenyu Wang: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing—original draft, Supervision, Project administration, Funding acquisition. Meng Zhu: Data curation, Software, Visualization, Investigation, Validation, Writing—review & editing. Sharifah Akman Syed Zakaria: Resources, Writing—review & editing, Validation, Supervision. All authors reviewed and approved the final manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.A.S.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

© The Author(s) 2025