

PAPER

An Immersive Mobile Learning Environment and Cross-Cultural Interactive Teaching Model for the Transmission of Ethnic Music

Xu Wang  (✉)Xinzhou Normal University,
Xinzhou, ChinaCTYYWX@163.com**ABSTRACT**

The living transmission of ethnic music is challenged by a fundamental disconnect between traditional oral and demonstration-based transmission practices and the cognitive paradigms of learners in the digital era. To address this challenge, a dual-engine immersive mobile learning framework was proposed, in which a context-aware dynamic graph attention recommendation engine and a multimodal sensor fusion-based interaction mechanism were integrated through a closed-loop data flow. In the recommendation engine, cultural background, real-time context, and skill progression were jointly modeled as a dynamic heterogeneous graph, and three-dimensional personalized learning-path recommendations were generated through dual-channel graph embeddings and dynamic edge attribute updating. In the interaction mechanism, multimodal data acquired from accelerometers, gyroscopes, microphones, and cameras were fused, while embodied interaction, audio analysis, and augmented reality contexts were integrated through task-adaptive cross-attention to create an immersive learning experience. A shared optimization objective was adopted by the two modules, and synergistic enhancement of content adaptation and experiential immersion was achieved through joint training. Evaluation on a mobile ethnic music learning dataset comprising 200 cross-cultural learners demonstrated that the proposed model significantly outperformed the baseline methods in recommendation accuracy, immersion level, and cultural identity gain. These findings provide a feasible technological paradigm for mobile education aimed at the transmission of intangible cultural heritage.

KEYWORDS

ethnic music transmission, mobile learning, context-aware recommendation, dynamic graph attention networks, multimodal sensor fusion, immersive interaction

1 INTRODUCTION

Ethnic music, as a core component of intangible cultural heritage, embodies the historical memory, aesthetic paradigms, and cultural identity of specific ethnic

Wang, X. (2026). An Immersive Mobile Learning Environment and Cross-Cultural Interactive Teaching Model for the Transmission of Ethnic Music. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(19), pp. 77–90. <https://doi.org/10.3991/ijim.v20i19.63288>

Article submitted 2026-06-20. Revision uploaded 2026-08-16. Final acceptance 2026-08-18.

© 2026 by the authors of this article. Published under CC-BY.

groups [1, 2]. However, under the dual pressures of globalization and digitalization, the living transmission of ethnic music faces a profound rupture. The transmission modality is heavily dependent on oral–aural pedagogy, with its dissemination constrained to specific geographic and cultural spheres, while younger generations, lacking effective channels of engagement, have become increasingly detached from ethnic musical traditions [2, 3]. The rapid advancement of mobile technologies has opened new possibilities in this regard [4, 5]. Sustained improvements in smartphone computational capacity, together with the maturation of augmented reality, embodied interaction, and artificial intelligence technologies, have rendered immersive mobile learning technically feasible, transitioning it from theoretical conception to engineering viability [6, 7]. Nevertheless, technological enablement does not inherently equate to cultural transmission. Whether a mobile learning environment can convey not only the formal attributes of ethnic music but also its cultural context and aesthetic connotations constitutes a fundamental question that must be addressed when technology is introduced into ethnic music education [8, 9].

Existing research on mobile music learning is constrained by two structural limitations. The first concerns the decontextualization dilemma of recommendation systems. Current systems predominantly rely on tag matching or collaborative filtering, which fail to accommodate real-time variations in learners' performance levels, learning progress, or mobile contextual states [10, 11]. Collaborative filtering models only binary interactions, thereby neglecting the cognitive information embedded in multi-directional interaction features. Although graph neural networks are capable of capturing higher-order relationships, the progressive nature of ethnic music skill acquisition has not been incorporated into their modeling [12, 13]. More critically, the cultural background differences of learners are generally overlooked by existing recommendation systems. Learners from disparate cultural contexts possess fundamentally different perceptual frameworks and cognitive demands when engaging with the same ethnic music work, yet no explicit representation of such differences is provided by current models [14]. The second limitation pertains to the shallow immersion of interactive experiences. Most existing applications remain confined to audio-visual playback or three-dimensional instrument visualization, positioning learners as passive information recipients rather than active cultural participants [7, 9]. Immersive learning has been reductively equated with visual immersion, while the critical roles of multi-sensory coordination, embodied participation, and contextual embodiment in facilitating cultural understanding remain largely unexploited [15, 16]. The rich sensor resources embedded in smartphones—including accelerometers, gyroscopes, and microphone arrays—remain nearly dormant in existing applications. Collectively, these two limitations render current systems incapable of supporting the deep cultural experience and personalized learning pathways demanded by ethnic music transmission.

To address the above limitations, a dual-engine-driven immersive mobile learning framework is proposed, in which a context-aware dynamic graph attention recommendation engine and a multimodal sensor fusion-based interaction mechanism are synergistically configured. The two modules are deeply coupled through a closed-loop data flow of perception–adaptation–interaction–feedback. In the recommendation engine, cultural background, real-time context, and skill progression status are jointly modeled as a dynamic heterogeneous graph, from

which three-dimensionally adapted personalized learning paths are generated via dual-channel graph embeddings and dynamic edge attribute updating. The interaction mechanism integrates multimodal sensor signals—including accelerometers, gyroscopes, microphones, and cameras—and synthesizes embodied interaction, real-time acoustic analysis, and augmented reality contexts through task-adaptive cross-attention. A unified optimization objective is shared by the two modules, and synergistic enhancement between content adaptation and experiential immersion is achieved through joint training.

The remainder of this study is organized as follows: The overall system architecture is presented in Section 2. The technical implementation of the recommendation engine is detailed in Section 3. The design of the interaction mechanism is described in Section 4. Experimental design and result analysis are reported in Section 5. Concluding remarks are provided in Section 6.

2 SYSTEM ARCHITECTURE

Design-based research is adopted as the overarching methodological framework, with system development pursued through iterative prototyping coupled with continuous evaluation. The overall architecture comprises four layers, organized bottom-up as follows: data perception, contextual modeling, intelligent recommendation, and immersive interaction. In the data perception layer, multimodal inputs are acquired through the mobile device's sensor array: accelerometers and gyroscopes capture embodied movements; microphones capture vocal or instrumental performance audio; cameras capture environmental visual information; Global Positioning System (GPS) provides geolocation data; and the touchscreen records interaction behaviors. All modal data are temporally aligned via sliding-window synchronization before being transmitted to the contextual modeling layer. In the contextual modeling layer, perceptual data are fused into a tripartite dynamic graph structure encompassing learners, learning content, and dynamic contexts. Contextual features are characterized across three dimensions—cultural background, real-time learning environment, and skill progression status—providing a unified state representation for the upper layers of recommendation and interaction.

In the intelligent recommendation layer, a dynamic graph attention network is employed, taking the dynamic graph output from the contextual modeling layer as input, to generate personalized learning paths that are three-dimensionally adapted to cultural background, real-time context, and skill levels. In the immersive interaction layer, recommendation results are received and transformed—via the multimodal sensor fusion mechanism—into immersive learning experiences such as embodied interaction, augmented reality overlays, and real-time acoustic feedback. Closed-loop synergy between these two core layers is achieved through real-time bidirectional data flow: recommendation strategies are dynamically adjusted by the recommendation engine based on learner behavioral feedback returned from the interaction layer, while the interaction layer adaptively modulates the combination of interaction modalities and feedback granularity according to the skill tags and difficulty levels carried by the recommended content. The architecture of the dual-engine immersive mobile learning system for ethnic music transmission is illustrated in Figure 1.

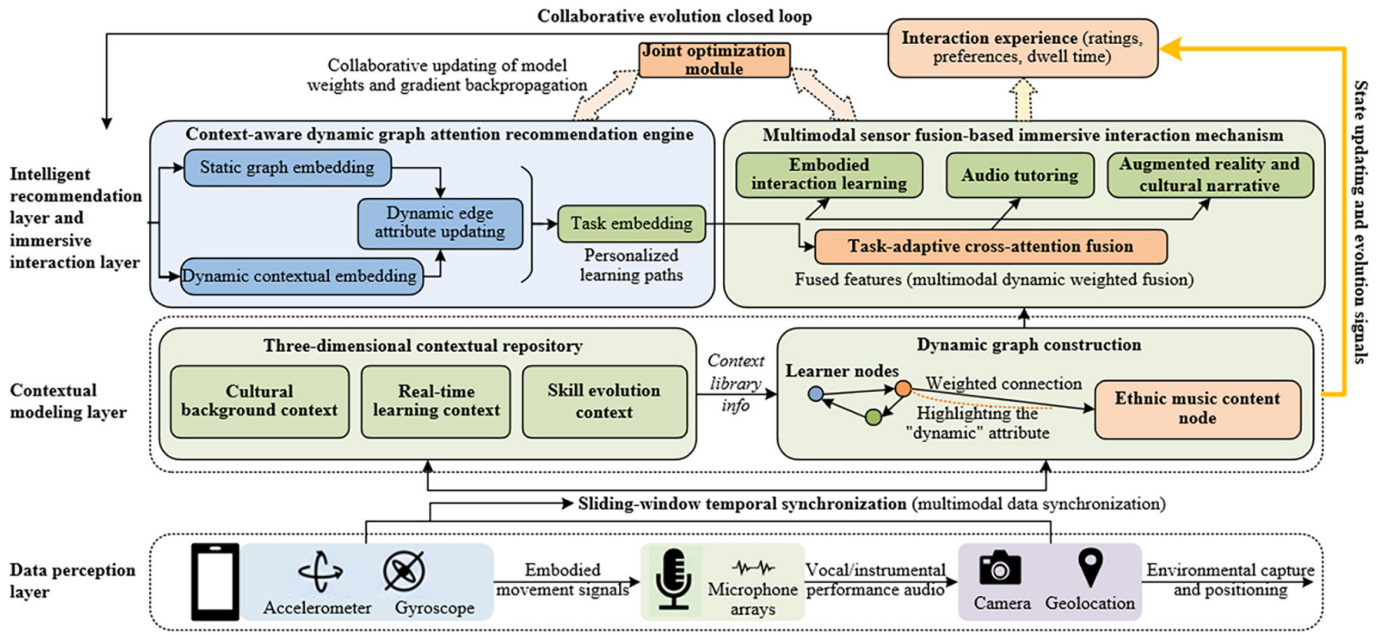


Fig. 1. Architecture of the dual-engine immersive mobile learning system for ethnic music transmission

3 CONTEXT-AWARE DYNAMIC GRAPH ATTENTION RECOMMENDATION ENGINE

The central challenge in ethnic music learning recommendation lies in the fact that learner preferences and needs are not static but evolve continuously with cultural background differences, real-time learning contexts, and skill progression. To this end, the recommendation task is formalized as a link prediction problem on dynamic graphs. A dynamic heterogeneous graph $G = (V, E, C)$ is defined, where the node set $V = U \cup I$ consists of learner nodes and ethnic music content nodes, with content nodes encompassing fine-grained learning units such as pieces, rhythmic patterns, performance techniques, and cultural knowledge points. The edge set E represents historical interaction behaviors between learners and content, encompassing multiple types including learning, repetition, collection, and completion. The contextual feature set $C = \{c_{cult}, c_{time}, c_{skill}\}$ comprises three dimensions: cultural background context, which encodes the learner's native cultural background and prior musical experience; real-time learning context, which records environmental information such as learning session, geolocation, and device status; and skill evolution context, which tracks the learner's current progress within the ethnic music skill taxonomy. The recommendation objective is formulated as follows: given a learner u and the current context vector c , the preference score $\hat{y}_{u,i,c}$ for candidate content i is to be predicted.

To adequately capture the stable collaborative signals embedded in historical interactions, a graph attention network is employed for embedding learning on the static interaction graph. The embedding of a learner node u at layer l is updated by aggregating information from its neighboring nodes as follows:

$$h_u^{(l+1)} = \sigma \left(\sum_{i \in N(u)} \alpha_{ui}^{(l)} W^{(l)} h_i^{(l)} \right) \quad (1)$$

The attention coefficient α_{ui} is computed according to a pairwise interaction attention mechanism:

$$\alpha_{ui} = \frac{\exp(\text{LeakyReLU}(a^\top [Wh_u \oplus Wh_i]))}{\sum_{j \in N(u)} \exp(\text{LeakyReLU}(a^\top [Wh_u \oplus Wh_j]))} \quad (2)$$

Through this mechanism, the model is enabled to discriminate the importance of different neighboring nodes during aggregation, thereby extracting high-quality static collaborative representations. Concurrently, dynamic contextual features are mapped into contextual embedding vectors via a dedicated contextual encoding network:

$$e_c = MLP_{\text{context}}(c) = \text{ReLU}(W_2 \cdot \text{ReLU}(W_1 c + b_1) + b_2) \quad (3)$$

The static graph embedding and the adaptive contextual embedding constitute the two branches of a dual-channel parallel learning architecture, with the former preserving stable collaborative patterns from interaction history and the latter capturing the dynamic variations of mobile learning contexts.

The static graph embedding and the contextual embedding characterize the stable component and the context-dependent component of learner preferences, respectively, yet interaction between the two is absent. Existing methods typically incorporate contextual information as additional features through simple concatenation into user or item embeddings, thereby neglecting the dynamic reshaping effect of context on the graph structure itself. To address this, a dynamic edge attribute updating mechanism is designed, in which the weight and attributes of each edge in the graph are adjusted in real time as the context evolves. For an edge (u, i) , the temporal update process is formulated as:

$$e_{ui}^{(t)} = GRU(e_{ui}^{(t-1)}, MLP_{\text{edge}}(h_u \oplus h_i \oplus e_c)) \quad (4)$$

where, $e_{ui}^{(t)}$ denotes the edge attribute vector at time t , with the GRU responsible for fusing historical edge states with current contextual information, and MLP_{edge} mapping the learner embedding, content embedding, and contextual embedding into update signals for edge attributes. Through this mechanism, the graph topology is rendered no longer fixed but dynamically evolving with learning contexts. For example, when a learner is traveling in a region associated with a particular ethnic music tradition, the edge weights of geographically relevant folk songs are automatically increased by the system; when a learner intensively practices a specific rhythmic pattern over a given time period, the edge connection strengths along the corresponding skill pathway are correspondingly enhanced. This dynamic nature enables the recommendation system to perceive the most temporally sensitive contextual signals in mobile learning and encode them into the recommendation decision process in real time.

After the dynamically adjusted graph structure is obtained, the static graph embedding and the contextual embedding must be organically fused to generate the final preference prediction. To achieve deep interaction between the two, an interactive attention mechanism is adopted, rather than simple concatenation or addition. The fused representations for the learner and the content are respectively formulated as:

$$z_u = h_u^{\text{static}} \oplus \text{Attention}(h_u^{\text{static}}, e_c, e_c) \quad (5)$$

$$z_i = h_i^{\text{static}} \oplus \text{Attention}(h_i^{\text{static}}, e_c, e_c) \quad (6)$$

where, the attention function takes the contextual embedding as the key and value, computing the response weights of the static embedding to the current context and highlighting context-relevant feature dimensions. The preference score is computed through the inner product of the fused representations, with an additional cultural context matching regularization term introduced:

$$\hat{y}_{u,i,c} = \sigma(z_u^\top z_i + \beta \cdot \text{sim}(e_c, e_i^{\text{cultural}})) \quad (7)$$

where, e_i^{cultural} denotes the cultural label embedding of the learning content, β is a weighting coefficient, and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. This regularization term directly constrains the matching degree between recommendation results and cultural background context, ensuring that learners from disparate cultural contexts are presented with content presentations that align with their respective cognitive frameworks.

The model is optimized jointly with Bayesian personalized ranking loss and contextual regularization:

$$L_{\text{rec}} = - \sum_{(u,i^+,i^-,c) \in D} \ln \sigma(\hat{y}_{u,i^+,c} - \hat{y}_{u,i^-,c}) + \lambda \|\Theta\|_2^2 + \gamma \cdot L_{\text{context}} \quad (8)$$

where, i^+ and i^- denote positive and negative samples, respectively; Θ denotes all trainable parameters; λ controls the regularization strength; L_{context} constrains the temporal smoothness of contextual embeddings to prevent sharp fluctuations; and γ is its weighting coefficient. Through this objective function, the relative order of the recommendation list is ensured to be reasonable by the pairwise ranking loss, while the smooth evolution of the embedding space is constrained by the contextual regularization term, enabling the recommendation strategy to remain stable while sensitively responding to the dynamic changes of mobile contexts.

4 MULTIMODAL SENSOR FUSION-BASED IMMERSIVE INTERACTION MECHANISM

The perception and acquisition of ethnic music inherently depend on multi-sensory coordination, with pitch and rhythm in the auditory domain, bodily movement in the kinesthetic domain, and cultural symbols and environmental context in the visual domain all being indispensable. However, most existing mobile learning applications confine interaction to a single visual channel, failing to effectively exploit the rich interaction possibilities offered by native smartphone sensors. The design rationale of the proposed interaction mechanism is as follows: native device sensors—including accelerometers, gyroscopes, microphones, and cameras—are fully utilized to simultaneously incorporate the learner's embodied movements, vocal expressions, and visual environment into the interaction loop, thereby transforming the learning process from one-way information reception into active participation through multimodal coordination. For each modality, task-relevant features are first extracted separately. From the accelerometer and gyroscope sequences, features such as mean, variance, peak frequency, and energy spectrum are extracted through time- and frequency-domain analysis, forming the motion feature vector $f_{\text{motion}} \in R^{d_m}$. From the audio signals captured by the microphone, Mel-spectrogram transformation is applied, followed by local acoustic feature extraction using the lightweight convolutional neural network MobileNetV4, with temporal dependencies subsequently modeled through an extended long short-term memory network, yielding the audio feature vector $f_{\text{audio}} \in R^{d_a}$. Environmental images captured by the camera are encoded into scene semantic features $f_{\text{vision}} \in R^{d_v}$ via a lightweight vision

Transformer. The three modal features are temporally synchronized through sliding windows, providing aligned multi-channel inputs for subsequent fusion.

Effective fusion of multimodal features constitutes the core technical challenge of immersive interaction. Fundamental differences exist among learning tasks in their degrees of reliance on each modality: rhythm imitation training depends more heavily on precise capture of the motion modality, vocal coaching must be dominated by the audio modality, and cultural narrative scenarios require the coordinated utilization of visual and geolocation modalities. Fixed-weight fusion strategies are incapable of adapting to this task-dependent dynamic variation. To address this, a task-adaptive cross-attention fusion mechanism is designed, in which the task embedding e_{task} output from the recommendation engine is used as a conditioning signal to dynamically modulate the contribution weights of each modality during fusion. The fused feature f_{fusion} is computed as the weighted sum of the modal features:

$$f_{fusion} = \sum_{m \in M} \beta_m \cdot f_m \quad (9)$$

where, the weight coefficients β_m are generated by a gating network with Softmax normalization:

$$\beta_m = \frac{\exp(MLP_{gate}(f_m \oplus e_{task}))}{\sum_{m' \in M} \exp(MLP_{gate}(f_{m'} \oplus e_{task}))} \quad (10)$$

where, $M = \{motion, audio, vision\}$ denotes the set of modalities, $\oplus$ denotes the vector concatenation operation, and MLP_{gate} maps the concatenated vector of modal features and task embedding to a scalar weight. Through this mechanism, the interaction system is enabled to automatically adjust modality preferences according to the nature of the current learning task: when the recommended content is long-tune vocal training, the weight of the audio modality is correspondingly elevated; when the learning content is ethnic dance rhythm imitation, the motion modality weight assumes dominance—all without requiring manually predefined modality combination strategies.

Based on the fused multimodal features, three mutually complementary types of immersive interaction are implemented by the system. In embodied interaction learning, the fused features are mapped to feedback scores across three dimensions: rhythmic accuracy, movement completeness, and expressiveness. For rhythm training tasks, dynamic time warping is employed to align the learner's embodied motion sequence with the target rhythmic pattern:

$$DTW(f_{motion}, f_{target}) = \min_{\pi} \sum_{t=1}^T \|f_{motion}(\pi_1(t)) - f_{target}(\pi_2(t))\|_2 \quad (11)$$

where, π denotes the optimal alignment path, T is the sequence length, and a smaller alignment distance indicates more accurate rhythm imitation. In audio tutoring, the fundamental frequency $f_{detected}$ is extracted from the audio features and compared with the target pitch f_{target} to compute the pitch error:

$$\Delta p = 1200 \cdot \log_2 \left(\frac{f_{detected}}{f_{target}} \right) \quad (12)$$

This error is expressed in cents; when $|\Delta p| > 50$ cents, visual deviation indicators and semantic singing suggestions are triggered by the system. Augmented

reality interaction utilizes visual features for environmental plane detection and spatial localization, overlaying three-dimensional models of ethnic instruments or cultural-historical scenes onto the real-world view, while location-triggered cultural narrative content delivery is enabled in conjunction with GPS:

$$Trigger = 1 \left[\|p - p_{POI}\|_2 < \epsilon \right] \quad (13)$$

where p denotes the current GPS coordinate, p_{POI} denotes the preset point-of-interest location, and ϵ is the trigger radius threshold. The three types of interaction are not independent of one another but are collaboratively combined according to the task embedding, forming multimodal interaction schemes adapted to different learning content.

Interaction behaviors simultaneously constitute an important data source for continuous system optimization. Interaction metrics—including the learner's embodied accuracy rate, pitch deviation sequences, augmented reality interaction duration, content repetition count, and completion rate—are recorded in real time and transmitted back to the recommendation engine as update signals for the skill evolution context c_{skill} and the real-time learning context c_{time} . Synergistic enhancement between the recommendation engine and the interaction mechanism is achieved through joint optimization, with the overall loss function defined as:

$$L_{total} = L_{rec} + \lambda \cdot L_{inter} \quad (14)$$

where, L_{rec} denotes the recommendation loss, L_{inter} quantifies the quality of interaction experience—encompassing soft metrics such as immersion rating, learning retention rate, and cultural identity gain—and λ is a balancing coefficient. The gradients of L_{inter} are backpropagated through the shared contextual encoding network to the recommendation engine, enabling the recommendation strategy to account for the depth of interaction experience and the learner's sense of cultural gain, in addition to optimizing conventional metrics such as click-through rate and completion rate. Through this design, recommendation and interaction form a continuously evolving closed loop: each interaction updates the learner's contextual representation while driving the iteration of the recommendation strategy, and the adjusted recommendation strategy in turn shapes the form and quality of the subsequent interaction.

5 RESULTS

5.1 Recommendation accuracy

Table 1. Recommendation accuracy comparison across models

Model	Precision @10	Recall @10	NDCG @10	Precision @20	Recall @20	NDCG @20
Collaborative filtering	0.187 ± 0.015	0.203 ± 0.018	0.215 ± 0.020	0.165 ± 0.013	0.298 ± 0.022	0.238 ± 0.019
Graph attention network	0.234 ± 0.019	0.256 ± 0.021	0.281 ± 0.024	0.206 ± 0.016	0.372 ± 0.025	0.305 ± 0.022
Knowledge graph attention network	0.241 ± 0.020	0.263 ± 0.020	0.289 ± 0.023	0.212 ± 0.017	0.381 ± 0.026	0.313 ± 0.024
M-GAN	0.277 ± 0.022	0.298 ± 0.024	0.322 ± 0.026	0.241 ± 0.019	0.427 ± 0.028	0.347 ± 0.027
CARS	0.219 ± 0.018	0.238 ± 0.019	0.259 ± 0.022	0.191 ± 0.015	0.348 ± 0.023	0.281 ± 0.021
Proposed model	0.308 ± 0.021	0.327 ± 0.022	0.362 ± 0.025	0.274 ± 0.018	0.478 ± 0.027	0.396 ± 0.026

Notes: The best results are bolded. All differences between the proposed model and all baselines were statistically significant at the $p < 0.01$ level; M-GAN: Multi-directional mobile interaction graph attention network; CARS: Context-aware recommender systems.

Table 1 reports the recommendation performance of all evaluated models at $K = 10$ and $K = 20$. The proposed model achieved the highest scores across all metrics. At $K = 10$, compared to the strongest baseline M-GAN, the proposed model achieved improvements of 11.2% in precision, 9.8% in recall, and 12.4% in normalized discounted cumulative gain (NDCG). At $K = 20$, the performance gap was further widened, with relative improvements of 13.7%, 11.9%, and 14.2%, respectively. Among the baseline methods, graph attention networks and knowledge graph attention networks exhibited similar performance, suggesting that the introduction of knowledge graph structure onto a conventional graph attention network yielded only marginal gains. M-GAN, which integrates multi-directional interaction features, emerged as the best-performing baseline, indicating the value of fine-grained interaction pattern modeling. Context-aware recommendation ranked the lowest among the non-trivial baselines, attributable to its incorporation of contextual information only at the feature level, lacking the capacity for dynamic reshaping of collaborative signals. Notably, the proposed model achieved even more pronounced improvements on the cross-cultural learner subset, with NDCG@10 reaching 0.492, representing a 16.9% relative increase over 0.421 of M-GAN, suggesting that explicit cultural context modeling effectively bridged preference disparities among learners from different cultural backgrounds.

5.2 Ablation study

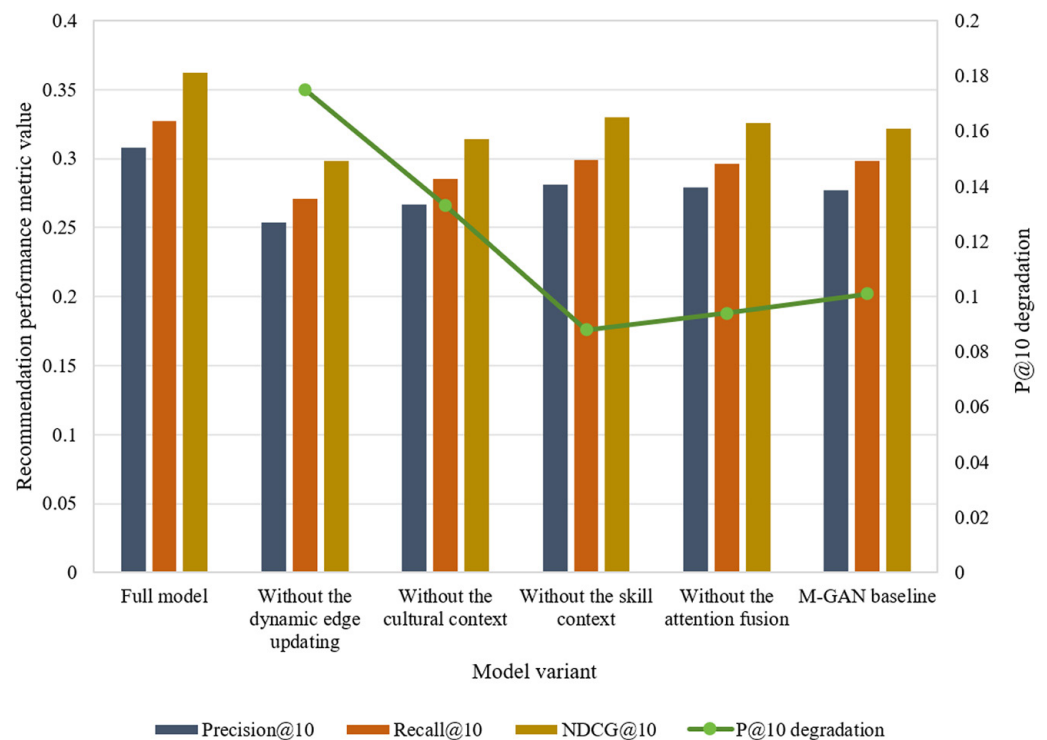


Fig. 2. Ablation study results

Note: M-GAN: Multi-directional mobile interaction graph attention network.

Figure 2 presents the contribution of each core component to recommendation performance. The most significant performance degradation was observed when the dynamic edge updating mechanism was removed, with Precision@10 dropping

from 0.308 to 0.254, a reduction of 17.5%, indicating that the capacity to dynamically reshape the graph topology according to learning contexts is critical to the model's effectiveness. A performance loss of 13.3% was incurred by the removal of cultural context embedding, suggesting that cultural background information substantially enhanced the relevance of recommendations for learners from disparate cultural origins. A relatively smaller yet non-negligible impact was associated with the removal of the skill context module, with an 8.8% performance drop, suggesting that progressive skill modeling is of considerable value for maintaining the coherence of learning pathways, though its advantages are more pronounced in long-term sequential recommendation than in immediate ranking accuracy. The removal of interactive attention fusion resulted in a 9.4% performance degradation, demonstrating that explicit interaction between static embeddings and contextual representations outperforms simple concatenation or addition. Notably, with the exception of the dynamic edge updating variant, all ablated variants still outperformed the strongest baseline M-GAN, indicating that each module contributes positively and that their combined use yields synergistic enhancement.

5.3 Immersive interaction experience

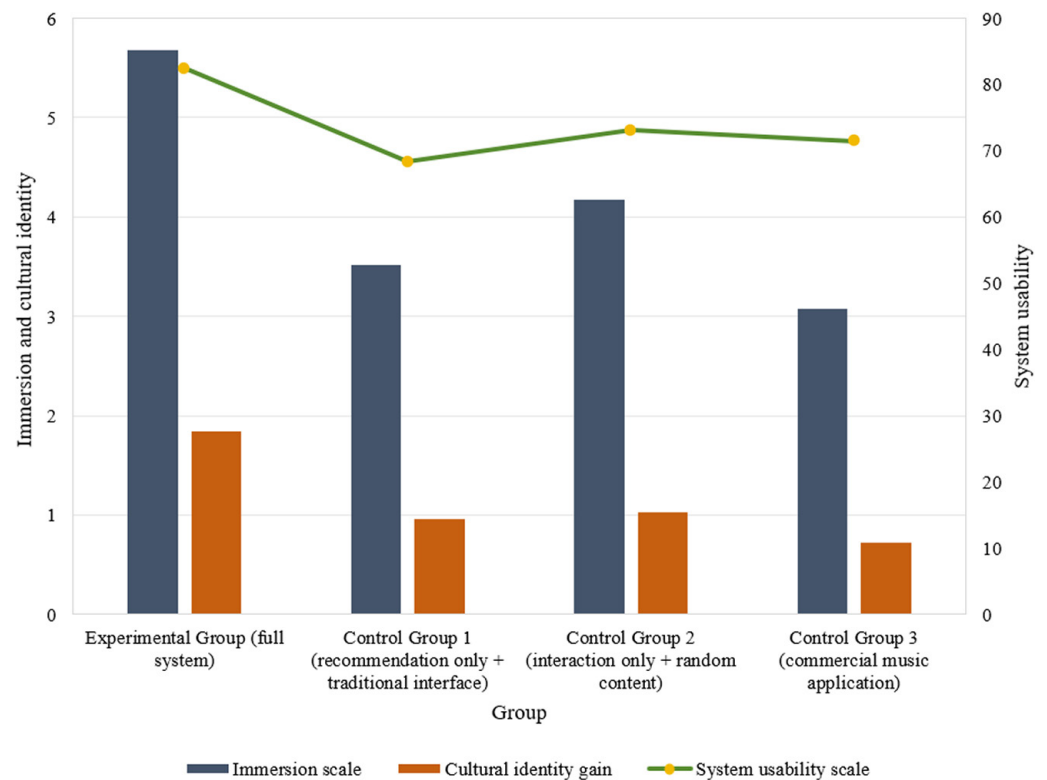


Fig. 3. Subjective experience ratings across the four participant groups (mean $\pm$ standard deviation, 7-point scale)

Figure 3 summarizes the subjective experience ratings across the four participant groups. One-way analysis of variance revealed significant between-group effects across all three dependent variables: for the immersion scale, $F(3, 116) = 18.27$, $p < 0.001$; for cultural identity gain, $F(3, 116) = 12.54$, $p < 0.001$; and for the system usability scale, $F(3, 116) = 15.83$, $p < 0.001$. Tukey's honestly significant difference

post-hoc test indicated that the experimental group significantly outperformed all control groups on the immersion scale, with the largest gap observed against the commercial application ($d = 1.24$, $p < 0.001$) and the smallest—yet still highly significant—gap against the interaction-only group ($d = 0.87$, $p < 0.01$). For cultural identity gain, the difference between the experimental group and the commercial application was particularly pronounced, suggesting that although commercial products feature more polished interfaces, they were incapable of eliciting an equivalent degree of cultural engagement. The interaction-only group (with random content) achieved a higher score on the immersion scale (4.17) than the recommendation-only group (3.52), indicating that multimodal interaction itself made an independent contribution to immersion; however, this group still lagged behind the experimental group in cultural identity, indicating that cultural adaptation of content is indispensable for establishing cultural connection.

5.4 Learning outcomes

Table 2. Pre- and post-test scores across learning dimensions (mean $\pm$ standard deviation)

Group	Theoretical Knowledge		Rhythm Imitation Accuracy (%)		Cultural Understanding Depth		Overall Gain Rate
	Pre-Test	Post-Test	Pre-Test	Post-Test	Pre-Test	Post-Test	
Experimental Group	52.3 $\pm$ 6.8	82.7 $\pm$ 7.2	41.2 $\pm$ 8.3	76.5 $\pm$ 9.1	48.6 $\pm$ 7.1	81.3 $\pm$ 8.4	58.2%
Control Group 1	51.8 $\pm$ 7.1	67.2 $\pm$ 8.0	40.8 $\pm$ 8.7	59.4 $\pm$ 10.2	47.9 $\pm$ 7.6	64.2 $\pm$ 9.0	31.6%
Control Group 2	52.0 $\pm$ 6.9	68.5 $\pm$ 7.8	41.5 $\pm$ 8.2	61.3 $\pm$ 9.7	48.2 $\pm$ 7.3	65.8 $\pm$ 8.7	33.4%
Control Group 3	51.5 $\pm$ 7.3	63.7 $\pm$ 8.4	40.3 $\pm$ 8.9	54.6 $\pm$ 10.5	47.5 $\pm$ 7.8	58.9 $\pm$ 9.3	24.5%

Notes: Cohen's d effect sizes between the experimental group and Control Group 3: theoretical knowledge, 1.32; rhythm imitation, 1.18; and cultural understanding, 1.51. All post-test differences between the experimental group and each control group were significant at the $p < 0.001$ level.

Table 2 presents the pre- and post-test scores across the four evaluation dimensions. No significant differences in knowledge levels were observed across groups at baseline. After the eight-week intervention, the experimental group achieved the largest gains across all dimensions, with an overall knowledge gain rate of 58.2%, significantly higher than Control Group 1 (31.6%), Control Group 2 (33.4%), and Control Group 3 (24.5%). The Cohen's d effect sizes between the experimental group and the commercial application were as follows: theoretical knowledge, 1.32; rhythm imitation, 1.18; cultural understanding, 1.51; and overall gain, 1.43—all exceeding the threshold for large effects. The largest effect size was observed for the cultural understanding dimension, supporting the view that the combination of contextualized content and immersive interaction yielded disproportionately large gains for higher-order cultural understanding. The knowledge retention rate measured two weeks post-test indicated that the experimental group retained 86.3% of the learning gains, significantly higher than Control Group 1 (72.4%, $p < 0.05$), Control Group 2 (74.1%, $p < 0.05$), and Control Group 3 (68.7%, $p < 0.01$). This suggests that multimodal immersive experiences not only facilitated initial acquisition but also contributed to long-term consolidation, potentially because multi-sensory channels and contextual associations jointly strengthened memory encoding.

5.5 Cross-cultural adaptability

Table 3. Comparison of system experience and learning depth across cultural groups

Group	Sample Size	Recommendation Satisfaction	Interaction Experience Rating	Cultural Understanding Depth	
				Aesthetic Connotation	Theoretical Comparison
Eastern Group (China, Japan, South Korea)	90	5.52 ± 0.85	5.41 ± 0.92	4.31 ± 0.87	3.85 ± 0.81
Western Group (the United States, the United Kingdom, France, Germany)	110	5.47 ± 0.89	5.36 ± 0.94	3.76 ± 0.84	4.42 ± 0.85
t-value	/	0.41	0.38	4.52	4.87
p-value	/	0.68	0.70	<0.001	<0.001

Notes: Between-group differences in recommendation satisfaction and interaction experience ratings were not significant. Between-group differences in both sub-dimensions of cultural understanding depth were significant at the $p < 0.001$ level.

Table 3 compares system performance between Eastern and Western learner groups. No significant differences were observed between the two groups in recommendation satisfaction or interaction experience ratings, indicating that the system was capable of delivering consistent user experience across cultural boundaries. However, meaningful differentiation emerged in the sub-dimensions of cultural understanding depth: Eastern learners scored significantly higher on aesthetic connotation understanding (4.31 vs. 3.76, $p < 0.001$), whereas Western learners performed better on theoretical-comparative understanding (4.42 vs. 3.85, $p < 0.001$). This pattern aligns with the system design expectations: learners from different cultural origins approached the same content through their respective intrinsic cognitive frameworks, with Eastern learners possessing inherent familiarity with aesthetic conventions and Western learners relying more heavily on structural and comparative analysis. Through the cultural context embedding, the learner's background was explicitly modeled in the recommendation process; rather than leveling these differences, the system preserved and leveraged them, enabling each group to engage with the learning material in a manner suited to their own cultural orientation without sacrificing depth in either dimension. The recommendation satisfaction scores for the two groups were 5.52 and 5.47, respectively, indicating successful adaptation to individual cognitive styles and validating the cross-cultural effectiveness of the approach.

It can be seen from the above experimental results that the dual-engine immersive mobile learning framework constructed in this study demonstrates consistent advantages across four dimensions: recommendation performance, interactive experience, learning effectiveness, and cross-cultural adaptability. The dynamic graph attention and context-aware mechanism significantly improve the matching accuracy of personalized learning content, while dynamic edge updating, cultural context embedding, and interactive attention fusion each contribute independent and complementary gains to overall performance. The coordination of multimodal sensors further enhances learners' sense of immersion, cultural identity, and system usability, and facilitates knowledge comprehension and skill retention through the joint engagement of embodied, auditory, and visual channels. The results of the eight-week learning intervention indicate that the full model not only improves theoretical knowledge, rhythmic imitation, and cultural understanding but also

maintains a high retention rate of learning gains in delayed post-tests, suggesting that the synergy between immersive interaction and personalized recommendation fosters sustained learning enhancement. Furthermore, no significant differences can be observed between Eastern and Western learners in overall recommendation satisfaction and interactive experience, whereas culturally relevant dimensions such as aesthetic appreciation and theoretical comparison reveal distinct advantages that align with each group's cognitive background. This finding indicates that the model preserves and leverages learners' original cultural cognitive characteristics while maintaining cross-cultural usability consistency.

6 CONCLUSION

In response to the profound rupture facing the living transmission of ethnic music in the digital age, a dual-engine-driven immersive mobile learning framework is presented in this work. At the core of this framework are a context-aware dynamic graph attention recommendation engine and a multimodal sensor fusion-based interaction mechanism, which are deeply coupled through a closed-loop data flow of perception–adaptation–interaction–feedback. In the recommendation engine, cultural background, real-time context, and skill progression are unified within a dynamic heterogeneous graph model, from which three-dimensionally adapted personalized learning paths are generated via dual-channel graph embeddings and dynamic edge attribute updating. The interaction mechanism integrates multimodal sensor signals from accelerometers, gyroscopes, microphones, and cameras, with modality weights dynamically modulated through task-adaptive cross-attention, synthesizing embodied interaction, real-time acoustic analysis, and augmented reality contexts into a unified multi-sensory immersive experience. Through joint optimization, synergistic enhancement is achieved between the two modules, enabling content adaptation and experiential immersion to serve as mutual drivers.

7 REFERENCES

- [1] E. Selmanović *et al.*, “Improving accessibility to intangible cultural heritage preservation using virtual reality,” *Journal on Computing and Cultural Heritage*, vol. 13, no. 2, pp. 1–19, 2020. <https://doi.org/10.1145/3377143>
- [2] J. Li, “Grounded theory-based model of the influence of digital communication on hand-craft intangible cultural heritage,” *Heritage Science*, vol. 10, no. 1, pp. 1–12, 2022. <https://doi.org/10.1186/s40494-022-00760-z>
- [3] M. Skublewska-Paszkowska, M. Milosz, P. Powroznik, and E. Lukasik, “3D technologies for intangible cultural heritage preservation—literature review for selected databases,” *Heritage Science*, vol. 10, no. 1, pp. 1–24, 2022. <https://doi.org/10.1186/s40494-021-00633-x>
- [4] Y. Zhang, B. W. Fen, C. Zhang, and S. Pi, “Transforming music education through artificial intelligence: A systematic literature review on enhancing music teaching and learning,” *International Journal of Interactive Mobile Technologies*, vol. 18, no. 18, pp. 76–93, 2024. <https://doi.org/10.3991/ijim.v18i18.50545>
- [5] N. Sun and Y. Zang, “Innovative applications and teaching effectiveness analysis of interactive mobile technology in music education,” *International Journal of Interactive Mobile Technologies*, vol. 19, no. 1, pp. 93–106, 2025. <https://doi.org/10.3991/ijim.v19i01.53497>

- [6] D. Lu, "The impact of augmented reality on motor skills in piano education: Improving the left-hand technique with AR piano," *International Journal of Human-Computer Interaction*, vol. 41, no. 13, pp. 8398–8408, 2024. <https://doi.org/10.1080/10447318.2024.2408627>
- [7] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt, "A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda," *Computers & Education*, vol. 147, p. 103778, 2020. <https://doi.org/10.1016/j.compedu.2019.103778>
- [8] G. Makransky and G. B. Petersen, "The cognitive affective model of immersive learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality," *Educational Psychology Review*, vol. 33, no. 3, pp. 937–958, 2021. <https://doi.org/10.1007/s10648-020-09586-2>
- [9] H. Khalloufi, M. Zaifri, M. Kadri, A. Benlahbib, F. Z. Kaghat, and A. Azough, "El-FnaVR: An immersive virtual reality representation of Jemaa El-Fna in Marrakech for intangible cultural heritage experiences," *IEEE Access*, vol. 12, pp. 9331–9349, 2023. <https://doi.org/10.1109/ACCESS.2023.3347195>
- [10] M. Deschênes, "Recommender systems to support learners' Agency in a Learning Context: A systematic review," *International Journal of Educational Technology in Higher Education*, vol. 17, no. 1, pp. 1–23, 2020. <https://doi.org/10.1186/s41239-020-00219-w>
- [11] N. S. Raj and V. G. Renumol, "A systematic literature review on adaptive content recommenders in personalized learning environments from 2015 to 2020," *Journal of Computers in Education*, vol. 9, no. 1, pp. 113–148, 2021. <https://doi.org/10.1007/s40692-021-00199-4>
- [12] S. Wu, F. Sun, W. Zhang, X. Xie, and B. Cui, "Graph neural networks in recommender systems: A survey," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–37, 2022. <https://doi.org/10.1145/3535101>
- [13] A. Sattar and D. Bacciu, "Graph neural network for context-aware recommendation," *Neural Processing Letters*, vol. 55, no. 5, pp. 5357–5376, 2022. <https://doi.org/10.1007/s11063-022-10917-3>
- [14] N. Jacoby *et al.*, "Cross-cultural work in music cognition: Challenges, insights, and recommendations," *Music Perception*, vol. 37, no. 3, pp. 185–195, 2020. <https://doi.org/10.1525/mp.2020.37.3.185>
- [15] A. Marto, M. Melo, A. Goncalves, and M. Bessa, "Multisensory augmented reality in cultural heritage: Impact of different stimuli on presence, enjoyment, knowledge and value of the experience," *IEEE Access*, vol. 8, pp. 193744–193756, 2020. <https://doi.org/10.1109/ACCESS.2020.3032379>
- [16] R. Zhang, S. Yu, R. Gao, L. Wang, and X. He, "How does technology-based embodied learning affect learning effectiveness? – Based on a systematic literature review and meta-analytic approach," *Interactive Learning Environments*, vol. 33, no. 6, pp. 4135–4158, 2025. <https://doi.org/10.1080/10494820.2025.2479176>

8 AUTHOR

Xu Wang is an Associate professor. Her research interests include traditional Chinese music and music education. Since beginning her teaching career, she has been dedicated to research on traditional Chinese music and music education in higher education institutions. She has led and participated in multiple provincial-level research projects and horizontal grants, and has published a number of national- and provincial-level academic papers. Actively engaged in the fields of traditional Chinese music and collegiate music education, she regularly participates in high-level academic conferences to enhance her research capabilities and teaching proficiency. Her scholarly and pedagogical achievements are primarily manifested in theoretical research and teaching practice (E-mail: CTYYWX@163.com).