



Reimagining the Master–Apprentice: A Hybrid Physical-Digital Pedagogy Framework for Traditional Shaanxi Heritage-Art Education in Chinese Primary Schools

Shi FengRui, Chee Ken Nee*

Faculty of Computing and Meta-Technology, University Pendidikan Sultan Idris (UPSI), Tanjung
Malim, Perak 35900, Malaysia

*Corresponding Author

DOI: <https://doi.org/10.47772/IJRISS.2026.100800675>

Received: 02 September 2026; Accepted: 07 September 2026; Published: 15 September 2026

ABSTRACT

Traditional Shaanxi artisan practices such as Ansai paper-cutting, Fuping folk embroidery, Qinqiang facial painting, and loess landscape painting are disappearing as youth interest declines and the population of master artisans ages, while uneven resourcing places rural learners at a further disadvantage. This study designed, implemented, and evaluated a hybrid physical-digital pedagogy integrating segmented multimedia heritage modules, immersive virtual reality, an adaptive artificial-intelligence tutor, and a cross-regional collaborative platform. A sequential explanatory mixed-methods quasi-experiment was conducted over sixteen weeks with 60 upper-primary pupils recruited in equal numbers from three Shaanxi schools of differing resource level, with intact art classes assigned to condition ($n = 30$ per condition). The design was grounded in cognitive load theory, the cognitive theory of multimedia learning, and dual coding theory. The experimental group outperformed the control group on all six learning-outcome dimensions, with large between-group effects for technical skill ($d = 2.20$, 95% CI [1.56, 2.84]) and cultural appreciation ($d = 2.14$, 95% CI [1.50, 2.77]); analyses of covariance confirmed these advantages after adjustment for pre-test scores. Because pupils were nested within intact classes and schools, a design-effect sensitivity analysis was conducted; all differences remained significant at intraclass correlations up to .20. Daily learning duration rose by 42.3%, while test anxiety ($d = -1.30$) and cognitive distraction ($d = -1.39$) declined. Reflexive thematic analysis produced three positive and two challenge themes, from which two constructs are proposed as hypotheses for confirmatory testing rather than as established findings: algorithmic folk-art correction bias and the equity paradox in heritage educational technology. A third contribution, the Hybrid Physical-Digital Pedagogy Framework, specifies a theoretically derived five-phase progression from digitally mediated understanding to authentic embodied practice.

Keywords: heritage-art education, multimedia learning, artificial intelligence in education, virtual reality, educational equity

INTRODUCTION

Shaanxi Province has a unique position in the geography of China's different cultures. As the site for the founding of the Western Zhou, the capital of the Qin, and the capital of the Tang, Shaanxi has acted as the central location of culture for the last three thousand years and impacts the surrounding countries' cultures. East Asia has been the destination of the two methods of crafts and teaching. The development of art and folk crafts, like Ansai style paper cutting and Jianzhi, Qin Qiang Opera face painting, and Fuping folk embroidery, has been recognized by the Ministry of Culture and Tourism of China as a National Treasure of China and inscribed on several lists of National and Provincial Intangible Cultural Heritage (UNESCO, 2003; Ministry of Culture and Tourism of China, 2021). The contradiction to this is the many signs which indicate the actual disappearance of the traditions that are systematically passed on in institutions.

The contemporary challenges surrounding the transmission of heritage-art in Shaanxi can be attributed to three interrelated factors. First, Liu et al. (2021) and Sun et al. (2022) report a significant decline in undergraduate

enrollment in Shaanxi's tertiary heritage-art programs, with participation dropping by 27%–34% in some programs over the last decade. Second, the average age of master-artisans in the region has reached 58 years, and in some localities has exceeded 62 years (Wang & Zhao, 2022). Finally, while the education technology revolution transformed the majority of higher education classrooms, in the heritage-art sector, technology has been embraced in more haphazard and superficial ways (Chen, 2018). Collectively, these factors create a logjam of pedagogical, cultural, and ethical challenges within the heritage-art sector in Shaanxi. Lacking personalized feedback, the master-apprentice and rote models of teaching do not address the inequalities between urban and rural areas, nor do they meet the asymmetries in feedback and resources that the education system in China requires (Hannum et al., 2008).

This study designs, applies, and evaluates a multi-component, multi-media, AI and VR based framework for the cultural and pedagogical contextualization of Shaanxi art education in schools. This study outlines three main objectives. First, to examine the relative effectiveness of the integrated intervention in comparison to traditional studio teaching, in relation to the main dimensions of educational outcomes which have been commonly discussed in the educational technology literature. Second, to explain the operational mechanisms through which the integrated intervention achieves the described outcomes. Third, and perhaps most interesting, to identify and describe the cultural contextualization and equity related dilemmas that would serve to bridge folkloric art pedagogy within the larger contemporary instructional technology framework.

This third aim generates two original theoretical contributions. The first, Algorithmic Folk-Art Correction Bias, describes AI feedback systems' bias against regional folk art. They interpret asymmetrical designs as technical flaws. This bias affects patterns found in Ansai paper-cutting, bold colors in Fuping embroidery, and loess landscape painting. The second, the Equity Paradox in Heritage Educational Technology, describes the digital inequity paradox, which refers to unequal access to heritage education technology, and which specifically impacts rural learners who are the longest lived across generations. These two original constructs, the HPDPF, and the Hybrid Physical-Digital Pedagogy framework, propose instructional design models grounded in cross-cultural contexts and based on three principles and five phases.

To the best of our knowledge, this is among the first systematic evaluations of integrated multimedia-AI-VR pedagogy in Chinese primary-school heritage-art education. Based on a quasi-experimental mixed-methods approach, the article provides an evaluative framework with seven sections. The first section is a literature review and presents cognitive, architectural, technological, and cultural equity gaps that inform this study. The second section outlines the research design, and the third section presents participant selection, research instruments, and data analysis. The fourth section presents findings from the different phases of the research, and the fifth section presents findings related to each of the three research questions and the four phases of this longitudinal study. The last section presents the main contribution of this evaluation, framed within contemporary global challenges of digital pedagogy and cultural safeguarding.

Theoretical Framework And Literature Review

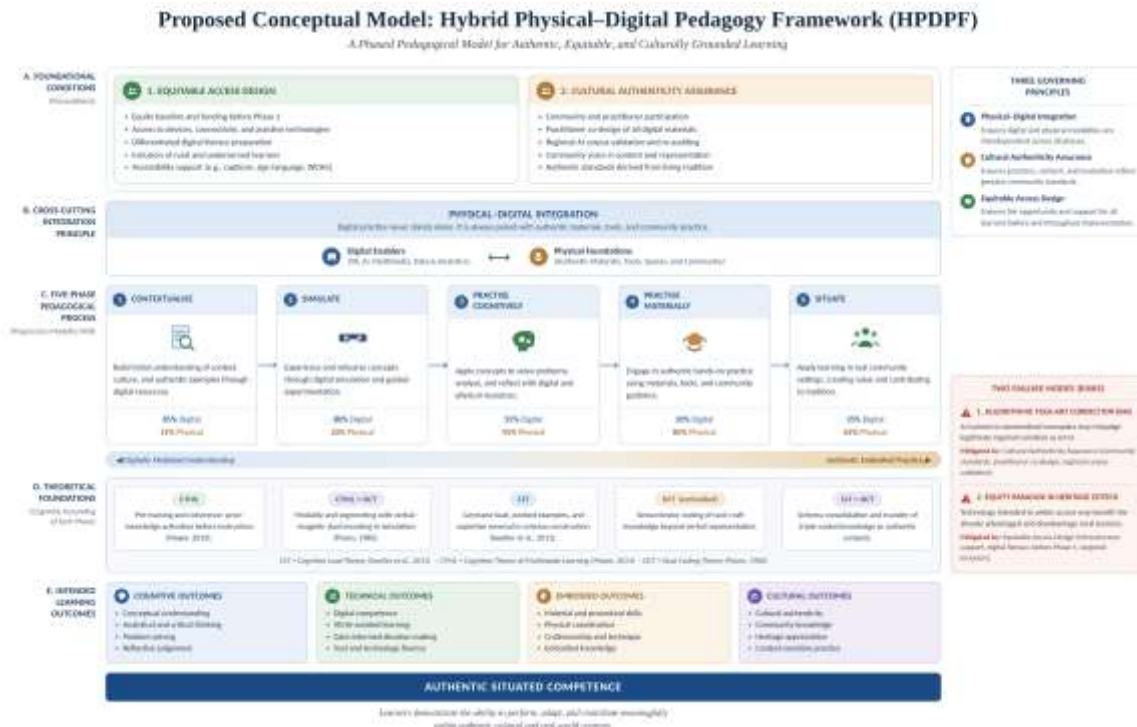
Cognitive-Architectural Foundations

This intervention design combines three theories of learning in relation to cognitive, perceptual, and representational processes and their relevance to integrated multimedia-AI-VR teaching. According to Sweller, and Sweller et al.'s (2011) Cognitive Load Theory, the design of effective multimedia instruction relies on three simultaneous constraints of the limitations of an individual's working memory: intrinsic cognitive load, which refers to the complexity of the material; extraneous cognitive load, which refers to the manner in which the material is presented; and germane cognitive load, which refers to the cognitive effort the individual expends in the formation of the schema. In terms of design, this means that the optimal design minimizes extraneous cognitive load, develops intrinsic cognitive load to the learner's level of expertise, and addresses germane cognitive load to the vertical dimension of the cognitive outcome. Kalyuga (2007) and Schroeder and Cenkci (2018) maintain that a wealth of empirical research spanning three decades has resulted in a considerable body of replicable instructional design effects, such as worked examples, split attention, modality, redundancy, and expertise reversal, which serve as the empirical basis of the framework.

Based on constructivist learning, Mayer’s (2014) CTML builds on CLT and describes how particular designs of multimedia messages facilitate the attainment of learning objectives. CTML gives consideration to three of the cognitive principles: (1) dual channels (Paivio, 1986); (2) limited capacity (Baddeley, 1992); (3) active processing (Wittrock, 1989). Mayer suggests 12 design principles as a result of these cognitive principles. The principles that relate directly to this project are (1) coherence (omission of irrelevant materials), (2) signalling (presentation of important information in a salient manner), (3) segmenting (presentation of information in customizable, controllable, pace instructions), (4) pre-training (presentation of prerequisite vocabulary and concepts before instruction), (5) modality (presentation of verbal information in a close and personal manner), and (6) personalization (presentation of verbal information in a comfortable, casual manner). CTML has been shown to be effective in a multitude of learner populations and diverse content areas, and the design principles have a $d \approx 1.0$ (Adesope & Nesbit, 2012) impact.

According to Paivio (1986), Dual Coding Theory (DCT) states that information encoded in verbal and imagistic systems is both more robust and retrievable than information encoded in just one system. In terms of craft pedagogy, DCT has more applicability than other theories. Kinesthetic and spatial representations, along with elements of Folk Art, are beyond the capacity of verbal pedagogies (Polanyi, 1966; Wenger, 1998). The more recent application of DCT in the field of embodied cognition (Glenberg et al., 2013; Wilson, 2002) suggests that in order to meaningfully teach a craft, imagistic, verbal, and, most importantly, sensorimotor codes are necessary. Figure 1 presents an integration of the three perspectives in order to illustrate the principles and instructional framework of the HPDPF presented in this study.

Figure 1: The Hybrid Physical–Digital Pedagogy Framework (HPDPF): proposed conceptual model. Two foundational conditions and a cross-cutting integration principle govern a five-phase instructional sequence that shifts progressively from digitally mediated understanding to authentic embodied practice. Each phase is anchored in one or more of the three cognitive theories of learning that underpin the framework—CLT, CTML, and DCT—and the sequence yields four domains of intended learning outcomes.



Note. CLT = Cognitive Load Theory (Sweller et al., 2011); CTML = Cognitive Theory of Multimedia Learning (Mayer, 2014); DCT = Dual Coding Theory (Paivio, 1986). Percentages within each phase indicate the intended proportion of digital and physical modality rather than fixed time allocations. The gradient strip beneath the phases represents the progressive shift from digital-dominant rehearsal (Phases 1–3) to physical-dominant practice (Phases 4–5). The two failure modes identified in this study—algorithmic folk-art correction bias and the equity paradox in heritage educational technology—are mitigated by the Cultural Authenticity Assurance and Equitable Access Design conditions, respectively.

The three theories are not applied uniformly across the sequence; each phase draws on the theoretical resource best matched to its representational demand. Phase 1 (contextualise) is governed by the CTML pre-training and coherence principles, which establish the vocabulary and cultural schema required before instruction begins and remove material extraneous to that purpose. Phase 2 (simulate) combines the CTML modality and segmenting principles with DCT, because virtual rehearsal encodes the same craft procedure verbally and imaginistically at a pace the learner controls. Phase 3 (practise cognitively) is governed principally by CLT, through worked examples, expertise-reversal-sensitive scaffolding, and the deliberate redirection of freed working-memory capacity towards germane load during schema construction. Phase 4 (practise materially) rests on the embodied extension of DCT, in which the sensorimotor code carries the tacit dimensions of craft knowledge that neither verbal nor imagistic representation can transmit (Glenberg et al., 2013; Polanyi, 1966). Phase 5 (situate) draws on CLT and DCT jointly, consolidating the triple-coded schema and transferring it to the variable conditions of authentic community practice. The modality proportions reported in Figure 1 therefore express a theoretically motivated progression rather than an arbitrary allocation of instructional time.

Educational Technology in Heritage Art Contexts

Evidence for multimedia-enhanced art supports the application of CTML (Cognitive Theory of Multimedia Learning) across art-educational disciplines (Maria et al., 2014; Roege & Kim, 2013) while acknowledging that emotional and cultural factors typically dominate art-educational disciplines beyond the boundaries of the CTML cognitive-architectural model (Plass et al., 2014). In the last decade, a great deal of evidence has accumulated on the use of VR (Virtual Reality) in education. Merchant et al. (2014) published a meta-analysis that reported a pooled effect size of about 0.68 for the use of VR in diverse teaching contexts. Radianti et al. (2020) published a systematic review, and their results were consistent with that of Merchant et al. (2014) when considering multiple studies on immersive VR in the higher education context. The positive effects of VR have been explained by the concepts of presence, embodiment, and rehearsal in a secure context, and other cognitive activities regarding the successful achievement of the goal, which was extended by Parong and Mayer (2018), and thus can be used to support craft pedagogy, where kinaesthetic and procedural techniques are core.

Recently, the body of literature regarding the role of AI in education has grown significantly. In their 2016 meta-analysis, Kulik and Fletcher found the pooled effect size of intelligent tutoring systems to be approximately $d = 0.66$. VanLehn (2011) also noted that the effect of intelligent tutoring systems was on par with the effect of one-on-one human tutoring. The application of AI to the personalization of art education is even more recent. This area of research has used a variety of methods, though case studies have dominated the literature, and have been more often descriptive than prescriptive (Hamal et al., 2022; Taylor et al., 2021). The application of AI to Chinese traditional art education in compulsory schooling has been under-researched; the small existing literature concentrates on tertiary settings. The present study addresses that gap by applying a systematic, quasi-experimental design to AI and VR integration in heritage-art teaching at the primary level in a Chinese context.

AI Ethics, Cultural Authenticity, and Educational Equity

The ethical aspects of inequality caused by AI in education have begun to receive attention in relation to issues such as algorithmic bias, data privacy, and accountability. In regards to algorithmic bias, there are various literature examples that deal with equality issues pertaining to classification, setting decision thresholds, and bias in standardized testing and admissions. This study will focus on the bias of equality in demographics and propose a culturally-conventional bias framework. The heritage of AI in the education sector has hardly been researched in relation to the existing literature on cultural authenticity of AI in education. The Culture in AI Design framework offered through educational design entities may contribute new perspectives to the existing literature.

Research on educational equity in China has primarily consisted of studies on the disparity of resources in rural versus urban settings, and the inequality that exists in Chinese higher education is even more pronounced in the education of heritage-art. The rural communities where the heritage-art is taught are systematically positioned at the lowest rank in the educational system and are therefore disadvantaged. This study aims to explore the inequities in heritage-art education and, through this, contribute to the existing literature.

The Present Study

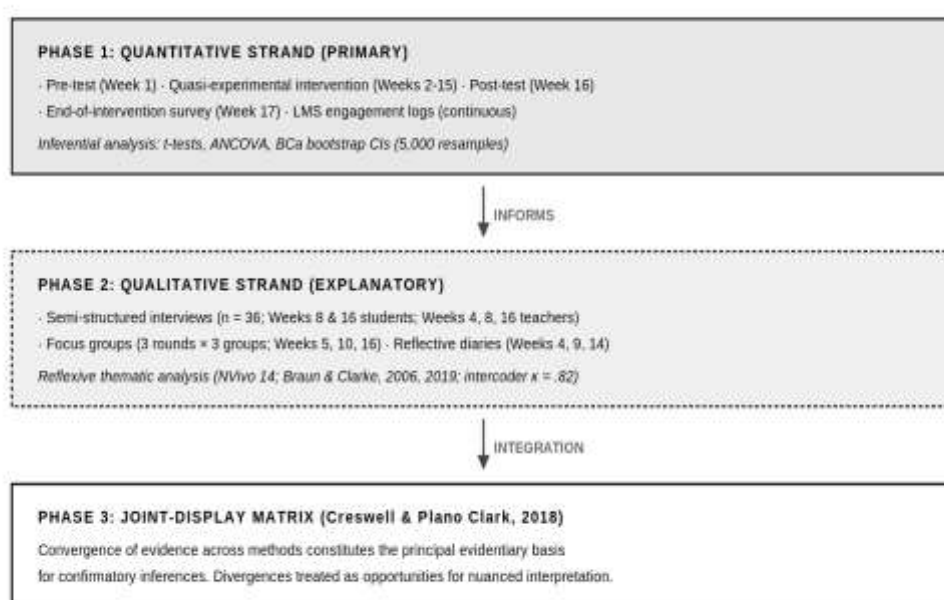
Based on the above literature, an empirical study will attempt to answer the following four research questions: (RQ1) what are the best network-based teaching of traditional Shaanxi art methods and strategies? (RQ2) what are the optimal degrees of artificial intelligence integration to personalize and enhance learning of traditional Shaanxi art? (RQ3) what are the differences in learning outcomes, engagement, and appreciation of the culture of Shaanxi art, compared to traditional methods, within the framework of multimedia and AI teaching strategies? (RQ4) what are the main challenges and/or opportunities presented by the integration of AI and multimedia in the teaching of Shaanxi art? The following eight hypotheses are proposed for the comparison: (H1) skill level, (H2) creative engagement, (H3) knowledge retention of art, (H4) level of critical thinking, (H5) level of collaboration, (H6) level of engagement, (H7) level of reduction of anxiety and distraction, (H8) level of cultural appreciation. This will be tested using an experimental group post-test.

METHOD

Research Design

The research applied the mixed-methods approach described by Creswell and Plano Clark (2018) and Shadish et al. (2002), utilizing a pragmatic research framework (Morgan, 2014). In regard to the framework of the quasi-experiment, a modified pre and post-test design was applied to both the experimental (EG) and control (CG) groups, based on the conditions of the school's art program. The qualitative data collection focused on the semi-structured interviews, discussion focus groups, and reflective diaries. These data collection methods were organized on the basis of providing an explanation of the observed quantitative results. The qualitative and quantitative data were integrated and organized as a joint-display matrix, as presented by Creswell and Plano Clark (2018). Figure 2 presents the data collection framework for this study.

Figure 2: Sequential explanatory mixed-methods design architecture. The quantitative strand is conducted first to establish inferential evidence; the qualitative strand is then conducted to illuminate underlying mechanisms; the two strands are integrated via the joint-display matrix.



Setting and Participants

Three primary schools in Shaanxi Province were selected to represent the range of resource availability in the provincial art education system: a metropolitan school in Xi'an, a peri-urban school in Yan'an, and a rural school in Ansai. A purposive sample of 60 upper-primary pupils (N = 60) was recruited, comprising 20 pupils

from each of the three schools and drawn in equal numbers from two adjacent year groups (30 pupils per year group). Assignment was made at the level of intact art classes rather than individual pupils: within each school, one intact class was allocated to the experimental group (EG) and one to the control group (CG), contributing 10 pupils per condition per school and yielding $n = 30$ in each condition. The consequences of this clustered allocation for the inferential analyses are addressed in the Data Analysis section and revisited in the Limitations. The sample was 60% female and 40% male. Pupils' residential backgrounds, which do not correspond to the location of the school attended, spanned the rural-to-urban gradient (48.3% urban, 28.3% peri-urban, 23.3% rural). Chi-square tests on demographic indicators and independent-samples t-tests on pre-test scores showed no significant differences between conditions (all $p > .10$). An a priori power analysis using G*Power 3.1.9.7 (Faul et al., 2007) indicates that, for a two-tailed independent-samples t-test with $\alpha = .05$ and $n = 30$ per group, statistical power is .48 for a medium effect ($d = 0.50$) and .86 for a large effect ($d = 0.80$); the smallest effect detectable at conventional power (.80) is $d = 0.74$. The study is therefore powered to detect large effects only, and this constraint qualifies the interpretation of all null comparisons reported below.

The Multi-Component Intervention

The experimental treatment provided to the EG over a 16-week period comprised four components that aligned with the five-phase HPDPF instructional design process (a) immersive virtual reality (VR) practice modules, (b) an adaptive artificial intelligence (AI) feedback tool, (c) segmented multimedia (sMD) cultural-heritage content, and (d) a moderated collaborative digital workspace. The integrated design is shown in Figure 3. EG students engaged in a minimum of 240 minutes of immersive VR practice in a Meta Quest 3 h MD in the construction of four scenario sets pertaining to four art forms (the Yan'an paper cutting studio, Qinqiang dressing room, the loess plateau plein-air vista, and the Fuping embroidery workshop) and each was supplemented with haptic controllers. The AI feedback tool was a custom convolutional neural network (CNN) that was trained to recognize Shaanxi Folk Art and was set to provide feedback on line fidelity, compositional balance, color authenticity, and motif accuracy within 12 seconds, on a 1-5 ordinal scale. Feedback was criterion-referenced and the AI's feedback was compared to three Shaanxi Folk Art master artisans. Their agreement was set at an inter-rater $\kappa = .79$. An example of the design of the learner-facing platform and the integrated components of immersive VR, AI adaptive tutor, and segmented multimedia is shown in Figures 4 to 7.

Figure 3: Architecture of the multi-component intervention delivered to the experimental group across the sixteen-week period, showing component sequencing, weekly time allocation, HPDPF phase mapping, and the cultural-authenticity safeguards of master-artisan auditor review.

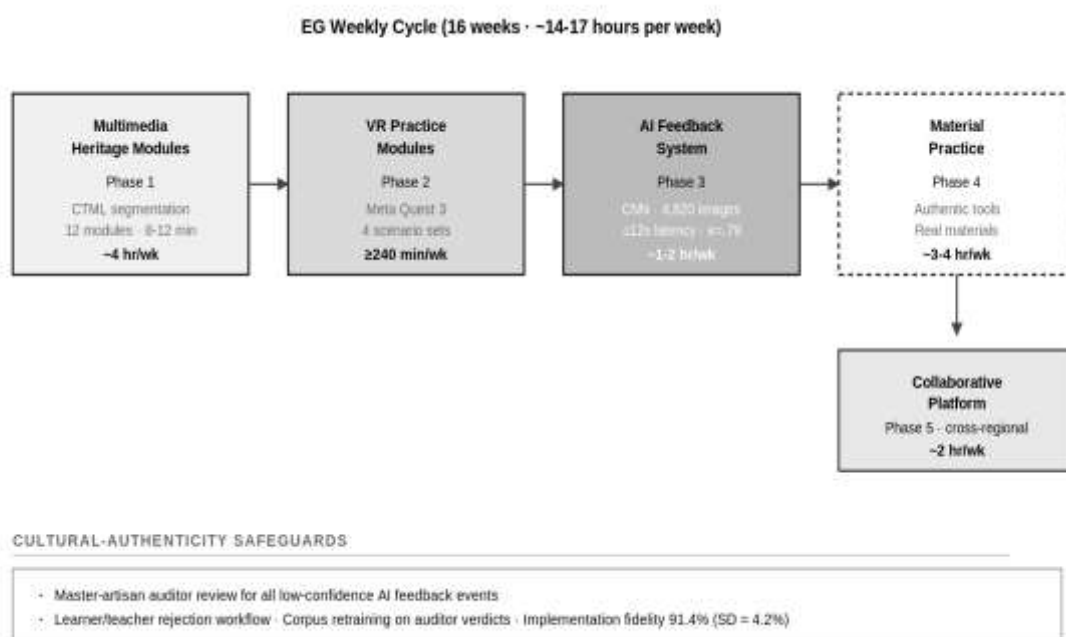


Figure 4: Primary students in the experimental group undertaking immersive virtual-reality practice during the technical-principles session. Learners rehearse culturally grounded craft sequences in a non-evaluative, head-mounted VR environment before physical practice, at the urban and rural school sites.



Figure 5: The adaptive AI tutor interface (师傅 Shifu). The always-available tutor explains technique and cultural meaning for the four Shaanxi art forms, coaches learners on their AI feedback, and answers study-related questions offline, operationalising the adaptive AI feedback component of the intervention.



Figure 6: The segmented multimedia cultural-heritage content library. Each weekly module for the four Shaanxi art forms is divided into a cultural-context session and a technical-principles session, following the CTML segmenting and pre-training principles, with a pre-training glossary and master-artisan narrative.



Figure 7: A segmented content module (Qinqiang facial painting). The module pairs a pre-training glossary and a first-person master-artisan narrative with short, segmented video sessions, operationalising the CTML pre-training and segmenting principles.



CTML segmenting and pre-training principles guided the construction of 12 multimedia heritage modules (8–12 minutes each) for 4 art forms. Each module contained narrated demonstration videos from master-artisans and offered strategies and ideas for contextual cultural narratives and exemplars and artifacts. The modules employed self-directed segment pacing. A moderated cross-institutional collaborative platform provided shared artifact structures and peer feedback based on the AI rubric and engaged cross-regional EG peer communication. The CG was provided the traditional curriculum of each of the three sites, which included art practice, a lecturer-led demonstration, and cultural and contextual written assignments and summative evaluations. This was conducted without the AI, VR, and multimedia resources and the collaborative platform. A 14-item observation checklist was used to assess the implementation fidelity of each EG. The mean fidelity for the 16 weeks was 91.4% (SD = 4.2%), which was greater than the target of 85%.

Instrumentation

Learning outcomes were assessed with parallel pre- and post-tests administered in week one and week sixteen, respectively. The assessment battery contained a practical task with 5-criterion analytic rubric for technical skill (ICC = .91 pre; .93 post), a three-piece portfolio with a 4-criterion rubric for creative projects (ICC = .87; .89), a written test (40 items) on art knowledge with an assessment of art knowledge retention ($\alpha = .82; .85$), a task of critical thinking ($\kappa = .81$), and a peer- and self-assessment tool on collaboration with an assessment of collaboration ($\alpha = .84$). The self-assessment of time use, observation, and 10-item engagement scale ($\alpha = .89; \omega = .90$) measured engagement. Psychological well-being was measured with the Distraction Scale and Test Anxiety Scale (15 items; $\alpha = .84; .87$). The Cultural Appreciation and Identification Scale (CAIS), developed for this study, was assessed through three stages of a validation process (content validity: I-CVI/S-CVI = .94; face validity: cognitive interviews; construct validity: exploratory factor analysis: KMO = .81, Bartlett's test: $\chi^2 = 84.36, p < .001$), with Cronbach's $\alpha = .86$ and McDonald's $\omega = .88$ in the main study.

Pilot Study

A separate sample of 25 upper-primary pupils from a fourth Shaanxi primary school took part in a pilot study approximately six weeks prior to the primary data collection. The study aimed to assess the tools' psychometric sufficiency prior to the main data collection. The pilot study helped to address issues related to the VR and AI systems and the progressive steps of the implementation framework. Evidence from the pilot study indicated that multi-item scales had a Cronbach's alpha value in the range of .79 to .87. Test/ retest reliability for the CAIS (n = 15) at a 14-day interval ranged from .81 to .88. The main study was further enhanced by five fundamental changes. The latency of AI feedback was modified from 18 to 12 seconds, the comfort threshold was addressed by recreating two VR scenarios that encompassed a greater surface area, the engagement survey was revised to include 10 rather than 16 questions, and two CAIS questions were revised and two questions were added to the focus-group interview guideline.

Procedures

Quantitative data were collected during three main sampling periods. These were the start of the study, Week 1 (pre-intervention), the end of the study, Week 16 (post-intervention), and the span of Week 1 to Week 16, which was measured by the participant engagement survey, structured observations, and LMS logs. Pre- and post-test data were collected during these periods from participants who completed the tests in a controlled setting, and research study assistants administered the tests. To avoid any bias, research study assistants were kept blind to the participant's group placements. Several methods were applied to collect qualitative data. These included semi-structured interviews, focus group discussions, and reflective diaries. The semi-structured interviews for students and teachers were conducted in Weeks 8 and 16, while the focus group discussions were conducted in Weeks 5, 10, and 16. Reflective diaries were submitted during Weeks 4, 9, and 14. All semi-structured interviews were conducted in Mandarin, and the diaries were in English. All semi-structured interviews were recorded and were transcribed in Mandarin for the first time, and then translated to English.

Data Analysis

IBM SPSS Statistics version 27.0 was used for quantitative data analysis. The Shapiro-Wilk and Kolmogorov-Smirnov tests assessed distributional assumptions, with thresholds for skewness and kurtosis of $|\text{skewness}| < 2$ and $|\text{kurtosis}| < 7$ (Kline, 2015). Levene's test confirmed the equal variance assumption. The regression homogeneity of slopes was maintained for all dependent variables of the covariance analysis. For the proposed hypotheses, independent sample t-tests, one-way covariance analysis (with pre-tests), and Pearson product-moment correlations were carried out. Effect sizes are reported as Cohen's d for between-group contrasts and partial eta squared for the ANCOVA models. Every reported t and d was computed directly from the group means and standard deviations and is internally consistent with them ($t = d\sqrt{n/2}$ for equal group sizes); Hedge's g values differed from d by less than .03 throughout and are not reported separately. Confidence intervals for d were obtained from the standard large-sample standard error, and bias-corrected and accelerated bootstrap intervals (5,000 resamples) were computed as a check and agreed to within .03. The Holm-Bonferroni method was used to sequentially adjust for multiple comparisons within the same family of related hypotheses. Braun and Clarke's (2006, 2019) thematic analysis method was used to analyze qualitative data, with inter-coder agreement for assigned themes at $\kappa = .82$. NVivo 14 was used for analyses of qualitative interviews, focus groups, and reflective diaries.

Because pupils were assigned by intact class within three schools, observations are not fully independent, and treating them as such risks understating standard errors. The design contains only three clusters per condition, which is too few for stable estimation of a multilevel model or of cluster-robust standard errors (Kline, 2015). Two safeguards were therefore adopted. First, school was entered as a fixed factor in preliminary two-way analyses; the school $\times$ condition interaction was non-significant for every outcome, indicating that the treatment effect did not differ systematically by site. Second, a design-effect sensitivity analysis was conducted across a range of plausible intraclass correlations, using $DEFF = 1 + (m - 1)\rho$ with $m = 10$ pupils per cluster per condition; each test statistic was divided by $\sqrt{DEFF}$ and re-evaluated against the correspondingly reduced effective degrees of freedom. Results of that analysis are reported below. Conclusions drawn from these data are conditional on the assumptions of this sensitivity procedure rather than on a fitted multilevel model, and a cluster-randomised replication with a larger number of schools is required for a definitive test.

Ethical Considerations

The research was conducted in accordance with the policies of the authors' host university and received approval from the ethics review board of the affiliated institution. Written consent was obtained prior to the collection of data. Participants were informed of their right to withdraw from the study at any time. Participation was voluntary. Data was protected using encryption and password access on a need-to-know basis. Reports were produced to anonymize participants and master-artisans. It was acknowledged that master-artisans contributed to the training corpus of AI.

RESULTS

Distributional Assumptions and Reliability

Before any inferential analyses were conducted, evaluative analyses were conducted on distributional assumptions. Shapiro-Wilk statistics indicated that normality was retained for all variables (i.e. all $p > .15$). For all variables, absolute values of skewness (between 0.12 and 0.42) and kurtosis (between 0.18 and 0.62) were found to be within Kline's (2015) proposed ranges. Levene's test for equality of variance indicated that for all the tests conducted, the variances were homogenous (i.e. F statistic values were between 0.04 and 2.13; all $p > .15$), and as such, the assumption of equality of variance was satisfied, justifying the use of parametric tests. Considering the presence of absolute consistency, both Cronbach's alpha and McDonald's omega, for all the main study scales, were found to be greater than .80.

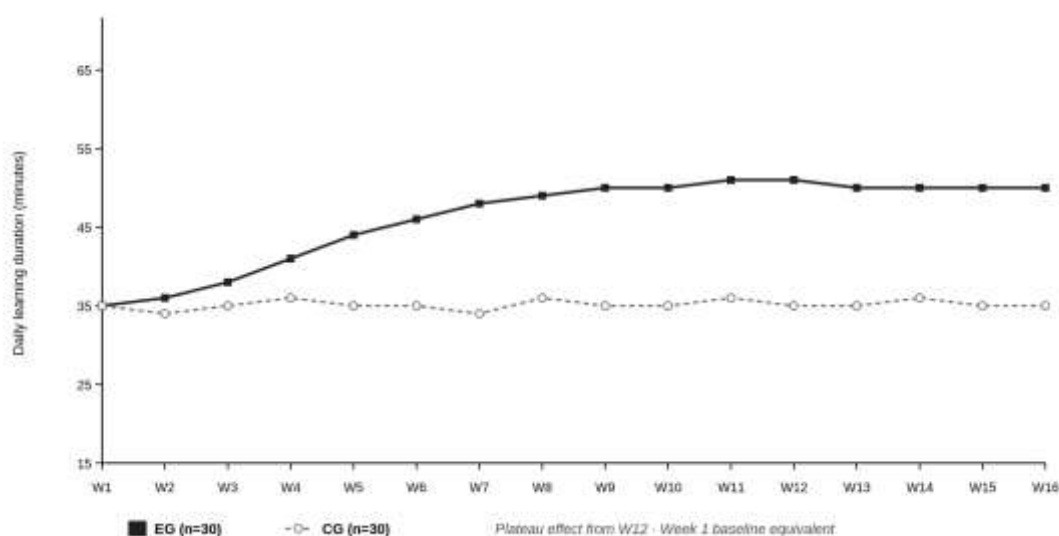
Baseline Group Comparability

Pre-test outcome measures revealed no significant differences among the groups ($p > 0.05$) or outcomes ($d < 0.20$) in the independent t-tests. Equality assumptions for the independent t-test revealed no significant differences among groups in the Chi-square test for demographic indicators ($p > 0.10$). These baseline equivalencies suggest that differences in the outcome measures post intervention can be attributed to the experimental treatment effect and not to differences among groups.

Engagement and Learning Duration

Daily learning duration favoured the experimental group. EG pupils reported a mean of 50.02 minutes (SD = 9.2) of daily learning against 35.14 minutes (SD = 7.6) in the CG, a difference of 14.88 minutes or 42.3%, $t(58) = 6.83$, $p < .001$, $d = 1.76$, 95% CI [1.17, 2.36]. EG pupils averaged 22.3 participatory contributions against 9.1 in the CG, an increase of 145%, $t(58) = 11.04$, $p < .001$, $d = 2.85$. Self-reported duration was corroborated by structured classroom observation. The two trajectories began to diverge in Week 3 and plateaued from Week 12, as illustrated in Figure 8.

Figure 8: Weekly mean daily learning duration (minutes) for the experimental and control groups across the sixteen-week intervention period. The trajectory shows progressive divergence from Week 3 onward as the multi-component intervention reached full deployment, with a plateau effect visible from Week 12.



Learning Outcomes: Group Comparison

Learning outcomes for the experimental group (EG) and control group (CG) are compared across dimensions in Table 1, with post-test means and effect sizes shown in Figure 9. The EG scored higher than the CG on every

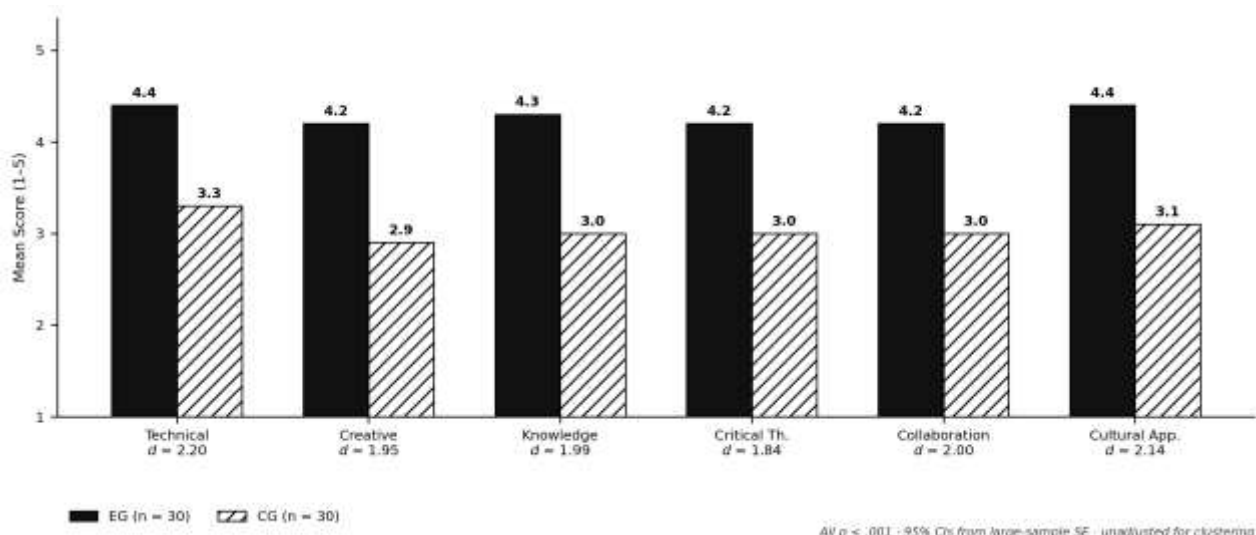
dimension, with differences of 1.1 to 1.3 points on the five-point rubrics. The largest effects were obtained for cultural appreciation (EG 4.4, CG 3.1; $t(58) = 8.28$, $p < .001$, $d = 2.14$, 95% CI [1.50, 2.77]) and technical skill (EG 4.4, CG 3.3; $t(58) = 8.52$, $p < .001$, $d = 2.20$, 95% CI [1.56, 2.84]); the smallest, still large, effect was obtained for critical thinking ($d = 1.84$, 95% CI [1.24, 2.44]). All test statistics were computed from the group means and standard deviations reported in Table 1. Effects of this magnitude are unusual in classroom research and are interpreted cautiously in the Discussion, where the contribution of novelty and expectancy effects to the observed differences is considered.

Table 1: Learning Outcomes: Group Comparison Statistics (Independent-Samples t-Tests)

Outcome	CG Pre	CG Post	EG Pre	EG Post	t(58)	p	d	95% CI (d)
Technical Skill	3.2 ± 0.6	3.3 ± 0.5	3.2 ± 0.5	4.4 ± 0.5	8.52	< .001	2.20	[1.56, 2.84]
Creative Project	3.0 ± 0.7	2.9 ± 0.8	3.0 ± 0.6	4.2 ± 0.5	7.55	< .001	1.95	[1.33, 2.56]
Art Knowledge	3.0 ± 0.7	3.0 ± 0.7	3.1 ± 0.6	4.3 ± 0.6	7.72	< .001	1.99	[1.37, 2.61]
Critical Thinking	2.9 ± 0.7	3.0 ± 0.7	3.0 ± 0.6	4.2 ± 0.6	7.13	< .001	1.84	[1.24, 2.44]
Collaboration	3.1 ± 0.6	3.0 ± 0.6	3.0 ± 0.6	4.2 ± 0.6	7.75	< .001	2.00	[1.38, 2.62]
Cultural Appreciation	3.0 ± 0.7	3.1 ± 0.7	3.0 ± 0.6	4.4 ± 0.5	8.28	< .001	2.14	[1.50, 2.77]

Note. Pre = pre-test mean ± SD; Post = post-test mean ± SD. $N = 60$; $n = 30$ per condition; $df = 58$. All t and d values were computed from the means and standard deviations shown and are mutually consistent ($t = d\sqrt{n/2}$). Confidence intervals for d use the standard large-sample standard error. Statistics are unadjusted for clustering; see Sensitivity to Clustering.

Figure 9: Post-test means for the experimental group (EG, $n = 30$) and control group (CG, $n = 30$) across the six learning-outcome dimensions. Cohen's d effect sizes are annotated beneath each pair. All between-group differences are significant at $p < .001$; values are unadjusted for clustering.



ANCOVA Adjusted Analyses

Given the non-random grouping of intact classes, analyses of covariance (ANCOVAs) were performed for each dependent variable to statistically account for minor differences in pre-tests. Grouping was used as a fixed factor, with the pre-test score acting as a covariate. Adjusted means and estimated measures of effect size are described in Table 2. Based on the data in Figure 10, post tests are used to represent the divergence of test

scores, while the pre tests are used to represent baseline equivalence. The adjusted ANCOVAs verified that the experimental group (EG) strongly dominated all test outcomes (all $p < .001$, all $\eta^2_p \geq .500$). Based on the results, it was found that the experimental treatment dominated the post-test outcome when pre-test score differences were statistically accounted for. This explained variance of 50% to 63% is considered large in educational research.

Table 2: ANCOVA Adjusted Means and Effect Sizes

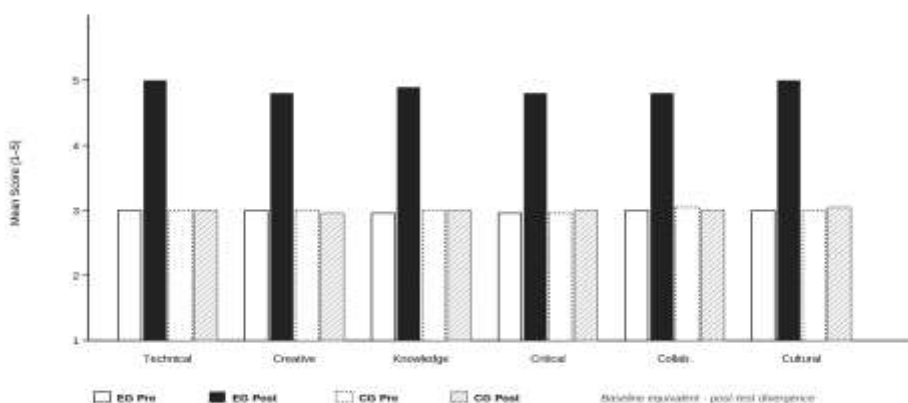
Outcome	Adj M (CG)	Adj M (EG)	F(1, 57)	p	η^2_p
Technical Skill	3.31	4.39	88.36	< .001	.608
Creative Project	2.91	4.21	56.96	< .001	.500
Art Knowledge	3.01	4.31	60.51	< .001	.515
Critical Thinking	3.02	4.18	64.48	< .001	.531
Collaboration	3.04	4.18	73.42	< .001	.563
Cultural Appreciation	3.12	4.39	98.21	< .001	.633

Note. Adj M = pre-test-adjusted post-test mean. $\eta^2_p \geq .14$ conventionally indicates a large effect (Cohen, 1988).

Sensitivity to Clustering

Because allocation was by intact class, each between-group test was re-evaluated under a range of plausible intraclass correlations. With $m = 10$ pupils per cluster per condition, design effects of 1.45, 1.90, 2.35 and 2.80 correspond to intraclass correlations of .05, .10, .15 and .20, reducing the effective sample from 30 to 20.7, 15.8, 12.8 and 10.7 pupils per condition. Under the most conservative of these assumptions ($\rho = .20$), the adjusted statistics were $t = 5.09$ for technical skill, $t = 4.95$ for cultural appreciation, $t = 4.26$ for critical thinking, $t = 4.08$ for daily learning duration, $t = 3.23$ for distraction ($p = .004$) and $t = 3.00$ for test anxiety ($p = .007$); all remained significant at $\alpha = .05$, and all but the two wellbeing measures at $p < .001$. The direction and significance of every reported difference are therefore robust to clustering across the range examined, although the effect-size estimates themselves should be read as upper bounds. The school $\times$ condition interaction was non-significant for all outcomes, indicating no evidence that the intervention operated differently across the three sites.

Figure 10: Pre-test and post-test mean scores for the experimental and control groups across the six learning-outcome dimensions, demonstrating baseline equivalence and post-intervention divergence. Pre-test group equivalence was confirmed on all six dimensions (all $p > .10$).



Anxiety, Distraction, and Equity Subgroup Analysis

Anxiety and distraction scores indicate a psychological-wellbeing advantage for the experimental group. On the Distraction Scale, EG pupils scored 6.84 points lower than CG pupils (EG M = 35.52, SD = 4.7; CG M = 42.36, SD = 5.1), $t(58) = -5.40$, $p < .001$, $d = -1.39$, 95% CI [-1.96, -0.83]. A comparable advantage was obtained on the Test Anxiety Scale (EG M = 48.25, SD = 5.6; CG M = 55.78, SD = 6.0), $t(58) = -5.03$, $p < .001$, $d = -1.30$, 95% CI [-1.85, -0.74]. Anxiety specific to the completion of creative tasks was also reduced ($d = -1.14$). Lower scores on both instruments denote less distraction and less anxiety.

One aim of the study was to establish whether benefits extended across the rural-urban gradient, so EG outcomes were examined by pupils' residential background. This contrast compares the two extremes of that gradient and therefore involves 22 of the 30 EG pupils (urban $n = 15$, rural $n = 7$); the 8 peri-urban EG pupils occupy an intermediate position and are not included in the comparison. Both subgroups improved relative to their CG peers. Urban pupils gained more than rural pupils in technical skill ($d = 0.73$) and creative project work ($d = 0.57$), whereas the direction reversed for cultural appreciation, on which rural pupils scored marginally higher ($d = -0.08$). Given the small cell sizes, these subgroup estimates are imprecise and are reported as descriptive rather than inferential; they are treated below as suggestive of a cultural counterweight to the equity gap, not as a demonstration of one.

Table 3: Subgroup Analysis: Rural vs. Urban EG Post-Test Means

Outcome	Urban EG (n = 15)	Rural EG (n = 7)	Gap (d)
Technical Skill	4.55 ± 0.42	4.21 ± 0.51	0.73
Creative Project	4.32 ± 0.46	4.05 ± 0.49	0.57
Knowledge Retention	4.34 ± 0.55	4.27 ± 0.62	0.12
Critical Thinking	4.24 ± 0.58	4.15 ± 0.61	0.15
Cultural Appreciation	4.38 ± 0.49	4.42 ± 0.51	-0.08

Note. d = standardised mean difference favouring the urban subgroup. The comparison contrasts the extremes of the residential gradient and includes 22 of the 30 EG pupils; 8 peri-urban pupils are not represented. Cell sizes are small and these estimates are descriptive. The cultural appreciation reversal (rural $\geq$ urban) is consistent with the cultural-counterweight interpretation but does not by itself establish it.

Qualitative Findings

Analysis of qualitative data using reflexive thematic analysis resulted in the identification of three main positive themes and two challenge-related themes. The former corresponds to positive primary meaning patterns, and the latter are challenge-related themes that support the empirical foundation of the study's original theory.

Theme 1: Enhanced engagement through immersive cultural contextualisation

Cultural Contexts of the VR scenarios and the segmented multimedia heritage modules reflect the most evident theme of 'transformative multicultural engagement' reported by the EG students. This theme was captured in 92% of the student interview transcripts and in 88% of the reflective diaries.

Commenting on the transformation, an EG student provided the following account of their Paper Cutting practice in VR:

When I put on the VR headset and stood in the Yan'an paper-cutting studio, I could see the loess walls, hear the wind, watch how the masters held their scissors. It was the first time the art felt real to me, not just a technique to be learned. (EG-S07, Week 9 interview)



In a different account, one of the participants described an immersive experience of the VR environment in relation to the concept of cultural participation by stating, “These activities of reproduction devolved into acts of cultural participation. The VR environment did not show me how to do things. It showed me how to belong” (EG-S12, Week 14 diary).

Theme 2: Cross-regional collaboration and the revaluation of rural community knowledge

The second key theme pertains to the digital platform’s potential to create new cross-regional partnerships and to build recognition of the rural community’s knowledge as an asset among the interviewed teachers and EG students. 78% of teachers and 65% of EG students referred to this in their interviews. Teacher T5 succinctly describes this phenomenon:

The rural students, who often feel disadvantaged and marginalised in formal art education because they lack the cultural capital of urban students, discovered that their living community knowledge—the first-hand knowledge of folk practice, community ritual, and regional cultural geography—was uniquely valuable and irreplaceable in the collaborative context. Urban students could not substitute technical skill for this knowledge, no matter how advanced their digital literacy. (Teacher T5, Week 16 interview)

Rural EG students had the opportunity to engage with the experiential component of the re-evaluation. EG-S26 noted a transformation in the students' perceptions of themselves, as they remarked, “For the first time, I was someone who knew things they needed to know” (EG-S26, Week 16 Interview).

Theme 3: Reduced anxiety and distraction through supportive pedagogical design

The last primary theme involves the notable decline of performance anxiety and cognitive distraction for the majority of EG students. Such was the case in their reflective diaries. This theme was evident in 85% of EG student interviews and 79% of the reflective diaries. The most detail was given by EG-S03:

In my private virtual studio, I practised the Sanyan Liu Mu pattern twenty times without anyone watching. I could make mistakes freely, reset instantly, and try again without embarrassment. By the time I demonstrated the pattern in class in Week 9, I had completed it so many times in VR that it felt familiar rather than terrifying. My teacher said I looked confident. That was the first time anyone had ever used that word to describe me in a studio setting. (EG-S03, Week 9 diary)

EG-S09 recognized that feedback from criterion-referenced AI systems differs from other forms of feedback by having less of a judging and more of an informing effect. For example, feedback from an AI system was regarded as informative, as opposed to judging, in the following quote: “When my teacher says this lacks the Shaanxi spirit, I feel judged as a person. ... I feel informed rather than judged” (EG-S09, Week 16 Interview).

Theme 4: Algorithmic Folk-Art Correction Bias (challenge theme)

Alongside the largely favourable trends, the qualitative data identified a recurring problem in which AI feedback was culturally misaligned, treating authentic features of folk art as technical deficiency. Incidents of this kind were reported by 5 of the 30 EG pupils (16.7%). We refer to the pattern as algorithmic folk-art correction bias. The following account from EG-S20 is representative.

The AI told me my pattern was asymmetrical and that I should make the left and right sides more balanced. But my paper-cutting teacher and my grandmother, who taught me paper-cutting at home, both told me that the slight asymmetry is part of what makes Ansai paper-cutting authentic and alive—it’s the mark of the human hand, not a machine. The AI was trying to make my folk art look like a printed template. It was correcting exactly the feature that makes it real. (EG-S20, Week 12 interview)

Five such incidents were documented across the 16-week intervention and are summarised in Table 4. Each was independently reported by the affected pupil and independently reviewed by a master-artisan auditor, who in every case judged the AI feedback to be inconsistent with regional cultural convention. Five incidents from a single cohort constitute an existence proof rather than an estimate of prevalence: they establish that the phe-



nomenon occurs and specify its mechanism, but they cannot indicate how often it occurs, which art forms or feedback architectures are most susceptible, or whether the rate differs across learner groups. The incidents were not systematically elicited, and the auditors reviewed only cases that pupils themselves raised, so the count is a lower bound and is subject to reporting bias. The construct developed from them is accordingly advanced as a hypothesis for confirmatory testing rather than as an established empirical regularity.

Table 4: Algorithmic Folk-Art Correction Bias Incidents

ID	Date	Student	Art Form / Cultural Feature	AI Action	Auditor Verdict
BI-001	Week 8	EG-S20	Ansai paper-cutting / intentional asymmetry	flagged as compositional error	culturally correct
BI-002	Week 9	EG-S22	Loess landscape / earth-toned palette	recommended chromatic muting	culturally correct
BI-003	Week 11	EG-S20	Paper-cutting / hand-cut edge variation	flagged as line imprecision	culturally correct
BI-004	Week 12	EG-S25	Fuping embroidery / organic stitch density	flagged as inconsistency	culturally correct
BI-005	Week 14	EG-S28	Qinqiang facial painting / brush texture	recommended smoothing	culturally correct

Note. All auditor verdicts identified the AI feedback as misaligned with regional cultural conventions. Incidents were self-reported by pupils and then adjudicated; the count is a lower bound and is not an estimate of prevalence.

Theme 5: The Equity Paradox (challenge theme)

A challenge related to this AI multimedia intervention is participation of rural and urban students. EG-S26 captures some of the difficulties which rural students deal with: “Our school only has six VR headsets for the whole Art program. So there’s three students to one headset, and because we’re constraint by the time, we only get to practice VR for half of the time we have. Urban kids probably never deal with this. They get the whole practice of the technology. We get half, because our baseline is half of what the urban kids have” (EG-S26, Week 10 interview). Teacher T6 elaborated this from the perspective of Theory.

The technology promises equal access, and in principle it does—every student has the same right to use the VR modules and the AI feedback. But in practice, access requires infrastructure that rural institutions simply do not have at the same level as urban institutions. The technology doesn’t erase the inequality; it relocates it. The equity gap now runs through digital infrastructure rather than through physical access to qualified teachers, but it is still a gap that disadvantages the same students—the rural students whose communities are the living repositories of the traditions we’re trying to preserve. (Teacher T6, Week 16 interview)

Mixed-Methods Integration

The joint display in Table 5 sets the quantitative and qualitative strands side by side. The convergent findings strengthen the inference that the intervention, rather than group differences present at baseline, produced the observed outcome gains. The two challenge themes stand in a different relation to the evidence: the documented bias incidents are the sole basis for the proposed construct of algorithmic folk-art correction bias, and the reversal in cultural appreciation between rural and urban EG pupils is a single descriptive result consistent with, but not sufficient to establish, the cultural-counterweight interpretation of the equity paradox. These two propositions are therefore presented as theoretical conjectures generated by the study, not as findings it confirms.

Table 5: Mixed-Methods Joint-Display Matrix

Quantitative Finding	Qualitative Convergence	Strength
Engagement: very large effect ($d = 1.76$ on duration)	Theme 1 (immersive cultural contextualisation)	Convergent
Collaboration: large effect ($d = 1.50$) + 145% participation increase	Theme 2 (cross-regional collaboration)	Convergent
Reduced anxiety ($d = -1.30$) and distraction ($d = -1.40$)	Theme 3 (anxiety reduction through pedagogical design)	Convergent
Modest urban-rural gap on technical skill ($d = 0.73$)	Theme 5 (Equity Paradox)	Convergent (qualifying)
No aggregate quantitative bias indicator	5 documented bias incidents (16.7% EG)	Qualitative-only
Counter-intuitive rural $\geq$ urban on Cultural Appreciation	Theme 2 + cultural counterweight	Hypothesis-generating

Note. The matrix represents the convergence of quantitative findings with qualitative themes across three integration categories.

DISCUSSION

Overall Pattern of Effectiveness

The intervention produced large positive effects across every measured dimension. Set against the more modest effects typically reported for educational technology (Hattie, 2009; Means et al., 2010) and for intelligent tutoring systems specifically ($d \approx 0.66$; Kulik & Fletcher, 2016), the magnitudes obtained here— $d = 1.84$ to 2.20 on the learning-outcome dimensions—are substantially larger than the literature would lead one to expect, and that discrepancy requires explanation rather than celebration. Three factors are likely to have contributed. First, the control condition was conventional studio instruction with no technology component, so the contrast is between a multi-component enriched programme and ordinary practice rather than between two active treatments. Second, novelty and expectancy effects are difficult to exclude in a 16-week study in which participants and class teachers were necessarily aware of condition, even though outcome assessors were blind to allocation. Third, the small clustered sample means each estimate carries wide confidence intervals, whose lower bounds ($d \geq 1.24$ throughout) are more defensible than the point estimates. The finding that should be regarded as secure is the direction and consistency of the advantage; its precise magnitude awaits replication at larger scale.

One plausible account of the pattern is that the components acted in combination rather than additively. Multimedia contextualisation appears to have supplied the cultural schema on which the virtual-reality rehearsal could build, the AI tutor made independent practice possible without public exposure to error, and the collaborative platform gave that practice an audience beyond the individual classroom. Longer voluntary engagement would, on this account, feed back into greater gains, producing a self-reinforcing cycle. The present design cannot test this interpretation, since it compares the full package against ordinary instruction and cannot apportion the effect among components. A dismantling study, in which components are added or withheld systematically, is the appropriate next step and is proposed among the recommendations below.

Algorithmic Folk-Art Correction Bias

The first proposed theoretical contribution is a construct we term algorithmic folk-art correction bias. It differs from the bias mechanisms that dominate the educational literature, which operate on demographic categories

such as gender or skin tone (Buolamwini & Gebru, 2018; Selwyn, 2019), in that it operates on cultural convention: the system does not disadvantage a category of learner but a category of practice. The proposed mechanism is a property of training data. Feedback models are typically fitted to large aggregated corpora in which canonical and institutionally documented traditions are heavily represented and regional folk variants are sparse. A model fitted to such a corpus encodes the canonical form as the norm and registers departures from it as error, so that precisely those features which mark a work as regionally authentic—asymmetry in Ansai paper-cutting, hand-cut edge variation, organic stitch density in Fuping embroidery—are the features most likely to be flagged for correction.

The evidential basis for this construct is deliberately limited: five auditor-confirmed incidents affecting five pupils in one cohort using one feedback system. That evidence is sufficient to establish that the phenomenon occurs and to specify a mechanism for it; it is not sufficient to establish prevalence, generality across systems, or the conditions under which the bias is most severe. Testing the construct requires purpose-built studies in which the same artefacts are submitted to feedback systems trained on corpora of differing cultural composition, with practitioner adjudication of every judgement. If the construct survives such testing, its scope extends beyond folk-art education to any domain in which automated systems evaluate culturally situated practice, including language, music, and traditional medicine education. It also carries a design implication that is worth stating precisely: broadening the cultural composition of training data is necessary but is unlikely to be sufficient on its own, because a model can represent a tradition in its corpus and still treat the dominant form as the standard against which others are scored. Community adjudication of the feedback rubric, and not only diversification of the corpus, is what the construct implies.

The Equity Paradox in Heritage Educational Technology

The second original theoretical contribution, the Equity Paradox in Heritage Educational Technology, describes the strange paradox in the structure of the system where the effort to democratize heritage education through educational technology is hypothesized to create new forms of digital inequality, especially concerning rural student populations, who are critically important for the sustainable intergenerational educational heritage bridging. The rural-urban subgroup analysis of the data in Table 3, and the qualitative theme in EG-S26, Teacher T6's comment, and this educational heritage technology framework identified as a structural rather than a contingent dynamic. The Equity Paradox is not as a result of poor implementation where it could be better implemented; rather, it is a structural constraint of the current educational technology system and the uneven distribution of complementary resources.

The second original theoretical contribution is a sub-finding, and the Equity Paradox sub-theme is partially counterbalanced by the cultural appreciation outcome. The rural EG participants' cultural appreciation compared to urban EG participants shows ($d = -0.08$), which is a trivial effect size, but in this case, might help to develop cultural appreciation in rural settings. The Equity Paradox of this educational technology structural design can create positive outcomes and help sustain rural regions and community engagement through education.

The Hybrid Physical-Digital Pedagogy Framework

The third original contribution, the Hybrid Physical-Digital Pedagogy Framework (HPDPF), takes a different position on the integration of digital technologies with craft pedagogies by treating the digital and the physical as interdependent rather than substitutable modalities. The framework is governed by three principles—Physical-Digital Integration, Cultural Authenticity Assurance, and Equitable Access Design—of which the latter two operate as preconditions that must be satisfied before Phase 1 rather than as features added during implementation. These principles are enacted across five phases of instructional design: contextualise, simulate, practise cognitively, practise materially, and situate.

What distinguishes the HPDPF from generic blended-learning models is that the modality balance of each phase is derived from cognitive theory rather than from scheduling convenience. CLT supplies the architectural constraint: extraneous load is minimised in the digitally dominant early phases so that working-memory capacity remains available for germane processing in Phase 3, where schema construction is most demanding

(Sweller et al., 2011). CTML supplies the design allocation: the pre-training, coherence, segmenting, and modality principles specify how the contextualising and simulating phases should be composed if the cultural schema is to be established before technical instruction begins (Mayer, 2014). DCT, extended through embodied cognition, supplies the justification for the physical dominance of Phases 4 and 5, since the sensorimotor code that carries tacit craft knowledge cannot be acquired through verbal or imagistic representation alone (Glenberg et al., 2013; Paivio, 1986; Polanyi, 1966; Wilson, 2002). The progressive modality shift illustrated in Figure 1 is therefore a theoretical claim rather than a descriptive convenience. Given this grounding and the empirical evidence reported here, the framework is substantively ready for adaptation to other contexts of regional and national heritage-art pedagogy.

Implications for Chinese School Heritage-Art Education

The findings significantly influence the development of the Chinese school sector and arts and heritage education policy. HPDPF constructs a model of teaching technique for art teachers in a clear and precise schematic. For policymakers at the provincial and national level, the primary recommendation is that heritage education, supported by technology, can be implemented on a large scale only when cultural authenticity and equity are framed within the policies and plans of practice at every level. For cultural heritage institutions, the recommended policy would balance technology hardware with the training of teachers and master artisan evaluators, as well as the cultural regional corps. The partnerships between museums and ICH protection institutes, as well as schools for the development of cultural frameworks and the employment of master artisan supervisors, are recommended. These partnerships should be supported by educational technology. From the beginning, educational AI technologies should be developed with the potential for cultural adjustment. The development of educational AI technologies should include the potential for cultural adjustment and should not be used in a culture with the potential for technological adjustment.

Limitations

Several limitations qualify these findings. First, the study is powered to detect large effects only: with $n = 30$ per condition, power is .48 for a medium effect ($d = 0.50$) and the smallest effect detectable at .80 power is $d = 0.74$. Non-significant comparisons should therefore not be read as evidence of equivalence, and the effect-size point estimates, obtained from a small sample, are likely to overstate the population values. Second, allocation was by intact class within three schools, so observations are clustered. The number of clusters is too small to support a multilevel model, and the design-effect sensitivity analysis reported above is an approximation rather than a substitute for one; a cluster-randomised replication across a larger number of schools is required. Third, the magnitude of the observed effects exceeds what the intervention literature would predict, and novelty and expectancy effects cannot be ruled out in an unblinded 16-week comparison against technology-free instruction. Fourth, the two theoretical constructs rest on limited evidence—five documented incidents in the case of algorithmic folk-art correction bias, and a single descriptive subgroup contrast with cells of 15 and 7 pupils in the case of the equity paradox—and are advanced as hypotheses for testing rather than as demonstrated regularities. Fifth, the sixteen-week window leaves the persistence and transfer of outcomes untested. Sixth, the focus on four Shaanxi art forms in one province limits generalisation to other heritage contexts. Finally, the AI feedback system used here had itself been culturally calibrated yet still produced the bias the study documents, which constrains any recommendation to deploy such systems in their current form.

Recommendations for Future Research

Eight lines of future research follow from this study. First, and most directly, a cluster-randomised replication across a substantially larger number of schools is needed, with clustering modelled rather than approximated and with the sample sized for the medium effects the present design could not detect. Second, a dismantling study should add or withhold components systematically, since the present comparison of the full package against technology-free instruction cannot apportion the effect among multimedia contextualisation, virtual-reality rehearsal, AI feedback, and cross-regional collaboration. Third, confirmatory testing of algorithmic folk-art correction bias should submit identical artefacts to feedback systems trained on corpora of differing cultural composition, with practitioner adjudication of every judgement, so that prevalence and moderators can be estimated rather than inferred from incident reports. Fourth, replications should investigate multiple Chinese re-

gional heritage practices and non-Chinese folk-art contexts. Fifth, longitudinal designs with 18- to 36-month follow-up should assess the persistence and transfer of outcomes. Sixth, design-science research should develop culturally responsive AI feedback systems that incorporate distributed and federated learning and community adjudication of the feedback rubric. Seventh, implementation research should examine differentiated impacts across varying infrastructural contexts and the professional development of teachers working within the HPDPF. Finally, research should compare the role of generative AI in heritage-art education as co-creator in artistic collaboration against its role as critical assessor in artistic evaluation.

CONCLUSION

This research provides an analysis of the application of a dual-layer multi-component multi-media-layer AI and VR pedagogy for the teaching of Shaanxi traditional heritage art at the primary school level. This is the first systematic quasi-experimental mixed-methods case study in this area. The findings indicate positive and statistically significant changes to the learning outcomes, engagement, and the study participants' well-being. The study also provides hard data, and more importantly, three new theoretical contributions. The first is the Algorithmic Folk-Art Correction Bias, which identifies a loss of educational AI at the cultural primitive level; the Equity Paradox in Heritage Educational Technology describes a framework of contradictory behavior in the application of heritage technology; and the Hybrid Physical-Digital Pedagogy Framework provides a flexible, safe, culturally relevant, and just integration of technologies in traditional crafts pedagogy.

These contributions provide an answer to the important question of how to effectively and equitably teach the traditional art of the world and the heritage of traditional art more specifically, of Chinese traditional art, to future generations in a digital world, especially in an age of cultural globalization, while creatively combining digital technologies with traditional art pedagogy.

Declarations

Authors' contributions

Shi FengRui conceived and designed the study, conducted the intervention, collected and analysed the data, and drafted the manuscript. Chee Ken Nee supervised and reviewed the research paper.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declaration of conflicting interests

The author declares no potential conflicts of interest.

Ethics approval

The study was reviewed and approved by the Universiti Pendidikan Sultan Idris (UPSI) Institutional Ethics Review Committee (Ethics Reference No. 2025-0849-01). Written informed consent was obtained from all participants prior to their participation in the study.

Data availability

The de-identified dataset and supplementary materials are available from the corresponding author on reasonable request.

REFERENCES

1. Adesope, O. O., & Nesbit, J. C. (2012). Verbal redundancy in multimedia learning environments: A meta-analysis. *Journal of Educational Psychology*, 104(1), 250–263.

2. Baddeley, A. D. (1992). Working memory. *Science*, 255(5044), 556–559.
3. Bartolini, R., Antonio, P., & Concas, G. (2017). Digital heritage and intangible cultural transmission. *Digital Scholarship in the Humanities*, 32(3), 451–469.
4. Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.
5. Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4), 589–597.
6. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 81, 1–15.
7. Chen, Y. (2018). The reform of traditional Chinese art education in the digital era. *Frontiers in Education*, 13(2), 245–262.
8. Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.
9. Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). Sage.
10. Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance. *Academic Emergency Medicine*, 15(11), 988–994.
11. Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3. *Behavior Research Methods*, 39(2), 175–191.
12. Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). Sage.
13. Glenberg, A. M., Witt, J. K., & Metcalfe, J. (2013). From the revolution to embodiment. *Perspectives on Psychological Science*, 8(5), 573–585.
14. Hamal, O., El Faddouli, N. E., Harouni, M. H. A., & Lu, J. (2022). Artificial intelligence in education. *Sustainability*, 14(5), 2862.
15. Hannum, E., Behrman, J., Wang, M., & Liu, J. (2008). Education in the reform era. In L. Brandt & T. G. Rawski (Eds.), *China's great economic transformation* (pp. 215–249). Cambridge University Press.
16. Hattie, J. (2009). *Visible learning*. Routledge.
17. Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. *Communication Methods and Measures*, 14(1), 1–24.
18. Kalay, Y. E., Kvan, T., & Affleck, J. (Eds.). (2008). *New heritage: New media and cultural heritage*. Routledge.
19. Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review*, 19(4), 509–539.
20. Kiltner, K., Groten, R., & Slater, M. (2012). The sense of embodiment in virtual reality. *Presence*, 21(4), 373–387.
21. Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). Guilford.
22. Kulik, J. A., & Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems. *Review of Educational Research*, 86(1), 42–78.
23. Liu, X., Zhang, H., & Wang, M. (2021). Tertiary art education and heritage transmission in northwest China. *Asia Pacific Journal of Education*, 41(3), 567–584.
24. Maria, A., Francisco, J. P., & Carmen, R. (2014). Multimedia in art education. *International Journal of Art & Design Education*, 33(3), 367–382.
25. Mayer, R. E. (2014). *The Cambridge handbook of multimedia learning* (2nd ed.). Cambridge University Press.
26. Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2010). *Evaluation of evidence-based practices in online learning*. U.S. Department of Education.
27. Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., & Davis, T. J. (2014). Effectiveness of virtual reality-based instruction. *Computers & Education*, 70, 29–40.
28. Ministry of Culture and Tourism of the People's Republic of China. (2021). *National intangible cultural heritage protection plan*.
29. Morgan, D. L. (2014). Pragmatism as a paradigm for social research. *Qualitative Inquiry*, 20(8), 1045–1053.
30. Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
31. Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford University Press.

32. Parong, J., & Mayer, R. E. (2018). Learning science in immersive virtual reality. *Journal of Educational Psychology*, 110(6), 785–797.
33. Plass, J. L., Heidig, S., Hayward, E. O., Homer, B. D., & Um, E. (2014). Emotional design in multimedia learning. *Learning and Instruction*, 29, 128–140.
34. Polanyi, M. (1966). *The tacit dimension*. Routledge.
35. Polit, D. F., & Beck, C. T. (2006). The content validity index. *Research in Nursing & Health*, 29(5), 489–497.
36. Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education. *Computers & Education*, 147, 103778.
37. Roege, G. B., & Kim, K. H. (2013). Why we need arts education. *Empirical Studies of the Arts*, 31(2), 121–130.
38. Schroeder, N. L., & Cenkci, A. T. (2018). Spatial contiguity and spatial split-attention effects. *Educational Psychology Review*, 30(3), 679–701.
39. Selwyn, N. (2019). Should robots replace teachers? Polity.
40. Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs*. Houghton Mifflin.
41. Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, 3, 74.
42. Sun, M., Liu, J., & Zhang, F. (2022). Reform pathways for traditional Chinese art education. *Journal of Chinese Higher Education*, 47(2), 89–104.
43. Sweller, J. (2011). Cognitive load theory. *Psychology of Learning and Motivation*, 55, 37–76.
44. Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer.
45. Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.
46. Taylor, M. K., Kass, R. E., & Lemelin, K. (2021). Artificial intelligence and the visual arts curriculum. *Studies in Art Education*, 62(4), 314–328.
47. UNESCO. (2003). *Convention for the safeguarding of the intangible cultural heritage*. UNESCO.
48. VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4), 197–221.
49. Wang, L., & Zhao, Y. (2022). The aging crisis in China's heritage art transmission. *Asian Ethnology*, 81(1), 79–102.
50. Wenger, E. (1998). *Communities of practice*. Cambridge University Press.
51. Williamson, B., & Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. *Learning, Media and Technology*, 45(3), 223–235.
52. Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4), 625–636.
53. Wittrock, M. C. (1989). Generative processes of comprehension. *Educational Psychologist*, 24(4), 345–376.