International Journal
Information Technology Applications (ITA)

Volume 15, Number 2, August 2026

LONGITUDINAL RECALL AND ENGAGEMENT VIA REAL-TIME 2D AI AVATARS IN VIRTUAL HERITAGE ENVIRONMENTS

Ján Lacko, Michal Drobný

Abstract:

Immersive virtual reality (VR) offers transformative potential for cultural heritage education, yet populating historical environments with authentic, believable pedagogical agents remains technically and financially prohibitive. Traditional 3D character pipelines require intensive modeling and frequently trigger the "uncanny valley", whereas static virtual museum exhibits risky cognitive overload via split-attention effects. This study evaluates an alternative paradigm: streaming real-time, 2D AI-generated neural talking-head avatars animated directly from archival portraits into an immersive Unreal Engine 5.7 environment.

A randomized controlled trial was conducted with university students ($N = 64$) assigned either to an experimental condition featuring an audio-driven 2.5D billboarded historical avatar or a control condition utilizing a traditional non-avatar VR guide with static placards and voice-over narration. Longitudinal knowledge retention was measured immediately post-intervention (T_0), after one week (T_1), and after three months (T_2), alongside psychometric assessments of cognitive load (NASA-TLX), social presence, and historical empathy. A 2×3 mixed-design ANOVA revealed a significant interaction effect ($F(1.78, 110.36) = 32.41, p < .001, \eta_p^2 = .343$). While initial recall was comparable (84.62% vs. 81.35%), the 2D AI avatar cohort demonstrated significantly buffered memory retention after three months (68.45% vs. 51.22%, $p < .001$). Furthermore, MANOVA results confirmed substantial reductions in extraneous cognitive load ($p < .001$) alongside higher social presence and affective engagement. Grounded in Social Agency Theory and Cognitive Load Theory, these findings establish real-time 2D generative media as a scalable, pedagogically robust framework for interactive historical storytelling in virtual heritage.

Keywords:

Virtual reality, cultural heritage, generative artificial intelligence, cognitive load, longitudinal evaluation.

ACM Computing Classification System:

Computing methodologies - Artificial intelligence - Natural language processing;
Applied computing - Education - Interactive learning environments;
General and reference - Cross-computing tools and techniques - Evaluation

Introduction

Immersive virtual reality (VR) has transformed cultural heritage education from passive observation into active, experiential engagement. By reconstructing lost architectural monuments, historical events, and archival artifacts in high-fidelity 3D spaces, VR enables learners to contextualize history with unprecedented spatial immersion. However, populating these digital environments with believable, historically accurate pedagogical agents remains a persistent challenge.

Historically, cultural heritage institutions have relied on two primary modalities to deliver narrative instruction in VR:

1. **Static, Non-Diegetic Expositions:** Utilizing ambient audio guides paired with static archival photographs, text placards, and 3D artifact scans.
2. **Polygonal 3D Historical Reconstructions:** Developing fully rigged, 3D animated avatars via manual 3D sculpting, motion capture, and blendshape lip-synchronization.

While 3D character modeling provides stereoscopic depth and 6-Degrees-of-Freedom (6-DoF) physical volume, it presents substantial hurdles. High-fidelity 3D historical reconstructions are cost-prohibitive, labor-intensive, and computationally demanding. Furthermore, budget-constrained 3D avatars frequently fall into the "uncanny valley"—where subtle deviations in facial micro-expressions, skin translucency, and ocular saccades evoke eerie, unnatural sensations among learners, directly compromising historical solemnity and cognitive engagement. Recent advancements in deep generative media and neural talking-head synthesis (e.g., audio-driven latent diffusion and motion-transfer models) offer a paradigm shift. By animating single, authentic 2D historical portraits in real time with synchronized speech, micro-expressions, and natural head poses, generative AI allows institutions to breathe lifelike motion directly into archival imagery.

Despite the rapid evolution of 2D neural video synthesis, its real-time integration as dynamic textures within immersive 3D game engines remains underexplored in educational literature. Existing research predominantly evaluates generative talking-head models on flat 2D screens (e.g., video calls, asynchronous e-learning video modules). Crucial questions remain regarding how learners perceive 2.5D planar video billboarding when embedded inside a fully immersive, 6-DoF spatial VR environment:

- Does the loss of true volumetric stereoscopic depth disrupt the learner's spatial presence?
- Does the absolute photorealism and authenticity of an archival 2D face outweigh the depth cues of a synthetic 3D avatar or the simplicity of traditional text-and-audio museum exhibits?
- How does embodied, personified historical narration impact **longitudinal retention** across extended academic intervals (e.g., 1 week and 3 months) compared to standard non-avatar VR learning?

Grounding this investigation within Mayer’s Cognitive Theory of Multimedia Learning, Social Agency Theory, and Sweller’s Cognitive Load Theory, this study evaluates the empirical and pedagogical viability of real-time 2D AI avatars within an immersive cultural heritage application. By activating social interaction schemas through gaze alignment and dynamic facial cues, embodied agents transform passive knowledge reception into an engaging episodic dialogue.

To address these gaps, this study investigates three central research questions:

- **RQ1 (Longitudinal Retention & Memorability):** How does real-time interaction with an AI-driven 2D archival portrait avatar in VR affect short-term (1 week) and long-term (3 months) factual and conceptual knowledge retention compared to traditional non-avatar VR learning environments?

- **RQ2 (Spatial & Social Presence):** How does a dynamic 2.5D billboarded neural talking-head avatar affect learners' sense of spatial presence, social presence (co-presence), and perceived historical authenticity within an Unreal Engine 5.7 virtual environment?

- **RQ3 (Cognitive Workload & Affective Resonance):** To what extent does real-time 2D avatar-mediated storytelling reduce extraneous cognitive load and enhance historical empathy compared to static text-and-audio museum guidance?

1 State of the art

The integration of real-time generative artificial intelligence (AI) with immersive Virtual Reality (VR) for cultural heritage education intersects multiple research domains. This chapter provides a critical synthesis of literature across four key pillars:

1. Virtual heritage pedagogy and intelligent pedagogical agents (IPAs);
2. Cognitive architectures, Cognitive Load Theory, and Social Agency Theory in VR;
3. The perceptual dynamics of character embodiment, authenticity, and the Uncanny Valley; and
4. The evolution of real-time neural talking-head synthesis and generative video streaming.

1.1 Virtual Heritage Pedagogy and Intelligent Pedagogical Agents

Virtual heritage (VH) environments have shifted from passive visual archives into interactive, experiential learning spaces [1][2]. High-fidelity real-time rendering engines have substantially lowered the technical barrier to rendering complex, photogrammetric reconstructions of historical monuments with photorealistic global illumination and digital twin architectures [3][4].

Despite visual advancements in environmental rendering, populating these spaces with historical actors remains challenging. Intelligent Pedagogical Agents (IPAs) and virtual guides foster learning motivation, guide navigational focus, and provide contextual anchoring within complex simulations [5][6]. Immersive cultural heritage environments benefit from virtual human guides that deliver contextualized historical narratives rather than static text displays [4][7].

However, historical education imposes unique constraints: historical figures require high fidelity. While fictional educational simulations tolerate stylized or cartoon-like agents, cultural heritage requires visual and documentary accuracy [1][7]. When historical personas are presented inaccurately or anachronistically, learner trust and historical solemnity erode [3][7].

1.2 Cognitive Foundations: Cognitive Load and Social Agency in VR

The pedagogical efficacy of immersive agents is rooted in Cognitive Load Theory (CLT) [8] and the Cognitive Theory of Multimedia Learning (CTML) [9][10]. The Split-Attention and Redundancy Effects: In traditional virtual museum exhibits, learners frequently suffer from the spatial split-attention effect [11]. Users must continuously divide visual attention between 3D artifacts, static archival photographs, and wall-mounted text placards while simultaneously processing ambient narration. This visual-visual channel competition increases extraneous cognitive load, impeding working memory allocation for knowledge schema construction [8][9][11]. Social Agency Theory: Social Agency Theory posits that social cues—such as natural eye gaze, responsive head movement, and synchronized vocal articulation—prime learners to process interaction as a quasi-social human dialogue rather than mere information processing [10]. This activation of social schemas promotes active cognitive sense-making and deeper semantic processing, leading to more resilient long-term memory encoding [6][10].

1.3 Character Embodiment, Historical Authenticity, and the Uncanny Valley

A central challenge in deploying human representations in VR is Mori's Uncanny Valley phenomenon [12]. As an artificial character approaches near-photorealistic human appearance, subtle imperfections in facial micro-expressions, ocular saccades, skin translucency, and viscoelastic lip movement trigger feelings of eeriness and cognitive dissonance [12]. Polygonal 3D Reconstructions: Creating high-fidelity 3D historical avatars requires photogrammetric facial scanning, high-resolution subsurface scattering shaders, manual blendshape sculpting, and motion-capture pipelines [13][14]. In non-commercial or regional cultural heritage projects, resource limitations often result in rigid, robotic 3D avatars that exacerbate the uncanny valley, drawing negative attention and increasing cognitive friction [7][12]. 2D Archival Authenticity vs. 3D Volume: Archival portraits and photographs carry inherent documentary value and historical authority. Virtual heritage implementations indicate that users exhibit high engagement when confronting the authentic likeness of a historical figure rather than an interpretive, stylized 3D artist reconstruction [3][7]. However, displaying planar media within 6-DoF interactive environments requires careful visual anchoring to avoid spatial incongruity [3].

1.4 Real-Time Neural Talking-Head Synthesis and Generative Video Streaming

The emergence of deep generative networks, neural radiance fields, and explicit 3D Gaussian representations has transformed dynamic portrait generation [13][14][15][16].

Table 1. Implementations of Neural Talking-Head.

Framework / Architecture	Core Methodology	Primary Application & Reference
Wav2Lip	GAN-based audio-to-lip synchronization	Accurate phoneme-to-viseme mapping [16]
Free-View Neural Heads	Latent 3D motion coefficients / Flow warping	High-fidelity micro-expressions from single portrait [13][15]
3D Gaussian Eigenmodels	Deformable 3D explicit Gaussians	Real-time dynamic neural head synthesis [14]

Single-Image Audio-Driven Generative Models: Modern generative frameworks extract 3D motion coefficients (expression, head pose, eye blink) from driving audio feeds, applying them to single archival portraits with minimal warping artifacts [13][16].

Real-Time Integration in Modern Game Engines: Historically, neural video generation was confined to offline batch processing. Recent developments have enabled low-latency inference pipelines directly within modern graphical engines, allowing a dynamic 2D AI-generated video stream to be integrated as an interactive surface running at interactive frame rates required for immersive VR comfort [3][13][14].

While real-time neural talking-head pipelines have demonstrated technical efficacy in screen-based applications [13][14][16], empirical research directly assessing their cognitive and pedagogical impact inside immersive 6-DoF VR environments remains scarce.

A dedicated empirical investigation comparing a real-time 2D AI archival avatar against a traditional non-avatar VR learning environment—measuring recall alongside Cognitive Load Theory [8] and Social Agency metrics [10]—addresses this gap in the educational technology and virtual heritage literature.

2 Methodology & Experimental Design

This chapter outlines the technical implementation, participant framework, experimental protocol, evaluation instruments, and statistical testing setup used to evaluate the pedagogical and affective impact of real-time 2D AI-driven avatars in immersive VR against a traditional non-avatar VR learning environment.

2.1 Technical Architecture & Virtual Environment Setup

The virtual heritage learning environment was built in Unreal Engine 5.7 leveraging Lumen dynamic global illumination and Nanite geometry streaming to ensure a high-fidelity, photorealistic museum setting. The application runs natively on standalone Meta Quest 3 HMDs via highbandwidth Wi-Fi 6E streaming. **Experimental Condition** (Group A – Real-Time 2D AI Avatar): A high-resolution 2D historical portrait is dynamically animated inside Unreal Engine 5.7 using an audio-driven neural head reenactment model. The generated synthetic video feed updates a dynamic material instance mapped to a camera-facing billboard quad with perspective correction, eye-gaze tracking, and HRTF spatialized audio via Unreal's MetaSounds subsystem. **Control Condition** (Group B – Traditional Non-Avatar VR Learning): Participants navigate the identical Unreal Engine 5.7 virtual museum hall containing static historical artifacts, high-resolution 2D archival portraits on display walls, and informational museum placards. The identical historical narrative is delivered via non-diegetic voice-over narration (ambient virtual audio guide) without any animated character or conversational agent.

2.2 Participant Demographics & Stratified Assignment

A cohort of $N = 64$ university students were recruited and assigned via double-blind stratified random assignment, controlling for baseline historical knowledge, gender distribution, and prior VR exposure as seen in (Tab.2).

Table 2. Participant demographics.

Demographic / Baseline Variable	**Group A (2D AI Avatar in VR)**	**Group B (Traditional non-Avatar VR)**	**Total Cohort (N=64)**
Sample Size (n)	32	32	64
Age ($M \pm SD$)	21.5 ± 2.2 years	21.7 ± 2.3 years	21.6 ± 2.2 years
Gender Balance	16 Female, 16 Male	16 Female, 16 Male	32 Female, 32 Male
VR Familiarity	43.8% Novice / 56.2% Exp.	40.6% Novice / 59.4% Exp.	42.2% Novice / 57.8% Exp.
Baseline Domain Knowledge	Pre-Test Score: 31.4%	Pre-Test Score: 30.9%	$p = .88$ (No significant diff.)

Inclusion criteria: Normal or corrected-to-normal sensory abilities, no history of severe simulator sickness or vestibular dysfunction, and no prior exposure to the specific cultural heritage topic.

2.3 Experimental Protocol & Timeline

The longitudinal experiment is organized into three distinct temporal phases to track the cognitive, affective, and pedagogical effects of the system over time. During Phase 1 on Day 0, participants first complete a 15-minute pre-intervention questionnaire to document demographic information and establish baseline historical knowledge. Following this, each student undergoes a standardized 5-minute tutorial to familiarize themselves with locomotion and navigation within the Unreal Engine 5.7 virtual environment. The core educational intervention then commences, immersing participants in a 12-minute historical simulation guided by either a real-time 2D AI avatar in Group A or a static virtual museum guide in Group B. Immediately following the virtual experience, participants spend 20 minutes completing the first post-test battery. This initial evaluation combines standardized psychometric scales, the NASA-TLX cognitive load index, and an immediate domain knowledge recall test. Phase 2 takes place seven days later (Day 7) to evaluate short-term memory retention. During this second phase, all students participate in a 15-minute standardized examination designed to evaluate both factual and conceptual recall. Phase 3 is conducted after a three-month interval (Day 90) to measure long-term cognitive consolidation. In this final phase, students complete an unannounced longitudinal knowledge retention test followed by structured qualitative interviews to gather in-depth feedback on their experience.

2.4 Measurement Instruments & Psychometric Scales

To address the core research questions (**RQ1**, **RQ2**, **RQ3**), six standardized instruments evaluate the cognitive, psychological, and usability dimensions:

1. **Longitudinal Knowledge Retention (RQ1):**
   - *Instrument:* 20-item standardized assessment (10 factual multiple-choice questions for rote recall, 10 thematic open-ended questions assessing historical context and causality). Tested at T_0 (immediate), T_1 (1 week), and T_2 (3 months).
2. **Spatial Presence & Environmental Immersion (RQ2):**
   - *Instrument:* Witmer & Singer Presence Questionnaire (PQ v3.0), evaluating Involvement, Sensory Fidelity, and Realism on a 7-point Likert scale.
3. **Social Presence & Co-Presence (RQ2):**
   - *Instrument:* Networked Mind Social Presence Questionnaire, assessing perceived mutual awareness, message comprehension, and emotional contagion between learner and avatar vs. narrator.
4. **Cognitive Load & Task Demand (RQ3):**
   - *Instrument:* NASA Task Load Index (NASA-TLX), comparing Mental Demand, Temporal Demand, Effort, Frustration, and Perceived Performance to detect split-attention or cognitive overload.
5. **Historical Empathy & Affective Connection (RQ3):**
   - *Instrument:* Geneva Emotion Wheel (GEW v3.0) and adapted Interpersonal Reactivity Index (IRI) to evaluate affective engagement and perspective-taking toward historical events.
6. **Visual Comfort & Usability:**
   - *Instrument:* System Usability Scale (SUS) alongside pre/post Simulator Sickness Questionnaire (SSQ) to control for visual fatigue and rendering discomfort.

2.5 Statistical Analysis & ANOVA Formulation

Data analysis is conducted in **R (v4.4)** / **SPSS (v29)** with a significance criterion of $\alpha = .05$.

1. Mixed-Design Factorial ANOVA (Longitudinal Retention)

To test **RQ1**, a 2×3 Split-Plot Mixed ANOVA is formulated:

- **Between-Subjects Factor (2 levels):** Learning Modality (Group A: 2D AI Avatar vs. Group B: Traditional Non-Avatar VR).
- **Within-Subjects Factor (3 levels):** Time Interval (Immediate T_0, 1 Week T_1, 3 Months T_2).

$$\text{Retention Score}_{ijk} = \mu + \alpha_i(\text{Modality}) + \beta_j(\text{Time}) + (\alpha\beta)_{ij} + \pi_{k(i)} + \epsilon_{ijk} \quad (1)$$

- **Diagnostic Checks:** Shapiro-Wilk for normality, Levene's test for homogeneity of variance, and Mauchly's test for sphericity (with Greenhouse-Geisser correction if $\epsilon <$ .75).
- **Effect Sizes:** Partial eta-squared (η_p^2) computed for main effects and the interaction effect (Modality $\times$ Time).

2. Multivariate Analysis of Variance (MANOVA)

To evaluate the psychological and psychometric variables (**RQ2**, **RQ3**):

- **Independent Variable:** Learning Modality (Group A vs. Group B).
- **Dependent Vector:** [Spatial Presence (PQ), Social Presence, Total Cognitive Load (NASA-TLX), Historical Empathy (IRI)]
- **Test Statistic:** Pillai's Trace used for multivariate robustness, followed by univariate ANOVAs with Bonferroni-adjusted post-hoc pairwise comparisons.

3 Results

This chapter presents the quantitative and qualitative findings evaluating the pedagogical efficacy, cognitive workload, and psychometric dimensions of real-time 2D AI avatars (Group A) compared to traditional non-avatar VR learning (Group B) within Unreal Engine 5.7.

3.1 Knowledge Retention & Longitudinal Memorability (RQ1)

To address **RQ1**, a 2×3 mixed-design ANOVA was conducted with **Learning Modality** (Group A vs. Group B) as the between-subjects factor and **Time** (T_0: Immediate, T_1: 1 Week, T_2: 3 Months) as the within-subjects repeated measure.

The dependent variable is the percentage score on the standardized 20-item cultural heritage assessment (max 100%) as seen in (Tab.3).

Table 3. Measurement of memorability in time frames.

Condition	T0: Immediate (M±SD)	T1: 1 Week (M±SD)	T2: 3 Months (M±SD)	Absolute Retention Drop (T0→T2)
Group A (2D AI Avatar, $n = 32$)	$84.62\% \pm 6.31$	$79.18\% \pm 7.12$	$68.45\% \pm 8.04$	-16.17%
Group B (Traditional VR, $n = 32$)	$81.35\% \pm 7.45$	$69.80\% \pm 8.23$	$51.22\% \pm 9.15$	-30.13%
Mean Difference (Δ)	$+3.27\%$ ($p = .062$)	$+9.38\%$ ($p < .001$)	$+17.23\%$ ($p < .001$)	—

Mauchly's test indicated that the assumption of sphericity had been violated for the within-subjects factor Time, $\chi^2(2) = 8.14, p = .017$; therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates ($\epsilon = .891$).

Table 4. Mixed ANOVA Summary Table.

Source of Variation	SS	df	MS
Between-Subjects			
Modality (Group A vs. B)	6281.42	1	6281.42
Error (Between)	8004.18	62	129.10
Within-Subjects			
Time (T_0, T_1, T_2)	18245.31	1.78	10250.17
Modality × Time	1892.64	1.78	1063.28
Error (Time)	4011.55	110.36	36.35

Main Effect of Modality: Statistically significant, $F(1,62) = 48.65, p < .001, \eta_p^2 = .440$.
Main Effect of Time: Statistically significant, $F(1.78,110.36) = 312.44, p < .001, \eta_p^2 = .834$.
Interaction Effect (Modality × Time**):** Highly significant, $F(1.78,110.36) = 32.41$, $p < .001, \eta_p^2 = .343$.

Pairwise comparisons revealed that while baseline recall immediately post-intervention (T_0) was comparable ($p = .062$), Group B experienced a rapid memory decay curve of -30.13% over 3 months, whereas Group A exhibited a significantly buffered memory decay of only -16.17%.

3.2 Multivariate Analysis of Psychometric Outcomes (RQ2 & RQ3)

A one-way MANOVA was performed to test differences across the multi-dimensional psychometric profile between Group A and Group B. Using Pillai's Trace, the combined dependent vector was significantly affected by the learning modality, $V = .682, F(5,58) = 24.89, p < .001, \eta_p^2 = .682$.

Table 5. Instruments and values.

Metric / Instrument	Group A (M±SD)	Group B (M±SD)	Univariate F(1,62)	p-value	ηp2
Spatial Presence (PQ 1–7)	5.82 ± 0.54	5.64 ± 0.61	1.58	$.213$	$.025$
Social Presence (1–7)	5.94 ± 0.62	2.31 ± 0.74	451.20	$< .001$	$.879$
Historical Empathy (IRI 1–5)	4.28 ± 0.49	3.12 ± 0.58	74.82	$< .001$	$.547$
Cognitive Load (NASA-TLX 0–100)	38.45 ± 8.20	47.60 ± 9.40	17.38	$< .001$	$.219$
System Usability (SUS 0–100)	82.50 ± 6.80	80.10 ± 7.10	1.89	$.174$	$.030$

Presence Dimensions (RQ2): Spatial presence remained statistically equivalent across conditions ($p = .213$), demonstrating that Unreal Engine 5.7's photorealistic lighting and spatial rendering provided equal environmental immersion. However, **Social Presence** was substantially higher in Group A ($\eta_p^2 = .879$), driven by eye contact and dynamic micro-expressions.

Cognitive Load & Empathy (RQ3): Group A demonstrated significantly lower cognitive load ($M = 38.45$) compared to Group B ($M = 47.60$), confirming that personified storytelling mitigated split-attention effects caused by reading placards while listening to audio. Historical empathy scores were significantly higher in the avatar condition ($F = 74.82, p < .001$).

3.3 Qualitative Themes & Perceptual Observations

Post-hoc thematic analysis of semi-structured exit interviews revealed three prominent themes:

- **Authenticity vs. 3D "Game-Like" Distortion:** 84.4% of Group A participants explicitly praised the "archival realism" of the 2D animated portrait, stating that a traditional 3D reconstructed face often feels "synthetic or like a video game character," whereas the live archival face maintained historical solemnity.
- **Spatial Alignment & 2.5D Billboarding:** 9 participants (28.1%) in Group A noted subtle flatness when viewing the avatar from extreme oblique angles ($> 60°$). However, dynamic billboard alignment and localized eye contact minimized this limitation during typical frontal viewing distances (1.5 m $-$ 2.5 m).

- **Narrative Coherence:** Group B participants reported feeling divided between locating referenced artifacts and reading wall placards, validating the higher Mental Demand sub-scores observed on the NASA-TLX.

4 Discussion

The empirical findings from Chapter 4 demonstrate that integrating a real-time, 2D AI-driven archival avatar inside an immersive Unreal Engine 5.7 environment yields a statistically significant retention advantage over a 3-month period (68.45% vs. 51.22%, $p < .001$), while simultaneously reducing subjective cognitive load (38.45 vs. 47.60, $p < .001$). This discussion contextualizes these outcomes within Social Agency Theory and Cognitive Load Theory (CLT), examining how affective resonance and multimedia design principles govern long-term knowledge consolidation in virtual heritage.

4.1 Social Agency Theory & The Mechanism of Longitudinal Retention

The marked divergence in longitudinal recall decay (-16.17% for Group A vs. -30.13% for Group B) provides strong empirical backing for Mayer's Social Agency Theory.

- **Social Cues as Cognitive Catalysts:** According to social agency models, pedagogical agents presenting authentic social cues—such as dynamic eye contact, natural head movements, and synchronized lip movements—prime learners to perceive the interaction as a human-to-human conversation rather than passive information retrieval. In Group A, the 2D neural talking-head avatar acted as an embodied historical witness.
- **Deep Semantic Encoding:** This perceived social partnership activates social schema, motivating learners to exert active cognitive effort in making sense of the dialogue. Rather than storing fragmented historical facts, participants in the avatar condition contextualized the narrative within an episodic, social encounter.
- **Emotional Resonance & The Memory Buffer:** The high historical empathy scores ($M = 4.28$ vs. 3.12, $\eta_p^2 = .547$) measured via the IRI indicate that seeing the authentic, archival face emote created an emotional bridge. Emotionally resonant episodic memories are known to undergo stronger hippocampal consolidation, explaining why the recall advantage was modest immediately post-intervention ($+3.27\%$, $p = .062$) but expanded drastically after 3 months ($+17.23\%$, $p < .001$).

4.2 Cognitive Load Theory: Split-Attention vs. Dual-Channel Scaffolding

The NASA-TLX results demonstrate a pronounced reduction in extraneous cognitive load for the avatar cohort. Sweller's Cognitive Load Theory and Mayer's Cognitive Theory of Multimedia Learning (CTML) explain this disparity across two key dimensions:

Mitigating the Spatial Split-Attention Effect: In the traditional condition (Group B), learners were forced to distribute visual attention across 3D artifacts, static portraits, and wall-mounted text placards while processing ambient voiceover narration. This visual-visual competition overloads the visual working memory channel. In contrast, Group A anchored visual attention directly onto the speaking figure, using gaze cues to direct the learner's focus toward surrounding artifacts.

Modality & Redundancy Principles: Group A received narration synchronized with facial visual cues rather than competing printed text. This offloaded informational weight cleanly across auditory and visual channels, maximizing working memory capacity for germane load—the mental processing dedicated to schema construction.

4.3 The Visual Trade-Off: Archival Authenticity vs. 3D Spatial Parallax

A core objective of this study was assessing whether a 2D neural projection compromises virtual presence compared to full 3D reconstructions.

Bypassing the 3D Uncanny Valley: Traditional 3D historical character reconstructions frequently suffer from the uncanny valley effect, where imperfect skin shaders or unnatural facial rigging trigger eerie perceptions that distract learners. By utilizing the original 2D archival photograph with neural reenactment, the avatar preserved absolute historical authenticity, texture fidelity, and solemnity.

Parallax Limitations vs. Perceptual Prioritization: The absence of full depth volume in a 2.5D billboard was detectable at extreme angles ($> 60°$). However, because human educational interactions naturally occur face-to-face at frontal offsets ($0°–30°$), the human perceptual system prioritized high-fidelity facial dynamics and micro-expressions over stereoscopic depth. Consequently, spatial presence was preserved ($p = .213$), while social presence increased significantly ($p < .001$).

4.4 Pedagogical & Museum Deployment Implications

Scalability for Cultural Institutions: 3D polygonal character modeling, rigging, and micro-expression sculpting are financially prohibitive for most regional museums and schools. The 2D AI reenactment pipeline requires only a single high-resolution archival photograph and an audio script, reducing digitization overhead while outperforming static exhibits.

Ethical Considerations & Synthetic Fidelity: Real-time deepfake technologies in heritage education require rigorous provenance disclosures. To prevent historical fabrication, generative frameworks must strictly adhere to verified historical source material, ensuring pedagogical integrity.

Conclusion

This study investigated the pedagogical, cognitive, and affective impact of deploying real-time, 2D AI-generated historical avatars within an immersive virtual reality learning environment. By comparing a dynamic 2D neural talking-head avatar against a traditional non-avatar VR museum layout inside Unreal Engine 5.7, the research provides empirical and theoretical foundations for the integration of generative synthetic media into cultural heritage education.

Longitudinal Knowledge Retention: The empirical findings demonstrate that personified historical storytelling via 2D AI avatars significantly curbs memory decay over time. While initial recall was comparable across conditions, students exposed to the 2D AI avatar retained 68.45% of domain knowledge after 3 months compared to 51.22% in the traditional VR condition ($\Delta = +17.23\%$, $p < .001$).

Validation of Social Agency & Cognitive Load Theories: The study validates Mayer's Social Agency Theory and Sweller's Cognitive Load Theory in real-time immersive generative settings. Dynamic facial reenactment, natural gaze cues, and synchronized diegetic speech acted as social catalysts, converting passive knowledge consumption into an episodic social interaction while eliminating the spatial split-attention effect caused by static wall placards.

Archival Authenticity Overcoming the Uncanny Valley: The results confirm that high-fidelity neural animation applied directly to authentic archival portraits preserves historical solemnity and avoids the uncanny valley distortions often associated with budget-constrained 3D character rigging. Spatial presence remained unaffected, while social presence and historical empathy increased substantially.

Acknowledgement

„Funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I03-03-V04-00495."

References

[1] Champion, E. (2023). *Playing with the past: Into the future* (2nd ed., Vol. 1). Springer. https://doi.org/10.1007/978-3-031-10932-4

[2] Bekele, M. K., Champion, E., McMeekin, D. A., & Rahaman, H. (2021). The influence of collaborative and multi-modal mixed reality: Cultural learning in virtual heritage. *Multimodal Technologies and Interaction*, 5(12), Article 79. https://doi.org/10.3390/mti5120079

[3] Vital, R., Sylaiou, S., Koukopoulos, D., Koukoulis, K., Dafiotis, P., & Fidas, C. (2023). Comparison of extended reality platforms and tools for viewing and exhibiting art. *Digital Applications in Archaeology and Cultural Heritage*, 31, Article e00298. https://doi.org/10.1016/j.daach.2023.e00298

[4] Shabani, A., Skamantzari, M., Tapinaki, S., Georgopoulos, A., Plevris, V., & Kioumarsi, M. (2022). 3D simulation models for developing digital twins of heritage structures: Challenges and strategies. *Procedia Structural Integrity*, 37, 314–320. https://doi.org/10.1016/j.prostr.2022.01.090

[5] Sá, M. J., & Serpa, S. (2023). Metaverse as a learning environment: Some considerations. *Sustainability*, 15(3), Article 2186. https://doi.org/10.3390/su15032186

[6] Johnson, W. L., & Lester, J. C. (2018). Pedagogical agents: Back to the future. *AI Magazine*, 39(2), 33–44. https://doi.org/10.1609/aimag.v39i2.2793

[7] Sylaiou, S., & Fidas, C. (2022). Virtual humans in museums and cultural heritage sites. *Applied Sciences*, 12(19), Article 9913. https://doi.org/10.3390/app12199913

[8] Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, 31(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

[9] Mayer, R. E. (2002). Multimedia learning. *Psychology of Learning and Motivation*, 41, 85–139. https://doi.org/10.1016/S0079-7421(02)80005-6

[10] Mayer, R. E. (2014). Incorporating motivation into multimedia learning. *Learning and Instruction*, 29, 171–173. https://doi.org/10.1016/j.learninstruc.2013.04.003

[11] Ayres, P., & Sweller, J. (2005). The split-attention principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (pp. 135–146). Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.009

[12] Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [From the field]. *IEEE Robotics & Automation Magazine*, 19(2), 98–100. https://doi.org/10.1109/MRA.2012.2192811

[13] Gao, Y., Zhou, Y., Wang, J., Li, X., Ming, X., & Lu, Y. (2023). High-fidelity and freely controllable talking head video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10214–10223). https://doi.org/10.48550/arXiv.2304.10168

[14] Zielonka, W., Bolkart, T., Beeler, T., & Thies, J. (2025). Gaussian eigen models for human heads. *arXiv preprint*, arXiv:2407.04545. https://doi.org/10.48550/arXiv.2407.04545

[15] Tewari, A., Fried, O., Thies, J., Zollhöfer, M., Garrido, P., Schaefer, B., Hadap, S., Theobalt, C., & Seitz, S. (2020). State of the art on neural rendering. *Computer Graphics Forum*, 39(2), 701–727. https://doi.org/10.1111/cgf.14022

[16] Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2020). A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia* (pp. 484–492). https://doi.org/10.1145/3394171.3413532

Authors

RNDr. Ján Lacko, PhD.
Faculty of Informatics, Pan-european university, Bratislava, Slovakia
jan.lacko@paneurouni.com
His research interests include the digitisation of objects from the fields of cultural heritage, healthcare, industry and urban planning, and their visualisation using various techniques, including virtual and augmented reality.

Mgr. Michal Drobný
Faculty of Informatics, Pan-european university, Bratislava, Slovakia
michal.drobny@paneurouni.com
His research interests are in the field of discrete event systems and in the field of web technologies using artificial intelligence.