informatics

MDPI

Article

Voice, Text, or Embodied AI Avatar? Effects of Generative AI Interface Modalities in VR Museums

Pakinee Ariya [1], Perasuk Worragin [1], Songpon Khanchai [2], Darin Poollapalin [2] and Phichete Julrode [2,*]

[1] College of Arts, Media and Technology, Chiang Mai University, Chiang Mai 50200, Thailand; pakinee.a@cmu.ac.th (P.A.); perasuk_w@cmu.ac.th (P.W.)
[2] Department of Library and Information Science, Faculty of Humanities, Chiang Mai University, Chiang Mai 50200, Thailand
* Correspondence: phichete.j@cmu.ac.th

Abstract

Virtual museums delivered through immersive virtual reality (VR) function as information environments where users access interpretive content while navigating spatially. With the integration of generative artificial intelligence (AI), conversational assistants can dynamically mediate information interaction; however, evidence remains limited regarding how different AI interface representations affect user experience. This study compares three generative AI interface modalities in a VR virtual museum: voice only, voice with synchronized text, and voice with an embodied AI avatar. A controlled experiment with 75 participants examined their effects on user engagement, perceived information quality, and subjective cognitive workload while holding informational content constant. The results indicate that the voice-and-text modality produced the highest perceived information quality, whereas the embodied AI avatar modality yielded the highest user engagement. No significant differences were observed in cognitive workload across modalities. These findings suggest that AI interface modalities play complementary roles in VR-based information interaction and provide design guidance for selecting appropriate AI representations in immersive information systems.

Keywords: information interaction; human–computer interaction; embodied conversational agent; virtual museums; generative AI assistant; digital cultural heritage

1. Introduction

Virtual museums and immersive virtual reality (VR) experiences have rapidly matured into practical information environments for cultural heritage interpretation, enabling visitors to explore digitized collections, navigate curated spaces, and access interpretive narratives beyond the constraints of physical exhibitions. Contemporary virtual museum implementations emphasize that user outcomes—such as understanding, satisfaction, and intention to continue exploring—are shaped not only by content authenticity but also by how interpretive information is delivered during the visit, particularly when users must alternate between navigation and information access. Recent work demonstrates that VR museum systems can be designed and evaluated as end-to-end experiences where interaction design and interpretation strategy jointly influence perceived educational value and overall experience [1,2].

In parallel, generative AI is increasingly positioned as an interaction layer capable of delivering responsive, context-aware explanations that go beyond static labels. A recent

Check for updates

Academic Editor: Yang Shi

Received: 8 January 2026
Revised: 14 February 2026
Accepted: 5 March 2026
Published: 11 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

study survey frames "Generative VR" as a shift from pre-scripted experiences toward AI-driven content generation and adaptive interaction, highlighting key challenges that align directly with museum contexts (e.g., cognitive load management, real-time performance constraints, and interaction design) [3]. Within cultural heritage VR, recent applied studies also continue to refine system setups and evaluation practices for museum-grade experiences, underscoring the importance of usability and interaction mechanisms in shaping visitor interpretation [1,4].

Despite the growing interest in AI-supported immersive heritage experiences, there remains limited controlled evidence addressing a core human–computer interaction and information systems design question: when informational content is held constant, how should a generative AI guide be represented to best support museum information interaction in VR? Recent cultural heritage research has examined the use of virtual humans and intelligent agents in museum settings [5], the effects of social presence associated with augmented or embodied avatars in immersive environments [6], and broader trends in embodied conversational agents across extended reality systems [7]. At the same time, museum-oriented VR studies have begun to explore generative AI assistants and gamified interaction strategies, reporting measurable impacts on engagement and cultural-heritage learning outcomes [8,9]. However, these studies rarely isolate the effects of AI interface representation itself, making it difficult to derive evidence-based design guidance for virtual museum applications.

To address this gap, the present study evaluates three generative AI interface modalities in a VR virtual museum—Voice-only, Voice + Text (transcript/subtitle), and Voice + Embodied AI Avatar—and examines their comparative effects on perceived information quality, cognitive workload, and user engagement. The research questions guiding this study are as follows:

RQ1: How do the three modalities differ in users' engagement with the VR museum experience?

RQ2: How do Voice-only, Voice + Text, and Voice + Embodied AI Avatar modalities differ in users' perceived information quality during VR museum information interaction?

RQ3: How do the three modalities differ in users' subjective cognitive workload during museum tasks?

2. Related Work

2.1. Virtual Museums and Information Interaction in Immersive Environments

Virtual museums delivered through immersive VR are increasingly conceptualized as information environments in which visitors must coordinate spatial exploration with situated access to interpretive content. Recent implementations and evaluations show that VR museum experiences can reinforce interpretation and improve educational value when interaction design supports purposeful navigation, attention management, and timely information access during exploration [10,11]. Empirical evidence from online VR exhibitions further suggests that interaction-related factors—particularly interactivity, immersion, and presence—are associated with stronger perceived usefulness/enjoyment and higher willingness to use (and even intention to visit on-site), emphasizing that "information interaction quality" is a central driver of downstream acceptance and experience outcomes [12].

At a finer design granularity, comparative studies indicate that the level of immersion and interaction fidelity (e.g., wearable VR versus mobile VR) meaningfully affects visitor immersion and experience consequences, implying that interaction constraints can shape how visitors seek, process, and retain museum information in VR settings [13]. Complementary models of VR exhibition experience also highlight relationships among presence, interaction, immersion, and satisfaction, reinforcing the need to treat interaction as a core

mechanism rather than a secondary usability detail [14]. More recent virtual museum research has begun to examine how adding interactive features changes audience behavior in replica-based environments, providing behavioral evidence that interaction design can reshape engagement patterns beyond self-report measures [15]. In parallel, virtual museums are increasingly integrating social and agent-like elements—such as virtual humans and metaverse-style social interaction layers—suggesting a shift toward more mediated information experiences where the interface itself becomes part of interpretation [16,17].

2.2. Generative AI and Intelligent Agents in VR and XR Systems

Recent advances in generative AI—particularly large language models (LLMs)—have accelerated the adoption of conversational assistants and intelligent agents in immersive VR/XR systems. Compared with scripted guides, LLM-driven agents can support open-ended queries, generate context-aware explanations, and adapt dialogue in real time, positioning generative AI as an interaction layer for information mediation in spatial experiences [18]. Recent scholarship has begun to operationalize these capabilities in immersive simulations and training contexts, demonstrating how generative AI can be embedded into VR scenarios to provide interactive coaching, scenario-based practice, and adaptive feedback while raising practical design constraints around latency, reliability, and user experience consistency [19,20].

In parallel, research on embodied conversational agents (ECAs) and virtual humans in VR/XR indicates that the form of the agent (e.g., disembodied voice vs. embodied avatar) can shape engagement, social presence, perceived naturalness, and workload—making interface representation a design variable rather than a purely aesthetic choice [21,22]. Systematic reviews further consolidate evidence that embodiment, multimodality, and backend AI integration jointly affect user outcomes across XR use cases, including information-assistant roles relevant to museum and cultural-heritage applications [23]. In cultural heritage VR specifically, recent work has surveyed and synthesized how virtual humans are used in museums [24] and has also demonstrated generative AI-based virtual assistants embedded within VR museums to enhance engagement with heritage content [25].

2.3. Multimodal and Embodied Interfaces for Information Delivery

Recent VR–HCI work increasingly treats multimodal interface design—combining speech with text, visual cues, or embodiment—as a key determinant of usability, cognitive workload, and interpretability in immersive systems. Large-scope reviews emphasize that effective VR information delivery requires careful coordination of multiple sensory channels, particularly when users must navigate 3D spaces while accessing interpretive content [26]. Empirical evidence further suggests that multimodal interaction can improve efficiency and perceived naturalness; MEinVR, for example, demonstrates that controller-plus-voice interaction supports more intuitive immersive exploration than single-channel interaction alone [27]. At the agent level, controlled studies indicate that how an agent is represented during speech can shape user cognitive responses. Chang et al. compared voice-only agents with embodied and gestural variants and reported differences in attentional demand indicators around speech onset, implying that embodiment may ease attentional transitions while additional gestures can increase demand depending on timing and salience [28].

More recent work has examined embodied conversational agents (ECAs) and LLM-powered avatars in VR, focusing on perceived latency, immersion, and social presence. Elfleet and Chollet show that multimodal feedback strategies can reduce perceived latency and enhance immersion for LLM-driven ECAs, highlighting the importance of interface-level feedback when responses are generated in real time [29]. Jolibois et al. similarly

demonstrate that response delay can significantly affect user impressions of an emotional ECA, even when informational content is comparable, underscoring the role of temporal interaction properties in perceived quality [30]. At a synthesis level, Yang et al. review ECAs in extended reality and identify embodiment and multimodal behaviors as central levers shaping user experience across XR applications, including museum guide scenarios [31].

2.4. Interface Modality, Cognitive Demand, and Engagement in Immersive VR

Interface modality is a critical design factor that shapes how users allocate cognitive resources and experience engagement in immersive VR systems. A substantial body of VR and HCI research has shown that immersive environments impose distinctive cognitive demands stemming from spatial navigation, sensory richness, and continuous interaction with virtual elements. A large-scale review of immersive VR applications highlights that interface design choices—rather than content alone—play a decisive role in determining usability, learning effectiveness, and perceived workload in VR-based information systems [32]. Consequently, interface modality should be understood as an integral component of interaction design that mediates how users process information and sustain attention during immersive experiences.

Earlier foundational studies in multimedia and virtual reality design also emphasized multisensory coordination and cognitive load management as central challenges in immersive systems [33,34]. These works demonstrated that modality configuration alone—independent of AI-generated content—can significantly influence attention, workload, and learning outcomes. Building on this foundation, the present study introduces generative AI as a dynamic information layer while experimentally isolating interface representation, thereby clarifying the added value of AI-driven content generation in VR contexts.

Prior empirical studies further demonstrate that different interface modalities introduce specific trade-offs between cognitive demand and engagement. Multisensory or multimodal interfaces can support performance and presence by distributing information across channels, yet they may also increase perceptual load and mental effort when visual complexity or sensory competition becomes excessive [35]. Similarly, avatar-based or embodied interfaces have been shown to influence users' cognitive task performance in VR, suggesting that embodiment can enhance social presence and engagement while simultaneously affecting attentional demands [36]. From an engagement perspective, constructs such as focused attention, perceived usability, and intrinsic reward are closely linked to how interaction modalities are experienced within immersive systems [37]. Together, these findings underscore the need for controlled comparisons of AI interface modalities in VR to better understand how alternative representations balance cognitive demand and user engagement in information-intensive contexts such as virtual museums.

3. Generative AI Interface Design in a Virtual Museum

3.1. Virtual Museum Context

The virtual museum employed in this study was adapted from the Wieng Yong House Museum in Thailand, a cultural heritage site dedicated to the preservation and interpretation of local textile and craft traditions. The VR environment digitally represents the architectural layout, exhibition spaces, and contextual artifacts of the Wieng Yong House, allowing visitors to explore the museum in an immersive and spatially coherent manner. This virtual museum system was previously developed and evaluated by the authors as an AI-assisted cultural heritage learning environment [8]. In the present study, the virtual museum's spatial layout, exhibit arrangement, and narrative structure were retained without modification to ensure continuity with prior implementations. The generative AI assistant was embedded as an informational guide within this established

museum context, enabling users to request explanations related to exhibits, materials, and cultural practices encountered during exploration. Grounding the experiment in the Wieng Yong House Museum ensures ecological validity and situates the investigation within a realistic digital heritage setting rather than an abstract or synthetic environment.

3.2. System Architecture Design

The system architecture, illustrated in Figure 1, follows the design framework established in the authors' prior work on generative AI-driven virtual museum assistants. The virtual museum application was developed using Unity 2022 LTS and deployed on Meta Quest 3 head-mounted displays. User interactions within the VR environment capture voice input at exhibit interaction points, initiating the AI processing pipeline. As shown in Figure 1, spoken input is converted to text via a speech-to-text API. Relevant exhibit descriptions are retrieved from the museum database and programmatically inserted into a structured prompt template before being transmitted to the external large language model (OpenAI ChatGPT, GPT-4o model accessed via the OpenAI API). This retrieval-based contextual grounding constrains the model's response to exhibit-specific knowledge without modifying or fine-tuning the underlying LLM parameters. The generated responses are returned to the VR application and presented as audio-only output, synchronized text, or speech delivered through an embodied AI avatar, depending on the interface modality. The AI generation logic and informational scope are identical across all conditions, ensuring that observed differences in user experience are attributable solely to interface representation. All prompts follow an identical template structure across conditions to ensure consistent semantic grounding from the museum database.

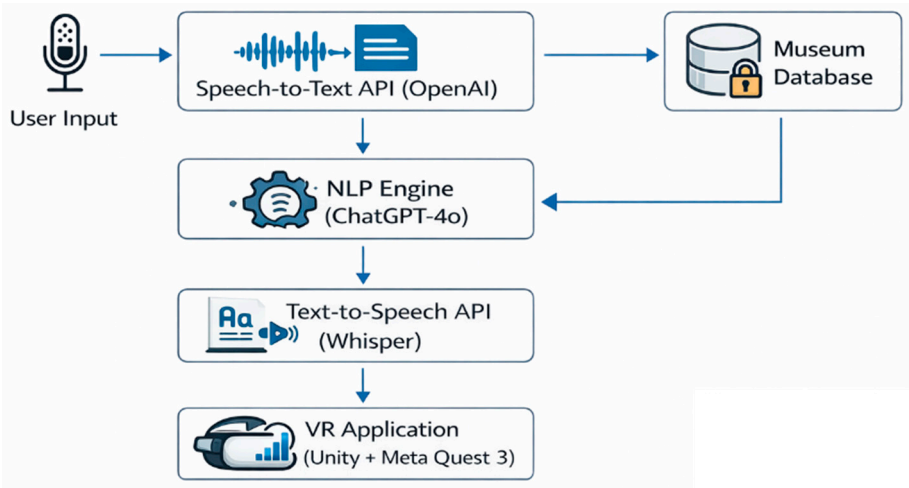


Figure 1. System architecture of the generative AI-driven virtual museum, showing the voice-based interaction pipeline from user input to AI response delivery.

From a scalability perspective, the architecture adopts a modular API-based design in which speech recognition, database retrieval, and LLM inference operate as independent service components. This separation enables horizontal scaling through parallel API calls and database access without altering the core VR application layer, while the museum knowledge base can be expanded by adding structured exhibit entries without reconfiguring the language model. During development, several technical challenges were encountered, including managing network-dependent latency associated with cloud-based LLM inference, synchronizing speech output with subtitle rendering and avatar lip movement, and maintaining stable speech-to-text performance within immersive environments. These challenges were mitigated through buffered response handling, synchronized playback control within Unity, and controlled environmental audio conditions during testing.

3.3. Interface Modalities

Three generative AI interface modalities were implemented within the same system architecture: voice only, voice and text, and voice plus embodied AI avatar. All modalities delivered identical AI-generated informational content, differing only in how the information was presented to users, thereby enabling a controlled examination of interface effects on information interaction in a VR virtual museum. In the Voice-only condition, AI responses were delivered exclusively through spatialized audio without any persistent visual representation of the assistant, allowing users to focus entirely on the exhibition environment while relying on auditory information. In the Voice + Text condition, spoken responses were supplemented with synchronized subtitles displayed in a fixed panel positioned below the primary field of view. The subtitles remained visible for the duration of each response and were formatted using standardized font size and contrast settings to ensure legibility without obstructing key visual elements of the exhibition space.

In the Voice + Avatar condition, a humanoid virtual guide appeared within the exhibition environment at conversational distance and delivered AI-generated responses with synchronized lip movement. The avatar remained stationary and did not include additional gestural animation, thereby preserving embodied presence while minimizing unintended motion-based distractions. The avatar did not introduce any additional social, expressive, or contextual cues beyond delivering the scripted AI-generated speech. No pointing, object-referencing gestures, or non-verbal emphasis behaviors were implemented. Across all three conditions, interface layout, response timing, and AI-generated semantic output were held constant. All modalities employed an identical prompt template structure, predefined question sequence, and exhibit-specific knowledge retrieved from the same museum database. No modality-specific modification, summarization, or paraphrasing was applied after generation. While responses were generated live via API calls, semantic consistency was maintained through structured prompting and controlled contextual grounding rather than fixed pre-generated outputs. Only the representational layer—audio-only, audio + text, or audio + embodied avatar—varied between modalities, ensuring that observed differences could be attributed solely to interface representation rather than informational content. As illustrated in Figures 2 and 3, these controlled variations support a direct comparison of modality effects on engagement, perceived information quality, and cognitive workload.



Figure 2. Comparison of AI interface modalities: (**A**) Voice + Text, displaying synchronized subtitles; (**B**) Voice + Avatar, featuring a humanoid AI guide delivering identical spoken responses within the exhibition space.



Figure 3. Voice-only interface modality, in which AI-generated responses are delivered exclusively through spatialized audio without subtitle display or embodied representation. No persistent visual AI elements are present in this condition.

3.4. AI Configuration and Runtime Control

The system employed GPT-4o via live cloud-based API calls integrated with speech recognition and text-to-speech services. Exhibit-specific contextual grounding was achieved through deterministic retrieval from a structured MySQL database triggered by exhibit interaction IDs within the VR environment. For each query, a single structured exhibit description block (approximately 150–250 words) was retrieved and programmatically inserted into a standardized prompt template prior to submission to the language model. Prompt structure, predefined question sequence, and retrieval rules were identical across all interactions. Although responses were generated live, semantic consistency was maintained through structured prompting and controlled contextual grounding rather than fixed pre-generated outputs. All prompts and responses were automatically logged for post hoc verification and runtime analysis.

Interaction latency for the present architecture was monitored through system log data and remained consistent with prior evaluation of the same system pipeline [8], which reported an average waiting time before AI response initiation (time-to-first-audio) of 6.34 s, an average response generation and delivery time of 12.39 s, and a system-level end-to-end response completion time of 21.97 s. In the present study, system log inspection confirmed that latency remained within this operational range during experimental sessions. These metrics represent processing and playback latency only; behavioral interaction duration per query is analyzed separately in Section 5.4. No automated retry mechanism was implemented. In cases of delayed API response, participants waited for playback to begin, and no generation failures requiring task termination were recorded.

4. Research Method

4.1. Research Design

This study employed an experimental design to examine the effects of generative AI interface modalities on information interaction in a VR virtual museum (Figure 4). Participants were randomly assigned to three groups: voice only, voice and text (transcript/subtitle), and voice plus avatar (embodied AI avatar). The informational content provided by the AI assistant was held constant across all conditions to isolate the effects of interface representation. The primary dependent variables were perceived information quality, cognitive workload, and user engagement, complemented by behavioral interaction data recorded by the system.

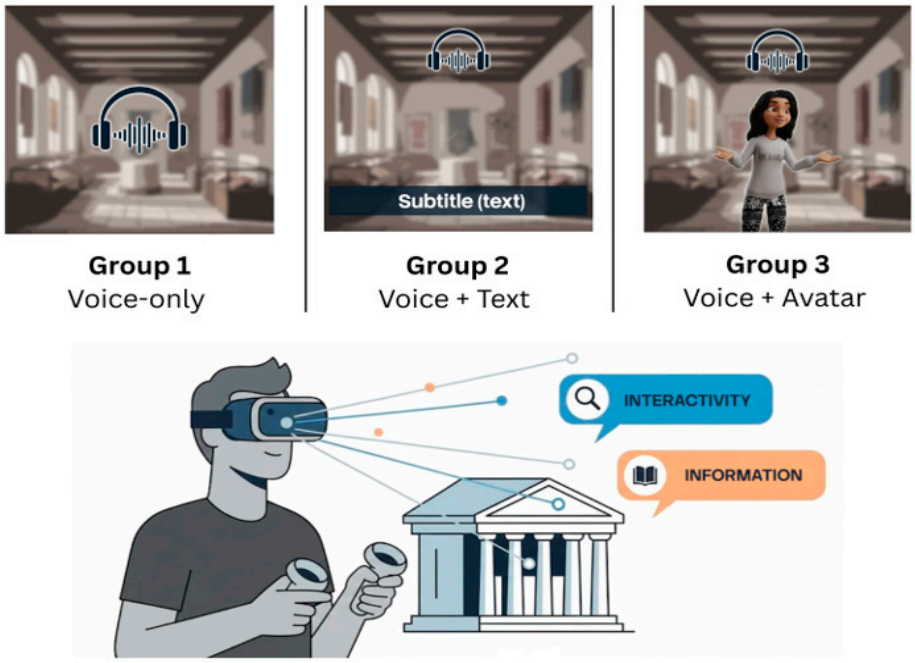


Figure 4. Overview of the experimental design comparing three generative AI interface modalities in a virtual museum.

4.2. Participants

A total of 75 participants were recruited and randomly assigned to one of three experimental groups (n = 25 per group) shown in Table 1. All participants had prior experience using virtual reality systems, including basic interaction skills such as controller use and navigation, and reported no history of motion sickness or discomfort when using VR environments. To control for prior knowledge effects, participants were required to have no prior familiarity with the textile museum featured in this study, including its exhibits or historical background. Random assignment ensured comparable demographic distribution across experimental conditions and minimized potential selection bias.

Table 1. Demographic characteristics of participants across the three experimental groups.

Gender	Voice-Only (n = 25)	Voice + Text (n = 25)	Voice + Avatar (n = 25)	Total (n = 75)
Male	13	12	14	39
Female	12	13	11	36

4.3. Instruments

4.3.1. User Engagement Scale

User engagement with the VR museum experience was measured using the short form of the User Engagement Scale (UES–SF) [38]. The instrument consists of four dimensions: Focused Attention, Perceived Usability (reverse-coded), Aesthetic Appeal, and Reward. All items were rated on a 5-point Likert scale ranging from 1 (Strongly Disagree) to 5 (Strongly Agree). The UES–SF has been widely adopted in interactive system and HCI research to capture users' cognitive, affective, and aesthetic engagement during digital experiences. The short form was selected to reduce participant burden in immersive VR experiments, where lengthy post-exposure questionnaires may affect response quality. The UES–SF has been validated as a reliable and psychometrically robust alternative to the full-length scale, making it suitable for controlled VR and HCI research contexts [38]. The full questionnaire is presented in Appendix A.1 Table A1.

4.3.2. Perceived Information Quality

Perceived information quality of the generative AI assistant was assessed using a self-report questionnaire measuring users' perceptions of information accuracy, clarity, relevance, and trustworthiness. All items were rated on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). To ensure content validity, the questionnaire items were reviewed by three domain experts in virtual museum design and human–computer interaction. Item–Objective Congruence (IOC) analysis yielded a value of 0.67. This measure was used to evaluate how effectively each AI interface modality supported users' understanding and confidence in the information provided during VR museum exploration. The full questionnaire is presented in Appendix A.2 Table A2.

4.3.3. NASA Task Load Index

Subjective cognitive workload was measured using a short version of the NASA Task Load Index (NASA-TLX) [39], adapted from the original instrument [40]. The questionnaire captures participants' perceived workload during museum tasks across key dimensions, including mental demand, effort, and frustration. Responses were recorded using standardized rating scales and aggregated to obtain an overall workload score. The NASA-TLX is a well-established instrument for assessing cognitive workload in immersive and interactive systems. A simplified (non-weighted) version of the NASA-TLX was employed to reduce participant burden following VR exposure while retaining the core workload dimensions relevant to immersive interaction. Prior research indicates that the raw NASA-TLX provides reliable and valid workload assessment comparable to the weighted version in controlled experimental settings. The full questionnaire is presented in Appendix A.3 Table A3.

4.3.4. Information-Seeking Behavior

Objective measures of information-seeking behavior were collected through system log data automatically recorded by the VR system. Logged variables included the number of AI queries, frequency of interactions, and time spent requesting information at exhibits. These behavioral metrics complemented the subjective questionnaire data by providing observable evidence of how participants interacted with the AI assistant under each interface modality.

4.4. Research Procedures

Participants were recruited through a social media recruitment campaign. Of the 98 individuals who initially applied, participants with a history of motion sickness or prior familiarity with the textile museum were excluded, and gender balance was considered during group assignment. A total of 75 eligible participants received a brief description of the study and provided informed consent. Participants were then randomly assigned to one of three experimental conditions (N = 25 per group): Voice only, Voice + Text, and Voice + Avatar, as illustrated in Figure 5.

Before the experimental task, participants completed a brief VR orientation session to familiarize themselves with basic controls and movement within the virtual museum; data from this phase were not included in the analysis. Participants subsequently performed a standardized AI interaction task at a predefined exhibit, following the task sequence and actions described in Table 2. Participants were provided with a written list of predefined questions prior to the experiment and were instructed to verbally submit the questions exactly as specified and in the prescribed order. No spontaneous or self-generated questions were permitted during the interaction. The task involved activating the generative AI assistant, asking a fixed set of predefined questions in a prescribed order, requesting a brief clarification or one-sentence summary, and terminating the interaction. Task content, question wording, and sequence were held constant across all conditions, with the AI interface modality as the only

experimental manipulation. Immediately after the VR session, participants completed post-test questionnaires assessing user engagement, perceived information quality, and cognitive workload, while system log data were automatically recorded during the interaction.

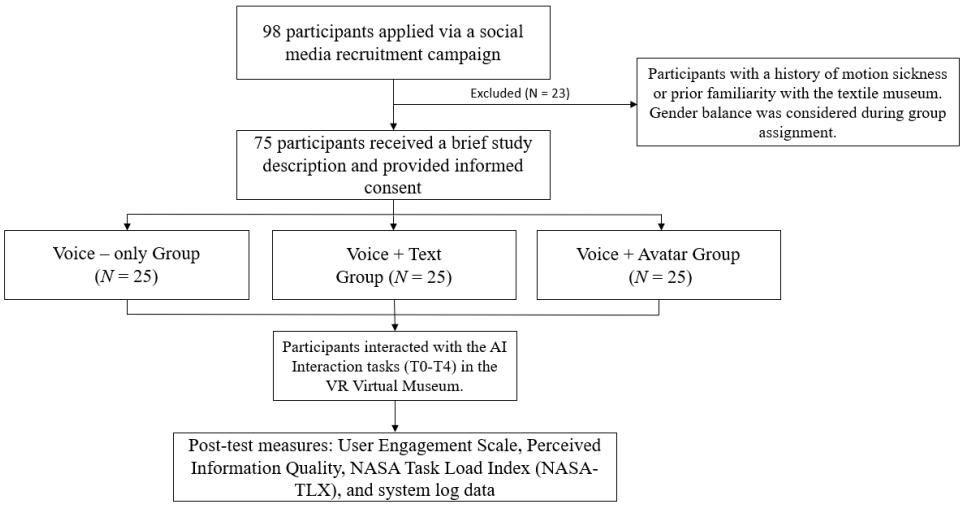


Figure 5. Experimental procedure and participant allocation across the three AI interface conditions in the VR virtual museum study.

Table 2. AI Interaction Tasks in the VR Virtual Museum Study Across Three Groups.

Task No.	Task Description	Purpose
T0	Participants familiarized themselves with basic VR controls and movement within the virtual museum. No data from this phase were included in the analysis.	Reduce novelty effects and ensure basic VR proficiency.
T1	Participants activated the generative AI assistant at a predefined exhibit using the standard interaction method provided by the system.	Ensure consistent initiation of AI interaction across participants.
T2	Participants asked a fixed set of predefined questions related to the exhibit (e.g., object identity, function, materials, and cultural or historical significance) in a prescribed order.	Isolate and compare AI interaction across interface modalities.
T3	Participants requested a brief clarification or one-sentence summary from the AI assistant following the initial responses.	Increase cognitive processing demands and assess perceived information quality.
T4	Participants ended the AI interaction using the system's standard exit mechanism.	Ensure consistent completion of the AI interaction session.

4.5. Data Collection and Data Analysis

Data collection included questionnaire responses and system-generated interaction logs. Quantitative data were analyzed using descriptive statistics and inferential tests to compare differences across the three experimental conditions. Depending on data distribution, one-way ANOVA or non-parametric equivalents were used, followed by post hoc comparisons where appropriate. Behavioral log data were analyzed to complement self-reported measures and provide objective evidence of information-seeking behavior in the VR environment.

5. Results

5.1. User Engagement

Mean user engagement scores differed across the three AI interface modalities, with the Voice + Embodied AI Avatar condition showing the highest engagement (M = 3.97,

SD = 0.29), followed by the Voice + Text condition (M = 3.63, SD = 0.39) and the Voice-only condition (M = 3.55, SD = 0.52), as illustrated in Figure 6. A one-way ANOVA indicated a statistically significant effect of interface modality on user engagement, F (2, 72) = 7.36, $p < 0.001$ (Table 3). Post hoc comparisons using the Games–Howell procedure revealed that engagement in the Voice + Avatar condition was significantly higher than in both the Voice-only condition (mean difference = 0.42, $p = 0.003$) and the Voice + Text condition (mean difference = 0.34, $p = 0.004$), whereas no significant difference was observed between the Voice-only and Voice + Text conditions ($p = 0.81$), as reported in Table 4.

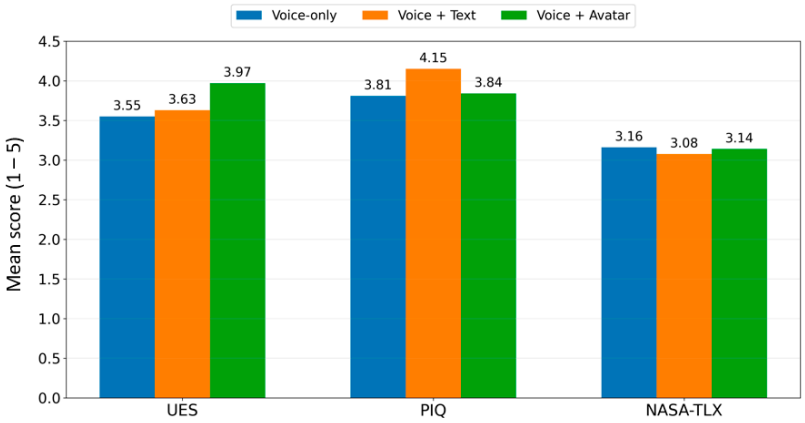


Figure 6. Comparison of User Engagement, Perceived Information Quality, and Cognitive Workload Across AI Interface Conditions.

Table 3. One-way ANOVA results for UES, PIQ, and NASA-TLX across the three AI interface modalities.

Questionnaire	Source	Sum of Squares	df	Mean Square	F	Sig.
UES	Between Groups	2.487	2	1.243	7.36	<0.001
	Within Groups	12.158	72			
	Total	14.645	74			
PIQ	Between Groups	1.772	2	0.886	54.68	<0.001
	Within Groups	1.166	72	0.016		
	Total	2.938	74			
NASA-TLX	Between Groups	0.082	2	0.041	1.28	0.28
	Within Groups	2.30	72	0.032		
	Total	2.38	74			

Table 4. Post hoc pairwise comparisons of UES, PIQ, and NASA-TLX across the three AI interface modalities.

Questionnaire	(I) Group	(J) Group	Mean Difference (I–J)	Sig.
UES	Voice only	Voice + Text	−0.08	0.81
(Post hoc:	Voice only	Voice + Avatar	−0.42	0.003
Games–Howell)	Voice + Text	Voice + Avatar	−0.34	0.004
PIQ	Voice only	Voice + Text	−0.34	<0.001
(Post hoc: Tukey	Voice only	Voice + Avatar	−0.03	0.74
HSD)	Voice + Avatar	Voice + Text	−0.31	<0.001
NASA-TLX	Voice only	Voice + Text	0.08	0.31
(Post hoc: Tukey	Voice only	Voice + Avatar	0.02	0.88
HSD)	Voice + Avatar	Voice + Text	0.06	0.41

5.2. Resulst of Perceived Information Quality

Descriptive statistics showed that perceived information quality was highest in the Voice + Text condition (M = 4.13, SD = 0.14), followed by the Voice + Embodied AI Avatar condition (M = 3.84, SD = 0.13) and the Voice-only condition (M = 3.81, SD = 0.14), as presented in Figure 7. A one-way ANOVA revealed a strong and statistically significant effect of AI interface modality on perceived information quality, F (2, 72) = 54.68, $p < 0.001$ (Table 3). Tukey HSD post hoc analyses indicated that the Voice + Text condition produced significantly higher perceived information quality than both the Voice-only condition (mean difference = 0.34, $p < 0.001$) and the Voice + Avatar condition (mean difference = 0.31, $p < 0.001$), while no statistically significant difference was found between the Voice-only and Voice + Avatar conditions ($p = 0.74$), as summarized in Table 4.

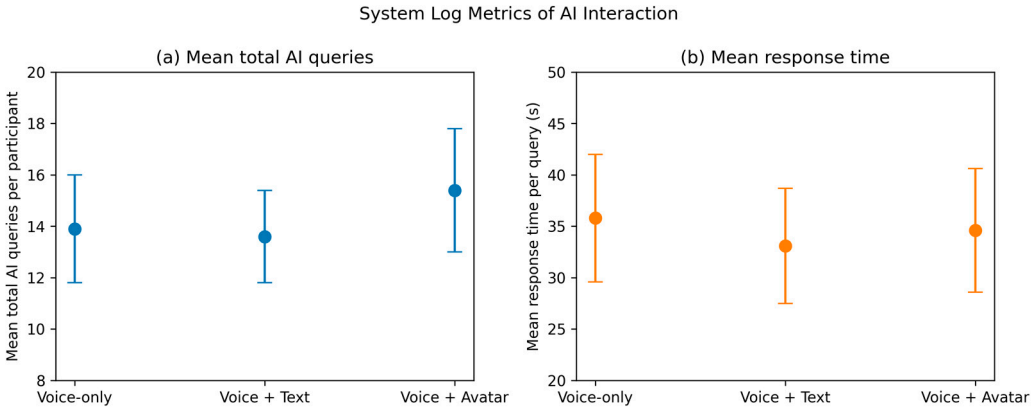


Figure 7. Mean total AI queries per participant across interface conditions.

5.3. Resulst of NASA Task Load Index

Mean NASA-TLX scores indicated comparable levels of perceived cognitive workload across all three AI interface modalities, with mean values of 3.16 (SD = 0.15) for the Voice-only condition, 3.08 (SD = 0.21) for the Voice + Text condition, and 3.14 (SD = 0.14) for the Voice + Embodied AI Avatar condition, as shown in Figure 6. A one-way ANOVA revealed no statistically significant effect of interface modality on subjective cognitive workload, F(2, 72) = 1.28, $p = 0.28$ (Table 3). Consistent with this result, Tukey HSD post hoc comparisons demonstrated that none of the pairwise differences among the three conditions reached statistical significance (all $p > 0.30$), as reported in Table 4.

5.4. Results of System Log Data

System log data were analyzed to examine participants' AI interaction behaviors across the three interface conditions. As shown in Figure 7a, the mean total number of AI queries per participant was highest in the Voice + Avatar condition (mean = 15.4 queries), followed by the Voice-only condition (mean = 13.9 queries) and the Voice + Text condition (mean = 13.6 queries), with comparable variability observed across groups. This pattern suggests that the presence of an embodied AI avatar encouraged more frequent user–AI interactions, reflecting higher levels of exploratory engagement rather than differences in informational demand across conditions.

Furthermore, Figure 7b illustrates the mean response time per query across the three modalities. The results indicate consistent system performance, with mean response times ranging from approximately 33 to 36 s across all groups. The overlapping confidence intervals suggest that there were no substantial differences in technical latency between the Voice-only, Voice + Text, and Voice + Avatar conditions. This stability confirms that the variations observed in user engagement and perceived information quality were driven by

the interface representation itself, rather than by discrepancies in system responsiveness or processing delays.

Figure 8 presents heatmaps of mean dwelling time across exhibition locations for each interface condition. Across all modalities, participants spent the greatest amount of time in the primary exhibition area (Zone B), particularly near the native weaving machine display, indicating consistent focal attention toward central artifacts regardless of AI representation. While overall spatial attention patterns were broadly similar across conditions, the Voice + Avatar modality exhibited slightly higher concentration intensity in key exhibit zones. This pattern aligns with the higher engagement scores observed in the embodied condition, suggesting that the presence of an avatar may encourage sustained spatial attention during AI-assisted exploration. However, no substantial redistribution of spatial attention across zones was observed, indicating that interface modality did not fundamentally alter navigational behavior.

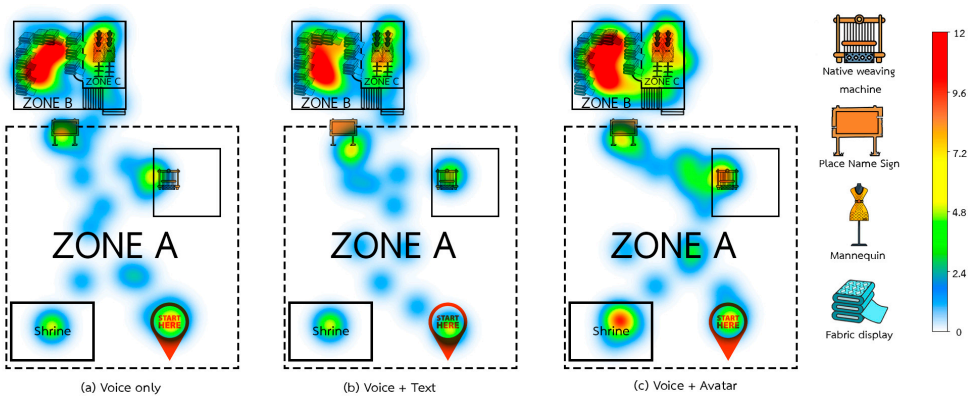


Figure 8. Heatmap of mean dwelling time per exhibition location across interface conditions (Voice only, Voice + Text, and Voice + Avatar).

6. Discussion

6.1. Effects of AI Interface Modality on User Engagement

Addressing RQ1, the results demonstrate that AI interface modality significantly influenced users' engagement during the VR museum experience. Among the three conditions, the Voice + Embodied AI Avatar modality elicited the highest overall user engagement, significantly outperforming both the Voice-only and Voice + Text conditions, while no significant difference was observed between the latter two. This pattern indicates that embodied AI guidance plays a critical role in shaping engagement, even when the informational content delivered by the AI assistant is held constant across modalities.

In immersive VR museum contexts, engagement is primarily driven by social presence and perceived interactivity rather than by mechanisms supporting information verification. The visual embodiment of the AI guide likely acted as a social and spatial anchor, fostering sustained attention and a stronger sense of companionship during exploration. Prior research in immersive and avatar-mediated environments shows that embodied agents enhance affective engagement, motivation, and attentiveness by simulating face-to-face interaction and social cues [36]. These findings suggest that embodied AI avatars are particularly effective for VR museum experiences that emphasize exploration, atmosphere, and prolonged visitor involvement.

6.2. Multimodal Effects on Perceived Information Quality

Regarding RQ2, the findings indicate that AI interface modality played a decisive role in shaping users' perceived information quality during VR museum exploration. Overall, the voice-and-text modality was perceived as providing clearer, more reliable,

and more trustworthy information than both the voice-only and the voice-plus-avatar modalities. This outcome suggests that textual reinforcement substantially supported users' interpretation and evaluation of AI-generated explanations in immersive settings.

In visually rich VR environments, spoken information alone is inherently ephemeral, whereas synchronized text provides reviewability and verification without disrupting exploration. Prior studies on subtitles, multimodal redundancy, and VR interaction design demonstrate that well-integrated textual cues enhance interpretability and perceived reliability while imposing minimal additional cognitive load [41–46]. In contrast, embodied and avatar-based interfaces primarily enhance experiential qualities such as engagement and social presence rather than directly improving informational clarity or credibility [36,47,48]. Together, these findings explain why text augmentation emerged as the most effective modality for perceived information quality in the present study.

6.3. Interface Modality and Cognitive Workload Stability

Regarding RQ3, the results indicate that AI interface modality did not produce meaningful differences in users' subjective cognitive workload during VR museum interaction. Across all three conditions—voice only, voice and text, and voice plus embodied avatar—participants reported comparable levels of mental demand, effort, and frustration while completing the information interaction tasks. This pattern suggests that, despite differences observed in perceived information quality and engagement, the overall cognitive demands imposed by the interaction remained stable. One plausible explanation is that the museum task structure, interaction sequence, and informational content were carefully controlled, allowing users to allocate cognitive resources consistently regardless of interface representation. In this sense, interface modality appeared to redistribute how information was experienced rather than increasing or decreasing overall workload.

This finding aligns with prior research on cognitive workload in immersive VR, which emphasizes that task structure and interaction pacing often play a more decisive role in workload than surface-level interface variations. Systematic reviews of immersive VR applications indicate that well-designed VR tasks can maintain stable workload levels even when multimodal or embodied interfaces are introduced, provided that sensory input is coordinated and does not compete excessively for attention [49]. Empirical studies on multisensory integration in VR further suggest that distributing information across complementary channels does not necessarily increase workload and may even help users manage attentional demands more effectively when interaction complexity is controlled [50]. Similarly, research comparing avatar-based and non-avatar-based interfaces has shown that embodiment does not inherently increase cognitive load, particularly in information-seeking or exploratory tasks [36]. Together, these findings help explain why the present study observed cognitive workload stability across interface modalities, supporting the view that multimodal and embodied AI interfaces can be incorporated into VR museum systems without imposing additional mental burden on users.

6.4. Design Implications for Generative AI-Driven Virtual Museums

The findings indicate that generative AI interface design in virtual museums should be context-sensitive and goal-driven, as each interface modality supports different experiential and informational objectives. Voice-only AI interfaces are most appropriate for museums that emphasize atmospheric exploration and visual immersion—such as contemporary art, architectural exhibitions, or site-based heritage environments—where uninterrupted visual attention is essential. In these settings, brief audio narration can enhance storytelling without introducing visual distraction; however, the absence of per-

sistent visual information limits their suitability for content involving complex facts or detailed historical interpretation.

Voice-and-text interfaces are particularly effective for information-intensive and educational virtual museums, including history, archaeology, science, and cultural heritage collections that involve specialized terminology or structured explanations. Synchronized text enables visitors to review and verify AI-provided information at their own pace, thereby enhancing clarity and trustworthiness. Nevertheless, textual elements must be carefully integrated to avoid visual clutter or divided attention in visually dense VR environments.

Voice-plus-embodied-AI-avatar interfaces are best suited for engagement-oriented and narrative-driven experiences, such as guided virtual tours, character-based storytelling exhibitions, or museums aimed at younger audiences. Embodied AI guides can enhance social presence, emotional connection, and motivation by simulating human-like guidance, yet embodiment alone does not necessarily improve informational precision and may distract users during detailed content inspection. Taken together, these implications suggest that future generative AI-driven virtual museums may benefit from hybrid or adaptive interface strategies that selectively combine voice, text, and embodiment according to exhibit type, visitor intent, and interaction phase to balance informational rigor, engagement, and cognitive workload. By enabling more accessible, inclusive, and engaging digital heritage experiences, such design strategies also support sustainable cultural education and community-oriented heritage dissemination aligned with the Sustainable Development Goals, particularly Quality Education and Sustainable Cities and Communities.

6.5. Limitations and Future Work

This study has several limitations that should be considered when interpreting the findings. First, the experiment was conducted within a single virtual museum context with a fixed exhibition type and interaction task, which may limit the generalizability of the results to other museum genres or content complexities. Different types of virtual museums—such as large-scale heritage sites, socially shared environments, or highly interactive exhibitions—may impose different cognitive and interaction demands that influence how AI interface modalities are experienced. In addition, the study focused on short, structured interactions with a generative AI assistant, which may not fully capture longer-term usage patterns, repeated visits, or evolving user expectations over time.

Future work should extend this research by examining AI interface modalities across a wider range of virtual museum contexts and usage scenarios. Longitudinal studies could explore how sustained interaction with generative AI assistants affects learning, engagement, trust, and potential fatigue over repeated visits. Further research may also investigate adaptive or hybrid interface strategies, in which the AI dynamically shifts between voice-only, text-augmented, and embodied representations based on exhibit characteristics, visitor intent, or interaction phase. Incorporating objective measures—such as eye-tracking, behavioral analytics, or physiological indicators—could provide deeper insight into the cognitive mechanisms underlying users' responses to different AI interface designs and support the development of more effective and personalized virtual museum experiences.

7. Conclusions

This study examined how different generative AI interface modalities—voice only, voice and text, and voice plus embodied AI avatar—shape information interaction in a VR virtual museum when informational content is held constant. The results demonstrate that interface representation plays a critical role: text augmentation primarily enhances perceived information quality by supporting clarity, reviewability, and trust, whereas embodied AI avatars primarily increase user engagement through social presence and expe-

riential richness, without imposing additional cognitive workload. These findings indicate that AI interface modalities serve complementary rather than interchangeable functions in immersive museum environments. From a design perspective, the study highlights the importance of purpose-driven interface selection in generative AI-driven virtual museums to balance informational rigor, engagement, and cognitive stability. By supporting accessible and engaging digital heritage experiences, this work also contributes to sustainable cultural education and knowledge dissemination aligned with the Sustainable Development Goals, particularly Quality Education and Sustainable Cities and Communities.

Author Contributions: Conceptualization, P.A. and P.J.; methodology, P.A. and P.J.; software, S.K.; validation, P.W.; formal analysis, P.A. and P.J.; investigation, P.W.; resources, S.K.; data curation, D.P.; writing—original draft preparation, P.A. and P.J.; writing—review and editing, P.A.; visualization, D.P.; supervision, P.J.; project administration, S.K.; funding acquisition, P.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by Chiang Mai University.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Committee of Research Ethics, Chiang Mai University Research Ethics Committee, Chiang Mai University (COE No. 029/67, date of approval 28 October 2024).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author due to restrictions. The data are not publicly available.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Appendix A.1

Table A1. The User Engagement Scale Questionnaire.

Dimension	Questionnaire
Focused attention	I was fully focused on the interaction with the AI assistant during the VR task.
	I lost track of time while interacting with the AI assistant.
	I was deeply concentrated on the information provided by the AI assistant.
Perceived Usability	The AI assistant was easy to interact with during the VR task.
	The interaction with the AI assistant felt smooth and well-organized.
	I could interact with the AI assistant without confusion or difficulty.
Aesthetic appeal	The presentation of the AI assistant was visually appealing in the VR environment.
	The AI interface enhanced my overall VR experience.
Reward	Interacting with the AI assistant was enjoyable.
	I felt motivated to continue interacting with the AI assistant.
	The AI assistant made the museum experience more interesting.

Appendix A.2

Table A2. Perceived Information Quality Questionnaire.

Dimension	Questionnaire
Accuracy	The information provided by the AI assistant was accurate.
Clarity	The information was clear and easy to understand.
Relevance	The information was relevant to the exhibit I was viewing.
Completeness	The explanations provided by the AI assistant were sufficiently detailed.
Appropriateness	I trusted the information provided by the AI assistant.
Usefulness	The AI assistant addressed my questions appropriately.
Trustworthiness	The information helped me understand the exhibit better.
Confidence	I felt confident relying on the information presented by the AI assistant.

Appendix A.3

Table A3. NASA Task Load Index (NASA-TLX) Questionnaire.

Dimension	Questionnaire
Mental Demand	How mentally demanding was the task while interacting with the AI assistant in the VR museum?
Physical Demand	How physically demanding was the interaction with the AI assistant during the VR task?
Temporal Demand	How hurried or rushed was the pace of the task?
Performance	How successful were you in accomplishing the task goals?
Effort	How hard did you have to work to accomplish your level of performance?
Frustration	How insecure, discouraged, irritated, stressed, or annoyed did you feel during the task?

References

1. Tsita, C.; Satratzemi, M.; Pedefoudas, A.; Georgiadis, C.; Zampeti, M.; Papavergou, E.; Tsiara, S.; Sismanidou, E.; Kyriakidis, P.; Kehagias, D.; et al. A Virtual Reality Museum to Reinforce the Interpretation of Contemporary Art and Increase the Educational Value of User Experience. *Heritage* **2023**, *6*, 4134–4172. [CrossRef]
2. Ruang-on, S.; Kidjaideaw, S.; Kaewchada, S.; Songsri-in, K. The Virtual Museum of Wat Wang Tawan Tok for Cultural Heritage Learning. *KKU Sci. J.* **2024**, *52*, 262–271. [CrossRef]
3. Rahimi, F.; Sadeghi-Niaraki, A.; Choi, S.-M. Generative AI Meets Virtual Reality: A Comprehensive Survey on Applications, Challenges, and Future Directions. *IEEE Access* **2025**, *13*, 94893–94909. [CrossRef]
4. Iacono, S.; Scaramuzzino, M.; Martini, L.; Panelli, C.; Zolezzi, D.; Perotti, M.; Traverso, A.; Vercelli, G.V. Virtual Reality in Cultural Heritage: A Setup for Balzi Rossi Museum. *Appl. Sci.* **2024**, *14*, 3562. [CrossRef]
5. Sylaiou, S.; Fidas, C. Virtual Humans in Museums and Cultural Heritage Sites. *Appl. Sci.* **2022**, *12*, 9913. [CrossRef]
6. Voinea, G.D.; Gîrbacia, F.; Postelnicu, C.C.; Duguleana, M.; Antonya, C.; Soica, A.; Stănescu, R.C. Study of Social Presence While Interacting in the Metaverse with an Augmented Avatar during Autonomous Driving. *Appl. Sci.* **2022**, *12*, 11804. [CrossRef]
7. Yang, F.-C.; Acevedo, P.D.; Guo, S.; Choi, M.; Mousas, C. Embodied Conversational Agents in Extended Reality: A Systematic Review. *IEEE Access* **2025**, *13*, 79805–79824. [CrossRef]
8. Ariya, P.; Khanchai, S.; Intawong, K.; Puritat, K. Enhancing Textile Heritage Engagement through Generative AI-Based Virtual Assistants in Virtual Reality Museums. *Comput. Educ. X Real.* **2025**, *7*, 100112. [CrossRef]
9. Sangamuang, S.; Wongwan, N.; Intawong, K.; Khanchai, S.; Puritat, K. Gamification in Virtual Reality Museums: Effects on Hedonic and Eudaimonic Experiences in Cultural Heritage Learning. *Informatics* **2025**, *12*, 27. [CrossRef]
10. Czimre, K.; Teperics, K.; Molnár, E.; Kapusi, J.; Saidi, I.; Gusman, D.; Bujdosó, G. Potentials in Using VR for Facilitating Geography Teaching in Classrooms: A Systematic Review. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 332. [CrossRef]
11. Theodoropoulos, A.; Antoniou, A. VR Games in Cultural Heritage: A Systematic Review of the Emerging Fields of Virtual Reality and Culture Games. *Appl. Sci.* **2022**, *12*, 8476. [CrossRef]
12. Li, J.; Lv, C. Exploring User Acceptance of Online Virtual Reality Exhibition Technologies: A Case Study of Liangzhu Museum. *PLoS ONE* **2024**, *19*, e0308267. [CrossRef] [PubMed]

13. Jangra, S.; Singh, G.; Mantri, A.; Ahmed, Z.; Liew, T.W.; Ahmad, F. Exploring the Impact of Virtual Reality on Museum Experiences: Visitor Immersion and Experience Consequences. *Virtual Real.* **2025**, *29*, 84. [CrossRef]
14. Chang, S.; Suh, J. The Impact of VR Exhibition Experiences on Presence, Interaction, Immersion, and Satisfaction: Focusing on the Experience Economy Theory (4Es). *Systems* **2025**, *13*, 55. [CrossRef]
15. Xu, H.; Li, Y.; Tian, F. Contrasting Physical and Virtual Museum Experiences: A Study of Audience Behavior in Replica-Based Environments. *Sensors* **2025**, *25*, 4046. [CrossRef]
16. Machidon, O.M.; Duguleana, M.; Carrozzino, M. Virtual Humans in Cultural Heritage ICT Applications: A Review. *J. Cult. Herit.* **2018**, *33*, 249–260. [CrossRef]
17. Alabau, A.; Fabra, L.; Martí-Testón, A.; Muñoz, A.; Solanes, J.E.; Gracia, L. Enriching User–Visitor Experiences in Digital Museology: Combining Social and Virtual Interaction within a Metaverse Environment. *Appl. Sci.* **2024**, *14*, 3769. [CrossRef]
18. Song, Y.; Wu, K.; Ding, J. Developing an Immersive Game-Based Learning Platform with Generative Artificial Intelligence and Virtual Reality Technologies—"LearningverseVR". *Comput. Educ. X Real.* **2024**, *4*, 100069. [CrossRef]
19. Lee, L.K.; Chan, E.H.; Tong, K.K.L.; Wong, N.K.H.; Wu, B.S.Y.; Fung, Y.C.; Fong, E.K.S.; Leong Hou, U.; Wu, N.I. Utilizing Virtual Reality and Generative AI Chatbot for Job Interview Simulations. In *Proceedings of the 2024 International Symposium on Educational Technology (ISET 2024), Macau, Macao, 29 July–1 August 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 209–212.
20. Hong, S.; Moon, J.; Eom, T.; Awoyemi, I.D.; Hwang, J. Generative AI-Enhanced Virtual Reality Simulation for Pre-Service Teacher Education: A Mixed-Methods Analysis of Usability and Instructional Utility for Course Integration. *Educ. Sci.* **2025**, *15*, 997. [CrossRef]
21. Ngo, B.; Harman, J.; Türkay, S. Using LLMs to Develop Personalities for Embodied Conversational Agents in Virtual Reality. In *Proceedings of OZCHI 2024: 36th Australasian Conference on Human–Computer Interaction*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 745–751.
22. Dasa, D.; Board, M.; Rolfe, U.; Dolby, T.; Tang, W. Evaluating AI-Driven Characters in Extended Reality (XR) Healthcare Simulations: A Systematic Review. *Artif. Intell. Med.* **2025**, *170*, 103270. [CrossRef]
23. Kiuchi, K.; Otsu, K.; Hayashi, Y. Psychological Insights into the Research and Practice of Embodied Conversational Agents, Chatbots and Social Assistive Robots: A Systematic Meta-Review. *Behav. Inf. Technol.* **2024**, *43*, 3696–3736. [CrossRef]
24. Li, Y.; Yang, R.; Zou, J.; Xu, H.; Tian, F. Human-Centric Virtual Museum: Redefining the Museum Experience through Immersive and Interactive Environments. *Int. J. Hum.–Comput. Interact.* **2025**, *41*, 8426–8437. [CrossRef]
25. Spyrou, O.; Hurst, W.; Krampe, C. A Reference Architecture for Virtual Human Integration in the Metaverse: Enhancing the Galleries, Libraries, Archives, and Museums (GLAM) Sector with AI-Driven Experiences. *Future Internet* **2025**, *17*, 36. [CrossRef]
26. Chen, M.-X.; Hu, H.; Yao, R.; Qiu, L.; Li, D. A Survey on the Design of Virtual Reality Interaction Interfaces. *Sensors* **2024**, *24*, 6204. [CrossRef] [PubMed]
27. Yuan, Z.; He, S.; Liu, Y.; Yu, L. MEinVR: Multimodal Interaction Techniques in Immersive Exploration. *Vis. Inform.* **2023**, *7*, 37–48. [CrossRef]
28. Chang, Z.; Bai, H.; Zhang, L.; Gupta, K.; He, W.; Billinghurst, M. The Impact of Virtual Agents' Multimodal Communication on Brain Activity and Cognitive Load in Virtual Reality. *Front. Virtual Real.* **2022**, *3*, 995090. [CrossRef]
29. Elfleet, M.; Chollet, M. Investigating the Impact of Multimodal Feedback on User-Perceived Latency and Immersion with LLM-Powered Embodied Conversational Agents in Virtual Reality. In *Proceedings of IVA 2024*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 1–9. [CrossRef]
30. Jolibois, S.C.; Ito, A.; Nose, T. The Development of an Emotional Embodied Conversational Agent and the Evaluation of the Effect of Response Delay on User Impression. *Appl. Sci.* **2025**, *15*, 4256. [CrossRef]
31. Potdevin, D.; Clavel, C.; Sabouret, N. Virtual Intimacy in Human–Embodied Conversational Agent Interactions: The Influence of Multimodality on Its Perception. *J. Multimodal User Interfaces* **2021**, *15*, 25–43. [CrossRef]
32. Radianti, J.; Majchrzak, T.A.; Fromm, J.; Wohlgenannt, I. A Systematic Review of Immersive Virtual Reality Applications for Higher Education: Design Elements, Lessons Learned, and Research Agenda. *Comput. Educ.* **2020**, *147*, 103778. [CrossRef]
33. Sutcliffe, A. *Multimedia and Virtual Reality: Designing Multisensory User Interfaces*; Psychology Press: London, UK, 2003.
34. Nelson, B.C.; Erlandson, B.E. Managing Cognitive Load in Educational Multi-User Virtual Environments: Reflections on Design Practice. *Educ. Technol. Res. Dev.* **2008**, *56*, 619–641. [CrossRef]
35. Marucci, M.; Di Flumeri, G.; Borghini, G.; Sciaraffa, N.; Scandola, M.; Pavone, E.F.; Babiloni, F.; Betti, V.; Aricò, P. The Impact of Multisensory Integration and Perceptual Load in Virtual Reality Settings on Performance, Workload and Presence. *Sci. Rep.* **2021**, *11*, 4831. [CrossRef] [PubMed]
36. Pan, Y.; Steed, A. Avatar Type Affects Performance of Cognitive Tasks in Virtual Reality. In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology (VRST)*; Association for Computing Machinery: New York, NY, USA, 2019. [CrossRef]
37. Kruijff, E.; Marquardt, A.; Trepkowski, C.; Schild, J.; Hinkenjann, A. Designed Emotions: Challenges and Potential Methodologies for Improving Multisensory Cues to Enhance User Engagement in Immersive Systems. *Vis. Comput.* **2017**, *33*, 471–488. [CrossRef]

38. O'Brien, H.L.; Cairns, P.; Hall, M. A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale (UES) and New UES Short Form. *Int. J. Hum.–Comput. Stud.* **2018**, *112*, 28–39. [CrossRef]
39. Hart, S.G. NASA-Task Load Index (NASA-TLX); 20 Years Later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*; Sage: Los Angeles, CA, USA, 2006; Volume 50, pp. 904–908. [CrossRef]
40. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*; Hancock, P.A., Meshkati, N., Eds.; North-Holland: Amsterdam, The Netherlands, 1988; pp. 139–183.
41. Rothe, S.; Tran, K.; Hußmann, H. Dynamic Subtitles in Cinematic Virtual Reality. In *Proceedings of the 2018 ACM International Conference on Interactive Experiences for TV and Online Video (TVX '18), Seoul, Republic of Korea, 26-28 June 2018*; ACM: New York, NY, USA, 2018; pp. 209–214. [CrossRef]
42. Rothe, S.; Hußmann, H. Guiding Visual Attention in Cinematic Virtual Reality by Subtitles. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video (TVX '19), Salford, UK, 5–7 June 2019*; ACM: New York, NY, USA, 2019; pp. 121–128.
43. Kaplan-Rakowski, R.; Gruber, A. An Experimental Study on Reading in High-Immersion Virtual Reality. *Br. J. Educ. Technol.* **2024**, *55*, 541–559. [CrossRef]
44. Yoshida, S.; Koyama, Y.; Ushiku, Y. Toward AI-Mediated Avatar-Based Telecommunication: Investigating Visual Impression of Switching between User- and AI-Controlled Avatars in Video Chat. *IEEE Access* **2024**, *12*, 113372–113383. [CrossRef]
45. Kim, H.; Lee, S.; Kang, C. From Controllers to Multimodal Input: A Chronological Review of XR Interaction across Device Generations. *Sensors* **2025**, *26*, 196. [CrossRef] [PubMed]
46. Wang, J.; Deng, Z.; Deng, D.; Wang, X.; Sheng, R.; Cai, Y.; Qu, H. Empowering Multimodal Analysis with Visualization: A Survey. *Comput. Sci. Rev.* **2025**, *57*, 100748. [CrossRef]
47. Horvat, N.; Kunnen, S.; Štorga, M.; Nagarajah, A.; Škec, S. Immersive Virtual Reality Applications for Design Reviews: Systematic Literature Review and Classification Scheme for Functionalities. *Adv. Eng. Inform.* **2022**, *54*, 101760. [CrossRef]
48. Karuzaki, E.; Partarakis, N.; Patsiouras, N.; Zidianakis, E.; Katzourakis, A.; Pattakos, A.; Zabulis, X. Realistic Virtual Humans for Cultural Heritage Applications. *Heritage* **2021**, *4*, 4148–4171. [CrossRef]
49. Liu, J.Y.W.; Yin, Y.H.; Kor, P.P.K.; Cheung, D.S.K.; Zhao, I.Y.; Wang, S.; Leung, A.Y. The Effects of Immersive Virtual Reality Applications on Enhancing the Learning Outcomes of Undergraduate Health Care Students: Systematic Review with Meta-Synthesis. *J. Med. Internet Res.* **2023**, *25*, e39989. [CrossRef]
50. El Iskandarani, M.; Bolton, M.; Riggs, S.L. Examining Dual-Task Interference Effects of Visual and Auditory Perceptual Load in Virtual Reality. *Int. J. Hum.–Comput. Stud.* **2025**, *205*, 103619. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.