An AI-Powered Multimodal Interaction System for Engaging with Digital Art: A Human-Centered Approach to HCI

Andrea Ferracani(✉), Simone Ricci, Filippo Principi, Giuseppe Becchi, Niccolò Biondi, Alberto Del Bimbo, Marco Bertini, and Pietro Pala

MICC - DINFO, University of Firenze, Florence, Italy
info@micc.unifi.it
https://www.micc.unifi.it

Abstract. In recent years, conversational systems have experienced significant advancements due to progress in Artificial Intelligence (AI) and Natural Language Processing (NLP). Models like ChatGPT, introduced by OpenAI in 2019, have transformed how users interact with machines, enabling real-time, natural language conversations. These systems have been deployed across various domains, from customer services to virtual assistants, and have opened new opportunities for human-machine interaction in fields such as art and education. Conversational Agents (CAs) have gained prominence in both academic research and practical applications, driven by their potential to enhance user interaction through natural language. While CAs have found success in contexts like healthcare, education, and commerce, challenges persist in ensuring that user experiences meet expectations, particularly in terms of engagement and emotional connection. A growing body of research in Human-Computer Interaction (HCI) emphasizes the importance of multimodal interaction, combining speech, movement, and other sensory stimuli, to foster more intuitive and engaging experiences. In interactive digital art, AI-driven systems are increasingly used to measure and influence emotional responses through sensory stimulation and multi-dimensional interactions. These systems aim to blur the boundaries between physical and digital environments, offering users immersive experiences. The integration of AI technologies into these installations highlights the potential of conversational agents to evoke deeper emotional engagement, promoting individual well-being and contributing to sustainable development in cultural and educational settings. Grounding from research on this topics [1–24], this article introduces a multimodal system that leverages conversational AI and body tracking to engage users with digital reproductions of artworks, advancing both interaction design and educational potential.

Keywords: HCI · Multimodal interaction · AI · Conversational systems · GPT · Computer vision · Cultural heritage · Education

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
H. Degen and S. Ntoa (Eds.): HCII 2025, LNAI 15821, pp. 281–294, 2025.
https://doi.org/10.1007/978-3-031-93418-6_19

1 State-of-the-Art: Recent Advancements in HCI

Recent advancements in Human-Computer Interaction (HCI) have increasingly focused on multimodal and emotional interactions as pivotal mechanisms for enhancing user engagement and learning experiences. This section highlights key research contributions that underline the potential of these approaches in various contexts, with particular emphasis on their application to cultural heritage and interactive systems.

1.1 Multimodal Interaction for Enhanced User Engagement

Multimodal systems leverage multiple sensory inputs, such as voice, gestures, and touch, to create richer and more engaging user experiences. These systems reduce cognitive load by aligning communication channels and offering intuitive interfaces. Sundar and Heck [16] outline a comprehensive taxonomy for multimodal conversational AI, emphasizing the importance of integrating modalities to enhance alignment, fusion, and co-learning between sensory inputs. This approach supports natural and fluid interactions, which are critical in both cultural and educational settings. In cultural heritage applications, multimodal interaction has been shown to foster a sense of presence and improve information retention. Tsepapadakis and Gavalas [12] demonstrate this with Exhibot, which combines Audio Augmented Reality (AAR), chatbots, and IoT devices to provide spatially registered soundscapes and conversational interaction. This system effectively blends auditory and contextual elements to create an immersive experience for museum visitors, highlighting the importance of synchronized multimodal inputs for engagement. Petrelli and Roberts [10] further underline the value of multimodal systems in their exploration of virtual reality (VR) and tangible interaction in museums. By integrating sight (VR) and touch (tangible interaction), these systems enable users to interact with artifacts in ways that evoke "cultural presence." This fusion of modalities not only enhances accessibility but also enriches the interpretive process, making historical narratives more engaging and meaningful.

1.2 Emotional Interaction and Flow State in Learning Environments

Emotional engagement plays a pivotal role in achieving a state of flow, where users are fully immersed in an activity. Csikszentmihalyi's flow theory [6] identifies the balance between challenge and skill as a key driver of this state. Digital technologies that promote emotional and psychological immersion have been increasingly used in cultural and educational domains to achieve this balance. Ding [6] highlights how digital technologies, such as VR, AR, and Mixed Reality (MR), can create a "person-environment" interaction, enabling visitors to experience exhibits on intellectual and emotional levels. This interaction fosters a complete psychological immersion, where users remain focused on content, receive immediate feedback, and derive pleasure from the experience. Chen

and Ibrahim [5] further emphasize the role of sensory stimulation and multidimensional interaction in eliciting emotional responses. In interactive installation art, these factors contribute to profound emotional engagement, promoting individual well-being and sustainable development. By understanding and measuring emotional responses, designers can refine systems to enhance user satisfaction and learning outcomes. In museum settings, emotional interaction through conversational agents has also been shown to improve user experiences. Schaffer et al. [2] found that conversational systems employing advanced natural language understanding (NLU) components, like ChiM, create emotionally resonant interactions that cater to individual preferences. These systems are particularly effective in engaging users who may not frequently use traditional audio guides, demonstrating their potential to reach broader audiences.

1.3 Advancing Multimodal and Emotional Interaction

Our proposed system builds on these foundational studies to integrate multimodal interaction and emotional design into a cohesive platform for engaging users with digital cultural artifacts. By combining conversational AI, body tracking, and spatial interaction, we aim to deliver experiences that align with Sundar and Heck's taxonomy [16] and the emotional engagement principles highlighted by Chen and Ibrahim [5]. The body tracking component leverages intuitive movement-based controls to simulate natural interaction with 3D artifacts, echoing the approach of Petrelli and Roberts [10], where physical interaction fosters cultural presence. Additionally, the integration of conversational AI ensures personalized interactions, as demonstrated by Exhibot [12] and Iris+ [27], which successfully engage users by providing contextually relevant information in natural, human-like dialogues. Finally, our system's focus on emotional design aims to achieve a flow state in users, ensuring sustained engagement and deeper learning. By incorporating principles of sensory stimulation, immediate feedback, and intuitive interaction, the system aligns with Csikszentmihalyi's theory [6] and extends its application to the digital exploration of cultural heritage.

2 The System

This paper presents a multimodal interaction system that leverages AI and natural interaction to engage users with digital reproductions of sculptures. To demonstrate the capabilities of the proposed system, we describe an installation at the Atelier Marco Bagnoli in Empoli, Italy, as a case study. The installation features two artworks by the contemporary Italian artist Marco Bagnoli: "Colui che sta" and "Come figura d'arciere". The latter is shown in the reference image (Fig. 1)[1].

[1] More info can be found at http://www.marcobagnoli.it/.

Artistic Highlights: Shadows and Profiles. A unique aspect of the sculptures *Come figura d'arciere* and *Colui che sta* is the shadow they cast when viewed from specific angles. These shadows reveal symbolic profiles, enriching the artworks' narrative.
Come figura d'arciere (see Fig.1): the shadow forms the profile of an archer releasing an arrow, where the arrow is symbolized by a disc in the sculpture. This composition draws inspiration from a Leonardo da Vinci study, creating a bridge between contemporary and Renaissance art.
Colui che sta (see Fig.2): the shadows depict two distinct profiles: Isabella d'Este, a Renaissance icon, and Shiva, the Hindu deity. These profiles symbolize the intersection of Eastern and Western traditions, fostering a dialogue between cultural and spiritual realms.

The system combines speech recognition, body tracking, and spatial interaction to offer an immersive experience, integrating verbal and non-verbal inputs to enhance user engagement and emotional connection. By allowing users to explore digital 3D sculptures using natural language and physical movement, the system represents an innovative approach in HCI.

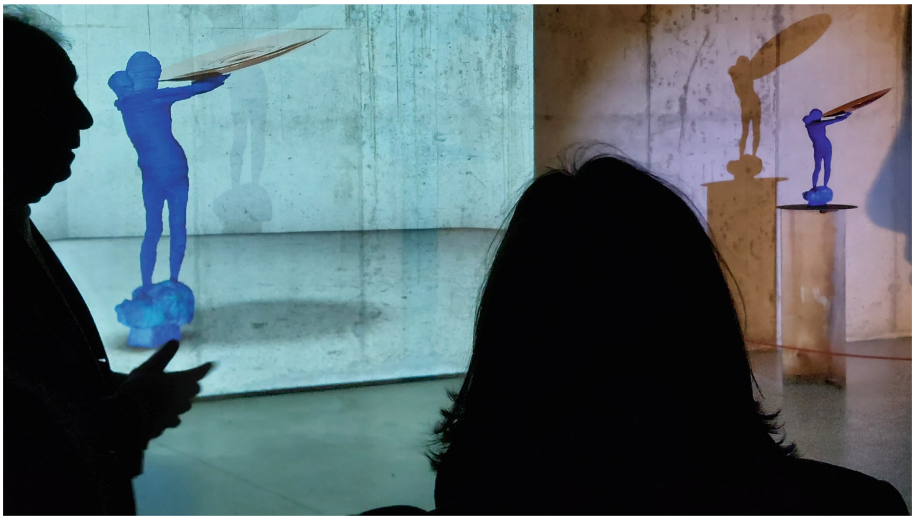

Fig. 1. User interacting through voice and movements with the natural interface showing the sculpture "Come figura d'arciere" of the italian artist Marco Bagnoli.

Users can ask questions about the sculptures via natural language through conversational AI powered by ChatGPT-mini and a Generative Augmented Retrieval (RAG) approach, providing contextually relevant responses. The sys-

tem uses ElevenLabs' APIs[2] for *text-to-speech* and Deepgram's APIs[3] *speech-to-text*, ensuring seamless, high-quality voice interaction. Text inputs are converted into spoken responses, enhancing interaction fluidity. In addition to voice interaction, the system incorporates tracking of body movements. Users can move around the installation, where their position and movement control the viewpoint of the 3D sculptures. By walking in an area along an elliptical path, users rotate the environment, changing their visual perspective. Body tracking is implemented using YOLOv11[4] pose estimation, which tracks head and torso movements for precise control without traditional input devices. This multimodal system provides a more natural way to explore digital art. By combining verbal and non-verbal inputs, it fosters a stronger emotional connection with the artwork, aligning with emotional design principles in HCI.

2.1 3D Data Acquisition and Digital Representation of Artworks

Two artworks by the contemporary Italian artist Marco Bagnoli, "Colui che sta" and "Come figura d'arciere" were digitized using high-resolution 3D scanning techniques, resulting in point clouds that included both geometric information and RGB color data for the surfaces. From these point clouds, detailed meshes were generated, with per-vertex textures accurately capturing the original coloration of the sculptures in order to support an engaging and emotionally resonant experience in the interactive video installation. The raw meshes obtained from the scanning process contained structural defects such as holes and cavities, which compromised the geometric consistency and visualization quality. To ensure high-fidelity and functional representations for the video installation, manual intervention was required to restore surface continuity.

2.2 System Architecture

The system consists of three core components:

- **Physical interaction engine**: This module performs body-tracking inputs using YOLOv11. The system synchronizes these inputs for an immersive experience (see Sect. 2.3).
- **Voice interaction engine Speech Interface**: This module processes verbal inputs via OpenAI and RAG for natural language queries. The system uses Deepgram's *speech-to-text* and ElevenLabs' *text-to-speech* APIs for voice interaction, converting speech to text and generating realistic spoken responses (see Sect. 2.4).
- **3D visualization module**: Developed using Qt Python[5], this module renders the sculptures in real-time. User movement adjusts the viewing angles, allowing exploration from multiple perspectives (see Sect. 2.5).

[2] https://elevenlabs.io/api.
[3] https://deepgram.com/product/speech-to-text.
[4] https://github.com/ultralytics/ultralytics.
[5] https://www.qt.io/.

2.3 Physical Interaction Engine

The Physical Interaction Engine leverages the YoloV11 Pose model, enhanced with a basic tracker to maintain consistent IDs for detected poses. The input is processed from a camera positioned centrally above the interaction area, directly above the 3D artwork. This placement ensures the precise alignment of the physical interaction area's midpoint with the virtual 3D sculpture's rotation axis, allowing a seamless 360° rotation of the artwork. The model identifies joints for each individual in the scene and stores them in a dictionary with their corresponding IDs. Key joints include the feet and hips. The feet are used to detect when a user enters the interaction area, while the center of the hips serves as a reference point to correlate the user's movement with the statue. The hips, being the body's center of gravity, exhibit minimal oscillations during movement or while stationary, ensuring stable tracking. When a user enters the interaction area, they are assigned a position in a FIFO queue. The first user in the queue controls the 3D artwork. If an active user leaves the interaction zone for more than five seconds (a temporal threshold), they are removed from the queue, and control passes to the next user. The user controls the artwork's rotation through movement (Fig. 2).

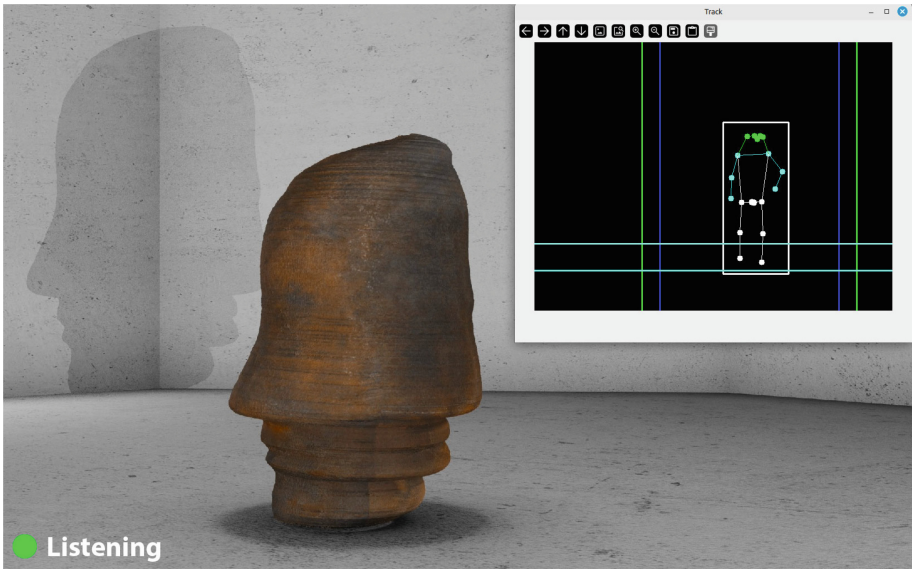

Fig. 2. The Physical Interaction Engine detects and tracks the user's skeletal structure using the YoloV11 Pose model, visualized on the right panel. Key joints, such as feet and hips, are identified to correlate user movements with the virtual 3D artwork's rotation. The user is positioned in the designated interaction area, where his movements control the viewing perspective of "Colui che sta" by Marco Bagnoli.

A full 360° rotation of the artwork is horizontally mapped to the interaction zone. As users traverse the area, they gain a comprehensive view of the piece. The rotation direction is opposite to the user's movement, simulating the natural experience of walking around an object in the real world. This design ensures an intuitive and cognitively seamless interaction. To enhance responsiveness, the 3D rendering is precomputed offline using Blender, ensuring higher frame rates and eliminating the need for online shadow and detail rendering. High-quality 4K images are preloaded into RAM for instantaneous access, minimizing disk reads and maximizing visual fidelity. When a user enters the interaction zone, the voice interaction system is activated. Conversely, when a user leaves, the system resets the context, enabling personalized and independent conversations for subsequent users.

2.4 Voice Interaction Engine

The Voice interaction module is designed to facilitate seamless natural language interaction between the user and the system. This module integrates several components, including speech-to-text (STT) and text-to-speech (TTS) systems, as well as a conversational AI model, to deliver a dynamic and contextually aware dialogue experience.

Speech-to-Text Processing. The speech recognition component employs Deepgram's[6] speech-to-text API for real-time transcription of user input. Configured to support Italian (*it-IT*), the system processes audio captured through a microphone and converts it into text with high accuracy. The STT system is optimized with features such as punctuation, smart formatting, and endpointing for fluid and structured transcription.

Conversational AI Model. The transcribed text is processed by a conversational AI model implemented using `OpenAI Chat Api` (based on model GPT-4.0-mini). This model is configured with a system-level prompt that defines the conversational style and ensures domain-specific knowledge relevant to the artwork and the installation context. The interaction memory is managed using a *ConversationBufferMemory*, which maintains the context of the dialogue, enabling coherent and contextually relevant responses over multiple interactions.

To initiate the interaction, the system pre-loads a welcome message, ensuring users are guided effectively. The AI processes the user input and generates responses in real time, which are then dynamically modified to correct specific terms for improved pronunciation, to align with phonetic conventions.

Text-to-Speech System. The generated response is converted to speech using ElevenLabs' text-to-speech API. The TTS system streams audio in real time, ensuring minimal latency between the user's query and the system's verbal

[6] https://deepgram.com/.

response. The selected voice model is multilingual and tailored to produce clear, natural, and engaging speech. The use of streaming audio enhances responsiveness, providing users with an uninterrupted and immersive conversational experience.

Multithreaded Architecture. The module operates within a multithreaded architecture to handle simultaneous audio processing and conversational management. One thread manages the dialogue loop, including transcription and response generation, while another controls rendering of the visual interaction, ensuring synchronization between physical movement and the voice-based interaction.

User Interaction Flow. The user interaction follows a structured flow, as described in Fig. 3: 1. The system detects the user entering the designated interaction area. 2. A microphone captures the user's speech, which is transcribed into text. 3. The AI processes the text and generates a contextually relevant response. 4. The response is streamed as audio via the TTS system. 5. The system maintains conversational context and adapts responses based on previous interactions.

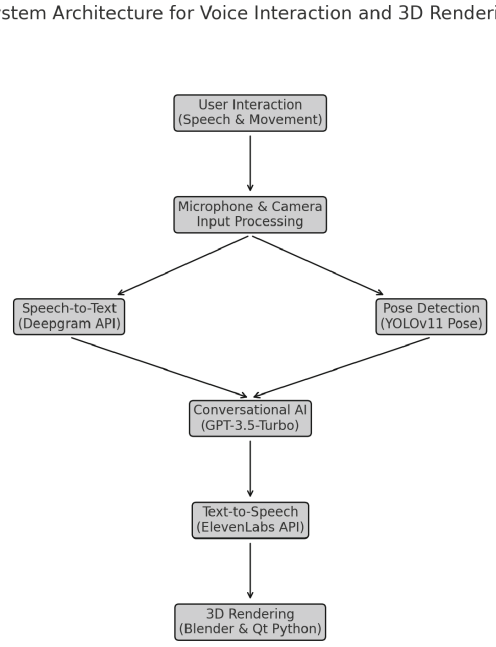

Fig. 3. System architecture illustrating the flow of user interaction through speech and movement, integrating modules for speech-to-text, pose detection, conversational AI, text-to-speech, and 3D rendering.

This module provides users with an accessible and engaging means of interacting with the digital artwork, combining real-time voice processing with domain-specific conversational capabilities.

2.5 3D Visualization Module

The 3D Visualization Module was developed to render the sculptures in real-time, enabling users to explore the artworks from multiple perspectives through natural interaction. This module integrates the optimized 3D meshes into a dynamic environment designed to maintain high visual fidelity while ensuring smooth and responsive user interactions. Built using Blender for scene design and Qt Python for rendering, the module combines advanced 3D rendering techniques with intuitive controls.

To enhance the viewing experience, the virtual environment includes carefully calibrated lighting and camera animations. These features not only emphasize the fine details of the sculptures but also provide an engaging and immersive exploration. The module ensures that the user's movements directly adjust the viewpoint, simulating the natural experience of observing physical artworks from different angles. The following subsections describe the integration of the sculptures into the 3D environment and the rendering techniques employed to achieve seamless and high-quality visualization.

Integration into the 3D Environment. Once processed, the meshes were imported into a 3D environment designed in Blender[7]. The sculptures were positioned at the center of a virtual scene equipped with a multi-layered lighting system and an animated camera. The lighting setup included: 1) A three-point lighting configuration to emphasize the details of the sculptures. 2) A general light source for global illumination. 3) A spotlight to create a static shadow on a specific portion of a virtual wall, selected as a conceptual element by the artist. The camera was programmed to simulate a 360° rotation around the sculptures, allowing viewers to observe the artwork from every angle. The shadow, however, remained fixed on a designated area of the wall, providing a dynamic yet conceptually consistent visual experience. The environment is described in Fig. 4.

Rendering for Smooth Visualization. To achieve a smooth and seamless simulation of the 360° rotation, the animation was rendered into approximately 1,280 individual high-resolution images. This approach ensured a continuous and high-fidelity visual experience, meeting both the technical and conceptual requirements of the project.

The combination of advanced 3D scanning, meticulous mesh optimization, and carefully designed virtual environments underscores the importance of integrating technological precision with artistic vision. These steps enabled the creation of an immersive and engaging installation that faithfully represents the artist's original works.

[7] https://www.blender.org/.

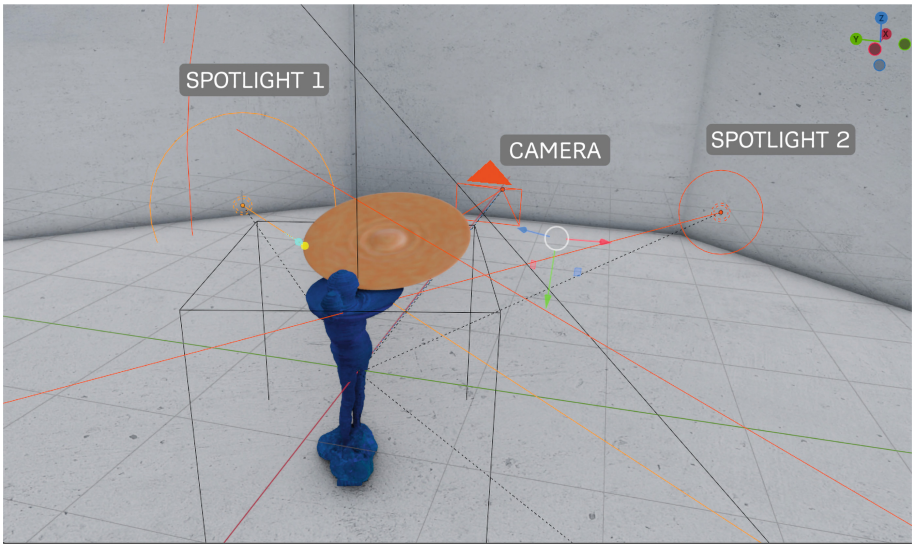

Fig. 4. Visualization of the 3D environment created in Blender. The scene includes a sculpture positioned at the center, surrounded by a three-point lighting system (Spotlight 1, Spotlight 2, and a global light source) and an animated camera. The setup allows a 360° rotation around the sculpture, emphasizing its details while maintaining a fixed shadow on a specific portion of the wall, as conceptualized by the artist.

2.6 Usability Testing and Educational Potential

To evaluate the system's ability to enhance user engagement, emotional response, and understanding of the artwork, comprehensive usability tests were conducted. The tests aimed to assess the educational and emotionally resonant experience offered by the multimodal system compared to traditional point-and-click interfaces.

Study Design and Procedure. The study followed a between-subjects design, with a total of 16 participants ($N = 24$) randomly assigned to two groups: the multimodal system group ($n = 12$) and the conventional interface group ($n = 12$). Participants explored digital representations of sculptures and answered questions about their cultural and historical context.

The multimodal system group used voice interaction and body movement to engage with the artwork, while the conventional interface group relied on mouse and keyboard inputs. The testing environment was controlled to ensure consistency across sessions, and participants were given identical tasks to complete.

Metrics and Data Collection. Key metrics were recorded to evaluate the effectiveness of the interaction modalities:

- **Task Completion Time:** Measured in seconds to determine the efficiency of the system.
- **Engagement:** Rated on a 5-point Likert scale to assess the user's level of interest and involvement.
- **Empathy:** Rated on a 5-point Likert scale to evaluate the user's emotional connection with the artwork.

Data from both groups were analyzed using a one-way ANOVA to identify significant differences in performance and user experience. A significance level of $p<0.05$ was used to determine statistical relevance.

Results and Analysis. The results, summarized in Table 1, indicate that the multimodal system outperformed the conventional interface in all measured metrics. Participants using the multimodal system completed tasks more efficiently, reported higher engagement levels, and demonstrated a stronger emotional connection with the artwork.

The one-way ANOVA revealed statistically significant differences across all metrics:

- **Task completion time:** f-ratio $= 4.29$, $p < 0.05$
- **Engagement:** f-ratio $= 102.52$, $p < 0.01$
- **Empathy:** f-ratio $= 154.09$, $p < 0.01$

Table 1. Comparison of performance metrics between the multimodal system and the conventional interface.

Metric	Multimodal system	Conventional interface
Task completion time (sec)	32.45	49.25
Engagement (1–5)	4.66	2.58
Empathy (1–5)	4.83	2.50

Discussion and Educational Implications. The findings highlight the potential of multimodal interaction to enhance the educational and emotional impact of digital cultural heritage. The integration of body movement and conversational AI fosters a deeper connection with the artwork, aligning with prior research on sensory engagement and emotional design in interactive art experiences. The system's ability to provide an immersive and intuitive exploration demonstrates its value as an innovative educational tool for museums and cultural institutions.

2.7 Conclusions

The findings from this study demonstrate that the proposed multimodal interaction system has significant potential as an educational tool, particularly in the domain of digital cultural heritage. By integrating conversational AI, body tracking, and spatial interaction, the system enables users to engage with digital representations of artworks in an intuitive and emotionally resonant manner. The case study conducted at the Atelier Marco Bagnoli, featuring the digital reproductions of "Colui che sta" and "Come figura d'arciere", provided valuable insights into the system's effectiveness. The results from usability tests highlight the advantages of multimodal interaction compared to conventional point-and-click interfaces. Specifically, users interacting with the multimodal system completed tasks more efficiently, reported higher levels of engagement, and exhibited a deeper emotional connection with the artworks. These outcomes align with prior research emphasizing the importance of sensory and emotional design in enhancing user experiences within cultural and educational contexts. The inclusion of conversational AI was particularly impactful in fostering a personalized dialogue with the user, providing contextually relevant information about the sculptures and adapting to individual queries in real-time. The body tracking and spatial interaction elements further enriched the experience by allowing users to explore the artworks naturally, simulating the sensation of walking around a physical sculpture. This design not only supports intellectual engagement but also creates an immersive and memorable experience, reinforcing the system's educational value. From a technical perspective, the system's architecture proved robust and adaptable. The seamless integration of advanced 3D scanning, real-time rendering, and multimodal interaction highlights the potential of such technologies to bridge the gap between physical and digital art experiences. Moreover, the approach taken to ensure high visual fidelity and smooth interaction, as demonstrated in the rendering pipeline and lighting setup, contributed to the overall success of the installation.

The system addresses key challenges in the field of digital cultural heritage by offering an accessible and engaging platform for exploring artworks. Traditional museum experiences often rely on static displays or audio guides, which can limit user interaction and emotional engagement. In contrast, the proposed system promotes active participation, encouraging users to explore and inquire about the artworks dynamically. This approach has the potential to attract a wider audience, including younger visitors and those less familiar with traditional museum settings.

References

1. Casheekar, A., Lahiri, A., Rath, K., Prabhakar, K.S., Srinivasan, K.: A contemporary review on chatbots, AI-powered virtual conversational agents, ChatGPT: applications, open challenges and future research directions. Comput. Sci. Rev. **52**, 100632 (2024)

2. Schaffer, S., Schwaetzer, E., Ruß, A., Gustke, O.: Chatbot in the museum-a field study of user experience and modality usage. In: Elektronische Sprachsignalverarbeitung 2024, Tagungsband der 35. Konferenz, Regensburg, 6–8 März 2024, pp. 14–21. TUDpress (2024)
3. Wang, J., Huang, Y., Chen, C., Liu, Z., Wang, S., Wang, Q.: Software testing with large language models: survey, landscape, and vision. IEEE Trans. Softw. Eng. (2024)
4. Maniscalco, U., Minutolo, A., Storniolo, P., Esposito, M.: Towards a more anthropomorphic interaction with robots in museum settings: an experimental study. Robot. Auton. Syst. **171**, 104561 (2024)
5. Chen, X., Ibrahim, Z.: A comprehensive study of emotional responses in AI-enhanced interactive installation art. Sustainability **15**(22), 15830 (2023)
6. Ding, Z.: Physical interaction and psychological immersion: analysis of digital technologies in museum exhibitions. Acad. J. Human. Soc. Sci. **6**(6), 9–14 (2023)
7. Goar, V., Yadav, N.S., Yadav, P.S.: Conversational AI for natural language processing: an review of ChatGPT. Int. J. Recent Innov. Trends Comput. Commun. **11**, 109–117 (2023)
8. Liu, Y., et al.: Summary of chatgpt-related research and perspective towards the future of large language models. Meta-Radiology, 100017 (2023)
9. Luther, W., Baloian, N., Biella, D., Sacher, D.: Digital twins and enabling technologies in museums and cultural heritage: an overview. Sensors **23**(3), 1583 (2023)
10. Petrelli, D., Roberts, A.J.: Exploring digital means to engage visitors with Roman culture: virtual reality vs. tangible interaction. ACM J. Comput. Cult. Herit. **16**(4), 1–18 (2023)
11. Ray, P.P.: ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet Things Cyber-Phys. Syst. **3**, 121–154 (2023)
12. Tsepapadakis, M., Gavalas, D.: Are you talking to me? An Audio Augmented Reality conversational guide for cultural heritage. Perv. Mob. Comput. **92**, 101797 (2023)
13. Diederich, S., Brendel, A.B., Morana, S., Kolbe, L.: On the design of and interaction with conversational agents: an organizing and assessing review of human-computer interaction research. J. Assoc. Inf. Syst. **23**(1), 96–138 (2022)
14. Park, D.M., Jeong, S.S., Seo, Y.S.: Systematic review on chatbot techniques and applications. J. Inf. Process. Syst. **18**(1), 26–47 (2022)
15. Singh, S., Beniwal, H.: A survey on near-human conversational agents. J. King Saud Univ.-Comput. Inf. Sci. **34**(10), 8852–8866 (2022)
16. Sundar, A., Heck, L.: Multimodal conversational ai: a survey of datasets and approaches. arXiv preprint arXiv:2205.06907 (2022)
17. Schaffer, S., Ruß, A., Gustke, O.: User experience of a conversational user interface in a museum. In: Brooks, A.L. (ed.) ArtsIT 2022. LNICST, vol. 479, pp. 215–223. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-28993-4_16
18. Allouch, M., Azaria, A., Azoulay, R.: Conversational agents: goals, technologies, vision and challenges. Sensors **21**(24), 8448 (2021)
19. Cao, Y., Han, Z., Kong, R., Zhang, C., Xie, Q.: Technical composition and creation of interactive installation art works under the background of artificial intelligence. Math. Probl. Eng. **2021**(1), 7227416 (2021)
20. Skantze, G.: Turn-taking in conversational systems and human-robot interaction: a review. Comput. Speech Lang. **67**, 101178 (2021)

21. Goncalves Freitas, A.R., et al.: Conversational user interfaces to support astronauts in extraterrestrial habitats. In: Proceedings of the 20th International Conference on Mobile and Ubiquitous Multimedia, pp. 169–178 (2021)
22. Raptis, G.E., Kavvetsos, G., Katsini, C.: MuMIA: multimodal interactions to better understand art contexts. Appl. Sci. **11**(6), 2695 (2021)
23. Savaş, E.B., Verwijmeren, T., van Lier, R.: Aesthetic experience and creativity in interactive art. Art Percept. **9**(2), 167–198 (2021)
24. Szubielska, M., Imbir, K., Szymańska, A.: The influence of the physical context and knowledge of artworks on the aesthetic experience of interactive installations. Curr. Psychol. **40**(8), 3702–3715 (2021)
25. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv. Neural. Inf. Process. Syst. **33**, 9459–9474 (2020)
26. Rajcic, N., McCormack, J.: Mirror ritual: human-machine co-construction of emotion. In: Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction, pp. 697–702 (2020)
27. Candello, H., Barth, F., Carvalho, E., Cotia, R.: Understanding how visitors interact with voice-based conversational systems. In: Marcus, A., Rosenzweig, E. (eds.) HCII 2020. LNCS, vol. 12201, pp. 40–55. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49760-6_3
28. Crovari, P., Pidó, S., Garzotto, F., Ceri, S.: Show, don't tell. reflections on the design of multi-modal conversational interfaces. In: Følstad, A., et al. (eds.) CONVERSATIONS 2020. LNCS, vol. 12604, pp. 64–77. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-68288-0_5
29. Saidi, H., Serrano, M., Irani, P., Hurter, C., Dubois, E.: On-body tangible interaction: using the body to support tangible manipulations for immersive environments. In: Lamas, D., Loizides, F., Nacke, L., Petrie, H., Winckler, M., Zaphiris, P. (eds.) INTERACT 2019. LNCS, vol. 11749, pp. 471–492. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-29390-1_26
30. Ferracani, A., Faustino, M., Giannini, G. X., Landucci, L., Del Bimbo, A.: Natural experiences in museums through virtual reality and voice commands. In: Proceedings of the 25th ACM International Conference on Multimedia, pp. 1233–1234 (2017)
31. Ferracani, A., Pezzatini, D., Del Bimbo, A.: A natural and immersive virtual interface for the surgical safety checklist training. In: Proceedings of the 2014 ACM International Workshop on Serious Games, pp. 27–32 (2014)
32. Weizenbaum, J.: ELIZA-a computer program for the study of natural language communication between man and machine. Commun. ACM **9**(1), 36–45 (1966)