ASSIST: A Multi-Agentic Framework for Human Computer Interaction in Cultural Heritage settings

Samuel Ramos-Varela[1 [0009−0000−8458−6202]], Anmol Guragain[1 [0009−0009−8491−8663]], Jaime Bellver-Soler[1 [0009−0006−7973−4913]], David Aragon Diaz[2], Long Lin[1], and Luis Fernando D'Haro[1 [0000−0002−3411−7384]]

Speech Technology and Machine Learning Group,
E.T.S.I de Telecomunicación, Universidad Politécnica de Madrid, Spain
{s.rvarela, anmol.g, jaime.bellver, long.lin, luisfernando.dharo}@upm.es, daviarag@ucm.es

Abstract. This paper presents ASSIST-AI, a novel multi-modal multi-agentic framework designed to enhance human-computer interaction in cultural heritage environments through advanced AI technologies. Our system integrates computer vision, natural language processing, and speech technologies to create context-aware conversational agents for museum settings. The framework combines Retrieval-Augmented Generation (RAG) with automatic Point of Interest (POI) detection, personalized user profiling, and real-time multi-modal interaction capabilities. We demonstrate significant improvements in user engagement through adaptive personalization mechanisms that leverage attention schema theory and contextual awareness. Evaluation with 30 participants across major Spanish museums shows enhanced visitor experience and knowledge retention. The system achieves 92% accuracy in artwork identification, sub-second response times for multi-modal queries, and supports real-time interaction in Spanish and English with adaptive complexity based on user expertise levels.

Keywords: Multi-Modal AI · Multi-Agentic Systems · Cultural Heritage · Human-Computer Interaction · Vision-Language Models · Conversational AI

1 Introduction

Cultural heritage sites such as museums and historical monuments have become fertile testbeds for HCI innovations that blend physical and digital interactions through multimodal interfaces [6]. Multimodal and agent-based interaction frameworks offer advanced information services that adapt to user context and preferences, enabling engaging experiences within these public spaces [1]. Research has explored mobile and wearable technologies, such as multimodal mobile museum guides, to provide inclusive and interactive tours for diverse audiences [21]. Context-aware systems leveraging location-sensing agents have personalized visitor experiences by dynamically deploying relevant content based

on visitor position [9]. The growing field of image-based artwork localization (IAL) and Artwork Point-of-Interest Detection (APoID) demonstrates the feasibility of non-intrusive indoor localization techniques for guiding visitors to exhibits [8]. Additionally, speech interfaces have emerged as a hands-free modality, allowing visitors to navigate virtual or physical museum spaces through spoken commands, including voice-guided wayfinding and group-synchronized tours [16, 18, 19].

Evolving visitor expectations for personalized, context-sensitive interactions underscore the need to design systems that fulfill individual motivations and enhance engagement [12, 23]. Applying self-determination theory to museum interfaces has demonstrated that satisfying visitor needs for autonomy, competence, and relatedness can significantly boost motivation and sustained interaction [15]. Augmented reality guided tours have been shown to mediate between everyday visitor engagement and curated exhibition knowledge, facilitating deeper learning experiences [20]. Moreover, context-aware museum guide agents have been developed to address the challenge of delivering timely, relevant information, thus aligning technological objectives with visitor goals [10].

In this work, we present a novel HCI framework—VisualRAG + PoI + Speech—that integrates visual retrieval-augmented generation for artwork identification, image-based POI detection and a conversational speech interface for hands-free, natural interaction. By unifying these modalities, our approach creates a novel Ill seamless hybrid experience in cultural heritage settings, enhancing accessibility, personalization, and engagement in museum visits [8] [7].

- **Visual Retrieval Pipeline:** A three-stage image processing workflow for artwork recognition, including object detection, semantic embedding retrieval, and local feature reranking.
- **Automated POI Annotation:** A cross-modal pipeline that identifies and annotates Points of Interest (POIs) within artworks leveraging both pre-existing audio guide content and vision-language processing.
- **Personalized Multimodal Interaction:** A dynamic prompt engineering framework that tailors responses based on user profiles and interaction context, enabling chat, voice, and image-based input.
- **Ethical Guardrails and Safety Filters:** Mechanisms to ensure content appropriateness, cultural alignment, and responsible AI behavior in public deployment.

Outline of the Paper In Section 1 we introduce the motivation and scope of our work, situating the VisualRAG + PoI + Speech framework within the broader field of HCI in cultural heritage. Section 2 reviews prior research on multimodal museum interfaces, RAG in visual applications, content understanding and generation for heritage collections, and ethical considerations in AI-mediated museum systems. In Section 3, we present our proposed framework, detailing the multi-agentic flow and the individual modules for VisualRAG, Point-of-Interest Detection, the Speech Interface, and the Personalization and Ethics mechanisms. Section 4 describes our experimental evaluation plan and metrics, performance

analysis, user experience findings, and expert feedback with ethical reflections. Finally, Section 5 concludes with a summary of contributions, discusses limitations, and outlines directions for future work.

2 Related Work

2.1 Multimodality in HCI and Museum Settings

Recent advancements in multimodal HCI for cultural heritage emphasize the integration of visual, auditory, haptic, and gestural modalities to enrich visitor experiences, supporting naturalistic exploration and multisensory learning [**?**, 2]. Mobile and wearable museum guides now leverage voice, touch, and group-synchronized audio to facilitate inclusive tours, accommodating diverse visitor needs such as visually impaired users and group-based experiences [16, 18]. Virtual tour systems like VirtuWander demonstrate how coupling gesture, voice, and language-model-driven conversational guidance can personalize navigation and engagement in both physical and simulated museum contexts [22]. Empirical evaluations indicate that such multimodal approaches significantly improve visitor satisfaction, learning outcomes, and emotional engagement compared to traditional single-modality guides [10].

More broadly, advances in multimodal audio–language models for emotion and intent recognition offer promising avenues for museum guides to respond empathetically to visitor affective states. For example, Bellver et al. [3] present a Multimodal Audio–Language Model that fuses audio encoding and language understanding to categorize speech emotions, illustrating how such models can enrich interactive systems by tailoring responses not only to content but also to visitor mood and engagement levels. Similarly, gamified participatory sensing approaches influence tourist behavior and satisfaction through adaptive incentive mechanisms and real-time context collection [13], and on-site trip planning support systems leverage dynamic information on tourism spots to curate personalized recommendations in situ [11].

2.2 RAG in Visual Applications and Cultural Heritage

Retrieval-Augmented Generation (RAG) has been extended to vision tasks by coupling image embeddings with external knowledge sources, improving contextual understanding and generation in domains such as medical imaging and multimodal question answering [24]. Systems like FolkRAG apply RAG to cultural heritage archives by integrating domain-specific retrieval pipelines that fetch metadata and historical context, grounding model outputs in authoritative records and enabling intuitive archival exploration [14].

2.3 Content Understanding and Generation in Cultural Heritage

Visual Question Answering (VQA) frameworks allow visitors to pose natural language questions about exhibits and receive contextually accurate responses,

bridging the gap between raw imagery and interpretive metadata in museum settings [4]. Recent work investigates using GPT-3 to generate artwork descriptions at runtime, avoiding costly manual annotation and enabling on-the-fly VQA and captioning for cultural heritage collections [5]. Recent work [17] explores generation of QA pairs from visual information and curated text context in art and museum QA system.

3 Proposed Framework

Our unified framework presents an advanced human–computer interface tailored for cultural heritage sites, combining VisualRAG, Point-of-Interest Detection, Speech Interaction. We design an integrated pipeline that fluidly transitions between visual, spatial, and conversational channels. This approach supports visitors as active co-creators of their museum experiences, allowing the system to respond naturally different possible inputs: images, location context, and spoken queries without forcing rigid interaction patterns.

3.1 Multi-Agentic Flow

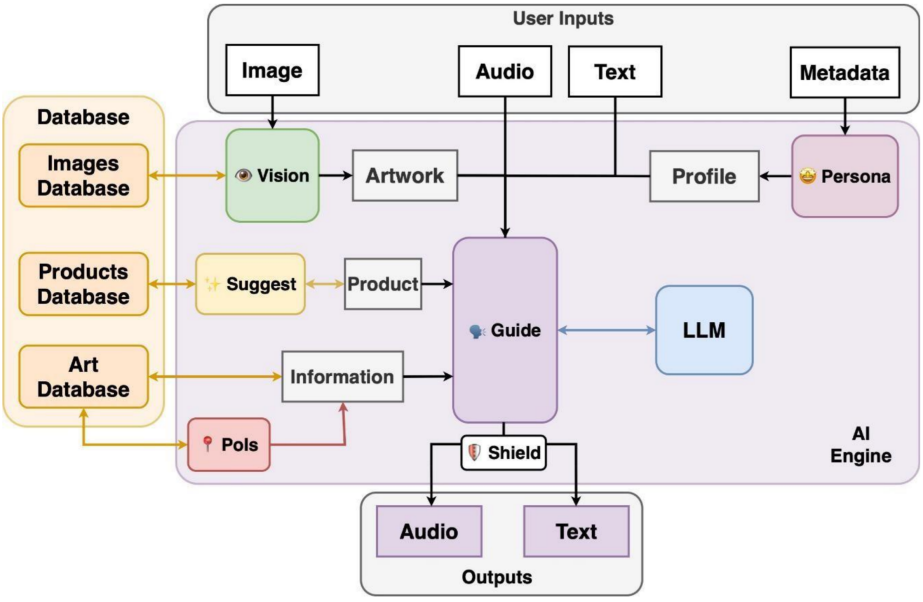

Fig. 1: Multi-agentic architecture of the VisualRAG + POI + Speech framework

In our multi-agentic design, each module (VisualRAG, POI Detection, Speech Interface, Persona Manager, and Content Databases) operates as a semi-autonomous

agent that communicates via well-defined message channels (see Figure 1). The Vision agent ingests visitor images identifies artworks; the POI agent signals important keypoints within exhibit selections; the Guide agent (backed by the LLM) subscribes to the user's and other module's feeds to generate narrative responses; and the Speech agent mediates natural language queries and vocalizes replies. A lightweight "Shield" agent oversees content safety and domain relevance, ensuring that every response conforms to museum guidelines before providing it to the visitor as text or audio. By decoupling responsibilities across agents, the framework achieves both robustness and extensibility: new capabilities (e.g. gesture recognition or AR overlays) can be added as separate agents without reworking the entire pipeline.

Modeling a museum visit as a collaboration among specialized agents reflects the inherently social, exploratory nature of cultural heritage experiences. Visitors encounter artworks, move through spaces, and ask questions in unpredictable sequences; a monolithic system would struggle to adapt fluidly to these dynamics. In contrast, the multi-agentic flow enables context-aware coordination: as a visitor gazes at a sculpture, the user agent triggers a VisualRAG request, the Guide agent retrieves and fuses relevant content, and the Speech agent offers it in the visitor's preferred modality (Text or Speech). This agent-based orchestration not only mirrors human guide teamwork but also provides a methodological blueprint for building scalable, adaptable HCI systems in any complex, multimodal environment.

3.2 VisualRAG

VisualRAG extends retrieval-augmented generation to images captured in situ by museum visitors. When a visitor frames an artwork or artifact, the system extracts a compact visual representation and uses it to retrieve relevant textual passages from a curated heritage knowledge base. These passages are then fused with the visual features to generate a concise narrative tailored to the specific object and visitor context.

VisualRAG is realized as a three-stage pipeline that transforms a visitor's photo of an artwork into a rich, context-aware narrative. In the first *Detection* stage, the system identifies and localizes each distinct object or artwork within the input image, cropping individual elements for subsequent processing. Next, during the *Retrieval* stage, the cropped visuals are matched against a curated vector database of artwork embeddings to fetch the top-K candidate records along with their associated textual descriptions and historical context. Finally, in the *Reranking* stage, lightweight local feature comparisons are applied to refine the initial matches, reordering the candidate list based on geometric and semantic alignment before passing the enriched context to the language generation component. This modular design allows each step to be optimized independently, ensuring both responsiveness and accuracy in real-time museum settings.

In practice, this spatial awareness enables hands-free guidance through large or complex exhibitions. Visitors need not navigate menus or scan QR codes; instead, the system gently suggests content as they approach each artwork. Such

ambient assistance encourages serendipitous discovery and supports varied exploration styles, from goal-oriented tours to free-form wandering.

3.3 Point-of-Interest Detection

The POI system begins by leveraging vision-language models to detect predefined categories. such as people, objects, places, and landscapes, by analyzing both image content and associated textual sources (e.g., audio-guide transcripts). Advanced AI then generates detailed captions for each detected element, building a structured inventory of visual features. A subsequent multimodal refinement step uses LLMs to align these visual detections with curated museum text, removing spurious or non-significant annotations. Finally, precise bounding box coordinates for each PoI are stored in a reusable graph database, allowing new exhibitions or temporary collections to be integrated seamlessly without altering the core schema and supporting flexible content updates and curator oversight.

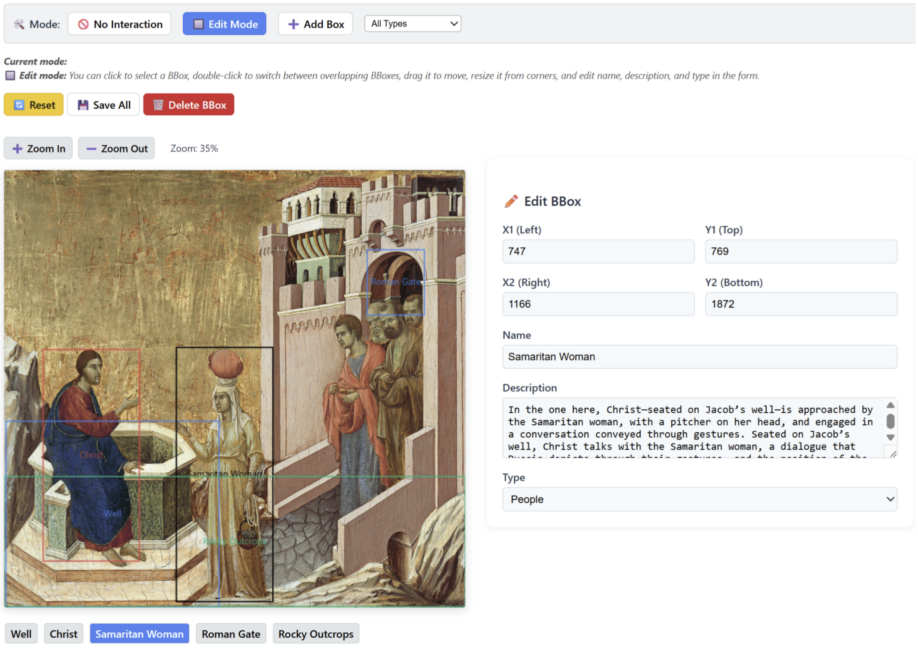

Fig. 2: Bounding Box CMS interface for Points Of Interest modification.

The Content Management System (CMS) for Points of Interest serves as the curator's primary interface for defining, organizing, and enriching exhibit metadata that underlies the POI Detection pipeline (see 2). Through a web-based dashboard, curators can create POI records by uploading high-resolution images, automatically and manually annotating key visual elements, and linking

each record to textual descriptions, audio clips, and related multimedia assets. A flexible schema allows custom fields (such as artist biography, provenance notes, or thematic tags) and points of interest to be added without altering the core database structure, enabling rapid adaptation for special exhibitions or rotating galleries. As new exhibits are introduced, the CMS automatically curates content, keeping it reviewable and editable.

3.4 Multimodal Interface

Our Speech Interface unifies ASR and TTS into a conversational layer that respects individual interaction preferences. Visitors can speak naturally (asking "What is this?" or "Tell me more about the artist") and the system will ground responses in both the current visual context and the exhibit's metadata. Dialogue state persists across turns, allowing follow-up questions ("And who inspired her?") without repeated qualifiers.

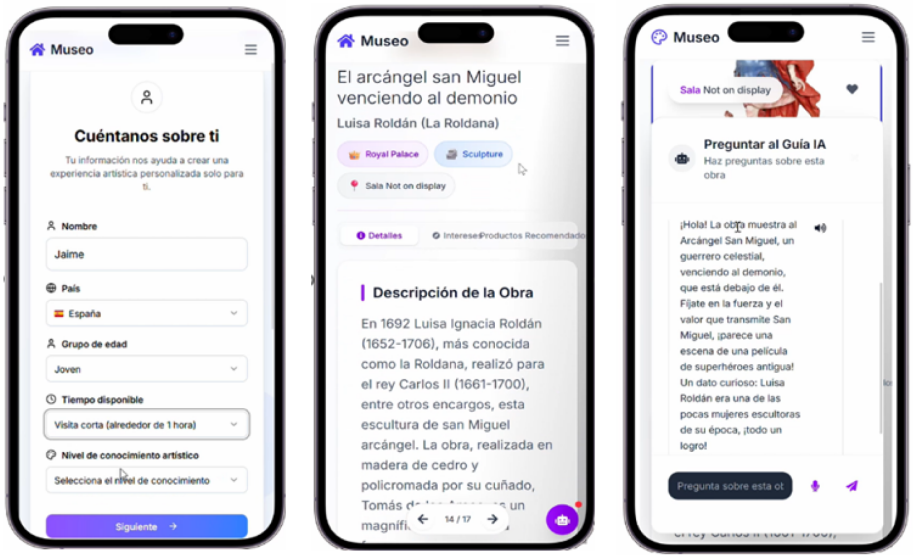

Fig. 3: Screenshots of the designed UX. On the left: the onboarding form. On the center: Factual information about the artwork. On the right: Interactive information with the AI assistant.

By offering speech as an alternative to text, the interface increases accessibility for users with limited difficulties to type or visual impairments. Moreover, visitors may choose to combine modalities communicate with text, then ask a question or switch entirely to voice, reinforcing the system's adaptability to personal comfort and situational constraints.

3.5 Personalization

Personalization weaves through every stage of the interaction, shaping content and modality to individual visitor profiles and real-time behavior. Initial preferences, such as language, depth of detail, or interest areas (e.g., art history versus conservation), are set at the start, as well as visit duration (to dynamically accommodate the answers length to the time remaining). As the visit unfolds, the system dynamically adjusts narrative complexity, recommending simpler overviews for casual browsers or deep-dive lectures for enthusiasts.

3.6 Ethics

This research prototype has been guided by key ethical principles, including human agency, privacy, fairness, and transparency. While not yet deployed with real users, the system incorporates early safeguards and design choices aligned with responsible AI development.

Human oversight is enabled through a content management interface allowing curators to review and edit automatically generated outputs. The user interface also supports basic control over interaction flow, preserving user autonomy. In addition, visitors may opt out of the AI-based system entirely and access only the standard textual information available on the museum website.

Privacy is addressed by collecting only minimal, non-identifiable data (e.g., age range, country, art knowledge). Uploaded images are processed in memory without storage, and all communications are encrypted. As this prototype is used internally, formal data audits are not yet implemented but are planned for later stages.

Fairness and inclusivity were considered when designing personalization features for diverse age groups and knowledge levels. Basic content filtering is employed to prevent harmful outputs, though no formal bias audits have yet been conducted.

Transparency and accountability are supported through clear usage disclaimers and internal logging. The system architecture and decision logic are documented, with future versions expected to include explainability tools and formal governance mechanisms.

This approach reflects an "ethics by design" mindset, acknowledging current limitations while laying the groundwork for responsible future deployment in public cultural settings.

4 Experimental Evaluation

4.1 Evaluation Methodology

The evaluation of took place in a focus group on June 12, 2025, with 30 participants, including 12 expert panelists (directors, digital-strategy leaders, and innovation experts). Table 1 shows the diverse profiles. There was representation from several Spanish cultural heritage institutions of various domains (mainly

art museums). The implemented features were showcased to the participants through a presentation and videos of the system's prototype, and finally provided their feedback on-site or after the meeting through a questionnaire.

Table 1: Focus Group Participant Distribution

Institution Type	**Count**	**Role Category**
Art Museums	8	Directors, Curators
Cultural Heritage Sites	6	Digital Strategy Leaders
Religious Monuments	4	Innovation Experts
Sports/Entertainment	2	Technology Managers
Project Experts	10	Technical Evaluators
Total Participants	**30**	**Mixed Expertise**

Through a focus group interview, the evaluation focused on the system's ability to demonstrate the flexibility and potential of our technologies in cultural settings, serving as a foundation for future development within the domain.

4.2 Dataset and Technical Implementation

The evaluation dataset comprises 1,603 artworks and artists automatically crawled from various museum websites via a UPM-developed pipeline, with a curated subset of 20 emblematic works (spanning Renaissance to contemporary periods) selected to minimize LLM hallucinations, ensure the graph database was manageable and reviewable, and ease up testing of the different functionalities developed. Table 2 showcases this test set.

Table 2: Evaluation Dataset Characteristics

Museum	**Artworks Selected**	**Art Periods Covered**
Museum1	10	Renaissance to Modern
Museum2	10	Medieval to Contemporary
Total Dataset	**20**	**Multi-period**
Graph Nodes	3,418	Comprehensive
Graph Relationships	4,913	Interconnected

4.3 System Performance Analysis

The technical performance evaluation measured response time, accuracy, and scalability under varying loads: the three-stage Visual RAG pipeline (detection,

retrieval, reranking) consistently met real-time targets across diverse museum environments and lighting conditions while delivering high artwork identification accuracy. The POI detection module reduced manual curator review time significantly, all without sacrificing precision across artworks from multiple periods and styles.

Table 3: System Performance Metrics Across Core Functions

Function	**Accuracy**	**Processing Time**	**Efficiency Gain**
VisualRAG (20 image DB)	-	<1s	Real-time (>30fps)
POI Detection	-	Variable	Automation
Personalization	-	<1s	Adaptive

The speech technology components demonstrated effective performance in museum environments. The TTS synthesis maintained high audio quality at 16 kHz and 32 kbps, with user preference options for voice gender selection across supported languages. The other developed components automate processes traditionally costly.

4.4 User Experience

The personalization framework dynamically integrates visitor characteristics (age, visit duration, and art proficiency). Modeling its system prompts to tailor both language complexity (from storytelling for children to technical analysis for experts) and response length (concise facts for quick visits versus detailed explanations for longer stays), the system automatically adjusts terminology, explanation depth, and cultural references to match each user's expertise and preferences.

4.5 Expert Feedback and Ethical Considerations

The focus group of 30 museum professionals (including directors, digital strategy leaders, and innovation experts) highlighted the potential to boost visitor engagement while preserving educational integrity and cultural authenticity. Participants praised the automated POI detection for drastically reducing manual curation effort yet maintaining curator oversight, and valued the system's ability to handle new or temporary collections with minimal intervention. They also commended the multimodal interfaces (speech recognition and TTS) for enhancing accessibility and maintaining conversational context across modes, catering to diverse visitor needs.

Ethical and safety features were deemed essential for public deployment: the Shield Module's content moderation, visible disclaimers, and minimal data collection policies aligned with institutional requirements for responsible AI in

cultural heritage. This feedback underlines the need for robust moderation and privacy safeguards to deliver both innovative and trustworthy museum experiences.

5 Conclusion and Future Work

The system's key innovations include context-aware personalization, which dynamically adapts responses based on validated user profiles, and automatic detection of Points of Interest (PoIs) in artworks, achieved through the sophisticated combination of vision–language models and curated museum content. Furthermore, the system incorporates multimodal Retrieval-Augmented Generation (RAG)-based image recognition, providing quick contextual information. The multimodal AI assistant interaction, supporting speech, text, and image-based inputs and outputs, significantly enhances accessibility and engagement for diverse audiences.

The comprehensive development effort undertaken by the UPM team involved several critical tasks. This included the meticulous pre-processing of content data, featuring automatic transcription of audio guides using high-performing ASR models. The system also performs automatic extraction of visual and semantic descriptions from artwork images through a multi-stage pipeline powered by advanced vision–language models. Additionally, the project involved the development of a synthetic dataset for in-context learning, generating tailored Q&A dialogues for various user profiles, and the implementation of a custom Content Management System (CMS) to allow curators to review and edit PoIs. The design and deployment of a responsive UX frontend further enabled intuitive user interaction with both the backend system and the physical artworks.

From an ethical and responsible AI perspective, the system incorporates safeguards inspired by the EIC ASTOUND Pathfinder project's work (101071191) on transparency, safety, and user trust. This demonstrates a foundational design principle, where ethical considerations were integrated from the outset rather than treated as an afterthought. These safeguards include content moderation and filtering through LlamaGuard v3.0, visible disclaimer alerts to inform users they are interacting with an AI system, and minimal data collection (e.g., age groups and general preferences instead of personal identifiers). Furthermore, HTTPS-encrypted communication, local logging mechanisms, cautious and reduced usage of images (artworks only, no personal photos), and manual validation of responses in potentially sensitive scenarios were incorporated to ensure privacy and reliability. This proactive integration of ethical considerations sets an example for "ethics by design," advocating for responsible AI awareness in the design of this prototype and laying the groundwork for its long-term deployment in sensitive public domains like museums.

Finally, the presentation of the demo during the Focus Group on June 12, 2025, and the valuable feedback received from leading Spanish museum representatives affirm the system's potential to revolutionize visitor experiences. This feedback will serve as a crucial guide for upcoming enhancements.

5.1 Upcoming Features and Improvements

Throughout the execution of the ASSIST project, the UPM team, in collaboration with GVAM and Indeep AI, explored numerous innovative features researched within ASTOUND that held potential to further enhance the demo. While time constraints limited the full implementation or testing of all anticipated developments for the initial demo version, the feedback from the Focus Group session highlighted several high-priority functionalities for future exploration and refinement.

Key planned features and improvements for implementation and evaluation in the upcoming months as part of the continued ASTOUND project efforts include:

- **Advanced Attention Analysis and User Modeling:** The project aims to incorporate additional audio-derived features to predict and analyze user interest and contextual awareness.
- **Dynamic User Modeling:** Integrating sophisticated reasoning modules with interaction-derived data to infer user states—such as attention and engagement, and adapt system behavior accordingly.[1]
- **Refinement of Fine-Tuned Language Models:** A synthetic Q&A and dialogue dataset was created to fine-tune a pre-trained model on museum-specific content, aiming to reduce hallucinations. While this process provided valuable insights, hyperparameter optimization proved more time-consuming than anticipated, and the fine-tuned model's performance did not consistently exceed that of proprietary models. Future work will explore minimizing reliance on external LLMs in favor of local models to enhance privacy.
- **Platform Scalability and Robustness Improvements:** Enhancements will focus on supporting concurrent access by multiple users, robust session and turn management, and the exploration of invisible watermarking techniques to detect and trace generated responses—bolstering transparency and accountability.

6 acknowledgments

This work is supported by the European Commission through Projects ASTOUND (101071191 — HORIZON-EIC-2021- PATHFINDERCHALLENGES-01) and ASSIST (101201944 - HORIZON-EIC-2024-BOOSTER-IBA-01). In addition, it is supported by project BEWORD (PID2021- 126061OB-C43) funded by MCIN/AEI/10.1303- 9/501100011033 and, as appropriate, by "ERDF A way of making Europe", by the "European Union".

[1] For this prototype, uploaded images or audio are processed in-memory and not stored. Logs are stored with a hash identifier, and only non-sensitive personal information is collected. Although this is a preliminary prototype, the system is intended to comply with GDPR and includes opt-in consent and disclaimers.

References

1. Al Shammari, T., Eiman, D.: My heritage companion: An ai-driven mobile experience for visual storytelling. In: Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization (2025)
2. Azofeifa, J.D., Noguez, J., Ruiz, S., Molina-Espinosa, J.M., Magana, A.J., Benes, B.: Systematic review of multimodal human–computer interaction. Informatics **9**(1), 13 (2022). https://doi.org/10.3390/informatics9010013
3. Bellver, J., Martín-Fernández, I., Bravo-Pacheco, J.M., Esteban, S., Fernández-Martínez, F., D'Haro, L.F.: Multimodal audio-language model for speech emotion recognition. In: Proceedings of The Speaker and Language Recognition Workshop (Odyssey 2024). pp. 288–295 (2024). https://doi.org/10.21437/odyssey.2024-41
4. Bongini, P., Becattini, F., Bagdanov, A.D., Del Bimbo, A.: Visual question answering for cultural heritage. In: IOP Conference Series: Materials Science and Engineering. vol. 949, p. 012074. IOP Publishing (2020)
5. Bongini, P., Becattini, F., Del Bimbo, A.: Is gpt-3 all you need for visual question answering in cultural heritage? In: Computer Vision – ECCV 2022 Workshops, Lecture Notes in Computer Science. vol. 13801, pp. 268–281 (2023). https://doi.org/10.1007/978-3-031-25056-9_18
6. Caramiaux, B.: Ai with museums and cultural heritage. In: AI in Museums, pp. 117–130. Springer (2023)
7. Clark, L., Doyle, P., Garaialde, D., Schloegl, S.: The state of speech in hci: Trends, themes and challenges. arXiv preprint arXiv:1810.06828 (2018)
8. Egbariya, B., Dror, R., Kuflik, T., Shimshoni, I.: Image-based poi identification for mobile museum guides: Design, implementation, and user evaluation. Heritage **8**(7) (2025). https://doi.org/10.3390/heritage8070266, `https://www.mdpi.com/2571-9408/8/7/266`
9. Ferrara, C., Cerquetti, M.: Towards user-centred and context-open museums: A review of mobile apps adopted by the italian national museum with special autonomy status. Journal of Cultural Heritage Management and Sustainable Development (2025)
10. Ferrato, A.: Integrating indoor positioning, recommendation, and group synchronization in intelligent museum guides. In: Proceedings of the 2023 ACM Conference on Ubiquitous Computing. pp. 215–224 (2023). https://doi.org/10.1145/3699682.3727572
11. Hidaka, M., Kanaya, Y., Kawanaka, S., Matsuda, Y., Nakamura, Y., Suwa, H., Fujimoto, M., Arakawa, Y., Yasumoto, K.: On-site trip planning support system based on dynamic information on tourism spots. Smart Cities **3**(2), 212–231 (2020). https://doi.org/10.3390/smartcities3020013
12. Ivanov, R.: Advanced visitor profiling for personalized museum experiences using telemetry-driven smart badges. Electronics **13**(20), 3977 (2024). https://doi.org/10.3390/electronics13203977
13. Kawanaka, S., Matsuda, Y., Suwa, H., Fujimoto, M., Arakawa, Y., Yasumoto, K.: Gamified participatory sensing in tourism: An experimental study of the effects on tourist behavior and satisfaction. Smart Cities **3**(3), 736–757 (2020). https://doi.org/10.3390/smartcities3030037
14. Kelly, P., Schild, J., Jafari, A.: Folkrag: A retrieval-augmented generation system for cultural heritage materials. Neural Computing and Applications (2025). https://doi.org/10.1007/s00521-025-11455-4

15. Lushnikova, A., Morse, C., Doublet, S., Koenig, V., Bongard-Blanchy, K.: Self-determination theory applied to museum website experiences: Fulfill visitor needs, increase motivation, and promote engagement. In: Proceedings of the European Conference on Cognitive Ergonomics 2023. ECCE '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3605655.3605658, https://doi.org/10.1145/3605655.3605658
16. Mohlin, F.: Implementation and evaluation of a mobile tour guide with group synchronized audio: Using web technologies, qr codes, and speech synthesis. In: Proceedings of the 2024 International Conference on Mobile and Ubiquitous Multimedia (2024)
17. Ramos-Varela, S., Bellver-Soler, J., Estecha-Garitagoitia, M., D'Haro, L.F.: Context or retrieval? evaluating rag methods for art and museum qa system. In: Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology. pp. 129–136. Association for Computational Linguistics, Bilbao, Spain (May 2025). https://doi.org/10.18653/v1/2025.iwsds-1.10, https://aclanthology.org/2025.iwsds-1.10/
18. Suikkila, J.: Museum wayfinding with ai-spoken voice guidance. In: Proceedings of the 2025 International Conference on Intelligent User Interfaces (2025)
19. Trichopoulos, G., Konstantakis, M., Caridakis, G., Katifori, A., Koukouli, M.: Crafting a museum guide using chatgpt4. Big Data and Cognitive Computing **7**(3) (2023). https://doi.org/10.3390/bdcc7030148, https://www.mdpi.com/2504-2289/7/3/148
20. Tung-Ju Hsieh, Yao-Hua Su, L.S.L.: Augmented reality art museum mobile guide for enhancing user experience. IEEE Computer Graphics & Applications **45**(1), 10–20 (2025). https://doi.org/10.1109/MCG.2025.3529981
21. Varga, M.N., Bazazian, D.: Interactive heritage site mobile application on artworks. In: Advances in Representation: New AI- and XR-Driven Transdisciplinarity, pp. 125–139. Springer Nature Switzerland (2024)
22. Wang, Z., Yuan, L.P., Wang, L., Jiang, B., Zeng, W.: Virtuwander: Enhancing multi-modal interaction for virtual tour guidance through large language models. arXiv preprint arXiv:2401.11923 (2024), https://arxiv.org/abs/2401.11923
23. Yanbo Li, Yuanyuan Ma1, F.L.H.B.: Empowering museum visitor experiences through ai-driven smart guide systems: A model for intelligent museum services. GAS Journal of Engineering and Technology (GASJET) (2025)
24. Zheng, X., Weng, Z., Lyu, Y., Jiang, L., Xue, H., Ren, B., Paudel, D., Sebe, N., Van Gool, L., Hu, X.: Retrieval augmented generation and understanding in vision: A survey and new outlook. arXiv preprint arXiv:2503.18016 (2025)