A Personalized, Multimodal AI Assistant for Enhancing Museum Visitor Experience

Anmol Guragain[1], Jaime Bellver Soler[1], Samuel Ramos Varela[1], Long Lin[1], David Aragón Diaz[1] and Luis Fernando D'Haro[1,*]

[1]Speech Technology and Machine Learning Group (THAU)- ETSI de Telecomunicación. Information Processing and Telecomunications Center (IPTC). Universidad Politécnica de Madrid. Av. Complutense 30, 28040. Madrid, Spain.

Abstract

This paper presents a functional demonstration of an AI assistant developed under the EIC funded ASSIST (101201944) and ASTOUND (101071191) projects, focused on enhancing the museum visitor experience through multimodal, personalized, and context-aware interactions. Taking into account the cognitive insights researched during the ASTOUND project, the system integrates multimodal vision-language models, a retrieval-augmented generation pipeline, and real-time ASR/TTS services. The system incorporates a Neo4j graph-based infrastructure for storing curated data from two large museums in Spain (Thyssen-Bornemizsa and Colecciones Reales). The modular system allows personalized recommendations, visual artwork recognition, and user profiling to enrich user engagement and personalization. The demo chatbot is intended for realistic world settings. Feedback from expert stakeholders from Spanish cultural institutions, highlights the potential of this system for next-generation museum experiences.

Keywords

Chatbot, personalization, cultural heritage, multimodal interaction, museum AI, Attention Schema Theory.

1. Introduction

The cultural heritage sector, particularly museums, is undergoing a significant transformation driven by the integration of digital technologies. This evolution moves beyond traditional static displays to embrace interactive and personalized visitor engagement. The burgeoning capabilities of Artificial Intelligence (AI) and Large Language Models (LLMs) offer unprecedented opportunities to create intelligent guides that can dynamically respond to diverse visitor needs, preferences, and learning styles [1, 2]. This shift aims to make cultural experiences more accessible, immersive, and deeply informative, effectively complementing the physical museum visit with rich digital layers. The application of LLMs, for instance, is already revolutionizing customer experience in the broader tourism industry by enabling nuanced understanding of feedback, generating personalized content, and facilitating multilingual communication [3].

This paper details the architecture and capabilities of our AST-inspired AI assistant, a functional demonstrator developed under the ASSIST project (GA Number: 101201944). This initiative builds directly upon the foundational research and expertise accumulated within the ASTOUND project (GA Number: 101071191), aiming to translate theoretical advancements into practical applications for the tourism sector. The primary objective of this paper is to showcase a proof-of-concept for ASTOUND technology within cultural settings, specifically a museum-focused AI assistant designed to engage visitors in natural, conversational interactions about selected artworks. A significant contribution of this work lies in its approach to augmenting the on-site visitor experience. While traditional virtual museums have focused on dematerializing the physical visit by offering remote access or digital replicas , our AI assistant follows a different paradigm. By integrating real-time visual recognition capabilities, such as identifying artworks from user-uploaded photos, and providing interactive Points

2025 Workshop on Awareness in Learning Agents (ALA), 25-30 October 2025, Bologna, Italy

*Corresponding author.

anmol.g@upm.es (A. Guragain); jaime.bellver@upm.es (J. B. Soler); s.rvarela@upm.es (S. R. Varela); long.lin@upm.es (L. Lin); s.daviarag@ucm.es (D. A. Diaz); luisfernando.dharo@upm.es (L. F. D'Haro)

0009-0006-7973-4913 (J. B. Soler); 0009-0000-8458-6202 (S. R. Varela); 0000-0002-3411-7384 (L. F. D'Haro)

of Interest (PoIs) that overlay information directly onto physical artwork images, the system creates a seamless interactive layer over the physical visit. This represents a progression beyond merely digitizing collections; it re-integrates the digital interaction within the physical space, providing a "hybrid" museum experience. This approach leverages the strengths of both physical presence and digital augmentation, thus increasing visitor engagement and learning through immediate, context-aware responses directly linked to exhibits.

The paper is organized as follows: in section 2 we present relevant background and related work. Section 3 describes the architecture and implementation; then in section 4 we present the results, and finally section 5 shows the conclusions and future work.

2. Background and Related Work

2.1. AI and Large Language Models in Cultural Heritage and Tourism

The rapid advancements in AI, particularly Large Language Models (LLMs), are also transforming the tourism and cultural heritage sectors. These models enable a better understanding of user queries, generate highly personalized content, and facilitate seamless multilingual communication [3, 4, 5, 6, 7]. Application of LLMs in tourism covers the more traditional sentiment analysis and automated customer service [8], evolving to more sophisticated tasks such as dynamic itinerary generation and the creation of culturally rich site descriptions [9, 10].

On the other hand, studies involving autonomous conversational robots in museum settings indicate a strong visitor interest in information retrieval and multilingual interaction [11, 12]. For instance, the android robot Andrea, deployed in a German museum [4], received numerous requests for museum-related information and language conversion, with visitors generally finding such robots acceptable as guides. Key desired improvements included broader language support and faster response times, highlighting the practical needs for effective conversational agents in these contexts.

Finally, LLMs are being used also for creating synthetic data that can be used to train or fine-tune language models for adaptation purposes to specific museums or types of artworks [13, 14].

2.2. Retrieval-Augmented Generation (RAG) Systems and Their Evolution

Retrieval-Augmented Generation (RAG) and Knowledge Graph systems have emerged as powerful paradigms to enhance LLMs by grounding their generative capabilities in external, up-to-date knowledge bases. Both approaches effectively mitigate common LLM limitations such as factual inconsistency and hallucination, where models generate inaccurate or nonsensical information [15, 16, 17].

Conventional RAG typically follows a static retrieve-then-generate pipeline, where information is retrieved once based on the initial query before generation begins. However, this approach can be suboptimal for complex tasks requiring multi-hop reasoning or adaptive information access [18]. Recent advancements have led to the development of Dynamic RAG, which adaptively determines when and what information to retrieve during the LLM's generation process, allowing for real-time adaptation to evolving information needs [19]. Another emerging direction is Parametric RAG, which focuses on injecting retrieved knowledge directly at the parameter level of the LLM for enhanced efficiency and deeper integration [20].

Beyond text-based applications, Multimodal RAG extends this concept to incorporate diverse data types, including images, audio, and video. This leverages multimodal LLMs (MLLMs) to process information from various modalities, leading to richer contextual understanding and more comprehensive responses [21, 22].

Our proposed system takes advantage of these recent progresses and research by integrating a RAG and a Knowledge-Graph (KG) to ensure that responses are grounded in curated museum data, thereby reducing hallucinations and providing accurate, current information. Additionally, the system stores pre-defined Q&A information for faster processing of users' questions [23]. Moreover, the KG is used to incorporate automatic relationships between artworks and artists which are used to provide

recommendations and expand contextual information provided to the generative LLM. Moreover, a Shield module filters irrelevant or biased data before integration, enhancing reliability.

2.3. Computer Vision Techniques for Art Analysis and Image Retrieval

Computer vision (CV) plays a pivotal role in the digital humanities, particularly in art history and cultural heritage, by enabling tasks such as object detection, form analysis, and iconography. Interdisciplinary projects between computer vision and art history have focused on studying iconography, identifying recurring objects and figures, and analyzing artistic styles across large image corpora [24, 25].

Content-Based Image Retrieval (CBIR) systems have evolved significantly to address the unique challenges of art and cultural heritage images, such as variations in artistic style, technique, and the absence of clear textual metadata [26]. Early CBIR systems relied on similarity measures derived from low-level image features, which could be computationally intensive for large databases. More advanced approaches now utilize pattern-based retrieval, defining similarity over higher-level patterns associated with clusters of low-level features, thereby improving efficiency and accuracy. These systems are crucial for overcoming limitations like language barriers and the lack of specific keywords when searching vast art collections [27].

On the other hand, recent advances on Vision-Language Models (VLMs) represent a key opportunity, fusing visual and textual data to perform complex tasks such as visual question answering (VQA), image captioning, and enhanced search [28]. VLMs are trained on massive datasets of image-text pairs, learning to associate visual elements with their corresponding language descriptions. This enables them to simultaneously comprehend and reason about both modalities, extending their utility beyond mere image retrieval to generating detailed descriptions or answering questions about visual content [29].

In our proposed system, we use LLMs for text analysis and VLMs for visual captioning and object detection. Then, the system intersects both analyses to match visible elements with historically significant descriptions, which human curators later refine through a Content Management System (CMS). This multi-modal, multi-stage, and human-in-the-loop pipeline is essential for generating high-quality, art-historically information presented to the visitor.

2.4. Personalization and User Modeling in Museum Visitor Experiences

Personalized content delivery is a most-needed factor for significantly enhancing museum visitor experiences, as it can adapt too technical or complex art information to individual interests, prior knowledge, and learning styles. AI-driven systems are increasingly being developed to deliver content to individual and group profiles, fostering both personalized learning and social interaction within museum settings [30].

Approaches to user profiling in museums typically involve a combination of explicit and implicit methods. Explicit methods rely on direct user participation, such as surveys, online forms, or direct input of preferences during an onboarding process. While useful for initial data collection, these methods can be limited in capturing the dynamic nature of visitor behavior and may be perceived as intrusive. Implicit methods, conversely, gather data based on user behavior and interactions without requiring direct input. This includes tracking browsing history, click patterns, dwell time at exhibits, and location via sensors or cameras.

For instance, systems utilizing Bluetooth Low Energy (BLE) smart badges or fixed cameras can track visitor movements and interactions to infer interests and build dynamic profiles [31]. The challenge lies in balancing the depth of personalization with user privacy and ensuring that data collection is not intrusive. Systems must maintain a strong focus on data privacy and ethical considerations throughout the profiling and content delivery process.

Our system addresses this by collecting minimal personal data during onboarding (i.e., knowledge level, range of age, time to visit the museum and country of origin) while enabling contextual adaptation,

aiming to provide a rich experience without compromising privacy. This information is then used by the LLM to adapt the generated responses to the user depending on the provided information.

2.5. Ethical AI Considerations for Public-Facing Conversational Agents

The deployment of AI assistants in public systems, particularly in sensitive domains like cultural heritage, raises ethical concerns that demand consideration. These include issues of privacy, algorithmic bias, transparency, and accountability.

For instance, algorithmic bias, often stemming from skewed or unrepresentative training data, can lead to discriminatory or unfair outputs from AI assistants. The lack of transparency in AI assistant decision-making processes, where models often operate as ”black boxes,” can erode user trust, especially in contexts where understanding the reasoning behind a response is crucial. Furthermore, determining accountability for AI assistant errors or misuse remains a complex challenge, raising questions about responsibility among developers, deploying organizations, and the AI system itself.

To mitigate these risks and build user trust, proactive measures are essential. These include robust content moderation and filtering mechanisms, such as incorporation of safe guardrails [32].

Several options are available such as Nemo-Guardrails [33] or LlamaGuard [34]. In our case, through LlamaGuard v3. In addition, clear user-facing disclaimers are needed to inform visitors that they are interacting with an AI system that, despite safeguards, RAG/KG-based approaches or contextualization information, may occasionally produce incorrect or imprecise answers.

Moreover, additional preventive steps such as usage of only curated data or secure data handling practices (e.g., HTTPS-encrypted communication, local logging) allow protecting user privacy. Finally, manual validation of responses in potentially sensitive scenarios provides an extra layer of oversight.

The integration of these ethical considerations from the outset, rather than as an afterthought, demonstrates a commitment to trustworthy AI development. This approach, where ethical considerations are a foundational design principle, is paramount for the successful and responsible long-term deployment of AI in public domains, including cultural institutions. It advocates for ”ethics by design” to foster user and stakeholder trust.

2.6. Attention Schema Theory as a Cognitive Framework for AI

The Attention Schema Theory (AST) offers a mechanistic explanation for subjective and contextual awareness, positing that it is the brain’s internal, simplified model of its own attention process [35, 36]. This internal model, analogous to the body schema that helps control physical movement, aids in regulating attention and understanding the attentional states of others. The theory suggests that the adaptive value of this internal model lies in its ability to improve the control of attention, allowing for flexible planning of attentional shifts and behavioral reactions. AST provides a compelling foundation for engineering artificial consciousness by focusing on the computational processes that lead a machine to claim awareness, rather than delving into metaphysical aspects. If a machine constructs internal information that describes the condition of consciousness, it can then ”believe” and ”act like it is conscious” in a manner analogous to human cognition.

The implications of AST for AI are significant[1]. It suggests pathways for building more human-like consciousness in machines, leading to enhanced social intelligence. An AI system equipped with an attention schema could better regulate its own computational resources and more effectively understand and predict the attentional states of other AI or human agents. This capacity to attribute consciousness to itself and others is considered crucial for socially competent human-AI interaction, potentially avoiding AI behavior that ignores human-like social cues or user intent, a challenge in current human-AI interaction.

The personalization features incorporated in our AI assistant are directly informed by AST and Theory of Mind research from the ASTOUND project [37], aiming to create more cognitively aligned

[1]When applied to machines, AST does not claim to create real consciousness, but rather a functional analog where systems can model their own focus and simulate awareness in interactions.

interactions (e.g., adapting tone/complexity to user knowledge or simulating attention to user's focus). This demonstrates a clear trajectory from theoretical cognitive science research to its practical application in AI system design.

The proposed AI assistant integrates the Attention Schema Theory (AST) as a guiding cognitive framework to inform the design of personalized and context-aware interactions in museum environments. Rather than attempting to simulate full attentional control or self-awareness, the system incorporates multimodal inputs—including user profile data, visual focus (e.g., selected artworks or Points of Interest), and conversational context—as sources of information that can be used to approximate a model of user awareness. These diverse signals are processed across multiple modules to build a situated understanding of the user's state, which enables the system to generate modulated responses in tone, depth of explanation, and vocabulary, based on user archetypes (such as age, knowledge level, and available time), simulating awareness of who the user is and what they are attending to. While real-time dynamic adaptation based on attentional shifts remains an open area of research, the current implementation demonstrates how foundational modeling of awareness-related information can be operationalized within a museum AI assistant.

Furthermore, the system is informed by Theory of Mind (ToM) principles to enhance its social intelligence, allowing it (in future developments) to infer and respond to inferred user intentions, interests, and knowledge gaps. By combining insights from AST and ToM, the assistant moves beyond simple rule-based personalization to offer interactions that are more aligned with human cognitive and communicative expectations. This approach represents a step toward building systems capable of more nuanced, context-sensitive behavior, while maintaining transparency and ethical design in public-facing applications.

Rather than claiming metaphysical consciousness, the system functionally models awareness as a mechanism for attentional control and social responsiveness, as proposed by Graziano's AST. The assistant's personalized responses, PoIs, and user profiles reflect our intention to simulate the user's attentional focus, fulfilling AST's goal of creating machines that "act as if they are aware." This computational modeling of awareness improves both the relevance and empathy of the AI's behavior, contributing to more human-aligned interactions in cultural and educational settings.

3. System Architecture and Implementation

3.1. Overall System Design and Data Flow

Our proposed system is guided by principles of flexibility and modularity, facilitating the seamless integration of diverse components and supporting future enhancements. The overarching architecture, as depicted in Figure 1, is organized around high-level functionalities, emphasizing clear data flows between modules to ensure seamless interaction.

At its core, the system comprises four main functional blocks: the Art Database, the User Interface, the LLM Layer, and the AI Engine. Data is systematically organized as a relational graph within a Neo4j database, allowing various system components to access information as needed. This design choice promotes a robust and scalable foundation for the AI assistant's operations.

3.2. Art Database: Structure and Content Curation

The Art Database forms the foundational knowledge base of the system, implemented using a Neo4j graph-based system and QDrant system. The database meticulously stores all pre-processed data curated by the Universidad Politécnica de Madrid (UPM) team [2]. Its content encompasses a rich array of information, including detailed artwork descriptions, artist biographies, associated artistic movements, visual salient elements, museum locations, multimedia content, and extensive metadata such as language, relationships, or creation year.

[2]For demo purposes we incorporated data from Thyssen-Bornemisza and Royal Collections museums

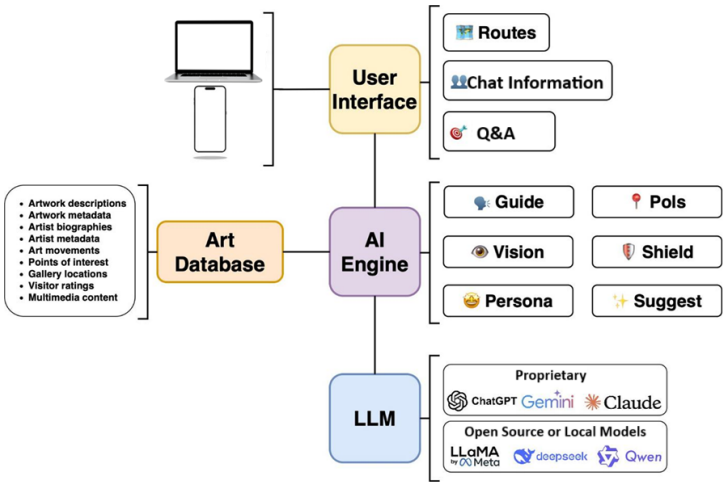

Figure 1: Overview of the architecture implementation for the developed demo.

The content for this database was meticulously sourced from official museum websites and audio guides provided by a collaborator partner GVAM. This prioritization of official and curated content ensures high factual accuracy and reflects the professional tone and institutional identity of the museums. Bilingual content in Spanish and English was specifically prioritized to test the system's multilingual capabilities. A custom pipeline, developed by UPM, was employed for data processing. This pipeline automatically crawled website content or transcribing audio-only content from audio guides using Whisper Large V3[3], which demonstrated a word error rate below 10% in both English and Spanish during internal evaluations (most errors were due to punctuation or some casing issues)[4]. The transcribed text was then segmented into paragraphs, grammatically corrected, and normalized for named entities. Subsequently, LLMs, such as GPT-4, were utilized to extract structured information, including proper nouns, dates, and artistic movements, enabling the automatic mapping of relationships among artworks, artists, and historical periods[5].

The graph database generation pipeline transforms each museum record into five primary node types: Artwork, Artist, Movement, Century, and Product. Each node retains its original metadata, including titles, dates, descriptions, and links to media stored in a separate image database accessible by components like the Visual Engine. An instruction-tuned LLM (LLaMA 3.1–8B-Instruct[6]) further processes artwork descriptions to extract entities such as people, places, techniques, and themes. These entities are added as labeled nodes and linked to relevant artworks and artists, with artworks also connected to their corresponding Century node based on creation date. All nodes and relationships are then exported in CSV format and imported into Neo4j, resulting in approximately 3,418 nodes and 4,913 relationships. This pipeline was designed for reusability, allowing new data to be integrated without disrupting the existing structure. For the purpose of the demo, a focused subset of 20 well-known artworks was selected from both museums. This deliberate selection aimed to reduce the risk of hallucinations by the LLMs used in generating synthetic datasets and to ensure the accuracy and reliability of the automated analysis employed for relationship generation within the Neo4j graph.

3.3. AI Engine Modules

3.3.1. Visual Retrieval-Augmented Generation (VisualRAG)

The VisualRAG module enables automatic artwork identification based on user-uploaded images. This functionality is achieved through a sophisticated three-stage retrieval pipeline. The image pre-

[3]https://huggingface.co/openai/whisper-large-v3
[4]We tested on +80 manually transcribed audios for English and Spanish
[5]In future activities, it will require curators to oversight the results.
[6]https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

processing phase is crucial for preparing artwork images for efficient retrieval. The system differentiates between 2D artworks (e.g., paintings, engravings) and 3D artworks (e.g., sculptures, functional items). Visual features are extracted using a pre-trained model (Qdrant/ResNet50-ONNX) and stored in a Qdrant vector database, facilitating efficient global embedding-based matching. During online operation, the system functions as an API endpoint with all models preloaded for real-time inference. When a user submits an image, the three-stage retrieval pipeline is executed:

- **DETECTION**: The detection module employs YOLOv8[7], a multi-class object detector, to identify and localize various types of artworks within a single frame. YOLOv8's flexibility, including support for custom object classes and natural language-based object queries, allows for adaptable deployments across different museum environments. Each detected object is individually cropped for subsequent processing.
- **RETRIEVAL**: Following detection, a Retrieval-Augmented Generation (RAG) step associates recognized artworks with relevant textual information, such as descriptions and historical context. This process begins with global embedding-based matching, which efficiently narrows the search space using semantic similarity. This step is vital for maintaining real-time performance, as direct use of local features over a large corpus would be computationally prohibitive. Once a shortlist of top-K candidates is identified, the system fetches corresponding textual content from the database for context-aware interaction.
- **RERANKING**: To refine retrieval accuracy, a reranking stage applies local feature matching. Additional features are extracted from each image at index time and stored in lightweight .npz files. At query time, new features are computed for the input image, and a matching score is used to re-rank the previously retrieved images in step two.

The system is optimized for real-time performance, targeting 30 frames per second (FPS) for streaming input. The full pipeline operates on a dedicated GPU-equipped server, utilizing approximately 1.5 GB of GPU memory. The compact.npz format for local features and the Qdrant vector database ensure scalability across thousands of artworks without overwhelming system resources.

3.3.2. Automatic Detection of Salient Visual Elements and Edition

This module is designed to automatically identify relevant visual elements, or Points of Interest (PoIs), within artworks. This capability enriches the visitor's experience by providing additional context, sparking curiosity, and deepening engagement, all while minimizing the manual effort required from museum staff. The system is built for high automation and scalability, particularly for newly acquired or temporary collections.

The full pipeline consists of three main components: automatic detection of PoIs and generation of associated textual descriptions, a Content Management Service (CMS) for reviewing and curating the automatically generated outputs, and integration with the User Experience (UX) interface for dynamic interaction with identified elements.

- **Automatic PoI Detection**: This module processes museum artwork images alongside their associated audio guide transcripts or crawled text through a multi-step pipeline:
  - Step 1: Visual Language Models are used to visually identify predefined elements (e.g., people, objects, places) and associate them with available text describing their significance, classifying each into different categories, e.g., People, Places, Landscapes, etc.
  - Step 2: Visual Caption Generation: Advanced vision AI models generate detailed textual descriptions of the artwork's visual content associated with the categories identified in the previous step.
  - Step 3: Iterative refining: A multimodal LLM matches the visual elements, descriptions and curated text to remove elements that are not relevant, as well as to associate an image with a descriptive text.

[7] Yolov8

  - Step 4: Bounding Box Storage: Post-Processing: Once the previous step is finished, the visual coordinates of the filtered elements are stored in the Neo4j dataset.
- **Part 2: Interactive Editing Interface (CMS)**: The Interactive Editing Interface is a comprehensive tool enabling museum staff and curators to review, verify, and enhance the automatically detected PoIs.
- **Part 3: Integration with the UX**: The graphical user interface for the visitor at the museum incorporates a button to display the Points of Interest for each artwork and shows the information associated to each element in the artwork.

3.3.3. Personalization and Adaptive Content Generation

The designed system dynamically adapts to user characteristics and uses contextual information, building upon findings from the ASTOUND project, particularly the exploration of Attention Schema Theory (AST), Theory of Mind, and contextualized awareness.

The AST proposes that subjective awareness aids in controlling attention and understanding the attentional states of others. This theoretical component guides the personalization approach in the current demo. An internal combination mechanism is implemented to incorporate information from different modules and use them to adapt the generative output based on previously defined user archetypes [13]. These archetypes include age-based categories (children, teenagers, young adults, adults, retirees), art knowledge levels (novice, intermediate, expert), and available time (short, standard, extended visits).

The onboarding process is streamlined and unobtrusive, and minimize the collection of identifiable personal data. In concrete, the system collects name, country of origin, age group (as predefined ranges), visit duration preference, and level of art proficiency. This ensures privacy while allowing contextual adaptation. Based on these inputs, the system automatically generates a customized system prompt that guides the AI assistant's behavior throughout the interaction, dynamically adapting its tone, vocabulary, and depth of information to the user's profile. For instance, a "Spanish Teen User" might receive casual, modern language with contemporary references, while an "English Retired User" with "Expert" art proficiency might receive professional language emphasizing historical context and artistic significance, with detailed, multi-paragraph explanations. Then, this information is dynamically combined with contextual information to improve the system's ability to simulate attentional focus and contextual adaptation, inspired by AST, which is manifested in its generated responses.

3.3.4. Suggestions and Recommendations Framework

This feature was identified as a potentially valuable addition for museums, aiming to maximize the return on investment (ROI) from the AI-based solution by linking it to revenue-generating activities, such as promoting museum gift shop items. The Neo4j graph system was specifically used to manage and present related products in a structured and easily retrievable manner, supporting future experimentation or deployment. The provided user's archetype is used to enable and display specific recommended items.

The current prototype includes a proof-of-concept integration using data from the Thyssen-Bornemisza Museum, where related products from their online store were crawled and stored in the Neo4j database, and a UX component was included to dynamically display these associated shop items. This case study provides valuable insight into the decision-making process for AI deployment in public-facing cultural institutions, demonstrating the importance of balancing commercial potential with the core values and user expectations of the specific domain.

3.3.5. Shield Module for Content Moderation and Safety

To ensure the reliability and appropriateness of AI-generated responses when interacting with real users, the demo incorporates basic safeguard mechanisms designed to analyze both input prompts and generated outputs before they are processed by the underlying generative models. Improper handling

of real-world input can lead to non-factual, misleading, or negatively biased content. These safeguards aim to detect irrelevant or biased user inputs, low-quality or grammatically incorrect content, and content not aligned with the museum's tone, factuality, or educational goals.

As part of this effort, the system integrates the LlamaGuard-V3 model for real-time content moderation and safety checks. Furthermore, a user-facing warning message is presented at the start of each AI assistant session, clearly indicating that the visitor is interacting with an AI-powered system that, despite safeguards, may occasionally produce incorrect or imprecise answers. The message also encourages users to keep their questions focused on factual information about the artworks and museum context. These initial safety mechanisms serve as a baseline, with future development plans aiming to incorporate more advanced trust and safety layers as the system scales or moves to public-facing deployments.

3.4. User Interface and Multimodal Interaction

A responsive User Experience (UX) frontend was developed to enable seamless and secure interaction between visitors and the backend AI components via socket connections. This interface supports a broad range of capabilities designed to enhance user engagement (see Figure 2). Key features of the user interface include:

- **Chat Interface**: Provides real-time messaging with typing indicators, voice integration for hands-free interaction, and support for multimodal inputs such as text, voice, and Points of Interest (PoI) selections. It also incorporates context awareness by considering artwork details and user preferences for personalized engagement.
- **Multimodal Input and Output**: Integrates Automatic Speech Recognition (ASR) capabilities, which can be provided by local ASR models like Whisper-V3-Large or via API calls (e.g., OpenAI or Microsoft Azure Cognitive Services Speech SDK). For our demo, the Azure solution demonstrated comparable performance to the local model in internal tests and offers phonetic transcriptions, custom vocabulary additions, and multi-language support (English and Spanish). Complementing ASR, Text-to-Speech (TTS) capabilities allow the AI assistant to deliver information both as text and through synthesized speech, with options for voice gender selection and playback controls.
- **Points of Interest (PoI)**: Through the UX interface, visitors can explore curated points of interest. This way, visitors can obtain detailed explanations through the chat interface upon selection.
- **Personalization Features**: Streamlined forms collect minimal user profile information—name, country of origin, age group, visit duration preference, and art proficiency level—to configure adaptive content. This data is used to dynamically adapt the AI assistant's tone, vocabulary, and depth of information.
- **Multimodal Identfication of Artworks**: The demo includes the possibility for the visitor to upload an image captured inside the museum, which the system processes to identify artworks and provide relevant localization data along with detailed artwork information.
- **Language Support**: A comprehensive manual translation process was undertaken to ensure cultural accuracy and contextually appropriate communication. All textual content stored in the Neo4j database, as well as UX interface elements, error handling messages, and system prompts, were translated between English and Spanish. Individual translations were completed and integrated, along with cultural localization features like adjustments to date formats and currency displays.

3.5. Core System Technologies and Development Philosophy

The final system relies on a modern and robust technology stack. The frontend is built with Next.js and React, utilizing Tailwind CSS for styling and Radix UI primitives for UI components. Google Gemini 2.0 Flash serves as the primary Large Language Model (LLM)[8], although the platform supports

[8]We decided to use Gemini instead of GPT-4o to test the adaptability of our system to different LLMs providers

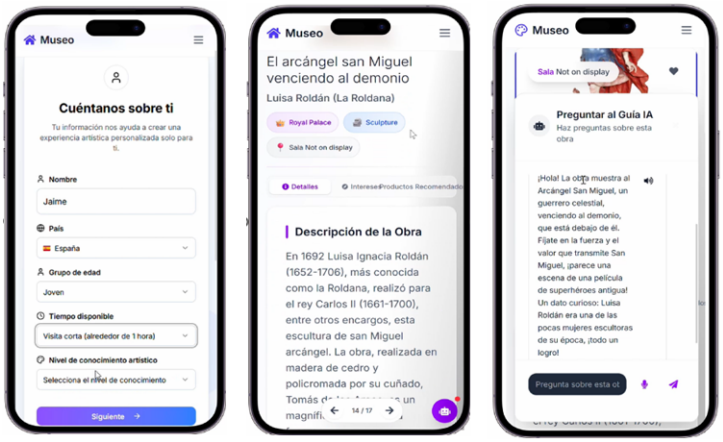

Figure 2: Screenshots of the designed UX. On the left: the onboarding form. On the center: Factual information about the artwork. On the right: Interactive information with the AI assistant.

other proprietary models (e.g., OpenAI's GPT, Anthropic's Claude) via API calls or using open-source models (e.g., Meta's LLaMA, DeepSeek, Qwen) deployed on-premises. Speech services integrate both proprietary options (e.g., Whisper, Microsoft ASR/TTS) and open-source alternatives (e.g., Whisper-v3-large, Chatterbox TTS). Framer Motion is used for animations, and TypeScript ensures full type safety throughout the codebase.

The design philosophy guiding the system's development emphasizes user-centricity, ensuring every feature adapts to individual user needs. Accessibility is prioritized through voice controls, keyboard navigation, and screen reader support. Progressive enhancement ensures core functionality works even without JavaScript, and mobile optimization provides touch-friendly interfaces with responsive design across various devices.

4. Results and Discussion

The described system was successfully presented during a Focus Group held on June 12th, 2025. The event, organized by our partner GVAM, was attended by 30 participants, including 12 expert panelists from prominent cultural institutions such as the Guggenheim Museum, Museo Thyssen-Bornemisza, Casa Batlló, Patrimonio Nacional, Real Madrid, and the Cathedrals of Barcelona and Seville.

The feedback garnered from this diverse group of directors, digital strategy leaders, and innovation experts was highly valuable and will directly inform upcoming enhancements during the final phase of the ASTOUND project.

Initial feedback from the focus group indicated a strong interest in the core visitor engagement features of the AI assistant. This choice underscores a pragmatic and user-centric development strategy. While the underlying technology could readily support many advanced features, the cultural institutions' representatives prioritized the educational and engagement value for visitors over other features. This demonstrates a nuanced understanding of the museum's primary mission, which revolves around cultural enrichment.

The system demonstrated robust real-time performance, particularly for its visual recognition capabilities, achieving a target of 30 frames per second (FPS) for streaming input. Efficient data retrieval from the Neo4j graph database further contributed to a seamless user experience. The multimodal capabilities, encompassing Automatic Speech Recognition (ASR), Text-to-Speech (TTS), and image input, provided a rich and accessible interaction.

The personalization engine effectively adapted AI assistant responses based on predefined user archetypes, showcasing its ability to generate context-aware content, tailored to individual preferences and profiles.

The integration of insights from the ASTOUND project is evident throughout the system. The AI

assistant's personalization features, for instance, are directly informed by the Attention Schema Theory (AST) and Theory of Mind research. This theoretical grounding aims to foster more cognitively aligned and empathetic interactions between the AI and museum visitors.

The architectural flexibility of the system is a deliberate design choice that allows for the future integration of advanced reasoning modules and attention analysis, as envisioned by the ongoing ASTOUND research. This forward-looking design ensures that the AST-chatbot can evolve towards even greater personalization by incorporating dynamic user modeling and the analysis of visual attention patterns, leading to more adaptive and intelligent responses.

5. Conclusions and Future Work

5.1. Conclusions

The described demonstrator, developed as part of the EIC ASSIST project and leveraging expertise from the EIC ASTOUND project, successfully integrates cutting-edge AI capabilities to deliver a personalized and interactive museum experience.

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. A graph-based storage layer, utilizing Neo4j for relational data and QDrant for semantic search with vector embeddings, provides the support for knowledge retrieval. 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 the backend system and the artworks.

From an ethical and responsible AI perspective, the system incorporates safeguards inspired by the measurements implemented through the ASTOUND's work on transparency, safety, and user trust. This demonstrates a foundational design principle, where ethical considerations were integrated from the outset, rather than being 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, not personal ones), and manual validation of responses in potentially sensitive scenarios were incorporated to ensure privacy and reliability[9]. This proactive integration of ethical considerations sets an example for responsible AI awareness during the design of this prototype, advocating for "ethics by design" to build trust with users and stakeholders, which is intended for the successful long-term deployment of AI in sensitive public domains like museums in the future.

Finally, the presentation of the demo during the Focus Group on June 12th, 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 in the final phase of the ASTOUND project.

[9]For this prototype uploaded images or audios are processed in-memory and not stored. Logs are stored with a hash number, and non sensitive personal information is collected. Although a preliminar prototype, the system is intented to comply with GDPR and includes opt-in consent and disclaimers.

5.2. Future Work

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 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 features from audio to predict and analyze user interest and contextualized awareness.
- Dynamic user modeling, integrating sophisticated reasoning modules and the information collected during the interaction. This progression directly stems from the Attention Schema Theory (AST) research within ASTOUND, in this case to allow the AI system to attempt proxy detection of attention via observable cues to infer user states (such as attention and engagement) and adapt its behavior accordingly, leading to more adaptive and intelligent responses. This demonstrates a clear commitment to advancing human-AI interaction beyond simple rule-based or data-driven responses towards systems that can "understand" and respond to human cognitive states.
- Refinement of Fine-tuned Language Models: During the project, a synthetic QA and Dialogue dataset was created to fine-tune a pre-trained model to specific museum content, with the goal of reducing 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. Additional work will be carried out to limit the need of using LLMs and instead use local models for privacy reasons.
- Platform Scalability and Robustness Improvements: Specific enhancements include better handling concurrent access for multiple users to the AI assistant, along with robust session and turn management. Additionally, invisible watermarking mechanisms will be explored for detecting responses generated by the fine-tuned model, enhancing transparency and accountability.

6. Ethics and Responsible AI

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, users could opt-out to not use the AI-based system and just got 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 without storage, and all communication is encrypted. As this prototype is used internally, formal data audits are not yet implemented but planned for later stages. Fairness and inclusivity were considered when designing personalization features for diverse age groups and knowledge levels. Basic content filtering (e.g., LlamaGuard) is used to prevent harmful outputs, though no bias audits have yet been conducted.

Transparency and accountability are supported through 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.

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".

Declaration on Generative AI

During the preparation of this work, the author(s) used Gemini-Flash 2.5 Flash in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] B. Caramiaux, Ai with museums and cultural heritage, AI in Museums: Reflections, Perspectives and Applications (2023) 117–130.
[2] H. Wang, Enhancing art museum experience with a chatbot tour guide, 2024.
[3] K. I. Roumeliotis, N. D. Tselikas, D. K. Nasiopoulos, Leveraging large language models in tourism: A comparative study of the latest gpt omni models and bert nlp for customer review classification and sentiment analysis, Information 15 (2024) 792.
[4] M. Heisler, C. Becker-Asano, Conversations with andrea: Visitors' opinions on android robots in a museum, arXiv preprint arXiv:2506.22466 (2025).
[5] C. Ouaddi, L. Benaddi, L. Naimi, M. Rahouti, A. Jakimi, R. Saadane, et al., Assessing the effectiveness of large language models for intent detection in tourism chatbots: A comparative analysis and performance evaluation, Scientific African 28 (2025) e02649.
[6] L. Secchi, et al., Knowledge graphs and large language models for intelligent applications in the tourism domain (2024).
[7] W. Villalobos, Y. Kumar, J. J. Li, The multilingual eyes multimodal traveler's app, in: International Congress on Information and Communication Technology, Springer, 2024, pp. 565–575.
[8] A. Kontogianni, E. Alepis, Ai in smart tourism: Llms & gpts leading the way, in: 2024 15th International Conference on Information, Intelligence, Systems & Applications (IISA), IEEE, 2024, pp. 1–8.
[9] M. Talakoti, A. BalaRam, D. Jagli, N. Deepika, R. L. Priya, K. Harshini, Trip tailor: Ai-powered travel planning with itinerary generation and chatbot assistance, International Journal of Environmental Sciences 11 (2025) 920–929.
[10] P. Krataithong, W. Buranasing, M. Buranarach, T. Wutthitasarn, P. Meeklai, P. Tumsangthong, Tourism chatbot framework: Enhancing visitor experience through graphrag and ai chatbot, in: 2025 IEEE International Conference on Cybernetics and Innovations (ICCI), IEEE, 2025, pp. 1–6.
[11] H. Lin, J. Wang, Y. Zhan, Y. Zhang, Tour guide robot based on large language model and user behavior analysis, in: 2024 5th International Conference on Artificial Intelligence and Computer Engineering (ICAICE), IEEE, 2024, pp. 794–798.
[12] G. Castellano, B. De Carolis, S. Ferilli, S. Massaro, G. Palestra, L. Perla, M. T. Santacroce, Towards social robot-based inclusive museum experiences for children with asd, Proceedings of AVICH (2024).
[13] I. Rachidi, A. Ezzakri, J. Bellver-Soler, L. F. D'Haro, Design, generation and evaluation of a synthetic dialogue dataset for contextually aware chatbots in art museums, in: Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, 2025, pp. 20–28.
[14] M. Gil-Martín, C. Luna-Jiménez, S. Esteban-Romero, M. Estecha-Garitagoitia, F. Fernández-

Martínez, L. F. D'Haro, A dataset of synthetic art dialogues with chatgpt, Scientific Data 11 (2024) 825.
[15] J. Li, Y. Yuan, Z. Zhang, Enhancing llm factual accuracy with rag to counter hallucinations: A case study on domain-specific queries in private knowledge-bases, arXiv preprint arXiv:2403.10446 (2024).
[16] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al., A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, ACM Transactions on Information Systems 43 (2025) 1–55.
[17] G. Agrawal, T. Kumarage, Z. Alghamdi, H. Liu, Can knowledge graphs reduce hallucinations in llms?: A survey, arXiv preprint arXiv:2311.07914 (2023).
[18] Y. Tang, Y. Yang, Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries, arXiv preprint arXiv:2401.15391 (2024).
[19] W. Su, Y. Tang, Q. Ai, Z. Wu, Y. Liu, Dragin: Dynamic retrieval augmented generation based on the information needs of large language models, arXiv preprint arXiv:2403.10081 (2024).
[20] W. Su, Q. Ai, J. Zhan, Q. Dong, Y. Liu, Dynamic and parametric retrieval-augmented generation, arXiv preprint arXiv:2506.06704 (2025).
[21] M. Riedler, S. Langer, Beyond text: Optimizing rag with multimodal inputs for industrial applications, arXiv preprint arXiv:2410.21943 (2024).
[22] J. Xue, Q. Deng, F. Yu, Y. Wang, J. Wang, Y. Li, Enhanced multimodal rag-llm for accurate visual question answering, arXiv preprint arXiv:2412.20927 (2024).
[23] S. Ramos-Varela, J. Bellver-Soler, M. Estecha-Garitagoitia, L. F. D'Haro, Context or retrieval? evaluating rag methods for art and museum qa system, in: Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, 2025, pp. 129–136.
[24] A. F. Foka, Computer vision applications for art history: Reflections and paradigms for future research, in: Proceedings of EVA London 2021, BCS Learning & Development, 2021, pp. 73–80.
[25] N. Rodríguez-Ortega, Image processing and computer vision in the field of art history, in: The Routledge companion to digital humanities and art history, Routledge, 2020, pp. 338–357.
[26] W. S. Tan, W. Y. Chin, K. Y. Lim, Content-based image retrieval for painting style with convolutional neural network, in: International Conference on Digital Transformation and Applications (ICDXA), volume 25, 2021, p. 26.
[27] D. Srivastava, S. S. Singh, B. Rajitha, M. Verma, M. Kaur, H.-N. Lee, Content-based image retrieval: A survey on local and global features selection, extraction, representation, and evaluation parameters, IEEE Access 11 (2023) 95410–95431.
[28] J. Zhang, J. Huang, S. Jin, S. Lu, Vision-language models for vision tasks: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).
[29] D. Li, L. Xia, L. Fan, Aug-creativity: Framework for human-centered creativity with vision language models, in: International Conference on Human-Computer Interaction, Springer, 2025, pp. 86–101.
[30] I. Vasic, H.-G. Fill, R. Quattrini, R. Pierdicca, Llm-aided museum guide: Personalized tours based on user preferences, in: International Conference on Extended Reality, Springer, 2024, pp. 249–262.
[31] R. Ivanov, Advanced visitor profiling for personalized museum experiences using telemetry-driven smart badges, Electronics 13 (2024) 3977.
[32] S. G. Ayyamperumal, L. Ge, Current state of llm risks and ai guardrails, arXiv preprint arXiv:2406.12934 (2024).
[33] T. Rebedea, R. Dinu, M. Sreedhar, C. Parisien, J. Cohen, Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails, arXiv preprint arXiv:2310.10501 (2023).
[34] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, arXiv preprint arXiv:2407.21783 (2024).
[35] M. S. Graziano, T. W. Webb, The attention schema theory: a mechanistic account of subjective awareness, Frontiers in psychology 6 (2015) 500.
[36] M. Graziano, The attention schema theory: A foundation for engineering artificial consciousness. frontiers in robotics and ai, 4, 60, 2017.
[37] J. W. Strachan, D. Albergo, G. Borghini, O. Pansardi, E. Scaliti, S. Gupta, K. Saxena, A. Rufo,

S. Panzeri, G. Manzi, et al., Testing theory of mind in large language models and humans, Nature Human Behaviour 8 (2024) 1285–1295.