UNIVERSITY OF PALERMO
PHD JOINT PROGRAM:
UNIVERSITY OF CATANIA - UNIVERSITY OF MESSINA
XXXVI CYCLE

DOCTORAL THESIS

Digital solutions for Immersive Virtual Reality exploration of Cultural Heritage sites

Author:
Mariella FARELLA

Supervisor:
Prof. Giosuè LO BOSCO

PhD Coordinator:
Prof. Maria Carmela LOMBARDO

Co-Supervisor:
Dott. Giuseppe CHIAZZESE

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in

Mathematics and Computational Sciences

November 11, 2025

UNIVERSITY OF PALERMO

Abstract

Department of Mathematics and Computer Sciences

Doctor of Philosophy

Digital solutions for Immersive Virtual Reality exploration of Cultural Heritage sites

by Mariella FARELLA

This thesis explores how digital technologies can be used to document, preserve, and improve the accessibility of cultural heritage through immersive experiences. The research proposes a complete process that integrates 3D data acquisition techniques, such as photogrammetry, LiDAR scanning, and Structure-from-Motion, with post-processing workflows that include mesh reconstruction, simplification, UV mapping, and texturing. The resulting models are optimized for interactive visualization within development engines such as Unity3D, making them suitable for implementation on multiple platforms, from VR headsets to CAVE environments.

Building on this technical foundation, the research also investigates how intelligent systems can enhance knowledge transfer within immersive environments. A question answering system, based on large language models, has been integrated into the virtual experience to provide users with dynamic and context-sensitive access to cultural content. This approach shifts the focus from providing static information to interactive dialogue, allowing users to explore works of art and monuments in a more engaging way. To evaluate its effectiveness, a pilot study was conducted in which participants interacted with a virtual museum and a qualitative assessment of the QA system was made. The results demonstrated the potential of combining immersive visualization with natural language interaction, opening up new perspectives for education, accessibility, and the public dissemination of cultural heritage.

Contents

List of Figures

List of Tables

List of Abbreviations

VR	**V**irtual **R**eality
AR	**A**ugmented **R**eality
XR	e**X**tended **R**eality
MR	**M**ixed **R**eality
AI	**A**rtificial **I**ntelligence
NLP	**N**atural **L**anguage **P**rocessing
TLS	**T**errestrial **L**aser **S**canning
SfM	**S**tructure **f**rom **M**otion
MVS	**M**ulti-**V**iew **S**tereo
UAV	**U**nmanned **A**erial **V**ehicle
LLMs	**L**arge **L**anguage **M**odels
QA	**Q**uestion **A**nswering
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers
MLM	**M**asked **L**anguage **M**odeling
NSP	**N**ext **S**entence **P**rediction
EM	**E**xact **M**atch
VQA	**V**isual **Q**uestion **A**nswering
xAPI	Experience **API**
STT	**S**peech-**T**o-**T**ext
TTS	**T**ext-**T**o-**S**peech
LRS	**L**earning **R**ecord **S**tore

6

Chapter 1

State of the art

In recent years, the rapid development of digital technologies has had a significant impact on the cultural heritage domain, completely changing the way we experience, preserve, and enhance it. In particular, the rise of virtual reality (VR), augmented and mixed reality (AR/XR), as well as artificial intelligence (AI), has opened up new possibilities for interaction with cultural heritage. This has led to immersive, personalized, and interactive experiences that can captivate a broader and more diverse audience.

At the same time, there has been a growing enthusiasm for using cultural heritage as an educational context to stimulate innovative teaching methods. Immersive and intelligent technologies have found application in museums, schools, universities, and informal learning environments, offering tools for cultural mediation and both formal and informal education.

This chapter aims to critically analyse the state of the art in this interdisciplinary field, offering an overview of key research directions and significant technological developments mainly in the field of virtual reality and artificial intelligence applied to cultural heritage, but also to explore the role of these technologies in educational contexts, with a focus on experiences that see heritage as a learning resource. This analysis provides the theoretical and methodological basis for the proposed research work, highlighting emerging trends, opportunities, and ongoing challenges in the field.

1.1 Theoretical framework: digital heritage and cultural mediation

The growing presence of digital technologies in the field of cultural heritage cannot be understood only as a technical development. It marks a deeper transformation in the way we relate to cultural memory, how it is created, shared, and experienced across time and space. As stated by UNESCO (UNESCO, 2003), digital heritage includes both the materials that have been digitized and those that are born digital. These are not merely data or images, but living expressions of human creativity that now form part of our shared memory. Preserving them, therefore, also means preserving the diversity of ways in which people and communities tell their stories.

In this sense, concepts such as *digital twin* and *virtual reconstruction* are not just technical tools for documentation, but instruments of interpretation and communication. A digital twin offers a dynamic model of an artifact or a site, able to evolve as new data are added. Virtual reconstruction, on the other hand, allows us to reimagine what is lost or inaccessible, connecting the tangible and intangible dimensions of heritage. Together, they transform the act of preservation into an act of storytelling,

inviting both experts and the public to engage with heritage as a living process rather than a fixed image (Mazzetto, 2024; Bekele et al., 2018).

However, the use of digital technologies in cultural heritage also calls for a careful balance between innovation and authenticity. According to ICOMOS (ICOMOS, 2017), digital representations should not replace the original object, but rather extend our ability to understand it, without oversimplifying its meaning or erasing its complexity. Every digital reconstruction is, in a sense, an interpretation; and with interpretation comes responsibility. The way a digital model is designed, narrated, or displayed inevitably reflects cultural perspectives and institutional choices.

Recent European initiatives, such as the European Commission's *Recommendation on a Common European Data Space for Cultural Heritage* (European Commission, 2021) and the *Europeana Strategy 2020–2025* (Europeana Foundation, 2020), have recognized this interpretive dimension. They emphasize openness, inclusivity, and community participation as essential values guiding digital transformation in the heritage sector. These frameworks encourage institutions not only to digitize collections but also to involve diverse audiences in shaping how cultural heritage is represented and understood. Similarly, the ICCROM report Sustaining Digital Heritage: Towards a Global Strategy (ICCROM, 2021) advocates for the creation of digital environments that are sustainable, transparent, and accessible to all, reflecting cultural diversity and ethical care.

From a broader perspective, this approach resonates with the ideas developed within the Digital Humanities, where technology is seen not simply as a means of efficiency, but as a medium for interpretation and meaning-making (Drucker, 2013). Digital heritage thus becomes a shared space of cultural mediation, a meeting point between curators, researchers, and communities, where past and present interact through digital narratives, virtual reconstructions, and participatory experiences. In this light, digitization is not an end in itself, but part of a continuous dialogue between people, memory, and technology.

1.2 Digital technologies in cultural heritage

Since the 1990s, digital technologies have fundamentally transformed the field of cultural heritage. This shift is not merely incremental but represents a significant reimagining of how heritage is managed, conserved, and experienced. The process of digitization has introduced novel approaches for preserving and documenting cultural artifacts, vastly improving public access to collections and archives that were once confined to limited audiences. Additionally, these technologies support preventive conservation and have paved the way for hybrid and interactive experiences, such as virtual museum tours and immersive digital exhibitions. One of the fields in which digital technologies have been applied concerns the digitalisation of cultural heritage, understood as the acquisition of visual and structural data using 2D and 3D scanners, photogrammetry, and advanced imaging techniques. These methodologies facilitate the creation of highly accurate digital replicas of objects, artifacts, and sites, many of which are otherwise inaccessible to the public, contributing to their protection and study. In line with these objectives, the Italian National Plan for the Digitisation of Cultural Heritage (PND), promoted by the Ministry of Culture, emphasises the strategic importance of digitalisation as a tool for the preservation, enhancement, and accessibility of heritage. The plan highlights the essential role of robust digital infrastructures in fostering a sustainable and inclusive cultural

ecosystem, ensuring that cultural resources are available to a broader and more diverse audience (Ministero della Cultura, 2022).

With the rise of digitalisation, digital storage systems have naturally evolved as well. In Italy, for example, databases and online repositories like Europeana(Europeana, 2008) or CulturaItalia (Ministero della Cultura, 2008), have emerged, offering organised and described digital content for both public and scholarly use. These platforms play a significant role in ensuring the long-term preservation of cultural materials and facilitating the dissemination of knowledge. At the same time, they introduce new challenges, particularly around system interoperability, standardisation, and the ongoing sustainability of the supporting technologies.

A natural evolution of digitisation is the emergence of virtual museums, digital environments that offer visitors experiences that use interactive technologies. These museums can be complementary to physical institutions or entirely digital, and are characterised by the use of 3D graphic interfaces, multimedia content, and sometimes even elements of gamification and social interaction. This evolution marks a significant shift in how cultural content is accessed and experienced.

Virtual museums are a great way to make culture more accessible and spread it around, as well as offering new opportunities for distance learning. In particular, their importance has become evident during crises such as the COVID-19 pandemic, when traditional museum visits were impossible. These digital tools have allowed a wider audience to continue to enjoy cultural and educational content, thus contributing to the digitisation of cultural heritage (Tserklevych et al., 2021). However, the adoption of virtual museums also presents a number of open questions and critical challenges. One of the main ones concerns the authenticity of the aesthetic experience: digital fruition, no matter how immersive or interactive, cannot fully replicate the sensory, spatial, and emotional perception resulting from direct comparison with the original work in a physical museum context. Secondly, the issue of mediation of curatorship emerges, i.e., the risk that the virtual experience reduces or simplifies the role of the curator as cultural mediator, turning a critical and contextualised narrative into a simple individual exploration guided by technology. Finally, there is the danger of a conceptual and symbolic disconnection from the original object: in the digital context, the work of art or artefact risks becoming images or models disconnected from their materiality, history, and cultural value, with a consequent loss of interpretative depth (Calise, 2022, Pietroni and Pagano, 2016).

Several European and international projects, such as the Virtual Museums Transnational Network (Commission, 2014) or Scan4Reco (Scan4Reco, 2018, have explored the potential of digital museums in redefining the very concept of cultural fruition, highlighting the importance of an integrated approach between technologies, content, and audience.

1.3 VR, AR and XR in cultural heritage

Initially, the focus was mainly on digitisation and remote access to cultural heritage. In recent years, however, there has been a clear shift towards experiential and immersive technologies that transform cultural enjoyment into a truly sensory and participatory interaction. In particular, virtual, augmented and extended reality applications have expanded the narrative and interpretative possibilities of museums (Carrozzino and Bergamasco, 2010), archaeological sites and temporary exhibitions, overcoming the traditional limitations imposed by the physical nature of collections and offering new models of engagement for the public.

Virtual reality is now an advanced tool for accessing fully simulated, three-dimensional, navigable spaces, digitally recreated to reproduce places, objects or historical scenarios, even those that no longer exist. Through the use of VR headsets, controllers and various sensory interfaces, users can immerse themselves in digital environments that offer detailed and faithful reconstructions of archaeological sites or museums (Kassahun et al., 2018). At the same time, augmented reality stands out for its ability to superimpose digital elements onto the physical world, enriching the perception of the real environment through mobile or wearable devices, such as tablets and smart glasses. This technology is particularly effective in exhibition and museum contexts, as it allows for the integration of additional information, three-dimensional models, translations and interactive content, enhancing the user experience while maintaining direct contact with the original object (Azuma, 1997, Cipresso et al., 2018). More generally, the term Extended Reality (XR) is used to describe the entire spectrum of experiences that combine the physical and digital worlds. This concept is effectively illustrated by the Virtuality Continuum proposed by Milgram et al. (Milgram and Kishino, 1994), which represents a scale ranging from the completely real environment to the completely virtual environment. At one end of the continuum lies the real environment, consisting exclusively of physical objects. At the opposite end is VR, where the entire environment is computer-generated. Between these two poles lies Mixed Reality (MR), which includes: AR, where digital content (such as 3D models, text, or video) is overlaid on the real world and enhances the user's perception of reality; Augmented Virtuality (AV), which is less common and consists of digital environments enriched by real-world elements (e.g., live video feeds integrated into a VR setting). This continuum helps conceptualize how different immersive technologies integrate real and virtual components to varying degrees, and it provides a theoretical framework for designing cultural heritage experiences with different levels of immersion and interactivity. The model is illustrated in Figure 1.1.

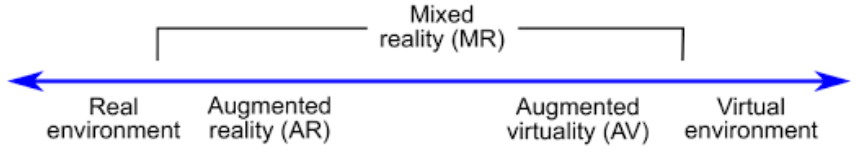


FIGURE 1.1: The Virtuality Continuum by Milgram et al. (1994)

This category now represents a rapidly expanding field that finds application in many areas, including cultural heritage, due to its ability to offer immersive, personalised and highly interactive experiences. These technologies not only support conservation or education, but also become active tools of cultural mediation, capable of bringing into play the role of the visitor, who, from passive observer, becomes an integral part of the experience. Immersive applications, in fact, propose multisensory fruition modes, based on interaction, historical reconstruction and personalised access to cultural information.

1.3.1 Museum and Archaeological Applications

As already mentioned, VR now allows the creation of three-dimensional digital museums, accessible both remotely and through dedicated physical installations. In these digital spaces, users can move freely, explore detailed historical reconstructions and interact with virtual objects. These environments, often created using photogrammetric surveys, laser scans and 3D modelling, guarantee remarkable realism

and significant emotional engagement (Carrozzino and Bergamasco, 2010). Some examples include The Virtual Museum of Iraq project or the VR installations at the National Archaeological Museum in Naples, where it is possible to experience an immersive reconstruction of the Villa of the Mysteries in Pompeii.

AR, on the other hand, represents a significant evolution from simple physical tours, superimposing layers of information and visuals directly onto the real world. Using devices such as tablets, smartphones or dedicated headsets, users can access multimedia content, animations, iconographic details and explanatory texts displayed alongside the works. This greatly enriches their understanding of the historical and artistic context and promotes a more immersive and personalised experience (Sylaiou et al., 2010). In this way, AR takes on the role of an augmented cultural mediator, capable of guiding visitors along the exhibition route, adapting the information offered in real time to the specific needs and curiosities of the user. In the context of archaeology, virtual reality is now an essential tool for the immersive reconstruction of ancient sites, which today are often only partially preserved or have disappeared altogether. Through collaboration between museums, universities and research centres, interactive experiences are being developed that allow users to "walk" virtually among the ruins of Pompeii, observe the daily life of a Roman settlement or explore iconic environments such as the Forum of Rome or the Palace of Knossos. These simulations are based on certified archaeological data, detailed topographical models and historically accurate reconstructions. The aim is not only to faithfully reproduce the physical environment, but also to stimulate empathy and cognitive engagement among users, thereby promoting a deeper understanding of the life and culture of ancient civilisations (Di Franco et al., 2015).

These technologies are not limited to visual representation, but contribute to museum educational activities, territorial enhancement and cultural accessibility, opening up new perspectives for active public participation, especially for young people, international tourists and people with disabilities.

1.3.2 Immersive Reconstructions and Sensory Experiences

Immersive reconstructions represent one of the most fascinating and immersive applications of extended reality technologies within cultural heritage. These experiences now go far beyond the simple three-dimensional representation of historical places or objects. They integrate a variety of sensory stimuli, not only visual, but also auditory and, increasingly, tactile and even kinesthetic, to recreate historical environments that are as faithful as possible from both an aesthetic and emotional point of view (Champion and Champion, 2011; Ch'ng et al., 2022). In this way, visitors are no longer limited to a passive role, typical of traditional museum models, but become active protagonists in a multisensory experience that fully engages them and transports them through different eras and places. The realism of the reconstructions is enhanced through the use of haptic interfaces, immersive audio devices with spatialised sounds, reactive flooring, smell stimulation and even tactile interaction with virtual surfaces. All these elements combine to generate a sense of presence and authority that has shown significant effects on engagement and teaching effectiveness (Riva, Waterworth, Murray, et al., 2014). In particular, in the educational sphere, these technologies can facilitate experiential learning, fostering a situated and more lasting memory than the mere frontal display of content.

These approaches are particularly effective in contexts dedicated to students, young people and people with disabilities, as they allow the experience to be tailored to individual needs and contribute significantly to improving overall accessibility.

For example, in several European projects like INCEPTION (Commission, 2020) or V-MUST.NET (Commission, 2014), the use of immersive environments has enabled visually impaired or mobility-impaired users to digitally explore archaeological sites and museum environments, enhancing the inclusive enjoyment of heritage. Looking at the process from a cognitive perspective, the integration of sensory stimuli with active interaction generates immersive learning. Here, emotions, curiosity and engagement become the driving force behind understanding historical content, facilitating not only empathy towards the past, but also critical reflection on the facts. In essence, direct experience promotes a deeper and more conscious approach to the study of history (Falk and Dierking, 2018).

1.3.3 Some Case Studies

The use of virtual and augmented reality in museums and archaeological sites has given rise to a number of exemplary projects that show the potential of these technologies to transform the cultural experience in an immersive, interactive and accessible way. Below are some case studies that concretely illustrate how XR is being applied to enhance and reinterpret cultural heritage in new and engaging ways:

1. The Virtual Museum of Pompeii allows visitors to explore the ancient city before the eruption of AD 79 through a detailed reconstruction in virtual reality, based on archaeological surveys, historical sources and 3D models. The digital environment is integrated with narrative paths, dynamic visualisations and customised interactions, offering an immersive experience designed for both the general public and education. This project has been investigated in projects like Forte and Kurillo's (Forte and Kurillo, 2010), which analyse the relationship between VR and archaeological reconstruction.

2. The virtual reconstruction project of the Parthenon in Athens offers an immersive experience that allows visitors to explore the building as it was originally, restoring the lost monumentality and aesthetics of the site. Using advanced three-dimensional modelling, texturing and real-time rendering techniques, this simulation allows detailed exploration for digital tourists as well as scholars and students, becoming an emblematic example of heritage enhancement through VR. The case is documented and analysed, for example, by Economou and Pujol-Tost (Economou and Tost, 2011), in the context of the virtual museum experience.

3. The MANN museum has adopted innovative augmented reality solutions to enrich the museum experience, integrating 3D animated objects, virtual historical contexts and ancient Roman settings superimposed on the physical museum space. This approach allows for a multi-layered and augmented enjoyment of the collections, enhancing the understanding and engagement of the public, particularly the younger generations.

1.4 Hardware and Software Solutions for Virtual Reality: From Low-Cost to Immersive Structures

The development and adoption of VR experiences for cultural heritage and education depend crucially on the availability and accessibility of hardware and software tools. In particular, the choice of technologies employed profoundly influences the

level of immersiveness, interactivity and scalability of the designed solutions. From a hardware perspective, the distinction between low-cost and high-end devices is clear. The former, such as Google Cardboard, are affordable solutions that use a smartphone inserted into a cardboard frame. Despite their simplicity, these devices provide inexpensive access to virtual reality and are suitable for basic experiences such as guided virtual tours, panoramic views, or 360-degree videos. Of course, these devices have limitations in terms of performance and immersion compared to more advanced solutions, but they are still a valid starting point for those who want to approach the world of VR without incurring high costs (Suh and Prophet, 2018). These devices are particularly useful in education, museum or school settings where economic resources are limited, or for large-scale projects requiring mass distribution. Furthermore, their simple configuration and ease of transport make these devices ideal for mobile installations or temporary events. That said, they have significant technical limitations: resolution and graphics quality are often inferior to high-end models. Motion tracking is generally limited to three degrees of freedom, which prevents true interaction in three-dimensional space. They also lack advanced sensors for gesture recognition or haptic feedback, thus reducing the depth and engagement of the immersive experience. One of the main limitations of low-cost displays is their dependence on mobile hardware, such as smartphones or integrated chips, which inevitably imposes restrictions in terms of computing power, graphics performance and battery life. These limitations compromise the ability to run particularly complex or interactive applications, such as real-time archaeological simulations or adaptive environments that leverage artificial intelligence. In an attempt to overcome these critical issues, several projects are experimenting with modular and open source solutions that aim to combine low costs and advanced functionality, often integrating WebVR or WebXR technologies to allow access via a browser (Fanini, Ferdani, and Demetrescu, 2021). Despite these advances, the gap between low-end and high-end devices remains significant, particularly for applications that require faithful reconstruction of historical contexts or works of art characterised by a high level of detail.

On the other hand, high-cost systems, such as Meta Quest 2 and 3, HTC Vive Pro, Valve Index or Varjo XR-3, provide greater immersiveness thanks to precision sensors, advanced environmental tracking (inside-out and outside-in tracking), haptic controllers and support for spatialised audio. These headsets enable detailed three-dimensional reconstruction and natural interactions, while maintaining low latency and high resolution. These features make them particularly suitable for complex museum projects, interactive art installations, and educational and scientific research activities that require visual fidelity, advanced interactivity, and spatial accuracy. In museum contexts, for instance, devices like the HTC Vive Pro have been used to develop detailed archaeological reconstructions, immersive tour in virtual galleries or multisensory experiences where the visitor can interact with ancient objects, manipulate them virtually or activate contextual information content. In academia, on the other hand, high-end devices are used in collaborative projects, educational laboratories and interdisciplinary initiatives between the humanities and computer science, facilitating experiments that require high interactivity and accuracy in spatial representation. However, their high cost, which can easily exceed €1,500-2,000 per complete workstation, coupled with the need for dedicated physical space, high-end PCs for real-time rendering, and specialised technical personnel for configuration and maintenance, still constitutes a significant barrier to large-scale deployment. This situation creates a digital divide between institutions with adequate resources and smaller or peripheral realities, raising the issue of equity in

access to cultural technology (Slater and Sanchez-Vives, 2016).

Next to stand-alone displays are immersive CAVE (Cave Automatic Virtual Environment) infrastructures, which represent one of the most advanced solutions for the collective and interactive enjoyment of VR content. These are closed or semi-open physical environments in which virtual images are projected onto multiple surfaces (walls, floor and sometimes ceiling), offering users a 360-degree visual immersion experience. Visitors, equipped with stereoscopic glasses and controllers, can move freely within the space and interact with virtual objects displayed in full scale, making the experience particularly immersive and collaborative. Initially introduced by Cruz-Neira, Sandin, and DeFanti, 2023, CAVE systems have been progressively adopted in scientific, engineering, archaeological and museum contexts, especially for activities requiring high visual fidelity, spatial accuracy and simultaneous interaction between multiple users. Compared to individual VR headsets, CAVEs offer a crucial advantage in terms of shared experience: multiple people can simultaneously observe and discuss the same virtual environment, making them ideal for training sessions, interactive guided tours and co-creation projects in cultural heritage. Significant examples include CINECA's Virtual Reality Theater, one of the most advanced infrastructures in Europe for immersive simulation, also used for the virtual reconstruction of complex archaeological contexts; or the CAVE installed at the Egyptian Museum in Turin, which allows visitors and scholars to navigate within 3D reconstructions of tombs and artefacts, enriching the visitor experience with contextual content and customised interactions (Guidazzoli and Liguori, 2011). Furthermore, the 3DLab-Sicilia project has created CAVE structures in Palermo, Catania and Troina, which are also used to enhance Sicilian cultural heritage. These infrastructures are used for immersive reconstructions and the collective enjoyment of 3D content, promoting a unique educational and learning experience through virtual reality (Barbera et al., 2022b; Barbera et al., 2022a). Despite their potential, CAVEs still present major critical issues: the high cost of implementation, the need for dedicated physical environments, the constant maintenance of projection surfaces and tracking systems, and the complexity of integration with development software limit their use to large academic institutions, advanced research centres or museums with dedicated funding. With the growing accessibility of 3D modelling tools and game engines such as Unity and Unreal Engine, lighter and more portable solutions are increasingly emerging, enabling the development of modular or semi-immersive CAVE systems even for educational and cultural tourism applications.

From a software perspective, Unity 3D has emerged as a leading platform for developing immersive virtual reality environments. Its flexibility allows for the creation of interactive 3D spaces that are compatible with a wide range of VR displays, from budget-friendly models to high-end equipment. This versatility has contributed to its widespread adoption among developers in the field. Unity 3D is known for its relatively accessible interface, which allows even developers with limited skills to quickly prototype immersive experiences. Its large developer community and extensive online documentation make it an ideal choice, while numerous plugins and open source packages dedicated to virtual reality extend its functionality, making the software adaptable to various domains, including cultural and educational. Unity also allows the integration of sensors, haptic devices and spatialised audio engines, enhancing user immersion. Unity's adaptability also allows for educational applications, where content can be designed to engage visitors in customised paths based on individual preferences or real-time interactions. This results in interactive educational experiences that combine immersive technologies and storytelling, crucial for teaching history and cultural disciplines. Unity stands

out for its ability to integrate seamlessly with artificial intelligence tools, offering the possibility to develop sophisticated virtual agents, such as chatbots or interactive guides, capable of responding dynamically to user requests. This synergy enables the creation of augmented or extended reality experiences that go beyond static storytelling, actively adapting to user behaviour and preferences and enriching the educational journey with personalised content (Buldu et al., 2025). Unity's versatility and advanced features make it a preferred development engine in a variety of contexts: from the creation of virtual museums and simulations of historical environments to archaeological research projects. The platform also offers remarkable scalability, allowing users to move from small-scale projects, such as temporary virtual reality exhibitions, to complex multimedia installations, such as immersive CAVE environments or other dedicated physical structures.

1.5 Artificial intelligence and cultural heritage

In recent years, artificial intelligence has assumed an increasing role in the valorisation, conservation and enjoyment of cultural heritage, offering tools capable of automating complex processes, increasing accessibility and generating new forms of interactive storytelling. The adoption of AI in the cultural sector is articulated in numerous applications, ranging from the automatic classification of works to digital reconstruction and the creation of interactive and personalised content. One of the most established applications concerns the automatic recognition and classification of artworks and archaeological artefacts through computer vision and supervised machine learning techniques. Deep learning systems, in particular convolutional neural networks (CNNs), are now able to identify authors, artistic schools, historical periods, iconographic subjects or materials with increasing precision, often superior to human analysis in the context of large datasets. These technologies find application both in the automatic classification of digital collections and in supporting attributive research and the fight against falsification. One example is the ArtGAN project, which uses adversarial generative networks for the stylistic recognition of paintings and the automatic generation of images in the style of artists such as Van Gogh or Picasso (Tan et al., 2017). Similarly, the Art Recognition system has demonstrated the ability to attribute works of classical art with high reliability, testing patterns on paintings by Rembrandt, Botticelli or Bruegel, with an accuracy exceeding 90% on balanced datasets (Ugail et al., 2023).

In the archaeological field, similar techniques have been employed for the automatic classification of ceramic fragments through visual patterns or decorative imprints, as in the case of the SmartScan Project, developed by the Fraunhofer Institute, or for the identification of architectural styles in the remains of ancient buildings through structural image analysis (Mathias et al., 2012). The integration of these tools with digital interfaces in museums has resulted in semantic and visual search engines capable of suggesting works similar in content or style, thus facilitating the exploration of collections by visitors and scholars. Systems like those developed by Europeana and Google Arts & Culture represent such applications on an international scale. A rapidly expanding field is the automatic reconstruction of objects, buildings or historical environments from physical fragments, partial 3D scans, archival images or incomplete archaeological documentation. In this context, artificial intelligence, in particular deep neural networks, generative models and completion techniques, plays a decisive role in interpolating missing data, generating

reconstructive hypotheses and producing plausible three-dimensional models consistent with historical, stylistic and architectural constraints. These techniques have taken on a fundamental role in digital archaeology, especially in post-war conservation projects, where significant portions of cultural heritage have been damaged or destroyed. A typical example is the reconstruction of the Temple of Bel in Palmyra, Syria: using photos taken by visitors and satellite images, it was possible to create a digital reconstruction using artificial intelligence models integrated with photogrammetric methods (New Palmyra Project, 2016). One of the most cutting-edge initiatives in this field is the AI-Assisted Cultural Heritage Reconstruction framework, developed at ETH Zurich, which combines deep learning algorithms and shape completion techniques to digitally reconstruct damaged artefacts, even in the presence of particularly severe deformation or erosion. Similarly, the EU-funded DigiArt platform (LocalProBook, 2024) applied AI technologies to support the reconstruction and visualisation of archaeological sites, such as the Lascaux Cave, with a multimodal approach integrating point clouds, laser data and semantic analysis. The application of AI in this field also has a strong educational and museum relevance: reconstructed models can be visualised in VR or AR environments, made available to the public through virtual museums or online platforms, and used in digital restoration or in the simulation of alternative interpretative hypotheses by the scientific community.

Today, artificial intelligence is establishing itself as a key element in the transformation of the museum experience and the enjoyment of cultural heritage. These are no longer just static tools, but systems capable of actively interacting with the public. A notable example is chatbots based on natural language processing technologies: these tools respond in real time to visitors' questions, guide them through the exhibition and personalise the experience according to individual preferences. In this way, each visit can be tailored to the specific needs of each user, making the museum a more accessible and engaging environment.

Another innovative aspect concerns adaptive storytelling systems. Thanks to advanced artificial intelligence algorithms, these systems adapt the narrative based on users' choices or reactions, offering dynamic and personalised content. This approach not only enriches the visit but also contributes to making culture more inclusive, encouraging the interest and participation of an increasingly broad and diverse audience. For example, in 'The Nightwatch' project at the Rijksmuseum in Amsterdam, artificial intelligence guides visitors on an interactive narrative journey around Rembrandt's famous painting, where the story evolves based on the audience's actions, creating a unique experience with each visit (Caramiaux, 2023). Artificial intelligence systems are fundamentally reshaping the educational landscape of museums, elevating visitor engagement by blending educational content with elements of entertainment. This approach encourages active participation, effectively fostering dialogue across the domains of history, art, and technology. The integration of artificial intelligence-based tools into virtual museums and interactive galleries greatly expands the reachable audience. These innovations are particularly beneficial for people with different cognitive abilities or specific accessibility needs. For example, chatbots offer simplified access to information, minimizing the need for physical interaction, which is an advantage for those with motor or cognitive difficulties. Furthermore, AI actively promotes inclusivity through multilingual narratives, allowing content to be adapted to a variety of cultural and linguistic contexts. In doing so, museums are not only diversifying their audiences but also cultivating environments where access to knowledge is equitable and responsive to the needs of all visitors.

1.6 Immersive technologies and AI in cultural heritage education

Immersive technologies are profoundly redefining the way museum education and the transmission of cultural heritage are conceived. These tools allow visitors not only to observe, but also to experience historical and artistic environments firsthand, thus overcoming the limits of passive enjoyment and promoting experiential learning. At the same time, gamification, the application of playful mechanisms such as goals, scores, and challenges, is playing an increasingly central role in educational experiences in museums and archaeological sites. This approach is particularly effective in increasing motivation and participation, especially among young people, encouraging more conscious and independent exploration. The integration of gamification not only stimulates interest and curiosity, but also promotes situated learning, in which knowledge is developed through direct contact with the environment and objects, rather than through simple theoretical transmission (Gee, 2003). One example is the "A Journey Through the World of Ancient Egypt" project at the Egyptian Museum in Turin, Italy, which combines gamification with augmented reality to allow students to have an immersive experience of historical exploration.

Artificial intelligence is radically transforming the educational experience, enabling a more personalized approach. These systems analyze user preferences, behaviors, and interactions: in essence, they “learn” from the student. As a result, educational content can adapt in real time, offering personalized recommendations and interactive paths.

Take, for example, the use of intelligent chatbots: they are not just static machines for frequently asked questions. On the contrary, they engage visitors in conversation, tailoring responses to the individual’s background and interests. This dynamic exchange encourages curiosity and independent exploration, making the educational process not only more engaging but often more effective. In essence, students are not just passive recipients, but are actively shaping their own path through the educational material (Gaia et al., 2019). The integration of immersive technologies and artificial intelligence into cultural heritage education represents a significant shift in pedagogical approaches. These advances are transforming traditional learning methods, enabling interactive and engaging experiences that go far beyond passive observation. In particular, these innovations are proving highly successful among younger audiences, who are already expert navigators of digital environments and increasingly expect educational experiences to be interactive and entertaining. By combining educational content with immersive and hands-on activities, these technologies are fostering deeper engagement and a more meaningful connection with cultural heritage.

1.7 Critical issues and challenges

Although immersive technologies and artificial intelligence applied to cultural heritage offer innovative prospects, numerous technical and ethical-social challenges remain. A crucial aspect concerns accessibility and inclusivity: while virtual and augmented reality could, in theory, overcome physical and cognitive barriers, effective access to these tools is still limited. There are many reasons for this, from the high cost of devices to complex technological requirements, to designs that often neglect the needs of older users, people with disabilities, or members of socially disadvantaged groups.

These critical issues raise fundamental questions about the possibility of making immersive experiences truly universal and accessible to a diverse audience. In the absence of concrete solutions, there is a risk that innovation will remain the preserve of a minority, relegating a significant part of the population to a marginal role with regard to new forms of cultural heritage enjoyment (Llamazares De Prado and Arias Gago, 2023).

Digital preservation poses a significant challenge, particularly for XR and AI-driven cultural initiatives. These projects often rely on formats, software, and hardware that become obsolete at an alarming rate, jeopardizing the long-term accessibility of digital works. Additionally, the absence of unified standards and sustainable archiving methods further complicates efforts to safeguard these experiences for future generations. Given these risks, the development of robust policies and frameworks for the digital preservation of technologically mediated cultural heritage has become an urgent priority (Vincent et al., 2017;Siliutina et al., 2024).

Finally, from an ethical perspective, significant tensions emerge related to the concept of authenticity of aesthetic experience and human-machine mediation. Virtual reconstruction of environments or objects, although based on scientific data, always involves interpretive choices that may influence the perception of the past. There are thus questions about who holds the authority to represent historical "truth" and how to avoid distortions in the transmission of heritage (Champion, 2016). In addition, the introduction of intelligent agents in museums raises serious questions about the possible dehumanization of the relationship with knowledge: there is a real risk that interaction will be delegated to automated systems, sacrificing human dialogue and direct educational relationships (Tiribelli et al., 2024). To summarize, for these technologies to truly contribute to the enhancement and digital transmission of cultural heritage, it is essential to adopt a critical and interdisciplinary approach that balances innovation, sustainability, and social responsibility.

Chapter 2

Digitalization of cultural heritage sites

The digitalization of cultural heritage sites has become an essential practice for the documentation, preservation, enhancement, and sustainable accessibility of tangible heritage. This approach directly addresses significant challenges, including physical fragility, technological obsolescence, and the growing demand for remote and inclusive access. In this context, 3D data acquisition represents the first and fundamental step in transforming real-world artifacts into interactive and durable digital resources. Advanced technologies such as terrestrial laser scanning (LiDAR) and aerial and terrestrial photogrammetry now make it possible to produce geometrically accurate and navigable digital models which can be enriched with metadata. These representations capture the shapes, dimensions, and material details of monuments, architectural structures, and cultural landscapes with millimetric precision, fundamentally transforming the methods used for surveying, studying, and communicating cultural heritage (Remondino and El-Hakim, 2006). These tools have significantly expanded the operational scope of archaeologists, architects, museum curators, and digital developers, enabling new forms of interaction between users and cultural assets, even within fully virtual environments.

This chapter analyzes the main technologies and methodologies for 3D acquisition, with a focus on the characteristics of the sensors used, a comparison between laser scanning and photogrammetry, and the software employed to process and convert data into usable 3D models.

2.1 Introduction to geometric digitalisation

The geometric digitalisation of cultural heritage today represents one of the most effective tools for the conservation, study and valorisation of material assets. The possibility of digitally acquiring and reproducing an object or site in three dimensions makes it possible not only to protect it in the event of damage, but also to make it accessible, analysable and usable at a distance, facilitating new forms of cultural mediation, interdisciplinary research and public participation (Guidi, Beraldin, and Atzeni, 2004). In recent years, advancements in 3D surveying, particularly through Terrestrial Laser Scanning (TLS) and digital photogrammetry, have significantly transformed the digital documentation and reconstruction of cultural heritage. These techniques, often employed in tandem, facilitate the creation of exceptionally dense point clouds, sometimes comprising millions or even billions of spatially referenced data points. Such datasets capture the precise geometry and metric

attributes of architectural monuments, artifacts, archaeological sites, and even extensive landscapes. The resulting digital models offer an unparalleled level of detail, supporting rigorous analysis and long-term preservation efforts within the field of cultural heritage. TLS is an active remote sensing technique based on the emission of coherent light pulses (laser) and the measurement of the time it takes for the signal to strike a surface and return to the sensor, which is the so-called time-of-flight principle. This process enables the direct and highly accurate measurement of distances, which are then converted into three-dimensional spatial coordinates. Modern TLS systems, developed by companies such as Leica Geosystems, FARO Technologies, and RIEGL, can perform high-resolution acquisitions with millimetric precision. Furthermore, the integration of RGB or multispectral sensors allows for the acquisition of real surface colors, expanding the documentation from purely geometric to chromatic and material dimensions. From an operational standpoint, laser scanning proves particularly effective in complex, articulated, or large-scale environments, including historical monuments, religious buildings, underground spaces, and stratified archaeological sites. Its ability to collect non-contact data at high sampling rates and with uniform surface coverage makes this technique especially suitable for preventive conservation, digital reconstruction, structural monitoring, and the creation of immersive and interactive content. In parallel, digital photogrammetry has experienced significant advancements, primarily due to the development of Structure from Motion (SfM) and Multi-View Stereo (MVS) algorithms. Distinct from TLS, photogrammetry operates as a passive method, reconstructing three-dimensional geometry from standard photographic images captured at multiple viewpoints. Modern algorithms automatically identify pixel correspondences across images, allowing for the extraction of spatial information and the semi-automated production of dense point clouds, 3D mesh models, and highly accurate metric orthophotos. This shift has notably streamlined the process and expanded photogrammetry's practical applications in various fields. A key advantage of photogrammetry lies in its affordability and versatility: accurate results can be obtained using low-cost equipment such as DSLRs, mirrorless cameras, or even smartphones, making it particularly suitable for projects with limited resources or for distributed documentation initiatives. In addition, integration with Unmanned Aerial Vehicles (UAVs) has significantly extended the range of photogrammetry, allowing the capture of archaeological sites, tall facades, rooftops, and otherwise inaccessible areas.

Despite their distinct characteristics, the integration of TLS and photogrammetry has become increasingly common within sophisticated processing workflows. By combining these techniques, researchers are able to capitalize on the unique advantages of each: TLS provides precise geometric data and extensive spatial coverage, while photogrammetry offers detailed texture and color information. The synergy between these methods leads to outputs that excel in both metric reliability and visual fidelity. This integrated approach has been widely implemented in various national and international initiatives focused on the preservation, appreciation, and communication of cultural heritage.

These methods have also been applied within the framework of the 3D Lab Sicilia project, where terrestrial laser scanning and UAV-based photogrammetry were used as foundational steps for the development of 3D models intended for immersive and interactive virtual reality experiences. The acquired data were then processed for mesh generation and optimized for VR platforms, demonstrating the feasibility and replicability of the proposed pipeline for both scientific research and cultural dissemination purposes.

2.2 Terrestrial laser scanning

TLS represents one of the most reliable and accurate technologies for the three-dimensional acquisition of structures and objects in the field of architectural and archaeological surveying. The technology operates by emitting laser pulses, which reflect off surfaces and return to the sensor; by recording the time taken for this round trip, TLS systems are able to determine precise distances between the instrument and various points in the environment. The result of this process is the generation of exceptionally dense 3D point clouds, which capture the intricate geometry and details of surveyed structures with remarkable resolution and accuracy.

Conceptually, TLS scanners can be classified mainly into:

1. Static scanners: they are installed on tripods and acquire from fixed positions (Figure 2.1). They are suitable for high-precision surveys and for objects or environments where multi-angle coverage can be planned. Among the most widely used instruments in the cultural sector are the Leica ScanStation P50, the FARO Focus3D X130 and the RIEGL VZ-400i, which are known for their ability to acquire millions of points per second with millimetre accuracy even at a distance of hundreds of metres.

2. Mobile Scanners (Mobile Mapping Systems): integrate LiDAR sensors and inertial positioning systems mounted on vehicles, backpacks or wearable devices. These instruments, such as the ZEB Horizon GeoSLAM or the Trimble MX9, allow rapid acquisitions on the move, being particularly useful for complex interiors or large sites. However, the geometric quality of point clouds may be slightly lower than with static scanners, due to approximations caused by movement and lower acquisition stability.

In the field of cultural heritage documentation and enhancement, TLS technology has established itself as a fundamental tool for acquiring highly detailed and accurate three-dimensional surveys. Today, its applications are widely recognized and used in various fields. A significant example concerns the metric documentation of historic buildings prior to restoration work, as was the case for the Basilica of St. Francis in Assisi or the Colosseum in Rome, where geometric precision is essential for planning compatible and reversible conservation strategies. Similarly, TLS has also been particularly effective in modeling complex archaeological contexts, including stratified and large-scale sites such as Pompeii or the Valley of the Temples in Agrigento. Another important field of application concerns the digital replication of monumental works for museum or educational purposes. The resulting 3D models can be used in interactive reconstructions, virtual exhibitions, or augmented reality systems, offering innovative ways of accessing and interpretingn cultural heritage. Equally significant is the use of TLS for the generation of navigable 3D models, designed for immersive environments or VR experiences, which allow for highly engaging and remote exploration, even by users with physical or geographical limitations. The main advantages of this technology lie in its high metric accuracy (often below the millimeter), the ability to capture large datasets in relatively short timeframes, and the capability to record complex geometries, even in poorly lit environments or on textureless surfaces. Furthermore, continuous technical advances have enabled the latest generation scanners to cover vast areas while maintaining excellent spatial resolution and point density. However, the adoption of TLS also poses several challenges that can hinder its broader dissemination. First of all, it is an expensive technology, both in terms of professional-grade hardware (such as

FIGURE 2.1: An example of static scanner installed on tripods

devices produced by Leica, FARO, or RIEGL) and in terms of the software required for processing and point cloud management, such as Leica Cyclone, FARO Scene, or CloudCompare. In addition, the workflow demands highly skilled workers, both during field acquisition and in the subsequent data processing and interpretation phases. Another significant obstacle processing stage, is often time-consuming, particularly when working with large data sets. Finally, the quality of the result can be affected by environmental conditions, such as reflective surfaces, physical obstacles, dense vegetation, or unfavorable weather. Despite these limitations, TLS remains a key technology for the surveying and digital preservation of cultural heritage, especially when integrated with complementary methods such as digital photogrammetry, within a broader framework of multimodal and interdisciplinary approaches.

A particularly significant development in the recent evolution of survey technologies is the integration of TLS with photogrammetric techniques and the use of unmanned aerial vehicles, commonly known as drones. This multimodal approach allows for the synergistic exploitation of the strengths of each methodology: on the

one hand, TLS provides high metric accuracy and the ability to capture detailed geometric data, even under complex environmental conditions; on the other hand, photogrammetry ensures rich color information, dense textures, and greater operational flexibility, especially in areas that are less accessible to laser scanners, such as roofs or elevated architectural elements. The use of drones further enhances the potential of this integrated workflow, allowing large areas or elevated structures to be covered without the need for scaffolding or aerial platforms. This technical complementarity is particularly valuable in the field of cultural heritage, where the morphological and typological diversity of objects requires adaptable solutions across different scales and environmental contexts. The integration of TLS, photogrammetric, and UAV data is typically managed within software environments through point cloud registration and merging techniques, resulting in 3D models that are metrically accurate and at the same time visually realistic and expressive. These models are especially effective for applications ranging from scientific documentation to virtual musealization, from restoration planning to digital dissemination, and are increasingly emerging as a standard in both professional practice and academic research dedicated to the valorization of cultural heritage.

2.3 Photogrammetry

Digital photogrammetry, whether conducted from aerial or terrestrial perspectives, stands as one of the most widely adopted and adaptable techniques for creating three-dimensional models of cultural heritage assets. By analyzing photographic images, this approach enables the reconstruction of the geometry of objects, architectural structures, and even expansive territorial contexts. The process relies on automated processing of photographs captured with calibrated digital cameras or unmanned aerial vehicles, such as multirotor or fixed-wing drones.

In recent years, advancements in photogrammetry have been closely tied to the development of Structure-from-Motion and Multi-View Stereo algorithms. Together, these methodologies form the cornerstone of contemporary photogrammetric workflows, significantly enhancing the precision and efficiency of 3D reconstruction in cultural heritage documentation (Remondino et al., 2014;Westoby et al., 2012).

SFM is an advanced method in the field of computer vision used to estimate both the spatial position of the camera and the three-dimensional structure of the scene, starting from a series of overlapping images acquired from different, initially unknown viewpoints. The process is based on the automatic identification of corresponding points between the various images—typically using algorithms such as SIFT or SURF. From these homologous points, both the initial positions of the camera and a sparse 3D reconstruction of the scene, commonly known as a "sparse point cloud," are obtained (Snavely, Seitz, and Szeliski, 2008). Once the sparse reconstruction is complete, the MVS stage comes into play, significantly increasing the model's density. MVS relies on the known camera positions obtained from SfM and matches pixels across the input images, resulting in a dense point cloud—often comprising millions of 3D points. This dense cloud is then processed into a triangulated mesh and textured with photorealistic images, yielding a model with high geometric detail and accurate color representation.

The synergy between SfM and MVS forms the foundation of most contemporary photogrammetric software solutions. Both open-source options (such as COLMAP, Meshroom, OpenMVG+OpenMVS) and commercial platforms (like Agisoft Metashape, Pix4D, and RealityCapture) (Luhmann et al., 2023) utilize this integrated approach.

As a result, advanced 3D documentation tools have become far more accessible to researchers, conservators, and cultural institutions, broadening the potential for high-quality digital preservation and analysis. In archaeological applications, SfM and MVS have been successfully used to reconstruct detailed epigraphic surfaces and to document stratigraphic layers in complex excavation sites. In architectural heritage, these methods are frequently adopted to produce digital twins of historical buildings, which can support both conservation efforts and public dissemination through online visualization or immersive virtual tours.

In this context, terrestrial photogrammetry involves the use of DSLR or mirrorless cameras, either mounted on tripods or operated handheld, following preplanned image acquisition paths to ensure homogeneous coverage and high overlap between photographs, typically exceeding 70–80%. This method is particularly well-suited for close-range documentation of architectural details, sculptural surfaces, museum artifacts, or indoor environments where access may be limited or lighting conditions variable.

In addition to traditional photographic equipment, integrated 3D scanners, such as Matterport cameras (Figure 2.2), have become increasingly popular in the documentation of cultural heritage. These devices combine RGB imaging with depth sensing technology to generate complete 3D models in a single acquisition workflow. Matterport systems, for example, enable rapid scanning of architectural interiors, museum artifacts, or archaeological structures, producing photorealistic and metrically accurate models with minimal post-processing effort. Unlike classical photogrammetry, which relies solely on image overlap, Matterport and similar devices simultaneously capture geometry and texture, bridging the gap between passive and active sensing techniques.

Aerial photogrammetry essentially involves the use of drones equipped with high-resolution cameras to survey large or complex areas that are difficult to reach, such as roofs, archaeological ruins, excavation sites, or extensive cultural landscapes. This type of acquisition allows for rapid coverage and provides an extremely detailed overview of the context being analyzed. When the collected data is accurately georeferenced, highly accurate spatial information is obtained, which can then be integrated with terrestrial surveys to produce complete and highly detailed 3D reconstructions.

The data obtained through photogrammetry achieves remarkable levels of detail, not only in terms of shape, but also in terms of the amount of radiometric information acquired, such as color, texture, and surface reflectance. This combination makes it possible to produce 3D models that are both metrically accurate and visually realistic. These characteristics are particularly relevant in areas such as scientific documentation, virtual visualization, and public engagement. High-resolution textures, obtained from accurate photographic sets, allow for the faithful reproduction of essential aspects of materials: patinas, signs of wear and inscriptions. These details are fundamental for the study and effective communication of cultural heritage. Compared to TLS, photogrammetry has some big technical differences and a few operational limitations that should not be underestimated. First, it's really sensitive to environmental conditions: the quality of the images can be seriously affected by poor lighting, reflective or transparent surfaces (like glass, polished stone, or water), and shadows cast by surrounding architectural elements. The accuracy of the resulting 3D reconstruction is strongly influenced by the resolution of the photographs, their degree of overlap, and the accuracy of the camera calibration.

FIGURE 2.2: An example of Matterport camera in action

However, despite these critical issues, photogrammetry offers significant advantages in terms of cost, portability, and ease of use. Unlike TLS, it does not necessarily require highly specialized or bulky equipment: in most cases, a standard digital camera or UAV, combined with widely available software, is sufficient. In summary, despite some technical limitations, photogrammetry is an accessible and versatile solution for numerous applications. In real-world contexts, photogrammetry and TLS are often used synergistically within advanced 3D surveying pipelines. This multimodal strategy allows the strengths of each technique to be exploited: TLS guarantees amazing geometric accuracy, thus turning out to be effective even in unfavorable lighting conditions or on surfaces lacking texture; photogrammetry, on the other hand, enriches the model with visual and chromatic details, improving the readability and communicative impact of the final result. When the two data sources are properly integrated and aligned, hybrid models are obtained that combine metric reliability and photorealism, proving to be robust tools for the preservation, analysis, and dissemination of cultural heritage. This integrated approach

is now an established practice in more complex documentation projects, especially when the variety and size of the site require flexible, multi-level solutions. A typical example concerns the documentation of architectural structures: TLS is used to acquire the overall geometry of the building, while photogrammetry focuses on decorative elements or painted surfaces, capturing textures and color nuances that are not accessible to LiDAR sensors alone. Together, these data produce high-fidelity digital models that are useful for scientific research as well as for communication and heritage enhancement.

Photogrammetry has significantly transformed the documentation and preservation of cultural heritage within the fields of archaeology and museology. Its capacity for detailed, non-invasive three-dimensional recording has proven indispensable, particularly in complex or fragile contexts. For instance, at sites such as Petra (Jordan) and Palmyra (Syria), both subject to substantial damage from conflict and natural deterioration, photogrammetry has enabled the digital preservation of threatened structures. This technology not only safeguards existing remains but also provides critical data to support virtual reconstruction hypotheses, marking a considerable advancement in heritage conservation and research practices (Fassi et al., 2013). In Italy, the Valley of the Temples in Agrigento represents a significant case study, where the integration of aerial and terrestrial photogrammetric surveys allowed the creation of high-resolution 3D models of Doric temples. These models are used both for research and conservation purposes and for public engagement through virtual tours and immersive installations (Di Salvo and Brutto, 2014). In the museum domain, photogrammetry is now systematically adopted for the digitization of permanent and temporary collections. This process allows the creation of virtual reproductions of delicate or otherwise inaccessible artifacts and supports the digital reconstruction of historical interiors, recontextualized archaeological contexts, or architectural spaces that are now lost. Furthermore, thanks to the compatibility of photogrammetric models with real-time 3D development platforms such as Unity 3D and Unreal Engine, the data can be integrated into interactive and immersive experiences. These include VR and AR applications, gamified educational pathways, and on-site multimedia installations (Bekele et al., 2018).

Alongside commercial solutions that offer advanced tools for processing large datasets and generating textured meshes, several open-source options have made this technology accessible to researchers, small museums, and independent projects. This transformation has reinforced photogrammetry as a fundamental element in the modern digitization of cultural heritage. The technology now offers not only reliable metric accuracy and high-quality images, but also financial feasibility and significant opportunities for public engagement.

2.4 Data processing software

Photogrammetric processing platforms play a key role in creating three-dimensional models derived from photographic datasets and are essential for the digital documentation of cultural heritage. Among the most widely used software solutions, Agisoft Metashape, RealityCapture, and 3DF Zephyr are each of which is recognized for their robust functionality, accessibility, and automated features. Agisoft Metashape is one of the most popular commercial solutions, renowned for its reliability, result accuracy, and user-friendly interface. The software supports the entire photogrammetric workflow, from automatic image alignment to dense point cloud generation,

mesh construction with texture mapping, and finally orthorectification and georeferencing of the 3D models. Its ability to balance performance with relatively modest hardware requirements is worthy of note, as is its flexibility in processing combined datasets from both aerial and terrestrial imagery. Additionally, Metashape natively supports the integration of Ground Control Points (GCPs), which is essential for producing georeferenced and metrically accurate models. RealityCapture, developed by Capturing Reality (now part of Epic Games), is distinguished by its remarkable processing speed and ability to handle extremely large datasets within short timeframes. Its highly optimized computation engine leverages GPU acceleration to perform Structure-from-Motion and Multi-View Stereo tasks efficiently. The combination of high-quality outputs and seamless export options for immersive environments makes it particularly suited to projects that require both metric precision and high-end visual presentation. 3DF Zephyr, developed by the Italian company 3Dflow, offers a strong alternative thanks to a user-friendly interface that makes it accessible to non-expert users while still providing a high level of control over photogrammetric parameters. The software includes advanced features such as batch processing, camera calibration, hybrid alignment, and support for standardized formats (e.g., E57, LAS), which have promoted its adoption in both academic and professional contexts. Zephyr also offers dedicated tools for model cleaning and optimization for web or mobile deployment.

Processing data collected through TLS requires a structured processing workflow to transform raw point clouds into accurate, clean, and usable 3D models for measurement, documentation, or visualization purposes. Several specialized software platforms are available for LiDAR data processing, including FARO Scene, Leica Cyclone, and CloudCompare, each offering specific features and advantages depending on their intended use. FARO Scene, for example, is the native software for managing data acquired with FARO Focus scanners. Its intuitive interface supports semi-automated registration of multiple scans, using advanced target recognition and cloud-to-cloud alignment algorithms (Figure 2.3). Scene enables data export in several standard formats (such as E57, LAS, or OBJ) and includes tools for cloud segmentation, extraction of orthogonal sections, and immersive visualization in VR environments through FARO WebShare. It is widely adopted in architectural and cultural heritage applications due to its efficiency in processing large datasets.

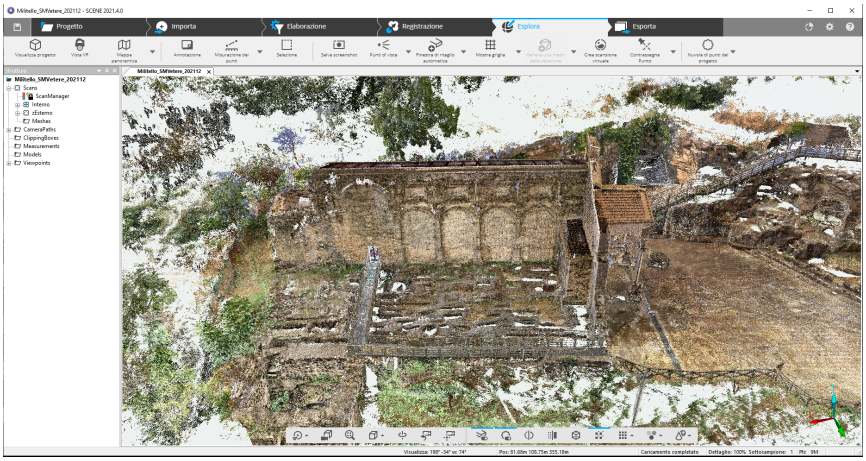


FIGURE 2.3: An example of a point cloud on Faro Scene

Leica Cyclone is one of the most comprehensive and powerful professional solutions for processing point clouds acquired with Leica scanners. The software consists of a series of specialized modules, each designed to address a specific stage of the workflow: from registration (Cyclone REGISTER 360) to editing, CAD model extraction, surface reconstruction, and meshing. It is particularly valued for its high-precision georeferencing capabilities, support for integrating auxiliary topographic data, and compatibility with BIM workflows. However, it typically requires more advanced expertise and presents a steeper learning curve than other platforms. CloudCompare is a robust and adaptable open-source platform widely used in academic and research contexts for point cloud processing, analysis, and comparison. Its strengths include broad compatibility with 3D data formats, advanced filtering tools, dataset comparison (cloud-to-cloud and cloud-to-mesh), scalar field coloring based on geometric properties, and the generation of profiles and cross-sections. Although it does not offer a fully automated pipeline, CloudCompare differs in its detailed control of processing parameters and the transparency of its operations, both of which are essential in scientific and diagnostic contexts. In addition to CloudCompare, another widely used open-source software is MeshLab. MeshLab allows the import, visualization, and processing of meshes and point clouds, providing tools for filtering, simplification, mesh repair, normal calculation, and texturing. Its modular interface and wide range of plugins make it a flexible tool for both preprocessing and optimizing 3D models for scientific, educational, or immersive visualizations. The combined use of these tools enables full management of the LiDAR data lifecycle: from field scan registration, cleaning, segmentation, and alignment, to the generation of exportable 3D models that can be integrated into CAD, GIS, or virtual reality environments. The selection of the most suitable software depends on the type of scanner used, the project objectives, and the desired level of automation.

In the context of digital documentation of cultural heritage, the integration of data acquired through different surveying techniques, most notably photogrammetry and terrestrial laser scanning (LiDAR), is a crucial phase for the construction of complete, consistent, and metrically accurate 3D models (Grussenmeyer et al., 2008). These datasets, often generated as high-density point clouds, must undergo a rigorous preprocessing workflow that includes spatial registration, semantic and geometric cleaning, and final data fusion. Registration of point clouds from different sources is typically carried out through iterative alignment algorithms such as Iterative Closest Point (ICP), which minimize the Euclidean distances between corresponding points to optimize the relative positioning of datasets (Besl and McKay, 1992). Alternatively, artificial targets or ground control points measured with total stations or GNSS can be used to ensure alignment within a global coordinate system (Barazzetti, Remondino, Scaioni, et al., 2010). The accuracy of this phase is particularly critical in complex or stratified architectural and archaeological contexts. After registration, the datasets are subjected to point cloud cleaning process, which involves the removal of outliers, spurious surfaces, and artifacts caused by reflections, transparency, or environmental noise. This process can be achieved through statistical filters, spatial segmentation, or manual editing in dedicated software such as CloudCompare or 3DReshaper. Cleaning is especially important in surveys conducted in urban or museum environments, where temporary elements (e.g., moving people, temporary objects, vegetation) are common. Once the point clouds have been preprocessed, the next step is the generation of a triangulated mesh. This involves reconstructing a surface from the point cloud using algorithms such as Delaunay triangulation or Poisson surface reconstruction (Kazhdan, Bolitho, and Hoppe,

2006), the latter being particularly effective in producing continuous, watertight surfaces. The resulting 3D mesh thus provides a geometrically accurate representation of the surveyed object or environment. The mesh is then enriched through photorealistic texturing, using images captured during the acquisition phase. This procedure, known as texture mapping or image projection, transfers radiometric information, such as color, brightness, and surface patterns, onto the 3D geometry, significantly enhancing visual readability. Advanced workflows may also apply multi-image blending techniques to correct for lighting inconsistencies or shadows, resulting in visually coherent representations. The final output is a highly detailed and metrically reliable 3D model, suitable not only for scientific documentation, but also for dissemination through immersive platforms, virtual environments, and Heritage Building Information Modeling (HBIM) systems (Murphy, McGovern, and Pavia, 2009). When preparing these models for deployment in web-based interfaces or real-time engines, such as those used in virtual and augmented reality, an additional step is often necessary: polygonal optimization and decimation. This process, also known as retopology, reduces geometric complexity while retaining essential surface characteristics. Techniques like normal mapping or displacement mapping are employed to preserve visual fidelity, resulting in models that are both lightweight and rich in information content.

Chapter 3

Digital solution for immersive virtual cultural heritage exploration

The digitalisation process is not limited to simple acquisition: it implies an integrated pipeline that starts from data collection (e.g., point clouds and photographic datasets), passes through the generation of 3D meshes and photorealistic textures, and finally ends with the development of interactive and immersive environments using real-time engines such as Unity 3D or Unreal Engine. These environments enable the creation of virtual tours, educational simulations, and practical tools for preventive conservation and risk management.

In recent years, the digitalisation of cultural heritage has evolved into a complex, multi-step workflow that combines technologies such as photogrammetry, LiDAR, 3D modelling, and immersive visualization. These technologies do not merely reproduce the visual appearance of objects and sites, but produce digital models that are rich in information, metrically accurate, georeferenced, and semantically structured. Such models can be reused in various fields (research, restoration planning, museum exhibitions, or virtual access), while ensuring long-term preservation and accessibility (Remondino and Stylianidis, 2016).

A concrete example of how this pipeline can be implemented is offered by the *3D Lab Sicilia* project, which applied advanced acquisition techniques, polygonal modelling, and immersive visualization to the conservation of architectural assets. In this case, the digital workflow was tested and adapted to real-world heritage contexts, highlighting both the technical challenges and the potential for innovative applications (Barbera et al., 2022b; Barbera et al., 2022a; Lo Bosco, Farella, and Di Piazza, 2023). Rather than being an isolated initiative, this project exemplifies how an integrated pipeline can support the documentation, preservation, and valorisation of cultural heritage in diverse settings.

3.1 Pipeline: from reality to 3D model

For the development of virtual environments in the field of cultural heritage, a structured implementation pipeline has been defined to guide the creation of explorable 3D models. The process, illustrated in Figure 3.1, is designed to ensure that cultural heritage can be accurately documented, digitally reconstructed, and integrated into immersive applications. The workflow is articulated into five main stages. First, the cultural site is digitally acquired using advanced techniques such as photogrammetry or LiDAR. The resulting data are then processed to clean, align, and optimize the point cloud. From this dataset, a complete 3D model is generated, including geometry, textures, and other key attributes. Finally, the optimized model is imported into Unity3D, the game engine used to develop immersive applications on different

platforms, including smartphones with cardboard displays, standalone VR headsets, and CAVE environments. This workflow provides a flexible foundation that can be applied to a variety of contexts. One concrete application is the creation of virtual environments aimed at improving accessibility to archaeological sites within the Etna Park and several Sicilian municipalities recognized as UNESCO World Heritage Sites. In these cases, the pipeline ensures that complex and fragile sites can be experienced virtually, preserving both scientific accuracy and immersive quality.

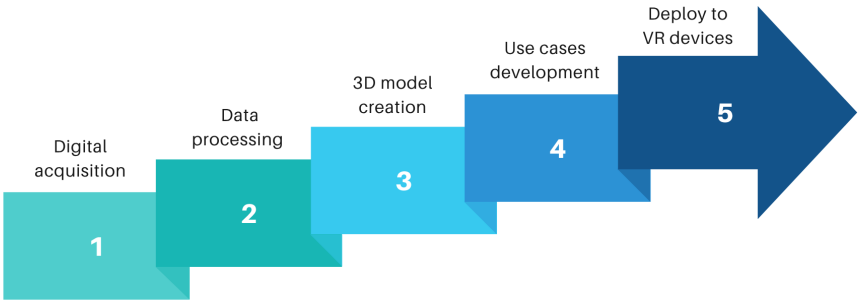


FIGURE 3.1: The proposed pipeline

3.1.1 Digital acquisition

A key step of the implementation pipeline is the data acquisition, which provides the foundation for all subsequent stages of 3D reconstruction and immersive visualization. Before a cultural asset can become a 3D model ready for immersive exploration, it must first be captured in all its detail and complexity. For this purpose, the pipeline makes use of a hybrid methodology, which combines SfM photogrammetry, based on photographs taken both at ground level and by drone flights, with LiDAR laser scanning. This mixed approach allows for obtaining a very complete dataset: photogrammetry contributes rich textures and visual fidelity, while LiDAR ensures millimetric accuracy in geometry. Together, these techniques provide a robust foundation for producing dense point clouds that reflect the real-world object with high precision.

The acquisition phase is not immediate or simple. It requires careful planning of scanning stations to guarantee continuity between indoor and outdoor areas, complete coverage of architectural details, and an adequate overlap between different scans. To support this, reflective spheres are strategically placed around the site: these act as reference points that facilitate the subsequent alignment of scans, ensuring that the different fragments of the survey can be recomposed into a coherent whole. This step is essential to minimize alignment errors and to address common difficulties, such as occlusions, missing or degraded surfaces, or complex geometries that can confuse even the most advanced instruments. It is not enough just to collect data. The raw point clouds generated by these instruments can reach enormous sizes, often exceeding 100 GB. This makes them unsuitable for direct use in VR or

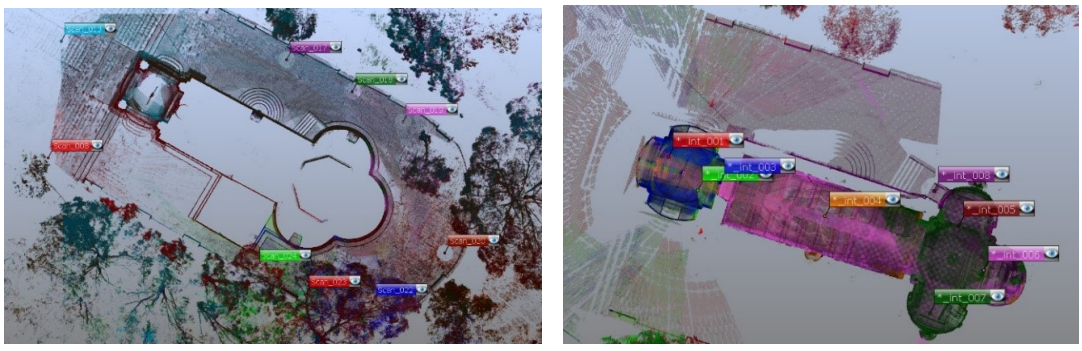
interactive systems. For this reason, post-processing is required: removing artifacts, simplifying geometries through controlled decimation, and optimizing topologies to make the models both lightweight and faithful to reality. Only by carefully balancing accuracy and efficiency can these models be transformed into usable assets for immersive experiences.

This pipeline has been put into practice in different cultural heritage contexts. Two examples are the Church of Santa Maria la Vetere (Figure 3.2a) and the Church of the SS. Crocifisso del Calvario, in Militello, Val di Catania (Figure 3.2b).

(A) Church of Santa Maria la Vetere (B) Church of the SS. Crocifisso del Calvario

FIGURE 3.2: Some cultural assets scanned

For the Church of Santa Maria la Vetere, the survey involved no fewer than 49 external and 9 internal scanning stations, covering elements such as the surviving nave, the ancient central portico, the surrounding courtyard with visible pit tombs, and even a Norman turret. The Church of the SS. Crocifisso del Calvario, instead, required a different strategy: 16 external and 8 internal stations distributed at regular intervals, with denser coverage near the apses and transition areas (Figure 3.3). In both cases, the use of reflective spheres and meticulous alignment procedures allowed residual registration errors to be kept below one centimeter, a level of accuracy that testifies to the robustness of the approach.



(A) Laser scanning 3D Survey of the outside: point of station (B) Laser scanning 3D Survey of the inside: point of station

FIGURE 3.3: Church of the SS. Crocifisso al Calvario point of station

These case studies highlight the practical value of the pipeline: what begins as a highly technical process of acquiring and processing billions of points ultimately becomes the basis for immersive environments that can be explored with VR headsets or CAVE systems. In addition to the immediate experience of visiting a site in virtual form, these digital resources are also reusable: they can support conservation

and restoration planning, structural analysis, educational outreach, or even experimental applications in multi-user environments. In this way, data acquisition is not just a technical phase, but the basis of a broader process of cultural heritage enhancement, capable of preserving fragile sites while opening them up to new audiences through immersive technologies.

3.1.2 Data processing

A critical step in the pipeline is the processing of the collected data, which transforms raw acquisitions into usable 3D assets ready for interactive visualization. This stage is not merely technical refinement: it is what ensures that the models are accurate, manageable in size, and visually effective in immersive environments. The integration of multiple techniques, photogrammetry and laser scanning, requires a carefully designed workflow to reconcile datasets with different characteristics and levels of detail. Photogrammetric surveys typically provide rich textures and broader coverage, while laser scans deliver geometric precision and high point density. By combining them, the resulting models achieve a balance between visual realism and metric reliability, which is crucial for both scientific documentation and VR deployment. During post-processing, both types of data must be harmonized. Photogrammetric datasets, constructed from hundreds of aerial and terrestrial images, must be aligned and optimized to generate dense point clouds and textured meshes. Point clouds obtained through laser scanning, which are often very large and detailed, must be registered, cleaned of noise, and organized into manageable clusters. Integrating these sources generally involves control point alignment, in which homologous features between different point clouds are matched to achieve accurate registration. To avoid redundant data and inconsistencies, selective filtering is applied: areas best captured by photogrammetry are retained, while areas where laser scanning excels are prioritized for accuracy. This ensures that the unified model reflects both geometric quality and efficiency, minimizing overlaps or conflicts between datasets. Another central challenge lies in data management. Point clouds can easily reach hundreds of gigabytes, making them unsuitable for direct use in immersive applications. As a result, segmentation and decimation become indispensable. Meshes must be simplified while preserving essential details, and textures must be processed to guarantee clarity without overwhelming computational resources. The workflow therefore often requires the integration of multiple tools, such as Faro Scene, Meshlab, CloudCompare, and Zephyr, each contributing specific strengths to optimize geometry, textures, and alignment. In some cases, particularly intricate architectural elements must be processed separately to preserve their readability before being recombined into a complete, multi-resolution model.

For the Church of Santa Maria la Vetere, approximately 500 images captured by both ground-based cameras and drones were processed in Zephyr to build a photogrammetric model. At the same time, scans acquired with a Faro Focus 350S laser scanner were managed and recorded using Faro Scene, then exported in .e57 format and imported into Zephyr. The integration of the two datasets was based on a control point alignment procedure, ensuring consistency between the photogrammetric and LiDAR sources. To avoid redundant overlaps, a selective filtering approach was adopted: the external context and roof were retained from the photogrammetric model, while the facades, interiors, and portico were retained from the LiDAR model due to their superior accuracy.

The subsequent mesh and texture processing involved comparing and optimizing results across different tools. Because of the building's complexity, standard

meshing workflows proved inadequate for maintaining detail, especially in religious architectural elements. For this reason, segmentation and controlled decimation were applied, with some parts such as the altar, the portals, and the external portico with bas-reliefs treated individually (Figure 3.4). These isolated point clouds were treated separately and then recombined into a comprehensive model with differentiated levels of detail. Particularly effective results were obtained with a hybrid workflow combining Scene and Meshlab, where meshes generated in Scene were re-positioned on the point cloud imported into Meshlab, producing a final model optimized for VR.

FIGURE 3.4: Santa Maria La Vetere Church, the individual elaborations of the altar, the arch and the portals

For the Church of the SS. Crocifisso al Calvario, the workflow followed a similar logic (Figure 3.5). The scans were organized into logical groups within Faro Scene, achieving a calibration error margin of less than one centimeter. Here too, laser scanning data were integrated with photogrammetric imagery, used mainly to reconstruct the roof and external surroundings. Special emphasis was placed on modularity and computational efficiency, so that the resulting model could be easily deployed on standalone headsets or mobile devices without sacrificing quality.

Taken together, these experiences highlight how data processing is an important and complex stage in the process, directly influencing the usability of models in VR environments. Beyond the technical aspects of point cloud alignment or mesh decimation, the real challenge lies in finding the right balance: preserving rich detail while ensuring performance in real-time applications. The use cases reported demonstrate how established workflows can be adapted and customized according to the architectural complexity of each site and the intended digital use, ensuring scientifically reliable models suitable for immersive dissemination.

3.1.3 3D model creation

After data acquisition and processing, another equally important step is the generation of the final 3D model, which transforms raw point clouds into objects that are accurate in terms of documentation and optimized for interactive visualization. This

FIGURE 3.5: Digital acquisition and data-processing of SS. Crocifisso al Calvario Church

stage requires a balance between geometric fidelity, computational manageability, and visual realism, ensuring that the final product can serve multiple purposes: scientific analysis, conservation planning, dissemination, and immersive exploration. Achieving this goal typically involves key steps: cleaning the dataset, reconstructing the mesh, simplifying the model to make it suitable for real-time environments, and finally applying UV mapping and texturing. Each step relies on specialized software tools with distinct strengths: Meshlab, which, as we mentioned earlier, is an open-source platform particularly effective for editing, filtering, and decimating meshes, and Blender, a professional-grade 3D environment widely used for UV mapping, advanced texturing, and rendering preparation. Together, these tools enable the creation of high-quality multi-resolution models that can be used in immersive applications.

In the case of the Church of SS. Crocifisso al Calvario, the following steps were taken during the creation of the 3D model:

1. Point Cloud Cleaning: The initial point cloud, although obtained with high-precision instruments, contained unwanted elements such as benches, scaffolding, or reflective targets, which could interfere with the clarity and usability of the final model. To address this, the dataset was processed in Meshlab, which offers a wide range of filters and manual editing functions specifically suited for mesh and point cloud management. The cleaning phase combined both manual and automated strategies: on the one hand, vertex and face selection tools made it possible to directly isolate and remove non-architectural objects; on the other, the Select Outliers filter was applied to automatically detect points whose distances significantly deviated from their local neighbors, identifying them as potential noise. To preserve flexibility and ensure non-destructive editing, all removed elements were not permanently deleted but transferred into separate layers, thereby keeping them available for possible reuse in alternative workflows. This careful cleaning step laid the groundwork for a reliable and geometrically consistent model, ready for the subsequent phases of reconstruction and optimization.

2. Mesh Reconstruction: The cleaned point cloud was subsequently converted

into a polygonal mesh using Meshlab's Screened Poisson Surface Reconstruction filter. This algorithm is particularly well-suited for cultural heritage modeling, as it generates smooth, watertight surfaces even when the input data contains gaps or imperfections. A reconstruction depth of 13 was selected, striking a balance between capturing fine surface details and maintaining computational efficiency, thereby avoiding excessively long processing times. Other parameters, such as the interpolation weight and samples per node, were left at their default values to ensure the stability of the reconstruction process. The resulting mesh consisted of over 240 million faces, providing a level of detail sufficient for precise documentation and analysis, even of the most intricate architectural elements, while preserving geometric continuity across the entire surface.

3. Mesh Simplification: Models of this size are generally unsuitable for real-time rendering due to their extremely high polygon count. To address this, a controlled decimation was performed in Meshlab using the Quadratic Edge Collapse Decimation algorithm. The target was to reduce the number of faces to approximately one million, striking a balance between visual fidelity and computational performance. Both the "Preserve Normals" and "Preserve Topology" options were enabled to prevent artifacts and ensure that the simplified geometry retained its structural integrity. Additionally, the quality threshold was set to 1, promoting a uniform simplification across the entire surface. Throughout this process, Meshlab's visual inspection tools, including wireframe overlays and curvature analysis, were employed to carefully monitor any loss of detail and to guarantee that the reduction was consistent while preserving the essential features of the model. This phase produced a lightweight but still visually detailed mesh, appropriate for immersive applications and web-based visualization platforms.

4. UV Mapping and Texturing: Once the geometry had been optimized, the model was prepared for visualization by projecting 2D photographic data onto its surface. This process was carried out in Blender, which provides more advanced UV mapping tools than Meshlab. The Smart UV Project operator was used to automatically unfold the geometry and generate a UV map, minimizing texture stretching and ensuring efficient use of texture space. Blender's UV/Image Editor then allowed for manual adjustments to refine the projection, particularly in areas with complex or intricate geometry, improving alignment and accuracy (Figure 3.6). After completing the UV mapping, the simplified mesh was exported in .OBJ format, including the UV coordinates. To apply realistic textures, the model was re-imported into Meshlab along with the original point cloud. Two key operations were used: Convert PerVertex UV into PerWedge UV, which adapts the UV layout to facilitate texture application, and Transfer: Vertex attributes to texture, which projects the RGB color values from the point cloud onto the UV coordinates. The resulting texture was generated at a resolution of 16,384 × 16,384 pixels, preserving subtle chromatic variations and surface details essential for both immersive visualization and rigorous scientific documentation.

Thanks to this integrated workflow, it was possible to obtain a detailed three-dimensional model, optimized both geometrically and visually, ready to be used for dissemination, scientific analysis, and immersive fruition of the cultural heritage asset.

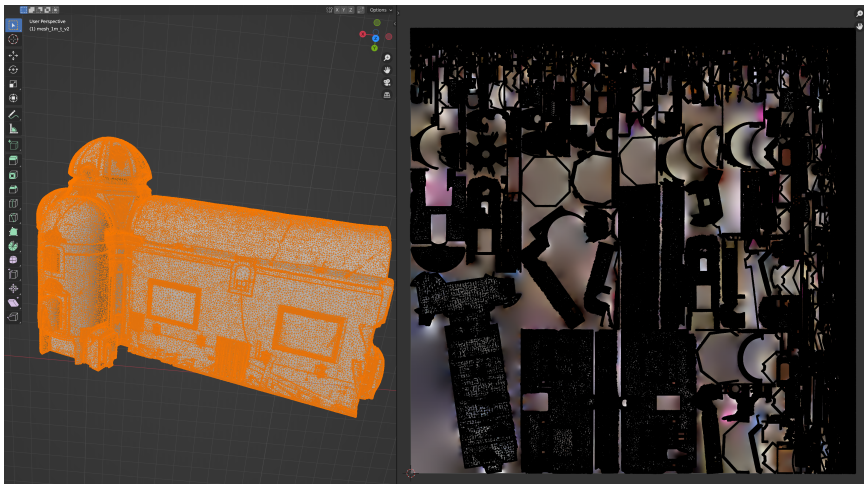


FIGURE 3.6: UV Mapping and Texturing using Blender of SS. Crocifisso al Calvario Church

3.1.4 Immersive experience and digital enhancement

The immersive system was designed to provide a versatile framework for exploring cultural heritage sites, leveraging the strengths of multiple platforms and devices. Unity 3D was chosen as the primary development environment due to its robust real-time rendering capabilities, cross-platform compatibility, and native support for virtual reality. The system architecture was conceived to deliver highly realistic and intuitive experiences, enabling users to navigate digital replicas of cultural spaces while examining architectural and decorative details with precision. Special attention was paid to designing interaction models and optimizing performance for different hardware, ranging from low-cost smartphones to high-end multi-wall immersive systems.

To achieve this, the development process involved several key strategies. For VR platforms with limited computational resources, such as smartphones, techniques like texture downsampling, light baking, and lightweight shaders were implemented to maintain smooth frame rates and visual consistency, minimizing the risk of VR-induced discomfort. Conversely, for standalone VR headsets, the system could exploit more advanced hardware capabilities, enabling natural hand-tracked interactions, precise object manipulation, and immersive locomotion. High-fidelity visuals were preserved through careful optimization of lighting, shadows, and texture resolution, while motion tracking and input handling were implemented using specialized Unity plugins like the Oculus XR Plugin and the XR Interaction Toolkit. For large-scale CAVE systems, the architecture included networked synchronization across multiple workstations, real-time rendering of multiple projection surfaces, and high-resolution textures with dynamic lighting, ensuring an immersive experience that maintained both spatial realism and responsive interactivity.

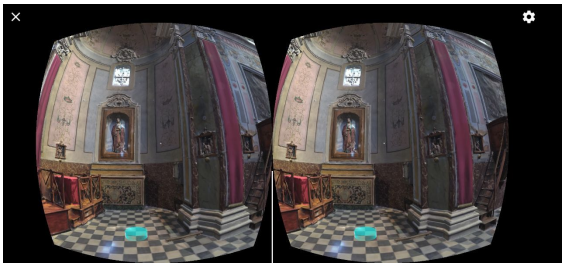
This general framework is applied to the 3D models obtained from the previous stages to explore them virtually, as in the case of the Church of SS. Crocifisso al Calvario ((Figure 3.7)) and the Church of Santa Maria La Vetere ((Figure 3.8)). The elaborated 3D model was integrated into three distinct platforms, each tailored to the capabilities and constraints of the target hardware. On entry-level smartphones, paired with Google Cardboard, the system prioritized accessibility and ease of use. Since Cardboard lacks physical input devices, a gaze-based interaction system was

implemented: key points of interest within the virtual environment were marked with interactive "trigger points." When a user fixates on a trigger for a set duration, they are automatically teleported to that location, allowing intuitive navigation without external controllers. Trigger points were carefully positioned near architectural highlights and significant spatial connections, such as entrances, aisles, and chapels, guiding the user along a coherent exploration path. For this smartphone-based implementation, performance optimizations included precomputed lighting, downscaled textures, and lightweight shaders, ensuring smooth frame rates and reducing the risk of VR sickness. Despite its simplicity, this configuration effectively democratizes access to cultural heritage, providing a compelling immersive experience to a broad audience, including educational institutions or museums.

The VR headset version provided a significantly more interactive and immersive experience. Its built-in handheld controllers enabled natural user movements, precise object interactions, and intuitive navigation. Development relied on specialized Unity plugins, including the Oculus XR Plugin for robust motion tracking and controller support, and the XR Interaction Toolkit to implement common VR interactions such as pointing, grabbing, and hovering over objects. Optimization strategies focused on reducing latency and preventing VR sickness while maintaining visual fidelity. This was achieved through prerendered lighting, careful optimization of shadows, and selective texture scaling, ensuring high frame rates and smooth responsiveness. The result is a highly realistic, fully interactive experience where users can freely explore the virtual church, manipulate objects, and engage with the environment in a natural and immersive way.

For the CAVE system, development was oriented toward creating a seamless, synchronized multi-wall immersive experience. The CAVE environment consists of three projection surfaces, front, left, and right walls, on which the virtual scene is displayed in real time. To manage this, the application was distributed across a networked architecture involving two high-performance workstations: one responsible for rendering the front wall and the other managing the two side walls. These machines communicate via a synchronization protocol that ensures perfect alignment of visual output across all three walls, avoiding tearing or desynchronization during user movement. User tracking within the CAVE is enabled through a VRPN (Virtual Reality Peripheral Network) system. This setup includes wearable devices, such as motion-tracking glasses and handheld controllers, worn by the user. The tracking data are continuously transmitted to Unity in real time, allowing the system to dynamically adjust the user's point of view and interactions based on physical movements. As a result, the user experiences coherent and immediate responses within the virtual environment, reinforcing the sense of immersion and spatial awareness. Thanks to the CAVE's dedicated hardware, which surpasses the computational limits of standalone headsets or smartphones, the application could support significantly higher-resolution textures and perform real-time lighting and shadow calculations without compromising performance. This enhances visual fidelity, especially in rendering architectural details and material qualities of the cultural site.

The result is a rich, immersive virtual tour that not only preserves spatial realism but also enables shared exploration experiences, ideal for educational or exhibition contexts. By tailoring the application to the unique capabilities of each platform, from mobile to high-end CAVE systems, the pipeline ensures that cultural heritage becomes more accessible, engaging, and understandable across a broad spectrum of audiences and devices.



(A) Cardboard

(B) Meta Quest 2

(C) CAVE system

FIGURE 3.7: Virtual exploration of the SS. Crocifisso al Calvario Church

3.2 Technical and ethical challenges of digitalisation

The digitalisation of cultural heritage represents a remarkable opportunity for the preservation, accessibility, and enhancement of historical and artistic assets. However, this process entails a series of complex challenges, both technical and ethical, that must be addressed with awareness and responsibility. From a technical perspective, the management of high-density 3D data, such as point clouds from laser scanning and complex meshes from reconstruction processes, requires advanced software tools, high-performance hardware, and specialized expertise. Making these models usable in real-time environments, such as VR applications, involves a delicate balance between visual quality and system performance, through techniques like geometry simplification, texture resolution adjustment, and pre-baked lighting.

Among the technical challenges encountered during the digitisation process, special attention was also devoted to managing motion sickness (or cybersickness). This discomfort, caused by the mismatch between physical movement and visual feedback in virtual environments, represents a significant limitation in the design of accessible and comfortable immersive experiences. Texture resolution, frame rate, and interaction dynamics had to be carefully calibrated, highlighting the technical complexity behind creating smooth and realistic digital experiences.

Another critical challenge is the interoperability across different platforms, ranging from smartphone-based VR solutions (like Google Cardboard) to standalone headsets (like Meta Quest), and fully immersive setups like CAVE systems. Each platform introduces distinct constraints in terms of computational resources, interaction methods, and rendering capabilities, requiring carefully tailored multi-platform development. In this context, user experience becomes a core concern, especially

(A) Cardboard

(B) Meta Quest 2

(C) CAVE system

FIGURE 3.8: Virtual exploration of the Santa Maria La Vetere Church

in relation to issues such as cybersickness, which may arise due to latency or mismatches between visual perception and physical movement.

On the ethical level, digitalisation raises fundamental questions about the authenticity and integrity of virtual representations. The process of creating digital models necessarily involves interpretative decisions: any virtual restoration, removal of secondary elements, or reconstruction of missing parts can influence the perception of the asset and potentially alter its historical meaning. For this reason, transparency is essential, documenting each step of the process and clearly distinguishing between surveyed data and hypothetical reconstructions.

Equally important are concerns around accessibility. While digitalisation promises to democratize access to culture, it may also generate new forms of exclusion, due to unequal access to appropriate devices, stable internet connections, or digital literacy. Designing inclusive experiences requires accounting for these barriers as well.

Further challenges include those of intellectual property, long-term digital preservation, and responsible content management. Who owns the rights to a digital reconstruction? How can we ensure that digital assets are stored in formats that remain usable and accessible over time? And how do we prevent unauthorized uses or commercial exploitation of digitised heritage?

Addressing these challenges calls for an interdisciplinary approach, combining technological expertise with historical-artistic sensitivity and a strong ethical commitment to cultural communication. Only in this way can digitalisation not just preserve cultural memory, but also strengthen its educational and inclusive potential for future generations.

Chapter 4

Generative AI and Cultural Heritage

In recent years, Generative Artificial Intelligence (Generative AI) has transformed the ways in which cultural heritage is accessed, enhanced, and interpreted. Large Language Models (LLMs) have made it possible to design advanced interfaces capable of providing articulate and relevant answers to complex questions specific to the cultural context. More recently, the emergence of multimodal and generative architectures has expanded the potential of AI systems to integrate text, images, and spatial data, opening new possibilities for immersive cultural experiences and interpretive applications.This progress allows for more interactive and meaningful engagement with cultural assets, overcoming traditional barriers to enjoyment. However, selecting an appropriate model requires a careful evaluation of the specific requirements of the cultural domain, the computational resources available, and the desired level of control over the generated outputs. In the context of this research, the objective was to develop a Question Answering (QA) system that could deliver accurate results while remaining lightweight, resource-efficient, and free from proprietary licensing constraints. To achieve this, the system was built upon BERT, a widely adopted, open-source, pre-trained model extensively documented in the scientific literature. Unlike many large-scale generative models, BERT operates primarily as an extractive model, producing responses strictly bound to the contextual information provided. This characteristic ensures a higher degree of determinism and transparency, while minimizing the risk of unverified or hallucinated content.

This chapter examines the current state of the art in language models applied to cultural heritage, compares the capabilities of BERT with those of more complex architectures, and discusses the rationale behind adopting a "light" model for the proposed system. Furthermore, it reflects on the trade-offs between computational efficiency, scalability, and the precision of domain-specific responses, highlighting the importance of controlled and context-aware QA approaches in the cultural heritage sector.

4.1 Language Models for Question Answering

Over the past ten years, Natural Language Processing (NLP) has experienced a profound transformation thanks to the introduction and development of LLMs. These models are artificial intelligence systems designed to understand, generate, and interact with human language in increasingly sophisticated ways by analyzing vast amounts of textual data (Brown et al., 2020). In the earlier stages of NLP, approaches were considerably more limited. Traditional systems often relied on keyword-matching

or rigid, rule-based grammatical structures (Manning, 2008). However, these methods had clear limitations: they were unable to grasp the complex meanings of sentences or understand the context in which certain words were used, often failing to handle the ambiguities and linguistic variations typical of natural language. The real revolution came with the introduction of the Transformer architecture (Figure 4.1), proposed by Vaswani et al. in 2017 (Vaswani et al., 2023), which radically changed how language models are built and trained. The key mechanism of Transformers is self-attention: each word can "look at" all the others and decide how strongly it is related to them, regardless of their position. This allows the model to capture connections that span long stretches of text, for example, linking a subject at the beginning of a sentence to the verb that appears much later. Another important feature is multi-head attention. Instead of focusing on a single type of relationship, the model learns to observe several patterns in parallel, so it can grasp different shades of meaning and structure at once. Because Transformers do not process words in a fixed order, they rely on positional encodings to keep track of where each token appears in the sequence. Finally, the architecture itself is highly modular: it is built by stacking layers of attention and feed-forward components, which can be combined in different ways depending on the task at hand. This combination of flexibility and depth explains why the Transformer quickly became the foundation for today's large language models. By choosing to use only the encoder, only the decoder, or both together, it is possible to tailor models either for understanding text, for generating it, or for combining both abilities.

Thanks to this innovation, it became possible to develop much larger language models capable of learning complex representations of language from massive text corpora, often consisting of billions of words drawn from books, articles, websites, and other sources. Models such as BERT (Devlin et al., 2019), GPT(Radford et al., 2018), LLaMA (Touvron et al., 2023), and many others are considered "large" not only because of the amount of data they are trained on, but also because of the number of parameters (the internal variables the model uses to interpret language) which can reach hundreds of millions or even billions. LLMs can perform a wide range of linguistic tasks: from automatic translation to text generation, document summarization to sentiment analysis, and notably, Question Answering systems that can understand and respond to complex queries posed in natural language. Modern QA systems are now able to interpret queries in natural language and generate contextually relevant responses, reflecting a genuine understanding of both the questions and source material. These advances have allowed for the creation of increasingly sophisticated and accurate applications across numerous fields, from customer service and medicine to cultural heritage, opening new possibilities for accessing, enhancing, and interpreting textual information. However, the success of LLMs also brings significant challenges, such as high computational costs, the complexity of model interpretability, and the necessity to ensure reliable and verifiable responses, especially in specialized domains like cultural heritage.

To better understand how these challenges manifest and how different models address them, it is essential to examine the underlying architectures of large language models. Despite their shared foundation on the Transformer architecture introduced by Vaswani, LLMs exhibit significant variations in design and functional emphasis, broadly categorized into three architectural types: encoder-only, decoder-only, and encoder-decoder models. Each architecture reflects specific trade-offs between language understanding and generation capabilities, and thus is suited to different classes of NLP tasks.

Encoder-only models, such as BERT, RoBERTa (Liu et al., 2019), and DistilBERT

(Sanh et al., 2019), are optimized primarily for tasks requiring deep comprehension of textual input rather than free-form text generation. Their key feature is the use of bidirectional self-attention mechanisms, which allow these models to consider the entire surrounding context of each word simultaneously during both training and inference. This approach enables models to comprehend complex aspects of language, including sentence structure, the nuances of words with multiple meanings, and intricate semantic relationships. Consequently, encoder-only models demonstrate superior performance in applications such as text classification, named entity recognition, sentiment analysis, and extractive question answering, contexts where the principal objective is to locate or identify specific information within a given passage. For instance, in extractive QA, the model simultaneously processes both the question and the relevant context, efficiently pinpointing the precise text span that answers the query. Because their predictions are grounded in the source material and driven by bidirectional attention, encoder-only models are less prone to generating unsupported or fabricated information. This property is particularly valuable in domains, such as cultural heritage or medicine, where factual reliability is paramount.

Decoder-only models, like those from the GPT family, operate by predicting each subsequent token in a sequence, strictly based on all prior tokens. This left-to-right, unidirectional attention structure allows them to generate coherent and contextually appropriate text, excelling in domains such as creative writing, conversational agents, and abstractive summarization.

Yet, despite their impressive generative fluency, these models face notable challenges. Because they generate responses one token at a time, without explicit mechanisms to ground their output in factual information, they occasionally produce content that is plausible-sounding but inaccurate or entirely fabricated, a phenomenon often described as "hallucination." While this generative flexibility is advantageous in creative contexts, it can be problematic in scenarios where accuracy, verifiability, and controlled output are essential.

Encoder-decoder models, often referred to as sequence-to-sequence architectures, form a hybrid strategy that merges the analytical power of encoder-only setups with the generative strengths of decoder-only systems. Take T5 and BART, for example. These models first process the input text, distilling its meaning into a compact latent representation. From there, they generate output sequences by translating that representation back into language.

This configuration is especially well-suited for tasks that demand substantial transformation between input and output-machine translation, abstractive summarization, or generative question answering come to mind. The approach offers a rare blend: deep comprehension of source material and the ability to generate fluent, context-aware responses. That dual capability makes encoder-decoder models particularly valuable for cultural heritage applications, where both faithful interpretation and engaging presentation are critical.

In summary, the choice among encoder-only, decoder-only, and encoder-decoder LLM architectures hinges on the specific requirements of the target application, whether it prioritizes interpretability and accuracy, generative creativity, or a combination of both. Understanding these distinctions is crucial when designing AI-driven systems for cultural heritage, where maintaining factual integrity while enabling engaging interactions often requires tailored trade-offs between model complexity, responsiveness, and control.

In the field of cultural heritage, the choice of a language model is not only a technical matter but also a reflection of how knowledge is constructed, accessed,

and validated. Unlike creative applications of language technology, cultural heritage work demands strict accuracy and reliance on authoritative sources to ensure reliability. Models like BERT or T5 are especially well-suited to this kind of work. Because they understand context so well, they can help extract key information, enrich digital archives, and support detailed cataloguing. Just as importantly, they keep a clear link back to the original source of every piece of data. That traceability is essential for preventing misinformation, something particularly critical when dealing with cultural heritage, where the authenticity of records and the integrity of history cannot be compromised.

In contrast, generative models, especially those employing decoder-only architectures, demonstrate significant strengths in creative and narrative-driven contexts within the cultural sector. These models facilitate the automatic composition of artwork descriptions, the development of immersive storytelling experiences for virtual museum tours, and even the reconstruction of historical events through coherent and engaging narratives. That said, the creative latitude these models possess often leads to a trade-off: factual accuracy may suffer, resulting in instances of hallucinated or unverified content. Thus, while the output is frequently compelling and contextually rich, careful fact-checking remains essential.

To address these issues, hybrid approaches play an increasingly important role. Retrieval-Augmented Generation (RAG) frameworks combine the strengths of both paradigms by constraining generative outputs through retrieved, domain-specific documents. By grounding generated text in trusted sources, this method helps reduce errors while still allowing for rich and expressive responses. In the cultural heritage field, such systems open the door to more interactive ways of exploring collections. They allow users to receive answers that are not only informative and engaging but also firmly supported by reliable documentation.

In cultural heritage, finding the right balance between creativity and accuracy is essential. Technology here has a special role: it's not just about making information easier to share, but also about protecting and verifying it. That's why choosing the right type of language model isn't only a technical detail. It's a careful decision, where innovation has to go hand in hand with responsibility.

Building upon the architectural considerations previously discussed, best practices emerging from recent literature emphasize several key strategies to effectively harness language models within the cultural heritage domain. Domain-specific fine-tuning stands out as a foundational step. General-purpose models like BERT or RoBERTa are already strong in understanding language, but they become far more effective when trained on texts created specifically for cultural heritage. These include curatorial notes, archival records, exhibition catalogues, and academic publications, sources rich in specialized terms and historical context. By learning from this material, the models can better understand the language and style typical of the heritage field. This makes them more reliable for tasks such as recognizing names and entities, improving search within collections, or answering questions directly from trusted texts. Without this kind of adaptation, generic models often struggle to grasp subtle meanings or risk misinterpreting terms that hold cultural significance.

Equally important is knowledge grounding, the practice of ensuring that all answers and generated content are explicitly traceable to trusted, verifiable sources. In the cultural heritage field, where accuracy and authenticity are essential, this approach helps prevent the risks of misinformation or "hallucinations," which are common challenges in generative AI. Well-designed systems do this by linking their responses to archival documents, curated databases, or scholarly works. This not only strengthens transparency but also builds trust (Izacard and Grave, 2021). In this way,

AI is not just producing content: it is helping to safeguard and share knowledge that has already been carefully verified.

Interoperability with established heritage data standards greatly increases the practical value of AI tools. By connecting NLP workflows to metadata frameworks like CIDOC CRM or Dublin Core, these applications can maintain consistent meaning and make it easier to share information across institutions. This approach supports scalable solutions for searching across collections, linking open data, and collaborative curation, all of which contribute to the sustainable management of cultural heritage. Using standard ontologies also ensures that AI-generated outputs integrate smoothly into existing digital systems, promoting reuse and helping preserve cultural assets over the long term.

Recent case studies illustrate how these principles translate into concrete implementations. Transformer-based QA systems have been successfully applied to automate metadata extraction from archival records, dramatically reducing manual effort and improving accessibility (Colla et al., 2022). Semantic search engines powered by fine-tuned models help users explore digitized collections more intuitively, allowing for precise retrieval of cultural artifacts. Even more engaging are interactive museum guides that use NLP-driven dialogue in VR or AR environments. These systems provide contextualized, reliable narratives that enrich the visitor experience, offering new ways to connect with cultural heritage beyond traditional displays.

In this context, BERT and its variants continue to be a reliable choice for building question-answering systems that are carefully controlled and tailored to cultural heritage. Their encoder-based architecture helps the models understand queries accurately and link answers to trusted sources, supporting priorities like transparency, reproducibility, and independence from proprietary APIs. This means cultural institutions can create AI tools that respect the specific needs and ethical considerations of the field, while still providing engaging and accessible ways for people to explore and learn from cultural collections.

4.2 Adopting BERT for Domain-Specific Question Answering

The selection of BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) as the foundation for the present Question Answering system comes from its unique combination of contextual understanding, precision, and open accessibility. BERT represented a major step forward in Natural Language Processing by showing how bidirectional attention can capture the meaning of a word based on both its preceding and following context, an improvement over earlier unidirectional or shallowly contextualized models. Unlike traditional embedding techniques such as Word2Vec (Mikolov et al., 2013) or GloVe (Pennington, Socher, and Manning, 2014), which generate static word vectors, BERT produces context-dependent representations that adapt to the surrounding text. This allows it to disambiguate words with multiple meanings and better capture the subtleties of specialized vocabularies, a crucial advantage in fields like cultural heritage, where terminology is dense and historically nuanced. Moreover, because BERT is available as a pre-trained open-source model, it has greatly accelerated research and application development. Institutions and researchers can start from this solid foundation and fine-tune it for specific tasks without the need for costly training from scratch, lowering barriers and making advanced NLP tools more accessible. At the same time, alternative models were

considered. DistilBERT, for example, offers a lighter and faster version of BERT, requiring fewer computational resources. However, this reduction in size comes at the cost of lower accuracy, especially in complex, context-rich domains such as cultural heritage. On the other end of the spectrum, RoBERTa provides higher performance thanks to optimized training and larger pre-training data, but its increased computational demands make it less suitable for lightweight, resource-efficient deployment in heritage institutions. BERT therefore represents a balanced compromise: it is robust enough to handle nuanced, domain-specific language while remaining accessible, open-source, and efficient for real-world applications.

4.2.1 Architecture and Core Principles

BERT is an encoder-only transformer model, which means it is designed entirely for understanding and representing language rather than generating it. Unlike decoder-based architectures, which process text in a left-to-right (autoregressive) manner and predict the next token based on previous context, BERT uses a bidirectional attention mechanism. This allows every token in the input to consider all other tokens, both before and after it, at every layer. The result is a rich, deeply contextualized embedding for each word, capturing its meaning and role within the wider sentence or document. This bidirectionality is particularly important for disambiguation. Many words in natural language have multiple meanings, and understanding the intended meaning often depends on the surrounding context. Earlier unidirectional or shallow models struggled with such cases, but BERT can capture subtle semantic and syntactic relationships across the whole sequence. This makes it particularly effective at modeling long-range dependencies, such as subject–object relationships or thematic links between clauses, with greater accuracy.

BERT's Transformer backbone is built from multiple layers of self-attention and feed-forward networks. In the self-attention mechanism, each token creates three learned vectors (queries, keys, and values) that interact to produce a weighted summary of information from all other tokens. Multi-head attention runs this process in parallel across several "heads," enabling the model to capture different types of relationships at the same time, for instance, one head might focus on syntactic structure while another captures semantic similarities.

BERT's pre-training process is key to its capabilities and relies on two complementary objectives:

- Masked Language Modeling (MLM) – A random subset of tokens (typically 15%) in the input sequence is replaced by a special [MASK] token, while the model must predict the original values based on surrounding, unmasked tokens. Because the missing words can occur anywhere in the sequence, the model learns to integrate information from both directions. This objective encourages a deep understanding of context and co-occurrence patterns, rather than depending only on local, sequential predictions.

- Next Sentence Prediction (NSP) – Pairs of sentences are fed to the model, which must predict whether the second sentence naturally follows the first in the source text. This task encourages comprehension at the discourse level, helping BERT excel in applications where sentence-to-sentence relationships matter, such as identifying cause–and-effect structures in historical narratives or following thematic continuity in archival documents.

These pre-training tasks produce a general-purpose language representation, which can then be adapted (fine-tuned) to specific tasks using relatively small amounts of

labelled data. Fine-tuning modifies the pre-trained weights slightly to specialise the model for applications such as extractive question answering, semantic search, named entity recognition, or document classification. In the cultural heritage domain, these capabilities have direct implications:

MLM enables BERT to learn the meaning of rare, domain-specific terms found in curatorial descriptions, cataloguing records, or historical archives, even when such terms appear infrequently in general-purpose corpora. NSP helps the model reason over sequences of related information, such as linking an artifact's description to its provenance, or connecting a historical event to contemporary records or scholarly interpretations. BERT's ability to model long-range dependencies ensures that references spanning multiple sentences, common in historical and academic texts, are captured accurately without losing coherence. The fact that BERT is open-source and available in many pre-trained variants makes it an appealing choice seeking transparency, reproducibility, and independence from proprietary systems. Moreover, the encoder-only architecture avoids the uncontrolled generative behaviours of decoder-based models, keeping outputs strictly grounded in the provided context which is a quality essential in domains where factual precision and traceability are critical.

4.2.2 Strengths for Cultural Heritage Applications

BERT is particularly well-suited for cultural heritage projects because it meets not only technical needs, but also methodological and ethical ones. In this context, the model acts as a bridge between historical evidence and its digital form, helping people access, understand, and explore knowledge in a reliable way. Using BERT means balancing careful language understanding with factual accuracy and long-term usability, qualities that are essential when working with historical and cultural materials.

1. Contextual Understanding – BERT's bidirectional encoding allows it to understand each word in the context of the entire surrounding text, which is especially important when dealing with archival descriptions, historical narratives, or specialized terminology. Texts in cultural heritage often include complex sentence structures, old or rare vocabulary, multiple languages, and references that stretch across long periods or locations. For instance, in a museum catalogue, the meaning of a date or place might only become clear when linked to an event mentioned several sentences later. BERT's architecture captures these connections accurately, helping to interpret relationships between people, places, events, and objects, connections that are often essential in history, archaeology, and art history research.

2. Controlled Responses - As an extractive rather than generative model, BERT is particularly effective at providing answers that are directly based on the given text or knowledge base. Instead of inventing information, it retrieves and highlights what is relevant, reducing the risk of errors or unverifiable claims. This feature is especially important in cultural heritage, where accuracy is critical. For example, a digital archive may be used by scholars or conservators as a primary research tool, and any misleading output could affect research or conservation decisions. BERT's reliance on explicit, source-linked data ensures that every response is transparent and traceable, allowing users to verify the information it provides.

3. Open-Source Availability - BERT's release as open source makes both its design and implementation transparent. Beyond the cost advantages, open-source access supports long-term digital preservation which is a foundamental aspect for public museums, archives, and cultural institutions that need to maintain service continuity and manage their data independently of commercial providers.

Together, these strengths make BERT not just a solid technical choice, but also a strategically suitable one for cultural heritage applications. Its deep understanding of context allows it to handle complex, historically rich language accurately, while its open-source nature promotes transparency, long-term preservation, and collaborative care of knowledge. These values are central to the work of museums, archives, and cultural institutions.

4.3 Comparison with other Generative Models

While BERT focuses on controlled, source-grounded outputs, more complex models like ChatGPT and other decoder-only LLMs offer a different set of capabilities. These generative systems not only understand input text but can also produce extended, coherent, and sometimes creative responses. Unlike encoder-based models, which excel at extracting and contextualizing information from a given passage, generative models are trained to continue language sequences, allowing them to create content that goes beyond the original source. This shift from understanding to generation brings impressive expressive power, but also raises challenges in terms of factual accuracy, computational demands, and control over the outputs.

4.3.1 Generative vs. Extractive Capacities

This distinction has important implications for designing and deploying AI systems. Generative models excel at producing fluent, contextually rich, and conversational outputs, making them well-suited for situations where user interaction, storytelling, or exploratory reasoning is needed (Raffel et al., 2020). For example, in a museum setting, a generative model could create engaging narratives, simulate dialogues with historical figures, or offer personalised explanations based on a visitor's interests. Their ability to go beyond simply repeating information allows them to synthesise knowledge and present it in an accessible and engaging way. However, this strength also comes with a major risk. Because generative models are not inherently tied to a specific source, they can produce content that sounds plausible but is not actually grounded in the original data (Ji et al., 2023). In cultural heritage, where accuracy and provenance are critical, this lack of guaranteed traceability can threaten reliability and even spread misinformation. By contrast, extractive models like BERT follow a very different approach. They focus on selecting or reformulating information that is explicitly present in the source material, ensuring factual traceability and interpretability (Min et al., 2019). This makes them especially suitable for tasks where transparency, reproducibility, and verifiability are essential. For instance, in digital cataloguing or archival enrichment, an extractive model guarantees that every response can be traced back to an authoritative document, preserving the integrity of cultural heritage data while still providing efficient access to information.

4.3.2 Advantages of Generative LLMs

The strengths of ChatGPT-like systems lie in their ability to carry on natural, flowing conversations, perform complex reasoning that connects seemingly unrelated pieces of knowledge, and generate engaging, coherent narratives. Unlike extractive models, which mainly retrieve information that is explicitly stated in the source, generative systems can synthesize content from multiple contexts, reshape information to suit the user's needs, and adjust their style dynamically. This flexibility makes them particularly powerful in situations where communication must be both accurate and captivating. In the context of cultural heritage, these capabilities open up a wide range of applications. For example, interactive museum guides can provide personalised responses based on a visitor's interests, and educational storytelling platforms can weave historical facts into compelling, memorable narratives (Colucci Cante et al., 2024). Generative LLMs can also support exploratory search experiences, where users are not just looking for a single fact but are navigating complex relationships between artefacts, events, and historical figures. Their ability to engage in open-ended dialogue enables a more intuitive, human-like interaction, making cultural content more accessible to non-specialists and creating opportunities for deeper engagement. Additionally, generative models excel at complex inferential reasoning. They can connect events across different periods or draw thematic parallels between cultural artefacts that are not explicitly linked in the underlying data. This capacity to bridge fragmented knowledge sources can enhance curatorial work, support interdisciplinary research, and help discover new interpretative insights in heritage collections. By combining fluent communication, adaptability, and deep inferential reasoning, generative LLMs add a layer of interaction that goes well beyond simple fact retrieval, positioning them as transformative tools for cultural education and engagement (Trichopoulos, 2023).

4.3.3 Limitations and Risks

The trade-offs of using decoder-only large language models are particularly significant in the context of cultural heritage. These models require substantial computational resources for both training and inference. For example, autoregressive generation in decoder-only architectures depends on extensive parallel corpora and significant computational power (Qorib, Moon, and Ng, 2024). This can present practical challenges for institutions with limited budgets or infrastructure, potentially slowing the adoption of these technologies. More critically, these models are prone to generating hallucinations, outputs that sound plausible but are factually incorrect. Research shows that LLMs can produce convincing yet inaccurate information, particularly when external verification is unavailable. In the field of cultural heritage, where accuracy and provenance are paramount, this tendency represents a serious risk. Xu et al. formalize this issue, demonstrating that LLMs cannot learn all computable functions and will inevitably hallucinate when applied as general problem solvers (Xu, Jain, and Kankanhalli, 2024). Additionally, the generative nature of these models makes it difficult to maintain precise control over the boundaries of contextual responses. This lack of control complicates efforts to ensure that outputs align with authoritative sources, a challenge highlighted by Huang et al., who emphasize the need for robust detection and mitigation strategies to manage hallucinations in LLMs. In summary, while generative LLMs offer advanced and flexible

capabilities, their limitations, particularly in terms of computational demands, susceptibility to hallucinations, and challenges in controlling contextual fidelity, highlight the necessity for careful deployment. In sensitive domains such as cultural heritage, these constraints suggest that hybrid or complementary approaches may be preferable to ensure accuracy, reliability, and responsible use of AI.

4.3.4 Extractive vs. Generative Models

For these reasons, while generative LLMs demonstrate remarkable versatility, controlled, extractive models like BERT remain the preferred choice in the cultural heritage domain. Their outputs are grounded in verifiable text, ensuring that every answer can be traced back to trusted documents or curatorial records, thereby safeguarding both scholarly integrity and public trust. In contexts where historical accuracy, provenance, and interpretative responsibility are paramount, the balance clearly favors architectures that prioritize control, reliability, and transparency over open-ended creativity. Extractive models also offer fine-grained control over the scope of responses, allowing systems to constrain answers strictly to the input text or curated datasets. This level of control is especially critical in heritage applications, where even minor deviations from verified sources could propagate misinformation, distort historical narratives, or mislead the public. By linking outputs directly to authoritative sources, these models enable reproducibility, auditing, and verification of each answer against original documents, a requirement essential for scholarly research, digital archiving, and museum information systems. Moreover, the open-source availability of models like BERT and its derivatives allows institutions to maintain full control over deployment, adaptation, and data governance. This independence ensures long-term sustainability, as institutions are not reliant on proprietary APIs or commercial platforms, and can customize models for domain-specific terminologies, languages, or collection types. It also facilitates cross-institutional collaboration, interoperability, and the creation of shared NLP tools tailored for heritage applications, reinforcing knowledge stewardship and digital preservation. Finally, while generative LLMs offer advantages in fluency, conversational interaction, and creative synthesis, their potential for hallucination, high computational demands, and reduced traceability make them less suitable for scenarios where factual accuracy and provenance are non-negotiable. Extractive models, in contrast, strike a strategic balance by providing precise, verifiable, and ethically responsible outputs, aligning closely with the core principles of cultural heritage management and scholarly communication.

4.3.5 Emerging Multimodal and Generative AI Models for Cultural Heritage

Recent developments in Artificial Intelligence have expanded beyond purely text-based architectures to multimodal and generative systems capable of processing and producing not only language but also images, sound, and spatial representations. Models such as GPT-4 (OpenAI, 2024), Gemini (Anil et al., 2024), and open-source frameworks like LLaVA (Liu et al., 2023) and Kosmos-2 (Peng et al., 2023) can jointly interpret textual and visual information, thus enabling more fluid and intuitive interactions between humans and machines. While GPT-4 and Gemini exemplify large-scale proprietary systems capable of reasoning across text, images, and even code, open frameworks such as LLaVA and Kosmos-2 focus on aligning vision

encoders with large language models through instruction tuning, enabling open research and domain-specific adaptation. In particular, LLaVA integrates CLIP-based vision encoders with conversational LLMs to produce detailed visual grounding, whereas Kosmos-2 extends multimodality toward grounding language in spatial and referential contexts. This technological evolution marks a significant shift from the linguistic logic of question answering toward a broader, sensory understanding of knowledge, a change that resonates deeply with the inherently multimodal nature of cultural heritage.

In museum and archival contexts, these systems open up new possibilities for how audiences can explore and interpret cultural artifacts. A multimodal AI agent, for example, could answer questions about a fresco by considering both its image and the related curatorial notes, or could automatically link decorative motifs in 3D scans to iconographic categories described in textual metadata. Similarly, generative models could assist in reconstructing fragmented artifacts or simulating historical environments, offering immersive experiences that blend factual data with plausible reconstructions. Such integration between perception and language may allow for a more embodied and emotionally engaging experience of heritage, bridging the gap between digital documentation and human interpretation. In this sense, multimodal AI systems have the potential to act not only as tools of access but also as mediators of meaning within immersive cultural environments. Moreover, the convergence of multimodal AI with technologies such as augmented reality, virtual reality, and haptic interfaces further amplifies the potential for experiential engagement. Imagine a visitor navigating a virtual reconstruction of a Roman villa, guided by an AI that not only narrates historical facts but also responds to visual cues, gestures, and spoken questions in real time. This kind of interaction transforms passive observation into active dialogue, fostering deeper cognitive and emotional connections with cultural content.

However, these advances also raise important challenges that extend beyond the technical domain. While multimodal and generative models can provide highly engaging, context-aware experiences, they inevitably introduce issues of control, transparency, and authorship. Unlike extractive systems such as BERT, which retrieve information directly from defined sources, generative architectures may synthesize or even invent details that were not present in the original materials. This capacity for "creative inference" can enrich storytelling but also risks distorting historical authenticity, a critical concern in heritage communication. Furthermore, the opacity of large-scale training datasets complicates the tracing of data provenance and the verification of interpretive claims, two pillars of scholarly practice in cultural heritage studies. These issues are especially important when cultural stories are sensitive, disputed, or not well represented. If AI systems are trained on biased or incomplete data, they can give a narrow view of culture. To prevent this, it is important to design them in an open and inclusive way, with clear documentation and attention to cultural diversity.

Technology should always go together with ethical awareness. AI tools for cultural heritage should respect the people and histories they represent, helping to share culture without changing or simplifying it. The aim is to support understanding and inclusion, not to replace human interpretation.

While these emerging systems point toward exciting future possibilities, the present research remains focused on BERT-based architectures. This choice ensures methodological transparency, reproducibility, and a clear connection between textual data and retrieved answers. The discussion of multimodal and generative AI is therefore

intended as a contextual and forward-looking reflection, outlining directions for future work rather than components of the implemented system. Within this scope, extractive models such as BERT continue to provide a reliable foundation for the controlled exploration of natural language interaction in virtual heritage environments.

4.4 Implementation in the Cultural Heritage QA System

The integration of BERT into a Question Answering system for cultural heritage requires careful and precise management of textual context. Each user query is paired with a segment of text drawn from archives, catalogs, or museum descriptions, which serves as the reference context for the model. By receiving both the question and the relevant passage as input, BERT acts as a precise extractor: rather than generating content freely, it identifies and returns only the information explicitly present in the provided text. This ensures both traceability and consistency in responses, which is essential when working with sensitive or scholarly heritage data. Accurate delivery of context is critical for producing relevant and reliable answers. One widely adopted approach is semantic retrieval, where the system ranks and selects the most pertinent documents, paragraphs, or passages based on the user's query. This can be achieved using embedding-based similarity measures, TF-IDF weighting, or hybrid retrieval strategies that combine lexical and semantic features. Text segmentation and indexing further enhance system performance by isolating the most informative portions of a corpus, ensuring that BERT focuses on the content most likely to contain the correct answer.

In more advanced implementations, multiple context sources can be concatenated or merged, enabling BERT to extract answers that draw on information from several documents while still maintaining relevance. Special attention must be given to input length limitations: long documents are typically split into overlapping chunks, with each chunk processed independently before the results are aggregated. Additionally, metadata such as provenance, date, or source type can be appended to the input, guiding the model and providing additional interpretive cues. By carefully structuring context and employing retrieval and segmentation techniques, BERT-based QA systems can deliver highly accurate and contextually grounded answers. This structured approach is especially valuable in cultural heritage applications, where even subtle misinterpretations can compromise the integrity of historical, artistic, or curatorial information, and where users, from scholars to the general public, expect both precision and reliability.

The quality of responses in a cultural heritage QA system is assessed through a combination of automated metrics and human evaluation, creating a comprehensive framework that ensures both technical performance and scholarly reliability. Standard quantitative measures, such as Exact Match (EM) and F1 score, evaluate the overlap between the model's predicted answer and the correct reference, providing an objective measure of accuracy, completeness, and alignment with annotated data. These metrics are particularly valuable during model training and iterative fine-tuning, allowing developers to systematically track improvements and identify areas where the system may struggle with complex queries or nuanced language. In the context of cultural heritage, relying solely on numerical metrics is never enough. Historical and cultural texts are full of subtleties, archaic terms, context-dependent meanings, and implicit references, that automated scores cannot fully capture. This is why expert review by domain specialists is indispensable. Curators, archivists,

and historians carefully examine the system's outputs to ensure factual accuracy, sound interpretation, and alignment with scholarly standards. They check details such as dates, attributions, relationships between events, and the correct use of terminology, making sure the QA system does not misrepresent or distort historical knowledge. In addition, qualitative evaluation makes it possible to assess factors such as relevance, clarity, and usability. Experts can spot answers that, even if factually correct, may be incomplete or ambiguously worded, providing guidance for iterative improvements in context selection, text preprocessing, or model fine-tuning. Combining automated metrics with expert review creates a feedback loop: quantitative measures enable rapid benchmarking and optimization, while qualitative assessment ensures that outputs align with the interpretive, ethical, and educational objectives of cultural heritage institutions. By incorporating both layers of evaluation, BERT-based QA systems strike a balance between technical efficiency and domain-specific responsibility. This hybrid framework not only ensures that responses are accurate and relevant but also builds trust among scholars, museum professionals, and the public. Ultimately, it guarantees that cultural heritage QA systems function as reliable, transparent, and ethically grounded tools, capable of enhancing access to knowledge while respecting the provenance, context, and interpretive complexity of historical and artistic materials.

A concrete example of BERT's application in the cultural heritage domain is provided by Md. Mahmud-uz-Zamanet al. (Uz-Zaman, Schaffer, and Scheffler, 2021). The study involved a dataset of 285 questions related to museum exhibits, classified into answerable and unanswerable, as well as factoid and open-ended types. The results illustrate the strengths and limitations of an extractive model like BERT in a real-world cultural context. For factoid questions, BERT achieved a high accuracy rate, demonstrating its ability to extract precise information when the relevant context is correctly provided. Conversely, performance on open-ended questions was more limited, with only 36% correct answers, highlighting the challenges of interpretative or exploratory queries that require reasoning beyond the supplied text. This example reinforces key principles of implementing BERT in heritage QA systems: the careful selection and structuring of textual context are essential for ensuring accurate, traceable responses; extractive models excel when the task requires fidelity to source material; and the combination of automatic metrics and expert evaluation remains critical to maintaining both factual reliability and cultural interpretative responsibility. By grounding responses in authoritative sources while acknowledging the model's boundaries, such implementations exemplify a controlled yet effective approach to AI-assisted access to cultural knowledge.

Another example of BERT's application in the cultural domain is presented by Bongini et al., 2020, who explore Visual Question Answering (VQA) for cultural heritage. In this approach, BERT-based models are combined with neural networks for image processing, enabling the system to answer questions expressed in natural language that also refer to visual content, such as photographs of artworks or images of museum artefacts. Compared to traditional audio guides, VQA introduces a richer and more flexible mode of interaction: visitors can ask open-ended questions and receive relevant answers that integrate textual knowledge with visual interpretation. In this framework, BERT plays a crucial role in understanding the user's language and semantically anchoring the question to the available cultural content, ensuring coherence between linguistic input and visual information. This type of application demonstrates how BERT's adoption is not limited to the management of archival or cataloguing texts, but can also extend to interactive scenarios that transform the

museum experience, bringing it closer to personalised and multimodal forms of dialogue. Looking ahead, such systems have the potential to evolve into advanced educational tools, combining accessibility, accuracy, and narrative engagement.

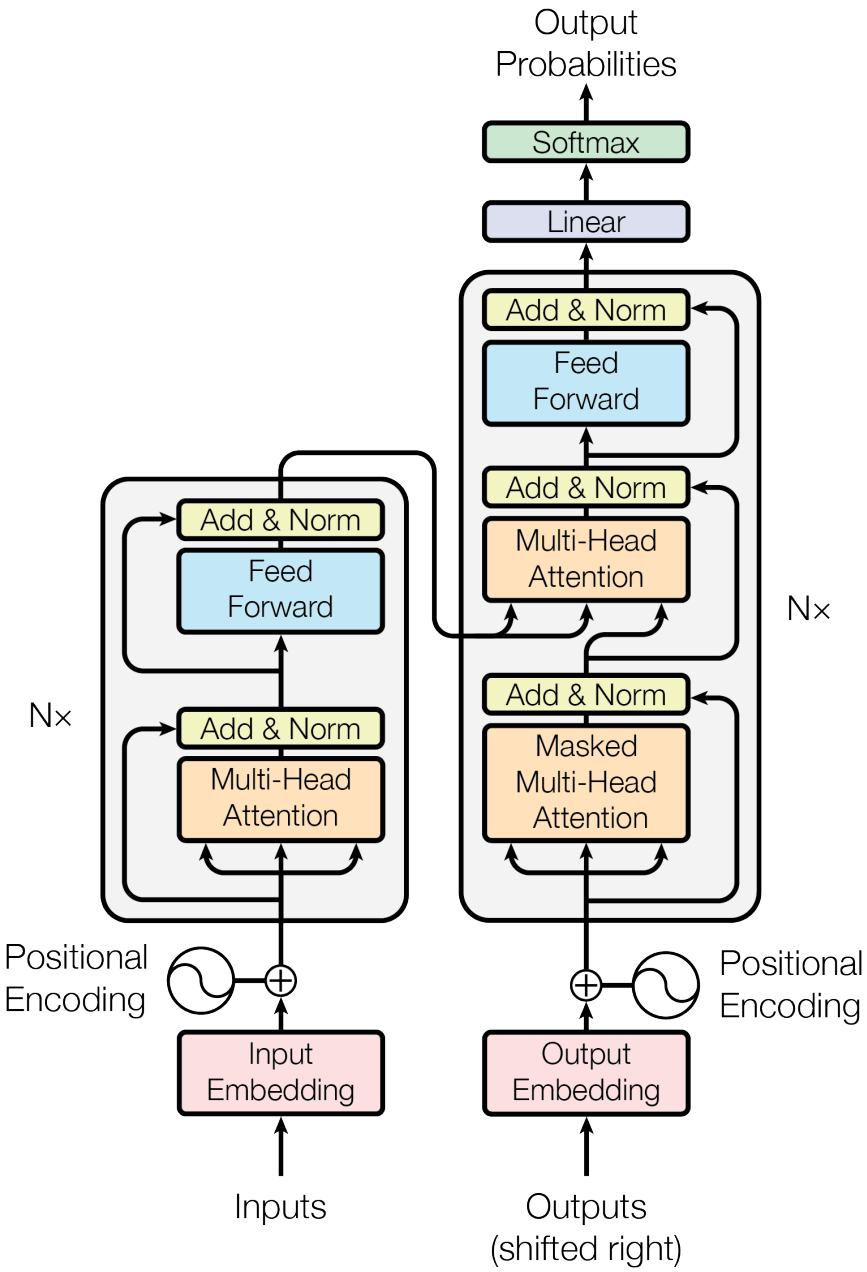


FIGURE 4.1: The Transformer - model architecture

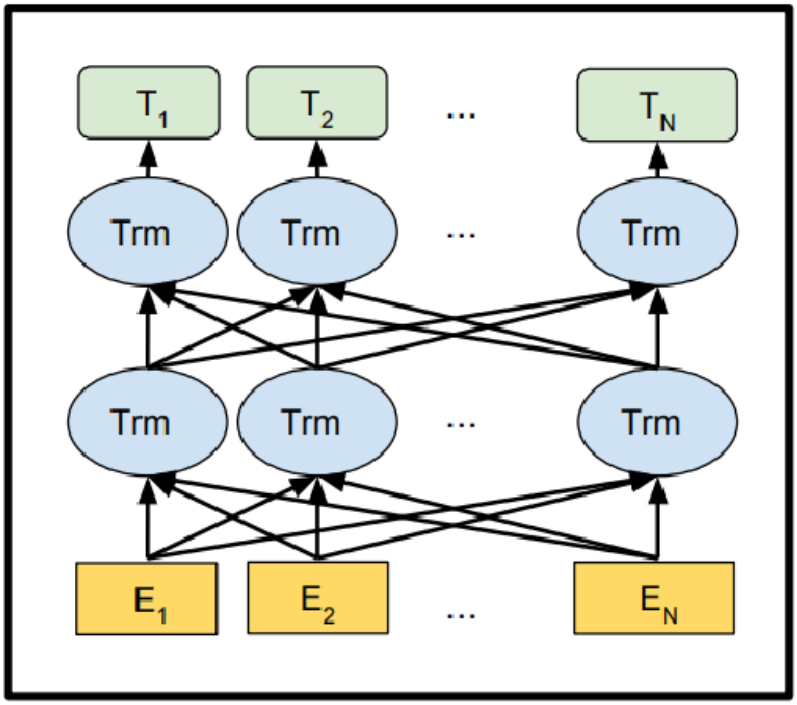


FIGURE 4.2: Bert Architecture

Chapter 5

AI powered virtual guide

This chapter presents the development and implementation of an AI-powered virtual guide for cultural heritage (Farella, Chiazzese, and Lo Bosco, 2022). The system integrates a BERT-based Question Answering model, specifically fine-tuned on the Italian SQuAD-it dataset, with both speech-to-text and text-to-speech technologies, thereby facilitating fluid spoken interaction. Within a museum setting, users can directly ask questions about artworks, receiving accurate and context-grounded answers provided by the virtual guide. To ensure comprehensive monitoring and ongoing enhancement of the system, all user experiences, both questions and generated answers, are systematically recorded via the Experience API (xAPI). Answer quality is assessed through a dual evaluation strategy, combining automatic metrics such as EM and F1 scores with contextual and domain-specific evaluation procedures, ensuring both technical robustness and the reliability of information provided.

5.1 System architecture

The architecture of the AI-powered virtual guide integrates several components into a pipeline that allows natural interaction between users and the cultural heritage (Figure 5.1). The system begins with user input, usually a voice query about a specific work of art or artifact. This signal is first processed through a speech-to-text (STT) module, implemented using Wit.ai, which converts the voice input into written natural language. The transcribed query is then passed to the Question Answering module, which is based on an Italian BERT model. This component plays a central role. As already mentioned, BERT operates in an extractive way, retrieving answers directly from specific contextual data sets. This component plays a central role. As already mentioned, BERT operates in an extractive manner, retrieving answers directly from specific contextual data sets. These data sets consist of contextual information related to works of art, including descriptive texts and historical documentation. This design allows the answers to remain accurate and verifiable. Once the QA system identifies the most relevant part of the context and formulates the corresponding response, the text is sent to the text-to-speech (TTS) component, also implemented via Wit.ai. This ensures that the response is provided to the visitor in natural spoken form, maintaining the flow of conversation. The voice output is embodied by a 3D virtual avatar, which acts as the visible interface of the system. The avatar enhances user immersion and engagement, acting as a guide that provides contextually relevant explanations and answers in real time. The voice output is embodied by a 3D virtual avatar, which acts as the visible interface of the system. The avatar enhances user immersion and engagement, acting as a guide

that provides contextually relevant explanations and answers in real time. To ensure that interactions are not only effective but also analysable, all user requests and system responses are recorded through an xAPI framework. This enables the structured collection of data related to user behaviour, frequently asked questions, and system performance. These records form the basis for subsequent evaluation and iterative improvements. This modular architecture balances natural interaction, a controlled knowledge base and traceability, creating a reliable and engaging virtual guide based on artificial intelligence for cultural heritage contexts.

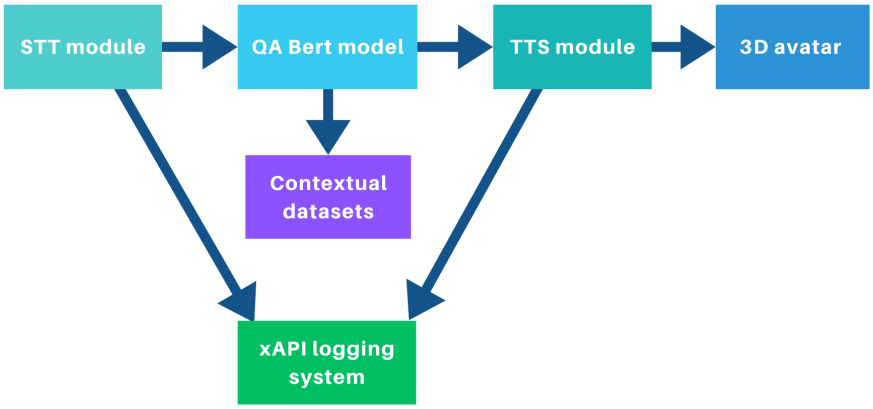


FIGURE 5.1: System architecture

5.2 Training and adaptation of the QA model

The effectiveness of a virtual guide system in the field of cultural heritage depends largely on the level of detail and flexibility of the question-answering model. The model must be able to balance two often conflicting objectives: linguistic versatility and scientific reliability. On the one hand, the system must understand and respond to questions asked by visitors in natural language, which can be different in style, complexity, or level of prior knowledge. On the other hand, its answers must be closely based on authoritative sources, making sure that the information it gives is verifiable and contextually appropriate. To satisfy this requirement, the model was trained and optimised through a series of interconnected steps. These include the use of robust linguistic resources such as large-scale annotated datasets for question-answering and lightweight implementation strategies, which ensure that the model can run efficiently within immersive environments without compromising responsiveness. This approach allows the system not only to perform well in controlled evaluations, but also to adapt dynamically to the demands of real-world interaction in museums and historical sites, where user engagement depends on both the fluidity of dialogue and the accuracy of content.

To improve accuracy and usability within the computational limitations of real-time interaction in immersive environments, the system was built using MobileBERT (Figure 5.2), a compact yet high-performance transformer architecture optimised specifically for mobile and embedded devices.

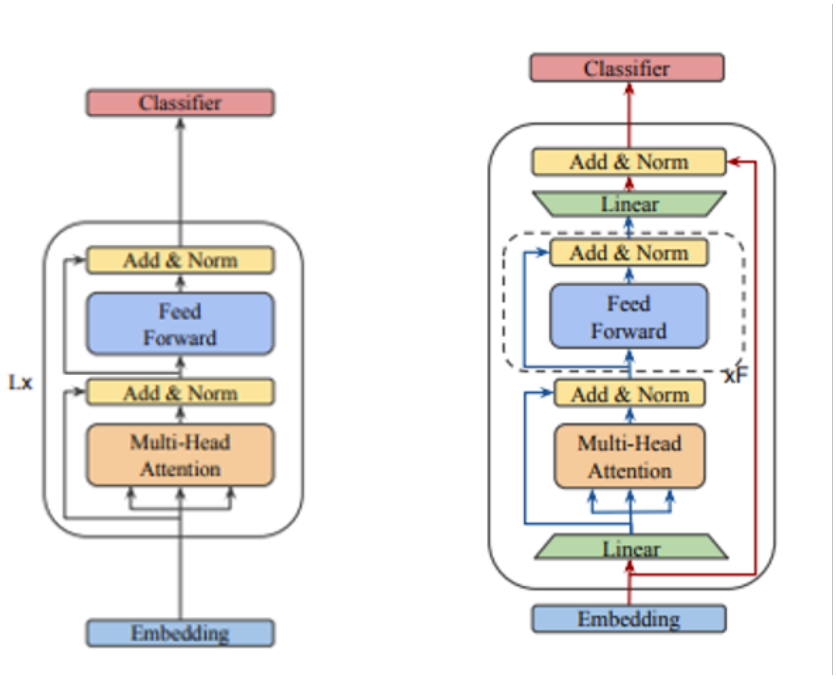


FIGURE 5.2: BERT Model vs. MobileBERT model

Unlike standard BERT models, which contain hundreds of millions of parameters and require substantial computational resources, MobileBERT is designed to retain much of BERT's representational power while significantly reducing model size and inference latency. MobileBERT introduces several architectural innovations:

- Bottleneck structures: the model uses smaller intermediate layers, which reduce the number of parameters without compromising the expressiveness of the representations.
- Inverted bottleneck embedding projections: These compress and expand hidden states, ensuring information flow is preserved while reducing memory usage.
- Task-specific distillation: MobileBERT has been pre-trained using a teacher-student concept, where a larger BERT model transfers knowledge to the smaller student model. This enables MobileBERT to capture detailed semantic and syntactic patterns despite its small size.

These characteristics make MobileBERT particularly well-suited for real-time applications such as virtual museum guides, where the system must process natural language queries, extract answers, and generate responses with minimal latency. Furthermore, its low computational impact allows for implementation on resource-limited platforms, including VR headsets and mobile devices, without the need for powerful external servers. Thus, MobileBERT enables the system to provide accurate, extractive responses while maintaining the responsiveness necessary for a seamless visitor experience.

The model was further converted into TensorFlow Lite (TFLite) format, enabling lightweight deployment while maintaining strong extractive QA capabilities. This process enables the model to be optimised for execution on devices with limited

computing power and memory. In fact, the TFLite version reduces both storage footprint and inference time without significantly compromising accuracy. This adaptation provides the virtual guide with the ability to process user queries, extract contextually relevant information, and generate responses with minimal latency directly within the device. In practice, the implementation of TFLite means that the avatar is able to process user queries, extract contextually relevant information, and generate responses with minimal latency directly within the VR environment. This design choice was essential to preserving the immersive experience, where even small delays in response can disrupt the fluidity of interaction and diminish user engagement. By combining the efficiency of MobileBERT with the portability of TensorFlow Lite, the resulting system achieves a balanced compromise between accuracy, speed, and deployment, making it particularly suitable for cultural heritage applications where interactive reactivity is fundamental.

5.2.1 Dataset: SQuAD-it

The training phase was based on SQuAD-it (Croce, Zelenanska, and Basili, 2018), the Italian adaptation of the Stanford Question Answering Dataset (Rajpurkar et al., 2016), which has become a reference resource for the evaluation and development of extractive QA systems in Italian. SQuAD-it contains more than 60,000 question-answer pairs derived from English Wikipedia articles and then semi-automatically translated into Italian. Its structure combines factoid questions (e.g., names, dates, places) with more comprehension-oriented questions that require the model to identify longer stretches of text or infer meaning from context. In fact, in museum and archival settings, visitors typically ask questions that are similar to the factoid and descriptive queries found in SQuAD-it, such as questions about dates, authorship, stylistic features, or historical significance. To provide a reference point for evaluating the QA model, the SQuAD-it dataset was benchmarked using the DrQA system (Chen et al., 2017). DrQA is an extractive question answering framework originally developed for English that retrieves relevant paragraphs from a knowledge source and predicts the span containing the answer. For SQuAD-it, DrQA was adapted to the Italian language, allowing for reliable performance measurement on this dataset. The evaluation was conducted by calculating two main metrics: Exact Match (EM), which indicates the percentage of predictions that exactly match the ground truth, and Macro F1, which measures the token-level overlap between the predicted and reference answers. On the SQuAD-it test set, the adapted DrQA system achieved an EM of 56.1 and an F1 score of 65.9. These reference values establish a baseline for QA in Italian and provide a useful point of comparison for evaluating the performance of optimised models on the same dataset. The results highlight both the challenge of adapting QA systems to Italian and the potential for further improvements through model selection, optimisation strategies, and context management.

5.2.2 MobileBERT Training on SQuAD-it

By combining MobileBERT with training on SQuAD-it, the system strikes a balance between performance and usability. On the one hand, the model benefits from the linguistic richness of SQuAD-it, which helps it handle the nuances of questions and answers in Italian; on the other hand, MobileBERT's efficiency ensures that these capabilities can be delivered quickly and smoothly, even in a real-time virtual environment. This means that the virtual museum guide is accurate in retrieving the right information, but is also responsive and reliable in practice. In other words, the

technical foundations translate directly into a more natural and engaging experience for visitors, where answers seem immediate and reliable.

The decision to use MobileBert as the starting model for fine-tuning on the SQuAD-it dataset was mainly driven by the availability of MobileBERT in TFLite format, a feature that, as we have said, makes it particularly suitable for running on mobile devices and in scenarios with limited computational resources. However, this decision raises some critical issues. MobileBERT was originally pre-trained on English-language corpora and uses an English-based vocabulary. Consequently, when fine-tuning on Italian data, the model must adapt to a language not covered in the pre-training phase, with possible losses in tokenisation efficiency and therefore in overall performance. Furthermore, compared to Italian or multilingual models available in the scientific community, such as UmBERTo (Parisi, Francia, and Magnani, 2020) or AlBERTo (Polignano et al., 2019), MobileBERT has the advantage of being compatible with TFLite, but does not have a vocabulary optimised for Italian. A possible alternative would be to convert existing Italian models to TFLite, but this is not always easy and often results in large models that are less efficient on mobile devices. For these reasons, the use of MobileBERT represents a compromise between the need for a lightweight, deployable model and the awareness that the linguistic performance in Italian may be lower than that achievable with models specifically trained on the language.

5.2.3 Training Results and Computational Setup

The model was trained using a workstation equipped with an Intel Core i7 CPU and an NVIDIA Titan V GPU. The initial training of the MobileBERT model on the SQuAD-it dataset required approximately 9 hours and 25 minutes, achieving an F1-score of 0.60 and an EM of 0.48. For comparison, a BERT-base model trained on the same dataset required 27 hours and 40 minutes, achieving slightly higher performance metrics, with an F1-score of 0.62 and an EM of 0.51. Further experiments were conducted to evaluate the impact of increasing the number of training epochs on MobileBERT. The results are summarised as follows:

- 3 epochs: Execution time $\approx$ 50,341 seconds ($\sim$14 hours), EM = 0.499, F1 = 0.619
- 4 epochs: Execution time $\approx$ 67,644 seconds ($\sim$18.8 hours), EM = 0.504, F1 = 0.628
- 5 epochs: Execution time $\approx$ 84,802 seconds ($\sim$23.6 hours), EM = 0.501, F1 = 0.623
- 6 epochs: Execution time $\approx$ 101,248 seconds ($\sim$28.1 hours), EM = 0.500, F1 = 0.621

These results show that MobileBERT strikes a good balance between speed and performance. The F1-score peaks at four epochs, and training for longer doesn't bring meaningful improvements, while it takes a lot more time. Compared to BERT-base, MobileBERT is much faster, though its performance is slightly lower. The EM staying around 0.50 suggests that the model often gets part of the answer right, which is reasonable given the complexity of the SQuAD-it dataset. Overall, MobileBERT appears to be a solid choice when computational efficiency matters."

Figure 5.3 shows a context passage on the painting The Starry Night, highlighting the answers given by the Bert QA model to the following questions: the name of the artwork, the year it was painted, and its current location.

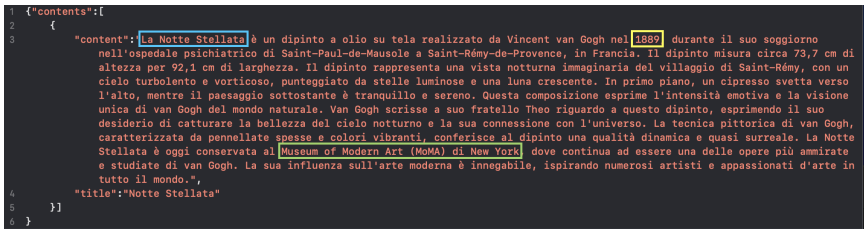


FIGURE 5.3: An example of a QA context related to an artwork, with the answers highlighted

5.3 Virtual Environment and user interaction

The virtual museum environment has been designed to offer an immersive and interactive experience, in which users can explore cultural heritage by interacting with a 3D avatar that acts as a tour guide. The environment includes a curated selection of artworks and historical objects, digitally reconstructed and positioned to simulate a realistic museum environment. Particular attention was dedicated to the spatial organisation, lighting and accessibility of the exhibits, in order to ensure an intuitive and engaging browsing experience. Within this context, the avatar serves as the primary interface for information retrieval, responding to users' questions with answers generated by the BERT-based Question Answering system. Each artwork or exhibit is linked to structured contextual data, allowing the avatar to provide accurate and relevant responses. This integration between the environment, the digital assets, and the QA system creates a coherent, informative, and interactive tour, forming the foundation for the subsequent discussion of avatar design, interaction modalities, and user scenarios.

5.3.1 Avatar design

Once the BERT-based Question Answering model was trained and integrated, a 3D avatar was developed to act as a virtual guide within the immersive museum tours. Its role is that of a touristic guide to whom users can pose questions about the artworks or historical objects they encounter in the virtual environment. The avatar is not merely a visual placeholder but a central interface for knowledge delivery, capable of engaging users in a natural and interactive way. An open-source humanoid model licensed under Creative Commons from Sketchfab, which is a widely used online repository for sharing and downloading 3D models, was used (Figure 5.4). The model is provided in T-Pose, which simplifies the rigging and animation process. It includes multiple blend shapes for facial expressions, allowing for nuanced movements such as blinking, subtle eye shifts, and basic mouth animations. These features make the avatar capable of conveying attention, emotion, and responsiveness, which are critical for maintaining user engagement in a virtual environment. The choice of this specific model was driven by several factors: its high compatibility with Unity 3D, its readiness for animation without extensive manual rigging, and the inclusion of facial blend shapes that facilitate synchronization with speech. By leveraging these existing resources, it was possible to ensure a high level of realism and expressiveness. The presence of the avatar in the VR environment was designed to enhance immersion and provide a sense of co-presence, making the educational experience more engaging.

FIGURE 5.4: 3D avatar

The avatar was implemented in Unity 3D, with the final development conducted on Unity version 2022.3.62, targeting the Meta Quest platform. Over the course of development, several Unity releases were adopted, requiring continuous adaptation of the project to maintain compatibility with updated features and to replace deprecated functions. This iterative process ensured that the application remained stable and took advantage of the latest improvements offered by the engine. In addition, the implementation integrated the official Meta Quest libraries and SDKs, such as the Meta XR Core SDK package, which provides access to device-specific functionalities like hand tracking, controller input and VR rendering optimizations. These resources were essential for achieving a smooth and responsive experience on a standalone VR headset, allowing the avatar to perform seamlessly in immersive environments without requiring external computing resources.

To enable natural voice interaction, Speech-to-Text and Text-to-Speech features were progressively developed and refined during the project. In the initial versions, IBM Watson's cloud API was used to manage both STT and TTS, offering reliable voice output in Italian. However, the service is only free up to a limited number of characters, after which it becomes expensive to use, making it less sustainable for large-scale implementation scenarios. Subsequent tests were conducted with OpenAI Whisper (Radford et al., 2023), a neural STT system known for its robustness across multiple languages and noisy conditions. While Whisper provides excellent transcription quality, its computational requirements make it less suitable for standalone VR devices such as the Meta Quest, where real-time performance is crucial.

In fact, Whisper's inference introduces noticeable latency on resource-constrained devices. For these reasons, the system ultimately adopted Wit.ai, a lightweight and efficient STT/TTS solution that integrates smoothly into Unity-based applications and offers good latency performance on VR headsets. Wit.ai ensures fast speech recognition and natural-sounding voice synthesis in Italian, allowing the pipeline to capture spoken input, transcribe it, process the question through the BERT QA model, and return the answer in audio form with minimal delay.

For enhanced realism, lip-sync functionality was progressively integrated into the avatar. In the early development stages, a free Unity3D asset called Lavstar was employed, which allowed basic lip movement by animating a single blend shape corresponding to mouth opening and closing. Although relatively simple, this approach already provided a noticeable improvement in user perception by making the avatar appear to "speak" when generating audio responses. Over time, more advanced solutions were tested. Lip-sync techniques vary in complexity: some libraries manage only binary mouth opening, while others can approximate viseme-based animations, synchronizing mouth shapes with specific phonemes for more natural results. However, these solutions often come with increased computational costs, which can negatively affect performance on standalone VR devices. In the final implementation, the system leveraged the Meta Quest lip-sync libraries, which natively support blend shape control optimized for the headset's hardware. This integration ensures efficient synchronization between generated speech and the avatar's facial animation, while maintaining real-time responsiveness. For deployment efficiency, the implementation currently focuses on a simplified open-and-close mouth animation synchronized with speech output, complemented by idle animations and eye-blinking scripts to simulate natural presence.

The combination of a fully animated 3D avatar, natural voice interaction, and real-time QA processing results in an immersive and engaging system that bridges cultural heritage knowledge and interactive virtual experiences.

5.3.2 Virtual environment design

The virtual environment was designed to simulate a museum-like space where users could explore artworks in an immersive way. As discussed in previous chapters, such environments can be created either from 3D scans of real locations or through computer-generated reconstructions. In the early stages of the project, scanned environments such as historical churches were used as testbeds (Figure 5.5). Within these reconstructions, the virtual avatar was placed directly inside the scanned heritage sites, allowing it to act as a guide and provide information about the cultural assets faithfully reproduced in 3D. These preliminary scenarios demonstrated the potential of combining real-world scans with interactive virtual guides, creating an engaging way to explore cultural heritage.

Based on this experience, a ready-made 3D museum model was subsequently selected from Sketchfab, which was then optimised and adapted to ensure compatibility and performance on standalone VR headsets. Optimisation involved modifications to textures, lighting, and polygon count to balance visual quality with real-time rendering constraints. The transition to a virtual museum environment offered greater flexibility in curating artworks and organising interactive experiences, while maintaining the immersive and educational nature of the previously scanned environments. Within this environment (Figure 5.6), ten representative works of art were integrated to enrich the user experience: L*a Passeggiata* by Monet, *David* by Michelangelo, *Mona Lisa* by Leonardo da Vinci, *The Starry Night* by Van Gogh,

FIGURE 5.5: A 3D avatar inside the SS. Crocifisso al Calvario Church

Nike of Samothrace, *The Thinker* by Rodin, *The Birth of Venus* by Botticelli, *Girl with a Pearl Earring* by Vermeer, *The Scream* by Munch, and *Cupid and Psyche* by Canova. These artworks were obtained from Sketchfab, which provides both reconstructed 3D models and photogrammetry-based scans. In fact, some of the imported assets are actual 3D scans, while others are reconstructed models, showing the potential of combining different sources within a single virtual museum setting.

5.3.3 User interaction

Inside the virtual museum, the user is accompanied by the 3D avatar, which acts as a virtual guide. As the visitor moves within the environment using the Meta Quest controllers, the avatar follows and stays close, reinforcing the feeling of being assisted during the tour. Each artwork included in the environment is associated with a descriptive context containing information about the title, artist, year, and cultural significance. When the user approaches a specific artwork, they can initiate an interaction with the avatar by pressing the B button on the controller. This deliberate activation ensures that the avatar listens only when the user is ready to ask a question, rather than being in constant listening mode. By requiring the user to explicitly initiate the avatar's listening phase, the system avoids unwanted input and allows interactions to occur at the user's pace, creating a more controlled and focused experience.

The user may then naturally ask the avatar questions about the selected artwork. Once the B button is pressed again, the recorded audio is processed:

1. Speech-to-Text converts spoken input into written text.

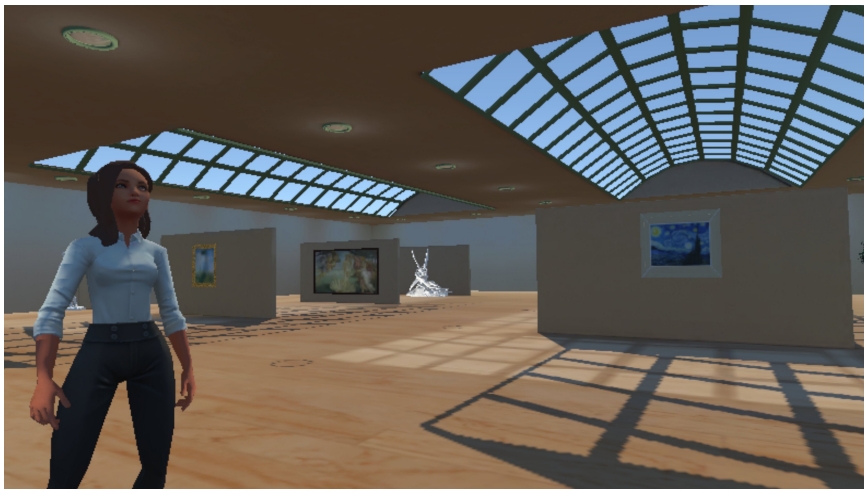


FIGURE 5.6: An example of a 3D avatar within the VR museum environment.

2. The BERT-based QA model analyzes the transcribed question and retrieves the most relevant answer from the descriptive context associated with that artwork.

3. The answer is converted back into speech through Text-to-Speech and vocalized by the avatar, which simultaneously animates its lips to match the spoken response.

This interaction pipeline enables a conversational and intuitive experience: instead of reading information panels or navigating menus, the user can engage in a natural dialogue with the avatar, similar to a real guided museum visit.

5.4 Evaluation strategy

The evaluation of the system has several goals that go beyond simply verifying the functioning of the technology. The idea is to understand how well the virtual guide performs its role, both from a technical point of view and from the perspective of the visitor's experience inside the museum. The main objective is to measure the accuracy of the Question Answering model. Since the system was designed to provide visitors with information about the artworks, it is important to verify that the answers it gives are correct and precise. If the information was incorrect or incomplete, the educational value of the experience would be compromised.

5.4.1 Metrics for evaluating Question Answering Systems

Therefore, to evaluate the model, it is not sufficient to simply check whether an answer is correct or incorrect. The system must be evaluated based on various accuracy parameters, including partial correctness and semantic proximity. For this reason, various automatic evaluation parameters can be used (Risch et al., 2021;Chiazzese and Lo, 2023). The Exact Match metric provides a strict measure: it checks whether the response returned by the model exactly matches the reference response in the dataset. While this is useful for measuring accuracy, it can sometimes be overly

strict, especially if the predicted response is semantically correct but phrased differently. Formally, Exact Match can be defined as:

$$\mathrm{EM}(\hat{y}, y) = \begin{cases} 1, & \text{if } \hat{y} = y, \\ 0, & \text{otherwise.} \end{cases}$$

where $\hat{y}$ is the predicted answer and y is the *ground truth*, the correct reference answer in the dataset, against which the prediction is evaluated.

To account for partial correctness, the F1 score is used. This metric balances precision and recall, effectively capturing cases where the model retrieves some of the correct information but omits certain details. It provides a broader view of performance, showing not only whether the response is completely right or wrong, but also how close it is to the objective truth. F1 is computed as the harmonic mean of precision and recall:

$$F1 = \frac{2 \cdot \mathit{Precision} \cdot \mathit{Recall}}{\mathit{Precision} + \mathit{Recall}}$$

with

$$\mathit{Precision} = \frac{|\hat{y} \cap y|}{|\hat{y}|}, \quad \mathit{Recall} = \frac{|\hat{y} \cap y|}{|y|}$$

BLEU (Papineni et al., 2002), a metric originally developed for machine translation, evaluates the similarity between predicted responses and reference responses based on overlapping n-grams. This provides an indication of whether the model is producing responses that are formulated similarly to human references. However, BLEU focuses primarily on surface-level matches and does not fully capture meaning. To address semantic similarity, METEOR (Banerjee and Lavie, 2005) extends BLEU by including synonym matching and leveraging WordNet. This means that if the model's response uses slightly different words but conveys the same meaning, it will receive a partial score, which is particularly important when dealing with natural language and the various ways of describing works of art or historical facts. Finally, BERTScore (Zhang et al., 2019), in both Vanilla and Trained variants, represents a more recent approach that leverages the embeddings of a pre-trained BERT model to compare answers. By computing the cosine similarity between the predicted and reference sentence embeddings, BERTScore captures deeper semantic relationships, allowing the evaluation to recognize when answers are meaningfully similar even if the exact words differ. The *Vanilla* variant uses a standard pre-trained BERT model without additional task-specific fine-tuning, providing a general measure of semantic similarity based on the knowledge learned during BERT's pre-training phase. The *Trained* variant, on the other hand, can be fine-tuned on a dataset specific to the domain or task, allowing the embeddings to better reflect the nuances of the target language, terminology, or context. In the case of cultural heritage QA, this means the Trained variant may better capture relationships between historical or artistic terms that might not appear frequently in general corpora. As shown in Table 5.1, different evaluation metrics behave differently depending on how the predicted answer compares to the ground truth. Exact Match requires a perfect match, F1 accounts for partial correctness, BLEU and METEOR handle wording variations, and BERTScore captures semantic similarity.

By combining selected metrics, it is possible to obtain a comprehensive view of the model's performance. This approach not only highlights where the system gets answers exactly right, but also where it produces responses that are semantically

Question	Who sculpted The David?
Context	The David, a masterpiece of Renaissance sculpture, was created by Michelangelo between 1501 and 1504.
Metric	**Example**
Exact Match	Ground truth: "The David was sculpted by Michelangelo." Predicted: "The David was sculpted by Michelangelo." → perfect match.
F1 Score	Ground truth: "The David was sculpted by Michelangelo in the early 16th century." Predicted: "The David was sculpted by Michelangelo." → correct information but incomplete.
BLEU	Ground truth: "The statue was created by Michelangelo." Predicted: "The sculpture was made by Michelangelo." → BLEU penalizes the difference in wording.
METEOR	Ground truth: "The statue was created by Michelangelo." Predicted: "The sculpture was made by Michelangelo." → METEOR recognizes synonyms (statue/sculpture, created/made).
BERTScore	Ground truth: "The work was made by Michelangelo." Predicted: "Michelangelo is the artist of the work." → words differ significantly, but BERTScore captures the semantic equivalence.

TABLE 5.1: Examples of model evaluation using different metrics.

accurate, even if phrased differently, an important consideration in cultural heritage contexts where descriptions may vary yet still convey correct information.

5.4.2 Design of the pilot study

The pilot study was designed to test the integration of the QA system within the immersive museum environment and to explore how different interaction strategies could affect the quality of the responses. Participants were asked to wear the VR headset and explore the virtual museum, interacting with the avatar guide by asking questions about the artworks they encountered inside the VR environment. Two experimental conditions were examined to observe how users interacted with the QA system in different circumstances. In the first condition, participants entered the museum and were free to look around, choose the works of art they found interesting, and ask their questions directly to the avatar. The QA model, based on a BERT architecture, attempted to provide answers based only on information contained in its contextual input. This meant that if a participant asked a question that was outside the scope of the available knowledge or phrased the question in a way that did not match the encoded data, the system returned an incomplete or simply incorrect answer. This situation highlighted a critical limitation: without sufficient contextual background, the model was unable to reliably meet user expectations.

In the second condition, the interaction was slightly different. Before leaving space for questions, the avatar itself took on a more proactive role by presenting each work of art through a brief introduction. This storytelling, expressed in natural language, outlined the essential information about the work, as if the participant were taking part in a guided tour. In this way, the avatar did not merely transfer knowledge, but also implicitly suggested the " limitations" of the available context. After listening to this description, users were then better prepared to ask focused and relevant questions. On the one hand, participants felt more engaged in the experience,

while on the other hand, the QA system had a clearer contextual reference point from which to generate its responses. In practice, this configuration simulated the cognitive process of "context reading" before starting a dialogue and helped bridge the gap between the technical limitations of the system and the natural expectations of human conversation.

The involvement of multiple participants was not primarily aimed at increasing the statistical representativeness of the study, but rather at enriching the linguistic diversity of the inputs. When faced with the same work of art, each user tended to formulate questions in their own way, sometimes with very similar content, but expressed through different words, syntactic structures, or levels of specificity. This variability, far from being a marginal detail, proved to be an important element of the experiment. It highlighted how interaction in natural language is rarely uniform: two individuals may ask essentially the same question, but the question-and-answer system may interpret them differently and produce answers that vary significantly in terms of accuracy or completeness. In this sense, the variety of participants worked almost like a stress test for the robustness of the model. To systematically evaluate the system's performance, all questions asked by users, along with the answers generated by the QA model, were carefully collected and archived. Each of these user-generated queries was then reviewed to determine what the correct answer would be in the given context, also asking users themselves what answer they expected. This process effectively created a "gold answer", a benchmark against which to compare the model's results. By comparing the model's results with these answers, it was possible to apply the evaluation framework described in the previous section. Measures such as METEOR or BERTScore made it possible to detect cases of partial correctness or semantic proximity. In this way, the evaluation was not reduced to a binary assessment of right or wrong, but reflected the more nuanced and interpretive nature of cultural heritage communication.

5.5 User Experience Tracking with xAPI

To gain a deeper understanding of how users interact with the virtual museum and the AI-powered guide, the system incorporates Experience API tracking. xAPI is a widely adopted standard for capturing and storing detailed records of user interactions across different platforms, applications, and environments. Unlike traditional logging systems, which typically record simple events or page views, xAPI allows for granular and structured tracking of learning experiences. Each recorded action is stored as a "statement" in the form of "actor – verb – object":

- *Actor* identifies who performed the action as the user or participant.
- *Verb* describes the action taken (for example, "asked", "answered", "viewed").
- *Object* specifies the target of the action (for example, a specific question, artwork, or exhibit in the virtual museum).

Beyond this basic triple, statements can include additional optional elements such as:

- *Context* that provides information about the environment or situation in which the action occurred (for example, which virtual room or which artwork was being observed).

- *Result* that captures the outcome of the action, such as whether the answer was correct or how long it took to respond.
- *Timestamp* that marks the exact time the action occurred.
- *Extensions* that allow for custom metadata specific to the application or research questions.

The flexibility and richness of xAPI make it particularly well-suited for immersive environments and interactive educational applications (Farella et al., 2021; Chiazzese et al., 2023), such as a virtual museum guided by an AI avatar. In fact, this structure allows to capture not just what users do, but also how and under what circumstances. For instance, in the context of the virtual museum, statements can record when a user asks a question, which artwork the question refers to, the answer returned by the AI model, and the timing of the interaction. This level of detail enables a comprehensive and structured record of user behavior, which is crucial for analyzing engagement, learning outcomes, and system effectiveness in immersive environments.

These statements are stored in a Learning Record Store (LRS), such as Learning Locker, which serves as a centralized and secure repository for all interaction data collected via xAPI. Learning Locker is one of the most widely used open-source LRS solutions and is fully compliant with the xAPI 1.0.3 specification. It can be installed locally or hosted in the cloud, and it offers RESTful APIs that make it easy to query, filter, and combine statements for analysis. In addition to storing data, Learning Locker provides built-in tools to create dashboards and visualizations, and it integrates seamlessly with platforms such as Moodle. Thanks to these features, it is well-suited for both research and practical use, allowing users to track user engagement, evaluate the effectiveness of learning activities, and gain meaningful insights from interactions within the virtual museum.

The LRS not only stores raw statements, but also ensures they are structured, searchable, and retrievable for later analysis. By centralizing data from multiple sources, the LRS enables a unified view of user interactions across sessions, users, and environments. In the context of the virtual museum, each statement, such as a question asked to the AI guide, the response provided, and the context in which it occurred, is recorded in the LRS. This allows to analyze patterns of user behavior, engagement, and learning outcomes, track how users interact with specific artworks or exhibits, and assess the effectiveness of the AI-powered guide. Additionally, the LRS supports integration with learning analytics tools, enabling visualization, reporting, and deeper insights into both individual and aggregate user experiences.

From a Learning Analytics perspective, xAPI enables the systematic analysis of user behavior and engagement. By aggregating and interpreting xAPI statements, it is possible to measure not only the accuracy of responses provided by the AI system but also patterns of exploration, types of questions asked, response times, and areas where users may struggle or lose engagement. This information can then be used to improve both the usability and educational effectiveness of the system, creating a feedback loop that supports evidence-based design decisions. In short, xAPI provides a standardized, detailed, and actionable view of user interactions, turning raw behavior data into meaningful insights for evaluating and enhancing learning experiences in immersive cultural heritage applications.

To represent the interaction between the user and the AI guide in a structured way, specific verbs are adopted within the xAPI framework. For instance, the verb "*asked*" (defined at https://xapi.elearn.rwth-aachen.de/definitions/generic/verbs/asked)

is used to indicate that a participant posed a question, while the verb "*answered*" (defined at https://xapi.elearn.rwth-aachen.de/definitions/generic/activities/answer) is used when the system provides a response. These verbs make it possible to consistently capture the flow of dialogue, recording not only the presence of a question or an answer, but also the exact moment in which the action occurred and its relation to a specific artwork.

Figure 5.7 shows the Learning Locker platform, where the xAPI statements are stored. Figure 5.8 provides an example of a statement that shows how a user asking a question about Munch's Urlo could be represented in xAPI format.

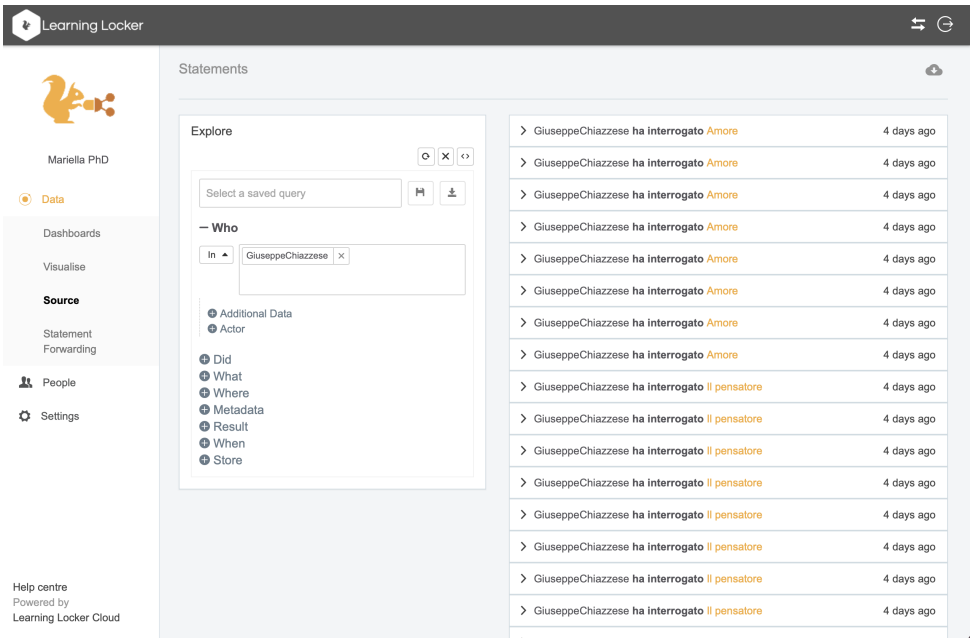


FIGURE 5.7: An example of xAPI statements stored in the Learning Locker platform

This structure clearly identifies the actor (the user), the verb (the action performed), the object (the artwork involved), the result (the actual content of the question) and the timestamp of the interaction.

5.6 Results

The data set analyzed in this study initially consisted of more than 300 questions related to the artworks in the virtual museum. These questions were collected from users interacting with an art information system and reflected the natural questions that a viewer might ask when encountering a work of art for the first time. During an initial analysis, it became evident that many questions were highly repetitive, particularly those concerning the author, the date of creation, or the title of the artwork. These standard questions, commonly posed by users immediately after interacting with a work of art, were consolidated and cleaned to avoid redundancy. The final data set therefore represents a curated selection of meaningful questions, preserving the diversity of inquiry while ensuring analytical clarity. However, even after this cleaning process, numerous questions remained similar in intent but were phrased differently. Interestingly, these variations often elicited different answers, reflecting subtle distinctions in interpretation or emphasis. For example, questions

```
{
  "actor": {
    "objectType": "Agent",
    "name": "GiuseppeChiazzese",
    "mbox": "mailto:giuseppe.chiazzese@itd.cnr.it"
  },
  "verb": {
    "id": "https://xapi.elearn.rwth-aachen.de/definitions/generic/verbs/asked",
    "display": {
      "en-US": "Asked",
      "it-IT": "ha interrogato"
    }
  },
  "object": {
    "objectType": "Activity",
    "id":
        "https://xapi.elearn.rwth-aachen
        .de/definitions/virtualReality/extensions/activity/vrObjectName",
    "definition": {
      "name": {
        "en-US": "Urlo",
        "it-IT": "Urlo"
      },
      "description": {
        "en-US": "Exhibit being questioned about",
        "it-IT": "Opera su cui l'utente ha fatto una domanda"
      }
    }
  },
  "result": {
    "response": "Q: da cosa è stato condizionato l'autore nella realizzazione dell'opera | A: la
        morte della madre e della sorella,"
  },
  "timestamp": "2025-09-08T13:37:38.563Z"
}
```


FIGURE 5.8: An example of xAPI statement

about what a work of art "represents" could generate responses focused on formal aesthetics, symbolic meaning, or emotional impact, depending on the phrasing. This phenomenon highlights the intrinsic complexity of natural language in art-historical inquiry, where minor differences in wording can influence the conceptual orientation of the response. This cleaning process was essential to allow a more precise examination of the different types of questions, the underlying conceptual clusters, and the variety of responses. By removing repetitive or trivial questions, the analysis could focus on both the factual and interpretative dimensions, revealing how users seek to understand not only the formal and technical aspects of artworks but also their symbolic, historical, and emotional significance.

5.6.1 Classification of Art-Related Questions

A systematic analysis of the questions posed to the system highlights recurring patterns in both form and content, which can be organized into distinct typological categories and conceptual clusters. This classification is not only useful for understanding how art-historical knowledge is structured in natural language, but also for identifying areas in which interpretation, ambiguity, or subjectivity play a significant role. The questions, in general, can be classified into six main categories:

1. Identification questions seek factual information such as authorship, date, or location (for example, "Who is the author of The Scream and in what year was the painting created?").

2. Descriptive questions request formal or visual accounts of artworks (for example, "How is the central figure represented in the painting?").

3. Symbolic questions inquire into meaning, allegory, and interpretative dimensions (for example, “What is the symbolic significance of the kiss between Cupid and Psyche?”).

4. Technical questions concern materials, techniques, and stylistic features (for example, “What material was used to construct this sculpture?”).

5. Historical-contextual questions situate the artwork within broader cultural and chronological frameworks (for example, “In which period was the work produced?”).

6. Comparative questions address multiplicity, alternative versions, or cross-artistic relationships (for example, “Does a second version of Cupid and Psyche exist, and if so, where is it located and how does it differ from the original?”).

Beyond typology, the questions naturally converge into broader conceptual clusters:

- Authorship and Historical Context (who, when, where).
- Formal and Technical Description (materials, techniques, use of light, anatomical precision).
- Symbolism and Interpretation (allegory, philosophical meaning, emotional resonance).
- Multiplicity and Comparison (versions, reproductions, intertextuality).
- Reception and Impact (influence on the history of art, audience perception).

These groups demonstrate how the investigation of works of art oscillates between empirical knowledge (date, author, place) and interpretative speculation (meaning, symbolism, cultural resonance).

One of the most notable observations is the coexistence of similar questions with divergent or heterogeneous answers. For example, the recurring question "What does the work represent?" elicits highly specific answers that depend entirely on the work in question: Botticelli’s Birth of Venus emphasizes idealized beauty, Vermeer’s Girl with a Pearl Earring foregrounds individuality, while Canova’s Cupid and Psyche emphasizes allegorical love. This highlights the need for contextualization in art-historical interpretation: identical linguistic forms conceal highly differentiated semantic contents. Equally important are cases where questions require open-ended answers, particularly in the areas of symbolism and meaning. In these cases, the answers vary in length, depth, and theoretical orientation, reflecting the interpretive plurality characteristic of art history as a discipline. In contrast, closed questions about materials, dates, or authorship tend to elicit short, unequivocal, and verifiable answers.

After the data cleansing process, which eliminated redundant or overly repetitive questions, the refined dataset provided a clearer basis for systematic analysis. With a consolidated and coherent set of questions, it was possible to examine recurring patterns in both form and content.

For example, in the case of Canova’s sculpture "Cupid and Psyche":

- Q: What does the sculpture represent? $\rightarrow$ A: “The artwork represents the pinnacle of idealized beauty.”

- Q: Who is depicted in the sculpture? → A: "Cupid awakening Psyche with a kiss."
- Q: How does Canova convey beauty in the sculpture? → A: "Through the gestures of the bodies and the delicacy of the faces."

In this case, even if all the questions concern the same sculpture, the answers focus on different aspects: aesthetic ideals, narrative content and expressive technique.

For example, in the case of Munch's painting "The Scream":

- Q: Why is the figure anguished? → A: "Due to the death of his mother and sister."
- Q: What does the distorted face signify? → A: "It represents existential anguish."
- Q: How is the figure depicted? → A: "A central figure with a distorted face in an expression of despair."

Here, small variations in wording lead to responses that emphasize biographical context, symbolic meaning, or visual description.

In the case of Vermeer's painting "Girl with a Pearl Earring":

- Q: Who is the girl in the painting? → A: "Her identity remains unknown."
- Q: What does the painting depict? → A: "A young woman with a turban and a large pearl earring."
- Q: What type of painting is this? → A: "A tronie, depicting a character in exotic costume or with a particular expression."

In this example, small variations in the question produce responses that highlight identity, visual description, or painting type.

Even when questions share the same semantic intent, the system may produce different answers depending on phrasing, focus, or context. Two representative examples are reported below:

1. Cupid and Psyche (Canova):
    - Q: "Who are the figures represented in the sculpture?" → A: "Cupid awakens Psyche"
    - Q: "Who are the characters in the artwork?" → A: "Cupido and Psiche"
2. The Scream (Munch)
    - Q: "Why is the painting called The Scream?" → A: "Profound anguish experienced by the artist."
    - Q: "What does the painting reflect?" → A: "Profound anguish."

These examples illustrate that even minor differences in question phrasing can lead to divergent answers, highlighting the interpretive sensitivity of art-historical inquiry.

5.6.2 Evaluation results

The evaluation of the model's answers against the gold standard was performed using BERTScore and METEOR. Although these metrics do not cover every aspect of answer quality, they provide a robust and interpretable foundation for analysis. METEOR can sometimes underestimate a correct response if the phrasing differs from the reference, while BERTScore better reflects the semantic correctness even when the wording varies. This combination allows us to balance the evaluation between lexical fidelity and conceptual accuracy.

The analyzed dataset included 272 question–answer pairs, covering all types of questions without distinction. The BERTScore results show a strong similarity between the answers generated by the model and the reference answers, with an average Precision of 0.818, a Recall of 0.799, and an F1-score of 0.807. In other words, the model generally captures most of the relevant information found in the gold answers and expresses it in a semantically meaningful way. The slightly higher Precision compared to Recall suggests that the model tends to provide accurate information, though occasionally it may miss some details from the reference answers. The F1-score confirms that overall, the model performs reliably across a wide range of questions.

The METEOR metric, which evaluates exact matches, word stems, synonyms, and word order, gave an average score of 0.474. While this is lower than the BERTScore F1, it is expected in open-ended question answering tasks. Even when the model provides a correct answer in meaning, different wording can reduce the METEOR score. This reflects the natural flexibility of language and shows that measuring quality purely by word choice can be limiting, especially in art history, where subtle differences in phrasing can convey similar ideas.

Taken together, these results highlight several important points:

1. The model reliably produces answers that are both accurate and relevant, addressing factual as well as interpretative questions.

2. Metrics like BERTScore, which focus on meaning, better capture the quality of responses when multiple wordings are possible. Metrics like METEOR, focusing on exact wording, reveal how variable language use can be.

3. The difference between the metrics shows that evaluating open-ended questions in art history requires looking at both the meaning and the specific words used, since small changes in phrasing can influence interpretation.

To show the evaluation of the answers generated by the model, let's consider a representative example from the dataset. The question, gold answer, model response, and their metric evaluations are reported below.

- Question: Who are the main characters in Monet's painting?
- Gold Answer: His wife Camille Monet and their son Jean
- Model Answer: Camille Monet and her son Jean
- METEOR Score: 0.677
- BERTScore: Precision = 0.914, Recall = 0.849, F1 = 0.880

In this case, the model-generated answer captures the main information correctly, identifying the protagonists of the painting. The METEOR score of 0.677 reflects the partial lexical overlap: minor differences in wording, such as omitting *"His wife"* and *"their"*, reduce the score even though the semantic content is preserved. METEOR is sensitive to exact matches, stemming, and synonyms, so small deviations in phrasing can result in lower values.

BERTScore, in contrast, shows a high semantic similarity, with Precision = 0.914, Recall = 0.849, and F1 = 0.880. This metric evaluates contextual meaning rather than exact word matching, recognizing that the model answer correctly identifies Camille Monet and her son Jean. The comparison highlights a key distinction between the two metrics: while METEOR may penalize lexical differences, BERTScore better captures the semantic correctness of paraphrased or concise responses.

Consider the following example illustrating differences between METEOR and BERTScore evaluations:

- Question: What size is the painting?
- Gold Answer: 77 cm high by 53 cm wide
- Model Answer: 77 cm
- METEOR Score: 0.120
- BERTScore: Precision = 0.791, Recall = 0.638, F1 = 0.706

In this case, the model answer provides only a partial response, mentioning the height but omitting the width. METEOR reflects this lexical mismatch strongly, producing a low score of 0.120. This demonstrates METEOR's sensitivity to exact word overlap: missing elements drastically reduce the metric even if some of the answer is correct.

BERTScore, on the other hand, produces higher values (F1 = 0.706), indicating partial semantic alignment. The metric recognizes that part of the information, the height, is correctly captured, even though the answer is incomplete. This example highlights how BERTScore can provide a more nuanced view of model performance, especially in cases of partial correctness, whereas METEOR is more stringent on lexical matching.

Finally, another interesting case is given by the following example:

- Question: How is Venus depicted in the scene?
- Gold Answer: On a large shell emerging from the waves
- Model Answer: Standing on a large shell emerging from the waves
- METEOR Score: 0.963
- BERTScore: Precision = 0.899, Recall = 0.944, F1 = 0.921

In this instance, the model answer closely matches the gold standard, with the only difference being the inclusion of "standing". METEOR reflects this strong lexical overlap with a very high score of 0.963, while BERTScore also indicates excellent semantic alignment (F1 = 0.921). This example demonstrates a case where both metrics agree that the model has captured the essential content accurately.

Overall, these examples illustrate how both METEOR and BERTScore provide complementary insights: METEOR is sensitive to exact lexical matches, while BERTScore captures semantic similarity, highlighting cases where the model produces answers that are correct in meaning even if phrased differently from the gold standard.

5.7 Advancing the State of the Art in Digital Cultural Heritage

This research is situated within an interdisciplinary context that, in recent years, has witnessed growing interest in the application of digital technologies to cultural heritage. However, as highlighted by several studies, most projects have addressed the dimensions of visualization, artificial intelligence, and interaction as separate domains, thereby limiting the transformative potential of an integrated approach.

Early contributions in the field of 3D digitization (Stylianidis and Remondino, 2016) highlighted the effectiveness of techniques like photogrammetry and laser scanning for the accurate documentation of archaeological sites and monuments. While these approaches are essential for preservation, they have often remained static and descriptive, focusing on geometric fidelity rather than fostering interpretation or user interaction. In parallel, projects such as Virtual Rome (Dylla et al., 2008) have demonstrated the potential of virtual reality for immersive reconstruction of historical environments. However, the interaction offered was predominantly passive: users explored the virtual space as spectators, without opportunities for dialogue or personalized insight. Similarly, immersive storytelling experiences, as analyzed by Champion, 2021, have sought to emotionally engage audiences through digital narratives, but often rely on predefined paths, lacking the cognitive flexibility that only AI-driven systems can provide.

In recent years, several studies have begun to explore the integration of immersive technologies and artificial intelligence. Fabbri, Collina, Barzaghi, et al., 2023 examines how AI can enhance visitor engagement in museums through personalization strategies, such as adaptive content delivery and intelligent recommendation systems. Their work highlights the potential of chatbots and machine learning to tailor experiences based on user profiles and behavior. However, the study remains conceptual and does not present a fully integrated system that merges immersive VR environments with real-time conversational AI. The emphasis is on personalization through data analytics rather than embodied interaction within virtual spaces. Zhang, Taib, and Taib, 2025 presents a comprehensive framework that leverages digital twin technology, AI-driven modeling, and blockchain-based authentication to support cultural heritage conservation. Their approach is innovative in terms of ensuring data integrity, provenance, and transparency through NFT mechanisms. Nonetheless, the user experience is secondary to the infrastructural and archival goals. The framework does not incorporate immersive storytelling or interactive dialogue, which limits its applicability for public-facing educational or museum contexts. European projects already mentioned, such as INCEPTION (Commission, 2020) and V-MUST.NET (Commission, 2014), have highlighted the importance of accessibility and inclusive cultural engagement by introducing virtual environments designed for users with disabilities. However, the absence of intelligent conversational components limits the ability of these systems to dynamically adapt to the cognitive and cultural needs of users.

More recently, generative and multimodal AI systems have begun to enter the field. Experimental initiatives, such as the EU-funded AI4Culture project (Consortium, 2023), have started to integrate conversational agents and large language models into heritage contexts, enabling users to query collections or explore narratives dynamically. Other prototype systems, developed in academic or research settings, are exploring multimodal interaction in immersive environments. While

these approaches are promising, they often face the dual challenge of ensuring factual accuracy and maintaining transparency in source attribution, issues that are particularly sensitive in heritage communication, where interpretive responsibility and data provenance are paramount.

In this context, the present thesis offers an original and methodologically innovative contribution. It integrates advanced 3D digitization techniques with immersive environments developed in Unity3D, alongside the implementation of a question answering system based on MobileBERT. This represents a significant advancement over existing approaches. The system not only enables immersive navigation of digitally reconstructed cultural environments but also provides dialogic and context-aware interaction, allowing users to ask questions and receive relevant answers tailored to the visualized content. Furthermore, the integration of user experience tracking through xAPI and Learning Locker enables both qualitative and quantitative evaluation of interaction, generating valuable data to refine experience design and assess the effectiveness of digital cultural mediation. The pilot study conducted, which included evaluation metrics and question classification, demonstrates the feasibility and impact of the proposed system. While the literature has produced substantial contributions in each of the individual domains (3D reconstruction, immersive visualization, and AI-based interaction), this thesis distinguishes itself by integrating them into a coherent, operational system that has been empirically tested. It offers a replicable model for museums, archaeological sites, and educational environments. This approach not only enriches the experience of cultural heritage, but also opens new perspectives for mediation, accessibility, and experiential learning. By bridging the gap between technical innovation and interpretive engagement, the system proposed in this research responds to the growing demand for intelligent, inclusive, and emotionally resonant cultural experiences. It contributes to a shift from static presentation to dynamic dialogue, where users become active participants in the exploration of heritage, supported by AI-driven tools that adapt to their curiosity and cognitive needs.

Beyond its methodological contribution, the system developed in this research is designed with scalability and sustainability in mind. Its modular architecture allows for flexible deployment across diverse cultural contexts, from small local museums to large institutional archives. The technical pipeline, from 3D acquisition to immersive visualization and AI-driven interaction, can be adapted to different levels of infrastructure and expertise, making it accessible to institutions with limited resources. In terms of sustainability, particular attention has been paid to optimizing hardware requirements. The immersive environments developed in this research are currently optimized for Meta headsets, ensuring high performance and visual fidelity. However, the system architecture is designed to be hardware-agnostic and easily adaptable to other types of VR devices, including tethered and standalone headsets. This flexibility supports broader accessibility and facilitates deployment across institutions with varying technological infrastructures. Efforts are also underway to extend portability to CAVE systems, enabling multi-user, room-scale immersive experiences in research centers and museums equipped with projection-based setups. By supporting both high-end and low-cost configurations, including mobile-based viewers and desktop simulations, the system reduces the environmental and financial footprint of implementation. Moreover, the use of open-source development tools and lightweight AI models such as MobileBERT contributes to a more energy-efficient and transparent process. This aligns with current sustainability goals in digital heritage and supports long-term maintainability, scalability, and ethical deployment across diverse cultural contexts.

Chapter 6

Conclusion

The work presented in this thesis demonstrates how immersive virtual reality, combined with advanced 3D reconstruction techniques and natural language interfaces, can become a powerful tool for cultural heritage documentation, dissemination, and engagement. Starting from a structured digital pipeline, from 3D scanning and photogrammetry to mesh optimization, texturing, and deployment in VR environments, the research has shown that it is possible to create high-quality, explorable models of complex sites and artworks. These models are not only accurate from a scientific perspective but also accessible and engaging for a wide audience.

In parallel, the integration of a QA system based on large language models has expanded the immersive experience beyond static visualization. Instead of passively observing virtual reconstructions, users could interact dynamically with an intelligent guide, asking questions about artworks and receiving immediate answers. The evaluation of this system, based on metrics such as BERTScore and METEOR, highlighted both its potential and its limitations. While the model demonstrated a strong capacity to provide semantically accurate answers, even when phrasing diverged from the reference, it also revealed the challenges of dealing with ambiguous, incomplete, or overly generic queries.

A key finding of the pilot study concerns the role of context. The system relies on the content made available within its knowledge base, which means that when users posed questions unrelated to the presented material, the responses were often incorrect or misleading. This limitation became particularly evident in the first experimental condition, where participants were not given prior contextual information and therefore tended to formulate questions with little or no connection to the artworks. Conversely, in the second condition, where the avatar introduced each artwork before allowing users to ask questions, the interactions proved more effective. Users, having been provided with a conceptual framework, formulated more precise and relevant questions, which in turn improved the reliability and relevance of the system's responses. This highlights that, although the QA system is context-based, the human capacity to situate oneself within the right frame of reference remains fundamental for meaningful interaction.

However, it is important to acknowledge that the pilot study was conducted with a relatively small and homogeneous group of participants, primarily university students and researchers familiar with digital technologies. While this allowed for a focused technical validation, future evaluations will involve a broader and more diverse sample, including secondary school students, museum visitors, and cultural heritage professionals, to ensure more representative and inclusive feedback.

The analysis of the collected dataset further reinforced this observation. Users frequently asked similar questions in different ways, and even slight variations in wording could trigger different responses from the system. This phenomenon is

particularly significant in the domain of art history, where questions often oscillate between factual requests and interpretive exploration. On the one hand, factual queries (for example, author, date, materials) were generally handled well, yielding concise and accurate answers. On the other hand, interpretive or symbolic questions led to a greater variability of answers, reflecting the system's sensitivity to nuances of formulation and the inherently open-ended nature of cultural interpretation.

Despite these encouraging results, several limitations emerged that must be acknowledged. First, the reliance on speech-to-text systems introduced errors, especially in the presence of background noise or when participants spoke with varying accents, intonation, or pronunciation. Such transcription inaccuracies sometimes are propagated into the QA process, leading to irrelevant or incorrect answers. Second, the system struggled with overly generic questions, such as "Tell me about this painting," where the expected answer could vary widely in length, focus, and content. Without a clearly defined scope, the model tended to provide incomplete or inconsistent responses. Third, the dependency on the availability of contextual information remains a structural limitation: when the knowledge base does not contain the requested detail, the model cannot invent accurate content, and responses may default to vague or erroneous formulations. Finally, the real-time nature of interaction in immersive environments requires not only semantic accuracy but also responsiveness.

In conclusion, this research confirms that combining immersive visualization with natural language interaction offers a promising direction for cultural heritage dissemination. The proposed pipeline enables the creation of scientifically accurate and visually compelling virtual environments, while the QA system fosters engagement through dialogue rather than passive observation. In conclusion, this research confirms that combining immersive visualization with natural language interaction offers a promising direction for cultural heritage dissemination. The proposed pipeline enables the creation of scientifically accurate and visually compelling virtual environments, while the QA system fosters engagement through dialogue rather than passive observation.

Looking ahead, there are several directions in which this research could naturally evolve. One of the first concerns the integration of more advanced speech-to-text systems. During the pilot, transcription errors were one of the main sources of misunderstanding, particularly in noisy environments or with users speaking in different accents. More robust solutions, possibly fine-tuned on art-related terminology, would greatly improve the fluidity of interaction and the overall reliability of the system. Equally important is the way the system handles very generic or open-ended questions. Future developments could make the avatar more proactive, guiding users toward more specific queries or suggesting alternative angles of exploration when a question is too vague. This would make the interaction feel less like a static question–answer exchange and more like a real conversation with a knowledgeable guide. In addition, future studies will adopt more systematic evaluation methods to assess user learning outcomes. These may include pre- and post-experience questionnaires, task-based assessments, and spatial understanding tests. Such approaches will help quantify the educational impact of the system and provide more robust evidence of its effectiveness in cultural mediation.

Another promising line of development lies in the adoption of new generative AI models that are able to manage context in more flexible ways. Unlike the current BERT-based approach, which requires a predefined knowledge base, recent large language models can integrate contextual information dynamically, generating answers that combine stored knowledge with interpretive flexibility. Exploring these

models, whether fine-tuned on art-historical data or connected to curated knowledge graphs, could open the way to richer, more adaptive dialogues.

On the experiential side, future work should aim at greater platform versatility. While this study focused primarily on VR headsets, the same system could be extended to augmented reality applications. This would allow users to experience cultural content directly in situ, overlaying digital reconstructions or interpretive narratives onto physical spaces such as archaeological sites or museum galleries. AR has the potential to complement VR by offering a more seamless integration between the real and the virtual, enhancing accessibility while preserving the embodied dimension of cultural heritage. At the other end of the spectrum, more advanced immersive infrastructures such as CAVE environments could be used to test collaborative exploration, where groups of visitors interact simultaneously with the artworks and the avatar guide. This would extend the experience beyond individual immersion, opening up possibilities for collective education and discussion in museum settings.

Beyond its technical contributions, this research also engages with broader cultural and ethical questions about how digital technologies shape our relationship with heritage. The integration of AI and immersive visualization should not be viewed merely as a matter of innovation or efficiency, but as part of an ongoing reflection on how we perceive, interpret, and share the past. In this sense, digital heritage is not only a field of technological experimentation but a space of cultural mediation, a place where knowledge, emotion, and experience converge.

The motivation behind this work lies precisely in exploring how digital tools can foster accessibility and inclusion without reducing the complexity of cultural narratives. By grounding technological design in transparency, authenticity, and interpretive responsibility, this research aims to demonstrate that innovation and cultural care are not opposing forces, but complementary dimensions of the same mission: preserving the past while re-imagining how it is experienced and understood.

The future of this research lies in finding the right balance between technological progress and cultural accessibility. With more advanced models capable of understanding context, and with the possibility of extending the system to different immersive formats such as augmented reality or collaborative environments, these tools can go beyond simple digital reconstructions. They can become living spaces where people can not only observe cultural heritage, but also experience it, interpret it, and share it in more engaging and meaningful ways.

Bibliography

Anil, R. et al. (2024). "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context". In: *arXiv preprint arXiv:2403.05530*. URL: https://arxiv.org/abs/2403.05530.

Azuma, Ronald T (1997). "A survey of augmented reality". In: *Presence: teleoperators & virtual environments* 6.4, pp. 355–385.

Banerjee, Satanjeev and Alon Lavie (2005). "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments". In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72.

Barazzetti, Luigi, Fabio Remondino, Marco Scaioni, et al. (2010). "Automation in 3D reconstruction: Results on different kinds of close-range blocks". In: *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 38.Part 5, pp. 55–61.

Barbera, Roberto et al. (2022a). "A case study for the design and implementation of immersive experiences in support of sicilian cultural heritage". In: *International Conference on Image Analysis and Processing*. Springer, pp. 174–185.

Barbera, Roberto et al. (2022b). "A pipeline for the implementation of immersive experience in cultural heritage sites in sicily". In: *International Conference Florence Heri-Tech: the Future of Heritage Science and Technologies*. Springer, pp. 178–191.

Bekele, Mafkereseb Kassahun et al. (2018). "A survey of augmented, virtual, and mixed reality for cultural heritage". In: *Journal on Computing and Cultural Heritage (JOCCH)* 11.2, pp. 1–36.

Besl, Paul J and Neil D McKay (1992). "Method for registration of 3-D shapes". In: *Sensor fusion IV: control paradigms and data structures*. Vol. 1611. Spie, pp. 586–606.

Bongini, Pietro et al. (2020). "Visual question answering for cultural heritage". In: *IOP Conference Series: Materials Science and Engineering*. Vol. 949. 1. IOP Publishing, p. 012074.

Brown, Tom et al. (2020). "Language models are few-shot learners". In: *Advances in neural information processing systems* 33, pp. 1877–1901.

Buldu, Kadir Burak et al. (2025). "Cuify the xr: An open-source package to embed llm-powered conversational agents in xr". In: *2025 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*. IEEE, pp. 192–197.

Calise, Anna (2022). "Il museo digitale. Analisi di un dispositivo tra curatela, spazialità e partecipazione." In: *Transmedialità e crossmedialità. Nuove Prospettive*. Mimesis, pp. 167–196.

Caramiaux, Baptiste (2023). "AI with Museums and Cultural Heritage". In: *AI in Museums: Reflections, Perspectives and Applications*, pp. 117–130.

Carrozzino, Marcello and Massimo Bergamasco (2010). "Beyond virtual museums: Experiencing immersive virtual reality in real museums". In: *Journal of cultural heritage* 11.4, pp. 452–458.

Champion, Erik (2016). *Critical gaming: Interactive history and virtual heritage*. Routledge.

Champion, Erik and Erik Champion (2011). *Playing with the Past*. Springer.
Champion, Erik Malcolm (2021). *Virtual Heritage*. Ubiquity Press.
Chen, Danqi et al. (2017). "Reading wikipedia to answer open-domain questions". In: *arXiv preprint arXiv:1704.00051*.
Chiazzese, Giuseppe and Giosué Lo (2023). "Design of a pilot study to evaluate a Question Answering model based on BERT". In: *BOOKOF*, p. 122.
Chiazzese, Giuseppe et al. (2023). "Using xAPIs for monitoring behavioral lessons in augmented reality". In: *Perspectives on Learning Analytics for Maximizing Student Outcomes*. IGI Global, pp. 144–167.
Ch'ng, Eugene et al. (2022). *Visual Heritage: Digital Approaches in Heritage Science*. Springer.
Cipresso, Pietro et al. (2018). "The past, present, and future of virtual and augmented reality research: a network and cluster analysis of the literature". In: *Frontiers in psychology* 9, p. 2086.
Colla, Davide et al. (Sept. 2022). "Bringing Semantics into Historical Archives with Computer-aided Rich Metadata Generation". In: *J. Comput. Cult. Herit.* 15.3. ISSN: 1556-4673. DOI: 10.1145/3484398. URL: https://doi.org/10.1145/3484398.
Colucci Cante, Luigi et al. (2024). "Automated storytelling technologies for cultural heritage". In: *International Conference on Emerging Internet, Data & Web Technologies*. Springer, pp. 597–606.
Commission, European (2014). *V-MUST.NET – Virtual Museum Transnational Network*. Rete di eccellenza finanziata dall'UE (FP7, Grant Agreement n. 270404). Accesso il 14 aprile 2025. URL: https://cordis.europa.eu/project/id/270404.
— (2020). *INCEPTION – Inclusive Cultural Heritage in Europe through 3D Semantic Modelling*. Progetto finanziato da Horizon 2020, Grant Agreement n. 665220. Accesso il 14 aprile 2025. URL: https://cordis.europa.eu/project/id/665220.
Consortium, AI4Culture (2023). *AI4Culture: Artificial Intelligence for Cultural Heritage*. https://ai4culture.eu. Accessed: 2025-11-07.
Croce, Danilo, Alexandra Zelenanska, and Roberto Basili (2018). "Neural learning for question answering in italian". In: *International Conference of the Italian Association for Artificial Intelligence*. Springer, pp. 389–402.
Cruz-Neira, Carolina, Daniel J Sandin, and Thomas A DeFanti (2023). "Surround-screen projection-based virtual reality: the design and implementation of the CAVE". In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 51–58.
Devlin, Jacob et al. (2019). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186.
Di Franco, Paola Di Giuseppantonio et al. (2015). "3D printing and immersive visualization for improved perception of ancient artifacts". In: *Presence: Teleoperators and Virtual Environments* 24.3, pp. 243–264.
Di Salvo, Fabio and Mauro Lo Brutto (2014). "Full-waveform terrestrial laser scanning for extracting a high-resolution 3D topographic model: A case study on an area of archaeological significance". In: *European Journal of Remote Sensing* 47.1, pp. 307–327.
Drucker, Johanna (2013). "Performative Materiality and Theoretical Approaches to Interface." In: *DHQ: Digital Humanities Quarterly* 7.1.
Dylla, Kimberly et al. (2008). "Rome reborn 2.0: A case study of virtual city reconstruction using procedural modeling techniques". In: *Computer Graphics World* 16.6, pp. 62–66.

Economou, Maria and Laia Pujol Tost (2011). "Evaluating the use of virtual reality and multimedia applications for presenting the past". In: *Handbook of research on technologies and cultural heritage: Applications and environments*. IGI Global, pp. 223–239.

European Commission (2021). *Commission Recommendation (EU) 2021/1970 of 10 November 2021 on a common European data space for cultural heritage*. Official Journal of the European Union L 401/5 (12 Nov. 2021). CELEX 32021H1970.

Europeana (2008). *Europeana Collections*. Accesso il 14 aprile 2025. URL: `https://www.europeana.eu`.

Europeana Foundation (2020). *Europeana Strategy 2020–2025: Empowering Digital Change*. `https://op.europa.eu/en/publication-detail/-/publication/2174be9e-ac58-11ea-bb7a-01aa75ed71a1/`. Europeana Foundation, The Hague. ISBN 978-92-76-17398-4. DOI 10.2759/524581.

Fabbri, Francesca, Federica Collina, Sebastian Barzaghi, et al. (2023). "AI and chatbots as a storytelling tool to personalize the visitor experience. The Case of National Museum of Ravenna". In: *ExICE23*.

Falk, John H and Lynn D Dierking (2018). *Learning from museums*. Rowman & Littlefield.

Fanini, Bruno, Daniele Ferdani, and Emanuel Demetrescu (2021). "Temporal lensing: an interactive and scalable technique for Web3D/WebXR applications in cultural heritage". In: *Heritage* 4.2, pp. 710–724.

Farella, Mariella, Giuseppe Chiazzese, and Giosué Lo Bosco (2022). "Question Answering with BERT: designing a 3D virtual avatar for Cultural Heritage exploration". In: *2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON)*. IEEE, pp. 770–774.

Farella, Mariella et al. (2021). "Integrating xAPI in AR applications for positive behaviour intervention and support". In: *2021 International Conference on Advanced Learning Technologies (ICALT)*. IEEE, pp. 406–408.

Fassi, Francesco et al. (2013). "Comparison between laser scanning and automated 3d modelling techniques to reconstruct complex and extensive cultural heritage areas". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 40, pp. 73–80.

Forte, Maurizio and Gregorij Kurillo (2010). "Cyberarchaeology: Experimenting with teleimmersive archaeology". In: *2010 16th international conference on virtual systems and multimedia*. IEEE, pp. 155–162.

Gaia, Giuliano et al. (2019). "Engaging Museum Visitors with AI: The Case of Chatbots". In: *ResearchGate*. URL: `https://www.researchgate.net/publication/336020755_Engaging_Museum_Visitors_with_AI_The_Case_of_Chatbots`.

Gee, James Paul (2003). "What video games have to teach us about learning and literacy". In: *Computers in entertainment (CIE)* 1.1, pp. 20–20.

Grussenmeyer, Pierre et al. (2008). "Comparison methods of terrestrial laser scanning, photogrammetry and tacheometry data for recording of cultural heritage buildings". In: *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* 37.B5, pp. 213–218.

Guidazzoli, Antonella and Maria Chiara Liguori (2011). "Realtà virtuale e Beni culturali: una relazione in evoluzione vista attraverso i progetti sviluppati presso il Cineca". In: *Storia e Futuro* 25.

Guidi, Gabriele, J-A Beraldin, and Carlo Atzeni (2004). "High-accuracy 3D modeling of cultural heritage: the digitizing of Donatello's" Maddalena"". In: *IEEE Transactions on image processing* 13.3, pp. 370–380.

ICCROM (2021). *Sustaining Digital Heritage: Towards a Global Strategy*. https://www.iccrom.org/sites/default/files/2021-12/en_0_sustainingdigitalheritage-findingsreport_iccrom_2021.pdf. International Centre for the Study of the Preservation and Restoration of Cultural Property (ICCROM).

ICOMOS (2017). *Principles for the Conservation of Heritage Sites in the Digital Era*. https://www.icomos.org/. ICOMOS Charter for the Interpretation and Presentation of Cultural Heritage Sites.

Izacard, Gautier and Edouard Grave (Apr. 2021). "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty. Online: Association for Computational Linguistics, pp. 874–880. DOI: 10.18653/v1/2021.eacl-main.74. URL: https://aclanthology.org/2021.eacl-main.74/.

Ji, Ziwei et al. (2023). "Survey of hallucination in natural language generation". In: *ACM computing surveys* 55.12, pp. 1–38.

Kassahun, Bekele Mafkereseb et al. (2018). "A survey of augmented, virtual, and mixed reality for cultural heritage". In: *ACM Journal on Computing and Cultural Heritage* 11.2, pp. 7–1.

Kazhdan, Michael, Matthew Bolitho, and Hugues Hoppe (2006). "Poisson surface reconstruction". In: *Proceedings of the fourth Eurographics symposium on Geometry processing*. Vol. 7. 4.

Liu, Haotian et al. (2023). "Visual Instruction Tuning". In: *arXiv preprint arXiv:2304.08485*. URL: https://arxiv.org/abs/2304.08485.

Liu, Yinhan et al. (2019). "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692*.

Llamazares De Prado, Jose Enrique and Ana Rosa Arias Gago (2023). "Education and ICT in inclusive museums environments". In: *International journal of disability, Development and Education* 70.2, pp. 186–200.

Lo Bosco, G, M Farella, and G Di Piazza (2023). "Immersive Virtual Reality for Cultural Heritage Exploration". In: *Memorie della Società Astronomica Italiana Journal of the Italian Astronomical Society*, p. 51.

LocalProBook (2024). *DigiArt Project: Revolutionizing Cultural Heritage Through Digital Art and 3D Technology*. URL: https://www.localprobook.com/article/DigiArt.html.

Luhmann, Thomas et al. (2023). *Close-range photogrammetry and 3D imaging*. Walter de Gruyter GmbH & Co KG.

Manning, Christopher D (2008). *Introduction to information retrieval*. Syngress Publishing.

Mathias, Markus et al. (2012). "Automatic architectural style recognition". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 38, pp. 171–176.

Mazzetto, Silvia (2024). "Integrating emerging technologies with digital twins for heritage building conservation: an interdisciplinary approach with expert insights and bibliometric analysis". In: *Heritage* 7.11, pp. 6432–6479.

Mikolov, Tomas et al. (2013). "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems* 26.

Milgram, Paul and Fumio Kishino (1994). "A taxonomy of mixed reality visual displays". In: *IEICE TRANSACTIONS on Information and Systems* 77.12, pp. 1321–1329.

Min, Sewon et al. (2019). "A discrete hard EM approach for weakly supervised question answering". In: *arXiv preprint arXiv:1909.04849*.

Ministero della Cultura (2008). *CulturaItalia*. Accesso il 14 aprile 2025. URL: https://www.culturaitalia.it.

— (2022). *Piano Nazionale di Digitalizzazione del Patrimonio Culturale. Versione 1.0*. Accesso il 4 maggio 2025. URL: https://docs.italia.it/italia/icdp/icdp-pnd-docs/it/v1.0-giugno-2022.

Murphy, Maurice, Eugene McGovern, and Sara Pavia (2009). "Historic building information modelling (HBIM)". In: *Structural Survey* 27.4, pp. 311–327.

New Palmyra Project (2016). *The #NEWPALMYRA Project*. Progetto open source per la ricostruzione digitale di Palmira, fondato da Bassel Khartabil. URL: https://www.newpalmyra.org.

OpenAI (2024). "GPT-4 Technical Report". In: *arXiv preprint arXiv:2303.08774*. URL: https://arxiv.org/abs/2303.08774.

Papineni, Kishore et al. (2002). "Bleu: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318.

Parisi, Loreto, Simone Francia, and Paolo Magnani (2020). *UmBERTo: an Italian Language Model trained with Whole Word Masking*. https://github.com/musixmatchresearch/umberto.

Peng, Zhenzhong et al. (2023). "Kosmos-2: Grounding Multimodal Large Language Models to the World". In: *arXiv preprint arXiv:2306.14824*. URL: https://arxiv.org/abs/2306.14824.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning (2014). "Glove: Global vectors for word representation". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Pietroni, E and A Pagano (2016). "Un metodo integrato per valutare i Musei Virtuali e l'esperienza dei visitatori. Il caso del Museo Virtuale della Valle del Tevere". In: *Proceedings of the Workshop at Etruscan National Museum of Villa Giulia, Rome, Italy*. Vol. 15.

Polignano, Marco et al. (2019). "AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets". In: *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*. Vol. 2481. CEUR. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074851349&partnerID=40&md5=7abed946e06f76b3825ae5e294ffac14.

Qorib, Muhammad, Geonsik Moon, and Hwee Tou Ng (2024). "Are decoder-only language models better than encoder-only language models in understanding word meaning?" In: *Findings of the Association for Computational Linguistics ACL 2024*, pp. 16339–16347.

Radford, Alec et al. (2018). "Improving language understanding by generative pre-training". In.

Radford, Alec et al. (2023). "Robust speech recognition via large-scale weak supervision". In: *International conference on machine learning*. PMLR, pp. 28492–28518.

Raffel, Colin et al. (2020). "Exploring the limits of transfer learning with a unified text-to-text transformer". In: *Journal of machine learning research* 21.140, pp. 1–67.

Rajpurkar, Pranav et al. (2016). "Squad: 100,000+ questions for machine comprehension of text". In: *arXiv preprint arXiv:1606.05250*.

Remondino, Fabio and Sabry El-Hakim (2006). "Image-based 3D modelling: a review". In: *The photogrammetric record* 21.115, pp. 269–291.

Remondino, Fabio and Efstrations Stylianidis (2016). *3D recording, documentation and management of cultural heritage*. Vol. 2. Whittles Publishing Dunbeath, UK.

Remondino, Fabio et al. (2014). "State of the art in high density image matching". In: *The photogrammetric record* 29.146, pp. 144–166.

Risch, Julian et al. (2021). "Semantic answer similarity for evaluating question answering models". In: *arXiv preprint arXiv:2108.06130*.

Riva, Giuseppe, John Waterworth, Dianne Murray, et al. (2014). *Interacting with Presence: HCI and the Sense of Presence in Computer-mediated Environments*. De Gruyter Open Berlin, Germany:

Sanh, Victor et al. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter". In: *arXiv preprint arXiv:1910.01108*.

Scan4Reco (2018). *Scanning for Cultural Heritage. Horizon 2020 – Grant Agreement No. 665091*. Progetto attivo dal 2015 al 2018. Accesso il 14 aprile 2025. URL: http://www.scan4reco.eu.

Siliutina, Iryna et al. (2024). "Cultural preservation and digital heritage: challenges and opportunities". In: *Amazonia Investiga* 13.75, pp. 262–273.

Slater, Mel and Maria V Sanchez-Vives (2016). "Enhancing our lives with immersive virtual reality". In: *Frontiers in Robotics and AI* 3, p. 74.

Snavely, Noah, Steven M Seitz, and Richard Szeliski (2008). "Modeling the world from internet photo collections". In: *International journal of computer vision* 80.2, pp. 189–210.

Stylianidis, Efstratios and Fabio Remondino (2016). *3D recording, documentation and management of cultural heritage*. Whittles Publishing Dunbeath, UK.

Suh, Ayoung and Jane Prophet (2018). "The state of immersive technology research: A literature analysis". In: *Computers in Human behavior* 86, pp. 77–90.

Sylaiou, Stella et al. (2010). "Exploring the relationship between presence and enjoyment in a virtual museum". In: *International journal of human-computer studies* 68.5, pp. 243–253.

Tan, Wei Ren et al. (2017). "ArtGAN: Artwork synthesis with conditional categorical GANs". In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 3760–3764.

Tiribelli, Simona et al. (2024). "Ethics of Artificial Intelligence for Cultural Heritage: Opportunities and Challenges". In: *IEEE Transactions on Technology and Society*.

Touvron, Hugo et al. (2023). "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971*.

Trichopoulos, Georgios (2023). "Large language models for cultural heritage". In: *Proceedings of the 2nd International Conference of the ACM Greek SIGCHI Chapter*, pp. 1–5.

Tserklevych, Viktoriia et al. (2021). "Virtual Museum Space as the Innovative Tool for the Student Research Practice". In: *International Journal of Emerging Technologies in Learning (iJET)* 16, pp. 213–231. DOI: 10.3991/ijet.v16i14.22975. URL: https://online-journals.org/index.php/i-jet/article/view/22975.

Ugail, Hassan et al. (2023). "Deep transfer learning for visual analysis and attribution of paintings by Raphael". In: *Heritage Science* 11.1, p. 268.

UNESCO (2003). *Charter on the Preservation of Digital Heritage*. https://unesdoc.unesco.org/ark:/48223/pf0000179529. Adopted at the 32nd session of the General Conference, Paris.

Uz-Zaman, Md Mahmud, Stefan Schaffer, and Tatjana Scheffler (2021). "Factoid and open-ended question answering with BERT in the museum domain". In.

Vaswani, Ashish et al. (2023). *Attention Is All You Need*. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

Vincent, Matthew L et al. (2017). *Heritage and archaeology in the digital age: Acquisition, curation, and dissemination of spatial cultural heritage data*. Springer.

Westoby, Matthew J et al. (2012). "'Structure-from-Motion'photogrammetry: A low-cost, effective tool for geoscience applications". In: *Geomorphology* 179, pp. 300–314.

Xu, Ziwei, Sanjay Jain, and Mohan Kankanhalli (2024). "Hallucination is inevitable: An innate limitation of large language models". In: *arXiv preprint arXiv:2401.11817*.

Zhang, Tianyi et al. (2019). "Bertscore: Evaluating text generation with bert". In: *arXiv preprint arXiv:1904.09675*.

Zhang, Weiyi, Nooriati Taib, and Mariati Taib (2025). "Reimagining cultural heritage conservation through VR, metaverse, and digital twins: An AI and blockchain-based framework". In: *PloS one* 20.11, e0335943.