

Article

# A Comparative Study of Vision Language Models for Italian Cultural Heritage

Chiara Vitaloni <sup>1,\*</sup>, Dasara Shullani <sup>2</sup> and Daniele Baracchi <sup>2</sup><sup>1</sup> Department Cultures and Societies, University of Palermo, 90128 Palermo, Italy<sup>2</sup> Department of Information Engineering, University of Florence, 50139 Florence, Italy; dasara.shullani@unifi.it (D.S.); daniele.baracchi@unifi.it (D.B.)

\* Correspondence: chiara.vitaloni@unipa.it

**Abstract:** Human communication has long relied on visual media for interaction, and is facilitated by electronic devices that access visual data. Traditionally, this exchange was unidirectional, constrained to text-based queries. However, advancements in human-computer interaction have introduced technologies like reverse image search and large language models (LLMs), enabling both textual and visual queries. These innovations are particularly valuable in Cultural Heritage applications, such as connecting tourists with point-of-interest recognition systems during city visits. This paper investigates the use of various Vision Language Models (VLMs) for Cultural Heritage visual question answering, including Bing's search engine with GPT-4 and open models such as Qwen2-VL and Pixtral. Twenty Italian landmarks were selected for the study, including the Colosseum, Milan Cathedral, and Michelangelo's David. For each landmark, two images were chosen: one from Wikipedia and another from a scientific database or private collection. These images were input into each VLM with textual queries regarding their content. We studied the quality of the responses in terms of their completeness, assessing the impact of various levels of detail in the queries. Additionally, we explored the effect of language (English vs. Italian) on the models' ability to provide accurate answers. Our findings indicate that larger models, such as Qwen2-VL and Bing+ChatGPT-4, which are trained on multilingual datasets, perform better in both English and Italian. Iconic landmarks like the Colosseum and Florence's Duomo are easily recognized, and providing context (e.g., the city) improves identification accuracy. Surprisingly, the Wikimedia dataset did not perform as expected, with varying results across models. Open models like Qwen2-VL, which can run on consumer workstations, showed performance similar to larger models. While the algorithms demonstrated strong results, they also generated occasional hallucinated responses, highlighting the need for ongoing refinement of AI systems for Cultural Heritage applications.



Academic Editors: Nicola Masini, Paolo Rosati, Marco Raoul Marini and Augusto Palombini

Received: 24 December 2024

Revised: 10 February 2025

Accepted: 27 February 2025

Published: 2 March 2025

**Citation:** Vitaloni, C.; Shullani, D.; Baracchi, D. A Comparative Study of Vision Language Models for Italian Cultural Heritage. *Heritage* **2025**, *8*, 95. <https://doi.org/10.3390/heritage8030095>

**Keywords:** visual question answering; cultural heritage; artificial intelligence; ChatGPT; human-centered approaches

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Cultural Heritage, as defined by the UNESCO during the Convention concerning the Protection of the World Cultural and Natural Heritage [1], includes monuments, ensembles, and sites of “outstanding universal value from a historical, artistic, or scientific perspective” (Art. 1). Italy's vast Cultural Heritage encompasses art cities, museums, archaeological and landscape sites, works of art, and elements of ethnological and anthropological heritage. Consequently, Cultural Heritage is recognized as one of the fundamental principles in the

Italian Constitution (Art. 9), which highlights the significance of the nation's cultural and landscape heritage and underscores the importance of intergenerational dialogue. Our initial approach was driven by curiosity and interdisciplinarity, bridging the humanities and computer science. Today, the use of artificial intelligence in a mainstream and accessible manner is increasingly finding applications in the field of Cultural Heritage. We therefore explored how Vision–Language Large Models (VLLMs) can address the challenge of recognizing specific objects or monuments, particularly in a country like Italy, renowned for its extraordinarily rich and diverse Cultural Heritage. The findings of this study could offer valuable insights into optimizing the use of VLLMs to enhance both the accessibility and overall experience of cultural sites. Additionally, they may help determine whether the language used to interact with these models influences their performance and whether elements that are distinctive to the human eye are similarly recognized as defining features by artificial intelligence. The aim of this work is to explore how artificial intelligence “sees and perceives” the Cultural Heritage of a place, and how it interprets it. By investigating these processes, we can gain a better understanding of how AI might reshape our interaction with heritage, blending preservation with innovative reinterpretation. The questions we posed in this research are as follows: Which model performs best in subject and city identification? To what extent does providing the city where the picture was taken enhance accuracy? Does the choice of language, between Italian and English, significantly affect the model's performance? And, finally, which subjects are most and least easily recognized by the models?

This study focuses on the evaluation of 10 different Vision–Language Large Models (VLLMs) and their corresponding visual question answering (VQA) systems. Through an experimental approach, the research aims to analyze the responses generated by these AI systems with two primary objectives: first, to assess the level of detail and accuracy achieved by the various algorithms; and, second, to explore how the broader field of Cultural Heritage identification is positioned within this technological landscape. This dual focus seeks to provide valuable insights into the current capabilities and limitations of VLLMs in applications related to Cultural Heritage, offering a foundation for future advancements in this interdisciplinary domain.

The paper is structured in the following way: Section 2 provides an overview of the literature review. Section 3 describes the proposed materials and methodology, which includes dataset curation, the selection of Visual Large Language Models, and the analysis protocol. Section 4 presents the results and discussions, addressing the methodological and experimental questions that motivated the study. It examines the accuracy of city and subject identification across the models, focusing on the identification process for cities and subjects, changes in evaluation based on the second question (which already provides the city of origin), linguistic performance between Italian and English, and issues such as hallucinations, errors, and peculiarities in the responses. Section 5 highlights some of the key limitations of the proposed method. Finally, in Section 6, a summary of the study is provided along with recommendations for future research and development.

## 2. Literature Review

Since the 1970s, the term “cultural asset” has been commonly used in Italy as a synonym for “Cultural Heritage” [2]. This growing interest is evidenced by the large number of projects funded by the European Commission in recent decades, particularly under Horizon 2020 and Horizon Europe. Many of these projects have focused not only on the digitization of Cultural Heritage [3], but also on its protection and conservation [4], restoration, research, and promotion [5–7], as well as improving accessibility to heritage assets [8], across various cultural institutions such as museums, libraries, archives, monuments, and galleries [9].

In the field of Cultural Heritage, there is a growing push to apply artificial intelligence (AI) models and algorithms in various ways. The AI technology available today is making remarkable strides month by month, with experts continually updating their knowledge to keep pace. Meanwhile, Cultural Heritage itself, and the perception around it, remains largely static, centered on preservation. This stability is its strength; heritage represents and preserves the identity of a community, a city, or even a nation. In general terms, heritage embodies the culture of a people, bearing witness to their history. The Colosseum, for instance, stands as an emblem of Italy, just as the Eiffel Tower represents France, and the Great Wall represents China. While these examples resonate globally, they also risk reinforcing stereotypes, potentially limiting a deeper understanding of a territory and its people.

Nowadays, interdisciplinary collaboration has become essential for multidisciplinary research groups. Humanities fields (such as archaeology, history, and archival studies) must work in synergy with computer science, statistics, and communication to expand the scope and applications of research. Dialogue across these disciplines is the true foundation for addressing the new challenges that technology is presenting to cultural studies.

In this context, AI applications have taken diverse directions, from 4D reconstructions of the history of the Old Continent that offer immersive time-travel experiences [10] to the recognition of ceramic fragments [11]. AI is also making inroads into augmented reality [12] and personalized museum tours, although human validation remains a crucial final step [13].

The Artificial Intelligence for Multimedia Information Retrieval research group at Consiglio Nazionale delle Ricerche has pioneered a variety of approaches to image analysis, deep learning, and scalable data structures for similarity algorithms. The prototypes developed by Vadicamo et al. [14] have been applied to tasks such as recognizing works of art or historical buildings, accessing augmented reality information, and generating automatic descriptions for digital materials that lack adequate annotation. Moreover, in the field of Cultural Heritage, the majority of methods have been centered around building AI models to classify and identify art pieces, but also generate art descriptions [15]. Bai et al. [16] proposed a ResNet-based model for art painting captioning where multiple artistic perspectives are provided in the artwork description. Zhang et al. [17] introduced a small generative model for emotion recognition and emotion-grounded explanations for artworks. Balauca et al. [18] showed how to fine-tune CLIP, a well-known foundation model, to provide structured data on museum artworks. Rachabatuni et al. [19] proposed a system that integrates detailed contextual descriptions with visual question answering to enhance both the understanding and accuracy of answers related to artworks.

Recently, the advent of the transformer architecture [20] marked a significant milestone in natural language processing, enabling unprecedented advances in modeling human languages. This innovation paved the way for the development of Generative Pretrained Transformers (GPTs) [21–23], which further revolutionized the field by leveraging large-scale pretraining on diverse datasets followed by task-specific fine-tuning. This led to the release and subsequent widespread adoption of multiple AI-driven conversational agents, such as ChatGPT, conventionally called large language models (LLMs). While early models primarily focused on text-based applications, recent research has extended this paradigm to multimodal large language models (MLLMs), which integrate information from multiple modalities such as text and images. By processing data from diverse sources, MLLMs enable richer interactions.

Given that much of the existing research has only focused on the description and identification of objects within museum settings, we sought to explore how these new

ways of interacting with computers could be applied to the Cultural Heritage of our cities. Specifically, we envisioned a virtual guide capable of assisting tourists during their explorations by leveraging the capabilities of MLLMs. To this end, we investigated freely accessible multimodal models within this context. To the best of our knowledge, this is the first study to address this topic.

### 3. Materials and Methods

To assess the capabilities of various AI models, we designed an experiment in which a series of structured questions related to a collection of images was posed to several conversational AI systems. The primary objective was to gain a deeper understanding of how computational models perceive and interpret Cultural Heritage landmarks. For this purpose, a dataset comprising 40 images was curated, representing iconic and widely recognized landmarks from different regions of Italy. The responses generated by the selected models were then categorized and analyzed to identify patterns and extract relevant statistical insights. The following subsections provide detailed information on the dataset construction process, the selected language models, the methodology for formulating the questions, and the approach employed for analyzing the results.

#### 3.1. Dataset Curation

The selection of subjects for recognition was conducted arbitrarily, focusing on elements integral to Italy's Cultural Heritage that are internationally renowned and easily recognizable as quintessential Italian landmarks. Efforts were made, where possible, to ensure a broad geographical representation across the Italian peninsula. Statues, commemorative plaques, and monuments represent a specific type of material heritage, designed to embody and celebrate collective memories, symbols, and values. Public monuments have an ostensive dimension, attracting tourists and becoming recognizable landmarks by which cities are identified and with which they themselves identify [24]. The selection comprised ten prominent national landmarks (the so-called Colosseum—the Flavian Amphitheater—in Rome, the Verona Arena, the Milan Cathedral, the Mole Antonelliana in Turin, Piazza del Plebiscito in Naples, Teatro Massimo and the Cathedral in Palermo, the Asinelli and Garisenda Towers in Bologna, the Temple of Neptune in Paestum, and the historic center of San Gimignano) and ten culturally significant landmarks within Florence (Santa Maria del Fiore, the Ponte Vecchio, the Baptistery of San Giovanni, the exterior of the Uffizi Gallery, the replica of Michelangelo's David in Piazza della Signoria, the Neptune Fountain, commonly referred to as the *Biancone*, the Church of the Santissima Annunziata, the historical sundial located on a building along the Ponte Vecchio, and the Monument to Bettino Ricasoli), for a total of 20 subjects. Among the selected sites, four are listed as UNESCO World Heritage Sites: the Colosseum in Rome (1980), the Temple of Neptune in Paestum (1988), the historic center of San Gimignano (1990), and Florence (1982), recognized for their outstanding universal value, highlighting Italy's cultural and historical richness<sup>1</sup>. Notably, Florence's Historic Center has been inscribed on the UNESCO World Heritage List since 17 December 1982 as a representation of the Renaissance. UNESCO's selection criteria define the site as "a unique artistic achievement, a masterpiece resulting from over six centuries of continuous creation" exerting "a predominant influence on architectural development and monumental arts" [25]. The remaining subjects are defined as cultural assets considered of particular interest, comprising monuments, archaeological sites, and areas of scenic value, according to article 10 of the Italian Code of Cultural Heritage and Landscape (Decreto Legislativo 22 gennaio 2004, n. 42, "Codice dei beni culturali e del paesaggio").

Images in which the subject is the “monument in pose” were selected [26]. These images were carefully chosen according to established “good photography” principles, ensuring that the cultural asset is sharply focused, centrally placed within the frame, and prominently featured in the foreground [27]. Given the experimental nature of this study, preference was given to subjects that are easily identifiable and widely recognized, at least to the human eye, with some specific exceptions requiring detailed recognition (e.g., the sundial on the Ponte Vecchio and the Monument to Bettino Ricasoli).

Since our goal was to assess the accuracy and recognition capabilities of neural-network-based algorithms, it was crucial to use images that were not part of their training data. If an algorithm has previously encountered an image, it may recognize it solely because it matches a known example. In contrast, accuracy measured on images that the algorithm has never seen provides a more reliable indication of its performance in real-world scenarios. Overall, the created dataset included 40 images (depicted in Figure 1) obtained from sources encompassing both publicly accessible and previously unpublished material. The pictures are categorized as follows:

- 20 photographs from Wikimedia: These images were sourced from the widely recognized Wikimedia platform, which provides indexed and searchable content. They served as the reference standard for both national and Florentine landmarks due to their high accessibility and consistent quality;
- 10 photographs from FloreView [28]: This subset originates from a specialized dataset dedicated to Florence’s historic center. While available to the scientific community, these images are not indexed by search engines, offering a unique and controlled dataset for research purposes;
- 10 photographs classified as “Others”: This category includes personal photographs captured exclusively for private use. These images have not been indexed or published online, ensuring their novelty and exclusivity within the dataset.



**Figure 1.** *Cont.*

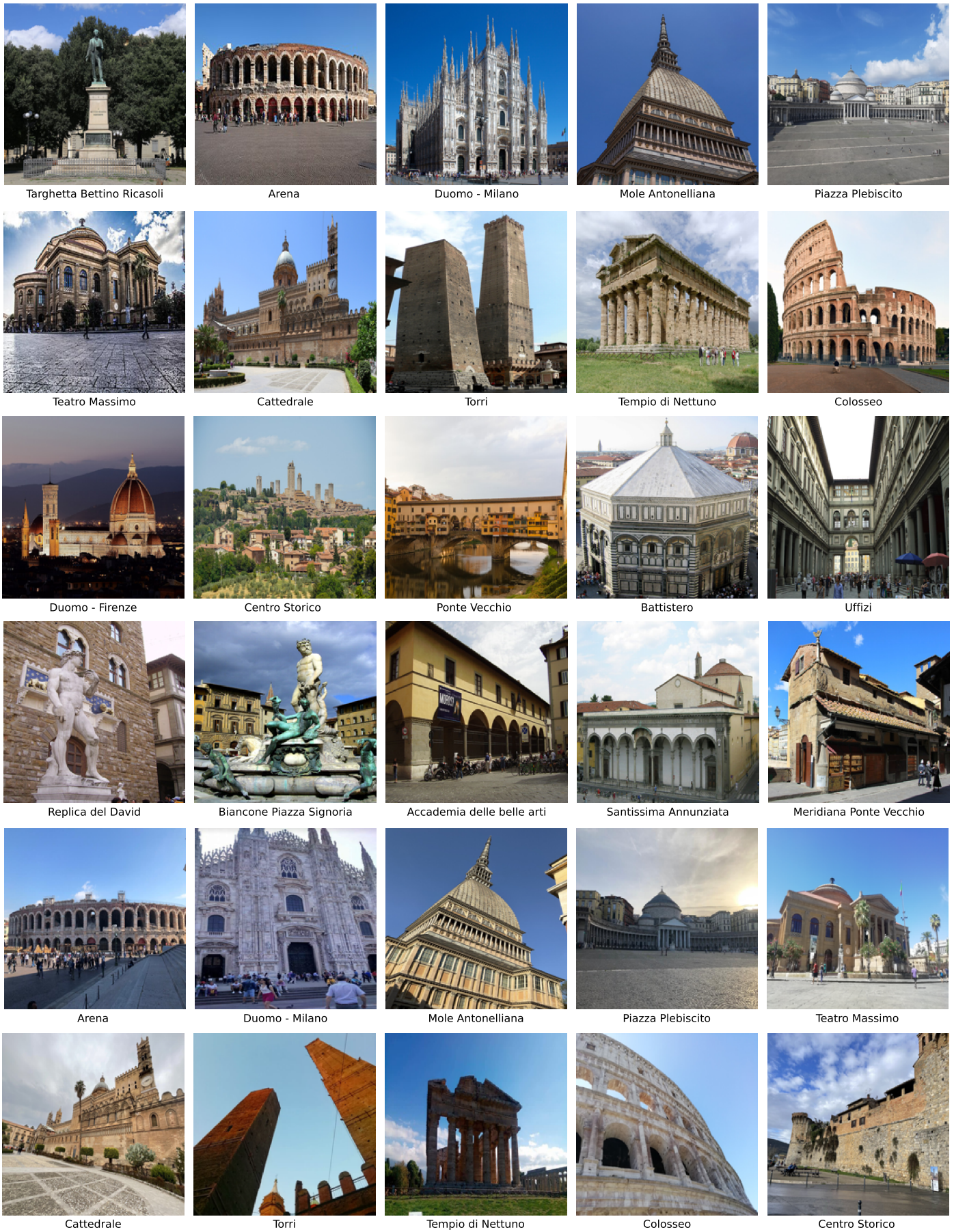


Figure 1. The considered images from FloreView [28], Wikimedia, and Other sources.

### 3.2. Visual Large Language Models Selection

In this study, we evaluated the capabilities of multimodal algorithms by selecting 10 publicly available Visual Large Language Models (VLLMs)—a subset of MLLMs designed to process both images and text. To ensure our analysis reflects the current state of the field, we focused exclusively on models released in 2023 or 2024. We included in this study Qwen2-VL [29] by Alibaba Cloud, CogVLM2 [30,31] by Tsinghua University, Deepseek-vl [32] by DeepSeek AI, Phi-3.5 [33] by Microsoft, InternVL2 [34,35] by the General Vision Team of Shanghai AI Laboratory, Pixtral [36] by Mistral AI, Llava 1.6 [37,38] by the University of Wisconsin–Madison and Microsoft, Molmo [39] by the Allen Institute for AI, SmolVLM [40] by Hugging Face, and Bing+ChatGPT-4 by Microsoft and OpenAI.

Of the selected models, all but one are freely available for download, enabling users to run them locally on their own computers. The exception is Bing+ChatGPT-4, which is accessible only online but remains free to use. Many of the downloadable models offer multiple variants, differing in their training objectives. When possible, we selected the variants specifically fine-tuned for conversational tasks. Additionally, many of the models are available with various levels of quantization—a technique that reduces memory requirements by compressing the model, albeit with some potential trade-offs in performance. In such cases, we opted for the lowest level of quantization that was compatible with our hardware. Finally, each model is defined by a specific number of parameters, which can be conceptualized as the “neurons” that constitute the artificial neural network. Generally, a larger number of parameters corresponds to enhanced reasoning capabilities and a broader knowledge base, albeit at the cost of increased hardware requirements. The majority of the selected models contain between 7 and 8 billion parameters, a commonly adopted size that facilitates seamless deployment on consumer-grade hardware. Pixtral and CogVLM2, however, are larger models, and the latter necessitated more aggressive quantization to ensure compatibility with our hardware. At the opposite end of the spectrum, Phi-3.5 and SmolVLM are smaller language models specifically designed for use on lower-end hardware. Details of the chosen Visual Large Language Models are reported in Table 1. For each model, we provide the number of parameters, the selected variant, and whether the parameter was quantized to reduce its memory footprint. We do not report details regarding Bing+ChatGPT4 as they are not publicly available.

**Table 1.** List of VLLMs used in the experimental setting. Cutoff dates were collected from the corresponding paper (†), the Hugging Face page (‡), or the official release (★). Bing+ChatGPT-4’s question marks indicate unavailable information.

Model	Variant	Parameters	Quantization	Cutoff Date
Qwen2-VL [29]	Instruct	7 B	–	June 2023 †
CogVLM2 [30,31]	llama3-Chat	19 B	int4	March 2023 ‡
Deepseek-vl [32]	base	7 B	–	May 2023 †
Phi-3.5 [33]	Vision-Instruct	4.2 B	–	March 2024 ‡
Pixtral [36]	–	12 B	float16	September 2024 ★
InternVL2 [34,35]	–	8 B	–	July 2024 ★
Llava 1.6 [37,38]	mistral-7b-hf	7 B	–	December 2023 †
Molmo [39]	O-0924	7 B	bfloat16	October 2024 ★
SmolVLM [40]	–	1.7 B	bfloat16	December 2024 ‡
Bing+ChatGPT-4	?	?	?	?

### 3.3. Evaluation Protocol

The evaluation procedure for each model involved presenting it with a series of questions related to the images in the dataset. For each image, a new conversation was initiated with the model, during which two questions were asked sequentially. The first

question required the model to identify the subject of the photo without being provided with any additional context, while the second question asked the model to identify the subject, this time specifying the city where the photo was taken. Both questions were posed within the same conversation for all models except Molmo, which, due to a limitation in the available implementation, required the creation of two separate conversations (one for each question). This approach enabled us to investigate language-related differences and assess the impact of the models' predominantly English-language training on their performance when responding in a different language.

The questions (reported in Table 2) were designed to minimize bias related to the subjects depicted in the images while providing only essential context to the visual language model. Although many images featured Italian landmarks, we opted for the term *subject* to avoid influencing the VLM's interpretation. In the second question, we included the name of the city to guide the VLM's response while keeping the provided information minimal.

**Table 2.** Questions utilized during the model evaluation process. For questions marked as "City", the placeholder CITY was replaced with the actual name of the city where the photo was taken.

Language	City	Question
ITA	–	Qual è il soggetto rappresentato in questa immagine?
ITA	✓	Sei a CITY, qual è il soggetto rappresentato in questa immagine?
ENG	–	Which is the subject captured in this picture?
ENG	✓	You are in CITY, which is the subject captured in this picture?

### 3.4. Results Analysis Protocol

The responses from all models were collected and analyzed separately. For each answer, human evaluators assigned three labels based on three key criteria: whether an indication of the city was provided ("City" criterion), whether the subject depicted in the picture was identified ("Subject" criterion), and whether a description of the picture was provided ("Description" criterion). When evaluating an answer for the "City" criterion, one of four possible labels was assigned to the answer: "OK" if the city was correctly identified, "NO" if no indication of the city was provided, "PAR" if the city was not explicitly mentioned but the region or country was correctly recognized, or "ERR" if an incorrect city was identified. As the city was explicitly mentioned in the second question for each picture, this criterion was only evaluated for answers to the first question. The "Subject" criterion evaluated the accuracy of the identified cultural asset, and responses were labeled with one of three possible values: "OK" if the subject was correctly identified, "NO" if no indication of the subject was provided, or "ERR" if the identified subject was incorrect. When evaluating the "Description" criterion, responses were labeled with "OK" if the answer contained a description consistent with the provided image, irrespective of its level of detail or length, "NO" if no description was provided, or "ERR" if the response contained an inconsistent or largely inaccurate description.

For instance, InternVL2 answered the first question on a picture depicting the Colosseum with the following: *The subject captured in this picture is the Colosseum.* While the subject of the picture was correctly identified, the VLM failed to mention the location where the picture was taken or provide a description. As a result, the image was scored as "NO" for the "City" criterion, "OK" for the "Subject" criterion, and "NO" for the "Description" criterion. In another case CogVLM2 provided the following answer for the second question regarding an image depicting a statue of Bettino Ricasoli: *Yes, the statue depicted in the image is the famous 'David' by the sculptor Michelangelo, located in Florence, Italy.* In this case, the model did not provide a description of the image and erroneously identified Bettino



Ricasoli as Michelangelo’s David. Therefore, we labeled this entry as “ERR” for the “Subject” criterion and “NO” for the “Description” criterion.

We report an example of response labeling for one of the images depicting the Colosseum in Table 3.

**Table 3.** An example of the answers to the first question and the corresponding evaluation for each tested VLM. The image used depicts the Colosseum in Rome, and the conversation was conducted in English. We highlighted in *italic* the occurrences of the subject and the city in each response.

Algorithm	Answer 1	City	Subject	Description
SmolVLM [40]	The <i>Colosseum</i> in <i>Rome</i> .	OK	OK	NO
Qwen2-VL [29]	The subject captured in this picture is the <i>Colosseum</i> , an ancient amphitheater located in <i>Rome</i> , Italy. The Colosseum is one of the most iconic and well-preserved structures of ancient Roman architecture.	OK	OK	OK
InternVL2 [35]	The subject captured in this picture is the <i>Colosseum</i> .	NO	OK	NO
Llava 1.6 [37,38]	The subject captured in this picture is the <i>Colosseum</i> , an iconic landmark located in <i>Rome</i> , Italy. It’s a large, oval amphitheater that was used for gladiatorial contests and other public spectacles during the Roman Empire.	OK	OK	OK
Pixtral [36]	The image depicts the <i>Colosseum</i> , an iconic ancient amphitheater located in <i>Rome</i> , Italy. The Colosseum is known for its distinctive elliptical shape and multiple levels of arches. The structure is made of stone and concrete, showcasing the architectural ingenuity of the Roman Empire. The Colosseum was used for gladiatorial games, public spectacles, and other events, and it remains one of the most famous landmarks in the world. The clear blue sky in the background highlights the grandeur and historical significance of this ancient structure.	OK	OK	OK
Phi-3.5 [33]	The <i>Colosseum</i> in <i>Rome</i>	OK	OK	NO
Bing+ChatGPT-4	The subject captured in the picture you sent is the <i>Colosseum</i> in <i>Rome</i> , Italy. The Colosseum is an ancient amphitheater, built of concrete and stone.	OK	OK	OK
Molmo [39]	The subject captured in this picture is the <i>Colosseum</i> in <i>Rome</i> , Italy. This iconic ancient amphitheater dominates the frame, showcasing its impressive architecture with multiple levels of arched windows and columns. The image focuses on the upper portion of the Colosseum, highlighting its grandeur and historical significance against a clear blue sky.	OK	OK	OK
Deepseek-vl [32]	The subject captured in this picture is the <i>Colosseum</i> , a famous landmark in <i>Rome</i> , Italy.	OK	OK	NO
CogVLM2 [31]	The subject captured in this picture is the <i>Colosseum</i> , an iconic ancient Roman amphitheater located in <i>Rome</i> , Italy.	OK	OK	OK

## 4. Results and Discussion

We analyzed the results with the aim of addressing the following experimental questions:

- Q1 Which model performs best in subject and city identification?
- Q2 To what extent does providing the city where the picture was taken enhance accuracy?
- Q3 Does the choice of language, between Italian and English, significantly affect the model’s performance?
- Q4 Which subjects are most and least easily recognized by the models?

In the following subsections, we will analyze the models’ responses to address these questions in detail. Additionally, we will provide a qualitative assessment of the errors made by the model.

### 4.1. City and Subject Identification

In order to evaluate the accuracy of the models in identifying the city and the subject, we considered the classification of responses according to the “City” and “Subject” criteria. For each model, the responses provided for the images belonging to the three subsets defined in Section 3.1 were analyzed separately, and the percentage of responses labeled as “OK” was calculated for each subset. Additionally, the accuracy in identifying the subject

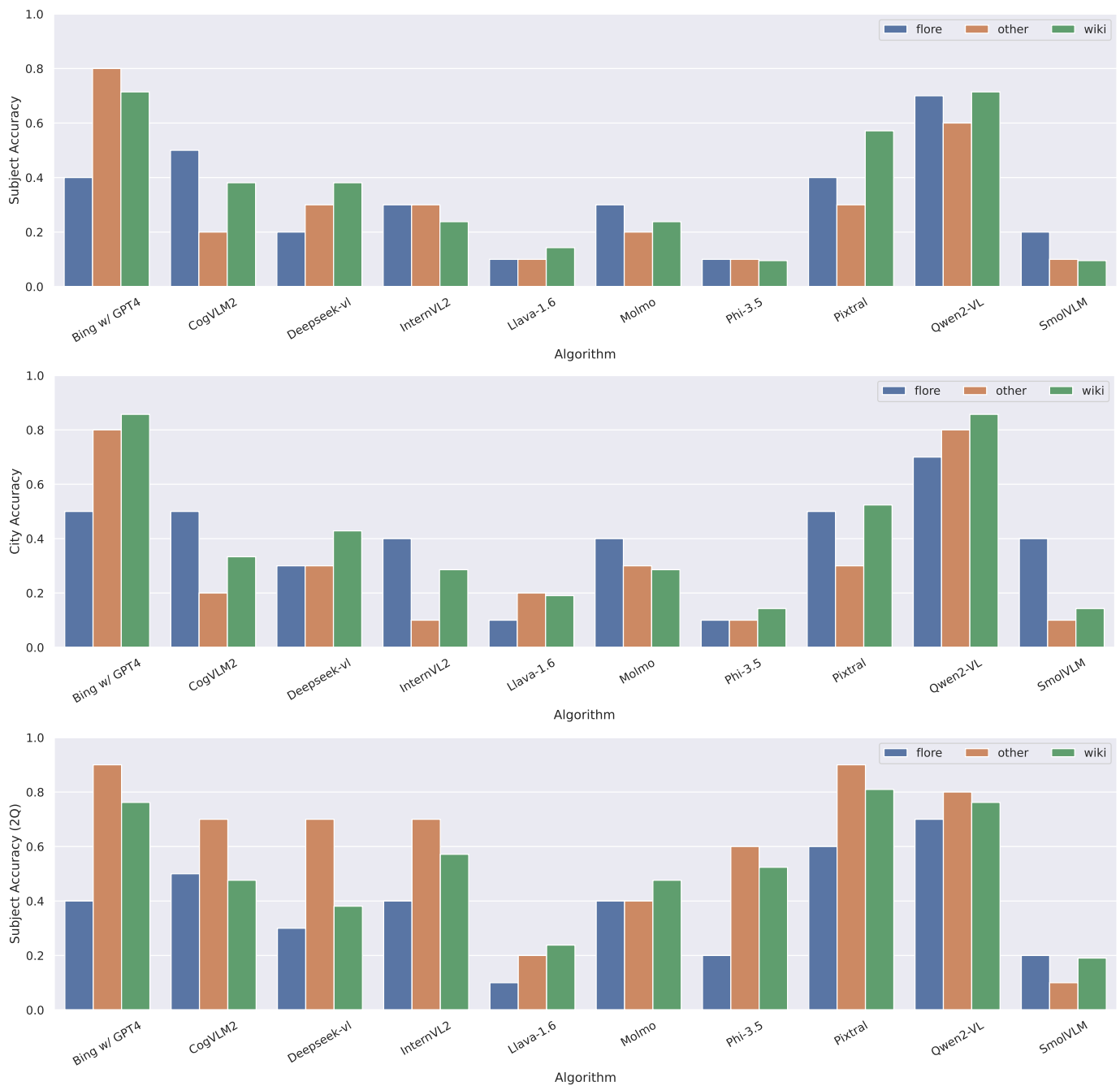
was assessed when information about the city was provided. In this case, all images for which the model was able to identify the subject on either the first or second attempt were counted as correct.

In Figure 2, we present the accuracy of identification in English for the correct detection of the subject and city in the first question and the subject in the second question. The results highlight the performance of each algorithm across the three available datasets. Overall, the best-performing algorithms were Bing+ChatGPT-4, achieving an accuracy of 0.8 in identifying the subject in the *Other* dataset, and Qwen2-VL, which reached 0.7 on both the *Flore* and *Wiki* datasets. Identifying the city proved to be easier, with both models achieving an accuracy of 0.85 on the *Wiki* dataset. To evaluate whether the differences in the models' responses were statistically significant, we conducted a McNemar's test [41] with Holm–Bonferroni correction [42], comparing the results obtained for all the images by each pair of models. The results indicate that Qwen2-VL and Bing+ChatGPT-4 significantly outperformed all other models, except for CogVLM2 and Pixtral, in subject identification in both English and Italian interactions. Similar results were observed for city identification; however, in this case, neither Qwen2-VL nor Bing with ChatGPT-4 achieved a statistically significant improvement over CogVLM2. While the answers to the first question did not indicate a dataset that was consistently easier to analyze, the second question (where the city in which the subject was captured was provided to the model) showed that the *Other* dataset yielded the best overall performance. Exceptions to this trend were Llava-1.6, Molmo, and SmolVLM. In general, the results demonstrate that providing additional contextual information improves the algorithms' ability to identify the subject. Notably, Llava-1.6, Phi 3.5, and Molmo had improved performance in the second question, while Pixtral exhibited the greatest improvement, achieving an accuracy of 0.9 in identifying subjects in the *Other* dataset. A McNemar's test performed on the results of each model before and after incorporating the city into the context revealed that Pixtral and Molmo demonstrated a statistically significant improvement with this additional information in both English and Italian interactions. In contrast, InternVL2, Phi-3.5, and CogVLM2 exhibited significant improvements only in English interactions, while Llava-1.6 showed a significant improvement exclusively in Italian interactions.

In Figure 3, we present the accuracy in identifying the subject and city in Italian for the first question and the subject for the second. Qwen2-VL and Pixtral performed best in Italian. When the city was included in the question, Bing+ChatGPT-4 also achieved an accuracy of 0.9 on the *Other* dataset. Qwen2-VL emerged as the top-performing algorithm in Italian, achieving an accuracy of 0.8 in the first question and 0.9 in the second. Pixtral showed stronger performance in Italian compared to English, identifying the subject in the *Other* dataset with an accuracy of 0.7 in Italian versus 0.3 in English. Notably, six out of the ten VLMs achieved an accuracy below 0.3 in the first question and 0.5 in the second. Pixtral and Qwen2-VL, as the latest large language models, likely benefited from training on multilingual datasets, distinguishing them from their predecessors and other algorithms.

#### 4.2. Impact of Providing Additional Context

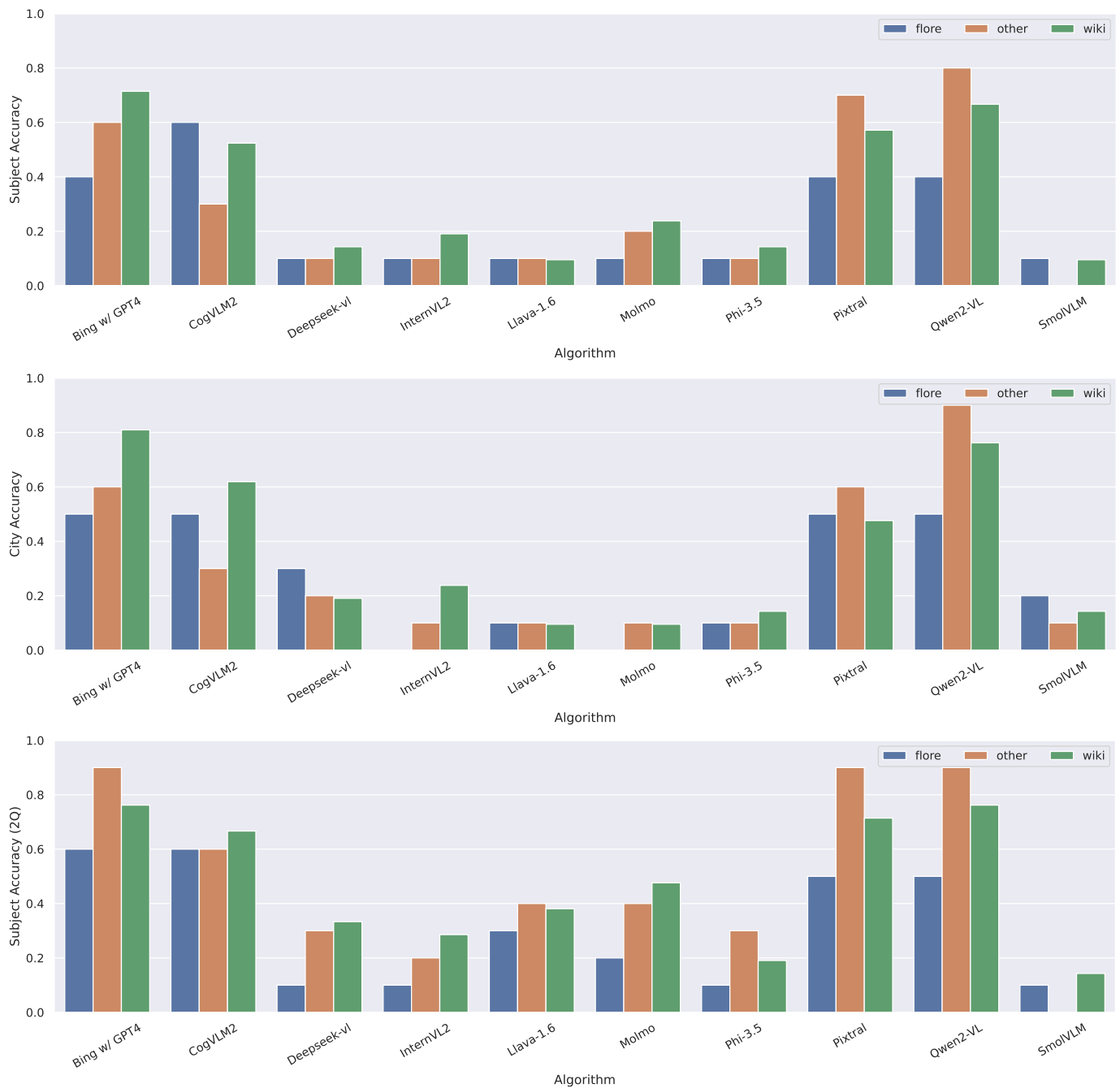
To assess the effect of providing additional context to the model, we analyzed how misclassified responses (“NO” or “ERR”) were impacted when supplementary information about the city was included. More specifically, we measured the increase or decrease in “NO” and “ERR” answers and the possible increase in “OK” answers, which could not decrease given that the additional context was only provided when the model failed to identify the subject on its first try.



**Figure 2.** Accuracy in identifying the city and subject in English. 2Q refers to the answers given to the second question.

In Figure 4, we present the analysis conducted on the models' second response—specifically, the one where we provided the city in which the image was taken. The first vertical plot displays the results in English, while the second shows the results in Italian.

Overall, providing the city tended to improve the detection of the depicted subject in both languages. Notably, this effect was particularly pronounced in English, where models like InternVL2, Qwen2-VL, and SmolVLM often converted hallucinated or unidentified subjects into correct ones. On the other hand, models like Phi-3.5, CogVLM2, and Pixtral tended to shift between correct detections and hallucinated subjects when the city was provided.



**Figure 3.** Accuracy in identifying the city and subject in Italian. 2Q refers to the answers given to the second question.

While InternVL2 was one of the best-performing models in English, this was not the case for Italian, where it showed the highest number of incorrect responses. For most models, knowing the city led to an increase in both correct and incorrect answers. Notably, Bing+ChatGPT-4 stood out as the most consistent across both languages, exhibiting an increase in correct answers in Italian while maintaining the same pattern of incorrect and no-subject responses as observed in English.



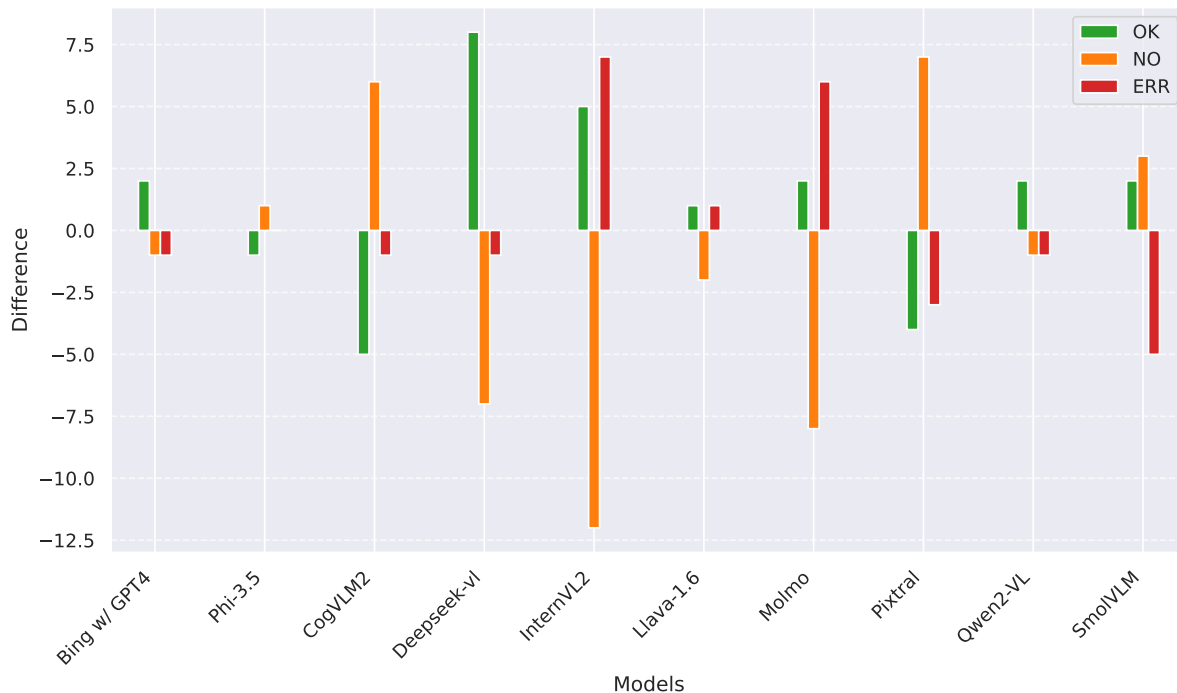
**Figure 4.** Impact of including the city in the second question on overall performance. The first plot reports results in English, while the second one reports those in Italian. Positive values indicate an increase when the additional context is provided, while negative values indicate a decrease.

#### 4.3. Language Performance

To evaluate the impact of choosing between English and Italian, we examined how the number of correct and incorrect responses for the “Subject” criterion varied across models. In this context, the counts for “OK”, “NO”, and “ERR” could increase or decrease. This variation arose because switching from Italian to English could either enhance or reduce the models’ accuracy.

In Figure 5, we exhibit the potential improvements in labeling accuracy for each model when the conversation was conducted in English compared to Italian. Among the best-performing models in our analysis (Qwen2-VL and Bing+ChatGPT-4), only Phi-3.5 and

Llava-1.6 showed minimal sensitivity to language changes. Deepseek-VL performed better in English, while CogVLM2 demonstrated higher accuracy in Italian. Molmo typically failed to provide information about the subject when responding in Italian, but, when switched to English, it exhibited a higher tendency toward hallucinations. Similarly, InternVL2 often provided little to no information in Italian, while, in English, its responses alternated between correct answers and numerous incorrect ones.

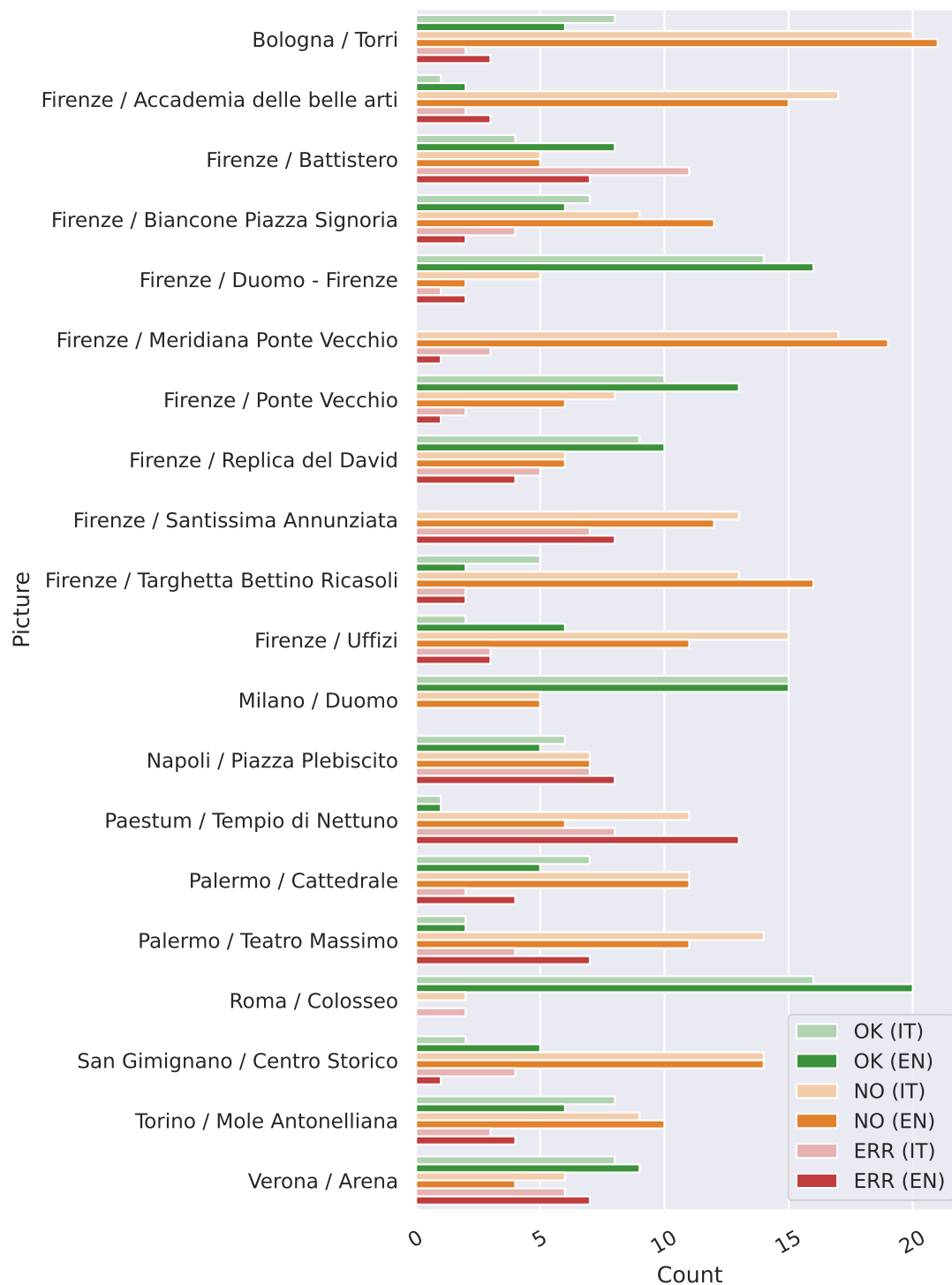


**Figure 5.** Accuracy improvement in the subject detection when using English instead of Italian. Positive values indicate an increase when the conversation uses the English language, while negative values indicate an increase when the conversation uses the Italian language.

#### 4.4. Subjects Most and Least Recognized

Finally, to evaluate the relative difficulty of identifying the provided images, we analyzed the number of models that produced responses classified as “OK”, “NO”, or “ERR” for each subject. This analysis was conducted for conversations in both Italian and English.

In Figure 6, we present the accuracy of identifying a specific subject, where light colors represent the analysis for Italian and bold colors represent the analysis for English. For the evaluated VLMs, the Colosseum was the easiest subject to identify, with all models successfully recognizing it in English. The three most recognizable landmarks in the dataset were likely its most iconic: the Colosseum, Florence’s Duomo, and Milan’s Duomo. Conversely, the most challenging subjects to recognize were Bologna’s towers, the sundial atop a building on the Ponte Vecchio, and the statue of Bettino Ricasoli. This is unsurprising, as the latter two are not widely recognized as symbols of Florence. Additionally, San Gimignano’s historic center also proved difficult to identify, likely because the selected image was too generic and resembled many small walled villages in Italy. The subjects most frequently misidentified included the Neptune Temple in Paestum, the Santissima Annunziata church, and Naples’ Piazza del Plebiscito. These misclassifications are understandable, as the Neptune Temple can easily be mistaken for a generic Greek temple, while the Santissima Annunziata church could be confused with a generic colonnade typical of Mediterranean architecture.



**Figure 6.** Accuracy in identifying each subject across all the analyzed models. Light colors refer to responses in Italian, while bold colors to those in English.

#### 4.5. Hallucinations, Errors, and Peculiarities in Responses

In the “ERR” label type, the algorithms provided peculiar responses. It seems that the trend is to invent a plausible answer, based on references that the algorithm itself recognizes as more popular. Below are some noteworthy examples from both the Italian and English responses. Following the *Wiki* and *Others* datasets, the notable responses of attention concern the following:

The Mole Antonelliana in Turin was often referred to as the Eiffel Tower of Paris (Llava 1.6—ITA), *il campanile del Duomo di Milano* (Molmo—ENG), the “Teatro alla Scala” in Milan

(InternVL2—ITA), the city of Domodossola (InternVL2—ITA), or, extending overseas, *the Basilica of Our Lady Aparecida, a famous Roman Catholic church located in Aparecida, Brazil* (Llava 1.6—ITA).

The Arena of Verona was frequently confused with the Colosseum (Llava 1.6—ITA/ENG, Molmo—ITA/ENG, InternVL2—ENG, SmolVLM—ENG), the Amphitheater of Parma (InternVL2—ITA), or the Cathedral of Turin (InternVL2—ITA).

The Piazza del Plebiscito in Naples was frequently classified as the Pantheon in Rome (SmolVLM—ENG, Phi3\_ITA), the Piazza Navona (InternVL2—ENG), the Piazza del Campidoglio (Llava 1.6—ITA), the Piazza del Popolo (Molmo ITA/ENG) in Rome, St. Peter’s Basilica in the Vatican (Molmo—ENG), the Piazza San Marco in Venice (Llava 1.6—ENG), and the Church of Santa Maria del Fiore in Florence (Deepseek—ENG).

The Temple of Neptune in Paestum was often confused with other temples in the Paestum Archaeological Park, such as the Temple of Hera (Phi-3.5—ENG, Qwen2VL—ENG) or the Temple of Zeus (Qwen2VL—ITA). It was also still confused with other monuments: the Pantheon in Rome (InternVL2—ENG), the Temple of Concordia in Agrigento (SmolVLM—ENG, InternVL2—ENG, Qwen2VL—ITA), the Temple of Segesta (Molmo—ITA, Deepseek—ITA, Pixtral—ITA), the Roman Temple of Capo Colonna (Molmo—ITA), the Temple of Jupiter, *also known as the Jupiter Temple, located in Baalbek, Lebanon* (Llava 1.6ITA/ENG), the Temple of Augustus in Pula, Croatia (Pixtral—ENG), and the most suggestive Parthenon in Athens (Phi 3.5—ENG).

The Teatro Massimo in Palermo was often confused with the nearest “Teatro Massimo Bellini”, which exists but is another theater located in Catania (Qwen2VL—ITA/ENG, Bing+ChatGPT-4—ITA), or with the Palazzo Normanni in Palermo (Llava 1.6—ITA) and the Palazzo dei Congressi in Naples (Molmo—ITA). It was also confused with international locations: Cibeles Square, a public square in Madrid, Spain (InternVL2—ENG), the National Museum of Beirut (Deepseek—ENG), the Monument to the Flag in Montevideo, Uruguay (Molmo—ENG), and the Metropolitan Cathedral of San Juan Bautista, located in San Juan, Puerto Rico (Molmo—ENG).

The Cathedral of Palermo was often referred to as the Chapel of St. Mark in the city of Amalfi (SmolVLM—ENG), the Cathedral of Cagliari (smolvlm\_ita), the Episcopal Palace in Seville, Spain (Molmo—ENG, Molmo—ITA), or the Episcopal Palace in Seville, Spain (InternVL2—ENG).

Regarding the Cultural Heritage present in the Florentine area and included in the ‘Wiki’ and ‘Flore’ datasets, the errors and hallucinations affected were the following:

The replica of the David located in Piazza della Signoria in Florence—the original marble statue by Michelangelo Buonarroti is preserved at the Galleria dell’Accademia—was considered “OK” in the responses, even though it was not emphasized that the subject photographed in Piazza della Signoria is the replica and not the original. Only one algorithm, Bing+ChatGPT-4—ITA, emphasized that it is a replica. Moreover, the David was recognized as another historically significant figure, such as the Statue of King Vittorio Emanuele II (Bing+ChatGPT-4—ENG), or the statue depicting the Greek god Apollo (Llava 1.6—ITA/ENG). It was also associated with more imaginative figures like Orso II, described as *Orso Benevolo. Orso II is a historical figure linked to the Other Republic of Genoa* (InternVL2—ITA), or Damiano di Gela (InternVL2—ITA).

The figure of Bettino Ricasoli, a well-known 19th-century Florentine political figure, was often confused with other historical Italian political figures, such as Giuseppe Garibaldi (Bing+ChatGPT-4—ENG, Deepseek—ITA, Bing+ChatGPT-4—ITA) or Betto Bocci (InternVL2—ENG).

The Fountain of Neptune was mistaken for the Fontana della Barcaccia, a famous Baroque fountain located in the Piazza di Spagna in Rome (Molmo—ENG) or the David



(InternVL2—ENG), the “Fiume di Acqua Santa” fountain (Molmo—ITA), or the fountain in the Piazza Navona (Molmo—ITA).

The Accademia degli Uffizi was recognized as the University of Siena (Pixtral - ENG) or the Faculty of Architecture in Venice (InternVL2—ENG).

The Church of the Santissima Annunziata in Florence was described as the Colosseum (smolvlm\_ita) or the Borghese Gallery Museum in Rome (InternVL2—ITA), the Palace of the Popes (Molmo—ENG), the Basilica of St. Alexander the Martyr in Genoa (InternVL2—ENG), the Basilica of San Lorenzo in Florence (Bing+ChatGPT-4—ENG, Qwen2VL—ENG), the Basilica of St. Francis of Assisi (Pixtral—ITA), or the Arch of Augustus, a triumphal arch located in Rimini (Pixtral—ITA). Incorrect interpretations of the inscription on the facade were also provided.

The sundial on the Ponte Vecchio was described as a building from the Sassi of Matera (Molmo—ENG), and, since a pigeon had landed on the sundial in the photo, the image was interpreted as follows: *The building appears to be made of stone and plaster and has a weathered look. A metal weather vane shaped like a rooster is mounted on the roof of the building* (Bing+ChatGPT-4—ENG).

## 5. Limitations

The methodology and data used, although appropriate for the study’s objectives, have some limitations that should be highlighted to provide a clearer understanding of the experimental context. In this section, we discuss two key aspects: potential training bias in the datasets and the evaluation methodology for the generated responses.

The dataset used to evaluate the VQA models was composed of three components: Wikimedia images, FloreView images, and Other images. Among these, the “Other” dataset consisted of the authors’ personal images, which were specifically chosen because they have never been shared on the Internet. This ensured that the models’ performance on this portion of the dataset was free from any training bias. In contrast, it was not possible to rule out potential training data overlap for the Wikimedia and FloreView datasets. Unfortunately, for most VQA models, it is unclear which specific images were used during their training process, with the exception of Molmo<sup>2</sup>, for which this information is available. Wikimedia images are particularly likely to introduce some degree of training bias, given their widespread use in publicly available datasets. On the other hand, the FloreView dataset was only released in October 2023. Therefore, it is certain that models such as Qwen2-VL, GogVLM2, Deepseek-vl, and Llava 1.6 did not use these images in their training, as their training cutoff dates precede FloreView’s publication, as reported in Table 1.

The evaluation of responses generated by generative models was conducted manually by the authors, making the interpretation inherently subjective. While this approach is suitable for small-scale evaluation, extending it to a larger set of models and images will require automated validation techniques (e.g., BLEU [43]). However, the automated evaluation of generated text remains an ongoing challenge for the scientific community.

## 6. Conclusions

In conclusion, this study presented a novel approach to understanding how artificial intelligence interprets and “perceives” the Cultural Heritage of a location, focusing on the responses generated by ten different open Vision–Language Large Models (VLLMs) and their corresponding visual question answering systems. Through a rigorous experimental framework, we evaluated not only the accuracy and level of detail of the responses provided by these algorithms, but also examined the broader implications of Cultural Heritage identification within the expanding technological landscape. To the best of our knowledge, this is the first study to explore the application of freely accessible multimodal

large language models to Cultural Heritage. By evaluating these models through a practical application, this study addresses a critical scientific gap, specifically assessing their performance in responding to direct inquiries that a tourist might realistically make in real-world scenarios. Results show that the easiest subjects to recognize are the primary landmarks in our dataset, such as the Colosseum, Florence's Duomo, and Milan's Duomo. We report that providing context for each image, such as the city in which the subject was captured, enhances the accuracy of subject identification. Interestingly, contrary to our initial intuition, the Wikimedia dataset—comprising publicly available images that might have been used during the training of our VLMs—was not the best-performing dataset. Instead, performance varied across datasets, with each algorithm excelling differently. It is worth noting that open models like Qwen2-VL, which can run on consumer workstations, achieved the same level of identification performance as larger, closed models like Bing+ChatGPT-4. Our evaluation, conducted in both English and Italian, highlights that Italian still requires significant improvements. However, in larger models such as Qwen2-VL and Bing+ChatGPT-4, which are trained on multilingual datasets, this performance gap is less pronounced. Moreover, despite the relative effectiveness of the evaluated algorithms, they often generated hallucinated responses, leading to arbitrary or sometimes fabricated answers, such as the incorrect references to historical figures linked to the statues of Michelangelo's David or Bettino Ricasoli. These responses not only involved similar elements (such as the apparent resemblance to Greek temples) or those found within the same city/region (as seen with Sicilian monuments), but also referenced more widely recognized contexts, such as the Italian capital, Rome, or extended to an international level. The models tended to avoid providing "blank" responses and, in fact, often offered insights that were not immediately apparent, such as the sundial rooster on the Ponte Vecchio. The findings offer important insights into the potential of artificial intelligence to contribute to the advancement of Cultural Heritage evaluation, suggesting new avenues for the application of AI in this rapidly developing field. One promising direction is the development of virtual assistants for Cultural Heritage engagement, which could not only enhance museum and tourism experiences but also provide new opportunities for increasing accessibility and fostering educational approaches to heritage preservation. Finally, potential future directions for our study include expanding the proposed dataset to incorporate a broader range of images featuring international landmarks. This would enhance the dataset's scope, offering a more comprehensive and global perspective on the functioning of visual question answering systems.

**Author Contributions:** Conceptualization, C.V., D.S. and D.B.; methodology, C.V., D.S. and D.B.; software, D.S. and D.B.; investigation, D.S.; resources, C.V., D.S. and D.B.; data curation, C.V., D.S. and D.B.; writing—original draft preparation, C.V.; writing—review and editing, C.V., D.S. and D.B.; visualization, D.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The raw data supporting the conclusions of this article are available at <https://github.com/IAPP-Group/VLM-Heritage> (accessed on 23 December 2024).

**Acknowledgments:** The author gratefully acknowledges the support of the prize in memory of Luca Restelli. The authors would like to thank Chiara Albisani and Francesca Nizzi for the valuable support during labeling.

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

## Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large language model
VQA	Visual Question Answering
VLLM	Visual Large Language Model
Bing+ChatGPT-4	Bing's search engine with GPT-4

## Notes

- <sup>1</sup> <https://whc.unesco.org/en/list/> (accessed on 9 February 2025).
- <sup>2</sup> <https://huggingface.co/collections/allenai/pixmo-674746ea613028006285687b> (accessed on 9 February 2025).

## References

1. Francioni, F.; Lenzerini, F. *The 1972 World Heritage Convention: A Commentary*; Oxford University Press: Oxford, UK, 2023.
2. Tosco, C. I beni culturali. In *Storia, Tutela e Valorizzazione*; Il Mulino: Bologna, Italy, 2014.
3. Hayhoe, S.; Carrizosa, H.G.; Rix, J.; Sheehy, K.; Seale, J. Accessible resources for cultural heritage ecosystems (arches): Initial observations from the fieldwork. In Proceedings of the Educational Research Association of Singapore (ERAS) Asia-Pacific Educational Research Association (APER) International Conference 2018: Joy of Learning in a Complex World, Singapore, 12–14 November 2018.
4. De Groote, I. D8. 2 Dissemination Plan (DIGIART-The Internet of Historical Things and Building New 3D Cultural Worlds). 2015. Available online: <http://researchonline.ljmu.ac.uk/id/eprint/3023/1/8%20%20Dissemination%20Plan%201%205.pdf> (accessed on 9 February 2025)
5. Frangakis, N. Social Platform for Heritage Awareness and Participation. In Proceedings of the EuroVR Conference 2016, Athens, Greece, 22–24 November 2016.
6. Sala, T.M.; Bruzzo, M. *I-Media-Cities. Innovative e-Environment for Research on Cities and the Media*; Edicions Universitat Barcelona: Barcelona, Spain, 2019.
7. Katifori, A.; Roussou, M.; Perry, S.; Drettakis, G.; Vizcay, S.; Philip, J. The EMOTIVE Project-Emotive Virtual Cultural Experiences through Personalized Storytelling. In Proceedings of the Cira@EuroMed, Nicosia, Cyprus, 29 October–3 November 2018; pp. 11–20.
8. Lim, V.; Frangakis, N.; Tanco, L.M.; Picinali, L. PLUGGY: A pluggable social platform for cultural heritage awareness and participation. In Proceedings of the Advances in Digital Cultural Heritage: International Workshop, Funchal, Portugal, 28 June 2017; Revised Selected Papers; Springer: Berlin/Heidelberg, Germany, 2018; pp. 117–129.
9. Tóth, Z. (Ed.) *Heritage at Risk: EU Research and Innovation for a More Resilient Cultural Heritage*; Working paper; European Commission, Publications Office: Luxemburg, 2018.
10. Maier, A.; Fernández, G.; Kestemont, M.; Fornes, A.; Eskofier, B.; Vallet, B.; van Noort, E.; Vitali, F.; Albertin, F.; Niebling, F.; et al. Time Machine: Big Data of the Past for the Future of Europe Deliverable D2.1 Science and Technology (Pillar 1) Roadmap—Draft. 2019. Available online: <https://pure.knaw.nl/portal/en/publications/time-machine-big-data-of-the-past-for-the-future-of-europe-delive> (accessed on 9 February 2025).
11. Anichini, F.; Dershowitz, N.; Dubbini, N.; Gattiglia, G.; Itkin, B.; Wolf, L. The automatic recognition of ceramics from only one photo: The ArchAIDE app. *J. Archaeol. Sci. Rep.* **2021**, *36*, 102788. [CrossRef]
12. Spallone, R.; Palma, V. Intelligenza artificiale e realtà aumentata per la condivisione del patrimonio culturale. *Boll. Della Soc. Ital. Fotogramm. Topogr.* **2020**, *1*, 19–26.
13. Trichopoulos, G.; Konstantakis, M.; Caridakis, G.; Katifori, A.; Koukouli, M. Crafting a Museum Guide Using ChatGPT4. *Big Data Cogn. Comput.* **2023**, *7*, 148. [CrossRef]
14. Vadicamo, L.; Amato, G.; Bolettieri, P.; Falchi, F.; Gennaro, C.; Rabitti, F. Intelligenza artificiale, retrieval e beni culturali. In Proceedings of the Ital-IA 2019, Rome, Italy, 18–19 March 2019.
15. Ishmam, M.F.; Shovon, M.S.H.; Mridha, M.F.; Dey, N. From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. *Inf. Fusion* **2024**, *106*, 102270. [CrossRef]
16. Bai, Z.; Nakashima, Y.; Garcia, N. Explain me the painting: Multi-topic knowledgeable art description generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 5422–5432.
17. Zhang, J.; Zheng, L.; Wang, M.; Guo, D. Training a small emotional vision language model for visual art comprehension. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 397–413.

18. Balauca, A.A.; Paudel, D.P.; Toutanova, K.; Van Gool, L. Taming CLIP for Fine-Grained and Structured Visual Understanding of Museum Exhibits. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 377–394.
19. Rachabatuni, P.K.; Principi, F.; Mazzanti, P.; Bertini, M. Context-aware chatbot using MLLMs for Cultural Heritage. In Proceedings of the 15th ACM Multimedia Systems Conference, Bari, Italy, 15–18 April 2024; pp. 459–463.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: New York, NY, USA, 2017; Volume 30.
21. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. OpenAI 2018. Available online: [https://cdn.openai.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf) (accessed on 9 February 2025).
22. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. OpenAI 2019. Available online: [https://cdn.openai.com/better-language-models/language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf) (accessed on 9 February 2025).
23. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165.
24. Martellozzo, N. Quando cadono le statue: Memorie contestate e counter-heritage nelle proteste di Black Lives Matter. *Dialoghi Mediterr.* **2020**, *45*, 1–10.
25. di Firenze, C. The Management Plan of the Historic Centre of Florence. Firenze: Ufficio Centro Storico UNESCO del Comune di Firenze. 2016. Available online: <http://www.firenzepatrimoniomondiale.it/wp-content/uploads/2015/12/Piano-gestione-enweb1.pdf> (accessed on 9 February 2025).
26. Dondero, M.G. Scenari del sé e monumenti in posa nella fotografia turistica. In *Espressione e Contenuto: Rivista dell'Associazione Italiana di Studi Semiotici*; Associazione Italiana di Studi Semiotici: Palermo, Italy, 2005.
27. Verhoeven, G. Basics of photography for cultural heritage imaging. In *3D Recording, Documentation and Management of Cultural Heritage*; Whittles Publishing: Dunbeath, UK, 2016; pp. 127–251.
28. Baracchi, D.; Shullani, D.; Iuliani, M.; Piva, A. FloreView: An image and video dataset for forensic analysis. *IEEE Access* **2023**, *11*, 109267–109282. [[CrossRef](#)]
29. Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv* **2024**, arXiv:2409.12191.
30. Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; et al. CogVLM: Visual Expert for Pretrained Language Models. *arXiv* **2023**, arXiv:2311.03079.
31. Hong, W.; Wang, W.; Ding, M.; Yu, W.; Lv, Q.; Wang, Y.; Cheng, Y.; Huang, S.; Ji, J.; Xue, Z.; et al. CogVLM2: Visual Language Models for Image and Video Understanding. *arXiv* **2024**, arXiv:2408.16500.
32. Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; et al. DeepSeek-VL: Towards Real-World Vision-Language Understanding. *arXiv* **2024**, arXiv:2403.05525.
33. Abidin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A.A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv* **2024**, arXiv:2404.14219.
34. Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv* **2023**, arXiv:2312.14238.
35. Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. *arXiv* **2024**, arXiv:2404.16821. [[CrossRef](#)]
36. Agrawal, P.; Antoniak, S.; Hanna, E.B.; Bout, B.; Chaplot, D.; Chudnovsky, J.; Costa, D.; Monicault, B.D.; Garg, S.; Gervet, T.; et al. Pixtral 12B. *arXiv* **2024**, arXiv:2410.07073.
37. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 34892–34916.
38. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved Baselines with Visual Instruction Tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024.
39. Deitke, M.; Clark, C.; Lee, S.; Tripathi, R.; Yang, Y.; Park, J.S.; Salehi, M.; Muennighoff, N.; Lo, K.; Soldaini, L.; et al. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Multimodal Models. *arXiv* **2024**, arXiv:2409.17146.
40. Allal, L.B.; Lozhkov, A.; Bakouch, E.; Blázquez, G.M.; Tunstall, L.; Piqueres, A.; Marafioti, A.; Zakka, C.; von Werra, L.; Wolf, T. SmoLLM2—With Great Data, Comes Great Performance. 2024. Available online: <https://arxiv.org/pdf/2502.02737> (accessed on 9 February 2025).
41. McNemar, Q. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika* **1947**, *12*, 153–157. [[CrossRef](#)] [[PubMed](#)]

42. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
43. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; pp. 311–318.

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