2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Cultural Heritage Assistant: A Lightweight Retrieval Augmented Generation Method Enhanced Vision-Language Model for Cultural Heritage

Shiyu Wang[1], Huanda Lu[2], Haibiao Yao[3], Zhiyu Wu[4], Chen Hu[5], Xiangjie Xie[5], and Xin Yu[1(✉)]

[1] College of Computer Science and Technology, Zhejiang University, Hangzhou, China
yuxin@zju.edu.cn
[2] School of Computer and Data Engineering, NingboTech University, Ningbo, China
huandalu@nbt.edu.cn
[3] Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, China
yaohb@mail.ustc.edu.cn
[4] School of Computer Science, Fudan University, Shanghai, China
wuzy24@m.fudan.edu.cn
[5] Hemudu Site Museum, Ningbo, China

Abstract. External knowledge makes the Vision-Language Model(VLM) more versatile. However, Traditional methods often fail to address the nuanced challenges of cultural heritage, mainly when dealing with unfamiliar or complex artifact-related queries. This limitation is evident in Vision-Language Models, which struggle to generate responses without exposure to domain knowledge. Frequent retraining to accommodate new artifacts or knowledge domains is computationally expensive and impractical. To overcome these limitations, we propose Cultural Heritage Assistant, a lightweight Retrieval-Augmented Generation (RAG) method designed for enhancing small-scale VLMs. Our approach integrates visual and textual retrieval modules to augment the input context, enabling the model to generate professional and accurate responses for cultural heritage queries. Experimental results on the constructed Hemudu Artifacts Visual Question-Answering dataset demonstrate the effectiveness of this approach. This method offers a solution for preserving and disseminating cultural heritage, bridging the gap between advanced VLM capabilities and domain-specific expertise.

Keywords: VLM, RAG, Cultural Heritage Assistant.

1 `Introduction

Recent years have witnessed the remarkable capability of the Large Language Model to generate high-quality responses without training [1], [15]. Through instruction fine-tuning and retrieval augmented generation techniques, LLM demonstrates exceptional performance in knowledge-intensive tasks. Subsequently, in computer vision, the development of visual instruction tuning [3] has led to significant achievements by Vision-Language Models, such as LLaVA-1.6-34B [5] and Qwen2-VL-72B [11], particularly in visual question-answering and optical character recognition tasks.

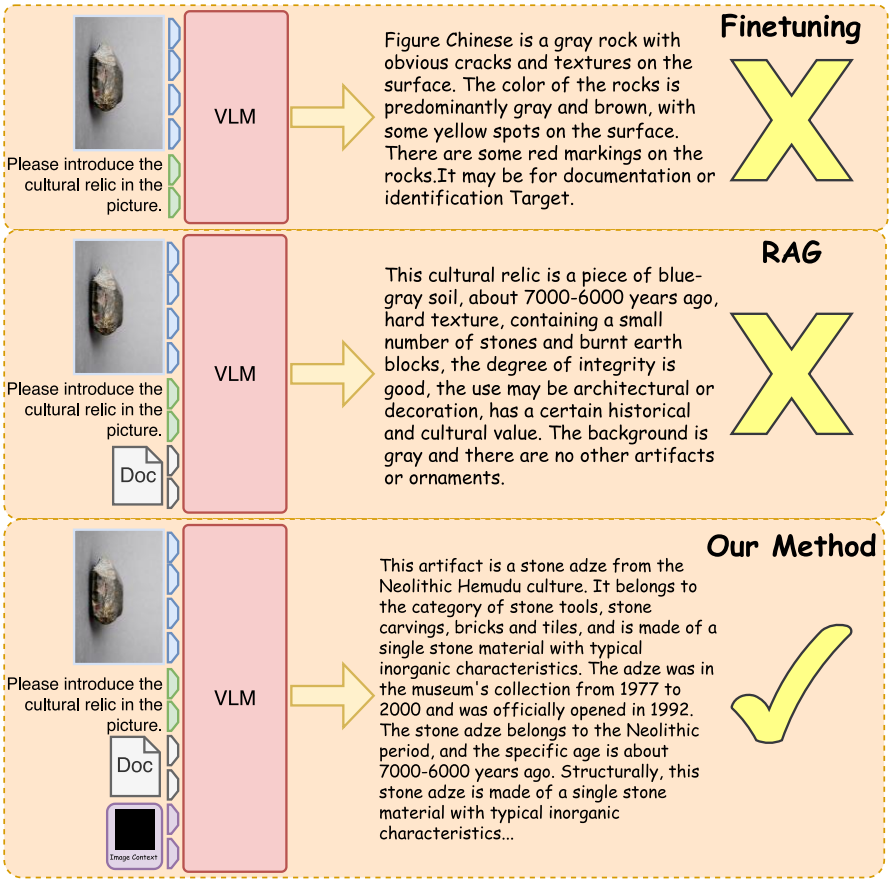

Fig. 1. Comparison between Finetuning, naïve RAG, and our method in describing cultural relics.

However, both LLMs and VLMs often struggle to provide domain-specific responses. While LLMs address this limitation through retrieval-augmented generation, accessing documents to adapt to domain expertise [18], VLMs face ongoing challenges in effective multimodal knowledge learning for specialized visual question-answering. Current multimodal retrieval-augmented approaches rely on processing documents as complete inputs to obtain embedded representations, followed by vector retrieval for

knowledge enhancement [29]. This methodology does not fully exploit vision and text information. The comparison in Fig 1. To address these challenges, this paper presents a lightweight retrieval augmented generation Method—Cultural Heritage Assistant, designed specifically for small-scale VLM to support artifact question-answering [22]. The system's architecture comprises two modules: Retrieval on Text and Retrieval on Image, which integrate information from both vision and text perspectives to enhance the VLM's comprehension and response capabilities.

Specifically, the visual retrieval module employs Vision Transformer to convert input images into high-dimensional embedding vectors, which are efficiently matched against a constructed vector database to identify professional descriptions of the most similar images. Concurrently, the text retrieval module collects relevant domain expertise within knowledge graphs based on VLM-generated image descriptions. These retrieved information elements are then fed into the VLM alongside the original question and image, significantly improving response quality without additional training.

To validate the framework's effectiveness, a specialized question-answering dataset focused on Hemudu culture artifacts is constructed. Extensive testing is performed using Qwen2-VL-2b [11] as the baseline model. The experimental results demonstrate both the superior performance and design rationality of the Cultural Heritage Assistant. The primary contributions of this work are summarized as follows:

1. Hemudu Artifacts Visual Question-Answering Dataset: This dataset provides a comprehensive visual question-answering resource for ancient Chinese archaeological studies.
2. Innovative Retrieval Augmented Generation: The proposed method demonstrates superior performance in domain-specific visual question-answering tasks.
3. Lightweight and Deployable Methodology: The method is specifically designed for seamless integration into cultural industry applications, contributing significantly to the preservation and dissemination of cultural heritage artifacts.

2 Related Work

Visual-Language Models. Large Language Models, exemplified by OpenAI's groundbreaking developments such as ChatGPT [27], have transformed human interaction. These advanced models demonstrate exceptional performance across natural language processing tasks through the implementation of pre-training, instruction tuning [28], and Reinforcement Learning from Human Feedback [6].Sub-sequently, the computer vision community has achieved modal alignment between vision and text through the development of VLMs, where image features are encoded and fed into large-scale models via visual encoders [10]. Early VLM implementations, such as BLIP-2 [4] and Flamingo [2], successfully integrated image features into LLMs, effectively bridging the gap between visual and textual modalities. Models like LLaVA-1.5 [3] and MiniGPT-

4 [7] further advanced this integration by incorporating region-level feature information, thereby enhancing their capacity for detailed image comprehension. Recent developments, including LLaVA-next [5] and Qwen2-VL [11], have extended support for broader resolution ranges while optimizing model architectures for improved efficiency and applicability through lightweight designs. However, despite continuous improvements in general-purpose tasks, these models exhibit limitations when processing domain-specific images, particularly in tasks requiring deep professional expertise.

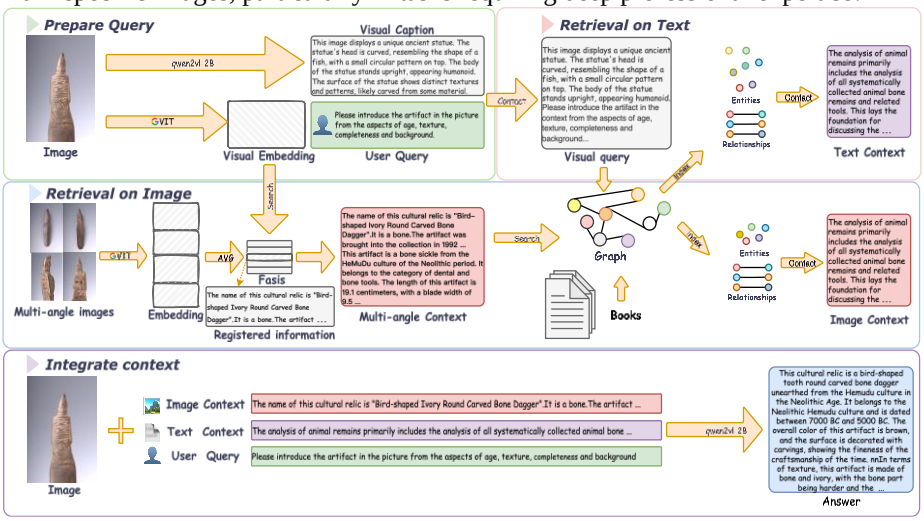

Fig. 2. Overview of Cultural Heritage Assistant. Illustration of the retrieval framework for integrating image context and text context to answer user queries effectively.

Retrieval-Augmented Generation. Despite their impressive performance in natural language processing tasks, Large Language Models frequently exhibit hallucinations, particularly in knowledge-intensive and domain-specific applications. To address this limitation and enhance LLM performance, several researchers have developed Retrieval-Augmented Generation systems. These systems expand the input space by extracting relevant textual passages from external sources [18], [19]. Unlike traditional fine-tuning approaches [9], RAG employs a retriever-generator architecture that enables real-time access to current and relevant information during inference, resulting in more accurate predictions. Recent years have witnessed significant improvements in RAG frameworks through enhanced retrieval mechanisms and reranking systems [16]. Several studies have incorporated knowledge graphs to augment RAG systems' retrieval capabilities, as demonstrated in Think on Graph [20], GraphRAG [14], and LightRAG [24]. Moreover, RAG applications have transcended textual boundaries, evolving into multimodal retrieval-augmented approaches that effectively utilize diverse information sources. For example, ColPali [29], treats each page of the PDF as a whole image and utilizes an embedding model for implicit segment retrieval. This approach has high computational requirements.

Fig. 3. Comparison with Qwen2-VL-2B Finetuning [9], GPT-4o-mini [21], Gemini-exp-1206 [17], with Cultural Heritage Assistant show that Cultural Heritage Assistant's answer is better.

Cultural Heritage Visual Question Answering. Initial experiments with general-purpose large language models for cultural heritage question-answering reveal significant limitations in domain expertise [23], as these models tend to generate generic responses. This performance proves inadequate for Visual Question-answering tasks that require a deep understanding of cultural heritage contexts. Even when augmented with web search capabilities [12], existing LLMs struggle to retrieve effective domain knowledge from external sources, which compromises response quality and accuracy.

3 Proposed Method

The primary objective of the research is to enable Vision-Language Models to perform effective visual question-answering tasks in the cultural heritage domain. This method is shown in Fig 2.

3.1 Knowledge Graph Construction

To enhance the information retrieval capabilities of knowledge bases, this study advances a knowledge graph-based text retrieval method, inspired by the LightRAG [24]. The construction process encompasses several key phases. Initially, large-scale books

are segmented into sections to optimize processing efficiency. This segmentation can be modeled as partitioning a corpus C into smaller subsets $S_1, S_2, \cdots, S_n$, where $\mathrm{C} = \cup_{i=1}^{n} S_i$, and each S_i is processed independently to minimize complexity. Subsequently, an automated system identifies and annotates critical entities, including historical figures, locations, and temporal markers, denoted as $E = e_1, e_2, \cdots, e_m$, where each e_i corresponds to an entity. These entities are connected through relationships R, forming a directed graph $G(E, R)$. For instance, "rice cultivation" and "pottery manufacturing" are interconnected, illustrating how the Hemudu people advanced their society. Each entity e_i and it's relationships $R(e_i)$ are then augmented with detailed descriptions $D(e_i)$ and linked to relevant background materials through keywords, enabling users to gain comprehensive insights into topics.

3.2 Retrieval on Text

Visual Context Representation. In visual question-answering tasks, VLMs generate responses based on image I and query Q. However, directly retrieving information solely based on query Q may lack sufficient context support, limiting the accuracy of VLM responses. To address this limitation, we propose an enhanced methodology that leverages VLM to extract visual caption C from the image I, expressed as:

$$C = VLM_{caption}(I) \tag{1}$$

where $VLM_{caption}$ denotes VLM specialized for generating image descriptions. The visual caption C is integrated with the original query Q to construct a visual query $\mathrm{Q_{visual}}$:

$$Q_{visual} = f(Q, C) \tag{2}$$

where $f(\cdot)$ represents the operation of concatenating the query Q and the visual caption C. We will use the visual query to perform retrieval of text.

Search on Graph. We use the LLM [11] to extract Keywords from $\mathrm{Q_{visual}}$, which we represent by the symbol:

$$K = LLM(Q_{visual}) \tag{3}$$

After extracting the keywords K. The function can represent the next step:

$$(E, R) = F(K) \tag{4}$$

where $F(K)$ returns the set of top entities E and relationships R based on the vector similarity between the query keywords and the knowledge graph entries [25]. And F is a word embedding-based model designed for semantic matching, which measures the similarity between word vectors to establish correspondences between keywords. The Text Context incorporates not only descriptions of entities but also details about the relationships between them. This can be expressed as:

$$\mathcal{C} = \bigcup_{i} \big(D(E_i) \cup D(R_i)\big) \tag{5}$$

where $\mathcal{C}$ represents the Text Context, E_i are the top entities, and R_i are the relationships linked to those entities, $D(.)$represents the linked relevant context of E_i and R_i.Finally, using the Text Context $\mathcal{C}$, the system combines the retrieved entity descriptions and their relationships to deliver comprehensive and accurate responses to the Visual Query, ensuring both precise retrieval and broad association capabilities are effectively integrated.

3.3 Retrieval on Image

Visual information plays an indispensable role in vision-language models. For a given image I, Vision Transformer(ViT) [13] is employed to extract its visual embedding E_I, where

$$E_I = ViT(I), \quad E_I \in R^d \tag{6}$$

and d is the dimensionality of the embedding space. Using this embedding, a vector database [26] retrieves similar artifact embeddings $\{E_{A_1}, E_{A_2}, \dots, E_{A_k}\}$. The corresponding structured artifact descriptions $\{D_{A_1}, D_{A_2}, \dots, D_{A_k}\}$. are then aggregated to form the Image Context, denoted as

$$\mathcal{C}(I) = \{D_{A_1}, D_{A_2}, \dots, D_{A_k}\} \tag{7}$$

To construct the database, multi-angle photographs of an artifact $\{\mathrm{I}_1, \mathrm{I}_2, \dots, \mathrm{I}_\mathrm{n}\}$ are encoded using the Vision Transformer to produce their respective embeddings $\{\mathrm{E}_{\mathrm{I}_1}, \mathrm{E}_{\mathrm{I}_2}, \dots, \mathrm{E}_{\mathrm{I}_\mathrm{n}}\}$. These embeddings are averaged to form a single artifact representation E_A:

$$E_A = \frac{1}{n}\sum_{i=1}^{n} E_{I_i} \tag{8}$$

which is stored in the vector database along with its structured description D_A. This approach ensures that the stored representation captures the artifact's visual features comprehensively from multiple angles. To mitigate hallucination, we subsequently leverage the multi-angle context obtained through visual retrieval to query the knowledge graph using the same search protocol described in previous sections. This approach not only effectively reduces model hallucination but also enables the extraction of visually grounded and task-relevant information from the graph structure. By leveraging visual embeddings, this method bridges a critical gap in cultural heritage information retrieval, uncovering functional or cultural connections that may not be apparent through traditional text-based techniques.

3.4 Integrate context

The integration of image context I, text context T, the input image x, and the user query q serves as the foundation for generating responses through the VLM. This can be formally expressed as:

$$Y = VLM(I, T, x, q) \tag{9}$$

where Y is the generated answer.

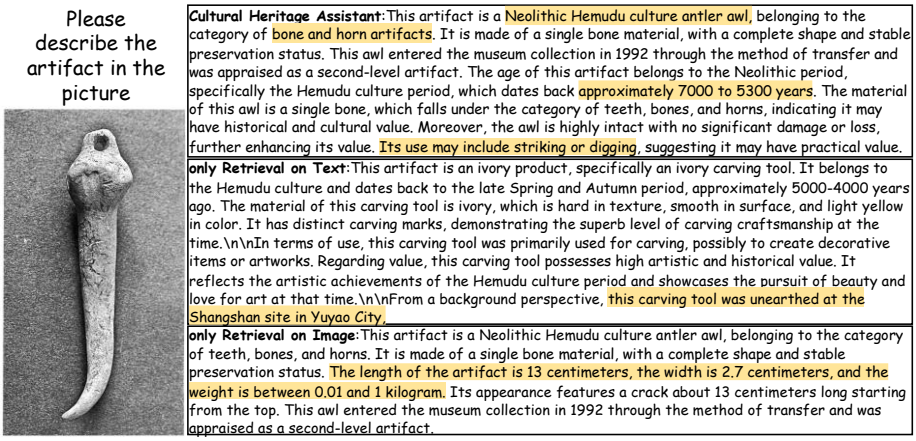

Fig. 4. The ablation study retrieves only text and only images. The approach we propose makes VLM's answer more comprehensive and professional.

4 Experiment

4.1 Dataset Setup

In addition to the 2,912 multi-view artifact images (forming our training set that covers 412 distinct cultural artifacts, with 7-8 photographs per item capturing multiple perspectives) utilized in this study, an additional corpus of 70 cultural artifact images is collected from professional publications and the Hemudu Culture Museum to constitute our test set. These supplementary artifact images encompass diverse material compositions, including lithic implements, bone tools, pottery vessels, and jade objects. The image collection spans various quality specifications, including monochrome, color, and high-resolution formats. Notably, these test images are captured from arbitrary viewpoints and correspond to 15 newly acquired artifacts from the museum's latest collections that never appeared in the training set, thereby providing comprehensive data support for model evaluation and enabling thorough assessment of our framework's generalization capability and performance metrics while ensuring reliable evaluation of cross-artifact generalization.

4.2 Implementation Details

Our method is constructed based on the Qwen2-VL-2B model. The experiments are being conducted on an NVIDIA GeForce RTX 3090 GPU. To achieve efficient image retrieval, a dataset of 2912 multi-view cultural relic images is embedded into a vector database(Fasis). For text retrieval, the method utilizes the Qwen2-VL-2B model to generate captions of images. During this process, the global mode functionality provided by LightRAG is adopted to retrieve knowledge fragments related to entity attributes from extensive textual data. The obtained image context and text context are then concatenated with the original input and fed into the Vision-Language Model to enrich and optimize the contextual information of generated content. To validate the effectiveness of the proposed method, we use finetuning of Qwen2-VL-2B, encompassing all 2,912 cultural relic images and their caption. Additionally, GPT-4o is introduced as a performance benchmark for comparative analysis.

Table 1. Dataset Configuration for Cultural Artifact Analysis.

Feature	Training Set	Test Set
Total Images	2,912	70
Unique Artifacts	412	15
Views per Artifact	7-8	1
Data Sources	Original Collection	Hemudu Museum
Material Types	Lithic, Bone Pottery, Jade	Lithic, Bone Pottery, Jade
Image Specifications	Multi-view Multi-quality	Arbitrary Views Multi-format

4.3 Evaluation Metrics

Providing expert-level responses to cultural artifact inquiries presents significant challenges in computer vision. Building upon existing work [8], [23], a multi-dimensional comparative approach based on Large Language Models is proposed. GPT-4o [21] is employed to score each baseline and compare it with our method. Since artifact analysis requires substantial domain expertise, three evaluation dimensions are established [24]:

- Veracity: Is the information provided accurate and reliable?
- Professionalism: Does the answer demonstrate a certain level of professional knowledge and tone?
- Comprehensiveness: Does the answer provide a complete description of the artifact's details?

Below is the prompt we use with GPT-4o. We provide a reference answer for GPT-4o to evaluate against, which helps improve the reliability of the evaluation and reduce hallucinations.

> Role: You are an expert in Hemudu culture archaeology. Your task is to evaluate the following description based on three criteria.

Answer: The provided response to be evaluated.

Objective: Evaluate the answer in the dictionary below according to the following three criteria and the reference answer, and assign aspecific score out of 10.

Evaluation Criteria:

Veracity: Is the information provided accurate and reliable?

Professionalism: Does the answer demonstrate a certain level of professional knowledge and tone?

Comprehensiveness: Does the answer provide a complete description of the artifact's details?

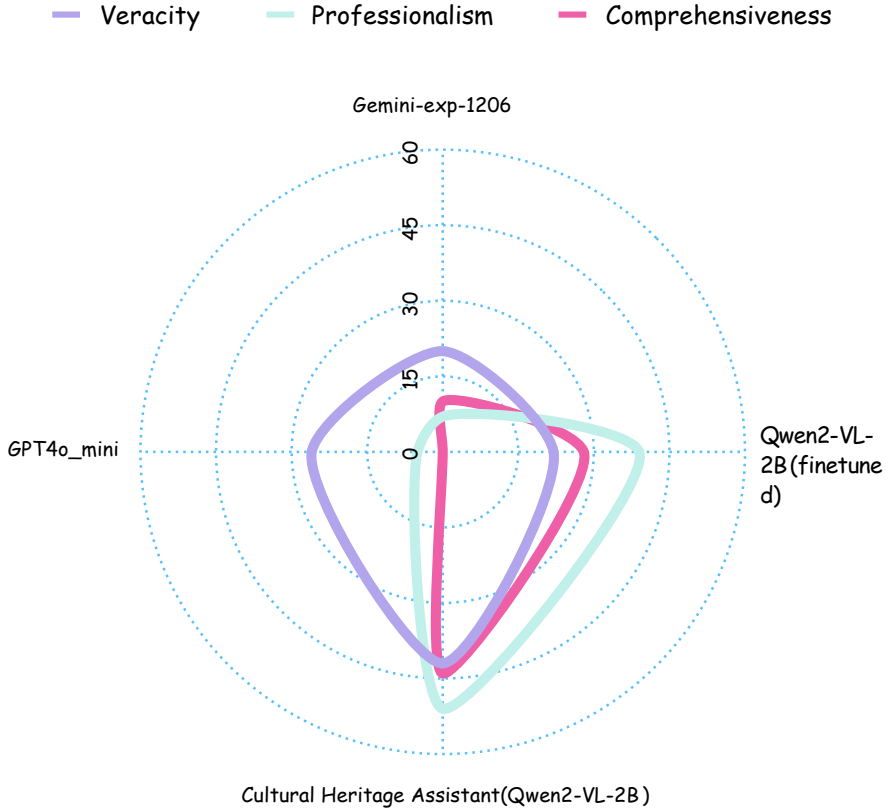

Fig. 5. Cultural Heritage Assistant significantly outperforms other models across all dimensions.

4.4 Experimental Results

In our experiment, the performance of different models on the Hemudu artifact visual question answering task were compared, including the fine-tuned Qwen2-VL-2B model [11], GPT-4o-mini [21], Gemini2-exp-1206 [17], and the Cultural Heritage Assistant (Qwen2-VL-2B). Evaluations were conducted based on four critical dimensions:

veracity, professionalism, comprehensiveness, and overall performance. The results are summarized in Table 1. The Cultural Heritage Assistant significantly outperforms other models across all dimensions. It achieves the highest scores in professionalism at 72.8%, veracity at 60%, and comprehensiveness at 62.8%, with an overall performance score of 67.1%. In comparison, Qwen2-VL-2B(Finetune) shows a professionalism score of 55.7%, veracity at 31.4%, and comprehensiveness at 40%, resulting in an overall score of 40%. GPT-4o-mini struggles significantly, particularly in comprehensiveness where it scores 0, and achieves an overall score of only 4.2%. Similarly, Gemini2-exp-1206 underperforms across all metrics, scoring 28.5% in veracity, 10% in professionalism, 14.2% in comprehensiveness, and an overall score of 10%.

Table 2. Performance comparison of various models on veracity (V), professionalism (P), comprehensiveness (C), and overall (O). The overall score is calculated by summing the scores of these three metrics, and the model's highest score is then used to determine the proportion of instances where each model achieves the highest score. Scores indicate the percentage of times each model achieved the highest score on the same artifact, based on GPT-4o evaluation.

Model	V (%)	P (%)	C (%)	O (%)
Qwen2-VL-2B (Finetune)	31.4	55.7	40.0	40.0
GPT-4o-mini	37.1	7.1	0.0	4.2
Gemini2-exp-1206	28.5	10.0	14.2	10.0
Cultural Heritage Assistant	**60.0**	**72.8**	**62.8**	**67.1**

Therefore, as shown in Fig 3, the proposed method works well in the visual question and answer task of cultural relics, and it is more professional than the answers of models such as GPT-4o-mini.

4.5 Ablation Study

Table 3. Ablation study. Performance comparison of various models on veracity (V), professionalism (P), comprehensiveness (C), and overall (O). We use GPT-4o [21] to assess the quality of answers. Text retrieval enhances comprehensiveness, while image retrieval improves professionalism.

Model	V	P	C	O (Total)
Qwen2-VL-2B (BaseLine)	5.47	4.90	4.13	14.50
only Retrieval on Image	7.13	6.76	6.81	20.70
only Retrieval on Text	6.39	6.21	6.20	18.80
Cultural Heritage Assistant	**8.16**	**8.39**	**8.47**	**25.01**

As shown in Fig 4, to more accurately capture the contribution of each module, the average scores of Veracity, Professionalism, and Comprehensiveness are used instead of relying solely on the highest scores. The results demonstrate that the complete dual enhancement strategy significantly outperforms the baseline model and the single-en-

hancement variants, with the total score increasing from 14.50 to 25.01 (+10.51). Specifically, Retrieval of Images effectively extracts visual features but struggles to capture contextual information. In contrast, Retrieval of Text excels in providing broad contextual and relational information but lacks the capability to identify fine-grained visual details. By integrating the strengths of both modules, the dual retrieval strategy achieves notable improvements in Veracity (+2.69), Professionalism (+3.49), and Comprehensiveness (+4.34). As shown in Table 3, these findings validate the complementary nature of image and text enhancement and their synergistic effect in optimizing model performance.

5 Conclusion

The strategy employs a dual retrieval mechanism that draws on external professional knowledge to improve visual question-answering performance in specific domains. By retrieving pertinent information from both text and image resources, this mechanism integrates the knowledge into the VLM's input context, thereby enriching the model's comprehension and boosting the accuracy of its generated responses. Experimental results demonstrate the broad applicability and effectiveness of this retrieval method within the cultural heritage domain, highlighting its capacity to significantly enhance VLM performance without imposing additional training demands. We believe that this paper will offer new insights into the adaptation and application of VLM in specialized domains, ultimately contributing to the preservation and inheritance of cultural heritage.

Acknowledges. This project was partially supported by Zhejiang Provincial Philosophy and Social Sciences Planning Project（25JCXK005YB）, Ningbo Public Welfare Science and Technology Plan (No. 2021S093), Zhejiang Province Cultural Relics Protection Science and Technology Project（2022008）.Philosophy and Social Science Planning Cross-disciplinary Key Support Subjects of Zhejiang Province (No. 22JCXK08Z)

References

1. K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, “Retrieval Augmented Language Model Pre-Training,” in ICML, 2020.
2. J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S.Samangooei,M.Monteiro,J.L.Menick,S.Borgeaud,A.Brock,A.Nematzadeh, S. Sharifzadeh, M. a. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, “Flamingo: a visual language model for few-shot learning,” in Advances in Neural Information Processing Systems (S. Koyejo, S. Mohamed, A. Agarwal, D.Belgrave,K.Cho,andA.Oh,eds.),vol.35,pp.23716–23736,CurranAssociates, Inc., 2022.
3. H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual Instruction Tuning,” in NeurIPS, 2023.
4. J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: Bootstrapping Language-Image Pretraining with Frozen Image Encoders and Large Language Models,” arXiv preprint arXiv:2301.12597, 2023.

5. H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "LLaVA-NeXT: Improved reasoning, OCR, and world knowledge," 2024.
6. L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., "Training Language Models to Follow Instructions with Human Feedback," in NeurIPS, 2022.
7. D.Zhu,J.Chen,X.Shen,X.Li,andM.Elhoseiny,"MiniGPT-4: Enhancing Vision Language Understanding with Advanced Large Language Models," arXiv preprint arXiv:2304.10592, 2023.
8. D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, L. Baraldi, M. Cornia, and R. Cucchiara, "The (R)Evolution of Multimodal Large Language Models: A Survey," arXiv preprint arXiv:2402.12451, 2024.
9. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," arXiv preprint arXiv:2106.09685, 2021.
10. H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023.
11. P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," arXiv preprint arXiv:2409.12191, 2024.
12. Z. Zhang, Y. Zhang, X. Ding, and X. Yue, "Vision search assistant: Empower vision-language models as multimodal search engines," 2024.
13. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," ICLR, 2021.
14. D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson, "From local to global: A graph rag approach to query-focused summarization," arXiv preprint arXiv:2404.16130, 2024.
15. G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière,T.Schick,J.Dwivedi-Yu,A.Celikyilmaz,et al.,"Augmentedlanguagemodels: a survey," arXiv preprint arXiv:2302.07842, 2023.
16. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," ArXiv, vol. abs/2005.11401, 2020.
17. Google, "Introducing gemini 2.0: our new ai model for the agentic era," 2024.
18. A. Asai, K. Hashimoto, H. Hajishirzi, R. Socher, and C. Xiong, "Learning to retrieve reasoning paths over wikipedia graph for question answering," arXiv preprint arXiv:1911.10470, 2019.
19. D. Caffagni, F. Cocchi, N. Moratelli, S. Sarto, M. Cornia, L. Baraldi, and R. Cucchiara, "Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 1818–1826, June 2024.
20. J. Sun, C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, L. M. Ni, H.-Y. Shum, and J. Guo, "Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph," 2024.
21. OpenAI, "Hello gpt4-o.," 2024.

22. P. Bongini, F. Becattini, A. D. Bagdanov, and A. Del Bimbo, “Visual question answering for cultural heritage,” in IOP Conference Series: Materials Science and Engineering, vol. 949, p. 012074, IOP Publishing, 2020.
23. P. Bongini, F. Becattini, and A. D. Bimbo, “Is gpt-3 all you need for visual question answering in cultural heritage?,” 2023.
24. Z. Guo, L. Xia, Y. Yu, T. Ao, and C. Huang, “Lightrag: Simple and fast retrievalaugmented generation,” 2024.