SARCH: Multimodal Search for Archaeological Archives

Nivedita Sinha
IIT Delhi, New Delhi
India

Bharati Khanijo
IIT Delhi, New Delhi
India

Sanskar Singh
IIT Delhi, New Delhi
India

Priyansh Mahant*
CSV Technical University, Bhilai
India

Ashutosh Roy*
CSV Technical University, Bhilai
India

Saubhagya Singh Bhadouria*
GGSIP University, New Delhi
India

Arpan Jain®
AK Technical University, Lucknow
India

Maya Ramanath
IIT Delhi, New Delhi
India

Abstract

In this paper, we describe a multi-modal search system designed to search old archaeological books and reports. This corpus is digitally available as scanned PDFs, but varies widely in the quality of scans. Our pipeline, designed for multi-modal archaeological documents, extracts and indexes text, images (classified into maps, photos, layouts, and others), and tables. We evaluated different retrieval strategies, including keyword-based search, embedding-based models, and a hybrid approach that selects optimal results from both modalities. We report and analyze our preliminary results and discuss future work in this exciting vertical.

ACM Reference Format:

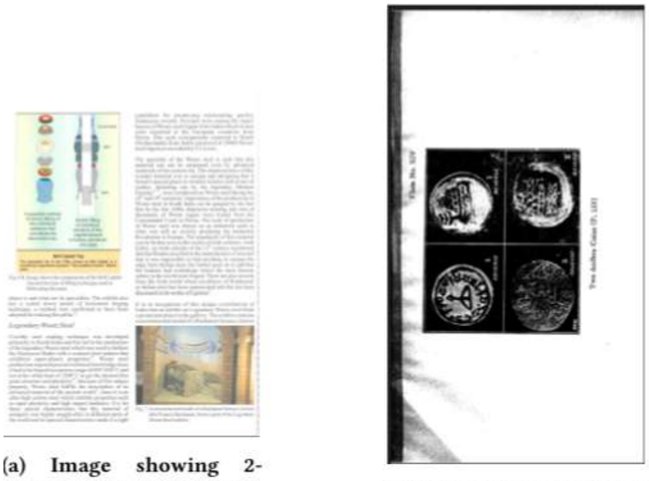
Nivedita Sinha, Bharati Khanijo, Sanskar Singh, Priyansh Mahant, Ashutosh Roy, Saubhagya Singh Bhadouria, Arpan Jain, and Maya Ramanath. 2018. SARCH: Multimodal Search for Archaeological Archives. In Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym ·XX). ACM, New York, NY, USA, 7 pages. https: i{doiorg/XXXXXXX.XXXXXXX

1 Introduction

Old archaeological archives, often consisting of field reports, survey documents, excavation records and books are invaluable resources for researchers. These documents are typically stored as scanned pdfs, consisting of multi-modal data, including textual descriptions, maps, photos and tables, that provide rich historical and geographical context. While many corpora have multi-modal content, there are a number of archaeology-specific challenges when dealing with old archaeological records.

- Complex layouts: Figure 1a shows a rather complicated layout with 2-column text with 2 figures, of which one Figure (left/top) has text of its own that describes the images and should be extracted. Further, Figure 1b shows a bad quality scan that prevents users from observing the details of the photos of coins. The photo itself is oriented differently then the rest of the document.
- Multiple modalities: While prior work has typically dealt with models for text, images and tables, we find that this categorisation is insufficient when dealing with archaeological data. First, "image" is too broad a category ·- we need to further break it down into specific kinds of images, such as maps, photos, layouts, all of which are meaningful in the archaeological domain. Second, textual content from within the image or table adds very important context. This can specially be seen in the case of maps - the locations in the maps are important context and serve to describe the map itself. Third, image or table captions are often short and not very descriptive. The actual description comes from within the text that refers to the image or table. Such an example can be seen in Figure 1b, where the actual description of those coins are described in a different page.

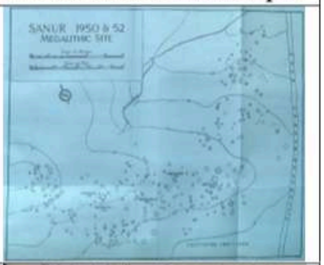
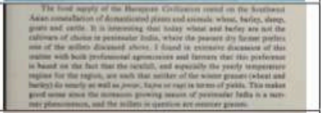
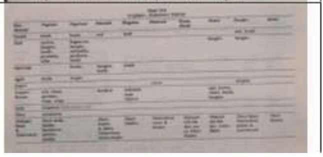
Table 1 shows examples of text queries and results. As can be seen, the modality of the result is different for each query and is to be inferred using the query itself.



(a) Image showing 2-column text, image with caption

(b) Page with shadows and rotated image

Figure Complex layouts and bad quality 1: scans

Query	Result type	Relevant result in the top-5
Map showing sites discovered during 1948 and 1957	Map	
Primary crops of the Harappan civilization	Text	 The food supply of the Harappan Civilization rested on the Southwest Asian constellation of domesticated plants and animals: wheat, barley, sheep, goats and cattle. It is interesting that today wheat and barley are not the cultivars of choice in peninsular India, where the peasant dry farmer prefers one of the millets discussed above. I found in extensive discussion of this matter with both professional agronomists and farmers that this preference is based on the fact that the rainfall, and especially the yearly temperature regime for the region, are such that neither of the winter grasses (wheat and barley) do nearly as well as *jowar*, *bajra* or *ragi* in terms of yields. This makes good sense since the monsoon growing season of peninsular India is a summer phenomenon, and the millets in question are summer grasses.
Raw material used for making beads?	Table	

returned by a Table Example result queries in benchmark and search engine 1: top-5 our our

In this paper, we describe our system, SARCH!, an end-to-end solution for pre-processing archival scans of archaeological content and a search engine to search the archive in different modalities. Our pre-processing pipeline separates text from images and tables, and further classifies images into maps, layouts, photos, and a generic "figure". Further, to facilitate the search functionality, the extraction pipeline looks to add context to: i) maps, by extracting locations mentioned in the maps, and tables, by extracting the contents of the table, ii) images and tables, by associating them with their captions and additionally searching surrounding text for descriptions of the image or table. As mentioned before, the extraction is challenging because of the varying quality of scans (see Section 2 for details).

While typical multi-modal search may receive queries in text, but return results from other modalities, the same idea comes with additional complexity in the case of archaeological search. As mentioned earlier, low-quality scans are quite prevalent. While applying standard corrections may help in extracting text, the same may not be true of images. Therefore, most of the description of the image comes not from the image itself, but from its context (that is, the surrounding text where the image is described), thereby limiting the utility of image embeddings alone. Hence, we applies three different retrieval strategies: keyword-only, embeddings-only and a hybrrid search that does both embeddings-based search as well as keyword search (see Section 3 for details).

A video demonstrating the proposed system can be accessed at https://tinyurl.com/sarchdemo.

Related Work. Around the world, several extensive digital archaeological archives such as The Digital Archaeological Record [5], Archaeology Data Service [4], Arachne [3], etc. help researchers document their research and search for related literature and artifacts within these websites. However, all these services rely extensively on manual curation and metadata for their search. In contrast, our corpus consist simply of old, scanned books and reports that are automatically processed. We rely on automated methods of text, image, and table extraction. Our search is based directly on the content of the documents, rather than metadata.

For our search system, we make use of MiniLM [19], CLIP [17], and TAPAS [14] for textual content, images and tables, respectively, and use the embeddings to conduct similarity searches on user queries. Further, we perform keyword search using the BM25 [18] scoring method. More details are described in Section 3.

1.1. Contributions

In this paper, we address the problem of building a robust, multi-modal search system that enables retrieval from archival archaeological legacy documents. Our contributions are as follows:

e a pipeline for extracting multi-modal content from old, scanned archaeological reports.

e three different search models: keyword only, embeddings only and a hybrid search system utilizing both embedded vector search as well as keyword search. A key feature of our system is that it is able to return results of types that are very specific to archaeology, such as maps, photos and layouts.

Organisation. The rest of this paper is organised as follows. Section 2 describes the offline processing of our system, including content extraction and cleaning for scanned archaeological texts. Section 3 gives an overview of our search engine along with sample results in each of the three modalities, while Section 5 concludes and outlines future work.

2 Offline Processing

Figure 2 shows the components of our offline processing system · content extraction, embeddings generation and indexing. The input to the system is the corpus of scanned pdf files. Our corpus currently has 296 scanned documents, consisting of 63,208 pages. We describe these three components next.

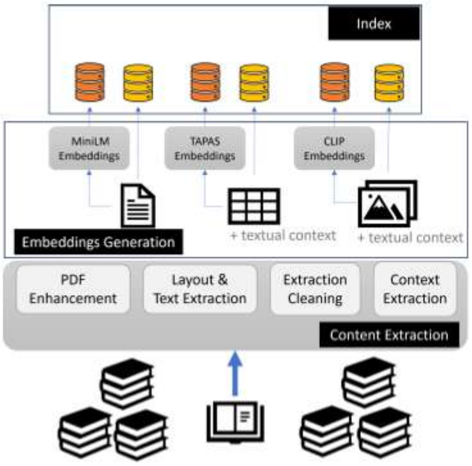


Figure Offline processing: extraction, embedding 2: content generation and indexing

2.1 Content Extraction

PDF enhancement. The scanned pages vary greatly in quality. The use of OCR tools are best done over good quality pdfs. To enhance quality of OCR results, we processed each page as follows: i) convert the page to grayscale, ii) denoise using bilateral filtering (this is chosen for its superior edge preservation, which in turn is advantageous for scanned textual content), iii) process with Otsu·s thresholding [15]. This workflow improves text clarity by producing a high-contrast binary image. All these steps are implemented with OpenCV [12].

Layout and Content Extraction. Parsing the layout of the page was a big challenge, as the layout could combine multiple modalities. We used Surya Document Analysis Toolkit [16] to identify text, images, and tables. This toolkit produces bounding boxes and classifies each bounding box as text, image, or table. However, the extraction of tabular structure and content was not robust. Therefore, for table content extraction, we used Qwen2.5-VL-7B-Instruct [11]. Once the layout was parsed, text regions were analyzed using OCR [16].

Additionally, we implement a CLIP [17] based zero shot classifier where an input image is classified into one of the 4 classes: map, photograph, site layout, and figure.

Extraction Cleaning. Despite efforts to improve the scan quality, quite often the OCR tools fail in correctly extracting the text. Additionally, formatting tags like bold, italics were also present in the text extracted by OCR. These tags affect search quality since they are not part of the semantic content of the page. Therefore, we had a "text cleaning" module that took as input whatever text was extracted by the OCR tool and attempted to correct the text. We used beautiful soup python library [8] for this purpose. Spell correction, using TextBlob [10] was done to improve text search.

For tabular data cleaning, we first used Qwen to extract the contents of the table, but noticed that further processing was required for more accurate extraction. Therefore, we implemented our own rule-based implementation that takes as input the extracted content from Qwen and further cleans the data to ensure cell alignment with the appropriate header and to handle empty cell values. During preprocessing, HTML tags are removed from the extracted content. Empty cells in tables are filled with NaN to create a consistent format. Rows are also adjusted so they always match the number of columns in the header, ensuring a uniform table structure for later processing.

Context Extraction. This module is the most important module for our search system. For images and tables, the context in which they appear is key to ensuring that they are returned as relevant results. The most obvious context for both images and tables is the caption. But, a more semantically rich context can be found in the text that describes the image or table, as well as text that occurs in the image or table itself.

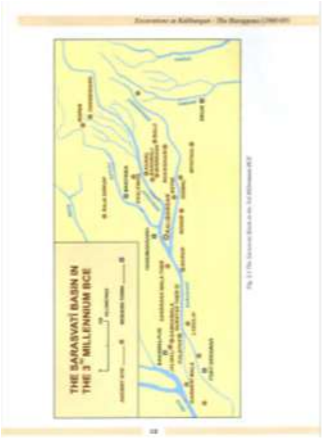
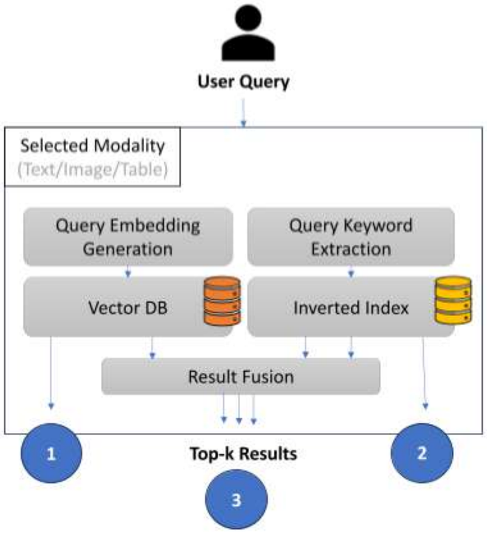


Figure Rotated Picture 3: on page

Therefore, we further enhance the context in the following ways.

(1) Sometimes a caption could be present in the text extracted within the image, especially where the image is rotated. Here, it is extracted by finding the location the word like "Fig" or "Figure" within the OCR text followed by number until the end of sentence as shown in Figure 3.
(2) We look through the textual content surrounding the images or tables on the page where they were found, as well as the previous and next page and identify the paragraphs that mention the ordinal number of that image (table), for example, "Figure 2 shows the map of..." through string matching. We identified the description containing the image ID, e.g. "Figure 2" present in the caption. The image may be referred to in multiple paragraphs. We use all these paragraphs as context.
(3) For maps, we extract the location names from the map. This is a challenging task since the location names could be spelled out with non-standard spacing as well as orientations. Therefore, a simple OCR will only extract a small subset of location names. We are using Surya OCR model [7] to read any text present inside the map. Future work includes trying more sophisticated text spotting methods for maps.
(4) For tables, the content of the tables is also used as context. Apart from row and column headers, the non-numerical values are valuable context and help in retrieval. To process



Search Pipelines: The first pipeline takes the Figure in 4: 1. user·and modality and searches the appropriate query vec- The second pipeline does keyword search using DB, 2. tor a the BM25 scoring model, The third pipeline is hybrid 3. a method that combines reranks results from the and previous searches. two

each table and produce matching captions, we use the Qwen [11] model. First, Qwen is prompted to extract any table-related captions or descriptions that may already exist. They are saved as the table·s caption if Qwen can find them. If not, we give Qwen instructions to summarize the table·s contents, which is subsequently utilized to generate a table caption.

Pipeline Output. Once a pdf file goes through the pipeline, we have the following content that is ready for indexing: i) textual content along with page number, ii) images and tables, image type (relevant only for images), associated with their contextual text and their page numbers.

In all, we have 63,208 pages, 22,349 images, and 1,824 tables that we extracted from our corpus of 296 documents.

2.2 Embeddings Generation and Indexing

As mentioned in the Introduction, our search system is a hybrid system and makes use of both embeddings search as well as keyword search. As shown in Figure 2, for each of the three kinds of content that we extract · text, image (further classified into map, photo, layout, and figure), and table - separate embeddings are generated using MiniLM [19], CLIP [17] and TAPAS [14], respectively and stored in a database to facilitate top-k nearest neighbor search. In addition, the textual content is stored as an inverted index to facilitate keyword search.

3 Search Architecture

The search takes a user query in natural language and can return results in three different modalities: text, image, and table. We have analyzed three different retrieval systems, a keyword-based search, an embedding-based search and a hybrid approach. Figure 4 describes these pipelines.

Keyword-Based Search. The query is analyzed to extract keywords, which are then used to search the corpus. The corpus, consisting of text extracted by the content extraction pipeline (Section 2) is indexed at the page level for each document, using Apache Solr [1]. The extracted images and tables are indexed using their extracted textual context.

The most important keywords from the user·s query are extracted using standard methods. For each query, the top-k results, scored using BM25 [2] [18] are returned. We expect this method to be beneficial for queries with archaeology specific words.

Embedding-Based Search. This is a multi-modal embeddings based search system. Embeddings are generated using modality-specific models. For textual data, embeddings are generated using pre-trained sentence transformer all-MiniLM-L6-v2 [6, 19]; for image data and its extracted textual context, embeddings for both images and their associated textual context are generated using CLIP (17]; for tabular data and its associated context, embeddings are generated using TAPAS [14]. These embeddings are stored in appropriated indexes - (Solr [1], Postgres+pgvector [9]) along with associated metadata (eg. document name, path, etc).

User query in natural language is encoded using a modality-specific model. Top-k results corresponding to most similar embeddings associated with the selected modality in the vector database are scored using cosine similarity and returned. *

Hybrid search. The results of these retrieval systems are ranked using their scores and are merged using reciprocal rank fusion [13]. Top-k of the merged results based on the rank are returned.

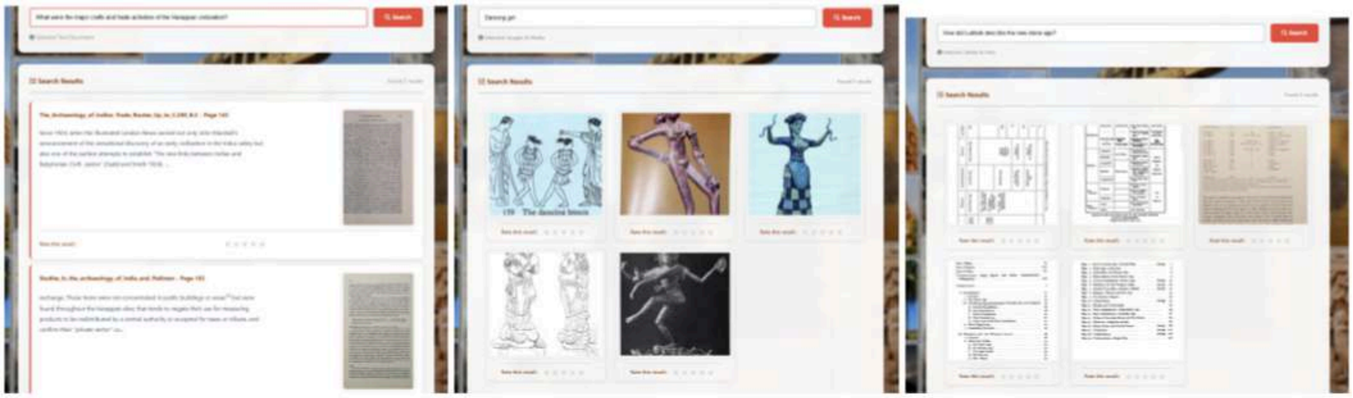
Examples. Figure 5 shows examples of the result by hybrid search for each of the three modality types. For text, the query "Major trade activities of the Harappan civilization" in Figure 5a returns snippets from a book on the Archaeology of Indian Trade Routes as the first result. For image modality, the query "photo: dancing girl" in Figure 5b shows at least 3 photos of dancing girls, including the famous dancing girl from Mohenjodaro (second result). For table modality, the query ·How did Lubbock describe the New Stone Age?· returns a comparative table of classifications of the stone age, showing how different scholars, including Lubbock, classified prehistoric periods (the first result in Figure 5c).

4 Evaluation

We are not aware of any standard benchmarks for archaeological search. Therefore, we created our own benchmarks and report on our preliminary results on these benchmarks.

Benchmark. We solicited the help of a professional archaeologist in creating and refining 15 text 11 image queries and 4 table queries that would be of interest to archaeologists.

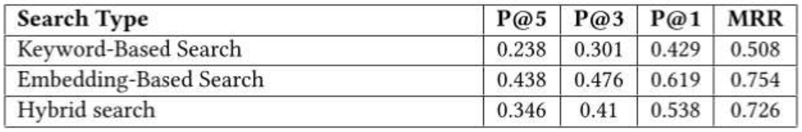
Results and Analysis. We generated the top-5 results and asked PhD students in our group to evaluate the relevance of each result.We calculated P@5, P@3, P@1 and MRR for each of the search pipelines:



(a) Text modality (Query: "Major trade activities of the Harappan civilization")

(b) Image modality (Query: "Dancing Girl")

(c) Table modality (Query: "How did Lubbock describe the New Stone Age")



Search Type	P@5	P@3	P@1	MRR
Keyword-Based Search	0.238	0.301	0.429	0.508
Embedding-Based Search	0.438	0.476	0.619	0.754
Hybrid search	0.346	0.41	0.538	0.726

Table Benchmark Evaluation results 2:

Keyword-based, embedding-based and hybrid search. The results are tabulated in Table 2. Our results show that the embedding-based search results are better overall. The MRR and P@1 numbers indicate that relevant results are ranked relatively high. However, in our preliminary analysis, we observed that for queries with hard archaeology-specific terms (example: "Harappa", "New Stone Age", etc,), the keyword-based search pipeline gave the best results.

5 Conclusion and Future Works

In this paper, we described SARCH, our multi-modal search system for archaeological archive data. The data consists of pdf scans of various documents and is of varying quality. We developed a content extraction system to extract text, images, and tables from these scans. To query this data, we implemented three search pipelines consisting of both embedding-based search as well as keyword search, Our preliminary results are promising.

Our future work will focus on both improved content and context extraction (especially from maps and tables) as well as improved search pipelines that use instruction-tuned models that eliminate the need for the user to select a modality. We plan to further investigate which queries perform well with just keyword context and which need image representations as well.

6 GenAII Usage Disclosure

We have used GenAI tools to help write code for the system described here. We have also used GenAI tools to correct typos and grammar in writing this paper.

References

(1) Apache solr. https://solr.apache.org/.

2) Apache solr reference guide. https://solr.apache.org/guide/solr/latest/indexing-guide/schema-elements.htmlésimilarity.

3] Arachne. https;//arachne.dainst.org.

4] Archaeology data service. https://archaeologydataservice.ac.uk.

5] The digital archaeological record. https://www.tdar.org.

6] Pretrained models - sentence transformer documentation. https://www.sbert.net/docs/sentence_transformer/pretrained_models.html. Last Accessed: 2025-06-16.

7) Vik paruchuri. 2024b. surya; Accurate line-by-line text detection and recognition in complex documents.

8] Beautiful soup. https://beautiful-soup-4.readthedocs.io/en/latest/, 2015. Last Accessed: 2025-06-17.

9] pgvector: Open-source vector similarity search for postgres. https://github.com/pevector/pgvector, 2021. Last Accessed: 2025-06-16.

[10] Textblob, https://textblob.readthedocs.io/en/dev/quickstart.html, 2025. Last Accessed: 2025-06-17.

{11]_ Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.

[12] Gary Bradski. The opencv library. Dr, Dobb's Journal: Software Tools for the Professional Programmer, 25(11):120-123, 2000.

(13] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Biittcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel, editors, Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July '9-23, 2009pages ome uM en

(**14**] Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Miller, Francesco Piccinno, and Julian Martin Eisenschlos. Tapas: Weakly supervised table parsing via pre-training. arXiv preprint arXiv:2004,02349, 2020,

[15] Nobuyuki Otsu. A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics, 9(1):62-66, 1979,

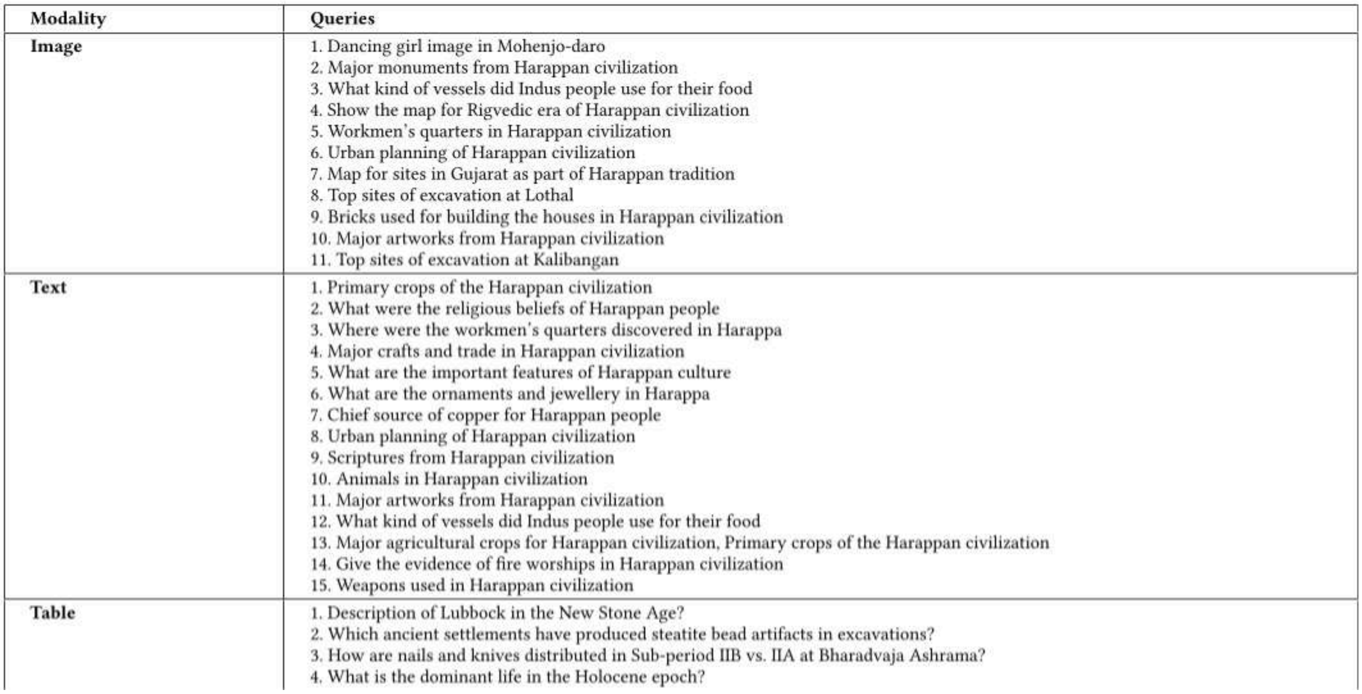
[16] Vikas Paruchuri and Datalab Team. Surya: A lightweight document ocr and analysis toolkit. https://github.com/VikParuchuri/surya, 2025. GitHub repository.

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of Proceedings of Machine Learning Research, pages 8748-8763. PMLR, 2021.

[18] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends®in Information Retrieval, 3(4):333-389, 2009.

[19] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. Advances in neural information processingsystems,33:5776·5788, 2020.

Appendix



Modality	Queries
Image	1. Dancing girl image in Mohenjo-daro 2. Major monuments from Harappan civilization 3. What kind of vessels did Indus people use for their food 4. Show the map for Rigvedic era of Harappan civilization 5. Workmen's quarters in Harappan civilization 6. Urban planning of Harappan civilization 7. Map for sites in Gujarat as part of Harappan tradition 8. Top sites of excavation at Lothal 9. Bricks used for building the houses in Harappan civilization 10. Major artworks from Harappan civilization 11. Top sites of excavation at Kalibangan
Text	1. Primary crops of the Harappan civilization 2. What were the religious beliefs of Harappan people 3. Where were the workmen's quarters discovered in Harappa 4. Major crafts and trade in Harappan civilization 5. What are the important features of Harappan culture 6. What are the ornaments and jewellery in Harappa 7. Chief source of copper for Harappan people 8. Urban planning of Harappan civilization 9. Scriptures from Harappan civilization 10. Animals in Harappan civilization 11. Major artworks from Harappan civilization 12. What kind of vessels did Indus people use for their food 13. Major agricultural crops for Harappan civilization, Primary crops of the Harappan civilization 14. Give the evidence of fire worships in Harappan civilization 15. Weapons used in Harappan civilization
Table	1. Description of Lubbock in the New Stone Age? 2. Which ancient settlements have produced steatite bead artifacts in excavations? 3. How are nails and knives distributed in Sub-period IIB vs. IIA at Bharadvaja Ashrama? 4. What is the dominant life in the Holocene epoch?

for Table Modality-wise Benchmark Queries used Evaluation 3: