

Article

RiverEcho-2.0: A Real-Time Interactive System for Yellow River Culture via Enhanced MultiModal Document RAG

Haofeng Wang^{1,2}, Yilin Guo³, Tiange Zhang¹, Zehao Li⁴, Tong Yue³, Yizong Wang³, Rongqun Lin⁵, Feng Gao^{6,*}, Shiqi Wang⁷ and Siwei Ma^{3,*}

¹ School of Electronic and Computer Engineering, Peking University, Shenzhen 518071, China

² Advanced Institute of Information Technology, Peking University, Hangzhou 310020, China

³ School of Computer Science, Peking University, Beijing 100091, China

⁴ The School of History, Renmin University of China, Beijing 100086, China

⁵ Pengcheng Laboratory, Shenzhen 518055, China

⁶ School of Arts, Peking University, Beijing 100080, China

⁷ Department of Computer Science, City University of Hong Kong, Hong Kong

* Correspondence: gaof@pku.edu.cn (F.G.); swma@pku.edu.cn (S.M.)

How To Cite: Wang, H.; Guo, Y.; Zhang, T.; et al. RiverEcho-2.0: A Real-Time Interactive System for Yellow River Culture via Enhanced MultiModal Document RAG. *Transactions on Artificial Intelligence* 2025, 1(1), 212–226. <https://doi.org/10.53941/tai.2025.100014>

Received: 28 August 2025

Revised: 8 September 2025

Accepted: 15 September 2025

Published: 22 September 2025

Abstract: The Yellow River culture is a cornerstone of Chinese civilization, embodying rich historical, social, and ecological significance. To conserve and promote this invaluable cultural heritage, we propose RiverEcho-2.0, a real-time interactive digital system designed to facilitate user engagement with Yellow River culture. As the foundation of our system, we curated and digitized a comprehensive collection of books and documents related to Yellow River heritage, constructing a dedicated multimodal corpus. To effectively leverage this corpus, we introduce a novel multi-modal Document Retrieval-Augmented Generation (RAG) framework that enhances document retrieval through context-aware image-text alignment and joint embedding. Experimental results demonstrate that our method achieves a large improvement over existing state-of-the-art multi-modal RAG baselines, leading to significant gains in downstream tasks.

Keywords: Yellow River culture; dataset construction; multi-modal document RAG

1. Introduction

Historical records from the Yellow River Basin document the millennia-long development of Chinese civilization, forming a solid foundation for the region's economic and cultural achievements [1]. However, the dissemination of Yellow River culture to the general public remains limited. One of the primary challenges lies in the difficulty of collecting and digitizing ancient materials, which are often fragmented, geographically dispersed, and stored in various non-digital formats. Furthermore, the complexity of historical carriers and the interpretive challenges of classical texts continue to hinder public accessibility and understanding. In addition to ancient literature, modern publications and materials related to the Yellow River also convey a wealth of knowledge through multimodal forms, offering inspiration across multiple domains. Therefore, it is urgent to leverage emerging technological advancements to overcome these barriers and promote the widespread preservation and dissemination of this cultural heritage.

Currently, Human-Computer Interaction (HCI) systems have been extensively implemented across various sectors including healthcare, education, and legal systems [2–4], demonstrating significant enhancements in both social productivity and operational efficiency while fundamentally transforming human-technology engagement. However, research and technological applications in traditional cultural preservation remain substantially underdeveloped [5], particularly regarding systematic conservation and innovative transmission approaches for the Yellow River culture. As a paramount cultural symbol of the Chinese nation, the preservation and transmission of Yellow River culture bear critical implications not merely for cultural continuity but more profoundly for sustaining national identity. In



this context, the progressive evolution of HCI technology presents promising potential to offer innovative methodological frameworks and technical pathways for safeguarding this invaluable traditional heritage.

This work presents RiverEcho-2.0, a real-time interactive intelligent system for ancient Yellow River culture, which integrates Automatic Speech Recognition (ASR), Multimodal Large Language Models (MLLM), Text-to-Speech (TTS), and Talking-head generation. The system is capable of recognizing users' spoken questions, generating culturally and historically informed responses, and dynamically presenting these answers through a talking-head virtual human. The entire system employs a streaming pipeline architecture, enabling low-latency responses and ensuring real-time interaction and response coherence.

To enable the system to accurately address historical and cultural inquiries raised by users and to effectively disseminate the historical and cultural knowledge related to the ancient Yellow River region, while also mitigating the risk of hallucinated content from the LLM, this work constructs a dedicated high-quality knowledge dataset for ancient Yellow River culture. Specifically, we collected over 100 ancient manuscripts from different Chinese dynasties and disciplinary domains related to the Yellow River, along with authoritative works authored by leading contemporary experts in Yellow River cultural studies. These materials were subjected to automated preprocessing and hierarchical annotation, followed by manual sampling verification and supplementary labeling conducted by history students. As a result, we obtained a curated text corpus containing over 20,000 fully annotated cultural and historical segments.

To fully leverage the capabilities of the proposed corpus, we integrated an Enhanced Multimodal Document Retrieval-Augmented Generation mechanism into the MLLM's inference pipeline. Within this framework, we introduce the Context-Aware Image-Text Matching and Embedding (CIME) module, which synthesizes visual content with its corresponding document context. For non-textual elements such as diagrams and formulas, we employ multimodal large language models to perform textual interpretation. These processed elements are subsequently integrated with textual data to construct a comprehensive knowledge graph. Our retrieval methodology employs a hybrid approach that combines graph-based and embedding-based retrieval techniques. The proposed system demonstrates enhanced retrieval precision on document-based VQA datasets, consequently improving the downstream MLLM's question-answering capabilities. When implemented in our framework, this approach significantly enhances the factual accuracy and cultural relevance of generated responses.

Meanwhile, considering Li Daoyuan's remarkable contributions to ancient waterway systems, we used his historical persona as an example to construct a digital human-driven system. This enhances the overall visual appeal of the framework while also stimulating user engagement and meeting their visual expectations.

Overall, our contributions are summarized as follows:

- We developed RiverEcho-2.0, a digital system capable of real-time interaction with users, which is driven by digital human images and can provide real-time and accurate answers to user-inputted questions, especially those related to the Yellow River.
- We have established a specialized knowledge corpus focusing on Yellow River culture, collecting historical documents and contemporary publications under the guidance of historical experts. This initiative has significantly advanced the digitization of Yellow River-related literature while establishing a foundational resource for addressing pertinent cultural inquiries, thereby facilitating the preservation and transmission of this cultural heritage.
- We propose a novel multimodal document RAG approach that leverages a Context-aware Image-Text Matching and Embedding (CIME) module to align and fuse images with their corresponding textual context within documents. To handle non-textual elements such as charts and mathematical formulas, we employ a multimodal large language model to convert them into textual representations. These outputs are then integrated with textual data to construct a unified knowledge graph. For the retrieval stage, we introduce an innovative hybrid retrieval strategy that combines graph-based and embedding-based methods. Experiments on document VQA datasets demonstrate improved retrieval efficiency and question-answering performance. When deployed in our system, this technique significantly enhances the factual accuracy and cultural relevance of the generated responses.

Building upon our previous work (RiverEcho) presented at the ICME AIART workshop, RiverEcho-2.0 introduces a Multimodal Document RAG framework that significantly extends the original text-based retrieval approach to incorporate the rich visual and tabular information found in contemporary Yellow River literature. These architectural improvements enable more comprehensive responses to user queries about Yellow River culture by effectively leveraging multimodal document contents. Additionally, we have expanded the experimental evaluation in response to reviewer feedback from our prior publication, providing further empirical validation of the system's capabilities in supporting cultural preservation efforts through advanced human-computer interaction technology.

2. Related Works

2.1. Large Language Model

Nowadays, Large Language Models (LLMs) [6–10] are emerging rapidly. They have demonstrated remarkable capabilities across a wide range of natural language processing (NLP) tasks. Through instruction tuning and RAG, contemporary LLMs have achieved not only impressive general intelligence but also notable expertise in specialized domains [11–14]. This has spurred growing interest in the development of domain-specific LLMs, a rapidly evolving area of research that continues to drive innovation and exploration within the field.

The rapid development of LLMs has contributed to some extent to the dissemination and preservation of Chinese traditional culture. Recent research has focused on Classical Chinese Understanding (CCU), aiming to enhance the comprehension and generation of classical texts. Early CCU systems were primarily designed for specific tasks, such as translation [15, 16], punctuation restoration [17], and named entity recognition (NER) [18, 19]. Recent advancements, such as GujiBERT [20], have utilized large-scale unlabeled classical Chinese corpora for masked pre-training, providing task-specific models with embeddings enriched with classical Chinese knowledge. Similarly, SikuGPT [21] has demonstrated the potential of generative pre-training for classical poetry and prose creation. Additionally, models like Bloom-7b-Chunhua [22] and Xunzi-Qwen-7B-CHAT [23] have combined open-source base models with extensive classical Chinese corpora, offering preliminary insights into the capabilities of LLMs in understanding and generating classical Chinese texts. Notably, TongGu [24], in the same work, proposed a two-stage fine-tuning approach, enabling the model to perform multiple tasks such as classical text reading comprehension, classical-to-modern Chinese translation, and classical poetry generation. These developments highlight the growing potential of LLMs in the preservation and inheritance of cultural heritage.

Nevertheless, integrating ancient Yellow River culture with LLMs is far from straightforward. The main challenge is the absence of a structured dataset specifically built around Yellow River classics and scholarly texts, making it difficult for LLMs to accurately capture and interpret the rich historical content embedded in this tradition. This gap underscores the necessity of developing dedicated datasets and tailored methods to facilitate a deeper understanding of this vital component of Chinese cultural heritage.

2.2. Multimodal Document RAG

LLMs suffer from hallucinations and outdated knowledge due to their reliance on static training data [25–27]. RAG mitigates these issues by dynamically integrating external knowledge bases, thereby enhancing factual grounding in generated outputs. Furthermore, RAG's flexible data repository enables efficient knowledge updates, accommodates long-tail knowledge, and reduces privacy leakage risks. Additionally, RAG optimizes computational efficiency through model size reduction [28], long-context handling [29], and eliminating redundant generation steps [30].

In practical implementations, RAG systems typically adopt a two-stage retrieve-then-generate workflow [31]. First, the retriever component utilizes dense embedding models [32, 33] to identify relevant documents from external knowledge sources, often enhanced by cross-encoder re-rankers [34]. The generator then conditions on these retrieved passages to produce more accurate and contextually grounded outputs. Recent innovations have introduced advanced techniques like retrieval planning [35], agent-based retrieval [36], and iterative retrieval-generation [37] to further optimize system performance.

Recent research has increasingly focused on Document RAG [38–40], with particular emphasis on multimodal document RAG due to the heterogeneous nature of modern documents containing images, tables, formulas, and other non-textual elements [41–43]. As shown in Figure 1, current multimodal document RAG architectures primarily adopt one of two paradigms:

1. Text-Centric Pipeline :

- Offline Processing : Multimodal content undergoes text extraction via OCR systems [44] or specialized parsers [45], followed by semantic encoding using text embedding models.
- Online Retrieval: User queries retrieve relevant text chunks via maximum inner product search (MIPS), which are then processed by downstream LLMs.

2. Vision-Integrated Pipeline:

- Offline Processing: Document pages are processed as images and encoded into unified representations using vision-language models.
- Online Retrieval: Query-relevant page embeddings are fed to multimodal LLMs for joint text-visual reasoning.

The first methodology exhibits a significant limitation: loss of semantic information inherent in multimodal data, particularly for non-textual images and semantically rich graphics such as node diagrams or flowcharts. The second approach suffers from cross-page information fragmentation. When processing long text passages spanning multiple pages or multimodal content (e.g., semantically related text and images distributed across pages), page-level embeddings generate fragmented representations that impair cross-page retrieval. This subsequently degrades downstream MLLMs' answer quality. Attempting to concatenate page embeddings introduces semantic entanglement, further compromising retrieval effectiveness while imposing substantial memory footprint burdens on the system.

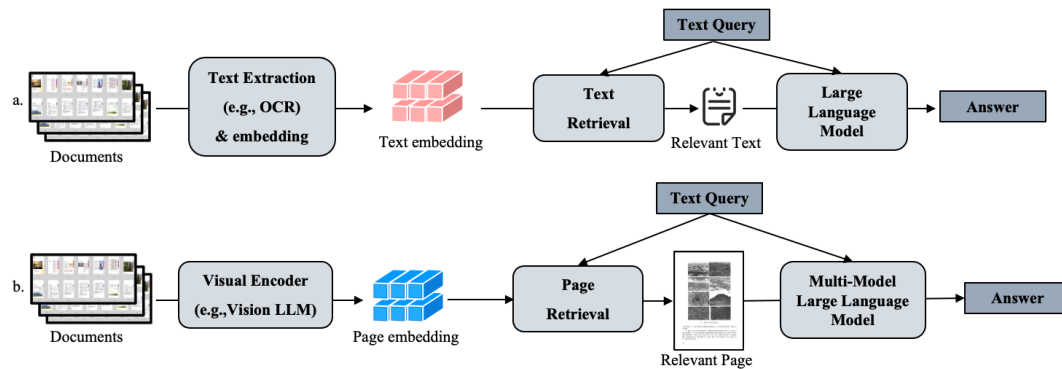


Figure 1. Architecture of text-centric and vision-integrated multimodal document RAG pipelines. Subfigure a illustrates the text-centric approach relying on textual extraction and encoding, while subfigure b demonstrates the vision-integrated method that preserves original document layout and non-textual elements through page-level embeddings.

2.3. Real-Time Interactive Digital Humans for Audio-Visual Dialogue

Recent advancements in real-time interactive digital human technology have enabled synchronous audio and video dialogue.

LiveTalking [46] introduces a framework for audio-video synchronization, supporting models like wav2lip [47] and musetalk [48] to address initial inference delays. Metahuman-stream [49] offers an open-source solution with voice cloning, speech interruption, and full-body video stitching, compatible with RTMP [50] and WebRTC [51] protocols. Synthesia [52] provides a platform for creating digital avatars with lifelike narration in over 140 languages, reducing video production time significantly. VTube Studio [53] enables real-time avatar control using face tracking, enhancing interactive experiences through open-source integration. These open-source technical frameworks lay a solid foundation for the development of domain-specific applications.

3. Methodology

To enable the system to deliver more specialized and information-rich responses pertaining to Yellow River culture, we constructed a Yellow River Cultural Corpus. To facilitate the effective utilization of the proposed corpus, we designed a multimodal document RAG approach to enhance the question-answering capabilities of downstream MLLMs. Furthermore, we developed a real-time interactive digital system that integrates the aforementioned corpus and RAG method, along with several open-source modules. This system is capable of recognizing user speech and providing professional responses in real time through a digital human avatar.

3.1. Yellow River Cultural Corpus

Book Collection: The historical documents related to the Ancient Yellow River culture have been preserved in various forms, such as manuscripts, bamboo slips, and inscriptions, with their textual content recorded in Classical Chinese, Tangut script, and other writing systems. The diversity of these media, along with the challenges in interpreting historical scripts, has created a significant barrier to public understanding of the Ancient Yellow River culture.

To facilitate the widespread dissemination of this cultural heritage, we have collected a diverse set of historical texts based on recommendations from scholars specializing in Chinese history. The proposed dataset includes ancient texts from various historical periods (Pre-Qin, Han, Wei-Jin and Northern and Southern Dynasties, Tang-Song, and Ming-Qing periods), primarily consisting of modern annotated editions. Additionally, it encompasses over a hundred historical and contemporary works related to the Yellow River, covering multiple themes: river

governance, technology and engineering, natural knowledge, socio-economic aspects, cultural heritage, historical narratives, disasters and their impacts, and interdisciplinary topics. Distribution of themes in the proposed dataset is shown in Table 1.

Data Structuring and Processing Pipeline: As shown in Figure 2, the construction of the dataset consists of three steps. First, the collected large-scale ancient books are preprocessed using vertical text OCR to obtain unstructured text. Next, an LLM is utilized to structure the unstructured text. Finally, the structured data undergoes a proofreading process to ensure accuracy. We elaborate on the last two steps in detail.

Unstructured Data Structuring: To improve the retrieval efficiency and enhancement capability of the LLM when processing the dataset, we performed structured processing on the collected Yellow River historical texts. We first divided the collected unstructured text into multiple unstructured chunks. Then, using a large LLM along with a corresponding structuring template, we converted these unstructured chunks into structured data, performed entity deduplication, and constructed a knowledge graph. The structured chunks consist of three key components:

- **Basic information:** Includes the original text, translation, and summary, as well as the corresponding book title and page number.
- **Entities:** Refer to the named entities and their types that appear within the paragraph.
- **Relations:** Represent the connections between different entities.

Data Proofreading: To enhance the professionalism of responses to Yellow River-related inquiries, we incorporated a human proofreading mechanism. Specifically, we invited professional reviewers to conduct randomized sampling and verification of the structured data. They were requested to identify, assess, and annotate instances of hallucination, including incorrect translations, overgeneralizations, and excessive information supplementation. To enhance the reliability and consistency of the verification process, each flagged hallucination case was subjected to a two-stage human review pipeline. In the first stage, a student reviewer annotated the detected errors and categorized them based on predefined error types. In the second stage, another reviewer reassessed the annotations, verified the discrepancies, and finalized the manual corrections.

The processed knowledge graph supports downstream graph-based retrieval, while the text chunks facilitate embedding-based retrieval in subsequent stages.

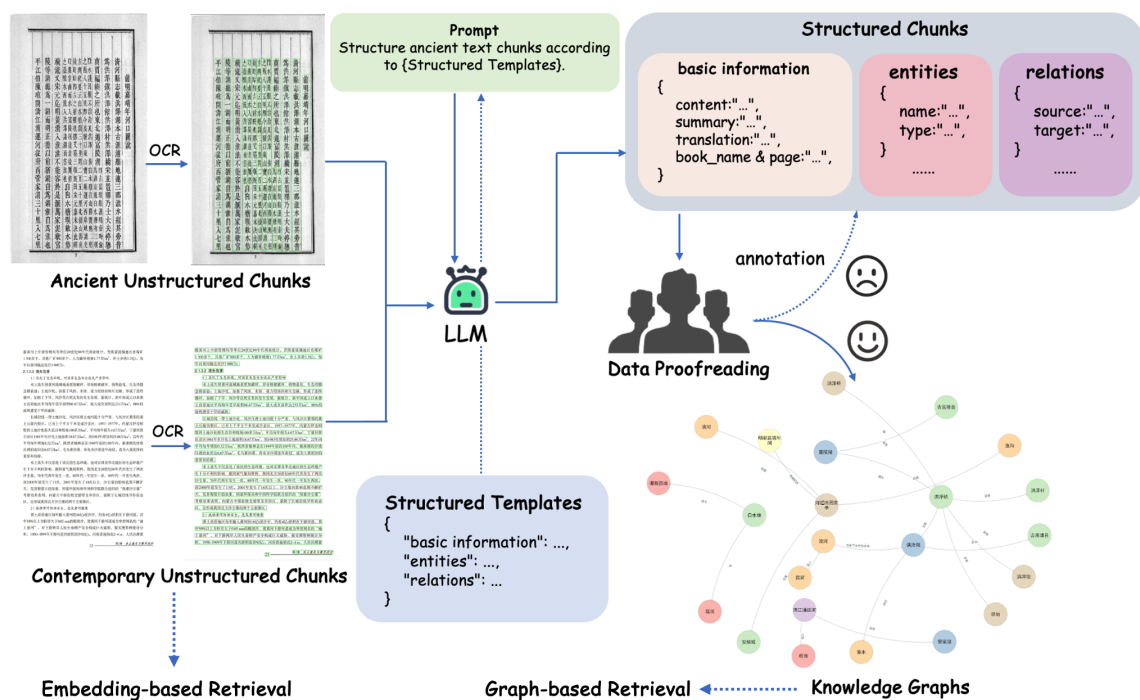


Figure 2. Overall process of corpus construction. First, collected ancient and contemporary books are processed into unstructured chunks. These chunks, along with structured-template prompts, are fed into an LLM to produce structured chunks. Next, the structured chunks undergo manual proofreading. Those failing quality standards are annotated and returned for re-proofreading. Finally, the verified structured chunks are converted into knowledge graphs and stored to support downstream graph-based retrieval, while the unstructured chunks serve downstream embedding-based retrieval.

Table 1. Distribution of Themes in Proposed Dataset.

Theme	Number of Trunks
Total	20408
River governance	6125
Technology and engineering	4369
Natural knowledge	2552
Socio-economic aspects	1649
Cultural heritage	1778
Historical narratives	1551
Disasters and their impacts	1268
Interdisciplinary topics	1116

3.2. Enhanced Multi-Modal Document RAG

In our previous conference paper, we merely employed an open-source RAG approach [54], which is capable of retrieving only textual information. Given that Yellow River cultural documents contain abundant valuable multimodal content—such as images, charts, and diagrams—we propose an Enhanced Multimodal Document RAG framework to better exploit these resources.

3.2.1. Overall Process

As illustrated in Figure 3, our method consists of three main stages. First, documents are processed using a PDF parsing tool to extract structural elements. Second, these elements are fed into the Context-aware Image-Text Matching and Embedding (CIME) module to generate multimodal fused embeddings. Concurrently, non-textual elements such as charts, figures, and mathematical formulas are converted into descriptive text using a large language model, and a knowledge graph is constructed to capture semantic relationships among the extracted content. Third, given a user query, a hybrid retrieval strategy is performed, combining both embedding-based similarity search and graph-based retrieval. Finally, the retrieved multimodal information is delivered to a downstream multimodal large language model for response generation.

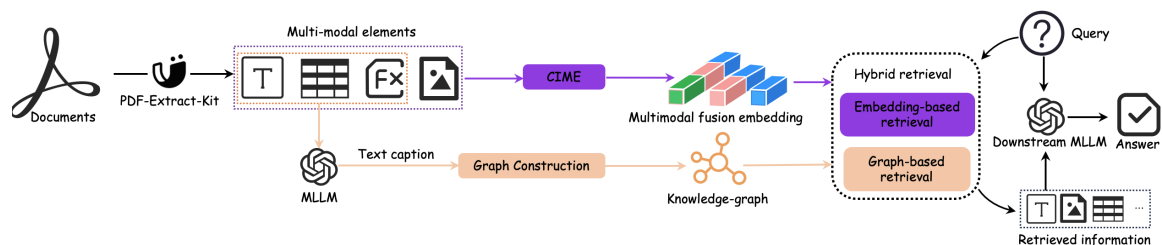


Figure 3. Architecture of the enhanced multimodal document RAG method. Documents are first decomposed into multimodal elements, which are then processed via CIME and graph construction into multimodal fusion embeddings and a knowledge graph. Finally, hybrid retrieval is performed, and the retrieved information is fed into the downstream MLLM.

3.2.2. Context-Aware Image-Text Matching and Embedding Module

To enhance the efficiency of offline embedding conversion and retrieval for multimodal documents, we propose CIME. As depicted in Figure 4, documents are parsed into four modalities: images, paragraphs, tables, and formulas. Tables and formulas are processed as paragraphs after being converted into text indices with predefined prompts.

During the training phase, a visual encoder transforms images into global and local features, while a text encoder converts queries and paragraphs into corresponding features.

Layout information is also encoded and fused with the text features. We feed query features (Q), local image features (K), and text features (V) into a query-aware fusion gate module to obtain fused features. A matching loss is then calculated between these fused features and the query features. Concurrently, we use contrastive learning between the image global features and text features to ensure semantic alignment.

The query-aware architecture is illustrated in Figure 5. First, the input features are concatenated and fed into an MLP-based gate to produce a gate score, which serves to measure the relevance between the query and the image

and text modalities, respectively. Subsequently, the query feature is jointly attended with the local image feature and text feature in one attention block, while in another attention block, the query feature is attended with the text feature only. Finally, the outputs from the two attention blocks are combined with the gate score through a weighted fusion mechanism to obtain the final fused feature.

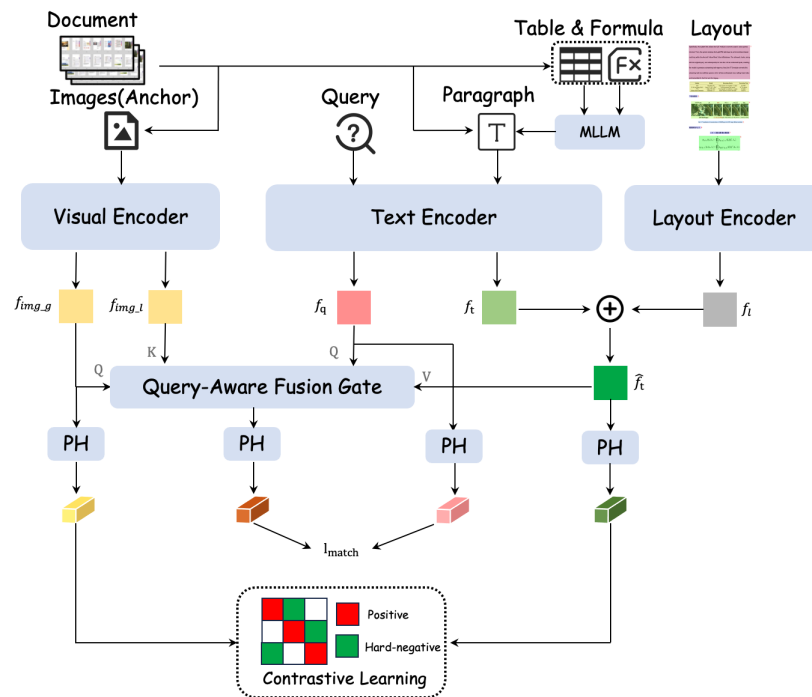


Figure 4. Architecture of the CIME module. Documents are decomposed into multimodal elements and transformed into features via encoders. These features are then fused through a query-aware fusion gate, where PH denotes the projection head.

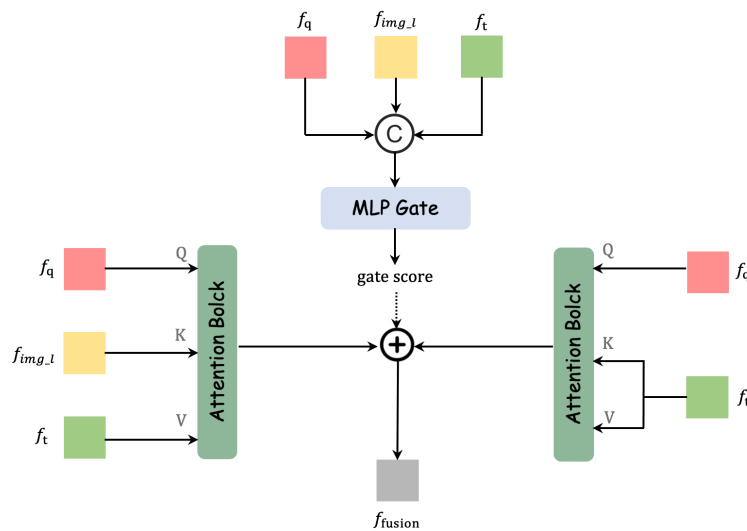


Figure 5. Architecture of the query-aware fusion gate.

For offline inference, the query features are replaced with image global features since user queries are absent. This substitution does not introduce unacceptable cross-modal discrepancies because the semantic alignment between the image global and text features was established during the training phase.

As depicted in Figure 6, the training data for our CIME module is an adapted version of the existing DocVQA dataset. We begin by using a PDF parsing tool to decompose the documents in the dataset into their layout and content. Next, we employ MLLMs with templated instructions for annotation. The final step involves a small-scale manual refinement to obtain positive and negative sample pairs, with images serving as the anchor.

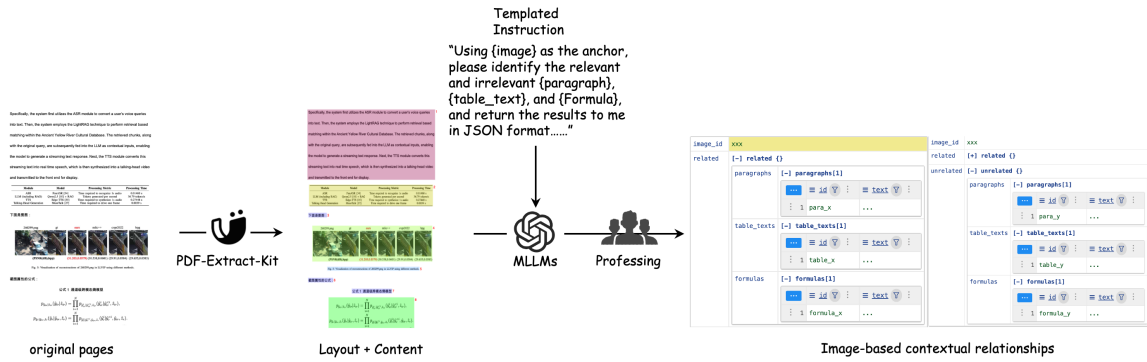


Figure 6. CIME training data processing pipeline.

Our CIME module is trained with a multi-task loss that jointly optimizes cross-modal alignment, query-aware fusion, and explicit gate supervision. The overall objective is formulated as:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{align}} + \beta \mathcal{L}_{\text{match}} + \delta \mathcal{L}_{\text{gate}}, \tag{1}$$

where α , β , and δ are balancing coefficients for the three components.

The image-text alignment loss ($\mathcal{L}_{\text{align}}$) ensures that the global representations of semantically related image and text pairs are aligned in the embedding space:

$$\mathcal{L}_{\text{align}} = -\log \frac{\exp\left(\frac{\text{sim}(f_{\text{img}_g}^{\text{emb}}, f_t^{\text{emb}})}{\tau}\right)}{\sum_{k=1}^K \exp\left(\frac{\text{sim}(f_{\text{img}_g}^{\text{emb}}, f_t^{(k)\text{emb}})}{\tau}\right)}, \tag{2}$$

where $f_{\text{img}_g}^{\text{emb}}$ and f_t^{emb} are the projected global image and text embeddings, $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, $\tau = 0.07$ is a temperature hyperparameter, and K is the batch size used for in-batch negative sampling.

The query-fused matching loss ($\mathcal{L}_{\text{match}}$) drives the fused representation f_{fused} to be semantically close to the input query:

$$\mathcal{L}_{\text{match}} = -\log \frac{\exp\left(\frac{\text{sim}(q^{\text{emb}}, f_{\text{fused}})}{\tau}\right)}{\sum_{k=1}^K \exp\left(\frac{\text{sim}(q^{\text{emb}}, f_{\text{fused}}^{(k)})}{\tau}\right)}, \tag{3}$$

where q^{emb} is the projected query embedding, and $f_{\text{fused}} = g \cdot f_{\text{vision}} + (1 - g) \cdot f_{\text{text}}$ is the gate-controlled fusion result, with f_{vision} and f_{text} computed via cross-attention over image patches and text tokens, respectively.

Finally, the gate supervision loss ($\mathcal{L}_{\text{gate}}$) explicitly guides the model to decide whether a query is image-relevant:

$$\mathcal{L}_{\text{gate}} = \text{BCE}(g, y_{\text{type}}),$$

$$g = \sigma\left(\text{MLP}_{\text{gate}}\left([q^{\text{emb}}, f_{\text{img}_g}^{\text{emb}}, f_t^{\text{emb}}]\right)\right), \tag{4}$$

where $g \in [0, 1]$ is the gate score, $y_{\text{type}} \in \{0, 1\}$ is the binary label indicating image relevance, and σ is the sigmoid function. BCE represents Binary Cross-Entropy Loss. This loss enables the model to adaptively attend to visual or textual content based on query semantics.

3.3. Real-Time Interactive Digital System

As illustrated in Figure 7, our real-time interactive system integrates multiple key modules, including ASR, MLLM, TTS, and talking-head generation.

Specifically, the system first utilizes the ASR module to convert a user’s voice queries into text. Then, the system employs the proposed multimodal document RAG technique to perform retrieval-based matching within the Ancient Yellow River Cultural corpus. The retrieved elements, along with the original query, are subsequently fed into the MLLM as contextual inputs, enabling the model to generate a streaming text response. Next, the TTS module converts this streaming text into real-time speech, which is then synthesized into a talking-head video and transmitted to the front end for display.

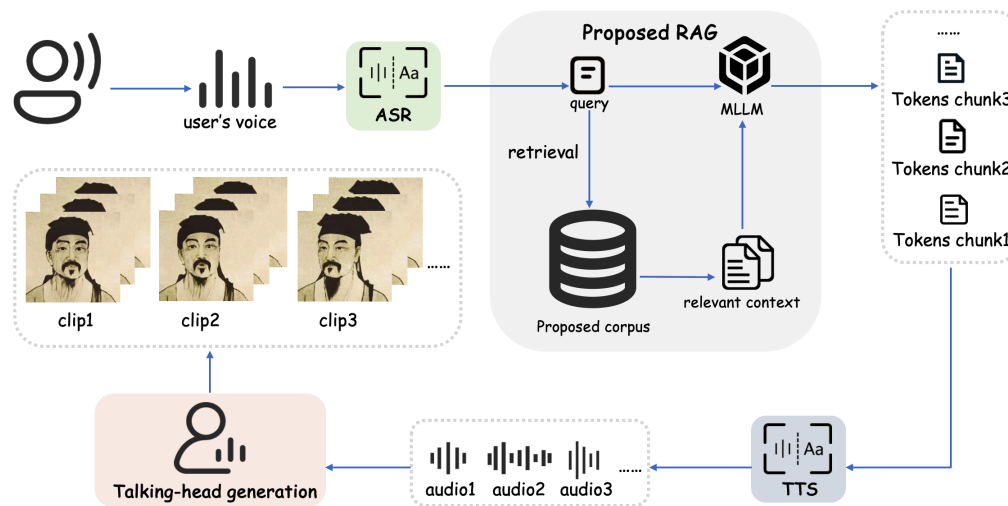


Figure 7. Overall architecture of RiverEcho-2.0. The user’s voice queries is first converted into a textual query by the ASR module. Using proposed multimodal document rag, relevant context is retrieved from the constructed corpus and combined with the query as input to the LLM, which then generates token chunks in a streaming manner. The TTS module receives these token chunks and synthesizes corresponding audio segments in real-time, which are then fed into the talking-head generation module to produce video clips for presentation.

Since the latency of existing ASR tools is generally negligible, we do not apply additional streaming processing to it. However, all other modules in the system operate in a streaming fashion, except for the MLLM. Given that both the inputs and outputs of all modules are streamed, the system follows a fully integrated streaming workflow. In other words, once the ASR module transcribes the user’s speech, all subsequent modules function in parallel. This design significantly reduces response latency, particularly when generating long sequences of tokens, ensuring a highly efficient real-time user interaction experience.

Specifically, we adopt Livetalking [46], an open-source real-time interactive digital human platform, as the basic framework of our system. For each module, we adopt the following configurations:

- **ASR** : We use FunASR [55], a high-performance open-source ASR model, capable of recognizing user input speech with both high speed and accuracy.
- **MLLM**: We employ Qwen2-VL [10] as the base model and integrate proposed RAG method for corpus retrieval and enhanced inference, optimizing the system’s performance in information retrieval tasks. The tokens generated by the MLLM will accumulate into chunks until encounter punctuation marks that indicate the end of a sentence, such as a period, exclamation mark, or ellipsis.
- **TTS**: We use Edge-TTS [56], an open-source, powerful few-shot speech synthesis model capable of generating high-quality speech with limited data. The sample rate is configured at 16,000 Hz, the voice is selected as “zh-CN-YunjianNeural”, and the synthesis speed is adjusted to 20% slower than the default.
- **Talking-Head Generation Module**: We implemented the training and synthesis of ancient digital characters using the MuseTalk-2.0 [48] framework, an audio-driven lip-synchronization system designed for real-time high-fidelity facial animation. Specifically, we gathered dozens of portraits and statues of Li Daoyuan from the Internet and extracted their facial features. Leveraging generative models, we synthesized a series of modernized character images. We then refined these images by manually selecting and processing their micro-expressions to construct a compact character dataset. Finally, utilizing MuseTalk-2.0, we developed and animated digital humans based on this dataset.

Our system seamlessly integrates these modules to enable efficient streaming processing, delivering a smooth and responsive low-latency interactive experience for users. Optimizing the coordination between components, ensures real-time performance with minimal delays, enhancing overall usability.

4. Experiments

4.1. Implementation Details

4.1.1. Hardware Configuration

The entire system was deployed and executed on a server equipped with five NVIDIA A800 GPUs. All quantitative experiments were conducted on this hardware configuration.

4.1.2. Data Engineering Pipeline

For the proposed Enhanced Multimodal Document RAG Framework, we utilized the MP-DocVQA dataset for data processing. Initial coarse-grained image-text matching was performed using Qwen-VL-MAX, followed by manual screening of 3600 PDF documents.

4.1.3. CIME Module Training

The selected 3600 PDF documents were partitioned into training, validation, and test sets with a 2:1:1 ratio for training the Cross-modal Information Matching and Enhancement (CIME) module. The model underwent extensive training for 3000 epochs to ensure convergence.

4.1.4. Retrieval System Implementation

For the graph-based retrieval component of our hybrid search system, we adopted the same processing methodology as LightRAG. All PDF parsing operations were consistently performed using the MinerU toolkit to maintain processing uniformity.

4.2. Evaluation of Enhanced Multimodal Document RAG Framework

4.2.1. MM-DocVQA Benchmark Performance

To evaluate the performance of the proposed Enhanced Multimodal Document RAG Framework, we conducted comprehensive experiments on the MMLongBench-Doc dataset. As shown in Table 2, we tested both text-based and multimodal pipelines, comparing native mllms with other RAG-enabled pipelines. Our method achieves significant improvements in both F1 score and answer accuracy over existing approaches. Notably, the framework delivers particularly substantial gains on image-related (IMG) and multimodal (MUL) questions, validating the effectiveness of our enhanced multimodal document retrieval and generation approach.

We also applied the proposed RAG method to both close-sourced and open-sourced large language models. As shown in Table 3, the proposed approach yields substantial improvements in answer accuracy across different models, further demonstrating the effectiveness of our method.

Meanwhile, disregarding the downstream large language model, we directly compared our proposed method with other RAG approaches. As shown in Table 4, in terms of page-level retrieval performance, our model outperforms existing text-based, graph-based, and multimodal RAG methods.

Table 2. Closed-domain DocVQA evaluation results on MMLongBench-Doc. We report generalized accuracy (ACC) and F1 score over five evidence source modalities: Text (TXT), Layout (LAY), Chart (CHA), Table (TAB), and Image (IMG), as well as three evidence locations: Single-page (SIN), Cross-page (MUL), and Unanswerable (UNA).

Method	# Pages	TXT	LAY	CHA	TAB	IMG	SIN	MUL	UNA	ACC	F1
Text Pipeline											
ChatGLM-128k [57]	-	23.4	12.7	9.7	10.2	12.2	18.8	11.5	18.1	16.3	14.9
Mistral-Instruct-v0.2 [58]	-	19.9	13.4	10.2	10.1	11.0	16.9	11.3	24.1	16.4	13.8
ColBERT v2 [59] + Llama 3.1 [60]	1	20.1	14.8	12.7	17.4	7.4	21.8	7.8	41.3	21.0	16.1
ColBERT v2 [59] + Llama 3.1 [60]	4	23.7	17.7	14.9	24.0	11.9	25.7	12.2	38.1	23.5	19.7
Multimodal Pipeline											
DeepSeek-VL-Chat [61]	-	7.2	6.5	1.6	5.2	7.6	5.2	7.0	12.8	7.4	5.4
Idefics2 [62]	-	9.0	10.6	4.8	4.1	8.7	7.7	7.2	5.0	7.0	6.8
MiniCPM-Llama3-V2.5 [63]	-	11.9	10.8	5.1	5.9	12.2	9.5	9.5	4.5	8.5	8.6
InternLM-XC2-4KHD [64]	-	9.9	14.3	7.7	6.3	13.0	12.6	7.6	9.6	10.3	9.8
mPLUG-DocOwl 1.5 [65]	-	8.2	8.4	2.0	3.4	9.9	7.4	6.4	6.2	6.9	6.3
Qwen-VL-Chat [66]	-	5.5	9.0	5.4	2.2	6.9	5.2	7.1	6.2	6.1	5.4
Monkey-Chat [67]	-	6.8	7.2	3.6	6.7	9.4	6.6	6.2	6.2	6.2	5.6
ColPali [42] + Idefics2 [62]	1	10.9	11.1	6.0	7.7	15.7	15.4	7.2	8.1	11.2	11.0
ColPali [42] + Qwen2-VL 7B [66]	1	25.7	21.0	18.5	16.4	19.7	30.4	10.6	5.8	18.8	20.1
ColPali [42] + Qwen2-VL 7B [66]	4	30.0	23.5	18.9	20.1	20.8	32.4	14.8	5.8	21.0	22.6
Ours + Qwen2-VL 7B [66]	-	31.1	24.0	19.3	24.6	24.6	31.7	20.2	5.8	25.8	27.2

Table 3. Multimodal RAG Performance Across MLLMs (ACC)

Model Type	Model Name	Baseline	+Ours	Δ (%)
Close-sourced	GPT-4V	28.4	41.7	+13.3 (46.8%)
	Gemini Pro 2.0	31.2	45.8	+14.6 (46.8%)
Open-sourced	LLaVA-7B	12.3	25.1	+12.8 (104.1%)
	Qwen-VL-7B	14.6	25.8	+11.2 (76.7%)
Overall	Average	21.6	34.6	+13.0 (60.1%)

Table 4. Retrieval Performance of RAG Methods on MP-DocVQA.

Method	Page Retrieval	
	R@1 (%)	R@4 (%)
Langchain-Chatchat	38.5	62.1
FlashRAG	42.3	65.8
GraphRAG	45.7	68.4
KAG	40.2	64.0
ColBERT v2	53.6	73.9
M3DocRAG	58.2	81.1
Ours	67.4	88.7

4.2.2. Ablation Study

We performed rigorous ablation experiments to verify the contribution of each proposed component, specifically examining proposed CIME module and hybrid search strategy.

Table 5 reveals that replacing the CIME module with conventional OCR+text embedding leads to an 40.1% reduction in recall (F1 score). Similarly, when substituting our hybrid search with either pure embedding-based or pure graph-based search, the overall F1 score decreases by 13.2% and 32.7% respectively. These results conclusively demonstrate the necessity of both proposed enhancements for optimal multimodal document retrieval.

Table 5. Ablation Study of Key Components.

Method	F1	Relative Change
Proposed System	27.2	-
w/OCR+text embedding	16.3	−40.1%
Embedding-based search only	23.6	−13.2%
Graph-based search only	18.3	−32.7%

Furthermore, to establish our framework’s model-agnostic nature, we evaluated its performance across multiple MLLMs using identical experimental settings. Table 3 shows consistent accuracy improvements in document VQA tasks for both proprietary and open-source models when integrated with our multimodal RAG system.

4.2.3. Qualitative Examples

Our multimodal RAG approach was evaluated on public datasets and our proposed corpus. As shown in Figure 8, it effectively retrieves relevant tables, texts, and images on the M3DocVQA benchmark. Figure 9 shows accurate retrieval of texts and tables from our corpus, yielding more precise answers than baselines. Additionally, Figure 10 demonstrates the model’s ability to handle cultural questions.

Question:
SIE Bend Studio's 2019 game cover has man leaning on what?

Retrieved elements:

Retrieved Table

Bend Studio	
Logo used since	2022
Formerly	Blank, Berlin & Co., Inc. (1992–1995) Eidetic, Inc. (1995–2000)
Company type	Subsidiary
Industry	Video games
Founded	1992, 33 years ago
Founders	Marc Blank Michael Berlin
Headquarters	Bend, Oregon, US
Key people	Christopher Reese (studio director)
Products	Bubay 3D Siphon Filter Days Gone
Number of employees	150 ^[1] (2022)
Parent	Sony Computer Entertainment (2000–2005) PlayStation Studios (2005–present)
Website	bendstudio.com ^l

Retrieved Text

Days Gone is a 2019 action-adventure video game developed by Bend Studio and published by Sony Interactive Entertainment. The game was released for the PlayStation 4 in April 2019, and Windows in May 2021. A remastered version for PlayStation 5 was released in April 2025 alongside a DLC version for Windows.

Retrieved Image

Ours:
Motorcycle.

Figure 8. An example of multimodal retrieval. Our method retrieves and integrates relevant tables, texts, and images from the M3DocVQA dataset.

Question:
描述一种利用农业废弃物的生态肉牛循环经济模式，列出其关键组成部分、步骤及经济效益。
What Describe the key components and economic benefits of an ecological cattle farming model that utilizes agricultural waste through a circular economy approach. List the steps and final outputs.?

Retrieved elements:

Retrieved text

畜牧业突出高效生态。高青县以“肉牛、肉鸭、奶牛”完善产业化经营为重点，以突出规模、提升档次为突破口，大力加强现代化畜牧业建设，全力实现畜牧大县向畜牧强县转变。畜禽程序化免疫、人工授精、青贮氨化、标准化生产、自然发酵等技术得到大面积推广应用，胚胎移植、环境控制、免疫监测、高效综合饲养管理等高新技术在试验示范的基础上正在由点到面逐步推广。2011年，全县肉牛存栏 2.8 万头，出栏 5.58 万头；奶牛存栏 3.2 万头；生猪出栏 18.5 万头；家禽出栏 1.980 万只。肉、蛋、奶总产量达到 18.5 万吨，实现畜牧业产值 20.5 亿元。预计到 2015 年，全县存栏肉牛 35 万头，其中奶牛 15 万头，年出栏 10 万头，认证 2 个无公害肉牛品牌，奶牛存栏 4.5 万头，生猪年出栏 30 万头，家禽存栏 1.700 万只（其中肉鸭 1.500 万只、肉鸡 100 万只、蛋鸡 100 万只），出栏 10 400 万只（其中肉鸭 10 000 万只、肉鸡 400 万只），肉、蛋、奶总产量 42 万吨，实现畜牧业产值 40 亿元（参见专栏：大地肉牛高效生态循环产业链）。

Retrieved Table

Ours:
该循环经济模式利用玉米秸秆、菌渣等农业废弃物，通过青贮发酵和沼气工程转化为肉牛养殖资源，同步生产有机肥并通过CDM创收。至2015年实现肉蛋奶产量42万吨、产值40亿元，并成功认证高端有机肉牛品牌。
This circular economy model utilizes agricultural waste like corn stalks and mushroom residue through fermentation and biogas engineering to support cattle farming, producing organic fertilizers and generating additional revenue via CDM. By 2015, it achieved an annual output of 420,000 tons of meat eggs and milk with 4 billion CNY revenue, while certifying premium organic beef brands

Figure 9. Another example of multimodal retrieval. Our method retrieves and integrates relevant tables and texts from the proposed corpus.

Question:
宋代黄河治理的困境有哪些？
What were the difficulties in managing the Yellow River during the Song Dynasty?

Qwen-2.5-max:
..... 泥沙淤积严重:黄河携带大量泥沙, 流经黄土高原后, 泥沙在下游沉积, 使河道逐渐淤塞, 增加了治理难度。.....
... Severe sediment accumulation: The Yellow River carries a large amount of sediment, and after flowing through the Loess Plateau, the sediment settles in the lower reaches, gradually silting up the river channel and increasing the difficulty of management ...

Deepseek-v3:
..... 泥沙淤积: 黄河含沙量高, 泥沙沉积导致河床抬高, 河道频繁改道, 增加了治理难度。.....
... Sediment accumulation: The Yellow River has a high sediment content, and the deposition of sediment elevates the riverbed, causing frequent changes in the river's course and increasing the difficulty of management. ...

ChatGPT-4o:
..... 由于黄河泥沙淤积严重, 大量泥沙沉积在河床之上, 导致河床不断抬高, 使得河水流速减缓, 水位逐渐升高, 最终形成高出周围地面的“地上悬河”.....
... Due to severe sediment accumulation in the Yellow River, a large amount of sediment settles on the riverbed, causing the riverbed to continuously rise. This slows the river's flow velocity and gradually elevates the water level, ultimately forming an "elevated river above ground" that is higher than the surrounding terrain. ...

Ours:
..... 泥沙淤积严重:黄河携带大量泥沙, 导致下游河道淤积严重。欧阳修指出, 黄河“本泥沙, 无不淤之理”, 下游淤积后, 水流受阻, 最终在上游低洼处决口。宋代黄河也曾多次决口, 如庆历八年(1048年)的商胡决口, 以及横陇河道的淤积, 导致水流不畅, 最终在上游决口。这种淤积现象使得黄河河道频繁改道, 治理难度极大。.....
... Severe sediment accumulation: The Yellow River carries a large amount of sediment, leading to serious siltation in its lower reaches. Ouyang Xiu pointed out that the Yellow River, "being full of sediment, has no reason not to silt up." After siltation occurs in the lower reaches, the water flow is obstructed, ultimately causing breaches in the low-lying areas upstream. During the Song Dynasty, the Yellow River experienced multiple breaches, such as the Shanghu breach in the eighth year of the Qingli era (1048) and the siltation of the Henglong river channel, which impeded water flow and eventually led to breaches upstream. This siltation phenomenon caused the Yellow River to frequently change its course, making it extremely difficult to manage. ...

Figure 10. An example related to the history and culture of the Yellow River: our model can cite historical allusions compared to other models, thanks to the Yellow River corpus we proposed.

4.3. Response Time Analysis

To validate the real-time performance of our system, we measured both the latency of individual modules and the end-to-end system response latency. Table 6 presents the latency of each module in our system. Specifically, the processing latency of the ASR module is negligible; the MLLM module, including the RAG process, can generate nearly 31 tokens per second; the TTS module requires only 0.27448 seconds to generate 1 second of audio; and the final Talking-Head Generation module achieves an audio-driven speed of 25 fps.

Crucially, to better evaluate the user's real-time interactive experience, we further measure the end-to-end system latency, defined as the time from when a user finishes speaking to when the generated talking-head video begins playback. Under streaming input and output, the overall system achieves an average response rate of approximately 28.36 tokens per second. This demonstrates that our system maintains high responsiveness despite the integration of multiple components, and remains well within the requirements for real-time interaction. Thanks to the streaming design of all modules, the system delivers smooth and low-latency responses, providing strong empirical support for its real-time capabilities.

Table 6. Processing Time of Each Module in the System

Module	Model	Processing Metric	Processing Time
ASR	FunASR [55]	Time required to recognize 1s audio	0.01460 s
MLLM (including RAG)	Qwen2-VL [10] + RAG	Tokens generated per second	30.09 tokens/s
TTS	Edge-TTS [56]	Time required to synthesize 1s audio	0.27448 s
Talking-Head Generation	MuseTalk [48]	Time required to drive one frame	0.0039 s

5. Conclusions

In this paper, we present RiverEcho, a real-time interactive digital system designed for the Ancient Yellow River culture. It processes user voice queries and delivers professional, informative responses in real-time via a digital human interface. Specifically, to enhance the output performance of the LLM module, we collected and processed historical texts and modern books related to the Ancient Yellow River from different dynasties and various topics, constructing a dataset for Ancient Yellow River culture. Finally, we conducted a subjective evaluation to validate the effectiveness of this system. We hope that artificial intelligence will contribute to the preservation, revitalization, and innovative dissemination of traditional Chinese culture.

Author Contributions

H.W.: Experiments, system implementation, and writing. Y.G.: System implementation, writing, and revision. Z.L.: Data guidance. T.Y.: Assistance with system implementation. Y.W.: Writing guidance. T.Z.: Data engineering. R.L.: Experimental design. F.G.: Artistic guidance. S.W.: Academic supervision and writing guidance. S.M.: Conceptualization, academic supervision, and writing guidance. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the National Key R&D Program of China 2022YFF0902400, NSFC 62025101, BNSF L242014 and New Cornerstone Science Foundation through the XPLOER PRIZE.

Data Availability Statement

The datasets generated during and/or analysed during the current study are available from the following sources: The M3DocVQA dataset is available at <https://github.com/bloomberg/m3docrag/tree/main/m3docvqa>, The MP-DocVQA dataset is available at <https://trc.cvc.uab.es/?ch=17&com=downloads>, The Yellow River Corpus is available at <https://pan.baidu.com/s/1uPo206qqeTDGaWRxZufG5w?pwd=tlpd>, The source code developed for this study is available at the GitHub repository: <https://github.com/hfwang2001/MMRAG>.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Cao, G. The Historical Inheritance and Contemporary Value of Yellow River Culture. *Jinyang Acad. J.* **2022**, *2*, 119–124.

2. Langote, M.; Saratkar, S.; Kumar, P.; et al. Human-computer interaction in healthcare: Comprehensive review. *Aims Bioeng.* **2024**, *11*, 343–390.
3. De Wet, L. Teaching Human-Computer Interaction Modules—And Then Came COVID-19. *Front. Comput. Sci.* **2021**, *3*, 793466.
4. Amato, F.; Barolli, L.; Cozzolino, G.; et al. An Intelligent Interface for Human-Computer Interaction in Legal Domain. In Proceedings of the In International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, Tirana, Albania, 27–29 October 2022.
5. Hirsch, L.; Paananen, S.; Lengyel, D.; et al. Human–Computer Interaction (HCI) Advances to Re-Contextualize Cultural Heritage toward Multiperspectivity, Inclusion, and Sensemaking. *Appl. Sci.* **2024**, *14*, 7652.
6. Achiam, J.; Adler, S.; Agarwal, S.; et al. Gpt-4 technical report. *arXiv* **2023**, arXiv:2303.08774.
7. Touvron, H.; Lavril, T.; Izacard, G.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971.
8. Yang, A.; Xiao, B.; Wang, B.; et al. Baichuan 2: Open large-scale language models. *arXiv* **2023**, arXiv:2309.10305.
9. Liu, A.; Feng, B.; Xue, B.; et al. Deepseek-v3 technical report. *arXiv* **2024**, arXiv:2412.19437.
10. Yang, A.; Yang, B.; Zhang, B.; et al. Qwen2. 5 technical report. *arXiv* **2024**, arXiv:2412.15115.
11. Roziere, B.; Gehring, J.; Gloeckle, F.; et al. Code llama: Open foundation models for code. *arXiv* **2023**, arXiv:2308.12950.
12. Li, Y.; Li, Z.; Zhang, K.; et al. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus* **2023**, *15*, e40895.
13. Zhang, H.; Qiu, B.; Feng, Y.; et al. Baichuan4-Finance Technical Report. *arXiv* **2024**, arXiv:2412.15270.
14. Cui, J.; Ning, M.; Li, Z.; et al. Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model. *arXiv* **2023**, arXiv:2306.16092.
15. Jiang, Z.; Wang, J.; Cao, J.; et al. Towards better translations from classical to modern Chinese: A new dataset and a new method. In Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing, Foshan, China, 12–15 October 2023.
16. Chang, E.; Shiue, Y.T.; Yeh, H.S.; et al. Time-aware ancient chinese text translation and inference. *arXiv* **2021**, arXiv:2107.03179.
17. Li, Z.; Sun, M. Punctuation as implicit annotations for Chinese word segmentation. *Comput. Linguist.* **2009**, *35*, 505–512.
18. Yu, P.; Wang, X. BERT-based named entity recognition in Chinese twenty-four histories. In Proceedings of the International Conference on Web Information Systems and Applications, Guangzhou, China, 23–25 September 2020.
19. Han, X.; Xu, L.; Qiao, F. CNN-BiLSTM-CRF model for term extraction in Chinese corpus. In Proceedings of the Web Information Systems and Applications: 15th International Conference, WISA 2018, Taiyuan, China, 14–15 September 2018.
20. Wang, D.; Liu, C.; Zhao, Z.; et al. GujiBERT and GujiGPT: Construction of intelligent information processing foundation language models for ancient texts. *arXiv* **2023**, arXiv:2307.05354.
21. Chang, L.; Dongbo, W.; Zhixiao, Z.; et al. SikuGPT: A generative pre-trained model for intelligent information processing of ancient texts from the perspective of digital humanities. *arXiv* **2023**, arXiv:2304.07778.
22. Wptoux. Bloom-7B-Chunhua. Available online: <https://huggingface.co/wptoux/bloom-7b-chunhua> (accessed on 1 October 2023).
23. XunziALLM. Available online: <https://github.com/Xunzi-LLM-of-Chinese-classics/XunziALLM> (accessed on 1 March 2024).
24. Cao, J.; Peng, D.; Zhang, P.; et al. TongGu: Mastering Classical Chinese Understanding with Knowledge-Grounded Large Language Models. *arXiv* **2024**, arXiv:2407.03937.
25. Mallen, A.; Asai, A.; Zhong, V.; et al. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. *arXiv* **2022**, arXiv:2212.10511.
26. Carlini, N.; Tramer, F.; Wallace, E.; et al. Extracting training data from large language models. In Proceedings of the 30th USENIX Security Symposium (USENIX Security 21), Virtual, 11–13 August 2021.
27. Huang, L.; Yu, W.; Ma, W.; et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Acm Trans. Inf. Syst.* **2025**, *43*, 1–55.
28. Izacard, G.; Lewis, P.; Lomeli, M.; et al. Atlas: Few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.* **2023**, *24*, 1–43.
29. Wu, Y.; Rabe, M.N.; Hutchins, D.; et al. Memorizing transformers. *arXiv* **2022**, arXiv:2203.08913.
30. He, Z.; Zhong, Z.; Cai, T.; et al. Rest: Retrieval-based speculative decoding. *arXiv* **2023**, arXiv:2311.08252.
31. Kang, M.; Gürel, N.M.; Yu, N.; et al. C-rag: Certified generation risks for retrieval-augmented language models. *arXiv* **2024**, arXiv:2402.03181.
32. Karpukhin, V.; Oguz, B.; Min, S.; et al. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the Empirical Methods in Natural Language Processing, Virtual, 16–20 November 2020.
33. Ni, J.; Qu, C.; Lu, J.; et al. Large dual encoders are generalizable retrievers. *arXiv* **2021**, arXiv:2112.07899.
34. Nogueira, R.; Cho, K. Passage Re-ranking with BERT. *arXiv* **2019**, arXiv:1901.04085.
35. Yoran, O.; Wolfson, T.; Bogin, B.; et al. Answering questions by meta-reasoning over multiple chains of thought. *arXiv* **2023**, arXiv:2304.13007.
36. Yao, S.; Zhao, J.; Yu, D.; et al. React: Synergizing reasoning and acting in language models. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.

37. Lewis, P.; Perez, E.; Piktus, A.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
38. Liu, Z.; Simon, C.E.; Caspani, F. Passage segmentation of documents for extractive question answering. In *European Conference on Information Retrieval*; Springer: Berlin/Heidelberg, Germany, 2025; pp. 345–352.
39. Laitenberger, A.; Manning, C.D.; Liu, N.F. Stronger Baselines for Retrieval-Augmented Generation with Long-Context Language Models. *arXiv* **2025**, arXiv:2506.03989.
40. Edge, D.; Trinh, H.; Cheng, N.; et al. From local to global: A graph rag approach to query-focused summarization. *arXiv* **2024**, arXiv:2404.16130.
41. Cho, J.; Mahata, D.; Irsoy, O.; et al. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv* **2024**, arXiv:2411.04952.
42. Faysse, M.; Sibille, H.; Wu, T.; et al. Colpali: Efficient document retrieval with vision language models. *arXiv* **2024**, arXiv:2407.01449.
43. Wang, Q.; Ding, R.; Chen, Z.; et al. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv* **2025**, arXiv:2502.18017.
44. Memon, J.; Sami, M.; Khan, R.A.; et al. Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). *IEEE Access* **2020**, *8*, 142642–142668.
45. Ingemarsson, P.; Daniel, P. PDF Parsing, Unveiling the Most Efficient Method. Bachelor’s Thesis, Linnaeus University, Växjö, Sweden, 2024.
46. LiveTalking: Real-Time Interactive Streaming Digital Human. 2024. Available online: <https://github.com/lipku/livetalking> (accessed on 16 March 2025).
47. Prajwal, K.R.; Mukhopadhyay, R. Wav2Lip: A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild. 2020. Available online: <https://github.com/Rudrabha/Wav2Lip> (accessed on 16 March 2025).
48. Zhang, Y.; Liu, M.; Chen, Z.; et al. Musetalk: Real-time high quality lip synchronization with latent space inpainting. *arXiv* **2024**, arXiv:2410.10122.
49. Metahuman-stream: Real-time Streaming Digital Human Based on NeRF. 2023. Available online: <https://github.com/tsman/metahuman-stream> (accessed on 16 March 2025).
50. Adobe Systems Incorporated. Real-Time Messaging Protocol (RTMP) Specification. 2002. Available online: https://web.archive.org/web/20201001140644/https://www.adobe.com/content/dam/acom/en/devnet/rtmp/pdf/rtmp_specification_1.0.pdf (accessed on 16 March 2025).
51. IETF and W3C. Web Real-Time Communication (WebRTC) Standard. 2011. Available online: <https://www.w3.org/TR/webrtc/> (accessed on 16 March 2025).
52. Synthesia. Synthesia: AI Video Generation Platform. 2017. Available online: <https://www.synthesia.io/> (accessed on 16 March 2025).
53. Diener, V. VTube Studio: Live2D VTuber Streaming Software. 2021. Available online: <https://github.com/mouwoov/VTubeStudio/wiki> (accessed on 16 March 2025).
54. Guo, Z.; Xia, L.; Yu, Y.; et al. Lightrag: Simple and fast retrieval-augmented generation. *arXiv* **2024**, arXiv:2410.05779..
55. Gao, Z.; Li, Z.; Wang, J.; et al. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv* **2023**, arXiv:2305.11013.
56. . Edge-tts: Use Microsoft Edge’s Online Text-to-Speech Service from Python WITHOUT Needing Microsoft Edge or Windows or an API Key. 2024. Available online: <https://github.com/rany2/edge-tts> (accessed on 16 March 2025).
57. Glm, T.; Zeng, A.; Xu, B.; et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv* **2024**, arXiv:2406.12793.
58. Chaplot, D.S.; Jiang, A.Q.; Sablayrolles, A.; et al. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825.
59. Jha, R.; Wang, B.; Günther, M.; et al. Jina-colbert-v2: A general-purpose multilingual late interaction retriever. *arXiv* **2024**, arXiv:2408.16672.
60. Vavekanand, R.; Sam, K. Llama 3.1: An in-depth analysis of the next-generation large language model. *Preprint* **2024**.
61. Lu, H.; Liu, W.; Zhang, B.; et al. Deepseek-vl: Towards real-world vision-language understanding. *arXiv* **2024**, arXiv:2403.05525.
62. Laurençon, H.; Tronchon, L.; Cord, M.; et al. What matters when building vision-language models? *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 87874–87907.
63. Guo, Z.; Xu, R.; Yao, Y.; et al. Llava-uhd: An lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 390–406.
64. Dong, X.; Zhang, P.; Zang, Y.; et al. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 42566–42592.
65. Hu, A.; Xu, H.; Ye, J.; et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv* **2024**, arXiv:2403.12895.
66. Bai, J.; Bai, S.; Chu, Y.; et al. Qwen technical report. *arXiv* **2023**, arXiv:2309.16609.
67. Li, Z.; Yang, B.; Liu, Q.; et al. Monkey: Image resolution and text label are important things for large multi-modal models. In proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024.