摘要

随着大模型技术在金融、教育、司法等领域的深度渗透与广泛应用，其为更多垂直领域如传统文化的数字化保护、传承与创新性传播提供了全新的技术路径。黄河文化作为中华文明的核心发祥地与精神象征，承载着数千年的历史积淀与精神内涵，但其数字化传播仍面临诸多瓶颈：一是垂域数据支撑不足，黄河文化相关古籍、专著分布零散且数字化程度低，缺乏系统化的多模态数据集，文旅视频资源也无针对性整理，难以满足大模型训练与增强需求；二是多模态检索能力有限，传统检索增强生成方法对包含文字、图像、表格、公式的黄河文化多模态文档检索精度不足，跨模态语义对齐效果欠佳，导致大模型回答易出现幻觉、信息量不足等问题；三是交互传播形式单一，现有系统多以单向文本输出为主，缺乏多模态实时交互能力，难以提升用户的沉浸式体验与文化认知参与度。尽管国内外已在检索增强生成技术、传统文化大模型及多模态检索领域开展相关研究，但针对黄河文化这一特定垂域的系统性大模型构建工作尚未开展，上述核心问题仍未得到有效解决，难以满足黄河文化数字化传播的实际需求。

针对黄河文化数字化传播中的上述问题，以检索增强生成技术为核心，从数据集构建、检索方法创新到交互系统开发三个层次展开系统研究，构建了由黄河文化多模态垂域数据集、多模态文档检索增强生成方法到黄河文化实时交互系统的多层技术体系。

首先，针对黄河文化垂域数据稀缺、数字化程度低、缺乏专用多模态数据集的问题，构建了两类多模态垂域数据集。一是黄河文化垂域知识语料库，通过系统化文献采集整合先秦至明清古籍与当代学术专著，覆盖河流治理、技术工程、文化遗产等八大主题，经 OCR 识别、大语言模型结构化处理与人工校对的三级流程，提取实体与关系信息并构建知识图谱，支持图结构检索与向量检索；二是文旅场景视频生成数据集，通过多渠道采集黄河流域九省区标志性自然与人文地标相关素材，经场景分割、水印与字幕去除、结构化标注等预处理，形成高质量视频片段，构建省、市、景区三级地理标签体系，并通过低秩适配技术验证了数据集对文生视频模型的微调有效性。两类数据集分别从文本知识与视觉场景维度，为黄河文化大模型提供了高质量数据支撑。

其次，针对传统检索增强生成方法对多模态文档检索效果不佳、跨模态语义对齐不足的问题，提出了基于文档内容对齐的多模态文档检索增强生成方法。该方法先通过文档内容拆解工具，将多模态文档解耦为文字、表格、公式、图像四类核心元素；再设计上下文感知的图文对齐和嵌入模块，将表格、公式转换为文本索引后按段落处理，通过视觉编码器与文本编码器提取特征并融合布局信息，结合对比学习实现跨模态语

义精准对齐；同时构建查询感知门控融合模块，通过门控网络生成相关性得分，结合双注意力模块交互查询与多模态特征，经加权融合得到最终特征；并采用多任务损失函数联合优化跨模态对齐、查询融合匹配与门控监督，保障特征表示准确性；最后采用嵌入表示与图表示结合的混合检索策略，实现多模态元素的段落级精准检索。实验结果表明，该方法显著优于现有文本与多模态检索增强方法，且在不同多模态大模型上均取得良好的泛化性。

最后，针对黄河文化传播交互形式单一、多模态实时交互缺失的问题，开发了黄河文化实时交互系统，包含实时对话数字人系统与文旅场景视频生成系统两大核心部分。实时对话数字人系统采用全流程流式处理架构，集成语音识别、多模态大模型、文本转语音、数字人生成等模块，实现语音唤醒、实时打断、唇形同步的自然交互，端到端响应速率满足实时交互需求；文旅视频生成系统采用前后端分离的 Web 架构，支持用户输入文本提示，生成多分辨率横版与竖版黄河文旅短视频，适配大屏展示与移动端传播等场景。该系统已在全息仓等终端设备部署并投入商业使用，实现了黄河文化从文本传播向多模态交互体验的转变。

综上所述，围绕黄河文化大模型构建中的数据支撑、检索方法和交互系统等关键技术问题展开系统性研究，提出了多模态垂域数据集构建方案、多模态文档检索增强生成方法以及实时交互系统实现路径。研究工作聚焦于黄河文化垂域数据整合、多模态检索精度提升及多模态交互传播创新，旨在实现黄河文化数字化传播从数据零散向系统支撑、从单一检索向多模态对齐、从被动输出向主动交互的转变。所形成的方法技术可用于黄河文化大模型完善、传统文化高效数字化传播以及文旅产业智能化升级。

关键词：检索增强生成，黄河文化，大模型，多模态检索

A Large Language Model for Yellow River Culture Based on Retrieval-Augmented Generation

Haofeng Wang (Computer Application Technology)
Directed by: Prof. Siwei Ma

ABSTRACT

With the deep penetration and wide application of large model technology in finance, education, justice and other fields, it provides a brand-new technical pathway for the digital protection, inheritance and innovative dissemination of traditional culture in more vertical domains. Yellow River culture, as the core birthplace and spiritual symbol of Chinese civilization, carries thousands of years of historical accumulation and connotation, yet its digital dissemination still faces multiple bottlenecks: insufficient domain-specific data support, limited multimodal retrieval capability, and monotonous interactive dissemination. Although relevant research on RAG technology, traditional-culture large models and multimodal retrieval has been carried out domestically and internationally, systematic construction of a large model specific to the Yellow River culture domain has not yet been undertaken, leaving the above core problems effectively unresolved.

To address the above problems, this paper takes retrieval-augmented generation technology as the core and conducts systematic research at three levels —dataset construction, retrieval method innovation, and interactive system development —establishing a multi-layer technical system composed of Yellow River culture multimodal domain datasets, a multimodal document RAG method, and a Yellow River culture real-time interactive system.

First, to solve the scarcity, low digitization and lack of dedicated multimodal datasets in the Yellow River culture domain, two multimodal domain datasets are constructed. One is the Yellow River culture domain knowledge corpus, which integrates ancient books and contemporary academic monographs covering eight themes including river governance, technical engineering and cultural heritage. Through a three-stage pipeline of OCR recognition, large language model structured processing and manual proofreading, entities and relations are extracted to build a knowledge graph supporting graph-based retrieval and vector retrieval. The other is the tourism video generation dataset, which collects materials from iconic natural and

cultural landmarks across nine provinces of the Yellow River basin and, after scene segmentation, watermark and subtitle removal and structured annotation, forms high-quality video clips with a province-city-scenic-spot three-level geographic tag system. Low-rank adaptation fine-tuning verifies the dataset's effectiveness for text-to-video models. These two datasets provide the Yellow River Culture Large Model with data support from both textual knowledge and visual scene dimensions.

Second, aiming at the poor performance of traditional RAG methods on multimodal documents and insufficient cross-modal semantic alignment, a multimodal document retrieval-augmented generation method based on document content alignment is proposed. The method decouples multimodal documents into four core elements —text, tables, formulas and images —and designs a context-aware image-text alignment and embedding module, which converts tables and formulas into text indices, processes them by paragraph, extracts features through visual and text encoders incorporating layout information, and achieves cross-modal semantic alignment via contrastive learning. A query-aware gated fusion module is constructed, generating relevance scores through a gating network and employing dual-attention to interact between queries and multimodal features for weighted fusion. A multi-task loss jointly optimizes cross-modal alignment, query-fusion matching and gate supervision. Finally, a hybrid retrieval strategy combining embedding and graph representations achieves paragraph-level precise retrieval. Experiments demonstrate that this method significantly outperforms existing text-based and multimodal RAG approaches and exhibits good generalization across different multimodal large models.

Third, to overcome the single interaction form and lack of multimodal real-time interaction in Yellow River culture dissemination, a real-time interactive system is developed, comprising a real-time conversational digital human and a tourism video generation system. The conversational digital human adopts an end-to-end streaming architecture, integrating speech recognition, a multimodal large model, text-to-speech and digital human generation modules, supporting voice wake-up, real-time interruption and lip synchronization, with an end-to-end response rate satisfying real-time interaction demands. The tourism video generation system uses a front-end and back-end separated web architecture, generating multi-resolution horizontal and vertical Yellow River tourism short videos from text prompts, suitable for large-screen display and mobile dissemination. The system has been deployed on holographic terminals and put into commercial use, shifting Yellow River culture dissemination from text-centered to multimodal interaction.

In summary, this research systematically addresses the key technical problems of data support, retrieval methods and interactive systems in the construction of the Yellow River Culture Large Model, proposing a multimodal domain dataset construction scheme, a multimodal document retrieval-augmented generation method, and a real-time interactive system implementation pathway. The work focuses on domain data integration, multimodal retrieval precision improvement, and multimodal interactive dissemination innovation for Yellow River culture, aiming to realize the transitions from scattered data to systematic support, from single-modal retrieval to multimodal alignment, and from passive output to active interaction in the digital dissemination of Yellow River culture. The proposed methods and technologies can be applied to the improvement of the Yellow River Culture Large Model, the efficient digital dissemination of traditional culture, and the intelligent upgrading of the cultural tourism industry.

KEY WORDS: Retrieval-Augmented Generation, Yellow River Culture, Large Language Model, Multimodal Retrieval