



“博古问津”:知识图谱增强的文化遗产领域 多模态大模型

赵万青^{1,2},徐朝阳¹,谢智伟¹,张少博^{1,2},张晓丹^{1,2},彭进业^{1,2}

(1. 西北大学 电子信息学院[人工智能学院],陕西 西安 710127;

2. 陕西省高等学校生成式人工智能与混合现实重点实验室,陕西 西安 710127)

摘要 近年来,大语言模型(LLMs)和多模态大模型(MLMs)在自然语言处理和多模态内容理解方面取得了显著成就,然而,这些通用模型在处理文化遗产相关任务时存在明显缺陷,如对领域专业术语的理解存在偏差、缺乏文化历史背景导致回答不够深入以及知识幻觉等问题,使得输出结果难以满足实际需求。针对这些挑战,首次提出了面向文化遗产领域的多模态大模型——“博古问津”。首先,设计半自动化策略构建大规模的多模态文化遗产数据集并形成多模态知识图谱;然后,利用构建的数据集对通用大模型进行图文对齐和指令微调两阶段训练,以适应文化遗产领域的特定需求;此外,还引入了知识图谱作为辅助知识库,通过图文检索和关系检索策略,有效提升了模型在文化遗产领域问答任务上的可信度和可解释性。实验结果表明,“博古问津”在文物图像描述、属性问题解答及关系问题理解等多个方面表现优异,相较于通用多模态大模型,对复杂文化内容的理解和回答能力提升效果显著,分别在文物图像描述、文物属性问题和文物关系问题3个不同任务的综合分值上高出次优模型21.4%、53%和20.6%。

关键词 多模态大模型;文化遗产;视觉问答;知识图谱;模型微调;知识增强

中图分类号:TP391.4 **DOI**:10.16152/j.cnki.xdxbr.2025-06-006

Bogu-Wenjin: A cultural heritage domain multimodal large model enhanced by knowledge graphs

ZHAO Wanqing^{1,2}, XU Chaoyang¹, XIE Zhiwei¹,
ZHANG Shaobo^{1,2}, ZHANG Xiaodan^{1,2}, PENG Jinye^{1,2}

(1. School of Electronic Information [School of Artificial Intelligence], Northwest University, Xi'an 710127, China;

2. Shaanxi Key Laboratory of Higher Education Institution of Generative Artificial
Intelligence and Mixed Reality, Xi'an 710127, China)

Abstract In recent years, large language models (LLMs) and multimodal large models (MLMs) have made significant achievements in natural language processing and multimodal content understanding. However, these general-purpose models have obvious shortcomings when dealing with tasks related to cultural heritage, such as biased understanding of domain-specific terminology, lack of cultural and historical background leading to su-

收稿日期:2025-07-07

基金项目:国家重点研发计划(2024YFF0907600);国家自然科学基金(62273275);陕西省自然科学基金基础研究计划青年项目(2025JC-YBQN-847)。

第一作者:赵万青,男,教授,从事机器学习、媒体数据分析与检索研究,zhaowq@nwu.edu.cn。

perficual answers, and knowledge hallucination issues, making it difficult for the results to meet actual needs. In response to these challenges, this paper first proposes a multimodal large model oriented towards the field of cultural heritage: Bogu-Wenjin. This study first designs a semi-automated strategy to construct a large-scale multimodal cultural heritage dataset and forms a multimodal knowledge graph. Using the constructed dataset, the general large model is trained in two stages: image-text alignment and instruction fine-tuning, to adapt to the specific needs of the cultural heritage field. In addition, a knowledge graph is introduced as an auxiliary knowledge base, and the credibility and interpretability of the model in the field of cultural heritage Q&A tasks are effectively improved through graph-text retrieval and relationship retrieval strategies. Experimental results show that Bogu-Wenjin performs excellently in various aspects such as artifact image description, attribute question answering, and relationship question understanding. Compared with general multimodal large models, it significantly improves the ability to understand and answer complex cultural content, with a comprehensive score increase of 21.4%, 53% and 20.6% in artifact image description, artifact attribute questions, and artifact relationship questions respectively over the second-best model.

Keywords multimodal large models; cultural heritage; visual question answering; knowledge graphs; model fine-tuning; knowledge enhancement

随着人工智能技术的飞速发展,大语言模型(large language models, LLMs)^[1-3]已成为自然语言处理领域的一个突破性进展。这些模型通过在海量文本数据上的预训练,展现出在语言理解、生成和翻译等方面的卓越能力。进一步地,多模态大模型(multimodal large models, MLMs)的出现,将文本处理能力扩展到了图像、声音和视频等其他数据类型,使得机器能够更全面地理解和生成内容。通过融合视觉、听觉以及结构化信息,多模态大模型^[4-7]为复杂场景下的信息解析提供了更加丰富且全面的新视角。

在信息化和数字化时代,各个领域都产生了大量的多模态数据^[8-9]。这些数据的多样性为运用先进的人工智能技术开辟了广阔的应用前景。通用多模态大模型,如 GPT-4V^[10] 和 LLaVA^[11],已经展示了生成式预训练的强大学习能力,它们通过自监督学习有效地利用海量的图像和文本数据进行视觉语言建模。并通过对模型进行基于多模态输入的指令微调,确保其与人类意图的一致性。这些多模态大模型在图像理解、推理以及其他面向用户的视觉语言任务中表现出色,同时展现了在零样本任务上的良好泛化性^[12-13]。特别是在金融领域,FinTral 模型通过整合文本、数值、表格和图像数据,显著提升了金融分析的深度和广度^[14]。在生物医学领域,LLaVA-Med 模型则利用大规模的医学图像-病例文本对,训练出能够准确理解和回答生物医学图像相关问题的模型^[15]。

尽管多模态大模型在很多领域取得了显著成

就,然而在文化遗产领域^[16],这一技术的研究和应用却明显滞后,鲜有类似的突破。目前,学术界和工业界尚未开发出专门针对文化遗产保护、研究与传播的多模态大模型,这一现状与文化遗产领域的迫切需求形成了鲜明对比。文化遗产领域具有其独特的复杂性和多样性,包含了大量的历史信息和文化符号,这些都需要模型具备深厚的文化背景知识。此外,文化遗产文本、图像等多模态数据之间存在千丝万缕的联系,难以被通用模型学习。通用多模态大模型在处理文化遗产任务时存在明显缺陷,如对领域专业术语的理解存在偏差、缺乏文化历史背景导致回答不够深入以及知识幻觉等问题,严重影响了模型在文化遗产领域的实际应用效果。与此同时,多模态大模型在文化遗产领域的潜在价值却不容忽视:一方面,它可以赋能文物修复,通过分析不同时期的图像与文本记录,辅助专家推断文物原始形态或降解路径;另一方面,它能提升公众参与度,例如通过自然语言交互生成个性化的文物解说,或结合 AR 技术实现沉浸式文化体验。此外,面对濒危的非物质文化遗产(如方言、传统工艺),多模态建模可以更完整地保存其工艺流程与文化语境。因此,建立一个面向文化遗产领域的多模态大模型是一个多维度、跨学科的复杂任务,它不仅涉及到语言文字的记录,还包括图像、音频、视频以及非物质文化遗产的关联。

针对该需求,本研究首次构建了一个面向文化遗产领域的多模态大模型(称之为“博古问

津”),并基于该大模型实现了文化遗产领域的多模态智能问答(见图 1)。

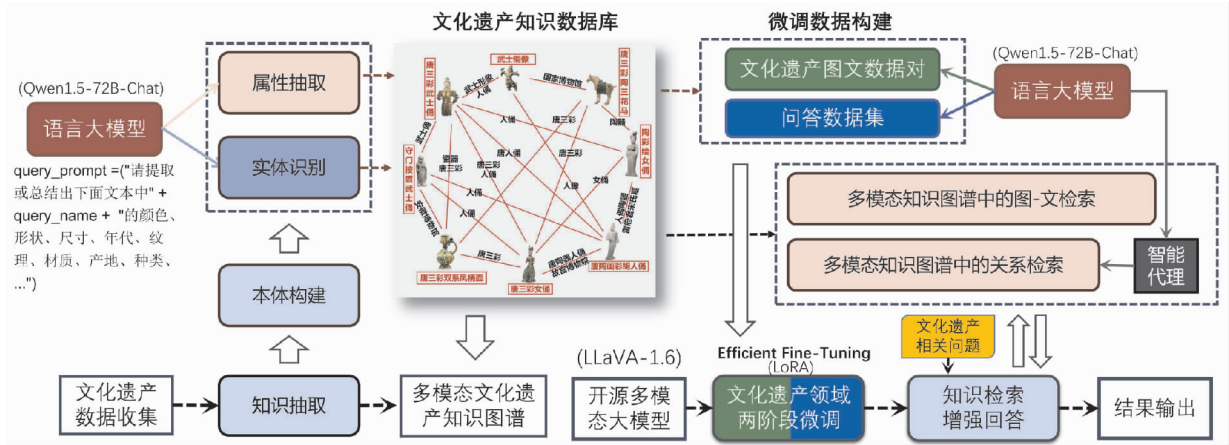


图 1 “博古问津”文化遗产多模态大模型构建过程

Fig. 1 The process of constructing the cultural heritage multimodal large model “Bogu-Wenjin”

如图 1 所示,该研究首先通过互联网爬取、文献 OCR 识别^[17]得到文化遗产原始数据,并经过基于大语言模型的数据清理和命名实体识别,构建了一个多模态文化遗产数据集。随后,基于该数据集,创建了文化遗产多模态知识图谱,包括图谱本体构建、基于大语言模型的实体识别和属性抽取、知识图谱关系构建。通过图文对齐学习与对话数据指令微调相结合的方式,有效地提升了模型在特定领域内的表现力,确保其能够准确理解用户意图并给出高质量的回答。此外,还引入了知识图谱作为辅助知识库,通过图文检索和关系检索策略,有效提升了模型在文化遗产领域问答任务上的可信度和可解释性。实验结果表明,“博古问津”在文物图像描述、属性问题解答及关系问题理解等多个方面均表现出色,相较于通用多模态大模型,在处理复杂文化内容时显示出明显的优势。

1 研究现状

1.1 多模态知识图谱

知识图谱技术^[18]是一种以符号形式描述物理世界中的概念及其相互关系的技术,其本质上是一种以实体为节点,以实体之间的语义关系作为边的大规模的语义网,已经被广泛用于医疗、金融等领域。但是现有的知识图谱大多只是传统的文本知识图谱,极大地影响了机器理解和表示现实世界的能力,人们从知识图谱中能获取到的语义信息也有限,例如“秦始皇兵马俑”这一文化遗产实体,单纯用文字难以表述其丰富的形态、纹理

以及细节雕刻等视觉表现。因此,有必要将文字和图像等模态结合起来,这就需要关注非结构化的数据,如视觉数据,多模态知识图谱应运而生。多模态知识图谱就是包含了多种模态数据(如视觉模态、声音态等)的知识图谱,不仅在知识的展示上更加丰富,而且在其他很多领域的应用上也越来越广泛,比如在关系提取任务中,图像模态的信息会大大提高一些文字难以识别的属性和关系,如“partOf”和“colorOf”等。同时,机器在文本生成任务中可以通过参考多模态知识图谱生成信息量更大的完整句子,而不是模糊的概念描述。

虽然多模态知识图谱的定义已经比较完善,但是其构建方法仍处于一个初步阶段。近年来,多模态视觉技术^[19]取得了显著进展,但是这些技术的主要应用领域还是图像生成和分类等任务,尚未能有效应用于多模态知识图谱的构建,尤其是在文化遗产领域。构建文化遗产多模态知识图谱面临的主要挑战包括:首先,文化遗产数据来源广泛且分散,不同机构和博物馆的数据格式各异,导致数据整合难度大,例如一些博物馆的数据以文本形式存储,而另一些则以图像或音频形式存在,这种异构性增加了数据预处理的复杂性;其次,文化遗产的多模态数据之间存在复杂的语义关联,如文物的图像与其历史背景、制作工艺等文本信息之间的关联,这些关联难以通过简单的技术手段自动识别和提取;此外,文化遗产领域的专业知识性强,对实体和关系的定义需要深厚的文化和历史背景知识,增加了知识抽取和图谱构建的难度。

1.2 多模态大模型

近年来,大型语言模型(LLM)的推出与发展标志着自然语言处理领域的一个重要里程碑。借助深度学习与自监督学习技术,LLM 已在理解和生成自然语言的任务中取得了长足的进步。诸如 BERT^[20]、GPT^[21]、LLaMa^[22]、Qwen^[23] 等模型通过在海量文本数据上的预训练,积累了广泛的语言知识和模式识别能力,使得它们能够胜任复杂的语言任务,例如文本分类、情感分析以及机器翻译等。人类通过视觉、语言等多种感官与环境互动,每一种感知通道都在表达和交流特定概念时展现出独特的优势,从而帮助我们更全面地理解世界。人工智能的一个核心目标是创建一个能够有效理解并执行多模态视觉和语言指令的通用助手,使之与人类意图相匹配。因此,多模态大模型(MLM)应运而生,这类模型能够处理和理解文本、图像等多种类型的数据,以更贴近人类交流的方式整合不同模态的信息,从而提升任务处理的准确性和效率。OpenAI 发布的首个多模态大模型 GPT-4V^[24] 自问世以来,凭借其卓越的多模态处理能力赢得了广泛的关注,尤其在处理复杂图像输入方面展现了前所未有的能力。微软研发的 LLaVA^[11] 则结合了 CLIP^[25] 的视觉编码器与大型语言模型,通过一个轻量级的映射层实现了图像特征与文本特征的对齐。LLaVA 采用两阶段训练策略,首先是特征对齐预训练,其次是针对具体任务如视觉聊天和科学问答的端到端微调,展现了简洁高效的架构设计。MiniGPT-4^[26] 与 LLaVA 模型结构相似,其通过利用更先进的大型语言模型 Vicuna^[27] 增强视觉语言理解能力,并探讨了 GPT-4 卓越多模态生成能力主要原因是利用了更高级的大型语言模型。国内研发的 Qwen-VL^[28] 则是在 Qwen 这一大型语言模型基础上,结合多模态多任务训练而提出的解决方案,旨在突破多模态模型在泛化能力上的限制,支持中英文等多种语言,并具备图像识别、描述、问答以及对话等功能,同时增加了视觉定位和图像中文本理解的新特性。智谱 AI 推出的 GLM-4V^[29] 是一款专注于图像和视频理解的模型,具备处理高分辨率图像和视频内容的能力,能够在时间维度上感知视频信息。

然而,当通用多模态大模型应用于具体细分领域时,由于难以准确理解特定领域的专业术语、行业规范或文化背景,导致输出结果不够精准或

相关。此外,知识幻觉问题在细分领域表现更加明显,即模型生成看似合理但实际上错误或误导性的信息。为了克服这些挑战,领域多模态大模型成为了一个研究热点^[3]。领域多模态大模型通常需要专门设计的数据集和训练策略,以确保模型能够充分理解特定领域的专业知识和背景信息。比如 LLaVA-med 通过在医学领域的多模态数据微调,展现出了在医学图像问答中的出色性能。在文化遗产保护这一领域,多模态大模型的应用尚处于起步阶段。文化遗产领域因其特有的复杂性和多样性,给多模态模型提出了更高的要求:一方面,文化遗产包含了大量的历史信息和文化符号,这些都需要模型具备深厚的文化背景知识;另一方面,文化遗产文本、图像等多模态数据之间存在千丝万缕的联系,难以被大模型学习。因此,针对文化遗产领域的多模态大模型及相关应用需要专门的设计与训练,不仅要克服知识幻觉等普遍问题^[31],还要深入理解并准确再现文化遗产的独特价值才能在文化遗产保护和推广中发挥更大的作用。

2 文化遗产领域数据集及多模态知识图谱构建

2.1 文化遗产领域数据收集

文化遗产是历史的见证,它们以历史建筑、艺术品、手工艺品、文献资料等多种形式存在,承载着丰富的历史和文化价值。随着计算机技术的不断发展,越来越多的文化遗产资源被数字化保存,包括高分辨率的图像、3D 模型、音频和视频记录等方式。目前,各个地方博物馆都对自己馆藏的文物进行数字化处理,并通过在线博物馆的方式进行互联网展陈。尽管文化遗产通过数字化方式记录了下来,但数据资源分散在不同的机构和博物馆,数字化资源的格式多样,导致这些数据难以充分利用。

本研究在遵守版权和保护政策的前提下,通过互联网对公开的文化遗产资源进行自动爬取,从指定的初始 Web 页面出发,采用广度优先搜索策略,系统地遍历相关网页,抓取页面信息。这些信息随后被保存为 HTML 文件,以保留原始的网页结构和内容。进一步地从保存的 HTML 文件中精确提取所需文化遗产数据。对于结构化数据,使用 XPath 选择器从 HTML 文件中提取所需

的文化遗产的图像及文本介绍。XPath 选择器是一种强大的文档检索定位工具,能够精确地定位 HTML 文档中的特定元素。对于非结构化数据,根据页面的具体布局和格式定制提取规则,通过使用正则表达式识别和提取文本描述,以及使用 HTML 标签属性定位和下载图片文件。为了丰富数据集,本工作还对文献等纸质资料进行数字化提取。首先,通过高精度扫描将纸质资料转化为数字文档;然后,采用 PDFMiner 库对 PDF 文档进行解析,以提取和分离图像与文本数据;此外,还应用 OCR(光学字符识别)进一步对文字进行提取,以获取可用的信息。

2.2 文化遗产数据本体构建

在文化遗产领域,数据的组织和分类至关重要,因为它们构成了知识体系的基础。为了使杂

乱无章的原始数据变得有序,本文首先着手于确立文化遗产数据本体的结构化一致性。通过对文化遗产领域内普遍接受的概念和知识的深入分析,确定构建本体所需的实体类、实体关系和实体属性等基本要素。

实体类是指文化遗产中的核心概念,例如“文物”、“历史事件”等,它们构成了知识体系的骨架。实体关系则描绘了这些实体类之间的相互联系,如“属于”、“发生在”等,它们是理解文化遗产内在联系的关键。而实体属性则详细描述了实体类的具体特征,例如“年代”、“材质”等,为每个概念提供了丰富的细节。通过这一过程,我们精心设计了 6 个实体类、11 个实体关系和 18 个实体属性,形成了一个全面而细致的本体结构,如图 2 所示。

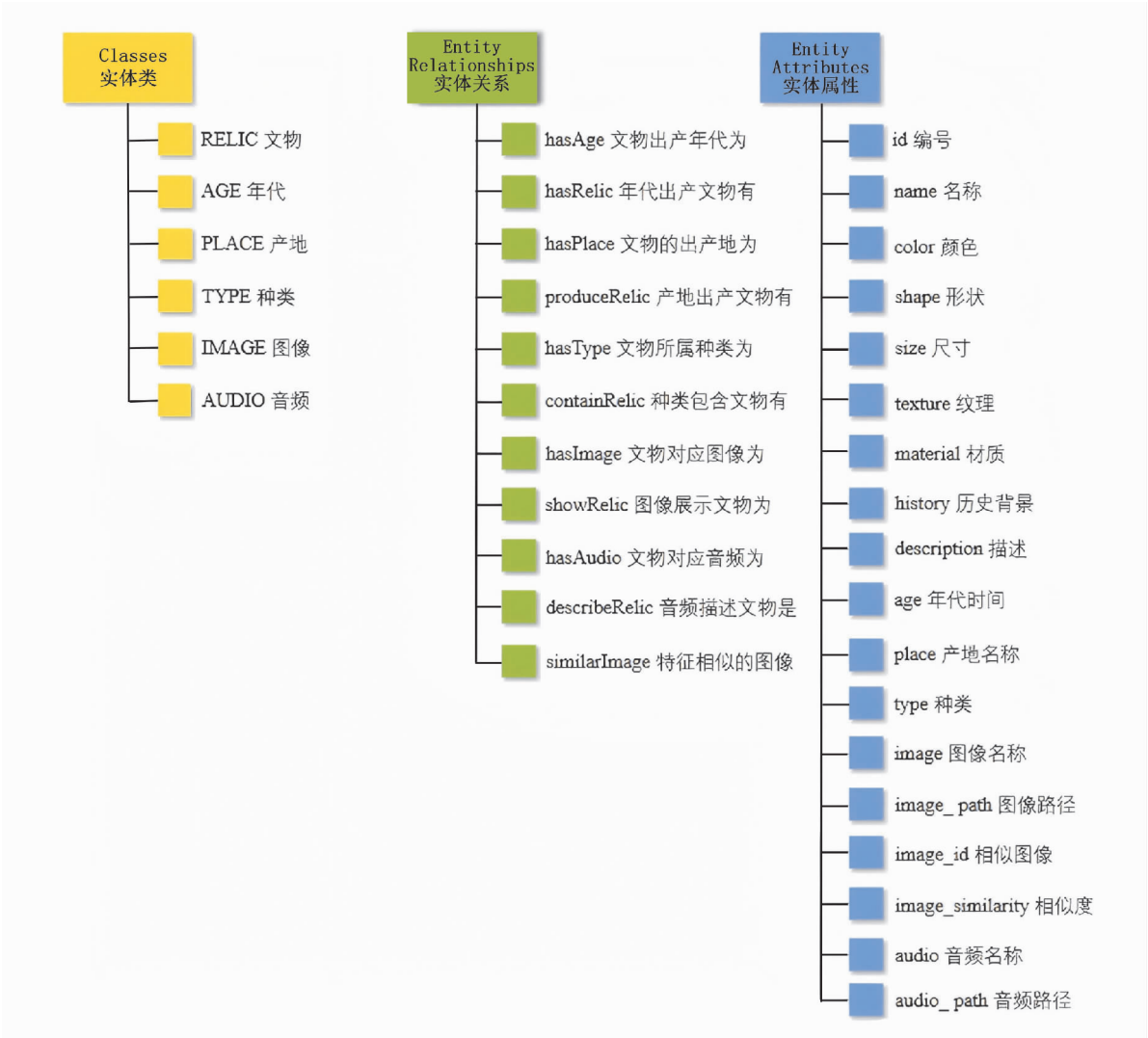


图 2 定义的文化遗产实体的类型、关系和属性

Fig. 2 Types, relationships and attributes of defined cultural heritage entities

其中,文物实体类指的是所有文物抽象出来的实体类别,也是所有实体类的中心类。文物是指具有着深厚历史、文化和艺术价值的物品,其存在不仅仅是一种实体形式,更是对历史、社会制度和整个文明的生动回溯。这一类物品涵盖了古代

手工艺品、书籍、绘画、器物等多个方面,所以将所有的文物抽象为文物实体类。该类包含文物编号、文物名称、文物颜色、文物形状、文物尺寸、文物纹理、文物材质、历史背景、文物描述共 9 个数据属性,如表 1 所示。

表 1 文物实体类的属性描述

Tab. 1 Attribute descriptions of heritage entity classes

实体类	属性名称	英文名称	数据类型
文物(RELIC)	文物编号	id	xsd:int
	文物名称	name	xsd:string
	文物颜色	color	xsd:string
	文物形状	shape	xsd:string
	文物尺寸	size	xsd:string
	文物纹理	texture	xsd:string
	文物材质	material	xsd:string
	历史背景	history	xsd:string
	文物描述	description	xsd:string

除了文物实体类,还定义了其他几个关键的实体类来丰富数据的维度和深度。年代实体类捕捉文物的历史时期,为研究者提供了将文物放置在历史时间轴上的能力;产地实体类揭示了文物的起源地,有助于追溯文物的流通路径和文化传播;种类实体类则根据文物的类型进行分类,如青铜器、玉器等,便于研究者深入研究特定类别的文物;图像实体类通过视觉数据,如照片和绘画,提供了文物的视觉特征,增强了文物信息的直观性;

音频实体类引入音频模态,尤其是对于乐器类文物,提供了独特的音响信息,为文物的多模态研究开辟了新途径。实体类之间的关系,即实体关系,以文物实体类为核心,构建了 11 种有向关系,具体的定义如表 2 所示。这些关系不仅明确了不同实体类之间的联系,而且为知识图谱的构建提供了结构化的框架,使得文化遗产的数据更加有序和易于理解。

表 2 实体关系类的属性描述

Tab. 2 Attribute description of entity relationship classes

实体关系类	关系含义	英文描述	定义域	值域
文物与年代	文物出产年代为	hasAge	RELIC	AGE
	年代出产文物有	hasRelic	AGE	RELIC
文物与产地	文物的出产地为	hasPlace	RELIC	PLACE
	产地出产文物有	produceRelic	PLACE	RELIC
文物与种类	文物所属种类为	hasType	RELIC	TYPE
	种类包含文物有	containRelic	TYPE	RELIC
文物与图像	文物对应图像为	hasImage	RELIC	IMAGE
	图像展示文物为	showRelic	IMAGE	RELIC
文物与音频	文物对应音频为	hasAudio	RELIC	AUDIO
	音频描述文物是	describeRelic	AUDIO	RELIC
图像与图像	特征相似的图像	similarImage	IMAGE	IMAGE

2.3 多模态文化遗产知识抽取

多模态文化遗产知识抽取涉及从文化遗产的原始数据中提取和整合信息,以构建一个全面、丰富且结构化的知识体系。这一过程不仅需要文化遗产的物理特征进行描述,如形状、尺寸、材质等,还需要对其历史背景、文化意义和艺术价值等信息进行提取。此外,多模态知识抽取还涉及到对文化遗产原始图像的信息提取,以便将这些非结构化的数据转化为可以被计算机处理和分析的结构化数据。

本文利用语言大模型的自然语言处理能力从

```
query_prompt = (“请提取或总结出下面文本中” + query_name +  
“的颜色、形状、尺寸、年代、纹理、材质、产地、”  
“种类(可选:石器、玉器、骨器、木器、竹器、铜器、铁器、金器、银器、铅锌器、瓷器、漆器、玻璃器、珐琅器、纺织品、纸  
类文物)、历史背景,”  
“没有提取出来的项输出 None。” + query_des)
```

对于图像数据,使用不同的图像特征提取文物的视觉元素,包括颜色直方图、纹理特征(灰度共生矩阵和局部二值模式)、形状特征(Canny 边缘检测)。这些特征从不同的角度描述文物的视觉表现,补充了文本信息的局限性,丰富了对文物视觉表现的理解和知识。

2.4 多模态文化遗产知识图谱

多模态知识图谱构建是在完成文化遗产数据的收集与本体构建的基础上,进一步将这些数据整合到一个统一的知识图谱中。图谱以结构化的方式表示文化遗产领域的知识,不仅包括了文物的文本信息,还包括了文物的图像及文物间的关联关系,从而形成了一个更加全面的文化遗产知识表示。

该图谱的核心组成部分包括实体类、实体关系和实体属性。实体类定义了文化遗产中的核心概念,如“文物”“历史事件”“地点”等,构成知识体系的骨架,例如“文物”实体类包含了一系列属性,包括文物编号、名称、颜色、形状、尺寸、纹理、材质、历史背景和描述等。实体关系描绘了实体类之间的相互联系,如“属于”“发生在”等,有助于理解文化遗产内在的联系,再如“文物”与“年代”之间存在“hasAge”的关系,表明某件文物产生于特定的历史时期。实体属性提供了实体类的具体特征描述,例如文物的材质、尺寸等,这些属性为每个概念提供了详细的细节信息。此外,多模态数据融合也是图谱的重要部分,图像数据通过视觉数据如照片和绘画提供文物的视觉特征,

文本数据中进行自动的知识抽取,通过精心设计的对话式查询引导语言大模型进行实体识别和属性抽取。具体地,将包含了文物的名称和描述信息的文本输入语言大模型,同时,将对话提示词设置为需要提取的关键的实体(如年代、产地、种类)和属性(如颜色、形状、尺寸、纹理、材质、历史背景)信息。最终构建的对文物数据进行实体识别和属性抽取的提示词(query_prompt)如下框中文字所示(query_name 是文物名称;query_des 是从半结构化数据中取出的文物描述)。

可以用来识别文物的形状、颜色、纹理等,并且可以用来建立视觉相似关系,即特征相似的图像之间的联系。

构建多模态知识图谱的过程主要是将抽取的多模态文化遗产知识依据预先定义的实体类、实体关系和实体属性进行数据填充。通过遍历文物数据,根据定义的实体类属性列表,将抽取的信息准确填入相应的属性字段,并通过实体知识将不同实体类之间建立联系,如通过“hasAge”关系将文物实体类与年代实体类关联,通过“produceRelic”关系将文物与产地实体类关联。此外,根据图像特征相似度建立视觉关系,例如,当 2 幅图像的形状特征相似度超过一定阈值时,在知识图谱中建立“similarImage_shape”关系,表明这 2 个文物在视觉形状上是相似的。

多模态文化遗产知识图谱通过系统化的结构整合了文化遗产的多样化信息,形成了一个全面且互联的知识体系。该图谱有助于文化遗产数据的组织和利用效率,还可以为文化遗产智能问答和个性化检索提供准确的知识支撑。

3 文化遗产领域多模态大模型

3.1 基础多模态大模型架构

本文选择 LLaVA-1.6 作为面向文化遗产领域的多模态基座大模型。LLaVA-1.6 通过端到端训练的方式,巧妙地融合了预训练的大语言模型(LLM)和视觉模型的能力。该模型的突出特点在

于,它将视觉编码器与语言模型通过 2 层多层感知器(MLP)相连,这一架构使得模型能够通过训练 MLP 层精确对齐视觉特征与语言模型,从而实现高效的多模态交互和理解。在中文多模态场景中,LLaVA-1.6 的表现也非常卓越,它在 MMBench-CN 基准测试中取得了业界领先的成绩。

本文提出的“博古问津”文化遗产多模态大模型构建过程如图 1 所示,在构建的文化遗产知识图谱基础上,进一步构建了针对多模态大模型的微调数据集,对 LLaVA-1.6 进行两阶段微调训练。在推理阶段,开发了基于知识检索的增强回答技术,通过在知识图谱中检索与问题相关的文化遗产多模态知识,提升微调模型的推理可靠性。

3.2 面向文化遗产领域的多模态大模型训练

LLaVA-1.6 作为一款先进的多模态基础模型,其设计初衷是为了整合大语言模型(LLM)与视觉模型的能力,以实现高效的多模态交互和深入理解。然而,尽管 LLaVA-1.6 在多模态场景下表现出色,但是直接将它应用于文化遗产领域时会存在一些挑战。文化遗产领域具有其独特的语境和专业术语,这要求模型不仅要有强大的多模态理解能力,还需要对文化遗产的特定知识有深入的掌握。

为了克服这些挑战,通过使用第 2 节描述的文化遗产数据集对 LLaVA-1.6 进行训练,优化模型以适应文化遗产领域。在微调 LLaVA-1.6 时,设计两阶段训练方法。第 1 阶段利用收集的文化遗产领域的结构化多模态数据进行图文对齐学习,增强模型对文化遗产多模态内容的理解和生成能力;第 2 阶段是文化遗产相关问答数据指令微调,这一阶段使模型能够更好地理解用户关于文化遗产的相关问题,并以更加自然和专业的方式与用户进行对话。以下是两阶段训练方法的具体描述。

第 1 阶段是文化遗产领域图文对齐学习。利用构建的文化遗产多模态文化遗产知识图谱,提取与每幅文物图像相关的实体信息和属性,例如文物的名称、年代、材质、颜色、形状等。利用语言大模型 Qwen,将提取到的知识转换为自然语言描述,这些描述在细致刻画文物外观特征的同时,也深入阐述了其历史和文化背景,为图像提供了更为丰富和全面的内容。在训练过程中,模型接收文物图像及其对应的文本描述信息 y_i 作为输入。为了提升模型对图像和文本描述的对齐能力,引

入了图文对比学习损失(contrastive learning loss, 式中简记 L_{CLL})优化图像和文本的编码器。

$$L_{CLL} = \frac{1}{N} \sum_{i=1}^N - \log \left[\frac{\exp(\text{sim}(x_i, y_i) / \tau)}{\sum_{j=1, j \neq i}^N \exp(\text{sim}(x_i, y_j) / \tau)} \right] \quad (1)$$

式中: x_i 和 y_i 分别表示图像编码器输出的图像特征和文本编码器输出的文本特征; N 为一个批次的样本数量; $\text{sim}(x_i, y_i)$ 表示 2 个特征向量之间的相似度度量; τ 是温度参数,用来调整概率分布的锐利程度。这一策略通过优化图像和文本编码器的特征表示,使得相关联的图文对在特征空间中更接近,而不相关的对则更远离。此外,为了使图像编码器更好地与语言模型协同工作,在它们之间加入多层感知器(MLP)层。保持视觉编码器和语言大模型的权重不变,通过只训练连接视觉特征与语言大模型的多层感知器(MLP)层,让语言大模型生成准确的图像描述。这种方法确保了大量的文化遗产信息及其视觉特征得以与预训练的通用 LLM 的文本词嵌入进行有效对齐。

第 2 阶段是对话数据指令微调。为了让语言大模型的回答更加贴合文化遗产领域的专业形式,利用对话数据微调的方式优化模型的输出。首先,采用少样本提示指令方式(few-shot prompt)生成丰富的文化遗产问答训练数据。针对每一幅文化遗产图像,精心设计少样本提示,引导通用多模态大模型产出 3~5 个问答对,这些问答对模拟了用户可能提出的问题,构建出一个既广泛又深入的对话数据集。该数据集不仅捕捉了文化遗产的丰富多样性,还反映了用户交互的多样性。然后,利用这些定制化的问答数据对语言大模型进行微调。为保留通用模型的理解能力并避免过拟合,采用 LoRA(low-rank adaptation)^[32]进行技术微调,LoRA 通过在模型的权重矩阵中引入低秩结构来实现对模型的调整,减少了微调过程中需要更新的参数数量,从而在保持预训练模型知识的同时,有效地适应新任务。通过这种策略,确保模型在文化遗产领域的对话能力得到显著提升,同时保持其在多模态理解方面的卓越性能。

3.3 基于知识图谱的文化遗产多模态问答和检索

在文化遗产领域,通过多模态问答不仅为公众提供了一个便捷的途径了解和探索文化遗产,还为研究人员和教育工作者提供了强大的工具,以支持他们的研究和教学活动。例如,用户上传

一张文物图像,多模态问答系统可以回答图像中文物的朝代、材质、历史等相关知识,亦或者用户可以通过文字描述检索某个文物的相关图像。虽然经过训练的文化遗产多模态大模型已经能够较专业地与用户进行文遗内容的交流,但对于未经学习过的内容,仍存在一定的答非所问、不符合事实等幻觉问题。

为了提升回答的准确性并有效减少幻觉问题,通过引入文化遗产多模态知识图谱作为模型的辅助知识库。模型在回答用户问题前,会先从知识图谱中检索与问题相关的可靠内容,进而结合检索内容进行回答,这一策略可以显著提高模型在回答问题时的可信度和可解释性。具体地,我们设计了 2 种策略在多模态知识图谱中进行检索。

1)多模态知识图谱中的图文检索。首先利用图像和文本编码器提取知识图谱实体的文本和图像特征。当收到用户问题时,提取问题中图像和文本的多模态特征,并与知识图谱中预先提取的特征进行相似度比较,使用余弦相似度作为度

量公式,检索并返回与问题相关的多模态实体信息。这些信息随后被作为附加输入提供给多模态大模型,以促进模型生成更准确的回答。

2)多模态知识图谱中的关系检索。在文化遗产领域,文物之间存在着丰富的联系,这些联系对于理解文物的背景和历史至关重要。为此,设计了一个基于语言大模型(LLM)的智能代理(Agent),通过 few-shot 学习方法,专门用于判断用户的问题是关于文物图像本身,还是涉及图像中文物与其他文物之间的某种关系。如图 3 所示,当用户提出问题时:智能代理首先根据预设的文物节点及其属性信息,以及多种预定义的关系类型(如相同朝代、相同类别、相同博物馆等),判断问题是否需要检索知识图谱中的其他文物信息及所需检索的关系类型;之后,智能代理根据判断结果在知识图谱中检索相应类型的文化遗产信息;最终,智能代理将这些检索到的信息汇总,作为额外的知识背景,增强多模态大模型的回答质量,从而提供更加全面和深入的服务。

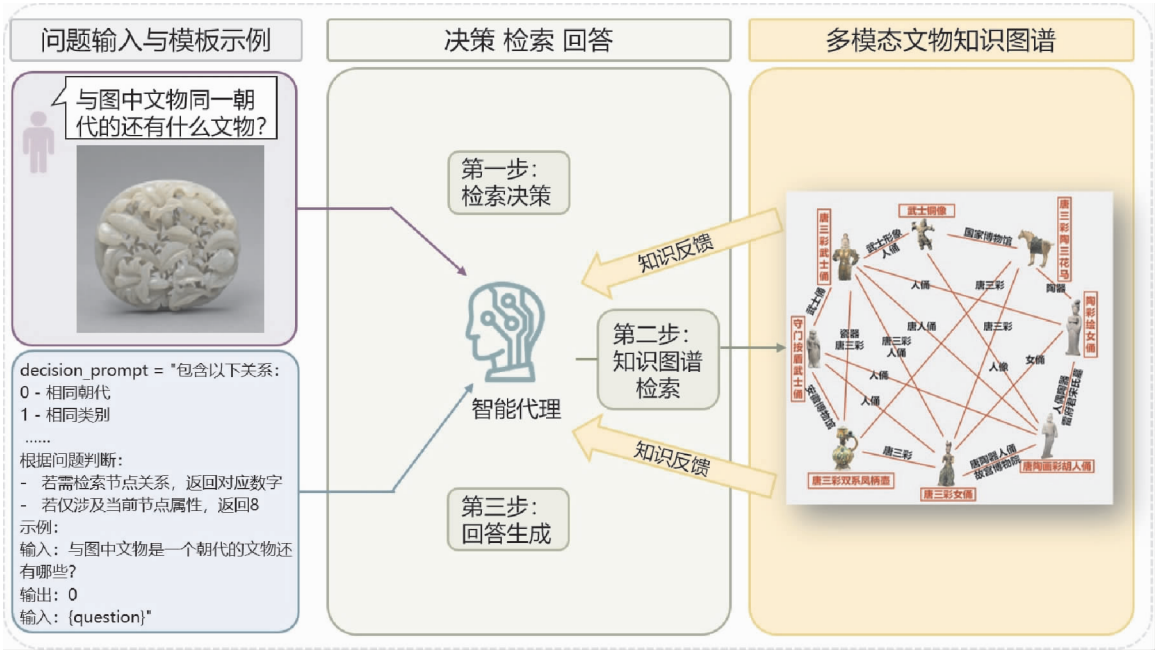


图 3 基于关系检索增强的文化遗产多模态智能问答
Fig. 3 Multimodal intelligent questioning of cultural heritage based on relational retrieval enhancement

利用以上检索策略,不仅能为文化遗产多模态大模型提供充分的可靠知识,减轻模型的推理负担,缓解模型的幻觉问题。同时还可以为用户提供文化遗产领域的跨模态检索功能,如从数据库中检索出与用户输入文物图像相似或者相关的其他文化遗产资源。

4 实验分析

4.1 数据集

本研究构建了一个专门针对文化遗产领域的多模态数据集,以支持和验证所提出的“博古问津”多模态大模型的性能。该数据集的构建方法

在第 2 节进行了详细介绍。此外,为了支持多模态大模型在文化遗产领域的两阶段微调,采用 Qwen 语言大模型对数据库内的文物描述进行格式化处理,转为统一的自然语言描述,最终形成 17 850 组精准匹配的图像与描述对,这些将作为第 1 阶段的微调数据集。随后,进一步利用这些文物的描述信息,生成了相应的问题和答案对,用于第 2 阶段的模型微调,确保数据集的全面性和

深度。在验证环节,选择了 90 个未经训练的图文对,用于针对文物图像描述任务的测试,此外,针对文物属性问题和文物关系问题分别设计了 60 个问答数据。表 3 为数据集及所构建知识图谱的统计信息,图 4 为各类型文物在数据集中的分布情况,表 4 为部分数据的示例。以上数据还经过人工审核,保证数据集的质量和与原始采集数据的一致性。

表 3 数据集统计表
Tab. 3 Dataset statistics

数据集名称	统计量	数量
文化遗产多模态知识图谱	实体	578 个
	关系	7 353 个
	关系类型	6 个
图文对数据集	训练集	17 580 对
	测试集	90 对(文物图像描述)
问答数据集	训练集	17 580 组(每组 3 ~ 5 个问答对话)
	测试集	120 个(文物属性问题和文物关系问题各 60 个)

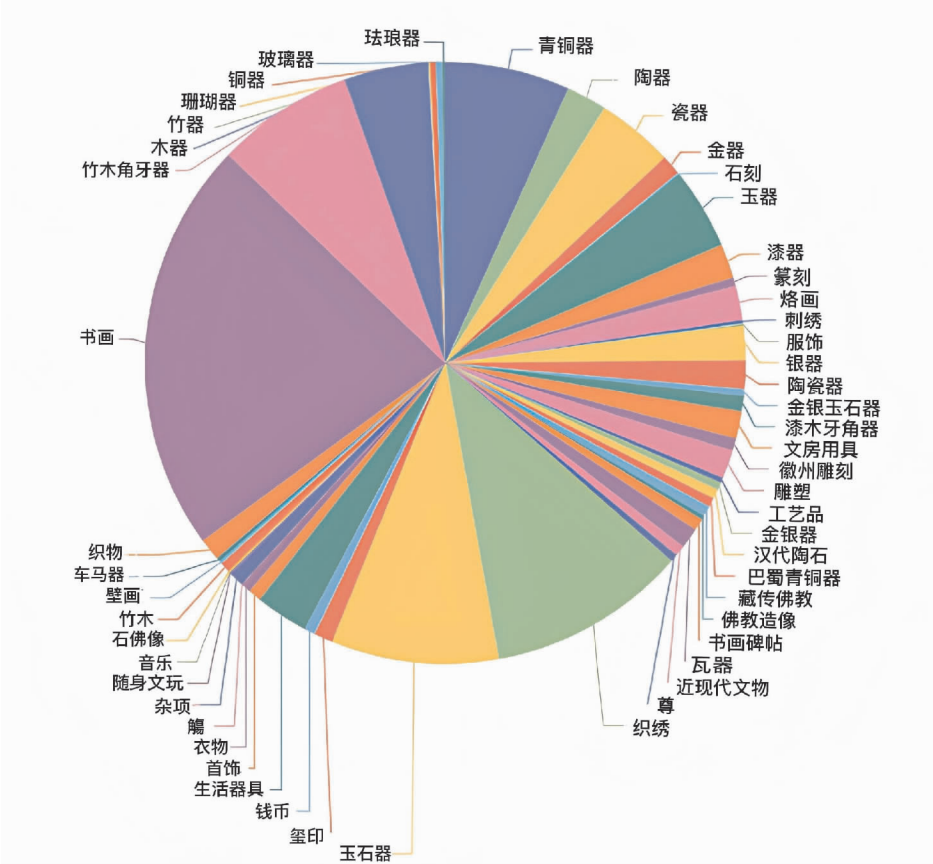


图 4 各类型文物分布图
Fig. 4 Distribution of various types of cultural heritage

表 4 数据集部分示例
Tab. 4 Examples of dataset

名称	图像	文本表述	问答数据
三羊铜罍		“三羊铜罍”是一件来自商朝的青铜器珍品,被首都博物馆精心收藏。这件文物以其独特的敛口和方唇设计,配以颈粗短、折肩、深腹和高圈足的造型,展现出商代中期的典型风格。颈部装饰有两道平行的凸弦纹,肩部则有一周目雷纹,更引人注目的是肩部等距离分布的三个大卷角羊首。……	{“问”:“图片上的青铜器它是什么时代的?”“答”:“这件青铜器是来自商朝的,具有商代中期的典型风格。”} {“问”:“它有什么独特的设计?”“答”:“它的特点是敛口和方唇,配合颈粗短、折肩、深腹和高圈足的造型,特别是肩部的三个大卷角羊首,是通过分铸法铸造的。”}……
鹰首短剑		“鹰首短剑”是一件珍贵的青铜器,源自西周时期,收藏于首都博物馆。这件短剑独具特色,展现了草原文化的鲜明风格。它是在北京昌平白浮村被发掘出土的,其具体尺寸和材质信息未详,但其功能和工艺价值无疑是研究古代历史文化的重要实物证据。	{“问”:“图中展示了什么类型的文物?”“答”:“是一件名叫鹰首短剑的青铜器,源自西周时期。”} {“问”:“它是在哪里被发现的?”“答”:“鹰首短剑是在北京昌平白浮村被发掘出土的。尽管具体尺寸和材质信息未详,但其考古价值非常高。”}……
白玉双螭耳杯		“白玉双螭耳杯”是一件出自明代的精美玉器,收藏于首都博物馆。这件玉杯充分展现了明代玉器的艺术特色,是该时期玉器的典型代表。杯体两侧精心镂雕了螭虎图案,其螭首上还刻有阴文“王”字,“透露出些许元代的遗风。……	{“问”:“介绍下它的雕刻样式?”“答”:“杯体两侧有螭虎图案,螭首上有阴文‘王’字,这是明代玉器的典型装饰。”} {“问”:“谁是这个玉杯的拥有者?”“答”:“这个玉杯的主人是万贵,明宪宗宠爱的万贵妃的父亲,他在 1392 年至 1475 年间生活。”}……

4.2 实验细节及环境

实验在配备有 2 个 RTX 3090 24G GPU 的服务器上,选择 llava-v1.6-mistral-7b 作为基础多模态大模型。在微调过程中使用 LoRA (low-rank adaptation) 技术^[21],通过在模型权重矩阵中引入低秩结构实现对模型的精细调整,减少微调过程中需要更新的参数数量。第 1 阶段图文对齐学习设置 LoRA 的 rank 为 128, alpha 为 256。学习率设置为 2×10^{-5} , 批量大小为 1, 训练轮数设为 10, 使用 bf16 精度。第 2 阶段对话指令微调,只冻结视觉编码器,其他设置与第 1 阶段相同。在选择两阶段训练的超参数时,综合考虑了多方面因素。首先,基于文化遗产数据集的特点,其包含丰富的文物图像与文本描述等多模态信息,且文物历史背景等信息具有独特性和复杂性,这就要求模型在训练过程中需要对细节有较高的关注度,因此选择较小的学习率,以便模型能够更细致地学习文物图像与文本的特征表示。其次,由于显存资源有限,设置批量大小为 1,避免出现显存不足。最后,选择 10 轮训练可以在模型的拟合能力和泛化能力之间取得较好的平衡,以及结合

bf16 精度计算的优势,能够有效减少计算资源的消耗,提高训练效率,使模型在训练过程中能够更快地进行参数更新和迭代,从而加速模型的收敛过程。在训练时间方面,基础多模态大模型经过两阶段训练,其中,第 1 阶段的图文对齐学习耗时约 9 天 7 小时,第 2 阶段的对话指令微调耗时约 8 天 14 小时。

4.3 评价标准

本节设计了一套评价标准衡量“博古问津”多模态大模型的性能。首先,选用具有 72 亿参数的语言大模型 Qwen2-72b,在给定文物的准确相关信息下,生成针对描述问题的标准答案,这些标准答案被视作模型性能的理论上限。随后,对于待评测的多模态大模型,针对同样的问题生成回答。最终,将标准答案作为提示,利用 Qwen2-72b 根据有用性、相关性、准确性和细节级别这 4 个指标对评测模型的回答进行评分,每个指标的评分范围是 1 ~ 10 分,分数越高代表性能越好。有用性评价的是回答对用户问题的实用性;相关性考察的是回答与问题的匹配度;准确性关注的是回答中事实信息的正确性;细节级别则衡量回答是

否提供了足够的信息来全面回答问题。Qwen2-72b 在 4 个方面评分的同时,还给出 1~100 的综合分值。为了消除大模型评分的随机性波动,并使评分结果更具可比性,将候选模型的得分除以标准答案的得分,得到一个标准化的百分比作为相对评分,其分值越大表示越接近标准答案。

4.4 消融实验

在本研究中,消融实验被用来评估“博古问津”多模态大模型中不同组件对模型性能的贡献。实验设计涵盖了基础模型 LLaVA 及其经过不同阶段训练和不同知识检索方式的变体,包括仅进行图像与文本对齐学习的微调模型,以及在此基础上进一步接受对话指令微调的全微调模型。此外,还对比了在这 2 个训练阶段之后引入不同知识检索机制的模型。实验结果如表 5 所

示,基础模型 LLaVA 在文物图像描述、文物图像属性问题和文物关系问题上的性能较低,通过图文对齐学习,LLaVA 在描述图像任务上性能有显著提升。引入对话指令的全微调进一步增强了模型在图像属性问题和图像关系问题上的性能。而引入知识检索机制后,模型在所有任务上的性能显著提升。其中,虽然关系检索会引入一些额外信息影响“博古问津”模型在文物属性问题的性能,但在文物关系问题上达到了 82.9% 的性能,表明关系检索在提升模型综合性能方面具有重要作用。这些消融实验证实了图文对齐学习和对话指令微调对于提升模型在文化遗产领域的多模态任务中的重要性,并验证了基于知识的检索机制能够有效提高模型的准确性和可靠性。

表 5 所提出方法消融实验结果
Tab. 5 Ablation experiments of the proposed method

模型	文物图像描述		文物属性问题		文物关系问题	
	综合分值	相对评分	综合分值	相对评分	综合分值	相对评分
基础模型 (LLaVA)	21.7	21.7	22.5	22.7	30.8	37.0
LLaVA + 图文对微调	32.6	32.8	23.3	23.6	28.3	33.7
LLaVA + 全微调	30.4	30.7	49.5	50.7	43.7	56.5
LLaVA + 全微调 + 图文检索	54.0	56.8	<u>73.0</u>	<u>75.3</u>	34.3	45.8
LLaVA + 全微调 + 关系检索	33.9	34.3	56.0	57.4	<u>59.0</u>	<u>77.6</u>
LLaVA + 图文对微调 + 全检索	65.0	70.5	64.2	71.1	56.3	74.6
博古问津	<u>55.0</u>	<u>57.8</u>	74.1	75.7	61.2	82.9

注:加粗表示每列最优结果;下划线表示次优结果。

4.5 方法比较

本节我们将“博古问津”多模态大模型与主流的 LLaVA^[14] 和 Qwen-VL^[12] 系列开源多模态大模型进行比较,包括 Qwen-VL-Chat-V1、Qwen-VL-V1、Qwen-VL-Plus 和 Qwen-VL-Max。所有模型都在相同的文物图像描述、文物属性问题和文物关系问题测试集上进行评估。从表 6 可见,与现有的开源多模态大模型相比,经过领域训练和知识

检索增强的“博古问津”模型在所有任务上均显示出了优越的性能,分别在文物图像描述、文物属性问题和文物关系问题 3 个不同任务的综合分值上高出次优模型 21.4%、53% 和 20.6%。这一结果表明,本文所提出的针对文化遗产领域定制的微调和知识增强策略能够有效提升多模态大模型对文化遗产领域适应性。

表 6 对比其他多模态大模型的实验结果
Tab. 6 Experimental results compared with other multimodal models

模型	文物图像描述		文物属性问题		文物关系问题	
	综合分值	相对评分	综合分值	相对评分	综合分值	相对评分
Qwen-VL-Chat-V1	21.9	22.0	19.5	19.6	39.8	54.6
Qwen-VL-V1	22.4	22.5	19.3	19.4	39.5	54.1
Qwen-VL-Plus	31.4	31.9	17.1	17.2	37.3	49.1
Qwen-VL-Max	<u>35.1</u>	<u>35.4</u>	18.5	18.5	<u>46.0</u>	<u>62.3</u>
LLaVA	21.7	21.7	<u>22.5</u>	<u>22.7</u>	30.8	37.0
博古问津	54.0	56.8	74.1	75.7	61.2	82.9

注:加粗表示每列最优结果;下划线表示次优结果。

图 5 是与其他多模态大模型在有用性、相关性、准确性和细节级别这 4 个方面的具体比较,结

果表面“博古问津”模型在 3 个任务的 4 个不同方面均明显优于对比的通用多模态大模型。

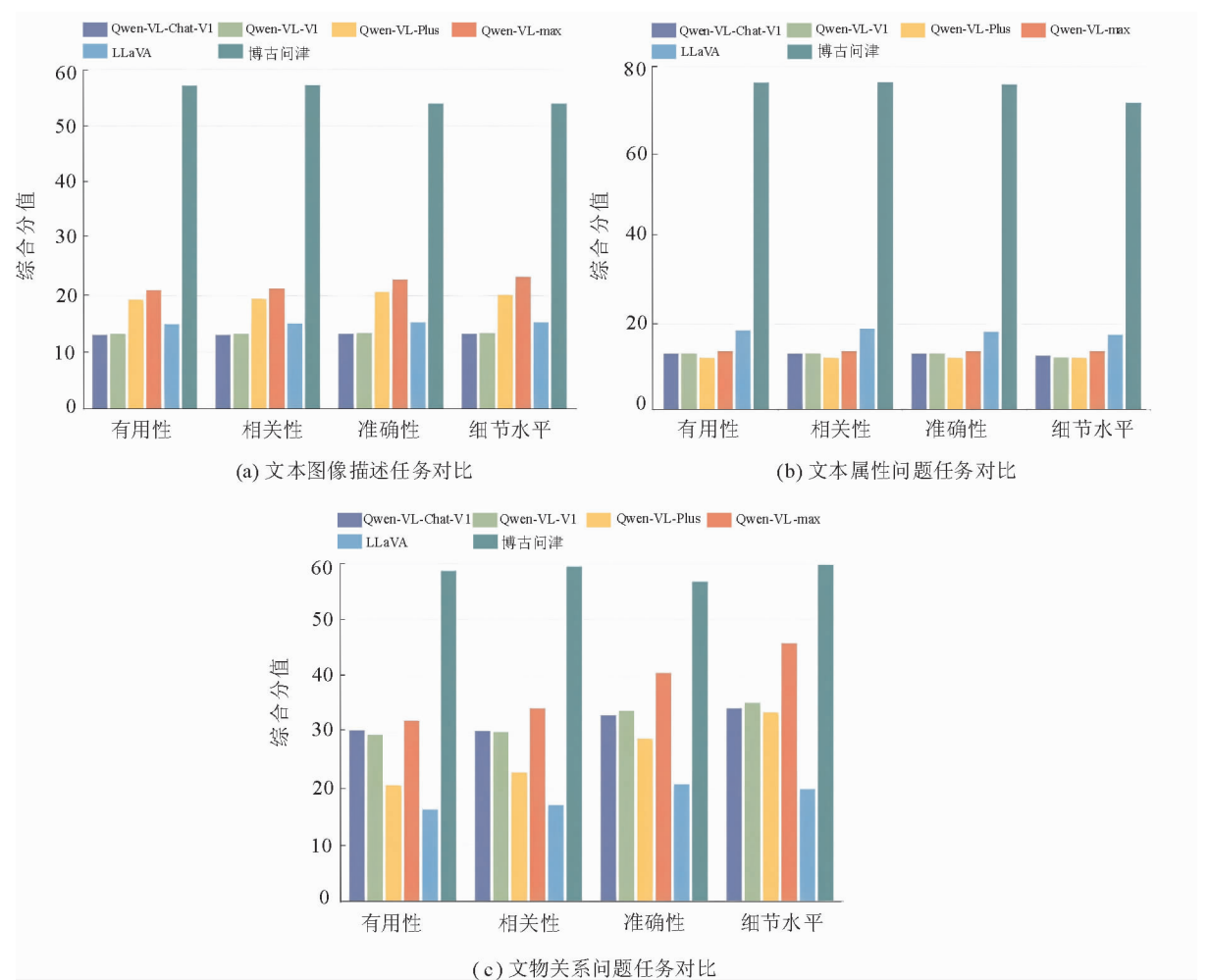


图 5 不同评价指标上的对比直方图
Fig. 5 Histograms on different evaluation metrics

4.6 示例分析

本节通过一系列具体的问答示例,展示“博古问津”多模态大模型在处理文化遗产相关问题时的有效性。这些示例包括对文物图像的描述(见表 7)、询问文物的具体属性(见表 8)以及探询文物间的关系(见表 9)等问题,均从测试集中选取。例如,当用户询问某件文物的材质时,“博古问津”能够精准地回答该文物为青白玉质。在涉及文物详细信息的问题时,模型不仅能准确报出文物的名称,还能一并提供其所属朝代、尺寸规格以及样式图案的深层含义,且表述流畅自然。此外,在处理文物关系问题时,如用户询问与某件文物同馆的其他展品,“博古问津”能列举出绿釉陶仓、母己爵等实例,并明确指出其回答依据,体现出回答的可信性。这些示例体现“博古问津”在回答的精准度与相关性上,与参考答案不相上

下,甚至在某些情境下,其回答更显直接与精炼。与通用模型 LLaVA 相比,经过文化遗产领域的深度微调后,模型对专业问题的理解与回应能力得到了显著提升。特别地,针对表 9 中“与图中文物同属一个历史时期且具有相似工艺特征的文物有哪些?”这一复杂问题时,“博古问津”多模态大模型融合多模态知识图谱,能够快速关联文物的图像特征、历史背景和工艺特征。这种能力使其在处理复杂问题时,能够从多个维度进行推理和判断,而不仅仅是依赖单一的文字信息或图像信息。与 LLaVA 仅能提供图中文物的名称和时代信息相比,“博古问津”能够列举出具有相似工艺特征的文物,如“三犀鼎”“三鸟鼎”“虎纹鼎”等,体现模型在推理过程中巧妙融合的多模态知识图谱,使其在应对文物关系等复杂问题时,展现出了卓越的处理能力。

表 7 文物图像描述示例对比

Tab. 7 Examples of cultural heritage image descriptions

文物图像		描述示例
	问题	非常详细地阐明所显示文物图像的内容
	LLaVA	这是一幅展示中华古代文物的图像。文物是一种古代的陶器,其表面上展现了一些人物和动物形象,以及一些花瓣和叶子的装饰。文物的颜色是黑色,而且有一些金色的装饰,使其更加吸引人的。这种文物通常被称为“玄琬”,是中华古代文化中非常重要的一种艺术形式 这件紫檀木镂雕会昌九老图笔筒是明代的,以紫檀木制成,通高 19 厘米,口径 14 厘米,底径 16 厘米。它的设计独特,笔筒口沿嵌银丝勾连菊花纹,近口处以螺钿镶嵌狮纹及葡萄纹,外壁镂雕《会昌九老图》。底座如岩石状,与筒身景物相呼应。壁身描绘的是唐会昌年间(841—846 年)白居易等 9 位文人在洛阳龙门香山寺宴集的情景。它以高浮雕和圆雕为主,刀法快利,磨工显得草率,螺钿装饰重厚,整体带有鲜明的明代紫檀雕刻风格
	博古问津	
	参考答案	所示文物图像为“紫檀木镂雕会昌九老图笔筒”,这是一件源自明代的珍贵艺术品,目前珍藏于故宫博物院。该笔筒的尺寸为通高 19 厘米,口径 14 厘米,底径 16 厘米。其口沿部分装饰有嵌银丝的勾连菊花纹,而近口处则镶嵌有螺钿制成的狮纹和葡萄纹,这些细节处理得极为精细,彰显出匠人的精湛技艺。笔筒的外壁采用高浮雕和圆雕技法,生动地再现了《会昌九老图》中的场景。此图描绘的是唐会昌年间(841—846 年),著名诗人白居易与八位文人雅士在洛阳龙门香山寺宴集的情景,他们围坐一桌,或谈笑风生,或举杯畅饮,或沉思冥想,画面充满了浓厚的文化氛围和生活气息。底座部分设计成岩石形状,巧妙地与筒身的景物融为一体,增强了整个作品的立体感和观赏性。尽管在雕刻过程中,刀法略显快利,磨工也有些许草率,但螺钿装饰的厚重感却赋予了笔筒一份庄重与华贵。从整体上看,无论是造型设计还是图纹刻划,这件“紫檀木镂雕会昌九老图笔筒”都充分展现了明代紫檀雕刻的独特风格,堪称一件不可多得的艺术瑰宝
	问题	精确地叙述文物图像的内容
	LLaVA	您好! 这是一件古代的宝藏,它是一种用于储存和保护文物的宝藏。这个宝藏的外观非常精美,它是由金和宝石制成的,并且有着精致的雕刻和墨画。宝藏的顶部有一个圆形的墙壁,上面有一些文字和图案,而下面则有一些花纹和花朵。宝藏的底部则是一种金属的材料,上面有一些花纹和花朵的设计。宝藏的颜色是金色和蓝色,它的表面非常光滑,表现出一种古老的美感。宝藏的设计非常精致,它的雕刻和墨画都非常精细,表现出一种高雅的文化气息。宝藏的材料和设计都是古代的特色,它是一种非常珍贵的文物
	博古问津	
		这件银烧蓝累丝圆盒是清代中期的银器珍品,以粗银丝作为骨架,细银丝累掐成精美的八瓣菊花和卷草纹样,填充了烧绿透明硬珐琅,显得极具艺术感和历史价值。出土地点是清宫旧藏,尺寸小巧,高 3.9 厘米,口径 9.5 厘米,足径 6.1 厘米,工艺细致,体现了清代工艺的高超技艺

续表 7

文物图像	描述示例
	参考答案

“银烧蓝累丝圆盒”是一幅展现清朝珍贵银器的图像,此文物由故宫博物院珍藏。图像中的圆盒以银为胎,利用粗银丝搭建骨架,再以细银丝精心编织出复杂的图案,展示了精湛的工艺技术。盒盖边缘分为八组,每组均饰以累丝卷草和八瓣菊花,掐丝卷须焊接于底部,其间填充了烧绿透明硬珐琅,色泽温润,晶莹剔透。下层装饰着累丝卷草花,而上层则呈现掐丝八瓣菊花与卷草的组合,其中叶子部分巧妙地填充了蓝色和绿色透明珐琅,色彩对比鲜明,层次分明。器物下方配有一圈圈足,稳固而雅致。整个器物将银累丝与透明珐琅完美融合,展现了清代中期的艺术风格,作为清宫内廷陈设品,其稀有性和艺术价值不言而喻。这件文物的具体尺寸为:口径 9.5 厘米,足径 6.1 厘米,高 3.9 厘米,是清宫旧藏中的一件瑰宝,极具观赏和研究价值

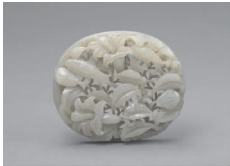
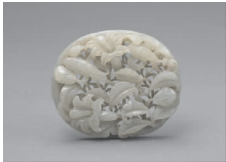
表 8 文物属性问答示例对比		
Tab. 8 Comparison of Q&A on cultural heritage attributes		
文物图像	文物属性问答示例	
	问题	图中的文物叫什么名字?
	LLaVA	这个文物被称为“古代的瓶子”
	博古问津	图中的文物名为“陵阳壶”
	参考答案	图中的文物名为“陵阳壶”,这是一件来自战国时期的珍贵青铜器,现收藏于首都博物馆
	问题	这个文物是什么材质?
	LLaVA	这个文物是白色的,可能是纯白色的瓷或玻璃
	博古问津	这件文物是青白玉质的
	参考答案	这个文物的材质是青白玉

表 9 文物关系问答示例对比		
Tab. 9 Comparison of Q&A on heritage relationships		
文物图像	文物关系问答示例	
	问题	与图中文物同一朝代的还有啥?
	LLaVA	在图中,你可以看到一种古老的中国文物——琢器。琢器是中国古代文化中非常重要的一种艺术品。它通常是由琥珀、玉石、玻璃等材料制成的,通常用于宴会、礼仪、祭祀等场合。琢器的形状和设计都非常多样,可以表现出各种风格和艺术形式。在这个图片中,琢器的设计非常精致,表现出了不同的花瓣和叶子形状,展现出了中国古代艺术的精美和诗意
	博古问津	根据知识图,同一朝代的其他文物包括剔红枫叶秋虫图盒
	参考答案	根据知识图谱信息,与图中文物在同一朝代的还有剔红枫叶秋虫图盒

续表 9

文物图像		文物关系问答示例	
	问题	与图中文物在同一博物馆的还有啥?	
	LLaVA	非常抱歉,我无法提供关于博物馆中的其他文物的信息,因为我没有访问博物馆的权限。我只能提供关于图片中的文物的信息。如果您有其他问题,我很乐意帮助您	
	博古问津	在同一博物馆的其他展品包括绿釉陶仓、母己爵、羊脂玉凌霄花饰、镶嵌红铜鸟兽纹壶、镶贝壳鎏金鹿	
	参考答案	根据知识图谱的信息,与图中文物在同一博物馆的还有绿釉陶仓、母己爵、羊脂玉凌霄花饰、镶嵌红铜鸟兽纹壶、镶贝壳鎏金鹿	
	问题	与图中文物同属一个历史时期且具有相似工艺特征的文物有哪些?	
	LLaVA	图中文物是三鸟簋,它是战国时期的	
	博古问津	在战国时期,与三鸟簋具有相似工艺特征的文物主要是指那些同样展现了高度发达的青铜铸造技术和复杂装饰艺术的作品。有三犀鼎、三鸟鼎、虎纹鼎	
	参考答案	战国时期的青铜器如三鸟簋一样,展现了高超的铸造技术和复杂的装饰艺术,同期的代表作品还包括三犀鼎、三鸟鼎、虎纹鼎、战国羽纹簋、曾侯乙编钟以及燕国特色的青铜器等,它们均体现了这一时期青铜器物在设计和工艺上的卓越成就	

5 结语

本文提出了一个面向文化遗产领域的多模态大模型——“博古问津”,通过构建大规模的多模态文化遗产数据集和知识图谱,实现了对文化遗产的全面理解和高效利用。利用 LLaVA-1.6 作为基础模型,经过两阶段微调训练,定制开发了适合文化遗产领域的多模态大模型,显著提升了模型在图文对齐学习和对话能力方面的表现。此外,引入知识图谱作为辅助知识库,通过图文检索和关系检索策略,有效提升了模型在问答任务中的可信度和可解释性。实验结果表明,与现有的开源多模态大模型相比,“博古问津”在文化遗产领域的多模态对话任务上展现出显著的性能优势,其准确度达到了 82.9%,与标准答案的匹配度极高。未来的工作将进一步提升模型的泛化能力,探索跨领域应用的可能性,并优化用户交互体验,以期在文化遗产保护和推广中发挥更大的作用。

参考文献

[1] KASNECI E, SESSLER K, KÜCHEMANN S, et al.

[7] CUI C, MA Y S, CAO X, et al. A survey on multi-

ChatGPT for good? On opportunities and challenges of large language models for education[J]. Learning and Individual Differences, 2023, 103: 102274.

[2] NAVEED H, KHAN A U, QIU S, et al. A comprehensive overview of large languagemodels[EB/OL]. (2023-11-27)[2024-12-20]. https://arxiv.org/abs/2307.06435.

[3] CHANG Y P, WANG X, WANG J D, et al. A survey on evaluation of large language models[J]. ACM Transactions on Intelligent Systems and Technology, 2024, 15(3): 1-45.

[4] KOE HWEE NGA J, WAI MUN S. The influence of MLM companies and agent attributes on the willingness to undertake multilevel marketing as a career option among youth[J]. Journal of Research in Interactive Marketing, 2011, 5(1): 50-70.

[5] YIN S K, FU C Y, ZHAO S R, et al. A survey on multimodal large language models[J]. National Science Review, 2024, 11(12): nwae403.

[6] ZHANG D Z, YU Y H, DONG J H, et al. MM-LLMs: Recent advances in multimodal large language models[EB/OL]. (2024-05-28)[2024-12-20]. https://arxiv.org/abs/2401.13601.

- modal large language models for autonomous driving [C]//2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). Waikoloa, HI, USA: IEEE, 2024: 958-979.
- [8] LIPKOVA J, CHEN R J, CHEN B W, et al. Artificial intelligence for multimodal data integration in oncology[J]. *Cancer Cell*, 2022, 40(10): 1095-1110.
- [9] SHARMA K, GIANNAKOS M. Multimodal data capabilities for learning: What can multimodal data tell us about learning? [J]. *British Journal of Educational Technology*, 2020, 51(5): 1450-1484.
- [10] OPENAI, ACHIAM J, ADLER S, et al. GPT-4 technical report [EB/OL]. (2024-03-04) [2024-12-20]. <https://arxiv.org/abs/2303.08774>.
- [11] LIU H T, LI C Y, WU Q Y, et al. Visual instruction tuning [EB/OL]. (2024-11-11) [2024-12-20]. <https://arxiv.org/abs/2304.08485>.
- [12] ASKELL A, BAI Y T, CHEN A N, et al. A general language assistant as a laboratory for alignment [EB/OL]. (2021-11-09) [2024-12-20]. <https://arxiv.org/abs/2112.00861>.
- [13] GAN Z, LI L J, LI C Y, et al. Vision-language pre-training: Basics, recent advances, and future trends [J]. *Foundations and Trends® in Computer Graphics and Vision*, 2022, 14(3/4): 163-352.
- [14] BHATIA G, NAGOUDI E M B, CAVUSOGLU H, et al. FinTral: A family of GPT-4 level multimodal financial large language models [EB/OL]. (2024-06-14) [2024-12-20]. <https://arxiv.org/abs/2402.10986>.
- [15] LI C Y, WONG C, ZHANG S, et al. LLaVA-med: Training a large language-and-vision assistant for bio-medicine in one day [EB/OL]. (2023-06-01) [2024-12-20]. <https://arxiv.org/abs/2306.00890>.
- [16] 程启玥. 我国文化遗产领域数字人文研究热点与趋势探究[J]. *现代信息科技*, 2024, 8(14): 106-111.
- CHEN Q Y. Research hotspots and trends of digital humanities in the field of cultural heritage in China [J]. *Modern Information Technology*, 2024, 8(14): 106-111.
- [17] MITTAL R, GARG A. Text extraction using OCR: A systematic review [C]//2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). Coimbatore, India: IEEE, 2020: 357-362.
- [18] 徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述 [J]. *电子科技大学学报*, 2016, 45(4): 589-606.
- XU Z L, SHENG Y P, HE L R, et al. A survey on knowledge graph techniques [J]. *Journal of University of Electronic Science and Technology of China*, 2016, 45(4): 589-606.
- [19] 杜鹏飞, 李小勇, 高雅丽. 多模态视觉语言表征学习研究综述 [J]. *软件学报*, 2021, 32(2): 327-348.
- DU P F, LI X Y, GAO Y L. Survey on multimodal visual-linguistic representation learning [J]. *Journal of Software*, 2021, 32(2): 327-348.
- [20] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [EB/OL]. (2019-05-24) [2024-12-20]. <https://arxiv.org/abs/1810.04805>.
- [21] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, BC, Canada: ACM, 2020: 1877-1901.
- [22] TOUVRON H, LAVRIL T, IZACARD G, et al. LLaMA: Open and efficient foundation language models [EB/OL]. (2019-05-24) [2024-12-20]. <https://arxiv.org/abs/2302.13971>.
- [23] BAI J Z, BAI S, CHU Y F, et al. Qwen technical report [EB/OL]. (2019-05-24) [2024-12-20]. <https://arxiv.org/abs/2309.16609>.
- [24] YANG Z, LI L, LIN K, et al. The dawn of LMMs: Preliminary explorations with GPT-4V (ision) [EB/OL]. (2023-10-11) [2024-12-20]. <https://arxiv.org/abs/2309.17421>.
- [25] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision [EB/OL]. (2021-02-26) [2024-12-20]. <https://arxiv.org/abs/2103.00020>.
- [26] ZHU D Y, CHEN J, SHEN X Q, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models [EB/OL]. (2023-04-20) [2024-12-20]. <https://arxiv.org/abs/2304.10592>.
- [27] CHIANG W L, LI Z H, LIN Z, et al. Vicuna: An open-source chatbot impressing GPT-4 with 90% * ChatGPT quality [EB/OL]. (2023-03-30) [2024-12-20]. <https://vicuna.lmsys.org>.
- [28] BAI J Z, BAI S, YANG S S, et al. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond [EB/OL]. (2023-10-13) [2024-12-20]. <https://arxiv.org/abs/2308.12966>.
- [29] ZENG A H, XU B, WANG B W, et al. ChatGLM: A family of large language models from GLM-130B to

GLM-4 all tools [EB/OL]. (2024-06-18) [2024-12-20]. <https://arxiv.org/abs/2406.12793>.

[30] HARTSOCK I, RASOOL G. Vision-language models for medical report generation and visual question answering: A review[EB/OL]. (2024-04-15) [2024-12-20]. <https://arxiv.org/abs/2403.02469>.

[31] HAN T Y, LIAN Q, PAN R, et al. The instinctive bias: Spurious images lead to illusion in MLLMs[C]// Proceedings of the 2024 Conference on Empirical

Methods in Natural Language Processing. Miami, FLorida, USA: ACL, 2024: 16163-16177.

[32] HU E J, SHEN Y, WALLIS P, et al. LoRA: Low-rank adaptation of large language models [EB/OL]. (2021-10-16) [2024-12-20]. <https://arxiv.org/abs/2106.09685>.

(编辑 李 静)

作者简介



赵万青,工学博士,西北大学电子信息学院(人工智能学院)教授,博士生导师。中国图象图形学学会数字文化遗产专委会委员、陕西计算机学会理事、陕西图象图形学学会理事、中国计算机学会会员。主要研究方向为智能媒体检索、机器学习和计算机视觉。已发表学术论文四十余篇,其中以第一作者或通信作者在人工智能国际顶级期刊会议(*IEEE Transactions on Image Processing*、CVPR、AAAI、IJCAI)上发表多篇学术论文并做大会报告,作为主持人或骨干成员参与多项国家自然科学基金和国家重点研发计划等项目,授权发明专利及软件著作权十余项,两次获陕西省高校科学技术奖一等奖,日本大阪大学访问学者,担任人工智能国际会议 IJCAI2021 程序委员会成员,国际会议 CVPR、AAAI、IJCAI 审稿人。