



上海大学

SHANGHAI UNIVERSITY

本科毕业论文（设计）

UNDERGRADUATE THESIS (PROJECT)

题 目：融合 RAG 机制的多模态医学大模型微调

学 院：计算机工程与科学学院

专 业：人工智能

学 号：21122890

学生姓名：黎思源

指导教师：张惠然

起讫日期：2025.02.17 - 2025.05.31



上海大学本科毕业论文（设计）

姓 名：黎思源

学号：21122890

论文题目：融合 RAG 机制的多模态医学大模型微调

原创性声明

本人声明：所呈交的论文是本人在指导教师指导下进行的研究工作。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

签 名：黎思源 日期：2025 05.22

本论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

（保密的论文在解密后应遵守此规定）

签 名：黎思源 指导教师签名：张 芝 日期：2025 05.22

摘 要

本研究面向医学人工智能中的多模态对话任务，构建了一个结合多模态 RAG 机制与大语言模型微调的医学内容生成系统。整体方案支持图文联合输入，通过构建多模态向量数据库、设计不同的 RAG 检索范式，并配合多模态语言模型的领域微调，支持医学场景下的图文联合理解与内容生成。系统已开源在 Github 平台。

系统融合了 RAG 的外部知识检索能力与模型微调的适应性，尝试缓解医学数据在准确性、表达多样性与样本分布等方面的挑战。针对医学任务特点，研究设计并实现了图文共嵌、标签代理检索与图文对绑定等多种 RAG 范式，其中“图文对绑定”方法在保持视觉与语言信息协同方面表现较好，提升了生成内容的语义一致性。

在模型训练方面，采用了分阶段微调策略。第一阶段基于中文医学问答数据提升语言理解，第二阶段则在图文对数据上强化视觉语义对齐，最终得到具备医学任务适应能力的 Qwen2.5-VL-Med 模型。实验在 OmniMedVQA 基准数据集上验证了模型性能，在 Question-Answering Score 和 Prefix-based Score 评估指标上分别以 51.36 和 30.08 分的成绩达到对比模型中第一和第二的水平。

系统采用模块化设计，支持本地部署与云端 API 调用，可适应不同部署需求。前端提供基于 Web 的交互界面，支持图文输入、流式响应与对话历史管理，初步具备实际医疗场景的应用潜力。整体而言，本研究在多模态 RAG 设计、模型适配与系统集成方面提供了一种可行方案，为医学 AI 在内容生成方向的探索提供了一定参考价值。

关键词：检索增强生成；多模态；大语言模型；医学

ABSTRACT

This study focuses on multimodal dialogue tasks within the field of medical artificial intelligence and presents a medical content generation system that integrates a multimodal RAG mechanism with fine-tuning of a large language model. The system supports joint image-text input by building a multimodal vector database, designing multiple RAG retrieval paradigms, and applying domain-specific fine-tuning of a multimodal language model, enabling joint understanding and content generation in medical contexts. The project has been open-sourced on GitHub.

By combining the external knowledge retrieval capability of RAG with the adaptability of model fine-tuning, the system seeks to address challenges in medical data related to factual accuracy, expressive diversity, and imbalanced distribution. To accommodate the characteristics of medical tasks, the study implements several RAG paradigms, including joint image-text embedding, label-guided retrieval, and image-text pair binding. Among them, the “image-text pair binding” approach demonstrates better coordination between visual and textual semantics, contributing to improved consistency in generated content.

For model training, a staged fine-tuning strategy is adopted. The first stage enhances the model’s language understanding using a large-scale Chinese medical QA dataset, while the second stage focuses on aligning visual and textual semantics using paired image-text data. The resulting model, named Qwen2.5-VL-Med, is evaluated on the OmniMedVQA benchmark, achieving 51.36 in the Question-Answering Score and 30.08 in the Prefix-based Score, ranking first and second respectively among compared methods.

The system is designed modularly, supporting both local deployment and cloud-based API access, allowing for flexible adaptation to different use scenarios. The web-based front-end interface supports image-text input, streaming responses, and dialogue history management, showing initial potential for real-world medical applications. Overall, this study offers a practical solution in the areas of multimodal RAG design, model adaptation, and system integration, and may serve as a reference for future exploration of medical AI in content generation.

Keywords: Retrieval-Augmented Generation; Multimodality; Large Language Model; Medicine

目 录

摘 要 I

ABSTRACT II

1 绪论 1

1.1 研究背景与意义 1

1.2 国内外研究现状 3

1.2.1 医学数据特点与挑战 3

1.2.2 RAG 及其医学应用 5

1.2.3 LLMs 及其医学应用 7

1.3 主要研究内容与创新点 11

1.4 论文章节结构安排 13

2 相关技术及算法 15

2.1 检索增强生成 15

2.1.1 RAG 架构 15

2.1.2 多模态 RAG 18

2.1.3 向量数据库 19

2.2 大语言模型相关算法 21

2.2.1 Transformer 架构 21

2.2.2 Vision Transformer 架构 24

2.2.3 Qwen2.5-VL 模型 27

3 系统架构 29

3.1 总体架构 29

3.2 RAG 系统架构 31

3.3 基于 Qwen2.5-VL 的大模型微调 36

3.3.1 医学语义理解能力微调 37

3.3.2 医学图像理解能力微调 39

3.4 API 调用 41

4 实验	42
4.1 基准测试	42
4.2 范式对比与消融实验	43
4.3 模型评估与分析	46
4.3.1 基于任务类型评估	46
4.3.2 基于成像模态评估	48
4.4 实验总结	53
5 交互设计	55
5.1 终端交互	55
5.2 前端交互实现	56
5.2.1 网页交互界面	56
5.2.2 Qwen2.5-VL 交互界面	58
6. 结论与展望	60
6.1 结论	60
6.2 展望	61
参考文献	63
致 谢	75

1 绪论

1.1 研究背景与意义

在医学知识快速增长、临床决策日益复杂的今天，如何高效获取并整合异质医学信息，辅助医生进行诊断、分析与判断，已成为医学信息学、人工智能乃至医学实践共同关注的课题[1]。传统医学知识获取主要依赖专家经验和文献查阅[2]，效率有限，且在处理跨模态、跨语义的信息时常面临理解偏差与信息孤岛的问题[1][3]。在此背景下，如何从复杂的医学数据中提取有意义的知识，为临床诊断、辅助决策乃至医学科研提供支持，成为医学人工智能领域的重要研究课题。近年来，伴随深度学习与自然语言处理技术的高速发展，大规模预训练模型（Pretrained Models）尤其是大语言模型（Large Language Models, LLMs）逐渐显示出在语义理解、推理生成等方面的强大潜力[5]。与此同时，基于多模态输入的智能系统也开始向更高层次的语义整合与决策支持迈进[7]。然而，医学场景中存在的数据来源的异质性、应用的高风险性与信息安全等问题，使得现有语言模型在面对真实医学问题时仍面临诸多挑战[2][4][6]。

传统基于语言模型的生成系统在处理复杂任务时往往依赖内建的知识能力，但这类模型的知识储备具有静态性和时效性限制，难以满足医疗信息日新月异、任务语境多变的需求[8][9]。尤其在医学场景中，患者症状、影像结果、检验数据等往往需同时考虑到[10]，而大语言模型由于训练语料的非特异性，在知识覆盖与事实一致性方面存在明显短板[8]。也即是，大语言模型缺乏在医学场景中进行判断的基本方法。以医学影像和电子病历为例，前者主要承载结构化或图像化的视觉信息[11]，而后者则承载医生的主观判断、患者的主诉与临床指标等语言信息[12]。两者融合后的数据空间不仅内容量多、模态复杂，而且隐含语义高度专业化、非通用，给信息处理提出了极高的要求。

近年来逐渐发展起“增强检索生成”架构（Retrieval-Augmented Generation, RAG）[15]成为一种解决上述部分问题的一种新型范式[13]。该范式基于“检索-生成协同”的思想，将传统知识库与大模型语义能力结合，在文本生成前通过向量匹配从知识源中动态检索相关文段，从而显著提升模型的上下文理解能力与响

应准确性[13][14]。通过本地的 RAG 架构,可以有效的规避信息安全性问题[16],并且可以随时间更新,有很强的时效性[13]。RAG 架构已经证明了其在通用问答[17]、法律文书[18]、科技文献[19]等领域的极佳效果,并取得了显著成果,但其大多仍局限于单一模态,难以适配医学中“图文共存”的典型场景[8],例如病例参考等。

与此同时,具备视觉-语言多模态感知能力的大语言模型不断涌现,如 LLaVA[20]、BLIP[21]、GPT-4V[22]、Qwen-VL[23]等,它们将视觉信息作为输入与语言共同建模,为多模态理解任务打开了新的空间。医学数据天然具有多模态属性[9][24],包括但不限于文本(病例描述、病历记录)、图像(X光、CT、MRI等)、结构化表格(化验单、指标趋势)以及语音等模态。因此,融合“RAG 结构”与“多模态大语言模型”的复合体系,成为构建智能医学内容生成与辅助诊断系统的关键路径之一。然而,尽管这些通用多模态大模型在开放领域取得了显著进展,其在医学数据理解方面仍面临诸多挑战。首先,医学领域的数据具有高度专业性和复杂性,涉及大量专业术语、缩略词和复杂的语义结构[25]。通用模型通常缺乏对这些专业知识的深入理解,导致在处理医学文本时可能出现误解或信息丢失。其次,医学图像如 X 光、CT 和 MRI 等,具有高分辨率和复杂的解剖结构,通用视觉模型在没有专门训练的情况下,难以准确识别和解读这些图像中的关键特征[26]。

可以看出,若能将 RAG 机制与多模态大语言模型有效结合,将为医学人工智能系统提供新的解决方案。一方面,RAG 机制能够动态接入医学知识库与图文样例,增强模型在面对复杂语境和专业需求时的知识广度与可解释性,缓解语言模型在事实一致性与时效性方面的局限;另一方面,多模态大语言模型具备联合理解图像与文本的能力[22][27],为医学图文数据的语义整合理解与生成推理提供了基础支持。二者的协同融合,将有望打破现有模型在信息封闭、模态割裂与语义漂移等方面的瓶颈,为医学图像诊断、临床问答、病历解读等多样化场景提供具备专业理解能力与强泛化能力的系统支撑。基于此,本文将围绕“融合 RAG 机制的多模态医学大模型微调”这一核心思路展开研究,提出并构建一个具备图文联合感知、知识动态调用与语义生成能力的医学内容生成系统,探索其在真实医疗任务中的可行性与性能表现,旨在为医学人工智能系统的落地提供新

的路径与实践范式。

1.2 国内外研究现状

近年来，随着人工智能技术的迅猛发展，特别是大语言模型（Large Language Models, LLMs）在自然语言处理领域的突破，医学人工智能研究也迎来了新的契机。在此背景下，检索增强生成（Retrieval-Augmented Generation, RAG）机制和多模态大模型，成为提升医学信息处理能力的重要方向。

1.2.1 医学数据特点与挑战

医学数据作为临床实践、科研分析和智能诊断的核心资源，其本身具备高度专业性、强语境依赖性和多模态复杂性的独特特点。这些特性既赋予医学人工智能系统以巨大的潜力，也带来了不同于通用场景的诸多挑战。针对多模态医学问答与生成任务而言，如何从结构复杂、语义模糊、来源异质的医疗信息中提取有用知识，是构建高效、多模态生成系统的前提条件。

多模态：实际的临床信息通常由结构化数据（如血压、心率、实验室检查指标）、非结构化文本（如病历、主诉、诊断描述）、图像数据（如 X 光、CT、MRI、病理切片）以及语音、视频等构成[24]。其中，医学图像常需结合医生的语言描述进行联合理解，如放射科报告中的“肺门模糊”必须联系 X 光图像具体解读[28]，而非仅靠文字或图像单一模态可获得。多模态信息在时间和语义上的对齐不一致，增加了数据融合的难度。模型需要具备对图文、结构化数值和上下文病情演变过程的联合建模能力，这对多模态语言模型的表征机制、上下文维持能力和推理能力都提出了较高要求。

隐私保护要求严格：电子病历、影像资料和生理监测数据等均属于患者隐私范畴，受到严格法规保护。这导致真实、可用的临床数据难以公开获取[29]，从而限制了模型的预训练与微调规模。如果在经过脱敏处理后，医学数据仍可能因样本稀缺、缺乏上下文而影响模型训练效果。因此，在医学领域构建大模型需要格外注意隐私合规机制。

知识高度复杂且存在多种描述：医学领域术语众多，诊断标准因时间、地域或权威机构不同而异，临床指南每年都会发生更新，导致数据分布随时间迁移。

此外，不同医院、科室之间的记录风格和数据标准存在较大差异[30]。例如，某些机构偏好使用“肺炎”，另一些则使用“肺部感染”表述同一病症。此类“语义异构”现象使得模型难以在不同医疗系统间泛化，必须设计跨语义对齐与知识更新机制，才能保持模型长期有效性与专业准确性。

数据质量少且不确定性高：相比通用 NLP 任务如情感分类或问答匹配，医学场景中的数据标注往往需由有资质的临床医生完成，且涉及病情理解、影像判读、药物识别等多任务判断，标注时间长、一致性差[32]。例如，在肿瘤分型任务中，即使是经验丰富的影像科医生也可能对同一病灶产生分歧，导致标签不一致，从而影响模型收敛与泛化[31]。因此，高质量医学标注数据的获取成为制约模型性能的重要瓶颈。当前主流做法包括引入多专家共识机制[35]、半自动标注[33]与主动学习策略[34]以降低成本。

类别不均衡：临床中常见疾病样本较多，而罕见病、特殊病例、边缘诊断的样本则极为稀缺，严重影响模型对低频事件的识别能力[38]。这与医学实践中“不能漏诊”的高风险需求形成明显冲突。因此，系统不仅需关注整体准确率，还需对考虑如何处理类间样本数量不均的问题，这对训练策略和系统结构设计都提出了特殊要求。

解释性要求极高：在临床应用中，人工智能系统的建议必须具有可追溯性和可解释性[36]，才能获得医生和患者的信任。相比于通用 NLP 任务，医学场景更强调模型生成内容的事实一致性与因果推理能力[37]。例如在辅助诊断系统中，若系统建议某种检查或治疗，其依据必须能回溯至图像中的病灶区域或病历中的关键描述。否则，即便生成文本语法流畅，也难以被医生采纳。

时效性与演变性：临床知识和疾病分类标准本身在不断更新，医学术语体系也随着指南调整而变化。一个训练良好的系统如果不具备动态更新能力，在不到一两年内就可能被新标准淘汰。

可以看到，医学数据在模态结构、隐私属性、语义复杂性、标注成本、类别分布与可解释性等方面具有高度特殊性。这些特点决定了将通用多模态大语言模型直接迁移至医学领域的难度高于其他领域，必须结合 RAG 机制、多模态对齐与高效微调等手段进行有针对性的改造与优化。

1.2.2 RAG 及其医学应用

1.2.2.1 RAG 的发展

自 2020 年 RAG 框架提出以来[14]，Retrieval-Augmented Generation（RAG）作为解决大语言模型（LLMs）知识局限和事实幻觉问题的关键方案，经历了持续而快速的发展。RAG 的核心思想在于引入外部知识库作为辅助信息源，通过检索器动态提取与输入查询相关的文档，并将其作为上下文输入生成器，从而提升生成内容的准确性、可靠性和知识广度[14]。

早期的 RAG 系统以稀疏检索方法（如 BM25[40]、TF-IDF[39]）为主，配合预训练生成模型（如 BERT[41]、T5[42]）构成基础的“检索-生成”流程。然而，这种组合方式往往面临两个挑战[17]：一是检索质量受限于关键词匹配，缺乏语义理解能力；二是生成器对输入的知识融合能力不足，难以进行深层的推理与信息整合。为此，后续研究逐步引入密集检索器，如 DPR[43]与 Contriever[44]，使用语义嵌入向量来实现更精确和可训练的检索。同时，在检索粒度上，从文档级扩展至段落、实体甚至 token 级单元，使模型更易定位关键信息。

随着 RAG 在开放领域问答、事实验证等任务中的广泛应用，人们开始关注其更深层次的协同优化机制。传统 RAG 架构将检索与生成视为独立模块，然而这种松耦合方式容易导致信息流断裂。为提升整体性能，研究者提出了端到端联合训练（Joint Training）的策略[45]，即将检索器与生成器置于统一框架中，通过联合优化目标函数来实现信息的高效传递与协同学习。此外，针对黑盒 LLM 无法微调的局限，更多方法转向增强阶段的优化[13]，例如引入压缩、排序、过滤等模块，在生成前精炼检索结果，提升与生成器的兼容性。

近年来，RAG 进一步演进多种增强策略，如推理路径驱动检索（planning-based retrieval）[46][47]、反馈机制迭代优化（feedback-driven refinement）[48]、检索触发判定机制（retrieval necessity estimation）[49]等。这些改进不仅提升了检索的针对性，还使得生成过程更加动态和自适应。例如，通过在生成初步答案后分析其置信度，模型可以判断是否需要追加检索，从而避免不必要的冗余计算，同时减小无关信息引发幻觉的风险。

1.2.2.2 多模态 RAG 的发展

随着人工智能应用需求的多元化，多模态表示学习因新算法和理论的不

现而蓬勃发展。与此同时，多模态大模型凭借庞大的参数规模和精妙的架构设计，在多模态语义理解与生成方面取得了重大突破。在这一背景下，检索增强生成（Retrieval-Augmented Generation, RAG）技术也突破了原有的单一文本模态限制，开始向多模态领域扩展，逐步演化为多模态检索增强生成（Multimodal Retrieval-Augmented Generation, MRAG）这一新兴研究范式[50]。

多模态 RAG 的核心在于融合跨模态检索与生成能力，能够从海量多模态数据中精准提取关键信息，并将其高效融入生成流程。这一机制不仅提升了如图像描述生成、视频理解等任务的表现，也为如智能驾驶中的视觉与路况融合[51]、医疗诊断中的影像与病历推理[52]等复杂场景提供了更优的解决方案。

近年来，多项算法创新推动了多模态 RAG 的发展。例如，RAFT 方法[53]引入可微分检索机制，使得模型能动态筛选最相关信息参与生成，显著提升了特定任务的效果。GME 框架[54]则构建了统一的多模态嵌入空间，借助注意力机制强化不同模态之间的信息交互，尤其在复杂数据集上展现出优越的检索性能。

在系统架构设计方面，研究者从多角度探索了多模态信息的融合路径。VISTA 方法[55]构建了共享语义空间，使图像中的视觉元素（如颜色、形状、物体）能与文本中的语义信息精确对齐，提升了图文匹配的准确率。MARVEL 项目[56]通过为原有文本检索系统引入视觉插件，拓展了其对图像输入的处理能力，实现了跨模态信息的有效联动。DocPrompting 研究[57]进一步探索文档驱动的代码生成机制，在实际场景中，如开发者根据 API 文档生成调用代码等任务中表现出良好的应用潜力。此外，MRAG 还被引入图像地理定位等领域[53]，结合视觉特征与外部知识库，实现更精确的位置推理。

1.2.2.3 RAG 在医学领域的应用

在 RAG 的医学应用中，国外研究起步较早，相关成果已逐步进入临床试验与原型系统开发阶段。例如，Google Research 提出的 Med-PaLM 框架[58]将医学领域的问答系统与结构化知识库融合，能够在面对复杂医学问诊时调用权威医学资料、文献与指南，显著提升了回答的可靠性与覆盖度。类似地，弗吉尼亚大学与伊利诺伊大学合作研发的 i-MedRAG 模型[59]则基于真实的临床电子病历系统构建知识索引，应用于住院医生的临床辅助决策流程中，通过 RAG 结构实现对既往病例、影像报告与医学指南的综合调用，改善了信息碎片化带来的诊疗偏

差问题。

在更偏工程与实践层面，国外已有医疗服务商尝试将本地 RAG 模型部署于医院内部知识系统中，用以实现隐私数据保护条件下的病例辅助分析与临床文献调阅。例如卡耐基梅隆大学开发的 DocPrompt 系统[60]，通过构建疾病文献与药物说明书的嵌入库，实现临床路径推荐与用药提醒功能，初步显示了 RAG 系统在辅助决策层面的可行性。

国内在 RAG 架构应用于医学场景方面的系统研究领域也进行了诸多探索。近年来，随着国产大模型的发展与相关算力平台的完善，一些高校和研究机构开始尝试将 RAG 思想引入中文医学问答系统、健康咨询平台与在线诊疗系统中。例如中科院提出基于中文医学知识图谱的 RAG 模型，在疾病分类任务中取得了超越传统 BERT 类模型的准确率；也有企业如腾讯、阿里[61]、百度等开始布局医学 RAG 产品线，探索如何在隐私合规的前提下对接医院信息系统，实现知识增强型对话与推荐服务。

1.2.3 LLMs 及其医学应用

1.2.3.1 LLMs 的发展

自从 Transformer 架构[82]通过自注意力机制（Self-Attention）实现了对长距离语义依赖的高效建模，取代了传统循环神经网络（RNN）和卷积神经网络（CNN）在序列建模中的主导地位，预训练大语言模型（Pretrained Large Language Models, PLLMs）便开始了蓬勃发展。

BERT（Bidirectional Encoder Representations from Transformers）[41]由 Google 在 2018 年提出，标志着预训练语言模型的第一个高峰。BERT 采用了 Transformer 的 Encoder-only 结构，并引入了“掩蔽语言模型”（Masked Language Model, MLM）和“下一句预测”（Next Sentence Prediction, NSP）作为联合训练目标。MLM 的核心思想是随机掩蔽输入序列中的一部分词语，让模型在全双向上下文条件下预测被遮蔽的词汇，从而学习到词与上下文之间的深层语义联系。BERT 通过预训练+下游微调的方式，在多个标准任务上显著超越此前所有基线模型。在 BERT 之后，研究者对预训练目标、网络结构、参数规模和微调策略进行了广泛探索，发展出一系列变体模型。RoBERTa[62]取消了 NSP 目标，并增加训练轮数与数据

规模,进一步提升性能;ALBERT[63]引入参数共享与因子分解以减少模型参数;ELECTRA[64]使用替换检测目标代替 MLM,提高了训练效率;SpanBERT[65]优化了短语级表示能力;DistilBERT[66]通过蒸馏压缩模型规模以便于部署。

与 BERT 相对,OpenAI 相继推出的 GPT 系列 (Generative Pre-trained Transformer) 开辟了 Decoder-only 架构模型的一种新的方法。GPT-1[67]首次验证了单向 Transformer 解码器架构的可行性,GPT-2[68]通过 45 亿网页数据训练展示了零样本学习能力。GPT-3[69]拥有 1750 亿参数,训练语料覆盖多领域,具备良好的零样本 (Zero-shot)、少样本 (Few-shot) 学习能力,其表现引发了学术界与产业界对通用人工智能的广泛讨论。在技术路线成熟后,GPT-4[22]进一步在多模态能力、指令对齐 (Instruction Tuning)、安全性控制方面进行突破。该模型采用大规模多轮对话与人工反馈优化 (RLHF) 结合训练,使其更善于遵循人类指令,语言生成更加自然可靠。GPT-4 (及其多模态改版 GPT-4o) 也成为多个后续系统的底座模型。

进入预训练大语言模型时代,更多的公司开始发布自己的预训练大语言模型。例如 Meta 发布的 LLaMA[70]系列提供了 7B 至 65B 多个参数规模的开源模型,其高训练效率和性能表现使其成为学术界与工业界的研究基线。LLaMA 2[71]加入了基于人类反馈的强化学习 (Reinforcement Learning with Human Feedback, RLHF) 机制,使模型在对话场景下表现更具人性化。LLaMA 3[72]和 LLaMA 4 系列模型进一步扩展了其开源大语言模型的能力,在多项基准测试中表现出色,特别是在多语言处理、代码生成和推理任务上。此外,Anthropic 推出的 Claude 系列[73],谷歌推出的 Gemini 系列[74],均采用大规模预训练+多阶段微调的标准范式,成为新一代通用人工智能系统的重要构成。

国内近年来大语言模型也进入快速发展阶段,百度推出的文心一言 (ERNIE Bot) 基于 ERNIE 系列[77]模型演进而来,融合了知识增强预训练 (Knowledge Enhanced Pre-training) 理念,通过引入知识图谱辅助语言理解,在中文问答、对话、生成等任务中取得了优秀表现。DeepSeek 发布的 DeepSeek-V3[75]和 DeepSeek-R1[76]模型在多个数学和推理基准测试中超越了行业领先,并且提出了包括蒸馏学习在内的多项学习优化方法。阿里巴巴达摩院推出的通义千问 (Qwen) 系列模型,以大规模中文语料为基础进行训练,支持高质量的中文生

成和多轮对话。其 Qwen-VL[23]系列进一步扩展为多模态版本，在图文理解和多模态问答任务上效果显著。在 Qwen 系列的基础上，阿里巴巴进一步开发了 QwQ（Qwen with Questions）模型，专注于增强模型的推理和逻辑思维能力。QwQ-32B 模型通过在生成答案前先输出思考过程，模拟人类的推理方式，在数学、编程和复杂问答任务中表现出色。此外，智谱 AI 与清华大学联合开发的 ChatGLM 系列[79]，商汤和上海人工智能实验室的书生系列[80]、讯飞的星火认知大模型、华为诺亚方舟实验室的 PanGu 系列[81]、IDEA 研究院的 Fengshenbang 模型等，也在不同参数规模和任务方向上开展了深入探索与开源实践。

1.2.3.2 多模态 LLMs 的发展

随着 LLMs 的不断发展，单一模态的语言模型已难以满足现实世界中复杂多样的信息处理需求。多模态大语言模型（Multimodal Large Language Models, MLLMs）应运而生，旨在融合文本、图像、音频等多种模态的信息，实现更全面的理解和生成能力。MLLMs 通过引入多种模态作为输入，使得模型能够处理图文混合指令、视觉问答、图像生成描述等复杂任务，广泛应用于对话系统、辅助诊断、图像检索等领域。

最早为多模态模型奠定基础的是 OpenAI 于 2021 年提出的 CLIP (Contrastive Language-Image Pretraining) [78]。尽管 CLIP 本身并非语言生成模型，其训练目标是通过对比学习在大规模图文对数据上对齐图像与文本的语义表示，使二者嵌入到一个共享的语义空间。CLIP 的结构包含一对编码器：图像编码器通常使用 ViT (Vision Transformer) [83]，文本编码器则为 Transformer[82]。训练中，模型学习最大化图文匹配对之间的相似度，同时最小化非匹配对之间的相似度。这一范式显著提升了图像语义理解能力，并为后续多模态系统提供了图像嵌入的直接来源。事实上，CLIP 生成的 image embedding 已成为多个后续 MLLM 中视觉 token 嵌入的标准起点。

在 CLIP 的基础上，BLIP (Bootstrapped Language-Image Pretraining) [21]系列工作进一步将视觉对齐与语言建模相结合。BLIP 提出了图文对比和文本生成的混合式预训练任务，结构上采用 ViT 与 Transformer 解码器的结合形式，将图像特征引入语言模型上下文中，从而实现端到端的多模态学习。BLIP-2[84]在此基础上引入了 Q-former 模块，这是一种基于 Transformer 的跨模态对齐器。Q-

former 将 CLIP 生成的图像向量与可学习的 Query token 结合,输出具有语言 token 形态的伪嵌入 (pseudo tokens), 其维度与 OPT[85]、Flan-T5[86]、LLaMA[70]等主流语言模型的 token 嵌入一致。借此,图像信息得以作为“语言提示”(prompt)自然输入至预训练语言模型中,从而建立“图像→伪 token→语言模型”的信息传递路径。这一设计不仅提升了图像与语言的对齐效率,也极大提高了视觉信息对语言生成的控制能力。

MiniGPT-4[87]则进一步继承并简化了 BLIP-2 的设计思路,其视觉编码器仍由预训练好的 ViT 和 Q-former 组成,输出视觉 token 后,通过一个线性层将其调整为与语言 token 相同的维度。随后,这些视觉 token 与文本 prompt 拼接组成完整的输入序列,输入冻结参数的大语言模型(如 LLaMA)。在输入结构中,视觉 token 通常以占位符“<Image>”的形式插入,使语言模型能够在上下文中感知和引用图像信息。该方法高度兼容传统语言建模流程,且只需微调视觉编码器与投影层,具备极高的迁移效率。

相较之下,LLaVA (Large Language and Vision Assistant) [20]在结构上更为简洁,取消了 Q-former,直接采用 CLIP 或 ViT 提取图像特征,并通过多层感知机 (MLP) 投影为语言模型可接受的 token 形式。这些 token 通常作为“视觉句子”插入语言输入中,构建诸如“Please describe the following image:”这样的上下文提示,后接图像 token。LLaVA 使用 Align 和 Instruction Tune 阶段对模型进行大规模图文数据微调,从而强化其在问答、描述、推理等任务中对视觉输入的响应能力。由于结构精简、训练高效,LLaVA 成为多个开源 MLLM 的技术参考模板。

在闭源模型方面,OpenAI 的 GPT-4V (GPT-4 with Vision) [22]代表了当前最强多模态能力。尽管其内部结构未完全公开,已有的信息显示 GPT-4V 使用了专门设计的图像编码模块(很可能为 ViT 结构),将图像编码为 patch 级别的视觉 token,并通过线性投影后与语言 token 共同输入 Transformer 模型中。在实际应用中,GPT-4V 支持复杂的图文混合输入,具备长上下文引用、跨模态记忆与视觉推理等能力,展现出高度拟人化的认知水平。

国内在该领域的代表作是阿里达摩院推出的 Qwen-VL[23]系列。该模型以 Qwen 语言模型为基础,引入视觉桥接模块,实现图文统一建模。Qwen-VL 采用 ViT 提取图像 patch 特征,并将其通过线性映射转化为符合语言模型输入维度的

token 形式，再将这些图像 token 嵌入语言序列中，通过显式的位置标记使模型识别其为视觉信息。在最新的 Qwen2.5-VL 中，这一机制进一步优化，图文输入可以自然融合，系统具备连续多轮图文对话的能力，具体内容和嵌入方法会在后续进行详细分析。

1.2.3.3 MLLMs 在医学领域的应用

多模态大语言模型的出现，为解决传统语言模型在复杂信息融合与语义跨模态对齐方面的瓶颈提供了新的技术路径。在医学领域，文本信息与视觉信息长期共存且紧密关联，如放射影像与诊断描述、病理图像与临床报告、手术视频与病例总结等。因此，能够理解并联合处理多模态医学数据的智能系统，被认为是提升 AI 医疗能力的关键突破口。随着模型规模扩大与多模态预训练技术的成熟，国外在 MLLMs 医学应用方面的研究已取得一系列重要进展。

国外较早一批关注该问题的研究工作包括 LLaVA-Med[88], Med-Flamingo[89], BioGPT-Vision[90]等。这些模型在视觉语言预训练架构的基础上，加入医学影像文本对进行联合训练，并在医疗影像问答、图像描述生成、诊断辅助等任务中取得了令人瞩目的效果。其中，Med-Flamingo 基于 OpenAI Flamingo 模型结构，结合大量医学图文对进行微调，在多个医疗 VQA 数据集上刷新了性能指标，显示出模型在图像引导下理解医学描述、推断病变情况的能力。

在国内，多模态医学模型的研究也有大量相关工作。例如，浙江大学与阿里巴巴合作开发的 Med-CLIP[91]尝试基于中文医学图文对构建语义匹配模型，用于图像分类与检索等任务；而百度与北大合作的医学视觉大模型则更侧重于图像侧的语义建模。尽管国内对 MLLM 在医学应用中的研究有所欠缺系统性方法，但相关研究已表现出浓厚兴趣，并在图像诊断、辅助问答、自动标注等场景中做出初步探索。当前的挑战主要包括：如何解决医学图像标注资源稀缺问题、如何建立符合语义一致性的图文对齐机制、以及如何在保证隐私安全前提下进行大规模模型训练等。

1.3 主要研究内容与创新点

本研究旨在构建一个融合多模态检索增强生成（RAG）机制与多模态大模型微调的医学内容生成系统，以解决前文提到的大模型在处理医学数据存在的专业

语义理解、图文对齐能力不足、缺乏专业知识注入渠道等问题。为实现这一目标，系统在结构设计上引入了多模态 RAG 机制以动态检索外部医学知识，从而提升模型回答的准确性、可解释性、安全性与时效性。同时，为弥补通用大模型在医学场景中的知识盲区与能力不足，本文基于 Qwen2.5-VL 模型[92]进行了分阶段的专业化微调，分别使用医学文本数据与医学图文对数据增强其语义理解与视觉分析能力。二者结合，使模型不仅能处理复杂的图文输入，还能通过 RAG 模块访问专业外部知识，最终形成一个可查询、可生成、可更新的医学内容生成系统。

主要创新及贡献总结如下：

1. 首先，本文首次在医学场景下将多模态 RAG 机制与大模型微调技术协同集成，构建了一个支持图文语义联合理解与外部知识动态调取的医学内容生成系统。系统在图文联合输入的基础上，结合向量检索与语义生成，实现了高准确度、高语境适应性的医学知识生成流程。

2. 其次，本文自主设计并实现了多种多模态 RAG 范式，涵盖图文共嵌、标签代理检索与图文绑定等策略，并进一步设计了一种更适合医学场景的图文对绑定检索范式，并通过系统性实验证明了其有效性。该范式能够兼顾语义准确性与信息完整性，在医学图像理解与检索生成任务中展现出更高的实用价值。

3. 最后，本文采用了基于语言数据与多模态数据的分阶段微调策略，对 Qwen2.5-VL 模型进行了多层次的医学适配训练，形成了具备强医学语言理解与医学图像分析能力的专业化模型 Qwen2.5-VL-Med。该模型与 RAG 系统结合，在下游评测内容中展现出优异的表现，验证了分阶段微调在提升模型专业能力方面的显著效果。

此外，在系统设计层面，本系统坚持以可扩展性为核心导向，对各功能模块均进行了高度模块化与解耦设计。无论是 RAG 检索机制、多模态嵌入方法，还是模型加载与调用流程，均采用独立组件化结构，确保各部分可以灵活替换、独立优化。这一设计不仅为后续接入更强大的嵌入模型、向量数据库或大语言模型提供了接口支持，也使得系统能够快速适配不同类型的医学数据与模型范式，具备良好的兼容性与演化能力。同时，为了提升系统的易用性与交互体验，本文开发了可视化前端界面，支持文本输入、图像上传与对话过程可视回显，用户无需掌握底层模型逻辑即可便捷地使用系统，极大地拓宽了该医学内容生成平台的适

用场景与实际推广价值。

综上，本文从系统架构设计、范式探索到模型训练策略均进行了多方面创新，为多模态医学内容生成系统的发展提供了一种切实可行、性能优越的技术路径。随着医疗智能化进程的加快，融合 RAG 架构的多模态医学模型有望成为临床辅助系统、远程诊断平台以及健康管理服务的重要组成部分，其研究与实现具有突出的理论意义与应用价值。

1.4 论文章节结构安排

具体而言，本文的研究内容与章节安排如下：

第一章为绪论部分，介绍了本研究的背景与意义，系统回顾了当前 RAG 技术、多模态大模型（MLLMs）及其在医学领域的研究现状，梳理了当前多模态医学内容生成系统面临的关键挑战，明确了本研究的创新点与研究目标。

第二章为相关技术与基础算法综述，详细介绍了 RAG 的基本结构与发展路径，涵盖稀疏/密集检索器、增强策略与生成模块，并系统回顾了多模态 RAG 的研究趋势。同时，介绍了 Transformer 及 Vision Transformer 架构、并以 Qwen2.5-VL 为例，分析了大语言模型与多模态模型的主流实现，为后续系统设计奠定理论基础。

第三章系统地介绍了本研究设计的多模态医学内容生成系统的整体架构，并对涉及到的技术内容进行较为详细的解释。首先，对模型的整体架构进行了分析，介绍了模型架构的模块化设计思路，并解释了不同模块拼接形成的三条路径。接下来详细阐述了系统中多模态 RAG 模块的检索逻辑与范式划分以及具体工程实现中的模块解耦设计。随后，分析了生成模块中 Qwen2.5-VL-Med 的微调方式，阐述了语言微调与图像微调两阶段策略的设计目标、训练设置与资源配置。该章节还介绍了系统如何通过 API 集成多平台大模型服务，以提升整体工程灵活性与可维护性。

第四章围绕实验评估展开，设计并实施了多项实验任务，包括使用 RAG 与微调前后性能提升分析、多模态 RAG 范式消融实验、基准问答准确率对比等。通过在 OmniMedVQA 基准数据集[93]上的实验验证，本文提出的范式 4 RAG 与 Qwen2.5-VL-Med 组合在医学图文推理、问答生成任务中表现总体优于其他范式

方法及模型，验证了本系统设计在医学场景中的实用价值。

第五章介绍了系统为用户交互层面的设计思路与实现细节。为满足调试、演示与部署等不同需求，系统支持三种交互方式：终端交互、网页交互界面以及 Qwen2.5-VL 原生 WebUI。不同交互方式均支持图文上传、流式输出与历史记录维护，并能适配 API 或本地模型调用。

第六章对全文工作进行了总结，回顾了本文提出的多模态医学内容生成系统在 RAG 机制设计、大模型微调策略、系统架构整合方面的创新点，指出当前系统的局限性，并对未来可扩展方向与研究演进趋势进行了展望。

2 相关技术及算法

2.1 检索增强生成

2.1.1 RAG 架构

RAG（检索增强生成，Retrieval-Augmented Generation）[15]是一种结合外部知识库与语言模型生成能力的架构。它在语言模型生成之前引入信息检索机制，根据用户输入从数据库中提取相关内容，再将这些检索结果作为上下文输入生成模块，从而提升输出的准确性与时效性。

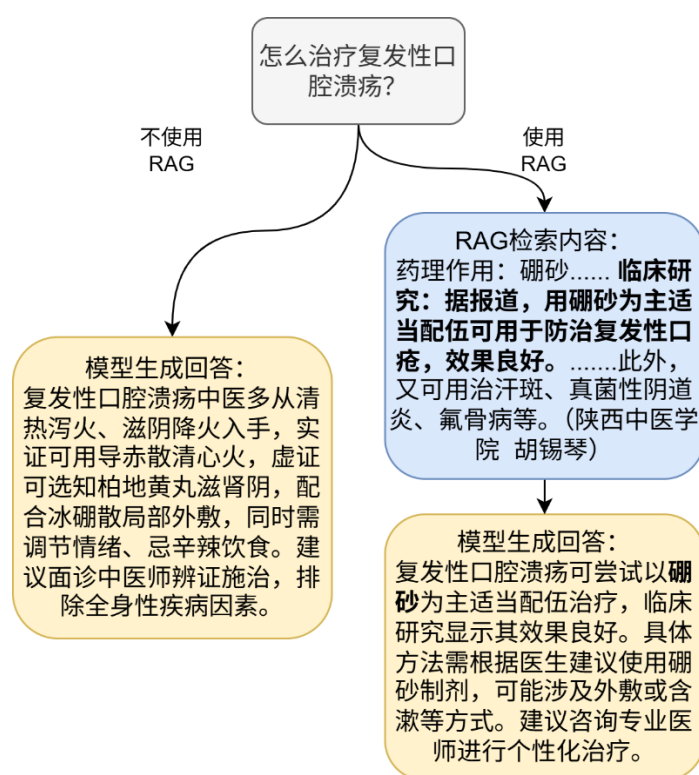


图 2.1 使用 RAG 前后回答对比

RAG 系统可以显著提升模型的事实性、可扩展性与灵活性[13]。首先，相比仅依赖参数内存的传统大模型，RAG 能够实时检索最新知识，有效缓解语言模型“知识陈旧”的问题，尤其适用于需要紧跟信息更新的场景。其次，RAG 提升了生成内容的可信度与可解释性——模型在回答问题时不仅给出结论，还能追溯所引用的检索来源，这在医疗、法律等对“信息溯源”要求高的领域尤为关键。

此外，RAG 具备良好的任务适配性：面对领域跨度大的问题时，它可以动态调整检索策略，灵活调取不同知识来源，增强模型在多任务场景下的泛化能力。因此，RAG 不仅是一种提升性能的技术手段，更是通往“知识驱动型智能”的关键路径之一。

RAG 的基本流程包含三个核心阶段：检索(Retrieval)、增强(Augmentation)和生成(Generation)，此外还包括判断是否需要触发检索的机制。每个阶段都有多种设计方式和实现路径，构成了一个高度可配置、灵活适配的生成框架。

在检索阶段，主流方法大致可分为稀疏检索与密集检索两类。稀疏检索依赖词项级别的匹配，如 TF-IDF 和 BM25[95]，主要通过倒排索引在原始文本中定位内容，适合处理段落级文本查询。这种方式简单高效，但性能高度依赖于数据库结构和查询质量，且难以捕捉语义相似性。相比之下，密集检索方法将查询与文档嵌入到统一的向量空间中，从而捕捉更深层的语义关联，具备更强的可训练性和适应性。在 RAG 系统中，常见的密集检索器采用双编码器结构，如基于 BERT 的 DPR (Dense Passage Retriever) [94]，在开放问答 (OpenQA) 等任务中表现突出。此外，单编码检索器则通过对比学习进行预训练，具有更高的通用性，适用于多种下游任务。实验表明，通用型预训练检索器在零样本或开放任务中更具弹性，但在特定任务中，经过微调的 DPR 通常能带来更优性能。

检索粒度也是影响模型性能的关键因素，它直接关系到数据库的空间占用与搜索效率。早期的检索增强模型多以整篇文档为单位进行处理，由机器阅读理解模块从中定位答案。而随着任务复杂性的增加，更细粒度的检索方式逐渐被采用，例如基于 token 或实体的检索。这些方式虽然加重了数据库的负担，但在处理领域外知识或罕见模式时更具优势。如今，文本块级的检索因信息集中、冗余少而成为主流选择。此外，面向实体任务的检索机制也逐步兴起，如 EAE 模型[96]通过学习实体表示增强实体识别效果，提出基于提及构建知识库的思路，使得 RAG 在实体识别和推理类任务中更加高效。

为了进一步提升检索效果，研究者们设计了多种增强策略，既包括检索前的查询扩展，也包括检索后的信息压缩与筛选。例如，Query2doc [97]能根据初步生成结果对原始查询进行扩展，提高召回率；PRCA 方法[98]则对前 K 条检索文档进行筛选与重排序；BlendFilter [99]则专门用于过滤噪声信息，提升知识的

可用性与相关性。此外，Xu 提出的 RECOMP [100]与 Hofstatter 的 FiD 模型 [101]也对文档向量或内容进行压缩，减少对生成模型的计算压力，提升处理效率。RAG 系统的知识源既可以是闭源数据库（以键值对形式存储原始文本），也可以是开源百科知识库，如维基百科，因其结构清晰、覆盖广泛而被广泛使用。

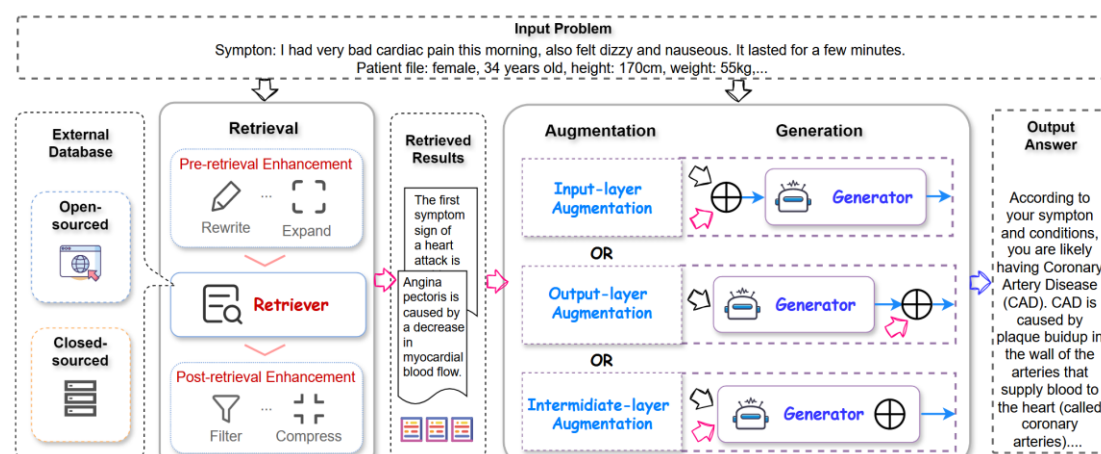


图 2.2 多种 RAG 增强的方式[17]

增强阶段主要负责将检索到的信息有效整合进生成流程，提升整体生成质量。这一阶段通常涵盖查询扩展、文档筛选、信息压缩等多种策略，并非简单拼接检索内容，而是需要对信息流的有效性与生成器的兼容性进行平衡。

在生成阶段，RAG 系统根据任务需求采用不同的生成器结构。常见结构包括编码器-解码器模型（如 T5、BART）与仅解码器模型（如 GPT 系列）。白盒模型如 BART 和 T5 具备开放的结构与参数，可根据任务进行微调，通过编码输入与目标之间的交叉注意力机制实现信息融合，是许多 RAG 实现的核心组件。相比之下，黑盒生成模型如 GPT、Claude、Codex 等则更难以调优，只开放输入输出接口，限制了对生成过程的深度控制。因此，在这类模型中，研究更侧重于优化检索与增强流程，确保生成器输入尽可能高质量。例如利用模型自身生成的数据来训练查询提示器，或者通过文档压缩来减少生成开销。

除了核心流程，是否需要检索、以及何时触发检索，也是 RAG 系统中的关键决策机制。当前一些方法尝试赋予模型“自我感知”的能力，用于判断是否进行外部知识查询。例如 Self-RAG [49]通过特殊标记来引导检索行为，SKR 方法[102]则利用语言模型自身的知识置信度动态决定是否调用检索器。而在

FLARE[103]这类系统中，模型甚至会先生成一个“启发式答案”，据此评估当前知识是否足够，若不足再发起查询重写，从而实现更加灵活和自适应的检索策略。

2.1.2 多模态 RAG

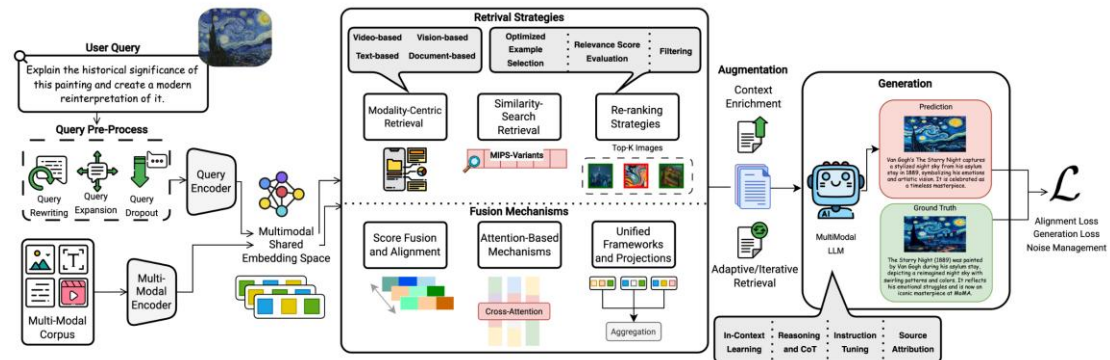


图 2.3 MRAG 系统[50]

现代多模态 RAG 系统主要包括以下四个核心模块：基于模态的检索、重排序与选择策略、融合机制以及动态检索控制。

基于模态的检索方法通过感知不同模态（如文本、图像、视频、文档）的独特特性，优化信息提取和匹配过程。在多模态检索生成系统中，文本仍是最为核心的模态，常用于各类任务的基础检索。传统的稀疏检索方法和密集表示模型在文本语义匹配中应用广泛，强调语义对齐的精度和细粒度建模[96]，提升了对用户查询的理解能力。在图像检索领域，有的方法直接基于视觉特征进行匹配，也有研究实现了图文之间的零样本关联，使系统具备更强的跨模态泛化能力。视频检索方面，随着对时序建模需求的提升，一些方法[99]结合了动态信息与语义理解，提升了长视频内容中的上下文处理能力与时间推理表现，使系统能更准确地定位并解析关键片段。文档检索则呈现出向端到端融合建模方向发展的趋势，多个方法[100][101]将文档中的文本、视觉、版面结构等多种信息整合，不依赖传统 OCR 流程，同时在效率与准确性之间寻求更优权衡，并探索了跨格式结构化学习的潜力。

其次，重排序与选择策略在检索阶段之后进行结果优化。除了基础相关性评分外，诸多方法还引入多轮示例筛选与动态排序机制，以提升检索结果的质量。例如，两阶段检索与重排序流程可显著优化生成质量，减少冗余与噪声干扰。

融合机制是 MRAG 系统中模态信息交互的核心。目前主流方法包括跨模态对齐模型、基于注意力的融合策略以及统一表征学习。部分方法通过跨编码器结构将文本、图像等信息统一转化为自然语言形式，借助 CLIP 或 BLIP 特征进行融合。其他方法则通过双流注意力机制实现信息动态加权。IRAMIG 等方法进一步通过迭代融合生成统一知识表示，DQU-CIR 则通过生成图像标题与文本叠加处理复杂查询，展现出融合策略的多样性与有效性[104]。

最后，动态检索控制机制使得系统在处理复杂任务时能够灵活评估是否需要外部知识并进行实时检索。例如，SKURG 根据输入复杂度动态控制检索步数；SAM-RAG 与 mR2AG 引入大模型评估机制，判断检索必要性并过滤无效内容；MMed-RAG 进一步提升了检索的精确度；OmniSearch 通过代理机制将多模态复杂问题分解为子问题链，并进行实时搜索规划。

2.1.3 向量数据库

向量数据库（Vector Database）作为支持大规模相似性检索与表示学习任务的关键基础设施，近年来随着大语言模型、多模态模型及 RAG 系统的兴起而迅速发展。与传统的关系型数据库不同，向量数据库专注于处理高维向量的存储、索引与检索问题，尤其适用于文本、图像、音频等非结构化数据的语义匹配与相似度查询任务。在 AI 系统中，它主要用于存储由嵌入模型生成的向量表示，并支持通过“最近邻搜索”快速查找语义上最相关的内容，构建知识增强生成的支撑能力。

向量数据库的发展经历了从简单线性搜索，到复杂近似最近邻（Approximate Nearest Neighbor, ANN）算法体系的演进。在早期阶段，向量检索依赖于暴力搜索（Brute-force），即对每个查询向量与数据库中所有向量逐一计算距离。这种方式虽然结果准确，但在数据量大时计算代价极高，难以满足实时性需求。为解决这一问题，研究者提出了一系列高效的 ANN 算法，其中最具有代表性的包括基于哈希的局部敏感哈希（LSH）、基于树结构的 KD-Tree 与 Ball-Tree、基于图的导航小世界图（NSW）与 Hierarchical Navigable Small World（HNSW），以及基于量化的产品量化（PQ）与向量量化（VQ）等方法[105]。

其中，HNSW[106]因其在高维空间中的搜索效率与精度表现最为出色，已成

为当前多数工业级向量数据库的主流索引结构之一。HNSW 的核心思想是构建一个分层的图结构，每一层都是一个“小世界”图，节点之间具有小平均路径长度和高度聚集的连接关系。在搜索时，算法从顶层开始，通过贪心策略逐层向下跳转，每层都寻找距离查询向量更近的节点，最终在底层图中进行精细搜索，快速找到近似最邻近点。由于其结构天然支持高效导航与剪枝操作，HNSW 在保持高精度的同时显著降低了搜索延迟。PQ 等量化方法则主要用于压缩向量表示，通过将高维向量编码为多个低维子空间的量化索引，大幅减少存储空间并加快距离计算速度，适用于存储成本敏感或内存受限的应用场景。

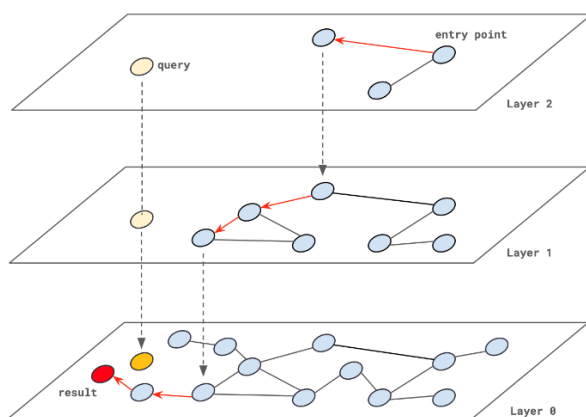


图 2.4 HNSW 算法

现代向量数据库不仅关注索引结构的优化，还在查询语义增强、元数据过滤、异构索引融合等方面取得了显著进展。例如，FAISS[107]支持将粗量化索引与精细 HNSW 搜索结合，实现粗筛+精排的两阶段检索机制；Milvus[109]则引入 Hybrid Search 机制，将结构化字段过滤与向量查询结合，提高多模态检索任务中的筛选能力；Weaviate[108]、Qdrant 等系统还支持使用自然语言描述进行查询，并提供上下文感知的检索接口，进一步增强与大模型协同工作的能力。

随着多模态 RAG 等生成系统的普及，向量数据库的作用越发凸显。未来，随着多模态数据融合与实时智能问答的需求不断上升，向量数据库可能在可解释性、时效性、扩展性等方面继续演进，为未来更先进的 AI 系统奠定底层技术基础。

2.2 大语言模型相关算法

2.2.1 Transformer 架构

Transformer 结构由来自谷歌的 Vaswani 等人于 2017 年在论文《Attention is All You Need》[82]中首次提出，开创性地摒弃了传统序列建模中常用的循环神经网络（RNN）与卷积神经网络（CNN）结构，完全依赖注意力机制（Attention）进行信息传递与特征抽取。其结构以高度并行化、长距离依赖建模能力强等优势迅速取代 RNN 类模型，成为当代自然语言处理乃至跨模态学习的基础架构。

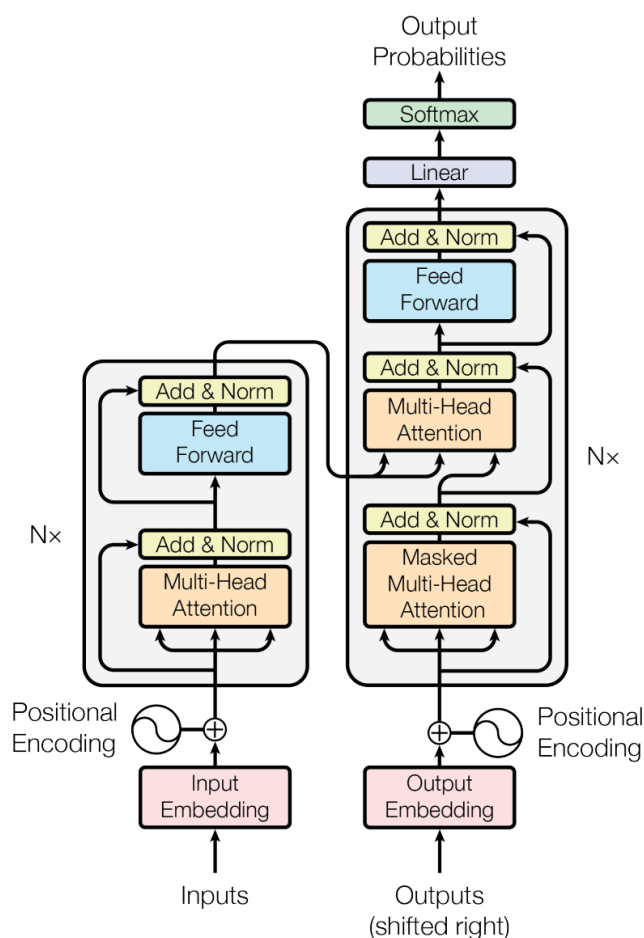


图 2.5 Transformer 架构[82]

Transformer 的基本架构由编码器（Encoder）和解码器（Decoder）两部分组成，每部分又包含多个重复堆叠的子层（Layer）。标准的 Transformer 通常采用 6 层编码器与 6 层解码器，每层结构相同但参数独立。在预训练语言模型（如 BERT、

GPT) 中, Transformer 结构被进一步裁剪或扩展, 如 BERT 仅保留编码器部分, 称为 Encoder-only 架构; 而 GPT 仅保留解码器部分, 称为 Decoder-only 架构。而如今的大部分 LLMs 都采用了和 GPT 相同的 Decoder-only 架构。每个编码器和解码器均由多个结构完全相同的子层 (Layer) 堆叠而成, 编码器由若干个“多头自注意力层+前馈网络”组成, 解码器在此基础上引入了掩蔽机制与跨注意力模块, 以实现目标序列的逐步生成。

Transformer 的核心是自注意力机制 (Self-Attention), 它允许模型在处理每一个位置的输入时, 都能考虑序列中所有位置的信息, 从而捕捉长距离依赖关系。具体地, 对于输入序列中每一个词向量 x_i , 通过三个可训练的线性变换矩阵 W^Q 、 W^K 与 W^V 分别得到其对应的查询 (Query)、键 (Key) 和值 (Value) 向量, 即

$$Q = XW^Q, K = XW^K, V = XW^V \quad (2.1)$$

其中 $X \in \mathbb{R}^{n \times d}$ 表示输入序列中 n 个词向量, 每个向量维度为 d 。然后, 计算所有位置之间的注意力权重, 通过缩放点积方式完成:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.2)$$

其中 d_k 是键向量的维度, 缩放因子 $\sqrt{d_k}$ 用于稳定梯度。Softmax 操作保证了每一行的注意力权重总和为 1, 表示某个查询位置对其它位置的关注程度。

为了增强模型的表达能力, Transformer 引入了多头注意力机制 (Multi-Head Attention)。该机制不是仅用一组 Q, K, V 进行计算, 而是通过 h 个不同的线性变换矩阵生成多组 Q_i, K_i, V_i , 分别计算各自的注意力后再拼接结果, 最后通过一个线性变换映射到原始维度。其公式表达为:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (2.3)$$

其中每个 $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, 而 W^O 是输出映射矩阵。多头机制的设计使得模型可以在不同的子空间中关注不同的语义信息, 有效提升了语义建模能力。

在注意力模块之后, Transformer 还包含一个前馈全连接层 (Position-wise Feedforward Network), 它对每一个位置的表示独立地进行非线性变换。形式如下:

$$\text{FFN}(x) = \text{ReLU}(xW_1 + b_1)W_2 + b_2 \quad (2.4)$$

其中 $W_1 \in \mathbb{R}^{d \times d_{ff}}$ 和 $W_2 \in \mathbb{R}^{d_{ff} \times d}$, d_{ff} 通常大于 d , 以提高非线性变换的容量。这一结构能够增强模型的表达非线性特征, 使其更适合捕捉复杂模式。在后续的大模型中, ReLU 激活函数可能会被替换成 LeakyReLU, GLU, GeLU 等其他的激活函数。

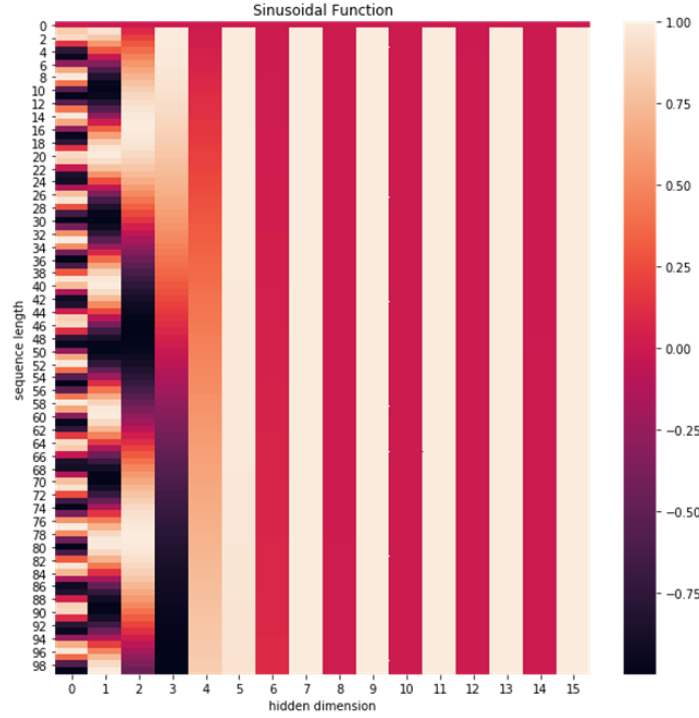


图 2.6 Transformer 位置编码的形象化展示

由于 Transformer 本身不包含任何递归或卷积结构, 缺乏对输入序列位置信息的建模能力。因此, 在输入阶段必须加入位置编码 (Positional Encoding) 以注入位置信息。最初的 Transformer 使用固定的位置编码, 其形式为:

$$\begin{aligned} PE_{(pos, 2i)} &= \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \\ PE_{(pos, 2i+1)} &= \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \end{aligned} \quad (2.5)$$

其中 pos 为位置索引, i 为维度索引。该设计使得不同位置间的编码具有规律性, 便于模型学习相对与绝对位置信息。后续也有研究引入可学习的位置编码、相对位置编码等改进方法, 以提升位置建模的灵活性。通过绘制的位置嵌入图, 纵向观察, 可见随着 d 序号增大, 位置嵌入函数的周期变化越来越平缓。

此外，为了提升模型训练的稳定性与收敛速度，Transformer 在每个子层后添加了残差连接（Residual Connection）与层归一化（Layer Normalization）机制。每个子层的输出为：

$$\text{LayerOutput} = \text{LayerNorm}(x + \text{SubLayer}(x)) \quad (2.6)$$

其中 $\text{SubLayer}(x)$ 是多头注意力或前馈网络的输出。这种结构有助于缓解梯度消失问题，并加速深层网络的训练过程。

在训练过程中，Transformer 常采用自回归语言建模目标或掩蔽语言建模目标，分别对应于 GPT 类模型和 BERT 类模型的结构设计。在解码阶段，为保证输出序列的生成具有因果性（即不能看到未来信息），会对注意力矩阵引入上三角掩蔽，使得当前位置只能关注当前及之前的位置信息。

Transformer 提出以来，迅速成为自然语言处理领域的主流框架，被广泛应用于语言建模、问答系统、文本生成、对话系统等多种任务中。在其基础上衍生出多个经典预训练模型，包括 BERT、GPT、T5 等。随着模型规模和训练数据的不断扩大，Transformer 展现出强大的泛化能力和跨任务迁移能力，成为大语言模型的技术基石。

2.2.2 Vision Transformer 架构

随着 Transformer 架构在自然语言处理中的成功，研究者开始尝试将其应用于计算机视觉任务。传统的图像识别模型主要基于卷积神经网络（Convolutional Neural Networks, CNNs），如 ResNet、DenseNet 等，它们依赖局部感受野与多层卷积核的堆叠来捕捉图像的空间结构与语义特征。尽管 CNN 在多个视觉任务中表现良好，但其感受野扩展效率较低，难以建模长距离像素间的依赖关系，且通用性与跨模态能力受限。

在此背景下，Dosovitskiy 等人于 2020 年提出了 Vision Transformer（ViT）结构[83]，首次在图像分类任务中完全替代卷积模块，使用纯 Transformer 架构处理图像。ViT 的核心思想是将输入图像划分为固定大小的图像块（Patch），并将每个 Patch 展平为向量后嵌入 Transformer 中处理，从而实现视觉任务的序列建模。这一方法不仅保留了 Transformer 在长距离建模中的优势，还使得图像任务能够借助自然语言处理领域的大规模预训练范式进行迁移学习。

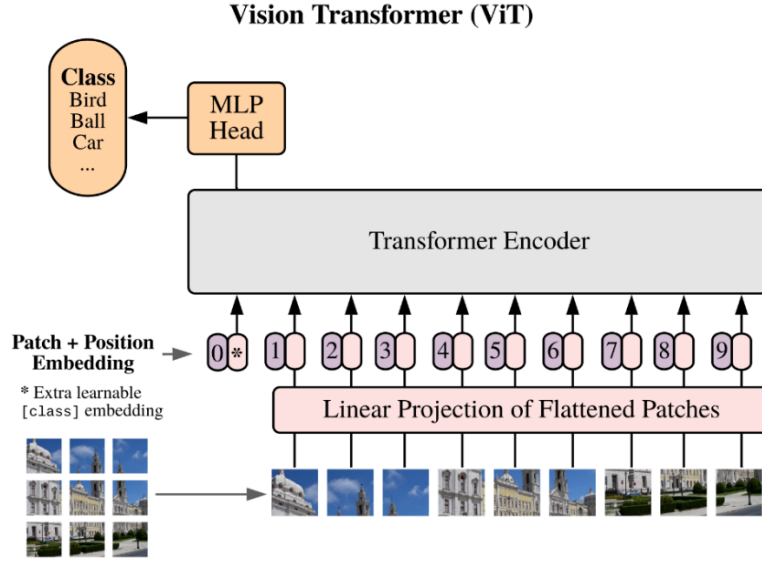


图 2.7 Vision Transformer 架构[83]

ViT 中处理图像的第一个关键步骤即为块嵌入 (Patch Embedding)。设输入图像为三维张量 $\mathbb{R}^{H \times W \times C}$ ，其中 H 、 W 分别为图像的高度与宽度， C 为图像的通道数。ViT 首先将输入图像划分为 $N = \frac{HW}{P^2}$ 个大小为 $P \times P$ 的不重叠图像块 (Patch)，并将每个图像块展平成一个向量：

$$x_i = \text{Flatten}(X_{(i)}) \in \mathbb{R}^{P^2 \cdot C}, i = 1, 2, \dots, N \quad (2.7)$$

之后，使用一个可学习的线性映射（通常为一个全连接层）将每个向量投影至维度为 D 的向量空间：

$$z_i = Ex_i + b \in \mathbb{R}^D, i = 1, 2, \dots, N \quad (2.8)$$

其中 $E \in \mathbb{R}^{D \times (P^2 \cdot C)}$ 是 Patch 嵌入矩阵， $b \in \mathbb{R}^D$ 是偏置项。经过该过程后，图像被表示为一个长度为 N 的向量序列，每个向量维度为 D ，即 $Z = [z_1, z_2, \dots, z_N] \in \mathbb{R}^{N \times D}$ 。

此外，ViT 还在输入序列前添加了一个额外的分类标记 ([CLS] Token)，其对应的嵌入向量是一个可学习参数，用于在 Transformer 处理结束后提取全局特征用于分类任务。

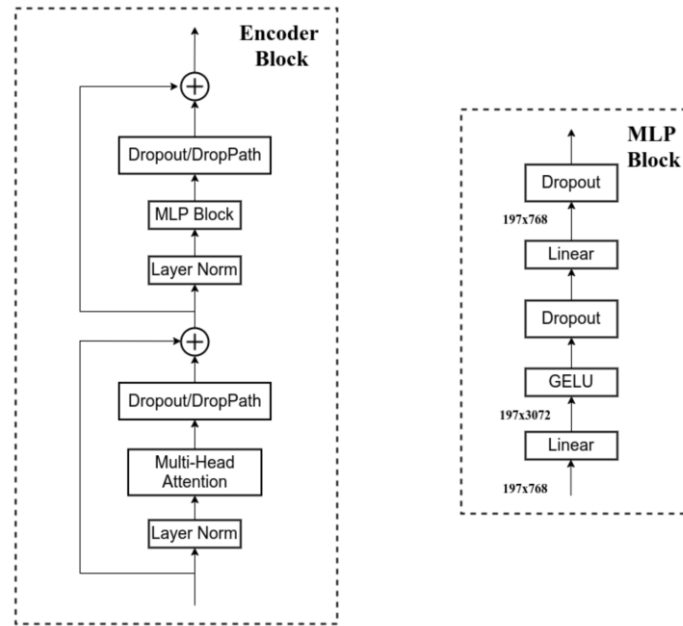


图 2.8 ViT 的 Encoder 部分

如图所示，ViT 结构中每个编码层均包含多头自注意力模块与前馈网络，与语言 Transformer 结构基本一致。

ViT 中使用的位置编码也是 Transformer 中的位置编码方法，即正弦余弦位置编码。由于上述章节已经对具体公式和计算方法做过说明，此处不再赘述。

在多模态任务中，Vision Transformer 已逐步成为图像模态的主流编码器，广泛应用于视觉问答、图文匹配、图文生成等任务中。尤其在医学图像与文本结合的任务中，ViT 提供了一种比 CNN 更易对齐语言表示的结构基础。医学图像如 X 光、MRI、CT 等存在着细节复杂、长程依赖强的视觉特征，ViT 的全局建模能力有助于捕捉图像中的异常特征区域并与文本信息进行语义对齐，提升模型的判别能力与可解释性。

近年来，多个多模态大模型均采用 Vision Transformer 作为图像输入主干，并通过与语言 Transformer 结构对接，实现图文双向建模与融合。本文所使用的 Qwen2.5-VL 模型亦在图像分支中基于 ViT 变体进行编码。

2.2.3 Qwen2.5-VL 模型

Qwen2.5-VL 是阿里巴巴 Qwen 团队于 2025 年发布的最新一代多模态大语言模型[92]，是 Qwen2-VL 系列的继任者。该模型在视觉编码、模态对齐、时空理解与视觉语言任务执行等方面均实现了显著提升。整体上，Qwen2.5-VL 延续了当前主流多模态大模型的架构范式，即由视觉编码器（Vision Encoder）、语言模型（LLM）和视觉语言融合模块（Merger）三部分组成，并在多项设计上进行了创新，以适应医学、办公、视频等高复杂度应用场景的需求。

在语言建模部分，Qwen2.5-VL 采用了 Qwen2.5 LLM 作为基础骨干。该语言模型基于 Decoder-only 架构，具备强大的文本理解与生成能力。在实现多模态对齐时，模型对原始的一维 RoPE（旋转位置编码）[110]进行了扩展，提出了多模态旋转位置编码（MRoPE），并进一步与绝对时间轴对齐，以增强模型对视频等时序数据的理解能力。这使得模型不仅能精准处理静态图文，还能识别视频中事件发生的具体时刻，具备更高的时间感知分辨率。

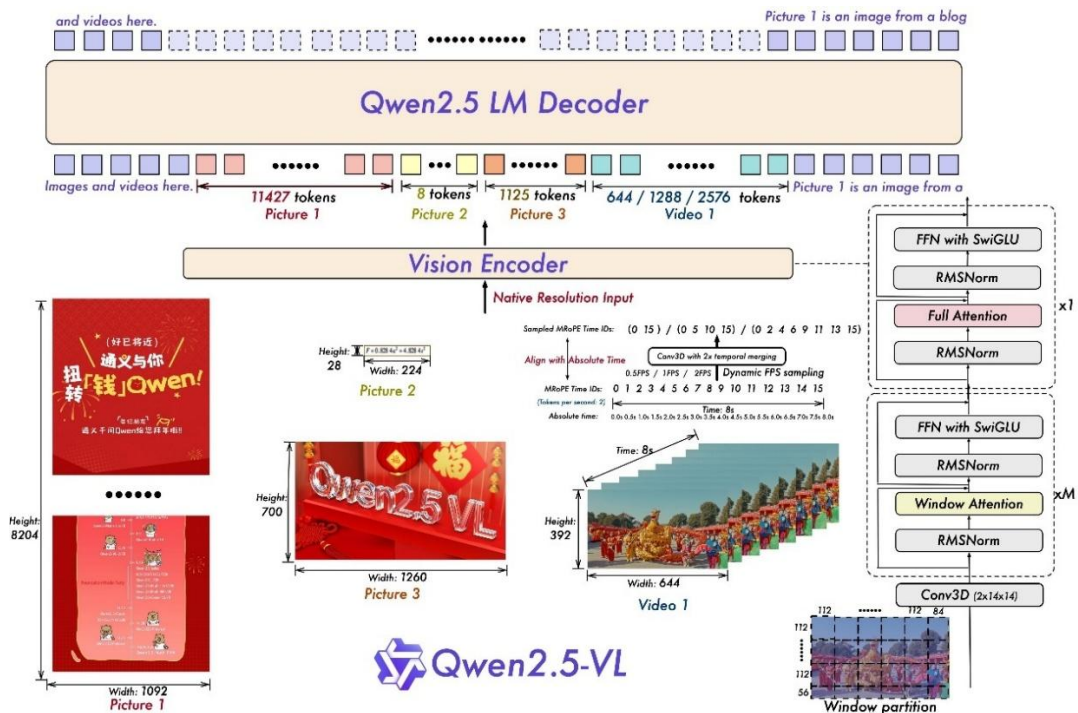


图 2.9 Qwen2.5-VL 模型架构[92]

在视觉编码部分，Qwen2.5-VL 重新设计了基于 Vision Transformer（ViT）的视觉模块，并引入了多个关键优化。首先，模型支持原生动态分辨率输入，图像

无需强制缩放至固定尺寸，而是保持长宽为 28 的倍数输入。其次，采用窗口注意力机制（Window Attention），在多数层中将计算限制在局部窗口内，仅在少数关键层采用全局注意力，从而显著降低视觉处理的计算复杂度。该机制使得视觉编码器计算量从原来的平方增长转变为线性增长，提高了训练和推理效率。

具体到图像处理流程，模型将输入图像以 14×14 像素为单位划分为 patches，并对相邻 4 个 patch 进行拼接，再通过两层多层感知机（Multi-Layer Processor, MLP）压缩为与 LLM 兼容的输入维度。这一“压缩融合”操作既保持了空间结构信息，又有效减少了传入语言模型的视觉 token 数量，从而平衡了计算负担。此外，在图像位置编码中，模型引入二维 RoPE，用于增强图像中空间关系的表达能力，并在视频处理中采用三维 patch 分组策略，即将相邻帧共同编码，进一步提升了序列建模效率。

为了提升跨模态的嵌入一致性，Qwen2.5-VL 引入了一套轻量级的视觉语言融合模块。该模块通过 MLP 将视觉编码输出压缩至语言模型的嵌入空间，完成跨模态对齐。这种“先视觉压缩、再语言处理”的策略有效避免了图像 token 对语言模型计算资源的挤占，使得整个系统在保持图文对齐能力的同时实现了可扩展性。

在多模态序列融合过程中，Qwen2.5-VL 采用了统一的输入格式：文本 token 与图像 token 被共同序列化，并加上统一的特殊标记（如<image>, <text>, <patch>等）以提示模型输入的模态边界。在训练中，模型支持纯文本任务、多模态问答、图文推理、表格解析等任务的统一处理，并支持最长 32768 长度的上下文序列，保证其在长文本和多图任务中的稳定性。

此外，Qwen2.5-VL 在数据构建和训练机制上也做了大量工程优化。其训练数据量从前代的 1.2 万亿 token 扩展至 4.1 万亿 token，涵盖了图文对话、结构化文件解析、OCR、学术知识、视频描述、图像标注、Agent 交互等多模态子任务。模型以阶段化方式进行预训练与指令微调，逐步提升其从通用视觉理解到复杂跨模态任务执行的能力。

3 系统架构

3.1 总体架构

本系统旨在构建一个融合多模态 RAG 检索机制与医学领域微调大语言模型（Qwen2.5-VL）的医学智能生成系统，支持多种模态输入形式与交互方式，实现医学图文信息的联合理解、知识检索与响应生成。系统架构如图所示，整体由三大核心模块构成：多模态 RAG 检索模块、下游生成模型模块，以及多端交互接口模块。各模块之间具备良好的解耦性与可组合性，能够适配不同的推理策略与应用需求，体现出系统较强的可扩展性与工程实用性。

系统首先接收用户输入，支持纯文本、医学图像以及图文混合三种输入形式。文本内容可包括用户提问、症状描述或电子病历片段；图像内容则涵盖 X 光、CT、MRI 等典型医学影像。系统会将输入中的文字与图像分别处理，并根据配置策略选择是否进入多模态 RAG 模块进行外部检索，或直接拼接形成下游模型可接受的输入格式。此阶段实现了输入模态的动态融合与路径分流，确保系统在不同任务场景下具备灵活的响应能力。

针对信息检索需求，系统内嵌了一个自研的多模态 RAG 模块，能够支持图像、文本双模态的向量化索引与语义检索。在该模块中，模型可以根据当前输入提取最相关的图文片段，从医学知识库或历史病例库中调用相关图像及文字说明。此 RAG 模块也设计了多种“嵌入-查找”方式，后续会进行详细介绍，并进行实验对比。检索到的 RAG 内容会与提问模板和原始输入共同综合成最终的输入内容，送入生成模块进行回复。

生成模块支持两种运行方式：其一为使用经过医学微调的本地模型 Qwen2.5-VL-Med，具备更强的图文语义对齐能力与医学知识表达能力；其二为调用云计算平台上已部署的大模型 API，如阿里云、百度云、DeepSeek 等平台提供的通用或专用模型。在资源允许的前提下，Qwen2.5-VL-Med 能够更准确地完成医学图像分析、结构化问答生成、病理解释等任务，尤其适用于对专业性与语义准确性要求较高的应用场景。同时，云端 API 方式具有更高的灵活性与性价比，用户可根据任务需求灵活选择合适的模型资源，同时系统架构亦对未来模型的快速替换

与升级保持开放。一旦有更强性能的模型发布，系统可通过更换 API 接入配置快速完成迁移，降低部署成本并提升可维护性。

系统最终将生成内容通过不同形式的前端交互方式进行展示，包括自定义构建的前端界面、命令行终端交互界面，以及 Qwen 官方提供的可视化 WebUI。不同的展示方式满足了从本地调试到在线演示的多种需求，进一步提升了系统的灵活性与可用性。

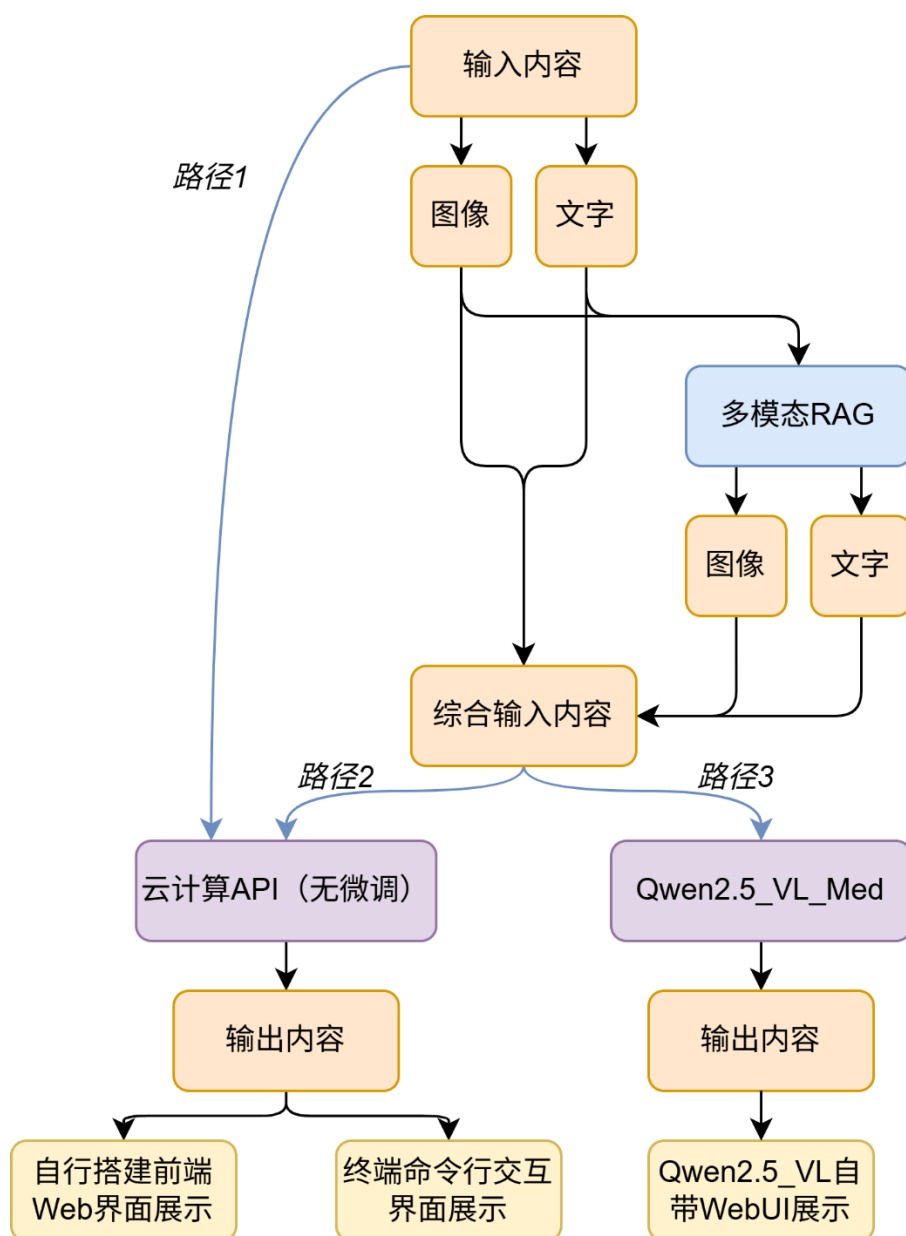


图 3.1 系统总体架构

系统提供了三种主要的推理路径，分别适配不同的模型能力、资源条件与实际应用场景，具备良好的灵活性与功能扩展性。三种路径在输入处理与模型调用流程上存在差异，能够实现从轻量通用模型调用到专业模型深度推理的渐进式处理策略。

路径一：是最基础的通用处理路径，直接将用户原始输入（包括图像与文本）送入通用大语言模型进行生成，不引入任何外部知识检索模块，即直接与多模态大模型进行对话。这一路径模拟了真实世界中用户在未参考任何背景材料的情况下，基于直觉或已有知识进行回答的方式，适用于模型预训练知识覆盖较全、问题复杂度较低的任务场景。该方式具有响应速度快、系统结构简单的优点，但在专业性强或背景依赖性高的问题上可能生成质量有限。

路径二：引入了系统自行实现的多模态 RAG 模块，先对用户输入进行图文联合检索，从知识库或历史案例库中提取出与输入内容相关的图像与文本信息作为参考，再与原始输入整合构建统一的输入序列送入下游模型生成。这种方式通过提供外部知识增强模型上下文理解能力，在不改变生成模型架构的前提下显著提升了答案的专业性与语义准确性。该路径可类比为“非专业人员在查阅参考资料后进行作答”，具有较好的通用性与知识适应能力。

路径三：在路径二的基础上进一步提升了生成模块的专业能力，采用经过医学领域微调的 Qwen2.5-VL-Med 模型替代通用大模型进行内容生成。在该路径中，模型不仅具备了更强的图文对齐能力，还拥有更丰富的医学语义理解能力与诊疗场景适应能力，能够实现图像分析、病理识别、医学解释等复杂任务。该路径模拟了“专业医生在参考临床样例后进行诊断回复”的行为过程，是系统中精度最高、能力最强的推理方式。

通过这三种路径的组合设计，系统实现了“无检索+通用模型”、“检索增强+通用模型”与“检索增强+专业模型”三种层级式问答方式，增强了系统的适用广度，也为后续任务性能评估与范式对比实验提供了良好的结构支持。

3.2 RAG 系统架构

在本研究构建的 RAG 系统架构中，针对检索触发机制的设计进行了重新审视与调整。在上文中提到，一些改进型 RAG 方法（如 Self-RAG[49]）引入了“是

否需要检索”的判断机制，即由大模型根据用户输入自动决定是否执行知识检索，以节省上下文空间和计算开销。这类方法在早期大模型上下文长度受限的情况下具有一定的实用价值。然而，随着新一代大语言模型上下文处理能力的迅速提升，尤其是本文所采用的 Qwen2.5-VL 模型已支持高达 32k token 的上下文长度，原本紧张上下文资源已不再是限制因素。在这种背景下，“按需检索”的机制反而可能因判断不准而漏检关键知识，导致模型回答的准确性下降。基于此考虑，本文所设计的 RAG 系统回归了经典的全检索策略，即在每一次对话中均默认触发外部知识检索，确保模型始终具备丰富的语义参考背景，从而最大程度提升医学问答任务中信息完整性与生成质量。该策略虽然在理论上引入了更多上下文信息，但在当前模型架构与推理能力支撑下，已不构成负担，反而更符合专业任务对信息准确性与覆盖性的需求。

根据 LangChain 官方文档，将多模态内容（以下都以图片代称）引入 RAG 系统主要有三种范式。在本系统中，三种范式都得到了实现。需要说明的是，尽管参考了 LangChain 的官方文档，本文并未使用 LangChain 的架构，而是从底层一步步搭建 RAG 系统。

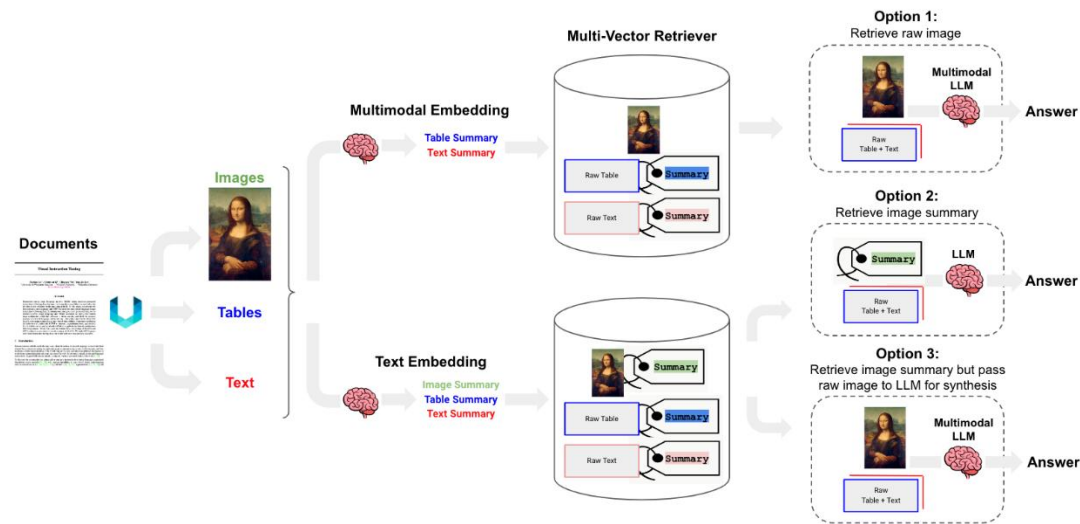


图 3.2 前三种多模态 RAG 实现范式

范式 1： 使用多模态对齐的嵌入系统，将图片和文字嵌入到同一空间中。检索返回的内容也是原始图片和文字。之后将查找到的图片和文字送入到多模态大

模型中进行回复。

这种范式使用的是一种可以同时处理图像和文字的嵌入系统，它把两种不同模态的信息映射到一个共享的语义空间中，特点在于保留了图像中原始的视觉细节，同时使图像和文字可以用相同的方式进行比对和检索。检索结果会直接返回原始图像和对应文字，从而为下游大模型提供最丰富的输入信息。在医学应用中，这种方式尤其有效，因为医学图像中的很多诊断依据往往体现在边缘模糊、形状异常等微小特征上，这些是文字描述难以完全表达的。因此，让模型直接处理原始图像更有助于生成准确的结果。不过，这种方法需要模型具有很强的图文对齐能力，否则嵌入不准会导致检索效果下降。此外，图像的向量表示比文字更占用存储空间，索引构建和检索过程也更复杂。

范式 2：利用大模型对图片生成缩略文字描述，或者图片本身自带缩略文字描述。将图片的缩略描述嵌入数据库中，检索返回的也是缩略文字描述，而非图片本身。之后将查找到的缩略文字描述和文字送入到多模态大模型中进行回复。这种方式主要依靠语言文字空间对图片的进行语义表达，本质上退化成了一种普通的文字单模态 RAG 系统。

这种范式采取了一种更轻量的方案：先使用大模型对图像生成一段简短的文字说明，或使用图像自带的描述，将这段文字作为图像的代表参与检索。由于整个过程只在文本空间中完成，因此系统结构更简单，运行效率也更高。这种方式适用于图像中没有特别多细节要求的任务，比如大致判断图文是否匹配。但它的问题在于，图像信息在转换为文字描述时会有不可逆的压缩。如果生成的描述缺少重要特征或者存在偏差，后续大模型在生成回答时就可能出现误判。这意味着模型的理解仅依赖于那段文字本身，而不是图像原貌，一旦文字表述不准确，输出的结果也可能失真。

范式 3：与第二步相同，首先获得图片的缩略文字描述。将图片的缩略描述嵌入数据库中，但检索返回的是图片本身。之后将查找到的图片和文字送入到多模态大模型中进行回复。这种方式利用缩略描述作为图片的标签，仍然依靠语言文字空间对图片的进行语义表达，但在输入模型前重新升维为原始图片。

这种范式在设计上试图平衡检索效率和输出质量：图像在构建数据库时仍然被转成文字描述用于检索，但在模型实际生成回答之前，会把对应的原始图像重

新取回，作为最终输入的一部分。这种方式在索引阶段仍然保持轻量化，不需要存储和计算图像向量，但又在生成阶段保留了原图，使得模型可以访问完整的图像细节。这对某些图文之间语义差异较大的情况特别有用，比如医学图像报告中那些依赖图像细节判断的任务。不过，这种方法的流程相对更复杂，图像和缩略描述之间的映射关系需要被严格管理，否则可能出现图文不一致的错误。同时，处理链条变长，系统在设计和维护上也更具挑战。

此外，本文还设计了一种新的多模态 RAG 范式，并且在实验时取得了最佳的效果（具体结果见实验部分）：

范式 4： 对于本身自带标签文字的图片，本文使用多模态嵌入模型，将图片本身和文字描述都进行嵌入。当查找到到图片时，返回图片连同其文字描述。同样，当查找到文字时，也返回文字及其代表的图片。在送入多模态大模型时，将图片及其文字描述一同作为参考送入。这种方法相当于绑定了文字-图片对，并且可以从文字语义和图片语义分别访问到这个文字-图片对。

这种范式在语义表达上结合了范式 1 的高保真与范式 3 的结构灵活，同时避免了语义对齐失败的风险，显著提升了图文检索的鲁棒性。例如，在用户上传一张 X 光片并提问其潜在病变时，该范式不仅能够检索到相似图像，还能同时返回清晰的文字说明，如“胫骨骨折”，从而为下游模型生成更准确的医学回答提供充分支撑。该范式结合了图像与语言两个模态的优势，是目前更适用于医学图文问答任务的一种 RAG 策略。

表 3.1 四种多模态 RAG 范式对比

	范式 1	范式 2	范式 3	范式 4
使用多模态嵌入	√	×	×	√
使用多模态大模型	√	×	√	√
使用图片描述	×	√	√	√
模型输入图片	√	×	√	√
模型输入图片描述	×	√	×	√

在系统的工程实现方面，本文构建了一个高度灵活、结构清晰的多模态 RAG

系统，不仅支持多种图文嵌入范式的切换，还具备良好的跨平台扩展能力与嵌入模型适配机制。整个系统在嵌入模块与检索模块之间通过接口规范进行解耦，将向量嵌入过程封装为独立的可配置组件，用户可根据使用场景与资源条件灵活选择本地模型或远程 API。具体而言，该系统支持本地部署的 Huggingface 模型、Sentence-Transformers 框架下的轻量模型，以及远程云服务平台（如 Cohere 等）所提供的文本或图像向量化服务，从而在保持准确性与性能的基础上，兼顾部署环境的弹性与经济性。

在模型嵌入空间的设计上，系统对文字与图像采用统一的语义向量空间进行建模，实现跨模态内容的可比性与一致性。实验中默认使用 Huggingface 中的“facebook/opt-2.7b”模型作为文字嵌入模型，图像部分则采用“Salesforce/blip2-opt-2.7b”所提供的视觉编码器，两者共享同一个 BLIP2 嵌入空间，使得图文信息可以在同一索引结构中进行语义对齐与联合检索。该设计有效提升了系统对跨模态信息融合与组织能力，为构建多模态医学知识库提供了技术基础。

为适配多样化的图文配对与检索模式，并实现四种范式，系统在 RAG 模块设计中引入了“图文对嵌入模式”参数，使得 RAG 模块不但可以解析并嵌入文本和图片，也可以解析图文对并根据设置选择其中所需要的模态进行嵌入，在查询时也可以根据图文对中的任意一种模态查询到整个图文对，提升了检索系统在不同任务设定下的灵活性与鲁棒性。与此同时，系统设计了一个独立的模式控制器，用于在上层任务中动态实例化不同的 RAG 结构，基于设定的任务范式自动构造合适的向量数据库与检索逻辑，形成一个高度可配置、组件独立、调用便捷的 RAG 工程框架。这种设计方法将 RAG 模块和具体的模态解耦，使得 RAG 模块有更强的泛用性，并且代码结构更加清晰。

此外，系统使用了上文中介绍的 FAISS 向量数据库进行向量存储及查找，具备优秀的近似最近邻搜索能力，尤其适合在大规模向量空间中执行快速相似度匹配，因此非常适用于本系统中对图像、文本及其组合表示的检索需求。系统还引入索引状态的持久化机制。具体而言，系统支持将当前的文档内容与对应的 FAISS 索引结构分别保存为本地文件，便于在不同运行阶段或部署场景中进行加载与恢复。用户在构建知识库后可将其存档，在后续使用时直接加载索引，无需重复计算嵌入，大幅缩短初始化时间。此外，该机制还为模型热更新、跨任务迁

移以及多环境部署提供了技术基础，使得知识库的维护与版本管理更加灵活高效。

与后续大模型微调过程不同，系统在构建知识库阶段更侧重于引入具有指导性、示范性和图文配合特征的内容，而非追求样本分布的均衡性与统计代表性。基于这一目标，所选数据涵盖了多种类型的信息来源，包括医学 Wiki、专业医学指导书籍等纯文本类文献，以及 PathVQA[122]、MedICaT[123]、MedPix[124]等公开医学图像问答与图文匹配数据集。此外，系统还整合了部分来自外部合作医疗机构的真实诊疗资料，进一步增强了知识库的临床适应性与语义多样性。

3.3 基于 Qwen2.5-VL 的大模型微调

由于通用大模型相较专用大模型泛化能力更佳且可塑性更强，而 Qwen2.5-VL 作为最新、社区支持最好的大模型，在国际多模态大模型竞技场中也有优异的成绩，同时作为国内模型对中文支持较好，因此本文选择 Qwen2.5-VL 模型作为基准模型进行微调。由于是医学领域微调，因此将微调的版本称为 Qwen2.5-VL-Med。

为适应医学图文问答任务对专业性和多模态语义理解能力的双重要求，本研究在 Qwen2.5-VL 模型基础上进行参数高效微调。考虑到全参数微调成本极高、硬件资源消耗大、泛化能力下降等问题，本文采用 LoRA (Low-Rank Adaptation) [111]作为主干微调策略，并且在文字和多模态上分别进行了微调，分别增强了模型的医学文字理解能力和医学多模态分析能力。

LoRA 微调是一种简单但高效的模型微调方式，传统全参微调方式通常需要对整个模型的参数进行更新，计算和存储开销极大，尤其在处理数十亿甚至上百亿参数的模型时，这一过程不仅耗费资源，还难以迁移和部署。LoRA 通过引入低秩矩阵对模型的部分权重进行近似替代，从而显著降低了训练所需的参数量，同时保持模型性能。

LoRA 的核心思想是，对于一个已经训练好的全连接层权重矩阵 $W \in \mathbb{R}^{d \times k}$ ，LoRA 并不直接对其进行更新，而是将其表示为一个低秩形式：

$$W' = W + \Delta W = W + BA \quad (3.1)$$

其中 $A \in \mathbb{R}^{r \times k}$ 、 $B \in \mathbb{R}^{d \times r}$ ，秩 $r \ll \min(d, k)$ 。也就是说，原始权重的更新部分 ΔW 被限制在一个低秩空间中。这种设计的关键在于：在训练过程中，只对

A 和 B 两个小矩阵进行学习，而原始权重 W 保持冻结不变。因此，大多数参数保持不动，只有极少数新引入的参数参与反向传播，显著减少了计算开销与显存需求。

训练完成后，推理阶段将 BA 与原始权重 W 合并，使得最终部署过程与原模型等价，不增加额外计算复杂度。这种高效性使得 LoRA 特别适用于低资源场景、多任务适配、边缘部署等需求场景。

在具体实现上，本项目选取了 Transformer 结构中的多个关键线性变换模块作为 LoRA 插入点，包括注意力机制中的 q_proj 、 k_proj 、 v_proj 、 o_proj ，以及前馈模块中的 $gate_proj$ 、 up_proj 与 $down_proj$ 等部分，确保模型对输入图文信息的表示路径在训练中具有足够的调整空间。所设定的 LoRA 秩为 32，表征矩阵的自由度较高，适合对图文联合语义进行细粒度调控；同时设置了适度的 Dropout 机制以提升模型在训练过程中的泛化能力。

作为 LoRA 的一种量化模型版本，尽管 QLoRA[112]微调速度更快，并且占用更少显存，同时 NF4 量化也能保持较高保真度，但本质上仍存在信息丢失的风险，因此本系统没有选择 QLoRA 微调方法。

3.3.1 医学语义理解能力微调

如前所述，为了分别提升模型在医学内容生成任务中的专业语义理解能力与图片理解能力，本系统采用了分阶段微调策略，依次在医学文本数据集与医学多模态数据集上进行训练，以实现从语言理解到图文联合推理能力的系统性强化。微调基于 Qwen2.5-VL 的 32B 参数版本，采用单张 NVIDIA H20 显卡（96GB 显存）进行训练。选择该卡是因为该卡基于 Ampere 架构，支持 FlashAttention2[125] 优化，具备出色的浮点性能和内存容量，同时也是大陆市场上可以使用到的显存容量最大的计算卡。在训练设置上，batch size 设为 8，并通过梯度累积机制每 4 个 batch 执行一次参数更新，相当于实现了 32 的等效 batch size，有效避免了大 batch 对显存的压力。同时，为进一步提升训练效率，系统启用了 FlashAttention2 优化加速机制，并使用 bfloat16 精度以充分发挥硬件特性，兼顾训练速度与数值稳定性。考虑到训练数据规模较大，且以覆盖多样语义为目标，因此每个数据集均仅训练一个 epoch，以加快迭代并防止过拟合。

表 3.2 CMD 数据集不同领域问答具体数量

内科	外科	儿科	妇产科	男科	肿瘤科
220,606	115,991	101,602	183,751	94,596	75,553

在第一阶段微调中，系统选用了 Chinese Medical Dialogue (CMD)数据集作为训练语料。该数据集由开源社区发布，是当前中文医疗问答领域最具代表性和应用广度的数据资源之一，广泛用于医疗问答系统开发、医学对话模型训练及各类大语言模型的医学适配研究。CMD 数据集涵盖内科、外科、妇科、儿科、口腔、精神等多个医学子领域，具有极高的专业性与语义多样性，共包含约 80 万个多轮问答对。为控制训练量并确保数据均衡分布，本文从其中每个领域中随机抽取了 2 万个样本，总计构建了一个包含 12 万个问答对的训练集，覆盖症状描述、诊断建议、药物使用、检验项目解释等多个典型医疗问答任务。这一阶段训练主要用于加强模型在医学语言理解、专业术语识别与问答结构生成方面的能力，为后续图文联合建模奠定语言语义基础。

表 3.3 CMD 数据格式样例

部门	标题	问题	回答
心血管科	高血压患者能吃党参吗？	我有高血压这两天女婿来的时候给我拿了些党参泡水喝，您好高血压可以吃党参吗？	高血压病人可以口服党参的。党参有降血脂，降血压的作用，可以彻底消除血液中的垃圾，从而对冠心病以及心血管疾病的患者都有一定稳定预防工作作用，因此平时口服党参能远离三高的危害。另外党参除了益气养血，降低中枢神经作用，调整消化系统功能，健脾补肺的功能。感谢您的进行咨询，期望我的解释对你有帮助。
		哪家医院能治胃反流	建议你用奥美拉唑同时，加用吗丁啉或莫沙必利或援生力维，另外还可以加用达喜片
		烧心，打隔，咳嗽低烧，以有 4 年多	

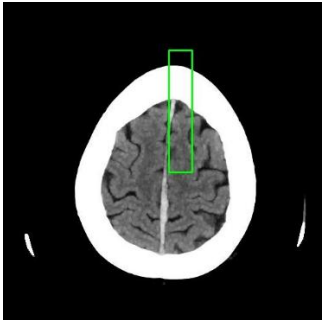
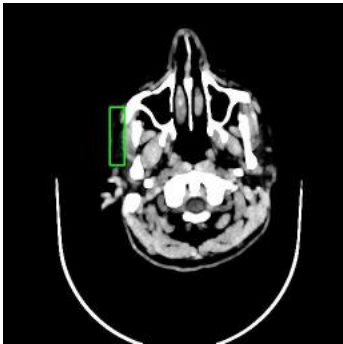
3.3.2 医学图像理解能力微调

在完成文本语义能力的基础微调后，系统进一步对模型的视觉感知能力进行针对性增强，特别是提升其在医学影像场景中的判读能力与多模态信息对齐能力。本阶段使用的训练数据为 MedTrinity-25M 数据集[113]中的 25M_demo 子集。MedTrinity-25M 是目前规模最大的医学图文三元组数据集之一，包含超过 2500 万个由“图像-感兴趣区域（ROI）-文字描述”构成的样本，涵盖 CT、MRI、X 光、超声、皮肤镜等 10 种以上成像模态，并涵盖 65 种以上疾病类别，具备高度的模态多样性与病理复杂性。25M_demo 子集包含约 16 万张高质量图像及其描述，本文从中随机抽取 2 万张图像与对应描述构建训练集，以控制训练时间并保持图像质量的一致性。

针对图像输入的特殊性，训练时启用了图像增强与规范化策略，确保不同模态图像在输入阶段统一分辨率与通道规范。batch size 维持在 4，并设置梯度累积步数为 8，以保持等效 batch size 为 32，适应图像处理所需的显存资源。

该阶段训练的主要目标是强化模型在医学图像语义建模、文字描述生成与图文一致性理解等方面的能力。通过对大规模图文对进行微调，模型获得了对图像中关键信息（如病灶位置、形态、密度等）的识别能力，并能够结合图像与文字信息进行联合推理，生成语义完整、符合医学逻辑的自然语言答案。结合前一阶段语言训练的成果，模型在多模态内容生成任务中展现出显著的表现提升，尤其是在图文协同诊断、影像问答与病例解释等任务上表现出较强的专业理解能力和语言生成质量。

表 3.4 MedTrinity 数据集格式样例

图片	Id	说明
	8031efe	The image is a non-contrasted computed tomography (CT) scan of the brain, showing the cerebral structures without any medical devices present. The region of interest, located centrally and in the middle of the image, exhibits an area of altered density, which is indicative of a brain hemorrhage. This area is distinct from the surrounding brain tissue, suggesting a possible hematoma or bleeding within the brain parenchyma. The location and characteristics of this abnormality may suggest a relationship with the surrounding brain tissue, potentially causing a mass effect or contributing to increased intracranial pressure.
	0-1b5c-11ef-8929-000066532cad	
	803c9df	The image is a computed tomography (CT) scan of the brain, showing the intricate structures within the cranial cavity. The region of interest, located left-center horizontally and in the middle vertically, occupies approximately 0.7% of the area. This region exhibits an abnormality, likely representing a subdural hemorrhage, characterized by a difference in density compared to the surrounding brain tissue. The affected area's position relative to other brain structures suggests a potential mass effect or displacement of adjacent brain tissue, which could be indicative of the hemorrhage's impact on surrounding areas.
	2-1b5c-11ef-a25b-000066532cad	

3.4 API 调用

考虑到完整部署原版 Qwen2.5-VL 模型（尤其是 32B 参数规模）在云计算环境中对 GPU 资源的占用极为高昂，其运行通常需要单张 A100 或更高级别的显卡，租赁成本不容忽视。为了在开发与实验过程中降低硬件成本，同时提升模型调用的灵活性与便捷性，系统在不依赖 Qwen2.5-VL-Med 微调版本的场景中，优先选择通过 API 方式调用模型服务进行交互。

采用 API 调用的方式不仅规避了本地部署过程中复杂的环境配置与模型加载延迟问题，还具备高度的开发友好性与跨平台可迁移性。借助 API 接口，开发者可以专注于系统功能和交互逻辑的设计，无需关心底层模型的硬件部署细节。此外，云平台所提供的模型服务通常拥有自动版本升级、负载均衡和容错机制，可在多种请求负载下保持稳定响应，这也为系统在多用户、并发访问场景中的可靠性提供了保障。

在实际工程实现中，系统已集成了多个主流平台的 API 服务调用能力，包括百度千帆平台、阿里云百炼平台、DeepSeek 模型云服务，以及本地兼容 OpenAI 接口的 Ollama 部署环境。这些平台基本覆盖了当前主流中文与多模态大模型的云端服务，开发者可通过统一接口配置进行调用，从而快速适配不同模型方案。

得益于模块化设计与平台抽象机制，API 调用方式也为系统的未来升级提供了良好的可扩展性。当出现新的开放 API 平台或更具性能优势的预训练模型时（如 Qwen3-VL、DeepSeek-VL 等），系统可通过简单配置完成模型替换，无需大幅改动代码结构或重构后端流程，从而提升系统在快速迭代和技术演进中的适应能力。同时，该机制也具备较好的跨平台迁移能力，可在不同计算环境中无缝迁移部署，满足多样化场景下的应用需求。整体而言，API 调用方式不仅作为高性价比的模型交互手段有效补充了本地部署方案，也体现了系统在资源控制与工程可维护性方面的综合设计优势。

4 实验

4.1 基准测试

本文采用 OmniMedVQA 数据集[93]作为对比测试的主要基准数据集。OmniMedVQA 由 Hu 等人提出，是当前规模最大、模态最丰富、覆盖解剖区域最广的医学多模态问答评测集，专为医学大模型（Medical LLM）能力评估而设计，适用于检验模型在医学图文语义理解、模态推理与跨模态生成方面的泛化能力。

该数据集包含来自 73 个真实医疗场景的子数据集，总计收录 118,010 张图像和 127,995 个问答项目，涵盖 12 种成像模态（包括 CT、MRI、X 光、眼底照、超声、内镜、数字照片等）与 20 余个解剖区域（如脑、眼、肺、消化道、骨骼等），并围绕五大任务类型设计了问答模板：成像模态识别（Modality Recognition）、解剖区域定位（Anatomy Identification）、疾病诊断（Disease Diagnosis）、病变评级（Lesion Grading）以及其他生物特征识别（Other Biological Attributes）。每个问答项目均被构造为多选题形式，并由 ChatGPT API 辅助生成干扰项与自然语言重写，保证了语言表达的多样性与干扰强度。

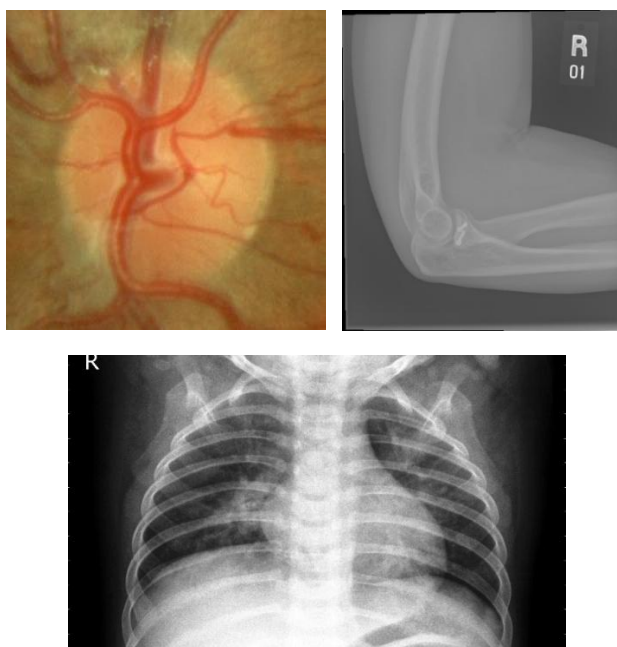


图 4.1 OmniMedVQA 基准数据集部分图片展示

在评测方法方面，OmniMedVQA 提供了两种互补的指标体系：

1. Question-Answering Score: 在这一评估框架下，研究者将每个问题及其候选选项整合为一个具备指令语义的 Prompt，输入给待测 LVLM 模型。例如，输入模板可能为：“This is a medical question with several options... Please select the correct answer. The question is: <question>. The candidate options are: <options>”。模型随后会生成一个自由文本的答案，而不是直接选择一个选项。为了判定其准确性，研究者会将模型生成的答案与所有选项分别计算相似度，并选择最相似者作为预测答案。最后，将该预测与真实标签进行比对以判断是否正确。这种方法模拟了真实使用场景中用户对模型的自然语言提问行为，检验模型对指令的理解、图文信息整合以及在选择任务中的可靠性，尤其反映其“指令跟随”能力（instruction-following）和上下文推理能力。

2. Prefix-based Score: 该方法更强调模型对图像与文本之间联合概率建模的内在能力。在这一设置中，图像首先被编码为视觉特征，然后以“前缀（prefix）”的方式与问题和每一个候选答案拼接，形成完整的文本输入序列，并将视觉特征作为上下文注入语言模型。随后，对每个答案候选项，模型会计算其生成该候选答案的概率得分（likelihood score）。最终得分最高的候选项被视为模型的预测结果。这种评估方式并不依赖模型真实输出的自然语言内容，而是从“哪个答案更可能被模型生成”的角度进行推断，体现了模型在内在知识表征与图文对齐方面的能力。它尤其适用于那些生成结果不稳定或对 Prompt 敏感性较高的模型，有助于避免因语言输出噪声而掩盖模型潜在能力的问题。

当嵌入 RAG 时，本文对两个指标体系分别进行了补充修改。对于 Question-Answering Score 体系，额外检索到的信息内容通常以附加段落的形式拼接到原始输入中，从而为模型提供额外的上下文信息。对于 Prefix-based Score 体系，系统在构造输入序列时，会将由检索模块返回的相关内容一并作为额外上下文，与视觉前缀和文本问题共同输入到模型中。

4.2 范式对比与消融实验

在本实验小节中，由于涉及到使用 API 这种外部封装方法，无法读取候选答案的生成概率，并且反复询问价格也较高。因此，只随机保留了 2000 个问答项

目（未区分所属模态和问题模板），得出的实验数据不能与后续实验小节中的完整实验结果进行对比。由于 API 无法读取候选答案的生成概率，因此只使用了 Question-Answering Score 体系作为评判标准。

首先对比的是三种处理路径，分别对应了完全无微调、增加多模态 RAG、增加多模态 RAG 并微调大模型三种情况。其中 RAG 方法使用的是范式 4。

表 4.1 处理路径消融实验

路径	RAG	生成模型	Question-Answering Score
路径 1	×	Qwen2.5-VL (API)	32.50
路径 2	√	Qwen2.5-VL (API)	43.15
路径 3	√	Qwen2.5-VL-Med	48.20

从处理路径的消融实验可以看出，本文的 RAG 系统和模型微调都发挥了一定作用。RAG 系统提供了任务上下文之外的外部医学知识，使得大模型在生成回答时不再仅依赖于图像与问题中的显性信息，而是可以基于检索到的相似图文对作为隐性提示完成语义迁移与知识补全。在路径 1 中，模型完全依赖于原始输入（图像与问题）作答，表现出一定的基础识别能力，但整体分数偏低，说明 Qwen2.5-VL 内置知识与多模态推理能力在医学专业领域仍有欠缺。而在路径 2 中引入 RAG 模块后，即便模型本身未进行任何微调，仅凭外部信息的补充与语言提示的重构，就显著提升了答案的匹配准确度，这说明医学图文问答任务中，RAG 系统的知识补全作用是实质性的。

进一步地，路径 3 在路径 2 的基础上引入了本文微调的模型 Qwen2.5-VL-Med，相较于通用模型更贴合医学语义场景。微调后的模型在相同 RAG 输入条件下进一步取得了较为显著的提升，表明大模型本体的领域适应性同样至关重要。尤其是在候选答案语义相近、题干歧义较强的问题样本中，微调模型表现出更强大的上下文理解与细节判别能力，体现了专业微调在问答精度方面的正向增益。此外，微调使用的 MedTrinity 数据集和实验所使用的基准测试数据集在医学模态上较为相似（都包含多种扫描图片），可能使模型相关生成能力产生了较大进步。

接下来将进一步探讨不同 RAG 范式之间的对比效果。不过为了实验的便捷，均选用了使用外部 API 进行生成的方法。此外为了对比，引入了仅使用文本内容进行构建的数据库作为参考。

表 4.2 不同 RAG 范式之间的对比效果

范式	Question-Answering Score
纯文本 RAG	34.45
范式 1	37.75
范式 2	31.80
范式 3	35.15
范式 4	43.15

从实验结果可以看出，不同 RAG 范式在 Question-Answering Score 上的表现存在明显差异，说明检索路径中图文组织方式、模态嵌入方式及返回结果的处理策略对最终内容生成效果具有显著影响。

范式 4 取得了最高的 Question-Answering Score，在所有策略中表现最为优异。其优势在于对图像与文字进行联合嵌入，并在数据库中存储图文对而非单一模态。检索结果不仅包括语义相近的图像，还附带与其绑定的文字标签。模型在生成回答时能够同时参考图像和精确语义，解决了范式 1 中存在的“仅有图像缺乏解释”的问题，也规避了范式 2 中压缩表达所带来的语义丢失问题。这种双向补全机制赋予了系统更强的表达力与上下文完整性，特别适用于图像语义高度专业、但又需要自然语言辅助定位的医学场景。

范式 1 作为一个对图像和文字进行对齐嵌入的典型方案，其表现优于纯文本 RAG 与范式 2。该策略能保留图像本身的原始视觉细节，有利于模型从中提取微小但关键的医学信号。然而，由于缺乏显式的图像语义标签，生成模型完全依赖对图像的自主感知能力，这在某些模态模糊或图像质量不佳的样本中可能造成理解困难，进而限制其上限性能。

相比之下，范式 2 退化为单模态文字 RAG，采用图像描述生成文本并以此构建数据库。甚至由于使用了通用大语言模型对部分图片生成了无用甚至错误的文

字标签，再叠加 API 调用的不确定性，使得 RAG 得分甚至意外得低于纯文字 RAG 的效果。由于整个流程仅在语言空间中完成，其实现代价最低，运行效率高，但模型的判断完全依赖于图像文字描述的质量。一旦图像描述缺失或简化过度，将导致模型无法恢复原始图像所包含的复杂临床特征，从而使得回答偏离真实诊断语义。该范式在指标上表现最差，验证了医学场景中对图像原始特征的保留具有不可忽视的重要性。

范式 3 作为一种折中策略，在检索阶段使用图像描述构建索引，提升了查询效率，同时在生成阶段恢复了图像输入，弥补了语义压缩带来的信息缺失。该方法在指标上略优于范式 2，但相比纯文本 RAG 并未拉开较大差距，原因可能在于图像与标签之间的索引映射存在误差，且提供的图片缺乏文字描述，导致其在推理过程中的贡献不稳定。

综上，实验验证了图像信息在医学问答中的重要作用，也说明了图文联合嵌入与图文对绑定检索策略的有效性。

4.3 模型评估与分析

在本实验小节中，将会使用范式 4 的多模态 RAG 与 Qwen2.5-VL-Med（路径 3）的总体系统，与其他相关模型进行比较。评测 OmniMedVQA 数据集中全部的问答项目。

4.3.1 基于任务类型评估

为系统性评估多模态 RAG 系统在不同任务维度下的适应性，本节首先从 OmniMedVQA 中划分的五类问答模板入手，对比不同模型在各任务类型下的表现，并基于 Question-Answering Score 与 Prefix-based Score 两个指标展开分析。

表 7 是 13 种不同模型在 OmniMedVQA 基准数据集的五类问答模板上的 Question-Answering Score 与 Prefix-based Score 两个指标具体得分情况，其中每列得分模板中最高分被加粗标注。这 13 种不同的模型包括了 8 种通用模型和医学专用模型，以及本文的 RAG+Qwen2.5-VL-Med 方法。从评测结果表格可以看出，本文系统“多模态 RAG+Qwen2.5-VL-Med”方法展现出全面而稳定的性能优势，验证了多模态检索增强生成机制在医学问答场景下的实用性与先进性。

表 4.3 模型对不同问答模板结果（斜杠前为 Question-Answering Score，斜杠后为 Prefix-based Score）

模型	成像模 态识别	解剖区域 定位	疾病诊 断	病变评 级	其他生物特 征识别	总计
MiniGPT-4 [87]	28.23/	28.41/	30.51/	32.85/	37.30/	30.53/
	28.71	30.26	20.24	38.53	43.28	25.68
BLIP-2 [84]	57.51/	49.19/	46.24/	30.52/	73.52/	50.69/
	37.85	64.75	23.19	25.03	37.70	33.43
InstructBLIP	70.62/	42.75/	33.62/	54.60/	48.16/	42.49/
[114]	24.83	51.78	22.41	54.71	28.85	28.71
mPLUG-owl	27.93/	24.44/	30.31/	38.50/	37.84/	29.90/
[115]	16.46	30.45	29.32	76.96	43.70	29.89
Otter [117]	25.62/	25.14/	27.12/	34.22/	32.12/	27.20/
	10.53	23.47	22.11	39.53	24.74	21.17
LLaVA [20]	27.21/	25.99/	27.35/	38.53/	32.68/	27.85/
	14.43	15.08	19.60	26.82	38.24	20.02
LLaMA-	44.51/	33.73/	29.17/	39.07/	36.80/	33.15/
Adapter-v2 [116]	33.48	40.67	24.93	38.65	34.07	29.88
VPGTrans [118]	29.32/	30.76/	26.91/	29.61/	32.65/	28.49/
	31.77	36.27	18.88	38.53	41.47	26.15
Med-Flamingo	28.67/	25.32/	41.47/	31.25/	35.27/	36.17/
[89]	21.21	25.16	26.10	49.60	34.97	26.54
RadFM [120]	21.31/	19.96/	28.46/	24.72/	37.58/	26.82/
	38.57	29.48	25.66	35.86	31.73	29.00
MedVInT	59.79/	41.36/	36.79/	15.49/	46.79/	41.50/
[121]	30.92	23.90	25.02	5.23	30.12	25.81
LLaVA-Med	31.38/	28.34/	28.01/	32.35/	29.23/	28.78/
[119]	12.02	22.06	27.25	30.98	25.88	24.06
RAG+Qwen2.5-	64.00/	40.44/	52.01/	55.08/	38.60/	51.36/
VL-Med	27.98	22.77	33.64	36.72	30.17	30.08

从整体性能来看，本文系统在 Question-Answering Score 评估框架中以 51.36 分的平均得分排名第一，明显优于包括 BLIP-2(50.69)、InstructBLIP(42.49)、MedVInT(41.50) 等领先模型，体现了其出色的指令理解能力、图文融合能力与专业内容生成能力。特别是在 疾病诊断(Disease Diagnosis) 和 病变评级(Lesion Grading) 两类对医学知识准确性和上下文建模要求较高的任务中，系统分别取得 52.01 和 55.08 的高分，大幅领先于其他模型。说明系统能够有效结合输入图像中的视觉线索与外部知识库中的诊疗参考内容，生成符合临床逻辑的答案。此外，在成像模态识别、解剖结构定位等基础任务上亦保持稳定表现，说明其在多模态输入条件下具备广泛的知识迁移能力与任务适应性。

在 Prefix-based Score 评估中，本文系统同样获得 30.08 分的第二高分，仅次于 BLIP-2(33.43)，高于 InstructBLIP、MedVInT 等通用或医学模型。这种评估方式更关注模型对于“候选答案”之间语言生成概率的建模能力。因此，本文系统虽不以语言模型精确排序见长，但依然展现出不俗的选项判断能力，表明其内部图文联合建模结构具有较强的选择倾向性。需要指出的是，Prefix-based Score 并不充分衡量模型的真实语义理解与回答生成能力，而偏向于测试“语言模型偏好性”。相比之下，本文系统在 Question-Answering Score 中的卓越表现，反映出其更真实、更贴近实际应用的任务能力。

此外，值得关注的是，与部分通用预训练模型（如 BLIP-2）相比，本文系统所使用的 Qwen2.5-VL-Med 是经过双阶段医学任务微调的模型，尤其强调语义理解和结构化生成能力，而非语言建模的概率极大化。在实际应用中，这种模型更倾向于通过深度图文理解与知识综合，生成具有逻辑性、可解释性和临床参考价值的内容，而非“猜出”最可能的选项。因此，其在 Prefix-based Score 略逊一筹，但在实际生成任务中的专业性与稳定性优势则更为突出。

4.3.2 基于成像模态评估

除了任务类型的划分，医学图像问答系统还需面对多种成像模态的输入挑战。医学影像不仅包括常见的 X 光、CT、MRI，还可能涉及超声、病理切片、内窥镜图像等，每种成像模态在图像分辨率、纹理特征、语义信息密度和结构表现形式上存在显著差异。例如，CT 和 MRI 偏重于三维结构的精细分析，而 X 光图像

则更强调二维平面中的密度变化。因此，本小节将从具体模态维度进一步分析不同模型在各类医学图像输入条件下的性能差异，验证系统在跨模态理解与具体结构认知上的稳健性。

表 4.4 模型在具体方向下的 Question-Answering Score

模型	Co	CT	DP	FP	IRI	MR
MiniGPT-4	23.67	25.84	18.05	42.37	38.51	27.6
BLIP-2	48.52	56.74	23.01	57.66	66.18	41.77
InstructBLIP	32.25	28.72	35.75	37.72	59.27	33.79
mPLUG-owl	36.69	24.54	19.81	41.81	38.44	29.82
Otter	33.73	18.53	18.2	37.7	30.7	26.37
LLaVA	12.72	17.73	22.25	32.16	31.23	26.99
LLaMA-Adapter-v2	38.46	21.41	28.93	36.85	35.7	27.23
VPGTrans	32.54	21.26	20.1	34.5	32.6	25.36
Med-Flamingo	18.4	38.47	21.48	27.61	39.69	40.01
RadFM	15.43	27.44	13.25	28.99	36.13	24.16
MedVInT	39.17	40.74	43.89	39.69	46.22	42.84
LLaVA-Med	28.99	18.69	18.34	35.14	30.68	27.49
RAG+Qwen2.5-VL-Med	41.02	28.91	40.27	58.33	66.00	44.57

在具体的成像模态表格中，Co 表示阴道镜检查（Colposcopy），CT 表示计算机断层扫描（Computed Tomography），DP 表示数码摄影（Digital Photography），FP 表示眼底摄影（Fundus Photography），IRI 表示红外反射成像（Infrared Reflectance Imaging），MR 表示磁共振成像（Magnetic Resonance Imaging），OCT 表示光学相干断层扫描（Optical Coherence Tomography），Der 表示皮肤镜检查（Dermoscopy），End 表示内窥镜检查（Endoscopy），Mic 表示显微镜图像（Microscopy Images），US 表示超声成像（Ultrasound）。由于表格过宽，因此表格被划分成两个部分展示。

从 Question-Answering Score 的整体表现来看, RAG+Qwen2.5-VL-Med 在多个模态任务上均处于领先地位, 特别是在眼底摄影 (FP)、红外成像 (IRI)、数码摄影 (DP) 和 内窥镜 (End) 等场景中取得了显著优势。例如, 本文系统在 FP 模态下达到 58.33 分, 远高于 BLIP-2 (57.66) 和 InstructBLIP (37.72); 在 IRI 模态下也基本追平了 BLIP-2 (66.00 vs 66.18), 大幅领先于其余模型。这些结果表明, 结合 RAG 模块对医学图文知识进行动态扩充, 能显著增强模型对低频、结构复杂模态的理解能力, 尤其在需要细粒度语义配对的图文任务中表现突出。

续表 4.4 模型在具体方向下的 Question-Answering Score

模型	OCT	Der	End	Mic	X-Ray	US
MiniGPT-4	31.4	40.09	30.26	28.05	39.75	25.5
BLIP-2	68.08	41.07	48.85	50.17	70.55	37.27
InstructBLIP	42.59	61.86	36.65	48.2	61.21	41.25
mPLUG-owl	43.76	35.98	24.45	25.99	28.29	21.4
Otter	29.64	42.66	33.94	22.94	31.73	23.49
LLaVA	33.73	49.67	38.2	27.95	31.35	18.66
LLaMA-Adapter-v2	33	51.43	46.62	34.78	46.7	34.05
VPGTrans	25.14	44.66	30.53	23.61	46.53	25.45
Med-Flamingo	26.51	32.33	30.97	46.6	28.3	24.64
RadFM	32.8	39.03	28.4	24.81	29.21	16.57
MedVInT	23.26	29.13	30.11	40.71	56.62	41.26
LLaVA-Med	34.61	44.9	41.88	26.33	31.26	29.88
RAG+Qwen2.5-VL-Med	39.04	51.54	52.80	47.76	63.12	39.65

尽管在 CT 和 MRI 等结构信息较强的成像模态下, BLIP-2 和 MedVInT 等模型略占优势, 但本文系统依然取得了接近甚至超越水平的成绩。例如, 在 MR 模态中, RAG+Qwen2.5-VL-Med 得分 44.57, 高于 BLIP-2 (41.77) 和 Med-Flamingo (40.01), 仅次于表现最好的 MedVInT (42.84), 说明本系统具备良好的跨模态

语义对齐与医学推理能力。

表 4.5 模型在具体方向下的 Prefix-based Score

模型	Co	CT	DP	FP	IRI	MR
MiniGPT-4	26.33	29.46	22.94	23.59	43.61	12.75
BLIP-2	32.25	38.87	25.95	20.01	43.4	20.45
InstructBLIP	17.46	35.66	10.01	26.74	27.26	15.02
mPLUG-owl	18.64	37	18.31	56.87	44.96	13.22
Otter	2.37	32.55	20.78	16.96	21.9	10.73
LLaVA	33.73	38.27	23.55	13.84	36.37	4.65
LLaMA-Adapter-v2	28.4	36.03	21.36	22.11	33.69	17.26
VPGTrans	35.21	30.3	27.71	23.93	42.35	11.04
Med-Flamingo	30.56	22.25	7.4	51.57	30.22	14.43
RadFM	22.55	45.47	13.43	21.48	25.16	24.52
MedVInT	60.24	37.77	39.04	9.45	32.92	30.1
LLaVA-Med	13.61	36.75	10.48	16.15	21.61	24.58
RAG+Qwen2.5-VL-Med	30.02	38.07	23.34	30.46	22.44	31.92

从各类模态的细分任务结果来看，RAG+Qwen2.5-VL-Med 在多个典型医学图像任务中展现出领先性能。在 X-Ray（46.90）和 Ultrasound（US，41.05）等主流临床成像模态上，其得分超过大多数对比模型，尤其显著优于未经过医学微调的通用模型如 Otter（27.8/20.88）、LLaVA（24.1/20.74）以及 RadFM 等医学基线模型。在内窥镜图（33.77）与 显微镜图（27.93）等结构复杂、语义粒度较细的图像任务上，本文系统也展现出较高的视觉理解与语义判断能力，体现了其在细粒度医学视觉任务中的稳定性。

值得特别指出的是，本文系统在磁共振成像（MR）任务中取得了最高分（31.92），超过了包括 MedVInT（30.10）、RadFM（24.52）与 BLIP-2（20.45）在内的所有模型。这一结果充分说明：在结构化程度较高、影像细节关键的成像

任务中，结合 RAG 的内容增强机制与微调后的医学视觉-语言模型能有效提高模型对解剖结构、病灶分布等医学特征的建模质量，提升判断准确性。

续表 4.5 模型在具体方向下的 Prefix-based Score

模型	OCT	Der	End	Mic	X-Ray	US
MiniGPT-4	30	25.18	26.9	27.6	40.11	27.2
BLIP-2	18.7	19.8	25.95	28.43	49.81	81.79
InstructBLIP	51.74	29.53	28.77	22.8	40.85	58.51
mPLUG-owl	41.76	18.75	35.32	31.04	32.18	29.14
Otter	45.98	22.52	23.39	21.99	27.8	20.88
LLaVA	51.23	14.17	22.16	17.65	24.1	20.74
LLaMA-Adapter-v2	54.22	21.61	32.23	30.6	41.71	47.51
VPGTrans	31.6	22.18	24.77	25.87	39.94	40.89
Med-Flamingo	58.57	39.48	46.61	23.04	38.13	17.42
RadFM	37.4	25.71	34.44	20.52	55.21	24.78
MedVInT	18.54	20.72	24.03	15.35	29.09	25.43
LLaVA-Med	51.51	25.35	27.92	17.4	25.54	16.75
RAG+Qwen2.5-VL-Med	50.23	19.15	33.77	27.93	46.90	41.05

RAG 机制在此过程中发挥了关键作用，它不仅补充了模型在专业知识覆盖方面的不足，还在面对如显微图（Mic）、内窥镜图（End）等低频训练样本场景中，通过引入语义一致的外部图文内容，有效缓解了模型“见图不识意”的问题。借助多模态 RAG 注入的提示，模型能够建立从图像内容到医学概念的结构映射，进而提高其对于低资源模态的预测能力。

进一步来看，RAG+Qwen2.5-VL-Med 模型在多模态能力与专业理解方面已具备跨模态泛化与结构认知优势。即便在面对如 BLIP-2、MedVInT 等基于大规模图文预训练的通用模型，其综合表现依然稳居前列，这也从侧面反映了 RAG 系统中“内容增强+结构注入”的方法对复杂场景有更好的适应能力。

此外，这些实验结果也揭示出当前医疗多模态模型仍存在“精度-模态迁移-任务广度”之间的平衡难题。部分模型如 Med-Flamingo 在 CT 类任务中表现较佳，但在 DP（数码摄影）、Co（阴道镜）等更偏非结构模态的任务中失分严重。相比之下，RAG+Qwen2.5-VL-Med 具备较强的任务稳定性与模态适应性，其设计思路为未来医学大模型的泛化性能优化提供了新的范式参考。

4.4 实验总结

通过在多个实验维度上的综合评估，本文提出的“多模态 RAG + Qwen2.5-VL-Med”系统在医学多模态问答任务中展现出全面的性能优势和良好的通用适应性。首先，在消融实验中，通过路径 1（无 RAG 无微调）、路径 2（RAG+未微调大模型）与路径 3（RAG+微调模型）的对比，系统性验证了多模态 RAG 机制与大语言模型专业微调在提升问答准确性与语义匹配度方面的有效性。其中路径 3 表现最佳，说明两者的组合产生了协同增益：RAG 机制强化了知识覆盖与上下文提示，大模型微调增强了专业语义建模与医学推理能力。

此外，本文通过四种多模态 RAG 范式的对比实验，进一步确认了范式 4（图文绑定检索）的性能优势，其联合嵌入和语义对齐能力显著提升了系统对高难度图像问答任务的适应力。相比之下，退化为文字 RAG 的范式 2 在医学任务中表现不佳，进一步强调了保留图像原始特征与语义补全的重要性。范式 1 与范式 3 则分别反映了纯图像嵌入与图像恢复策略的局限性与折中效果，从侧面印证了范式 4 的图文对的绑定方式兼顾了语义完整性与视觉细节保持，在医学图像任务中具有更优的上下文建模能力。

在整体系统评估方面，本文系统在 OmniMedVQA 医学问答基准数据集上的测试结果显示，在 Question-Answering Score 体系中取得了最高分（51.36），在 Prefix-based Score 体系中获得第二高分（30.08）。这一结果表明，系统在真实语义理解、图文融合及任务生成能力方面已达先进水平，尤其在疾病诊断与病变评级等对专业知识建模要求较高的子任务中优势突出。

具体到对每种模态任务的细粒度分析，本文系统在 X-Ray、US、Mic、End 等典型医学图像任务中均取得优异成绩，且在 MR 任务中以 31.92 的 Prefix-based Score 达到了全模型最高，说明其在结构化视觉信息建模方面具备更强的能力。

尤其是在复杂模态与低频任务中，RAG 机制与微调模型协同增强了模型的可解释性与稳定性，缓解了样本分布不均和模态知识稀疏问题。

总体而言，本实验章节全面展示了“多模态 RAG+医学大模型微调”体系在多个医学图文任务中的卓越表现与实践价值，验证了 RAG 机制与专业微调在医学智能问答中的协同优势，并为构建可扩展、可落地的医学多模态问答系统提供了坚实的实证支撑与范式借鉴。

5 交互设计

为了便于不同用户群体在多种使用场景下顺利使用系统并进行开发调试，本文设计并实现了三类具备互补功能的交互界面，分别面向终端调试、网页操作与远程部署等使用场景，兼顾开发效率与用户体验：（1）针对 API 开发过程中的快速验证需求，设计了基于命令行的终端交互方式，支持多轮对话与流式输出；（2）针对图文对话任务的可视化需求，构建了基于 Web 前端技术栈与 Flask 的网页交互界面，支持图像上传、Markdown 渲染与消息预览；（3）针对模型远程部署后的可视化交互需求，集成了 Qwen2.5-VL 官方 WebUI 界面，依托 Gradio 构建，适用于快速图文测试与展示。三类界面在功能设计上遵循奥卡姆剃刀法则：“如无必要，勿增实体”，在实现核心功能的同时，最大程度保持界面简洁与交互直观，避免复杂操作对用户造成认知负担。

5.1 终端交互

在模型调用与 API 调试阶段，为开发者提供了便捷的命令行交互通道。该方式支持模型响应内容的原始展示与结构化解析，适用于开发过程中的功能验证与问题定位。终端交互模式具备响应快、可控性强、依赖少等特点，尤其适合早期开发与快速测试场景。

```
choices=[Choice(delta=ChoiceDelta(content=None, function_call=None, refusal=None, role='assistant', tool_calls=None, reasoning_content=''), finish_reason='stop', index=0), Choice(delta=ChoiceDelta(content=None, function_call=None, refusal=None, role=None, tool_calls=None, reasoning_content='好的'), finish_reason='stop', index=1), Choice(delta=ChoiceDelta(content=None, function_call=None, refusal=None, role=None, tool_calls=None, reasoning_content='。'), finish_reason='stop', index=2), Choice(delta=ChoiceDelta(content=None, function_call=None, refusal=None, role=None, tool_calls=None, reasoning_content='用户'), finish_reason='stop', index=3), Choice(delta=ChoiceDelta(content=None, function_call=None, refusal=None, role=None, tool_calls=None, reasoning_content='问'), finish_reason='stop', index=4), Choice(delta=ChoiceDelta(content=None, function_call=None, refusal=None, role=None, tool_calls=None, reasoning_content='的是'), finish_reason='stop', index=5), Choice(delta=ChoiceDelta(content=None, function_call=None, refusal=None, role=None, tool_calls=None, reasoning_content='胆碱'), finish_reason='stop', index=6), Choice(delta=ChoiceDelta(content=None, function_call=None, refusal=None, role=None, tool_calls=None, reasoning_content='能'), finish_reason='stop', index=7), Choice(delta=ChoiceDelta(content=None, function_call=None, refusal=None, role=None, tool_calls=None, reasoning_content='性'), finish_reason='stop', index=8), Choice(delta=ChoiceDelta(content=None, function_call=None, refusal=None, role=None, tool_calls=None, reasoning_content='荨'), finish_reason='stop', index=9), Choice(delta=ChoiceDelta(content=None, function_call=None, refusal=None, role=None, tool_calls=None, reasoning_content='麻疹'), finish_reason='stop', index=10), Choice(delta=ChoiceDelta(content=None, function_call=None, refusal=None, role=None, tool_calls=None, reasoning_content='。'), finish_reason='stop', index=11), Choice(delta=ChoiceDelta(content=None, function_call=None, refusal=None, role=None, tool_calls=None, reasoning_content='我需要'), finish_reason='stop', index=12), Choice(delta=ChoiceDelta(content=None, function_call=None, refusal=None, role=None, tool_calls=None, reasoning_content='先'), finish_reason='stop', index=13), Choice(delta=ChoiceDelta(content=None, function_call=None, refusal=None, role=None, tool_calls=None, reasoning_content='确认'), finish_reason='stop', index=14)]
```

图 5.1 不对 API 返回内容进行解析的输出（流式输出，部分内容）

该交互方式支持与多个平台的 API 对接，并根据调用的模型类型自动适配响应处理逻辑，可以支持区分推理模型与非推理模型。系统在处理推理型模型时，

会在模型输出推理信息时自动提示，并在实际回复开始后进行视觉标记，帮助用户分辨生成过程的不同阶段。此外，终端交互支持流式输出模式，模型响应可以按 Token 实时打印至终端界面，模拟实际对话过程中的连续生成体验，可以避免完全输出后再给出答案造成的极大延迟。

```
Test LLM API:
Question: 简短回答，什么是胆碱能性荨麻疹？
Reasoning Content:
好的，用户问的是胆碱能性荨麻疹是什么。首先，我需要确认这个术语的正确性，确保拼写和术语准确。胆碱能性荨麻疹，英文是Cholinergic Urticaria，属于荨麻疹的一种类型。接下来，我要解释发病机制。胆碱能性荨麻疹可能与体内乙酰胆碱的释放有关，导致皮肤中的肥大细胞释放组胺，进而引发症状。症状方面，常见的有小而痒的红疹，周围可能有红晕，通常出现。接下来，要区分它和其他类型的荨麻疹的不同之处。比如，普通荨麻疹可能由过敏原引发，而胆碱能性更侧重于体温变化。此外，用户可能想知道治疗方法，比如抗组胺药物，避免诱因如剧烈运动。需要检查是否有常见的误区需要澄清，比如是否传染，是否属于严重疾病。通常，胆碱能性荨麻疹虽然不适，但多数情况下不危及生命，但会影响生活质量。用户可能还关心日常注意事项，比如。最后，确保回答简洁明了，符合用户要求的简短回答，同时涵盖关键点：定义、诱因、症状、治疗和注意事项。避免使用专业术语过多，用通俗易懂的语言解释。
Reply Content:
胆碱能性荨麻疹是一种因体温升高（如运动、情绪紧张、热水澡等）引发的荨麻疹，表现为皮肤出现小而痒的红疹，可能伴随刺痛感。通常由乙酰胆碱释放触发，抗组胺药可缓解，需避免诱因。
```

图 5.2 对 API 返回内容进行解析的输出（推理模型）

尽管终端界面无法直接展示图像内容，系统仍支持通过文件路径输入图像，实现图文联合输入的实验目的。该限制在后续的网页交互界面中被进一步克服，使用户能够获得更完整的图文交互体验。

5.2 前端交互实现

5.2.1 网页交互界面

为提升系统的可视化交互能力，系统基于 Web 前端技术（HTML、CSS、JavaScript）与后端框架 Flask，设计并实现了一套轻量化的图文对话界面。该界面不仅满足图文输入、流式响应、历史记录查看等基本需求，还对 Markdown 渲染、多模态展示做了特别支持，适合模型在线部署后的交互体验与展示演示需求。该界面适用于在浏览器端直接调用大模型 API 的对话任务，能够满足日常演示、调试和轻量级部署场景下的交互需求。

整个界面布局采用上下结构，顶部区域为对话历史展示区域，中部为滚动式聊天框，用于呈现用户与模型之间的问答过程。由于大模型倾向于输出 Markdown 格式内容，因此为了增强可读性，系统支持 Markdown 语法的渲染，能够清晰展示模型生成的结构化文本，包括段落、列表、代码块以及图像回显等内容。用户输入框设置在底部，配有文本输入区域与文件选择按钮，可上传本地图像作为辅

助信息参与对话。此外，网页的外观和配色遵循了作者的其他开源网页的家族化设计，保证了本文系统与作者其他开源项目的配色一致性（例如<https://liaoyanqing666.github.io/game.html>）。



图 5.3 网页交互界面与图片预览

在交互逻辑设计方面，网页界面支持多轮连续对话，用户提交输入后，系统会立即显示消息内容，并在模型生成回复的过程中以流式方式实时更新响应，提升交互的即时性与沉浸感。对于上传的图像，界面可实现即时预览，并将图像附加至对应的对话消息中，便于用户直观追踪输入信息的完整性与上下文关系。整个交互过程采用气泡式布局，明确区分用户提问与模型回复，使对话内容层次分明、视觉逻辑清晰。值得一提的是，网页界面延续了终端交互中的流式输出解析机制，支持将模型生成的每个 **Token** 实时显示，模拟自然会话过程中的连续响应体验，有效避免了等待全部内容生成后才统一展示所造成的延迟问题，从而提升了系统在人机交互场景下的响应效率与使用流畅度。



图 5.4 图文结合多轮对话

此外，界面采用响应式设计，能够适配不同分辨率设备，包括移动端与桌面端，确保在不同使用场景下均具备良好的交互体验。整体风格以简洁柔和为主，配色温和，具备较高的可读性与可用性，为多模态医学内容生成任务提供了直观、友好的交互通道。

5.2.2 Qwen2.5-VL 交互界面

除了在调用 API 时自定义的网页前端交互界面外，在调用远程部署的 Qwen2.5-VL-Med 微调模型时，系统还支持直接使用 Qwen2.5-VL 模型自带的官方 WebUI 进行交互。该交互界面由 Qwen 模型开发团队提供，使用 gradio 库开发，作为其多模态模型的默认可视化入口，具有高度集成与即用即启的特点，便于在模型部署完成后立即进行图文对话测试与演示。

由于微调模型部署在云端计算平台，因此若需在本地浏览器中访问其 WebUI 界面，需要通过网络端口映射进行“穿刺”操作。在本项目中，采用了 AutoDL 平台提供的 AutoDL SSH 隧道工具，将远程服务器的 WebUI 监听端口通过本地端口进行映射，使得用户可以在本地浏览器中访问远程部署的 Qwen2.5-VL-Med 模型交互界面。具体而言，通过一条包含本地端口与远程地址绑定的 SSH 命令，

即可在不中断云端服务的前提下，实现图形界面的无缝访问，确保微调模型的调试与展示能够在本地完成。



图 5.5 使用 AutoDL SSH 隧道工具进行端口映射

Qwen2.5-VL 交互界面与前述本文自行实现的网页交互界面功能基本相同，以多轮对话形式组织交互内容，用户可以直接在浏览器中输入自然语言问题，并上传图像文件或通过拖拽方式添加视觉输入。系统会自动将图像编码并与问题合并作为统一输入送入模型处理，模型生成的回复内容将以气泡式结构实时展示在页面中。整体 UI 设计简洁明了，用户操作无需额外学习成本，适合用户快速上手使用。



图 5.6 Qwen2.5-VL 交互界面

6. 结论与展望

6.1 结论

本文围绕“融合 RAG 机制的多模态医学大模型微调”这一课题，构建了一个面向医学场景的内容生成系统，具备图文联合理解能力，并能够结合外部知识进行动态调用。针对当前医学人工智能应用中常见的图文协同不足、知识更新滞后和医学知识表达复杂等问题，本文从系统结构设计、技术实现方法到实验验证等多个方面展开研究，提出了一种集成多模态信息处理与知识增强生成能力的解决方案，为医学大模型应用提供了新的实践路径和工程框架。

在系统设计上，本文首次在医学场景中将多模态 RAG 检索机制与大语言模型微调技术结合，建立了统一的信息处理流程。系统能够将图像和文本共同送入 RAG 模块进行向量检索，并将检索结果与原始输入拼接后输入至微调后的大模型中进行生成。这种方案不仅保留了医学图像中关键视觉信息，还引入了与用户问题相关的参考内容，从而提升了输出结果的专业性和针对性。相比传统仅依赖语言生成的问答系统，该方法引入的外部知识动态补全机制，有效缓解了训练语料不足和样本不均的问题，特别适用于高专业要求的医学任务。实验证明，该系统在 OmniMedVQA 标准数据集上的整体表现优于几乎所有现有方法，体现了该架构在真实应用中的可行性与实用价值。

在 RAG 机制设计上，本文系统性地实现了多种多模态范式，包括图文共嵌检索、文本代理检索和图文对绑定方式。特别是本文提出的范式四——图文绑定检索机制，在保持图像原始信息的同时，将其标签性描述一并作为索引内容，形成语义与视觉信息相结合的检索体系。该范式适用于图像主导或文本主导的不同查询需求，提升了系统在多种输入条件下的检索准确性和生成一致性。实验对比显示，该范式在多个子任务上的性能表现均优于其他方案，证明了其在医学信息检索生成场景中的有效性。

针对模型训练，本文提出了基于语言数据和多模态数据的分阶段微调方法，用于对 Qwen2.5-VL 通用大模型进行专业适配。第一阶段以 CMD 中文医学问答数据为基础，通过语言层面的预训练提升模型对医学术语和问诊语境的理解能力；

第二阶段则使用 MedTrinity-25M 图文对数据进一步训练，增强模型在图文联合输入场景下的响应能力和视觉信息提取能力。这种训练策略能够充分利用不同类型数据的优势，使模型更好地适应医学领域中复杂多变的输入模式。最终形成的 Qwen2.5-VL-Med 模型在多项任务中均展现出优异性能，尤其在医学图像分析和跨模态推理方面取得显著提升，验证了双阶段微调策略在专业大模型训练中的实际效能。

本文在系统部署与工程实现层面同样注重开发友好性与平台适应性，引入了模块化设计理念，对嵌入模型、检索模式、模型推理方式等核心模块进行接口抽象与解耦，使系统具备良好的可拓展性和跨平台移植能力。系统支持多种嵌入方法、检索策略、模型部署形式（包括 API 调用与本地部署），并提供可视化前端支持图文上传与生成回显。系统整体架构稳定，能够适配不同推理路径与模型资源，在不同医学应用场景中具备良好的实用性与推广潜力。

6.2 展望

尽管本文提出的融合多模态 RAG 机制与大语言模型微调的医学内容生成系统在多个任务中展现出优异性能，但仍存在一些尚未解决的问题和进一步研究的空间。首先，目前系统中的 RAG 模块主要依赖静态构建的向量索引库，虽然能够满足多数场景下的检索需求，但缺乏对复杂医学知识结构的支持。实际临床决策过程中往往涉及疾病演进路径、诊疗流程以及结构化检查指标等高度关联的信息，仅靠向量语义匹配难以全面覆盖这些异构知识。因此，未来可以考虑将结构化医学知识图谱、标准电子病历（EMR）模板或病理树引入到 RAG 模块中，与当前的图文检索机制形成互补，从而提升知识召回的覆盖率与准确性。

在模型微调方面，本文采用了分阶段的 LoRA 微调策略，虽已实现了基础的语言理解与多模态对齐能力，但仍属于“离线训练、静态模型”的范式。当前最新的语言模型训练范式，如 RLHF（基于人类反馈的强化学习）与 DPO（Direct Preference Optimization）等方法，已在提升模型输出稳定性、专业性及对偏好任务的适应性方面展现出显著优势。未来的研究可以尝试将此类人类偏好学习机制引入到医学图文生成任务中，使模型更贴近临床实际所需，例如生成更符合医生语言风格的诊断建议、病情描述或检查解读文本。

系统目前的评估主要集中在 OmniMedVQA 等公开数据集上，尚未在真实医疗场景中部署进行测试。考虑到医疗系统的实际工作流程复杂多变，未来需开展面向医院信息系统（HIS）、放射科 PACS 系统等平台的适配实验，并评估系统在医生交互、患者支持和教学训练等场景下的泛化能力、稳定性和响应效率。同时，进一步引入医务人员的主观评估与实际任务参与，将有助于对系统功能进行针对性的优化迭代。未来研究可从以下几个方向展开：一是引入结构化医学知识图谱与电子病历系统，构建具备语义关联与事实一致性的多模态知识源；二是探索多轮对话与上下文记忆机制，增强系统的交互能力与语义连贯性；三是结合大模型低成本推理与边缘部署方案，推动系统在基层医疗、移动终端等场景中的落地应用。

综上所述，本文所构建的多模态 RAG 医学内容生成系统不仅在技术路径上具有创新性，更在实际性能与适应性上展现出强大的应用潜力，为构建面向医疗行业的智能内容生成系统提供了有益的技术范式与实践探索，力求搭建更符合医学实际需求、更具智能性与实用性的多模态 AI 系统，推动医学大模型在临床应用、医学教育、科研辅助等多个维度的落地发展，服务于智慧医疗与健康生活的远大目标。

参考文献

- [1] Elhaddad M, Hamam S. AI-driven clinical decision support systems: an ongoing pursuit of potential[J]. Cureus, 2024, 16(4).
- [2] Khosravi M, Zare Z, Mojtabaieian S M, et al. Artificial intelligence and decision-making in healthcare: a thematic analysis of a systematic review of reviews[J]. Health services research and managerial epidemiology, 2024, 11: 23333928241234863.
- [3] Krishnan G, Singh S, Pathania M, et al. Artificial intelligence in clinical medicine: catalyzing a sustainable global healthcare paradigm[J]. Frontiers in artificial intelligence, 2023, 6: 1227091.
- [4] Torab-Miandoab A, Samad-Soltani T, Jodati A, et al. Interoperability of heterogeneous health information systems: a systematic literature review[J]. BMC medical informatics and decision making, 2023, 23(1): 18.
- [5] Matarazzo A, Torlone R. A Survey on Large Language Models with some Insights on their Capabilities and Limitations[J]. arXiv preprint arXiv:2501.04040, 2025.
- [6] Ye J, Hai J, Song J, et al. The role of artificial intelligence in the application of the integrated electronic health records and patient-generated health data[J]. medRxiv, 2024: 2024.05. 01.24306690.
- [7] Wu J, Gan W, Chen Z, et al. Multimodal large language models: A survey[C]//2023 IEEE International Conference on Big Data (BigData). IEEE, 2023: 2247-2256.
- [8] Clusmann J, Kolbinger F R, Muti H S, et al. The future landscape of large language models in medicine[J]. Communications medicine, 2023, 3(1): 141.
- [9] Karabacak M, Margetis K. Embracing large language models for medical applications: opportunities and challenges[J]. Cureus, 2023, 15(5).
- [10] AlSaad R, Abd-Alrazaq A, Boughorbel S, et al. Multimodal large language models in health care: applications, challenges, and future outlook[J]. Journal

- of medical Internet research, 2024, 26: e59505.
- [11] Sepehri M S, Fabian Z, Soltanolkotabi M, et al. MediConfusion: Can you trust your AI radiologist? Probing the reliability of multimodal medical foundation models[J]. arXiv preprint arXiv:2409.15477, 2024.
- [12] Gebreab S A, Salah K, Jayaraman R, et al. Llm-based framework for administrative task automation in healthcare[C]//2024 12th International Symposium on Digital Forensics and Security (ISDFS). IEEE, 2024: 1-7.
- [13] Gao Y, Xiong Y, Gao X, et al. Retrieval-augmented generation for large language models: A survey[J]. arXiv preprint arXiv:2312.10997, 2023, 2: 1.
- [14] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[J]. Advances in neural information processing systems, 2020, 33: 9459-9474.
- [15] Salemi A, Zamani H. Evaluating retrieval quality in retrieval-augmented generation[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024: 2395-2400.
- [16] Zhou P, Feng Y, Yang Z. Privacy-Aware RAG: Secure and Isolated Knowledge Retrieval[J]. arXiv preprint arXiv:2503.15548, 2025.
- [17] Fan W, Ding Y, Ning L, et al. A survey on rag meeting llms: Towards retrieval-augmented large language models[C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024: 6491-6501.
- [18] Matak M, Chudziak J A. GAIus: Combining genai with legal clauses retrieval for knowledge-based assistant[C]//Proceedings of the 17th International Conference on Agents and Artificial Intelligence. 2025, 3: 868-875.
- [19] Lála J, O'Donoghue O, Shtedritski A, et al. Paperqa: Retrieval-augmented generative agent for scientific research[J]. arXiv preprint arXiv:2312.07559, 2023.
- [20] Liu H, Li C, Wu Q, et al. Visual instruction tuning[J]. Advances in neural information processing systems, 2023, 36: 34892-34916.
- [21] Li J, Li D, Xiong C, et al. Blip: Bootstrapping language-image pre-training for

- unified vision-language understanding and generation[C]//International conference on machine learning. PMLR, 2022: 12888-12900.
- [22] Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report[J]. arXiv preprint arXiv:2303.08774, 2023.
- [23] Bai J, Bai S, Yang S, et al. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv preprint arXiv:2308.12966, 2023.
- [24] Mohsen F, Ali H, El Hajj N, et al. Artificial intelligence-based methods for fusion of electronic health records and imaging data[J]. Scientific Reports, 2022, 12(1): 17981.
- [25] Grossman Liu L, Grossman R H, Mitchell E G, et al. A deep database of medical abbreviations and acronyms for natural language processing[J]. Scientific Data, 2021, 8(1): 149.
- [26] Liang P, Pu B, Huang H, et al. Vision Foundation Models in Medical Image Analysis: Advances and Challenges[J]. arXiv preprint arXiv:2502.14584, 2025.
- [27] Zhu J, Ding X, Ge Y, et al. Vl-gpt: A generative pre-trained transformer for vision and language understanding and generation[J]. arXiv preprint arXiv:2312.09251, 2023.
- [28] Sarkar S, Jash D, Maji A, et al. Approach to unequal hilum on chest X-ray[J]. The Journal of Association of Chest Physicians, 2013, 1(2): 32-37.
- [29] Olatunji I E, Rauch J, Katzensteiner M, et al. A review of anonymization for healthcare data[J]. Big data, 2024, 12(6): 538-555.
- [30] Calvo F, Karras B T, Phillips R, et al. Diagnoses, syndromes, and diseases: a knowledge representation problem[C]//AMIA annual symposium proceedings. 2003, 2003: 802.
- [31] Anwar R W, Abrar M, Ullah F. Transfer learning in brain tumor classification: challenges, opportunities, and future prospects[C]//2023 14th International Conference on Information and Communication Technology Convergence (ICTC). IEEE, 2023: 24-29.

- [32] Wac M, Santos-Rodriguez R, McWilliams C, et al. Evaluation of a data annotation platform for large, time-series datasets in intensive care: Mixed methods study[J]. arXiv preprint arXiv:2410.16959, 2024.
- [33] Zhang W, Zhu L, Hallinan J, et al. Boostmis: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 20666-20676.
- [34] Schmarje L, Grossmann V, Zelenka C, et al. Annotating Ambiguous Images: General Annotation Strategy for High-Quality Data with Real-World Biomedical Validation[J]. arXiv preprint arXiv:2306.12189, 2023.
- [35] Wang H, Jin Q, Li S, et al. A comprehensive survey on deep active learning in medical image analysis[J]. Medical Image Analysis, 2024: 103201.
- [36] Hou J, Liu S, Bie Y, et al. Self-eXplainable AI for Medical Image Analysis: A Survey and New Outlooks[J]. arXiv preprint arXiv:2410.02331, 2024.
- [37] Holzinger A, Carrington A, Müller H. Measuring the quality of explanations: the system causability scale (SCS) comparing human and machine explanations[J]. KI-Künstliche Intelligenz, 2020, 34(2): 193-198.
- [38] Liu S, Zhang J, Xiang Y, et al. A study of data pre-processing techniques for imbalanced biomedical data classification[J]. International Journal of Bioinformatics Research and Applications, 2020, 16(3): 290-318.
- [39] Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of documentation, 1972, 28(1): 11-21.
- [40] Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond[J]. Foundations and Trends® in Information Retrieval, 2009, 3(4): 333-389.
- [41] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers).

- 2019: 4171-4186.
- [42] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of machine learning research, 2020, 21(140): 1-67.
- [43] Karpukhin V, Oguz B, Min S, et al. Dense Passage Retrieval for Open-Domain Question Answering[C]//EMNLP (1). 2020: 6769-6781.
- [44] Izacard G, Caron M, Hosseini L, et al. Unsupervised dense information retrieval with contrastive learning[J]. arXiv preprint arXiv:2112.09118, 2021.
- [45] Zhou J, Chen L. OpenRAG: Optimizing RAG End-to-End via In-Context Retrieval Learning[J]. arXiv preprint arXiv:2503.08398, 2025.
- [46] Cornelio C, Petruzzellis F, Lio P. Hierarchical Planning for Complex Tasks with Knowledge Graph-RAG and Symbolic Verification[J]. arXiv preprint arXiv:2504.04578, 2025.
- [47] Verma P, Midigeshi S P, Sinha G, et al. Plan-RAG: Planning-guided Retrieval Augmented Generation[J]. 2024.
- [48] Lyu Y, Niu Z, Xie Z, et al. Retrieve-plan-generation: an iterative planning and answering framework for knowledge-intensive LLM generation[J]. arXiv preprint arXiv:2406.14979, 2024.
- [49] Asai A, Wu Z, Wang Y, et al. Self-rag: Learning to retrieve, generate, and critique through self-reflection[C]//The Twelfth International Conference on Learning Representations. 2023.
- [50] Abootorabi M M, Zobeiri A, Dehghani M, et al. Ask in Any Modality: A Comprehensive Survey on Multimodal Retrieval-Augmented Generation[J]. arXiv preprint arXiv:2502.08826, 2025.
- [51] Hernandez-Salinas B, Terven J, ChaveZ-Urbiola E A, et al. IDAS: Intelligent Driving Assistance System using RAG[J]. IEEE Open Journal of Vehicular Technology, 2024.
- [52] Xia P, Zhu K, Li H, et al. Mmed-rag: Versatile multimodal rag system for medical vision language models[J]. arXiv preprint arXiv:2410.13085, 2024.

- [53] Yu D, Bao R, Mai G, et al. Spatial-rag: Spatial retrieval augmented generation for real-world spatial reasoning questions[J]. arXiv preprint arXiv:2502.18470, 2025.
- [54] Zhang T, Patil S G, Jain N, et al. Raft: Adapting language model to domain specific rag[C]//First Conference on Language Modeling. 2024.
- [55] Cheng M, Sun Y, Wang L, et al. Vista: Vision and scene text aggregation for cross-modal retrieval[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5184-5193.
- [56] Zhou T, Mei S, Li X, et al. MARVEL: unlocking the multi-modal capability of dense retrieval via visual module plugin[J]. arXiv preprint arXiv:2310.14037, 2023.
- [57] Zhou S, Alon U, Xu F F, et al. Docprompting: Generating code by retrieving the docs[J]. arXiv preprint arXiv: 2207.05987, 2022.
- [58] Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models[J]. Nature Medicine, 2025: 1-8.
- [59] Xiong G, Jin Q, Wang X, et al. Improving retrieval-augmented generation in medicine with iterative follow-up questions[C]//Biocomputing 2025: Proceedings of the Pacific Symposium. 2024: 199-214.
- [60] Zhou S, Alon U, Xu F F, et al. Docprompting: Generating code by retrieving the docs[J]. arXiv preprint arXiv: 2207.05987, 2022.
- [61] Wang Q, Ding R, Chen Z, et al. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents[J]. arXiv preprint arXiv:2502.18017, 2025.
- [62] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.
- [63] Lan Z, Chen M, Goodman S, et al. Albert: A lite bert for self-supervised learning of language representations[J]. arXiv preprint arXiv:1909.11942, 2019.
- [64] Clark K, Luong M T, Le Q V, et al. Electra: Pre-training text encoders as discriminators rather than generators[J]. arXiv preprint arXiv:2003.10555, 2020.

- [65] Joshi M, Chen D, Liu Y, et al. Spanbert: Improving pre-training by representing and predicting spans[J]. Transactions of the association for computational linguistics, 2020, 8: 64-77.
- [66] Sanh V, Debut L, Chaumond J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter[J]. arXiv preprint arXiv:1910.01108, 2019.
- [67] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training[J]. 2018.
- [68] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.
- [69] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
- [70] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models[J]. arXiv preprint arXiv:2302.13971, 2023.
- [71] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models[J]. arXiv preprint arXiv:2307.09288, 2023.
- [72] Grattafiori A, Dubey A, Jauhri A, et al. The llama 3 herd of models[J]. arXiv preprint arXiv:2407.21783, 2024.
- [73] Bai Y, Kadavath S, Kundu S, et al. Constitutional ai: Harmlessness from ai feedback[J]. arXiv preprint arXiv:2212.08073, 2022.
- [74] Akter S N, Yu Z, Muhamed A, et al. An in-depth look at Gemini's language abilities[J]. arXiv preprint arXiv:2312.11444, 2023.
- [75] Liu A, Feng B, Xue B, et al. Deepseek-v3 technical report[J]. arXiv preprint arXiv:2412.19437, 2024.
- [76] Guo D, Yang D, Zhang H, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning[J]. arXiv preprint arXiv:2501.12948, 2025.
- [77] Sun Y, Wang S, Feng S, et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation[J]. arXiv preprint arXiv:2107.02137, 2021.
- [78] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from

- natural language supervision[C]//International conference on machine learning. PmLR, 2021: 8748-8763.
- [79] Du Z, Qian Y, Liu X, et al. Glm: General language model pretraining with autoregressive blank infilling[J]. arXiv preprint arXiv:2103.10360, 2021.
- [80] Team I L M. Internlm: A multilingual language model with progressively enhanced capabilities[EB/OL].(2023-9-27)
- [81] Zeng W, Ren X, Su T, et al. PanGu-alpha: Large-scale autoregressive pretrained Chinese language models with auto-parallel computation[J]. arXiv preprint arXiv:2104.12369, 2021.
- [82] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- [83] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
- [84] Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models[C]//International conference on machine learning. PMLR, 2023: 19730-19742.
- [85] Zhang S, Roller S, Goyal N, et al. Opt: Open pre-trained transformer language models[J]. arXiv preprint arXiv:2205.01068, 2022.
- [86] Chung H W, Hou L, Longpre S, et al. Scaling instruction-finetuned language models[J]. Journal of Machine Learning Research, 2024, 25(70): 1-53.
- [87] Zhu D, Chen J, Shen X, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models[J]. arXiv preprint arXiv:2304.10592, 2023.
- [88] Li C, Wong C, Zhang S, et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day[J]. Advances in Neural Information Processing Systems, 2023, 36: 28541-28564.
- [89] Moor M, Huang Q, Wu S, et al. Med-flamingo: a multimodal medical few-shot learner[C]//Machine Learning for Health (ML4H). PMLR, 2023: 353-367.

- [90] Luo R, Sun L, Xia Y, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining[J]. Briefings in bioinformatics, 2022, 23(6): bbac409.
- [91] Wang Z, Wu Z, Agarwal D, et al. Medclip: Contrastive learning from unpaired medical images and text[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing. 2022, 2022: 3876.
- [92] Bai S, Chen K, Liu X, et al. Qwen2. 5-vl technical report[J]. arXiv preprint arXiv:2502.13923, 2025.
- [93] Hu Y, Li T, Lu Q, et al. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 22170-22183.
- [94] Karpukhin V, Oguz B, Min S, et al. Dense Passage Retrieval for Open-Domain Question Answering[C]//EMNLP (1). 2020: 6769-6781.
- [95] Robertson S E, Walker S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval[C]//SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University. Springer London, 1994: 232-241.
- [96] Févry T, Soares L B, FitzGerald N, et al. Entities as experts: Sparse memory access with entity supervision[J]. arXiv preprint arXiv:2004.07202, 2020.
- [97] Wang L, Yang N, Wei F. Query2doc: Query expansion with large language models[J]. arXiv preprint arXiv:2303.07678, 2023.
- [98] Yang H, Li Z, Zhang Y, et al. PRCA: Fitting black-box large language models for retrieval question answering via pluggable reward-driven contextual adapter[J]. arXiv preprint arXiv:2310.18347, 2023.
- [99] Wang H, Li R, Jiang H, et al. Blendfilter: Advancing retrieval-augmented large language models via query generation blending and knowledge filtering[J]. arXiv preprint arXiv:2402.11129, 2024.

- [100] Xu F, Shi W, Choi E. Recom: Improving retrieval-augmented lms with compression and selective augmentation[J]. arXiv preprint arXiv:2310.04408, 2023.
- [101] Hofstätter S, Chen J, Raman K, et al. Fid-light: Efficient and effective retrieval-augmented text generation[C]//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023: 1437-1447.
- [102] Wang Y, Li P, Sun M, et al. Self-knowledge guided retrieval augmentation for large language models[J]. arXiv preprint arXiv:2310.05002, 2023.
- [103] Jiang Z, Xu F F, Gao L, et al. Active retrieval augmented generation[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 7969-7992.
- [104] Wen H, Song X, Chen X, et al. Simple but effective raw-data level multimodal fusion for composed image retrieval[C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024: 229-239.
- [105] Jafari O, Maurya P, Nagarkar P, et al. A survey on locality sensitive hashing algorithms and their applications[J]. arXiv preprint arXiv:2102.08942, 2021.
- [106] Malkov Y A, Yashunin D A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 42(4): 824-836.
- [107] Douze M, Guzhva A, Deng C, et al. The faiss library[J]. arXiv preprint arXiv:2401.08281, 2024.
- [108] Yadav N, Monath N, Zaheer M, et al. Adaptive retrieval and scalable indexing for k-nn search with cross-encoders[J]. arXiv preprint arXiv:2405.03651, 2024.
- [109] Wang J, Yi X, Guo R, et al. Milvus: A purpose-built vector data management system[C]//Proceedings of the 2021 International Conference on Management of Data. 2021: 2614-2627.
- [110] Su J, Ahmed M, Lu Y, et al. Roformer: Enhanced transformer with rotary

- position embedding[J]. Neurocomputing, 2024, 568: 127063.
- [111] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models[J]. ICLR, 2022, 1(2): 3.
- [112] Dettmers T, Pagnoni A, Holtzman A, et al. Qlora: Efficient finetuning of quantized llms[J]. Advances in neural information processing systems, 2023, 36: 10088-10115.
- [113] Xie Y, Zhou C, Gao L, et al. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine[J]. arXiv preprint arXiv:2408.02900, 2024.
- [114] Panagopoulou A, Xue L, Yu N, et al. X-instructblip: A framework for aligning x-modal instruction-aware representations to llms and emergent cross-modal reasoning[J]. arXiv preprint arXiv:2311.18799, 2023.
- [115] Ye Q, Xu H, Xu G, et al. mplug-owl: Modularization empowers large language models with multimodality[J]. arXiv preprint arXiv:2304.14178, 2023.
- [116] Gao P, Han J, Zhang R, et al. Llama-adapter v2: Parameter-efficient visual instruction model[J]. arXiv preprint arXiv:2304.15010, 2023.
- [117] Li B, Zhang P, Yang J, et al. Otterhd: A high-resolution multi-modality model[J]. arXiv preprint arXiv:2311.04219, 2023.
- [118] Zhang A, Fei H, Yao Y, et al. Vpgtrans: Transfer visual prompt generator across llms[J]. Advances in Neural Information Processing Systems, 2023, 36: 20299-20319.
- [119] Li C, Wong C, Zhang S, et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day[J]. Advances in Neural Information Processing Systems, 2023, 36: 28541-28564.
- [120] Wu C, Zhang X, Zhang Y, et al. Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data[J]. arXiv preprint arXiv:2308.02463, 2023.
- [121] Zhang X, Wu C, Zhao Z, et al. Pmc-vqa: Visual instruction tuning for medical visual question answering[J]. arXiv preprint arXiv:2305.10415, 2023.

- [122] He X, Zhang Y, Mou L, et al. Pathvqa: 30000+ questions for medical visual question answering[J]. arXiv preprint arXiv:2003.10286, 2020.
- [123] Subramanian S, Wang L L, Mehta S, et al. Medicat: A dataset of medical images, captions, and textual references[J]. arXiv preprint arXiv:2010.06000, 2020.
- [124] Siragusa I, Contino S, La Ciura M, et al. Medpix 2.0: a comprehensive multimodal biomedical dataset for advanced AI applications[J]. arXiv preprint arXiv:2407.02994, 2024.
- [125] Dao T. Flashattention-2: Faster attention with better parallelism and work partitioning[J]. arXiv preprint arXiv:2307.08691, 2023.
- [126] 丁宁, 宋雨欣, 单泽田, 等. 基于检索增强生成 (RAG) 技术的医学教学辅助智能问答系统的构建探索[J]. China Medical Education Technology, 2025, 39(1).
- [127] 任海玉, 刘建平, 王健, 等. 基于大语言模型的智能问答系统研究综述[J]. Journal of Computer Engineering & Applications, 2025, 61(7).

致 谢

感谢一路走来所有帮助过我的导师们，包括课程老师。特别感谢校内的论文导师张惠然和全程导师余航。此外，感谢西湖大学吴泰霖老师对我科研道路的指导。

作者落款：黎思源

完成地点：上海大学

2025 年 5 月 12 日

