E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(2025)05-1286-17

论文引用格式:Li H T, Liu H, Chen G G, Yan Z S, Sheng Y R and Zhang C M. 2025. Semantic fine-tuning and cross-modal retrieval-augmented Chinese medical report generation. Journal of Image and Graphics, 30(5):1286-1302(李恒泰, 刘慧, 陈公冠, 闫子申, 盛玉瑞, 张彩明. 2025. 语义微调和跨模态检索增强的中文医学报告生成. 中国图象图形学报, 30(5):1286-1302)[DOI:10.11834/jig.240451]

语义微调和跨模态检索增强的中文医学报告生成

李恒泰[1,2],刘慧[1,2*],陈公冠[1,2],闫子申[1,2],盛玉瑞[3],张彩明[2,4]

1. 山东财经大学计算机与人工智能学院,济南　250014; 2. 山东省数字经济轻量智算与可视化重点实验室(山东财经大学),济南　250014; 3. 山东第一医科大学第一附属医院放射科,济南　250013; 4. 山东大学软件学院,济南　250101

摘　要: **目的**　医学报告生成旨在根据医学影像生成准确的诊断结果,以减轻医生负担、提高临床工作效率。然而,中文医学报告生成在准确理解医学影像及规范描述医学报告方面仍存在局限,并存在幻觉问题。为应对上述挑战,提出一种基于语义微调和跨模态检索增强的中文医学报告生成模型FRCM。**方法**　基于多模态大模型LLaVA,对其视觉编码器和大语言模型进行领域适配与微调,并提出一种通用数据与垂域数据协同训练策略:利用通用数据提高模型对复杂指令的理解能力,利用垂域数据使模型具备医学图像—文本对齐能力及专业的中文医学报告生成能力。在推理阶段,提出一种新的跨模态检索增强策略,利用引导知识有效缓解模型的幻觉问题,进一步提高模型生成医学报告的准确性和鲁棒性。**结果**　在中文数据集MIMIC-CXR上,与XrayGLM和XrayPULSE模型相比,FRCM在双语评估替代指标BLEU-*n*的BLEU-1、BLEU-2、BLEU-3、BLEU-4、基于最长公共子序列的召回率指标ROUGE-L、显式顺序翻译评价指标METEOR和基于共识的图像描述评估指标CIDEr这7个指标上分别提升了10.4%、10.1%、9.7%、9.1%、6.6%、9.4%和38.4%。与LLaVA和Qwen-VL上微调过的模型相比,FRCM在BLEU-1、BLEU-2、BLEU-3、BLEU-4和CIDEr这5个指标上的得分分别提升了4.1%、3.1%、3.3%、3.6%和25.1%。消融实验结果表明,FRCM使用的训练方法和关键组件能够有效提升模型的性能。实验通过3个案例分析,进一步证明FRCM生成的中文医学报告在准确性和信息丰富度上优于对比模型。**结论**　本文通过设计多模态大模型训练与推理策略,综合了语义微调和检索增强的优点,生成了更加详细且准确的中文医学报告。

关键词: 中文医学报告生成;多模态大模型(LMM);检索增强;语义微调;知识引导

Semantic fine-tuning and cross-modal retrieval-augmented Chinese medical report generation

Li Hengtai[1,2], Liu Hui[1,2*], Chen Gongguan[1,2], Yan Zishen[1,2], Sheng Yurui[3], Zhang Caiming[2,4]

1. School of Computing and Artificial Intelligence, Shandong University of Finance and Economics, Ji'nan　250014, China; 2. Shandong Key Laboratory of Lightweight Intelligent Computing and Visualization for Digital Economy (Shandong University of Finance and Economics), Ji'nan　250014, China; 3. Radiology Department of the First Affiliated Hospital of Shandong First Medical University, Ji'nan　250013, China; 4. School of Software, Shandong University, Ji'nan　250101, China

Abstract: **Objective** The task of generating medical reports involves producing accurate and comprehensive examination

收稿日期:2024-08-12;**修回日期**:2024-11-12;**预印本日期**:2024-11-19

*通信作者:刘慧　liuh_lh@sdufe.edu.cn

基金项目:国家自然科学基金项目(62072274, U22A2033);中央引导地方科技发展项目(YDZX2022009);山东省泰山学者特聘专家计划项目(tstp20221137);济南市人才发展专项资金项目(202333037)

Supported by: National Natural Science Foundation of China(62072274, U22A2033); Central Guidance on Local Science and Technology Development Project(YDZX2022009); Special Funds of Taishan Scholars Project of Shandong Province (tstp20221137)

results based on symptoms observed in medical images. This technology can ease the workload of radiologists, reduce diagnostic errors due to a lack of experience, and expedite clinical workflows. Medical report generation shares similarities with image captioning, but it presents two unique challenges: generating long texts and handling the imbalance in medical data distribution. Current approaches often train models specifically for medical report generation from scratch, relying on limited publicly available data. Due to their insufficient capability to fuse visual and textual features and generate rich, detailed information, these models frequently underperform. Large multimodal models (LMMs), which combine visual encoders and large language models (LLMs), are well-suited for image-based text generation tasks. They have the capability to recognize images and generate high-quality, knowledge-rich text, making them promising candidates for medical report generation. However, the application of LMMs in Chinese medical report generation remains in its early stages, particularly in terms of accurately understanding medical images and generating normatively correct medical reports. Moreover, these models often suffer from hallucination issues, where the generated responses appear logical but are factually incorrect or unfounded. This paper proposes a Chinese medical report generation model that integrates semantic fine-tuning and cross-modal retrieval-augmented (FRCM) to address these challenges. **Method** Building upon the LMM framework of LLaVA, this paper fine-tunes and adapts the visual encoder and LLM to the medical domain. A collaborative training strategy that incorporates general data and domain-specific data is proposed, and a novel cross-modal retrieval-augmented strategy is introduced during the inference phase. The paper also translates the largest dataset in the medical report generation domain, MIMIC-CXR, into Chinese and uses it as in-domain data for research on Chinese medical report generation. First, considering the unique characteristics of medical images and Chinese medical reports, the corresponding modules of LLaVA are replaced with a medical visual encoder that is trained on a large volume of medical images and a medical LLM optimized for Chinese language processing. This adaptation enhances the model's capability to effectively handle medical data. Second, a two-phase training strategy is employed, using general and domain-specific data. In the first phase of training, only the projection layer is trained. The domain-specific data facilitates medical image-text alignment, ensuring the model can accurately link medical images to their corresponding textual descriptions. In the second phase of training, the parameters of the projection layer are further updated, and a low-rank adaptation method is applied to fine-tune the LLM. The domain-specific data enhances the model's capacity to generate professional Chinese medical reports, while the general data improves the model's understanding of complex instructions. Throughout the entire training process, medical images are processed by the visual encoder to extract global and local feature vectors. These local feature vectors are then projected into visual embeddings that align with the dimensionality of the LLM's embedding space. Medical reports and instructions are tokenized into text embeddings by the tokenizer of the LLM. These text embeddings, along with the visual embeddings, are input into the LLM during training. Finally, a cross-modal retrieval-augmented strategy is proposed to further mitigate the hallucination problem in the model. This strategy incorporates a cross-modal similar report retrieval module. During inference, the global feature vectors produced by the visual encoder are layer-normalized and input into the report retrieval module, which performs a cross-modal retrieval from the image to relevant reports. The retrieved similar reports are then used as supplementary knowledge, providing additional context to the LLM. This approach helps reduce hallucinations, thereby improving the accuracy and robustness of the generated medical reports. **Result** On the Chinese MIMIC-CXR dataset, the FRCM model outperformed other Chinese medical report generation models, such as XrayGLM and XrayPULSE, achieving notable improvements in several evaluation metrics. Specifically, FRCM showed increases of 10.4%, 10.1%, 9.7%, 9.1%, 6.6%, 9.4%, and 38.4% in BLEU-1, BLEU-2, BLEU-3, BLEU-4, recall-oriented understudy for gisting evaluation-longest common subsequence (ROUGE-L), metric for evaluation of translation with explicit ORdering (METEOR), and consensus-based image description evaluation (CIDEr) scores, respectively. When compared to models fine-tuned on LLaVA and Qwen-VL, FRCM also achieved score improvements of 4.1%, 3.1%, 3.3%, 3.6%, and 25.1% in BLEU-1, BLEU-2, BLEU-3, BLEU-4, and CIDEr, respectively. Ablation experiments revealed the effectiveness of the proposed approach. Data ablation demonstrated that adding diverse general data during training enhances the model's capability to follow complex instructions, resulting in better utilization of additional knowledge and improving the quality of the generated medical reports. Module ablation further highlighted the importance of key components within FRCM, which substantially enhance its performance. Three case studies demonstrated that the Chinese

medical reports generated by FRCM were superior to those produced by other models in terms of accuracy and information richness. **Conclusion** This paper proposes FRCM, a model designed to generate Chinese medical reports from medical images. Unlike traditional medical report generation methods, FRCM leverages LMMs to effectively address the challenges associated with long text generation and the imbalance of medical data in the report generation task. While LMMs are typically pre-trained on extensive general datasets, they face limitations in recognizing medical images and generating specialized medical reports. The paper builds upon the LLaVA model framework, utilizing a medical visual encoder and a medical LLM, fine-tuning them semantically. This study introduces a similar report retrieval module to further mitigate the inherent hallucination problem of LMMs. This module supplies additional knowledge during the inference stage, aiding the model in generating more accurate reports. Experimental results show that FRCM performs satisfactorily in generating Chinese medical reports.

Key words: Chinese medical report generation; large multimodal model (LMM); retrieval enhancement; semantic fine-tuning; knowledge guidance

0 引 言

医学报告是指由经验丰富的医生结合放射学影像表现出具的全面检查结果，在疾病诊断和治疗中起着重要作用。然而，我国当前面临放射科医生数量不充足、医师培训项目分布不均匀以及各地域医疗水平不平衡等问题（Zhang等，2021），使临床医生面临无法得到放射学专业意见支持的风险，进而影响临床诊断的准确性和及时性（Petinaux等，2011）。在这样的背景下，结合计算机视觉和自然语言处理（natural language processing，NLP）技术的医学报告生成技术逐渐受到关注，有望减轻放射科医生的负担，减少由于经验不足而出现的漏诊和误诊，全面提高临床工作效率。

医学报告生成任务中需要解决的两个关键问题是长文本生成问题和数据分布不平衡问题。一方面，与图像字幕任务（Liu等，2022；杜海骏和刘学亮，2020）类似，医学报告生成任务也遵循“为图像生成对应文本”的基本流程。然而，与图像字幕要求生成简短的一句话的描述不同，医学报告生成任务需要生成多个句子组成的详细且连贯的长句描述。为了解决这个问题，Jing等人（2018）利用报告的组成性质，采用分层长短期记忆网络（long short-term memory，LSTM）来生成高级主题，然后根据主题并结合共同注意机制生成详细的长文本描述。Chen等人（2020）设计了关系记忆网络来记录生成过程中的关键信息，并应用记忆驱动的条件层归一化将记忆纳入解码器中来增强模型生成长文本的能力。邢素霞等人（2024）利用记忆驱动模块，解决医学影像数据中的长距离依赖问题。另一方面，医学数据中正常和异常病例数量分布不均衡，导致模型在学习过程中偏向于学习数量更多的正常样本，从而限制其检测异常病变的能力。Li等人（2018）提出一种带有模板检索和文本生成模块的混合模型，通过在生成过程中结合模板句子和自动生成句子的方式，以增强模型生成多样化医学报告的能力，同时提高对异常情况的描述精度。Harzig等人（2019）为了解决数据分布不平衡问题，在分层LSTM模型中引入双词LSTM，分别生成正常和异常的句子，从而提高生成报告的多样性。Liu等人（2021a）提出一种基于先验—后验知识的报告生成方法，通过数据驱动的方法缓解数据分布不平衡的问题。尽管先前的研究者们对这些问题已经做出很多努力，但是这两个难点依然是医学报告生成领域的重大挑战，特别是对于中文医学报告的生成，由于数据的稀缺和独特的语言特征，导致这两个难点更加显著。

预训练大语言模型（large language model，LLM）在医疗、编程等各领域的多种任务中展现出卓越能力（严昊 等，2023；Yao等，2024；Zhang等，2024）。大多数LLM基于Transformer（Vaswani等，2017）架构，在来源广泛的文本数据集上经过自回归语言建模的方式训练，具备强大的语言理解和生成能力，对处理医学报告生成等长文本生成的任务有着天然优势。同时，使用大量语料预训练的LLM拥有丰富的知识，不易受到不平衡数据的影响。多模态大模型（large multimodal model，LMM）代表了传统LLM在纯文本数据以外的模态中的扩展（Liu等，2024b；Bai等，2023），使用视觉编码器和LLM来解决涉及多种模态的任务，为医学报告生成任务提供了一个有希

望的解决方案。但LMM通常在大量通用领域数据上进行预训练，因此不能够很好地理解医学图像特征和生成专业的医学报告，并且其固有的幻觉问题，为医学报告生成任务增加了不可控的风险。

为解决上述问题，本文提出一种基于语义微调和跨模态检索增强的中文医学报告生成（semantic fine-tuning and cross-modal retrieval-augmented Chinese medical report generation，FRCM）模型。与一般的LMM相比，FRCM针对中文医学报告生成任务进行了特别优化。首先，在处理医学图像方面，利用预训练医学视觉编码器强大的特征提取能力提升病灶识别的准确性；其次，针对中文医学报告的特点，采用领域内强中文能力的LLM作为解码器进行报告生成；最后，为了缓解LMM的幻觉问题，在推理阶段，提出一种跨模态检索增强策略，对LMM生成医学报告过程提供额外的知识引导。通过使用通用数据与垂域数据协同训练，使模型能够更好地完成中文医学报告生成任务。

本文主要贡献如下：1）使用多模态大模型生成中文医学报告的研究工作还处于起步阶段，本文首次以中文文献形式阐述主要工作和实现方法，为该领域的研究提供了新的思路和参考。2）不同于目前开源的中文医学报告生成的多模态大模型，本文对视觉编码器和大语言模型进行领域适配，并通过语义微调提升其对医学图像内容的理解和专业医学报告的生成能力。3）为了缓解幻觉问题，设计了一种跨模态的相似报告检索模块，该模块通过从图像到报告的检索，在推理阶段提供额外的文本知识对模型进行引导，提高生成报告的准确性。4）与较先进的开源模型在多个指标上进行对比，实验表明本文方法能够取得更好的效果，展现了其在中文医学报告生成任务中的潜力和应用价值。

1 相关工作

1.1 医学领域的大语言模型和多模态大模型

自OpenAI发布ChatGPT（chat generative pre-trained Transformer）以来，LLM的发展迎来爆炸式增长，在各种自然语言任务中有着卓越表现，但其应用于生物医学领域时会遇到困难（Tian等，2024a）。通过使用医学领域的特定知识对LLM进行微调，可以显著提升其在生物医学应用中的表现（Wang等，2024）。Med-PaLM2（Singhal等，2023）是在医学数据集上对PaLM2进行微调得到的，是第一个通过美国医师执照考试的模型，在理解和回答复杂医学问题方面达到了人类医生水平。TourSynbio-7B（Shen等，2024）是一个专为蛋白质工程设计的LLM，无需依赖外部蛋白质编码器即可直接处理蛋白质序列，在蛋白质理解和工程任务中表现出卓越的性能。DoctorGLM（Xiong等，2023）、BenTaso（Wang等，2023a）、Zhongjing（Yang等，2024）、ChiMed-GPT（Tian等，2024b）和Medchatzh（Tan等，2024）等中文医学大模型也在各种医疗知识问答、诊断支持以及个性化治疗建议等任务中表现出不俗的能力。然而只能理解离散文本显然不能满足全面的医疗场景，面向多模态医疗问题的处理才是未来发展方向。Med-Flamingo（Moor等，2023）基于OpenFlamingo-9B，使用来自出版物和教科书的配对和交错医学图像文本数据继续进行预训练得到，视觉信息通过交叉注意力层交织到冻结的LLM中进行模态对齐。LLaVA-Med（Li等，2024）利用更简单的线性层将视觉表示从视觉编码器的输出映射到LLM的词嵌入空间，并在GPT-4生成的指令调整数据集和下游生物医学数据集上微调线性层和整个LLM。MLeVLM（multi-level visual language model）（Xu等，2024a）通过构建多级指令数据集使LLM具备多层次分析和逐步推理的能力，同时设计了一个多级特征对齐模块，有效地对齐从视觉到语义的不同级别的特征。尽管它们在医学视觉问答任务中有着不错表现，但是对于医学报告生成这样的垂域任务尚未展现出成熟能力。

1.2 医学报告生成

医学报告生成与图像字幕任务类似，都是生成与图像对应的文本，因此许多医学报告生成模型都是基于图像字幕任务的编码器—解码器框架（Vinyals等，2015）设计的。Zhang等人（2020）和Wang等人（2022a）使用预训练的卷积神经网络（convolutional neural network，CNN）作为编码器、双层LSTM作为解码器，并引入图卷积神经网络模块，以此生成长文本描述。Wang等人（2022a）还在其架构中引入知识图谱，通过向模型提供医学知识来协助生成过程。Wang等人（2021）在CNN-LSTM架构上提出一种自提升框架，该框架基于报告生成的主要任务和图像文本匹配的辅助任务相互协作来改进医学图像报告的生成。Liu等人（2021b）引入对比注意力机

制，通过对比输入图像和正常图像来捕获异常区域，以生成更精细的胸部X射线报告。Transformer的发展为医学报告生成领域创造了新的探索途径。Chen等人（2020）通过内存驱动的Transformer生成放射学报告，该Transformer采用关系内存来捕获整个生成过程中的关键信息。Huang等人（2023）开发了一种知识注入U-Transformer框架，专注于提取和提炼多层次信息和多重注入知识。Wang等人（2022b）将视觉编码器和解码器全部使用Transformer，共同实施更好的视觉文本对齐、多标签诊断分类和单词重要性加权，以促进医学报告的生成。在以上方法中，编码器和解码器同样重要。随着LLM的出现，由于其巨大的参数量，因此诞生了以LLM解码器为核心的医学报告生成方法。R2-GenGPT（Wang等，2023b）使用高效的视觉对齐模块将视觉特征投射到LLM之中，并提出浅对齐、增量对齐和深对齐3种训练方法来探究视觉编码器不同量级的参数对医学报告生成结果的影响。XrayGPT（Thawakar等，2023）使用简单的线性层将视觉编码器和经过医疗数据微调的LLM连接起来，并在经过预处理的交互式X光放射报告和图像上进行微调。Liu等人（2024a）设计了一种域内实例归纳模块使LLM适应医学报告生成任务，并通过一种从粗到细的解码方法使LLM生成更加精确的报告。InVERGe（intelligent visual encoder for bridging modalities in report eneration）（Deria等，2024）使用自监督学习训练视觉编码器提取图像的高层次语义特征，并引入一种跨模态查询融合层来增强图像与文本的对齐，从而帮助LLM生成高质量的医学报告。然而，这些模型都是基于英文医学数据进行研究，在中文应用环境中可能不适用。为了促进中文领域医学LMM的研究发展，澳门理工大学开发了XrayGLM，以VisualGLM-6B作为基座模型，并在翻译的中文版MIMIC-CXR和OpenI数据集上进行微调。虽然XrayGLM考虑了中文数据的特点，但在处理医学图像和医学文本方面仍存在不足。

本文提出的FRCM模型，同样使用翻译后的中文MIMIC-CXR数据集进行微调。然而，与XrayGLM不同的是，FRCM不仅深入考虑了中文数据的特性，还针对医学图像和医学报告的特点选择了合适的处理模块，这与局限在通用LMM原有框架中的XrayGLM有本质区别。

2 本文方法

本文提出一种基于语义微调和跨模态检索增强的中文医学报告生成模型FRCM，旨在对医学图像进行分析并生成中文医学报告，模型的架构如图1所示。为了方便叙述，在后文中均以垂域数据，即医学图文数据为例进行介绍。

2.1 模型架构

FRCM由3部分组成：医学视觉编码器、医学LLM以及相似报告检索模块。训练过程中，医学图像通过医学视觉编码器编码为全局特征和局部特征，然后图像的局部特征经过一个可训练的投影层，将其转化为与医学LLM嵌入层相同维度的图像嵌入。指令和医学报告经过分词器转化为文本嵌入。随后这些嵌入一起被送入医学LLM中进行训练。在推理过程中，医学图像经过相同的处理变为图像嵌入。此外，医学图像编码后的全局特征会进行归一化处理，随后送入相似报告检索模块进行从图像到报告的跨模态检索。检索到的相似报告经过分词器处理后，作为增强知识，与图像嵌入和指令嵌入一起送入医学LLM中，指导其生成最终的医学报告。

2.1.1 医学视觉编码器

因为X光图像无法直接作为LLM的输入，所以生成医学报告时，首先需要从医学图像中提取视觉特征。本文使用视觉Transformer（vision Transformer，ViT）（Dosovitskiy等，2021）架构的医学视觉编码器。首先，将输入图像$\boldsymbol{I} \in \mathbf{R}^{H \times W \times C}$分割成$N$个固定大小的非重叠图像块，然后将这些图像块展平为图像块序列$[\boldsymbol{I}_{\mathrm{p}}^{1}; \boldsymbol{I}_{\mathrm{p}}^{2}; \cdots; \boldsymbol{I}_{\mathrm{p}}^{N}]$，其中，每个图像块序列是一个一维的向量，用$\boldsymbol{I}_{\mathrm{p}}^{i} \in \mathbf{R}^{(P^2 \cdot C)}$表示，$C$是图像的通道数，$H$和$W$分别是图像的高和宽，$P$是图像块的高和宽。随后经过一个全连接层$\boldsymbol{E} \in \mathbf{R}^{(P^2 \cdot C) \times D}$将$\boldsymbol{I}_{\mathrm{p}}^{i}$映射到维度为$D$的向量空间，得到图像块嵌入$\boldsymbol{I}_{\mathrm{p}}^{i}\boldsymbol{E} \in \mathbf{R}^{D}$，同时在图像块嵌入首部加入一个可学习的图像分类嵌入$\boldsymbol{I}_{\mathrm{cls}}$。因为Transformer对输入序列的顺序不敏感，ViT模型为图像分类嵌入和每个图像块嵌入添加了位置编码$\boldsymbol{E}_{\mathrm{pos}} \in \mathbf{R}^{(N+1) \times D}$，以保留每个图像块的位置信息。最终得到Transformer编码器的输入$\boldsymbol{z}$，可表示为

$$\boldsymbol{z} = [\boldsymbol{I}_{\mathrm{cls}}; \boldsymbol{I}_{\mathrm{p}}^{1}\boldsymbol{E}; \boldsymbol{I}_{\mathrm{p}}^{2}\boldsymbol{E}; \cdots; \boldsymbol{I}_{\mathrm{p}}^{N}\boldsymbol{E}] + \boldsymbol{E}_{\mathrm{pos}} \tag{1}$$

Transformer编码器由多个堆叠的Transformer层

组成，每一层又由多头自注意力模块和前馈神经网络（feedforward neural network，FNN）两个子层组成。FNN 通常是一个两层的多层感知机（multi-layer perceptron，MLP）。在每个子层后都接有一个残差连接和层归一化。在编码过程中，每一层的输出作为下一层的输入，逐层提取特征，最后将 Transformer 编码器的隐藏层状态作为输出。隐藏层状态是指经过每一层 Transformer 编码器处理后的嵌入表示，这些嵌入捕捉了输入图像丰富的特征信息。此过程可表示为

$$\boldsymbol{e} = f_{\text{Transformer}}(\boldsymbol{z}) = [\boldsymbol{e}_{\text{cls}}; \boldsymbol{e}_1; \boldsymbol{e}_2; \cdots; \boldsymbol{e}_N] \quad (2)$$

式中，$\boldsymbol{e}_{\text{cls}}$ 是图像分类嵌入 $\boldsymbol{I}_{\text{cls}}$ 经过编码得到的特征，$\boldsymbol{e}_i (i = 1, \cdots, N)$ 是每个图像块嵌入 $\boldsymbol{I}_{\text{p}}^i (i = 1, \cdots, N)$ 经过编码得到的特征。本文参考 Zhang 等人（2022）和 LLaVA-Med（Li 等，2024）方法，将 $\boldsymbol{e}_{\text{cls}}$ 用于表示整个图像的全局特征，因为其聚合了来自所有图像块的全局信息，将 $\boldsymbol{e}_i (i = 1, \cdots, N)$ 用于表示整个图像的局部特征，即使它们在自注意力机制下相互进行信息交互，融合了一定的全局信息，但仍保留了各自对应图像块的特征细节。在下文中使用 $\boldsymbol{e}_{\text{others}}$ 来表示图像的局部特征，即 $\boldsymbol{e}_{\text{others}} = [\boldsymbol{e}_1; \boldsymbol{e}_2; \cdots; \boldsymbol{e}_N]$。

本文采用 BiomedCLIP（Zhang 等，2023）的视觉编码器用于提取医学图像的视觉特征。BiomedCLIP 是一种在 PMC-15M 数据集上训练的生物医学视觉语言模型，由于其视觉编码器已经具备充分的医学图像特征提取能力，因此在整个训练过程中，其模型权重保持冻结状态。

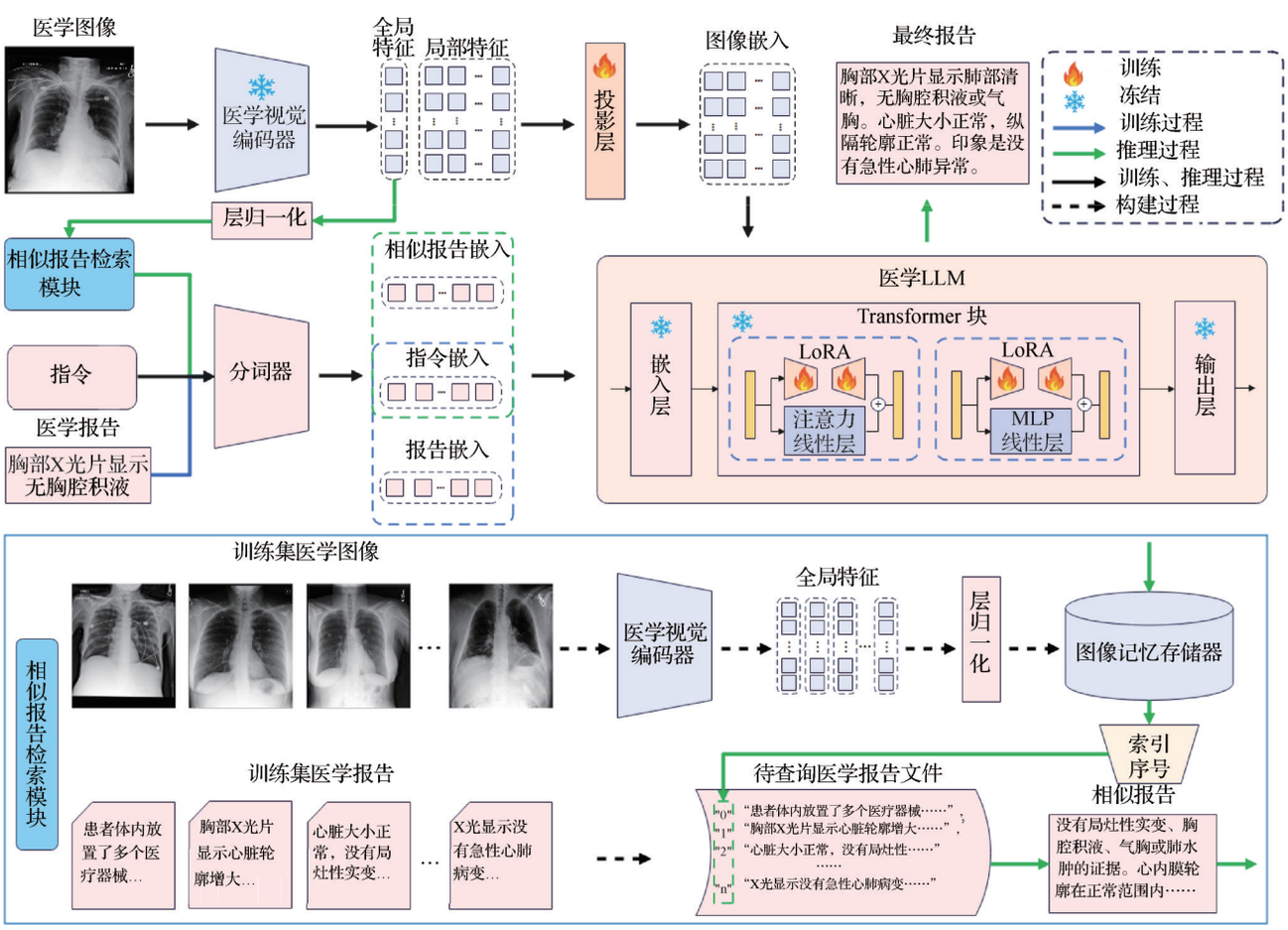


图 1　FRCM 模型架构图

Fig. 1　The architecture diagram of the FRCM model

2.1.2　医学大语言模型

在 FRCM 模型中，医学 LLM 作为生成报告的解码器。医学图像经过医学视觉编码器编码后，提取出图像的局部特征 $\boldsymbol{e}_{\text{others}}$，再通过一个可训练的投影层 $\boldsymbol{f}_{\text{p}}$，将 $\boldsymbol{e}_{\text{others}}$ 转换为与医学 LLM 的嵌入层空间维度相同的视觉嵌入，计算为

$$\boldsymbol{v}=\boldsymbol{f}_{\mathrm{p}}\left(\boldsymbol{e}_{\mathrm{others}}\right) \tag{3}$$

式中，$\boldsymbol{v}$是能够作为医学LLM输入的视觉嵌入，它包含了图像的更多细节，并用于指导医学报告的生成。同时，使用医学LLM的分词器将医学报告和指令变为文本嵌入，与视觉嵌入一起集成到训练的多模态模板中，并送入医学LLM中进行训练。

本文使用简单的两层MLP网络作为投影层，这样既保证模型的计算效率，又能有效地学习视觉嵌入到医学LLM词嵌入空间中的映射关系。选取Taiyi-LLM（Luo等，2024）作为医学LLM组件。该模型以Qwen-7B-base模型为基座，并使用生物医学数据进行了两个阶段的监督指令微调。因此，Taiyi-LLM不仅继承了Qwen-7B-base模型强大的中文理解和生成能力，还具备了相较于通用LLM更适合医学报告生成任务的医学知识。

2.1.3　相似报告检索模块

受Liu等人（2024c）启发，本文设计了一个相似报告检索模块，用于推理阶段对模型提供额外知识引导，从而进一步减轻模型的幻觉问题。输入的医学图像经过医学视觉编码器后，得到全局特征$\boldsymbol{e}_{\mathrm{cls}}$和局部特征$\boldsymbol{e}_{\mathrm{others}}$。在推理过程中，分离出$\boldsymbol{e}_{\mathrm{cls}}$并将其进行层归一化$\boldsymbol{f}_{\mathrm{LN}}$处理，具体为

$$\boldsymbol{y}=\boldsymbol{f}_{\mathrm{LN}}\left(\boldsymbol{e}_{\mathrm{cls}}\right) \tag{4}$$

式中，$\boldsymbol{y}$为归一化后得到的特征向量。随后将$\boldsymbol{y}$送入相似报告检索模块中，以检索出最相似的图像$\hat{\boldsymbol{I}}$。然后将$\hat{\boldsymbol{I}}$对应的医学报告$\hat{\boldsymbol{T}}$作为额外的文本知识，与医学图像的视觉信息一起输入医学LLM中，从而使医学LLM生成的报告更精确且与真实报告的语义更相似。相似报告检索模块的结构如图1下半部分所示。

Faiss（facebook AI similarity search）（Douze等，2024）是一个专为向量检索而设计的高效开源库。它为各种向量索引结构和相似度计算方法提供了强大的支持，在处理大规模向量场景下有着显著的优势。本文使用Faiss来构建相似报告检索模块中图像的记忆存储器。由于医学图像$\boldsymbol{I}$和医学报告$\boldsymbol{T}$以$\left(\boldsymbol{I}_i,\boldsymbol{T}_i\right)$的形式成对存在，因此首先使用与FRCM模型相同的医学视觉编码器对$\boldsymbol{I}_i$进行编码，获取其全局特征$\boldsymbol{e}_{\mathrm{cls}\,i}$，经过归一化处理后得到$\boldsymbol{y}_i$并存储到图像记忆存储器中，得到待查询医学图像向量$[\boldsymbol{y}_1;\ \boldsymbol{y}_2;\ \cdots;\ \boldsymbol{y}_M]$。每个待查询向量都有一个索引序号$i$来表示其在图像记忆存储器中的位置，并按照索引序号的顺序将报告$\boldsymbol{T}_i$存储到待查询的医学报告文件中。所有医学报告均不用编码成向量，直接以自然语言的形式存储，这样不仅简化了相似报告检索模块的构建过程，而且提高了图像到报告的跨模态检索效率。在进行医学图像检索过程中，为了确保在图像记忆存储器中查找到最相似的向量，本文使用IndexFlatL2结构对这些向量进行索引，该结构采用欧氏距离来进行向量检索，是一种基于精确匹配的索引方法。给定查询向量$\boldsymbol{y}$，计算为

$$\hat{\boldsymbol{y}}=\underset{i=0,\cdots,M}{\arg\min}\left\|\boldsymbol{y}-\boldsymbol{y}_i\right\|_2 \tag{5}$$

式中，$\hat{\boldsymbol{y}}$是检索到最相似的图像向量。通过使用IndexFlatL2索引，相似报告检索模块高效地计算查询向量和存储向量之间的欧氏距离，从而实现基于距离的相似度排序和最近邻搜索。随后，根据$\hat{\boldsymbol{y}}$的索引序号在待查询文件中寻找与之对应的报告$\hat{\boldsymbol{T}}$，并将其与用户指令一起送入分词器中，提取语义嵌入表示。这些相似报告的嵌入将在推理中进一步减轻模型的幻觉，引导模型生成更高质量的报告。

2.2　训练策略

为了在中文医学报告生成任务上提升模型的性能和适应性，模型基于构造的混合数据进行两个阶段的训练。下面的内容详细描述了两个阶段的训练过程。混合数据的构造细节在第2.3节介绍。

第1阶段的训练称为图文对齐训练，目的是将医学图像的视觉特征与医学LLM的嵌入层空间对齐。因为对未充分对齐的模型进行大规模视觉指令微调，会导致模型的效果更差（Chen等，2024），所以在此训练阶段使用第1阶段混合数据对用于视觉转换的投影层进行充分训练。此阶段使用的多模态模板为<*image*>[*caption*]，其中<*image*>代表的是医学图像，[*caption*]代表的是对医学图像的描述。在训练过程中，忽略<*image*>部分，仅使用[*caption*]部分来计算损失。此外，将医学视觉编码器和医学LLM的所有参数冻结，只更新投影层的参数，旨在优化视觉特征到医学LLM嵌入层空间的映射能力，从而提高模型对医学图像的理解能力。

第2阶段的训练称为指令微调，目的是让医学LLM能够学会遵循提供的指令，生成专业的中文医学报告。在此阶段，使用第2阶段混合数据对模型进行微调。与第1阶段训练方式不同，此阶段使用

多样化的指令来提示模型生成答案。具体而言，此阶段使用的多模态模板为[*system*]*USER*:<*image*>\ *n* [*instruction*]*ASSISTANT*:[*caption*]<*endoftext*>。其中[*system*]表示系统提示，赋予模型角色并定义模型的能力。*USER*是用户标签，<*image*>和[*instruction*]代表用户输入的医学图像和从10条候选指令中随机选择的一条指令，例如“给出这张胸部X光片的细节描述”。实验中使用的系统提示和候选指令如图2所示。*ASSISTANT*是助手标签，表示接下来是系统的响应部分，[*caption*]是对这幅医学图像的描述，<*endoftext*>是模型回答结束的停止标记。在这一阶段，损失计算仅涉及[*caption*]<*endoftext*>部分，使模型能够学习如何根据指令生成专业的医学报告，并学会在适当时停止生成。此阶段中同样保持医学视觉编码器参数和医学LLM参数冻结，继续训练模型中的投影层，并使用低秩适配（low-rank adaptation，LoRA）（Hu等，2022）微调方法对医学LLM进行高效参数微调。具体来说，本文微调了医学LLM中所有Transformer块的注意力模块和MLP模块中的线性层。

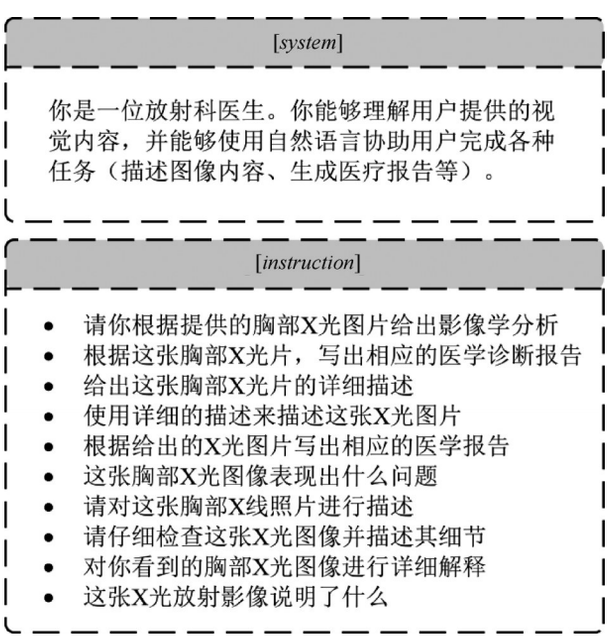


图2　系统提示和用户指令

Fig. 2　The system prompt and user instructions

在模型原有权重矩阵的基础上引入低秩矩阵分解，不仅有效降低了微调的参数数量，显著减少了计算资源的消耗，而且保证了微调后的模型能够有效适应下游任务。LoRA的原理如图3所示，假设模型某一层的权重矩阵为 $\boldsymbol{W}_0 \in \mathbf{R}^{d\times k}$，LoRA将要微调的增量参数矩阵 $\Delta\boldsymbol{W} \in \mathbf{R}^{d\times k}$ 低秩近似为两个参数量更小的矩阵 $\boldsymbol{A} \in \mathbf{R}^{d\times r}$ 和 $\boldsymbol{B} \in \mathbf{R}^{r\times k}$，参数更新过程可视为

$$\boldsymbol{W}_0 + \Delta\boldsymbol{W} = \boldsymbol{W}_0 + \boldsymbol{AB} \tag{6}$$

式中，d 和 k 是预训练权重的维度，r 是增量矩阵的秩，且 $r \ll \min(d,k)$。此时，微调的参数量从原来 $\Delta\boldsymbol{W}$ 的 $d\times k$ 变成 $d\times r + r\times k$，显著减少了需要更新的参数量。在微调时，保持 $\boldsymbol{W}_0$ 被冻结，用随机高斯分布对 $\boldsymbol{A}$ 初始化，用零矩阵对 $\boldsymbol{B}$ 初始化，这样能保证 $\Delta\boldsymbol{W} = \boldsymbol{AB}$ 在训练时逐渐学习新的知识，保证整个过程的稳定性。因此给定输入 $\boldsymbol{x} \in \mathbf{R}^d$，对于输出 $\boldsymbol{h} = \boldsymbol{xW}_0$，它的前向计算变为

$$\boldsymbol{h} = \boldsymbol{xW}_0 + \boldsymbol{x}\Delta\boldsymbol{W} = \boldsymbol{xW}_0 + \boldsymbol{xAB} \tag{7}$$

在推理时，可将 $\Delta\boldsymbol{W} = \boldsymbol{AB}$ 直接添加到原参数 $\boldsymbol{W}_0$ 上，不会引入额外的计算延迟。如果需要切换附加矩阵以适应不同的应用场景，也可以通过减去 $\boldsymbol{AB}$ 然后添加不同的 $\boldsymbol{A'B'}$ 的方式简单实现。

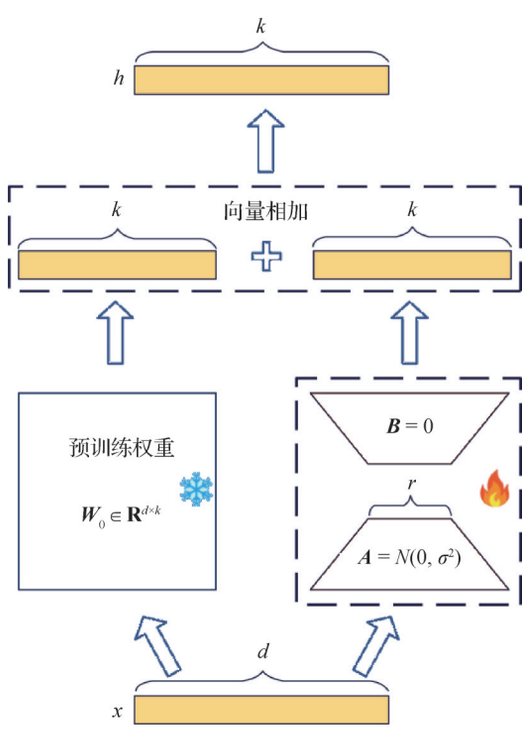


图3　LoRA微调策略

Fig. 3　The fine-tuning strategy for LoRA

通过这两个阶段的训练，使模型在中文医学报告生成任务中的表现得到显著提升。第1阶段的图文对齐训练确保了图像视觉特征与文本嵌入的有效匹配，而第2阶段的指令微调则使模型学会根据指令生成专业的医学报告。这种双阶段的训练策略为

FRCM理解医学图像的视觉信息并遵循指令生成医学报告提供了可行的解决方案。

2.3 数据集构建

为了使模型具备生成专业医学报告的能力并提升模型对复杂指令的理解能力，本文构造了混合数据用于模型训练。混合数据包括垂域数据和通用数据。垂域数据选择医学报告生成领域最大的公开数据集MIMIC-CXR（Johnson等，2019）作为研究对象。MIMIC-CXR数据集中的原始医学报告是非结构化的，并且包含一些如“__”等官方预处理留下的符号，这些问题会影响模型的精度。此外，报告的长度和完整度也会对训练效果产生影响。先前的研究（Longpre等，2024）已证明，训练数据的质量对下游任务的性能至关重要，低质量的数据会对LLM的表现产生不利影响。Thawakar等人（2023）利用GPT-3.5模型将MIMIC-CXR非结构化的原始医学报告转化为更符合人类语言习惯的高质量交互式报告，以此来训练XrayGPT，使其具有优秀的医学报告能力生成。腾讯翻译在国际翻译大赛WMT2018的中英翻译中取得世界第1名，并在国内率先使用神经网络翻译引擎，全面部署Transformer与循环神经网络算法。其翻译系统拥有亿级大数据资源支持，融合领域自适应技术，可满足多领域、多场景翻译需求。因此，本文选择使用腾讯翻译将用于训练XrayGPT的高质量医学报告从英文翻译成中文，最终获得中文MIMIC-CXR数据集，如图4所示。图4(a)展示了一个正常案例；图4(b)展示了常见肺部疾病案例；图4(c)展示了带有外置器械的案例；图4(d)展示了骨骼问题的案例。在数据划分方面，随机选取了2 000份数据作为测试集，其余数据作为训练集。通过这种方法，开展了LMM在中文医学报告生成领域的研究。

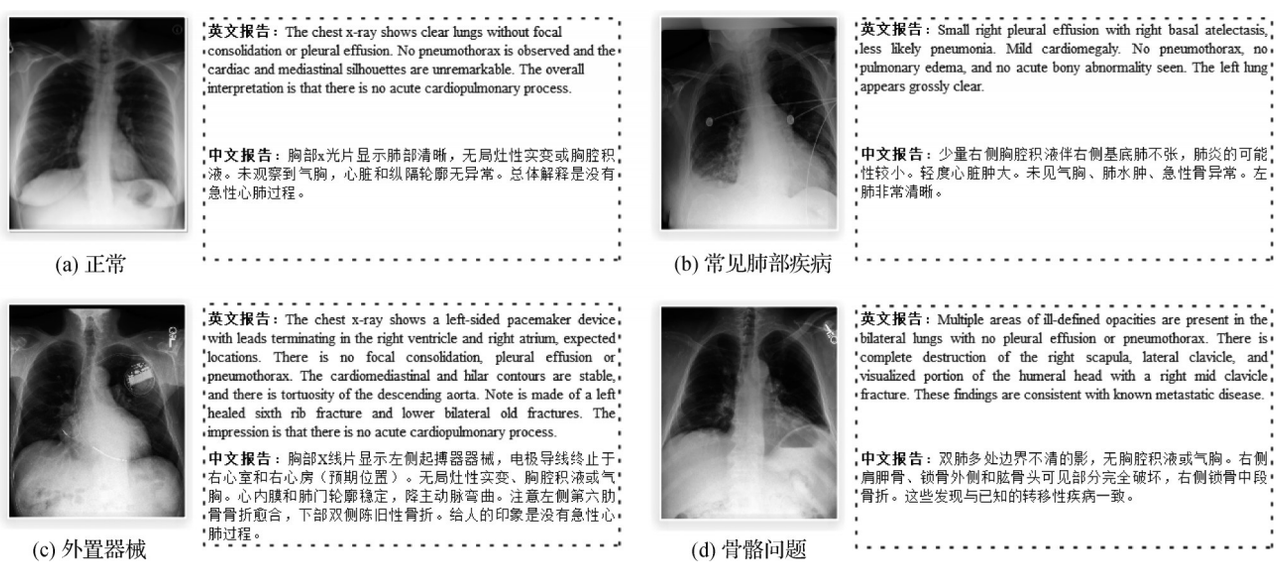


图4 中文MIMIC-CXR数据集展示

Fig. 4 Chinese version of the MIMIC-CXR dataset presentation

((a)normal;(b)common lung diseases;(c)external instrument;(d)bone problem)

通用数据使用整合了来自101个开源数据集的191个视觉问答（visual question answering，VQA）任务的Vision-FLAN（Xu等，2024b）作为研究对象。在Vision-FLAN数据集中，每个任务都附有专家编写的指令。这些多样化的图像和任务指令对提升LMM的泛化能力及对复杂指令的理解能力至关重要。得益于ALLaVA（Chen等，2024）项目的贡献，本文对其提供的中文版Vision-FLAN数据集进行了进一步预处理，去除了质量较低和不利于模型训练的数据，最终得到VFLAN-Caption和VFLAN-Instruct两部分数据。VFLAN-Caption部分是图像字幕数据集，用于第1阶段训练投影层，提升其图文对齐的泛化能力；VFLAN-Instruct部分是多样化的VQA数据集，用于第2阶段微调医学LLM，提升其对复杂指令的理解能力。添加通用数据所带来的模型能力提升将在消融实验中详细介绍。实验中使用的数据集具体情况

如表1所示。

表1　实验数据集

Table 1　Experimental dataset

阶段	数据集	数量	语言
第1阶段	中文MIMIC-CXR	212 539	中文
	VFLAN-Caption	170 872	中文
第2阶段	中文MIMIC-CXR	212 539	中文
	VFLAN-Instruct	162 730	中文

3　实验分析

3.1　实验设置

本文所有实验均在两块NVIDIA 4090D GPU上进行，每块GPU有24 G显存。编程语言选择PyThon 3. 10，并使用Pytorch深度学习框架。在训练中，图像的尺寸设置为224 × 224像素，模型的最大输入长度为2 048。数据处理过程中，将每个批次中的序列填充至相同长度，以便于张量计算，同时采用梯度累积策略进行参数更新。FRCM模型具体的训练参数如表2所示。

表2　训练参数

Table 2　Training parameters

序号	训练参数	第1阶段取值	第2阶段取值
1	设备批处理大小	4	2
2	梯度累积步数	32	32
3	全局批次大小	256	128
4	优化器	AdamW	AdamW
5	权重衰减	0	0
6	学习策略	余弦退火	余弦退火
7	预热比例	0.03	0.03
8	LoRA的秩	–	64
9	LoRA的缩放因子	–	128
10	学习率	0.001	0.000 2
11	精度	BF16	BF16

注：“–”表示无对应数据。

3.2　评价指标

本文采用广泛应用于NLP领域的评价指标来评估模型生成的医学报告的质量，包括双语评估替代指标BLEU-n（Papineni等，2002）、基于最长公共子序列的召回率指标ROUGE-L（Lin，2004）、显式顺序翻译评价指标METEOR（Banerjee和Lavie，2005）和基于共识的图像描述评估指标CIDEr（Vedantam等，2015）。其中，BLEU-n通过计算生成文本与参考文本中的n元词组的重叠概率，同时考虑长度惩罚因子来衡量生成的文本与参考文本之间的相似度。ROUGE-L基于生成文本和参考文本之间的最长公共子序列来判断两者之间的相似性。METEOR不仅考虑生成文本和参考文本之间的精确匹配，同时还考虑词干匹配和同义词匹配。CIDEr是一种专门用于图像描述生成任务的评价指标，该指标特别考虑了词频—逆文档频率的加权，以提高评价的准确性和一致性。

3.3　消融实验

为了验证本文方法的有效性，在本节进行数据消融和模块消融，实验结果如表3和表4所示。

3.3.1　数据消融

本节探讨了在添加与不添加相似报告检索模块两种情况下，使用不同数据训练模型的效果，具体结果见表3。第1、4行展示了仅使用垂域数据（中文MIMIC-CXR）在图文对齐训练和指令微调两个阶段训练模型的效果；第3、6行是两个阶段全部使用混合数据训练模型的效果。以上结果均为模型在两个阶段各经过一轮训练得出。为了最大限度地消除由于训练数据规模不一致带来的影响，表中增加了使用垂域数据在两个阶段分别进行2轮训练的结果（第2、5行）作为对比。

结果表明，在未添加相似报告检索模块时，仅使用垂域数据训练2轮的模型效果与使用混合数据训练1轮的模型效果几乎一致，并在所有指标上均优于仅使用垂域数据训练1轮的模型。另外，对3个模型的输出结果进行分析，它们都能够描述医学图像的基本情况，但使用垂域数据训练1轮的模型主要关注当前的检查结果，没有对病情进行深入分析。其他两个模型对医学图像都进行了较为细致的分析，与真实报告语义相似性更高。

添加相似报告检索模块后，仅使用垂域数据训练的模型在各指标上都远低于使用混合数据训练的模型。特别是仅使用垂域数据训练2轮的模型效果更差。这种现象的原因在于增加相似报告检索模块后，输入模型的指令变得更加复杂，包含了更多的文

本信息。模型不仅要依赖医学图像提供的视觉信息和自身学习到的知识生成医学报告，还需使用提供的相似报告作为额外的引导知识，进行辅助生成。在模型训练过程中，通用数据中的多样化VQA数据提升了模型对于复杂指令的理解能力，这有助于模型在生成医学报告时更有效地结合医学图像提供的信息和相似报告提供的额外知识引导，从而做出更优的选择和组合。

3.3.2 模块消融

为了证明FRCM模型中各个模块的重要性，本节进行了模块消融实验。表4展示了本文的关键组件在逐步集成到基线模型LLaVA（Liu等，2024b）框架中的表现。第1行为经过微调的基线模型的结果；第2、3、5行分别展示了单独替换视觉编码器、替换LLM以及加入相似报告检索模块的结果；第4行为同时替换视觉编码器和LLM，但未添加相似报告检索模块的结果，称之为中间模型。第6行是本文完整FRCM模型的结果。实验结果表明，相较于基线模型，替换视觉编码器和LLM后，模型各项指标均有显著提升。由此可以说明，针对中文医学报告生成任务的特点选择合适的处理模块，引入对下游任务特定领域的适应，能够有效提升模型的性能。具体而言，医学视觉编码器能够提供更加细粒度的视觉信息，医学LLM则能通过这些信息更好地理解医学图像中的细微病灶，结合自身具有的专业医学知识，从而生成更准确的医学报告。

表3 不同训练数据的消融实验

Table 3 Ablation study of different training data

训练数据	训练轮次	相似报告检索模块	指标						
			BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr
垂域数据	1	×	0.349	0.213	0.144	0.102	0.308	0.317	0.192
垂域数据	2	×	0.358	0.217	0.146	0.103	0.307	**0.320**	0.222
混合数据	1	×	0.353	0.215	0.146	0.104	**0.308**	**0.320**	0.223
垂域数据	1	√	0.303	0.190	0.139	0.110	0.272	0.264	0.341
垂域数据	2	√	0.292	0.179	0.128	0.099	0.264	0.257	0.275
混合数据	1	√	**0.375**	**0.234**	**0.171**	**0.135**	0.293	0.304	**0.443**

注：×表示未添加，√表示添加，加粗字体表示各列最优结果。

表4 模型组件的消融实验

Table 4 Ablation study of model components

视觉编码器	大语言模型	相似报告检索模块	指标						
			BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr
×	×	×	0.332	0.201	0.136	0.097	0.302	0.306	0.192
√	×	×	0.352	0.215	0.146	0.103	0.307	0.318	0.214
×	√	×	0.350	0.212	0.143	0.102	0.305	0.315	0.189
√	√	×	0.353	0.215	0.146	0.104	**0.308**	**0.320**	0.223
×	×	√	0.357	0.224	0.164	0.130	0.288	0.294	0.420
√	√	√	**0.375**	**0.234**	**0.171**	**0.135**	0.293	0.304	**0.443**

注：×表示未替换或添加，√表示替换或添加，加粗字体表示各列最优结果。

从第5、6行的结果可以看出，额外的知识引导在缩小生成报告与真实报告之间的差距方面起到了关键作用。基线模型添加相似报告检索模块后，在BLEU-1、BLEU-2、BLEU-3、BLEU-4和CIDEr等5个指标上分别提升2.5%、2.3%、2.8%、3.3%和22.8%。这种情况说明本文设计的跨模态相似报告检索模块使模型在使用专业术语和短语方面的能力进一步提升，同时也降低了模型的幻觉并增强了生

成报告与真实报告之间的语义相似性。ROUGE-L和METEOR指标分别下降了1.4%和1.2%，这种情况则可能源于相似报告在结构上与真实报告的差异。尽管相似报告的借鉴提高了词汇一致性，但限制了词汇多样性和同义词的使用，并导致生成的医学报告在某些情况下的顺序和布局与真实报告存在差异。将3个组件全部添加的FRCM模型在中文医学报告生成任务中达到了一个最好的平衡状态。

3.4 对比实验

为了证明FRCM模型在中文医学报告生成任务中的强大性能，进行对比实验。选取XrayGLM（https://github.com/WangRongsheng/XrayGLM）和XrayPULSE（https://github.com/openmedlab/XrayPULSE）这两个可以生成中文医学报告的LMM作为比较对象，它们都已经在翻译后的中文MIMIC-CXR数据集上进行了微调。其中在与XrayGLM的比较中，使用的模型权重是此项目开源的checkpoints-XrayGLM-3000。同时，本文还在LLaVA和Qwen-VL这两个主流LMM上进行了微调，使用的微调参数均为官方提供的默认参数。比较结果如表5所示。

结果显示，在BLEU指标（BLEU-1、BLEU-2、BLEU-3、BLEU-4）上，FRCM模型的得分分别为0.375、0.234、0.171和0.135，比其他模型最好结果分别提高了4.1%、3.1%、3.3%和3.6%，这表明FRCM模型在生成的报告与真实报告的词汇和短语匹配度上更高，生成的医学报告更加专业。在CIDEr指标上，FRCM模型的得分高达0.443，显著优于所有对比模型，这一指标上的卓越表现，表明其生成的报告内容与真实报告的内容高度一致。在ROUGE-L和METEOR指标上，FRCM的表现略低于微调的Qwen-VL和LLaVA，其原因如3.3.2节所述，相似报告与真实报告的结构差异影响了其最长公共子序列的生成。即便如此，在其他指标上的提升依然使FRCM表现出强劲的竞争力。

表5 不同模型的实验结果对比

Table 5 Comparison of experimental results from different models

模型	指标						
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR	CIDEr
XrayGLM	0.172	0.047	0.019	0.010	0.167	0.121	0.019
XrayPULSE	0.271	0.133	0.074	0.044	0.227	0.210	0.059
LLaVA（微调）	0.332	0.201	0.136	0.097	0.302	0.306	0.192
Qwen-VL（微调）	0.334	0.203	0.138	0.099	**0.303**	**0.310**	0.189
FRCM（本文）	**0.375**	**0.234**	**0.171**	**0.135**	0.293	0.304	**0.443**

注：加粗字体表示各列最优结果。

此外，对不同模型生成的医学报告长度进行了对比分析，结果如图5所示。图5中，横轴表示医学报告的字数区间，纵轴表示每个区间内报告的数量。可以观察到，XrayGLM生成的短报告数量最多，这可能是导致其在各项评价指标上表现相对较差的原因之一。Qwen-VL、LLaVA和XrayPULSE生成的报告长度主要集中在50～69字的区间，说明这些模型能够描述医学图像中的基本信息，但在细致刻画疾病特征和复杂病例分析方面存在局限性，生成的内容相对简洁。FRCM生成的报告长度分布与真实报告最为接近，表明该模型在生成医学报告时能较为全面地覆盖医学图像的关键信息。

3.5 案例分析

为了更好地展示本文模型在生成中文医学报告方面的有效性，在测试集中选取3幅医学图像，并使用基线模型、中间模型和FRCM模型进行报告生成，结果如图6所示，其中生成的报告中关键部分的正确描述用绿色背景标注，不正确描述用红色背景标注。

在第1个案例中，基线模型和中间模型只能识别出正常的肺部情况，对“主动脉结节钙化”和“沿脊柱可见退行性变化”等特征不明显的病变缺乏识别能力。而添加相似报告检索模块的FRCM模型，能够利用额外的知识引导，生成更准确的医学报告。另外，没有对视觉编码器和LLM进行医学领域适配

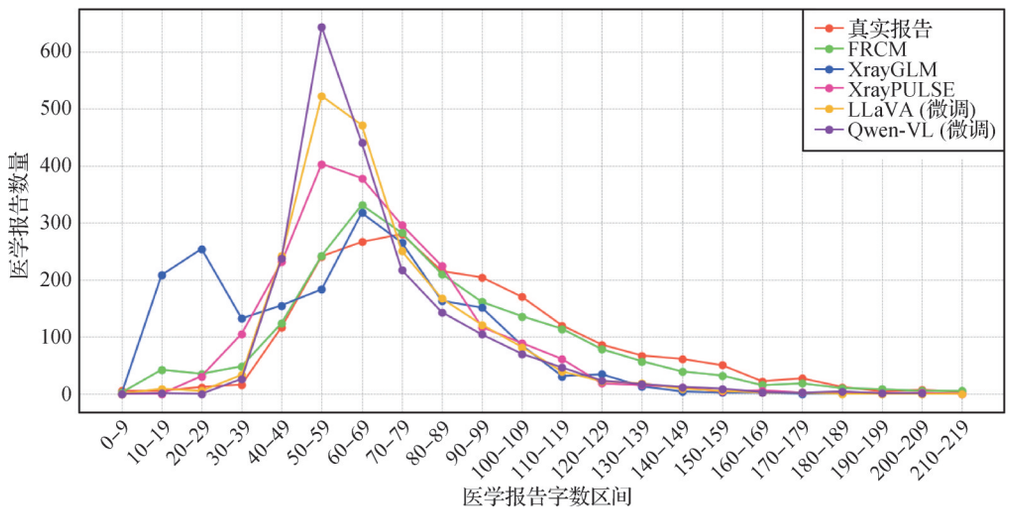


图5　不同模型生成医学报告长度

Fig. 5　Medical report lengths generated by different models

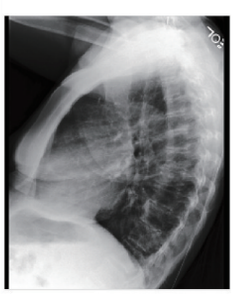
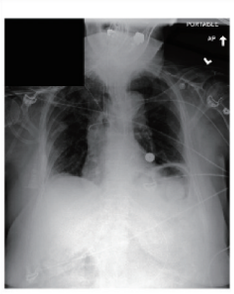
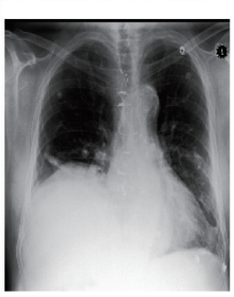
医学图像	真实报告	基线模型生成报告	中间模型生成报告	FRCM生成报告
	胸部x光片显示没有急性心肺病变的迹象。未见局灶性实变、胸腔积液或气胸。心脏和纵膈轮廓稳定，主动脉结节钙化。沿脊柱可见轻度退行性变化。	胸部x光片显示肺部清晰，无胸腔积液、气胸或局灶性空变。心脏和纵隔轮廓稳定，不明显。总体印象是没有急性心肺过程。	胸部x光显示无急性心肺异常。心脏大小正常，纵隔和肺门轮廓正常。肺血管系统正常，肺部清晰，无胸腔积液或气胸。未发现急性骨质异常。	胸部X线图像显示没有局灶性实变、胸腔积液或气胸的迹象。心脏和纵膈轮廓稳定，主动脉结节钙化。沿脊柱可见退行性变化。
	胸部X光片显示气管导管位于隆突上方约2 cm处，但这可能是由于患者颈部屈曲和脊柱前凸定位所致。左侧的中心静脉导管终止于上下腔静脉，可以看到一个肠管突出于胃。心脏轮廓外观稳定，肺部清晰，无局灶性实变、积液或气胸迹象。	胸部x光片显示胸腔积液伴双基底动脉肺不张。肺容积较低，但心脏和纵隔轮廓正常。气管内导管终止于隆突上方约5.2cm处。	胸部X线片显示气管导管终止于隆突上方5.6cm处，NG管终止于胃内。肺容积低，双基底动脉肺不张，心脏大小正常。没有气胸或胸腔积液的迹象。给人的印象是气管插管的位置合适。	胸部X光片显示气管导管终止于隆突上方约4.8厘米处，肠导管盘绕在膈水平以下胃的预期位置。心脏和纵隔轮廓稳定，未见胸腔积液或气胸，也未见明显肺水肿。胸部X线图像上未观察到明确的局灶性实变。
	患者既往接受过正中胸骨切开术和冠状动脉搭桥术。心脏轮廓尺寸轻度增大。此外，还存在一个小到中等大小的食管裂孔疝，以及双侧钙化胸膜斑块。主动脉钙化扭曲。右半膈慢性抬高也存在。在两个肺基底处均检测到线状影，可能是由于肺不张所致。没有胸腔积液或气胸的迹象。胸部x光片显示双基底肺不张，双侧胸膜斑块钙化，小到中等大小的食管裂孔疝。	胸部X线片显示胸骨正中切开钢丝完整，肺部清晰，无胸腔积液或气胸迹象。心脏大小正常，纵隔轮廓正常。根据这些发现，印象是没有急性胸内突起。	胸部X线图像显示胸骨切开钢丝和纵隔夫完整，以及人工心脏瓣膜和左胸壁双腔中心静脉导管头端终止于右心房。右半膈抬高伴肺不张，但未见局灶性实变或胸腔积液。心脏大小正常，纵隔轮廓无异常。印象是没有急性胸内病变。	胸部X光片显示右半膈持续抬高，伴少量胸腔积液和双基底动脉肺不张。注意到正中胸骨切开术导丝和CABG夹。多个胸膜斑块存在，但没有局灶性实变或气胸。心脏、肺门和纵隔轮廓在正常范围内。印象是右半膈持续抬高伴双基底动脉肺不张。

图6　FRCM生成报告性能案例对比

Fig. 6　Comparison of report generation performance cases for FRCM

的基线模型将“局灶性”和“空变”搭配在一起，这种不准确的医学术语会造成一定的误导。在第2个案例中，3个模型都能识别出“气管导管终止于隆突上方”，但基线模型由于缺乏细粒度的医学视觉指导，

因此未能生成更加全面的医学报告，并且幻觉问题较为严重，错误地报告了“胸腔积液”、“双基底动脉肺不张”和“肺容积较低”这3个不真实的内容。中间模型虽然生成了较为全面的报告，但同样存在着较为严重的幻觉问题，生成了“肺容积低”和“双基底动脉肺不张”这样真实报告中不存在的内容。相比之下，FRCM模型生成的医学报告与真实报告内容基本相符，语义更加丰富，但缺乏准确的空间位置感知能力，不能准确生成导管所在的位置。在第3个案例中，3个模型均能发现患者接受过正中胸骨切开术，但均未正确判断心脏的状态。基线模型没有识别出患者的其他任何症状，错误输出患者“肺部清晰”和“心脏大小正常”。中间模型识别出了“右半隔抬高”和“肺不张”症状，虽然其也注意到心脏进行过手术，但是将冠状动脉搭桥手术的移植物错误识别成“人工心脏瓣膜”。FRCM模型除了能够发现前两个模型已识别出的症状，还发现患者存在“胸膜斑块”，同时也正确识别了患者进行过“冠状动脉搭桥手术（coronary artery bypass grafting，CABG）”。但是FRCM模型误判患者存在“少量胸腔积液”。

经过3个案例的展示，可以体现本文所提出的FRCM模型在生成中文医学报告任务中的有效性，不仅能够全面描述医学图像中的常规特征，还能够发现基线模型和中间模型难以识别的症状。相较于基线模型，中间模型能够识别出更多的病变情况，且在报告的用语方面也更加专业，这得益于视觉编码器和LLM在领域适配下进行的语义微调；相较于中间模型，FRCM不仅发现了更多的病变情况，还进一步缓解了幻觉问题，这得益于相似报告检索模块在模型推理阶段提供的额外知识引导。然而，相似报告检索模块受限于其存储的医学图像以及报告的种类和数量，另外，视觉编码器能够处理的图像分辨率大小和LLM的参数规模也对生成报告的质量产生一定影响。因此，对于如第3个案例中情况较为复杂的医学图像，FRCM仍然存在一定的误判和漏判情况。

4 结 论

本文提出一种基于语义微调和跨模态检索增强的中文医学报告生成模型FRCM，用于医学图像分析并生成中文医学报告。近年来，医学报告生成任务逐渐成为研究热点，但大多数研究集中在英文领域。为此，本文将当前领域中最大的MIMIC-CXR数据集翻译成中文，并针对中文医学报告生成展开研究。针对医学报告生成任务中的长文本生成和医学数据分布不平衡这两个难点，采用了具有图像识别能力、强大文本生成能力和丰富知识的LMM技术。为了克服通用LMM在理解医学图像内容和生成专业中文医学报告方面的不足，本文基于LLaVA模型的框架，采用医学视觉编码器和具有强中文处理能力的医学LLM，并对其进行了语义微调。为进一步缓解LMM的幻觉问题，本文提出一种跨模态检索增强策略，通过设计一种跨模态相似报告检索模块，在推理阶段提供相似医学图像的医学报告作为额外的引导知识。另外，还对模型的指令理解能力进行研究，在训练数据中添加多样化的通用数据，采用通用数据与垂域数据协同训练的策略，能够使模型理解更加复杂的指令，从而利用额外的引导知识，使生成的中文医学报告更加准确。通过与两个开源中文医学LMM和两个微调的通用LMM对比，FRCM在BLEU、CIDEr等指标上取得了明显提升，充分体现了FRCM在此任务中的优势。案例分析中与基线模型进行定性对比，生成的报告在医学术语、症状识别等方面得到了提升，证明了FRCM在生成中文医学报告任务中的潜力。

当然，本文通过翻译得到的数据可能不足以表达报告的最精确含义，但是目前在国内医院中收集的医学数据数量不足以支持实验，因此仍需依赖翻译后的数据来进行初步研究。在未来的研究中，将继续收集中文医学图像文本对数据，并且将任务扩展到CT（computer tomography）、MRI（magnetic resonance imaging）等模态的医学图像中，通过不断地改进和创新，力求在中文领域生成更准确、更全面的医学报告。

参考文献（References）

Bai J Z, Bai S, Yang S S, Wang S J, Tan S N, Wang P, Lin J Y, Zhou C and Zhou J R. 2023. Qwen-VL: a versatile vision-language model for understanding, localization, text reading, and beyond [EB/OL]. [2024-07-31]. https://arxiv.org/pdf/2308.12966.pdf

Banerjee S and Lavie A. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evalua-

tion Measures for Machine Translation and/or Summarization. Ann Arbor, USA: Association for Computational Linguistics: 65-72

Chen G H, Chen S N, Zhang R F, Chen J Y, Wu X B, Zhang Z Y, Chen Z H, Li J Q, Wan X and Wang B Y. 2024. ALLaVA: harnessing GPT4V-synthesized data for lite vision-language models [EB/OL]. [2024-07-31]. https://arxiv.org/pdf/2402.11684.pdf

Chen Z H, Song Y, Chang T H and Wan X. 2020. Generating radiology reports via memory-driven Transformer//Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [s. l.]: Association for Computational Linguistics: 1439-1449 [DOI: 10.18653/v1/2020.emnlp-main.112]

Deria A, Kumar K, Chakraborty S, Mahapatra D and Roy S. 2024. InVERGe: intelligent visual encoder for bridging modalities in report generation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, USA: IEEE: 2028-2038 [DOI: 10.1109/CVPRW63382.2024.00208]

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J and Houlsby N. 2021. An image is worth 16 × 16 words: Transformers for image recognition at scale//Proceedings of the 9th International Conference on Learning Representations. [s. l.]: ICLR: 1-22

Douze M, Guzhva A, Deng C Q, Johnson J, Szilvasy G, Mazaré P, Lomeli M, Hosseini L and Jégou H. 2024. The faiss library [EB/OL]. [2024-07-31]. https://arxiv.org/pdf/2401.08281.pdf

Du H J and Liu X L. 2020. Image description generation method based on inhibitor learning. Journal of Image and Graphics, 25(2): 333-342 (杜海骏, 刘学亮. 2020. 融合约束学习的图像字幕生成方法. 中国图象图形学报, 25(2): 333-342) [DOI: 10.11834/jig.190222]

Harzig P, Chen Y Y, Chen F and Lienhart R. 2019. Addressing data bias problems for chest x-ray image report generation//Proceedings of the 30th British Machine Vision Conference 2019. Cardiff, UK: BMVC

Hu E J, Shen Y L, Wallis P, Allen-Zhu Z, Li Y Z, Wang S A, Wang L and Chen W Z. 2022. LoRA: low-rank adaptation of large language models//Proceedings of the 10th International Conference on Learning Representations. [s.l.]: ICLR

Huang Z Z, Zhang X F and Zhang S T. 2023. KiUT: knowledge-injected U-Transformer for radiology report generation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE: 19809-19818 [DOI: 10.1109/CVPR52729.2023.01897]

Jing B Y, Xie P T and Xing E. 2018. On the automatic generation of medical imaging reports//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics: 2577-2586 [DOI: 10.18653/v1/P18-1240]

Johnson A E W, Pollard T J, Berkowitz S J, Greenbaum N R, Lungren M P, Deng C Y, Mark R G and Horng S. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data, 6(1): #317 [DOI: 10.1038/s41597-019-0322-0]

Li C Y, Liang X D, Hu Z T and Xing E P. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal, Canada: Curran Associates Inc.: 1537-1547

Li C Y, Wong C, Zhang S, Usuyama N, Liu H T, Yang J W, Naumann T, Poon H and Gao J F. 2024. LLaVA-med: training a large language-and-vision assistant for biomedicine in one day//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #1240

Lin C Y. 2004. ROUGE: a package for automatic evaluation of summaries//Proceedings of Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics: 74-81

Liu C, Tian Y H, Chen W D, Song Y and Zhang Y D. 2024a. Bootstrapping large language models for radiology report generation//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 18635-18643 [DOI: 10.1609/aaai.v38i17.29826]

Liu F L, Wu X, Ge S, Fan W and Zou Y X. 2021a. Exploring and distilling posterior and prior knowledge for radiology report generation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 13748-13757 [DOI: 10.1109/CVPR46437.2021.01354]

Liu F L, Yin C C, Wu X, Ge S, Zhang P and Sun X. 2021b. Contrastive attention for automatic chest x-ray report generation//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. [s.l.]: Association for Computational Linguistics: 269-280 [DOI: 10.18653/v1/2021.findings-acl.23]

Liu H T, Li C Y, Wu Q Y and Lee Y J. 2024b. Visual instruction tuning//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #1516

Liu M F, Hu H J, Li L J, Yu Y and Guan W L. 2022. Chinese image caption generation via visual attention and topic modeling. IEEE Transactions on Cybernetics, 52(2): 1247-1257 [DOI: 10.1109/TCYB.2020.2997034]

Liu Z Y, Sun Z Y, Zang Y H, Li W, Zhang P, Dong X Y, Xiong Y J, Lin D H and Wang J Q. 2024c. RAR: retrieving and ranking augmented MLLMs for visual recognition [EB/OL]. [2024-07-31]. https://arxiv.org/pdf/2403.13805.pdf

Longpre S, Yauney G, Reif E, Lee K, Roberts A, Zoph B, Zhou D, Wei J, Robinson K, Mimno D and Ippolito D. 2024. A pretrainer's guide to training data: measuring the effects of data age, domain

coverage, quality, and toxicity//Proceedings of 2024 Conference of the North American Chapter of the Association for Computational Linguistics. Mexico City, Mexico: Association for Computational Linguistics: 3245-3276 [DOI: 10.18653/v1/2024.naacl-long.179]

Luo L, Ning J Z, Zhao Y W, Wang Z J, Ding Z Y, Chen P, Fu W R, Han Q Y, Xu G T, Qiu Y Z, Pan D H, Li J R, Li H, Feng W D, Tu S B, Liu Y Q, Yang Z H, Wang J, Sun Y Y and Lin H F. 2024. Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks. Journal of the American Medical Informatics Association, 31(9): 1865-1874 [DOI: 10.1093/jamia/ocae037]

Moor M, Huang Q, Wu S, Yasunaga M, Dalmia Y, Leskovec J, Zakka C, Reis E P and Rajpurkar P. 2023. Med-flamingo: a multimodal medical few-shot learner//Proceedings of the 3rd Machine Learning for Health Symposium. New Orleans, USA: PMLR: 353-367

Papineni K, Roukos S, Ward T and Zhu W J. 2002. BLEU: a method for automatic evaluation of machine translation//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, USA: Association for Computational Linguistics: 311-318 [DOI: 10.3115/1073083.1073135]

Petinaux B, Bhat R, Boniface K and Aristizabal J. 2011. Accuracy of radiographic readings in the emergency department. The American Journal of Emergency Medicine, 29(1): 18-25 [DOI: 10.1016/j.ajem.2009.07.011]

Shen Y Q, Chen Z, Mamalakis M, Liu Y G, Li T B, Su Y Z, He J J, Liò P and Wang Y G. 2024. TourSynbio: a multi-modal large model and agent framework to bridge text and protein sequences for protein engineering [EB/OL]. [2024-07-31]. https://arxiv.org/pdf/2408.15299.pdf

Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, Clark K, Pfohl S, Cole-Lewis H, Neal D, Schaekermann M, Wang A, Amin M, Lachgar S, Mansfield P, Prakash S, Green B, Dominowska E, Arcas B A Y, Tomasev N, Liu Y, Wong R, Semturs C, Mahdavi S S, Barral J, Webster D, Corrado G S, Matias Y, Azizi S, Karthikesalingam A and Natarajan V. 2023. Towards expert-level medical question answering with large language models [EB/OL]. [2024-07-31]. https://arxiv.org/pdf/2305.09617.pdf

Tan Y, Zhang Z X, Li M C, Pan F, Duan H, Huang Z J, Deng H, Yu Z H, Yang C, Shen G Y, Qi P, Yue C Y, Liu Y X, Hong L, Yu H Q, Fan G S and Tang Y. 2024. MedChatZH: a tuning LLM for traditional Chinese medicine consultations. Computers in biology and medicine, 172: #108290 [DOI: 10.1016/j.compbiomed.2024.108290]

Thawakar O C, Shaker A M, Mullappilly S S, Cholakkal H, Anwer R M, Khan S, Laaksonen J and Khan F S. 2023. XrayGPT: chest radiographs summarization using large medical vision-language models//Proceedings of the 23rd Workshop on Biomedical Natural Language Processing. Bangkok, Thailand: Association for Computational Linguistics: 440-448 [DOI: 10.18653/v1/2024.bionlp-1.35]

Tian S B, Jin Q, Yeganova L, Lai P T, Zhu Q Q, Chen X Y, Yang Y F, Chen Q Y, Kim W, Comeau D C, Islamaj R, Kapoor A, Gao X and Lu Z Y. 2024a. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Briefings in Bioinformatics, 25(1): #bbad493 [DOI: 10.1093/bib/bbad493]

Tian Y H, Gan R Y, Song Y, Zhang J X and Zhang Y D. 2024b. ChiMed-GPT: a Chinese medical large language model with full training regime and better alignment to human preferences//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand: Association for Computational Linguistics: 7156-7173 [DOI: 10.18653/v1/2024.acl-long.386]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I. 2017. Attention is all you need//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 6000-6010

Vedantam R, Lawrence Zitnick C and Parikh D. 2015. CIDEr: consensus-based image description evaluation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 4566-4575 [DOI: 10.1109/CVPR.2015.7299087]

Vinyals O, Toshev A, Bengio S and Erhan D. Show and tell: a neural image caption generator//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE: 3156-3164 [DOI: 10.1109/CVPR.2015.7298935]

Wang C, Li M Y, He J J, Wang Z R, Darzi E, Chen Z, Ye J, Li T B, Su Y Z, Ke J, Qu K L, Li S X, Yu Y, Liò P, Wang T Y, Wang Y G and Shen Y Q. 2024. A survey for large language models in biomedicine [EB/OL]. [2024-07-31]. https://arxiv.org/pdf/2409.00133.pdf

Wang H C, Liu C, Xi N W, Qiang Z W, Zhao S D, Qin B and Liu T. 2023a. HuaTuo: tuning LLaMA model with Chinese medical knowledge [EB/OL]. [2024-07-31]. https://arxiv.org/pdf/2304.06975.pdf

Wang S, Tang L Y, Lin M Q, Shih G, Ding Y and Peng Y F. 2022a. Prior knowledge enhances radiology report generation. AMIA Joint Summits on Translational Science Proceedings, 2022: 486-495

Wang Z Y, Han H W, Wang L, Li X and Zhou L P. 2022b. Automated radiographic report generation purely on Transformer: a multicriteria supervised approach. IEEE Transactions on Medical Imaging, 41(10): 2803-2813 [DOI: 10.1109/TMI.2022.3171661]

Wang Z Y, Liu L Q, Wang L and Zhou L P. 2023b. R2genGPT: radiology report generation with frozen LLMs. Meta-Radiology, 1(3): #100033 [DOI: 10.1016/j.metrad.2023.100033]

Wang Z Y, Zhou L P, Wang L and Li X. 2021. A self-boosting framework for automated radiographic report generation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE: 2433-2442 [DOI: 10.1109/CVPR46437.2021.00246]

Xing S X, Fang J Z, Ju Z H, Guo Z and Wang Y. 2024. Research on automatic generation of multimodal medical image reports based on memory driven. Journal of Biomedical Engineering, 41(1): 60-69 (邢素霞, 方俊泽, 鞠子涵, 郭正, 王瑜. 2024. 基于记忆驱动的多模态医学影像报告自动生成研究. 生物医学工程学杂志, 41(1): 60-69) [DOI: 10.7507/1001-5515.202304001]

Xiong H L, Wang S, Zhu Y T, Zhao Z H, Liu Y X, Huang L L, Wang Q and Shen D G. 2023. DoctorGLM: fine-tuning your Chinese doctor is not a herculean task [EB/OL]. [2024-07-31]. https://arxiv.org/pdf/2304.01097.pdf

Xu D X, Chen Y Y, Wang J Y, Huang Y, Wang H P, Jin Z, Wang H X, Yue W H, He J, Li H and Huang Y. 2024a. MLeVLM: improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering//Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand: Association for Computational Linguistics: 4977-4997 [DOI: 10.18653/v1/2024.findings-acl.296]

Xu Z Y, Feng C, Shao R L, Ashby T, Shen Y, Jin D, Cheng Y, Wang Q F and Huang L F. 2024b. Vision-flan: scaling human-labeled tasks in visual instruction tuning//Findings of the Association for Computational Linguistics: ACL 2024. Bangkok, Thailand: Association for Computational Linguistics

Yan H, Liu Y L, Jin L W and Bai X. 2023. The development, application, and future of LLM similar to ChatGPT. Journal of Image and Graphics, 28(9): 2749-2762 (严昊, 刘禹良, 金连文, 白翔. 2023. 类ChatGPT大模型发展、应用和前景. 中国图象图形学报, 28(9): 2749-2762) [DOI: 10.11834/jig.230536]

Yang S H, Zhao H J, Zhu S B, Zhou G Y, Xu H F, Jia Y X and Zan H Y. 2024. Zhongjing: enhancing the Chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 19368-19376 [DOI: 10.1609/aaai.v38i17.29907]

Yao Y F, Duan J H, Xu K D, Cai Y F, Sun Z B and Zhang Y. 2024. A survey on large language model (LLM) security and privacy: the good, the bad, and the ugly. High-Confidence Computing, 4(2): #100211 [DOI: 10.1016/j.hcc.2024.100211]

Zhang C, Liwicki S and Cipolla R. 2022. Beyond the CLS token: image reranking using pretrained vision Transformers//Proceedings of the 33rd British Machine Vision Conference 2022. London, UK: BMVA Press: #80

Zhang J F, Han X X, Yang Z H, Wang Z C, Zheng J J, Yang Z M and Zhu J M. 2021. Radiology residency training in China: results from the first retrospective nationwide survey. Insights into Imaging, 12(1): #25 [DOI: 10.1186/s13244-021-00970-2]

Zhang L Y, Shu J H, Hu J L, Li F F, He J J, Wang P and Shen Y Q. 2024. Exploring the potential of large language models in radiological imaging systems: improving user interface design and functional capabilities. Electronics, 13(11): #2002 [DOI: 10.3390/electronics13112002]

Zhang S, Xu Y B, Usuyama N, Xu H W, Bagga J, Tinn R, Preston S, Rao R, Wei M, Valluri N, Wong C, Tupini A, Wang Y, Mazzola M, Shukla S, Liden L, Gao J F, Lungren M P, Naumann T, Wang S and Poon H. 2023. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs [EB/OL]. [2024-07-31]. https://arxiv.org/pdf/2303.00915.pdf

Zhang Y X, Wang X S, Xu Z Y, Yu Q H, Yuille A and Xu D G. 2020. When radiology report generation meets knowledge graph//Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto, USA: AAAI: 12910-12917 [DOI: 10.1609/aaai.v34i07.6989]

作者简介

李恒泰,男,硕士研究生,主要研究方向为医学图像处理和机器学习。E-mail:232115003@mail.sdufe.edu.cn

刘慧,通信作者,女,教授,主要研究方向为医学图像处理、数据挖掘和机器学习。E-mail:liuh_lh@sdufe.edu.cn

陈公冠,男,博士研究生,主要研究方向为图像处理和机器学习。E-mail:cgg970411@163.com

闫子申,男,硕士研究生,主要研究方向为图像处理。E-mail:232215050@mail.sdufe.edu.cn

盛玉瑞,女,住院医师,主要研究方向为医学影像诊断。E-mail:shengyurui168@sina.com

张彩明,男,教授,主要研究方向为数据挖掘、计算机图形学、信息可视化和医学图像处理。E-mail:czhang@sdu.edu.cn