

基于细粒度检索增强的医学报告生成

牛庆港^{1,2}, 张晓伟^{1,2}, 吕千一^{1,2}, 李 真³, 闫中敏^{1,2}, 王 峻², 郭茂祖⁴, 余国先^{1*}

(1. 山东大学软件学院, 山东济南 250101; 2. 山东大学人工智能国际联合研究院, 山东济南 250101;

3. 山东大学齐鲁医院, 山东济南 250012; 4. 北京建筑大学电气与信息工程学院, 北京 100044)

摘 要: 医学诊断报告的自动生成能减轻医生工作负担、提高报告生成效率,并降低主观偏差。现有检索增强方法将相似病例报告作为参考文本引入生成过程,为模型提供诊断语义提示,来提升生成报告的质量,但多数方法仍停留在粗粒度文本检索层面,易出现“全局相似而局部病灶不一致”的问题,导致参考报告中的无关病灶描述干扰生成结果。此外,不同医生的报告书写风格差异显著,直接引入多个检索文本会因表述风格不一致带来噪声干扰,影响生成结果的准确性。为解决上述问题,本文提出一种基于细粒度检索增强的医学报告生成框架(Fine-grained Retrieval-enhanced Medical Report Generation, FRMRG)。首先,设计解剖感知细粒度校正模块,对检索得到的粗粒度文本进行细粒度校正,保留与当前影像相关的有效信息,削弱不一致病灶描述的干扰;其次,构建结构化提示驱动的初始报告生成模块,将多个相关报告映射到统一语义空间,并通过固定提示模板组织相关诊断报告,削弱报告书写风格差异的影响,生成初始报告;最后,设计多模态最终报告生成模块,将初始报告作为语义先验,结合视觉特征引导大语言模型生成连贯且符合临床事实的医学报告。通过上述设计,FRMRG能够充分挖掘检索报告中的有效诊断信息,同时降低无关病灶描述及书写风格差异带来的干扰,并结合视觉信息对报告生成过程进行约束,从而提升生成内容与医学影像之间的语义一致性。实验结果显示,FRMRG在印第安纳大学胸部X线数据集(Indiana University Chest X-Ray Collection, IU X-Ray)、重症监护医学信息数据库胸部X线数据集(Medical Information Mart for Intensive Care Chest X-Ray, MIMIC-CXR)和私有临床胃镜数据集上的BLEU-1分别较次优方法提升2.8%、2.1%和1.7%,METEOR分别提升1.4%、3.0%和3.2%,表明FRMRG能够提升医学报告生成质量,并具有较好的泛化能力。

关键词: 医学报告生成;检索增强生成;解剖区域对齐;细粒度校正;结构化提示;视觉语言模型;多模态学习

基金项目: 国家自然科学基金重点项目(No.62531013);国家自然科学基金项目(No.62272276)

中图分类号: TP391.4;R445 **文献标识码:** A **文章编号:** 0372-2112(XXXX)XX-0001-14

电子学报 URL: <http://www.ejournal.org.cn>

DOI: 10.12263/DZXB.20260621

Medical Report Generation With Fine-grained Retrieval-based Refinement

NIU Qinggang^{1,2}, ZHANG Xiaowei^{1,2}, LV Qianyi^{1,2}, LI Zhen³, YAN Zhongmin^{1,2},
WANG Jun², GUO Maozu⁴, YU Guoxian^{1*}

(1. School of Software, Shandong University, Jinan, Shandong 250101, China;

2. Joint Centre for Artificial Intelligence Research, Shandong University, Jinan, Shandong 250101, China;

3. Qilu Hospital of Shandong University, Jinan, Shandong 250012, China;

4. College of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing 100044, China)

Abstract: Automatic generation of medical diagnostic reports can reduce physicians' workload, improve reporting efficiency, and mitigate subjective variations in report writing. Existing retrieval-enhanced methods usually introduce reports from similar cases as reference texts into the generation process to provide diagnostic semantic cues and improve report quality. However, most existing methods remain at the level of coarse-grained text retrieval, which can lead to the problem of “global similarity but local lesion inconsistency”, causing irrelevant lesion descriptions in retrieved reports to interfere with the generated results. In addition, substantial differences in terminology and writing styles among physicians may introduce additional noise when multiple retrieved reports are directly incorporated, thereby affecting the accuracy of report generation. To address these issues, this paper proposes a fine-grained retrieval-enhanced medical report generation framework, termed FRMRG(Fine-grained Retrieval-enhanced Medical Report Generation). First, an anatomy-aware fine-grained correction module is designed to refine the coarse-grained retrieved texts, preserve information relevant to the

current medical image, and suppress interference from inconsistent lesion descriptions. Second, a structured prompt-driven initial report generation module is constructed to map multiple relevant reports into a unified semantic space and organize diagnostic information using fixed prompt templates, thereby reducing the influence of differences in report-writing styles and generating an initial report. Finally, a multimodal final report generation module is developed, in which the initial report serves as a semantic prior and is integrated with visual features to guide a large language model in generating coherent medical reports consistent with clinical findings. Through the above design, FRMRG can effectively exploit diagnostic information contained in retrieved reports, reduce interference from irrelevant lesion descriptions and writing-style variations, and further constrain the report generation process with visual information, thereby improving the semantic consistency between generated content and medical images. Experimental results demonstrate that FRMRG consistently outperforms the second-best methods on the IU-X-Ray(Indiana University Chest X-Ray Collection), MIMIC-CXR(Medical Information Mart for Intensive Care Chest X-Ray), and a private clinical gastroscopy dataset, with BLEU-1 improvements of 2.8%, 2.1%, and 1.7%, respectively, and METEOR improvements of 1.4%, 3.0%, and 3.2%, respectively. These results indicate that FRMRG can effectively improve the quality of medical report generation and exhibits good generalization capability.

Keywords: medical report generation; retrieval enhancement; region alignment; fine-grained refinement; structured prompting; vision-language model; multimodal learning

Foundation Item(s): Key Program of the National Natural Science Foundation of China (No.62531013); National Natural Science Foundation of China (No.62272276)

0 引言

由于医学影像在临床诊断中的应用日趋广泛,放射科医生需要判读的影像数量以及报告撰写的工作量都急剧增加,工作负担不断加重^[1]。为缓解这一问题,智能医学报告生成成为一种具有前景的解决方案。该类方法把视觉编码器与语言生成模型结合起来,从医学图像中自动提取视觉特征,再据此生成语义明确、有临床价值的诊断文本,故能切实减轻医生负担,也有利于提高诊断效率^[2-3]。然而,生成符合临床需求的高质量诊断报告仍面临重大挑战,模型不仅要准确理解医学图像内容,还必须生成符合临床书写规范的诊断文本。为提高报告生成的准确性与可靠性,基于对齐的医学报告生成方法被提出来。现有基于对齐的生成方法大致可分为两类。第一类为粗粒度对齐方法^[4-7],其将整幅医学图像与完整诊断报告直接关联,由此很好地建模全局语义一致性,但难以捕捉局部区域中的细微病灶信息,容易造成信息冗余或关键细节丢失。第二类为细粒度对齐方法^[8-9],其明确建立了视觉区域与文本句子之间的对应关系,因而能更充分、更准确地分析局部病灶特征。然而,此类方法通常依赖代价较高的像素级或短语级标注,在弱监督或标注信息稀缺的情形下,其性能容易波动。因此,当前对医学报告生成的研究亟需设计一种更高效、更鲁棒的对齐策略,在不过度依赖密集标注的前提下,可靠地建立细粒度的视觉-文本对应关系。

除了局部视觉语义不一致的问题,参考报告的表达差异也会影响报告生成质量,由于不同临床医生撰

写的诊断报告在语言风格、表述方式上存在差异,即使是相同或相似的临床发现,不同医生也可能采用不同的句式、用词习惯和语法结构来描述,所以表达方式上的不一致容易引入一定程度上的语言噪声,干扰模型建立稳定的视觉-文本对应关系,也直接降低了基于提示的生成方式的可靠性。因此,当前亟需引入一种标准化的处理机制,在保证临床语义等价性的前提下对多样化语言表达加以统一规范。

医学报告生成相关研究目前绝大多数都以放射学场景为出发点,且以胸部X线影像为主,故而现有方法处理此类任务时一般使用标准投照体位下的单视图或双视图影像来生成诊断报告,即前后位及侧位图像。由于一幅此类图像常包含若干解剖结构,因此利用全局特征表示自然更有利于挖掘影像内容与文本描述之间的关联。相比之下,其他影像场景,尤其是内镜图像(如胃镜图像)有十分明确、不同于放射学的特性,单张图像一般仅对应一个具体解剖部位。因此,现有依赖于整体多区域图像建模的方法很难直接泛化到内镜场景。由此亟需设计一种符合内镜影像特点的对齐与生成策略,方能在内镜场景下输出符合临床规范、又高度准确的医学报告。

针对上述问题,本文提出一种基于细粒度检索增强的医学报告生成框架(Fine-grained Retrieval-enhanced Medical Report Generation, FRMRG)。该框架首先设计了解剖感知细粒度校正模块,以解剖区域为基本单元对粗粒度检索文本进行区域级语义修正,因此能很好地解决“全局相似而局部不一致”问题。

该设计既有利于建模胸部X线图像中多解剖区域与相应诊断描述的对应关系,也十分适合于胃镜等内镜图像中单张图像仅对应某一部位或局部病灶的特点;其次,框架还设计了结构化提示驱动的初始报告生成模块,把若干相关报告映射到同一语义空间,再用固定形式的提示模板将语义一致的诊断信息加以聚合,由此弱化书写风格差异带来的干扰;最后,将结构化初始报告作为语义先验,结合视觉特征来引导大语言模型生成真正临床一致的医学报告。实验证明,FRMRG在多个数据集上的主流评价指标上都优于目前几种有代表性的医学报告生成方法^[2,4-5],本文的主要贡献概括如下:

(1)提出了一种基于细粒度检索增强的医学报告生成框架FRMRG。该框架利用解剖引导的区域对齐和细粒度检索增强策略,充分挖掘医学报告中所涉的诊断语义信息,最终得到准确且具备临床一致性的医学报告。

(2)提出了解剖感知细粒度校正范式。该范式能够减少人工细粒度标注依赖,并且实现对检索报告中的局部不一致描述进行校正,同时该模块能够很好地适应内镜影像“一张图片对应一个主要解剖区域”的场景,增强了模型的泛化能力。

(3)构建结构化提示驱动的初始报告生成模块。该模块将若干有关联的报告映射到统一的语义空间,再用标准化提示模板把语义一致的诊断线索汇集起来,减少不同医生书写风格差异带来的文本噪声,生成表达相对规范的初始报告,为后续报告生成提供了稳定的语义基础。

1 相关工作

1.1 基于检索增强的报告生成方法

基于检索增强的医学报告生成(medical report generation, MRG)需要解决的一个根本问题是如何将图像中的视觉信息与文本中的描述信息对应起来以实现精确检索。根据对应粒度的不同,现有方法大体可以分为粗粒度对齐和细粒度对齐。具体而言,粗粒度对齐处理的是整张图像与全篇报告之间的相似性,而细粒度对齐处理的是局部区域与相应语句之间的对应关系。

基于粗粒度对齐的报告生成方法。粗粒度对齐方法出现较早,原理简单,因此被大量报告生成模型所采用。这类方法一般先提取整幅医学图像的视觉特征,再与自身报告或历史报告做粗粒度对齐,选出语义相近的医学报告作为生成参考^[4-5,10-11]。为增强医学报告生成的准确性,研究者已在粗粒度图文匹配的基础上从多个角度有层次地增强生成模型的语义表

达能力。一部分方法从上下文推理的角度增强视觉—语言之间的一致性,以提升视觉模态与文本模态的语义对应关系。例如,Chen等^[12]在Transformer中引入记忆机制,使模型在生成报告时能够保存并更新图像与文本之间的语义信息。而Yang等^[13]把历史临床语句作为上下文提示,让生成过程可以直接参考已有诊断表达。除了直接建模图像和文本之间的关系外,也有研究引入医学知识或多模态信息,Shentu等^[14]用医学知识图谱对诊断描述施加明确约束;Li等^[5]更进一步地构造了动态知识图谱,从检索报告中抽取临床知识并在生成过程中动态更新图谱结构。周涛等^[15]利用正电子发射断层成像/计算机断层扫描(Positron Emission Tomography/Computed Tomography, PET/CT)跨模态信息做肺部肿瘤检测,论证了不同模态影像之间的互补信息对病灶识别的实际帮助。还有若干研究从多任务学习或对比学习的角度巧妙地增强了模型的表征能力,Jing等^[16]把疾病预测和报告生成放在同一模型中联合学习。而面向疾病自动预测与放射学报告生成的联合学习框架(Jointly learn to align: automated disease Prediction and radiology report Generation, JPG)把特征对齐、疾病预测都设计为辅助任务,直接增强了跨模态表征能力^[17]。

基于粗粒度对齐的报告生成方法有助于增强报告的整体语义表达,但其匹配对象是整幅图像和完整报告,故很容易忽略局部病灶的细节。因此当检索到的报告与当前图像整体相似而局部描述不同时,所生成的报告内容可能与当前样本不符。

基于细粒度对齐的报告生成方法。由于粗粒度对齐局部描述不一致的问题比较突出,故目前学界对细粒度方法已有一些探索。例如,Tanida等^[8]提出区域引导的报告生成框架,先检测解剖区域,再就各区域分别生成描述。Chen等^[18]提出AdaMatch方法,在胸部X光图像块和报告词元之间直接建立匹配关系,因而能更合理地关注具体的图文对应关系。Yin等^[19]提出面向放射报告自动生成的疾病知识增强细粒度图文对齐框架(Disease-Knowledge-Enhanced Fine-Grained Image-Text Alignment for Automatic Radiology Report Generation, DKA-RG),将疾病知识引入病灶区域和文本描述的对齐过程。与粗粒度方法不同的是,细粒度对齐更有利于直接约束局部图像内容与诊断语句的关系,因此对“整体相似但局部不一致”的问题有更切实有效的缓解作用。

然而,现有细粒度对齐方法仍有不足,显式区域标注一般要依靠人工标注或预定义的解剖结构,故标注成本很高,而检索所得的各参考报告语句又常有描述顺序、术语习惯、书写风格等诸种差异,因此简单

拼接必然引入额外噪声。

基于以上问题,本文提出FRMRG。该方法不直接依赖大量人工细粒度标注,而是利用大语言模型对检索文本中的解剖区域和病灶描述进行结构化标记。随后,模型根据这些标记引导图像区域和文本语义进行细粒度校正。这既保留了报告的整体结构又实现了粗粒度报告的局部修正,提升了检索报告的语义一致性。

1.2 基于大语言模型的报告生成方法

近年来,研究者开始将大语言模型(large language models, LLMs)引入医学报告生成任务来考察其在复杂临床场景中的适应能力以及泛化能力。具体而言,Google提出的Med-PaLM2在医学问答基准上已经达到了接近医生水平的表现^[20]。与此形成极好补充的是BioGPT等生物医学领域专用模型^[21],其在大量医学语料上做了充分预训练,因此在医学实体识别、医学推理两项任务中都表现突出。除医学文本理解任务外,刘青海等^[22]面向大语言模型开展医学图像语义挖掘与跨模态哈希研究,说明大语言模型在医学图像与文本语义关联建模中也具有较好的应用潜力。进一步地,R2GenGPT和BootstrappingLLM等方法^[23-24]将视觉编码器与大语言模型巧妙地结合起来,成功实现了胸部X线报告的端到端生成。故而目前已有研究从医学文本理解、医学图像语义挖掘、报告生成诸种角度完备地验证了大语言模型在医学场景中的有效性。

受前文所论方法的启发,本文提出的FRMRG框架把细粒度校正后的检索文本与大语言模型的生成能力很好地结合起来,因而有利于提高医学报告生成的诊断准确性及临床一致性。更重要的是,与现有方法不同,FRMRG没有直接把检索文本作为非结构化提示输入模型,而是引入结构化建模过程,把解剖区域信息及检索语义统一地组织起来,减少不同医生书写风格给大语言模型提示的干扰,从而提升生成报告的准确性。

2 方法

FRMRG是以编码器-生成器为基础构造的,主要分为解剖感知细粒度校正模块、结构化提示驱动的初始报告生成模块和多模态最终报告生成模块三个部分。各模块协同工作、彼此衔接,分别承担局部语义修正、初始报告构建和最终报告细化生成的功能。整体框架如图1所示。

2.1 解剖感知细粒度校正

参考相似历史报告有利于提升医学报告生成的临床准确性,但单纯从整幅图像的全局相似性出发进

行检索所得的结果一般是粗粒度匹配结果,容易出现“全局相似而局部不一致”的问题。因此,本文设计了解剖感知细粒度校正模块。解剖感知是指以预定义解剖部位为基本建模单元,建立局部影像特征与相应报告描述之间的语义对应关系。在此基础上,模型以粗粒度检索报告为基础,结合当前影像的区域视觉特征,对其中的局部描述进行区域级修正。针对不同影像模态,胸部X光按多个解剖区域分别建模,胃镜图像则依据单一局部视野进行整体图文对齐。

胸部X光图像处理。对于第*i*个胸部X光样本,其输入视图集合为 $\{\mathbf{X}_{i,v}^I\}_{v=1}^{V_i}$,其中用于区域级建模的正位图像记为 $\tilde{\mathbf{X}}_i^I$ 。本文采用全局-局部双分支视觉编码方法。用全局分支提取整幅图像的整体语义表示,用局部分支提取正位图像各解剖区的视觉特征,分别用来支持全局图文对齐及解剖区域级文本对齐。

在全局分支中,本文采用Swin Transformer^[25]作为视觉编码器。具体而言,对于每个输入视图,首先利用全局视觉编码器提取视觉特征,然后用多视图平均聚合与投影映射的方法得到样本级全局视觉表示:

$$\mathbf{z}_i^{lg} = \text{Norm} \left(\mathbf{W}_g \left(\frac{1}{V_i} \sum_{v=1}^{V_i} f_g(\mathbf{X}_{i,v}^I) \right) \right) \quad (1)$$

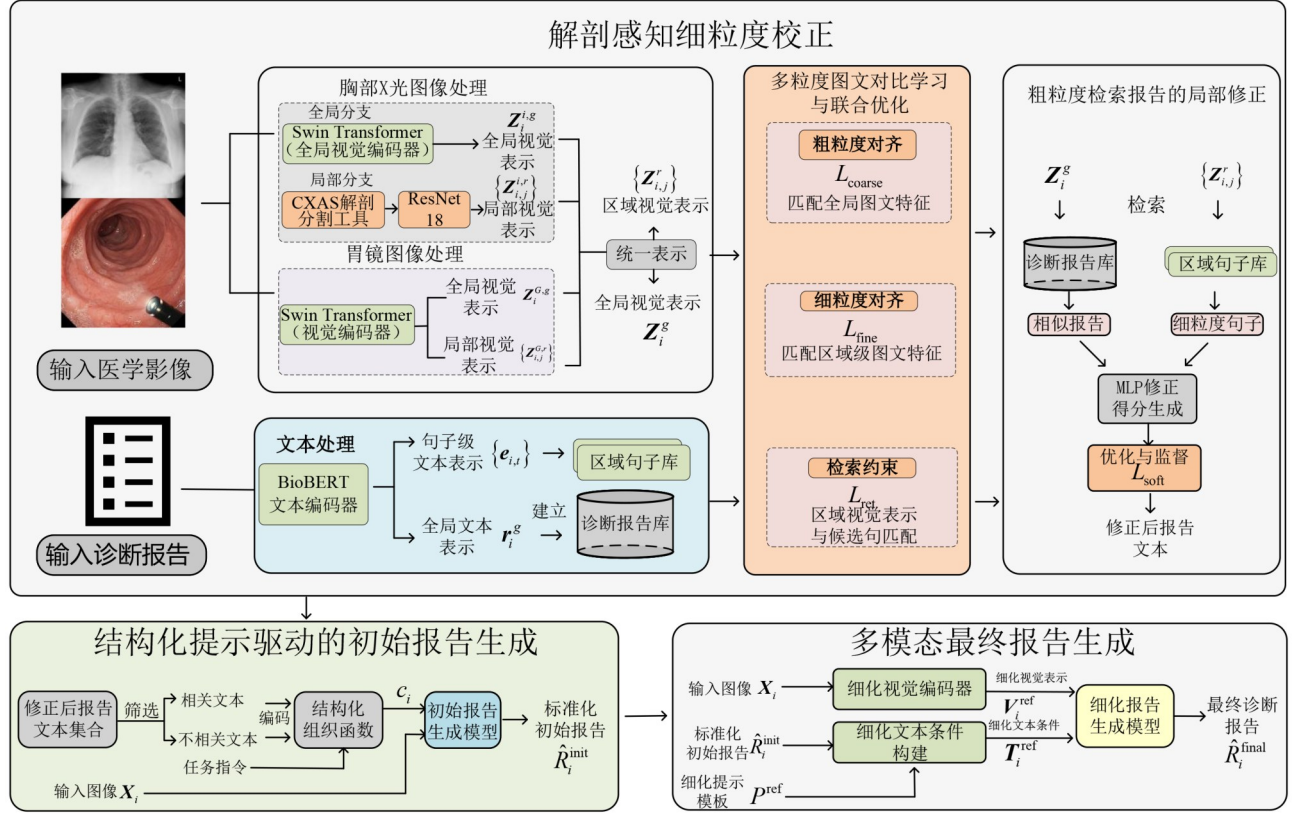
其中, $f_g(\cdot)$ 代表全局视觉编码器,用来提取单一视图的全局语义特征; V_i 表示样本*i*的输入视图数;平均操用来融合多视图信息; $\mathbf{W}_g(\cdot)$ 表示投影头,把视觉特征映射到联合嵌入空间; $\text{Norm}(\cdot)$ 表示归一化操作。

在局部分支中,本文引入CXAS^[26]解剖分割工具获得正位胸片中各预定解剖区域的掩码 $A_{i,j}$ 以及区域可用性标记 $a_{i,j} \in \{0,1\}$,并采用ResNet18^[27]对正位图像提取空间特征图, $\mathbf{H}_i = f_r(\tilde{\mathbf{X}}_i^I)$ 。随后,将区域掩码映射至特征图尺度,并通过掩码编码器生成软区域权重 $\mathbf{S}_{i,j}$,用于突出目标区域的重要视觉响应并抑制无关区域干扰。基于软区域权重进行加权池化,得到第*j*个解剖区域的局部视觉特征。相关的CXAS区域信息失效鲁棒性分析见附录A。

$$\mathbf{v}_{i,j} = \frac{\sum_{h,w} \mathbf{H}_i(:,h,w) \odot \mathbf{S}_{i,j}(h,w)}{\sum_{h,w} \mathbf{S}_{i,j}(h,w) + \epsilon} \quad (2)$$

其中, ϵ 是预防分母为零的平滑项。这个加权池化过程聚合了目标区域内的视觉特征,并抑制了其他区域对该区域特征提取的影响。

最后,将局部视觉特征送入区域投影头,映射到联合嵌入空间,得到第*j*个解剖区域的视觉表示:



注:该框架主要由三个关键模块组成。(1)解剖感知细粒度校正模块,首先分别对医学图像和诊断文本进行编码,获得全局视觉表示、区域级视觉表示以及文本表示。随后通过粗粒度对齐、细粒度对齐和检索约束进行多粒度图文对比学习,以增强图像区域与诊断语句之间的语义对应关系;最后基于区域句子库对粗粒度检索报告进行局部修正,缓解“全局相似而局部不一致”的问题。(2)结构化提示驱动的初始报告生成模块,将多个相关报告映射到统一语义空间,并通过固定提示模板整合语义一致的诊断线索,从而削弱不同医生书写风格差异带来的干扰。(3)多模态最终报告生成模块,将初始报告作为语义先验,结合视觉特征引导大语言模型生成具有临床一致性的医学报告。

图1 FRMRG整体框架流程图

Figure 1 Overall framework of FRMRG

$$\mathbf{z}_{i,j}^{L,r} = a_{i,j} \cdot \text{Norm}\left(\mathbf{W}_j(\mathbf{v}_{i,j})\right), j=1,2,\dots,M \quad (3)$$

其中, $\mathbf{W}_j(\cdot)$ 表示第 j 个解剖区域对应的投影头, $\text{Norm}(\cdot)$ 表示归一化操作。由此,能够得到胸部X光图像的区域级视觉表示集合 $\{\mathbf{z}_{i,j}^{L,r}\}_{j=1}^M$, 并用来支撑后续细粒度图文对齐、检索修正的过程。

胃镜图像处理。与胸部X光图像不同,一张胃镜图像通常只呈现一个主要观察部位。因此,在处理胃镜图像时,本文不再对单张图像进行多区域划分,而是针对每个预定义解剖部位选取一张有代表性的图像进行特征提取。设第 i 个样本在第 j 个解剖部位对应的代表性胃镜图像为 $\mathbf{X}_{i,j}^G$, 其中 $j=1,2,\dots,M$, M 表示解剖部位总数。

对于每幅胃镜图像,本文采用 Swin Transformer 提取其视觉特征,并通过全局平均池化与区域投影得到对应解剖部位的区域级视觉表示:

$$\mathbf{z}_{i,j}^{G,r} = \text{Norm}\left(\mathbf{W}_j\left(\text{GAP}\left(f_g^G(\mathbf{X}_{i,j}^G)\right)\right)\right) \quad (4)$$

其中, $f_g^G(\cdot)$ 表示胃镜图像视觉编码器, $\text{GAP}(\cdot)$ 表示全局平均池化操作, $\mathbf{W}_j(\cdot)$ 表示第 j 个解剖部位对应的投影头, $\text{Norm}(\cdot)$ 表示归一化操作。

进一步地,对所有解剖部位的区域特征进行聚合,以构建病例级全局视觉表示:

$$\mathbf{z}_i^{G,g} = \text{Norm}\left(\mathbf{W}_g\left(\frac{1}{M} \sum_{j=1}^M \text{GAP}\left(f_g^G(\mathbf{X}_{i,j}^G)\right)\right)\right) \quad (5)$$

由此,可得到胃镜图像的病例级全局视觉表示 $\mathbf{z}_i^{G,g}$ 及其解剖部位级视觉表示 $\{\mathbf{z}_{i,j}^{G,r}\}_{j=1}^M$ 。为便于后续模块统一表述,不再区分胸部X光图像和胃镜图像的特征记号,分别将全局视觉表示和区域级视觉表示统一记为 $\mathbf{z}_i^g$ 和 $\{\mathbf{z}_{i,j}^r\}_{j=1}^M$ 。

文本处理。与图像处理过程相对应,本文对第 i 个样本的诊断报告 R_i 进行全局与局部两级文本建模。首先,利用 BioBERT 对完整报告进行编码,得到报告级文本表示 $\mathbf{r}_i^g$, 用于全局图文对齐和粗粒度报告检索。随后,将报告切分为句子集合 $S_i = \{s_{i,t}\}_{t=1}^{T_i}$, 并

对每个句子进行编码,得到句子级文本表示 $\{e_{i,t}\}$,用于细粒度图文对齐与区域级检索修正。

为建立句子与解剖区域之间的对应关系,本文利用大语言模型为每个句子自动标注一个或多个对应的解剖区域,并将标签表示为多热向量 $y_{i,t} \in \{0, 1\}^M$,其中 $y_{i,t,j} = 1$ 表示句子 $s_{i,t}$ 与第 j 个解剖区域相关,否则为0。在此基础上,本文构建两个文本库。一是诊断报告库,用于保存完整报告及其报告级文本表示 r_i^s ,为粗粒度检索提供整体语义参考;二是区域句子库,将带有区域标签的句子按照解剖区域进行归类,用于后续区域级候选句检索和细粒度校正。

多粒度图文对比学习。为实现对粗粒度检索报告的区域级修正,模型不仅需要获得与当前影像整体语义相近的参考报告,还需要判断报告中的局部描述是否与相应解剖区域一致,并从候选句子中选择更合适的替换内容。因此本文在得到的全局视觉表示、区域级视觉表示及文本表示的基础上,从粗粒度对齐、细粒度对齐和检索约束三个层面联合对模型做了优化,来分别建模图像与报告之间的全局语义对应、句子与解剖区域之间的局部语义对应,以及区域表示对相关文本的检索判别能力。

在粗粒度层面,利用全局图像表示与完整报告表示进行双向对齐。设样本 i 的全局视觉表示为 z_i^s ,样本 j 的报告级文本表示为 r_j ,则图像 i 与报告 j 之间的匹配分数定义为

$$s_{ij} = \alpha z_i^s r_j^T \quad (6)$$

其中 α 是可以学习的温度系数。基于一个批次内所有图像与报告的匹配分数,可以构造匹配分数矩阵 $S = [s_{ij}]$,其中第 (i,j) 个元素表示第 i 个图像与第 j 个报告之间的匹配程度;对角线元素对应真实匹配样本对,非对角线元素对应非匹配样本对。粗粒度对齐损失定义为

$$\mathcal{L}_{\text{coarse}} = \frac{1}{2} [\text{CE}(S, m) + \text{CE}(S^T, m)] \quad (7)$$

其中 m 表示批次内的真实匹配标签。这个损失分别从“图像到报告”和“报告到图像”两个方向进行约束,用来拉近同一样本的图像和文本的嵌入距离,拉远不同样本图像和文本的嵌入距离,并为后续粗粒度报告检索提供依据。

在细粒度层面,在前文得到的句子级文本表示 $e_{i,t}$ 及其解剖区域标签 $y_{i,t}$ 的基础上,本文建立句子与解剖区域之间的多标签对齐关系。对于样本 i 的第 j 个解剖区域,设其区域视觉表示为 $z_{i,j}^r$,则句子表示与该区域之间的匹配分数定义为

$$p_{i,t,j} = \sigma \left(\frac{(z_{i,j}^r)^T e_{i,t}}{\tau_1} \right) \quad (8)$$

其中 τ_1 为温度系数。 $\sigma(\cdot)$ 表示Sigmoid函数。这个分数用于刻画句子表示与区域表示之间的语义一致性,分数越高表示句子越可能描述该区域。

进一步地,设区域可用性掩码为 $a_i \in \{0, 1\}^M$,其中 $a_{i,j} = 1$ 表示该区域在当前样本中有效,则细粒度对齐损失定义为

$$\mathcal{L}_{\text{fine}} = \frac{1}{|B|} \sum_{i \in B} \frac{1}{T_i} \sum_{t=1}^{T_i} \frac{\sum_{j=1}^M a_{i,j} \text{BCE}(p_{i,t,j}, y_{i,t,j})}{\sum_{j=1}^M a_{i,j} + \epsilon} \quad (9)$$

其中 B 表示一个批次内的样本集合, M 表示解剖区域类别数, $\text{BCE}(\cdot)$ 表示二元交叉熵损失, ϵ 为平滑项。这个损失利用解剖区域掩码屏蔽无效区域,并通过句子级标签学习文本描述与解剖区域之间的对应关系。

在检索约束层面,考虑到细粒度对齐主要建模区域视觉表示与区域文本表示之间的语义一致性,而没有直接对区域表示在候选文本空间中的检索判别能力加以约束,因此本文进一步引入了检索式监督,以增强区域视觉表示对相关文本的区分及排序能力。利用前文构建的区域句子库,本文为每个解剖区域构造正、负候选句集合。具体来说,对于样本 i 的第 j 个解剖区域,设其区域视觉表示为 $z_{i,j}^r$ 。若该样本真实报告中存在与第 j 个区域对应的句子,则将其作为正候选集合 $P_{i,j}$;若不存在对应句,则认为该区域未见异常描述,即处于正常状态,并从预定义的正常句模板库中选取一条与该区域对应的正常描述句作为正候选。进一步地,从该样本其他区域对应的句子以及区域句子库中采样与当前区域不匹配的句子,构成负候选集合 $N_{i,j}$ 。于是,该区域对应的候选句集合记为 $Q_{i,j} = P_{i,j} \cup N_{i,j}$ 。

对于任一候选句 $s \in Q_{i,j}$,其与区域表示之间的匹配分数定义为

$$r_{i,j}(s) = \sigma \left(\frac{(z_{i,j}^r)^T e(s)}{\tau_2} \right) \quad (10)$$

其中 $e(s)$ 表示候选句 s 的文本表示, τ_2 为温度系数。进一步地,记候选集合 $Q_{i,j}$ 上的预测概率向量为 $r_{i,j} = \{r_{i,j}(s)\}_{s \in Q_{i,j}}$,对应的监督标签向量为 $y_{i,j}^{\text{ret}} = \{y_{i,j}^{\text{ret}}(s)\}_{s \in Q_{i,j}}$,其中 $y_{i,j}^{\text{ret}}(s) = 1$ 表示正候选句, $y_{i,j}^{\text{ret}}(s) = 0$ 表示负候选句。于是,检索约束损失定义为

$$\mathcal{L}_{\text{ret}} = \frac{1}{|B|} \sum_{i \in B} \frac{1}{M} \sum_{j=1}^M \text{BCE}(r_{i,j}, y_{i,j}^{\text{ret}}) \quad (11)$$

该约束能够进一步增强区域表示对相关文本与不相关文本的区分能力,从而为后续区域级局部修正

提供更可靠的候选排序基础。

最终,整体训练目标由粗粒度对齐损失、细粒度对齐损失与检索约束损失加权构成:

$$\mathcal{L} = \mathcal{L}_{\text{coarse}} + \lambda_f \mathcal{L}_{\text{fine}} + \lambda_r \mathcal{L}_{\text{ret}} \quad (12)$$

其中, λ_f 和 λ_r 分别表示损失的权重。粗粒度对齐损失解决样本级图文匹配问题,细粒度对齐损失处理句子与解剖区域的对应关系,检索约束损失用来提高区域候选文本检索的准确性,三者结合便能兼顾全局图像特征检索及区域级文本匹配。

粗粒度报告检索的局部修正:粗粒度检索基于病例级全局相似性,能够提供整体语义参考,但难以保证各解剖区域的描述均与当前影像一致。为此,本文利用前述多粒度图文对齐建立的区域-句子对应关系,将整份报告的修正转化为区域级候选选择。针对每个解剖区域,模型将粗粒度报告中的原始描述、区域句子库检索得到的细粒度候选句以及“跳过修正”选项组成候选集合,并判断保留原始描述或替换为更匹配的候选句,从而在保持报告整体结构的同时修正局部不一致内容。

对于查询样本 i 的第 j 个解剖区域,设其候选集合 $C_{i,j} = \{c_{i,j}^{(k)}\}_{k=1}^{K_{i,j}}$, 其中 $K_{i,j}$ 表示该区域的候选项总数,候选集合 $c_{i,j}^{(k)}$ 包括粗粒度检索报告中的原始区域描述、区域句子库检索得到的细粒度候选句以及“跳过修正”选项。对于每个候选文本,先利用文本编码函数得到其表示:

$$e_{i,j}^{(k)} = e(c_{i,j}^{(k)}) \quad (13)$$

随后,将候选文本表示输入多层感知机,得到候选选项的修正得分:

$$o_{i,j}^{(k)} = \text{MLP}(e_{i,j}^{(k)}) \quad (14)$$

在此基础上,汇总该区域所有候选选项的得分,可得区域级修正打分向量 $\mathbf{o}_{i,j} = [o_{i,j}^{(1)}, o_{i,j}^{(2)}, \dots, o_{i,j}^{(K_{i,j})}]$ 。对于不可用候选,其得分置为极小值,以避免误差。

训练阶段,若当前区域在真实报告中存在对应句,则将与该真实句最匹配的候选项类别作为监督标签 $y_{i,j}$; 否则,将“跳过修正”对应的类别作为监督标签。由此,局部修正模块的训练目标定义为

$$\mathcal{L}_{\text{slot}} = \frac{1}{|B|} \sum_{(i,j) \in B} \text{CE}(\mathbf{o}_{i,j}, y_{i,j}) \quad (15)$$

其中, B 表示一个批次内的区域样本集合, $\text{CE}(\cdot)$ 表示交叉熵损失。

通过上述设计,模型能够在保留粗粒度检索报告整体语义框架的基础上,对局部不匹配区域进行针对性修正,从而提升区域级描述的准确性与一致性。

2.2 结构化提示驱动的初始报告生成

粗粒度检索报告局部修正的根本目的是为当前图像的诊断提供更可靠的语义参考,以引导之后的报告生成。但不可避免的是,不同医生在书写风格、表达粒度上有不同特点,故单一检索报告不能充分覆盖诊断信息,而单纯引入多条检索报告又容易造成相关性下降、语义冲突的问题。因此,本文提出了结构化驱动的初始报告生成模块,对多篇相关报告中的诊断信息进行重新组织,并按照固定模板生成结构更统一的初始报告,为后续最终报告生成提供参考。

具体而言,对于样本 i , 设输入图像为 X_i 。根据前一阶段粗粒度报告检索及区域级局部修正结果,本文筛选得到 K 个与当前图像语义一致的修正后相关报告,记为 $\mathcal{T}_i^+ = \{t_{i,m}^+\}_{m=1}^K$ 。为了约束模型避免引入与当前图像不匹配的诊断描述,本文选取 K 个与当前图像语义不一致的不相关报告,记为 $\mathcal{T}_i^- = \{t_{i,m}^-\}_{m=1}^K$ 。其中,相关文本给出正向语义线索,不相关文本给出抑制性约束。

随后,利用 BioBERT 对筛选出的文本进行编码,将它映射到统一的语义表示空间,用来减弱不同文本表述方式带来的差异。在此基础上利用结构化组织函数将编码后的相关文本表示、不相关文本表示以及任务指令 x_{instr} 组织为统一的文本引导上下文:

$$C_i = \Phi(E_{\text{Bio}}(\mathcal{T}_i^+), E_{\text{Bio}}(\mathcal{T}_i^-), x_{\text{instr}}) \quad (16)$$

其中, $E_{\text{Bio}}(\cdot)$ 表示 BioBERT 文本编码函数, $\Phi(\cdot)$ 表示结构化组织函数,用于按照预定义模板对编码后的相关文本、不相关文本及任务指令进行组织,从而形成统一的文本引导表示。

在此基础上,把图像 X_i 与文本引导上下文 C_i 分别编码为视觉 token 和文本 token,并共同输入多模态大模型,得到标准化初始报告:

$$\hat{R}_i^{\text{init}} = G_0(X_i, C_i) \quad (17)$$

其中, $G_0(\cdot)$ 表示初始报告生成模型, $\hat{R}_i^{\text{init}}$ 表示生成的标准化初始报告。

通过上述设计,模型能够先利用 BioBERT 对多条文本线索进行统一语义编码,减弱不同表述方式下相同诊断语义的差异,再结合图像信息生成结构规范、语义连贯的标准化初始报告,从而为后续模块提供更稳定的文本参考。

2.3 多模态最终报告生成

在获得标准化初始报告后,本文进一步构建多模态最终报告生成模块,对初始报告进行细化修正,并得到准确、完整、符合临床逻辑的最终诊断报告。模块的基本思想是把标准化初始报告当作稳定的文本先验,再根据图像细粒度视觉证据对之加以补充、修

正,从而使得最终报告更加细化和完整。

具体而言,对于样本 i ,设输入图像为 X_i ,上一阶段生成的标准化初始报告为 $\hat{R}_i^{\text{init}}$ 。首先,利用细化视觉编码器对输入图像进行编码,并通过线性映射得到细化阶段的图像表示

$$V_i^{\text{ref}} = W_v^{\text{ref}} f_v^{\text{ref}}(X_i) + b_v^{\text{ref}} \quad (18)$$

其中, $f_v^{\text{ref}}(\cdot)$ 表示细化阶段的视觉编码函数, V_i^{ref} 表示用于最终报告生成的图像特征表示。

随后,将标准化初始报告 $\hat{R}_i^{\text{init}}$ 与细化提示模板 P^{ref} 共同构成最终报告生成阶段的文本条件:

$$T_i^{\text{ref}} = \Psi(\hat{R}_i^{\text{init}}, P^{\text{ref}}) \quad (19)$$

其中, $\Psi(\cdot)$ 表示提示组织函数,用于构建文本条件。该文本条件用于结合图像特征对已有诊断内容进行补充、修正与细化。

在此基础上,将图像表示与文本条件共同输入细化生成模型,得到最终诊断报告:

$$\hat{R}_i^{\text{final}} = G_{\theta}^{\text{ref}}(V_i^{\text{ref}}, T_i^{\text{ref}}) \quad (20)$$

其中, $G_{\theta}^{\text{ref}}(\cdot)$ 表示多模态最终报告生成模型, $\hat{R}_i^{\text{final}}$ 表示生成的最终诊断报告。

训练阶段,以真实报告 R_i 作为监督信号,对该模块采用自回归语言建模目标进行优化:

$$\mathcal{L}_{\text{final}} = - \sum_{t=1}^T \log p_{\theta}(r_{i,t} | V_i^{\text{ref}}, T_i^{\text{ref}}, r_{i,<t}) \quad (21)$$

其中, T 表示目标报告长度, $r_{i,t}$ 表示第 t 个目标词元, $r_{i,<t}$ 表示第 t 个词元之前的真实前缀序列。

通过上述过程,模型能够基于标准化初始报告提供的文本信息,结合图像细节对报告进行补充、修正与扩展,从而生成更加准确完整的最终诊断报告。

3 实验结果与分析

3.1 实验设置

实验在三个数据集上进行,包括两个公开胸部X光医学报告生成数据集以及一个私有临床胃镜数据集。印第安纳大学胸部X线数据集(Indiana University Chest X-Ray Collection, IU X-Ray)和重症监护医学信息数据库胸部X线数据集(Medical Information Mart for Intensive Care Chest X-Ray, MIMIC-CXR)用来评估模型在胸部X光医学报告生成任务中的能力,私有临床胃镜数据集用来验证模型在不同医学影像场景下的泛化能力。具体而言,IU X-Ray是医学影像报告生成领域广泛使用的公开基准数据集,包含7 470张胸部X光图像和3 955份放射学报告。在这个数据集中,一个样本通常包含同一病例的正位和侧位胸部X光图像,并对应一个由专家撰写的报告文本。按照常用的数据划分方式,本文将IU X-Ray数据集划分为训练集、验证集和测试集,比例分别为70%、

10%和20%。MIMIC-CXR是目前规模较大的公开胸部X光报告生成数据集,包含368 960张胸部X光图像和222 758份结构化放射学报告。在MIMIC-CXR中样本规模大、疾病类型覆盖广、报告文本形式更加复杂,因此MIMIC-CXR被广泛用作医学报告生成任务的重要基准。为了保证实验结果的公平性,本文遵循推荐的数据划分方式进行实验。私有临床胃镜数据集由国内某知名医院采集,共包含217名患者的数据。每名患者对应一份诊断报告,同时,从胃部八个解剖区域中分别选取一张具有代表性的胃镜图像,最终得到1 736张胃镜图像和217份诊断报告。为了验证小样本数据集评估结果的可靠性,本文采用五折交叉验证进行实验。在实现时,每次实验选取其中一折作为测试集,其余四折作为训练集,共进行五次实验,并计算各项指标的平均值和标准差。

为了评价生成报告和真实报告之间的接近程度,本文采用BLEU^[28]、ROUGE-L^[29]和METEOR^[30]三种常用指标。BLEU主要看生成报告和参考报告之间有多少n-gram片段是重合的,用来反映报告在词语层面的相似程度。ROUGE-L基于最长公共子序列,主要关注生成文本是否覆盖了参考报告中的重要内容。METEOR在unigram匹配的基础上,同时考虑精确率、召回率、词干和同义词匹配的匹配情况,因此能够更全面地反映生成文本与参考文本之间的相似性。通过这些指标,可以从不同角度衡量模型生成报告与真实报告之间的一致程度。

在实验中,对于胸部X光图像,所有图像都被统一调整为224像素×224像素。本文采用预训练的Swin Transformer^[25]作为全局视觉编码器,用于提取整幅图像的全局语义特征;同时,利用CXAS^[26]工具获得胸部X光图像的解剖区域掩码,并采用ResNet18提取区域级视觉特征。文本编码方面,本文采用BioBERT^[31]对完整报告和句子级文本进行编码,分别用来支持粗粒度图文对齐和细粒度句子-区域对齐。在大语言模型设置方面,本文采用Qwen-Max对诊断报告中的句子级解剖区域进行自动标注,构建句子-区域对应关系;在报告生成阶段,IU X-Ray数据集采用Qwen1.5, MIMIC-CXR数据集采用Qwen1.5、LLaMA2。两个数据集均基于相关文本、不相关文本及任务指令生成标准化初始报告,并进一步结合视觉特征对初始报告进行细粒度校正,得到最终诊断报告。对于私有临床胃镜数据集,每个解剖区域图像同样被调整为224像素×224像素,并通过拼接所有区域特征生成整体特征表示。训练设置方面,模型训练与评估在2张NVIDIA L40 GPU(48 GB)上进行。IU X-Ray数据集训练15个epoch, MIMIC-CXR数据集训练5个epoch,私

有临床胃镜数据集训练 10 个 epoch。实验中 batch size 设置为 8, 优化器采用 AdamW, 初始学习率为 1×10^{-4} , 并使用余弦退火策略调整学习率。报告生成阶段采用 beam search 策略, beam size 设置为 3。

3.2 定量结果与分析

为了检验 FRMRG 的有效性, 故本文将 FRMRG 与三类具有代表性的医学报告生成方法做了对比实验, 具体包括: 粗粒度对齐方法 R2Gen^[2]、R2GenCMN^[12]、

PPKED^[3]、DCL^[5]和 PromptMRG^[4]、LLM-RG4^[32]; 细粒度对齐方法 RGRG^[8]、AdaMatch^[18]以及基于大语言模型的方法 R2GenGPT^[23]、BootstrappingLLM^[24]。本文在 MIMIC-CXR 和 IU X-Ray 数据集上对 FRMRG 进行评估, 并根据各对比方法的公开实验结果和适用数据集设置进行比较。对于仅在 MIMIC-CXR 数据集上报告结果, 或依赖特定细粒度标注而难以在 IU X-Ray 上直接复现的方法, 本文仅在 MIMIC-CXR 数据集上进行对比分析, 结果如表 1 所示。

表 1 在 IU X-Ray 和 MIMIC-CXR 数据集上, 将本文方法与当前最先进的医学报告生成方法进行比较

Table 1 Performance comparison of FRMRG with state-of-the-art medical report generation methods on the IU X-Ray and MIMIC-CXR datasets

数据集	方法	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
IU X-Ray	R2Gen ^{†[2]}	0.336	0.200	0.137	0.099	0.313	0.137
	R2GenCMN ^{†[12]}	0.475	0.309	0.222	0.170	0.375	0.191
	PPKED ^[3]	0.483	0.315	0.224	0.168	0.376	0.187
	DCL ^{†[5]}	—	—	—	0.153	0.374	0.186
	PromptMRG ^{†[4]}	0.419	0.252	0.162	0.109	0.313	0.157
	R2GenGPT ^{†[23]}	0.478	0.311	0.225	0.172	0.378	0.211
	BootstrappingLLM ^{†[24]}	0.495	0.323	0.235	0.177	0.380	0.206
	FRMRG	0.509	0.337	0.242	0.185	0.383	0.214
MIMIC-CXR	R2Gen ^[2]	0.353	0.218	0.145	0.103	0.277	0.142
	R2GenCMN ^{†[12]}	0.361	0.220	0.148	0.106	0.278	0.142
	PPKED ^[3]	0.360	0.224	0.149	0.106	0.284	0.149
	DCL ^[5]	—	—	—	0.109	0.284	0.150
	PromptMRG ^{†[4]}	0.377	0.227	0.149	0.104	0.267	0.147
	AdaMatch ^[18]	0.379	0.235	0.154	0.106	0.286	0.163
	LLM-RG4 ^[32]	0.374	—	—	0.124	0.286	—
	R2GenGPT ^{†[23]}	0.362	0.232	0.156	0.110	0.275	0.134
	RGRG ^{†[8]}	0.369	0.245	0.172	0.125	0.261	0.149
	BootstrappingLLM ^{†[24]}	0.386	0.259	0.171	0.120	0.283	0.164
	FRMRG	0.394	0.263	0.175	0.127	0.291	0.169

注: 粗体表示最佳结果, †表示本文根据公开代码复现得到的结果, —表示未测试该指标。

由表 1 可以看出, 本文提出的 FRMRG 在 IU X-Ray 和 MIMIC-CXR 两个数据集的所有评价指标上都取得了最优性能, 说明本文方法能够很好地适应不同复杂度的医学报告生成任务。具体而言, 与粗粒度对齐方法相比, FRMRG 在 IU X-Ray 数据集上表现出稳定优势, 其中 BLEU-1、BLEU-4 和 ROUGE-L 分别提升 2.8%、7.6% 和 1.3%。这说明, 单纯根据整幅图像相似性检索参考报告尚不足, 因为参考报告可能整体相似, 但是所涉解剖区域及病灶描述可以完全不同。因此, FRMRG 在检索粗粒度参考报告之后, 又利用解剖区域视觉特征及区域级候选文本做了细致的局部校正, 由此自然、合理地解决了局部诊断描述不匹配的问题, 所生成的报告也更符合当前图像的实际内容。

与 RGRG^[8]和 AdaMatch^[18]等细粒度对齐方法相

比, 本文方法在 MIMIC-CXR 数据集上取得了更优性能, 其中 BLEU-1、BLEU-3、BLEU-4 和 ROUGE-L 分别提升 4.0%、1.7%、0.8% 和 1.7%。尽管提升幅度相对有限, 但已有细粒度方法通常需要借助解剖区域级标注、区域检测器或额外局部对齐先验等人工标注来建立图文对应关系。相比之下, 本文不依赖额外人工标注, 而是通过结构化标记区域级视觉特征, 并结合细粒度检索校正机制, 对粗粒度检索文本中的局部语义偏差进行修正, 从而在降低标注依赖的同时获得了较好的医学报告生成效果。

与基于大语言模型的 R2GenGPT^[23]和 BootStrappingLLM^[24]相比, 本文方法在 IU X-Ray 数据集上仍表现出十分明确的性能优势。其中, BLEU-1、BLEU-2 和 BLEU-4 分别提升 2.8%、4.3% 和 4.5%。这

一结果说明,大语言模型固然有极强的文本生成能力,但是医学报告生成绝不是单纯生成语句通顺的文字,更根本的要求是所写内容必须与图像表现严格一致。R2GenGPT^[23]主要依赖全局视觉特征,缺少解剖区域约束;BootstrappingLLM^[24]基于整份报告的全局相似性进行检索,难以保证局部病灶与当前影像一致。本文通过粗粒度报告提供整体语义参考,并利用解剖区域级匹配校正局部描述,从而减少不一致信息,增强生成报告与影像的局部语义一致性。

3.3 解剖感知细粒度校正模块有效性验证

为验证解剖感知细粒度校正模块对检索文本的校正效果,本文在IU X-Ray测试集上进行对比实验。实验中,以测试样本的图像特征作为查询,从训练集报告库中检索与当前图像视觉语义相似的历史报告,作为粗粒度检索结果;同时,基于解剖区域视觉表示,从区域句子库中检索与各解剖区域相关的细粒度候选文本。在相同检索库和评价设置下,本文比较了三种文本检索策略:直接采用粗粒度检索报告、基于区域检索候选句直接拼接生成报告,以及本文提出的解剖感知细粒度校正策略,评价时,将三种策略得到的文本结果分别与对应测试样本的人工真实报告进行比较,并采用BLEU-1、BLEU-4、ROUGE-L和METEOR指标衡量其与真实报告之间的一致性。实验结果如表2所示。

表2 IU X-Ray测试集上不同检索结果处理策略与真实报告的一致性比较

Table 2 Comparison of different retrieval-result processing strategies in terms of consistency with ground-truth reports on the IU X-Ray test set

方法	BLEU-1	BLEU-4	ROUGE-L	METEOR
粗粒度检索结果	0.333	0.063	0.266	0.134
直接拼接结果	0.292	0.004	0.093	0.105
细粒度修正结果	0.449	0.115	0.337	0.193

注:粗体表示最佳结果。

由表2可以看出,三种检索结果处理策略与真实报告之间的一致性表现出较大差别。整体来看,直接用粗粒度检索结果可以获得较好的文本一致性,而区域检索直接拼接结果导致性能下降,本文所提出的解剖感知细粒度校正策略取得了最优结果。这说明,检索信息能否有效提升报告质量,不仅取决于检索结果本身,更要看后续对检索文本的判别、筛选与组织方式。具体而言,粗粒度检索结果在BLEU-1、BLEU-4、ROUGE-L和METEOR上分别取得0.333、0.063、0.266和0.134的指标,说明基于全局图像特征检索所得的相似历史报告确实能为当前测试样本提供有价值的整体语义参考。但也毋庸讳言,该策略是以整幅图像的全局相似性为基础做匹配的,故所得检索报告虽在

整体诊断语义上与当前图像相近,但是其中所含的局部病灶描述很可能与当前图像在解剖区域层面不一致,因此其与人工真实报告之间尚有可改进之处;相比之下,区域检索直接拼接结果低于粗粒度检索结果,其BLEU-1和METEOR分别下降12.3%和21.6%。该结果表明,区域检索虽能获得与解剖区域相关的候选句,但是若不作进一步的判别、筛选就直接将候选句拼接成报告文本,就容易引入语义重复或与当前图像无关的局部描述,因而破坏报告的整体结构及语义连贯性;本文提出的解剖感知细粒度校正策略取得了最优结果。与粗粒度检索结果相比,BLEU-1和METEOR分别提升34.8%和44.0%;与区域检索直接拼接策略相比,BLEU-1和METEOR分别提升53.8%和83.8%。这说明,本文方法在保留粗粒度检索所得报告整体语义框架的基础上,巧妙地引入解剖区域视觉表示对候选文本做判别、校正,由此来解决局部区域描述与当前图像不一致的问题,也因而大大提高检索文本与人工真实报告之间的一致性。

3.4 消融实验

为了验证FRMRG中各个组成模块对报告生成性能的实际贡献,本文在IU X-Ray数据集上进行了消融实验。由于本文方法不是单一模块直接生成报告,而是以粗粒度检索结果为基础,依次通过解剖感知细粒度修正、结构化提示驱动的初始报告生成以及多模态最终报告生成逐级优化诊断文本,因此有必要分析各阶段模块是否能够带来有意义的性能增益。具体而言,本文以粗粒度检索结果作为基线,依次加入细粒度修正模块、初始报告生成模块和最终报告生成模块,用来考察不同模块对报告生成质量的影响。实验结果如表3所示。

表3 不同模块对报告生成性能的影响

Table 3 Effects of different modules on report generation performance

方法	BLEU-1	BLEU-4	ROUGE-L	METEOR
粗粒度基线	0.441	0.158	0.364	0.181
+细粒度校正	0.457	0.163	0.369	0.191
+标准化报告生成	0.497	0.180	0.380	0.203
FRMRG	0.509	0.185	0.383	0.214

由表3可以看出,随着各模块的逐步引入,模型在BLEU-1、BLEU-4、ROUGE-L和METEOR指标上整体呈现持续升高的趋势。相对于粗粒度基线,加入细粒度修正模块后,四项指标分别相对提高3.6%、3.2%、1.4%和5.5%。这说明,单纯依靠粗粒度检索得到的相似报告,不能保证局部病灶描述与当前图像的一致性;而细粒度校正模块能合理利用解剖区域信息对局部描述予以校正,来解决粗粒度检索中的局部语义偏差问题。

在引入细粒度校正模块的基础上,进一步加入结构化提示驱动的初始报告生成模块后,BLEU-1、BLEU-4、ROUGE-L和METEOR分别继续提升8.8%、10.4%、3.0%和6.3%。这一结果表明,虽然细粒度修正能很好地缓解局部语义偏差,但是经细粒度修正所得的检索文本仍来自不同报告,因而可能存在表述风格差异、诊断线索分散诸种问题。因此,结构化提示驱动的初始报告生成模块把有效诊断信息组织于统一语义空间中,用固定提示模板予以呈现,由此减少风格干扰,令所生成的初始报告语义更集中、表达更规范,故能提高报告生成质量。

最终,完整模型FRMRG在所有指标上都取得了最优结果。与加入结构化驱动提示之后的初始报告生成模块的结果相比,完整模型在BLEU-1、BLEU-4、ROUGE-L和METEOR上进一步提升2.4%、2.8%、0.8%和5.4%;而与粗粒度基线相比,各个指标的相对提升达到15.4%、17.1%、5.2%和18.2%。这表明,多模态最终报告生成模块能够融合标准化初始报告与图像特征,对初始报告中存在的遗漏信息和偏差信息进行补充与校正,从而提升文本的生成质量。

综合来看,细粒度校正模块主要用来纠正粗粒度

检索中的局部描述偏差,结构化驱动提示的报告生成模块用来组织和规范化检索文本线索,多模态最终报告生成模块则进一步结合图像信息完成最终报告细化。因此各模块彼此衔接、互为补充,共同提升了医学报告生成质量。

为从更多角度评估本文整体方法的有效性与稳健性,并分析关键设计对报告生成结果的影响,本文补充开展了自动标注准确性、候选文本数量、区域信息失效、解剖区域可视化和细粒度校正分析等实验,详细结果与分析见附录A。整体而言,这些实验结果证明了方案设计的合理有效性。

3.5 胃镜数据集上的泛化能力验证

为进一步验证本文方法在不同医学影像报告生成场景中的泛化能力,本文在前述私有临床胃镜数据集上进行实验验证。与胸部X光图像不同,胃镜检查通常以解剖部位为基本采集单元,每个病例包含多个解剖部位对应的白光图像,最终需要将各部位的观察结果整合为完整诊断报告。因此,该任务能够用于检验FRMRG在解剖部位驱动的多图像报告生成场景中的适用性。不同方法在私有临床胃镜数据集上的报告生成结果比较如表4所示。

表4 不同方法在私有临床胃镜数据集上的报告生成结果比较

Table 4 Comparison of report generation performance of different methods on the private clinical gastroscopy dataset

方法	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
R2Gen ^[2]	0.745 0±0.014 5	0.673 0±0.016 5	0.621 0±0.019 9	0.583 0±0.020 8	0.676 0±0.018 5	0.375 0±0.011 5
R2GenCMN ^[12]	0.744 0±0.016 5	0.671 0±0.020 5	0.618 0±0.022 3	0.577 0±0.023 3	0.675 0±0.017 5	0.366 0±0.013 5
R2GenGPT ^[23]	0.669 0±0.004 5	0.609 0±0.007 5	0.563 0±0.010 5	0.528 0±0.012 5	0.681 0±0.014 5	0.365 0±0.006 5
CvT2DistilGPT2 ^[33]	0.658 0±0.037 5	0.602 0±0.030 5	0.559 0±0.027 5	0.522 0±0.025 5	0.692 0±0.027 5	—
FRMRG	0.758 0±0.004 5	0.679 0±0.012 5	0.624 0±0.011 8	0.590 0±0.013 5	0.694 0±0.016 2	0.387 0±0.014 5

注:粗体表示最佳结果;—表示未测试该指标。

由表4可以看出,FRMRG在私有临床胃镜数据集上取得了最优的整体性能,在BLEU-1、BLEU-2、BLEU-3、BLEU-4、ROUGE-L和METEOR六项指标上均优于对比方法。相较于各指标上的次优结果,FRMRG的相对提升幅度分别为1.7%、0.9%、0.5%、1.2%、0.3%和3.2%。相比公开胸部X光数据集,本文方法在胃镜数据集上的提升幅度较小,主要由于该数据集报告模板化程度较高、表达变化有限,传统方法已能较好学习常见模式,从而压缩了不同方法之间的性能差异。R2GenGPT在该数据集上也未明显优于传统方法R2Gen,进一步说明大语言模型的开放式生成能力在高度模板化场景中难以充分转化为文本匹配指标优势。尽管如此,FRMRG仍在各项指标上取得最优结果,该提升可能得益于本文方法对胃镜数据

组织特点的有效利用。胃镜图像通常以解剖部位为基本单元,每幅图像对应一个具体部位的局部观察结果。相比粗粒度检索或直接拼接多部位文本的方式,FRMRG以解剖部位为线索,分别建模不同部位图像与相关诊断文本之间的对应关系,并通过细粒度校正和标准化初始报告生成过程对候选诊断信息进行筛选与整合。该过程能够减少部位间语义混杂和无关描述干扰,从而为最终报告生成提供更准确的参考信息,提升生成报告与真实报告之间的一致性。此外,表4中的标准差结果显示,FRMRG在取得较高平均性能的同时,整体波动处于较小范围内,说明本文方法在胃镜数据集上具有较稳定的生成表现。综合来看,该实验进一步验证了FRMRG在不同医学影像报告生成场景中的泛化能力。

4 结论

本文提出了一种基于细粒度检索增强的医学报告生成框架FRMRG,通过解剖感知细粒度校正模块、结构化提示驱动的初始报告生成模块和多模态最终报告生成模块的协同设计,实现了高质量医学报告生成。在IU X-Ray、MIMIC-CXR以及自建胃部数据集上的实验结果表明,所提方法能够有效提升医学报告生成质量,并具有良好的跨领域泛化能力。

参考文献

- [1] Chen M, Gholamrezanezhad A. Burnout in radiology[J]. *Academic Radiology*, 2023, 30(6): 1031-1032.
- [2] Chen Zhihong, Song Yan, Chang T H, et al. Generating radiology reports via memory-driven transformer[C]// *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. ACL, 2020: 1439-1449.
- [3] Liu Fenglin, Wu Xian, Ge Shen, et al. Exploring and distilling posterior and prior knowledge for radiology report generation[C]// *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 2021: 13748-13757.
- [4] Jin Haibo, Che Haoxuan, Lin Yi, et al. PromptMRG: Diagnosis-driven prompts for medical report generation [C]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, 38(3): 2607-2615.
- [5] Li Mingjie, Lin Bingqian, Chen Zicong, et al. Dynamic graph enhanced contrastive learning for chest X-ray report generation[C]// *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 2023: 3334-3343.
- [6] Wu Ruiqi, Zhang Chenran, Zhang Jianle, et al. MM-retinal: Knowledge-enhanced foundational pretraining with fundus image-text expertise[C]// *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024*. ChamSpringer Nature Switzerland2024: 722-732.
- [7] Liu Fenglin, You Chenyu, Wu Xian, et al. Auto-encoding knowledge graph for unsupervised medical report generation[C]// *Proceedings of NeurIPS*. 2021: 16266-16279. https://proceedings.neurips.cc/paper_files/paper/2021/hash/876e1c59023b1a0e95808168e1a8ff89-Abstract.html.
- [8] Tanida T, Müller P, Kaissis G, et al. Interactive and explainable region-guided radiology report generation[C]// *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 2023: 7433-7442.
- [9] Wang Siyuan, Peng Bo, Liu Yichao, et al. Fine-grained medical vision-language representation learning for radiology report generation[C]// *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: ACL, 2023: 15949-15956.
- [10] 樊琳, 龚勋, 郑岑洋. 基于文本引导下的多模态医学图像分析算法[J]. *电子学报*, 2024, 52(7): 2341-2355.
Fan Lin, Gong Xun, Zheng Cenyang. A multi-modal medical image analysis algorithm based on text guidance [J]. *Acta Electronica Sinica*, 2024, 52(7): 2341-2355. (in Chinese)
- [11] 才华, 易亚希, 付强, 等. 基于跨模态引导和对齐的多模态预训练方法[J]. *电子学报*, 2024, 52(10): 3368-3381.
Cai Hua, Yi Yaxi, Fu Qiang, et al. Multimodal pretraining with cross-modal guidance and alignment[J]. *Acta Electronica Sinica*, 2024, 52(10): 3368-3381. (in Chinese)
- [12] Chen Zhihong, Shen Yaling, Song Yan, et al. Cross-modal memory networks for radiology report generation [C/OL]// *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. 2021: 5904-5914[2026-04-30]. <https://aclanthology.org/2021.acl-long.459/>.
- [13] Yang Shuxin, Wu Xian, Ge Shen, et al. Radiology report generation with a learned knowledge base and multi-modal alignment[J]. *Medical Image Analysis*, 2023, 86: 102798.
- [14] Shentu J, Al Moubayed N. CXR-IRGen: An integrated vision and language model for the generation of clinically accurate chest X-ray image-report pairs[C]// *2024 IEEE/CVF Winter Conference on Applications of Computer Vision*. Piscataway: IEEE, 2024: 5200-5209.
- [15] 周涛, 叶鑫宇, 刘凤珍, 等. 利用跨模态轻量级YOLOv5模型的PET/CT肺部肿瘤检测[J]. *电子与信息学报*, 2024, 46(2): 624-632.
Zhou Tao, Ye Xinyu, Liu Fengzhen, et al. CL-YOLOv5: PET/CT lung cancer detection with cross-modal lightweight YOLOv5 model[J]. *Journal of Electronics & Information Technology*, 2024, 46(2): 624-632. (in Chinese)
- [16] Jing Baoyu, Xie Pengtao, Xing E. On the automatic generation of medical imaging reports[C]// *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg: ACL, 2018: 2577-2586.

- [17] You Jingyi, Li Dongyuan, Okumura M, et al. JPG-Jointly learn to align: automated disease prediction and radiology report generation[C]//Proceedings of ACL. 2022: 5989-6001. <https://aclanthology.org/2022.coling-1.523/>.
- [18] Chen Wenting, Shen Linlin, Lin Jingyang, et al. Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2024: 9494-9509.
- [19] Yin Heng, Wu Wei, Hao Yongtao. DKA-RG: Disease-knowledge-enhanced fine-grained image-text alignment for automatic radiology report generation[J]. Electronics, 2024, 13(16): 3306.
- [20] Singhal K, Tu Tao, Gottweis J, et al. Toward expert-level medical question answering with large language models [J]. Nature Medicine, 2025, 31(3): 943-950.
- [21] Luo Renqian, Sun Liai, Xia Yingce, et al. BioGPT: Generative pre-trained transformer for biomedical text generation and mining[J]. Briefings in Bioinformatics, 2022, 23(6): bbac409.
- [22] 刘青海, 吴钱林, 罗佳, 等. 面向大语言模型的医学图像语义挖掘跨模态哈希[J]. 电子与信息学报, 2026, 48(3): 1167-1179.
- Liu Qinghai, Wu Qianlin, Luo Jia, et al. Cross modal hashing of medical image semantic mining for large language model[J]. Journal of Electronics & Information Technology, 2026, 48(3): 1167-1179. (in Chinese)
- [23] Wang Zhanyu, Liu Lingqiao, Wang Lei, et al. R2GenGPT: Radiology report generation with frozen LLMs[J]. Meta-Radiology, 2023, 1(3): 100033.
- [24] Liu Chang, Tian Yuanhe, Chen Weidong, et al. Bootstrapping large language models for radiology report generation[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(17): 18635-18643.
- [25] Liu Ze, Lin Yutong, Cao Yue, et al. Swin transformer: Hierarchical vision transformer using shifted windows [C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 9992-10002.
- [26] Seibold C, Jaus A, Fink M A, et al. Accurate fine-grained segmentation of human anatomy in radiographs via volumetric pseudo-labeling[EB/OL]. 2023[2026-04-30]. <https://arxiv.org/abs/2306.03934>.
- [27] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.
- [28] Papineni K, Roukos S, Ward T, et al. Bleu: A method for automatic evaluation of machine translation[C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Morristown: ACL, 2002: 311-318.
- [29] Lin C Y. ROUGE: A package for automatic evaluation of summaries[C]//Text Summarization Branches Out. Kerrville: Association for Computational Linguistics 2004: 74-81.
- [30] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments[C]//Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Kerrville: Association for Computational Linguistics 2005: 65-72.
- [31] Lee J, Yoon W, Kim S, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.
- [32] Wang Zhuohao, Sun Yihua, Li Zihan, et al. LLM-RG4: Flexible and factual radiology report generation across diverse input contexts[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2025, 39(8): 8250-8258.
- [33] Nicolson A, Dowling J, Koopman B. Improving chest X-ray report generation by leveraging warm starting[J]. Artificial Intelligence in Medicine, 2023, 144: 102633.

附录 A

具体内容见附件链接 <https://dx.doi.org/10.12263/DZXB.20260621>

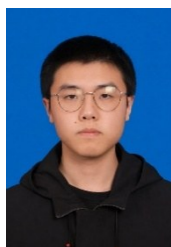
作者简介



牛庆港 男,1997年生,山东临沂人。现为山东大学硕士研究生。主要研究方向为机器学习和计算病理学。
E-mail: qgniu@mail.sdu.edu.cn



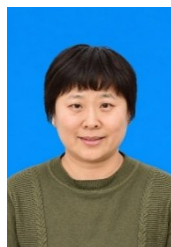
张晓伟 男,1997年生,山东临沂人。现为山东大学博士研究生。主要研究方向为 AI for Science。
Email: zhangxiaowei@mail.sdu.edu.cn



吕千一 男,2002年生,山东临沂人。现为山东大学硕士研究生。主要研究领域为计算病理学、机器学习。
Email: qylv@mail.sdu.edu.cn



李 真 女,1984年生,山东济南人。现为山东大学齐鲁医院主任医师。主要研究方向为胰腺疾病的超声内镜介入诊疗、人工智能在消化内镜质量控制中的应用。
E-mail: qilulizhen@sdu.edu.cn



闫中敏 女,1977年生,山东济宁人。现为山东大学副教授。主要研究方向为 Web 数据集成与智能数据分析。
E-mail: yzm@sdu.edu.cn



王 峻 女,1983年生,重庆人。现为山东大学人工智能国际联合研究院教授。主要研究方向为数据挖掘和生物信息学。
E-mail: kingjun@sdu.edu.cn



郭茂祖 男,1966年生,山东德州人。现为北京建筑大学电气与信息工程学院教授。主要研究方向为机器学习、数据挖掘和生物信息学。
E-mail: guomaozu@bucea.edu.cn



余国先 男,1985年生,湖北孝感人。现为山东大学软件学院教授。主要研究方向为机器学习、数据挖掘和生物信息学。
E-mail: gxyu@sdu.edu.cn