Med-GLIP：利用大规模落地数据集推进医学语言-图像预训练

Ziye Deng*
Zhejiang University
Hangzhou, China
ziye.21@intl.zju.edu.cn

Ruihan He*
Zhejiang University
Hangzhou, China
ruihanhe021@gmail.com

Jiaxiang Liu*
Zhejiang University
Hangzhou, China
jiaxiang.21@intl.zju.edu.cn

Yuan Wang*
Zhejiang University
Hangzhou, China
yuanwang23@zju.edu.cn

Zijie Meng
Zhejiang University
Hangzhou, China
zijie.22@intl.zju.edu.cn

Songtao Jiang
Zhejiang University
Hangzhou, China
songtao.22@intl.zju.edu.cn

Yong Xie
Nanjing University of Posts and Telecommunications
Nanjing, China
yongxie@njupt.edu.cn

Zuozhu Liu
Zhejiang University
Hangzhou, China
zuozhuliu@intl.zju.edu.cn

***Abstract*—医学图像定位旨在将自然语言短语与医学图像中特定区域对齐，是智能诊断、视觉问答（VQA）和自动化报告生成（MRG）的基础任务。然而，现有研究受限于有限的模态覆盖范围、粗粒度标注以及缺乏统一、可推广的定位框架。为应对这些挑战，我们构建了一个大规模医学定位数据集 Med-GLIP-5M，该数据集涵盖七种成像模态，包含超过 530 万条区域级标注，涉及多样的解剖结构和病理发现。该数据集支持分割和定位任务，具有分层的区域标签，从器官级边界到细粒度的病变。在此基础上，我们提出了 Med-GLIP，一种基于 Med-GLIP-5M 训练的模态感知定位框架。Med-GLIP 不用依赖明确设计的专家模块，而是从多样的训练数据中隐式获取分层语义理解，使其能够识别多粒度结构，例如区分肺部与肺炎病变。大量实验表明，Med-GLIP 在多个定位基准测试中持续超越最先进的基线。此外，将其空间输出整合到下游任务中，包括医学 VQA 和报告生成，带来了显著的性能提升。我们的数据集将很快发布。**

***Index Terms*—Medical Dataset, Medical Image Grounding, Grounded Language-Image Pre-training**

随着医学人工智能的快速进展，建立自然语言描述与医学图像中特定区域之间的精确对应已成为智能诊断、手术导航和多模态问答等任务的基础步骤。医学图像定位的目标是根据语言输入定位解剖结构或病理发现，实现跨模态的空间语义对应。例如，当放射科医生说”肿瘤位于右上角”时，模型必须准确识别对应的图像区域。该任务不仅提高了模型的可解释性，还为视觉问答（VQA）和医学报告生成（MRG）等下游应用提供了重要的空间先验，在临床决策支持中发挥着关键作用。然而，较之自然图像领域，医学图像定位面临着几个独特的挑战。首先，公开可用的定位数据集极其稀少，尤其是那些具有大规模、多器官和多模态标注的数据集，这显著限制了该领域的进展。其次，医学图像高度专业化，并在 CT、MRI 和超声等模态之间表现出显著的异质性，无论是空间分辨率、解剖外观还是对比度，这使得跨模态对齐本质上变得复杂。此外，许多目标区域，如小病变或血管分支，通常缺乏清晰的边界，并表现出高度的解剖变异性，对模型的精度和泛化能力提出了更高的要求。

为了应对医学图像语义对齐中的数据稀缺和语义错位问题，先前的工作探索了数据集构建和跨模态建模。在数据集方面，SLAKE [1] 为胸部 X 光片提供了区域-短语注释，这是早期的一个对齐尝试，但只有 6k 样本且模态覆盖率有限。MedTrinity-25M [2] 提供了跨任务的大规模但松散对齐的图文对，其缺乏细粒度的区域级监督。这些数据集普遍存在以下问题：(1) 模态和器官多样性有限；(2) 缺乏密集的区域级注释；(3) 更多关注于分类或问答任务而非空间语义对齐。在建模方面，最近的方法将自然图像跨模态模型适应到医学领域（例如，LLaVA-Med [3] 、MedKLIP [4] 、MedSAM [5] ），通常伴随有结构化的医学知识。LLaVA-Med [3] 引入了多模态大语言模型，但依赖于图像-字幕对而不具备细粒度的对齐。MedKLIP [4] 采用区域-短语对比学习，但仍然特定于 X 光图像。MedSAM [5] 基于 SAM [6] 引入了结构感知先验用于零次器官分割，但依赖提示且缺乏深层语言对齐。OntoRay 更进一步结合放射学本体来捕捉术语和区域之间的因果关系 [7] 。尽管有这些努力，仍没有一个统一的、模态适应性强的对齐框架，能够实现多尺度语义对齐和跨模态泛化，其对下游任务如视觉问答或报告生成的影响尚未得到深入研究。

为了解决这些挑战，我们首先提出了 Med-GLIP-5M，这是一个大规模的医学定位数据集，包括超过 530 万个区域级标注，涵盖七种成像方式，涉及多样的解剖结构和病理发现。我们首先构建了一个综合数据集，包含来自七种成像方式和数十种解剖及病理类别的超过 530 万个区域级标注（Table I ）。该数据集是从多个公共医学分割来源精心整理而来，具有三个关键特性：可分割性、可定位性和多

TABLE I: 医疗图像定位数据集的统一比较。BB = 边界框, 3D-BB = 体积边界框, EP = 极点监督, ROI = 感兴趣区域。✓ / ✗ ：支持 / 不支持。

Dataset	Year	RoI Scale	Annotation Type	Multi-modal	Seg.	Ground	≥ 100K	Coverage / Highlight
VQA-RAD	2018	N/A	None	✗	✗	✗	✗	Common pathologies; QA
SLAKE	2021	642 images	Mask	✗	✓	✗	✗	7 organ categories; QA + segmentation
MS-CXR	2022	1162 images	BB	✗	✗	✓	✗	8 thoracic findings
CT-RAD	2023	10410 studies	3D-BB	✗	✗	✓	✗	Volumetric abnormalities
MedTrinity-25M	2024	–	BB / Mask	✓	✓	✓	✗	> 15 organs, multimodal reports
Med-GLIP-5M	2025	5.3M pairs	BB / Mask	✓	✓	✓	✓	7 modalities, 30+ anatomical regions

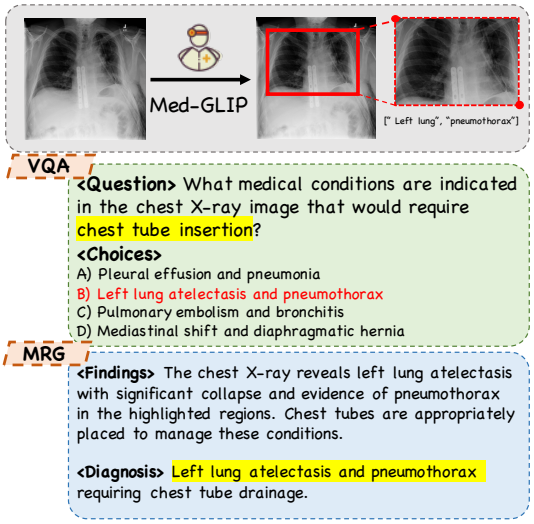

Fig. 1: 用 Med-GLIP 加强 VQA 和 MRG。

粒度性。我们已对数据处理进行了标准化和精细化，以确保其符合数据使用法规和伦理考虑。在此基础上，我们还提出了 Med-GLIP，该方法采用分层专家范式，为每种成像方式训练专门的定位模块，从而实现对细粒度解剖结构的精确定位（Figure 4 ）。此外，我们证明将 Med-GLIP 的定位输出整合到后续任务中，例如医学 VQA 和报告生成，可以显著提高性能，突显了明确的空间定位对多模态医学推理的重要性。总之，我们的贡献如下：

- 据我们所知，我们介绍了迄今为止最大和最多样化的医学定位数据集 Med-GLIP-5M，涵盖 7 种模态、超过 30 个解剖区域和超过 530 万对图像-文本区域组合，解决了该领域长期存在的数据稀缺问题。
- 我们提出了 Med-GLIP，一种模态感知的分层专家框架，通过建模细粒度结构差异，实现对多种医学影像类型的高精度定位。
- 通过广泛的实验，我们验证了空间定位在下游任务中的有效性，展示了 Med-GLIP 在 VQA 和报告生成性能上的增强。我们的工作在数据、建模和应用视角上提供了一个统一的解决方案，推进了通用化医疗视觉语言模型的发展。

I. Med-GLIP-5M 构建

在本研究中，我们汇总了来自多个平台的资源，以策划一个综合的医学图像数据集，包括 GitHub、Grand Challenge、OpenNeuro、SYNAPSE 和 BraTS 等公共库，以及附属医院的合作临床医生提供的专家标注。收集的数据集涵盖了广泛的成像模式、空间分辨率和解剖区域，确保了不同器官和临床任务的多样性，如 Figure 2 所示。我们的数据收集策略中的一个关键原则是尽可能包含具有多个分割掩膜的图像，以实现多标签学习和增强的上下文理解。

在数据获取之后，我们进行了系统的预处理流程。数据集首先按照器官类型分类，并按数据量升序排序。重复的样本和低质量的数据——例如掩码覆盖不明显的图像或损坏的注释——被过滤掉。为了解决标签不一致的问题，语义相似的类别名称标准化为统一的命名方案。对于剩余的干净数据，我们从分割掩码中计算边界框（BBox），并通过包括标签名称、模态类型、数据集大小和掩码图像配对完整性在内的元数据来丰富每个数据集，如 Figure 3 和 Figure 4 所示。

最后，所有处理过的数据集被合并并重新格式化为一个符合 COCO 标准的统一元数据存储库。该整合的数据集作为训练我们 Med-GLIP 模型的基础，使得在异构的临床数据源中进行强大且可扩展的医学视觉语言学习成为可能。

A. 数据收集

我们策划了一个多源医学影像数据集用于视觉-语言对齐。每张图像的选择都偏向于多实例的空间注释——每个图像有多个掩膜或边界框——从而确保生成的数据集非常适合于视觉-语言对齐任务。数据集从多个平台提取，包括但不限于：GitHub、Grand Challenge、OpenNeuro、BraTS、SYNAPSE 和 CodaLab。选择标准：（1）DICOM/NIfTI 可访问性；（2）空间注释（掩膜/边界框）；（3）涵盖多个解剖区域。最终生成了 2720979 张图像，涵盖了多个器官的 X 射线、CT、MRI 等，包括腹部（32.8 % ）、骨骼（23.9 % ）、头部（17.3 % ）、胸部（7.6 % ）等。最终数据集平均每张图像有 4.13 个掩膜，包含各种模态，包括 CT（61.84 % ）、MRI（26.71 % ）、超声波（1.44 % ）、X 光（0.22 % ）、皮肤镜（0.21 % ）、内镜（0.09 % ）、眼底（0.09 % ）。

B. 数据质量控制

采集后实施了一个三层质量控制（QC）管道。首先，所有的 DICOM 图像都使用 `dcm2niix` (v1.0) 和 `--check-mosaic` 标志进行了处理。无法读取或损坏的文件被丢弃。3D 数据集在不同轴上切割成 2D，以提高可用数据的数量。我们将数据集中的图像格式标准化为 PNG，同时保留原始分辨率以保持与源注释的一致性。其次，验证每张图像都拥有对应的分割掩膜。我们实施了特定格式的匹配脚本以确保注释与视觉的一致性。缺少有效关联掩膜的样本不包含在进一步处理之中。第三，过滤不一致的注释：具有结构不完整、未定义标签语义或存在模糊映射的条目被排除。此外，我们删除了标注掩膜仅覆盖图像中极小部分的样本（根据 PASCAL VOC 最小 IOU 标准实验证明为 1.5 % 区域排除），因为这类情况无法为下游任务

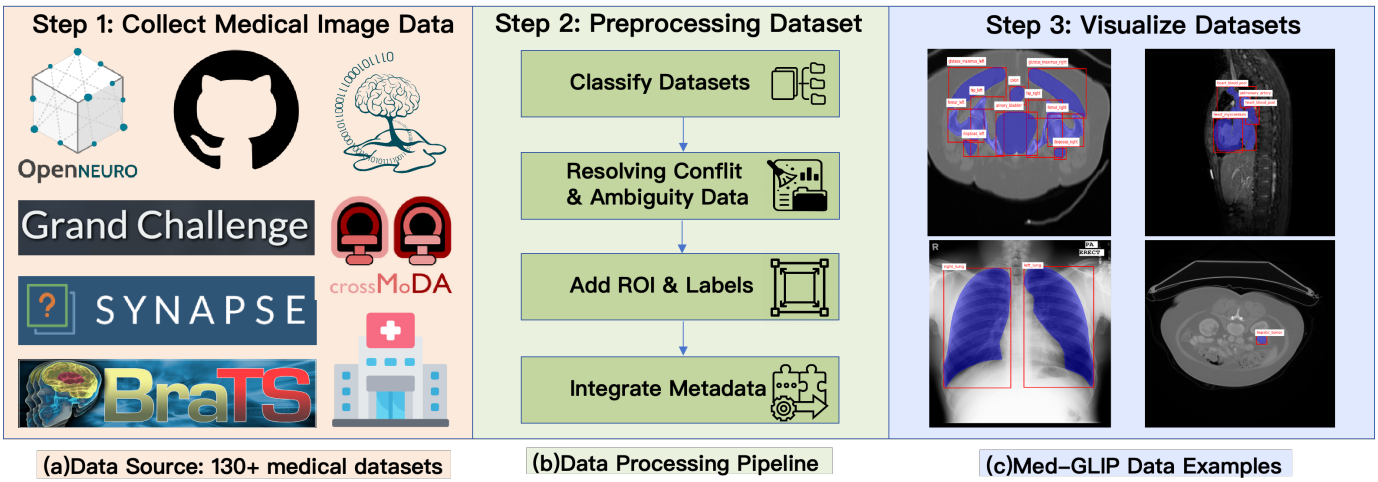

Fig. 2: Med-GLIP-5M 数据集构建过程概述。(a) 数据来源: 我们从包括 OpenNeuro、Grand Challenge 等多平台收集 129 个公共医学数据集。(b) 数据预处理流程: 数据集被分类，消除歧义，加入 ROI 和文本标签，元数据集成到统一格式。(c) 已处理数据的示例可视化: 我们的数据集包含细致的 ROI 标注和跨多种模式的相应文本描述，支持区域级别的定向和结构化理解。

[8] 提供有意义的空间基础。这个质量保证工作流程确保仅保留结构良好、标签明确和空间上有意义的样本用于模型训练和评估。

C. 数据预处理

Med-GLIP-5M 的数据预处理阶段旨在确保数据的高质量，同时遵循伦理标准和监管要求。在获取数据后，数据集经历了多阶段的细化过程。最初，数据根据器官类型和体积进行分类，并删除了重复和低质量样本。通过标准化命名解决了语义标签不一致的问题。然后，通过几何和强度变换来增强多样性，而不影响诊断完整性。图像格式被标准化以确保在机器学习框架中兼容。严格的验证协议验证了图像和分割掩码之间的一致性。此外，整合了可用的诸如患者历史和诊断结果之类的临床元数据。这种结构化的方法生成了一个既大规模又高质量的数据集，为医学人工智能研究提供了坚实的基础。

D. 数据统计

该研究中使用的数据集包含 198 个细粒度的标注标签，这些标签经过人类分析并根据标签语义和相应的数据集文档组织为 38 个更广泛的解剖类别。这一过程由大型语言模型促进并由人类专家验证，旨在增强解剖的连贯性和解释性，以用于下游建模（例如，将特定的髋关节映射到“骨盆（髋部）”，将光学组件映射到“头部（眼部）”）。这 38 个层级类覆盖了 6 个主要人体区域，累计图像数量超过 1100 万张。值得注意的是，脑、脊柱和腹部代表了注释量最高的系统。这种层级标签结构允许在器官级别分割、多器官检测等任务中进行灵活实验，并促进领域适应和多标签学习。

该数据集具有显著的异质性，包含 7 种不同的成像模式。计算机断层扫描（CT）和磁共振成像（MRI）是主要模式，分别贡献了约 420 万和 108 万张图像。这两种模式共同占整个数据集总量的 80 % 以上。CT 扫描在一次分析中被报告在所有模式中占 78 %，涵盖了多个器官系统，包括脊柱、肺、肝和肾。MRI 占类似情境图像的 20 %，特别丰富于高分辨率的脑和神经成像数据，以及前列腺、心脏和血管区域。其余的模式，如超声波，提供了重要的多样性。

这种多模态组合使数据集非常适合于广泛的下游应用，包括跨模态学习、领域自适应、模态感知分割以及在疾病检测和预后中的多模态融合。它还使得研究能够进行统一框架的探索，连接传统上不同的临床领域。

II. Med-GLIP

在我们的方法中，医学目标检测被重新构思为一个短语定位任务，其中每个被识别的图像区域对应于其匹配的医学短语。给定一组与特定成像模态相关的预定义医学概念，例如，对于 X 光图像，“肺炎”、“结节”、“骨折”，构建一个文本提示，例如：

$$\text{Prompt} = \text{“Detect: pneumonia, nodule, fracture”.}$$

按照与 GLIP [9] 类似的方法，我们可以使用预训练的语言模型，例如 BERT [10]，来编码更多语义丰富的提示（例如，“肺炎。结节。骨折。”），这已经展示了经验上的优势。在我们特定模态的定位框架中，计算医学图像区域特征 $\boldsymbol{F}$ 和提示中的编码单词或标记特征 $\boldsymbol{T}$ 之间的对齐分数 S_{ground}。这正式表示为：

$$\boldsymbol{F} = \text{Enc}_I(\text{Img}), \boldsymbol{T} = \text{Enc}_L(\text{Prompt}), S_{\text{ground}} = \sigma(\boldsymbol{F}\boldsymbol{T}^\top) \tag{1}$$

其中 $\boldsymbol{F} \in \mathbb{R}^{N\times d}$ 表示由图像编码器 Enc_I 提取的区域特征，$\boldsymbol{T} \in \mathbb{R}^{M\times d}$ 表示来自语言编码器 Enc_L 的上下文词/标记特征。$\sigma(\cdot)$ 表示用于归一化的 sigmoid 函数。每种成像模态（例如，X 光、CT、MRI）使用专用的图像编码器 $\text{Enc}_{I,\text{modality}}$，而通用语言编码器 Enc_L 是跨模态共享的。每种模态的基础模型通过端到端训练，通过最小化方程 (2) 中定义的损失函数，该方程经过调整，以便将方程 (3) 中的分类 logits S_{cls} 替换为计算的对齐分数 S_{ground}。

$$L_{cal} = L_{cal}^{\text{cls}} + L_{cal}^{\text{loc}}. \tag{2}$$

$$\boldsymbol{F} = \text{Enc}_I(\text{Img}), S_{\text{cls}} = \boldsymbol{F}W^T, L_{cal}^{\text{cls}} = loss(S_{\text{cls}}; T), \tag{3}$$

公式 (2) 和 (3) 概述了一个基线检测模型的损失计算。模型通过图像编码器 Enc_I 提取特征 $\boldsymbol{F}$ ，并使用分类器权重矩阵 $\boldsymbol{W}$ 计算分类逻辑值 S_{cls} 。总训练目标 L_{cal} 是一个由两个组成部分构成的复合损失：(i) 从逻辑值 S_{cls} 和真实目标 $\boldsymbol{T}$ 计算的分类损失 L_{cal}^{cls} ，以及 (ii) 用于边界框回归的定位损失 L_{cal}^{loc} 。

为了解决编码提示中的标记数量（M ）与目标医疗概念数量（c ）之间可能存在的不匹配，我们采用与 GLIP 中类似的扩展策略。对于二元 sigmoid 损失，真实目标矩阵 $\boldsymbol{T} \in \{0,1\}^{N\times c}$ 被扩展为 $\boldsymbol{T}' \in \{0,1\}^{N\times M}$ 。此扩展涉及将构成一个正向医疗短语的所有子词标记对齐为正匹配（即，目标值为 1），同时将任何附加引入的标记（例如，标点符号、填充）指定为负匹配（即，目标值为 0）。在推理阶段，每个医疗短语的概率通过聚合其构成标记的概率来得出，通常通过平均来实现。这允许从文本提示到视觉检测的灵活且稳健的映射。

在每种医学成像模式中，我们的重构确立了对象检测和短语定位之间的理论等效关系。通过在特定模式的数据和提示上训练定位模型，我们可以直接将其以零样本的方式应用于该模式下的检测任务，利用语言提示中编码的丰富语义信息。

III. 实验

A. 实验设置

区域级配准的数据集。为了评估我们 GLIP-experts 的配准性能，我们选择了涵盖五种成像模态的六个数据集：SegRap2023_ct（CT）、AMOS2022_MR（MRI）、sz_cxr（X 射线）、CETUS2014（超声）和 CVC-ClinicDB（内窥镜）。除了 CVC-ClinicDB 以外，所有数据集都来自于 IMed-361M 基准，其中提供了标准化的训练-测试划分（90

MRG 数据集。我们采用了两个广泛使用的基准：MIMIC-CXR 和 IU-Xray。MIMIC-CXR 包含来自 227,835 项研究的 377,110 份胸部 X 光片及相应的自由文本放射学报告。使用 MLRG 框架进行评估以计算自然语言生成（NLG）度量和概念基础分数（例如 CE）。IU-Xray 有 7,470 个图像-报告对，提供了一个较小且更结构化的替代方案，并使用 R2Gen 模型在标准 NLG 度量上进行评估。

Med-VQA 数据集。我们通过三个医学 VQA 数据集评估模型对视觉语言推理的泛化能力：VQA-RAD、SLAKE 和 PathVQA。这些数据集跨越放射和病理学领域，包含英文和中文注释，并提供超过 50,000 对问答对。在评估期间，GLIP 专家支持：（1）通过验证短语与检测区域之间的空间对齐来进行封闭式多项选择推理；（2）通过基础语义推理生成开放式答案。所有评估均遵循官方的训练-测试分割，以确保可重复性和公平比较。

基线和模型。为了评估我们的医学边界框数据集在下游放射报告生成任务中的有效性，我们使用了两个代表性模型进行实验：R2Gen 和 MLRG。R2Gen 是一个用于医学报告生成的基于 Transformer 的模型，使用关系记忆和条件层归一化。它是自然语言生成任务中一个强大且广泛使用的基线。MLRG 是一个最近的最新模型，它利用多视角纵向数据和对比学习。通过整合患者特定的先验知识和先进的多模态信息，它关注临床准确性。对于 Med-VQA 任务，我们采用了 LLaVA-Med [11] ，这是专为生物医学应用设计的领先大型语言和视觉助手。该模型通过有效结合医疗图像的视觉信息与临床知识，在封闭式 & 开放式医疗视觉问答方面表现优异。对于基于内容的任务，我们比较了五种不同的模型：（1）未进行微调的原始 GLIP 模型，(2) 在特定模态子集上进行微调的 GLIP 专家，(3) 在 10 个% 特定模态子集上进行微调的 GLIP 专家，（4）未进行微调的原始 CO_DETR 模型，以及（5）在特定模态子集上进行微调的 CO_DETR 模型。GLIP [9] 是一个视觉语言模型，通过在大规模图像文本对上进行预训练，实现了开放词汇检测和灵活的区域文本对齐，统一了目标检测与基于内容的标记。Co-DETR [12] 建立在 DETR 架构之上，并通过加入语言查询来使用变压器直接链接自由文本与特定图像区域，执行指称表达标记。

实现细节。 所有实验均在配备 8 个 NVIDIA RTX 3090 GPU（每个 24 GB）的 Ubuntu 服务器上进行。对于定位任务，GLIP 和 Co-DETR 在 Med-GLIP 数据集上进行了 30 个周期的微调，使用 Adam 优化器，学习率为 2×10^{-4} ，批处理大小为 5。对于医学报告生成，R2Gen 使用 StepLR 调度进行了 100 个周期的训练，视觉编码器的学习率为 5×10^{-5} ，其他参数的学习率为 1×10^{-4} （批处理大小 =32）。MLRG 使用 AdamW 和 ReduceLROnPlateau 进行了 50 个周期的训练，学习率为 5×10^{-5} ，批处理大小为 6。

评估指标。 为了评估基础模型 GLIP 的精度，我们采用了一种称为平均精度（AP）的指标。AP 是在对象检测和基础任务中广泛使用的评估指标。它通过计算精度-召回曲线下的面积，总结了在不同置信度阈值下精度和召回之间的权衡关系。更高的 AP 表示更好的总体检测或定位性能。

我们采用了两个指标来评估 MRG 结果的质量：自然语言生成（NLG）指标和临床效果（CE）指标。NLG 指标衡量生成报告在语言相似性方面与参考文本的匹配程度。它包括 BLEU-n、METEOR 和 ROUGE-L 等指标。CE 指标评估生成报告是否正确捕捉临床相关发现及其关系。它侧重于医学准确性而不是语言重合度。它包括 F1 得分、精确率（P）和召回率（R）。由于 MLRG 模型的内置评估模块，本项目还采用了 RadGraph F1 和 CheXpert F1。RadGraph F1 衡量报告在多大程度上准确再现临床实体及其关系。CheXpert F1 是通过标记生成的医学报告中十四种胸部疾病的存在与否，然后将这些预测与参考标签进行比较而计算得出的。对于 Med-VQA 任务，我们采用封闭式问题的准确性和开放式问题的召回率，与现有工作如 LLaVA-Med [11] 一致，以进行公平比较。

B. 结果分析 &

各模型的定位性能。 如 **??** 所示，我们首先比较了 GLIP 和 CO-DETR 在所有模态和数据集上的零样本定位性能。两种模型在这种设置下几乎未能达到准确性，这表明它们都无法在没有领域特定微调的情况下有效定位医学实体。这表明了巨大的领域差距，并强调了在医学影像定位任务中进行细化调整的必要性。当我们用 10 % 和 100 % 的训练数据微调 Med-GLIP 时，我们观察到在各模态下的 AP 和 AP50 指标都有明显且一致的提升。例如，在 CT（SegRap）数据集上，AP 从 0.0（零样本）增加到 41.2（10 %）再进一步提升到 82.7（100 %）。在 X-ray（Sz_cxr）、

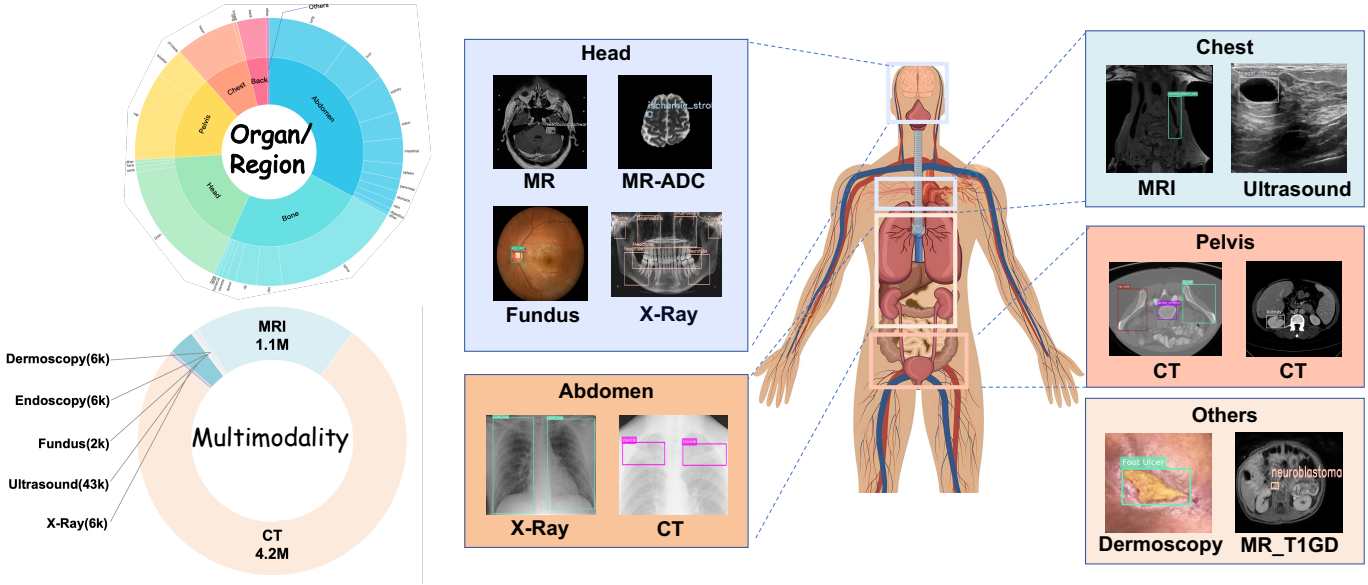

Fig. 3: Med-GLIP-5M 总共有 7 种模态类别，数据集中多个器官包含子器官。

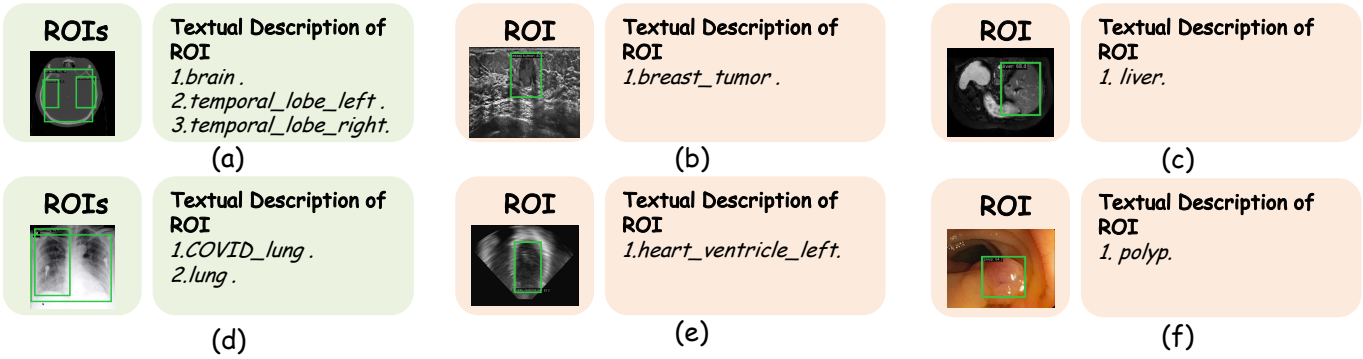

Fig. 4: 跨模态的层级区域级注释示意图。每个子图 (a–f) 展示了在 CT、X 光、超声波、内窥镜和 MRI 图像上的绿色边界框和文本描述。多级框反映了层级语义，为结构化的医学定位提供了细粒度的区域文本监督。

超声波（CETUS2014）、内镜（Clinic-DB）以及其他模态中也观察到类似的趋势，证实了我们的数据集和训练策略在逐步增加注释数据的情况下有效提升模型性能。

经过全面微调后，Med-GLIP 在七种模式中的五种模式下表现优于 CO-DETR，包括 CT、MRI、X 光、超声和皮肤镜。例如，Med-GLIP 在 CT（82.7 对 31.8）、MRI（28.0 对 7.6）、X 光（86.9 对 69.8）、超声（59.2 对 1.0）和皮肤镜（61.7 对 43.0）上获得了显著更高的 AP 得分。虽然 CO-DETR 在内镜和眼底摄影方面的表现略优，但 Med-GLIP 在更广泛的模式范围内表现出稳健且稳定的性能，突显了其作为医学定基任务统一框架的强大泛化能力和有效性。总体而言，这些结果表明，我们的数据集和适应策略有利于各种定基模型，尤其是 Med-GLIP 实现了最佳性能（Figure 5 ）。

下游 MRG 任务的性能如 Figure 6 所示，增强的 R2Gen 模型在 BLEU 和 METEOR 分数上比其基线变体取得了更高的成绩，特别是在 BLEU4 和 METEOR 上，证实了其生成更语义一致报告的能力。然而，由于我们使用的 MIMIC-CXR 数据集缺乏与结构化临床注释的整合，R2Gen 在此设置下不适用 CE 指标（如 Precision、Recall、F1 和 RadF1）。

当增强了 Med-GLIP 后，MLRG 在 CE 指标上表现出了显著的改进。这些结果归因于 MLRG 能够利用由边界框衍生的视觉特征作为其多视角对比学习和对齐机制的一部分。模型编码临床依赖关系和疾病进展路径的能力进一步增强了其事实一致性，这在 CE 指标中得到了反映。下游 Med-VQA 任务的性能。如 Figure 7 所示，Med-GLIP 可以增强下游 Med-VQA 任务的性能。在以准确性衡量的封闭式任务中，Med-GLIP 在所有数据集上均有提升，其中在 VQA-RAD 上的最高提升为 2.52 %。在以 Rouge-L 进行评估的开放式任务中，它也提供了一致的提升，在 VQA-RAD 上达到 1.68 的最大提升。这些结果突出了 Med-GLIP 在改进 Med-VQA 性能方面的有效性。

我们提出了 Med-GLIP，一个统一的、具有模态感知能力的医学图像定位框架，并附有 Med-GLIP-5M——一个大型且多样化的定位数据集。我们的方法有效地弥合了不同成像模式间的语义差距，并显著提升了医学问答和报告

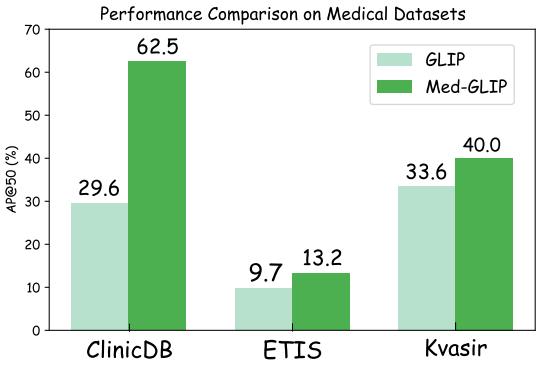

Fig. 5: 在三个医学图像数据集（ClinicDB、ETIS 和 Kvasir）上，GLIP 和 Med-GLIP 在 AP@50 方面的性能对比。Med-GLIP 在所有数据集上均显著优于原始 GLIP 模型，展示了领域适应对于医学视觉定位任务的有效性。

生成等下游任务的表现。大量实验验证了其卓越的定位准确性和泛化能力。Med-GLIP 展示了具有空间定位的大规模预训练在构建可广泛应用的医学视觉-语言模型中的潜力，为更广泛的临床应用和未来与大型语言模型的整合铺平了道路。

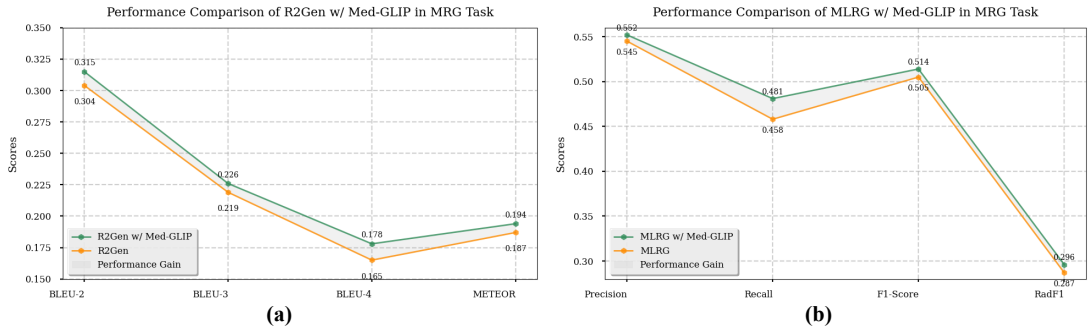

Fig. 6: 在基线 R2Gen 和 MLRG 的 MRG 任务中，带和不带 Med-GLIP 的性能比较。

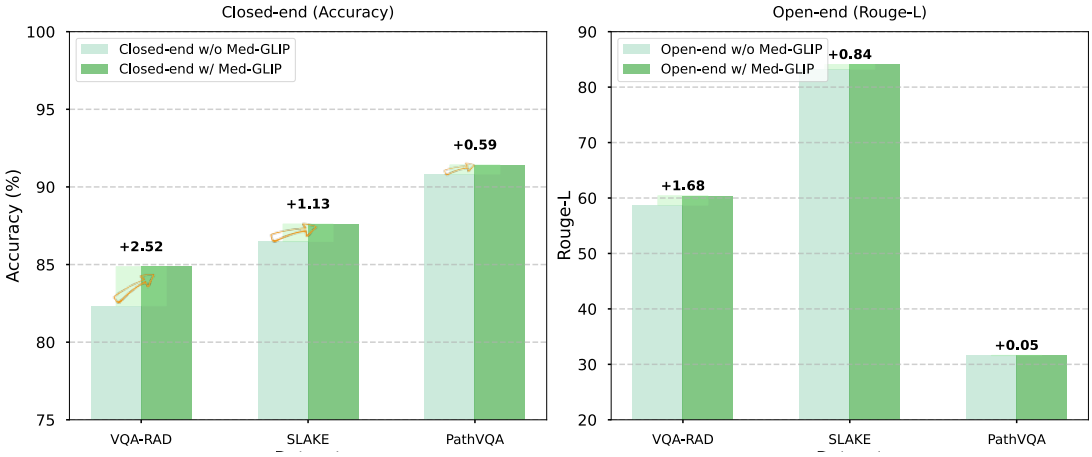

Fig. 7: 在 VQA-RAD、SLAKE 和 PathVQA 数据集上的医学视觉问答中，含有和不含有 Med-GLIP 的性能比较。

References

[1] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu, "Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," in *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*. IEEE, 2021, pp. 1650–1654.

[2] Y. Xie, C. Zhou, L. Gao, J. Wu, X. Li, H.-Y. Zhou, S. Liu, L. Xing, J. Zou, C. Xie *et al.*, "Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine," *arXiv preprint arXiv:2408.02900*, 2024.

[3] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *Advances in Neural Information Processing Systems*, vol. 36, pp. 28 541–28 564, 2023.

[4] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 372–21 383.

[5] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Communications*, vol. 15, no. 1, p. 654, 2024.

[6] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[7] L. L. Chepelev, D. Kwan, C. E. Kahn, R. W. Filice, and K. C. Wang, "Ontologies in the new computational age of radiology: Radlex for semantics and interoperability in imaging workflows," *RadioGraphics*, vol. 43, no. 3, p. e220098, 2023.

[8] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, Jan. 2015.

[9] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang *et al.*, "Grounded language-image pre-training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 965–10 975.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," 2023. [Online]. Available: https://arxiv.org/abs/2306.00890

[12] Z. Zong, G. Song, and Y. Liu, "Detrs with collaborative hybrid assignments training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 6748–6758.