IPLab 智能感知与信息处理实验室

论文分享会

余海军　2025 年 2 月 3日

#多模态模型

#多层对齐

nature biomedical engineering

Article

https://doi.org/10.1038/s41551-025-01574-7

A multimodal vision–language model for generalizable annotation-free pathology localization

Received: 5 March 2024

Accepted: 24 October 2025

Published online: 06 January 2026

Check for updates

Hao Yang[1,2,3,14], Hong-Yu Zhou[4,14], Jiarun Liu[1,2,3], Weijian Huang[1,2,3], Cheng Li[1], Zhihuan Li[5,6], Yuanxu Gao[5], Qiegen Liu[7], Yong Liang[2,8], Qi Yang[9,10], Song Wu[11], Tao Tan[12], Hairong Zheng[1], Kang Zhang[5,6,13] ✉ & Shanshan Wang[1] ✉

Existing deep learning models for defining pathology from clinical imaging data rely on expert annotations and lack generalization capabilities in open clinical environments. Here we present a generalizable vision–language model for Annotation-Free pathology Localization (AFLoc). The core strength of AFLoc is extensive multilevel semantic structure-based contrastive learning, which comprehensively aligns multigranularity medical concepts with abundant image features to adapt to the diverse expressions of pathologies without the reliance on expert image annotations. We conducted primary experiments on a dataset of 220,000 pairs of image–report chest X-ray images and performed validation across 8 external datasets encompassing 34 types of chest pathology. The results demonstrate that AFLoc outperforms state-of-the-art methods in both annotation-free localization and classification tasks. In addition, we assessed the generalizability of AFLoc on other modalities, including histopathology and retinal fundus images. We show that AFLoc exhibits robust generalization capabilities, even surpassing human benchmarks in localizing five different types of pathological image. These results highlight the potential of AFLoc in reducing annotation requirements and its applicability in complex clinical environments.

Accurate diagnosis and precise pathology localization in medical images facilitate customized treatment approaches that improve patient outcomes and mitigate the possibility of diagnostic errors. By pinpointing the exact location and extent of abnormalities, clinicians can make informed decisions that lead to more-targeted therapies and improved prognoses for patients[1–4].

Over the past decade, supervised deep learning methods have accelerated advancements in disease localization[5–8]. However, the efficacy of these methods heavily relies on extensively annotated training datasets, which require domain experts to invest considerable time[9,10]. Specifically, clinical localization tasks often require experienced clinicians to meticulously annotate numerous precise bounding boxes or perform pixel-wise delineations of localized pathology areas. This annotation process is costly, particularly in resource-constrained clinical settings, and algorithms frequently struggle to generalize to diverse datasets.

Several methods have been proposed to reduce the reliance on large annotated datasets[10–13]. Initially, these methodologies acquire general visual representations through self-supervised learning from image datasets, followed by fine tuning on smaller annotated datasets. This approach enables models to achieve high performance on specific tasks while decreasing the need and cost of data labelling[13].

A full list of affiliations appears at the end of the paper. ✉e-mail: kang.zhang@gmail.com; ss.wang@siat.ac.cn

论文基本信息

- 论文题目：
  - multimodal vision–language model for generalizable annotation-free pathology localization Transformer
    - 方法
    - 特色1
    - 特色2
    - 任务
  - 一种用于可泛化、无需标注的病理定位的多模态视觉–语言 Transformer 模型

- 作者信息：中国科学院深圳先进技术研究院
- 期刊分区：
  - 2026 nature biomedical engineer
  - IF = 26.7

Abstract

Existing deep learning models for defining pathology from clinical imaging data rely on expert annotations and lack generalization capabilities in open clinical environments. Here we present a generalizable vision–language model for Annotation-Free pathology Localization (AFLoc). The core strength of AFLoc is extensive multilevel semantic structure-based contrastive learning, which comprehensively aligns multigranularity medical concepts with abundant image features to adapt to the diverse expressions of pathologies without the reliance on expert image annotations. We conducted primary experiments on a dataset of 220,000 pairs of image–report chest X-ray images and performed validation across 8 external datasets encompassing 34 types of chest pathology. The results demonstrate that AFLoc outperforms state-of-the-art methods in both annotation-free localization and classification tasks. In addition, we assessed the generalizability of AFLoc on other modalities, including histopathology and retinal fundus images. We show that AFLoc exhibits robust generalization capabilities, even surpassing human benchmarks in localizing five different types of pathological image. These results highlight the potential of AFLoc in reducing annotation requirements and its applicability in complex clinical environments.

开门见山：现有基于临床影像数据的病理定义深度学习模型高度依赖专家标注，且在开放式临床环境中泛化能力有限。

所提方法：本文提出了一种具有良好泛化能力的视觉－语言模型——无标注病理定位模型（Annotation-Free pathology Localization，AFLoc）。

核心创新：AFLoc 的核心优势在于基于多层级语义结构的对比学习机制，该机制在无需专家影像标注的情况下，将多粒度医学语义概念与丰富的图像特征进行全面对齐，从而适应病理在不同影像中的多样化表达形式。

无标注性能：我们首先在包含 22 万对影像－报告的胸部 X 光数据集上进行了主要实验，并在 8 个外部数据集上进行了验证，覆盖 34 种胸部病理类型。实验结果表明，AFLoc 在无标注病理定位与分类任务中均显著优于当前最先进的方法。

泛化性能：此外，我们还在其他影像模态上评估了 AFLoc 的泛化能力，包括组织病理图像和眼底彩照图像。结果显示，AFLoc 具备稳健的跨模态泛化性能，在五类病理影像的定位任务中甚至超过了人工基准水平。

引言读解

A multimodal vision–language model for generalizable annotation-free pathology localization Transformer

第一段：任务重要性：精确的病理定位

Accurate diagnosis and precise pathology localization in medical images facilitate customized treatment approaches that improve patient outcomes and mitigate the possibility of diagnostic errors. By pinpointing the exact location and extent of abnormalities, clinicians can make informed decisions that lead to more-targeted therapies and improved prognoses for patients[1–4].

（总说）：医学影像中准确的诊断与精确的病理定位有助于制定个性化治疗方案，从而改善患者预后并降低误诊风险。

（具体而言）：通过精确确定异常的具体位置及其范围，临床医生能够做出更加科学的决策，进而实施更具针对性的治疗策略，提高患者的整体预后效果。

引言读解

A multimodal vision–language model for generalizable annotation-free pathology localization Transformer

第二段：痛点：监督深度学习方法严重依赖标注数据

Over the past decade, supervised deep learning methods have accelerated advancements in disease localization[5–8]. However, the efficacy of these methods heavily relies on extensively annotated training datasets, which require domain experts to invest considerable time[9,10]. Specifically, clinical localization tasks often require experienced clinicians to meticulously annotate numerous precise bounding boxes or perform pixel-wise delineations of localized pathology areas. This annotation process is costly, particularly in resource-constrained clinical settings, and algorithms frequently struggle to generalize to diverse datasets.

总说：在过去十年中，监督式深度学习方法显著推动了疾病定位领域的发展。然而，这类方法的有效性在很大程度上依赖于大规模、精细标注的训练数据集，而这些数据需要领域专家投入大量时间与精力来完成。

具体而言：具体而言，临床定位任务通常需要经验丰富的临床医生对大量影像进行精确的边界框标注，或对局部病理区域进行像素级精细勾画。

问题：这一标注过程成本高昂，尤其在资源受限的临床环境中尤为突出，同时相关算法也往往难以在多样化的数据集之间实现良好的泛化性能。

引言读解

A multimodal vision–language model for generalizable annotation-free pathology localization Transformer

第三段：深入痛点：微调/显著性方法 仍然依赖下游任务的标注

Several methods have been proposed to reduce the reliance on large annotated datasets[10–13]. Initially, these methodologies acquire general visual representations through self-supervised learning from image datasets, followed by fine tuning on smaller annotated datasets. This approach enables models to achieve high performance on specific tasks while decreasing the need and cost of data labelling[13].

Moreover, saliency-based methods[14–17] have been developed to reduce annotation costs in pathology localization tasks by allowing coarse localization of target categories in models trained with image-level annotations. However, these methods still require annotations for specific downstream tasks. This requirement is particularly challenging in flexible and dynamic clinical environments, especially for emerging diseases (for example, COVID-19), where deployed models may fail to perform effectively[18–20].

总说：为降低对大规模标注数据集的依赖，已有多种方法被提出。

自监督微调：早期方法通常先通过自监督学习在大规模影像数据集上获取通用视觉表征，随后在规模较小的标注数据集上进行微调。该策略在减少数据标注需求与成本的同时，仍能在特定任务上取得较高性能。

基于显著性算法：此外，基于显著性的定位方法也被用于降低病理定位任务中的标注成本，这类方法允许在仅使用图像级标注训练的模型中，对目标类别进行粗粒度定位。

仍然有不足：然而，这些方法在具体下游任务中仍然需要相应的标注支持。

在灵活且动态变化的临床环境中，这种依赖尤为棘手，尤其是在新发疾病（如 COVID-19）场景下，已部署的模型往往难以有效发挥作用。

引言读解

A multimodal vision–language model for generalizable annotation-free pathology localization Transformer

第四段：深入痛点：现有无监督学习的痛点

In recent years, unsupervised deep learning methods have gained increasing attention due to their independence from annotated datasets, particularly in the field of anomaly detection[21–24]. These methods typically train models using only healthy samples, learning the distribution of normal anatomical structures, which enables the identification of abnormal pathology samples during the testing phase[21,22]. They are particularly effective for data with simple structures and low intersample variance, allowing them to learn normative distributions and achieve excellent anomaly detection performance[23–25]. However, challenges such as the high heterogeneity of pathology images, similarities between different pathologies and large variations in contrast for the same lesions reduce the usability of these methods in complex scenarios, thereby hindering their practical application in real medical environments[25].

总说：无监督深度学习方法因其不依赖标注数据集受到越来越多的关注，尤其是在异常检测领域。

无监督方法介绍：这类方法通常仅使用健康样本来训练模型，从而学习正常解剖结构的分布，并在测试阶段据此识别异常的病理样本。

无监督方法缺陷：它们对于结构简单、样本间差异较小的数据尤为有效，能够学习规范化分布并取得优异的异常检测性能。

然而，病理图像的高度异质性、不同病变之间的相似性，以及同一病灶在对比度上的大幅变化等挑战，降低了这些方法在复杂场景中的可用性，从而阻碍了其在真实医疗环境中的实际应用。

引言读解

A multimodal vision–language model for generalizable annotation-free pathology localization Transformer

第五段：挖坑：VLM的潜力与挑战

A promising approach is the development of medical vision–language pre-training methods[19,26–31]. These methods establish effective correlations between medical reports and medical images, allowing them to flexibly localize disease types not encountered during pre-training without requiring additional customized annotations[19]. However, achieving precise pathology localization solely through the combination of medical images and clinical reports remains challenging. A primary obstacle is the lack of explicit pathology localization markers in clinical reports, which often provide only coarse information such as 'upper' or 'left' to indicate disease location. Moreover, clinical descriptions by clinicians are subjective and variable, further complicating the task of accurately extracting and localizing diseases in medical images. To address this challenge, several methods have been proposed to integrate finer-grained information. For instance, GLoRIA[30] extracts the correlation of the image's regions and paired words in reports to learn global and local representations of images. MedKLIP[19] uses well-defined medical vocabulary knowledge bases to provide supervision at the entity level through triplet training paradigms. However, these fine-grained methods typically focus on individual levels of medical concepts and may overlook the variable meanings of concepts in different contexts. Therefore, these approaches may struggle to adapt to the diverse expressions of disease descriptors in clinical practice, often requiring customized textual cues to enhance localization performance.

总说：一种很有前景的方向是发展医学VLM。这些方法在医学报告与医学图像之间建立有效的关联，使模型无需额外的定制化标注，就能灵活定位在预训练阶段未见过的疾病类型。

然而，仅依靠医学图像与临床报告的结合来实现精准的病灶定位仍然具有挑战性。1）临床报告中缺乏明确的病灶定位标记，报告往往只提供诸如“上部”或“左侧”等较粗略的信息来指示疾病位置。2）临床医生的描述具有主观性且差异较大，进一步增加了从医学图像中准确提取并定位疾病的难度。

解决以上问题的综述：为应对这一问题，已有多种方法尝试融入更细粒度的信息。例如，GLoRIA30 提取图像区域与报告中对应词语之间的相关性，以学习图像的全局与局部表示；MedKLIP19 则利用定义良好的医学词汇知识库，通过三元组训练范式在实体层面提供监督。

挖坑，现有细粒度VLM的缺陷：然而，这些细粒度方法通常只关注医学概念的某一单独层级，可能忽略同一概念在不同语境下语义的可变性。因此，它们在适应临床实践中多样化的疾病描述表达时可能存在困难，并且常常需要定制化的文本提示来提升定位性能。 。

引言读解

A multimodal vision–language model for generalizable annotation-free pathology localization Transformer

第六段：本文方法与性能

In this study, we propose AFLoc, a vision–language model based on contrastive learning aimed at alleviating the need for costly pathology localization annotations. AFLoc can autonomously perform pathology localization and clinical diagnosis with medical images. Unlike traditional global semantic alignment strategies[28,29], AFLoc introduces a contrastive learning framework with a multilevel semantic alignment component, facilitating the comprehensive alignment of medical concepts from reports with image features. Specifically, the image encoder generates three levels of features: shallow local features, deep local features and global features, which are aligned with word-level, sentence-level and report-level features extracted by the text encoder. We extensively validated AFLoc across three types of medical image dataset, including chest X-ray (8 external datasets), histopathology (3 external datasets) and retinal fundus images. Our results show that AFLoc outperforms state-of-the-art methods in localization and clinical diagnostic tasks across different modalities. We hope that this study can help address the challenges posed by annotation scarcity and modality diversity in clinical environments, while providing insights for the design of future clinical open-environment methods.

总说：我们提出了 AFLoc，这是一种基于对比学习的视觉—语言模型，旨在缓解对昂贵病灶定位标注的需求。AFLoc 能够仅凭医学图像自主完成病灶定位与临床诊断。

技术创新：不同于传统的全局语义对齐策略，AFLoc 引入了带有多层级语义对齐组件的对比学习框架，从而促进报告中的医学概念与图像特征之间的全面对齐。

具体而言，图像编码器生成三种层级的特征：浅层局部特征、深层局部特征和全局特征；这些特征分别与文本编码器提取的词级、句级以及报告级特征进行对齐。

性能展示：我们在三类医学图像数据集上对 AFLoc 进行了广泛验证，包括胸部 X 光（8 个外部数据集）、组织病理图像（3 个外部数据集）以及视网膜眼底图像。结果表明，AFLoc 在不同模态的定位与临床诊断任务上均优于当前最先进的方法。

展望：我们希望本研究能够帮助应对临床环境中标注稀缺与模态多样性带来的挑战，并为未来临床开放环境方法的设计提供启示。

Result

我们主要在两项任务上评估了 AFLoc：病灶定位和临床诊断。我们考虑了三种模态的数据，包括胸部 X 光、组织病理图像和视网膜眼底图像。

此外，我们还进行了大量消融实验，以评估不同提示词、文本粒度以及文本编码器对 AFLoc 性能的影响。相关实验将在后续章节中进行详细讨论。

Result：三个模态的无标注定位性能

（作者再次强调）AFLoc 能够在不需要任何标注的情况下执行零样本定位任务。在本节中，我们展示 AFLoc 在不同临床模态上的无标注定位性能。

实验设置：这里我们将“病灶定位”同时定义为热力图预测与二值掩膜结果，因为在某些临床应用中，相比离散的分割结果，热力图提供的可视化信息可能更具价值。我们使用对比度—噪声比（CNR） 对每个预测热力图进行定量评估，并使用交并比（IoU）32 与 Dice 相似系数等指标评估阈值化后的分割掩膜。

1) 指标一：CNR（对比度-噪声比）
用来评估热力图/显著图有没有把病灶“凸显出来”。
直觉：病灶区域应该比背景“更亮/更高”，同时这种差异要**稳定**，别只是噪声波动。

2) IoU（Intersection over Union，交并比）
用来评估二值分割掩膜和真实标注重合得有多好。
直觉：你画出来的病灶区域（预测）和医生标的病灶区域（真值）重叠越多越好，同时别画太多多余区域。

3) Dice（Dice Similarity Coefficient，Dice 相似系数）
Dice 和 IoU 很像，也是衡量重叠，但**更强调重叠像素本身**，常被认为对小目标更“友好”一些。

Result：三个模态的无标注定位性能

1) CNR（Contrast-to-Noise Ratio，对比度-噪声比）

用来评估热力图/显著图有没有把病灶"凸显出来"。

直觉：病灶区域应该比背景"更亮/更高"，同时这种差异要稳定，别只是噪声波动。

常见计算形式

选两块区域：

- **病灶区域**（ROI_lesion）
- **背景区域**（ROI_bg）

记：

- 病灶区域平均值：μ_L
- 背景区域平均值：μ_B
- 噪声（常用背景标准差）：σ_B（有的定义用两者方差合并，但核心思想一致）

常用定义之一：

$$\mathrm{CNR} = \frac{|\mu_L - \mu_B|}{\sigma_B}$$

2) IoU（Intersection over Union，交并比）

用来评估二值分割掩膜和真实标注重合得有多好。

直觉：你画出来的病灶区域（预测）和医生标的病灶区域（真值）重叠越多越好，同时别画太多多余区域。

定义

- 预测掩膜：P
- 真值掩膜：G

交集：$P \cap G$（重叠部分）

并集：$P \cup G$（两者覆盖到的所有区域）

$$\mathrm{IoU} = \frac{|P \cap G|}{|P \cup G|}$$

3) Dice（Dice Similarity Coefficient，Dice 相似系数）

Dice 和 IoU 很像，也是衡量重叠，但更强调重叠像素本身，常被认为对小目标更"友好"一些。

定义

$$\mathrm{Dice} = \frac{2|P \cap G|}{|P| + |G|}$$

数值范围

- 0 到 1
- 1 = 完美重合

Result：三个模态的无标注定位性能

胸部 X 光上的定位性能

（先强调了一下任务的重要性）在临床胸部疾病实践中，医生往往需要对胸部 X 光片进行细致评估，以用于早期筛查和手术规划。例如，气腔样致密影（airspace opacity）可能提示肺部感染、炎症或肿瘤。 肺不张（atelectasis）会导致气体交换受限，引起低氧血症，并严重影响呼吸功能。 然而，由于这些疾病往往边界不清、且与周围组织对比度相近，医生通常需要投入大量精力才能加以区分。因此，自动化诊断系统在辅助诊断方面发挥着关键作用。

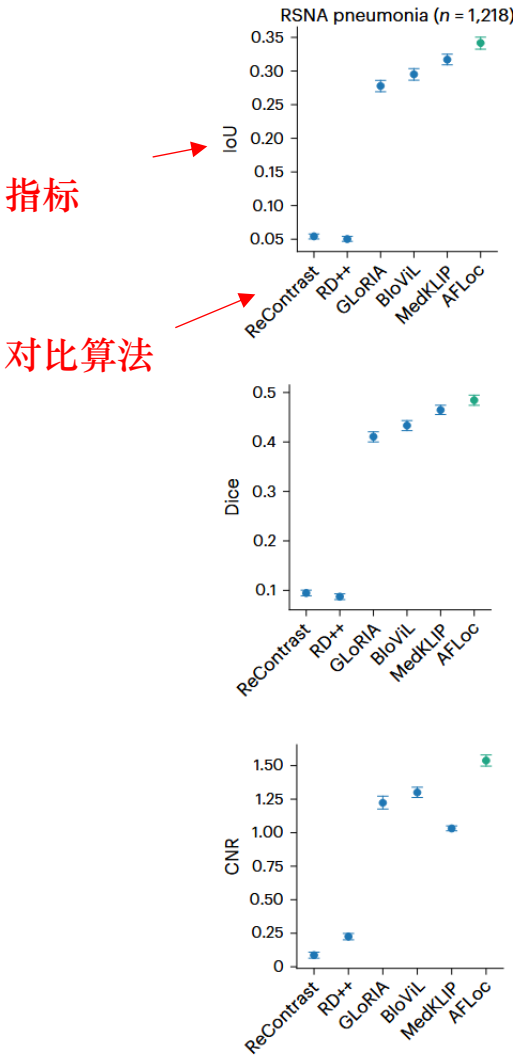
（继续介绍数据集）四个外部数据集评估了 AFLoc 在胸部 X 光上的定位性能：RSNA Pneumonia35、MS-CXR28、CheXlocalize 以及 COVID Rural。

这些数据集涵盖 13 种常见胸部病理：肺炎（pneumonia）、气腔样致密影（airspace opacity）、肺不张（atelectasis）、心脏增大（cardiomegaly）、实变（consolidation）、水肿（oedema）、纵隔/心纵隔增宽（enlarged cardiomediastinum）、肺部病灶（lung lesion）、胸腔积液（pleural effusion）、气胸（pneumothorax）、肺部致密影（lung opacity）、支持装置（support devices）以及 COVID-19。

Result：三个模态的无标注定位性能

胸部 X 光上的定位性能

◆ RSNA Pneumonia 数据集上AFLoc 的定位表现



- 在三项评估指标上，AFLoc 均优于所有对比方法。
  - 具体而言，与各指标上表现最佳的对比方法相比：

  AFLoc 将 IoU 提升 7.9%，
  Dice 系数提升 4.1%，
  CNR 提升 18.3%，
- 表明 AFLoc 具有更强的病灶区域定位能力。

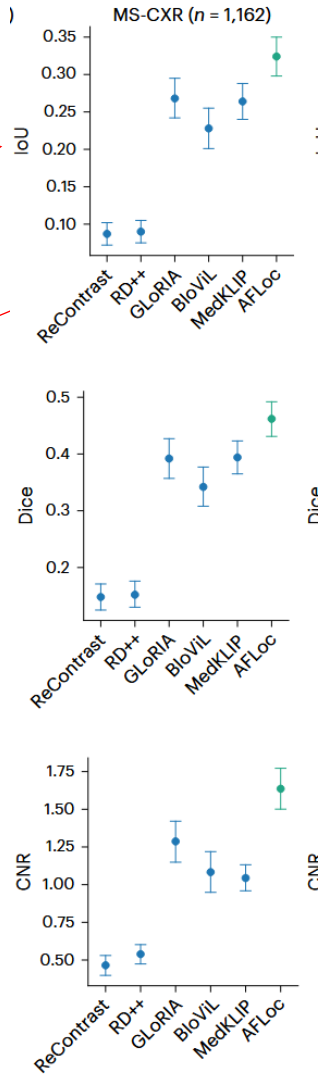
Result：三个模态的无标注定位性能

胸部 X 光上的定位性能

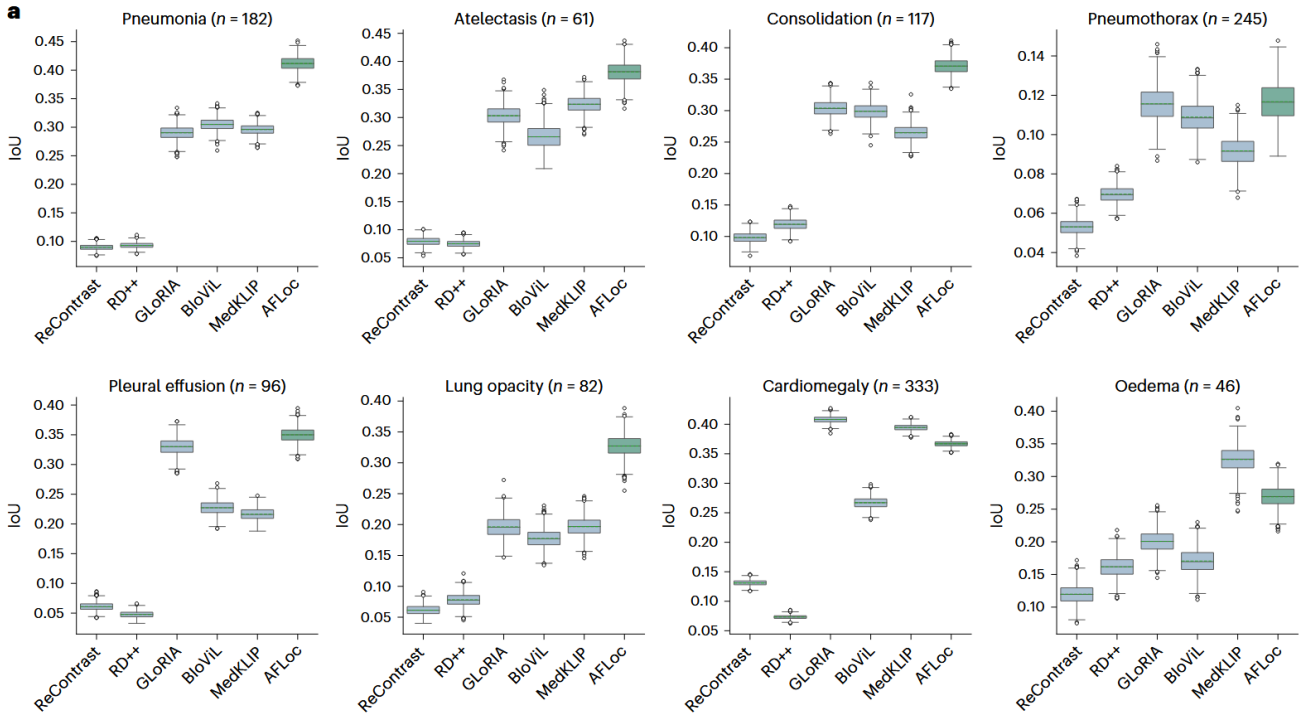
◆ MS-CXR 数据集上AFLoc 的定位表现

指标

对比算法



➢ AFLoc 明显优于现有的视觉—语言预训练方法

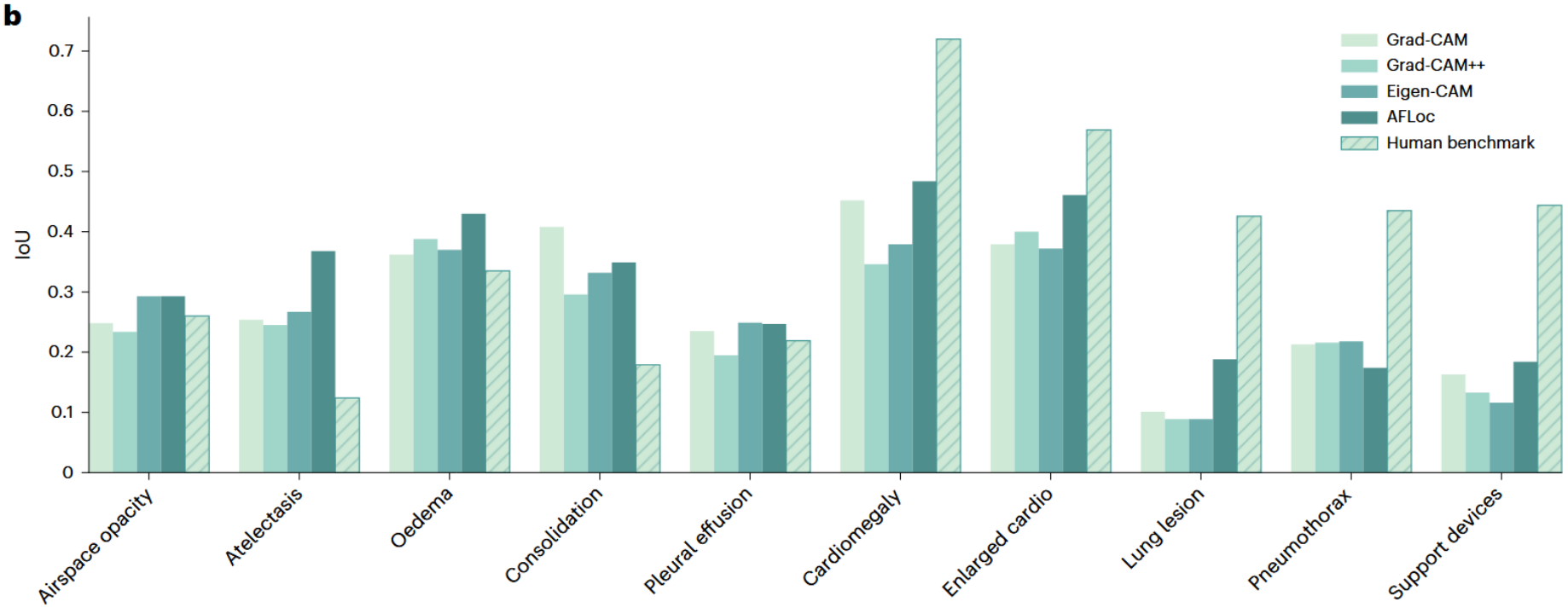


➢ AFLoc 在多数病理（8 种中的 6 种）上都表现出一致的提升

Result：三个模态的无标注定位性能

胸部 X 光上的定位性能

◆ CheXlocalize 数据集上AFLoc 的定位表现

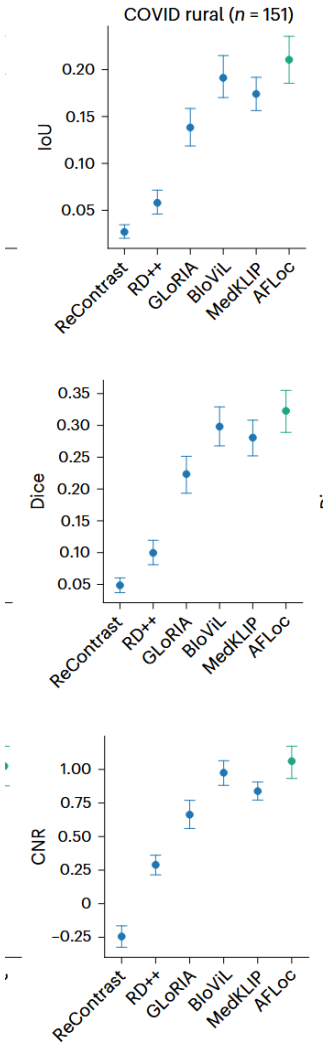


我们将 AFLoc 与基于显著性的方法（Grad-CAM、Grad-CAM++、Eigen-CAM）以及文献报告的人类基线进行了对比。在大多数情况下，AFLoc 的表现优于这些方法，其平均 IoU 比 Grad-CAM 高 12.8%（0.318 对 0.282）。AFLoc 在定位气腔样致密影和胸腔积液等病理时也超过了人类基线，显示出其在医学图像分析中的潜力。

Result：三个模态的无标注定位性能

胸部 X 光上的定位性能

◆ COVID Rural 数据集上AFLoc 的定位表现 (未见疾病泛化能力)



- 展示了 AFLoc 对未见疾病的泛化能力。AFLoc 达到 0.211 的 IoU（95% CI：0.185－0.236）
- 与所有对比模型相比，AFLoc 也取得了更高的 Dice 系数（0.323，95% CI：0.289－0.355）和 CNR（1.062，95% CI：0.929－1.173）。这些结果凸显了 AFLoc 在面对新发情况等真实临床场景中的应用潜力。

Result：三个模态的无标注定位性能

视网膜眼底图像上的定位性能

- 介绍数据集：为了进一步验证我们提出的 AFLoc 的泛化能力，我们将其应用于视网膜眼底图像。我们构建了一个包含三种常见视网膜病变的眼底图像数据集：
  - **脉络膜新生血管（choroidal neovascularization, CNV）、**
  - **玻璃膜疣（drusen）**
  - **视网膜内出血（intraretinal haemorrhages）。**

- 介绍三个任务的重要性：检测 CNV 对于湿性年龄相关性黄斑变性（AMD）的早期诊断至关重要。Drusen 是干性 AMD 的重要指征。视网膜内出血则可能反映糖尿病视网膜病变或高血压性视网膜病变等严重疾病。对这些病变的早期检测和识别能够显著减缓疾病进展，并有助于实施更积极的干预措施。

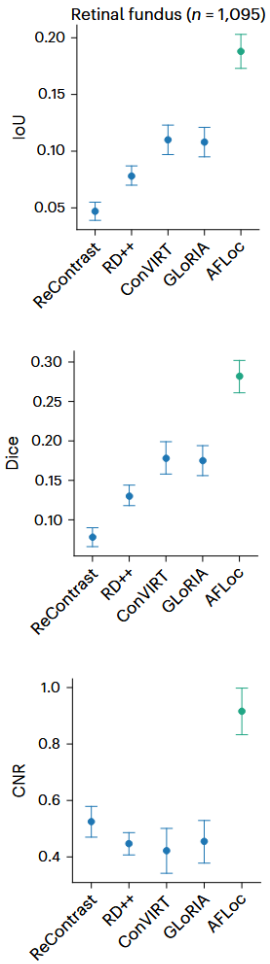
Result：三个模态的无标注定位性能

视网膜眼底图像上的定位性能

◆ Retinal fundus 数据集上AFLoc 的定位表现

指标

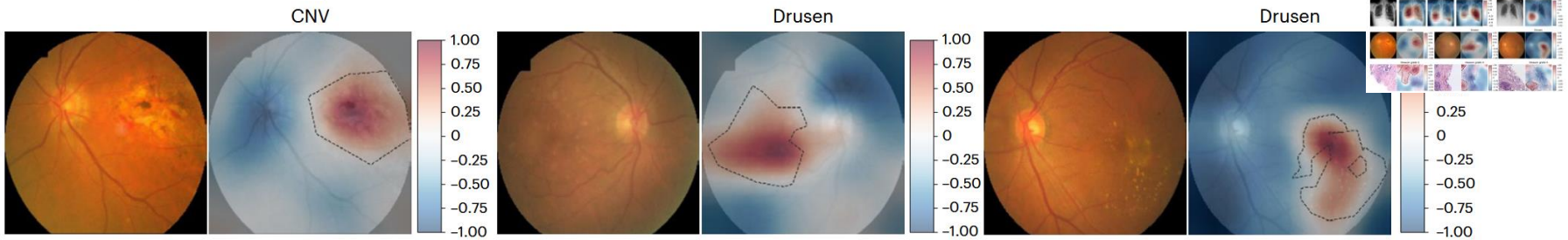
对比算法



- 值得注意的是，AFLoc 在不同病变上均取得了最高的 IoU、Dice 和 CNR 指标。 表明该方法在定位细微特征方面具有良好的能力。

- 尽管 AFLoc 在所有方法中取得了最优结果，但其在视网膜内出血（intraretinal haemorrhages）的定位能力相较于另外两种病变仍然较弱，其 IoU 为 0.097（95% CI: 0.091, 0.102）。这表明在定位弥散性且边界不明显的病灶方面仍有进一步改进的必要。

- 可视化结果：其中颜色越深的红色区域表示模型预测的病变位置，而黑色虚线框表示临床医生的标注区域。AFLoc 能够准确定位诸如 CNV 和 drusen 等较为集中分布的病变。
- Tip:但是作者没有给出视网膜内出血的可视化结果，因为作者提前说明了在视网膜出血定位不佳。



Result：三个模态的无标注定位性能

组织病例学的定位性能

- 介绍数据集：本研究中，我们使用 Quilt-1M 数据集训练 AFLoc 模型用于组织病理学任务，并在 **SICAPv2 测试集**（n = 2,122）上，在无标注定位（annotation-free localization）设置下评估其性能

- 介绍组织病理学任务的重要性：在组织病理学中，准确定位异常组织对于癌症的诊断和分级至关重要。例如，在前列腺癌中，Gleason 分级系统通过观察腺体结构、细胞核特征以及腔结构形成等组织学特征来评估组织样本。病理学家根据组织的分化程度赋予 Gleason 评分，以反映癌症的侵袭性。对这些病变进行早期检测和识别能够显著减缓疾病进展，并有助于实施预防性干预。

- 我们将 AFLoc 与多种先进方法进行了比较，包括 CLIP、BiomedCLIP 以及三个专门针对组织病理学的基础模型：PLIP、QuiltNet 和 CONCH。具体而言，对于所有对比方法，我们均采用 GradCAM 进行可视化定位，因为已有研究表明该方法能够提升其定位性能。

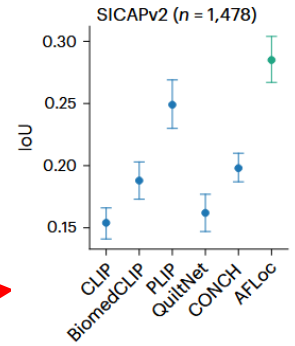
Result：三个模态的无标注定位性能

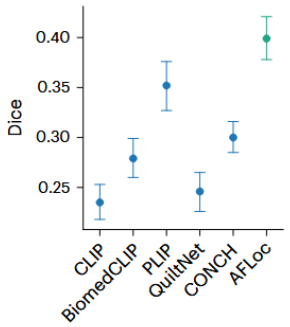
组织病理学的定位性能

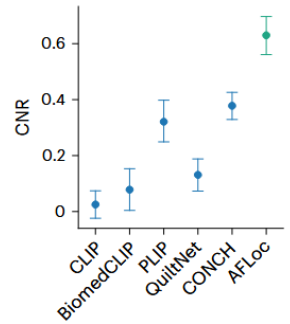
◆ Retinal fundus 数据集上AFLoc 的定位表现

指标

对比算法







- 得AFLoc 在三项评估指标上均取得了最佳的定位性能。具体而言，与表现最好的对比方法 PLIP 相比，各个指标分别提升了xxxxx（列数据）。
- 这些结果表明，与通用模型和专门的组织病理学模型相比，AFLoc 能够更准确地定位异常组织。
- 尽管结果令人鼓舞，但需要指出的是，**由于组织形态的复杂性和高度变异性，组织病理学定位仍然是一项极具挑战性的任务**。尽管如此，AFLoc 在所有指标上的优异表现表明，其作为一种无需标注的组织病理学定位工具具有良好的鲁棒性和跨场景泛化潜力。（强调自己的方法在很难的问题上，无需标注都能取得很好的性能！）

Result：两种提示的消融实验

在上一节中，我们通过为三种影像模态设计特定提示词，验证了 AFLoc 在无需标注情况下进行定位（annotation-free localization） 的能力。在临床环境中，提示词的生成通常有两种方式：
(1) 基于不同影像模态的通用规则自动生成提示词；
(2) 为了提高定位精度而人工设计的提示词，例如 MS-CXR 数据集 中提供的提示词。

这就引出了一个问题：定制化提示词在多大程度上可以提升 AFLoc 的性能?

简单提示（Simple prompt）

只说 **疾病是什么。**

例子：

- *findings suggesting pneumonia*
  → “提示肺炎的影像学表现”

精确提示（Precise prompt）

直接描述 **影像上具体看到什么异常。**

例子：

- *severe bibasilar consolidation*
  → 双肺底严重实变
- *airspace opacity in a right infrahilar location*
  → 右肺门下区气腔不透明影

Result：两种提示的消融实验

Table 1 | Comparisons of localization performance on the MS-CXR dataset with different descriptive granularities

Description	BioViL	GLoRIA	AFLoc
IoU			
Simple description	0.187 (0.162, 0.213)	0.240 (0.214, 0.265)	**0.289 (0.262, 0.314)**
Precise description	0.228 (0.201, 0.255)	0.268 (0.242, 0.295)	**0.324 (0.298, 0.350)**
Dice			
Simple description	0.287 (0.252, 0.321)	0.357 (0.322, 0.392)	**0.418 (0.383, 0.450)**
Precise description	0.342 (0.308, 0.377)	0.392 (0.357, 0.427)	**0.462 (0.431, 0.492)**
CNR			
Simple description	0.825 (0.694, 0.948)	1.128 (1.003, 1.254)	**1.351 (1.219, 1.481)**
Precise description	1.083 (0.949, 1.219)	1.287 (1.149, 1.421)	**1.636 (1.501, 1.772)**

Numbers within parentheses indicate 95% CI. Bold values represent the highest performance score among the compared methods.

- 当使用精确提示词时，所有模型——包括我们的 AFLoc 以及另外两种对比方法（BioViL 和 GLoRIA）——在定位性能上都得到提升。
- 对于 AFLoc，使用精确提示词使 IoU 提高了 12.1%、Dice 提高了 10.5%、以及 CNR 提高了 21.1%。这些改进表明，**更详细且具有临床相关性的提示词能够显著提升模型的定位能力**。

Result：三个模态的无标注诊断性能

组织病理学的定位性能

在本节中，我们通过**零样本分类（zero-shot classification）任务，在胸部 X 光（CXR）、视网膜眼底图像和组织病理学**数据集上展示 AFLoc 的**无标注诊断能力**。

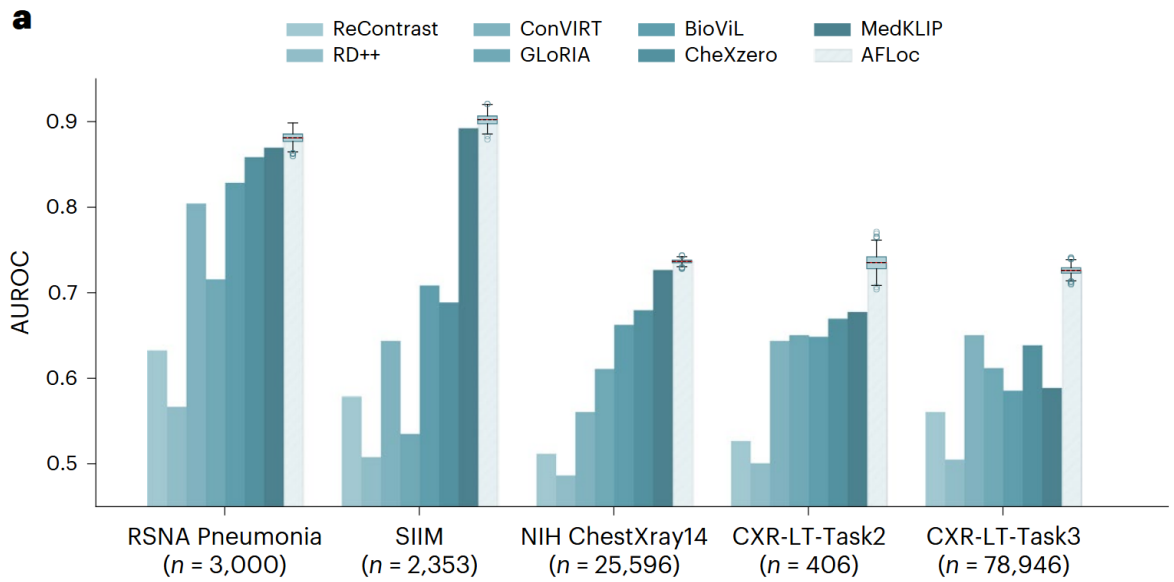
作为对比，我们还报告了两种**无监督异常检测方法**——ReContrast 和 RD++ 的结果。

由于这些方法在评估过程中**未使用带标签的数据进行微调**，因此为方便起见，我们将其统称为 **“零样本（zero-shot）方法”**。

Result：三个模态的无标注诊断性能

组织病理学的分类性能

（介绍数据集）：在我们的研究中，我们在 RSNA Pneumonia、SIIM、NIH ChestXray14 和 CXR-LT 数据集上进行了 零样本（zero-shot）分类任务。



- 与视觉—语言多模态预训练方法相比，基于显著性（saliency-based）的方法（如 ReContrast 和 RD++）得到的分数较低。这表明 引入包含丰富专家知识的文本报告 具有重要意义。
- MedKLIP 在该任务中表现突出，这可能得益于其对 文本报告进行独特的语义简化处理，从而有助于模型有效学习与分类相关的信息。
- 然而，我们的 AFLoc 在四个数据集上取得了最高的受试者工作特征曲线下面积（AUROC，roc曲线下面积）。
- AFLoc 能够有效地对齐多模态信息，通过结合视觉特征与文本特征来捕捉复杂的分类模式。

Result：三个模态的无标注诊断性能

视网膜眼底图像的分类性能

（介绍数据集）：我们进一步将 AFLoc 应用于 视网膜眼底图像，评估涵盖了 九种视网膜疾病：黄斑变性（MD）、视网膜病变（retinopathy）、近视（myopia）、青光眼（glaucoma）、先天性视盘异常（CODA）、视网膜动脉硬化（RAS）、白内障（cataract）、黄斑前膜（MEM） 和 黄斑病变（ML）。

Table 2 | Comparisons of AUROC on retinal fundus datasets for the zero-shot classification task

Methods	Label-free	MD	Retinopathy	Myopia	Glaucoma	CODA	RAS	Cataract	MEM	ML	Mean
ReContrast	✓	0.423	0.689	0.921	0.693	0.468	0.408	0.341	0.546	0.812	0.589
RD++	✓	0.569	0.518	0.291	0.431	0.634	0.671	0.819	0.569	0.176	0.520
ConVIRT	✓	0.421	0.759	0.870	0.848	0.854	0.466	0.424	0.862	0.698	0.689
GLoRIA	✓	0.914	0.592	0.774	0.926	0.803	0.636	0.610	0.834	0.864	0.772
AFLoc	✓	**0.978**	**0.837**	0.962	**0.939**	0.899	0.644	0.988	**0.978**	**0.946**	**0.908**
RETFound	✗	0.935	0.758	**0.971**	0.918	**0.900**	**0.901**	**0.995**	0.908	0.883	**0.908**

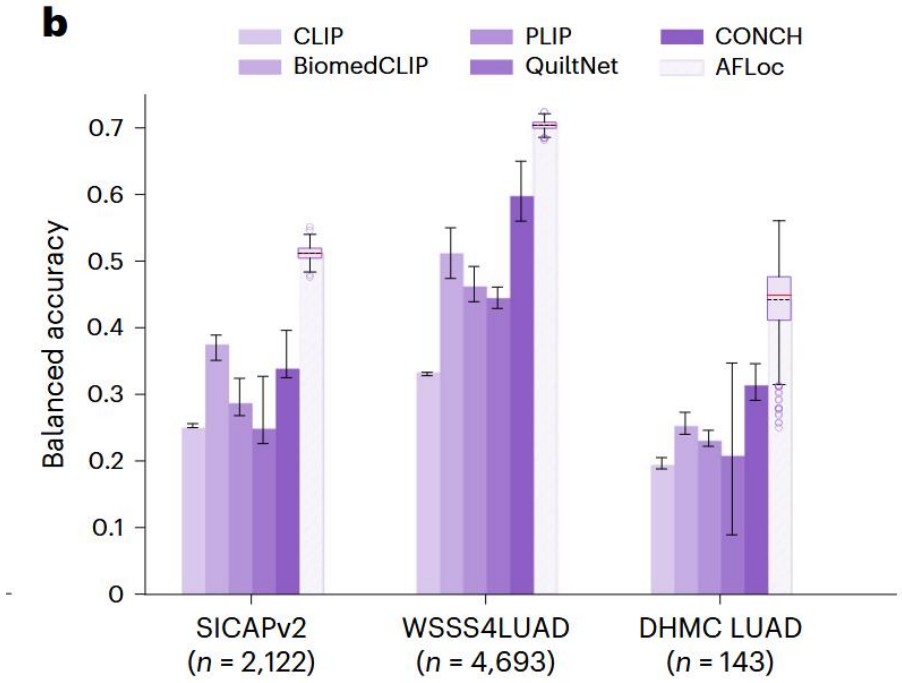
RETFound uses human-annotated classification labels during fine tuning. Bold values represent the highest performance score among the compared methods.

- 虽然 **无监督异常检测方法**（ReContrast 和 RD++）在某些特定疾病上表现较好，例如 **近视**（0.921 AUROC） 和 **白内障**（0.819 AUROC），但它们的**平均得分**（分别为 0.589 和 0.520）仍低于 **视觉—语言预训练方法**。
- 在所有 无需标签 的方法 中，AFLoc 在所有疾病上都取得了最高的成绩，平均 AUROC 达到 0.908，远高于排名第二的方法 GLoRIA（0.772）。此外，AFLoc 的性能与 RETFound 相当，而 RETFound 是使用带标签数据进行微调（fine-tuning）得到的模型。
- AFLoc 具有卓越的诊断能力，并且在多种不同疾病之间具备良好的泛化潜力。

Result：三个模态的无标注诊断性能

组织病理学的分类性能

（介绍数据集）：对于许多疾病而言，组织病理图像的检查仍然是诊断的金标准。下面展示了在 SICAPv2、WSSS4LUAD 和 DHMC LUAD 数据集上，不同方法在 零样本分类任务（zero-shot classification） 中的 平衡准确率（balanced accuracy） 对比。



为什么指标使用平衡准确率，而不是AUROC?

答：之前的胸部X片和眼底图像都是二分类任务，但是病理往往有多种分类时一个多分类任务。且病例类别往往面临数据不平衡的问题，因此需要使用平衡准确率。

Balanced Accuracy= $(Recall_1 + Recall_2 + ... + Recall_n) / n$

- 在所有对比方法中，AFLoc 在所有数据集上均表现出更优性能，分别在 SICAPv2、WSSS4LUAD 和 DHMC LUAD 数据集上取得了 最高的平衡准确率：0.512、0.704 和 0.442。
- AFLoc 持续稳定的性能提升可以归因于其强大的多模态语义对齐能力，该能力能够有效地同时捕捉 组织病理图像中的视觉特征 以及 文本描述中的语义上下文信息。

Result：通过有限定位标注 提升性能

SIIM 数据集 分割性能

为了展示 AFLoc 在临床应用中的更强实用性，我们给出了在 定位任务中利用少量标注数据进行模型微调（fine-tuning） 所得到的结果。

Supplementary Tables 7: Comparisons of Dice scores with state-of-the-art methods on the fine-tuning segmentation task. Performance variations are reported for different data amounts using 1%, 10%, and 100% of the data.

Data Portion	ConVIRT	GLoRIA	BioViL	MedKLIP	AFLoc
1%	0.541	0.567	0.627	0.666	**0.772**
10%	0.612	0.578	0.700	0.721	**0.781**
100%	0.735	0.769	0.785	0.794	**0.809**

Bold values represent the highest performance score among the compared methods.

- SIIM 数据集 上，随着用于微调的标注数据增加，AFLoc 的 分割性能持续提升。
- 在不同规模的标注数据比例下，AFLoc 始终取得高于所有对比方法的 Dice 得分。这些结果表明，AFLoc 能够有效利用有限的标注数据来提升分割性能。

Result：通过有限定位标注 提升性能

ChestXray14 数据集定位性能

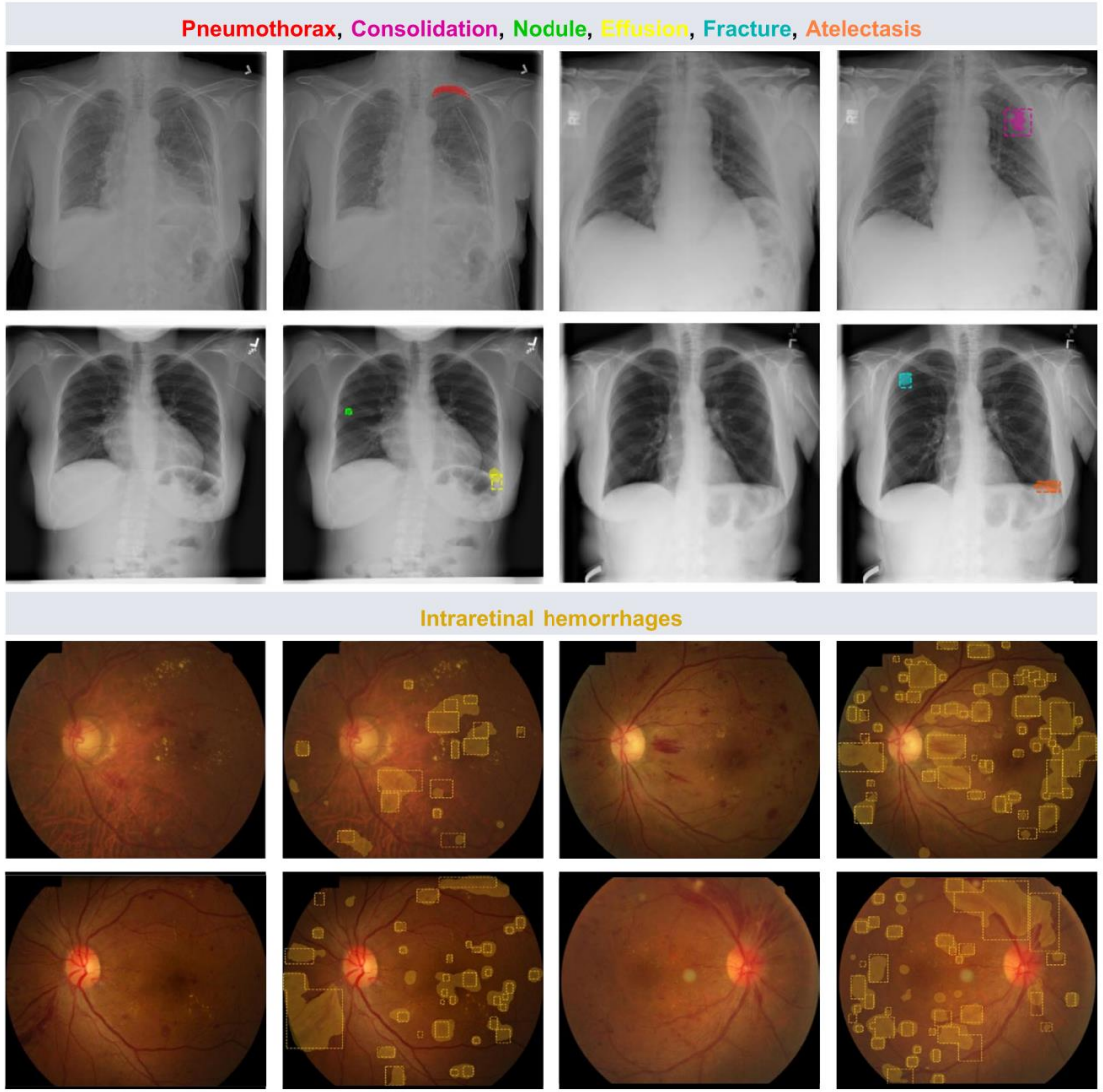
Supplementary Tables 8: Quantitative results of disease localization accuracy at various T(IoU) thresholds for eight pathologies in the NIH ChestXray14 dataset. Compared results were obtained from [55].

T(IoU)	Method	Atelectasis	Cardiomegaly	Effusion	Infiltration	Mass	Nodule	Pneumonia	Pneumothorax	Mean
0.1	Wang et al. [54]	0.69	0.94	0.66	0.71	0.40	0.14	0.63	0.38	0.57
	Li et al. [55]	0.71	0.98	0.87	0.92	0.71	0.40	0.60	0.63	0.73
	AFLoc	0.72	1.00	0.86	0.91	0.78	0.68	0.90	0.82	**0.83**
0.2	Wang et al. [54]	0.47	0.68	0.45	0.48	0.26	0.05	0.35	0.23	0.37
	Li et al. [55]	0.53	0.97	0.76	0.83	0.59	0.29	0.50	0.51	0.62
	AFLoc	0.59	1.00	0.69	0.86	0.66	0.57	0.78	0.65	**0.73**
0.3	Wang et al. [54]	0.24	0.46	0.30	0.28	0.15	0.04	0.17	0.13	0.22
	Li et al. [55]	0.36	0.94	0.56	0.66	0.45	0.17	0.39	0.44	0.49
	AFLoc	0.43	1.00	0.49	0.72	0.57	0.52	0.73	0.45	**0.62**
0.4	Wang et al. [54]	0.09	0.28	0.20	0.12	0.07	0.01	0.08	0.07	0.12
	Li et al. [55]	0.25	0.88	0.37	0.50	0.33	0.11	0.26	0.29	0.42
	AFLoc	0.30	1.00	0.26	0.52	0.46	0.34	0.57	0.31	**0.47**
0.5	Wang et al. [54]	0.05	0.18	0.11	0.07	0.01	0.01	0.03	0.03	0.06
	Li et al. [55]	0.14	0.84	0.22	0.30	0.22	0.07	0.17	0.19	0.27
	AFLoc	0.19	0.99	0.12	0.32	0.33	0.25	0.38	0.21	**0.35**
0.6	Wang et al. [54]	0.02	0.08	0.05	0.02	0.00	0.01	0.02	0.03	0.03
	Li et al. [55]	0.07	0.73	0.15	0.18	0.16	0.03	0.10	0.12	0.19
	AFLoc	0.09	0.94	0.05	0.21	0.22	0.19	0.20	0.11	**0.25**
0.7	Wang et al. [54]	0.01	0.03	0.02	0.00	0.00	0.00	0.01	0.02	0.01
	Li et al. [55]	0.04	0.52	0.07	0.09	0.11	0.01	0.05	0.05	0.12
	AFLoc	0.03	0.78	0.02	0.07	0.14	0.08	0.07	0.02	**0.15**

- 在不同 IoU 阈值 下，与两种对比方法相比，AFLoc 均表现出稳定的性能提升。
- 当 IoU = 0.3 时，AFLoc 的 平均定位准确率达到 0.62，超过了两种对比方法。
- 值得注意的是，即使在更严格的阈值条件下（如 IoU = 0.5 和 IoU = 0.7），模型依然表现优异，尤其是在 肺不张（atelectasis） 和 浸润（infiltration） 等更具挑战性的病变上。？？

Result：通过有限定位标注 提升性能

微调可视化定位性能



- 图展示了 AFLoc 在 四个数据集（JSRT、SIIM、ChestX-Det10 和 Retinal Fundus） 上的定位可视化示例。

- 这些示例包括 含有常见微小病灶的胸部 X 光图像 以及 视网膜出血图像。模型预测的区域与 真实标注（ground truth） 表现出良好的一致性，证明 AFLoc 即使在仅使用少量标注数据的情况下，也能有效提升定位准确性。

- 这些结果表明，AFLoc 在使用带标注数据进行微调后能够进一步受益，在分割任务和定位任务中都能取得具有竞争力的结果。这种能力对于 拓展 AFLoc 在临床中的潜在应用 至关重要。

Result：多粒度 消融实验

Supplementary Tables 9: Quantitative results (IoU) from the ablation study to investigate the importance of aligning image features with text features at different levels.

Different Levels of Features			RSNA Pneumonia	COVID Rural	MS-CXR	CheXlocalize	Mean
Word–Local (Shallow)	Sentence–Local (Deep)	Report–Global					
✓			0.239	0.147	0.180	0.133	0.138
	✓		0.281	0.169	0.318	0.301	0.272
		✓	0.191	0.056	0.105	0.186	0.145
✓	✓		0.329	0.177	0.264	0.285	0.271
✓		✓	0.254	0.159	0.175	0.200	0.179
	✓	✓	0.285	0.211	**0.325**	0.305	0.281
✓	✓	✓	**0.342**	**0.211**	0.324	**0.318**	**0.299**

Bold values represent the highest performance score among the compared methods.

- 当只使用 单一层级的文本特征 时，AFLoc 的性能明显下降，尤其是在 **词级特征** 和 **报告级特征** 的情况下。一个可能的解释是：词级特征缺乏足够的上下文信息，难以捕捉病变的特征；而 报告级特征由于过于宏观，可能无法捕捉关键的诊断细节。
- 相比之下，使用 **句子级特征** 的 AFLoc 表现出 相对更好的性能，这表明 句子级特征能够提供更丰富的上下文信息，而这些信息对于医学影像中的 病变定位（pathology localization） 非常重要。
- 当 不同层级的文本特征结合使用 时，可以观察到明显的性能提升；而当 **三种层级（词级、句子级和报告级）同时使用** 时，模型取得了 最佳性能。这一结果充分支持了我们的假设：多层级信息的结合对于提升医学影像中病变定位的准确性至关重要。

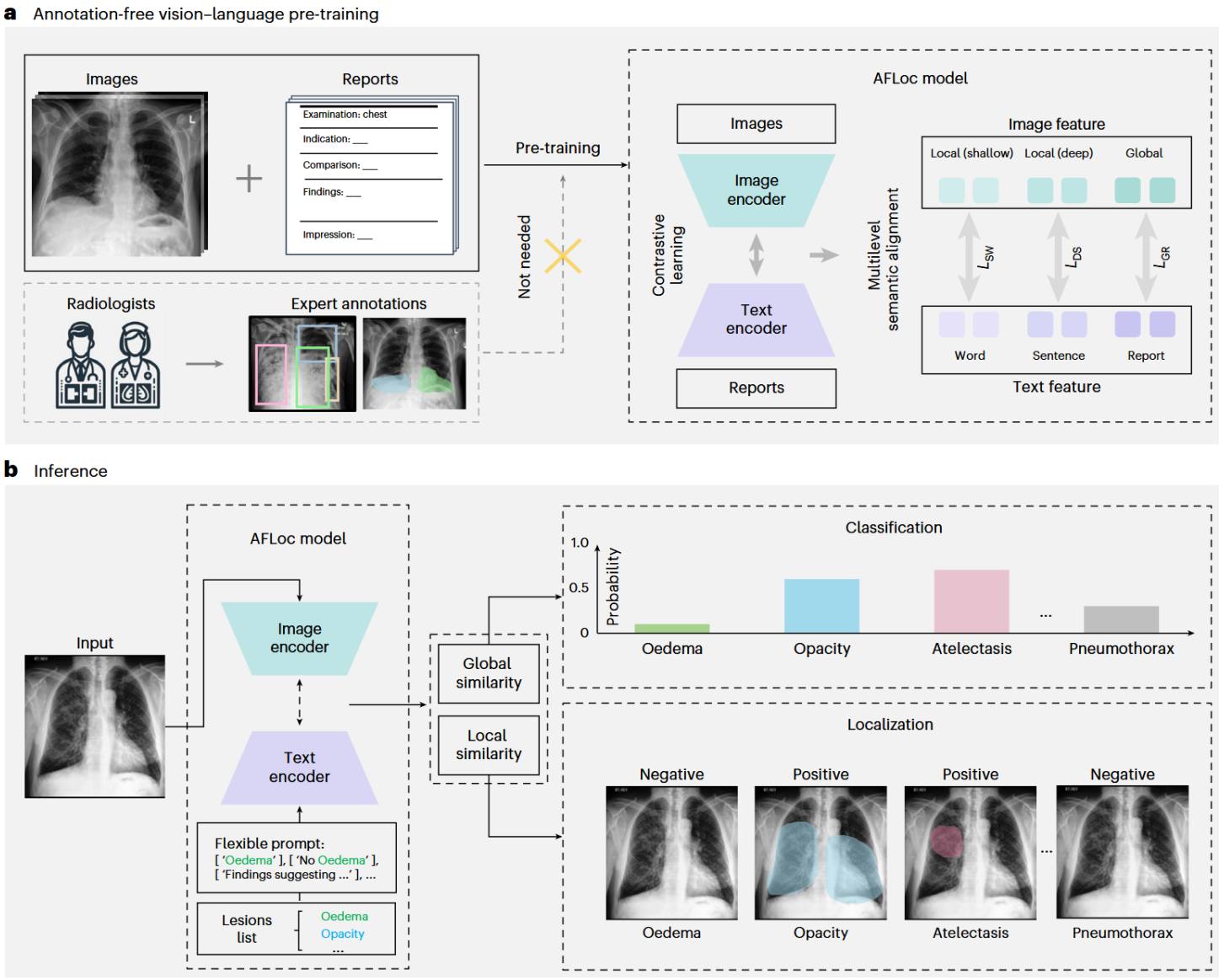
Result：文本编码器消融实验

Text encoder	IoU	Dice	CNR
LLaVA-Med-v1.5-Mistral-7B	0.307	0.441	1.483
BioClinicalBERT	**0.324**	**0.462**	**1.636**

- 为了研究文本编码方式对病变定位性能的影响，我们实现并测试了来自最新的生物医学大语言与视觉助手 LLaVA-Med 的文本编码器。LLaVA-Med 在开放式生物医学问答任务中已经表现出性能提升。
- 然而，如补充表 10 所示，其性能仍然不及 BioClinicalBERT。这种差异可以归因于：在特定领域数据上训练的语言模型通常在对应任务上表现更好。
- LLaVA-Med 是在 **PMC-15M 数据集** 上训练的，该数据集来源于 **PubMed Central 的科学论文**；而 **BioClinicalBERT** 则是在 **重症监护病房（ICU）患者的电子病历（EHR）** 上训练的，这与本研究中使用的**临床报告数据**更加一致。

Method：算法方面

文本编码 | 图像编码 | 多级语义对齐 | 无标注定位与诊断



Method：算法方面

动机描述

在临床实践中，医学文本可能来源于多种类型的信息，包括**临床观察、诊断描述以及相关信息**。这些文本通常由**多个句子或短语**组成。

在 GLoRIA 中，采用了一种 **块级分词技术**，通过聚合多个 subword（子词） 来构成完整的词语。而 BioViL 则使用了一个由多个数据集构建的**自定义词典**，旨在减少词语被拆分成子词的频率。

然而，将**单词作为独立实体**进行处理可能会忽略完整的语义信息，从而导致**文本与图像之间出现错误的语义对齐**。

我们提出的 AFLoc 中，文本特征从**三个粒度层级**进行提取：

- **词级**（word level）
- **句子级**（sentence level）
- **报告级**（report level）

通过这种**多粒度语义表示**（multigranularity semantics），可以更加**全面且精确地表达医学报告内容**。

Method：算法方面

科普：BERT 对词的处理

BERT 的全称是 **Bidirectional Encoder Representations from Transformers**。可以理解为一个基于transformer的强大的文本编码器。且它不是只从左到右读，或者只从右到左读，而是**双向地看上下文**。

The patient has no lung opacity.

BERT 在理解 opacity 这个词时，会同时看到前面的 no lung，所以它更容易理解这是一个否定语义，而不是单独把 opacity 当成正向异常。

BERT 的处理流程：

第一步：分词（tokenization）： BERT 先把文本切成一个个 token。注意，这里的 token 不一定等于完整单词。 unremarkable 可能被拆成：un+##remark+##able。（三个**token/subword**）

第二步：转成编号：每个 token 都会映射成词表里的一个 id。

第三步：变成向量：每个 token id 会被映射成一个向量，这叫 **embedding**。

第四步：上下文编码：这些向量进入 BERT 的多层 Transformer Encoder。

Method：算法方面

具体技术 采用文本编码器：**BioClinicalBERT**

设一份医学报告为 x_t，它有：Q个词；P个句子。

1、词会先被拆成 token / subword

对于第 i个词，它会被切成 q_i个 token/subword。

结果：所以整篇报告最后会变成一个 token 序列，长度上限记作 H。 $e_t^I \in \mathbb{R}^H$

2、输入 BioClinicalBERT

tokenized report 被送入 text encoder $e_t^O \in \mathbb{R}^{L \times H \times D}$

L是bert的层数，H是token长度，D是维度

3、得到 subword-level feature

只取后四层得到平均特征，获得子词级别特征 subword $t_{sub} \in \mathbb{R}^{H \times D}$

Method：算法方面

具体技术 采用文本编码器：**BioClinicalBERT**

设一份医学报告为 x_t，它有：Q个词；P个句子。

4、获得词级特征

一个词可能对应多个 **subword**。作者把属于同一个词的所有 **subword** 特征汇总，得到一个词向量。因此获得：**Q个词级特征**。

5、获得句级特征

把属于同一个句子的所有 **subword 特征**做平均，得到一个句子向量。因此有**P个句级特征**。

6、得到报告级特征

再把整份报告的所有相关 **subword 特征**聚合，得到一个全局向量。即获得**一个报告级特征**。

Method：算法方面

具体技术 采用图像编码器：**ResNet-50**

分别提取：浅层局部特征、深层局部特征和全局特征。

1、获得浅层局部特征

从 **第三个下采样阶段** 提取浅层局部特征。 $v_s \in \mathbb{R}^{D \times M}$

D是特征维度，**M**是图像的子区域数量

2、获得深层局部特征

从 **第四个下采样阶段** 提取深层局部特征 $v_d \in \mathbb{R}^{D \times \frac{M}{4}}$

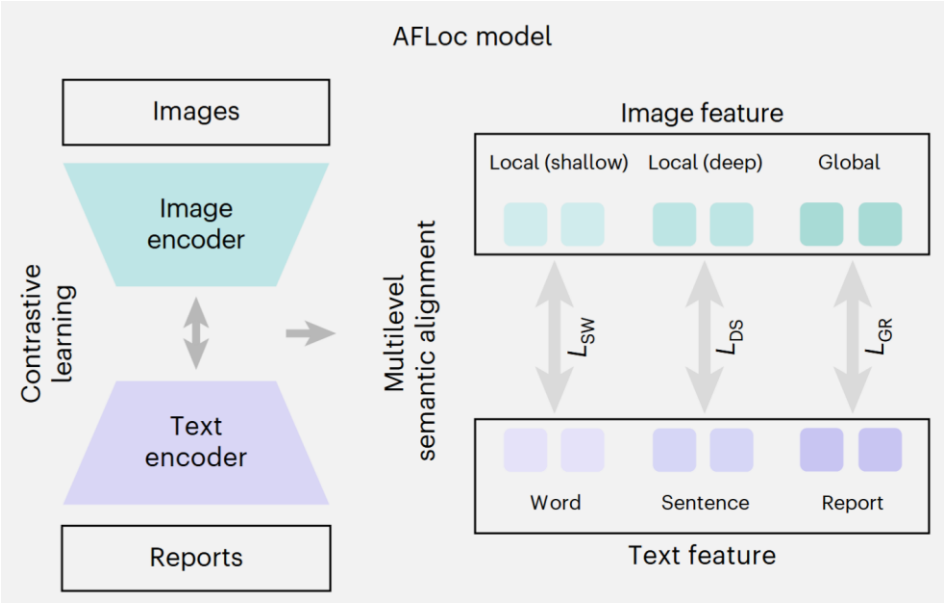
3、得到全局特征

取最后一层卷积输出做平均池化。 $v_g \in \mathbb{R}^{D}$

为了让它们能匹配图像和文本的特征维度，作者对三种图像特征都加了一个 **projection layer（投影层）**。

Method：算法方面

动机描述



医学报告是对应医学图像的文本描述，其中包含了不同粒度层面的丰富信息。理想情况下，医学图像及其对应报告中的语义信息应当在这些不同层次上保持一致。

这促使我们利用多层级语义信息来增强跨模态对齐。

具体而言，AFLoc 学习将图像中的

- 浅层局部特征 v_s和文本中的词级特征 t_w

- 深层局部特征 v_d和句子级特征 t_s对齐。

除了局部对齐之外，我们还保持

- 图像的全局特征 v_g与报告级文本特征 t_r之间的全局语义对齐。

Method：算法方面

具体技术 句子级对齐：深层局部图像 ↔ 句级文本

$v_d \in \mathbb{R}^{D \times \frac{M}{4}}$ P*D

1、计算相似度：

$$s = v_d^T t_s$$

第i个句子所对应的视觉特征

第i个句子和第j个区域的相似度的softmax

2、计算句子所对应上下文特征：

$$c_i = \sum_{j=1}^{\frac{M}{4}} \log \frac{\exp(s_{ij})/\tau_1}{\sum_{k=1}^{\frac{M}{4}} \exp(s_{ik})/\tau_1} v_{dj}$$

作为注意力权重聚合视觉特征

所有句子的视觉特征 和 所有句子特征 的匹配程度

第i个句子的视觉特征 和 第i个句子特征的 余弦相似度

3、计算局部匹配函数：

$$Z(v,t) = \log\left(\sum_{i=1}^{N} \exp(\Phi(c_i, t_i)/\tau_2)\right)$$

在句子级对齐中N=P,单词级对齐中N=Q，报告级对齐中N=1

Method：算法方面

具体技术　句子级对齐：深层局部图像 ↔ 句级文本

$v_d \in \mathbb{R}^{D \times \frac{M}{4}}$　　P*D

4、计算句子级对比损失：

第i个句子和图像的匹配分数 的 softmax

Text to image infoNCE

$$L_{DS} = -\frac{1}{B}\sum_{i}^{B}\left(\log\frac{\exp(Z(v_d^i, t_s^i)/\tau_3)}{\sum_{k=1}^{B}\exp(Z(v_d^i, t_s^k)/\tau_3)} + \log\frac{\exp(Z(v_d^i, t_s^i)/\tau_3)}{\sum_{k=1}^{B}\exp(Z(v_d^k, t_s^i)/\tau_3)}\right)$$

固定图像 v_d^i，让它更接近正确文本 t_s^i，远离 batch 内其他文本 t_s^k

Image to text infoNCE

Method：算法方面

具体技术

单词级对齐：浅层局部图像 ↔ 词级文本

$v_s \in \mathbb{R}^{D \times M}$　　Q*D

计算句子级对比损失：

Text to image infoNCE

$$L_{SW} = -\frac{1}{B}\sum_{i}^{B}\left(\log\frac{\exp(Z(v_s^i, t_w^i)/\tau_3)}{\sum_{k=1}^{B}\exp(Z(v_s^i, t_w^k)/\tau_3)} + \log\frac{\exp(Z(v_s^i, t_w^i)/\tau_3)}{\sum_{k=1}^{B}\exp(Z(v_s^k, t_w^i)/\tau_3)}\right)$$

固定图像 v_s^i，让它更接近正确文本 t_w^i，远离 batch 内其他文本 t_w^k

Image to text infoNCE

Method：算法方面

具体技术

单词级对齐：浅层局部图像 ↔ 词级文本

$v_s \in \mathbb{R}^{D \times M}$　　Q*D

计算单词级对比损失：

Text to image infoNCE

$$L_{SW} = -\frac{1}{B}\sum_{i}^{B}\left(\log\frac{\exp(Z(v_s^i, t_w^i)/\tau_3)}{\sum_{k=1}^{B}\exp(Z(v_s^i, t_w^k)/\tau_3)} + \log\frac{\exp(Z(v_s^i, t_w^i)/\tau_3)}{\sum_{k=1}^{B}\exp(Z(v_s^k, t_w^i)/\tau_3)}\right)$$

固定图像 v_s^i，让它更接近正确文本 t_w^i，远离 batch 内其他文本 t_w^k

Image to text infoNCE

Method：算法方面

具体技术

报告级对齐：全局图像 ↔ 报告级文本

$v_g \in \mathbb{R}^D$ 1*D

计算报告级对比损失：

Text to image infoNCE

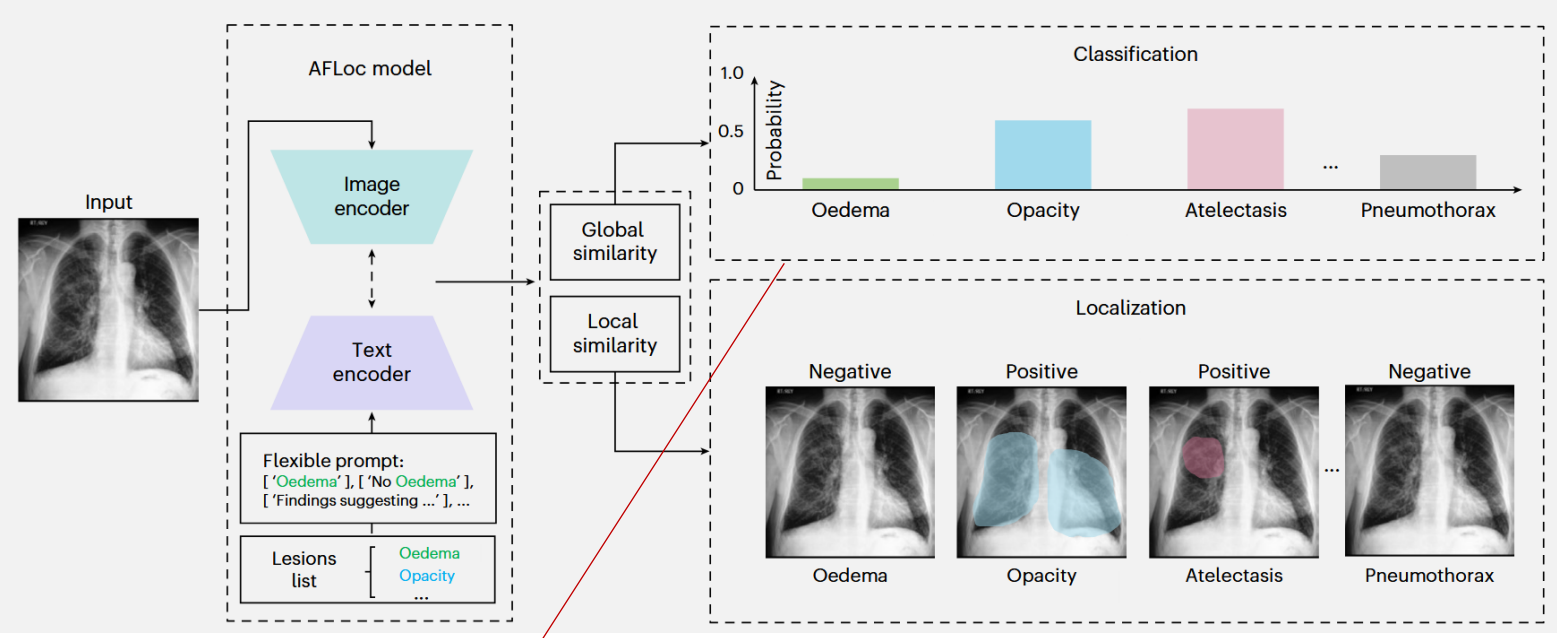
$$L_{GR} = -\frac{1}{B}\sum_{i}^{B}\left(\log\frac{\exp(Z(v_g^i, t_r^i)/\tau_3)}{\sum_{k=1}^{B}\exp(Z(v_g^i, t_r^k)/\tau_3)} + \log\frac{\exp(Z(v_g^i, t_r^i)/\tau_3)}{\sum_{k=1}^{B}\exp(Z(v_g^k, t_r^i)/\tau_3)}\right)$$

Image to text infoNCE

$$L = L_{SW} + L_{DS} + L_{GR}$$

Method：算法方面

具体技术



我们将该任务转化为一个**文本-图像匹配问题。（定位与诊断）**

定位：要由于其需要更细致的表示，我们计算深层局部图像特征 v_d 与文本特征 t_r 之间的相似度图。

随后，通过双线性插值和归一化操作，将该相似度图上采样到原始图像大小，作为病灶定位生成的热力图。

最后，我们使用硬阈值化从热力图中得到二值分割掩码。为了保证定位结果的可靠性，我们只输出那些**诊断预测为阳性**的定位结果。

诊断：比如你想检测“气胸”，那就写两段文本：

•阳性文本：

“Findings suggesting pneumothorax”

•阴性文本：

“No findings suggesting pneumothorax”然后计算相似度。

discussion

本文贡献总结

AFLoc 的主要贡献是：

在不依赖病灶框或像素级标注的情况下，仅利用图像和临床报告，在预训练阶段学到用于病灶定位与诊断的多粒度跨模态表示。

相比以前只做全局对齐，或只关注词级局部对齐的方法，AFLoc 引入了：

- 词级
- 句级
- 报告级

三层文本语义，与图像不同层次特征进行对齐，因此更适合医学场景中“小病灶、细粒度、强上下文依赖”的任务。

discussion

实验结果总结

作者在三种差异很大的医学影像模态上验证了它：
•胸部 X 光
•眼底图像
•组织病理图像

说明这个方法不是只对单一任务有效，而是具有**跨模态通用性**。

特别是在胸片任务中，它不仅超过了 MedKLIP、BioViL 等方法，还能较好处理：
•低对比度病灶
•细微病灶
•新发疾病（如 COVID-19 肺炎）
这说明 AFLoc 的优势不仅是“平均指标更高”，而且在临床上真正困难的病例上也更有价值。

discussion

临床意义总结

作者专门强调了几点临床价值:

•在一些任务上超过了人工基准
•可以帮助医生减少漏诊和误诊
•能够缩短阅片时间
•可以作为决策支持工具，而不是替代医生

特别是放射科医生辅助实验很重要，可以改善临床流程:

•准确率提高
•阅片速度更快
所以作者想表达的是:

AFLoc 不只是一个“算法上更好”的模型，而是一个有机会进入临床工作流的系统。

discussion

强调其可扩展性

AFLoc 虽然强调“无标注”，但作者并没有把它局限在完全零标注场景。
他们指出：

- 在零标注下已经表现很好
- 加少量标注微调后还能继续提升

这意味着它既适合：
- 资源丰富的大医院
- 也适合缺少标注资源的小医院、基层机构

AFLoc 兼具“低资源可用性”和“高资源可扩展性”。

discussion

三个局限

局限 1：模型结构还不够先进

目前虽然做了多粒度对齐，但还可以进一步做：

•分层多尺度融合

•在不同下采样阶段逐步融合图像和文本

也就是说，当前 AFLoc 还不是架构上的最终形态。

局限 3：更多模态还没验证

虽然已经验证了三种模态，但更复杂的数据类型还没研究透：

•超声

•MRI

•基因组学

•ECG

这些模态差异更大，可能需要专门的编码器或更复杂的统一表示空间

局限 2：定位依赖阳性分类

现在只有当分类为阳性时，才输出定位结果。

这更符合临床流程，但也带来问题：

•如果分类错了，定位就可能直接失效

•对“不确定病例”的可解释性还不够

因此作者提出未来要加：

•不确定性估计

•低置信度提示

•融合临床先验知识

•利用医生反馈做强化学习式迭代优化

好词好句

- 1) The core strength of AFLoc is extensive multilevel semantic structure-based contrastive learning, which comprehensively aligns multigranularity medical concepts with abundant image features to adapt to the diverse expressions of pathologies without the reliance on expert image annotations.

点评

用 comprehensively 修饰 对齐，强调所提方法“对齐的全面性”

好词好句

- 2) Over the past decade, supervised deep learning methods have accelerated advancements in disease localization.

点评

句式：over the past decade, xxx（方法） have accelerated advancements in xxx（领域）
过去十年里，xx方法加速了xx领域的发展

- 3) Specifically, clinical localization tasks often require experienced clinicians to meticulously annotate numerous precise bounding boxes or perform pixel-wise delineations of localized pathology areas.

点评

句式：meticulously 修饰标注，强调标注的成本很大。

好词好句

- 4) Initially, these methodologies acquire general visual representations through self-supervised learning from image datasets, followed by fine tuning on smaller annotated datasets.

点评

副词，Initially 可以用在综述解决某一科学问题场景中，起初，xxx方法致力于....

- 5) This requirement is particularly challenging in flexible and dynamic clinical environments, especially for emerging diseases (for example, COVID-19), where deployed models may fail to perform effectively.

点评

形容词：形容临床环境复杂且多变，使用 flexible 和 dynamic

好词好句

- 6) In recent years, unsupervised deep learning methods have gained increasing attention due to their independence from annotated datasets, particularly in the field of anomaly detection.
- They are particularly effective for data with simple structures and low intersample variance, allowing them to learn normative distributions and achieve excellent anomaly detection performance

点评

副词，particularly + 名词/名词短语　particularly in/for/among...：强调某个对象特别重要
particularly + 形容词/副词：提高　形容词的强度

好词好句

- 7) A promising approach is the development of medical visionlanguage pre-training methods

点评

用A promising approach 来叙说解决某一科学问题的可能思路

- 8) A primary obstacle is the lack of explicit pathology localization markers in clinical reports, which often provide only coarse information such as ‘upper’ or ‘left’ to indicate disease location.

点评

用A primary obstacle来叙说某个领域的关键问题，可以替换“difficulties”等基础表达

好词好句

- 9) We hope that this study can help address the challenges posed by annotation scarcity and modality diversity in clinical environments, while providing insights for the design of future clinical open-environment methods.

点评

Intro的最后一句，用展望来展示自己的格局
我们希望我们的研究能够帮助解决xxx难题，同时为未来xxx提供参考。

好词好句

- 10) However, existing methods encounter various challenges, notably stemming from the scarcity of annotated data related to generalization

点评

encounter various challenges：面临挑战，尤其源于:notably stemming from

观点论据

- 1) However, the efficacy of these methods heavily relies on extensively annotated training datasets, which require domain experts to invest considerable time.
- **深度学习方法依赖专家标注数据集**

翻译

然而，这些方法的有效性在很大程度上依赖于大量带标注的训练数据集，而这需要领域专家投入大量时间。

观点论据

- 2) However, these methods still require annotations for specific downstream tasks. This requirement is particularly challenging in flexible and dynamic clinical environments, especially for emerging diseases (for example, COVID-19), where deployed models may fail to perform effectively.
- **微调算法 缓解标注压力 的 局限性**

翻译

然而，这些方法仍然需要针对特定下游任务进行标注。在灵活且动态变化的临床环境中，这一要求尤为具有挑战性，尤其是面对新发疾病（例如 COVID-19）时，已部署的模型可能无法有效发挥作用。

观点论据

- 3) A primary obstacle is the lack of explicit pathology localization markers in clinical reports, which often provide only coarse information such as ‘upper’ or ‘left’ to indicate disease location. Moreover, clinical descriptions by clinicians are subjective and variable, further complicating the task of accurately extracting and localizing diseases in medical images.
- VLM方法 缓解标注压力 的 局限性

翻译

一个主要障碍在于，临床报告中缺乏明确的病灶定位标记，通常只会提供诸如“上部”或“左侧”这类较为粗略的信息来指示疾病位置。此外，临床医生的描述具有主观性且存在差异性，这进一步增加了从医学图像中准确提取并定位疾病的难度。

观点论据

- 4) However, these fine-grained methods typically focus on individual levels of medical concepts and may overlook the variable meanings of concepts in different contexts. Therefore, these approaches may struggle to adapt to the diverse expressions of disease descriptors in clinical practice, often requiring customized textual cues to enhance localization performance.
- **现有细粒度VLM的局限性**

翻译

然而，这些细粒度方法通常只关注医学概念的单一层级，可能会忽视概念在不同语境中的含义变化。因此，这些方法往往难以适应临床实践中疾病描述符的多样化表达，常常需要定制化的文本提示来提升定位性能。

观点论据

- 5) Unlike traditional global semantic alignment strategies, AFLoc introduces a contrastive learning framework with a multilevel semantic alignment component, facilitating the comprehensive alignment of medical concepts from reports with image features.
- **介绍多级细粒度对齐的好处**

翻译

不同于传统的全局语义对齐策略，AFLoc 引入了一种带有多层级语义对齐组件的对比学习框架，从而促进报告中的医学概念与图像特征之间实现更全面的对齐。

观点论据

- 5) In contrast, multitask deep learning enables the simultaneous analysis of different tasks within a single model. By sharing feature representations and interactions among related tasks, multitask learning is more data efficient and has been shown to reduce overfitting and improve model generalization across various applications, including computer vision, disease diagnosis, and drug discovery.
- **多任务学习的好处**
  - 相比之下，多任务深度学习可在单个模型中同时分析不同的任务。 通过共享相关任务之间的特征表征和交互，多任务学习的数据效率更高，并已证明可在计算机视觉、疾病诊断和药物发现等各种应用中减少过度拟合，提高模型泛化能力。