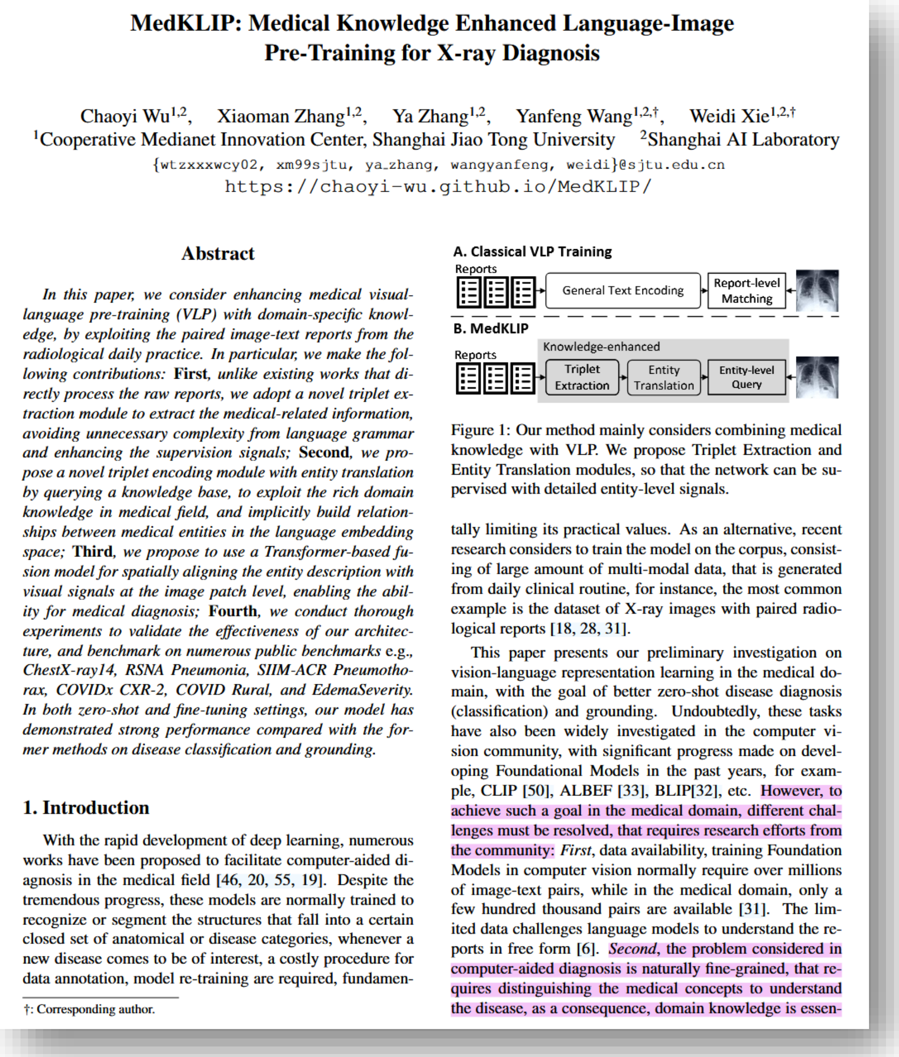
IPLab 智能感知与信息处理实验室

JC-club

余海军　2025 年 12 月 7日

#知识增强

#MedKLIP



MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training for X-ray Diagnosis

Chaoyi Wu[1,2], Xiaoman Zhang[1,2], Ya Zhang[1,2], Yanfeng Wang[1,2,†], Weidi Xie[1,2,†]
[1]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University [2]Shanghai AI Laboratory
{wtzxxxwcy02, xm99sjtu, ya_zhang, wangyanfeng, weidi}@sjtu.edu.cn
https://chaoyi-wu.github.io/MedKLIP/

Abstract

In this paper, we consider enhancing medical visual-language pre-training (VLP) with domain-specific knowledge, by exploiting the paired image-text reports from the radiological daily practice. In particular, we make the following contributions: **First**, *unlike existing works that directly process the raw reports, we adopt a novel triplet extraction module to extract the medical-related information, avoiding unnecessary complexity from language grammar and enhancing the supervision signals;* **Second**, *we propose a novel triplet encoding module with entity translation by querying a knowledge base, to exploit the rich domain knowledge in medical field, and implicitly build relationships between medical entities in the language embedding space;* **Third**, *we propose to use a Transformer-based fusion model for spatially aligning the entity description with visual signals at the image patch level, enabling the ability for medical diagnosis;* **Fourth**, *we conduct thorough experiments to validate the effectiveness of our architecture, and benchmark on numerous public benchmarks e.g., ChestX-ray14, RSNA Pneumonia, SIIM-ACR Pneumothorax, COVIDx CXR-2, COVID Rural, and EdemaSeverity. In both zero-shot and fine-tuning settings, our model has demonstrated strong performance compared with the former methods on disease classification and grounding.*

Figure 1: Our method mainly considers combining medical knowledge with VLP. We propose Triplet Extraction and Entity Translation modules, so that the network can be supervised with detailed entity-level signals.

1. Introduction

With the rapid development of deep learning, numerous works have been proposed to facilitate computer-aided diagnosis in the medical field [46, 20, 55, 19]. Despite the tremendous progress, these models are normally trained to recognize or segment the structures that fall into a certain closed set of anatomical or disease categories, whenever a new disease comes to be of interest, a costly procedure for data annotation, model re-training are required, fundamentally limiting its practical values. As an alternative, recent research considers to train the model on the corpus, consisting of large amount of multi-modal data, that is generated from daily clinical routine, for instance, the most common example is the dataset of X-ray images with paired radiological reports [18, 28, 31].

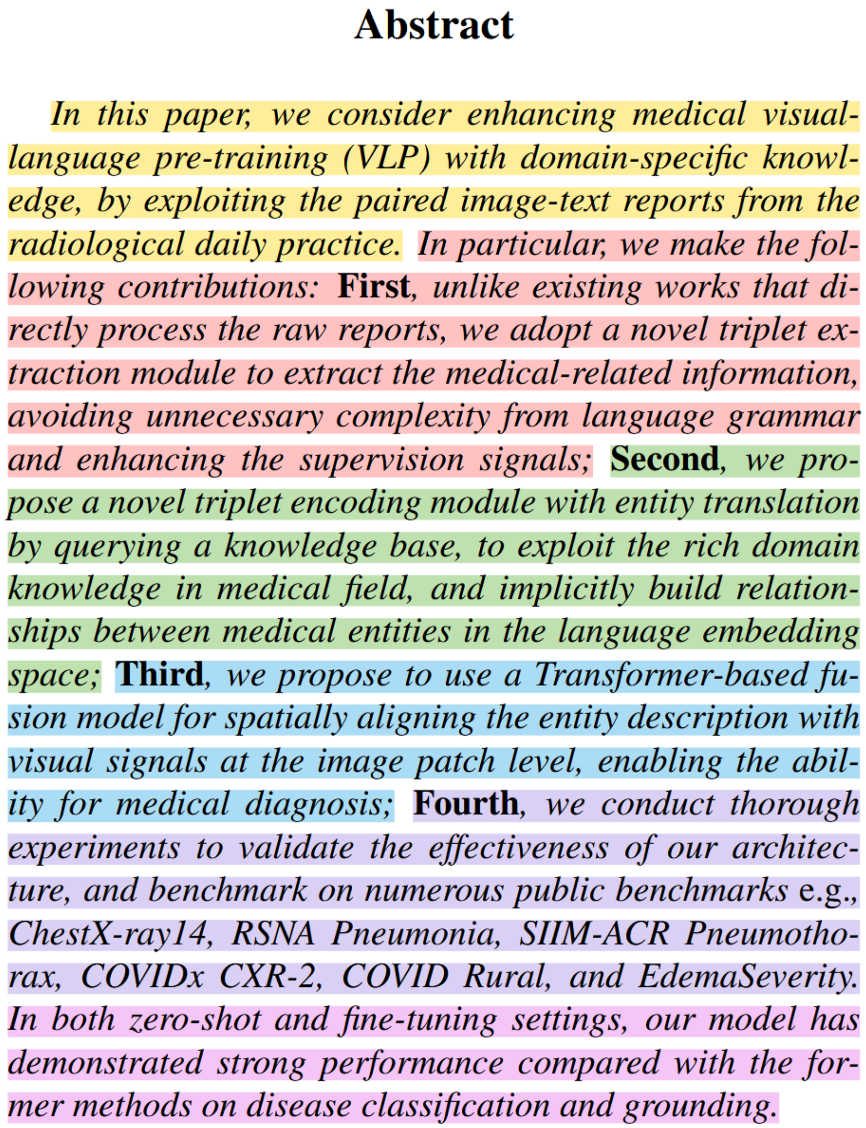
This paper presents our preliminary investigation on vision-language representation learning in the medical domain, with the goal of better zero-shot disease diagnosis (classification) and grounding. Undoubtedly, these tasks have also been widely investigated in the computer vision community, with significant progress made on developing Foundational Models in the past years, for example, CLIP [50], ALBEF [33], BLIP[32], etc. However, to achieve such a goal in the medical domain, different challenges must be resolved, that requires research efforts from the community: *First*, data availability, training Foundation Models in computer vision normally require over millions of image-text pairs, while in the medical domain, only a few hundred thousand pairs are available [31]. The limited data challenges language models to understand the reports in free form [6]. *Second*, the problem considered in computer-aided diagnosis is naturally fine-grained, that requires distinguishing the medical concepts to understand the disease, as a consequence, domain knowledge is essen-

†: Corresponding author.

论文基本信息

- 论文题目：
  - **MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training for X-ray Diagnosis**
  - MedKLIP：基于医学知识增强的语言-图像预训练模型用于X光诊断
- 作者信息：上海交通大学、上海*AI lab*
- 会议分区：ICCV CCFA

Abstract



Abstract

In this paper, we consider enhancing medical visual-language pre-training (VLP) with domain-specific knowledge, by exploiting the paired image-text reports from the radiological daily practice. In particular, we make the following contributions: **First***, unlike existing works that directly process the raw reports, we adopt a novel triplet extraction module to extract the medical-related information, avoiding unnecessary complexity from language grammar and enhancing the supervision signals;* **Second***, we propose a novel triplet encoding module with entity translation by querying a knowledge base, to exploit the rich domain knowledge in medical field, and implicitly build relationships between medical entities in the language embedding space;* **Third***, we propose to use a Transformer-based fusion model for spatially aligning the entity description with visual signals at the image patch level, enabling the ability for medical diagnosis;* **Fourth***, we conduct thorough experiments to validate the effectiveness of our architecture, and benchmark on numerous public benchmarks* e.g., *ChestX-ray14, RSNA Pneumonia, SIIM-ACR Pneumothorax, COVIDx CXR-2, COVID Rural, and EdemaSeverity. In both zero-shot and fine-tuning settings, our model has demonstrated strong performance compared with the former methods on disease classification and grounding.*

一句话概括本文贡献：本文旨在利用放射科日常实践中产生的成对图像—文本报告，引入领域特定知识，以增强医学视觉-语言预训练（VLP）模型的能力。

创新一：第一，与直接处理原始报告的现有工作不同，我们设计了一种新颖的三元组抽取模块，用于提取与医学相关的信息，从而避免自然语言语法所带来的不必要复杂性，并强化监督信号。

创新二：第二，我们提出了一种结合知识库实体翻译的三元组编码模块，通过查询医学知识库充分利用丰富的领域知识，并在语言嵌入空间中隐式建模各医学实体之间的关系。

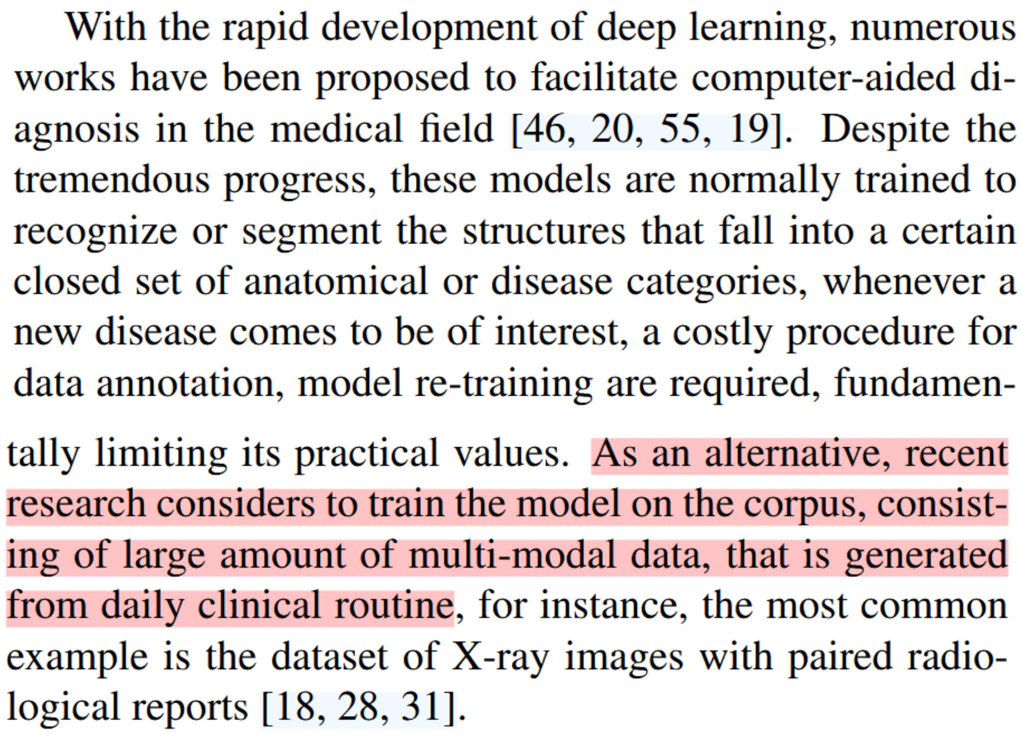
创新三：第三，我们采用基于 Transformer 的多模态融合模型，在图像补丁（patch）层面实现实体描述与视觉信号的空间对齐，从而赋予模型医学诊断的能力。

第四，我们在多个公开数据集上对所提出的架构进行了系统而全面的实验验证。

在零样本和微调两种设置下，与现有方法相比，我们的模型在疾病分类和疾病定位等任务上均展现出优异的性能。

引言读解

第一段：【背景写作】：计算机辅助诊断进入到多模态大模型时代



With the rapid development of deep learning, numerous works have been proposed to facilitate computer-aided diagnosis in the medical field [46, 20, 55, 19]. Despite the tremendous progress, these models are normally trained to recognize or segment the structures that fall into a certain closed set of anatomical or disease categories, whenever a new disease comes to be of interest, a costly procedure for data annotation, model re-training are required, fundamentally limiting its practical values. As an alternative, recent research considers to train the model on the corpus, consisting of large amount of multi-modal data, that is generated from daily clinical routine, for instance, the most common example is the dataset of X-ray images with paired radiological reports [18, 28, 31].

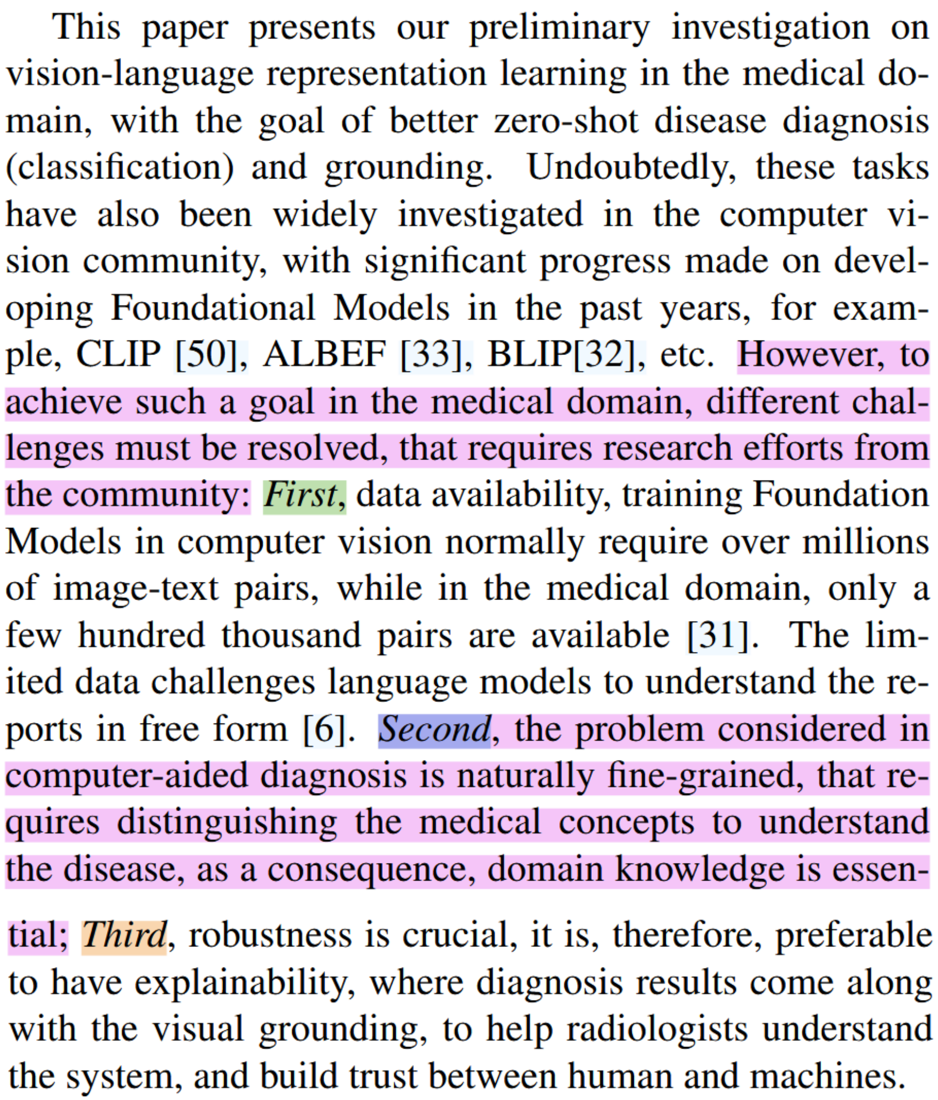
随着深度学习的快速发展，大量研究被提出以推动医学领域的计算机辅助诊断。

【小模型的缺点】：尽管取得了巨大进展，这类模型通常被训练用于识别或分割预先定义的有限解剖结构或疾病类别；一旦出现新的感兴趣疾病，就必须进行昂贵的数据标注与模型重新训练，从根本上限制了其实际应用价值。

【引入多模态和放射学报告】：作为替代方案，近期研究开始考虑使用来源于临床日常实践的大规模多模态语料库进行模型训练，例如最常见的就是包含 X 线影像及其对应放射学报告的数据集。

引言读解

第二段：【挖坑写作】：医学视觉语言模型的挑战



This paper presents our preliminary investigation on vision-language representation learning in the medical domain, with the goal of better zero-shot disease diagnosis (classification) and grounding. Undoubtedly, these tasks have also been widely investigated in the computer vision community, with significant progress made on developing Foundational Models in the past years, for example, CLIP [50], ALBEF [33], BLIP[32], etc. However, to achieve such a goal in the medical domain, different challenges must be resolved, that requires research efforts from the community: *First*, data availability, training Foundation Models in computer vision normally require over millions of image-text pairs, while in the medical domain, only a few hundred thousand pairs are available [31]. The limited data challenges language models to understand the reports in free form [6]. *Second*, the problem considered in computer-aided diagnosis is naturally fine-grained, that requires distinguishing the medical concepts to understand the disease, as a consequence, domain knowledge is essential; *Third*, robustness is crucial, it is, therefore, preferable to have explainability, where diagnosis results come along with the visual grounding, to help radiologists understand the system, and build trust between human and machines.

【本文任务】：本文对医学领域中的视觉-语言表征学习进行了初步探索，目标是实现更优的零样本疾病诊断（分类）和疾病/视觉定位。

【阐述挑战】：毫无疑问，这类任务在计算机视觉领域也已被广泛研究，并在近几年基础模型的发展上取得了显著进展，例如 CLIP 、ALBEF 、BLIP 等。然而，要在医学场景中实现同样的目标，仍需解决一系列不同的挑战，这也需要整个社区的共同努力

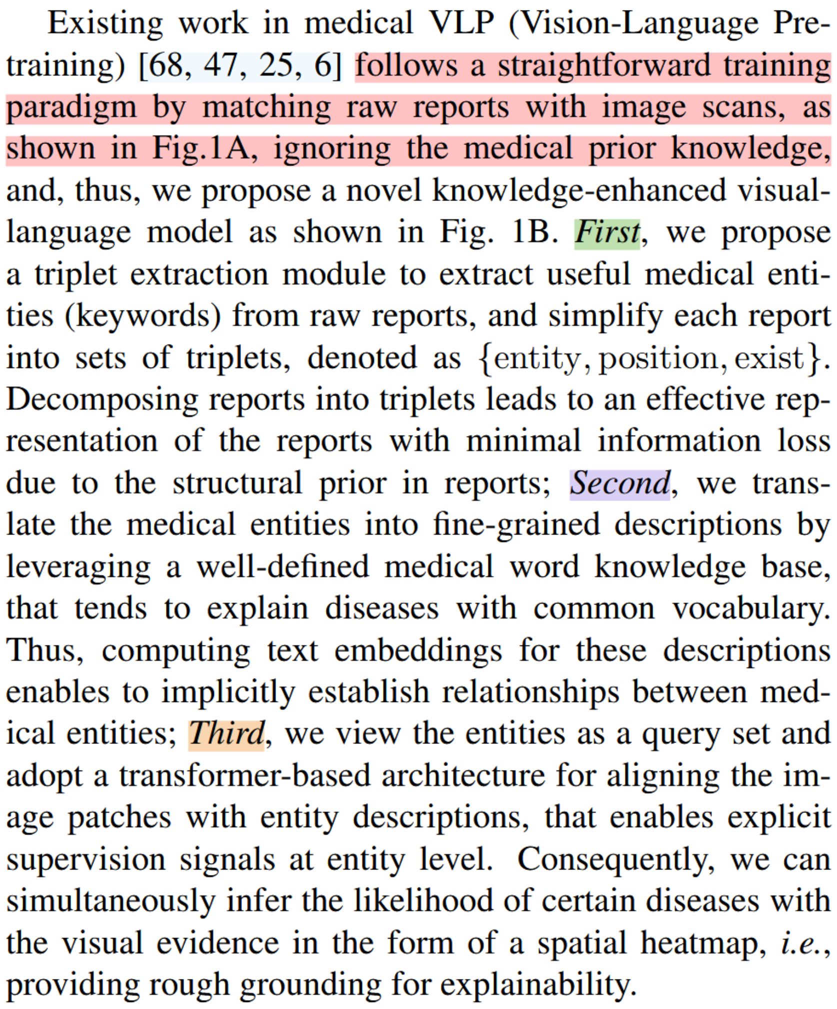
【坑一】数据可得性问题——在通用计算机视觉任务中，训练基础模型通常需要上百万对图文样本，而在医学领域，往往只能获得几十万对左右的图像 - 文本配对数据 [31]，有限的数据规模使得语言模型难以充分理解自由文本形式的报告 。

【坑二】细粒度领域知识：计算机辅助诊断中所处理的问题天然属于细粒度范畴，需要区分精细的医学概念以理解疾病本身，因此领域知识至关重要。

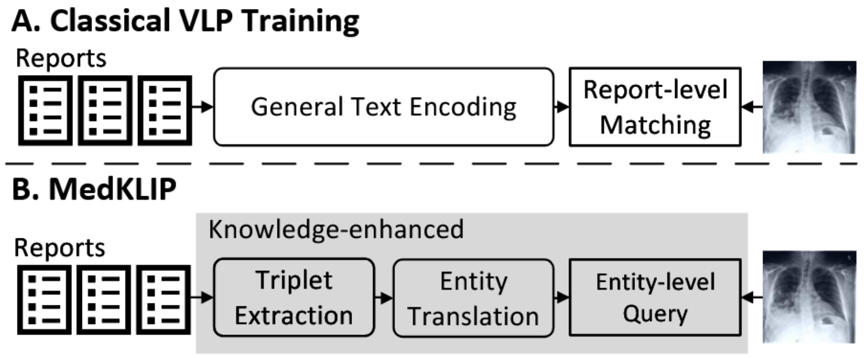
【坑三】可解释性：鲁棒性同样关键，因此模型最好具备可解释性，即诊断结果应伴随相应的视觉定位，以帮助放射科医生理解系统的决策过程，并在医生与智能系统之间建立信任。

引言读解

第三段：【填坑写作】：介绍本文方法



Existing work in medical VLP (Vision-Language Pre-training) [68, 47, 25, 6] follows a straightforward training paradigm by matching raw reports with image scans, as shown in Fig.1A, ignoring the medical prior knowledge, and, thus, we propose a novel knowledge-enhanced visual-language model as shown in Fig. 1B. *First*, we propose a triplet extraction module to extract useful medical entities (keywords) from raw reports, and simplify each report into sets of triplets, denoted as $\{\text{entity}, \text{position}, \text{exist}\}$. Decomposing reports into triplets leads to an effective representation of the reports with minimal information loss due to the structural prior in reports; *Second*, we translate the medical entities into fine-grained descriptions by leveraging a well-defined medical word knowledge base, that tends to explain diseases with common vocabulary. Thus, computing text embeddings for these descriptions enables to implicitly establish relationships between medical entities; *Third*, we view the entities as a query set and adopt a transformer-based architecture for aligning the image patches with entity descriptions, that enables explicit supervision signals at entity level. Consequently, we can simultaneously infer the likelihood of certain diseases with the visual evidence in the form of a spatial heatmap, *i.e.*, providing rough grounding for explainability.



【核心坑】现有的医学视觉-语言预训练（VLP）工作通常采用一种直接的训练范式，即将原始报告与图像扫描匹配，如图1A所示，忽略了医学领域的先验知识。

因此，我们提出了一种新的知识增强视觉-语言模型，如图1B所示。

【填坑一】我们提出了一个三元组提取模块，用于从原始报告中提取有用的医学实体，并将每个报告简化为一组三元组，表示为{实体，位置，存在性}。将报告分解为三元组可以有效表示报告内容，同时由于报告结构的先验知识，信息损失最小

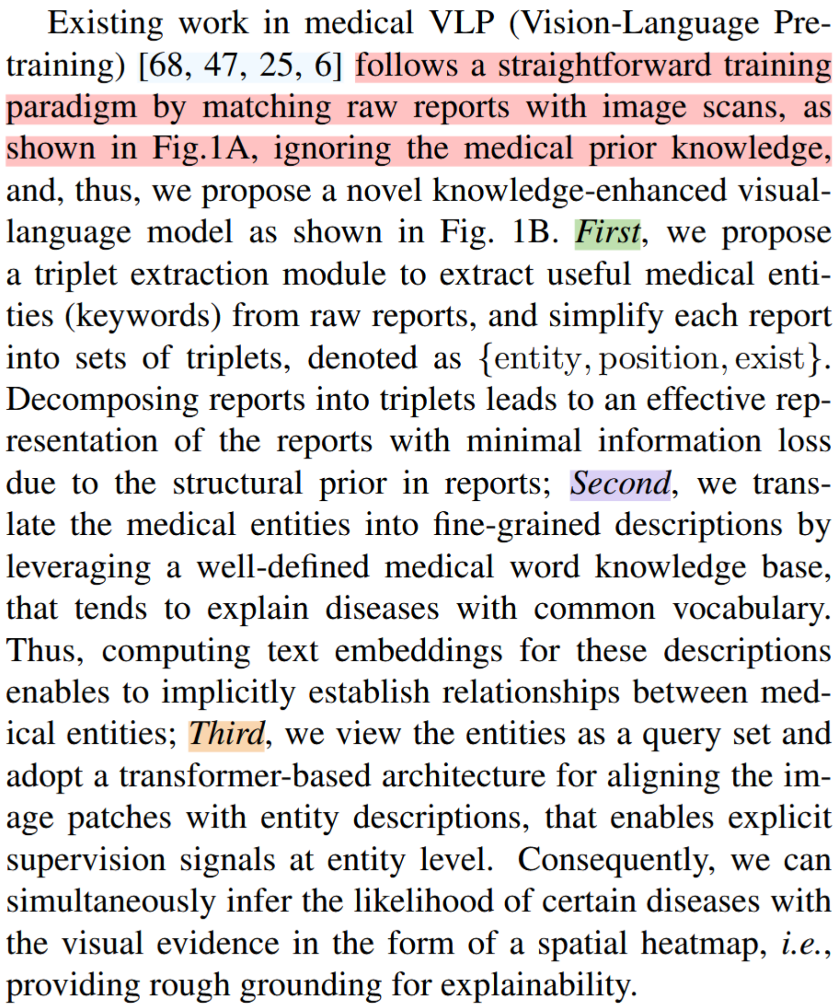
【坑一】数据可得性问题

【坑二】细粒度领域知识

【坑三】可解释性

引言读解

第三段：【填坑写作】：介绍本文方法



Existing work in medical VLP (Vision-Language Pre-training) [68, 47, 25, 6] follows a straightforward training paradigm by matching raw reports with image scans, as shown in Fig.1A, ignoring the medical prior knowledge, and, thus, we propose a novel knowledge-enhanced visual-language model as shown in Fig. 1B. *First*, we propose a triplet extraction module to extract useful medical entities (keywords) from raw reports, and simplify each report into sets of triplets, denoted as $\{\text{entity}, \text{position}, \text{exist}\}$. Decomposing reports into triplets leads to an effective representation of the reports with minimal information loss due to the structural prior in reports; *Second*, we translate the medical entities into fine-grained descriptions by leveraging a well-defined medical word knowledge base, that tends to explain diseases with common vocabulary. Thus, computing text embeddings for these descriptions enables to implicitly establish relationships between medical entities; *Third*, we view the entities as a query set and adopt a transformer-based architecture for aligning the image patches with entity descriptions, that enables explicit supervision signals at entity level. Consequently, we can simultaneously infer the likelihood of certain diseases with the visual evidence in the form of a spatial heatmap, *i.e.*, providing rough grounding for explainability.

【填坑一】我们提出了一个三元组提取模块，用于从原始报告中提取有用的医学实体，并将每个报告简化为一组三元组，表示为{实体，位置，存在性}。将报告分解为三元组可以有效表示报告内容，同时由于报告结构的先验知识，信息损失最小

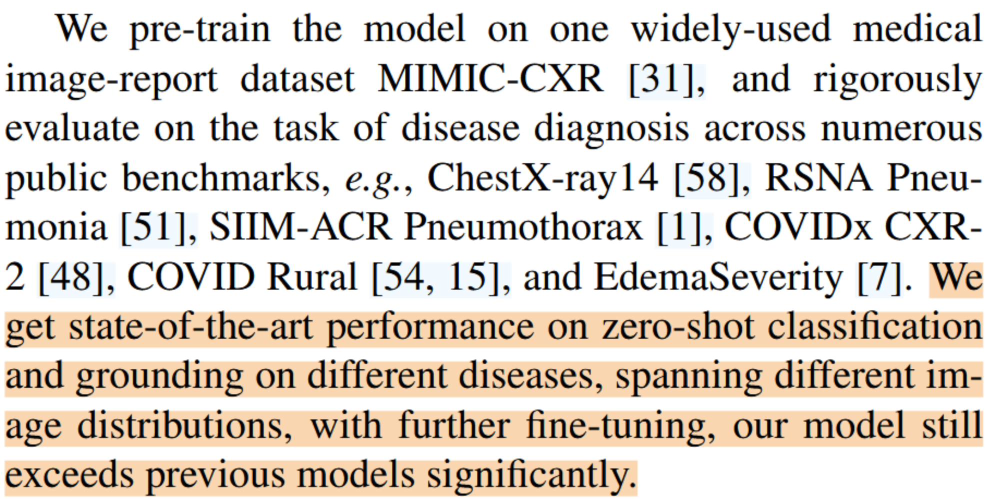
【填坑二】其次，我们利用一个定义明确的医学词汇知识库，将医学实体转化为细粒度的描述，该知识库通常使用常见词汇解释疾病。通过对这些描述计算文本嵌入，可以隐式地在医学实体之间建立关系；

【填坑三】第三，我们将实体视为查询集合，并采用基于 Transformer 的架构，将图像补丁与实体描述对齐，从而在实体层面提供明确的监督信号。因此，我们可以通过空间热图的形式同时推断某些疾病的可能性，即为可解释性提供粗略的定位。

【坑一】数据可得性问题　**【坑二】细粒度领域知识**　**【坑三】可解释性**

引言读解

第四段：实验性能涨点



We pre-train the model on one widely-used medical image-report dataset MIMIC-CXR [31], and rigorously evaluate on the task of disease diagnosis across numerous public benchmarks, *e.g.*, ChestX-ray14 [58], RSNA Pneumonia [51], SIIM-ACR Pneumothorax [1], COVIDx CXR-2 [48], COVID Rural [54, 15], and EdemaSeverity [7]. We get state-of-the-art performance on zero-shot classification and grounding on different diseases, spanning different image distributions, with further fine-tuning, our model still exceeds previous models significantly.

我们在一个广泛使用的医学图像-报告数据集 MIMIC-CXR [31] 上对模型进行了预训练，并在多个公共基准数据集上对疾病诊断任务进行了严格评估

在不同疾病的零样本分类和定位任务中，我们的模型在不同图像分布下表现出最先进的性能，并且通过进一步微调，模型的性能仍显著超越了之前的模型。

方法学-技术路线图

3. Method

3.1 problem scenario

3.2 report pre-processing

3.3 architecture

3.3.1 visual encoding

3.3.2 knowledge-enhanced triplet encoding

3.3.3 fusion module

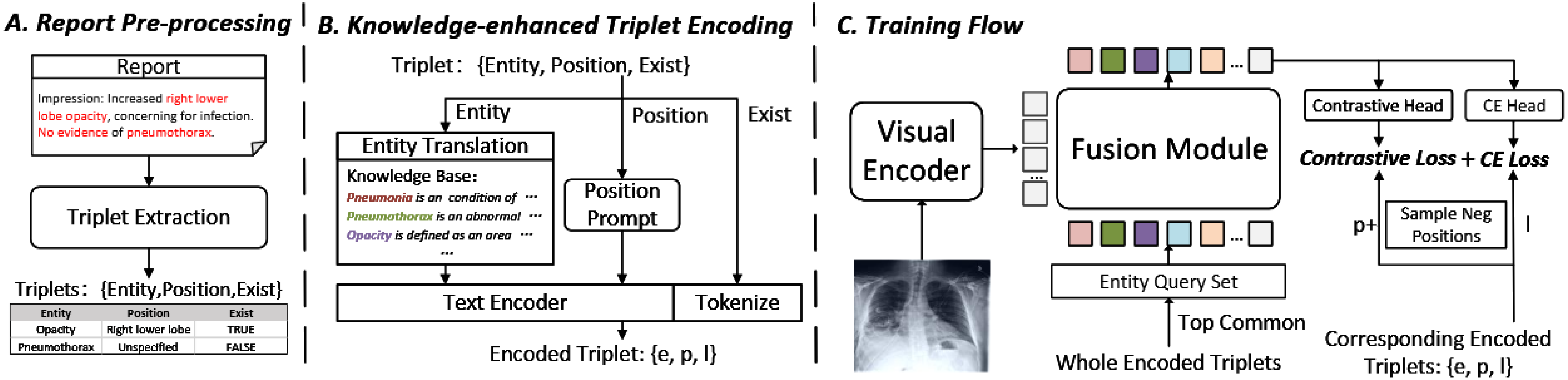
3.4 training

3.5 inference

【坑一】数据可得性问题

【坑二】细粒度领域知识

【坑三】可解释性


A. Report Pre-processing
Report
Impression: Increased right lower lobe opacity, concerning for infection. No evidence of pneumothorax.
Triplet Extraction
Triplets: {Entity,Position,Exist}
B. Knowledge-enhanced Triplet Encoding
Triplet: {Entity, Position, Exist}
Entity
Position
Exist
Entity Translation
Knowledge Base:
Pneumonia is an condition of …
Pneumothorax is an abnormal …
Opacity is defined as an area …
…
Position Prompt
Text Encoder
Tokenize
Encoded Triplet: {e, p, l}
C. Training Flow
Visual Encoder
Fusion Module
Contrastive Head
CE Head
Contrastive Loss + CE Loss
p+
l
Sample Neg Positions
Entity Query Set
Top Common
Whole Encoded Triplets
Corresponding Encoded Triplets: {e, p, l}

Entity	Position	Exist
Opacity	Right lower lobe	TRUE
Pneumothorax	Unspecified	FALSE

Report Pre-processing

Knowledge-enhanced Triplet Encoding

Training Flow

方法学

3.1 problem scenario

数据：假设我们有一个包含 N 个样本的训练集： $D_{train} = \{(X_1, T_1), \ldots, (X_N, T_N)\}$，其中 X_i 是 X射线图像， T_i 是相应的医学报告。

任务：该模型的输出包括：

$\hat{s}_i$：这是推断的疾病发生概率，反映了患者是否可能患有输入描述中提到的疾病。

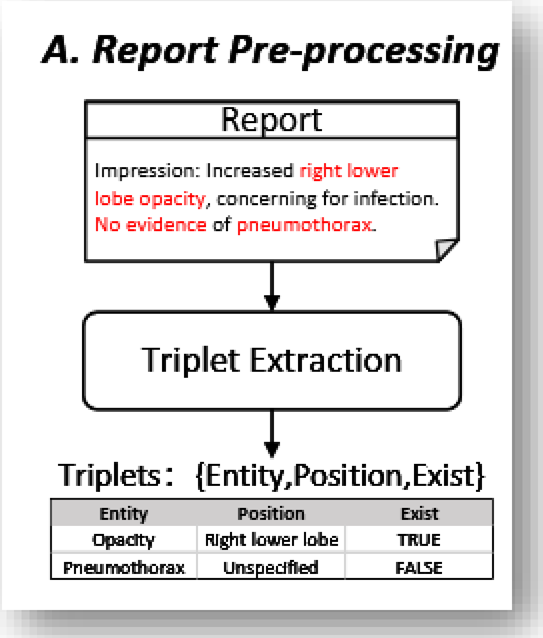
$\hat{m}_i$：这是一个空间热图，指示了图像中可能与疾病相关的区域。

$$\hat{s}_i, \hat{m}_i = \Phi_{\text{fusion}}(\Phi_{\text{visual}}(\mathcal{X}_i), \Phi_{\text{textual}}([\text{description}]))$$

- $X_i \in R^{H \times W \times 3}$：表示图像样本，其中 H 和 W 分别为图像的高度和宽度。
- $\hat{s}_i \in [0, 1]$：表示推断出的疾病发生概率。
- $\hat{m}_i \in R^{H \times W \times 1}$：表示预测的空间热图，其中每个像素的激活值指示可能与疾病相关的区域。

方法学

3.3.2 Knowledge-enhanced Triplet Encoding



A. Report Pre-processing

Report

Impression: Increased right lower lobe opacity, concerning for infection. No evidence of pneumothorax.

Triplet Extraction

Triplets：{Entity,Position,Exist}

Entity	Position	Exist
Opacity	Right lower lobe	TRUE
Pneumothorax	Unspecified	FALSE

动机

【填坑一】我们提出了一个三元组提取模块，用于从原始报告中提取有用的医学实体，并将每个报告简化为一组三元组，表示为 {实体，位置，存在性}。将报告分解为三元组可以有效表示报告内容，同时由于报告结构的先验知识，信息损失最小

技术

医学关键词通过命名实体识别（NER）方法被提取和分类为“实体”或“位置”，NER模块还会为每个实体提供一个“存在性”标签，用来判断该实体是否存在于报告中。基于此，我们可以使用一组三元组形式表示，例如{entity，position，exist}，以重新构造报告中的句子。因此，给定一个包含多个句子的报告 $T = \{s_1, s_2, \ldots, s_M\}$，提取模块会独立处理每个句子，并从报告中构建一组三元组

$$\Phi_{\mathrm{ex}}(s_j) = \{\mathrm{entity}_n, \mathrm{position}_n, \mathrm{exist}_n\}, n \in [0, t_j]$$

动机

讨论：与自然语言处理中的通用文本相比，医学报告内容更加专业，并且通常在特定的词汇表内（大多列在UMLS [5]中）。专门设计的NER方法 [29] 在报告中表现出色。因此，在医学视觉-语言预训练中采用三元组提取操作，能够避免因理解语法带来的不必要复杂性，同时保留报告中的关键信息。

方法学

3.2 report pre-processing

动机

【填坑二】其次，我们利用一个定义明确的医学词汇知识库，将医学实体转化为细粒度的描述，该知识库通常使用常见词汇解释疾病。通过对这些描述计算文本嵌入，可以隐式地在医学实体之间建立关系；

技术

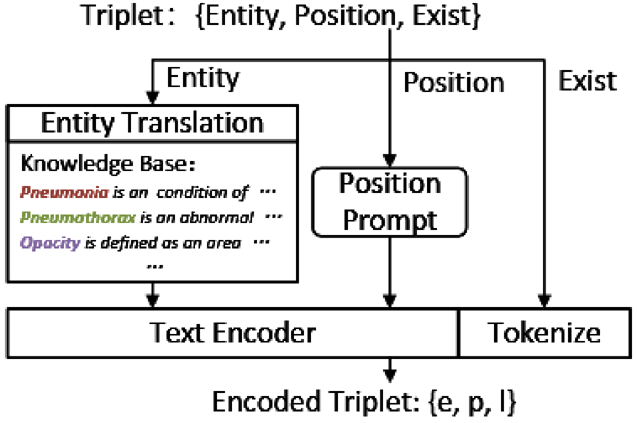
Exist 编码：我们使用 $l \in \{0,1,-1\}$，$1 \in \{0,1,-1\}$ 来表示其中的“存在性”，其中 1 表示存在（True），0 表示不存在（False），-1 表示不确定。

Entity 编码：我们通过查询一些容易访问的医学知识库来将其转化为详细描述，例如，Description([“Pneumonia”]) = “它是一种主要影响肺部的疾病……表现为不透明和胸腔积液……”

动机

尽管这种转换方法简单，但将实体转化为描述对于更可靠的零样本诊断至关重要，因为它进一步将专业医学实体分解为不同疾病共享的基本属性，促使模型对视觉证据进行更深入的理解。

B. Knowledge-enhanced Triplet Encoding


Triplet：{Entity, Position, Exist}
Entity
Position
Exist
Entity Translation
Knowledge Base:
Pneumonia is an condition of …
Pneumothorax is an abnormal …
Opacity is defined as an area …
…
Position Prompt
Text Encoder
Tokenize
Encoded Triplet: {e, p, l}

技术

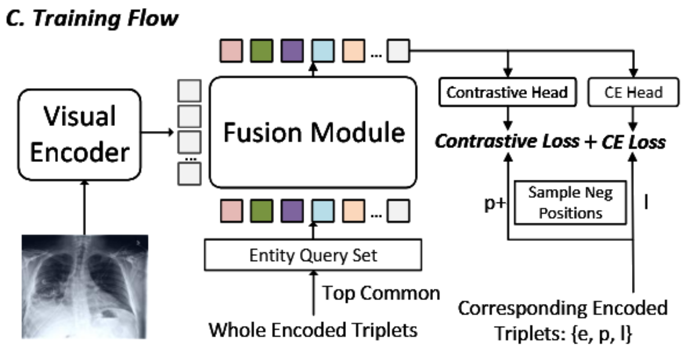
Position 编码：对于“位置”词汇，我们使用一个提示语句：“它位于{位置}”来形成句子。最后，我们使用预训练的文本编码器 ClinicalBERT 来计算“实体”和“位置”的嵌入向量，并使用线性多层感知机（MLP）将嵌入映射到所需的维度：

动机

讨论：提取的实体是医学术语，只有具备医学背景的观众才能理解，而通过详细的描述丰富这些实体有助于模型深入理解疾病的视觉证据。这样的模式可以跨疾病进行泛化，因为许多属性描述通常是共享的，从而使得模型能够建立已见类之间的隐式关系，并理解未见类的描述。

方法学

3.3 fusion module



动机

【填坑三】第三，我们将实体视为查询集合，并采用基于 Transformer 的架构，将图像补丁与实体描述对齐，从而在实体层面提供明确的监督信号。因此，我们可以通过空间热图的形式同时推断某些疾病的可能性，即为可解释性提供粗略的定位。

动机

通过报告中的三元组，我们可以在实体级别对模型进行监督，而不是在整个报告级别进行监督。三元组中的“位置”和“存在性”部分可以自然地视为更细粒度的监督标签。

技术

具体来说，我们采用基于 Transformer 的架构，使用实体的嵌入作为查询，迭代地关注图像的嵌入，并输出实体的存在性和位置预测。

$$\hat{s}, \hat{p}, \hat{m} = \Phi_{\text{fusion}}(V, Q),$$

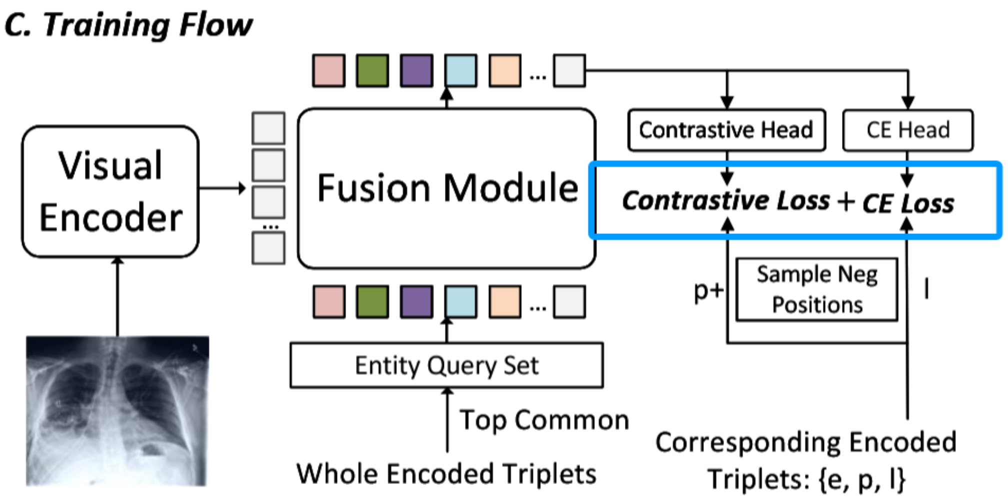
（就是把实体描述作为query，图像嵌入作为key和value，做一个交叉注意力机制，输出就是一个融合特征和一个注意力得分，再加个预测头，用融合特征分别预测实体的存在s、实体的位置p）

动机

讨论：采用 Transformer 解码器使得能够在图像的补丁级别计算实体与图像之间的对应关系。因此，图像特征 V 更适合下游的分割任务，并且每一层中的交叉注意力图的平均值可以直接用于零样本定位，为诊断提供可解释性

方法学

3.4 training



（部分BCE：）对于存在性预测 $\hat{s}$，我们使用与对应“存在”标签 l的二元交叉熵，如果 l为 -1，则直接跳过该样本，记为 L_{cls}

为了监督每个实体查询的位置信息预测，我们采用对比学习。我们首先从位置集合中选择出现频率最高的 $|P|$ 个位置嵌入作为位置集合， $P=\{p_1, p_2, \ldots, p_{|P|}\}$。然后随机从中抽取 M个负样本的位置嵌入，并将三元组中的相应位置嵌入 p作为正样本：

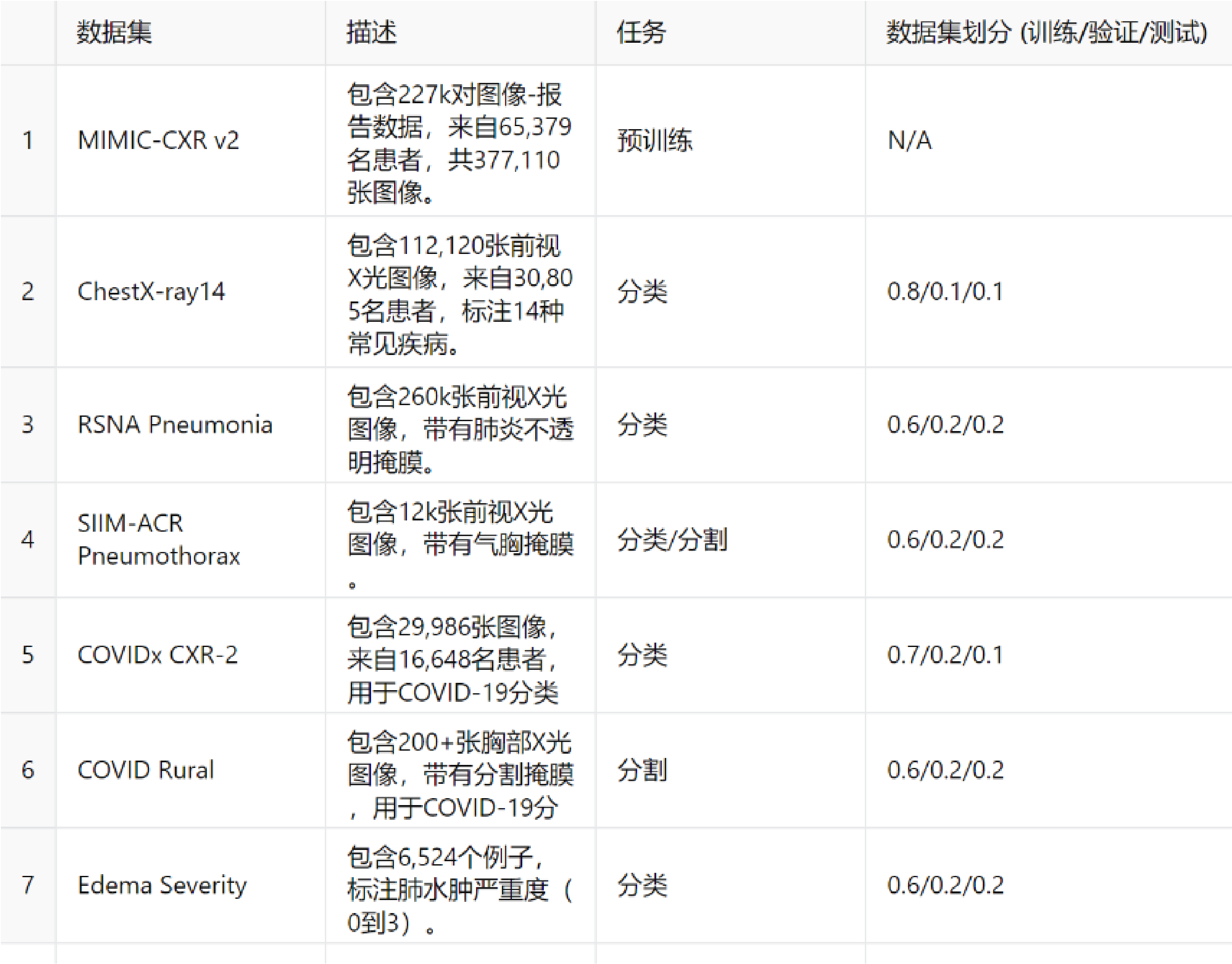
$$\mathcal{L}_{\text{loc}} = -\frac{1}{|Q|}\sum_{k=1}^{|Q|}\frac{e^{\langle\hat{p}_k, p_k\rangle}}{e^{\langle\hat{p}_k, p_k\rangle}+\sum_{u=1}^{M}e^{\langle\hat{p}_k, P_{\mathcal{I}(k,u)}\rangle}},$$

方法学

3.5 inferfence

在推理时，给定一张测试图像，我们可以直接推断出某些实体/疾病的存在性，并定位它们的视觉证据。具体来说，对于在实体查询集 Q 中出现的实体，我们直接采用 Q 中对应的元素；而对于那些未见过的实体，我们用用户提供的简短描述替换该实体，并将其作为一个额外的查询添加到实体查询集 Q 中，类似于零样本推理。存在性输出 $\hat{s}$ 可以直接用于分类，目标实体与视觉特征之间的平均交叉注意力 $\hat{m}$ 用于定位。

实验结果-数据集



	数据集	描述	任务	数据集划分 (训练/验证/测试)
1	MIMIC-CXR v2	包含227k对图像-报告数据，来自65,379名患者，共377,110张图像。	预训练	N/A
2	ChestX-ray14	包含112,120张前视X光图像，来自30,805名患者，标注14种常见疾病。	分类	0.8/0.1/0.1
3	RSNA Pneumonia	包含260k张前视X光图像，带有肺炎不透明掩膜。	分类	0.6/0.2/0.2
4	SIIM-ACR Pneumothorax	包含12k张前视X光图像，带有气胸掩膜。	分类/分割	0.6/0.2/0.2
5	COVIDx CXR-2	包含29,986张图像，来自16,648名患者，用于COVID-19分类	分类	0.7/0.2/0.1
6	COVID Rural	包含200+张胸部X光图像，带有分割掩膜，用于COVID-19分	分割	0.6/0.2/0.2
7	Edema Severity	包含6,524个例子，标注肺水肿严重度（0到3）。	分类	0.6/0.2/0.2

实验结果-实施细节

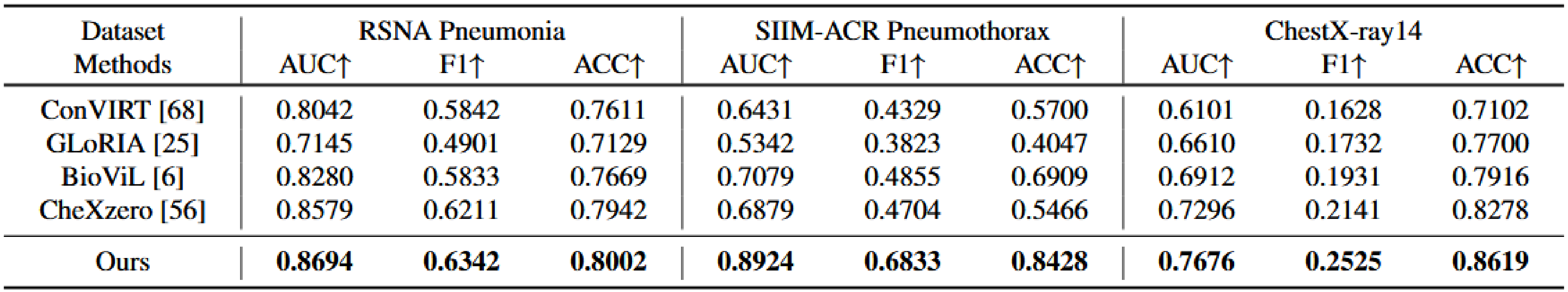
在预训练阶段，三元组提取模块和用于三元组编码的文本编码器都是固定的，而视觉编码器和融合模块则在图像-文本对上进行端到端训练。

在微调阶段，我们采用初始化为图像编码器的ResNet50 [24]进行分类，并使用我们的预训练图像编码器初始化ResUNet [16]的编码器进行分割。

我们与现有的多种最先进的医学图像-文本预训练方法进行比较，具体包括ConVIRT [68]、GLoRIA [25]、BioViL [6]和CheXzero [56]。由于ConVIRT和GLoRIA是在内部数据集上预训练的，为了公平比较，我们在MIMIC-CXR数据集上重新训练了它们的模型。对于BioViL，我们使用作者公开发布的模型。

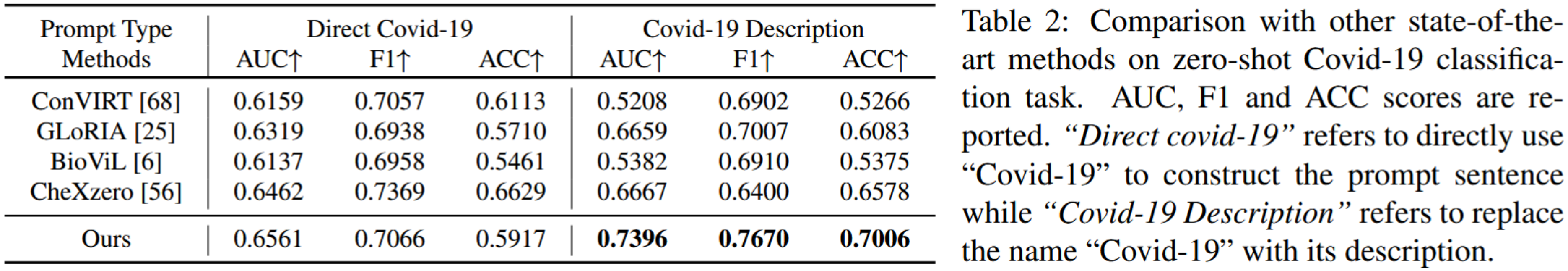
在零-shot设置下，我们使用BioViL [6]中提到的提示，并与最近的方法（CheXzero [56]）进行比较，后者在零-shot诊断能力上已被证明优于放射科医生。

实验结果-零样本能力（分类）



Dataset Methods	RSNA Pneumonia AUC↑	F1↑	ACC↑	SIIM-ACR Pneumothorax AUC↑	F1↑	ACC↑	ChestX-ray14 AUC↑	F1↑	ACC↑
ConVIRT [68]	0.8042	0.5842	0.7611	0.6431	0.4329	0.5700	0.6101	0.1628	0.7102
GLoRIA [25]	0.7145	0.4901	0.7129	0.5342	0.3823	0.4047	0.6610	0.1732	0.7700
BioViL [6]	0.8280	0.5833	0.7669	0.7079	0.4855	0.6909	0.6912	0.1931	0.7916
CheXzero [56]	0.8579	0.6211	0.7942	0.6879	0.4704	0.5466	0.7296	0.2141	0.8278
Ours	**0.8694**	**0.6342**	**0.8002**	**0.8924**	**0.6833**	**0.8428**	**0.7676**	**0.2525**	**0.8619**

已见疾病：们的模型将RSNA肺炎数据集的AUC得分从0.83提升至0.87，将SIIMACR气胸数据集的AUC得分从0.71提升至0.89，如表1所示。这表明我们的方法能更好地处理医学中的多中心和多疾病数据分布。



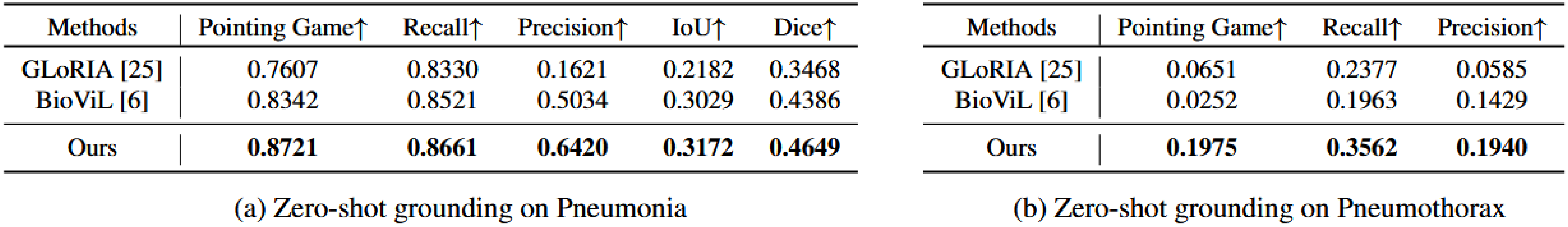
Prompt Type Methods	Direct Covid-19 AUC↑	F1↑	ACC↑	Covid-19 Description AUC↑	F1↑	ACC↑
ConVIRT [68]	0.6159	0.7057	0.6113	0.5208	0.6902	0.5266
GLoRIA [25]	0.6319	0.6938	0.5710	0.6659	0.7007	0.6083
BioViL [6]	0.6137	0.6958	0.5461	0.5382	0.6910	0.5375
CheXzero [56]	0.6462	0.7369	0.6629	0.6667	0.6400	0.6578
Ours	0.6561	0.7066	0.5917	**0.7396**	**0.7670**	**0.7006**

Table 2: Comparison with other state-of-the-art methods on zero-shot Covid-19 classification task. AUC, F1 and ACC scores are reported. *"Direct covid-19"* refers to directly use "Covid-19" to construct the prompt sentence while *"Covid-19 Description"* refers to replace the name "Covid-19" with its description.

未见疾病：COVID-19是一种新疾病，首次出现于2019年，2015年收集的MIMIC-CXR报告中没有涉及COVID-19的任何数据实体，因此它要求系统具备诊断真正未见疾病的能力。如表2所示，仅依赖疾病名称的现有方法在做出正确诊断时遇到困难。而我们提出的方法，在引入医学知识后，即使用实体描述，可以理解训练集中未见过的复杂医学实体描述，并显著提高性能：AUC从0.66提升至0.74，ACC从0.59提升至0.70，证明了实体翻译对于未见疾病的诊断至关重要。

实验结果-零样本能力（定位grounding）

除了简单的诊断外，可解释性在医疗保健中同样至关重要，它可以提高机器学习系统的可靠性和可信度。在这里，我们考虑通过在预测中定位异常来提供可解释性，并与现有方法进行比较



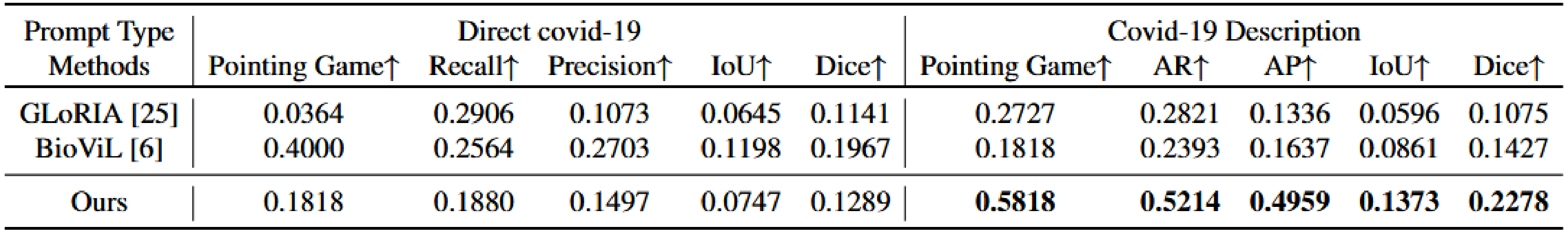
Methods	Pointing Game↑	Recall↑	Precision↑	IoU↑	Dice↑
GLoRIA [25]	0.7607	0.8330	0.1621	0.2182	0.3468
BioViL [6]	0.8342	0.8521	0.5034	0.3029	0.4386
Ours	**0.8721**	**0.8661**	**0.6420**	**0.3172**	**0.4649**

(a) Zero-shot grounding on Pneumonia

Methods	Pointing Game↑	Recall↑	Precision↑
GLoRIA [25]	0.0651	0.2377	0.0585
BioViL [6]	0.0252	0.1963	0.1429
Ours	**0.1975**	**0.3562**	**0.1940**

(b) Zero-shot grounding on Pneumothorax

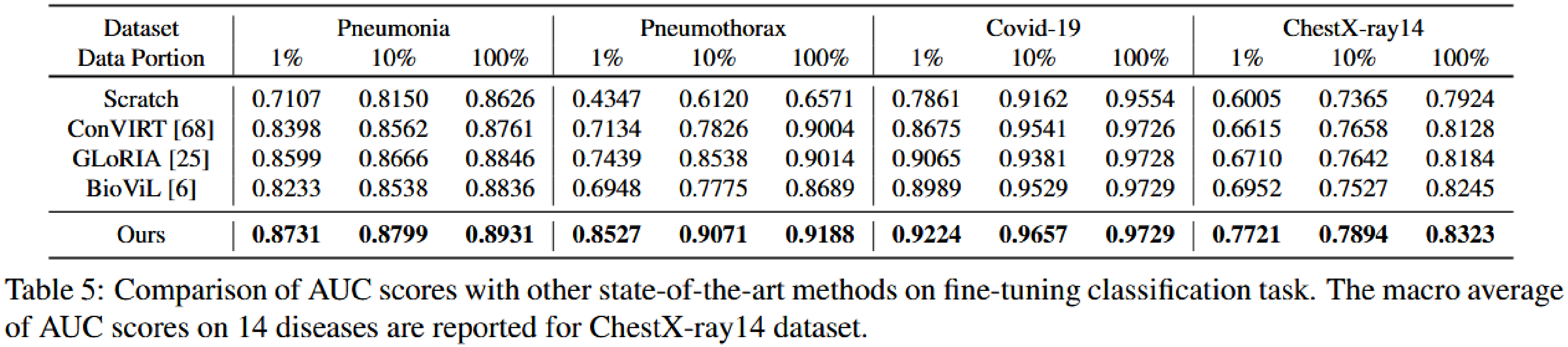
已见疾病：我们将指示游戏得分从0.83提升至0.87，检测召回率从0.85提升至0.87，检测精度从0.50提升至0.64，IOU从0.30提升至0.32，Dice系数从0.44提升至0.46。而在SIIM-ACR数据集（表3b）上，气胸区域往往较薄且狭窄，定位其位置通常比不透明性定位更具挑战性[6]，因此我们只考虑指示游戏得分、召回率和精度。类似地，我们的方法在这些指标上表现明显优于之前的方法。（删指标）



Prompt Type	Direct covid-19					Covid-19 Description				
Methods	Pointing Game↑	Recall↑	Precision↑	IoU↑	Dice↑	Pointing Game↑	AR↑	AP↑	IoU↑	Dice↑
GLoRIA [25]	0.0364	0.2906	0.1073	0.0645	0.1141	0.2727	0.2821	0.1336	0.0596	0.1075
BioViL [6]	0.4000	0.2564	0.2703	0.1198	0.1967	0.1818	0.2393	0.1637	0.0861	0.1427
Ours	0.1818	0.1880	0.1497	0.0747	0.1289	**0.5818**	**0.5214**	**0.4959**	**0.1373**	**0.2278**

未见疾病：我们还对未见疾病——即COVID-19进行了零-shot基础定位实验，如表4所示。我们的模型在所有指标上均表现出一致的提升，例如，指示游戏得分从0.40提升至0.58。

实验结果-微调能力（分类）

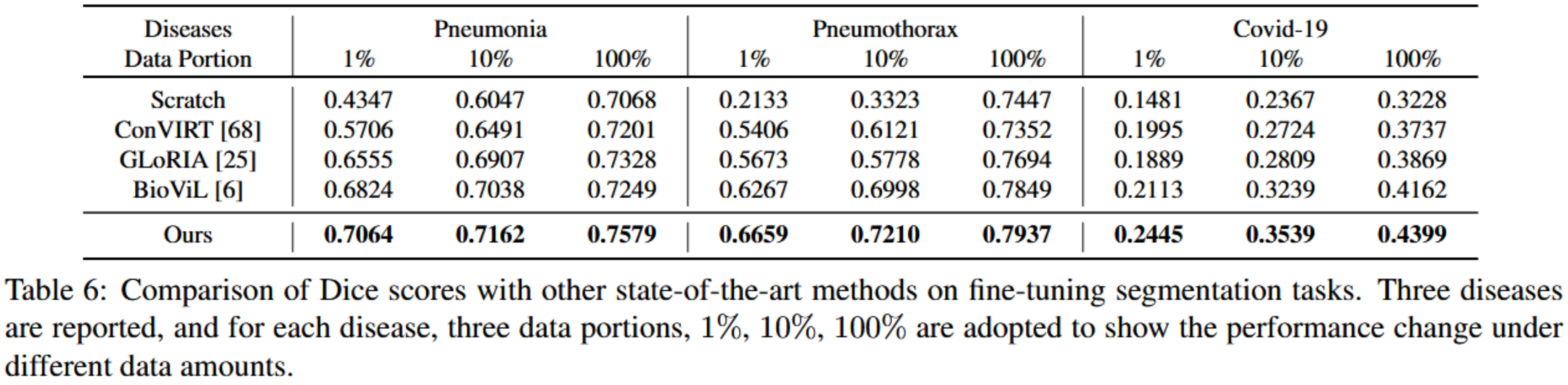


Dataset	Pneumonia			Pneumothorax			Covid-19			ChestX-ray14		
Data Portion	1%	10%	100%	1%	10%	100%	1%	10%	100%	1%	10%	100%
Scratch	0.7107	0.8150	0.8626	0.4347	0.6120	0.6571	0.7861	0.9162	0.9554	0.6005	0.7365	0.7924
ConVIRT [68]	0.8398	0.8562	0.8761	0.7134	0.7826	0.9004	0.8675	0.9541	0.9726	0.6615	0.7658	0.8128
GLoRIA [25]	0.8599	0.8666	0.8846	0.7439	0.8538	0.9014	0.9065	0.9381	0.9728	0.6710	0.7642	0.8184
BioViL [6]	0.8233	0.8538	0.8836	0.6948	0.7775	0.8689	0.8989	0.9529	0.9729	0.6952	0.7527	0.8245
Ours	**0.8731**	**0.8799**	**0.8931**	**0.8527**	**0.9071**	**0.9188**	**0.9224**	**0.9657**	**0.9729**	**0.7721**	**0.7894**	**0.8323**

Table 5: Comparison of AUC scores with other state-of-the-art methods on fine-tuning classification task. The macro average of AUC scores on 14 diseases are reported for ChestX-ray14 dataset.

我们在四个不同的数据集上进行了实验，使用1%、10%和100%的数据进行微调，这与现有的工作[68, 25, 6]一致。如表5所示，我们的模型在所有数据集上展示了显著的AUC得分提升，反映出我们的预训练表示相比现有模型具有更高的质量。

实验结果-微调能力（分割）

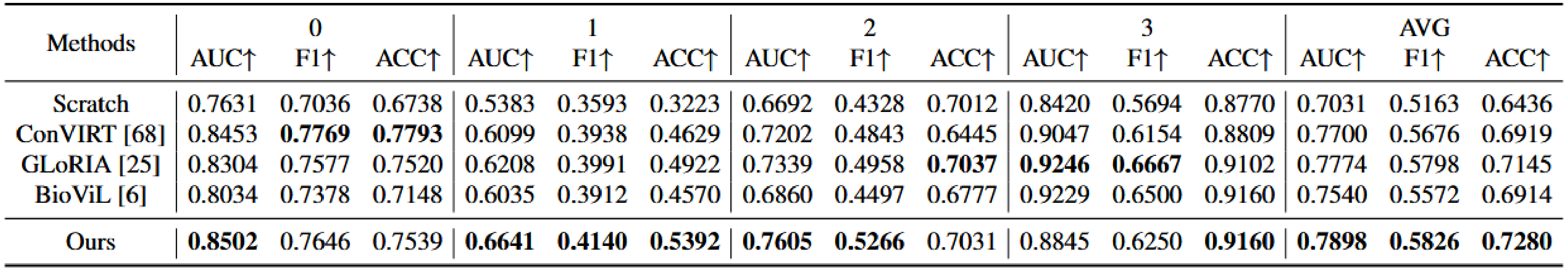


Diseases	Pneumonia			Pneumothorax			Covid-19		
Data Portion	1%	10%	100%	1%	10%	100%	1%	10%	100%
Scratch	0.4347	0.6047	0.7068	0.2133	0.3323	0.7447	0.1481	0.2367	0.3228
ConVIRT [68]	0.5706	0.6491	0.7201	0.5406	0.6121	0.7352	0.1995	0.2724	0.3737
GLoRIA [25]	0.6555	0.6907	0.7328	0.5673	0.5778	0.7694	0.1889	0.2809	0.3869
BioViL [6]	0.6824	0.7038	0.7249	0.6267	0.6998	0.7849	0.2113	0.3239	0.4162
Ours	**0.7064**	**0.7162**	**0.7579**	**0.6659**	**0.7210**	**0.7937**	**0.2445**	**0.3539**	**0.4399**

Table 6: Comparison of Dice scores with other state-of-the-art methods on fine-tuning segmentation tasks. Three diseases are reported, and for each disease, three data portions, 1%, 10%, 100% are adopted to show the performance change under different data amounts.

我们对三种不同疾病进行了分割的微调实验。我们选择了1%、10%和100%的数据进行微调。对于这三种具有不同图像分布的疾病，我们的方法在所有指标上都显著超越了现有的最先进方法，尤其是在低数据量的情况下。

实验结果-微调能力（分级,多分类）



Methods	0			1			2			3			AVG		
	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑	AUC↑	F1↑	ACC↑
Scratch	0.7631	0.7036	0.6738	0.5383	0.3593	0.3223	0.6692	0.4328	0.7012	0.8420	0.5694	0.8770	0.7031	0.5163	0.6436
ConVIRT [68]	0.8453	**0.7769**	**0.7793**	0.6099	0.3938	0.4629	0.7202	0.4843	0.6445	0.9047	0.6154	0.8809	0.7700	0.5676	0.6919
GLoRIA [25]	0.8304	0.7577	0.7520	0.6208	0.3991	0.4922	0.7339	0.4958	**0.7037**	**0.9246**	**0.6667**	0.9102	0.7774	0.5798	0.7145
BioViL [6]	0.8034	0.7378	0.7148	0.6035	0.3912	0.4570	0.6860	0.4497	0.6777	0.9229	0.6500	0.9160	0.7540	0.5572	0.6914
Ours	**0.8502**	0.7646	0.7539	**0.6641**	**0.4140**	**0.5392**	**0.7605**	**0.5266**	0.7031	0.8845	0.6250	**0.9160**	**0.7898**	**0.5826**	**0.7280**

Table 7: Comparison with other state-of-the-art methods on fine-tuning edema severity grading multi-class classification task. AUC score is reported in the Table. “0,1,2,3” in the table represents the severity level and final average scores are reported.

此外，疾病严重程度的分级也发挥着重要作用。在这里，我们采用我们的预训练特征，并将其用于多分类任务，0到3代表不同的严重程度级别。如表7所示，对于每个级别，AUC、F1和ACC得分是作为一个类别与其他类别进行计算的，例如，0与{1, 2, 3}比较。最终计算四个级别的宏观平均得分。在大多数严重程度级别上，我们的方法能够取得最佳结果。

总结

本文提出一种医学知识增强型视觉 - 语言预训练（VLP）模型，核心工作包括三部分：

1）通过三元组提取模块挖掘医疗相关三元组作为增强监督信号，简化原始报告并减少信息损失；

2）将三元组实体转化为详细医学描述并经文本编码器嵌入，助力网络理解专家级医学知识；

3）设计基于 Transformer 的结构实现局部区域对齐。

实验验证表明，该模型在不同数据集和设置下表现优异，具备强零样本分类与定位能力（可应对未见疾病），微调后仍显著优于现有最先进方法，凸显其技术优越性。

但是这个论文为什么没有可视化的图?

好词好句

- 1) With the rapid development of deep learning, numerous works have been proposed to facilitate computer-aided diagnosis in the medical field [46, 20, 55, 19].
  - With the rapid development of … numerous works have been proposed to …
  - 随着…的迅速发展，大量的工作被提出以…

- 2) Despite the tremendous progress, these models are normally trained to recognize or segment the structures that fall into a certain closed set of anatomical or disease categories
  - 描述现有技术的转折表达， tremendous（巨大的）

好词好句

- 3) As an alternative, recent research considers to train the model on the corpus, consisting of large amount of multi-modal data, that is generated from daily clinical routine
    - 创新性表达
    - As an alternative：作为一个替代的选择

- 4) This paper presents our preliminary investigation on vision-language representation learning in the medical domain, with the goal of better zero-shot disease diagnosis (classification) and grounding. Undoubtedly, these tasks have also been widely investigated in the computer vision community
    - 一个副词搞定自然图像域和医学图像域
    - 我们研究了医学的xx问题，毋庸置疑，这个问题在自然图像中已经广泛被研究，但…

好词好句

- 5) Consequently, we can simultaneously infer the likelihood of certain diseases with the visual evidence in the form of a spatial heatmap, i.e., providing rough grounding for explainability.
    - Consequently：因此
    - Simultaneously： 同时

- 6) We pre-train the model on one widely-used medical image-report dataset MIMIC-CXR [31], and rigorously evaluate on the task of disease diagnosis across numerous public benchmarks
    - rigorously evaluate
    - 严格评估，表达有气势

好词好句

- 7) One observation to be noticed is that, results in Tab. 4 are mostly consistent with those in Tab. 2
  - 强调实验发现表达
  - One observation to be noticed is that

- 8) First, we propose a triplet extraction module to extract useful medical-related triplets as more useful supervision signals, simplifying complex raw reports with minimal information loss. Second, we translate the entities in extracted triplets into detailed medical descriptions and embed them with a text encoder enabling the network to understand complex medical expert-level knowledge. Finally, a transformer-based structure is proposed to do local region alignment.
  - 表达的变换，用被动语态防止语言过于单调

好词好句

- 9) Despite the tremendous progress, these models are normally trained to recognize or segment the structures that fall into a certain closed set of anatomical or disease categories
  - 用fall into 表达现有结构陷入到xxx困境

- 10) a costly procedure for data annotation, model re-training are required, fundamentally limiting its practical values.
  - 根本上限制了xxx的价值
  - fundamentally

观点论据

- 1) First, data availability, training Foundation Models in computer vision normally require over millions of image-text pairs, while in the medical domain, only a few hundred thousand pairs are available . The limited data challenges language models to understand the reports in free form .

- **现有医学多模态模型的困境一：数据稀缺**
    - 首先是数据可用性问题：计算机视觉领域训练基础模型通常需要数百万张图像-文本配对数据，而在医学领域仅有数十万对数据可用。数据的局限性使得语言模型难以理解自由格式报告。

观点论据

- 2) Second, the problem considered in computer-aided diagnosis is naturally fine-grained, that requires distinguishing the medical concepts to understand the disease, as a consequence, domain knowledge is essential;
- **现有医学多模态模型的困境二：医学知识**
    - 其次，计算机辅助诊断所涉及的问题本质上具有细粒度特征，需要区分医学概念以理解疾病本质，因此领域知识至关重要；

观点论据

- 3) Third, robustness is crucial, it is, therefore, preferable to have explainability, where diagnosis results come along with the visual grounding, to help radiologists understand the system, and build trust between human and machines.
- 现有医学多模态模型的困境三：可解释性（鲁棒性）
  - 第三，鲁棒性至关重要。因此，具备可解释性更为理想——即诊断结果需附带可视化依据，这有助于放射科医生理解系统运作机制，从而建立人与机器之间的信任。

观点论据

- 4) Existing work in medical VLP (Vision-Language Pretraining) follows a straightforward training paradigm by matching raw reports with image scans, as shown in Fig.1A, ignoring the medical prior knowledge.

- First, unlike existing works that directly process the raw reports, we adopt a novel triplet extraction module to extract the medical-related information, avoiding unnecessary complexity from language grammar and enhancing the supervision signals;

- In contrast to natural texts, information in medical reports tends to be more condensed, with radiologists pointing out the existence of abnormality and their positions in the image. Meanwhile, medical terminologies tend to be professional.

- **为什么要在诊断报告中提取概念？而不是直接对齐整个报告。**
  - 目前现有医学视觉语言预训练（VLP）研究采用直接匹配原始报告与图像扫描的简单训练范式（如图1A所示），忽略了医学先验知识。
  - 首先，不同于现有直接处理原始报告的方法，我们采用创新的三元组提取模块提取医学相关信息，既规避了语言语法带来的冗余复杂性，又增强了监督信号；
  - 相较于自然文本，医学报告信息更为凝练——放射科医师仅指出异常存在及其在图像中的位置，且医学术语具有高度专业性。

观点论据

- 5) Despite its simplicity, converting the entities into descriptions is crucial for more reliable and zero-shot diagnosis, as it further decomposes the professional medical entities into basic attributes that are shared by different diseases, encouraging the model to capture a deep understanding of the visual evidence.
- **为什么要把概念转换成描述**
  - 尽管过程简单，将实体转化为描述对实现更可靠的零样本诊断至关重要。此举能将专业医学实体进一步分解为不同疾病共有的基本属性，从而促使模型深入理解视觉证据。

问题

1. MedKLIP在医学图像诊断中的主要目标是什么？它与传统的医学图像诊断方法有何不同？
2. MedKLIP如何通过增强医学视觉语言预训练（VLP）来提高X光图像的诊断性能？
3. 在MedKLIP方法中，Triplet提取模块的作用是什么？它如何帮助简化放射学报告的处理？
4. 实体翻译模块在MedKLIP中的作用是什么？它是如何增强模型对医学领域的理解的？
5. MedKLIP的Transformer融合模型如何在空间上对医学实体描述和视觉信号进行对齐？
6. MedKLIP使用了哪些医学数据集进行模型评估？这些数据集的特点是什么？
7. 在医学领域应用VLP时，MedKLIP面临了哪些主要挑战？如何克服这些挑战？
8. 如何通过Triplet提取模块简化放射学报告的复杂语言，同时保留报告中的有效信息？
9. MedKLIP在数据稀缺的情况下是如何训练模型的？它如何利用有限的数据提高模型的性能？
10. 与传统医学VLP方法相比，MedKLIP在哪些方面展现了显著的改进？
11. MedKLIP中的Triplet编码过程是如何增强模型对医学图像和报告的理解的？
12. 什么是“知识增强的Triplet编码”模块？它如何利用医学知识库来改善模型的表现？
13. Transformer解码器在MedKLIP的融合模块中扮演了什么角色？它如何帮助实现更精细的实体级监督？
14. 在疾病分类和定位任务中，MedKLIP使用了哪些评估指标？这些指标如何帮助衡量模型的性能？
15. 与现有的医学图像-文本预训练方法相比，MedKLIP在零样本分类和定位任务中表现如何