# 多模态大模型在医疗领域的应用研究

高朗 CS2110 U2021XXXXX
计算机科学与技术学院，华中科技大学

## 一、前言

2022-2023 年，大模型(Large Model,LM)逐渐普及，LlAMA,GLM 等系列大语言模型以及 LlAVA 等多模态大模型开源。这些模型的出现，为提高社会生产效率和生产水平产生了重要作用。大语言模型(Large language model,LLM)能够接受自然语言文本，并且输出相应的文本回复；多模态大语言模型(Multimodal large language model,MLLM，限于本文讨论范围以下简称为多模态大模型)能够同时处理多种模态（如图像、文本、语音等）的数据。这些模型所具备的强大表示能力和泛化能力让大模型在医疗领域的应用成为可能。

在 2023 年，如 Zhongjing（国内）、MedAlpaca(国际)等医疗大语言模型发行。这些模型知晓医疗领域的知识，能够根据用户的描述，解决一些基本的医学问题，并给出合理的医学建议。然而，这些模型目前能够适用的领域仅限于日常医疗问答，在临床投入使用存在较大困难。这主要是由医疗领域的特殊性造成的。

首先，医学数据的多样性和复杂性要求模型能够从多个角度对数据进行分析和理解。传统的大语言模型往往只能处理一种类型的数据，难以充分利用医学数据中蕴含的丰富信息。其次，临床医疗面临的问题往往难以用文字准确表述，大语言模型目前无法对问题给出细粒度、高专业性的分析；最后，医疗领域具有高度的敏感性，模型的输出关乎诊断结果和患者后续治疗工作的开展，大语言模型存在的幻觉(hallucination)问题和垂直领域的可解释性分析方法尚不充分的问题，均对模型给出精确有效的回答造成了阻碍。因此，将多模态大模型应用于医疗领域成为了一种可行的解决方案。

多模态大模型能够同时处理多种类型的数据，从而更好地挖掘出数据中的关联性和潜在规律。其次，多模态大模型可以提高医疗诊断和治疗的准确性和效率。通过将不同类型的数据进行融合和整合，多模态大模型可以提供更全面、准确的信息，帮助医生做出更准确的诊断和制定更有效的治疗方案。多模态大模型由于集成了图像分析功能，可以图像可视化并且文字描述判断依据，这也使得相比大

语言模型，多模态大模型拥有更强的可解释性。

然而，多模态大模型在医疗领域的应用也面临着一些挑战。具体包含以下几个方面：

**训练数据：**医疗领域多模态大模型的训练需要大量的数据。通常这类模型使用来自会议、书籍和诊断记录中收集到的图文数据作为训练语料，但是直接收集得到的数据存在包含噪音、图文不匹配、内容混乱等问题，这对模型训练后的性能带来了消极影响。如何恰当地清洗数据、能否寻找到更加高质量的医疗数据集，成为相对重要的课题。

**模型：**医疗领域大模型作为多模态大模型的一种，本身基本沿袭了多模态大模型的基本结构，但是考虑到医疗领域高精确度、高专业性的需求，部分研究也尝试调整模型结构，以获得某一方面能力的显著提升；同时，多模态大模型拥有许多可能的训练方式，如继续预训练(continue-pretraining)、微调(fine-tune)和仅进行预训练(pretraining)等。采用何种训练手段和使用方法能够让模型在疾病分类、病灶定位和诊断等任务上的表现更加优秀，也成为相关研究尝试的方向。

**评估方式：**医疗领域的专业性决定了其需要采用更加准确和客观的评价标准。但是，现有的评估数据集存在重复、答案错误和内容区分度不高等问题，导致模型的评估效果虚高。此外，医疗领域的评估更加看重模型关键信息提取的能力，这使得研究者重新考虑使用文本相似度匹配来反映模型性能的合理性。另外，针对不同的问题，研究者也需要设计不同的评估指标，以更加客观准确地描述模型的性能。

本综述主要涉及 2022-2024 年间有关医疗领域多模态大模型的研究工作。文章后续部分将大体按照论文发表时间先后顺序，首先介绍相关研究的研究成果，其次介绍这些研究采用的模型与训练方式与训练数据两个方面的工作，最后将对比这些研究，总结优势与不足，并且给出医疗领域多模态大模型的机遇和挑战。对于文中提到的专业概念，可以根据下标于附录中查找。

## 二、医疗领域多模态大模型研究现状

### 2.1 相关研究成果

MGCA(Multi-Granularity Cross-modal Alignment)[6]是一种多粒度的跨模态对齐方案。该研究的提出主要是为了解决多模态模型在放射图像识别中存在的精度不高的问题。在使用此类框架后，多模态模型在放射图像识别、分类以及语义分割等方面均表现出了优异的性能，这说明多粒度语义信息的综合提取确实有利于使多模态大模型获得优秀的图像理解能力，从而提升模型的回答精度。

PLIP（Pathology Language-Image Pre-training，病理学语言-图像预训练）[3]模型是斯坦福大学研究团队于2023年4月推出的一款病理学多模态大模型。PLIP具有理解文本和图像的能力，同时也被发现拥有 zero-shot[1]理解能力和迁移学习的能力，能够在不同的任务背景下对新的病理学图像进行分类。此外，PLIP 模型还支持检索功能，用户可以用病理学图像或自然语言来检索相似的案例。该研究的亮点在于，使用了从病理学领域社交平台收集到了高质量图文数据集 OpenPath 来训练模型，证明了在公众平台上传播的病理学知识是潜在的高质量数据集。

BiomedGPT[7]是 2023 年 5 月推出的多模态大模型，也是首个开源的生物医学领域通用大模型。BiomedGPT 在 26 个数据集上，针对一些重要的生物医学任务进行了充分的测试，并且在其中 16 个任务中达到了最好水准(state-of-the-art)，具体来说，BiomedGPT 在放射图像人类评估任务中超越了 OpenAI 发布的 GPT-4V。BiomedGPT 同样具有优秀的 zero-shot 和迁移学习能力。该研究说明了：使用不同的数据集针对不同的任务分别训练能够获得更加实用的生物医学模型。

LLaVA-Med[8]是微软于 2023 年 6 月推出的一款生物医学多模态大模型，在下游任务进行微调之后，性能超过了此前监督训练的最优模型。更加重要的是，LLaVA-Med 仅实用 15 个小时便训练完毕，这得益于其独创性的训练方法以及 self-instruct 开放式数据的数据处理思想。

CONCH(constructive learning from captions for histopathology,从组织病理学的图片标头对比学习)[4]是一个组织病理学的优秀基座模型。CONCH 可以通过微调开展广泛的下游任务，包括组织病理学图像和文本相关任务。其在组织学图像分类、分割、字幕、文本到图像和图像到文本检索方面取得了最先进的性能(优于 PLIP)。CONCH 代表了对用于组织病理学的视觉语言预训练系统的重大飞跃，有可能直接促进广泛的基于机器学习的工作流，使得它们在仅进行一次或者不进

行微调就能胜任广泛的下游任务。

Med-Flamingo[1]是斯坦福大学于 2023 年 7 月推出的医药学多模态大模型，其主要能力在于含图像的问答任务(VQA[2])。本研究同时首次采用专家评估的方式检验模型的性能。此外，该模型有能力根据 few-shot[3]的提示改变生成范式，比如在 few-shot 条件下给自己的分析结果提供合理解释 rationale[4]等等。其表现如图 1 所示。

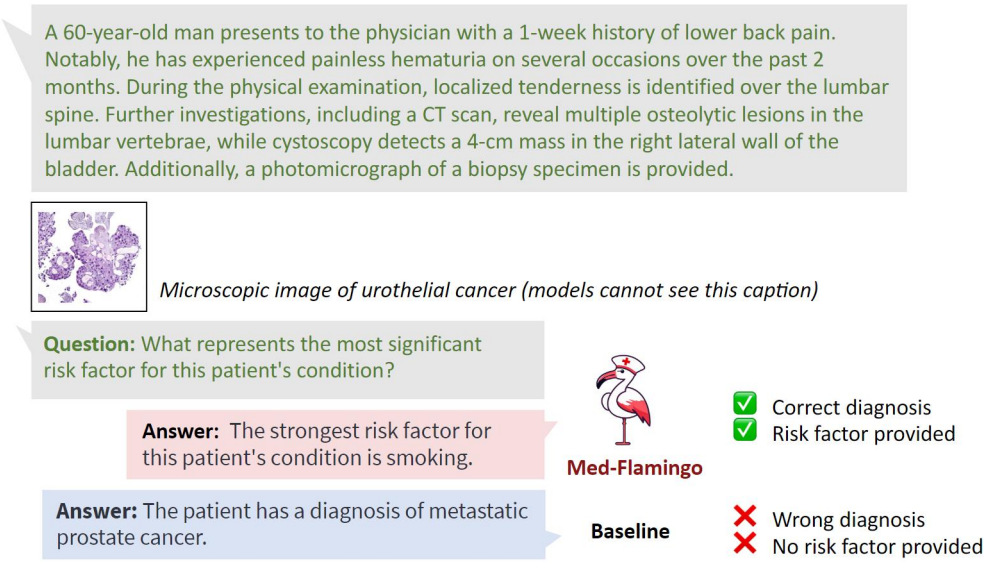



图 1：Med-Flamingo 表现举例

PathChat[5]是 2023 年 12 月推出的病理学多模态大模型，使用了内置视觉编码器，结合海量数据训练获得了优异的性能。该模型擅长交互式对话，其对话能力经过专家测试后，被证明更加具有参考价值和准确性。此外，PathChat 也在多项任务中表现出全面而优秀的强大性能，具体来说，在多项任务中超越或接近 GPT-4V，几乎完全超越 LLaVA-Med。该研究有望为医疗事业、医学教育事业等起到重要作用。图 2 展示了在 4 种不同任务中，PathChat 与对照组的性能对比。

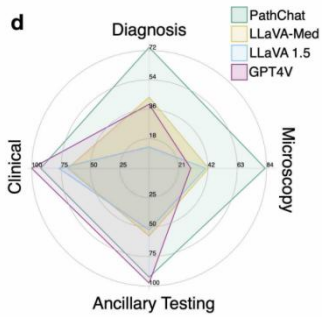



图 2 PathChat 与对照实验结果对比

PeFoMed[2]是 2024 年 1 月由哈尔滨工程大学推出的多模态大模型，其采用了高效参数微调的方式训练模型，在诸多任务中表现出比 GPT-4V 更强大的性能。PeFoMed 主要专注于解决的问题是医药领域的 VQA 任务，专家评测其总体准确率达到 81.9%，表现出强大的能力。

## 2.2 模型与训练方式

### 2.2.1 CLIP[13]

医疗垂直领域多模态大模型，主要都是根据经典多模态大模型架构改进而来。经典多模态模型的代表是 CLIP 模型,CLIP 的全称为 Contrastive Language-Image Pre-Training（语言-图像对比预训练），是通过对比学习的方式让模型同时学习文本、图像信息，以及二者之间的对应关系。CLIP 的训练和使用方式简要描述如下：

首先，选择合适的文本编码器(可以使用可训练的模型 ResNet[10])和图像编码器(在 Visual Transformer[12]问世之后，模型将图像像文本一样编码为向量成为可能)；其次，准备若干图片-文本对，并且分批次进行编码，得到编码表征(encoder representation)；之后，将同一个批次下的文本表征和图像表征线性投影到多模态嵌入空间(multi-modal embedding space)：可以理解为多模态嵌入空间为一个 batchsize*batchsize 的矩阵，矩阵中每一个单元格的横纵坐标代表一个图片的嵌入向量和一段文本的嵌入向量。矩阵中存放的内容为这两个嵌入向量的内积，我们此处认为，内积反映了两个定长向量的重合程度，也就是图文的相关程度，内积越大，图文相关性越强。从图 3 可以看出，如果文本和图片按相同顺序编码，那么得到的嵌入空间除了对角线的单元象征图文关联之外，其他均表示图文不关联，因此可以使用对比学习的方法进行训练。最终的训练目标是：希望对角线元素尽可能大，非对角线的元素尽可能小。在经过海量数据训练之后，两个编码器基本能够做到语义相关。此时可以使用 CLIP 开展诸多下游任务。图 3 右以图像分类为例，将不同类别扩写成文本，然后输入到 CLIP 与图像的嵌入向量计算内积，取内积最大者为模型判断的类别，因为模型认为二者相关性最大。这样就实现了模型功能的迁移。

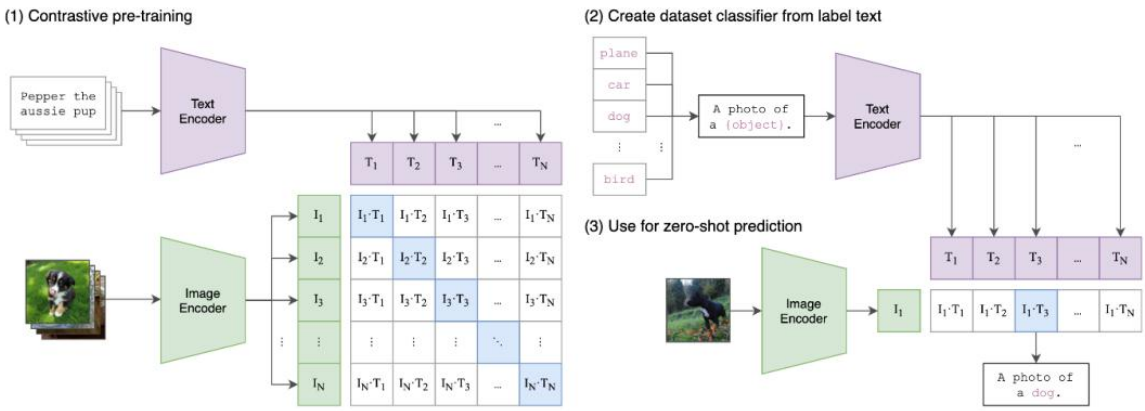



图 3 CLIP 基本训练和测试原理

### 2.2.2 MGCA

在 CLIP 中，模型训练考虑的是文本和图片的配对，然而在医学领域，尤其是放射图像的识别领域，用户更加关注图片在局部的表征，以及图片域内和文本域内的样本之间的区分等更加复杂且细粒度的信息。因此，MGCA 在经典多模态模型训练的基础之上添加了不同粒度的模块，从而丰富了模型的表示信息。

具体来说，MGCA 除了实例级别的对齐（与 CLIP 一致）之外，还添加了词级别的对齐，以及图片域文本域内部聚类两个粒度的训练模块。如图 4 所示。

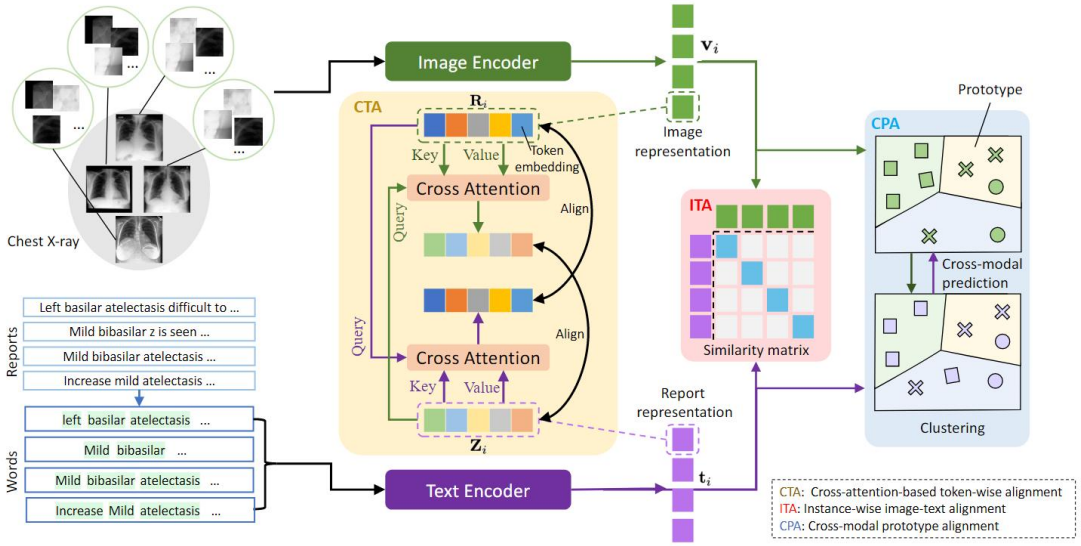



图 4 MGCA 基本原理

具体来说，MGCA 的训练目标除了 CLIP 中尽可能区分配对图文和不配对图文之外，还引入训练目标：图片描述信息中某个词(token)和图像对应的区域(image token)要尽可能对齐，与不对应的区域要尽可能分散；同类图像和相应文

本的嵌入向量之间应该尽可能重合而不同类之间应该尽可能分散。前者的具体表现为：使用交叉注意力机制，分别把 image token 作为文本嵌入向量的 Query，分别把 token 作为图片嵌入向量的 Query。这样做模型确实能够有效地学习到图片局部的信息，图 5 展示了关键词提示下模型赋予图片不同区域的权重。

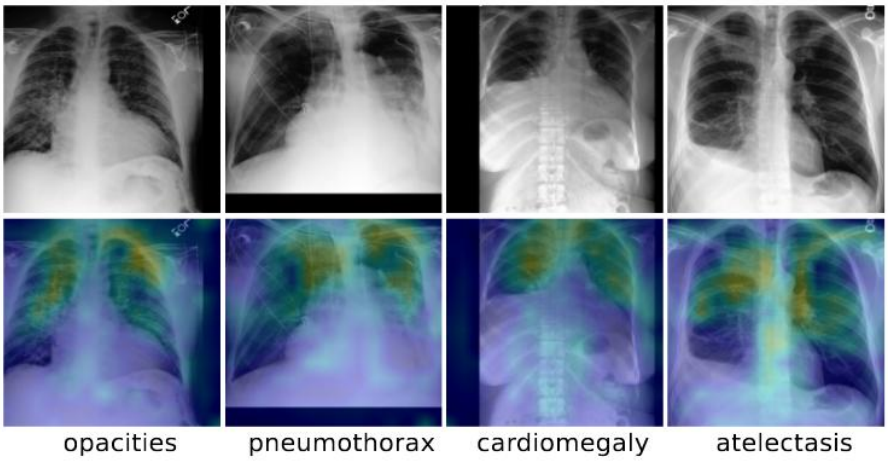



图 5 模型赋予图片不同区域的权重

### 2.2.3 BiomedGPT

BiomedGPT 希望能够有机整合用户输入模型的提示、文本和图像信息，因此沿用 OFA[11]模型结构，其使用 BART[9]作为基座模型，使用 BERT 风格编码器和 GPT 风格的解码器。该架构比较具有特色的一点在于合并了三类输入特征向量。不过其对文本和图像分别进行位置编码以解耦。除此之外，分别设计了文本部分的位置偏差和图像部分的位置偏差。如图 6 所示。

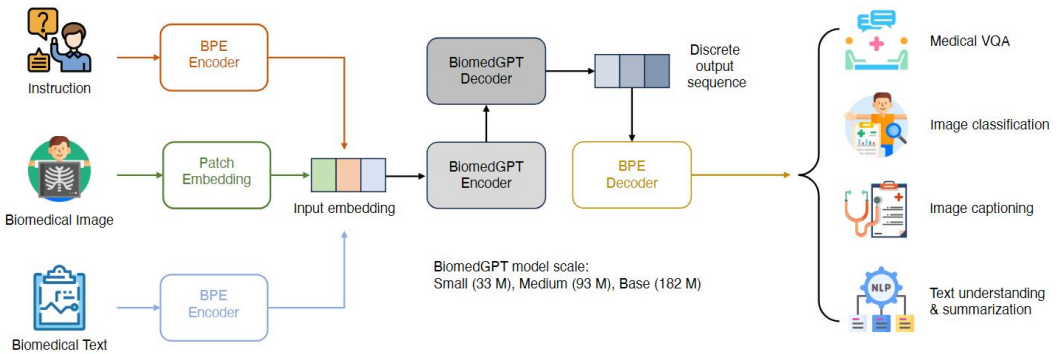



图 6 BiomedGPT 结构

### 2.2.4 LLaVA-Med

LLaVA-Med 以 LLaVA 为基模型，先后进行了通用基模型在医药学领域的对

齐训练和医药学领域的指令微调对齐两个阶段，进一步强化了通用多模态大模型所拥有的医药学知识，并且强调了模型的输出范式能够更加符合医药工作者的要求。整个训练过程仅仅持续 15 个小时，所得到的模型 LLaVA-Med 能够胜任医药专业领域的多种下游任务。如图 7 所示。

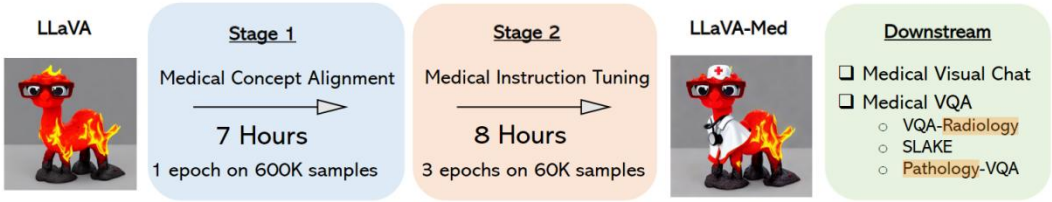



图 7 LLaVA 训练范式

### 2.2.5 CONCH

CONCH 的模型结构类似于 MGCA，其为了捕捉更加细粒度的图文配对信息，并且自回归地生成回复，采用了如图 8b 所示的编码器-解码器架构。其中，解码器接受来自两个编码器的内容，并且自回归地生成答案。因此，在训练期间，除了图文配对的目标之外，解码器还具备一个自回归的训练目标，希望其输出能够尽可能接近图文相关的内容。经过训练，模型获得了相当不错的性能，如图 8c 所示。

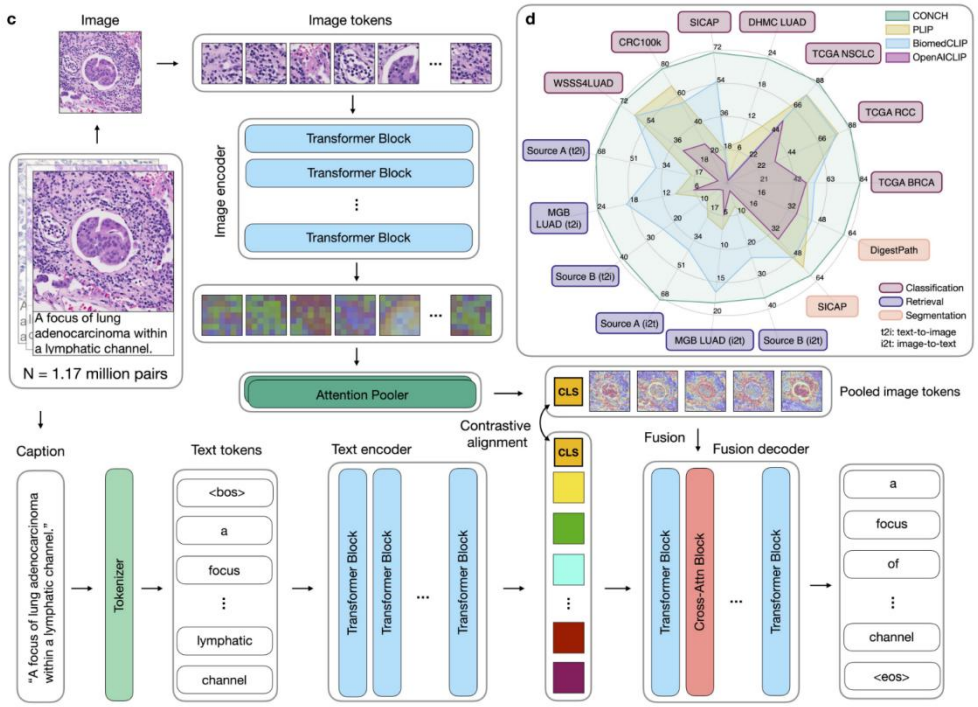


图 8 CONCH 训练框架示意以及 CONCH 模型表现

### 2.2.6 其他模型

PLIP 的主要创新点集中在数据集的构建上，模型训练几乎完全沿用 CLIP；Med-Flamingo 与 BiomedGPT 模型架构十分相似，其创新点主要集中在使用了需要制定特殊处理规则的数据上；PeFoMed 的输入特征构建同样沿用 BiomedGPT 的连接逻辑，其主要特点在后期使用了高效参数微调 LoRA[5]，此处略去。PathChat 沿用 CONCH 预训练框架，但使用私有数据进行了编码器的训练，模型结构上没有变化，如图 17。

## 2. 3 训练数据集

### 2.3.1 PLIP

PLIP 最大的创新点在于充分利用了病理学交流社区的内容。PLIP 从 twitter 中收集 2006 年 3 月 21 日至 2022 年 11 月 15 日的病理学数据作为模型的训练数据，用 2022 年 11 月 15 日至 2023 年 1 月 15 日的数据作为验证集。主要检索方式是使用 twitter 标签检索。针对检索结果进行仔细的清洗之后得到数据集 OpenPath。具体来说，数据的清洗包括删除二义性的样本（帖子包含不明确观点的或者包含矛盾观点的，如问句），删除高相似度的样本（twitter 的数据很可能存在转发和复制，这容易导致训练数据存在重复，或者验证集和训练集存在重复，这会影响模型的性能判定。基于此，团队设计了深度学习二分类模型来判断验证集中的某张图片是否出现在训练集里）。图 9 展示了数据集的提取流程，图 10 展示了数据集的构成。

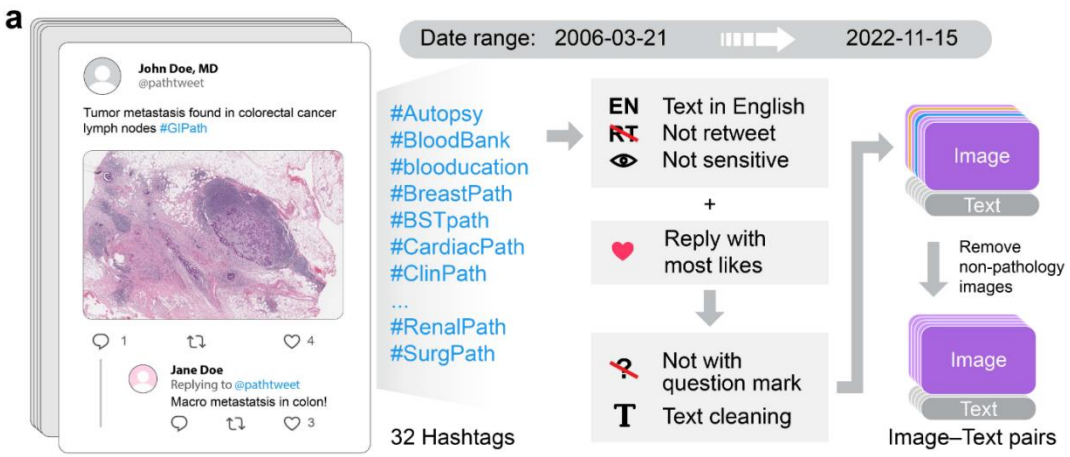


图 9 数据集抽取流程：删除转发、敏感的消息，将原帖和高赞评论进行组合后进行数据清洗，之后清洗掉不相关图像，得到最终的 twitter 数据集。


图 10 数据集构成：根据时间划分训练集和测试集，为了保持稳定性还引入了几个病理学常用的图片-图片附注数据集。

### 2.3.2 BiomedGPT

BiomedGPT 使用了横跨医疗多个领域的数据进行了广泛的预训练，因此形成了通用的能力。其数据集的描述如图 11 所示。除了常见的图文对数据集之外，还采用了目标检测数据集以及图像建模数据集，这两个数据集能够进一步提升模型对生物医学图像局部信息的理解程度，从而进一步提升其表征能力。

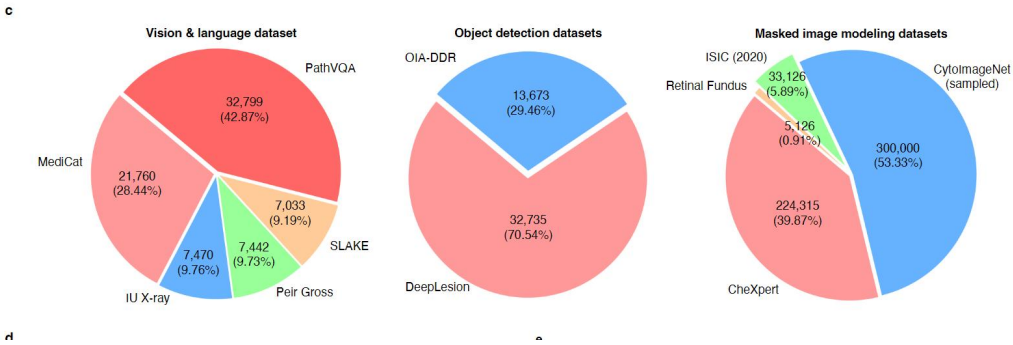



图 11 数据集构成，加粗字体为数据集类型，非加粗字体为数据集名称、样本数和比例。

### 2.3.3 LLaVA-Med

该研究最大的创新点在于使用 self-instruct 方式生成数据。首先我们收集了大量的带有附注的图片，附注基本上可以认为是在描述图片信息。接下来利用 GPT-4 辅助生成数据：我们让 GPT-4 根据图片的附注生成一段问答内容，此时，GPT-4 并不能真正看到图片，但是其根据附注生成的问答内容与图片紧密相关，看起来如同看到了图片。那么在实际训练时用生成的对话结合实际图片送入模型微调，同样能够获得接近 GPT-4 的优秀性能。下图 12 展示了这种 self-instruct 生成数据的经典示例，下图 13 展示了研究人员要求 GPT-4 生成问答内容中问题

和回复的主要类型。

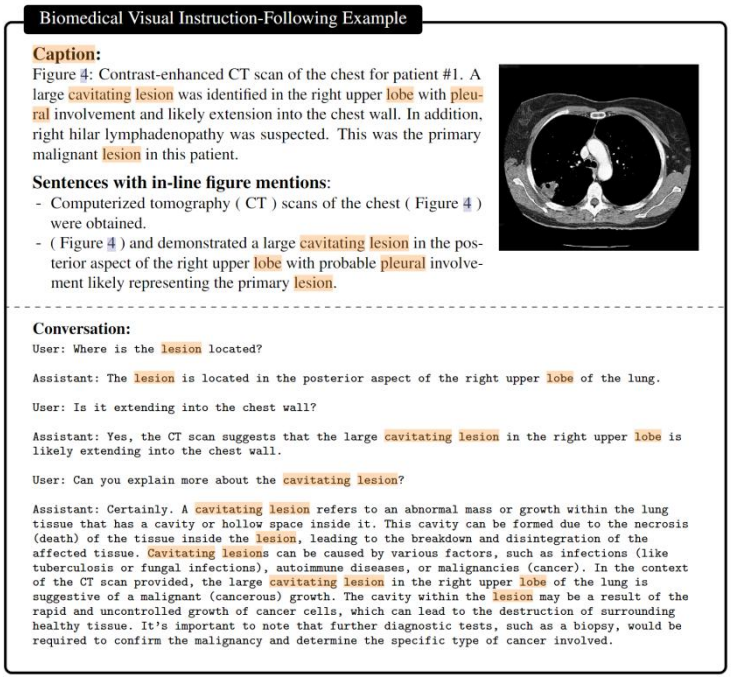
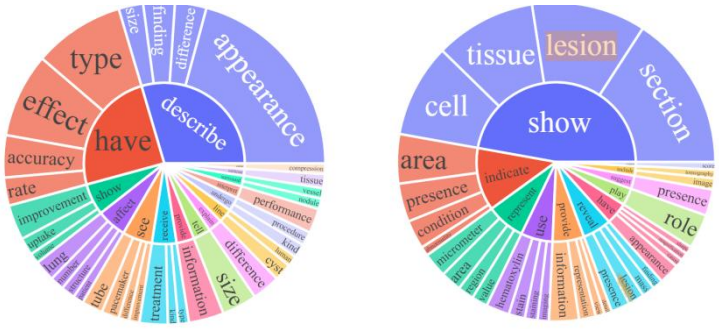



图 12 self-instruct 生成数据




(a) Instruction (b) Responses

图 13 问题和回复的主要类型，同一颜色的内圈词语加外圈词语构成一种问题或一种回答。

### 2.3.4 Med-Flamingo

该研究除了与其他研究一样使用图像-图像附注作为图像文本对之外，还考虑到使用图文交叉的内容，因为这类图片即便没有附注，文本也充分体现了图片的信息；同时，图文交叉的形式与实际使用时用户输入的内容格式十分相似，因此考虑到图文交叉数据是明智的。具体来说，对于图像-图像附注，我们会将图像和附注按一致的顺序分别编码；而对于图文交叉数据，无法准确分割图片和其对应描述，研究者选择将大量的图文交叉进行分块(chunk)，每一个块中包含比较少的图文交叉信息，然后对这些块使用文本编码器和图像编码器进行混合编码，构成混合嵌入向量的输入。如图 14 所示。

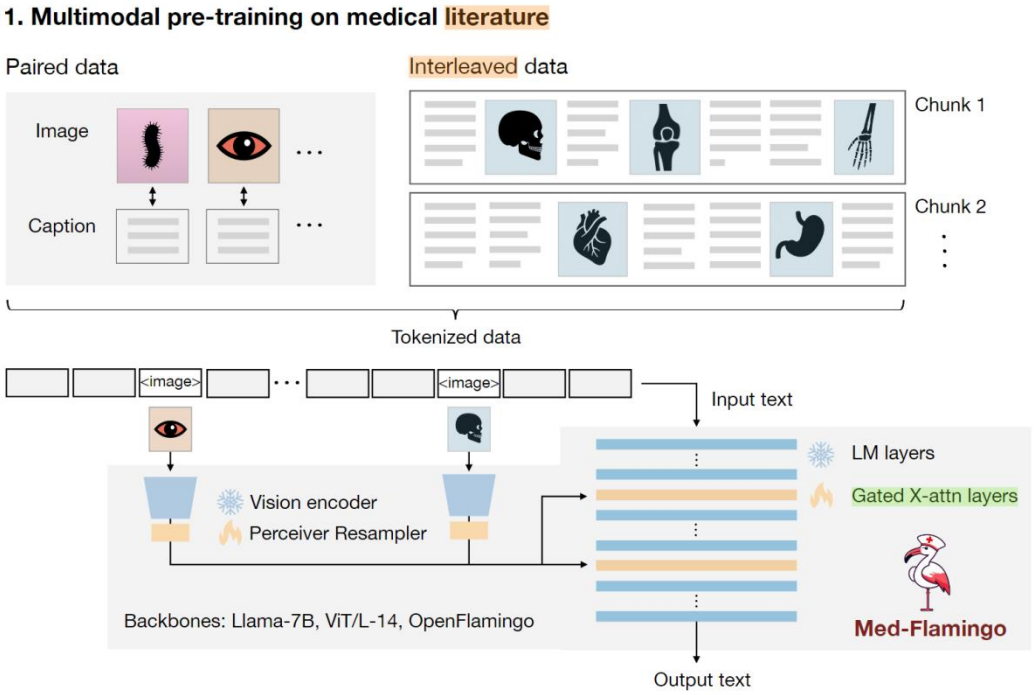



图 14 Med-Flamingo 数据处理示意

### 2.3.5 PathChat

PathChat 从不同的源收集并且制作了目前最大的多模态模型微调数据集，涵盖了多种 VQA 任务，包含对话、多选、描述等等。值得注意的是，研究人员发现单纯使用病理学数据集训练之后，如果给模型输入与病理学毫不相关的信息，模型的回答会出现混乱，这容易导致安全隐患。因此，研究人员同时设计了一个无关数据集 Guardrails,以便模型在判定该内容与病理学无关的时候告知用户自己无法作答。数据集的具体形式如图 15 所示。

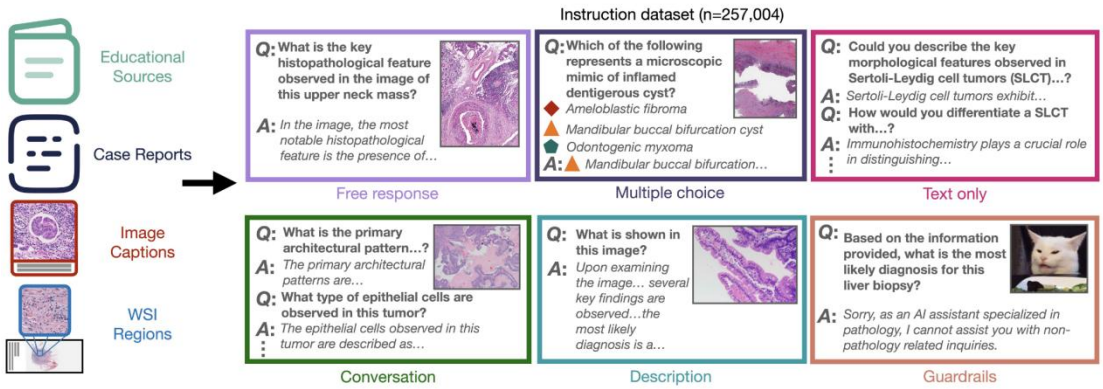



图 15　PathChat 数据集构建，每一个框代表一类任务，此处 Guardrails 中的图像为无关图像，除此之外，Guardrails 中也包含图片相关而文本不相关的内容，以及二者均不相关的内容。

### 2.3.6 CONCH

CONCH 作为病理学领域的基座模型，需要海量的数据进行预训练，为了解决数据噪声等问题，CONCH 设计了一套同一的自动清洗流程。具体来说，CONCH 从大规模医学数据集中手动构建了一个子集，包含若干图片-文本对。之后训练三个模型：图像目标检测模型：用来把多个子图结合为一张的图像进行分割；文本拆分模型：对于多子图的图像和附注，将附注按照语义分割为不同的子句；配对模型：将一张图和适合于它的子句进行拼接，这样就形成了大量的图像-文本对。之后，手动筛选掉不是描述人类病理的图像-文本对。下图 16 展示了 CONCH 流水线的基本流程，以及构建好的数据集的类别分布。

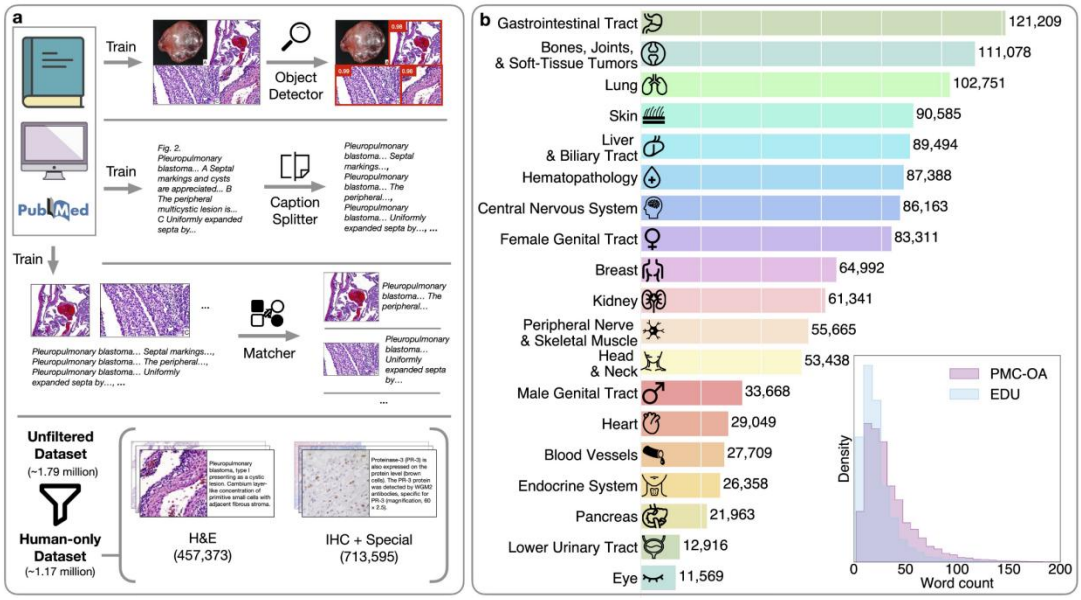



图 16 CONCH 数据集基本流程以及数据集类别分布图示。

### 2.3.6 其他数据

MGCA 采用放射图像识别的数据集 MIMIC-CXR 2.0.0，仅进行了基本的清洗和预处理；PeFoMed 使用常用的图像-图像附注数据集 ROCO，进行了简单的高效参数微调。此处均略过。

## 三、优势与不足


图 17 PathChat 的训练方式

模型结构和训练语料集中了垂直领域多模态大模型的主要创新点。在本综述所涉及的 8 项研究中，有部分存在前驱和后继的关系。例如 PathChat 与 LLaVA-Med（具体参考图 2）。图 2 鲜明地展现了 PathChat 在各个方面均强于 LLaVA-Med，这可能是由以下几部分原因导致的：

**预训练：**训练方式上，LLaVA-Med 从基座模型开始，进行了继续预训练和微调；而 PathChat 直接采用了强大的 CONCH 模型作为其文本编码器和最优的视觉编码器 UNI 作为自己的图像编码器，如图 17 所示。CONCH 模型已经被证明在多项任务中表现出强大的性能，而 LLaVA-Med 的基座模型 LLaVA 是一个通用多模态模型，本身没有经过医学领域知识的特殊训练。因此，虽然 LLaVA-Med 同样进行了大量数据的训练，结果依旧无法企及 PathChat。这说明预训练的语料对模型的整体性能有决定性影响。如果预训练语料具有明确的任务导向，模型在该任务的表现将会十分出色。

**任务设定：**LLaVA-Med 的继续预训练数据集主要包含通用开源图文数据集以及 GPT-4 生成的图片问答数据集。这些数据相比通用数据集有很大的局限性：LLaVA-Med 数据集的主要构成是简单图文对和图文问答，这决定了 LLaVA-Med 不可能适用于多种多样的医药学任务。而 PathChat 包含六种人为设计的不同任务的数据集 ，任务方向明确，同时引入了 Guardrails，保证了其安全性，因此在更多的任务中都能够表现出优秀的性能。这说明了多模态模型训练数据的构建上需要明确自身需求，从而更加有针对性地构造数据。

该结论同样能够从 PLIP 性能比较中体现，如图 18.

| Dataset | Number of candidates | Metric | PLIP | CLIP | Random | Fold change (PLIP vs. CLIP) |
|---|---|---|---|---|---|---|
| Twitter | 2,023 | Precision@10 | 0.271 | 0.061 | 0.005 | 4.5 |
| | | Precision@50 | 0.527 | 0.128 | 0.025 | 4.1 |
| PathPedia | 210 | Precision@10 | 0.409 | 0.167 | 0.048 | 2.5 |
| | | Precision@50 | 0.752 | 0.476 | 0.238 | 1.6 |
| PubMed | 1,419 | Precision@10 | 0.069 | 0.015 | 0.007 | 4.7 |
| | | Precision@50 | 0.206 | 0.082 | 0.035 | 2.5 |
| Books | 558 | Precision@10 | 0.265 | 0.045 | 0.018 | 5.9 |
| | | Precision@50 | 0.659 | 0.165 | 0.090 | 4.0 |


图 18 PLIP 性能比较：以图像检索为例

可以看出，与没有经过专门的病理学数据训练的 CLIP 相比，PLIP 检索性能不论在何类别均能达到更优的水平，这说明数据所对应的任务指向明确时，模型可以发挥出更强大的能力。

另外，PeFoMed 相比 LLaVa-Med 也表现出了优越的性能，如图 19 所示。

| Methods | Decoder Type | Accuracy Measurement | Trainable Parameters | Open | Closed | Overall |
|---|---|---|---|---|---|---|
| MMQ | Classifier | Exact Match | 28.3M | 52.0% | 72.4% | 64.3% |
| MTL | | | - | 69.8% | 79.8% | 75.8% |
| VQA-Adapter | | | 2.1M | 66.1% | 82.3% | 75.8% |
| M3AE | BERT | Exact Match | - | 67.2% | 83.5% | 77.0% |
| M2I2 | | | 262.2M | 66.5% | 83.5% | 76.8% |
| MUMC | | | 211.1M | 71.5% | 84.2% | 79.2% |
| ARL | | | 362.0M | 67.6% | 86.8% | 79.2% |
| LLaVA-Med(LLaVA) | Generative LLM | Open: Token Recall Closed: Exact Match | 7B | 61.5% | 84.2% | 75.2% |
| LLaVA-Med(Vicuna) | | | | 64.4% | 82% | 75% |
| LLaVA-Med(BioMed CLIP) | | | | 64.8% | 83.1% | 75.8% |
| GPT-4v | | Exact Match | - | - | 61.4% | - |
| **PeFoM-Med (ours)** | | Human Evaluation | 33.6M | **73.7%** | **87.4%** | **81.9%** |

图 19 PeFoMed 性能对比

虽然 LLaVA-Med 主要的优势在于整个预训练+微调流程仅持续不到 1 天，但是 PeFoMed 引入了高效参数微调 LoRA，用时更短，且使得小参数量也能够比 LLaVa-Med 收获更强的性能。这是否说明高效参数微调在某些情况下优于全参数微调和预训练尚有待研究，但可以肯定的是，不论是预训练还是微调，如果希望增强模型的整体性能，必须保证数据的全面性。

## 四、机遇和挑战

医疗领域所具有的数据类型复杂、分析精度高、多学科交融等特征，均揭示了其时代呼声在多模态大模型，而多模态大模型应用于医疗的主要瓶颈在于数据的构建，以及更加优秀的模型性能。现有的数据虽然规模足以训练新的大模型，但是因为其体量庞大，目前尚无一种有效的手段识别并且清理数据的重复和领域的相关性（目前几乎都需要一定人力）；此外，随着大模型的落地和逐渐普及，医疗领域的文本数据中自动生成的内容势必会越来越多。这些数据是否能够被人类信服，或者这些数据是否具有与人相同的效力，仍旧是没有被解决的问题。此外，医疗领域的特性要求我们对大模型生成的回复保持严格的判断和批判的态度，因此，医疗领域大模型的可解释性研究将会很大程度上影响人类能否信任模型生成数据，并且作为语料训练更加先进的模型。模型层面，所谓通用基座模型并不是真正意义上的完全通用，其可使用的范围不够精细，因此只能减少一些低成本的劳动（比如医学知识科普，病患自诊等等），无法设计更加精细的领域。如果多模态大模型要真正在医疗领域发挥推动性作用，就不能一味采用 CLIP 的范式，而是应该根据模型的使用场景，对模型的结构做细微的调整，即，基于特征（feature-based）和基于微调（finetuning-based）两种模式结合训练模型（如 MGCA）。此外，有关多模态大模型在医疗领域的研究目前仍局限于模型本身，当前大语言模型领域的诸如检索增强生成 RAG 和混合专家系统 MoE 并没有移植的迹象，这两种技术都旨在让模型在特定领域给出高水平回答，因此这是一个比较有前景的研究方向。

# 参考文献

[1] MOOR M, HUANG Q, WU S, et al. Med-Flamingo: a Multimodal Medical Few-shot Learner[J]. 2023.

[2] HE J, LI P, LIU G, et al. PEFOMED: PARAMETER EFFICIENT FINE-TUNING ON MULTIMODAL LARGE LANGUAGE MODELS FOR MEDICAL VISUAL QUESTION ANSWERING[J]. 2024.

[3] HUANG Z, BIANCHI F, YUKSEKGONUL M, et al. Leveraging medical Twitter to build a visual-language foundation model for pathology AI[Z/OL]. (2023-04). http://dx.doi.org/10.1101/2023.03.29.534834. DOI:10.1101/2023.03.29.534834.

[4] LU MingY, CHEN B, WILLIAMSON DrewF K, et al. Towards a Visual-Language Foundation Model for Computational Pathology[J]. 2023.

[5] LU MingY, CHEN B, WILLIAMSON DrewF K, et al. A Foundational Multimodal Vision Language AI Assistant for Human Pathology[J]. 2023.

[6] WANG F, ZHOU Y, WANG S, et al. Multi-Granularity Cross-modal Alignment for Generalized Medical Visual Representation Learning[J]. 2022.

[7] ZHANG K, YU J, YAN Z, et al. BiomedGPT: A Unified and Generalist Biomedical Generative Pre-trained Transformer for Vision, Language, and Multimodal Tasks[J]. 2023.

[8] LI C, WONG C, ZHANG S, et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day[J]. 2023.

[9] LEWIS M, LIU Y, GOYAL N, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.[C/OL]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online. 2020. http://dx.doi.org/10.18653/v1/2020.acl-main.703. DOI:10.18653/v1/2020.acl-main.703.

[10] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[11] WANG P, YANG A, MEN R, et al. UNIFYING ARCHITECTURES, TASKS, AND MODALITIES THROUGH A SIMPLE SEQUENCE-TO-SEQUENCE LEARNING FRAMEWORK[J].

[12] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[13] RADFORD A, KIM J, HALLACY C, et al. Learning Transferable Visual Models From Natural Language Supervision[J]. Cornell University - arXiv,Cornell University - arXiv, 2021.

# 附录

[1] zero-shot:zero-shot 是机器学习领域的一种先进方法，它旨在使模型不经过训练就能够识别、分类或理解在训练过程中未见过的类别或概念。这种学习方法对于解决现实世界中常见的长尾分布问题至关重要，即对于一些罕见或未知类别的样本，传统的监督学习方法可能无法很好地处理。

[2] VQA:即 Visual Question Answering，视觉问答。这是一个涉及计算机视觉和自然语言处理的复杂学习任务。具体来说，VQA 系统接收一张图像和关于该图像的自由形式的开放式自然语言问题作为输入，并产生一个自然语言答案作为输出。此任务的目标是开发一种系统，能够精确地回答与输入图像相关的特定问题，答案可以采用多种形式，包括单词、短语、二元答案、多项选择答案或文本填空。

[3] few-shot:是机器学习领域的一种学习方法，旨在让模型在只有少量训练样本的情况下进行分类或预测。few-shot learning 中，模型通过学习少数几个样本的特征和标签之间的关系，从而能够对新的、未见过的数据进行分类或预测。这种学习方法通常使用元学习（meta-learning）技术来实现，其中模型会学习如何快速适应新任务，并从少量样本中提取有用的信息。

[4] rationale:在人工智能（AI）中，rationale 指的是一个决策或行动的理由或原因。它通常是指基于逻辑、经验、数据或其他相关信息而得出的结论或推论，用于支持某个决策或行动的合理性和正确性。rationale 可以有多种形式，此处指模型生成的选择某图像的原因。

[5] LoRA:全称 Low-Rank Adaptation of Large Language Models，直译为大语言模型的低阶适应，是微调的一种，其通过在原始模型的基础上添加少量参数，从而改变模型行为或者在某任务上的性能。