M3D：利用多模态大型语言模型推进 3D 医学图像分析

摘要

摘要..... 医学图像分析对临床诊断和治疗至关重要，而多模态大语言模型（MLLM）也越来越多地支持临床诊断和治疗。然而，以往的研究主要集中在二维医学图像上，尽管三维图像具有更丰富的空间信息，但仍未得到充分探索。本文旨在利用 MLLMs 推动三维医学图像分析。为此，我们提出了一个大型三维多模态医疗数据集 M3D-Data，其中包括 120K 个图像-文本对和 662K 个指令-响应对，专门用于各种三维医疗任务，如图像-文本检索、报告生成、视觉问题解答、定位和分割。此外，我们还介绍了一种新的三维多模态医疗基准–M3D-Bench，它有助于对八项任务进行自动评估。通过综合评估，我们的方法被证明是一种强大的三维医学图像分析模型，其性能优于现有的解决方案。并且在将其迁移到 3D 脑 CT 的视觉问答任务的时候，通过在构建的两组公开数据集上的实验结果表明，M3D 仍可以取得不错的效果。

关键词：多模态; 大型语言模型; 3D 医学图像分析

1 引言

医疗场景 [13] 包含大量多模态信息，包括患者信息、诊断报告和各种模态的医学影像。诊断报告与医学图像配对后，可提供准确详细的描述、发现和诊断，被视为高质量的注释。这些医学影像和文本与医生的诊断工作流程一起被大规模保存到数据库中，无需额外费用。如何充分利用这些图像和文本数据建立医学影像诊断模型是一个关键问题。

在最近的研究中 [1, 9, 11, 20]，多模态大型语言模型（MLLMs）在各种多模态任务中表现出色，有效地整合了图像和文本数据。通过将视觉模型 [14] 的感知能力与大型语言模型（LLMs）[3] 的生成能力相结合，MLLMs 引起了研究人员的极大关注，尤其是在医学图像分析领域。现有的医学 MLLM [8] 对医学图像和文本数据上公开可用的二维 MLLM 进行了微调，以完成图像-文本检索、报告生成和视觉问题解答等任务。这些模型被认为是理解和推理二维医学图像的强大工具。然而，当面对 CT 和 MRI 等包含丰富空间信息的广泛三维医学图像时，这些方法往往力不从心，要么需要昂贵的逐片分析，要么完全失败。

在这项工作中，我们将重点放在三维医学图像上，并将 MLLMs 的使用扩展到对这些图像的分析。为此，我们收集了一个大型三维多模态医疗数据集 M3D-Data，其中包括 120K 个图像-文本对和 662K 个指令-响应对，涵盖各种疾病和任务。

该数据集是迄今为止最大的公开三维多模态医疗数据集，可推动相关研究。此外，我们还提出了用于医学图像分析的多功能 3D MLLM–M3D-LaMed。它可以执行图像文本检索、报告

生成和视觉问题解答等任务，还首次包含了视觉语言定位和分割等任务。利用类似 CLIP [15] 策略的预训练三维视觉编码器和高效的三维空间池化感知器，它可以直接理解和推理三维图像。M3D-LaMed 首次与三维可提示分割模型相结合，实现了三维医学图像的指代表达分割。此外，为了评估该模型在三维医学分析中的能力，我们提出了一个多模态医学基准–M3D-Bench，其中包括 8 个任务，涵盖了三维医学图像分析的各个方面。这是首个全面的三维医学图像分析基准。除传统指标外，我们还引入了基于 LLM 的评估，使 M3D-Bench 能够自动、准确地评估模型的性能。我们将 M3D-LaMed 迁移至脑 CT 分析的任务中，并基于两组公共数据 CQ500 和 PhysioNet 构建了关于脑疾病诊断的 QA 对，M3D-LaMed 在处理这些问题的时候也取得非常好的性能。

总之，我们的贡献如下：

- 建立 M3D-Data，这是一个大型 3D 医学数据集，包含 120K 个图像-文本对和 662K 个指令-响应对。
- 提出 M3D-LaMed，一种用于三维医学图像分析的多功能 MLLM，可应用于各种三维多模态任务。
- 创建 M3D-Bench，这是一个针对 8 项任务的综合性 3D 多模态基准。
- 基于公共数据集 CQ500 和 PhysioNet 构建了 QA 对，并且 M3D-LaMed 展现了很好的性能。

2 相关工作

2.1 医学多模态数据

在医疗场景中，获取丰富的图像和文本数据面临隐私和限制因素的挑战，导致构建大规模多模态医疗数据集变得困难。PMC-OA 通过网络爬虫从医学论文中获取了 160 万个图像-文本对 [10]，而 MedMD 则致力于整合公共数据集并抓取三维图像和文本数据 [17]。RP3D 数据集包含 51K 个三维图像-文本对和 142K 个由 LLM 生成的 VQA 数据 [17]。我们的工作主要集中在通过抓取医疗专业网站构建大规模三维医疗数据集，M3DData 包括 120K 个三维图像-文本对和 662K 个指令-响应对，且所有数据均通过低成本的自动化管道生成。此外，M3D-Seg 组件从 25 个公共医疗分割数据集中收集了近 6K 张三维图像，以支持视觉语言定位和分割任务。

2.2 医学 MMLMs

医学 MLLM [18] 通常是利用医学多模态数据集，从强大的 2D 开源 MLLM 中微调出来的。例如，LLaVA-Med [8]、Med-PaLM M [16] 和 MedFlamingo [12] 分别基于 LLaVA [11]、PaLM-E [4] 和 Flamingo [2] 等模型。PMCVQA [19] 等大规模数据集的可用性使得医学 MLLM 可以从头开始训练，尽管最初仅限于二维图像。虽然 RadFM [17] 支持二维和三维图像，但它主要用于文本生成任务，如 VQA，而且性能较差。在我们的工作中，M3D-LaMed 是用于三维医学图像分析的通用 MLLM。它不仅能处理报告生成和 VQA 等文本生成任务，还能处理

视觉任务，如三维医学图像中的视觉语言定位和分割，这对医学图像分析中的识别和定位至关重要。

3 方法

由于 3D 编码器不可靠，我们从头开始训练视觉编码器。如图1(a) 所示，我们在 M3D-Cap 上使用类似 CLIP 的策略预训练三维医疗视觉编码器。随后，我们引入端到端调整，利用指令数据将三维信息整合到 LLM 中，确保视觉和语言之间的无缝交互，如图1(b) 所示。

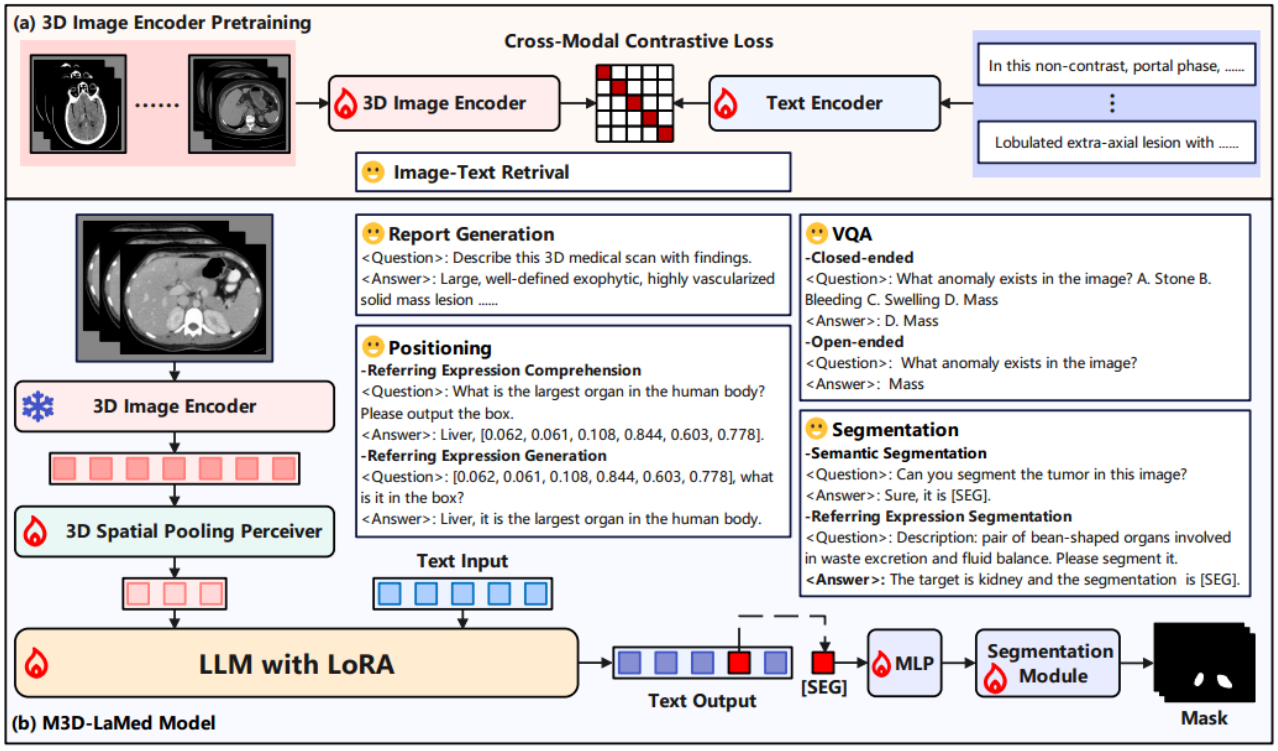

图 1. M3D-LaMed 框架结构

3.1 模型架构

3.1.1 3D 图像编码器

给定一个特定的三维图像 $I \in R^{C \times D \times H \times W}$ ，C、D、H、W 分别代表通道、深度、高度和宽度，我们得出图像嵌入 $v = E_{img}(I) \in R^{n \times d}$ 。这里，E_{img} 表示图像编码器，n 表示图像标记数，d 表示标记维数。为了提高通用性和通用性，我们使用 3D Vision Transformer（3D ViT）作为视觉编码器。3D ViT 包括一个具有注意力机制的 N 层变换器。每一层对从输入图像中提取的补丁进行操作，其中补丁大小为 $P_D * P_H * P_W$。我们可以直接从 MONAI 库中导入标准 3D ViT。

3.1.2 3D 感知器

由于三维图像本身具有高维度和大量标记，直接输入 LLM 会产生大量计算成本。为了缓解这一难题，我们提出了一种简单高效的三维空间池化感知器，旨在减少嵌入的数量和维度，

如图2所示。首先，将视觉编码器输出的标记重构到三维空间进行池化。这一步骤在保留原始空间信息的同时，有效减少了标记数量。其次，我们采用一系列多层感知器（MLP）来调整嵌入维度，使其与 LLM 所需的维度保持一致。通过实施这些步骤，三维感知器不仅降低了计算成本，还确保了空间信息的保留。

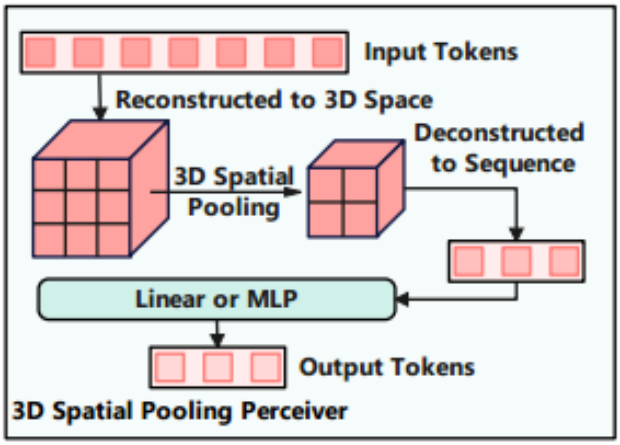

图 2. 三维空间汇集感知器的结构

3.1.3 LLM

在广泛的自然语言语料库中训练出来的大型语言模型提供了通用的嵌入表征和强大的生成能力。在我们的研究中，我们直接使用 LLaMA-2-7B 模型作为我们的基础 LLM，因为该模型在捕捉语言模式和生成跨领域的连贯文本方面的有效性已得到证实。

3.1.4 可提示的细分模块

受 LISA [7] 的启发，我们利用 MLLM 的功能，使用可提示的分割模块来实现指代表达分割。具体来说，如果输出标记中有 [SEG] 标记，我们就提取 [SEG] 标记的最后一层嵌入作为特征。随后，我们将这一特征映射为一个提示，通过 MLP 驱动分割模块，最终产生分割掩码。由于 SegVol [5] 性能强大，而且与我们的框架兼容，因此我们选择 SegVol 作为可提示的分割模块。

3.2 预训练的视觉编码器

由于缺乏稳健的 3D 医学图像编码器，我们采用了 CLIP 的架构和训练方法对 M3D-Cap 进行预训练。如图 1 (a) 所示，我们在预训练中使用了跨模态对比学习损失。视觉编码器从头开始预训练，而文本编码器则使用预训练的 BERT [6] 作为初始化。

3.3 MLLM 训练

在获得预先训练好的三维医学视觉编码器后，我们使用三维感知器将其集成到 LLM 中，进行端到端的训练。我们的训练过程包括两个主要步骤。首先，我们冻结视觉编码器和 LLM，仅使用图像文本对三维感知器进行微调。随后，我们使用指令数据对视觉编码器、三维感知

器、LLM 和分割模块进行微调。如果 [SEG] 标记出现在输出标记中，则使用与 SegVol 类似的 Dice 和 BCE loss 进行分割训练。

4 复现实验

4.1 实验配置

我们采用最小-最大归一化方法对作为输入的三维 CT 图像进行预处理。此外，我们还将三维图像调整大小并裁剪为 32 × 256 × 256 的标准尺寸。我们的 3D 视觉编码器采用 3D ViT，使用 12 层变换器，贴片大小为 4 × 16 × 16。输出嵌入值为 2048 × 768，代表 2048 个标记和 768 个特征维度。经过我们的三维空间池化感知器后，最终输入 LLM 的视觉标记为 256 × 768。我们使用 LLaMA2-7B 作为 LLM 基础，并加载预训练参数。

在视觉编码器的预训练中，我们采用了带有 12 层变换器的 BERT 作为文本编码器，最大文本长度为 128。视觉编码器和文本编码器的 [CLS] 标记作为全局特征表示，线性层用于跨模态对比训练投影。此外，我们还采用了 6 × 8 的批量大小，在 8 个 GPU 上进行并行训练，学习率为 10-4，热身和余弦衰减时间表为 10-4。MLLM 训练分为两个阶段。起初，我们冻结视觉编码器和 LLM，仅使用图像-文本对微调 3D 感知器，采用的批量大小为 12 × 8，学习率为 10-4，并应用热身和余弦衰减。具体来说，我们探讨了两种情况：感知器中的 1 层线性和 2 层 MLP。随后，我们利用指令数据对视觉编码器、三维感知器、LLM 和分割模块进行微调，使用的批量大小为 12 × 8，学习率为 2 × 10-5，并采用热身和余弦衰减。在复现的过程中，我们采用参数高效的 LoRA 方法对 LLM 进行微调，LoRA 参数设置为 r = 16、 = 32 和 0.1 的丢弃率。最大上下文长度定义为 512，并且只考虑有关脑 CT 的问答数据集。

4.2 数据构建

我们基于两组脑 CT 任务的公开数据集 CQ500 和 PhysioNet 去构建有关脑诊断任务的问答 (QA) 对。**CQ500**：这个数据集包含了多个指标：intracerebral hemorrhage、intraparenchymal hemorrhage、intraventricular hemorrhage、subdural hemorrhage、epidural hemorrhage、subarachnoid hemorrhage、Bleed Location-Left、Bleed Location-Right、chronic bleed?、fracture、calvarial fracture、mass effect、midline shift。可以划分为出血类型、出血位置以及其他相关指标。以此为基础构建了 1 个开放式问题和 5 个封闭式问题，具体的开放式问题为" 患者是否出现脑出血？如果出现，具体是哪种类型（如脑内出血、脑室出血）以及发生在哪个位置（左侧、右侧或双侧等）？"，封闭式问题为：1. " 患者是否存在慢性出血？"，2. " 患者是否患有骨折？"，" 患者是否有颅骨骨折？"，" 患者是否存在肿块效应？" 和" 患者是否存在中线偏移？"。**PhysioNet**：数据标签文件（Patient demographics.csv）中标注了 Intraventricular Hemorrhage、intraparenchymal hemorrhage、subarachnoid hemorrhage、epidural hemorrhage、subdural hemorrhage 和 fracture。可以划分为出血类型和骨折的问题，为此设计了一个开放式问题和封闭式问题，开放式问题为" 患者是否存在脑出血？如果存在，具体是哪种类型的脑出血？"，封闭式问题为" 患者是否患有骨折？"。根据上面的构建规则，借助 GPT-4 生成 QA 对，用于视觉问答任务。

4.3 实验结果

模型在 PyTorch 中实现，并通过 8 个配备 24 GB 内存的英伟达 RTX 4090 GPU 使用 AdamW 优化器进行并行训练 3 个 epoch。实验结果如下表1。图3和图4分别展示了模型在 CQ500 和 PhysioNet 构建的 QA 数据集上随机回答示例。

表 1. lora 训练后的 M3D-LaMed 在 CQ500 和 PhysioNet 的 QA 数据集上的测试结果

Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rougle	Meteor
CQ500	0.540541	0.33333	0.257143	0.205882	0.297521	0.187829
PhysioNet	0.395833	0.106383	0.043478	0.022222	0.257426	0.214133

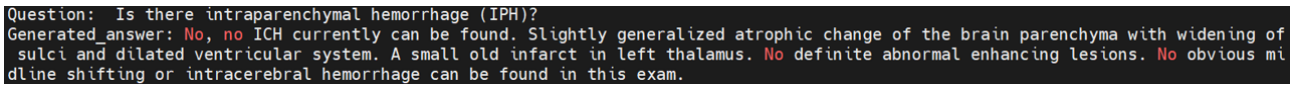

图 3. 在 CQ500 上的回答示例

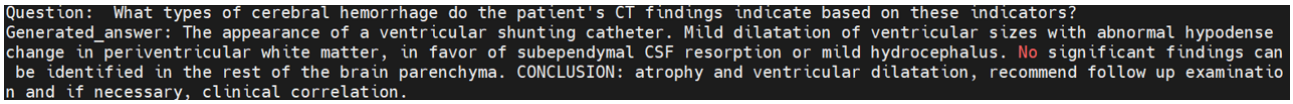

图 4. 在 PhysioNet 上的回答示例

4.4 实验结果分析

从表1来看，在由 CQ500 和 PhysioNet 数据集上构建的 QA 数据集上进行 lora 微调训练后的 M3D-LaMed 模型，在对应的 QA 测试集上可以达到比较不错的性能。

5 总结与展望

总之，我们的研究利用 MLLM 推进了三维医学图像分析。具体来说，初始的研究中构建了一个大型三维多模态医疗数据集 M3D-Data，其中包括 120K 个三维图像文本对和 662K 个为三维医疗任务定制的指令-响应对。此外，还提出了 M3D-LaMed，这是一个处理图像文本检索、报告生成、视觉问题解答、定位和分割的通用模型。此外，还引入了一个综合基准 M3D-Bench，该基准针对八项任务进行了精心设计。设计的方法为 MLLMs 理解 3D 医疗场景的视觉和语言奠定了坚实的基础。在涉及脑 CT 的 QA 任务中进行迁移研究后，发现仍可以取得比较不错的效果。

参考文献

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[3] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[4] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[5] Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao. Segvol: Universal and interactive volumetric medical image segmentation. *arXiv preprint arXiv:2311.13385*, 2023.

[6] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

[7] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024.

[8] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

[9] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[10] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023.

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[12] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal

medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.

[13] Xiangdong Pei, Ke Zuo, Yuan Li, and Zhengbin Pang. A review of the application of multi-modal deep learning in medicine: bibliometrics and future directions. *International Journal of Computational Intelligence Systems*, 16(1):44, 2023.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[16] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024.

[17] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023.

[18] Kai Zhang, Jun Yu, Eashan Adhikarla, Rong Zhou, Zhiling Yan, Yixin Liu, Zhengliang Liu, Lifang He, Brian Davison, Xiang Li, et al. Biomedgpt: a unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv e-prints*, pages arXiv–2305, 2023.

[19] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

[20] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.