MEDFACT-R1：通过伪标签增强实现事实性医学推理

Gengliang LI[1,6†], Rongyu CHEN[2,5†], Bin LI[3], Linlin YANG[4*], Guodong DING[2]

[1]Baosight, [2]NUS, [3]SIAT, CAS, [4]CUC, [5]Microsoft, [6]ANU

ABSTRACT

确保事实一致性及可靠的推理仍然是医疗视觉语言模型面临的重大挑战。我们引入了 MedFact-R1，这是一种两阶段框架，结合外部知识关联与强化学习以提升事实性医学推理。第一阶段使用伪标签监督微调（SFT）来整合外部事实性专业知识；而第二阶段则应用带有四个定制化事实奖励信号的组相对策略优化（GRPO），鼓励自我一致性的推理。在三个公共医疗问答基准测试中，MedFact-R1 相比之前的最先进的方法，在事实准确性上最多提升了 **22.5**% 个百分点。消融研究强调了伪标签 SFT 冷启动的必要性，并验证了每个 GRPO 奖励的贡献，突出了知识关联与基于 RL 驱动推理之间的协同作用，以建立可信的医疗 AI。代码发布在 https://github.com/Garfieldgengliang/MEDFACT-R1。

Index Terms— 医学视觉语言模型，事实性医疗推理，伪标签，GRPO

1. 介绍

医学诊断代表了信号处理和机器学习中最重要的前沿领域之一，体现了技术服务于人类的愿景。这需要卓越的专业知识、严格的准确性以及对本质上复杂的数据进行推理的能力。然而，高质量诊断数据的匮乏阻碍了进展，这些限制受专业要求和隐私约束的影响。这些局限性使模型训练变得复杂，并且经常导致不可靠的行为——例如依赖虚假的相关性、误判和漏诊。在医学领域，决策直接影响人类生命，因此这种错误是不可接受的。克服这些挑战对于开发能够实现可靠性能的深度学习系统至关重要，在现实世界的临床实践中 [1]。

最近，大规模的视觉语言模型（VLMs）[2, 3, 4] 迅猛发展，正在改变各行各业。它们在医疗领域的扩展显示出了巨大的潜力，近期的努力 [5] 整理了医学数据集，并对 VLMs 进行了微调以适应专业应用。然而，事实可靠性仍然是一个主要障碍：现有模型常常会在高风险场景中生成幻觉和事实错误。为了解决这一问题，RULE [1] 引入了一种风险控制的检索增强生成（RAG）方法，该方法平衡了外部检索与内部知识，从而显著提高了事实准确性。

同时，许多后训练工作成功地挖掘了视觉语言模型的知识和潜力，其中先进的强化学习（RL）已成为一个显著的例子。与通过下一个标记预测的监督学习不同，强化学习使用奖励信号优化任务策略而不依赖于详细的标注。GRPO [6, 7] 是最先进的 RL 后训练方法之一，在推广方面超越了监督微调，因为它解锁了推理中的“顿悟”时刻 [4]，这使其与先前的方法如 PPO [8]、DPO [9] 和传统方法 [10, 11] 区分开来。尽管取得了令人印象深刻的结果，但发现足够的领域知识是防止产生不现实或冗长输出的关键，这通常依赖于昂贵的多样化标注医疗数据集的整理 [12, 13]。

为了解决强化学习中数据稀缺性和对外部知识的依赖问题，我们提出了 MedFact-R1 (figure 1)。这是一个两阶段框架，将基于事实伪诊断数据的 SFT 与激励推理相结合，以激活事实能力而无需外部参考。在首先阶段，我们应用具有事实风险控制的生成器来生成用于 SFT 的伪诊断数据，从而更有效地扩展医学知识暴露并强化模型的基础。在第二阶段，我们采用 GRPO 后训练帮助模型全面消化和反思有价值的诊断

† 这些作者对此工作有同等贡献。

* 通讯作者。

数据。为了更好地适应医疗诊断场景，我们精心设计了四个奖励组件，考虑答案正确性、正式呈现、上下文相关性和基于事实推理的自洽性。我们的 GRPO 鼓励以事实为基础的推理，并允许模型超越观察案例进行泛化并提高医学事实准确性。

实验结果揭示了在多个公共医学问答基准测试中存在显著且一致的改进。

MedFact-R1 增强了所有评估指标，生成了更加事实性和可靠的诊断输出，并验证了我们框架在推进医疗事实性推理领域的有效性。此外，我们的消融研究提供了对伪标签 SFT 开始的必要性的更深入见解，并阐明了每个事实奖励的单独贡献。

总结而言，我们的贡献如下：1) 我们率先通过结合 SFT 和 RL 后训练范式，采用两阶段训练流程将先进的强化学习 GRPO 整合到医疗问答中，以增强 VLM 的事实推理能力；2) 我们使用事实伪标签初始化训练软烤温度，有效吸收外部医学知识；3) 为解决稀疏奖励信号并进一步提高事实性，对 GRPO 后训练定制了事实奖励，减少了幻觉现象，并促进了自我论证；4) MedFact-R1 设立了一个新的最先进的标准，在各种医疗问答基准的所有评估指标上均取得了超过 95% 的分数，比之前的方法提高了高达 22.5%。

2. MEDFACT-R1

2.1. 任务表述

医学视觉问答（QA）是基于图像回答医学问题的任务 [5] 。形式上，给定一张图像 I 和一个问题 Q，模型生成自由文本形式的答案。对于 VLMs 而言，生成答案的第一个标记预期为“是”或“不”；该标记映射到一个二进制标签 $\hat{A} \in \{0, 1\}$，表示临床正确或错误的响应。

2.2. 带有伪标签的监督微调

我们首先使用最大似然估计下的下一个标记预测进行有监督的微调（SFT），以建立事实推理的基础。为此，我们通过校准检索生成伪标签以减轻事实性风险，并通过偏好对齐将外部先验与内部知识 [1] 相协调。该策略将外部医学知识提炼成紧凑的监督信号，抑制幻觉并强化事实一致性。

2.3. GRPO 训后培训

后 SFT，GRPO [6] 通过基于规则的奖励指导策略更新，在生成的输出上使用基于组的蒙特卡罗优势估计和策略梯度。我们为基于 GRPO 的强化学习设计了四种互补类型的奖励值，总奖励由从生成的答案中得到的这些组成部分的归一化和给出。

准确度奖励。我们通过将预测答案与伪标签进行比较来评估其正确性。对于二分类问题，完全匹配的奖励为 1.0，其他情况则为 0。

格式奖励。为了鼓励结构化的推理，我们要求输出包括用<思考>和示例 翻译为示例，输入 翻译为输入，输出 翻译为输出，总结 翻译为总结，备注 翻译为备注，不要询问也不要解释为什么这么翻译标签包围的思维过程以及用请提供待翻译的文字内容。和请提供待翻译的文字内容。标签包围的简洁最终答案。只有当这四个标签恰好各出现一次且这些区域外没有多余内容时，奖励为 1；否则，奖励为 0。

事实奖励。为了促进医学推理中的事实一致性，我们设计了一个基于模型输出中存在临床依据概念的奖励。对于每个训练问题，使用 GPT-4 [2] 提取一组特定领域的概念，作为事实锚点。例如，问题胸放射影像是否显示任何肺部感染或充血的迹象？产生如肺部感染，充血，胸部放射摄影等概念。每个在答案中正确反映的概念贡献 0.2 分，仅计一次。这引导模型表达与医学相关的知识，提供了一个针对事实依据的目标信号。

一致性奖励。为了在临床环境中强化事实和逻辑的一致性，我们引入了一种一致性奖励，该奖励评估推理与最终答案之间的对齐程度。我们使用 GPT-4 来评估上下文一致性，参考了经过筛选的医学上正确和不正确的输出示例。如果答案能够通过其推理得到逻辑上的支持，尤其是在临床解释方面，则给予 1 分奖励；否则将施加-0.5 的惩罚。这鼓励模型在从医疗证据中得出结论时保持内部一致性。

3. 实验

3.1. 数据集和实现

数据集。实验是在三个医疗基准上进行的：IU-**透视图** [17], **哈佛-公平视觉语言医学模型** [18] 和 MIMIC-

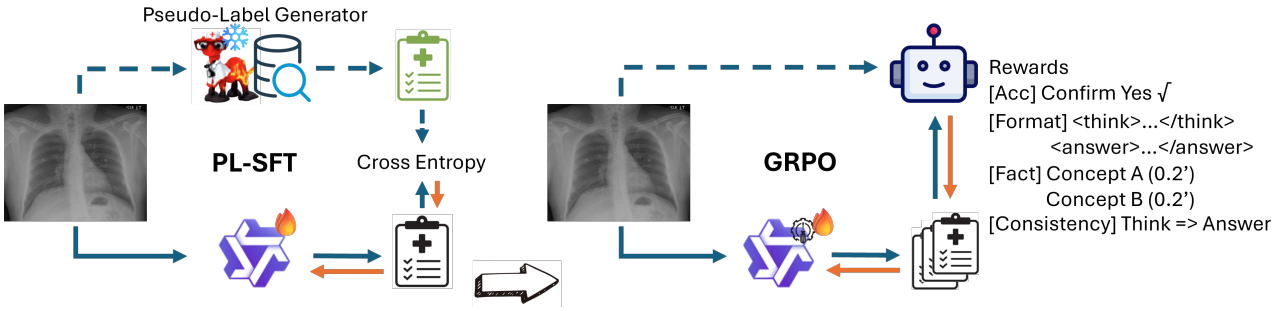

Fig. 1. **我们的方法** MedFact-R1 **概述，包括两个阶段：伪标签监督微调与基于** GRPO **的强化学习**。蓝色和橙色箭头分别表示前向和反向传播流程

Models	Venues	IU-Xray				Harvard-FairVLMed				MIMIC-CXR			
		Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1	Acc.	Pre.	Rec.	F1
LLaVA-Med v1.5 (7B) [5]	arXiv'24	75.47	53.17	80.49	64.04	63.03	92.13	61.46	74.11	75.79	81.01	79.38	80.49
+ Greedy	-	76.88	54.41	82.53	65.59	78.32	91.59	82.38	86.75	82.54	82.68	81.73	85.98
+ Beam Search	-	76.91	54.37	84.13	66.06	80.93	93.01	82.78	88.08	81.56	83.04	84.76	86.36
+ DoLa	-	78.00	55.96	82.69	66.75	76.87	92.69	79.40	85.53	81.35	80.94	81.07	85.73
+ OPERA	-	70.59	44.44	100.0	61.54	71.41	92.72	72.49	81.37	69.34	72.04	79.19	76.66
+ VCD	-	68.99	44.77	69.14	54.35	65.88	90.93	67.07	77.20	70.89	78.06	73.23	75.57
MedDr [14]	arXiv'24	83.33	-	-	67.80	70.17	-	-	80.72	55.16	-	-	56.18
RULE [1]	EMNLP'24	87.84	75.41	80.79	78.00	87.12	93.57	96.69	92.89	83.92	87.01	82.89	87.49
MMed-RAG [15]	ICLR'25	89.54	-	-	80.72	87.94	-	-	92.78	83.57	-	-	88.49
FactMM-RAG [16]	NAACL'25	84.51	-	-	68.51	83.67	-	-	87.21	77.58	-	-	81.86
Qwen2.5-VL-3B [3]	arXiv'25	61.21	37.32	66.91	47.91	42.83	85.83	37.32	52.00	53.57	80.45	30.81	44.53
Ours	-	97.63	95.62	95.48	95.55	96.54	96.17	99.97	98.03	95.36	93.13	99.91	96.40

Table 1. **与三个医疗基准的最新技术进行比较**。最佳结果以红色 标色。

CXR [19]。IU-胸部 X 光片包含胸部 X 光图像与诊断报告配对的数据集，提供了一组 2,573 个推理样本用于评估图文对齐和报告生成。哈佛-公平视觉语言医学模型侧重于多模态眼底成像中的公平性评估，有 4,285 个样本涵盖了多样的人口统计学和临床场景。MIMIC-CXR 是一个大型胸部放射图像与自由文本报告关联的数据集，包含了 3,470 个经过整理的样本，用于全面评估事实性和推理能力。我们采用 [1] 提供的官方分割和标准化评估协议。

指标。我们使用典型的分类指标进行评估，准确率 (**准确率**)，精确率 (**预备**。)，召回率 (**推荐**) 和 F1 分数。准确率衡量正确预测样本的比例，而精度和回忆分别量化模型识别相关实例和恢复所有真正疾病的能力。F1 分数提供了精确率和召回率的调和平均值，在类别不平衡场景中为模型性能提供平衡评估。

实现。我们的模型基于 Qwen-2.5-VL-3B 构建，使用开源的 GRPO 框架 [1] 进行训练。我们将提示的最大长度设置为 8,192，完成的最大长度设置为 2,048，允许进行长上下文建模。最大图像输入尺寸设置为 501,760 像素。我们对每个输入采样 6 次生成，以平衡探索与收敛，并将学习率设为 $5e^{-5}$，其他超参数保持默认值。模型训练了 2 个周期，使用了 bf16 混合精度、梯度检查点以及闪存注意力 [20]。训练在 4 个 NVIDIA A100 GPUs 80 GB 上耗时 10 小时，每个设备的批处理大小为 1，并且设置了梯度累积为 1。

[1] https://github.com/StarsfieldAI/R1-V

3.2. 与最新技术的比较

如 Tab. 1 所示，基线包括在医学数据上微调的 Qwen2.5-VL-3B [3] 和 LLaVA-Med v1.5 [5]，以及各种传统的后验增强（例如. 贪婪搜索）。它们在不同数据集上的表现各异，突显了一致泛化的挑战。在这些最先进的模型中，代表性的规则 [1] 通过校准检索和直接偏好优化强化学习 [9] 来解决其中一些限制，但受到推理深度的约束 [6]。虽然与基于 RL 的最先进的模型共享增强事实推理的目标，我们的方法是通过促进 GRPO 的事实伪标签来实现这一点的，从而避免了昂贵的标注，并且保持兼容性。值得注意的是，即使使用较小的 30 亿参数模型，我们的模型在所有模态和临床领域上的准确性也超过了更大的 70 亿参数最先进的模型超过 10%，这表明我们整体框架的优势。观察到的精确度和召回率提升表明，我们的训练策略不仅减少了幻觉现象，还增强了模型捕捉细微临床线索的能力，从而平衡了敏感性和特异性，并推进了事实可靠性和临床可解释性。定性案例展示在 figure 2 中。

3.3. 消融研究

3.3.1. 训练策略

PL-SFT	GRPO				Acc.	Pre.	Rec.	F1
	Acc.	Format	Fact	Cons.				
					61.21	37.32	66.91	47.91
✓					87.45	69.75	93.44	79.88
	✓	✓	✓	✓	82.95	67.94	67.62	67.77
✓	✓	✓			94.84	91.82	89.11	90.44
✓	✓	✓	✓		96.77	94.74	93.25	93.99
✓	✓	✓	✓	✓	97.63	95.62	95.48	95.55

Table 2. **训练策略和奖励在 IU-Xray 数据集上的消融研究**。PL 和 Cons. 分别代表伪标签和一致性。

Models	Acc.	Pre.	Rec.	F1
Base VLM [3]	61.21	37.32	66.91	47.91
RULE [1] (Ours)	97.63	95.62	95.48	95.55
Human	98.64	97.96	96.97	97.46

Table 3. **伪标签在 SFT 中的选择研究**。

在医学数据上的训练对于在这个专业领域中取得良好表现至关重要。如公式 Tab. 2 所示，基础通用模型 Qwen2.5-VL-3B 缺乏足够的医学专业知识，并且经常产生假阳性误诊，导致准确率仅为 37.32。监督学习与伪标签 (section 2.2) 和带有奖励信号的强化学习 (section 2.3) 显著改善了基准线 Qwen2.5-VL-3B。

有趣的是，这两种方法导致了不同的模型行为。监督微调（软交换技术）使模型能够通过模仿可靠且事实性的伪诊断 [1] 有效地识别阳性疾病，实现了 39.7% 的召回增益，并达到了高达 0.9344 的效果（参见第 2^{nd} 行）。相比之下，强化学习中我们采用最有效的变体群组 [6] 促进保守预测，正如精度、召回率和 F1 分数约为 0.67 的相似性所示（见第 3 行 $^{\text{rd}}$）。

该组合增强了基于事实的推理 [4]，带来了额外的 11.6% 增益，并将准确性推至 0.9763。值得注意的是，尽管事实伪标签存在不完美之处，但它们在 GRPO 框架中的整合与通过人工标注 SFT 实现的结果相匹敌，突显了我们方法的稳健性和可扩展性（Tab. 3）。

3.3.2. 奖励

奖励设计被广泛认为是影响强化学习性能的关键因素。如表 $^{\text{th}}$ 的第 4Tab. 2 行所示，在伪标签 SFT 的基础上，仅对二元诊断的准确性和简单的输出格式进行奖励，使多模态学习能够捕捉典型的疾病特征，从而获得显著的 +10.56F1 分数提升。这突显了强化学习后训练在医学应用中的有效性。事实奖励进一步丰富了输出内容，增加了领域特定的术语，而一致性奖励则通过使中间推理与最终答案保持一致来增强逻辑连贯性。每一项分别带来了 +3.55 和 +1.56 的进一步改进，强调了它们在加强事实推理和结构化生成响应中的互补作用。

4. 讨论与结论

MedFact-R1 建立了一个稳健的两阶段框架用于事实性医疗推理，集成了伪标签生成和基于 GRPO 的强化学习。它在各种医疗问答基准测试中显著提升了事实性和可靠性，突显了结合外部知识与自适应策略优化的价值。实验还表明，SFT 初始化和奖励设计对于确保稳定有效的训练至关重要。尽管取得了这些进步，但仍然存在扩展复杂现实临床场景以

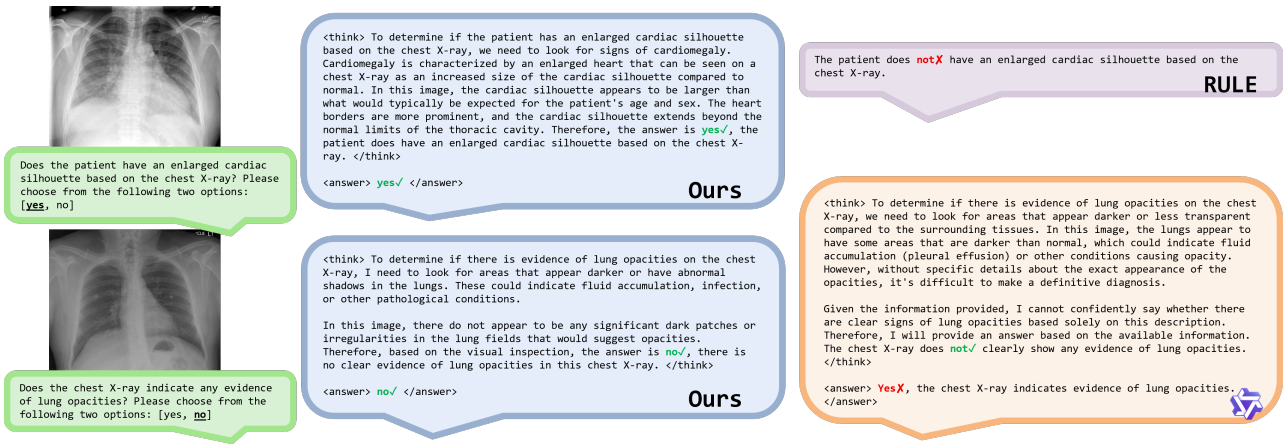

Fig. 2. **案例分析表明，我们的医学诊断相比规则和** Qwen2.5-VL **具有更高的准确性**。正确答案用 ✓ 标记，而错误答案则用 ✗ 表示。

及应对罕见或模糊案例时保持稳健性的挑战。未来的研究应探索更丰富的奖励函数、视频推理及部署挑战，包括安全性和公平性问题。我们相信，在知识整合和 RL 激励方面的持续创新将有助于推进可信且通用的医疗 AI。

参考文献

[1] Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao, "RULE: Reliable multimodal RAG for factuality in medical vision language models," in EMNLP, 2024.

[2] 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, 2023.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al., "Qwen2. 5-VL technical report," arxiv, 2025.

[4] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al., "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," arXiv, 2025.

[5] 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 NeurIPS, 2023.

[6] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al., "DeepSeekMath: Pushing the limits of mathematical reasoning in open language models," arXiv, 2024.

[7] Kaixuan Fan, Kaituo Feng, Haoming Lyu, Dongzhan Zhou, and Xiangyu Yue, "SophiaVL-R1: Reinforcing MLLMs reasoning with thinking reward," arXiv, 2025.

[8] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov, "Proximal policy optimization algorithms," arXiv, 2017.

[9] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and

Chelsea Finn, "Direct preference optimization: Your language model is secretly a reward model," in NeurIPS, 2023.

[10] Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P Langlotz, and Dan Jurafsky, "Improving factual completeness and consistency of image-to-text radiology report generation," in NAACLW, 2021.

[11] Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis P Langlotz, "Improving the factual correctness of radiology report generation with semantic rewards," in EMNLP, 2022.

[12] Peng Xia, Jinglu Wang, Yibo Peng, Kaide Zeng, Xian Wu, Xiangru Tang, Hongtu Zhu, Yun Li, Shujie Liu, Yan Lu, et al., "MMedAgent-RL: Optimizing multi-agent collaboration for multimodal medical reasoning," arXiv, 2025.

[13] Huihui Xu, Yuanpeng Nie, Hualiang Wang, Ying Chen, Wei Li, Junzhi Ning, Lihao Liu, Hongqiu Wang, Lei Zhu, Jiyao Liu, et al., "MedGround-R1: Advancing medical image grounding via spatial-semantic rewarded group relative policy optimization," in MICCAI, 2025.

[14] Sunan He, Yuxiang Nie, Zhixuan Chen, Zhiyuan Cai, Hongmei Wang, Shu Yang, and Hao Chen, "MedDr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning," arXiv, 2024.

[15] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao, "MMed-RAG: Versatile multimodal RAG system for medical vision language models," in ICLR, 2025.

[16] Liwen Sun, James Zhao, Megan Han, and Chenyan Xiong, "Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation," in NAACL, 2025.

[17] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," JAMIA, 2016.

[18] Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al., "FairCLIP: Harnessing fairness in vision-language learning," in CVPR, 2024.

[19] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng, "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," arXiv, 2019.

[20] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré, "FlashAttention: Fast and memory-efficient exact attention with IO-awareness," in NeurIPS, 2022.