MEDFACT-R1: TOWARDS FACTUAL MEDICAL REASONING VIA PSEUDO-LABEL AUGMENTATION

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

Ensuring factual consistency and reliable reasoning remains a critical challenge for medical vision-language models. We introduce MEDFACT-R1, a two-stage framework that integrates external knowledge grounding with reinforcement learning to improve the factual medical reasoning. The first stage uses pseudo-label supervised fine-tuning (SFT) to incorporate external factual expertise; while the second stage applies Group Relative Policy Optimization (GRPO) with four tailored factual reward signals to encourage self-consistent reasoning at deployment time without relying on external RAG. Across three public medical QA benchmarks, MEDFACT-R1 delivers up to $\mathbf{22.5}\%$ absolute improvement in factual accuracy over previous state-of-the-art methods. Ablation studies highlight the necessity of pseudo-label SFT cold start and validate the contribution of each GRPO reward, underscoring the synergy between knowledge grounding and RL-driven reasoning for trustworthy medical AI. Codes are released at https://github.com/Garfieldgengliang/MEDFACT-R1.

***Index Terms*—** Medical Vision-Language Models, Factual Medical Reasoning, Pseudo-Labeling, GRPO

1. INTRODUCTION

Medical diagnosis represents one of the most critical frontiers in signal processing and machine learning, embodying the vision of technology serving humanity. It demands exceptional expertise, stringent accuracy, and the ability to reason over inherently complex data. Yet, progress is hindered by the scarcity of high-quality diagnostic data, restricted by both professional requirements and privacy constraints. These limitations complicate model training and often result in unreliable behavior - such as reliance on spurious correlations, misjudgments, and missed diagnoses. In the medical domain, where decisions directly affect human lives, such errors are unacceptable. Overcoming these challenges is essential for developing deep learning systems that can achieve reliable performance in real-world clinical practice [1].

Recently, large-scale Vision-Language Models (VLMs) [2, 3, 4] have advanced rapidly, transforming various industries. Their extension to the medical domain has shown promising potential, with recent efforts [5] curating medical datasets and fine-tuned VLMs for specialized applications. However, factual reliability remains a major obstacle: existing models often generate hallucinations and factual errors in high-stake scenarios. To address this, RULE [1] introduces risk-controlled Retrieval-Augmented Generation (RAG), which balances external retrieval and internal knowledge, yielding notable factuality improvements.

Concurrently, many post-training efforts have successfully exploited the knowledge and potential of vision-language models themselves, among which advanced Reinforcement Learning (RL) has emerged as a prominent example. Unlike supervised learning via next-token prediction, reinforcement learning optimizes task policies using reward signals without relying on detailed annotations. GRPO [6, 7], one of the most advanced RL post-training methods, surpasses supervised fine-tuning in generalization by unlocking "aha" moments in reasoning [4], which sets it apart from prior approaches such as PPO [8], DPO [9] and traditional ones [10]. Despite the impressive results, sufficient domain knowledge is found essential to prevent unrealistic or verbose outputs, which often rely on the costly curation of diverse annotated medical datasets [11, 12].

To address the challenges of data scarcity in RL and reliance on external knowledge, we propose MEDFACT-R1 (Fig. 1). It is a two-stage framework that combines SFT on factual pseudo-diagnosis data with RL-incentivized reasoning to activate factual capabilities without external reference. In the *first* stage, we apply a generator with factuality risk control to generate pseudo-diagnosis data for SFT, thereby more effectively expanding medical knowledge exposure and strengthening the model's foundation. In the *second* stage, we adopt GRPO post-training to help the model fully digest and reflect on valuable diagnostic data. To better align with the medical diagnosis scenario, we carefully design four reward components, considering answer correctness, formal presentation, contextual relevance, and self-consistency with factual reasoning. Our GRPO encourages reasoning grounded in facts and allows the model to generalize beyond the observed cases and improve medical factuality.

The experimental results reveal strong and consistent improvements across several public medical Question Answer-

[†]These authors contributed equally to this work.
[*]Corresponding author.

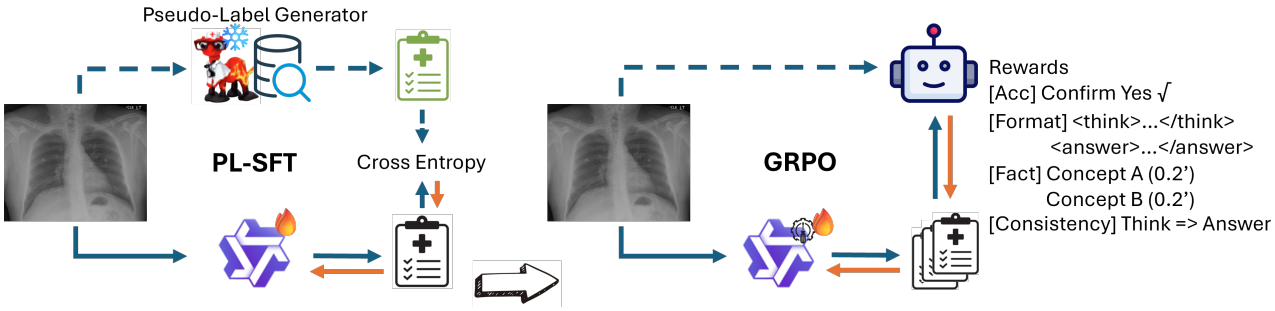

Fig. 1. **Overview of our method MEDFACT-R1 consisting of two stages, Pseudo-Label Supervised Fine-Tuning & GRPO-based Reinforcement Learning.** The **blue** and **orange** arrows represent the feedforward and backpropagation flow, respectively.

ing benchmarks. MEDFACT-R1 enhances all evaluation metrics, yielding more factual and reliable diagnostic outputs and validating the effectiveness of our framework in advancing the state-of-the-art medical factuality reasoning. In addition, our ablation studies provide deeper insights into the necessity of a pseudo-label SFT start and clarify the individual contributions of each factual reward.

In summary, our contributions are as follows: **1)** We propose better integration of advanced reinforcement learning GRPO into medical QA via a two-stage training pipeline combining SFT and RL post-training paradigms to enhance the VLM's own factual reasoning; **2)** We initialize training with SFT on factual *Pseudo-labels* generated via retrieval augmentation, to effectively absorb external medical knowledge; **3)** To address sparse reward signals and further improve factuality, *factual rewards* are tailored for GRPO post-training, reducing hallucinations and promoting self-justification; **4)** MEDFACT-R1 sets a new state-of-the-art, achieving scores exceeding 95% on all evaluation metrics across diverse medical QA benchmarks, with gains of up to 22.5% over prior methods.

2. MEDFACT-R1

2.1. Task Formulation

Medical visual Question Answering (QA) is the task of answering medical questions based on images [5]. Formally, given an image I and question Q, the model generates an answer in free-form text. For VLMs, the first token of the generated answer is expected to be either "*yes*" or "*no*"; this token is mapped to a binary label $\hat{A} \in \{0, 1\}$, indicating the presence or absence of disease.

2.2. Supervised Fine-Tuning with Pseudo-Labels

We begin with Supervised Fine-Tuning (SFT) using next-token prediction under maximum likelihood estimation to establish a foundation for factual reasoning. To this end, we generate pseudo-labels via calibrated retrieval to mitigate factuality risks and preference alignment to harmonize external priors with internal knowledge [1]. This strategy distills external medical knowledge into compact supervision signals that suppress hallucinations and reinforce the factual consistency.

2.3. GRPO Post-Training

Post-SFT, GRPO [6] guides policy updates via rule-based rewards on generated outputs using group-based Monte Carlo advantage estimation and policy gradient. We design four complementary types of reward values for GRPO-based reinforcement learning, with the total reward given by the normalized sum of these components from generated answers:

Accuracy Reward. We assess the correctness of the predicted answers by comparing them with pseudo-labels. For binary classification, the reward is 1.0 for exact matches and 0 otherwise.

Format Reward. To encourage structured reasoning, we require the outputs to include a thought process enclosed within `<think>` and `</think>` tags and a concise final answer within `<answer>` and `</answer>` tags. A reward of 1 is assigned only if all four tags appear exactly once and no extraneous content exists outside these regions; otherwise, the reward is 0.

Fact Reward. To promote factual alignment in medical reasoning, we design a reward based on the presence of clinically grounded concepts in the model output. For each training question, a small set of domain-specific concepts is extracted using GPT-4 [2], serving as factual anchors. For example, the question *"Does the chest radiograph show any signs of lung infection or congestion?"* yields concepts such as *"lung infection, congestion, chest radiography"*. Each concept correctly reflected in the answer contributes 0.2 points, counted once. This guides the model to express medically relevant knowledge, offering a targeted signal for factual grounding.

Consistency Reward. To reinforce factual and logical coherence in clinical contexts, we introduce a consistency reward that evaluates the alignment between the reasoning and final answer. We use GPT-4 to assess contextual consistency, referencing curated examples of medically sound and unsound outputs. A reward of 1 is given if the answer is logically

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 [13]	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 [14]	ICLR'25	89.54	-	-	80.72	87.94	-	-	92.78	83.57	-	-	88.49
FactMM-RAG [15]	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. Comparisons with the state-of-the-art across three medical benchmarks. The best results are colored in red.

supported by its reasoning, particularly in terms of clinical interpretation; otherwise, a penalty of -0.5 is applied. This encourages the model to maintain internal consistency when drawing conclusions from the medical evidence.

3. EXPERIMENTS

3.1. Datasets and Implementation

Datasets. Experiments are conducted on the three medical benchmarks: **IU-Xray** [16], **Harvard-FairVLMed** [17] and **MIMIC-CXR** [18]. IU-Xray consists of chest X-ray images paired with diagnostic reports, providing a set of 2,573 inference samples for evaluating image-text alignment and report generation. Harvard-FairVLMed focuses on fairness assessment in multimodal fundus imaging, with 4,285 samples spanning diverse demographic and clinical scenarios. MIMIC-CXR is a large-scale dataset of chest radiographs associated with free-text reports, with 3,470 samples curated for a comprehensive evaluation of factuality and reasoning. We adopt the official splits and standardized evaluation protocols provided by [1].

Metrics. We evaluate using typical classification metrics, Accuracy (**Acc.**), Precision (**Pre.**), Recall (**Rec.**), and **F1** score. *Accuracy* measures the proportion of correctly predicted samples, whereas *Precision* and *Recall* quantify the model's ability to identify relevant instances and recover all true positive diseases, respectively. The *F1* score provides a harmonic mean of Precision and Recall, offering a balanced assessment of the model performance in scenarios with class imbalance.

Implementation. Our model is built upon Qwen-2.5-VL-3B, trained with the open-sourced GRPO framework[1]. We set the maximum prompt length to 8,192 and the maximum completion length to 2,048, allowing for long-context modeling. The maximum image input size is set to 501,760 pixels. We sample 6 generations per input to balance exploration and convergence with a learning rate of $5e^{-5}$ and keep the other hyperparameters as default. The model is trained for 2 epochs with `bf16` mixed precision, gradient checkpointing, and FlashAttention [19]. Training takes 10 hours on 4 NVIDIA A100 GPUs 80 GB with a per-device batch size of 1 and gradient accumulation set to 1.

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

3.2. Comparisons with the State-of-the-Art

As shown in Tab. 1, the baselines include Qwen2.5-VL-3B [3] and LLaVA-Med v1.5 [5] fine-tuned on medical data, along with various traditional post-hoc enhancements (*e.g.* Greedy Search). They exhibit variable performance across datasets, highlighting the challenges of consistent generalization. Among the SOTAs, the representative RULE [1] addresses some of these limitations through calibrated retrieval and Direct Preference Optimization RL [9], yet is constrained by reasoning depth [6]. While sharing the goal of enhancing factual reasoning with RL-based SOTAs, our method achieves this via factual pseudo-labels that facilitate GRPO, avoiding costly annotations and remaining compatible with them. Notably, even with a smaller 3B model, our model significantly outperforms the larger 7B SOTA counterparts by over 10% in accuracy across modalities and clinical domains, demonstrating the superiority of our overall framework. The observed gains in both precision and recall suggest that our training strategy not only reduces hallucinations but also enhances the model's ability to capture subtle clinical cues, thereby balancing sensitivity and specificity while advancing factual reliability and clinical interpretability. Qualitative cases are shown in Fig. 2.

3.3. Ablation Studies

3.3.1. Training Strategies

Training on medical data is crucial for achieving strong performance in this specialized domain. As shown in Tab. 2, the base generalist model Qwen2.5-VL-3B lacks sufficient medical expertise and frequently produces false-positive misdiagnoses, yielding a low precision of 37.32. Supervised learning with pseudo-labels (Sec. 2.2) and reinforcement

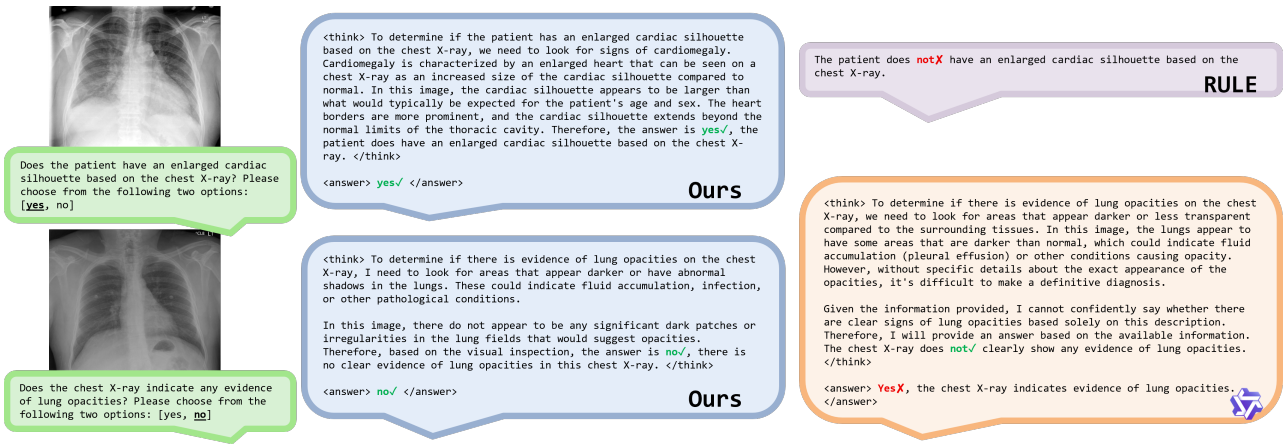

Fig. 2. **The case analyses show enhanced factuality of our medical diagnosis compared to RULE and QWEN2.5-VL.** Correct answers are marked with ✓, while incorrect ones are denoted by ✗.

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. **Ablation studies of the training strategies and rewards on the IU-Xray dataset.** PL and Cons. stand for Pseudo-Label and Consistency, respectively.

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. **The choice study of pseudo-labels for SFT.**

learning with reward signals (Sec. 2.3) significantly improve the QWEN2.5-VL-3B baseline.

Interestingly, these two approaches lead to distinct model behaviors. Supervised fine-tuning (SFT) enables the model to effectively identify positive diseases by imitating reliable and factual pseudo-diagnoses [1], achieving a recall gain of 39.7% and reaching up to 0.9344 (see the 2^{nd} row). In contrast, reinforcement learning, where we adopt the most effective variant GRPO [6] promotes conservative predictions, as evidenced by the similar precision, recall, and F1 scores of approximately 0.67 (see the 3^{rd} row).

The combination enhances fact-grounded reasoning [4], yielding an additional 11.6% gain and pushing accuracy to 0.9763. Notably, despite imperfections in the factual pseudo-labels, their integration within the GRPO framework rivals results achieved via human-annotated SFT, underscoring the robustness and scalability of our approach (Tab. 3).

3.3.2. Rewards

Reward design is widely recognized as a critical factor influencing the performance of reinforcement learning. As shown in the 4^{th} row of Tab. 2, building on pseudo-label SFT, rewarding only the *accuracy* of the binary diagnosis and a simple output *format* enables the VLM to capture typical disease characteristics, yielding a strong $+10.56$ F1 score gain. This highlights the effectiveness of RL post-training for medical applications. The *fact* reward further enriches the outputs with domain-specific terminology, whereas the *consistency* reward enhances logical coherence by aligning intermediate reasoning with final answers. Each contributes further improvements of $+3.55$ and $+1.56$, respectively, underscoring their complementary roles in strengthening factual reasoning and the structured generation of responses.

4. DISCUSSIONS & CONCLUSIONS

MEDFACT-R1 establishes a robust two-stage framework for factual medical reasoning, integrating pseudo-label generation and GRPO-based reinforcement learning. It delivers substantial gains in factuality and reliability across diverse medical QA benchmarks, highlighting the value of combining external knowledge with adaptive-policy optimization. The experiments also reveal that SFT initialization and reward design are critical for ensuring stable and effective training. Despite these advances, challenges remain in scaling complex real-world clinical scenarios and ensuring robustness against rare or ambiguous cases. Future research should explore richer reward functions, video reasoning, and deployment challenges, including safety and fairness. We believe that continued innovation in knowledge integration and RL incentivization will be promising for advancing trustworthy and generalizable medical AI.

5. ACKNOWLEDGMENTS

The work was supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2022), the National Natural Science Foundation of China (No. 62406298), and the Beijing Natural Science Foundation (No. L244043). We would also like to thank the ACs, reviewers, Dr. Angela Yao, and Dan Wang for their valuable suggestions.

References

[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] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

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