arXiv:2603.01301v1 [cs.CV] 1 Mar 2026

强化学习何时能助力医疗视觉语言模型？剖析视觉、监督微调与强化学习带来的收益

Ahmadreza Jeddi[1, 2, 3*], Kimia Shaban[1, 2, 3*], Negin Baghbanzadeh[2, 4*], Natasha Sharan[1], Abhishek Moturu[1,2,3], Elham Dolatabadi[2,4], 和 Babak Taati[1,2,3]

[1] 多伦多大学，加拿大 [2] Vector研究所，加拿大 [3] KITE研究所，大学健康网络，加拿大 [4] 约克大学，加拿大
ajeddi@cs.toronto.edu

摘要。 强化学习（RL）越来越多地用于对医疗视觉语言模型（VLM）进行后训练，但强化学习究竟能否提升医学视觉推理，还是主要强化监督微调（SFT）已形成的行为模式，仍不明确。我们开展了一项受控研究，从视觉、监督微调和强化学习三个维度剖析这些影响。我们使用 MedMNIST 作为多模态测试平台，通过将视觉语言模型的视觉塔与纯视觉基线进行基准测试来探究视觉感知；通过比较 Accuracy@1 与 Pass@K 来量化推理支持度和采样效率；并评估强化学习何时能弥合支持度差距，以及收益如何迁移至不同模态。我们发现，当模型已有非微不足道的支持度（Pass@K 较高）时，强化学习效果最佳：它主要通过锐化输出分布来提升 Acc@1 和采样效率，而监督微调则扩展支持度，使强化学习得以发挥作用。基于这些发现，我们提出一种边界感知方案，并以此为基础，使用 PMC 多项选择视觉问答数据的一个小型均衡子集，对一个以 OctoMed 初始化的模型进行强化学习后训练；该模型在六个医学视觉问答基准上取得了较强的平均表现。项目页面：https://medbridgerl.github.io

关键词： 医学视觉语言模型 · 强化学习后训练 · 推理能力。

1 引言

医学视觉语言模型正逐渐成为临床影像的统一交互界面。通过联合处理图像和文本查询，生成准确且可解释的回答，它们可应用于视觉问答、报告辅助和决策支持。然而，仅有准确性不足以满足临床部署的要求，临床部署还要求模型在不同模态和机构中具备可靠性、透明度和鲁棒性。受近期“推理型”大语言模型进展的启发，人们对采用后训练方法的兴趣日益浓厚，尤其是监督微调（SFT）和强化

* 贡献相同

学习与可验证奖励（RLVR），以鼓励形成连贯的推理轨迹并提升可信度 [3,18,20,6,29,21]。

现有的大多数医学视觉语言模型强化学习流程都紧密遵循更广泛的可验证奖励强化学习文献：从一个基于互联网规模多模态数据预训练的通用开源视觉语言模型开始，然后由于难以获取高质量的可验证奖励和经过筛选的推理数据，只在相对较小的医学子集上进行有限的监督微调和/或强化学习。尽管有几种方法报告了可喜的收益，但结果往往因模态和任务而异，跨模态泛化仍不稳定，即使是用多种成像类型训练的模型也是如此 [11,5,18]。这留下了一些尚未解答的根本问题：观察到的提升有多少来自视觉感知而非语言侧对齐，有多少可归因于监督微调而非强化学习，以及在什么条件下强化学习在医学场景中确实值得付出成本。

与此同时，近期大语言模型领域的研究对一种常见假设提出了质疑：可验证奖励强化学习会自然地产生超越基础模型的新推理能力。通过探究 Pass@K 的表现，Yue 等人 [26] 认为，强化学习主要是重塑输出分布，以更高效地采样正确答案，而往往并未扩展其底层支持集。Zhang 等人 [27] 的受控实验进一步表明，后训练收益在很大程度上取决于基础模型的能力边界所在，以及训练数据对该边界的覆盖情况。在这项工作中，我们将这些问题带入医学视觉语言模型领域，并旨在刻画：（i）不同模态下的感知极限；（ii）通过 Pass@K 衡量的推理能力边界；（iii）强化学习何时能够提高采样效率，以及这些效果能否迁移到不同医学模态。

贡献。 我们使用 MedMNIST-v2 [24] 作为受控的多模态测试平台，以区分视觉、监督微调和强化学习在医学视觉语言模型后训练中的作用。具体而言，我们：

1. 通过纯视觉线性评估探测视觉感知，以识别表征瓶颈；2. 比较各模态下的 Accuracy@1 和 Pass@K，刻画每个模型的支持边界；3. 在基础模型和经监督微调初始化的模型上进行测试，评估 GRPO 风格强化学习何时能助益医学视觉语言模型，包括域内收益以及模态内和跨模态迁移；4. 将这些发现转化为分阶段方案（先提供桥接支持，再通过强化学习提升能力），并通过对 Octo Med 基线模型进行后训练，在 PMC-VQA [30]，上验证该方案，使其在多个医学视觉问答基准上取得优异结果。

2 相关工作

医学视觉语言模型与后训练。 医学视觉语言模型经由大规模多模态预训练，再经领域适配而不断发展，最常见的做法是在医学图像-文本对、视觉问答指令数据以及报告式语料上采用监督微调 [16,17,3,23]。最近，若干工作采用了可验证奖励强化学习/群体相对策略优化式后训练，以鼓励结构化推理轨迹并提高答案

在有限监督下的可靠性 [18,11,8,28,22,14]。尽管其中许多流程遵循一种先监督微调再强化学习模式，但数据混合、奖励设计与模态覆盖方面的差异，使得我们难以分离出驱动改进的因素，也难以预判强化学习何时会有帮助。

强化学习是否能在基础模型之外带来推理能力？近期通用大语言模型研究提出疑问：可验证奖励强化学习是拓展了推理能力，还是主要通过重新加权基础模型分布中已有的解来提高采样效率 [26,10,27]。我们将这些问题带到医学视觉语言模型领域。与标准的监督微调→强化学习 流程相比，我们（i）结合边界视角（Accuracy@1 与 Pass@K）和按模态进行的感知探测，判断强化学习何时能够发挥作用；（ii）提出一种优先确保支持集的方案：仅在需要时通过有针对性的桥接提高 Pass@K，然后将强化学习作为锐化阶段，而非假定强化学习普遍有益。

3 医学视觉语言模型中的视觉、监督微调与强化学习解耦分析

我们研究监督微调和强化学习后训练如何影响医学视觉语言模型，以及这些阶段何时能带来超越强大基础模型的收益。我们的分析围绕三个研究问题展开，每个问题均通过受控评估进行探究，并以简短结论概述。我们首先介绍共同的实验设置。

测试平台。 我们使用 MedMNIST-v2 [24], ，这是一个受控套件，涵盖三种成像模态和十二项任务，并采用统一的输入分辨率（224×224）。为确保公平比较，我们在官方测试集划分中的平衡子集上进行评估。MedMNIST 非常适合我们的目标：它能够对表征进行纯视觉探测，基于类别标签开展视觉语言模型风格的选择题评估，并凭借其规模和标准化格式支持高效的可验证奖励强化学习实验。

模型。 我们的基础模型是 Qwen2.5-VL-7B-Instruct [2], ，记为 M_{Base}。为单独考察医学监督微调的效果，我们使用 OctoMed [17] 作为 M_{SFT}。作为经过强化学习后训练的代表性医学视觉语言模型，我们纳入 QoQ-Med [5], ，记为 M_{RL}。评估最终训练方案和检查点时，我们会在第 §4 节介绍其他基线。

3.1 研究问题 1：医学视觉语言模型的视觉表征有多强?

医学视觉语言模型的性能可能受视觉塔本身的限制。尽管以往研究曾探测对比编码器（例如 CLIP [19]）在医学迁移方面的表现，但针对 Qwen 等端到端自回归视觉语言模型的类似分析并不常见。因此，我们首先测量其底层视觉表征是否已能区分医学任务。

表1：在 MedMNIST-v2 各模态任务上对冻结的视觉编码器进行线性探测。MedViT-v2 被纳入作为强大的纯视觉参考模型。

方法	放射学						显微镜检查			可见光摄影		
	乳腺	胸部	肺炎	器官A	器官C	器官S	血液	病理	组织	皮肤	光学相干断层扫描	视网膜
MedViT-V2 [15]	91.00	96.70	97.30	97.30	96.10	85.10	98.70	97.10	71.60	81.70	95.20	57.80
M_{Base}	80.95	13.33	92.52	85.27	79.31	65.70	96.19	89.12	48.26	65.22	84.30	44.00
M_{SFT}	84.52	16.51	93.35	89.01	83.59	70.90	96.19	93.35	51.99	67.08	91.10	52.00
M_{RL}	75.00	16.19	92.95	85.52	80.35	64.96	96.86	88.50	52.88	69.57	84.80	51.00

我们冻结每个模型的视觉编码器（视觉变换器），并在 MedMNIST-v2 上通过线性探测评估表征质量。我们还纳入 MedViT-v2 [15]，作为强有力的纯视觉参考模型。

表1报告了 M_{Base}、 M_{SFT}和 M_{RL} 在不同模态上的探测准确率。我们利用这些分数来判断视觉语言模型的失败是否主要受限于感知：当模型冻结的视觉变换器无法对某个数据集的类别进行线性区分时，监督微调或强化学习带来的收益自然有限；而较高的探测准确率则表明，错误源于语言侧对齐以及解码/采样，而非视觉特征。

发现 **1.** 基础模型在许多 MedMNIST 任务上已经具备区分度尚可的视觉特征，而医学监督微调能进一步改善这些特征，尤其是在较弱的数据集上。强化学习并未持续提升视觉变换器的探测准确率，这表明其作用主要体现在采样和对齐方面，而非视觉表征。若干数据集的表现仍远低于 MedViT-v2 的参考水平，表明存在限制下游收益的感知瓶颈。

3.2 研究问题 2：医学视觉语言模型的推理能力如何?

即使视觉表征足够充分（研究问题 1），医学视觉语言模型仍可能表现不佳，因为模型的概率分布中虽然包含正确答案，但在贪心解码下无法稳定地生成这些答案。我们通过比较单样本准确率与多样本成功率，将此现象概括为支持边界，并探究后训练是扩大模型的支持范围，还是主要提升采样效率。

对于每个包含 C 个类别的 MedMNIST 任务，我们将分类转化为包含 C 个选项的视觉语言模型多项选择提示，并要求模型输出选项字母。我们报告 (i) 贪心解码下的 Accuracy@1，以及 (ii) Pass@K，即至少一个 K 独立样本答对的概率。我们使用$K \in \{1, 2, 4, 8, 16\}$，温度为 0.7，top-p 0.9。输出限定为 <think> 和 <answer> 标签；我们使用 Qwen2.5-VL-32B-Instruct 验证预测选项是否正确。

表 2 报告了各项任务的 Accuracy@1，图 1 则展示了按模态分组的 Pass@K 曲线。二者共同揭示了贪心解码之外存在多少潜在支持，以及不同后训练阶段如何改变这一边界。

表 2：按模态分组的 MedMNIST-v2 任务中，视觉语言模型在贪心解码下的 Accuracy@1 评估结果。

方法	放射学						显微镜检查			可见光摄影		
	乳腺	胸部	肺炎	器官A	器官C	器官S	血液	病理	组织	皮肤	光学相干断层扫描	视网膜
M_{Base}	54.99	8.70	50.53	14.39	10.94	13.14	13.16	26.48	10.39	33.54	32.62	20.19
M_{SFT}	79.17	13.22	86.00	18.66	16.03	19.38	76.00	64.53	12.66	52.33	82.78	22.19
M_{RL}	69.79	8.03	62.41	10.91	8.94	10.75	13.95	34.09	8.96	27.48	28.72	20.62

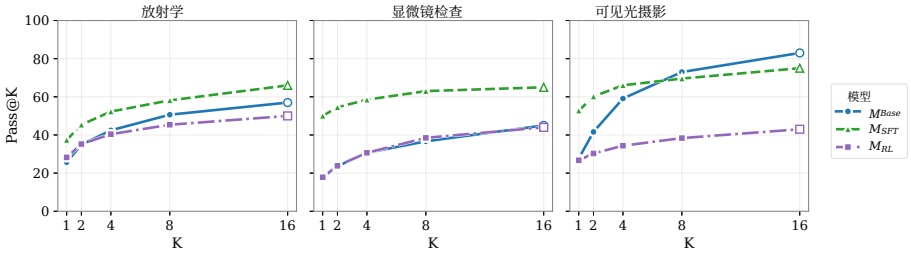

图 1：按模态分组的 MedMNIST-v2 上的 Pass@K 曲线。

发现 **2.** 在各项任务中，Accuracy@1 往往远低于 Pass@K，这表明存在大量贪心解码无法实现的潜在支持。医学监督微调提高了 Accuracy@1 和 Pass@K，这与覆盖范围和对齐能力的提升相符。相比之下，强化学习后训练基线未能持续提高 MedMNIST 上的 Accuracy@1，且经常降低 Pass@K，这表明其在未拓展支持范围的情况下进行了分布锐化，有时还使边界变得更窄。

3.3 研究问题 3：强化学习何时能助力医学视觉语言模型?

研究问题 1–研究问题 2 展现出一幅简单的图景。基础视觉语言模型在大规模预训练期间接触过许多这类模态，但覆盖并不均衡。大规模医学监督微调阶段（OctoMed）提升了视觉感知能力，并提高了各项任务中的 Pass@K。相比之下，现成的强化学习医学检查点（QoQ-Med）无法稳定迁移到 MedMNIST，甚至可能降低 Pass@K，这与强化学习往往会重塑分布而非增加新能力的观点一致 [26]。在此，我们探讨在受控设置下强化学习何时真正有用，以及它对准确率和支持度有何影响。

实验设置。 我们使用训练集，针对每种模态的一项任务进行针对性强化学习后训练：OrganAMNIST（放射学）、PathMNIST（显微镜检查）和 OCTMNIST（可见光摄影）。对于每项任务，我们训练两种强化学习变体，分别初始化自（i）基础模型 M_{Base} 和（ii）监督微调桥接模型 M_{SFT}。我们使用一致性感知 GRPO 流程 [4,9] 而非原始 GRPO 来

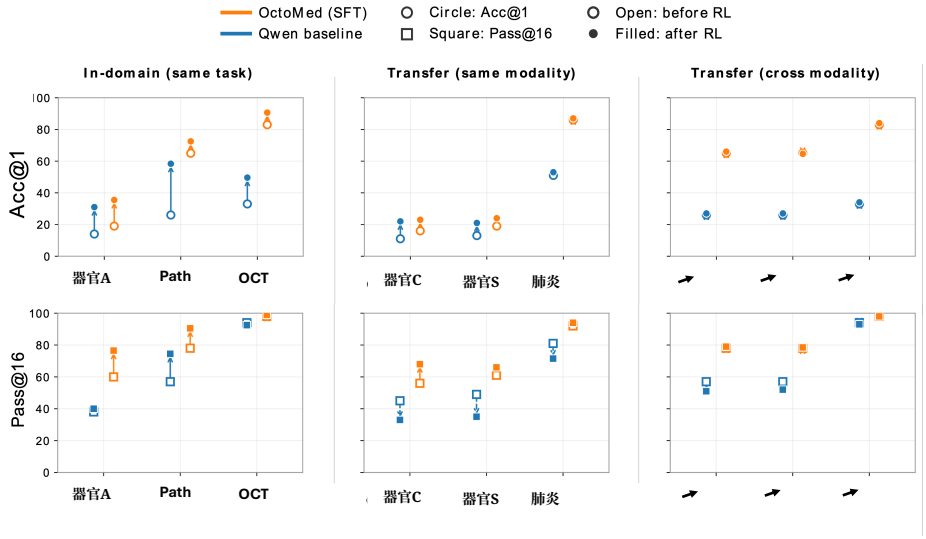

图2：强化学习前后 Acc@1 与 Pass@16 的变化，来自 M_{Base} 和 M_{SFT}，涵盖域内、模态内和跨模态评估。

在小规模医学数据上稳定优化。我们评估三种情形：（1）域内（与强化学习训练相同的任务）；（2）模态内迁移，即在 OrganAMNIST 上训练，并在 OrganCMNIST/OrganSMNIST（同一系列、不同视图）和 PneumoniaMNIST（更大偏移）上测试；（3）跨模态迁移，包含三种偏移：光学相干断层扫描→病理、器官A→病理，以及器官A→光学相干断层扫描。在整个过程中，我们报告强化学习前后的 Acc@1 和 Pass@16。

结果。 图2表明，强化学习主要起到锐化作用：它通过将更多概率质量放在模型支持集中已有的答案上，从而提高 Acc@1。与 [27],一致，即使是适度的桥接也足以产生明显的域内收益，但收益幅度受起始 Pass@16 的限制。对于较小的模态内偏移（器官A→器官C/器官S），强化学习通常会提高 Acc@1，而从基础模型初始化的强化学习则更常以 Pass@16 为代价。对于较大的模态内偏移（器官A→肺炎），收益较小，且 Pass@16 下降对 M_{Base} 更为常见。在跨模态偏移下，强化学习对 Acc@1 几乎没有改变，并可能略微降低 Pass@16，尤其是从未经桥接的基础模型开始时。

发现 **3.** 当模型已经具备一定的支持能力时，强化学习最为有效：它能够锐化输出分布、提升 Acc@1（采样效率），并缩小域内及小幅模态内分布偏移时的 Acc@1–Pass@K 差距。当支持能力较弱时（偏移较大或发生跨模态偏移），准确率提升有限，Pass@K 甚至可能下降；对于未经桥接的基础模型，这一现象尤为明显。

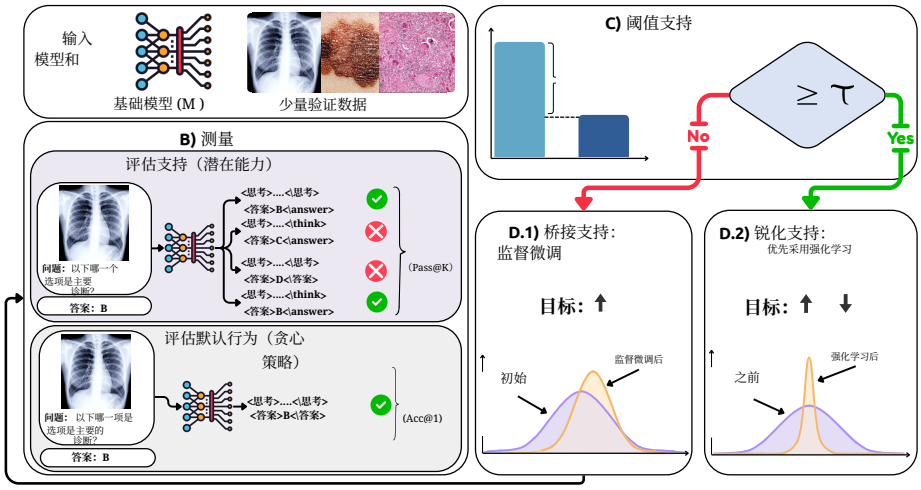

图 3：边界感知方案概览。我们首先使用 Pass@K 和 Acc@1 诊断支持度，然后决定采用桥接还是强化学习锐化。

4 从分析到实践：强化学习后训练方案

上一节表明，只有当模型对目标分布具备一定的支持能力后，强化学习才最有用。在实践中，强化学习可以通过锐化模型的输出分布，高效提升 Acc@1，但其效果仍受模型底层 Pass@K 的限制。接下来，我们将这些观察归纳为一条简单的决策规则，并将其应用于真实的医学视觉问答基准（图 3）。

支持度与锐化。对于任务分布 $\mathcal{D}$ 和固定的采样预算 K，我们定义：

$$S_K(\mathcal{D}) = \text{Pass@K}(\mathcal{D}), \qquad A(\mathcal{D}) = \text{Acc@1}(\mathcal{D}), \qquad G_K(\mathcal{D}) = S_K(\mathcal{D}) - A(\mathcal{D}), \tag{1}$$

其中，S_K 衡量潜在支持度，A 衡量贪心解码下的默认行为，G_K 是支持度差距。

配方（边界感知后训练）。给定任务分布 $\mathcal{D}$ 上的目标可靠性阈值 τ，我们采用以下分阶段规则：

1. **诊断支持度**。在一个小型验证集上估计 $S_K(\mathcal{D})$ 和 $A(\mathcal{D})$。2. **如果支持度较弱，则进行桥接**。如果 $S_K(\mathcal{D}) < \tau$，应优先进行桥接：添加任务邻近或模态邻近数据，并进行监督微调（或持续训练），以提升S_K。此步骤旨在扩大覆盖范围，而非锐化。3. **支持度充足后再进行锐化**。如果 $S_K(\mathcal{D}) \geq \tau$，则应用强化学习来提高采样效率，在提升 $A(\mathcal{D})$ 的同时监测 S_K 和 G_K，以避免过度锐化导致支持度崩溃。

表 3：医学视觉语言模型在多模态医学基准上的表现。最佳结果以粗体显示；次佳结果以下划线标出。

方法	PMC [30]	MMMU [25]	MedX-M [31]	PathVQA [7]	SLAKE [13]	VQA-Rad [12]	Avg.
Qwen2.5-VL-7B-Instruct [2]	53.00	55.29	23.50	65.00	68.00	74.00	56.47
QoQ-Med-7B [5]	51.00	<u>61.18</u>	25.00	63.50	68.50	74.00	57.20
MedVLThinker-7B [8]	55.00	54.12	27.50	57.50	63.50	64.00	53.60
OctoMed-7B [17]	<u>55.50</u>	56.47	**35.00**	63.00	<u>84.00</u>	**79.00**	<u>62.16</u>
MedVLM-R1-2B [18]	50.50	39.41	22.50	61.50	71.00	63.50	51.40
Med-R1-2B [11]	53.00	41.18	23.50	**66.50**	69.50	62.00	52.61
我们的方法（7B）	**59.00**	**62.94**	<u>34.50</u>	<u>65.50</u>	**88.0**	**79.00**	**64.91**

这一表述与 RQ3 中观察到的行为一致：桥接会增加支持度，而强化学习主要将支持度转化为更高的 Acc@1。

实例化。我们基于 OctoMed-7B 实践这一方案。该模型已通过大规模医学监督微调打下了坚实基础。随后，我们使用从 PMC-VQA [30] 训练划分中抽取的 8,000 道小规模、平衡的选择题进行强化学习后训练。为平衡不同模态，我们首先使用 Qwen2.5-VL-32B-Instruct 将每道题分类为常见且临床认可的成像类别（例如磁共振成像、计算机断层扫描、超声、X射线、显微镜检查、皮肤科、光学相干断层扫描）；如果模态不常见、含糊不清或未明确说明，则标记为 none。然后，我们根据这些预测构建模态平衡的子集。最后，我们采用具备一致性感知变体的 GRPO 风格强化学习，在数据有限的情况下稳定训练。

医学视觉问答评估。我们在六个多模态医学基准上评估经过强化学习后训练的模型，并将其与强大的公开基线进行比较（表 3）。结果符合我们的分析所预期的模式：从一个已经弥合差距的检查点出发，能够让强化学习提升准确率，同时避免在支持较弱的基础模型上应用强化学习时观察到的失效模式。总体而言，在相同的解码设置下，我们的模型在所列基于 Qwen2.5-VL 的医学视觉语言模型基线中取得了最强的平均表现。

5 结论

我们通过区分视觉、医学监督微调和强化学习各自的作用，考察了强化学习后训练在何种情况下有助于医学视觉语言模型。在 MedMNIST 上，我们观察到明显的支持边界：Accuracy@1 可能远低于 Pass@K，这表明贪心解码未能实现潜在能力。医学监督微调能够稳定地提升感知能力和支持度，而强化学习主要使输出分布更加集中；当基线支持度已达到一定水平时，它能提升 Accuracy@1，但在分布偏移较大时收益有限。

这些发现促使我们采用一种边界感知方案：通过 Pass@K 诊断支持度，在支持度不足的情况下进行有针对性的微调以弥补短板，然后应用强化学习来提高采样效率。

参考文献

1. Baghbanzadeh, N., Fallahpour, A., Parhizkar, Y., Ogidi, F., Roy, S., Ashkezari, S., Khazaie, V.R., Colacci, M., Etemad, A., Afkanpour, A., 等：通过高质量数据推进医学表征学习。载于：医学图像计算与计算机辅助干预国际会议。第24–33页。施普林格 (2025) 2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., 等：Qwen2.5-VL 技术报告。arXiv 预印本 arXiv:2502.13923 (2025) 3. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Cai, Z., Ji, K., Wan, X., 等：探索大规模向多模态大语言模型注入医学视觉知识。载于：2024年自然语言处理实证方法会议论文集。第7346–7370页 (2024) 4. Chen, Y., Ge, Y., Wang, R., Ge, Y., Cheng, J., Shan, Y., Liu, X.: GRPO-Care：面向多模态推理的一致性感知强化学习。arXiv 预印本 arXiv:2506.16141 (2025) 5. Dai, W., Chen, P., Ekbote, C., Liang, P.P.: QoQ-Med：通过领域感知型GRPO训练构建多模态临床基础模型。arXiv 预印本 arXiv:2506.00711 (2025) 6. Fan, Z., Liang, C., Wu, C., Zhang, Y., Wang, Y., Xie, W.: ChestX-Reasoner：通过逐步验证和推理推进放射学基础模型。arXiv 预印本 arXiv:2504.20930 (2025) 7. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: PathVQA：用于医学视觉问答的 $30000+$ 问题。arXiv 预印本 arXiv:2003.10286 (2020) 8. Huang, X., Wu, J., Liu, H., Tang, X., Zhou, Y.: MedVLThinker：多模态医学推理的简单基线。arXiv 预印本 arXiv:2508.02669 (2025) 9. Jeddi, A., Karaimer, H.C., Nguyen, H., Wang, Z., Zhao, K., Rajabi, J., Zhang, R., Goyal, R., Taati, B., Grzeszczuk, R.: 面向视觉中心推理的拼图课程GRPO。arXiv 预印本 arXiv:2512.14944 (2025) 10. Karan, A., Du, Y.: 通过采样进行推理：你的基础模型比你想象的更聪明。arXiv 预印本 arXiv:2510.14901 (2025) 11. Lai, Y., Zhong, J., Li, M., Zhao, S., Li, Y., Psounis, K., Yang, X.: Med-R1：面向视觉语言模型中可泛化医学推理的强化学习。IEEE医学影像汇刊 (2026) 12. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: 关于放射影像的临床生成视觉问答数据集。Scientific Data 5(1), 180251 (2018) 13. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: SLAKE：用于医学视觉问答的语义标注知识增强数据集。载于：2021年电气电子工程师学会第18届生物医学成像国际研讨会 (ISBI)。第1650–1654页。电气电子工程师学会 (2021) 14. Liu, C., Wang, H., Pan, J., Wan, Z., Dai, Y., Lin, F., Bai, W., Rueckert, D., Arcucci, R.: 超越蒸馏：借助极简基于规则的强化学习，突破医学大语言模型推理的极限。arXiv 预印本 arXiv:2505.17952 (2025) 15. Manzari, O.N., Asgariandehkordi, H., Koleilat, T., Xiao, Y., Rivaz, H.: 采用集成KAN的变换器和扩张邻域注意力进行医学图像分类。Applied Soft Computing，第114045页 (2025) 16. Myronenko, A., Yang, D., Turkbey, B., Aboian, M., Azamat, S., Akcicek, E., Yin, H., Molchanov, P., Edgar, M., He, Y., 等：用于胸部X射线分析的推理视觉语言模型。arXiv 预印本 arXiv:2510.23968 (2025)

17. Ossowski, T., Zhang, S., Liu, Q., Qin, G., Tan, R., Naumann, T., Hu, J., Poon, H.：OctoMed：最先进多模态医学推理的数据配方。arXiv 预印本 arXiv:2511.23269（2025） 18. Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Ouyang, C., Rueckert, D.：Medvlm-r1：通过强化学习激励视觉语言模型（VLMs）的医学推理能力。载于：医学图像计算与计算机辅助干预国际会议。第337–347页。施普林格（2025） 19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 等：从自然语言监督中学习可迁移视觉模型。载于：国际机器学习会议。第8748–8763页。PmLR（2021） 20. Wang, W., Ma, Z., Ding, M., Zheng, S., Liu, S., Liu, J., Ji, J., Chen, W., Li, X., Shen, L., 等：大语言模型时代的医学推理：增强技术与应用的系统综述。arXiv 预印本 arXiv:2508.00669（2025） 21. Xia, P., Wang, J., Peng, Y., Zeng, K., Wu, X., Tang, X., Zhu, H., Li, Y., Liu, S., Lu, Y., 等：Mmedagent-rl：优化多模态医学推理中的多智能体协作。arXiv 预印本 arXiv:2506.00555（2025） 22. Xu, H., Nie, Y., Wang, H., Chen, Y., Li, W., Ning, J., Liu, L., Wang, H., Zhu, L., Liu, J., 等：Medground-r1：通过空间语义奖励的群组相对策略优化推进医学图像定位。载于：医学图像计算与计算机辅助干预国际会议。第391–401页。施普林格（2025） 23. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., 等：Lingshu：用于统一多模态医学理解与推理的通用基础模型。arXiv 预印本 arXiv:2506.07044（2025） 24. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.：Medmnist v2——用于二维和三维生物医学图像分类的大规模轻量级基准。Scientific Data **10**（1），41（2023） 25. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., 等：MMMU：面向专家级通用人工智能的大规模多学科多模态理解与推理基准。载于：IEEE/CVF计算机视觉与模式识别会议论文集。第9556–9567页（2024） 26. Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Song, S., Huang, G.：强化学习真的能在基础模型之外激励大语言模型的推理能力吗？arXiv 预印本 arXiv:2504.13837（2025） 27. Zhang, C., Neubig, G., Yue, X.：预训练、中期训练与强化学习在推理语言模型中的相互作用。arXiv 预印本 arXiv:2512.07783（2025） 28. Zhang, S., Liu, Q., Qin, G., Naumann, T., Poon, H.：Med-rlvr：通过强化学习从3B基础模型中涌现医学推理能力。arXiv 预印本 arXiv:2502.19655（2025） 29. Zhang, W., Zhang, P., Guo, J., Cheng, T., Chen, J., Zhang, S., Zhang, Z., Yi, Y., Bu, H.：Patho-r1：基于多模态强化学习的病理学专家推理模型。arXiv 预印本 arXiv:2505.11404（2025） 30. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.：PMC-VQA：用于医学视觉问答的视觉指令微调。arXiv 预印本 arXiv:2305.10415（2023） 31. Zuo, Y., Qu, S., Li, Y., Chen, Z., Zhu, X., Hua, E., Zhang, K., Ding, N., Zhou, B.：Medxpertqa：专家级医学推理与理解基准。arXiv 预印本 arXiv:2501.18362（2025）