arXiv:2603.01625v1 [cs.CL] 2 Mar 2026

衡量视觉语言模型（VLM）未言之事：验证指标掩盖了放射学报告生成中的临床术语消除

Aditya Parikh、Aasa Feragen、Sneha Das 和 Stella Frank

丹麦技术大学 DTU Compute

摘要。 要在放射学领域可靠地部署视觉语言模型（VLM），必须采用超越表层文本相似度的验证指标，以确保临床忠实度和人口统计公平性。本文探讨了当前模型评估中的一个关键盲点：采用能够获得较高总体词元重叠分数的解码策略，却因此陷入**模板坍塌**，模型只生成重复、安全且泛泛的文本，并省略临床术语。若不解决这一盲点，便可能导致**指标投机**：模型在基准测试中表现出色，临床上却毫无参考价值。因此，我们主张采用词汇多样性度量，检验模型生成内容是否具有临床特异性。我们提出临床关联偏移（CAD），这是一种词汇层面框架，用于量化生成的报告中基于人口统计特征的词语关联变化。加权关联擦除（WAE）汇总这些变化，以衡量不同人口统计学群体中的临床信号损失。我们发现，确定性解码会造成严重的语义擦除，而随机采样虽能生成多样化的输出，却可能引入新的偏见；这促使我们从根本上重新思考如何定义“最优”报告。代码见 [1, 2]

关键词： 放射学报告生成 · 验证 · 指标 · 公平性 · 偏见 · 语义擦除 · 视觉语言模型 · ReXGradient。

1 引言与相关工作

放射学报告生成（RRG）旨在自动将医学影像转换为可用于临床决策的文本，减轻文档记录负担，并支持诊断决策 [7, 23]。尽管近期的视觉语言模型让放射学报告生成在临床工作流程中的应用更具可行性 [18],，但本文表明，在标准文本相似度指标上取得高分，与提供更优质的临床报告或为不同患者群体带来更公平的结果之间存在重大差距。

易被博弈的指标。 当代放射学报告生成评估依赖于词重叠指标，如BLEU [20], ROUGE [15],BERTScore [28], ，用于衡量生成报告与参考（真实）报告之间的总体一致性。依据此类指标评估的模型容易出现**模板坍塌**：生成高频“正常所见”模板，以最大化词元重叠，却忽略患者特异性细节 [26]。由于正常X线片占主导

数据集，此类泛化输出可以获得具有欺骗性的高分，但诊断价值较低 [9]。标准指标**无法检测术语再分布**，即群体特异性临床术语被抑制、过度生成，或在不同人口统计学群体之间被重新分配。随着视觉语言模型已记录的倾向——即较少利用视觉证据而偏向语言先验 [12],——该风险会增加，这可能导致更泛化的输出并再现人口统计学偏见。当前评估方案侧重于总体准确性，从而使此类偏见未被发现。

解码作为混淆变量。推理策略（确定性/随机解码）直接影响生成报告的多样性和特异性 [22]。确定性解码优化模式选择，最终收敛于一小组高概率模板；随机解码则引入受控随机性，以探索概率较低的词元。既有研究 [14] 基于 BLEU 和 ROUGE 得分的边际提升，推动了随机解码的应用。然而，仅依据平均指标的改进来选择解码策略，会掩盖关键的公平性权衡：随机性虽然可能减少模板坍缩，但也可能同时引入或放大参考报告中不存在的人口统计学关联 [22]。现有评估框架无法审查这些词汇层面的变化。

对互补指标的需求。近期研究呼吁采用超越表层相似度的评估方法，以衡量临床效用、人口群体公平性和语义忠实度 [8]。与临床医生判断相一致的评分（例如 RaTEScore [30]）提供了一种途径，但无法直接量化模型如何在人口统计学群体之间重新分配临床术语。自然语言处理中的公平性指标通常依赖群体间 F1 差距或反事实扰动测试。当前缺失的是一种**词汇层面的诊断方法**，能够明确追踪临床语言的人口统计学关联结构在生成过程中是否得到保留、被抹除或遭到扭曲。

我们的贡献。 我们指出，解码策略是放射学报告生成（RRG）临床评估中的一个隐藏混杂因素，因此将推理配置视为首要实验因素：（i）我们提出**临床关联偏移**，这是一种词汇层面框架，用于量化人类撰写的报告与模型生成的报告之间群体关联词汇用法的变化，并识别被消除的术语以及新出现和保留下来的偏见；（ii）我们提出**加权关联抹除**，这是一项汇总统计量，用于衡量关联偏移的整体幅度，诊断模型是否系统性地回避具有临床信息价值的语言；（iii）我们以**我们的放射学视觉语言模型 ReX-QWEN**为测试平台，开展推理敏感性分析，表明确定性解码会通过模板坍塌 （高消除率）抬高分数，而随机解码虽能恢复词汇，却可能加剧偏见；（iv）我们的基线比较揭示了传统指标无法察觉的公平性问题。

我们的工作证实，解码策略与指标的选择会共同混淆临床评估。通过审查模型说了什么以及系统性地没有说什么，我们能够透明地评估放射学报告生成的进步究竟反映了真正的临床效用，还是评估伪象。

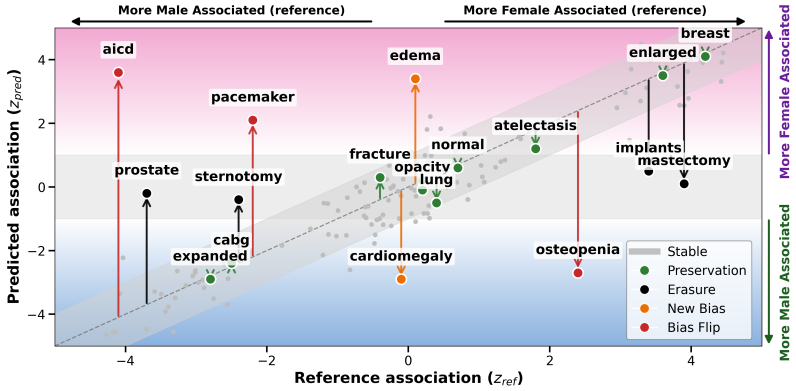

图1：**CAD 的临床解读**。每个点代表一个词汇项，展示其人口统计学关联从参考语料库（z_{ref}，横轴）到预测（z_{pred}，纵轴）的变化。靠近对角线（灰色带）的点表示关联稳定，而偏离对角线的点则突出显示人口统计学关联发生变化的词汇项，反映出消除、新偏见或偏差反转。

2 方法

我们提出一种全局词汇偏移的语料库级诊断方法，用以量化词汇在生成过程中如何得到保留、被消除或新近引入。这类偏移可能涉及具有临床意义、且与不同人口统计学群体存在差异性关联的术语，因此与诊断解读相关。为此，我们还提出特定人口统计群体偏移的诊断方法，揭示汇总指标未能捕捉的词汇用法偏移。我们将其形式化为：临床关联偏移（CAD）（第 2.1 节）和加权关联擦除（WAE）（第 2.2 节）。

2.1 临床关联偏移（CAD）

CAD 是一种词汇级诊断框架，用于量化单个临床术语在人口统计学关联的极性和幅度方面，如何在人工参考语料库与生成输出之间发生变化（见图1）。

问题形式化与对数比：令 $\mathcal{D} \in \{\text{ref},\text{pred}\}$ 分别表示参考语料库和预测语料库，其中每份报告都与一个二元群体相关联，此处为性别：$g \in \{F, M\}$。对于每个语料库 $\mathcal{D}$，我们针对词汇表 $\mathcal{V}$中的每个词 w 计算词元计数 $c_g^{(\mathcal{D})}(w)$，统计该词 w 在性别为 g的报告中的出现次数。每个群体的词元总数为 $N_g^{(\mathcal{D})} = \sum_{w \in \mathcal{V}} c_g^{(\mathcal{D})}(w)$。

为了衡量某个词 w 在语料库 $\mathcal{D}$中与女性报告相比和男性报告的关联强度，我们采用经过狄利克雷平滑的一元概率 [17]对数比。设 $\alpha > 0$ （默认值：$\alpha = 0.1$）为平滑参数（其中 $\sim$ 表示经过 α 平滑的估计值），$V = |\mathcal{V}|$ 为词汇量。我们定义：

$$\forall g \in \{F, M\}: \quad p_g^{(\mathcal{D})}(w) = \frac{c_g^{(\mathcal{D})}(w)+\alpha}{N_g^{(\mathcal{D})}+\alpha V} = \frac{\tilde{c}_g^{(\mathcal{D})}(w)}{\sum_{w' \in \mathcal{V}} \tilde{c}_g^{(\mathcal{D})}(w')}, \quad s^{(\mathcal{D})}(w) = \log \frac{p_F^{(\mathcal{D})}(w)}{p_M^{(\mathcal{D})}(w)}.$$

$s^{(\mathcal{D})}(w) > 0$ 表示与 w存在女性关联，而 $s^{(\mathcal{D})}(w) < 0$ 则表示 w 与男性关联更强。我们采用多项式模型下的 Delta 方法近似来估计 $s^{(\mathcal{D})}(w)$ 的方差：

$$\tilde{\mathrm{Var}}[s^{(\mathcal{D})}(w)] \approx \frac{1}{\tilde{c}_F^{(\mathcal{D})}(w)} + \frac{1}{\tilde{c}_M^{(\mathcal{D})}(w)}, \quad z^{(\mathcal{D})}(w) = \frac{s^{(\mathcal{D})}(w)}{\sqrt{\tilde{\mathrm{Var}}[s^{(\mathcal{D})}(w)]}},$$

得到经频率方差调整后的标准化关联 $z^{(\mathcal{D})}(w)$。我们将参考中的关联记为 $s^{(\mathrm{ref})}(w)$，将预测中的关联记为 $s^{(\mathrm{pred})}(w)$ 和 $z^{(\mathrm{ref})}(w)$， $z^{(\mathrm{pred})}(w)$。

量化偏移：我们将**关联偏移** $z_{\mathrm{disp}}(w)$定义为预测关联与参考关联之间的标准化差异，并将显著性检验的临界阈值记为 z_{crit}：

$$z_{\mathrm{disp}}(w) = \frac{s^{(\mathrm{pred})}(w) - s^{(\mathrm{ref})}(w)}{\sqrt{\tilde{\mathrm{Var}}[s^{(\mathrm{pred})}(w)] + \tilde{\mathrm{Var}}[s^{(\mathrm{ref})}(w)]}}, \quad z_{\mathrm{crit}} = \Phi^{-1}\left(1 - \frac{p^\star}{2}\right).$$

将语料库视为相互独立，则在零假设下， $z_{\mathrm{disp}}(w) \sim \mathcal{N}(0, 1)$。我们将 $z_{\mathrm{disp}}(w)$ 转换为双侧 p 值，采用 Benjamini–Hochberg FDR 校正 [4], ，并将满足 $|z_{\mathrm{disp}}(w)| > z_{\mathrm{crit}}$ 的词标记出来，显著性水平为 $p^\star$ （默认值：$p^\star = 0.05$），其中 Φ^{-1} 是标准正态分位数函数。

临床偏移类别：对于每个词汇项，我们计算 $z^{(\mathrm{ref})}$、$z^{(\mathrm{pred})}$ 和 z_{disp}。具有 $|z_{\mathrm{disp}}| < |z_{\mathrm{crit}}|$ 的词被视为**稳定**（无偏移），并排除在后续审查之外。偏移词分为三类： **（i）消除：**某个词在参考中具有较强关联（$|z^{(\mathrm{ref})}| > z_{\mathrm{strong}}$），但在预测中关联减弱、变为中性或未被生成（$|z^{(\mathrm{pred})}| < z_{\mathrm{neutral}}$），这表明临床信号有所丢失。**（ii）涌现偏差：**模型引入了参考中不存在的群体关联（**新偏差**： $|z^{(\mathrm{ref})}| < z_{\mathrm{neutral}}$，$|z^{(\mathrm{pred})}| \geq z_{\mathrm{neutral}}$），或使极性反转（**偏差反转**：$\mathrm{sign}(z^{(\mathrm{ref})}) \neq \mathrm{sign}(z^{(\mathrm{pred})})$，且两者均非中性），这表明模型产生了幻觉式人口统计学相关性。 **（iii）保留：**极性保持不变（$\mathrm{sign}(z^{(\mathrm{ref})}) = \mathrm{sign}(z^{(\mathrm{pred})})$，且人口统计学关联未丢失。

我们设定 $z_{\mathrm{neutral}} = 1$ & $z_{\mathrm{strong}} = 2$ （对应双尾正态分布截断值 $p \approx 0.32$ & $p \approx 0.05$），因此只有在统计上可靠的偏离中性状态的情况，才被视为具有临床意义的关联。

2.2 加权关联消除（WAE）

CAD 跟踪单个词语的偏移，而 WAE 则汇总这些词项级偏移，通过临床相关性加权，提供一个全局指标，衡量整个词汇中的临床信号损失。

$$\mathrm{WAE} = \frac{\sum_{w \in \mathcal{V}} \omega(w) \cdot z_{\mathrm{disp}}(w)^2}{\sum_{w \in \mathcal{V}} \omega(w)}$$

其中，分子表示按词项重要性（以词频近似）缩放的聚合位移能量 。分母是一个归一化因子 ，使该指标不受预测词元总数的影响。

我们定义权重 $\omega(w) = c^{(\mathrm{pred})}(w)$（基于预测）或 $\omega(w) = c^{(\mathrm{ref})}(w)$（基于参考）。基于预测的加权量化模型实际生成的文本内部的变化，从而避免因遗漏词汇而受到惩罚。基于参考的加权揭示了哪些临床重要术语（在人类报告中频繁出现）被系统性地回避。对参考中与群体 g 相关的术语进行汇总，可得群体特定分数 WAE$_g$；差异定义为 ΔWAE（例如， $\Delta\mathrm{WAE} = \Delta\mathrm{WAE}_F - \Delta\mathrm{WAE}_M$）。

不确定性通过非参数自助法估计；统计校准通过标签置换检验确认，在随机人口统计学分配下，观察到的 WAE 始终高于零分布。

3 实验

数据集： 我们使用**ReXGradient-160K**，这是目前公开可用的最大多中心胸部 X 线数据集 [29], ，其中包含临床医生撰写的所见和印象。我们使用基础模型 [16] 筛选正位视图，并从中分层抽取，000 例样本，确保性别代表性均衡。

模型：ReX-QWEN：我们使用 LoRA [11]（秩 $r = 16$， $\alpha = 32$）对 Qwen3-VL-8B-Instruct [25]进行微调，训练数据为 ReX- Gradient，并将其应用于注意力层和多层感知机层。输入图像会调整大小为 512×512 RGB。我们的主模型 **ReX-BB**（平衡基线）使用 6 万个性别平衡样本（男性占 50%，女性占 50%）进行训练。两个性别比例偏斜的变体：ReX-MB（男性占 67%）和 ReX-FB（女性占 67%），用于验证 CAD/WAE 对人口统计偏移的敏感性。我们在相同的评估方案下，将 ReX-QWEN 与三个模型进行比较: **CXR-LLaVA** [13],**LLaVA-Rad** [27]（用于放射学报告生成的医学视觉语言模型）；以及 **OpenAI GPT-4**。**1** [19],采用零样本提示进行报告生成。

推理策略研究：我们假设，语义擦除源于推理策略系统性地抑制临床术语，以尽量降低生成风险。我们将推理策略分为**确定性**（贪心搜索、束搜索）和**随机性** （温度 T、top-k、核采样）：前者选择概率最高的序列，可能导致模板坍塌；后者则从截断分布中采样，以提高词汇多样性。

具体而言，我们评估了：(i) 贪心搜索（argmax）、(ii) 束搜索（$k = 4$）、(iii) 保守采样（$T = 0.3$，top-$p = 0.95$）、(iv) 丰富采样（T =0。7，top-$p = 0.95$）、(v) Top-k （$k = 50$， $T = 1.0$），以及 (vi) 核采样（top-$p = 0.9$， $T = 1.0$）。所有方法均使用最大词元数=512 并启用缓存。

评估指标：我们从三个维度评估模型：**临床准确性：**我们使用 CheXbert 作为骨干模型，报告 BERTScore（F1、精确率、召回率）

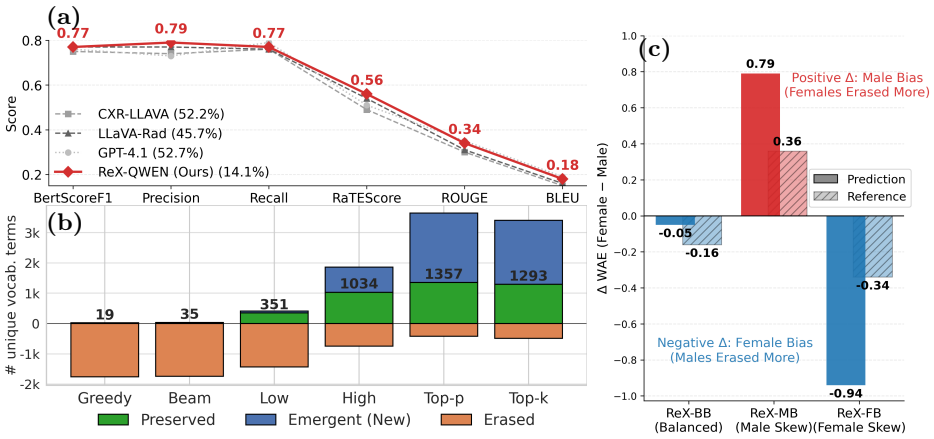

图2：**(a) 临床表现比较**（含模板多样性%）。**(b) 不同推理策略下的全局词汇偏移**(c) ΔWAE 成功捕捉了性别偏斜训练造成的语义偏移。

以及 ROUGE-1/L 和 RateScore，以衡量与参考报告在语义和表层上的相似度。**文本多样性：**按照 [10], ，我们计算：(i) 模板多样性（报告的唯一性比例）、(ii) 类型-词元比（TTR）、(iii) 1-Self-BLEU，以及 (iv) 语义多样性。**我们的指标：**我们报告：(i) 我们提出的 CAD 标记的临床类别（抹除、新偏差、偏差反转），用于量化词汇层面的关联偏移；(ii) 我们提出的 WAE，用于汇总全局位移幅度；(iii) 现有的 WEAT [5, 3] 及其效应量（ES），作为对比性的词语关联基线；以及 (iv) 差异影响比（ΔDIR），用于衡量参考文本与预测文本中男性与女性词元比例的变化。计算所有基于关联的指标前，都会移除停用词，以确保重点关注临床术语。

4 结果与讨论

我们发现，尽管 ReX-QWEN 达到了具有竞争力的基线性能（见图 2a），严重的模板坍塌依然显而易见（模板多样性为 14.1–52.7%）。这一坍塌促使我们在后续进行词汇分析，具体考察解码选择如何驱动语义擦除与涌现偏差之间的根本权衡。为评估这些动态变化，我们比较了 ReX-BB 的六种不同推理策略（表 1）。

全局词汇坍塌。 确定性策略表现出严重的词汇流失，生成的词汇表仅包含 22–38 个独特术语，占原始参考词汇的 5%（见图 2b），并转而坍缩为“正常所见”模板。随机方法保留了词汇表规模，却引入了参考中不存在的涌现术语，从而引发有效性疑虑：这些是恢复出的临床信号，还是幻觉噪声？为了确定这些变化是否会不成比例地影

表1：**推理策略敏感性**。（**粗体**：最佳；下划线：最差）。

指标类别	Ref.	确定性		随机性（熵递增）			
		贪心搜索	Beam	$T = 0.3$	$T = 0.7$	排名-k	排名-p
表层相似度							
BERTScore F1 ↑	–	**0.78**	**0.78**	0.76	0.73	0.71	0.69
BERTScore P/R ↑	–	0.80/0.75	**0.81**/**0.76**	0.78/0.75	0.73/0.74	0.68/0.74	0.70/0.73
ROUGE (1/L) ↑	–	0.34/0.29	**0.35**/**0.30**	0.32/0.26	0.26/0.19	0.20/0.14	0.21/0.15
RateScore ↑	–	**0.55**	**0.55**	0.54	0.48	0.44	0.44
文本多样性与坍缩							
模板多样性 (%) ↑	82.45	0.3	0.5	36.3	97.05	**100.0**	**100.0**
TTR (%) ↑	2.915	0.006	0.131	0.654	2.128	2.540	**2.781**
1 - Self-BLEU（%） ↑	35.70	1.05	0.20	10.78	33.78	46.73	**46.84**
语义多样性（%） ↑	85.23	12.16	5.41	55.51	83.75	85.80	**86.19**
临床类别（我们的方法）							
消除（#） ↓	–	32	35	17	4	5	**3**
新偏差（#） ↓	–	**0**	3	41	45	25	65
偏差反转（#） ↓	–	**0**	1	16	33	44	69
公平性							
WAE（预测）（我们的方法） ↓	–	**2.49**	2.50	4.95	33.81	12.90	13.27
WAE（参考）（我们的方法） ↓	–	**0.19**	0.60	3.43	9.06	4.60	14.70
WEAT（S/ES） ↓	–	-1.46/-0.64	1.10/0.03	-1.35/-0.01	4.30/-0.07	-2.13/-0.02	0.27/0.00
ΔDIR ↓	–	-0.03	**0.01**	-0.08	-0.01	-0.20	-0.32

响特定患者群体，我们转而使用 CAD，审查与性别相关术语中的语义擦除和涌现偏差。

确定性解码通过坍缩抬高相似度。贪心搜索和束搜索生成的报告中，独特报告仅占 0.3–0.5%，呈现出模板坍塌。尽管 BERTScore F1=0.78 表现良好，CAD 仍揭示了严重的性别关联抹除：有 32–35 个术语被抹除，这几乎涵盖了参考文本中所有偏见程度较高的术语。例如，自动植入式心律转复除颤器 （与男性相关）和乳房切除术（与女性高度相关）的预测关联均转为中性。这种钻指标空子的伪象证实，高总体得分可能掩盖系统性的临床信息抹除。

随机解码恢复了词汇多样性，却引入了偏见。温度缩放提高了多样性（36.3–100%）并减少了信息抹除，但也引入了 25–65 个新的偏见术语，并因新偏见术语的出现而推高了 WAE^{pred} （2.50 → 33.81）。 ΔDIR 差距扩大，表明不同性别群体之间存在冗长程度失衡。例如，病理术语气胸 在人工参考文本中是中性的（$z_{ref} = 0.15$）；然而，低温生成却主动捏造了该术语与男性患者之间的强关联（$z_{pred} = -5.69$）。WEAT 评估的是静态潜在嵌入，因此完全无法识别解码诱发的偏见。

启示。 CAD 识别出两种对立的失效模式： **（i）语义抹除**发生于确定性解码时：系统性词汇抑制无法通过 BERTScore/ROUGE 检测； **（ii）涌现偏差**发生于高温采样时：

表2：**外部验证证实，使用不同解码策略的 LLaVA-Rad 仍存在消除与偏见之间的权衡**文本相似度指标（BERTScore、ROUGE 等）相近（$\Delta \leq 2\%$，未显示），而多样性指标则显示模板使用（Div. Temp.）和擦除率存在明显差异。（NB：新偏差；BF：偏差反转）

指标	ReXGradient		CheXpert	
	贪心搜索	Rich	贪心搜索	Rich
划分。（温度/语义%） ↑	45.70/70.32	91.05/79.49	94.05/80.60	100/83.13
消除/NB/BF ↓	19/6/5	11/5/11	6/1/0	1/1/4
WAE（预测/参考） ↓	1.17/0.84	1.03/0.93	0.76/0.60	0.84/0.69
WEAT（S/ES） ↓	1.13/0.01	-0.36/-0.00	-0.78/0.11	-0.14/0.17

出现参考中不存在的新人口统计学关联。没有任何单一策略能够同时优化准确性、多样性和公平性，因此需要选择公平性感知的解码策略，并以临床准确性和CAD/WAE 诊断指标为联合调优依据，而非默认采用贪心推理。**我们建议**在标准评估中加入 CAD 类别、WAE 和词汇多样性指标，以免将模板坍塌伪影误认为临床效果确有改善。

泛化能力： LLaVA-Rad 在 ReXGradient 和 CheXpert [6] 上的结果证实了这一模式（表2）：确定性解码降低了模板多样性（45.7% 对比 91.1%），同时消除程度更高（19 个术语）；而随机采样则会引入偏差反转（5→12）。WAE 略有下降（在 ReXGradient 上为 1.17→1.03 ），证实这种权衡在不同架构和数据集上依然存在。

稳健性： 性别偏斜训练（ReX-MB、ReX-FB）使传统指标保持稳定（BERTScore F1$\in [0.76, 0.78]$、RateScore$\in [0.54, 0.57]$、WEAT-ES≈ 0）。然而，ΔWAE$_{\text{pred}}$ 从接近零（ReX-BB：-0.05）变为 ReX-MB 的 $+0.79$ 和 ReX-FB 的 -0.94 ，证实 WAE 能捕捉现有指标无法察觉的语义偏移（参见图 2c ）。

消融实验： 将 $p^{\star}_{\text{disp}}$ 扫过其整个范围，严格保持了所有解码策略的相对排名（参见图 3）。这项消融实验证实，擦除与偏差之间的权衡是一种根本性的生成机制，而非指标阈值设定造成的假象。

5 结论与未来工作

未来工作： 我们的分析聚焦于二元性别关联；未来工作应将 CAD 扩展至年龄、族裔和交叉性群体。还需由临床医生进行验证，以确认 CAD 标记的擦除是否对应具有临床意义的遗漏。

结论： 我们证明，解码策略与指标选择共同混淆了放射学报告生成的评估。通过引入 CAD 和 WAE，我们提供了一个框架，用于审查传统指标完全无法揭示的语义擦除和涌现偏差。这些工具共同实现

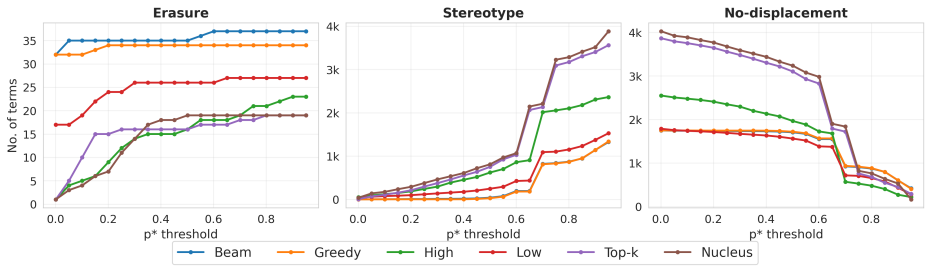

图 3：**消融实验**：扫描 $p^{\star}_{\mathrm{disp}}$ 可保留模型的相对排名：在所有显著性水平下，CAD 都能检测到随机输出中更多的涌现刻板印象。

透明评估，使该领域能够区分真正的临床效用与由模板坍塌造成的性能假象。

参考资料

1. 匿名作者：录用后发布。https://github.com/ tobereleased/CAD-Framework（2026年）2. 匿名作者：录用后发布。https://huggingface.co/ tobereleased/ReX-QWEN（2026年）3. Badilla, P., Bravo-Marquez, F., Zambrano, M.J., Pérez, J.：Wefe：一个用于测量和缓解词嵌入中偏见的 Python 库。《机器学习研究杂志》**26**(156), 1–6（2025年），http://jmlr.org/papers/v26/22-1133.html 4. Benjamini, Y., Hochberg, Y.：控制假发现率：一种实用且有效的多重检验方法。《英国皇家统计学会杂志：B辑（方法论）》**57**(1), 289–300（1995）5. Caliskan, A., Bryson, J.J., Narayanan, A.：从语言语料库中自动提取的语义包含类人偏见。《科学》**356**(6334), 183–186（2017）6. Chambon, P., Delbrouck, J.B., Sounack, T., Huang, S.C., Chen, Z., Varma, M., Truong, S.Q., Chuong, C.T., Langlotz, C.P.：CheXpert Plus：通过文本放射学报告、患者人口统计信息和其他图像格式扩充大型胸部X线数据集。arXiv 预印本 arXiv:2405.19538（2024年）7. Demner-Fushman, D., Chapman, W.W., McDonald, C.J.：自然语言处理能为临床决策支持做些什么？《生物医学信息学杂志》**42**(5), 760–772（2009）8. Du, X., Zhou, Z., Wang, Y., Chuang, Y.W., Li, Y., Yang, R., Zhang, W., Wang, X., Chen, X., Guan, H. 等：生成式大语言模型在电子健康记录应用中的测试与评估：系统综述。《美国医学信息学会杂志》p. ocaf233（2026年）9. Gao, L., Zhang, L., Liu, C., Wu, S.：处理不平衡医学图像数据：一种基于深度学习的单类分类方法。《医学人工智能》**108**, 101935（2020）10. Guo, Y., Shang, G., Vazirgiannis, M., Clavel, C.：语言多样性令人好奇的衰退：在合成文本上训练语言模型。载于：计算语言学协会 NAACL 2024 论文集。第3589–3604页（2024年）

10 Parikh 等人。

11. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. 等：LoRA：大语言模型的低秩适应。ICLR1(2), 3（2022）12. Lee, K.i., Kim, M., Yoon, S., Kim, M., Lee, D., Koh, H., Jung, K.：Vlind-bench：衡量大型视觉语言模型中的语言先验。载于：计算语言学协会 NAACL 2025 论文集。第4129–4144页（2025年）13. Lee, S., Youn, J., Kim, H., Kim, M., Yoon, S.H.：CXR-LLaVA：用于解读胸部X光图像的多模态大语言模型。《欧洲放射学》**35**(7), 4374–4386（2025年）14. Li, Q.：利用预训练视觉语言模型的强大能力，高效生成医学报告。载于：第32届美国计算机协会国际信息与知识管理会议论文集。第1308–1317页。CIKM '23，美国计算机协会，纽约，美国（2023年）。https://doi.org/10.1145/3583780.3614961, https://doi.org/10.1145/3583780. 3614961 15. Lin, C.Y.：ROUGE：用于摘要自动评价的软件包。载于：文本摘要的新进展。第74–81页（2004年）16. Ma, D., Pang, J., Gotway, M.B., Liang, J.：应用于胸部放射摄影的完全开放人工智能基础模型。《自然》**643**(8071), 488–498（2025年）17. Monroe, B.L., Colaresi, M.P., Quinn, K.M.：Fightin'words：用于识别政治冲突内容的词汇特征选择与评估。《政治分析》**16**(4), 372–403（2008）18. Nam, Y., Kim, D.Y., Kyung, S., Seo, J., Song, J.M., Kwon, J., Kim, J., Jo, W., Park, H., Sung, J. 等：医学影像中的多模态大语言模型：现状与未来方向。《韩国放射学杂志》**26**(10), 900（2025年）19. OpenAI：在 API 中推出 GPT-4.1。https://openai.com/index/gpt-4-1/（2025 24 20 2026 02），访问日期：- -。Papineni, K., Roukos, S., Ward, T., Zhu, W.J.：BLEU：一种用于机器翻译自动评价的方法。载于：第40届计算语言学协会年会论文集。第311–318页（2002年）21. Parikh, A., Das, S., Feragen, A.：研究医学分割中的标签偏差及年龄相关差异的表征来源。arXiv 预印本 arXiv:2511.00477（2025年）22. Presacan, O., Nik, A., Thambawita, V., Ionescu, B., Riegler, M.：医学文本生成解码策略的比较研究。载于：多媒体建模国际会议。第1–15页。Springer（2026年）23. Rocha, D.M., Brasil, L.M., Lamas, J.M., Luz, G.V., Bacelar, S.S.：结构化放射学报告的益处、优势与潜力的证据：一项整合性综述。《医学人工智能》**102**, 101770（2020）24. Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A.Y., Lungren, M.P.：CheXbert：结合自动标注器与专家标注，利用 BERT 实现准确的放射学报告标注（2020）25 09388 26 3 2025 2505 . Team, Q.：Qwen 技术报告（），https://arxiv.org/abs/ . . Yun, L., An, C., Wang, Z., Peng, L., Shang, J.：格式的代价：大语言模型中的多样性坍缩。arXiv 预印本 arXiv:2505.18949（2025年）27. Zambrano Chaves, J.M., Huang, S.C., Xu, Y., Xu, H., Usuyama, N., Zhang, S., Wang, F., Xie, Y., Khademi, M., Yang, Z. 等：一种临床易用的小型多模态放射学模型及用于胸部X光征象的评估指标。《自然-通讯》**16**(1), 3108（2025年）28. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.：BERTScore：使用 BERT 评估文本生成。arXiv 预印本 arXiv:1904.09675（2019年）

29. Zhang, X., Acosta, J.N., Miller, J., Huang, O., Rajpurkar, P.: ReXGradient-160K：一个大规模公开胸部X线片自由文本报告数据集。arXiv 预印本 arXiv:2505.00228（2025年） 30. Zhao, W., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: RateScore：一种用于放射学报告生成的指标。载于：2024年自然语言处理实证方法会议论文集。第15004–15019页（2024年）