arXiv:2611.32993v1 [cs.CV] 24 Nov 2026

通过解剖学定位与置信度校准实现决策原生多模态临床建模

Juntao Jiang*[1,2], Wenqi Zhong*[2,3,1], Guangyu Guo*[2,3,1], Dexia Chen2, Yilong Hu2, Yunhao Bai2,4, Zhenyu Yi2,5, Jianwei Xu2,3,1, Tony C. W. MOK2, Muxuan Gao1, Ling Zhang2, Yong Liu1, Jiangning Zhang†[1], Zhongwei Qiu†[2,3,1]

[1]Zejiang University [2]达摩院，阿里巴巴集团[3]湖畔实验室 [4]华东师范大学[5]上海交通大学

*共同贡献， †通讯作者

基于体数据计算机断层扫描（CT）进行结构化临床决策，需要具备解剖学定位的预测和经过校准的置信度估计。然而，现有的三维医学多模态模型通常受到未经校准的自回归解码以及依赖无关解剖区域的捷径学习影响。我们提出**Radev**，这是一种决策原生多模态模型，融合了解剖学引导的视觉压缩、直接候选预测和反事实置信度监督。为连接全局上下文与细粒度切片信息，冻结的三维编码器提取可复用的器官级特征，并由查询动态选择这些特征，以关注预测解剖边界内的二维切片特征。由此得到的紧凑视觉词元使语言主干网络能够在单次前向传播中估计候选项概率，无需进行自回归解码。我们进一步提出反事实校准训练（CCT），比较器官表征在完整状态和干预状态下的预测结果。当移除目标器官证据导致决策质量下降时，CCT促使置信度降低；而在移除无关器官时，则惩罚过度的置信度下降。为促进严格评估，我们构建了**CT-Verdict**，这是一个全面的基准测试，涵盖分布内和外部队列中的二值征象存在性、多项选择题和有序严重程度评估。大量实验表明，Radev在决策准确率和置信度校准方面均显著优于现有最先进方法，并且在分布偏移下仍具备稳健的泛化能力。项目主页：https://juntaojianggavin.github.io/projects/Radev/。

日期：2026年11月25日

1 引言

医学视觉语言模型（VLMs）将视觉输入与自然语言查询相结合，以支持多种临床任务（Jiang 等人，2025；Qiu 等人，2025；Xu 等人，2026b；Yuan 等人，2026年；Li等人，2026a）。在体数据计算机断层扫描（CT）中，临床查询通常针对特定器官或所见，以评估其是否存在、解剖位置、属性或严重程度。由于这些判断涉及重大的临床风险——漏诊发现会延误干预，而假阳性则会导致不必要的操作——临床医生需要可靠的不确定性估计，以标记模糊病例并交由人工审核（Geifman 和 El-Yaniv，2017；Angelopoulos 和 Bates，2023）。然而，过度自信的错误会使错误预测看起来可信，从而破坏这一安全保障。因此，要实现安全的临床部署，不仅需要较高的诊断准确率，还需要严格的置信度校准，确保模型报告的置信度能够真实反映其经验准确率。

如图所示，现有体积 CT 视觉语言模型主要遵循生成式范式1，采用三维或混合编码器，以自回归方式生成自由文本答案或放射学报告（Wu等人，2025；Xin等人，2025；Li等人，2026b；Yuan 等人，2026年）。尽管这些模型可以通过文本生成或候选项似然评分适配封闭集临床决策，但这种形式存在三个根本性局限。第一，对于结构化决策而言，自回归解码会带来不必要的计算开销，并且通常会产生校准不佳的词元级概率。第二，检测局部病灶需要在三维器官上下文中对细粒度的二维切片

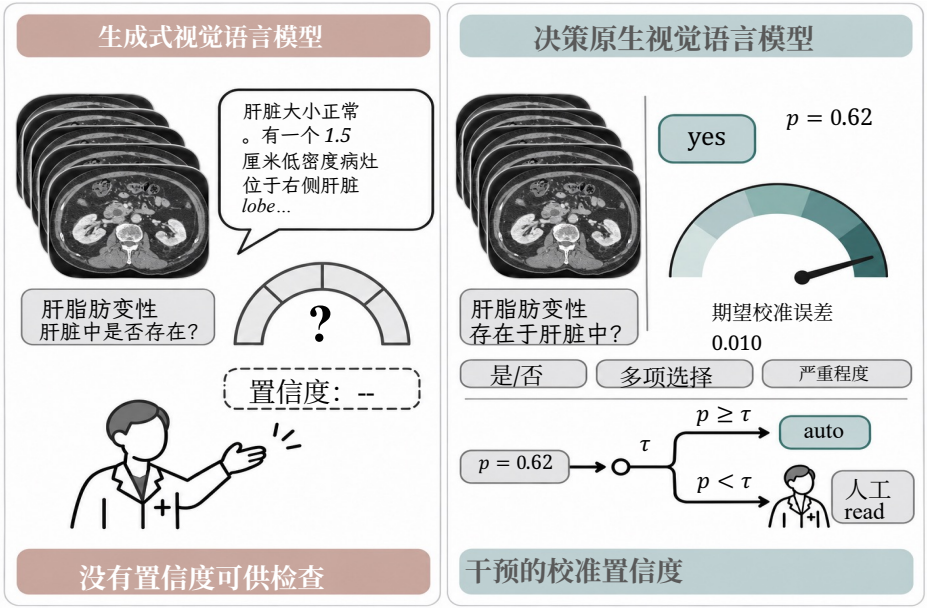


图1 生成式视觉语言模型与决策原生视觉语言模型的对比。 传统生成式视觉语言模型面向开放式推理和报告生成而设计，但往往依赖缺乏依据的视觉上下文，且置信度未经校准。**Radev** 是一种决策原生模型，专为结构化、以器官为依据且具备置信度意识的临床判断而设计。

细节进行定位，而相关的解剖学证据会随查询动态变化。第三，标准的答案级监督只奖励与真实标签的一致性，却不验证预测或其置信度是否真正基于目标解剖结构，而非源于虚假的背景相关性。在近期决策建模进展的基础上（Typesafe.ai，2025；Tang和Chen，2026年；Tommasino和Rinaldi，2026年），我们将结构化 CT 解读重新表述为一种决策过程，如图1 所示，其核心是直接的候选项概率估计、查询驱动的解剖学证据选择，以及反事实置信度监督。

为此，我们推出**Radev**，这是一种决策原生多模态模型，将基于解剖结构的视觉压缩与反事实置信度校准融为一体。具体而言，Radev 使用冻结的三维编码器提取可复用的器官级特征，并将其投射到持久化的器官账本中。给定临床查询后，文本引导的路由机制会选出目标器官，其词元会关注预测的三维解剖边界内的高分辨率二维切片特征。由此生成的紧凑视觉词元序列既保留了全局器官上下文，又针对查询涉及的解剖结构注入局部切片证据。将这些词元与问题及候选项结合后，语言主干网络便可通过一次前向传播估计候选项概率，无需进行自回归解码。在针对二分类、多项选择和序数任务完成多模态对齐与裁决微调后，我们提出反事实校准训练 （CCT）。通过对比模型在器官表征完整、目标器官被消融以及对照器官被消融时的预测，CCT 明确促使模型在移除目标器官证据会损害决策质量时降低置信度，同时惩罚模型对移除无关器官表现出无根据的敏感性。

为开展严谨的评估，我们构建了**CT-Verdict**，这是一个全面的结构化 CT 解读基准测试，涵盖二值征象存在性、多项选择的位置与属性查询，以及有序严重程度评估。CT-Verdict 包含分布内（IID）和外部测试队列，为评估不同任务类型及分布偏移下的准确率与校准建立了标准化测试平台。大量实验表明，Radev-2B 在 CT-Verdict-IID 上取得了 83.75% 的整体准确率，达到当前最佳水平，同时显著提升了置信度校准和分布外泛化能力。我们的主要贡献总结如下：

- **决策原生架构：** 我们提出 **Radev**，这是一种结构化 CT 解读模型，将可复用的三维器官账本与由查询选取的二维切片证据相结合，无需自回归解码即可通过单次处理估计候选项概率。

- **反事实置信度监督：** 我们设计了 反事实校准训练，这是一种基于干预的训练目标，通过对比完整、目标器官消融和对照器官消融的器官状态，使预测置信度与因果解剖学证据相对齐。

- **基准测试与实证验证：** 我们发布了**CT-Verdict**，这是一个包含内部和外部队列的多任务体积 CT 基准测试，并证明 Radev 在诊断准确率、置信度校准以及应对分布偏移的稳健性方面，始终优于现有生成式视觉语言模型。

2 相关工作

2.1 决策模型

通过候选答案直接进行预测，已在语言建模和视觉语言建模领域得到探索。BERT（Devlin 等人，2019）会为多项选择推理中的候选续写打分，而 LXMERT（Tan 和 Bansal，2019）则根据图像和问题，从预定义词汇表中预测答案。这些判别式形式为无需自由文本生成即可进行决策奠定了成熟基础。近期的原生决策语言模型通过一种接口重新审视这一原则：该接口接受任务描述和明确的候选项集合，并返回候选项概率。Jev（Typesafe.ai，2025）便是此类系统之一，它能通过一次前向传播，将类型化状态和离散候选项映射为决策分布。开放复现版本会对通用骨干网络进行微调，并采用决策头或指针读出层，例如 Kev（Palmer，2026年）、decider（Marosi，2026年）、OneJev（OmniJev 团队，2026年）、ReJev（Joe-rq，2026年）。这些模型保留了自回归骨干网络，但会对候选项进行打分。

近期，决策接口已扩展至视觉及其他模态。Intern-Decision（Zhu等人，2026年）将 Qwen3.5 骨干网络结合起来，在一次前向传播中返回文本状态上的答案分布，并可选地输入图像。OmniJev（Xu等人，2026a）和 Clef（Cloudflare，2026年）支持文本、图像、音频和视频。OmniMed-Jev（Tang 和 Chen，2026年）将视觉语言骨干网络与原生决策校准相结合，以回答涉及不同医学模态的问题。PathJev（Tommasino 和 Rinaldi， 2026年）在组织病理学图像块上训练决策头，并表明，使用专用头进行候选项评分，优于读取生成模型语言头输出的概率。然而，它们全部输入二维图像，且都无法处理三维体积影像检查，而这类检查在临床医学影像中无处不在且不可或缺。

2.2 医学视觉语言模型

视觉语言模型如今已广泛应用于医学影像，承担报告生成和视觉问答任务。医学图像的领域特殊性，使医学视觉语言模型形成了一系列独特的设计选择。一类模型保留了标准二维架构。LLaVA-Med（Li等人，2023）通过概念对齐和指令微调确立了这一范式。HealthGPT（Lin 等人， 2025）引入 H-LoRA 插件，以统一医学理解与生成。尽管取得了这些进展，它们的视觉归纳偏置仍局限于二维，缺乏体积监督。因此，CT 检查等体积图像会被简化为二维切片；Hulu-Med（Jiang等人， 2025）将这一折衷方案形式化，通过旋转位置自适应编码将切片作为图像块序列进行处理。为减少切片间冗余，CARVE（Yi等人， 2026年）将无需训练的词元压缩构建为预算约束的 2.5D 分配问题，以保留显著视觉证据。

第二类方法整合三维编码器，旨在解决两个关键瓶颈：体积令牌爆炸，以及缺乏网络规模的三维语料库。为控制词元复杂度，RadFM（Wu等人，2025）采用交错式二维/三维生成式预训练，M3D-LaMed（Bai 等人，2024）通过空间池化感知器实施固定令牌预算，而 Med3DVLM（Xin 等人，2025）则以高效的分解卷积取代密集三维注意力。为利用丰富的监督信号弥补数据匮乏，CT-CHAT（Hamamci，2026年）复用冻结的 CT-CLIP 编码器，而 Merlin（Blankemeier，2024）则在报告和电子健康记录编码上联合预训练三维 ResNet。混合模型则连接这两类方法，在保留二维先验的同时，以模块化方式注入三维线索：OmniCT（Lin 等人，2026b）和 TIF-GRPO（Lin 等人，2026a）将三维位置编码与掩码引导的器官令牌聚合相结合；ClinFusion（Yuan 等人，2026年）则通过局部交叉注意力融合三维体积与二维基础特征。尽管取得了这些进展，针对三维医学影像的决策视觉语言模型仍鲜有探索。

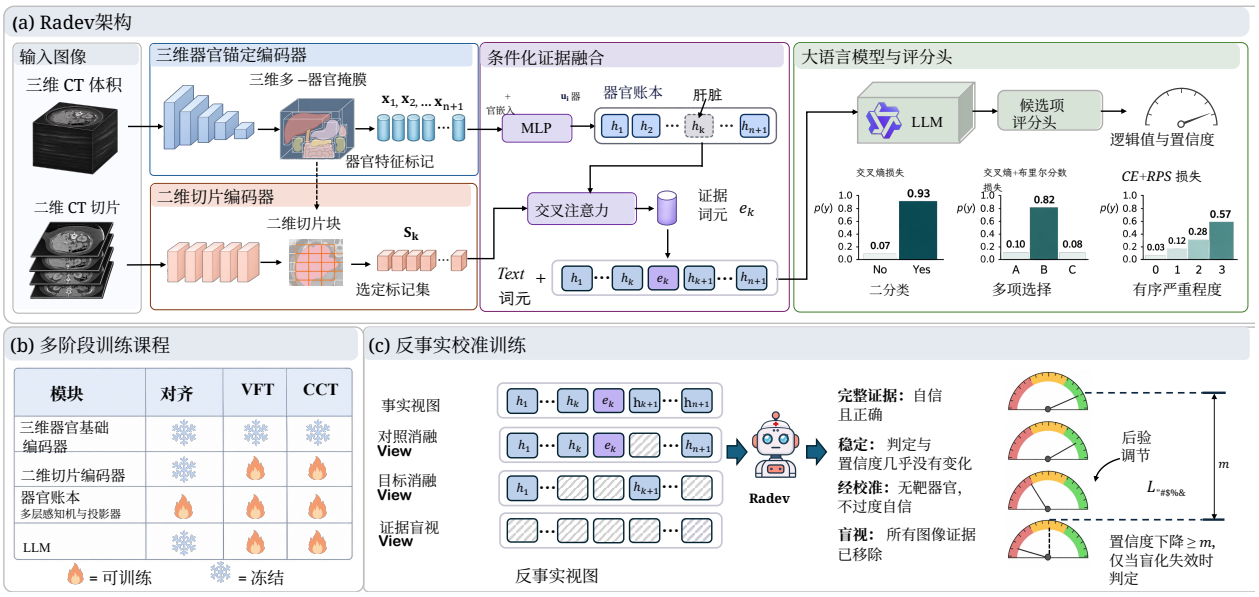


图 2 所提出的 Radev 框架。(a) 器官锚定视觉编码器利用器官掩膜和学习得到的掩膜查询提取三维器官token。问题选定的器官token会在其解剖边界内关注二维切片特征。语言主干网络和候选评分头通过一次前向传播预测答案分布。(b) 训练依次经过投影器-证据对齐、裁决微调和反事实校准训练（CCT）。(c) CCT 通过对视觉特征进行干预，构建事实视图、目标消融视图、对照消融视图和证据盲视图。前三种视图接受任务监督，而证据盲视图则接受置信度间隔惩罚。

3 方法

3.1 任务定义与概述

给定一个三维 CT 体积 V 和一个临床问题 q 及其候选答案 $\mathcal{C}(q)$，Radev 会预测候选项上的概率分布 $p_\theta(c \mid V,\ q)$。我们考察二值征象存在性、关于征象位置或属性的多项选择题，以及有序严重程度评估。每个问题都只有一个参考答案。

Radev 将器官锚定视觉编码器与 Qwen3.5 语言主干网络及候选评分头相结合（图2）。编码器利用器官掩膜引导三维器官词元读取，并通过区域受限注意力纳入与问题相关的二维切片证据。投影器-证据对齐和裁决微调（VFT）直接监督候选项概率。随后，反事实校准训练（CCT）结合完整表示和消融表示上的任务监督，并对证据盲视图施加置信约束。

3.2 器官对齐视觉编码器

掩膜引导器官标记化。 给定一个三维 CT 体积 V，冻结三维分支提供多尺度视觉特征token和器官掩膜 $\{\Omega_i\}_{i=1}^{n}$，对应于 n 个器官类别。每个掩膜 Ω_i 将空间支持范围限制为$\mathbf{T}_i$，即空间单元与 Ω_i重叠的多尺度三维特征token。器官关联掩膜token$\mathbf{m}_i$作为学习得到的读出查询，通过共享多头注意力$\mathbf{T}_i$ 、器官特定线性投影$\mathbf{W}_i$以及 ℓ_2 归一化进行聚合，从而得到 d维器官特征$\mathbf{x}_i$。

$$\mathbf{f}_i = \mathbf{W}_i\,\mathrm{MHA}(\mathbf{m}_i, \mathbf{T}_i, \mathbf{T}_i), \qquad \mathbf{x}_i = \frac{\mathbf{f}_i}{\|\mathbf{f}_i\|_2}. \tag{1}$$

单独的全局读出提取全卷特征 $\mathbf{x}_{n+1} \in \mathbb{R}^d$。未检测到、被视野截断或未完全包含在任何处理窗口中的器官，使用共享的学习特征占位符

代替 $\mathbf{x}_i$，再进行投影。共享多层感知机将所有 $n+1$ 特征映射到语言嵌入空间，而学习的槽位嵌入 $\mathbf{u}_i$ 则用于标识各个解剖学槽位或全局槽位，

$$\mathbf{h}_i = \mathrm{MLP}(\mathbf{x}_i) + \mathbf{u}_i, \qquad i = 1, \ldots, n+1. \tag{2}$$

由此得到的器官账本保留所有 $n+1$ 槽位，且顺序固定。它仅由 V构建，可在不同问题间重复使用；主干网络和窗宽窗位的详细信息见第5.1节。

问题条件证据融合。Qwen3.5 原生视觉编码器从均匀采样的轴位切片中提取词元$\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_L)$。基于确定性文本的映射会选择与 q相关的器官集合 $\mathcal{B}(q)$；当没有器官与之关联时，则使用 $\mathcal{B}(q) = \{n+1\}$。对于每个选定的器官 k，$\mathbf{S}_k$ 包含空间中心位于 Ω_k的三维边界框内的切片令牌；而全局回退使用$\mathbf{S}_{n+1} = \mathbf{S}$。将投影后的器官词元$\mathbf{h}_k$ 用作查询，并将$\mathbf{S}_k$ 用作键和值，通过残差注意力将器官上下文与区域切片细节融合为一个证据令牌，

$$\mathbf{e}_k = \mathbf{h}_k + \mathrm{MHA}(\mathbf{h}_k, \mathbf{S}_k, \mathbf{S}_k), \qquad k \in \mathcal{B}(q). \tag{3}$$

因此，问题决定选择哪些器官，掩码边界限定符合条件的切片令牌，而学习到的注意力则聚合其内容。注意力输出投影采用零初始化；对于空的词元集合$\mathbf{S}_k$，则返回$\mathbf{e}_k = \mathbf{h}_k$，不进行二维更新。每个证据令牌$\mathbf{e}_k$ 都会紧接着插入$\mathbf{h}_k$之后，从而保留账本中的所有基础槽位。语言主干网络接收 $n+1+K$ 视觉词元，其中 $K = |\mathcal{B}(q)|$，而非原始切片令牌。

3.3 决策原生预测与监督

Qwen3.5 将交错排列的视觉词元与问题及候选答案一同处理。每个候选项 c 均由单个词表词元表示。其在提示末尾位置的 logit z_c 会在候选项集合上归一化，

$$p_\theta(c \mid V, q) = \frac{\exp z_c}{\sum_{c' \in \mathcal{C}(q)} \exp z_{c'}}. \tag{4}$$

预测答案及其置信度为

$$\hat{c} = \arg\max_{c \in \mathcal{C}(q)} p_\theta(c \mid V, q), \qquad \hat{p} = p_\theta(\hat{c} \mid V, q). \tag{5}$$

这种读取方式无需自回归式答案解码，每个问题只需进行一次语言骨干网络前向传播。训练从投影器-证据对齐开始，随后进行裁决微调（VFT）。两个阶段都使用相同的提示和读取方式，直接监督候选分布。

对于一个包含 N 个样本的批次，$\mathcal{L}_{\text{task}} = N^{-1} \sum_{i=1}^{N} \ell_i$。设 $p_{i,\,k}$ 表示候选项 k 的概率，y_i 表示参考索引，K_i 表示候选项数量。各任务特定的损失为

$$\begin{aligned}
\ell_i^{\text{bin}} &= -\log p_{i,y_i}, \\
\ell_i^{\text{ord}} &= -\log p_{i,y_i} + \lambda_{\text{rps}} \sum_{j=1}^{K_i - 1} (\widehat{F}_{i,j} - F_{i,j})^2, \\
\ell_i^{\text{mc}} &= -\log p_{i,y_i} + \lambda_{\text{b}} \sum_{k=1}^{K_i} (p_{i,k} - \mathbb{I}(k = y_i))^2.
\end{aligned} \tag{6}$$

二元问题使用交叉熵。序数问题还使用排序概率得分，其中 $\widehat{F}_{i,\,j} = \sum_{k=1}^{j} p_{i,\,k}$ 和 $F_{i,\,j} = \mathbb{I}\,(y_i \le j)$，以考虑严重程度等级的顺序。多项选择题还会加入布里尔分数，用于衡量预测与独热参考分布之间的平方误差。

3.4 反事实校准训练

对完整输入的监督，并不能直接约束解剖学证据缺失时的置信度。以 VFT 检查点为初始化，反事实校准训练（CCT）结合了不同证据条件下的监督，并对证据盲视图施加明确的置信约束。

反事实视图。对于选定槽位，CCT 会在式（2）中的多层感知机之前，将其三维特征替换为共享缺失特征占位符，并禁用这些槽位的二维证据更新。随后，每个受影响的证据令牌都会简化为其干预后的台账令牌 $\mathbf{e}_k = \mathbf{h}_k$，同时保留交错令牌布局。对于映射到器官集合 $\mathcal{B}(q)$ 的体数据 V 和问题 q，CCT 会评估四种视图，它们共享相同的问题文本和标准答案。事实视图保留所有提取的三维特征和二维证据更新。目标消融视图会一并替换 $\mathcal{B}(q)$ 中的所有已绑定器官及全局槽位 $n{+}1$，同时保留未绑定器官。对照消融视图会替换 $\mathcal{B}(q) \cup \{n{+}1\}$ 之外随机抽取的一个已出现器官，且绝不消融全局槽位。证据盲视图会将所有 $n{+}1$ 视觉槽位替换为占位符，并禁用所有二维更新，从而移除所有图像衍生特征，同时保留占位符令牌序列。如果目标视图或对照视图不可行，则在任务平均值中以事实读出代替；而未绑定问题仍可采用证据盲视图。

多视图置信度监督。在 CCT 期间，$\mathcal{L}_{\text{task}}$ 使用相同的参考答案，对事实视图、目标消融视图和控制消融视图的读出应用式（6）中的损失并取平均。这样可以在不同证据条件下应用相同的评分规则，而无需指定合成软目标。证据盲视图不参与任务监督，而是接收置信度间隔惩罚。令 c 和 $\check{c}$ 分别表示事实视图和证据盲视图下的最大候选项概率。它们的决策分数 r 和 $\check{r}$ 对于二分类和多项选择题，表示完全匹配正确性；对于有 $K_i - 1$ 个等级的序数问题，则表示 1 减去绝对等级误差与 K_i 的比值。对于单个样本，辅助损失为

$$\mathcal{L}_{\text{blind}} = \mathbb{I}(r > \check{r})\big[m_{\text{blind}} - (\text{sg}[c] - \check{c})\big]_+, \tag{7}$$

其中 $[x]_+ = \max(0, x)$。当去除所有图像衍生特征会降低决策分数时，该指示函数才会激活损失。合页项鼓励盲置信度至少比事实置信度低 m_{blind}。停止梯度 $\text{sg}[\cdot]$ 会阻断经由 c的盲视图损失梯度直接传播，但经由盲视图预测的梯度仍然有效。整体 CCT目标函数为

$$\mathcal{L}_{\text{CCT}} = \mathcal{L}_{\text{task}} + \lambda_{\text{blind}}\mathcal{L}_{\text{blind}}, \tag{8}$$

与 $\lambda_{\text{blind}} = 0.2$ 和 $m_{\text{blind}} = 0.20$。

4 CT-Verdict 基准

CT-Verdict 将三维CT体积与结构化临床问题及候选答案配对，用于评估模型在分布内数据和外部数据上的决策性能。它涵盖第 3.1 节定义的三种任务类型。

4.1 问题构建与验证

首先将放射学报告解析为结构化发现记录，其中包含所见、相关器官或解剖区域及其极性，并可选地记录严重程度、大小、数量、侧别以及与邻近结构的关系等属性。固定合成规则将这些记录转换为问题文本、候选集合和参考标签。二分类问题采用“否/是”答案，多项选择题包含两个至五个候选项，序数问题则使用从 0 到 4 的五个严重程度等级。

两名独立的大语言模型读者根据每个问题的来源句子作答，并将答案与规则推导标签进行比较。对于不一致的答案，由第三名读者进行复核。此外，人工阅片者还会对随机抽取的样本进行审核。未达到读者一致性标准的问题将被弃用，而非重新标注。问题合成与标签验证是两个独立阶段：前者确定问题文本和参考标签，后者则检查这些内容与来源报告是否一致。

4.2 数据来源与划分

CT-Verdict-IID。 Merlin-CT（Blankemeier 等人，2024）提供了 25,270 项增强和非增强腹部 CT 检查及其配对报告，用于训练和分布内评估。同一项检查中的所有问题均分配至同一数据划分。测试集包含来自 5,080 项检查的 56,825 个问题，其中包括 37,925 个二分类（Noul）、16,210 个多项选择（选择）和 2,690 个序数（得分）问题，涵盖 17 个器官系统和 30 项临床发现。

CT-Verdict-OOD。 外部测试集采用了来自 AMOS-MM 的 421 项多中心 CT 检查及其报告（Ji等人，2024）。该数据集包含 2,055 个问题，其中包括 1,195 个二分类、801 个多项选择和 59 个序数问题，涵盖 12 个器官系统和 24 项临床发现。

4.3 评估方案

对于二分类和多项选择题，我们报告逐题准确率（Acc）、平衡准确率（mBAcc）和 ROC 曲线下面积（mAUC）。后两项指标是对问题所涉及的疾病类别进行宏平均后得出的。对于有序严重程度，我们报告平均绝对误差（MAE）、归一化排序概率评分（KRPS）（Epstein，1969）和二次加权卡帕（QWK）（Cohen，1968），分别用于衡量预测误差、有序概率准确度和机会校正的有序一致性。我们还使用最高标签期望校准误差（ECE）（Guo等人，2017）和多分类布里尔评分（Glenn等人，1950）评估概率校准。为将校准不足与区分能力区分开来，我们进一步采用 Murphy 分解，并报告可靠性（REL）和分辨率（RES）（Murphy，1973）。

5 实验

表1 代表性通用及医疗视觉语言模型在*CT-Verdict-IID*上的比较。准确率（Acc.）为逐题完全匹配准确率。mBAcc 和 mAUC 是对所发现疾病进行宏平均后得到的结果。得分 MAE、KRPS 和 QWK 根据 0–4 级计算。**粗体** 表示各列中的最佳结果，下划线表示次优结果。

模型	Noul						选择				得分				
	准确率↑	mB准确率↑	宏平均曲线下面积↑	期望校准误差↓	布里尔分数↓		准确率↑	mB准确率↑	宏平均曲线下面积↑	期望校准误差↓	布里尔分数↓	平均绝对误差↓	KRPS↓	二次加权 kappa ↑	期望校准误差↓
布里尔分数↓															
通用视觉语言模型															
LLaVA-Med-7B (Li等人，2023)	39.94	50.61	61.31	0.35	0.71	19.69	49.82	50.15	0.42	1.05	0.75	0.14	0.15	0.19	0.93
RadFM（Wu等人，2025）	58.88	50.03	51.72	0.11	0.47	39.36	49.92	50.42	0.25	0.89	0.93	0.21	0.07	0.54	1.19
CT-Chat（Hamamci etal.，2026年）	41.92	50.09	52.07	0.34	0.71	22.30	50.21	52.22	0.41	1.11	0.96	0.18	-0.08	0.46	1.03
M3D-Phi3-4B（Bai etal.，2024）	47.48	54.55	57.85	0.50	1.01	9.71	50.01	50.95	0.85	1.72	0.99	0.21	0.11	0.51	1.13
M3D-Llama2-7B（Bai etal.，2024）	52.31	50.45	52.27	0.18	0.61	8.72	50.01	50.39	0.81	1.67	1.00	0.22	-0.01	0.60	1.32
Merlin（Blankemeier 等人，2024）	66.56	56.06	79.70	0.16	0.43	38.22	50.56	61.49	0.14	0.77	0.70	0.14	0.13	0.24	0.89
HealthGPT-M3（Lin 等人，2025）	43.14	49.79	55.40	0.32	0.71	23.12	50.19	52.93	0.25	0.89	1.04	0.18	0.06	0.12	0.83
Med3DVLM（Xin 等人，2025）	63.13	50.44	54.94	0.28	0.65	28.74	50.89	52.75	0.52	1.21	0.80	0.17	0.29	0.44	1.04
InternVL3.5-8B (Wang 等人，2025)	43.94	53.67	72.44	0.24	0.57	20.23	49.34	49.62	0.48	1.16	1.05	0.17	-0.02	0.09	0.78
Hulu-med-4B (Jiang 等人，2025)	50.71	57.72	74.71	0.13	0.47	30.31	52.86	64.08	0.27	0.94	1.06	0.17	0.24	0.11	0.80
Hulu-med-7B (Jiang 等人，2025)	64.67	61.73	78.71	0.07	0.41	29.00	56.00	67.44	0.17	0.83	1.39	0.24	0.22	0.25	0.95
Qwen3.5-4B(Qwen团队，2026年)	53.16	58.44	75.02	0.15	0.51	29.69	52.54	64.87	0.29	0.92	1.12	0.19	0.23	0.14	0.85
Qwen3.5-9B(Qwen团队，2026年)	57.90	63.08	76.58	0.09	0.44	36.64	54.52	67.60	0.24	0.79	1.27	0.22	0.19	0.16	0.88
OmniCT-3B(Lin 等人，2026b)	44.79	52.28	56.73	0.16	0.54	23.79	50.25	55.46	0.37	1.00	1.17	0.21	0.07	0.30	0.93
OmniCT-7BLin 等人 (2026b)	44.86	51.86	53.20	0.18	0.56	19.44	50.57	54.89	0.33	0.97	0.88	0.15	0.02	0.19	0.81
HealthGPT-Pro-4B (Lin 等人，2025)	70.96	65.90	76.68	0.10	0.41	37.50	55.73	69.57	0.07	0.74	1.05	0.17	0.30	0.13	0.80
HealthGPT-Pro-8B (Lin 等人，2025)	62.03	68.04	77.31	0.02	0.45	35.93	55.89	67.80	0.08	0.76	0.93	0.15	0.34	0.09	0.77
ClinFusion-8B (Yuan 等人，2026年)	65.66	56.58	64.68	0.08	0.44	37.65	51.26	56.51	0.11	0.72	1.02	0.17	0.22	0.09	0.81
GPT-5.4 (OpenAI, 2025)	70.23	69.24	–	–	–	45.92	**64.31**	–	–	–	0.75	–	0.46	–	–
决策视觉语言模型															
OmniMed-Jev（Tangand Chen，2026年）	59.68	50.55	50.66	0.39	0.78	36.06	51.39	57.91	0.37	0.93	0.79	0.17	0.11	0.31	0.93
Intern-Decision-2BZhu等人 （2026年）	56.81	60.04	70.13	0.02	0.49	38.05	58.92	68.82	0.03	0.72	0.57	0.10	0.40	0.07	0.68
Intern-Decision-4BZhu等人 （2026年）	64.57	60.42	72.04	0.03	0.43	45.49	57.37	65.17	**0.02**	0.68	0.52	0.10	0.35	0.13	0.73
Jev-1.14（Typesafe.ai，2025）	71.39	65.14	79.31	0.04	0.37	50.99	58.52	71.78	0.02	0.63	0.60	0.12	0.36	0.21	0.79
Radev-0.8B（我们的）	88.20	66.59	85.48	0.01	0.17	68.61	57.73	78.19	0.02	0.44	0.40	0.06	0.53	0.03	0.42
Radev-2B（我们的模型）	**90.08**	**71.79**	**87.79**	**0.01**	**0.15**	**71.33**	62.06	**81.91**	0.02	**0.41**	**0.39**	**0.06**	**0.54**	**0.02**	**0.41**

5.1 实现细节

模型配置。 我们使用冻结的RADAR分割与编码网络（Zhang等人，2026年）实例化3D分支，该网络从未在*CT-Verdict* 测试集的数据源上进行训练。它提供256维的器官读出（$d = 256$），对应36 个器官槽位和一个全局槽位（$n = 36$）。每个器官都从第一个完整包含该器官的处理窗口中读出。全局特征使用其自身的查询、注意力模块和投影，从对体数据的另一份预处理副本执行的独立无窗口处理中聚合所有多尺度令牌。2D视觉编码器和语言主干网络均来自Qwen3.5（Qwen团队，2026年）。

监督配置。 在投影器-证据对齐期间，Qwen3.5 语言主干网络和原生 2D 视觉编码器被冻结，同时训练器官投影器和证据注意力。VFT 会联合更新语言主干网络、2D 视觉编码器和桥接模块。整个训练过程中，3D 分支始终保持冻结。在公式（6）中，我们使用 $\lambda_{\mathrm{b}} = 0.1$ 和 $\lambda_{\mathrm{rps}} = 0.5$。

训练。 所有模型均在四块 96 GB H20 GPU 上运行，并采用研究级数据并行分片，因此同一项研究中的所有问题均由同一个进程处理。两个监督阶段均使用大小为 32 的有效批次、AdamW（Ilya 等人，2019），且每个阶段均重新初始化优化器状态。我们训练 1,000 步，对齐阶段的学习率为 10^{-3}；VFT 阶段以 2×10^{-5} 训练一个周期；CCT 阶段以 4×10^{-6} 训练 700 步。

推理。 我们的模型完全按照训练时的方式进行读出，即通过公式（4）中的受限 softmax。对于基线模型，似然读出会对各模型聊天模板下 K 教师强制候选字符串的词元对数概率求和后进行 softmax，从而获得答案和完整的候选分布。决策模型不提供生成路径，而 GPT-5.4 只返回文本，不提供候选项对数概率。若有官方三维配置，则切片数量遵循该配置；否则默认设为 32，若模型仅支持少量图像，则将数量限制在其最大输入限制以内。

5.2 结果

CT-Verdict-IID 上的主要结果。 表 1 比较了 Radev 与具有代表性的一般视觉语言模型和决策原生视觉语言模型在两种读出模式下的表现。首先，Radev 在全部三项任务中均确立了显著优势：即使是我们的 2B 版本，整体准确率也达到 83.75%，超过了最强的开源基线模型；这是因为二维决策原生模型缺乏体积上下文，而生成式视觉语言模型难以处理有序严重程度。闭源模型的泛化能力优于开源基线模型，但仍落后于 Radev ∼19 %，凸显了领域特定对齐的必要性。除了原始准确率外，临床部署更关键的要求是模型报告的置信度是否可信。即使是准确率最高的开源和闭源基线模型，仍然过度自信且校准失配严重，这证实了仅有任务准确率并不能保证不确定性估计可靠。相比之下，Radev 在所有任务类别中的布里尔分数均最低，并将最高标签 ECE 降至0.01–0.02 范围。这确保了 Radev 的预测概率可直接用于阈值化，以实现选择性预测和临床弃权。

表2 通用视觉语言模型与医学视觉语言模型在 *CT-Verdict-OOD* 上的对比。

模型	Noul		选择		得分	
	Acc↑	宏平均曲线下面积↑	Acc↑	宏平均曲线下面积↑	MAE↓	KRPS↓
通用视觉语言模型						
LLaVA-Med-7B（Li等人，2023）	35.80	55.80	17.60	58.50	0.653	0.152
RadFM（Wu等人，2025）	55.70	43.50	55.40	46.00	0.582	0.130
CT-Chat（Hamamci等人，2026年）	40.00	39.70	8.90	51.30	1.440	0.229
M3D-Llama2-7B（Bai等人，2024）	41.80	53.10	15.70	51.40	0.993	0.222
Merlin（Blankemeier 等人，2024）	61.30	69.90	11.20	62.70	0.501	0.141
HealthGPT-M3（Lin 等人，2025）	42.30	47.90	22.60	50.60	1.125	0.185
Med3DVLM（Xin 等人，2025）	56.40	60.30	9.10	53.90	0.910	0.197
InternVL3.5-8B（Wang 等人，2025）	42.00	64.70	18.10	56.70	0.945	0.159
Hulu-med-7B（Jiang 等人，2025）	64.90	66.20	19.60	71.40	1.417	0.237
Qwen3.5-9B（Qwen团队，2026年）	54.31	65.60	25.34	61.76	1.191	0.185
OmniCT-7B（Lin 等人，2026b）	41.80	45.40	14.00	57.00	0.943	0.147
HealthGPT-Pro-8B（Lin 等人，2025）	56.70	71.60	35.00	70.20	0.990	0.162
ClinFusion-8B（Yuan 等人，2026年）	55.73	47.40	58.18	57.17	1.143	0.177
GPT-5.4（OpenAI，2025）	63.93	–	36.70	–	0.763	–
决策视觉语言模型						
OmniMed-Jev（Tang和Chen，2026年）	59.90	43.80	54.20	61.80	0.558	0.123
Intern-Decision-4B（Zhu等人，2026年）	63.30	64.60	53.70	58.60	0.321	0.106
Jev-1.14（Typesafe.ai，2025）	62.90	71.70	53.30	67.90	0.504	0.122
Radev-0.8B（我们的）	69.96	72.89	65.79	72.31	0.279	0.032
Radev-2B（我们的）	**72.72**	**77.61**	**69.29**	**75.16**	**0.234**	**0.028**

在 CT-Verdict-OOD 上的泛化能力。表2 评估了模型在外部分布偏移下的鲁棒性。尽管大多数基线模型的选择准确率和序数得分估计表现均有所下降，Radev-2B 仍展现出强大的跨中心泛化能力，在所有指标上始终优于其他模型。具体而言，Radev-2B 在NOUL上达到 72.72% 的准确率，在选择上达到 69.29% 的准确率，在二分类和多项选择准确率上分别比最强的闭源模型高出 9% 和 16%。Radev-2B 的序数 KRPS 和平均绝对误差也最低，KRPS 比最强基线低约 3.8× （95% 研究自助法置信区间 $[2.44, 7.20]$），表明其严重程度分布在外部分布偏移下仍保持良好的校准。

表3 Radev 与最先进方法在 *CT-Verdict-IID* 和 *CT-Verdict-OOD* 中的诊断任务上的比较。DD 表示疾病鉴别，OAD 表示器官异常诊断，SG 表示严重程度分级，AT 表示急诊分诊，LA 表示侧别评估。 ✓ 表示支持， × 表示不支持。

模型	临床任务					IID		OOD	
	DD	OAD	SG	AT	LA	Acc	mAUC	Acc	mAUC
Merlin-ZS	✓	×	×	×	×	66.00	79.80	63.23	71.67
RADAR	✓	×	×	×	×	66.73	80.10	60.50	74.71
Radev-2B	✓	✓	✓	✓	✓	**90.08**	**87.79**	**72.72**	**77.61**

与零样本表征匹配的比较。表3对比了决策原生监督与零样本表征匹配。嵌入相似度读出仅能支持疾病级别判别：器官级异常诊断、严重程度分级、分诊和侧别评估均需要结构化判定监督，向量匹配无法提供此类监督，而且相似度分数并非校准概率。Radev 通过单一的校准候选项读出支持全部五类任务，并且在两个队列上均取得更高的宏平均曲线下面积。

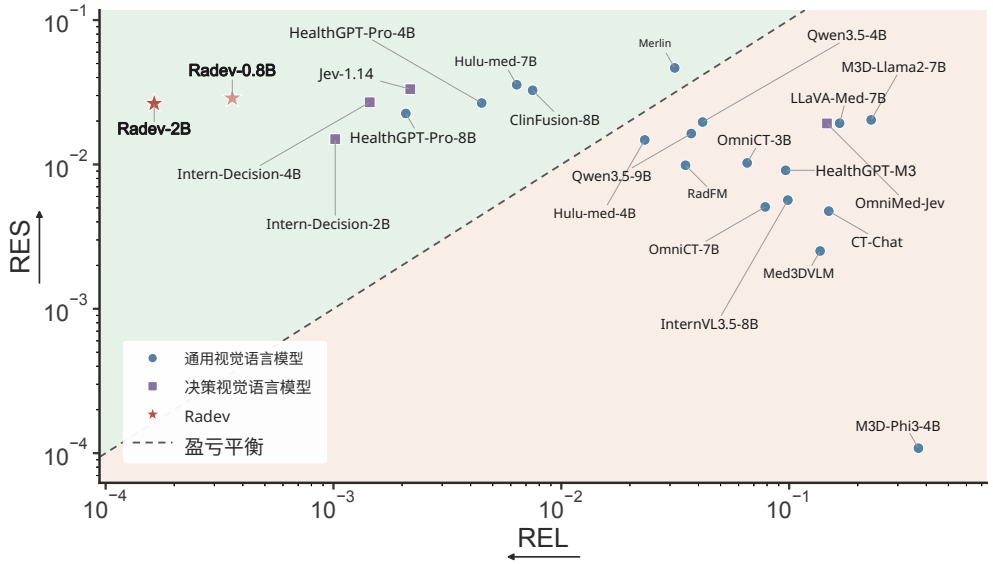


图3 可靠性与分辨率分析。每个点代表一个模型，其位置由可靠性误差（REL）和分辨率（RES）确定；二者根据顶级标签布里尔分数的分箱墨菲分解计算得出。虚线表示 RES ＝REL 边界。

5.3 置信度分析

置信度估计有多可靠? 模型可能预测正确答案，却给出不准确的置信度分数。因此，我们评估其置信度是否与预测正确的频率相符。图3 总结了置信度质量的两个方面。可靠性误差（REL）衡量置信度组内平均置信度与准确率之间的差距，数值越低表示校准效果越好。分辨率（RES）衡量这些组之间准确率的差异。这两个指标共同评估：为不同的单项预测赋予不同置信度分数，是否优于对每个预测都仅报告模型的整体准确率。在这项评估中，22 个基线中有 14 个的表现不如恒定置信度基准；而 Radev 的布里尔技巧分数为正，达到 0.193，最高标签 ECE 为 0.010。与最强基线相比，其校准误差降低了一个至两个数量级，表明只有 Radev 将最先进的诊断准确率与可信赖的置信度估计结合起来。

通过基于置信度的拒绝实现选择性预测。 图 4 展示了基于置信度的拒绝如何将 Radev 报告的概率转化为可执行的临床安全控制。将置信度最低的

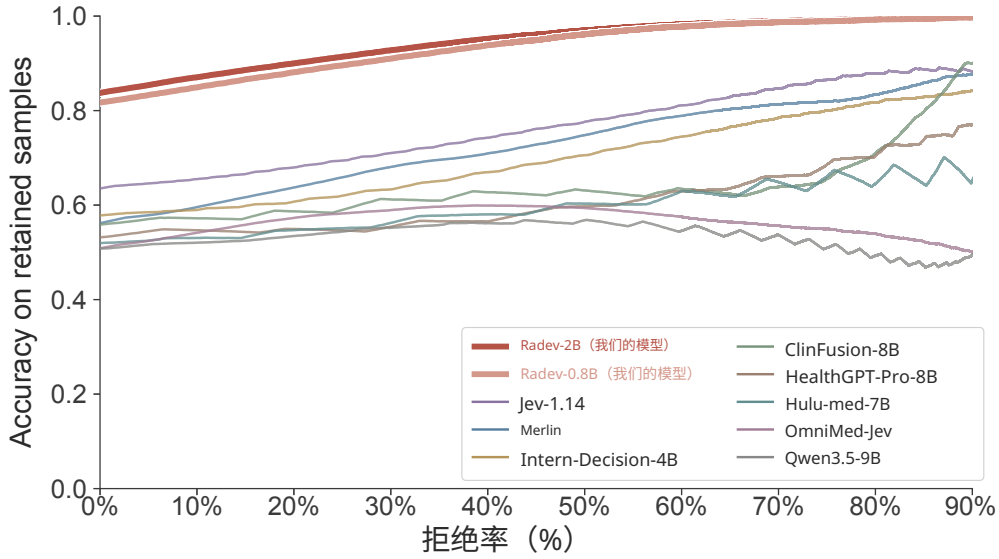


图4 基于置信度的拒绝下的准确率。 按预测置信度从低到高依次拒绝样本。拒绝率是被拒绝样本占全部样本的百分比，准确率则基于保留样本计算。

20%、40% 和 50% 的病例转交人工阅卷者处理，可使 Radev-2B 的保留准确率稳步从 83.8% 提升至 90.0%、95.2% 和 97.0%。相比之下，在最高达 90% 的任何转诊率下，基线模型的可靠性都无法与之匹敌：表现最强的基线（Jev-1.14）在 50% 的转诊率下准确率也仅为 77.6%。由于 Radev 的准确率–拒绝率曲线平滑，临床工作流程可以在风险–工作负荷前沿上选择任意工作点，从而对判定进行评分，并优先将不确定病例转交临床医生复核。

表4 消融　　研究针对 Radev-2B 在*CT-Verdict-IID*上经过 VFT 后的视觉编码器架构进行。

变体	二维切片	Noul		选择		得分	
		Acc↑	宏平均曲线下面积↑	Acc↑	宏平均曲线下面积↑	MAE↓	KRPS↓
二维编码器	16	83.20	69.07	63.63	62.49	0.43	0.07
二维编码器	32	83.10	67.21	62.85	60.83	0.43	0.07
三维编码器	0	88.11	83.10	67.52	77.40	0.41	0.07
标记拼接	32	88.69	84.19	68.29	77.03	0.41	0.07
无对齐	32	89.33	85.95	69.51	80.36	0.39	**0.06**
无器官嵌入	32	88.36	85.24	67.61	77.38	0.40	0.06
无槽位 $n{+}1$	32	86.93	82.53	67.81	75.25	0.40	0.06
器官对齐视觉	4	88.79	85.53	69.10	77.99	0.39	0.06
器官对齐视觉	8	88.66	84.91	68.67	78.92	0.39	0.06
器官对齐视觉	16	89.14	86.24	69.63	80.27	**0.39**	0.06
器官对齐视觉	32	**89.91**	**87.89**	**70.30**	**81.41**	0.39	0.06

5.4 消融研究

视觉编码器架构。表4 在*CT-Verdict-IID* 上对器官对齐视觉编码器进行了消融研究，评估在 VFT 之后进行。纯二维切片编码器难以捕捉体积连续性。相反，纯三维编码器显著提升了体积判别能力，而简单的二维–三维标记拼接仅带来微弱增益。用我们的设计替换简单融合后，所有任务的整体表现均达到最佳。组件级消融证实了各项设计在器官账本表示中的作用：移除三维空间框对齐、学习式器官身份嵌入或全局上下文槽位，都会持续削弱诊断判别能力。最后，在我们的设计下将二维流扩展至 32 个切片，使宏平均曲线下面积在Noul和选择两项任务上均单调提升，表明器官门控交叉注意力能够有效利用细粒度切片线索，同时避免标记稀释。为将架构带来的提升与训练数据的影响区分开来，表5比较了 Radev-2B-VFT 与采用完全相同的 VFT 配置进行微调的基线模型，这些模型均在*CT-Verdict*上评估。

表5 比较 Radev-2B-VFT 与基线在 CT-Verdict-IID 上采用相同训练配置进行 VFT 后的结果 e *CT-Verdict-IID*。

模型	Noul		选择		得分	
	Acc↑	宏平均曲线下面积↑	Acc↑	宏平均曲线下面积↑	MAE↓	KRPS↓
Qwen3.5-2B（Qwen团队，2026年）	83.10	67.21	62.85	60.83	0.43	0.07
Qwen3.5-4B（Qwen团队，2026年）	82.96	67.45	62.93	61.42	0.43	0.07
ClinFusion-8B（Yuan 等人，2026年）	80.36	51.23	59.34	50.20	0.46	0.07
Radev-2B-VFT（我们的模型）	**89.91**	**87.89**	**70.30**	**81.41**	**0.39**	**0.06**

Radev-2B-VFT 在所有指标上均显著优于这些规模更大的模型，证实我们性能领先的主要驱动因素是所设计的架构。

表6 Radev-2B 在 *CT-Verdict-IID* 上训练阶段的消融研究，分别从准确率（上）和校准（下，$\times 100$）层面进行评估。

对齐	VFT	CCT	Noul		选择		得分	
			Acc↑	宏平均曲线下面积↑	Acc↑	宏平均曲线下面积↑	MAE↓	KRPS↓
✓	×	×	86.93	80.78	66.15	72.27	0.46	0.07
×	✓	×	88.93	85.50	68.26	77.65	0.40	0.06
✓	✓	×	89.91	**87.89**	70.30	81.41	**0.39**	0.06
✓	✓	✓	**90.08**	87.79	**71.33**	**81.91**	0.39	**0.06**

对齐	VFT	CCT	Noul		选择		得分	
			布里尔分数↓	ECE↓	布里尔分数↓	ECE↓	布里尔分数↓	ECE↓
✓	×	×	19.05	1.96	47.70	2.88	50.39	5.63
×	✓	×	16.47	1.10	44.33	2.57	43.27	6.89
✓	✓	×	14.96	2.00	41.78	2.98	43.34	9.06
✓	✓	✓	**14.57**	**0.57**	**40.61**	**2.09**	**41.03**	**2.04**

训练策略。 表 6 考察了器官-账本对齐、VFT 和 CCT 的逐步效果。将对齐与 VFT 结合，相较于单独采用任一阶段，能够显著提升诊断准确率。加入 CCT 后，不仅保持或进一步提升诊断性能，还能在每项任务中大幅降低布里尔分数并显著削减 ECE。表 7 进一步通过 Murphy 分解（可靠性 REL 和分辨率 RES）细分了 CCT 如何实现这一校准增益。在事实视图、目标消融视图和对照消融视图之间引入多视图反事实与联合任务监督，带来了主要的校准跃升，使 ECE 降至 $0.60/2.26/2.20$，同时保持区分能力。这一实证结果验证了第3.4节所述的机制：使用原始独热金标准标签监督视觉上无法区分的目标消融视图，会自然地将过度自信的预测拉向条件经验先验。最后，纳入证据盲视图和证据依赖铰链 $\mathcal{L}_{\text{blind}}$ 能够抑制证据无关置信度，实现最低的 REL 和 ECE。

表7 在*CT-Verdict-IID* 上进行CCT消融实验，以评估校准效果（Murphy分解，$\times 100$）.

变体	Noul			选择			得分		
	REL↓	RES↑	ECE↓	REL↓	RES↑	ECE↓	REL↓	RES↑	ECE↓
继续 VFT	0.09	33.52	0.99	0.79	**23.65**	3.13	0.71	**18.46**	4.17
+ 多视图	0.05	**33.54**	1.15	0.66	23.63	3.45	0.45	18.17	3.07
+全视图监督	**0.02**	33.40	0.60	0.56	23.56	2.26	0.38	18.34	2.20
+ 盲视图与合页	0.02	33.39	**0.57**	**0.55**	23.58	**2.09**	**0.35**	18.25	**2.04**

6 结论

我们介绍了**Radev**，这是一个用于结构化三维腹部CT解读的决策原生多模态框架，并推出了**CT-Verdict** 基准测试。Radev 将可复用的三维器官台账与问题绑定的二维切片证据相结合，在单次前向传递中预测封闭集答案分布，无需自回归解码；与此同时，CCT 通过对比完整视图、消融视图和证据盲视图，将置信度锚定于解剖学证据。在*CT-Verdict-IID* 和外部*CT-Verdict-OOD* 评估中，Radev 均取得了最先进的诊断准确率和校准能力，为选择性预测和临床弃权提供了可直接信赖的概率。未来，我们将把器官锚定台账和反事实校准框架扩展到其他体积成像模态，尤其是多序列磁共振成像，以及更广泛的解剖区域。

参考文献

Anastasios N Angelopoulos 和 Stephen Bates。保形预测：入门简介。*Foundations and Trends in Machine Learning*，16(4):494–591，2023。Fan Bai、Yuxin Du、Tiejun Huang、Max Q-H Meng 和 Bo Zhao。M3d：借助多模态大语言模型推进三维医学图像分析。*arXiv* 预印本 *arXiv:2404.00578*，2024。Louis Blankemeier、Joseph Paul Cohen、Ashwin Kumar、Dave Van Veen、Syed Jamal Safdar Gardezi、Magdalini Paschali、Zhihong Chen、Jean-Benoit Delbrouck、Eduardo Reis、Cesar Truyts 等。Merlin：一种面向三维计算机断层扫描的视觉语言基础模型。*Research square*，第 rs–3 页，2024。Cloudflare。推出具备完整多模态能力的 clef-omni。https://blog.cloudflare.com/ clef-faster-cheaper-multimodal/，2026。扩展 Clef系列，使其能够在单一流程中原生处理音频、视频、图像和文本；访问于2026年10月。Jacob Cohen。加权卡帕系数：名义尺度一致性指标，用于衡量分级分歧或部分得分。*Psychological Bulletin*，70(4):213，1968。Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。BERT：用于语言理解的深度双向Transformer预训练。载于*2019*年计算语言学协会北美分会年会论文集：人类语言技术，第*1*卷（长文与短文），第4171–4186页，2019。Edward S Epstein。针对有序类别概率预测的评分系统。*Journal of Applied Meteorology (1962-1982)*，8(6):985–987，1969。Yonatan Geifman 和 Ran El-Yaniv。深度神经网络的选择性分类。*Advances in Neural Information Processing Systems*，30，2017。W Brier Glenn 等人。以概率形式表达的预测验证。*Monthly Weather Review*，78 3，1950 1 1( ): – 。Chuan Guo、Geoff Pleiss、Yu Sun 和 Kilian Q Weinberger。现代神经网络的校准。载于国际机器学习大会，第1321–1330页。PMLR，2017。Ibrahim Ethem Hamamci、Sezgin Er、Chenyu Wang、Furkan Almas、Ayse Gulnihan Simsek、Sevval Nil Esirgun、Irem Dogan、Omer Faruk Durugol、Benjamin Hou、Suprosanna Shit 等。基于多模态数据集的三维计算机断层扫描通用基础模型。*Nature Biomedical Engineering*，第1–19页，2026。Loshchilov Ilya、Hutter Frank 等。解耦权重衰减正则化。*ICLR*论文集，7:11，2019。Yuanfeng Ji、Chongjian Ge、Ruijiang Li 和 Ping Luo。AMOS-MM：腹部多模态分析挑战赛：挑战赛设计的结构化描述，2024年4月。https://doi.org/10.5281/zenodo.10992155。Songtao Jiang、Yuan Wang、Sibo Song、Tianxiang Hu、Chenyi Zhou、Bin Pu、Yan Zhang、Zhibo Yang、Yang Feng、Joey Tianyi Zhou 等。Hulu-Med：一种面向整体医学视觉语言理解的透明通用模型。*arXiv* 预印本 *arXiv:2510.08668*，2025。Joe-rq。ReJev：受JEV启发的 MiniCPM5-2B 决策模型实验。https://github.com/Joe-rq/ReJev，2026。

Chunyuan Li、Cliff Wong、Sheng Zhang、Naoto Usuyama、Haotian Liu、Jianwei Yang、Tristan Naumann、Hoifung Poon 和 Jianfeng Gao。LLaVA-Med：用一天时间训练生物医学领域的大型语言与视觉助手。*Advances in Neural Information Processing Systems*，36:28541–28564，2023。Sijing Li、Zhongwei Qiu、Jiang Liu、Wenqiao Zhang、Tianwei Lin、Yihan Xie、Jianxiang An、Boxiang Yun、Chenglin Yang、Jun Xiao 等。Tumorchain：用于可追溯临床肿瘤分析的交错式多模态思维链推理。载于国际学习表征会议，第2026卷，第119790–119819页，2026a。Sijing Li、Zhongwei Qiu、Zhuoya Wang、Boxiang Yun、Zhenyu Yi、Jianwei Xu、Wenqiao Zhang、Yingda Xia 和 Ling Zhang。E-mrl：用于可靠三维肿瘤分析的跨视图对齐的证据驱动多模态强化学习。载于国际医学图像计算与计算机辅助干预会议，第238–248页。施普林格，2026b。Tianwei Lin、Wenqiao Zhang、Sijing Li、Yuqian Yuan、Binhe Yu、Haoyuan Li、Wanggui He、Hao Jiang、Mengze Li、Xiaohui Song 等。HealthGPT：通过异构知识适配统一理解与生成的医学大型视觉语言模型。*arXiv* 预印本 *arXiv:2502.09838*，2025。Tianwei Lin、Zhongwei Qiu、Jie Cao、Jiang Liu、Wenjie Yan、Bo Zhang、Yu Zhong、Wenqiao Zhang、Yingda Xia 和 Ling Zhang。通过轨迹积分反馈调节解剖结构感知奖励，以进行体积计算机断层扫描分析。*arXiv* 预印本 *arXiv:2605.20277*，2026a。Tianwei Lin 、 Zhongwei Qiu 、 Wenqiao Zhang 、 Jiang Liu 、 Yihan Xie 、 Mingjian Gao、Zhenxuan Fan、Zhaocheng Li、Sijing Li、Zhongle Xie 等。OmniCT：迈向用于全面CT分析的统一切片-体积大型视觉语言模型。*arXiv* 预印本 *arXiv:2602.16110*，2026b。Mark Marosi。decider：单次处理的类型化决策与校准概率，2026。https://github.com/Mapika/ decider。

Allan H 应用气象学概率预报的一种新颖的评分。*月刊*，101(6):595–600，1973。 .

OmniJev 团队。OneJev：一个多模态系统单决策模型。https://github.com/OmniJev/OneJev，2026年。

OpenAI . GPT-5 简介 .4. https://openai.com/index/introducing-gpt-5-4/，202 5.

Jared Palmer。Kev：类似 Jev 的小型决策模型，可由你自行训练和运行。https://github.com/jaredpalmer/ kev，2026年。

Zhongwei Qiu、Hanqing Chao、Tiancheng Lin、Wanxing Chang、Zijiang Yang、Wenpei Jiao、Yixuan Shen、Yunshuo Zhang、Yelin Yang、Wenbin Liu等。利用 pixel-mamba 弥合局部归纳偏置与长程依赖，实现端到端全切片图像分析。载于*2025*年*IEEE/CVF*国际计算机视觉大会（*ICCV*），第22738–22747页。IEEE，2025年。

Qwen团队。Qwen3.5：迈向原生多模态智能体，2026年2月。https://qwen.ai/blog?id=qwen3.5。

Hao Tan 和 Mohit Bansal。LXMERT：从 Transformer 中学习跨模态编码器表示。载于*2019*年自然语言处理实证方法会议和第*9*届国际自然语言处理联合会议（*EMNLP-IJCNLP*）论文集，第5100–5111页，2019年。

Luyao Tang 和 Cheng Chen。OmniMed-Jev：通过系统一校准大型视觉语言模型置信度，实现可信的医学多模态决策。*arXiv* 预印本 *arXiv:2610.00381*，2026年。Cristian Tommasino 和 Antonio M. Rinaldi。PathJEV：面向组织病理学的决策原生视觉语言建模，2026年。https://arxiv.org/abs/2610.32478。Typesafe.ai。JEV 1.14.0。https://typesafe.ai/blog/jev-1-14，2025年。Weiyun Wang、Zhangwei Gao、Lixin Gu、Hengjun Pu、Long Cui、Xingguang Wei、Zhaoyang Liu、Linglin Jing、Shenglong Ye、Jie Shao 等。InternVL3.5：提升开源多模态模型的通用性、推理能力和效率

.

*arXiv*预印本 *arXiv:2508.18265*，2025年。Chaoyi Wu、Xiaoman Zhang、Ya Zhang、Hui Hui、Yanfeng Wang 和 Weidi Xie。借助网络规模的二维与三维医学数据，迈向放射学通用基础模型。《自然-通讯》，16(1):7866，2025年。Yu Xin、Gorkem Can Ates、Kuang Gong 和 Wei Shao。Med3dvlm：一种用于三维医学图像分析的高效视觉语言模型。*IEEE*生物医学与健康信息学汇刊，2025年。Tianrun Xu、Hongbang Fan、Jiahao Lin、Zilin Zhu、Zhenxin Diao、Longteng Guo 和 Jing Liu。Omnijev：一个全模态系统一决策模型，2026a。https://github.com/tinnel123666888/OmniJev。

Weiwen Xu、Hou Pong Chan、Long Li、Mahani Aljunied、Ruifeng Yuan、Jianyu Wang、Chenghao Xiao、Guizhen Chen、Chaoqun Liu、Zhaodonghui Li 等。Lingshu：用于统一多模态医学理解与推理的通用基础模型。*IEEE*模式分析与机器智能汇刊，2026b。Zhenyu Yi、Qiang Hu、Zhenhao Li、Jiaxuan Zhao、Yusong Sun 和 Lichi Zhang。CARVE：通过跨切片各向异性重分配视觉证据，实现高效的三维医学体数据理解。*arXiv*预印本 *arXiv:2608.04515*，2026年。Hangjie Yuan、Yichen Qian、Zhiwei Tang、Xianzhe Xu、Lirong Wu、Sicheng Yang、Jinwang Wang、Pengju Wang、Zhitao Zeng、Yizeng Han 等。ClinFusion：一种用于整体医学理解的以视觉为中心的多模态大语言模型系统。*arXiv*预印本 *arXiv:2607.24743*，2026年。Qi Zhang、Jianpeng Zhang、Weiwei Cao、Zilin Lu、Wanxing Chang、Haonan Ding、Cao Chen、Zhi Li、Xing Xue、Sinuo Wang 等。一种用于腹部CT诊断的专家级通用人工智能。《科学》，393(6817):eaec6129，2026年。Tong Zhu、Zhe Xu、Zefeng He、Yunkai Chai、Xuyang Hu、Dehong Zheng、Yun Luo、Yafu Li、Qi Zhang 和 Yu Cheng。Intern-Decision：通过自回归掩码语言建模实现高效的多模态决策。*arXiv*预印本 *arXiv:2610.22131*，2026年。

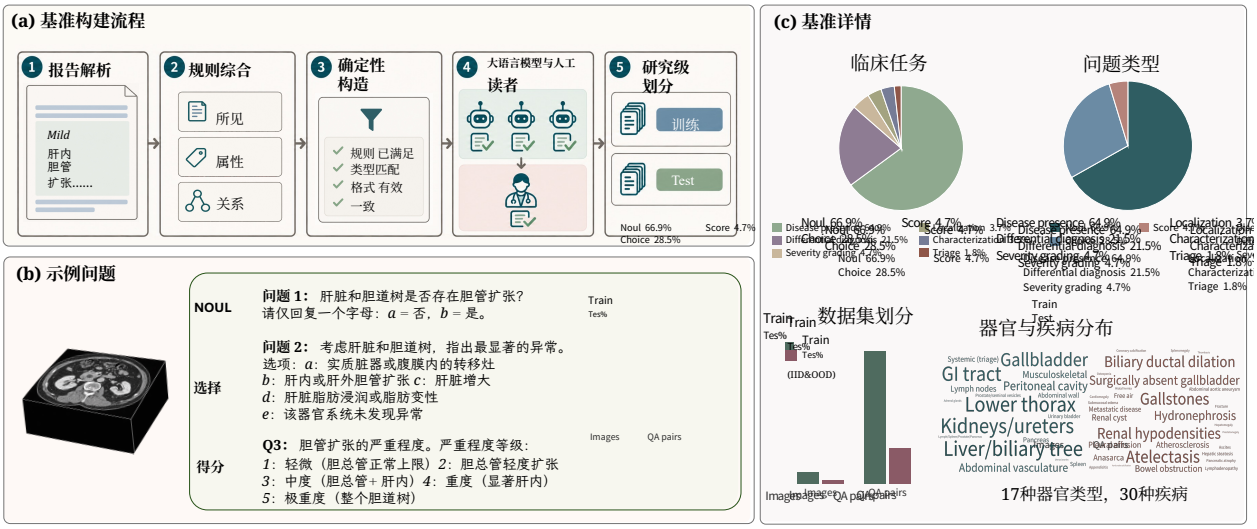
附录

补充大纲

- **章节A**：其他 CT-Verdict 详情
  – 第A.1节：数据集统计（解剖学、临床任务与疾病分布）– 第A.2节：验证与质量控制（多读者大语言模型审计、人工审计、纯文本审计）– 评估方案详情（分类、有序与校准指标）

- **第B节：补充方法详情**
  – 第B.1节：按器官分层的课程采样策略。– 第B.2节：反事实置信度调节的群体层面分析（命题1与证明）– 第B.3节：视觉编码器中的2D-3D特征融合（空间归属规则）

A AdditionalCT-Verdict详情



图A1 CT-Verdict概述。(a) 基准构建流程：将放射学报告解析为结构化发现；基于存在性、属性与关系的合成规则将每项发现转化为候选问题；问题以确定性方式生成，并由三位大语言模型读者对其标签进行验证，同时辅以人工抽查；最后，研究级划分将训练集与测试集分开。**(b)** 三种问题类型示例：二元存在性（Noul）、多选题和有序严重程度（得分）。**(c)** 数据集详情：问题类型与临床任务的分布、数据集划分，以及器官/疾病覆盖范围。

本节概述了对正文中基准测试描述的补充细节。基准测试概览见图 A1。

A.1 数据集统计

解剖结构覆盖范围。 *CT-Verdict* 涵盖 17 种解剖器官和区域类别，各划分的详细统计数据见表 A1。训练集（223,286 个问题）和 *CT-Verdict-IID* 测试集（56,825 个问题）的解剖结构分布几乎相同，主要涉及肝脏/胆道树、下胸部、肾脏/输尿管、胃肠道和胆囊。相比之下，外部 *CT-Verdict-OOD* 测试集侧重于核心腹部解剖结构（以肾脏/输尿管、腹膜腔和肝脏/胆道树为主），同时不包含下胸部、全身性/分诊、腹壁、膀胱和子宫类别。

临床任务和疾病分布。表 A2 和 A3 报告了训练集、IID 测试集和 OOD 测试集在五种主要临床能力（另设“其他”这一细粒度属性问题的剩余类别）以及 30 种标注疾病/发现中的分布。训练集和 IID 测试集在任务能力和各项单独发现方面都高度一致，呈现自然的长尾临床分布。相反，*CT-Verdict-OOD* 侧重于疾病鉴别和器官异常识别（不包含急诊分诊或侧别项目），省略了六种胸部邻近或心血管发现，并将权重重新分配至特定腹部病变，例如淋巴结肿大、腹水、脾肿大和肾囊肿。

表A1 训练集、IID 测试集和 OOD 测试集划分中主要解剖器官/区域类别的分布。数值为 n (%)。

器官/区域	训练	IID	OOD
肝脏/胆道树	33,483 (15.0%)	8,568 (15.1%)	248 (12.1%)
下胸部	33,427（15.0%）	8,414（14.8%）	0 (0.0%)
肾脏/输尿管	32,937（14.8%）	8,308（14.6%）	381（18.5%）
胃肠道	30,443（13.6%）	7,540（13.3%）	197（9.6%）
胆囊	24,295（10.9%）	6,041（10.6%）	143（7.0%）
腹膜腔	21,454（9.6%）	5,494（9.7%）	316（15.4%）
脉管系统	18,089 (8.1%)	4,740 (8.3%)	75 (3.6%)
剩余类别	29,158（13.1%）	7,720（13.6%）	695（33.8%）

表 A2 训练集、IID 测试集和 OOD 测试集划分中临床任务能力的分布。数值为 n （%）。每个问题对应一个能力标签；“其他”包含细粒度属性问题，百分比以各划分的总数为基准。

临床能力	训练集	IID	OOD
疾病鉴别	138,828（62.2%）	35,265（62.1%）	1,194（58.1%）
器官异常	47,836（21.4%）	12,178（21.4%）	769 (37.4%)
严重程度分级	10,521 (4.7%)	2,690 (4.7%)	59 (2.9%)
急诊分诊	3,963 (1.8%)	1,034 (1.8%)	0 (0.0%)
侧别评估	984 (0.4%)	259 (0.5%)	0 (0.0%)
其他	21,154 (9.5%)	5,399 (9.5%)	33（1.6%）

A.2 验证与质量控制

多读者验证与筛选。为确保标签保真度较高，我们的基于规则的流水线生成候选问题后，在发布前须通过严格的多读者验证关卡。如表 A4所示，两位基础大语言模型读者对经审核的报告单元达成了高度共识，在 *CT-Verdict-IID* 上的读者间一致性为 97.8%，在*CT-Verdict-OOD*上的一致性为 96.7%。基础读者投票分歧的条目会交由第三位裁决读者处理；任何未能获得三者中两者多数的条目都会被丢弃，而非重新标注。验证完成后，发布级筛选会移除仅含模板的陈述并平衡选项分布，最终形成包含 280,111 个问题的已发布语料库，用于 *CT-Verdict-IID*（223,286 个训练/验证问题和 56,825 个测试问题）。

人工审计与数据集划分完整性。为验证最终发布标签，我们对照源放射学报告，人工审计了*CT-Verdict-IID* 中的 400 个随机抽样条目，以及*CT-Verdict-OOD* 中的 200 个随机抽样条目。在 IID 样本中，全部 400 个条目均被确认为临床上可接受，且与源文本不存在任何事实矛盾。在 OOD 样本中，经第二轮整理后，抽取的 200 个条目也均被确认为与源报告事实一致；两个已发布基准数据集的审计准确率均达到 100%。最后，我们确认训练集划分与测试集划分之间不共享任何研究标识符、报告文本或图像文件，并且 OOD 队列仅用于零样本评估。

纯文本审计

为验证*CT-Verdict* 确实需要视觉解读，而非依赖问题表面捷径，并确认 Radev 的高性能并非仅仅源于记忆训练集语言先验，我们在完整的*CT-Verdict-IID* 测试集划分上开展了严格的无图像审计。在推理时，我们将视觉通路完全置零，保留问题文本和读出流程不变。

表 A3 训练集（150,333 个带疾病标签的项目）、IID 测试集（38,214 个项目）和 OOD 测试集（1,253 个项目）中，30 种标注疾病/发现的问题数量及占比；仅统计带疾病标签的问题。数值为 n （%）。

疾病 / 所见	训练	IID	OOD
肺不张	13,687（9.1%）	3,434（9.0%）	0 (0.0%)
胆结石	12,096（8.0%）	2,996（7.8%）	59（4.7%）
肾脏低密度影	9,594（6.4%）	2,370（6.2%）	110（8.8%）
胆管扩张	9,443（6.3%）	2,492（6.5%）	82 (6.5%)
胆囊术后缺如	9,392 (6.2%)	2,298 (6.0%)	33 (2.6%)
肠梗阻	9,063（6.0%）	2,240（5.9%）	35 (2.8%)
肾积水	8,990 (6.0%)	2,368 (6.2%)	78 (6.2%)
肾囊肿	7,418（4.9%）	1,867（4.9%）	123 (9.8%)
动脉粥样硬化	7,406（4.9%）	1,953（5.1%）	21 (1.7%)
胸腔积液	7,369（4.9%）	1,828（4.8%）	0 (0.0%)
全身性水肿	7,128（4.7%）	1,781（4.7%）	0 (0.0%)
转移性疾病	6,934（4.6%）	1,732（4.5%）	11 (0.9%)
腹腔游离气体	4,842（3.2%）	1,237（3.2%）	17 (1.4%)
腹主动脉瘤	4,549（3.0%）	1,250（3.3%）	17（1.4%）
肝脂肪变性	4,378（2.9%）	1,099（2.9%）	34（2.7%）
淋巴结肿大	3,829（2.5%）	1,051（2.8%）	155（12.4%）
腹水	3,323 (2.2%)	882 (2.3%)	102 (8.1%)
黏膜下水肿	2,746（1.8%）	711（1.9%）	57（4.5%）
胰腺萎缩	2,610（1.7%）	696 (1.8%)	60 (4.8%)
骨折	2,573 (1.7%)	674 (1.8%)	17 (1.4%)
阑尾炎	2,400 (1.6%)	556 (1.5%)	22 (1.8%)
冠状动脉钙化	1,766 (1.2%)	445 (1.2%)	0 (0.0%)
心脏肥大	1,709 (1.1%)	412 (1.1%)	0 (0.0%)
食管裂孔疝	1,487（1.0%）	363（0.9%）	10（0.8%）
脾肿大	1,439（1.0%）	434（1.1%）	96（7.7%）
骨量减少	1,323（0.9%）	371（1.0%）	9（0.7%）
肝肿大	1,249 (0.8%)	320 (0.8%)	44 (3.5%)
血栓形成	723 (0.5%)	155 (0.4%)	20 (1.6%)
前列腺肿大	666 (0.4%)	143 (0.4%)	41 (3.3%)
主动脉瓣钙化	201（0.1%）	56（0.1%）	0（0.0%）

禁用视觉输入后，所有判别性指标和临床分级指标均出现明显崩溃。在二元存在性和多项选择题上，基于概率的 mAUC 分别下降 22.67 和 22.63 个百分点（分别从 87.79% 降至 65.12%，以及从 81.91% 降至 59.28%），而平衡准确率（mBAcc）则从 71.79% 降至接近随机水平（53.32%）。在有序严重程度分级任务中，移除图像使平均绝对误差（MAE）从 0.392 增至 0.470，并使秩一致性（二次加权 kappa，QWK）从 0.537 减半至 0.264。对于无法根据常规报告统计数据推断的任务，性能下降更加明显，例如急诊分诊任务的准确率下降 45.36 个百分点（从 82.88% 降至 37.52%）。对前沿闭源模型（GPT-5.4）运行相同的无图像对照实验，也得到类似的崩溃结果：其有序严重程度 QWK 从 0.462 降至 0.105（MAE 则从 0.748 升至 0.983）。总体而言，这些结果证实，$CT\text{-}Verdict$ 同样无法仅凭文本先验解决，而 Radev 的性能源于真正的解剖视觉定位。

A.3 评估方案详情

分类指标（二分类与多项选择）。 我们报告每个问题的完全匹配准确率（**准确率**）、宏平均平衡准确率（**mBAcc**）以及宏平均 ROC 曲线下面积（**mAUC**）。对于每个疾病类别的问题，首先在一对其余的框架下分别计算 mBAcc 和 mAUC：mBAcc 为敏感度与特异度的平均值；mAUC 则使用预测概率（二分类使用“是”的概率，多项选择使用选项概率）与一对其余的标准标签进行计算。随后，分别在二分类的 30 个临床发现和多项选择的 26 种选项疾病之间进行宏平均，使罕见和常见的发现获得同等权重。总体 mBAcc 在宏平均之前，先对每个发现的两个分支取平均。相比之下，mAUC 按分支分别报告，因为两种概率尺度不可直接比较。

表 A4 经审核的发现存在性单元的多读者验证统计。

指标	CT-Verdict-IID	CT-Verdict-OOD
经审核的报告单元	82,836	613
基础读者共识	79,494 (96.0%)	567 (92.5%)
裁决单元	1,854	19
裁决保留	1,564	9
读者间一致性	97.8%	96.7%

一个报告单元是针对某项研究的一条疾病存在断言，并与其来源句子和规则推导的极性标签相关联；两位基础读者分别独立判断每个单元。

序数严重程度指标。 序数问题使用平均绝对误差（**MAE**）、归一化排序概率评分（**KRPS**）Epstein (1969）和二次加权卡帕（**QWK**）Cohen(1968）进行评估。KRPS 比较预测的累积等级分布与真实累积等级分布，

$$\mathrm{KRPS}_i \;=\; \frac{1}{K-1}\sum_{k=1}^{K-1}\big(\hat{F}_i(k)-F_i(k)\big)^2, \tag{A1}$$

对每个错误等级上的概率质量进行惩罚，而非仅惩罚最大值索引。对于缺乏累积结构的无序问题，不计算此项。二次加权 kappa 会根据边缘分布所预期的偶然一致性，对二次加权一致性进行校正，并为仅基于患病率的预测器赋予零分。

校准与 Murphy 分解。我们使用十个等宽分箱（去除空分箱），从两个互补层面评估校准：

- **决策层面：**最高标签期望校准误差（**期望校准误差**）Guo等人（2017）评估最大值索引预测的置信度：

$$\mathrm{ECE} \;=\; \sum_b \frac{n_b}{N}\,\big|\mathrm{acc}_b-\mathrm{conf}_b\big|, \tag{A2}$$

  其中，n_b 是分箱 b的大小，acc_b 和conf_b 分别是该分箱的准确率和最高标签置信度均值。

- **分布层面：**多分类布里尔评分（**BS**）Glenn等人（1950）是一种用于评估整个候选分布的严格适当评分规则：

$$\mathrm{BS} \;=\; \frac{1}{N}\sum_i\sum_k\big(p_{ik}-y_{ik}\big)^2, \tag{A3}$$

  其中，p_{ik} 是候选项 k 的预测概率，y_{ik} 是其真实独热指示变量。为区分校准不足与区分能力，我们将 Murphy 分解Murphy （1973）应用于其分箱形式：

$$\mathrm{BS} \;=\; \mathrm{REL}-\mathrm{RES}+\mathrm{UNC}, \tag{A4}$$

  其中三个项分别为

$$\begin{aligned}\mathrm{REL} &= \sum_b w_b\,(\bar{p}_b-\bar{y}_b)^2,\\ \mathrm{RES} &= \sum_b w_b\,(\bar{y}_b-\bar{y})^2,\\ \mathrm{UNC} &= \bar{y}\,(1-\bar{y}).\end{aligned} \tag{A5}$$

  可靠性 衡量纯粹的校准不足（置信度与准确率差值的平方，越低越好），**分辨率** 衡量区分能力（分箱准确率相对于基准率的离散程度，越高越好），而 **不确定性**则是独立于模型的数据内在不确定性。

分解恒等式对分箱评分严格成立。与其并列报告的原始逐行 BS 仅因箱内方差项而与之不同。对于多候选问题，我们针对每个候选项分别应用一对其余分解，这与多分类 BS 投影相符，而 ECE 针对的是最高标签决策。随后，将结果求和并按候选数量分箱加权，得到报告的 REL 和 RES。当且仅当RES >REL （布里尔技巧分数为正）时，模型才优于恒定流行率基线；这定义了图3中的盈亏平衡边界。

B 附加方法细节

B.1 CCT 的器官分层课程学习

由于多视图校准直接作用于消融后的槽位，均匀采样会导致罕见解剖结构的校准不足。因此，我们以概率 n_g 对 CT-Verdict 中大小为 n_g 的 17 个器官组分别进行采样

$$\pi_g \propto n_g^{\tau}, \qquad \tau = 0.5, \tag{A6}$$

在实例均匀分布与器官均匀分布之间进行插值。组配额至少占预算的0.5%，每组最多进行八次遍历，并在研究运行级别交错安排，从而即使在较短的 CCT 预算下（第 5.1 节），也能在所有器官之间保持均衡校准。

B.2 反事实置信度温度调节分析

我们提供了群体层面的分析，以说明根据公式（6）和公式（7）监督反事实器官视图，如何将预测置信度与解剖学证据而非捷径线索联系起来；在实践中，这种效应的大小取决于有限样本优化和模型平滑性，而对照消融的零信息条件是一种理想化设定，仅能近似成立，因为未绑定器官可能携带关于判定的残余信息。

命题1（目标消融下的后验温度调节）。令 $Y \in \mathcal{Y}(q)$ 表示问题q的参考判定。将事实器官账本划分为 $\mathbf{E} = (\mathbf{E}_T、\mathbf{E}_R)$，其中$\mathbf{E}_T$汇集目标消融所移除的所有组件：问题绑定的目标器官 $\mathcal{B}(q)$ 的账本令牌、其绑定的二维证据词元以及全局槽位 $n+1$，而$\mathbf{E}_R$ 表示其余未绑定器官及其证据。

令$\tilde{\mathbf{E}}$ 表示目标消融后的清单：将整个区块 $\mathbf{E}_T$ 替换为共享占位符序列 $\mathbf{a}$（每个受影响的证据令牌都简化为其干预后的台账令牌，见第 3.4节），同时保留 $\mathbf{E}_R$、并令 $\hat{\mathbf{E}}$ 表示对照消融后的清单：将一个无关器官 $\mathbf{E}_c \in \mathbf{E}_R$（其台账令牌及绑定证据）替换为 $\mathbf{a}$，同时保持 $\mathbf{E}_T$ 不变。

对于每个视图 $\mathbf{V} \in \{\mathbf{E}, \tilde{\mathbf{E}}, \hat{\mathbf{E}}\}$，式 ℓ（6）中的任何严格适当评分规则 的唯一总体风险最小化解都是条件分布 $p^*(\cdot \mid \mathbf{V}, q) = P(Y = \cdot \mid \mathbf{V}, q)$。此外，由于 $\mathbf{E}$ 是 $\mathbf{E}$的确定性粗化，这些条件分布的熵满足：

$$H(Y \mid \tilde{\mathbf{E}}, q) - H(Y \mid \mathbf{E}, q) = I(Y; \mathbf{E}_T \mid \mathbf{E}_R, q) \geq 0. \tag{A7}$$

因此：

1. 非冗余目标证据（$I(Y; E_T \mid E_R, q) > 0$）导致熵严格增加 $H(Y \mid E, q) > H(Y \mid E, q)$。

2. 非确定性残余上下文（$H(Y \mid E, q) > 0$）意味着 $\max_{m \in \mathcal{Y}(q)} P(Y = m \mid E, q) < 1$。

3. 对于对照消融，若 $I(Y; E_c \mid E^{\setminus c}, q) = 0$，则 $P(Y \mid E, q) = P(Y \mid E, q)$。

证明。我们分三步进行。

步骤 *1*：严格适当评分规则。对于每个固定视图 $\mathbf{V}$，交叉熵、多分类布里尔评分和排序概率得分都是严格适当评分规则，其条件风险 $\mathbb{E}_{Y \mid \mathbf{V}, q}[\ell(p, Y)]$ 在 $p^*(\cdot \mid \mathbf{V}, q) = P(Y = \cdot \mid \mathbf{V}, q)$ 处取得唯一最小值。

步骤 *2*：熵恒等式。注意，$\mathbf{E}$ 是 $\mathbf{E}$ 的确定性函数，因此 $(\mathbf{E}, \mathbf{E}, q) = H(Y \mid \mathbf{E}, q)$ 和 $H(Y \mid \mathbf{E}, q) \geq H(Y \mid \mathbf{E}, q)$；因此

$$H(Y \mid \tilde{\mathbf{E}}, q) - H(Y \mid \mathbf{E}, q) = I(Y; \mathbf{E} \mid \tilde{\mathbf{E}}, q).$$

由于 $\mathbf{E}$ 恰好保留了 $\mathbf{E}_R$，根据链式法则可得 $I(Y; \mathbf{E} \mid \mathbf{E}, q) = I(Y; \mathbf{E}_T \mid \mathbf{E}_R, q)$，得到公式（A7）。

步骤 *3*：推论。条件互信息严格为正会导致熵严格增加，而有限标签集上的熵严格为正意味着条件分布不是独热分布，从而确保 $\max_m P(Y = m \mid \tilde{\mathbf{E}}, q) < 1$。最后，$I(Y; \mathbf{E}_c \mid \mathbf{E}^{\setminus c}, q) = 0$ 直接意味着 $P(Y \mid \hat{\mathbf{E}}, q) = P(Y \mid \mathbf{E}, q)$。□

注 1（平滑性假设下向事实校准的迁移）。假设共享参数读出 $p_\theta(c \mid \mathbf{V}\ q)$ 关于完整消融块 $\mathbf{E}_T$（账本令牌、有界二维证据和全局槽位）在范数 $\|\cdot\|$ 下是 L-Lipschitz 的，因此

$$|p_\theta(c \mid \mathbf{E}, q) - p_\theta(c \mid \tilde{\mathbf{E}}, q)| \le L\|\mathbf{E}_T - \mathbf{a}\|.$$

˜当针对目标消融视图进行训练并近似达到其总体极小化解 $p_\theta(c \mid \mathbf{E}\ q)\ \approx P(Y = c \mid \mathbf{E}_R,\ q)$ 时，结合这两个关系可得

$$p_\theta(c \mid \mathbf{E}, q) \lesssim P(Y = c \mid \mathbf{E}_R, q) + L\|\mathbf{E}_T - \mathbf{a}\|, \tag{A8}$$

误差至多为有限样本优化误差。

解释。 在读出层满足平滑性的条件下，对事实视图和目标消融视图进行联合监督，会使事实置信度与经校准的消融后验相耦合：当事实目标表征接近器官缺失基线时，事实置信度就不易大幅超过非目标上下文所支持的置信度（表7）。

互补机制。 当隐去所有图像衍生特征会使决策分数降低时（$r > \check{r}$），式（7）中的证据盲铰链项会促使 $\check{c} \le \mathrm{sg}[c] - m_{\text{blind}}$。由于 $\mathrm{sg}[\cdot]$ 会阻断经由c的梯度，这一项会直接降低盲置信度，使其低于当前事实置信度；并通过共享参数，间接抑制事实输入上独立于证据、仅由查询决定的置信度。

B.3 视觉编码器中的二维-三维特征融合

二维通道和三维通道通过确定性的三维空间归属规则共享器官身份，而不是通过学习式对齐共享。设填充后的体积平面为 $w_0 \times h_0$，切片 i 携带体积索引 s_i。二维塔为每个切片输出一个图块网格 $(1,\ h_p,\ w_p)$。以因子 m进行空间合并后，合并单元 $(r,\ c)$（其中$0 \le r < h_p/m$ 和 $0 \le c < w_p/m$）的中心被反向投影至原始切片坐标：

$$y_r = \left(r + \tfrac{1}{2}\right) m\, w_0/h_p, \qquad x_c = \left(c + \tfrac{1}{2}\right) m\, h_0/w_p. \tag{A9}$$

将器官框写为 $R_k = (z_0,\ z_1,\ y_0,\ y_1,\ x_0,\ x_1)$，采用包含式边界，其中全为-−1 的行表示缺失器官。当且仅当以下条件成立时，单元属于器官 k

$$\begin{aligned} M_k(i, r, c) = \mathbb{I}[z_0 \le s_i \le z_1]\ \mathbb{I}[y_0 \le y_r \le y_1] \\ \times\ \mathbb{I}[x_0 \le x_c \le x_1]\,. \end{aligned} \tag{A10}$$

因此，切片选择（$s_i \in [z_0, z_1]$）和平面内框测试处于同一坐标系中，无需配准。命中集合 $\mathbf{S}_k = \{\mathbf{ie}_{i,\ r,\ c} :\ M_k(i,\ r,\ c)\ = 1\}$ 构成公式（3）中的精确键值集合。具体而言，几何关系决定 哪些切片令牌表示器官 k，而以查询 $\mathbf{h}_k$ 进行的残差交叉注意力则决定其权重。当 $\mathbf{S}_k$ 为空时（例如采样切片未覆盖 z-范围、框位于平面之外或器官缺失），证据令牌回退至 $\mathbf{h}_k$ 自身，仅增加零信息，而不会虚构某个区域。在同一掩码上进行均值池化，可得到供器官残差变体使用的无参数器官级二维摘要。两个使用者共享这一坐标约定。

C 补充实验结果

C.1 定性案例研究

图 A2 展示了模型在预测正确和错误时的行为。在正确案例中，Radev-2B始终给出明确的置信度（在(a)中，对一项未发现胆结石的IID研究，置信度为0.994；在(b)中，对一项发现肾囊肿的外部 AMOS 研究，置信度为0.993），这表明基于器官的证据有助于模型在分布内和分布偏移设置下都做出可靠的判定。相比之下，基线模型要么会幻觉出阳性发现，要么在接近随机水平时仍无法提供有用信息（例如，Jev-1.14的得分为0.540）。关键的是，在一个共同的失败案例中



图 A2 CT-Verdict-IID 和 CT-Verdict-OOD 上的定性判定。每一行展示一张轴位切片、包含标准标签的完整数据集问题，以及 GPT-5.4（自由文本）、Jev-1.14（似然头）和 Radev-2B（我们的方法）的判定。行 (a) 和 (b) 分别是 Radev 在独立同分布和分布外数据上的成功案例；行 (c) 则是三个模型共同失败的案例，在该案例中，Radev 给出的置信度最低。

三种模型都漏检了阳性肾囊肿（c），但只有Radev-2B的置信度骤降至0.062，实际上将自身的错误标记为不可靠。与此同时，GPT-5.4未能提供可用得分，而Jev-1.14仍无法提供有用信息，得分为0.540，因此这两种模型的错误都无法在后续环节被拦截。这表明Radev-2B不仅总体上出错更少，而且会将错误显著暴露出来：简单的置信度阈值就能在错误案例(c)中触发弃权，同时保留(a)和(b)中的正确判定，通过选择性预测将隐蔽的失败转化为可拒绝的预测。