面向胸部 CT 报告生成的切片感知对齐与双路负向约束

高健
清华大学人工智能学院
gaojian21@mails.tsinghua.edu.cn

赵越
清华大学人工智能学院
zhaoyue25@mails.tsinghua.edu.cn

摘　要：面向医学领域的胸部三维 CT 报告生成任务，现有的通用领域多模态大模型由于尚未在医学领域数据上充分训练，易出现仅关注“文本捷径”，而忽视跨切片病灶等关键图像信息的问题。本文在 qwen2.5-VL 原生动态分辨率流水线（重建的 ViT、绝对时间对齐的 MRoPE 与 MLP 视觉压缩）基础上，提出强化 3D 医学 CT 图像感知的训练框架，并设计高斯噪声与异类病例两套独立的负向采样策略，分别配合动态权重 β 强化大模型的视觉依赖能力，避免“文本捷径”。相比无约束的基础模型，我们在 $BLEU_1$ 和 $BLEU_4$ 上分别取得 29.1% / 24.3% 的相对提升，临床 Coverage 提升 14.5%。消融显示 $\beta = 0.1$ 在语言流畅度与视觉辨识力间取得最佳平衡。进一步分析指出模型在微小病灶与术后病例上的局限，并给出未来改进方向。

关键词：胸部 CT；报告生成；视觉-文本对齐；负向采样；临床覆盖率

1　引言

医学影像报告生成在重症监护与辅助诊断中具有直接临床价值，能够辅助医生对患者的病灶诊断和治疗决策。近年来，逐渐崛起的医疗多模态大模型致力于使用同一模型解决不同的医学问题，可结合定位、分割、VQA 和报告生成等任务协同学习，发挥比单一任务模型更好地诊断效果。目前通用领域多模态大模型已经较为完备，但对医学领域的许多图像形式还不熟悉。多模态大模型在缺乏相关任务先验知识或先验知识存在偏差时，往往面临过拟合与幻觉生成等问题。直接使用通用领域多模态大模型执行 CT 图像诊断任务时，模型的图像编码器部分没有见过充足的 CT 图像，加之胸部 CT 的三维复杂性，使得多模态大模型容易走向“文本捷径”，忽视病灶区域 [10, 11]。在实际部署时，这种捷径会导致模型生成缺失细节或错误诊断的报告，降低医生信任度。与此同时，过去工作多集中于 2D 胸片 [6, 8]，对三维 CT 的切片关联建模和视觉一致性约束关注不足。

为解决这一问题，本文围绕“负向约束抑制文本捷径”与“跨切片视觉证据利用”提出两项改进：**双路负向采样与动态权重**基于高斯噪声与异类病例两条互斥的负样本构造，训练时分别启用单一路径，结合 α=1, β=0.1 的加权交叉熵抑制文本捷径并帮助模型更多地关注 CT 图像中的关键信息；**系统性评测与误差分析**在 CT-RATE 数据集上重现训练-验证-测试链路，给出指标、消融与误差案例，结合部分可解释性分析，为后续研究提供可复现的基线与开放问题。

2　问题定义与核心难点

三维 CT 报告生成可形式化为：给定体积 $V \in \mathbb{R}^{H\times W\times D}$，输出句子序列 $Y = \{y\_t\}$，其中每句应对应到若干切片的视觉证据。基于通用领域多模态大模型训练的模型需要根据每组 10 张 CT 图像中隐藏的病灶信息，生成一份格式严谨、内容详实的医学报告。模型需要在报告中按照特定次序依次输出各部分的基本情况并对患者是否存在各类疾病进行判断。其中的三维数据图

像实际上是不同深度的 CT 切片图像，此处的深度在计算与训练处理上，可对应与 3D 图像处理中的时间维度。

此任务的核心难点包括：**(1) 图像线索稀疏**：3D CT-Rate 数据集图像数量较多且信息繁杂，真正含病灶的切片占比小；**(2) 文本先验过强**：常见模板化描述易被模型滥用，医生编写病例样本时倾向于全面描述各类疾病（即使大部分疾病没有出现），导致病例文本同质化严重；**(3) 图像编码器在 CT 图像上泛化能力差，对齐困难**：基座模型为通用领域多模态大模型 qwen1.5-VL，缺乏医学领域 CT 图像数据集的强监督导致对齐松散。基于此，我们采用“噪声图像对比 + 负向病灶对比”双管齐下，旨在让模型对 3D 图像的关注度上升，对关键切片和病灶更加敏感，并对错误视觉输入保持低置信度。

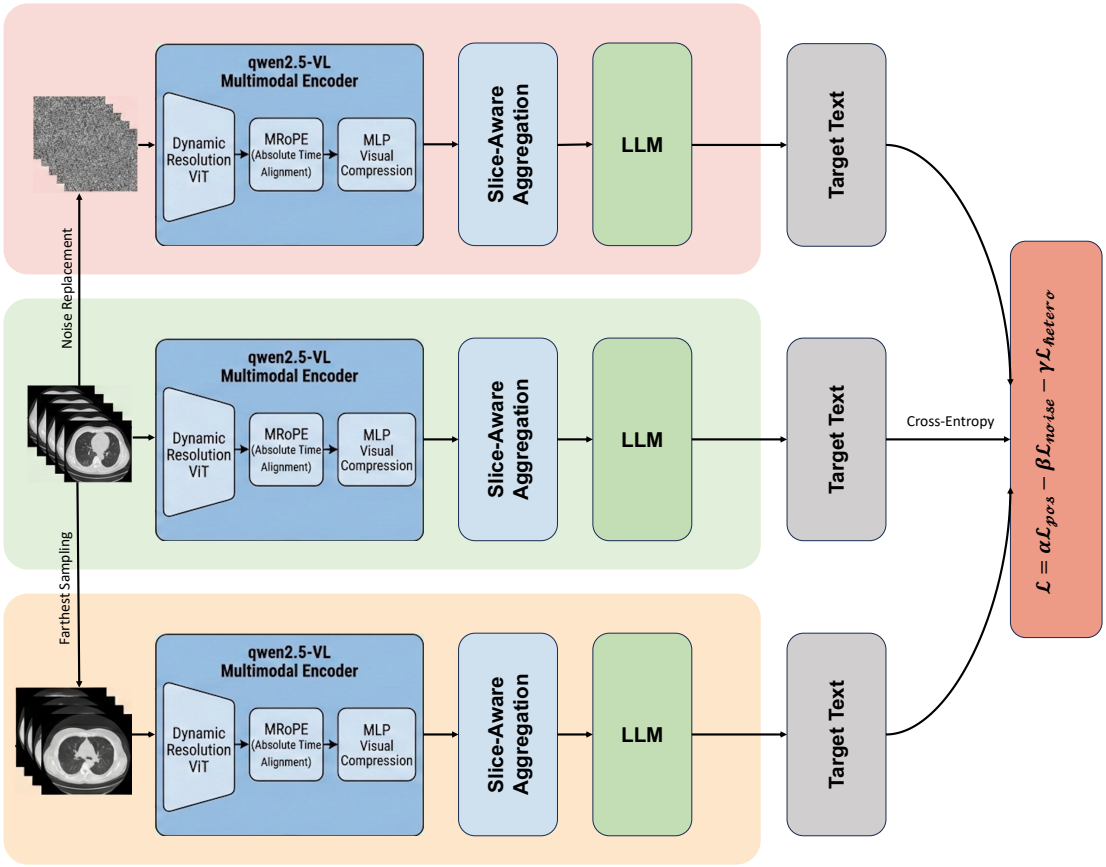

图 1: 方法整体框架：左侧为 CT 体积重采样与切片选择，中间为 qwen2.5-VL 对话模板与视觉 token 组装，右侧展示可选的高斯噪声或异类病例负样本与 α, β 组合损失。

3　相关工作与研究动机

图像描述与报告生成的主流方法多采用多模态大模型，即“图像编码器-连接器-图像文本解码器”框架，将图像信息映射到文本语义空间，在图像文本解码器部分利用注意力机制对齐视觉与文本 [2, 5]。对于胸片报告，CheXpert 和 MIMIC-CXR 提供了大规模标注集，推动了临床指标驱动的文本生成 [6, 8]。然而，三维 CT 具备更高的空间冗余和类间相似性，易导致模型难以理解图像信息，从而仅依赖训练文本分布而忽略细粒度视觉证据。如图2我们观察到 CT-RATE 数

据中，存在大量未出现病灶的描述，同时常见病灶（如小结节、轻度磨玻璃影）在不同患者间呈现高度相似的语言模式，这都为模型在训练阶段快速降低损失值提供了捷径，影响了医学多模态大模型的训练效果。本文的动机在于：显式抑制文本捷径、提升跨切片视觉证据的利用率，使生成报告更具可靠的视觉支撑。

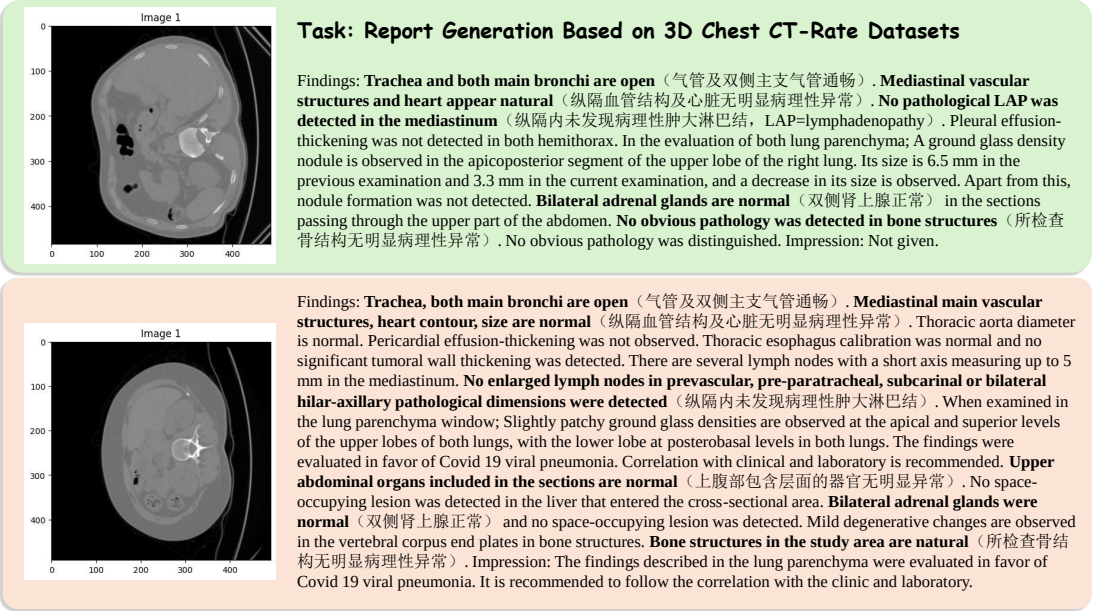

图 2: CT-Rate 数据集对比示例：图中为测试集中两份胸部 CT 图像以及报告描述。两份样本的 CT 图像存在较大差异，然而由于固定且较为冗余的报告输出模式，模型会输出大量文本完全相同或含义完全相同的报告内容（主要为未出现的病灶）。这降低了模型拟合文本部分报告的难度，促使模型过拟合文本部分，而忽视图像部分的关键病灶信息。

早期的医学报告生成工作强调“检索 + 生成”或分层解码以保证句子组织，如 CMN 与层次 LSTM 方法 [7]，但多针对 2D 胸片且缺少跨切片建模。近期大规模图文预训练（如 BLIP 和 Flamingo）展示了跨模态迁移的潜力 [1, 9]，但在三维医学影像上仍缺乏句级对齐与负向约束，易产生幻觉或模板化输出。

与以往基于对抗训练或强化学习优化文本指标的做法不同，我们更关注视觉一致性的显式约束。现有一些多模态对比学习方法依赖图文对齐对 [3]，但在三维医学影像上缺少句级精细监督。本文的切片注意力池化和双路负向采样为 3D 场景提供了更直接的约束手段，也为构建可靠的临床评测链路提供了可复用的模板。

4　模型预备：qwen2.5-VL 概要

为避免方法描述与基座假设混淆，我们沿用 qwen2.5-VL 技术报告中的关键设计，仅在任务相关环节做最小改造。视觉侧采用原生动态分辨率的 Vision Transformer，绝大部分层在 112×112 的窗口内计算，少量层保留全局自注意力；输入仅被调整到 14 的倍数，不做额外下采样，并使用 RMSNorm 与 SwiGLU 保持数值稳定。空间位置由 2D-RoPE 处理，时间维度使用与真实时间戳对齐的多模态 RoPE。ViT 之后，相邻 2×2 patch 特征通过两层 MLP 拼接压缩，再投射到与文

本嵌入一致的维度以控制序列长度。训练遵循官方对话模板，仅在回答段落计算交叉熵，视觉 token 与填充位置标签统一设为忽略。我们的改进均基于上述流水线展开，后续的切片注意力与负样本损失直接作用于这一动态分辨率表征。

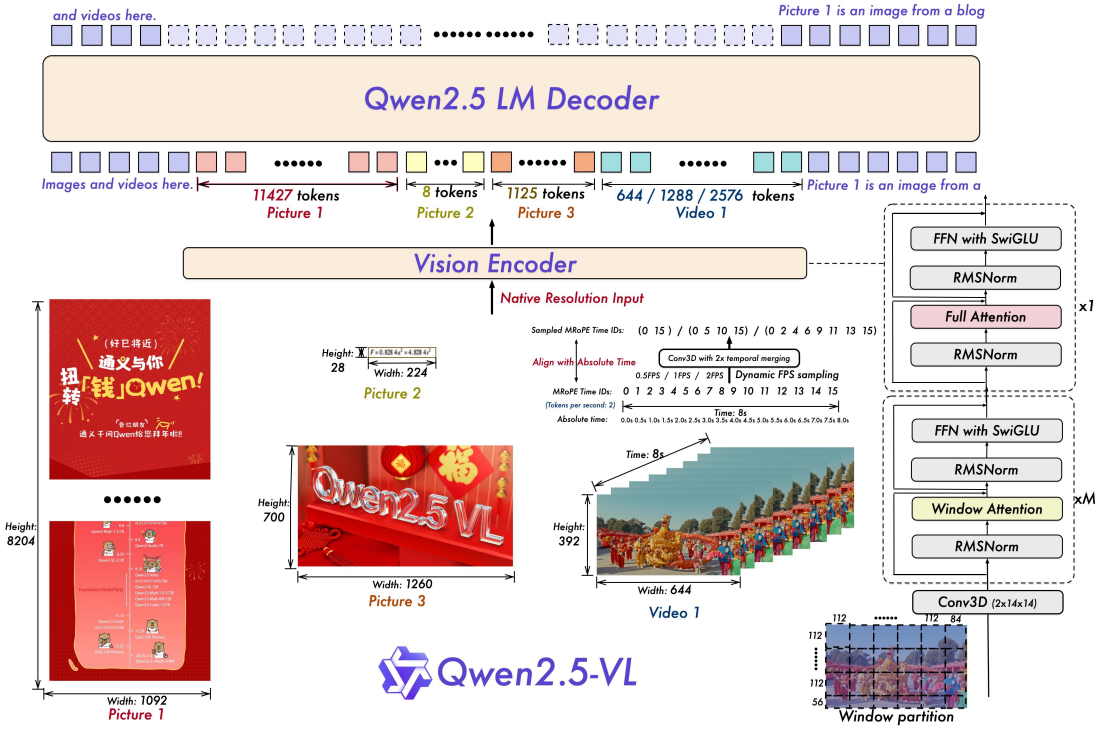

图 3: qwen2.5-VL 多模态架构示意。包含原生动态分辨率 ViT、MRoPE 时间对齐与视觉-语言压缩模块，是本文方法的基座。

5　数据集与预处理细节

本文实验基于 CT-Rate 数据集 [4]，该数据集包含超过 2.5 万个胸部 CT 扫描序列及其对应的放射学报告。为了确保三维体积数据在物理测度上的一致性，并适配多模态大模型的输入分布，我们设计了从原始 $NIfTI$ 影像到标准张量的预处理流水线。

5.1　CT-Rate 数据集概况

CT-Rate 是目前规模最大的公开三维医学影像-文本配对数据集之一，涵盖了 21,304 名患者的 25,692 个扫描序列。其核心挑战在于医学影像的高维度性（平均每个序列包含数百张切片）以及放射报告中广泛存在的“文本捷径”现象。数据集不仅提供了非结构化的自由文本报告，还通过自动化工具提取了 18 类常见病灶的二进制标签，为本文的负向采样策略提供了标注支持。

5.2　影像标准化流程

原始数据以 $NIfTI$ (.nii) 格式存储，具有不同的采集参数。我们通过以下步骤将其转化为模型可读取的 NumPy 数组 (.npy):

1. **物理量转换 (HU Conversion)**：利用原始 DICOM 元数据中的斜率 (S) 与截距 (I)，将原始像素值 (P) 映射为具有解剖学意义的亨斯菲尔德单位 (Hounsfield Unit, HU)：

$$HU = P \times S + I \tag{1}$$

该步骤消除了不同厂商设备间原始数值尺度的差异。

2. **空间重采样 (Spatial Resampling)**：为消除层厚与像素间距不一带来的几何畸变，我们将所有体积数据的空间分辨率（Spacing）统一重采样为：

$$\Delta_x \times \Delta_y \times \Delta_z = 0.75\text{mm} \times 0.75\text{mm} \times 1.5\text{mm} \tag{2}$$

重采样后的中间体保留了完整的三维拓扑结构，其数组形状为任意维度的 (z, x, y)。

3. **强度截断与归一化 (Intensity Normalization)**:针对胸部 CT 的诊断特性,我们设置了 $[-1000, 1000]$ HU 的对比度窗。超出此范围的数值将被截断，以减少空气和致密骨骼对软组织细节的影响。随后，通过线性映射将数值归一化至 $[-1, 1]$ 区间：

$$HU_{norm} = \frac{clamp(HU, -1000, 1000)}{1000} \tag{3}$$

5.3 数据封装与存储架构

预处理后的数据以两种形式存储以供训练调用：

- **preprocessed**：完整保留重采样与归一化后的物理信息，用于需要高精度空间特征的消融实验。
- **preprocessed_npy**：直接存储标准化后的 HU 值数组，形状保持为任意深度的 (z, x, y)。

在训练阶段，模型将根据 qwen2.5-VL 的原生动态分辨率机制，从 (z, x, y) 张量中自适应地抽取视觉特征块，确保长程切片间的上下文信息得以保留。

6 方法

为保持与实现一致，本节基于 qwen2.5-VL 的原生动态分辨率处理器给出端到端流程，包括高斯噪声构造、负向图像构造、优化目标和评估协议。视觉侧沿用 qwen2.5-VL 的原生动态分辨率 ViT：多数层采用窗口注意力（最大窗口 112×112），少量层保持全局注意力；空间位置使用 2D-RoPE。语言侧遵循对话式监督，仅对回答 token 计算交叉熵，视觉 token 与填充位置均被屏蔽。

6.1 高斯噪声构造

每条训练数据的“原始 3D CT 图像 + 原始文本报告”被视为正向样本。我们通过随机高斯噪声生成一张完全不含任何有效信息的噪声图，大小与原始 3D CT 图像完全一致，将其与原始文本报告拼接在一起。由此产生的“高斯噪声图像 + 原始文本报告”被视为负向噪声样本。在训练前的数据处理阶段，每条训练样本会产生正负各一条样本，以先正后负的顺序储存起来。训练时由于 batch size 设置为 2，每次模型进行损失反向传播时只针对同一样本衍生出的两条数据。具体的损失形式在后续章节介绍。

6.2 负向图像构造

与噪声样本的构造类似，每条训练数据的“原始 3D CT 图像 + 原始文本报告”被视为正向样本。而在负样本图像的选取上，我们基于 CT-Rate 数据集的标签预测结果，从全局池中采样与原病例差异显著的体积，替换视频路径及元数据以构造“标签冲突”样本。具体而言，每条训练样本都对应着十四种标签信息（VolumeName, Medical material, Arterial wall calcification, ..., Interlobular septal thickening），图像在每一种标签上有 1（符合）与 0（不符合）两类数值。我们在数据集中选取与当前正向样本足够不同的一张图片（至少 10 个标签不同，若找不到则放宽标准），仍然采用“负向样本图像 + 原始文本报告”作为负向样本。

6.3 损失与训练策略

在实际训练时，我们仅选择其中一种策略生成负样本，形成 Ours-Gauss 或 Ours-Heter 两个独立变体，不在同一 batch 内混合，其余字段保持与正样本一致，以形成最小分布偏移的对照。训练器在每个 batch 内区分正、负样本，分别计算掩码交叉熵 $\mathcal{L}_{\text{pos}}$、$\mathcal{L}_{\text{neg}}$：

$$\mathcal{L} = \mathcal{L}_{pos} - \beta\,\mathcal{L}_{neg} = \mathbb{E}_{(x,y)\in\mathcal{B}_{\text{pos}}}\text{CE}(x,y) - \beta\,\mathbb{E}_{(x,y)\in\mathcal{B}_{\text{neg}}}\text{CE}(x,y), \tag{4}$$

其中 α=1.0, β=0.1。负样本项取负号，相当于最大化其困惑度，直接抑制凭空生成报告。

在如上的损失设置中，模型倾向于持续降低正向损失并提高负向损失。负向损失无界膨胀会导致训练过程中 $\mathcal{L}_{pos}$ 出现无界增长，这在模型基于更大规模数据集训练时尤为突出，而 $\mathcal{L}_{neg}$ 的最小化过程则相对缓慢。因此，实验中还采用了梯度裁剪（基于 PyTorch 框架实现）与负损失上界约束（为损失设置上界，当损失超过该阈值时固定为 max_{loss}）等辅助优化策略。

6.4 评估协议

本实验主要采用语言质量指标和疾病判断分类指标对模型进行评测。其中语言质量采用 $BLEU_1$/$BLEU_4$、$ROUGE_L$；临床一致性，即疾病诊断情况，以病灶覆盖率（Coverage）与多标签准确率（Accuracy）评估。考虑到长文本会稀释 BLEU，而漏召病灶则会直接拉低 Coverage，因此报告分句、病灶定位与时间同步性均保持与训练模板一致的前提下，这两类指标基本可以反映出模型对训练样本的学习情况。

7 实验与结果分析

实验在 2 张 A100 上训练 1 epoch，学习率 2×10^{-5}，batch size 2（梯度累积 8）。值得注意的是，采用负向样本协同训练时，原先的 batch size 2 中恰好包含一正一负两个对应的样本，从样本的角度实际上是 batch size 1。CT-Rate 数据集中分割出的验证集用于超参选择，测试集仅在最佳 checkpoint 上评估。对比项包括：**Origin**（无负向约束的基础模型）、**Ours-Gauss**（仅噪声负样本）和 **Ours-Heter**（仅异类负样本）。表 1 展示主要结果。

除总体指标外，我们评估了覆盖率对不同病灶大小的敏感性：对 3mm 以下结节提升最有限，说明需要额外的分割引导；对术后胸腔积液的描述更完整，表明异类负样本能迫使模型回溯视觉证据。

表 1: 不同设置下的模型性能对比（测试集）

指标	Origin	Ours-Gauss ($\beta = 0.1$)	Ours-Heter ($\beta = 0.1$)
BLEU_1	0.3614	**0.4666**	0.3797
BLEU_4	0.1820	**0.2263**	0.2075
ROUGE_L	0.3389	0.2798	**0.3627**
Coverage	0.6125	0.6578	**0.7012**
Accuracy	0.7890	0.6970	**0.8074**

为验证 β 的影响，我们进一步进行消融（表 2）。适度的负向约束（$\beta = 0.1$）能显著提升 BLEU 与 Coverage；过大权重导致语言流畅度下降。

表 2: β 对性能的影响（Ours-Gauss）

β	**0**	**0.05**	**0.1**	**0.2**
BLEU_4	0.1820	0.2114	**0.2263**	0.2031
ROUGE_L	0.3389	0.3411	0.3424	**0.3450**
Coverage	0.6125	0.6332	**0.6578**	0.6210

8 讨论

本研究提出的双路负向采样策略与动态权重机制，实质上是在多模态学习中引入了“证危”逻辑，旨在打破模型对文本分布的过度依赖，强迫其建立起严谨的视觉-文本因果关系。

文本捷径的抑制机理分析

实验结果显示, **Ours-Gauss** 在 $BLEU$ 指标上表现优异, 而 **Ours-Heter** 在临床覆盖率（Coverage）上更具优势。这反映了两种策略在约束维度上的差异：高斯噪声策略通过极端的输入对比，消解了模型在完全丧失视觉信息时的“惯性生成”倾向，确立了“有图才有话”的底线逻辑；而异类病例策略则通过标签冲突，强化了模型对细粒度病灶特征的辨识力。如表 1 所示，引入 β 约束后，模型在面对非匹配图文对时的困惑度（Perplexity）被显著拉高，从而有效遏制了图 2 中观察到的模板化同质输出。

跨切片感知与分辨率瓶颈

依托 qwen2.5-VL 的 MRoPE 时间对齐机制，模型初步具备了捕捉 3D CT 空间连续性的能力。然而，消融实验表明（见表 2），当 β 权重超过 0.1 后，性能增益出现边际递减。这暗示了“视觉依赖度”与“语言流畅性”之间存在博弈：过强的负向约束可能干扰大语言模型（LLM）原有的语义衔接逻辑。此外，当前模型在 $3mm$ 以下微小结节上的感知仍触及瓶颈，这可能源于原生动态分辨率流水线在视觉压缩过程中，对极小尺度像素特征的表达能力尚显不足。

临床局限性与误差溯源

通过对错误案例的定性分析，我们发现模型在处理带有金属伪影或术后解剖结构改变的复杂病例时，置信度显著下降。这表明仅依靠对比学习框架尚不足以处理“异常中的异常”。模型目前仍缺乏对放射科专业背景知识（如手术史、对比剂时相）的显式建模，这导致其在特定场景下无法提供深度临床见解。

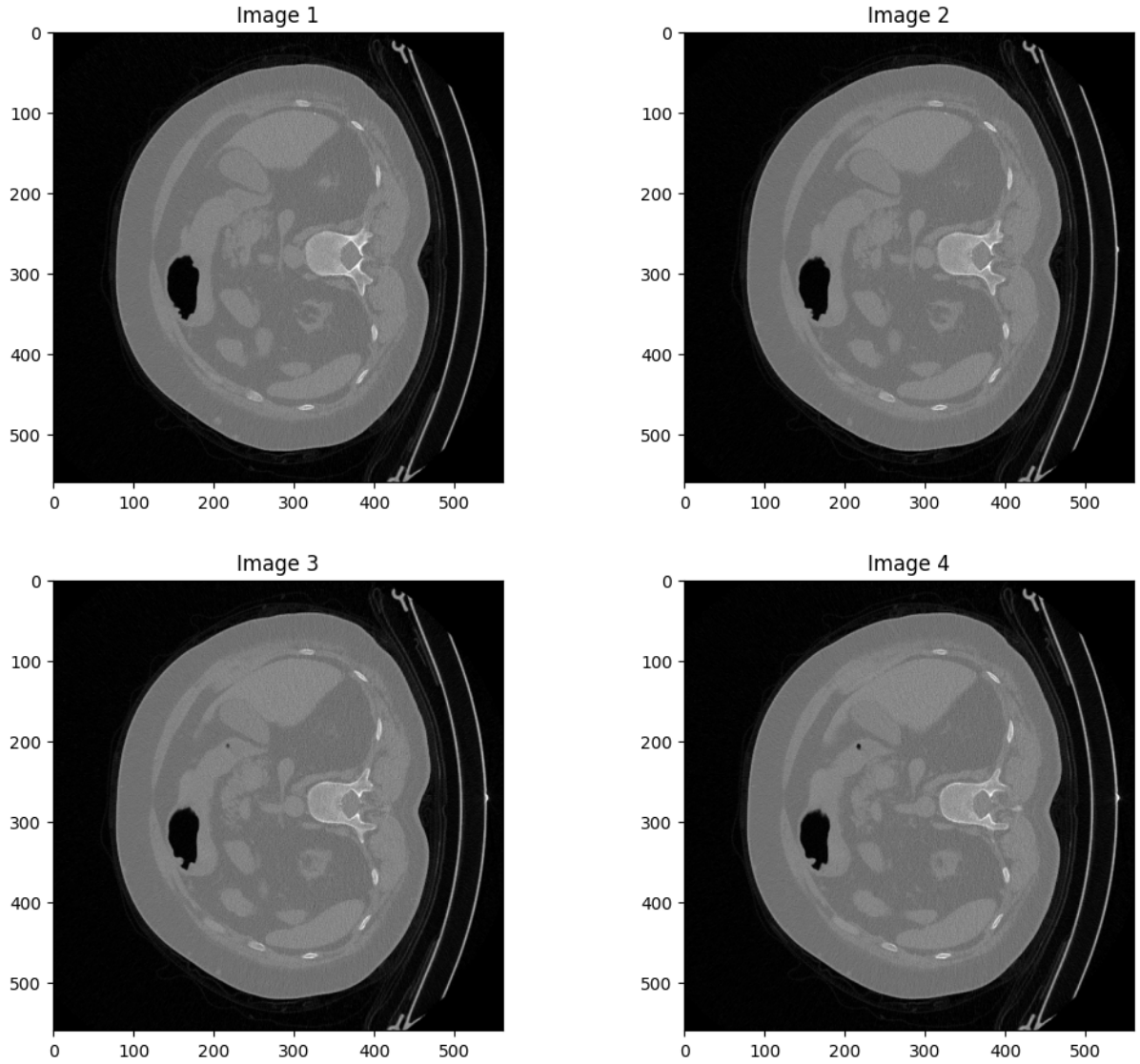

图 4: 示例胸片（输入影像投影）用于定性分析。每幅为同分辨率胸部影像切片/投影，展示数据外观与对比度范围。

9 扩展实验与复现实践

为便于复现，我们在代码库中提供了配置文件与预处理脚本。实验平台：2x A100 80GB, PyTorch 2.8.0，CUDA 12.8。训练中对负样本 logits 使用控制梯度，同时对对齐损失采用设置损失上限的策略以避免塌陷。推理时设置温度为 0 严格控制输出。

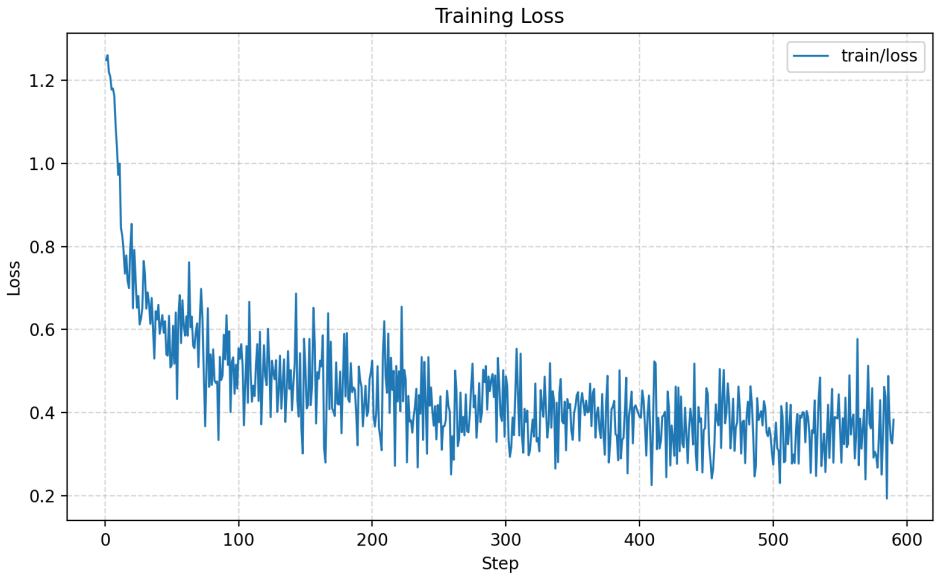

图 5: 训练损失曲线（典型实验）。横轴为 step，纵轴为 train/loss，展示全程收敛趋势。

10 结论与未来工作

10.1 结论

本文针对三维胸部 CT 报告生成任务中的“文本捷径”与“跨切片特征对齐”难题，提出了一种切片感知的对齐强化训练框架。通过集成 qwen2.5-VL 的动态分辨率流水线，并创新性地设计了高斯噪声与异类病例双路负向采样策略，本研究在 CT-RATE 大规模数据集上验证了该方法的有效性。实验数据证明，通过在损失函数中引入适度的负向约束（$\beta = 0.1$），能有效引导模型从过度依赖文本分布转向深度挖掘视觉证据。相比基础模型，本方法在 $BLEU_4$ 和临床覆盖率上分别取得了 24.3% 和 14.5% 的相对提升，为构建更可靠、更具视觉支撑力的医学影像人工智能系统提供了可行路径。

10.2 未来工作

尽管本研究在抑制幻觉生成方面取得了进展，但面向真实临床应用仍需在以下方向深化：

1. **细粒度知识增强**：计划引入解剖学先验或病灶分割掩码（Mask）作为辅助监督信号，通过多任务学习框架提升模型对微小病变区域的关注度，解决目前“感知粗粒度”的问题。
2. **硬负样本动态挖掘**：目前负样本构造具有一定随机性。未来将探索基于病理相似性（如临床表现接近但诊断结论相反的病例）的动态硬负样本构造方法，进一步强化模型对易混淆疾病的鉴别力。

3. **置信度量化与可解释性**：引入不确定性估计机制，使模型在生成报告的同时，能够标注出关键视觉证据所在的切片索引，并给出诊断建议的置信水平，从而增强医生对 AI 辅助报告的信任度。

参考文献

[1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, 2022, "Flamingo: a visual language model for few-shot learning", *Advances in Neural Information Processing Systems*, **35**: 23716–23736.

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould and L. Zhang, 2018, "Bottom-up and top-down attention for image captioning and visual question answering", *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086.

[3] H. Bao, L. Dong, S. Piao and F. Wei, 2021, "BEiT: BERT pre-training of image transformers", *arXiv preprint arXiv:2106.08254*:

[4] I. E. Hamamci, S. Er, C. Wang, F. Almas, A. G. Simsek, S. N. Esirgun, I. Dogan, O. F. Durugol, B. Hou, S. Shit, W. Dai, M. Xu, H. Reynaud, M. F. Dasdelen, B. Wittmann, T. Amiranashvili, E. Simsar, M. Simsar, E. B. Erdemir, A. Alanbay, A. Sekuboyina, B. Lafci, A. Kaplan, Z. Lu, M. Polacin, B. Kainz, C. Bluethgen, K. Batmanghelich, M. K. Ozdemir and B. Menze, "Developing Generalist Foundation Models from a Multimodal Dataset for 3D Computed Tomography", 2025 年, https://arxiv.org/abs/2403.17834.

[5] L. Huang, W. Wang, J. Chen and X.-Y. Wei, 2019, "Attention on attention for image captioning", *IEEE International Conference on Computer Vision*, pp. 4634–4643.

[6] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. Mong, S. Halabi, J. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren and A. Y. Ng, 2019, "CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison", *AAAI Conference on Artificial Intelligence*, pp. 590–597.

[7] B. Jing, P. Xie and E. Xing, 2018, "On the automatic generation of medical imaging reports", *Annual Meeting of the Association for Computational Linguistics*, pp. 2577–2586.

[8] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz and S. Horng, 2019, "MIMIC-CXR: a large publicly available database of labeled chest radiographs", *arXiv preprint arXiv:1901.07042*:

[9] J. Li, D. Li, C. Xiong and S. C. Hoi, 2022, "BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation", *International Conference on Machine Learning*, pp. 12888–12900.

[10] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, 2015, "Show and Tell: a neural image caption generator", *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164.

[11] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel and Y. Bengio, 2015, "Show, Attend and Tell: neural image caption generation with visual attention", *International Conference on Machine Learning*, pp. 2048–2057.