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　赵越

清华大学人工智能学院

2025 年 12 月 24 日

任务背景

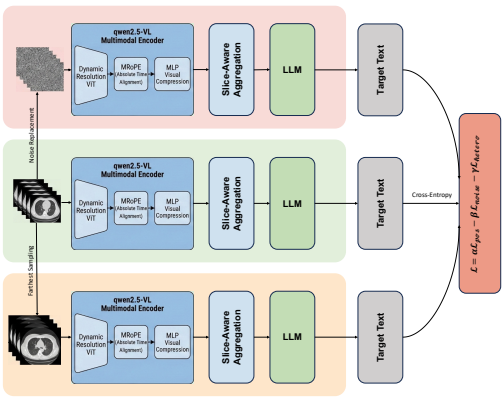
- 面向胸部三维 CT 的自动报告生成，需对跨切片病灶给出有依据的描述。
- 通用多模态大模型易走“文本捷径”，忽视稀疏的视觉线索，导致幻觉或模板化输出。
- 目标：提升模型对关键切片的依赖度，抑制无视觉证据的生成。

核心挑战

- 图像线索稀疏：有效病灶切片比例低，跨切片关联弱。
- 文本先验过强：报告模板同质化，易被模型滥用。
- 视觉对齐困难：基座 ViT 在医学 CT 上经验不足，时间/空间对齐不稳。

方法概览

- 基于 qwen2.5-VL 动态分辨率流水线，保持视觉-语言压缩与 MRoPE 时间对齐。
- 增加切片注意力池化与两类负向采样，显式强化视觉依赖。
- 以最小改动接入原对话模板，兼容现有训练/推理链路。



数据示例

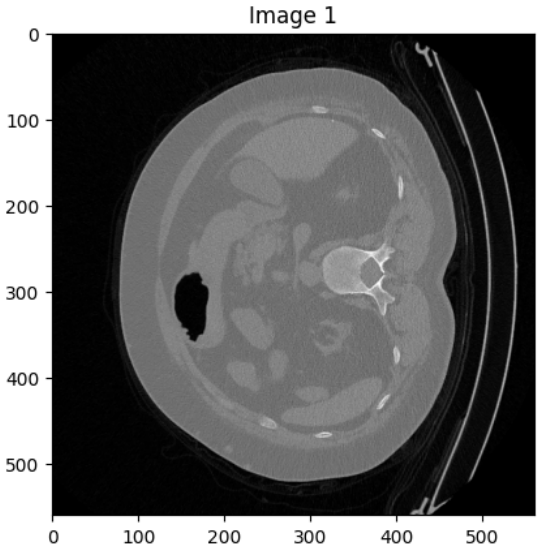


Image 1
0
100
200
300
400
500
0
100
200
300
400
500

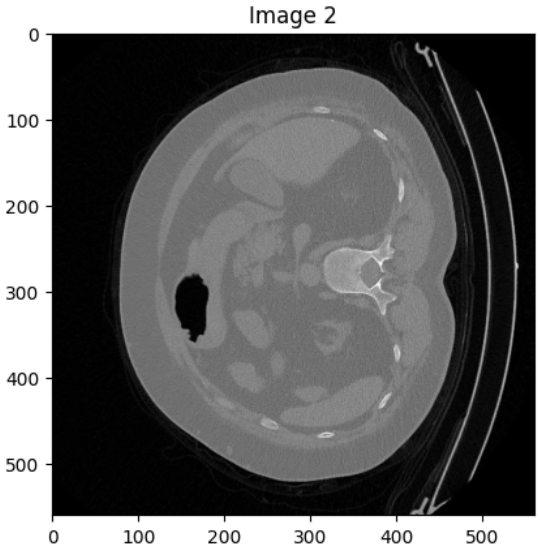
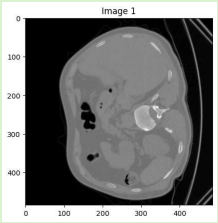


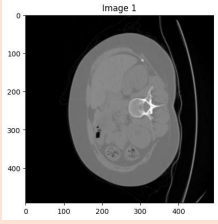
Image 2
0
100
200
300
400
500
0
100
200
300
400
500

对比现象



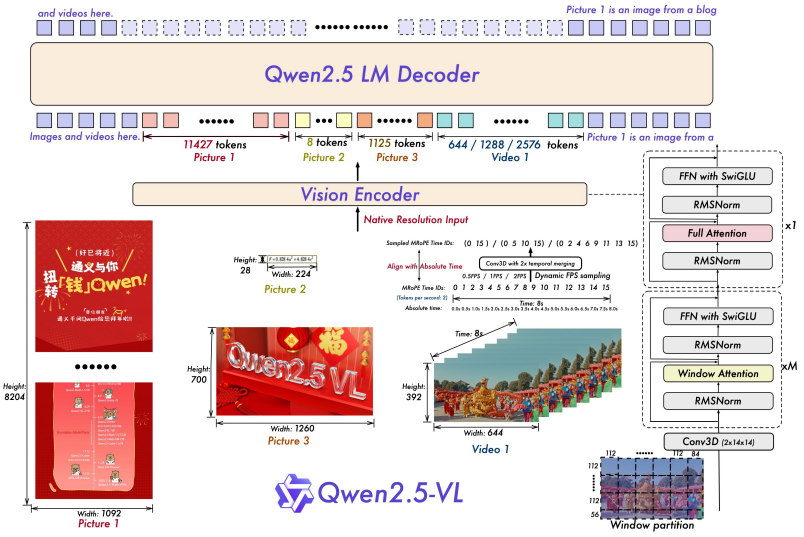
Task: Report Generation Based on 3D Chest CT-Rate Datasets

Findings: **Trachea and both main bronchi are open**（气管及双侧主支气管通畅）. **Mediastinal vascular structures and heart appear natural**（纵隔血管结构及心脏无明显病理性异常）. **No pathological LAP was detected in the mediastinum**（纵隔内未发现病理性肿大淋巴结，LAP=lymphadenopathy）. Pleural effusion-thickening was not detected in both hemithorax. In the evaluation of both lung parenchyma; A ground glass density nodule is observed in the apicoposterior segment of the upper lobe of the right lung. Its size is 6.5 mm in the previous examination and 3.3 mm in the current examination, and a decrease in its size is observed. Apart from this, nodule formation was not detected. **Bilateral adrenal glands are normal**（双侧肾上腺正常） in the sections passing through the upper part of the abdomen. **No obvious pathology was detected in bone structures**（所检查骨结构无明显病理性异常）. No obvious pathology was distinguished. Impression: Not given.



Findings: **Trachea, both main bronchi are open**（气管及双侧主支气管通畅）. **Mediastinal main vascular structures, heart contour, size are normal**（纵隔血管结构及心脏无明显病理性异常）. Thoracic aorta diameter is normal. Pericardial effusion-thickening was not observed. Thoracic esophagus calibration was normal and no significant tumoral wall thickening was detected. There are several lymph nodes with a short axis measuring up to 5 mm in the mediastinum. **No enlarged lymph nodes in prevascular, pre-paratracheal, subcarinal or bilateral hilar-axillary pathological dimensions were detected**（纵隔内未发现病理性肿大淋巴结）. When examined in the lung parenchyma window; Slightly patchy ground glass densities are observed at the apical and superior levels of the upper lobes of both lungs, with the lower lobe at posterobasal levels in both lungs. The findings were evaluated in favor of Covid 19 viral pneumonia. Correlation with clinical and laboratory is recommended. **Upper abdominal organs included in the sections are normal**（上腹部包含层面的器官无明显异常）. No space-occupying lesion was detected in the liver that entered the cross-sectional area. **Bilateral adrenal glands were normal**（双侧肾上腺正常） and no space-occupying lesion was detected. Mild degenerative changes are observed in the vertebral corpus end plates in bone structures. **Bone structures in the study area are natural**（所检查骨结构无明显病理性异常）. Impression: The findings described in the lung parenchyma were evaluated in favor of Covid 19 viral pneumonia. It is recommended to follow the correlation with the clinic and laboratory.

模型架构


and videos here.
Picture 1 is an image from a blog
Qwen2.5 LM Decoder
Images and videos here.
11427 tokens
Picture 1
8 tokens
Picture 2
1125 tokens
Picture 3
644 / 1288 / 2576 tokens
Video 1
Picture 1 is an image from a
Vision Encoder
Native Resolution Input
FFN with SwiGLU
RMSNorm
Full Attention
RMSNorm
x1
FFN with SwiGLU
RMSNorm
Window Attention
RMSNorm
xM
Conv3D (2x14x14)
Height: 8204
Width: 1092
Picture 1
Height: 28
Width: 224
Picture 2
Height: 700
Width: 1260
Picture 3
Sampled MRoPE Time IDs:
(0 15) / (0 5 10 15) / (0 2 4 6 9 11 13 15)
Align with Absolute Time
Conv3D with 2x temporal merging
0.5FPS / 1FPS / 2FPS
Dynamic FPS sampling
MRoPE Time IDs:
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Tokens per second: 2
Time: 8s
Absolute time:
0.0s 0.5s 1.0s 1.5s 2.0s 2.5s 3.0s 3.5s 4.0s 4.5s 5.0s 5.5s 6.0s 6.5s 7.0s 7.5s 8.0s
Time: 8s
Height: 392
Width: 644
Video 1
112
112
112
56
Window partition
Qwen2.5-VL

双路负向采样

- 噪声负样本（Ours-Gauss）：构造全局高斯噪声体积 + 原报告，迫使模型在无图像证据时输出高困惑度。
- 异类负样本（Ours-Heter）：替换为标签差异显著的 CT 体积 + 原报告，打破文本模板依赖。
- 训练时正负样本成对出现，保持其他字段一致以减少分布偏移。

损失与优化

$$\mathcal{L} = \mathbb{E}_{(x,y)\in\mathcal{B}_{\text{pos}}}\mathsf{CE}(x,y) - \beta\,\mathbb{E}_{(x,y)\in\mathcal{B}_{\text{neg}}}\mathsf{CE}(x,y), \quad \alpha{=}1,\ \beta{=}0.1$$

- 负样本项取负号，直接最大化其困惑度，抑制无依据的生成。
- 辅助策略：梯度裁剪 + 负损失上界，避免训练早期发散。

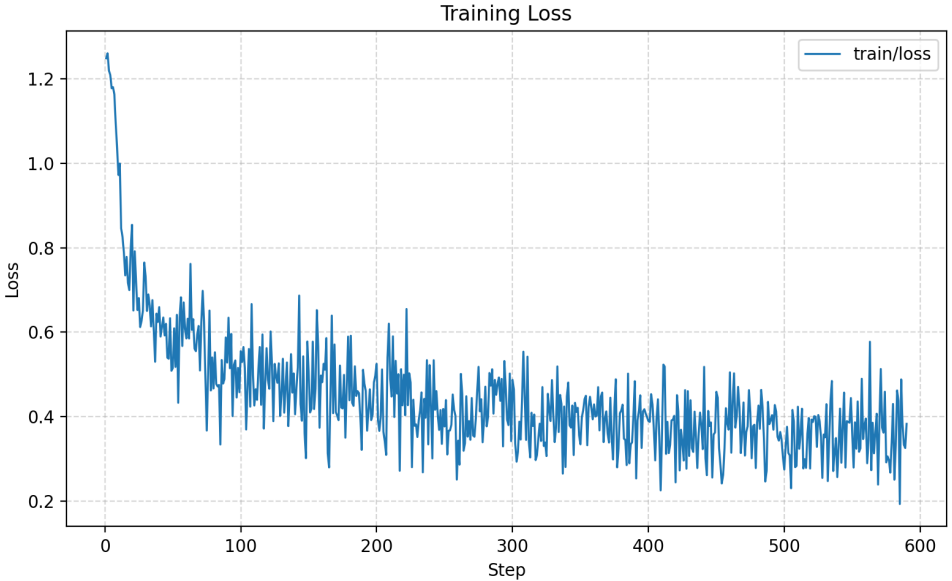
数据与预处理

- 数据集：CT-RATE（25k+ 胸部 CT 序列，21k 患者），提供报告与 18 类病灶标签。
- 预处理：HU 转换；重采样至 $0.75 \times 0.75 \times 1.5$ mm；窗宽 [-1000, 1000] 并归一化到 [-1, 1]。
- 存储形态：完整 preprocessed 与 numpy 版 preprocessed_npy 方便快速加载。

训练设置

- 设备：2 张 A100，训练 1 epoch；学习率 2×10^{-5}，batch size 2（梯度累积 8）。
- 与基座一致的视觉压缩与对话模板，仅调整负样本构造与损失。
- 验证集挑选超参，最佳 checkpoint 在测试集评估。

训练曲线


Training Loss
train/loss
Loss
Step
1.2
1.0
0.8
0.6
0.4
0.2
0
100
200
300
400
500
600

主要结果

方法	BLEU_1 相对提升	BLEU_4 相对提升	Coverage 提升
无负向约束（基础）	–	–	–
Ours-Gauss	+29.1%	+24.3%	+14.5%
Ours-Heter	+28.7%	+23.5%	+13.9%

负向采样显著降低幻觉，提升临床相关性。

局限与讨论

- 正负样本需成对加载，训练吞吐降低。
- 负损失上界的手工设定可能限制进一步收益。
- 仍缺少对病灶局部解释性的显式约束。

未来工作

- 引入切片级对比/检索，增强跨切片显式对齐。
- 结合局部病灶提示或掩码监督，提升小病灶敏感度。
- 探索自适应负样本权重调度，平衡语言流畅度与视觉依赖。