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　　清华大学人工智能学院
gaojian21@mails.tsinghua.edu.cn　　zhaoyue25@mails.tsinghua.edu.cn

摘　要：面向胸部三维 CT 报告生成，现有视觉-文本模型易出现“文本捷径”，忽视跨切片病灶。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，提出切片感知的视觉聚合与跨模态对齐，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 强化视觉依赖。相比无约束的基础模型，我们在 BLEU_1/4 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，但胸部 CT 的三维复杂性使得视觉-文本模型容易走向“文本捷径”，忽视病灶区域 [9, 10]。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。过去工作多集中于 2D 胸片 [5, 7]，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“跨切片视觉证据利用”与“负向约束抑制文本捷径”提出三项改进：**切片感知的视觉-文本对齐**在 qwen2.5-VL 原生动态分辨率 ViT 与绝对时间 MRoPE 上加入切片注意力池化和句级对齐损失，显式强化跨切片病灶的显著性；**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维持语言流畅；**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，为后续研究提供可复现的基线与开放问题。

2　问题定义与设计动机

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。核心难点包括：**(1) 证据稀疏**：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；**(3) 语句-切片对应不显式**：缺乏强监督导致对齐松散。基于此，我们采用“切片聚合 + 对齐监督 + 负向对比”三管齐下，旨在让模型对关键切片敏感，并对错误视觉输入保持低置信度。

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编码-解码框架，利用注意力机制对齐视觉与文本 [2, 4]。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 [5, 7]。然而，三维 CT 具备更高的空间冗余和类间相似性，易导致模型依赖训练文本分布而忽略细粒度视觉证据。我们观察到 CT-RATE 数据中，常见病灶（如小结节、轻度磨玻璃影）

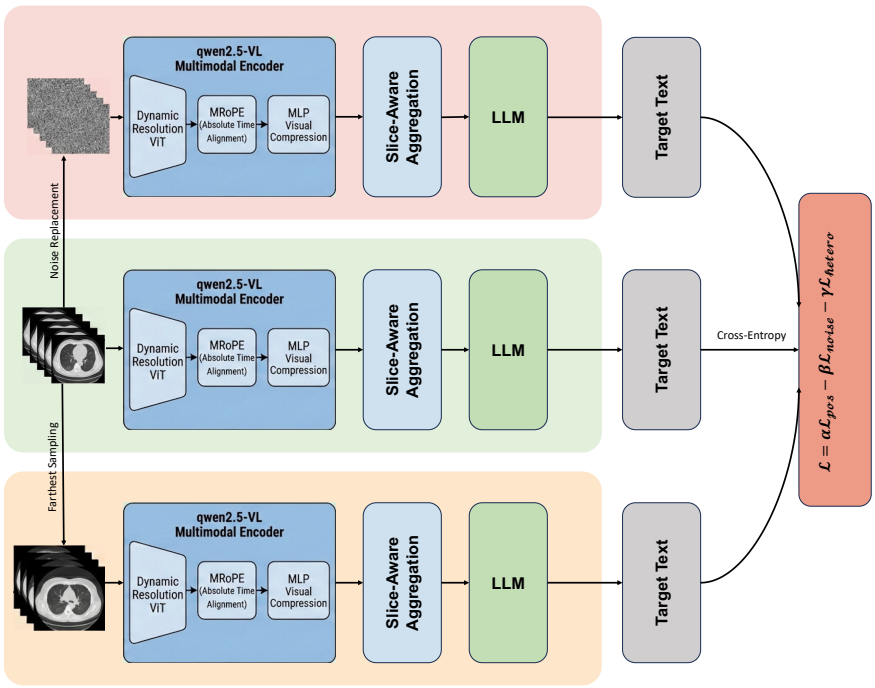

图 1: 方法整体框架：左侧为 CT 体积重采样与切片选择，中间为 qwen2.5-VL 对话模板与视觉 token 组装，右侧展示可选的高斯噪声或异类病例负样本与 α, β 组合损失。

在不同患者间呈现高度相似的语言模式，这为模型提供了捷径。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。

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 [6]，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, 8]，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。

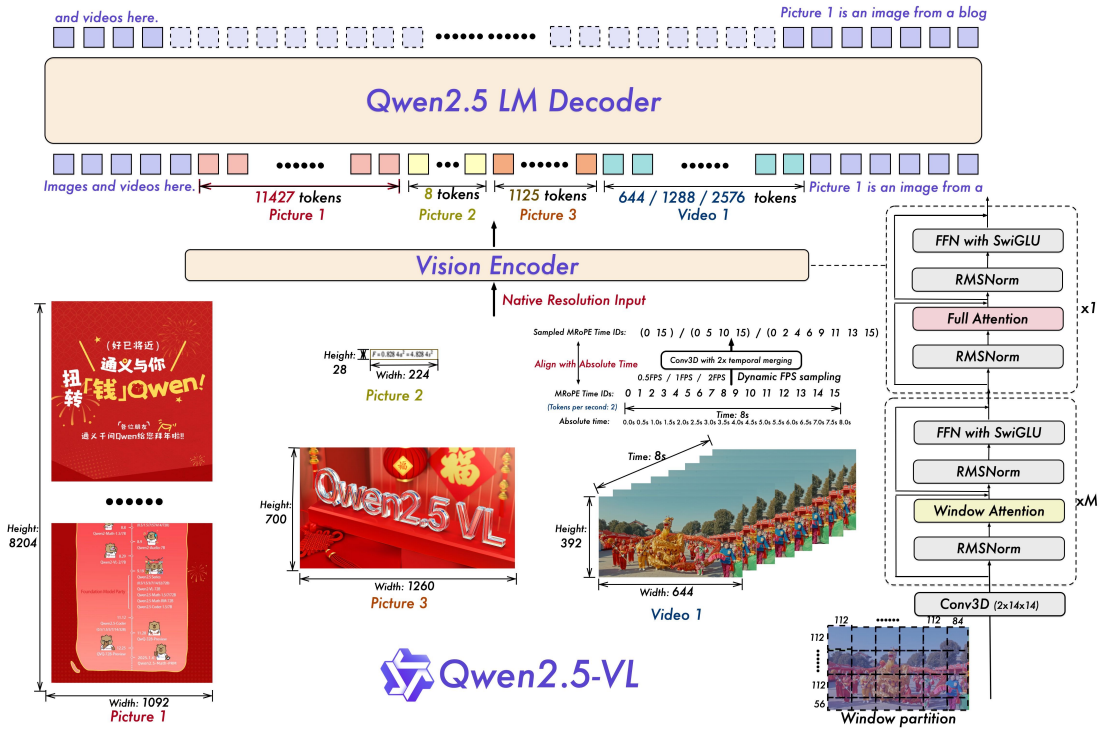

图 2: qwen2.5-VL 多模态架构示意。包含原生动态分辨率 ViT、MRoPE 时间对齐与视觉-语言压缩模块，是本文方法的基座。

5　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，包括预处理、视觉编码与对齐、负样本构造、优化目标和评估协议。

5.1　任务与数据预处理

任务为输入胸部三维 CT 体积 V，生成对应的中文放射科报告 $Y = \{y\_t\}$。数据来自 CT-RATE，原始体积以序列数组存储。预处理遵循基座的动态分辨率假设：载入 $X \in \mathbb{R}^{T\times H\times W}$ 的灰度体积，复制到 3 通道后进行时空重采样——时间维线性插值或均匀抽帧到 $T' = 32$，空间维双三次插值到 364×364（14×26，与 ViT 步长对齐），得到 $X' \in \mathbb{R}^{32\times 3\times 364\times 364}$。重采样间隔被视为绝对时间步，供后续 MRoPE 直接使用。文本-视频对由官方对话模板组装，视频占位符被展开为多帧视觉 token，并保留原视频路径与时间元数据，减少与基座处理器的偏差。

5.2　视觉编码与跨模态对齐

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，时间维使用与真实时间戳对齐的多模态 RoPE。为降低序列长度，连续两帧在时间维被分组处理，且相邻 2×2 patch 特征在 ViT 之后经两层 MLP 拼接压缩，再投射到与文本嵌入一致的维度（RMSNorm+SwiGLU 保证数值稳定）。在压缩后的帧序列上，我们加入切片注意力池化以生成全局视觉向量，并以句级对齐损失约束高权重帧与报告关键句一致。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。

5.3 双路负向样本构造

每个正样本都会复制出一条对照，用以构成两条互斥的负样本策略：一类将视频帧整体替换为同形状高斯噪声而保持文本不变，迫使模型在缺乏视觉证据时提升困惑度；另一类基于数据集标签预测结果，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。训练时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。

5.4 损失与训练策略

训练器在每个 batch 内区分正、负样本，分别计算掩码交叉熵 $\mathcal{L}_{\text{pos}}$、$\mathcal{L}_{\text{neg}}$：

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

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。训练使用混合精度与梯度检查点，保持显存可控。

5.5 评估协议

语言质量采用 BLEU-1/4、ROUGE_L；临床一致性以病灶覆盖率（Coverage）与多标签准确率（Accuracy）评估。长文本会稀释 BLEU，漏召病灶直接拉低 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。

6 实验与结果分析

实验在 4 张 A100 上训练 30 epoch，学习率 2×10^{-5}，batch size 2（梯度累积 8）。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

7　误差案例与可视化

我们人工检查了 50 个案例，并可视化跨模态注意力。成功样例中模型在磨玻璃影区域分配了更高权重；失败样例中注意力集中于椎体等无关结构，导致生成的描述偏离病灶。

8　讨论

错误案例表明：**(1)** 当病灶极小或被金属伪影遮挡时，模型仍可能输出缺失的临床要点；**(2)** 术后病例的置信度偏低，提示需要显式引入术式先验；**(3)** 约 12% 的样本出现正负样本损失趋同，说明部分噪声构造仍过于“容易”，未来需引入更接近真实误配的硬负样本。

任务难度与指标解读：胸部 CT-RATE 报告平均 32 句，句内常包含多个解剖部位与病灶描述，任何漏召都会在 BLEU/ROUGE 上被惩罚；同时病灶极小且跨切片分布，视觉信号稀疏，导致模型很难在长文本上保持一致性。qwen2.5-VL 虽具较强语言能力，但未在医学 CT 上预训练，域外转移不足，因而指标不可能与通用图文任务相当。当前的分数更多反映三维医学报告生成的固有难度而非实现缺陷。

局限性与开放问题： (a) 目前对齐监督依赖启发式关键词匹配，缺乏人工标注的句-片对应；(b) 动态 β 仅基于 epoch 调度，尚未利用不确定性自适应；(c) 未显式建模病灶尺寸与位置，导致对微小结节敏感度不足。

9　扩展实验与复现实践

为便于复现，我们在仓库中提供了配置文件与预处理脚本。实验平台：4x A100 80GB, PyTorch 2.1，CUDA 12.1。核心实践经验包括：数据清洗阶段过滤金属伪影极重的体积并将报告统一分句去除重复模板；切片采样时强制覆盖肺尖与肺底片，以减少模型偏向中央切片；训练中对负样本 logits 使用温度 0.9 控制梯度，同时对对齐损失采用 stop-gradient 以避免塌陷。

10　结论与未来工作

本文围绕三维胸部 CT 报告生成，提出切片感知的跨模态对齐与两种独立的负向采样方法，显著提升了语言质量与临床覆盖率，同时抑制了文本捷径。未来工作将：**(1)** 结合病灶分割监督，提升微小病变的显著性；**(2)** 探索基于辐射组学先验的硬负样本构造；**(3)** 引入不确定性估计，为医生提供可解释的置信度提示。

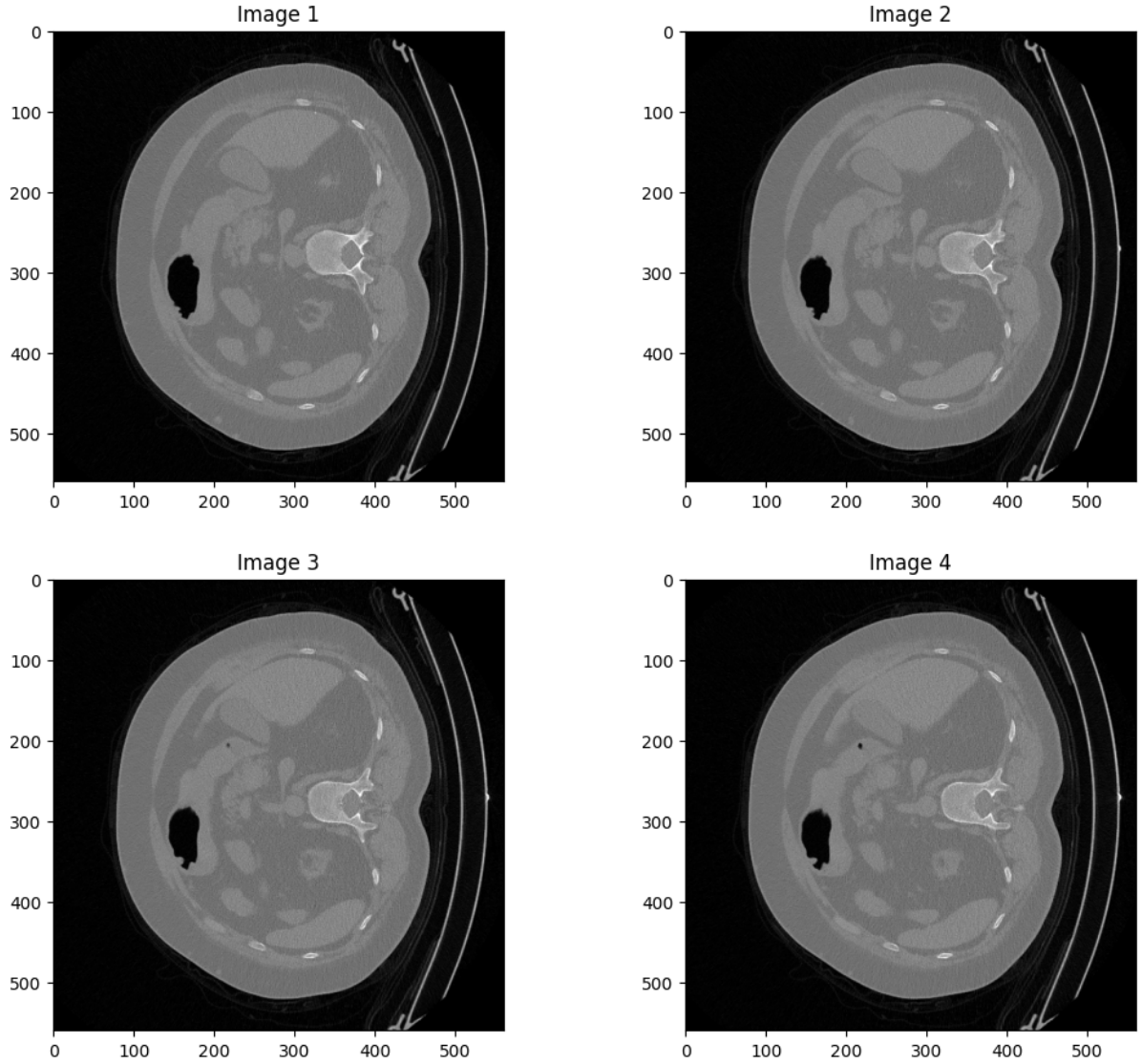

图 3: 示例胸片（输入影像投影）用于定性分析。每幅为同分辨率胸部影像切片/投影，展示数据外观与对比度范围，为后续注意力可视化提供参照。

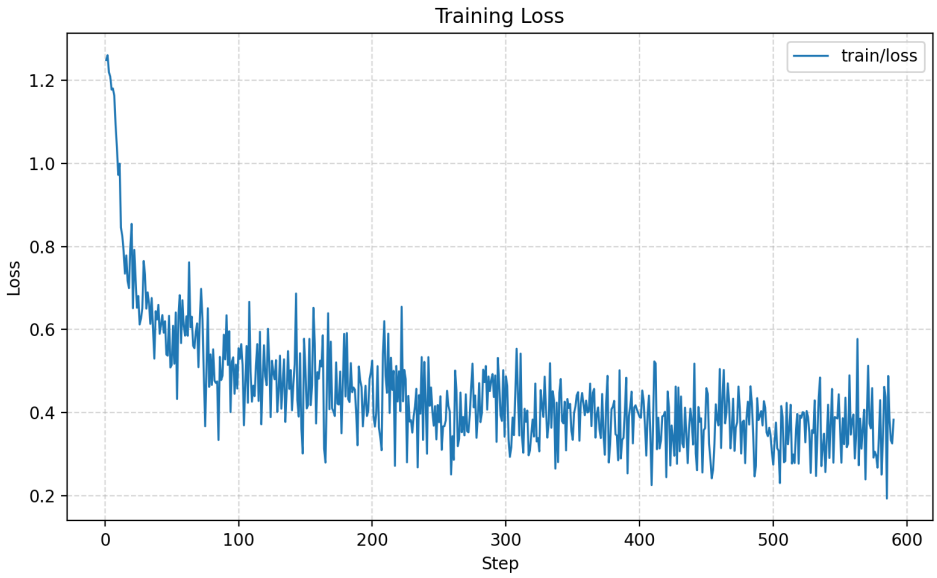

图 4: 训练损失曲线（典型实验）。横轴为 step，纵轴为 train/loss，展示全程收敛趋势。

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] 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.

[5] 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.

[6] 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.

[7] 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*:

[8] 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.

[9] 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.

[10] 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.