Hans Journal of Biomedicine 生物医学, 2026, 16(5), 914-924
Published Online September 2026 in Hans. https://www.hanspub.org/journal/hjbm
https://doi.org/10.12677/hjbm.2026.165093

Hans汉斯

基于原型对比学习的可解释放射学报告生成方法

张颜秀祺，李慧敏*

云南民族大学数学与计算机科学学院，云南 昆明

收稿日期：2026年8月17日；录用日期：2026年9月15日；发布日期：2026年9月24日

摘　要

自动放射学报告生成的目标是根据胸部X射线图像自动写出临床描述，以减轻放射科医生的报告撰写负担。已有方法虽取得一定进展，但生成结果缺乏可解释性，同时图像特征与文本语义之间没有显式对齐，限制了报告质量的提升。为此，我们提出ProtoR2Gen框架，引入一组可学习的疾病原型作为语义锚点，利用对比学习将图像特征向对应的原型聚集，从而在视觉编码和文本生成之间建立明确的对应关系。在推理时，原型的激活分布可转化为图像上的热力图，直观展示模型关注的解剖区域。在IU X-Ray数据集上，ProtoR2Gen的CIDEr指标达到0.656，显著优于对比方法。消融实验和可视化结果也证明了原型机制对语义对齐和可解释性的提升作用。

关键词

放射学报告生成，原型学习，对比学习，跨模态对齐

Interpretable Radiology Report Generation via Prototype-Guided Contrastive Learning

Yanxiuqi Zhang, Huimin Li*

School of Mathematics and Computer Science, Yunnan Minzu University, Kunming Yunnan

Received: August 17, 2026; accepted: September 15, 2026; published: September 24, 2026

Abstract

Automatic radiology report generation aims to automatically produce clinical descriptions from chest X-ray images to reduce the documentation burden on radiologists. Existing methods have

*通讯作者。

文章引用: 张颜秀祺, 李慧敏. 基于原型对比学习的可解释放射学报告生成方法[J]. 生物医学, 2026, 16(5): 914-924.
DOI: 10.12677/hjbm.2026.165093

made some progress, but the generated results lack interpretability, and visual features are not explicitly aligned with textual semantics, which limits further improvement in report quality. To address this, we propose ProtoR2Gen, a framework that introduces a set of learnable disease prototypes as semantic anchors. Through contrastive learning, image features are pulled toward their corresponding prototypes, establishing a clear correspondence between visual encoding and text generation. During inference, the activation distribution of prototypes can be converted into heatmaps over the image, intuitively showing which anatomical regions the model focuses on. On the IU X-Ray dataset, ProtoR2Gen achieves a CIDEr score of 0.656, significantly outperforming comparison methods. Ablation studies and visualization results also demonstrate the effectiveness of the prototype mechanism in improving semantic alignment and interpretability.

Keywords

Radiology Report Generation, Prototype Learning, Contrastive Learning, Cross-Modal Alignment

1. 引言

放射学报告直接关乎诊疗决策，胸部 X 射线检查因便捷廉价而广泛使用，但其巨大需求量使放射科医师持续超负荷。自动报告生成(RRG)旨在缓解这一矛盾，近年备受关注[1] [2]。

早期研究沿袭图像描述范式，以 CNN 提取视觉特征、RNN 逐词生成报告[1]-[3]，但 RNN 在长程语义建模方面存在不足。Transformer 架构的引入改善了文本连贯性[4]-[6]，然而在类别不平衡条件下对罕见病灶的学习仍不充分。近期，状态空间模型[7]与大型语言模型[8] [9]被引入该任务，但幻觉和事实不一致问题依然突出。

现有方法普遍缺乏可解释性，其单阶段对齐策略也忽略了读片过程固有的层次性——即从疾病模式识别到解剖区域定位、再到综合诊断的递进逻辑[10]，导致语义漂移。在评估方面，BLEU 和 ROUGE 仅衡量表层 n-gram 重叠，而 CIDEr [11]和 RadGraph [12]等结构化指标更贴近临床需求。

上述问题的根源在于缺少连接视觉特征与文本描述的显式语义中间表示。为此，我们借鉴原型学习的思想[13] [14] [15]。原型学习通过度量查询样本与可学习原型之间的距离进行分类，在医学图像领域已用于提供可解释的诊断依据。同时，对比学习通过拉近正样本对、推远负样本对来学习判别性表征，在跨模态对齐中展现出优势[16]-[19]。

基于这些思想，我们提出 ProtoR2Gen 框架，在 VMamba 视觉编码器[7] [20]与 Llama2-7B 语言解码器[21]之间插入可学习疾病原型作为语义锚点，通过对比损失聚合图像特征，并采用 LoRA [22]进行参数高效微调。

本文的主要贡献包括：

1) 提出 ProtoR2Gen，一个将可学习疾病原型作为跨模态语义锚点的放射学报告生成框架。

2) 利用原型的内部结构构建决策可视化解译机制，将模型推理映射至图像解剖区域。

3) 在 IU X-Ray 基准上的系统实验显示，ProtoR2Gen 的 CIDEr 达到 0.656；原型相似性分析与疾病亚组分析进一步揭示了模型的行为特性与当前局限。

2. 方法

ProtoR2Gen 的总体结构如图 1 所示，包括视觉编码器、语言解码器和原型对比学习模块三部分。视觉编码器基于 VMamba 提取多尺度图像特征；语言解码器采用 Llama2-7B 并通过 LoRA 微调；原型对比学习模块在二者之间建立语义锚定。

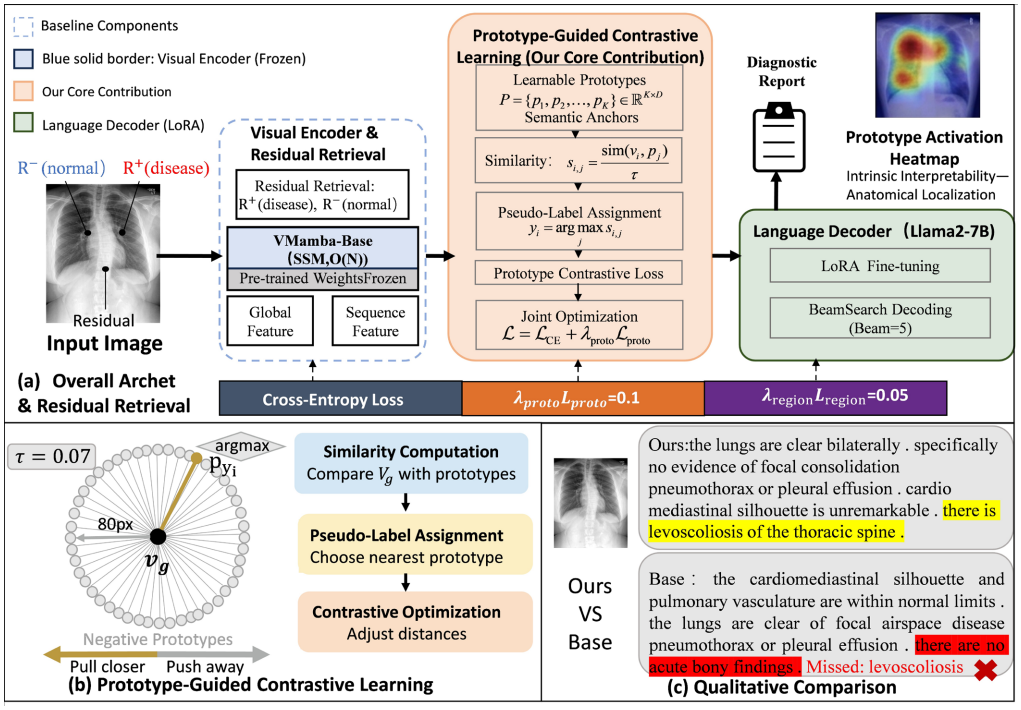

Figure 1. Overview of the ProtoR2Gen architecture
图 1. ProtoR2Gen 整体架构示意图

2.1. 问题定义

给定一张胸部 X 射线图像 I，目标是生成一段自然语言报告 $R=\{r_1,r_2,\cdots,r_T\}$。训练集包含图像–报告对 $\mathbb{D}=\{(I_i,R_i)\}_{i=1}^{N}$，其中报告使用 Findings 部分作为生成目标。该任务可形式化为一个条件语言建模问题：最大化 $p(R\mid I)$。

2.2. 基础框架

2.2.1. 视觉编码器

为高效处理高分辨率医学图像，我们采用 VMamba-Base 作为视觉主干网络。VMamba 基于状态空间模型(SSM)，具有线性计算复杂度 $O(N)$，相比 ViT 等 Transformer 变体的 $O(N^2)$ 复杂度，在处理长序列视觉标记时更具效率优势[7]。Vision Mamba [23]通过双向扫描与位置嵌入将 SSM 适配至二维图像，而 VMamba 进一步提出二维选择性扫描(SS2D)模块，沿四条遍历路径展开图像块以捕获全局上下文，为医学影像分析提供了兼具全局建模能力与计算效率的特征提取方案[24]。

输入图像 $x\in\mathbb{R}^{3\times H\times W}$，经过 VMamba 编码后，得到特征图 $v\in\mathbb{R}^{\hat{H}\times\hat{W}\times\hat{C}}$。随后，我们生成两种视觉表示形式：

$$v_g=\mathrm{Proj}\big(\mathrm{AvgPool}(v)\big),\ v_s=\mathrm{Proj}\big(\mathrm{Flatten}(v)\big) \tag{1}$$

其中 $v_g \in \mathbb{R}^E$ 为全局特征， $v_s \in \mathbb{R}^{L\times E}$ 为序列特征， E 为 LLM 的嵌入维度，Proj 为线性投影层。

2.2.2. 语言模型与 LoRA 微调

我们采用 Llama2-7B 作为报告生成的解码器。Hu 等[22]提出 LoRA 时指出：“我们提出了低秩适应方法，通过冻结预训练模型权重并在 Transformer 架构的每一层中注入可训练的低秩分解矩阵，从而实现大语言模型的参数高效微调，显著降低训练开销”[22]。为避免因全量微调导致的灾难性遗忘和计算资源浪费，我们引入 LoRA 进行参数高效微调。LoRA 通过低秩分解 $W = W_0 + \Delta W = W_0 + BA$ 冻结预训练权重 W_0，仅对注意力层中的 q_{proj} 和 v_{proj} 权重进行低秩矩阵更新(设定秩 $r = 16$)，将可训练参数量降至 0.12%左右。经线性投影映射后的视觉序列特征 v_s 与文本提示词拼接，共同作为 LLM 的输入。

2.2.3. 上下文残差机制

在特征提取前，我们从训练集中为每个样本检索含有疾病的正样本 R^{+i} 和正常的负样本 R^{-i}，并计算视觉和文本两个维度的残差标记。具体而言，我们通过正负样本特征相减来计算残差标记：

$$r_v^i = c_g^{+i} - c_g^{-i},\ r_t^i = t^{+i} - t^{-i} \tag{2}$$

其中 c_g^{+i} 和 c_g^{-i} 分别为正负样本的全局特征，t^{+i} 和 t^{-i} 为对应疾病提示文本的嵌入。这种残差检索机制能够引导模型关注正常样本与疾病样本之间的特征差异，从而强化对疾病的判别能力。

2.3. 原型引导的对比学习

2.3.1. 可学习原型库

受原型网络[15]启发，我们在跨模态潜在空间中引入 K 个可学习的原型嵌入，代表常见的疾病模式或临床语义单元：

$$P = \{p_1, p_2, \cdots, p_K\},\ p_j \in \mathbb{R}^D \tag{3}$$

其中 $D = 4096$ 为 Llama2 的隐层维度。原型参数采用标准正态分布 $\mathcal{N}(0,1)$ 初始化，并在训练中通过反向传播更新。训练过程中，我们对原型进行 L2 归一化 $p_j \leftarrow p_j / \|p_j\|$，以稳定后续的余弦相似度计算。

2.3.2. 图像 - 原型相似度计算与伪标签生成

对于每一个图像样本，首先将视觉特征序列 v_s 沿序列维度平均池化得到全局图像特征 $v_i \in \mathbb{R}^D$。模型计算全局特征与所有原型的余弦相似度：

$$s_{ij} = \frac{v_i \cdot p_j}{\|v_i\| \cdot \|p_j\|} \tag{4}$$

为增强对比效果并控制分布的锐度，我们引入温度系数 $\tau = 0.07$ 对相似度进行缩放：

$$\tilde{s}_{ij} = \frac{s_{ij}}{\tau} \tag{5}$$

由于缺乏对原型的人工标注，我们采用自监督策略为当前样本分配一个伪标签 y_i，即将该图像分配至相似度最高的正样本原型：

$$y_i = \arg\max_j \tilde{s}_{ij} \tag{6}$$

2.3.3. 原型对比损失

我们提出原型对比损失来优化特征空间。该损失函数等价于 InfoNCE 的形式，最大化图像特征与其正原型之间的互信息，同时推远与其他负样本原型的距离：

$$\mathcal{L}_{\text{proto}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(\tilde{s}_{i,y_i}\right)}{\sum_{j=1}^{K}\exp\left(\tilde{s}_{ij}\right)} \tag{7}$$

其中 B 为批大小。通过该损失的约束，视觉编码器提取出的特征将被强制聚合到特定的病理原型周围，相同或相似疾病的图像将在原型空间中彼此靠拢，而不同类别的疾病特征被有效推开。

2.4. 联合优化与可解释性热力图

2.4.1. 联合优化损失

除了上述原型对比损失外，基础框架中包含标准的报告生成交叉熵损失 $\mathcal{L}_{\text{CE}}$。总损失函数定义为：

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{proto}} \cdot \mathcal{L}_{\text{proto}} \tag{8}$$

其中，超参数 λ_{proto} 用于平衡两个优化目标，经过消融实验验证，我们将最优权重设定为 0.1。

2.4.2. 可解释性热力图生成

我们在推理阶段并不依赖外部的注意力解释模块，而是利用原型结构的内在可解释性。具体而言，我们在视觉编码器的输出特征图 ν 上，计算每个空间位置的视觉特征 $z_k \in \mathbb{R}^C$ 与所有 K 个原型之间的相似度，生成一个 K 通道的相似度响应图。将该响应图通过上采样插值恢复至原始图像分辨率 $H \times W$，并叠加对应疾病原型 p_j 的激活权重，即可得到原型激活热力图。该热力图直观地显示了模型在生成特定诊断报告时，所依据的解剖区域特征，从而为临床医生提供可视化的决策依据。

2.5. 训练算法

基于上述方法，ProtoR2Gen 的整体训练流程如算法 1 所示。

Algorithm 1 ProtoR2Gen 联合训练过程 (Algorithm 1 Joint Training Procedure)

Require: $\mathbb{D} = \{(I_i, R_i)\}$，原型数量 K，批大小 B
Ensure: 训练模型参数 θ 和原型库 P

初始化：视觉编码器 VMamba，LLM Llama2，原型 $P \sim \mathcal{N}(0,1)$
for epoch = 1 到 25 **do**
  for 每个 mini-batch $\{(I_b, R_b)\}_{b=1}^{B}$ **do** ▷ 视觉编码
    $\nu_b = \text{VMamba}(I_b)$
    $v_b^g, v_b^s = \text{Proj}(\nu_b)$ ▷ 核心：原型对比学习
    $\tilde{s}_{bj} = \text{cosine}(v_b^g, p_j)/\tau$
    $y_b = \arg\max_j \tilde{s}_{bj}$
    $\mathcal{L}_{\text{proto}} = -\frac{1}{B}\sum_b \log \frac{\exp(\tilde{s}_{b,y_b})}{\sum_j \exp(\tilde{s}_{bj})}$ ▷ 语言生成
    $R_b^{\text{pred}} = \text{Llama2}(v_b^s)$
    $\mathcal{L}_{\text{CE}} = -\sum_t \log p(R_t \mid R_{<t}, I_b)$ ▷ 联合优化
    $\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_{\text{proto}} \cdot \mathcal{L}_{\text{proto}}$ ▷ 原型维护
    $p_j \leftarrow p_j/\|p_j\|$，$\forall j \in [1, K]$
  end for
end for

3. 实验

3.1. 数据集与评估指标

实验采用 IU X-Ray 数据集[2]，包含 7470 张图像和 3955 份报告，仅使用 Findings 部分，按 7:1:2 划分训练/验证/测试集。评估指标包括 BLEU-1/2/3/4 [2]、ROUGE-L [25]、METEOR [26]和 CIDEr [11]，以

CIDEr 为主要指标，RadGraph F1 [12] [27]为补充。

3.2. 实现细节

视觉编码器使用 VMamba-Base，输入尺寸 224×224；语言解码器为 Llama2-7B，采用 LoRA (秩 16)微调，仅更新 q_proj 和 v_proj。原型数 $K = 64$，温度 $\tau = 0.07$，原型损失权重 $\lambda_{\text{proto}} = 0.1$。优化器 AdamW，学习率 1×10^{-4}，训练 25 轮，批次大小 8。推理时束搜索宽度 5，长度惩罚 2.0，输出上限 100 token。实验在单卡 RTX 4090 上完成，使用 DeepSpeed 和 bfloat16 加速。

3.3. 实验结果

表 1 报告了 ProtoR2Gen 在 IU X-Ray 测试集上的性能。我们与多个现有最优方法进行了全面对比，同时复现了 R2GenCSR 基线[8]作为参照。ProtoR2Gen (Epoch 8)的 CIDEr 达到 0.656，较最新的 KIA [28] (0.559)提升 17.4%，较 R2GenGPT [8] (0.542)提升 21.0%，较同骨干的 R2GenCSR 基线(0.580)提升 13.1%，表明原型引导的对比学习能够有效改善视觉–语义对齐质量。BLEU-4 达到 0.173，略低于 Token-Mixer [29] (0.190)和 KIA [28] (0.183)，但优于 R2GenGPT [8] (0.161)。这一格局印证了原型机制的核心优势在于语义层面对齐而非表层 n-gram 复制。

Table 1. Performance comparison on the IU X-Ray test set
表 1. IU X-Ray 测试集上的性能对比

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	CIDEr
R2Gen [4]	0.470	0.304	0.219	0.165	0.371	–
R2GenCMN [5]	0.475	0.309	0.222	0.170	0.375	–
PPKED [30]	0.483	0.315	0.224	0.168	0.376	0.351
METransformer [31]	0.483	0.322	0.228	0.172	0.380	0.435
DCL [18]	–	–	–	0.163	0.383	0.586
R2GenGPT† [8]	0.465	0.299	0.214	0.161	0.376	0.542
Token-Mixer [29]	0.483	0.338	0.250	0.190	0.402	0.482
SILC [32]	0.472	0.321	0.234	0.175	0.379	0.368
KIA [28]	0.501	0.325	0.240	0.183	0.375	0.559
R2GenCSR† [8]	0.456	0.287	0.205	0.162	0.366	0.580
ProtoR2Gen (Epoch 6)	0.480	0.312	0.226	0.173	0.371	0.655
ProtoR2Gen (Epoch 8)	0.480	0.312	0.226	0.173	0.371	0.656

†表示仅使用 Findings 部分进行评估的方法。

此外，为评估生成报告的事实一致性，我们采用 RadGraph F1 指标[12] [27]对 ProtoR2Gen (Epoch 8)进行补充评估。如表 2 所示，模型在 complete、simple 和 partial 三个粒度上分别取得 0.225、0.344 和 0.314 的 F1 分数。其中 simple 粒度表现最优，表明模型在抽取核心临床实体方面具备一定能力；complete 粒度相对较低，反映出生成报告在完整关系链上仍有提升空间。

Table 2. RadGraph F1 factual consistency evaluation
表 2. RadGraph F1 事实一致性评估

粒度	Complete	Simple	Partial
RadGraph F1	0.225	0.344	0.314

3.4. 消融实验

3.4.1. 上下文残差机制的消融实验

为量化上下文残差机制(CRM)对模型性能的独立贡献，我们在 VMamba + Llama2-7B 的固定配置下，对比了启用与禁用 CRM 的实验结果。CRM 通过从训练集中检索正样本(含病灶)和负样本(正常)图像，计算视觉与文本维度的残差标记，引导模型关注疾病相关特征差异(见第 2.2.3 节)。

实验结果如表 3 所示。启用 CRM 后，CIDEr 从 0.612 提升至 0.656 (相对提升 7.2%)，BLEU-4 从 0.168 提升至 0.173，ROUGE-L 从 0.362 提升至 0.371。该结果表明，CRM 能够有效增强模型对病灶区域的判别性关注，改善视觉–语义对齐质量。

Table 3. Ablation study of the context residual mechanism
表 3. 上下文残差机制消融实验

配置	BLEU-4	ROUGE-L	CIDEr	RadGraph-F1 (Simple)
禁用 CRM (context_pair = 0)	0.168	0.362	0.612	0.318
启用 CRM (context_pair = 3)	0.173	0.371	0.656	0.344

3.4.2. 原型损失权重的选择

为确定联合损失中原型对比损失的最优权重系数 λ_{proto}，我们选取 0、0.05、0.1 和 0.2 进行对比分析，结果如表 4 所示。

Table 4. Ablation study of prototype loss weight λ_{proto}
表 4. 原型损失权重 λ_{proto} 的消融实验

λ_{proto}	BLEU-4	CIDEr
0 (基线)	0.162	0.580
0.05	0.170	0.656
0.1 (本文)	0.173	0.655
0.2	0.165	0.620

原型监督的收益呈非单调规律。当 $\lambda_{\text{proto}} = 0.05$ 时，BLEU-4 和 CIDEr 均较基线提升；当 $\lambda_{\text{proto}} = 0.1$ 时，BLEU-4 达到最大值 0.173，CIDEr 保持在 0.655；当 $\lambda_{\text{proto}} = 0.2$ 时，两项指标均低于基线，表明过强的原型约束对主语言建模目标产生了干扰效应。基于以上分析，我们在所有主实验中采用 $\lambda_{\text{proto}} = 0.1$ 作为默认配置。

3.4.3. 公平骨干消融实验

为回应“性能提升可能源于强大的视觉骨干(VMamba)而非原型机制本身”这一潜在质疑，我们将视

觉编码器替换为 ResNet-101，并保持语言解码器(Llama2-7B + LoRA)、训练策略和超参数完全一致，结果如表 5 所示。

Table 5. Ablation results of prototype module under different visual backbones
表 5. 不同视觉骨干下原型模块的消融实验结果

视觉骨干	原型模块	BLEU-4	ROUGE-L	CIDEr
ResNet-101	×	0.148	0.371	0.508
ResNet-101	✓	0.147	0.360	**0.543**
VMamba	×	0.162	0.366	0.580
VMamba	✓	0.167	0.367	**0.656**

即使在 ResNet-101 上，原型模块依然带来了清晰且稳定的性能增益(CIDEr 从 0.508 提升至 0.543，相对提升 6.77%)，表明原型模块的增益并非源于对训练集 n-gram 分布的简单模仿，而是源于更本质的语义层面对齐。

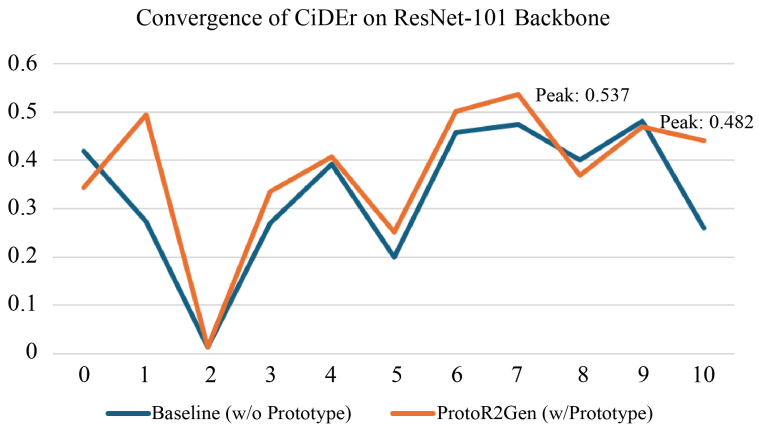

Figure 2. CIDEr curves over epochs under two settings on ResNet-101
图 2. ResNet-101 上两种设定下 CIDEr 随训练轮次的变化曲线

图 2 绘制了 ResNet-101 上 CIDEr 随训练轮次的变化曲线。带原型的模型在第 1 个 epoch 即达到 CIDEr 0.495，而无原型基线在第 4 个 epoch 仅达到 0.391；原型模型在第 7 个 epoch 达到峰值(0.537)，早于基线在第 9 个 epoch 才达到的峰值(0.482)。这表明原型作为结构化的语义锚点，在训练早期即为视觉编码器提供了有效的判别性梯度信号，加速了跨模态对齐的收敛过程。此外，原型模型在验证集峰值(0.537)与测试集(0.543)之间的偏差仅为 0.006，显著小于基线的 0.026，表明其抑制了对验证集特定样本的过拟合倾向。

3.5. 定性分析

3.5.1. 模型效率

原型模块仅增加 0.6M 参数量(<0.65%)，训练时间和推理延迟增加约 3.5%，显存占用几乎不变。结果表明原型引导的对比学习以可忽略的计算开销实现了显著的性能提升。

Table 6. Model efficiency comparison
表 6. 模型效率对比

模型	参数量(M)	训练时间(小时/epoch)	推理时间(ms/图像)	内存(GB)
基线	91.7	3.98	112	75.7
ProtoR2Gen	92.3	4.12	116	76.1

3.5.2. 生成样例

为直观展示模型性能，我们从测试集中随机抽取三个样本，对比基线与 ProtoR2Gen 的生成结果，如图 3 所示。在正常病例中，两者均能准确生成肺部清晰的阴性描述。在包含脊柱侧弯的病例中，两者均未能检出该骨骼异常，均输出“无急性骨性异常”的结论，构成明显的漏报——训练集中骨骼病变样本稀少，模型倾向于学习占多数的肺部正常模式，导致对非肺部异常的识别能力不足。

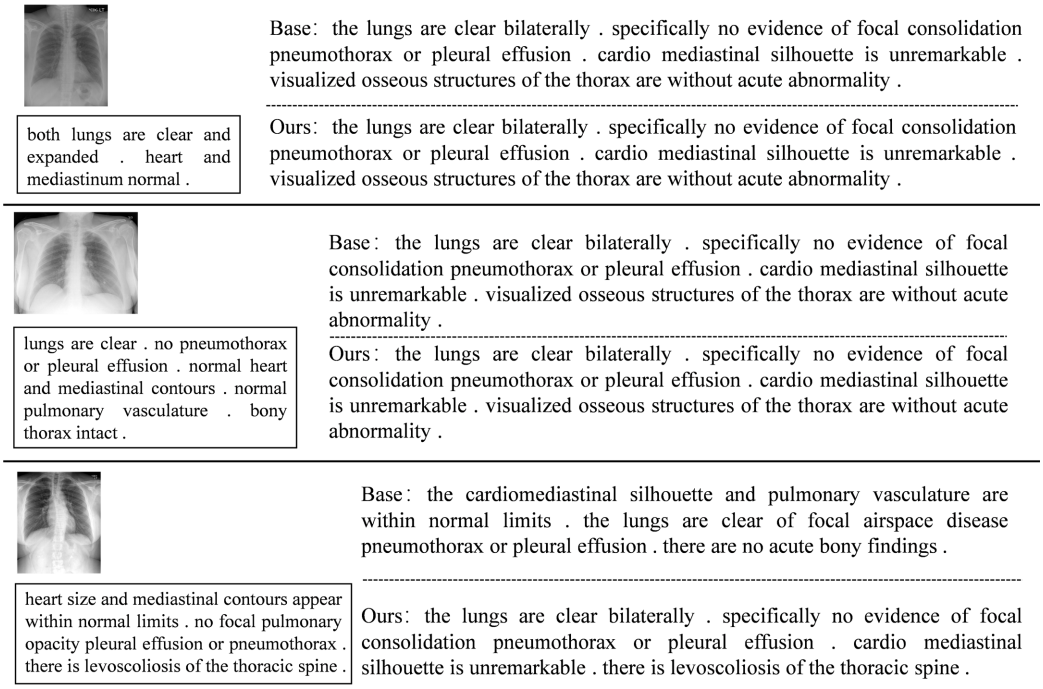

Figure 3. Comparison of generated report samples
图 3. 生成报告样例对比

4. 讨论

ProtoR2Gen 以可学习原型作为视觉与语言的语义中介，其有效性可从表征学习角度解释。现有 RRG 方法将视觉编码器输出直接输入解码器，缺乏中间语义层级的显式建模，而原型层强制图像特征先投影至原型张成的语义空间，每个原型相当于一个可微分的模式检测器，通过对比损失聚类，使视觉编码器学习判别性嵌入，避免少数类被淹没。上下文残差机制通过构造正常-异常残差提供额外判别信号，与原型损失形成互补：前者增强输入层面特征判别性，后者构建结构化语义空间，两者联合实现像素到语义的层次化对齐。然而，当前工作存在若干局限：(1) 原型数量 $K = 64$ 为经验值，未自适应不同病种；(2) 仅在 IU X-Ray 上验证，泛化性待测；(3) 原型可视化缺乏医师外部验证；(4) 骨骼漏报说明对非肺部病变敏感度低；(5) CRM 依赖正负样本检索，罕见病可能缺少足够正样本。未来将探索知识图谱初始化、

区域级对比及多模态迁移。

5. 结论

本文提出 ProtoR2Gen，一个基于原型引导对比学习的放射学报告生成框架。该方法以可学习疾病原型作为视觉与语言之间的语义锚点，通过对比学习实现跨模态的结构化对齐。在 IU X-Ray 数据集上的实验表明该方法有效，CIDEr 达 0.656；原型可视化分析提供了解剖层面的决策解释，这是区别于现有黑盒方法的显著特点。后续工作将聚焦于知识增强的原型初始化、区域级对齐以及在更大规模数据集上的验证。

基金项目

本研究得到云南省数字智能民族计算重点实验室(项目编号：20254916CE340003)和云南省民族多语言智能融合与应用国际联合实验室(项目编号：202403AP140014)资助。

参考文献

[1] Shin, H.C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J. and Summers, R.M. (2016) Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 2497-2506. https://doi.org/10.1109/cvpr.2016.274

[2] Jing, B., Xie, P. and Xing, E. (2018) On the Automatic Generation of Medical Imaging Reports. *Proceedings of the* 56*th Annual Meeting of the Association for Computational Linguistics*, Volume 1, 2577-2586. https://doi.org/10.18653/v1/p18-1240

[3] Li, Y., Liang, X., Hu, Z., *et al*. (2018) Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. *Advances in Neural Information Processing Systems*, Montréal, 3-8 December 2018, 1537-1547.

[4] Chen, Z., Song, Y., Chang, T. and Wan, X. (2020) Generating Radiology Reports via Memory-Driven Transformer. *Proceedings of the* 2020 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), 16-20 November 2020, 1439-1449. https://doi.org/10.18653/v1/2020.emnlp-main.112

[5] Chen, Z., Shen, Y., Song, Y. and Wan, X. (2021) Cross-Modal Memory Networks for Radiology Report Generation. *Proceedings of the* 59*th Annual Meeting of the Association for Computational Linguistics and the* 11*th International Joint Conference on Natural Language Processing*, Volume 1, 5904-5914. https://doi.org/10.18653/v1/2021.acl-long.459

[6] Cornia, M., Stefanini, M., Baraldi, L. and Cucchiara, R. (2020) Meshed-Memory Transformer for Image Captioning. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 10575-10584. https://doi.org/10.1109/cvpr42600.2020.01059

[7] Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., *et al*. (2024) VMamba: Visual State Space Model. *Advances in Neural Information Processing Systems* 37, Vancouver, 9-15 December 2024, 103031-103063. https://doi.org/10.52202/079017-3273

[8] Wang, Z., Liu, L., Wang, L. and Zhou, L. (2023) R2GenGPT: Radiology Report Generation with Frozen LLMs. *Meta-Radiology*, **1**, Article ID: 100033. https://doi.org/10.1016/j.metrad.2023.100033

[9] Liu, C., Tian, Y., Chen, W., Song, Y. and Zhang, Y. (2024) Bootstrapping Large Language Models for Radiology Report Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**, 18635-18643. https://doi.org/10.1609/aaai.v38i17.29826

[10] Tanida, T., Müller, P., Kaissis, G. and Rueckert, D. (2023) Interactive and Explainable Region-Guided Radiology Report Generation. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-21 June 2023, 7433-7442. https://doi.org/10.1109/cvpr52729.2023.00718

[11] Vedantam, R., Zitnick, C.L. and Parikh, D. (2015) CIDEr: Consensus-Based Image Description Evaluation. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 4566-4575. https://doi.org/10.1109/cvpr.2015.7299087

[12] Jain, S., Agrawal, A., Saporta, A., *et al*. (2021) RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. *Proceedings of the* 2021 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, 6-11 June 2021, 3172-3183.

[13] Chen, C., Li, O., Tao, D., *et al*. (2019) This Looks like That: Deep Learning for Interpretable Image Recognition.

Advances in Neural Information Processing Systems, Vancouver, 8-14 December 2019, 8930-8941.

[14] Kim, E., Kim, S., Seo, M. and Yoon, S. (2021) XProtoNet: Diagnosis in Chest Radiography with Global and Local Explanations. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 15719-15728. https://doi.org/10.1109/cvpr46437.2021.01546

[15] Snell, J., Swersky, K. and Zemel, R. (2017) Prototypical Networks for Few-Shot Learning. *Advances in Neural In-formation Processing Systems*, Long Beach, 4-9 December 2017, 4077-4087.

[16] Chen, T., Kornblith, S., Norouzi, M., *et al*. (2020) A Simple Framework for Contrastive Learning of Visual Representations. *International Conference on Machine Learning*, 13-18 July 2020, 1597-1607.

[17] He, K., Fan, H., Wu, Y., Xie, S. and Girshick, R. (2020) Momentum Contrast for Unsupervised Visual Representation Learning. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 9726-9735. https://doi.org/10.1109/cvpr42600.2020.00975

[18] Li, M., Lin, B., Chen, Z., Lin, H., Liang, X. and Chang, X. (2023) Dynamic Graph Enhanced Contrastive Learning for Chest X-Ray Report Generation. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-21 June 2023, 3334-3343. https://doi.org/10.1109/cvpr52729.2023.00325

[19] Zhang, Z. and Jiang, A. (2024) Interactive Dual-Stream Contrastive Learning for Radiology Report Generation. *Journal of Biomedical Informatics*, **157**, Article ID: 104718. https://doi.org/10.1016/j.jbi.2024.104718

[20] Gu, A. and Dao, T. (2023) Mamba: Linear-Time Sequence Modeling with Selective State Spaces. https://arxiv.org/pdf/2312.00752

[21] Touvron, H., Martin, L., Stone, K., *et al*. (2023) Llama 2: Open Foundation and Fine-Tuned Chat Models. https://arxiv.org/abs/2307.09288

[22] Hu, E.J., Shen, Y., Wallis, P., *et al*. (2022) LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations*, 25-29 April 2022. https://openreview.net/forum?id=nZeVKeeFYf9

[23] Zhu, L., Liao, B., Zhang, Q., *et al*. (2024) Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. *Proceedings of the* 41*st International Conference on Machine Learning*, Vienna, 21-27 July 2024, 62429-62442.

[24] Azad, R., Kazerouni, A., Heidari, M., Aghdam, E.K., Molaei, A., Jia, Y., *et al*. (2024) Advances in Medical Image Analysis with Vision Transformers: A Comprehensive Review. *Medical Image Analysis*, **91**, Article ID: 103000. https://doi.org/10.1016/j.media.2023.103000

[25] Lin, C.Y. (2004) ROUGE: A Package for Automatic Evaluation of Summaries. *Text Summarization Branches Out*, Barcelona, 25-26 July 2004, 74-81.

[26] Banerjee, S. and Lavie, A. (2005) METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, 29 June 2005, 65-72.

[27] Delbrouck, J., Chambon, P., Bluethgen, C., Tsai, E., Almusa, O. and Langlotz, C. (2022) Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards. *Findings of the Association for Computational Linguistics*: *EMNLP* 2022, Abu Dhabi, 7-11 December 2022, 4348-4360. https://doi.org/10.18653/v1/2022.findings-emnlp.319

[28] Park, S., Kim, J., Lee, H., *et al*. (2025) KIA: Knowledge-Infused Attention for Accurate Radiology Report Generation. *Proceedings of the* 31*st International Conference on Computational Linguistics* (*COLING*), Abu Dhabi, 19-24 January 2025, 1-12.

[29] Yang, Y., Yu, J., Fu, Z., Zhang, K., Yu, T., Wang, X., *et al*. (2024) Token-Mixer: Bind Image and Text in One Embedding Space for Medical Image Reporting. *IEEE Transactions on Medical Imaging*, **43**, 4017-4028. https://doi.org/10.1109/tmi.2024.3412402

[30] Liu, F., Wu, X., Ge, S., Fan, W. and Zou, Y. (2021) Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 13753-13762. https://doi.org/10.1109/cvpr46437.2021.01354

[31] Wang, Z., Liu, L., Wang, L. and Zhou, L. (2023) METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-21 June 2023, 11558-11567. https://doi.org/10.1109/cvpr52729.2023.01112

[32] Liu, A., Guo, Y., Yong, J-H. and Xu, F. (2024) Multi-Grained Radiology Report Generation with Sentence-Level Image-Language Contrastive Learning. *IEEE Transactions on Medical Imaging*, **43**, 2657-2669. https://doi.org/10.1109/tmi.2024.3372638