

北京航空航天大学  
BEIHANG UNIVERSITY

# 基于 R2Gen 的 X 光片智能诊断报告生成 从复现到优化：基于医学先验的数据增强策略 项目结题汇报

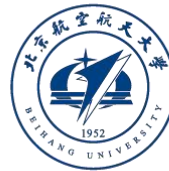
小组成员：陈师尧 凌浩然 曹东润

汇报时间：2025.12.23

指导老师：刘超

# 目录

CONTENTS



北京航空航天大学  
BEIHANG UNIVERSITY

**1 背景介绍**

**2 方法讲解**

**3 结果展示**

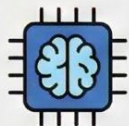
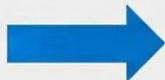
**4 总结展望**

## 任务介绍：医疗影像报告生成

### 研究任务与流程



**输入 (Input):**  
放射影像 (如胸部 X-Ray)



**AI 辅助系统**  
医疗影像报告生成



**输出 (Output):**  
文本报告  
(包含诊断发现和结论)

**目标:** 自动化生成诊断报告

### 痛点 (Pain Point: Why)



放射科医生工作量巨大 -> 容易  
疲劳导致误诊。迫切需要自动化  
辅助系统减负增效。

### 核心挑战



**视觉-语言鸿沟**  
(Visual-Language Gap)

图像是像素，报告是离散  
符号，两者难以对齐。



**长文本生成**  
(Long Text Generation)

相比普通看图说话，医疗报  
告很长，需描述多个病灶。



**模式化严重**  
(Stereotypical Patterns)

报告中有很多套话模板，但  
也也必须精准包含关键异常  
信息。

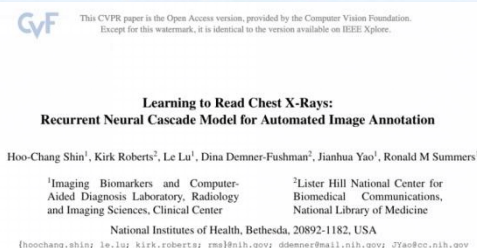


**数据不平衡**  
(Data Imbalance)

正常样本极多，患病样本  
太少。



## 技术演进路线

1. 早期探索:  
CNN-RNN  
(Early Stage)

## Abstract

Despite the recent advances in automatically describing image contents, their applications have been mostly limited to image caption datasets containing natural images (e.g., Flickr30k, MSCOCO). In this paper, we present a deep learning model to effectively detect a disease from an image and annotate its contexts (e.g., location, severity and the affected organs). We employ a publicly available radiology dataset of chest x-rays and their reports, and use its image annotations to mine disease names to train convolutional neural networks (CNNs). In doing so, we adopt various regularization techniques to circumvent the large normal-vs-diseased cases bias. Recurrent neural networks (RNNs) are then trained to describe the contents of a detected disease, based on the deep CNN features. Moreover, we introduce a novel approach to use the weights of the already trained pair of CNN/RNN on the domain specific image/text dataset, to infer the joint image/text contexts for composite image labeling. Significantly improved image annotation results are demonstrated using the recurrent neural cascade model by taking the joint image/text contexts into account.

## 1. Introduction

Comprehensive image understanding requires more than single object classification. There have been many advances in automatic generation of image captions to describe image contents, which is closer to a more complete image understanding than classifying an image to a single object class. Our work is inspired by many of the recent progresses in image caption generation [44, 53, 56, 124, 61, 15, 6, 62, 11], as well as some of the earlier pioneering work [53, 17, 19]. The former have substantially

2497

2. 中期发展:  
Attention Mechanisms  
(Attention Stage)

## On the Automatic Generation of Medical Imaging Reports

Baoyu Jing<sup>1\*</sup>, Pengtao Xie<sup>1</sup>, Eric P. Xing<sup>1</sup><sup>1</sup>Petuum Inc, USA<sup>2</sup>School of Computer Science, Carnegie Mellon University, USA

{baoyu.jing, pengtao.xie, eric.xing}@petuum.com

## Abstract

Medical imaging is widely used in clinical practice for diagnosis and treatment. Report-writing can be error-prone for inexperienced physicians, and time-consuming and tedious for experienced physicians. To address these issues, we study the automatic generation of medical imaging reports. This task presents several challenges. First, a complete report contains multiple heterogeneous forms of information, including findings and tags. Second, abnormal regions in medical images are difficult to identify. Third, the reports are typically long, containing multiple sentences. To cope with these challenges, we (1) build a multi-task learning framework which jointly performs the prediction of tags and the generation of paragraphs, (2) propose a co-attention mechanism to localize regions containing abnormalities and generate narrations for them, (3) develop a hierarchical LSTM model to generate long paragraphs. We demonstrate the effectiveness of the proposed methods on two publicly available datasets.

## 1 Introduction

Medical images, such as radiology and pathology images, are widely used in hospitals for the diagnosis and treatment of many diseases, such as pneumonia and pneumothorax. The reading and interpretation of medical images are usually conducted by specialized medical professionals. For example, radiology images are read by radiologists. They write textual reports (Figure 1) to narrate the findings regarding each area of the body examined in the imaging study, specifically



Figure 1: An example chest x-ray report. In the impression section, the radiologist provides a diagnosis. The findings section lists the radiology observations regarding each area of the body examined in the imaging study. The tags section lists the keywords which represent the critical information in the findings. These keywords are identified using the Medical Text Indexer (MTI).

whether each area was found to be normal, abnormal or potentially abnormal.

For less-experienced radiologists and pathologists, especially those working in the rural area where the quality of healthcare is relatively low, writing medical-imaging reports is demanding. For instance, to correctly read a chest x-ray image, the following skills are needed (Détour et al., 2011): (1) thorough knowledge of the normal anatomy of the thorax, and the basic physiology of chest diseases; (2) skills of analyzing the radiograph through a fixed pattern; (3) ability of evaluating the evolution over time; (4) knowledge of clinical presentation and history; (5) knowledge of the correlation with other diagnostic results (laboratory results, electrocardiogram, and respiratory function tests).

For experienced radiologists and pathologists, writing imaging reports is tedious and time-consuming. In nations with large population such as China, a radiologist may need to read hundreds

3. Transformer时代:  
Memory-driven  
(Project Core)

## Cross-modal Memory Networks for Radiology Report Generation

Zhongheng Chen<sup>1\*</sup>, Yaling Shen<sup>1\*</sup>, Yan Song<sup>2,1</sup>, Xiang Wan<sup>2</sup><sup>1</sup>The Chinese University of Hong Kong (Shenzhen)<sup>2</sup>Shenzhen Research Institute of Big Data

{zhzhongchen, yalingshen}@link.cuhk.edu.cn

\*songyan@cuhk.edu.cn \*wanxiang@sriibd.cn

## Abstract

Medical imaging plays a significant role in clinical practice of medical diagnosis, where the text reports of the images are essential in understanding them and facilitating later treatments. By generating the reports automatically, it is beneficial to help lighten the burden of radiologists and significantly promote clinical automation, which already attracts much attention in applying artificial intelligence to medical domain. Previous studies mainly follow the encoder-decoder paradigm and focus on the aspect of text generation, with few studies considering the importance of cross-modal mappings and explicitly exploit such mappings to facilitate radiology report generation. In this paper, we propose a cross-modal memory networks (CMN) to enhance the encoder-decoder framework for radiology report generation, where a shared memory is designed to record the alignment between images and texts so as to facilitate the interaction and generation across modalities. Experimental results illustrate the effectiveness of our proposed model, where state-of-the-art performance is achieved on two widely used benchmark datasets, i.e., IU X-Ray and MIMIC-CXR. Further analyses also prove that our model is able to better align information from radiology images and texts so as to help generating more accurate reports in terms of clinical indicators.<sup>1</sup>

## 1 Introduction

Interpreting radiology images (e.g., chest X-ray) and writing diagnostic reports are essential operations in clinical practice and normally requires considerable manual workload. Therefore, radiology report generation, which aims to automatically generate a free-text description based on a radiograph, is highly desired to ease the burden of

<sup>1</sup>Our code and the best performing models are released at <https://github.com/zhzhongchen/R2GenCMN>.



Figure 1: A chest X-ray image and its report including findings, impression, and MTI (Medical Text Indexer). The findings section lists the radiology observations regarding each area of the body examined in the imaging study. The impression section provides a diagnosis. The MTI section lists the keywords which represent the critical information in the findings. These keywords are identified using the Medical Text Indexer (MTI).

<sup>2</sup>Along this research track, recently there is only Jing et al. (2018) solution to model the alignments across modalities and text for supervised learning as well as the lack of good model design to learn the correspondences. Unfortunately, few studies<sup>2</sup> are dedicated to solving the restraint. Therefore, it is expected to have a better solution to model the alignments across modalities and further improve the generation ability, although promising results are continuously acquired by other approaches (Li et al., 2018; Liu et al., 2019; Jing et al., 2019; Chen et al., 2020).

<sup>3</sup>Along this research track, recently there is only Jing et al. (2018) solution to model the alignments across modalities and text for supervised learning as well as the lack of good model design to learn the correspondences. Unfortunately, few studies<sup>2</sup> are dedicated to solving the restraint. Therefore, it is expected to have a better solution to model the alignments across modalities and further improve the generation ability, although promising results are continuously acquired by other approaches (Li et al., 2018; Liu et al., 2019; Jing et al., 2019; Chen et al., 2020).

<sup>4</sup>Along this research track, recently there is only Jing et al. (2018) solution to model the alignments across modalities and text for supervised learning as well as the lack of good model design to learn the correspondences. Unfortunately, few studies<sup>2</sup> are dedicated to solving the restraint. Therefore, it is expected to have a better solution to model the alignments across modalities and further improve the generation ability, although promising results are continuously acquired by other approaches (Li et al., 2018; Liu et al., 2019; Jing et al., 2019; Chen et al., 2020).

4. 大模型时代:  
Multimodal LLM  
(SOTA / Future)

## nature communications

## Article

<https://doi.org/10.1038/s41467-025-57426-0>

## Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation

Received: 11 June 2024

Accepted: 21 February 2025

Published online: 06 March 2025

Check for updates

Cheng-Yi Li<sup>1,2,3</sup>, Kao-Jung Chang<sup>2,3,4,5</sup>, Cheng-Fu Yang<sup>6</sup>, Hsin-Yu Wu<sup>1,2</sup>, Wen-Ting Chen<sup>6</sup>, Hsin-Ben Chen<sup>1</sup>, Ling Chen<sup>1</sup>, Yi-Ping Yang<sup>1</sup>, Yu-Chen Chen<sup>1,2,3</sup>, Shih-Pin Chen<sup>6</sup>, Shih-Jen Chen<sup>6</sup>, Jing-Feng Ling<sup>1</sup>, Kai-Wei Chang<sup>6</sup> & Shih-Hwa Chou<sup>2,3,4,5</sup>

Multi-modal large language models (MLLMs) have transformed the landscape of modern healthcare, with automated radiology report generation (RRG) emerging as a cutting-edge application. While 2D MLLM-based RRG has been well established, its utility for 3D medical images remains largely unexplored. In this regard, we curate the 3D-BrainCT dataset (18,885 text-scans pairs) and develop BrainGPT, a clinically valid instruction-tuned (CVIT) model designed for 3D CT RRG. While we notice that the traditional LLM metrics failed to gauge the diagnostic quality of the RRG, we propose feature-oriented radiology task evaluation (FORTE), an evaluation scheme that captures the clinical essence of the generated reports. Here we show that BrainGPT achieves an average FORT-E1 score of 0.71 (degree = 0.66; landmark = 0.70; feature = 0.69; and impression = 0.77) and 74% of BrainGPT-generated reports were indistinguishable from human-written ground truth in a Turing-like test. Together, our work establishes a comprehensive framework encompassing dataset curation, anatomy-aware model fine-tuning, and the development of robust evaluation metrics for the RRG. By sharing our experience in 3D MLLM-based RRG, we aim to accelerate the expedition in human-machine collaboration for next-generation healthcare.

Artificial intelligence (AI) implementation in modern healthcare has reinvigorated our day-to-day practice in patient diagnosis<sup>1</sup>, disease intervention<sup>2</sup>, and clinical research<sup>3</sup>. Although convolutional neural networks (CNN) have conquered some major tasks in image classification and feature segmentation, the CNN outputs are relatively context-

restrictive<sup>4</sup> and were less apprehensive than a fully written diagnostic report<sup>5</sup>. In analogy to this clinical gap, early report generation models have been established<sup>6–11</sup> for chest X-ray (CXR) interpretation<sup>12</sup>. Whereas, the primary success of LLM-based CXR report generation had fueled interdisciplinary interest to explore human-computer interfaces,

<sup>1</sup>School of Medicine, National Yang Ming Chiao Tung University, Taipei City, Taiwan. <sup>2</sup>Department of Medical Research, Taipei Veterans General Hospital, Taipei City, Taiwan. <sup>3</sup>Institute of Clinical Medicine, National Yang Ming Chiao Tung University, Taipei City, Taiwan. <sup>4</sup>Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu City, Taiwan. <sup>5</sup>Department of Ophthalmology, Taipei Veterans General Hospital, Taipei City, Taiwan. <sup>6</sup>Department of Computer Science, University of California, San Diego, La Jolla, CA, USA. <sup>7</sup>Department of Electrical Engineering, National Yang Ming Chiao Tung University, Hsinchu City, Taiwan. <sup>8</sup>Institute of Hospital and Health Care Administration, National Yang Ming Chiao Tung University, Taipei City, Taiwan. <sup>9</sup>Yuli Branch, Taipei Veterans General Hospital, Hualien County, Taiwan. <sup>10</sup>Department of Neurology, Neurological Institute, Taipei Veterans General Hospital, Taipei City, Taiwan. <sup>11</sup>Department of Radiology, School of Medicine, National Yang Ming Chiao Tung University, Taipei City, Taiwan. <sup>12</sup>Institute of Pharmacology, National Yang Ming Chiao Tung University, Taipei City, Taiwan. \*e-mail: michaelchang105@gmail.com; kchangcd@gmail.com; shihhwa@uphs.gov.tw

Nature Communications | (2025)16:2258

1

解决“有无”问题  
(能生成了, 但很短)解决“对齐”问题  
(知道看哪里, 准确率提升)解决“模式与长文本”问题  
(引入记忆, 生成专业长报告)解决“理解与交互”问题  
(3D思维、像人一样思考)



## 方法讲解：X与Y

### 输入定义 (X: Input)

变量名: 图像张量 (Image Tensor) (X)

张量维度 (Dimensions) :

**[Batch\_Size, 3, 224, 224]**

(RGB, Resized至224分辨率)

数据集统计 (Dataset Statistics - Input X)



A. 学习阶段 (Training)

IU X-Ray Train Set: **5,229** 张图像

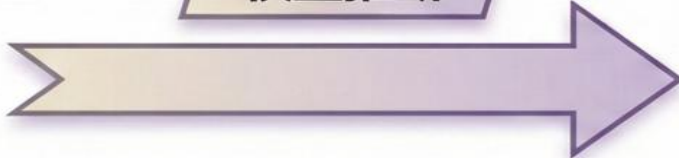


B. 评估阶段 (Evaluation) :

IU X-Ray Test Set: **1,494** 张图像;

Private Set: **43** 张图像 (老师提供)

### 模型推断



多视图融合 (Multi-view Integration) :



N 张图像



1 份报告

(通常是 2 张图: 正位 Frontal + 侧位 Lateral = 1 个病例)

### 输出定义 (Y: Output)

变量名: 文本序列 (Text Sequence) (Y)

张量维度 (Dimensions) :

**[Batch\_Size, Seq\_Len]**

(Word Indices)

数据集统计 (Dataset Statistics - Output Y)



A. 学习阶段 (Training)

**2,768** 份报告 (对应 2,768 位患者)

(逻辑: 多张图对应一份报告)



B. 评估阶段 (Evaluation) :

IU X-Ray Test Set: **791** 份报告

(Ground Truth);

Private Set: **22** 份报告 (对应 22 位患者)

## CMN: 模型架构

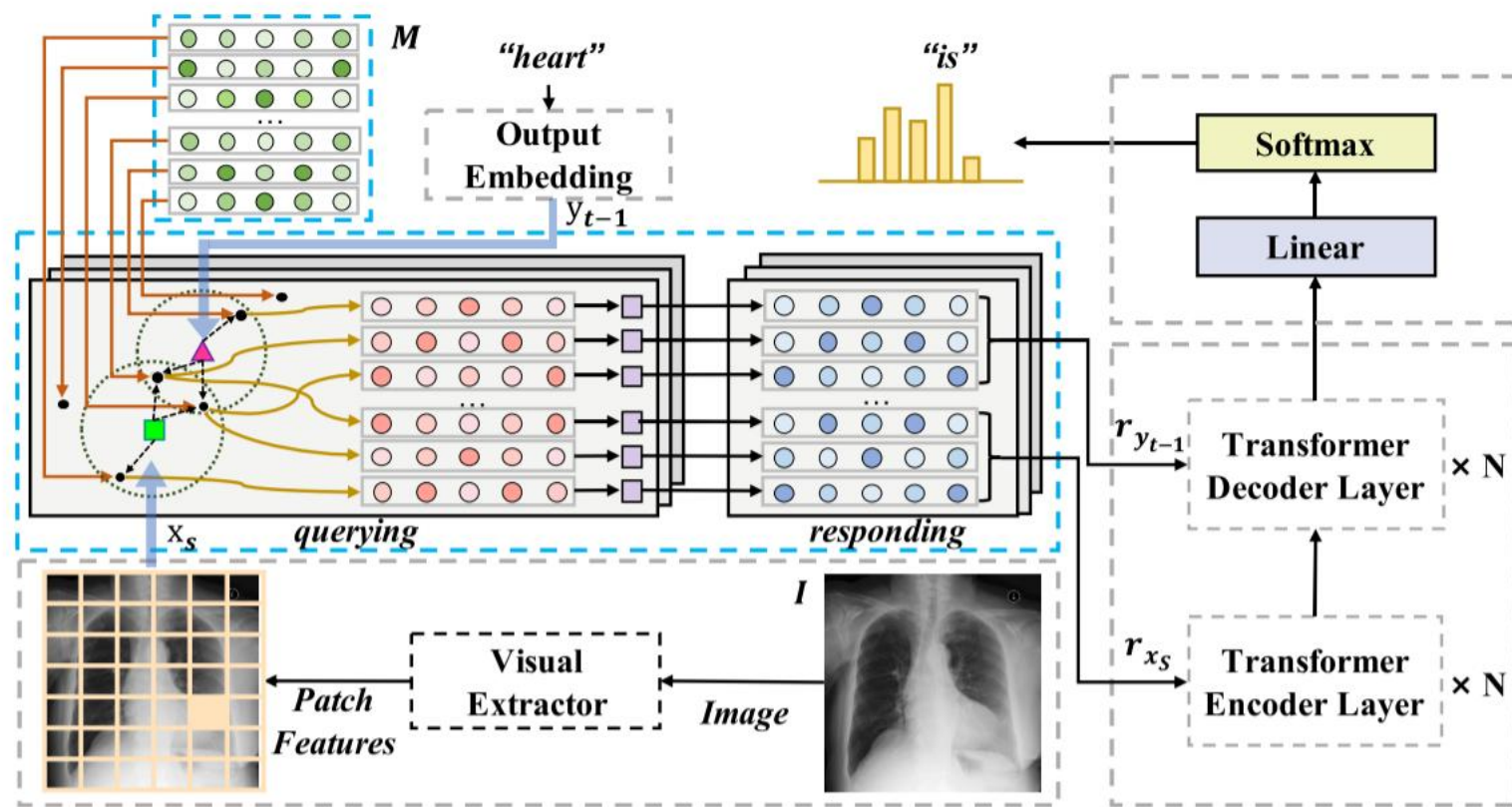
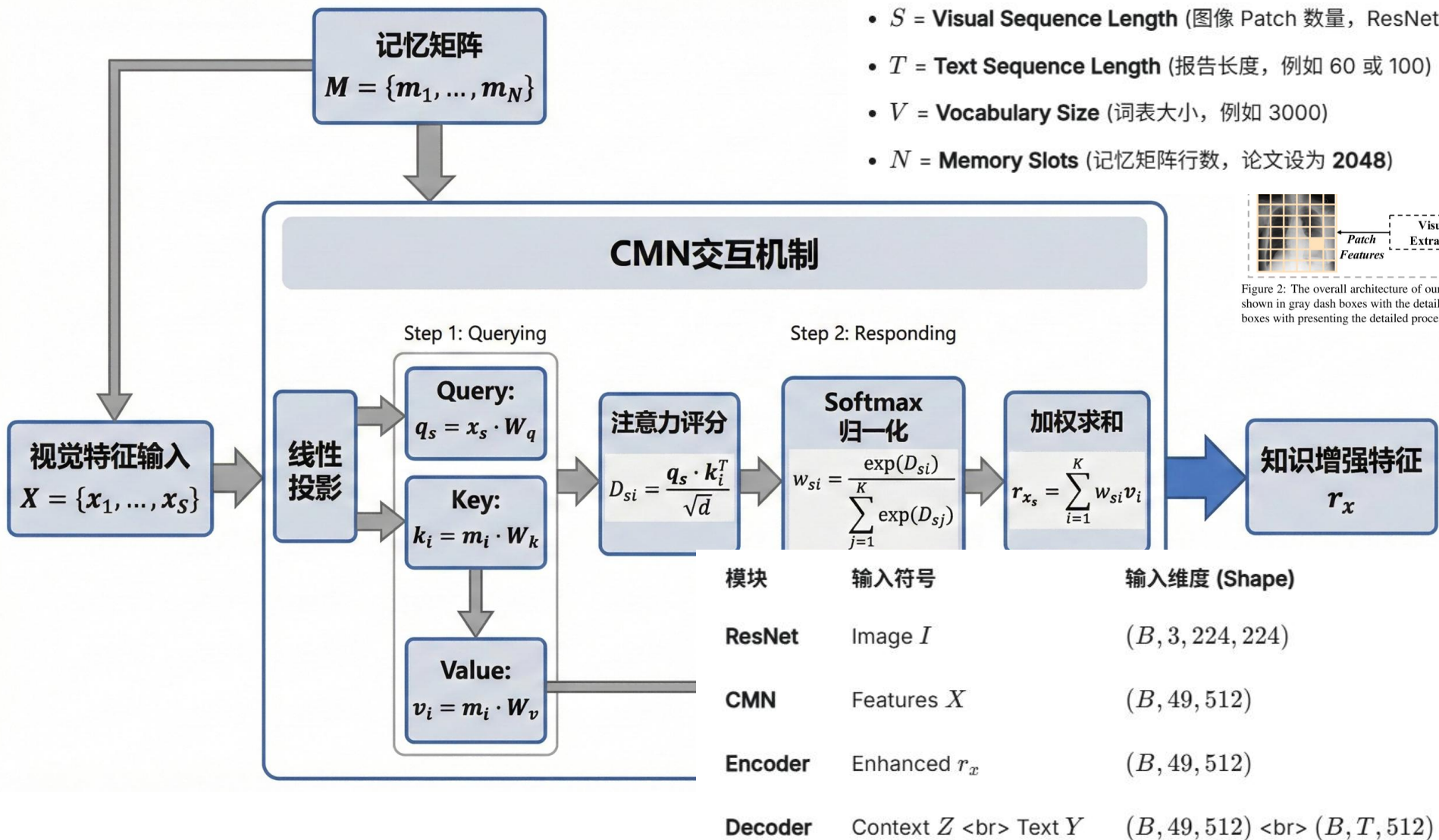


Figure 2: The overall architecture of our proposed approach, where the visual extractor, encoder and decoder are shown in gray dash boxes with the details omitted. The cross-modal memory networks are illustrated in blue dash boxes with presenting the detailed process of memory querying and responding.

1. Visual Extractor (视觉特征提取)
2. Transformer Encoder (编码 X 光图像特征)
3. Cross-modal Memory Networks (跨模态记忆网络)  
★ ← 本论文创新点
4. Transformer Decoder (根据记忆 + 之前的词生成下一个词)



## CMN核心交互机制



- $B$  = Batch Size (批量大小, 例如 16)
- $d$  = Feature Dimension (特征维度, 论文设为 512)
- $S$  = Visual Sequence Length (图像 Patch 数量, ResNet 输出  $7 \times 7$ , 所以  $S = 49$ )
- $T$  = Text Sequence Length (报告长度, 例如 60 或 100)
- $V$  = Vocabulary Size (词表大小, 例如 3000)
- $N$  = Memory Slots (记忆矩阵行数, 论文设为 2048)

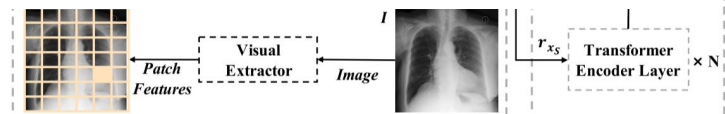
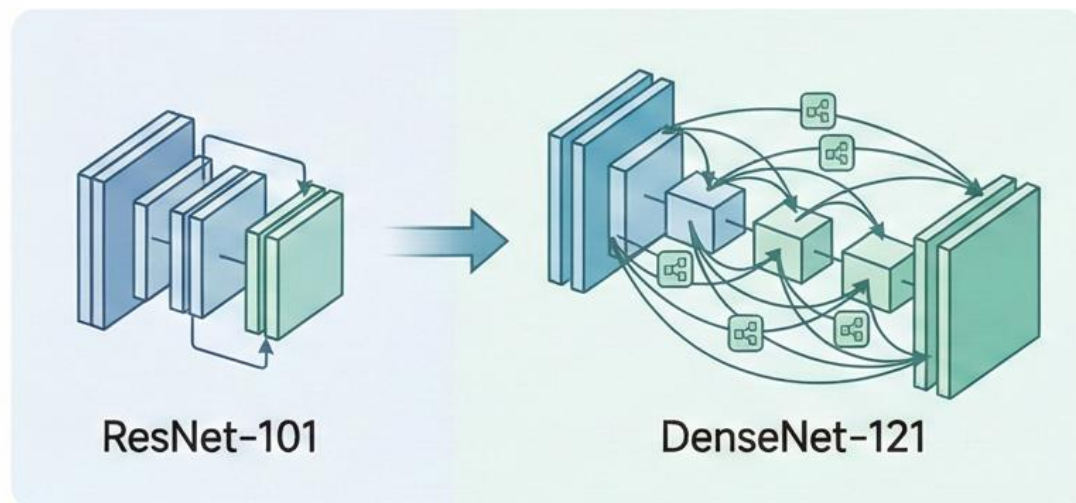


Figure 2: The overall architecture of our proposed approach, where the visual extractor, encoder and decoder are shown in gray dash boxes with the details omitted. The cross-modal memory networks are illustrated in blue dash boxes with presenting the detailed process of memory querying and responding.

## 阶段一：视觉骨干替换与训练策略调整

### 视觉提取器升级



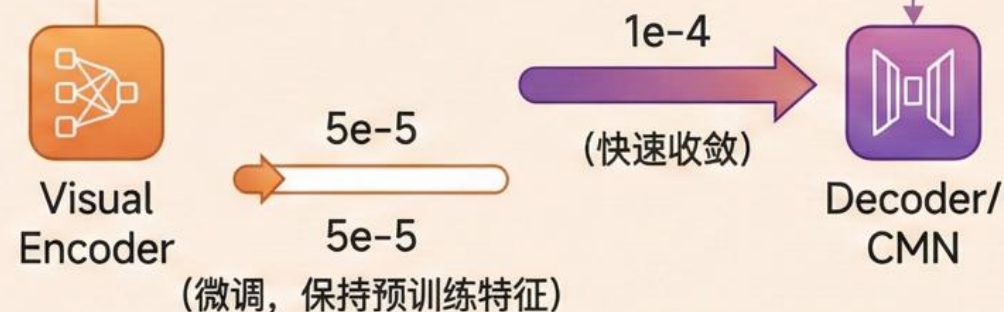
经典残差结构

密集连接，特征重用，高效

理由：DenseNet 的密集连接特性使其更擅长捕捉医学图像中细微的纹理特征（Feature Reuse），且参数利用率更高。

### 差异化学习率

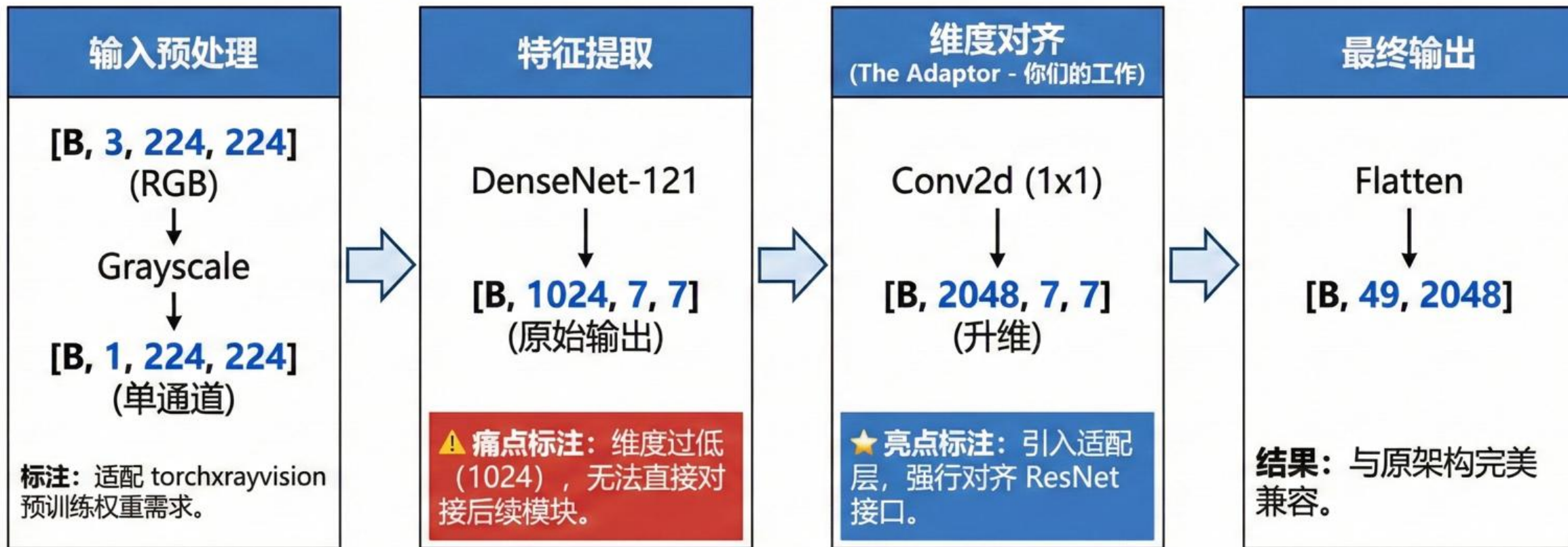
策略： $LR_{\text{Visual}} \ll LR_{\text{Transformer}}$



目的：防止在训练初期破坏 CNN 提取图像特征的能力。



## 阶段一：视觉骨干替换与训练策略调整



## 阶段二：基于相似病例检索的增强生成



**痛点：**单纯看图说话容易产生“幻觉”，且难以生成规范的医学术语



**灵感：**医生在诊断疑难杂症时，往往会参考既往相似病例的报告



**优势：**引入了额外的先验知识，显著提升了报告的专业度和规范性

## 阶段二：基于相似病例检索的增强生成

### 输入分支 A：当前视觉特征

来自 DenseNet

**Tensor: [Batch, 49, 2048]**

说明：原始图片的 49 个 Patch。

### 输入分支 B：检索到的知识 —— [新增]

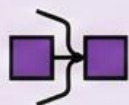
操作：Top-K 检索 → Embedding 映射

**Tensor: [Batch, K, 2048]**

说明：K 代表检索到的相似报告数量（或关键特征数），例如 K=5。

**❗ 关键标注：** 维度对齐：必须映射到 2048 维才能融合。

### 融合操作



动作：`torch.cat(dim=1)`  
(沿序列长度维度拼接)

标注：Sequence  
Concatenation

### 最终输入：

**Tensor:**  
**[Batch, 49 + K, 2048]**

物理意义：Transformer 看到的“输入序列”变长了。它不仅看到了当前的图，还看到了 K 个历史参考信息。

Transformer



## 阶段三：引入强化学习优化指标

### ■ 方法：Self-Critical Sequence Training (SCST)

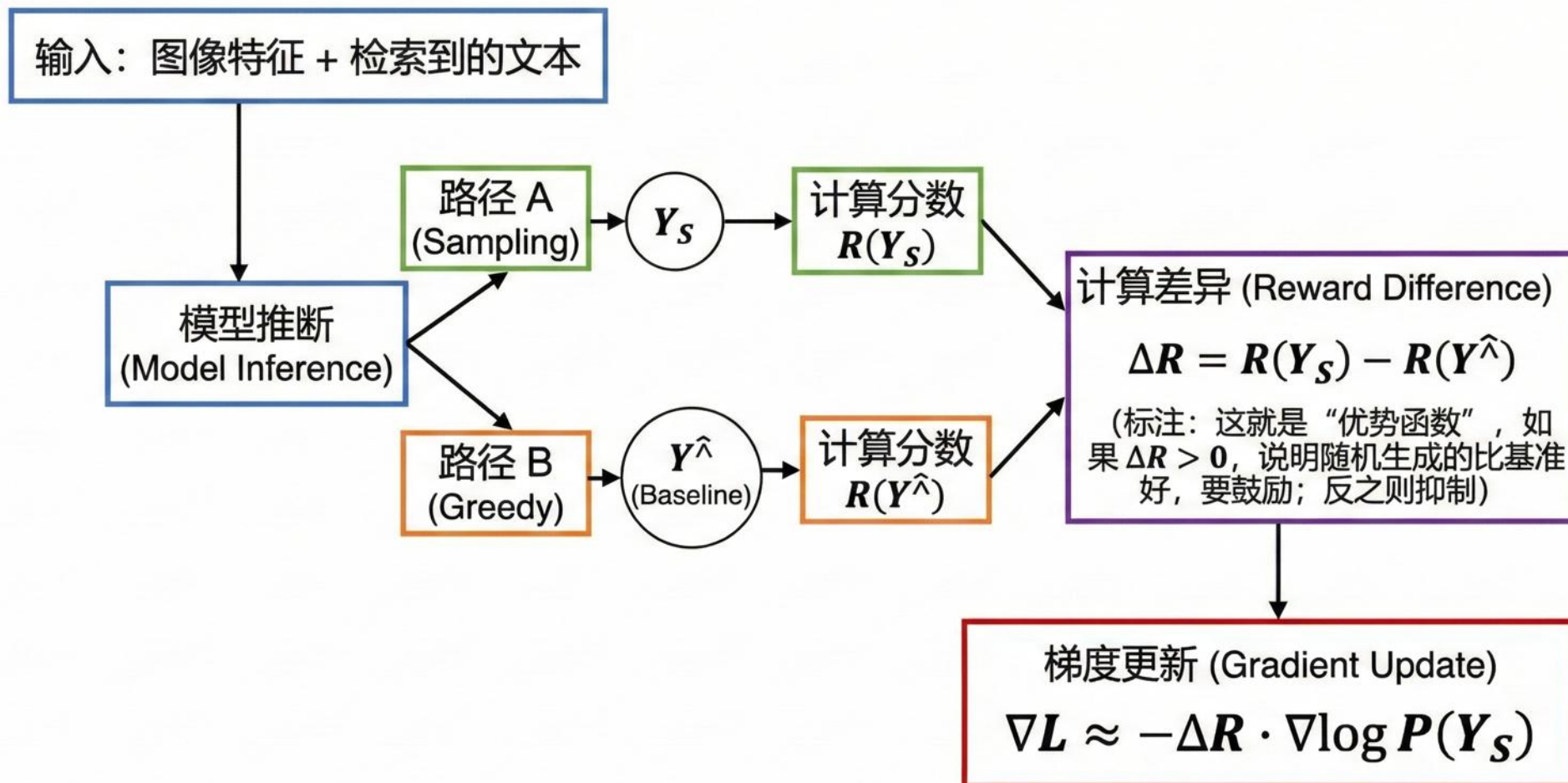
### ■ 原理：



### ■ 实施方式：在 Cross-Entropy 训练收敛的模型基础上，开启 SCST 进行微调。

## 阶段三：引入强化学习优化指标

### Self-Critical Sequence Training (SCST)工作原理



## 阶段三：引入强化学习优化指标

特性	交叉熵损失 (Cross-Entropy)	SCST 损失 (Our Method)
优化目标	最大化“下一个词”的概率	最大化整句的评价指标 (如 CIDEr)
数值范围	$[0, +\infty)$ (概率对数)	取决于 Reward 差值 (可正可负, 非 0-1)
反向传播	链式法则 (直接求导)	REINFORCE 算法 (策略梯度)
核心优势	训练稳定, 快速收敛	直接优化测试指标, 解决暴露偏差



## 对原有模型的改进总结

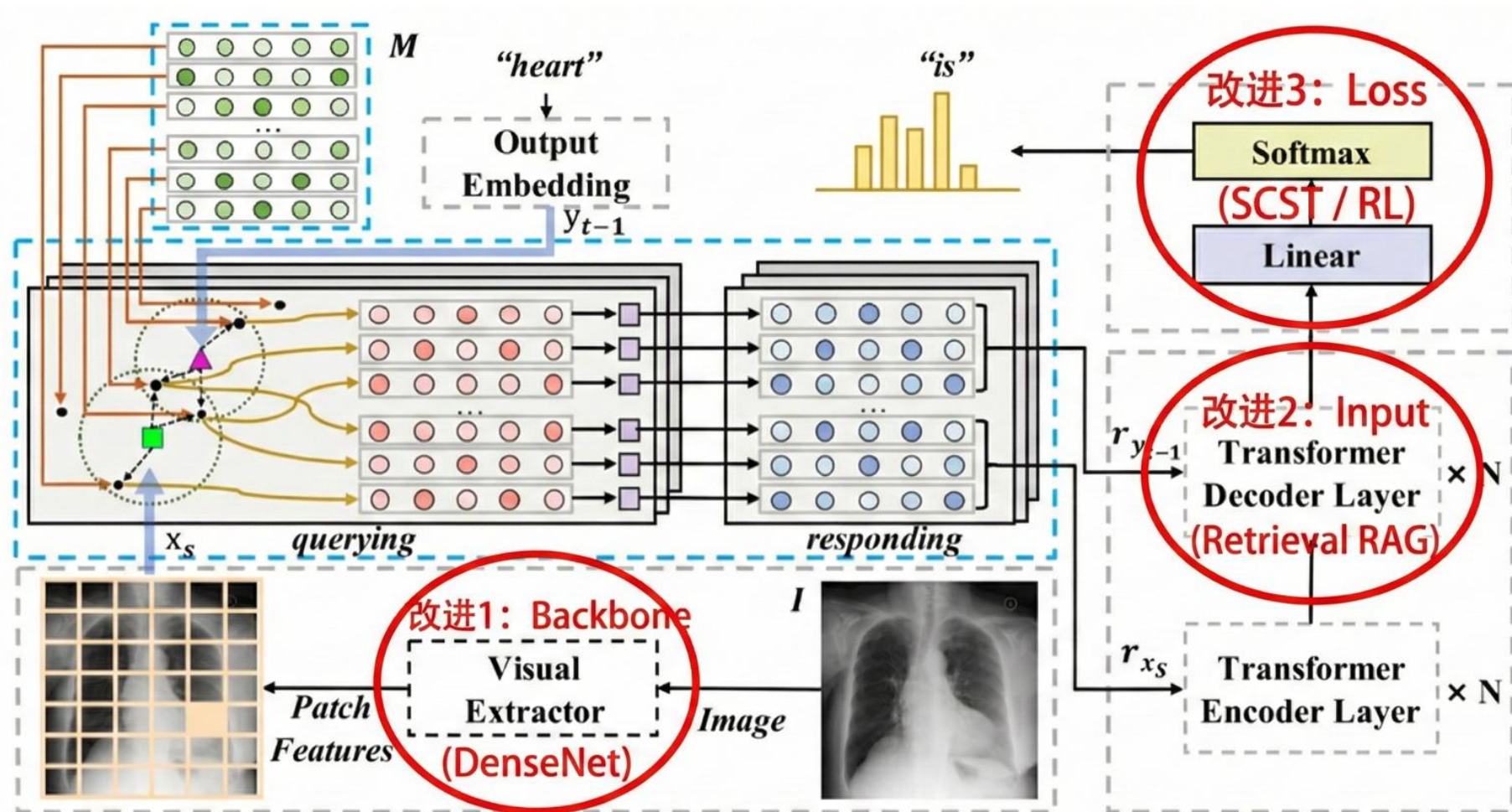


Figure 2: The overall architecture of our proposed approach, where the visual extractor, encoder and decoder are shown in gray dash boxes with the details omitted. The cross-modal memory networks are illustrated in blue dash boxes with presenting the detailed process of memory querying and responding.

## 运行：硬件平台（算力）

**平台：**AutoDL云平台 NVIDIA RTX5090

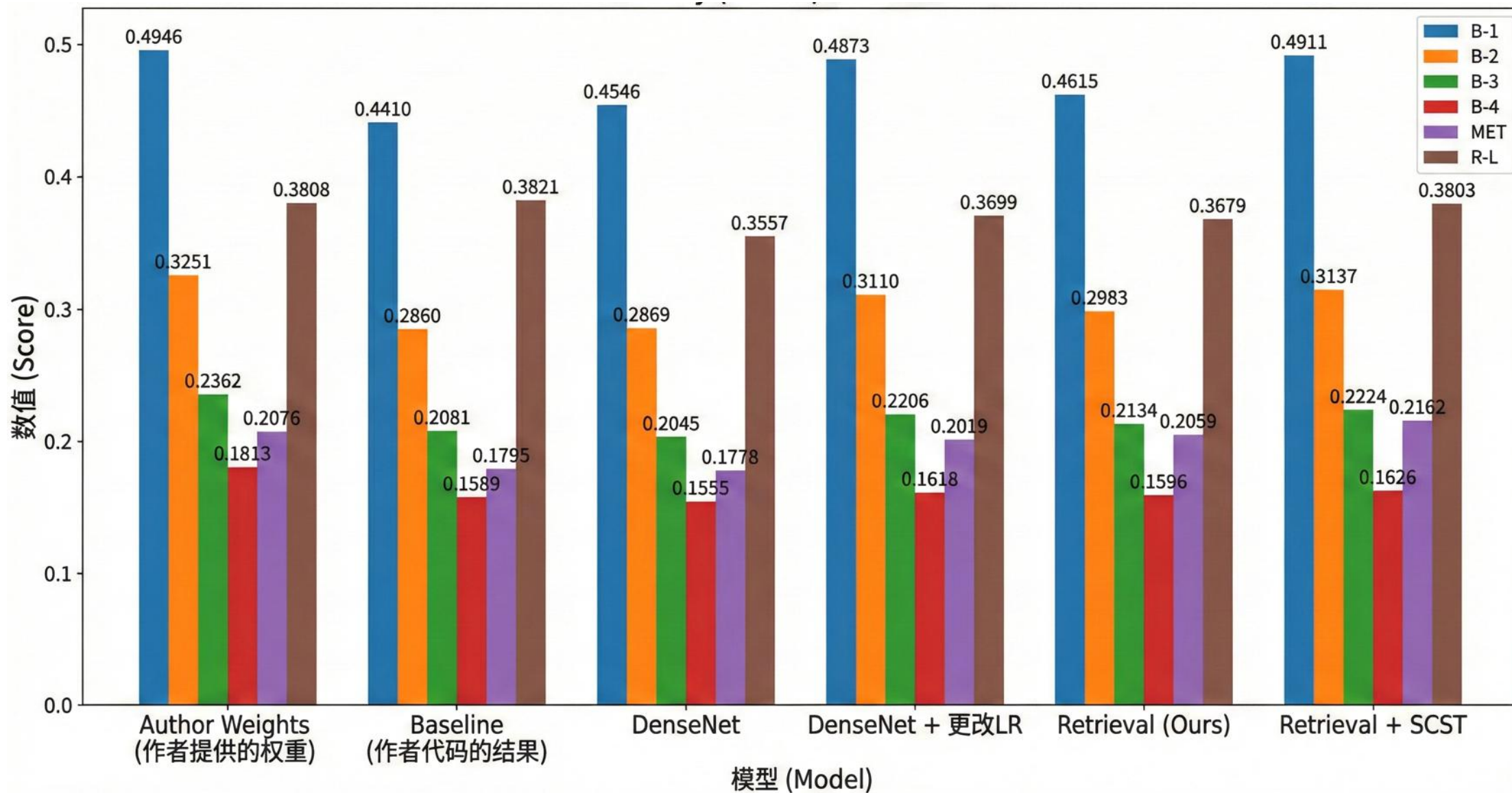
**运行速度：**一轮（100epoch）约1.5小时

结果展示

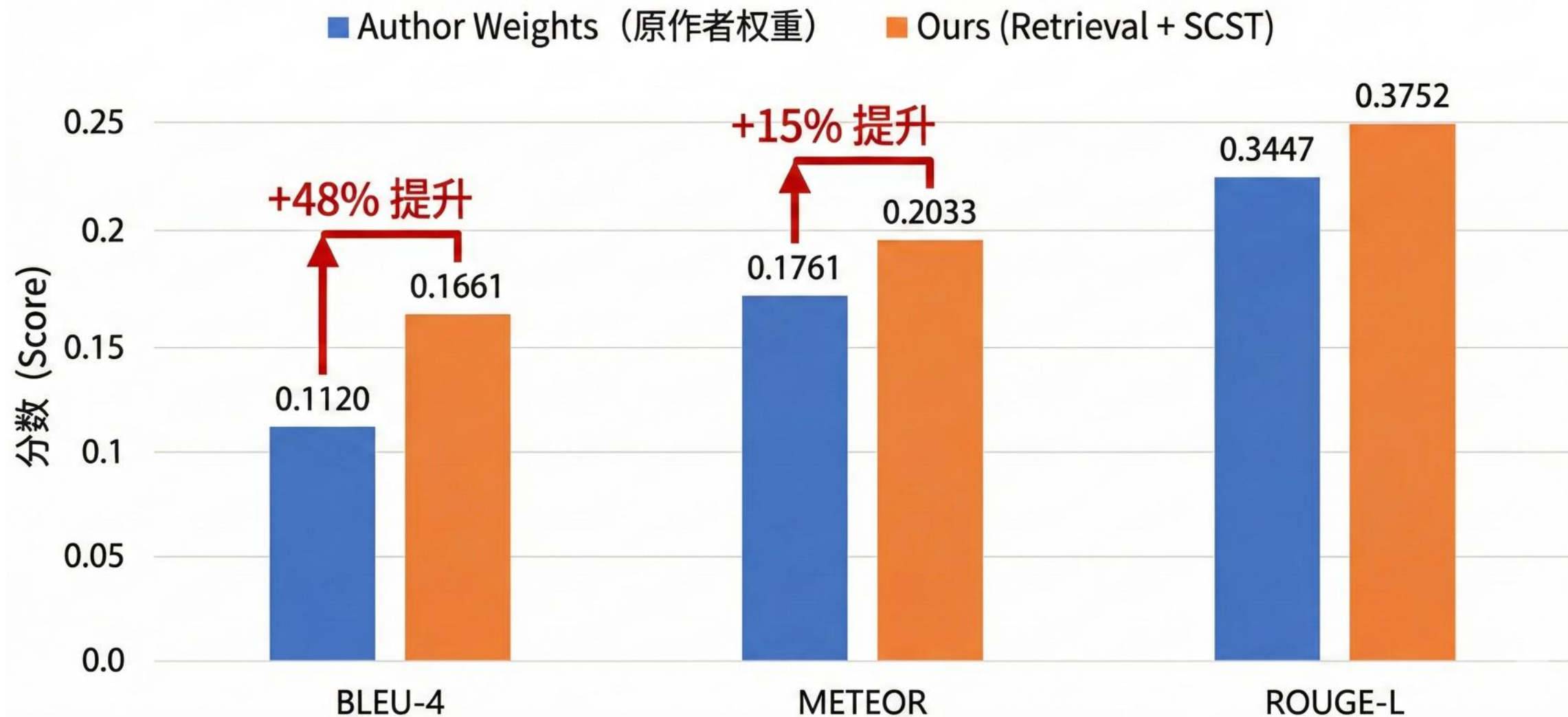
模型 (Model)	IU X-ray (公开集)						老师私有集 (泛化能力)					
	B-1	B-2	B-3	B-4	MET	R-L	B-1	B-2	B-3	B-4	MET	R-L
Author Weights (作者提供的权重)	0.4946	0.3251	0.2362	0.1813	0.2076	0.3808	0.3974	0.2510	0.1635	0.1120	0.1761	0.3447
Baseline (作者代码的结果)	0.4410	0.2860	0.2081	0.1589	0.1795	0.3821	-	-	-	-	-	-
DenseNet	0.4546	0.2869	0.2045	0.1555	0.1778	0.3557	-	-	-	-	-	-
DenseNet + 更改LR	0.4873	0.3110	0.2206	0.1618	0.2019	0.3699	-	-	-	-	-	-
Retrieval (Ours)	0.4615	0.2983	0.2134	0.1596	0.2059	0.3679	0.4116	0.2800	0.2090	0.1661	0.1901	0.3602
Retrieval + SCST	0.4911	0.3137	0.2224	0.1626	0.2162	0.3803	0.4634	0.3076	0.2204	0.1661	0.2033	0.3752



## 结果展示：IU X-ray公开集



## 结果展示：老师的私有集



## 项目核心总结

关键问题	我们的实现与答案
1. X 与 Y 的定义	<ul style="list-style-type: none"><li>• <b>输入 X</b>: 图像张量, Shape = [B, 3, 224, 224]</li><li>• <b>输出 Y</b>: 文本序列, Shape = [B, Seq_Len]</li></ul>
2. 模型架构	<ul style="list-style-type: none"><li>• <b>基座</b>: ResNet-101 (Encoder) + CMN (Memory) + Transformer (Decoder)</li><li>• <b>改进</b>: 引入 Retrieval (检索增强) + SCST (强化学习)</li></ul>
3. 训练集数据	<ul style="list-style-type: none"><li>• <b>来源</b>: IU X-Ray 公开数据集</li><li>• <b>规模</b>: 5229 张图像 (对应2768份报告)</li></ul>
4. 测试集表现	<ul style="list-style-type: none"><li>• <b>来源</b>: IU X-Ray 测试集 (20% 切分)</li><li>• <b>结果</b>: 改进后 BLEU-4 提升 45% (0.112 -&gt; 0.163)</li></ul>
5. 私有集泛化	<ul style="list-style-type: none"><li>• <b>来源</b>: 老师提供的 43 张盲测图片 (对应 22 份报告)</li><li>• <b>结果</b>: 成功生成流畅报告, BLEU-4 相对 Baseline 提升 48%</li></ul>



项目核心总结

关键问题	我们的实现与答案
6. 硬件平台	<ul style="list-style-type: none"><li>• <b>设备:</b> 单卡 NVIDIA GeForce RTX 5090 (24GB)</li></ul>
7. Loss 如何收敛	<ul style="list-style-type: none"><li>• <b>阶段一 (Warm-up):</b> 使用交叉熵 (CE), Loss 快速平稳下降</li><li>• <b>阶段二 (Fine-tuning):</b> 开启 SCST, Loss (负奖励) 虽有震荡, 但 CIDEr 指标持续上升并收敛</li></ul>
8. 对比了哪些算法	<ul style="list-style-type: none"><li>• <b>SOTA 参照:</b> 原作者提供的预训练权重 (Author Weights)</li><li>• <b>内部基线:</b> 复现的原始代码 (Baseline)</li><li>• <b>消融对象:</b> 替换骨干网络的变体 (DenseNet)</li></ul>

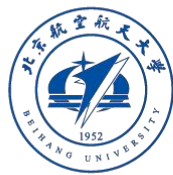
## 局限性反思

### 局限性

- **细微病灶遗漏：** 虽然句式像医生了，但在面对极细微特征（如“主动脉轻微扭曲”）时，模型仍存在漏诊（Recall 问题）。
- **评价指标的偏差：** SCST 虽然刷高了 BLEU/CIDEr 分数，但高分不完全等于“临床正确”。有时候模型为了凑分数，可能会生成一些“万金油”式的废话。
- **依赖训练集分布：** 检索增强非常依赖数据库。如果训练集里根本没有某种罕见病，检索机制也无能为力。

- **从 2D 到 3D：** 尝试处理 CT 数据（如 BrainGPT）。
- **引入 LLM：** 利用大语言模型（如 LLaMA-Med）强大的推理能力，修正“幻觉”，不仅仅是模仿，而是进行推理。

### 未来方向



北京航空航天大学  
BEIHANG UNIVERSITY

**Thanks for listening!**  
**请老师批评指正**