北京航空航天大學
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

北京航空航天大學
BEIHANG UNIVERSITY

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)

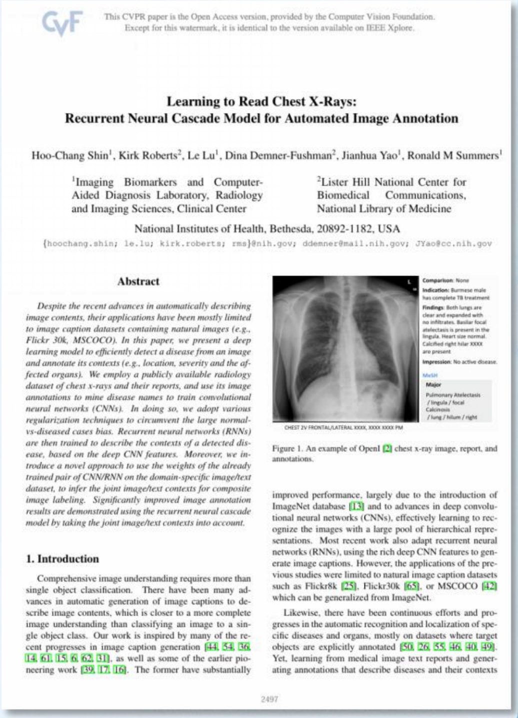
报告中有很多套话模板，但也也必须精准包含关键异常信息。

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

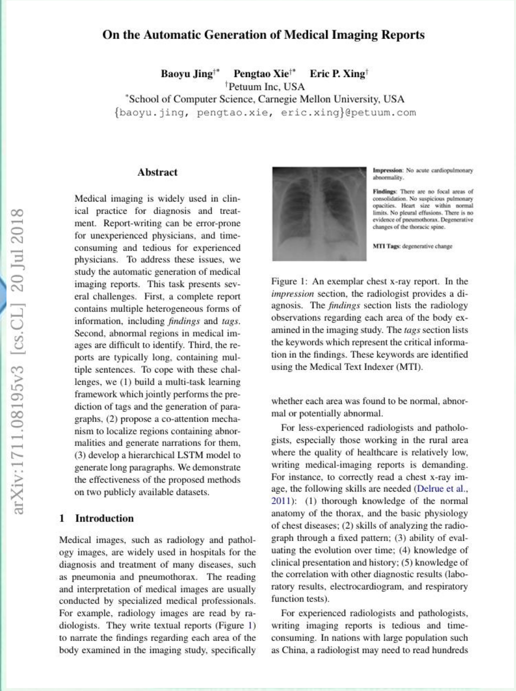
1. 早期探索：CNN-RNN (Early Stage)



解决“有无”问题

（能生成了，但很短）

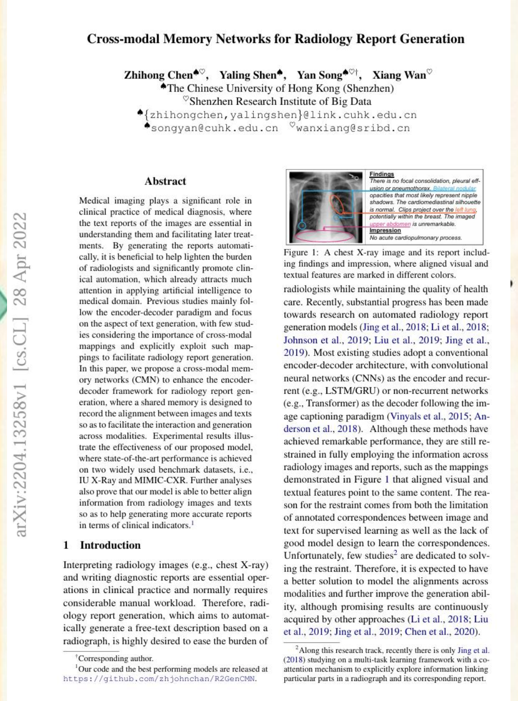
2. 中期发展：Attention Mechanisms (Attention Stage)



解决“对齐”问题

（知道看哪里，准确率提升）

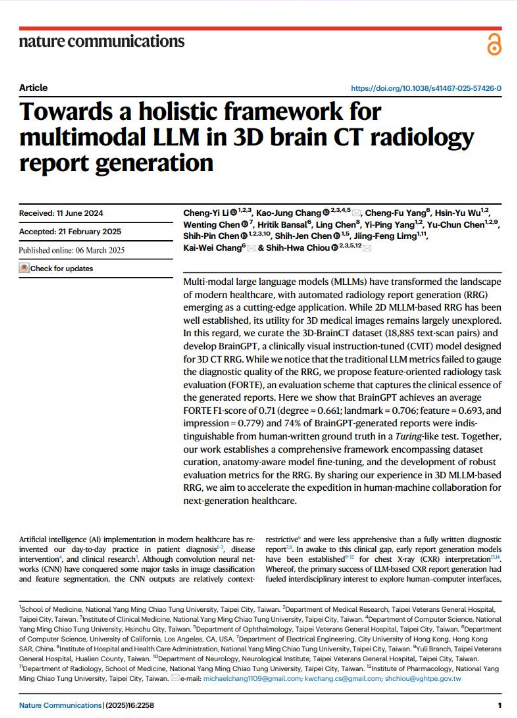
3. Transformer 时代：Memory-driven (Project Core)



解决“模式与长文本”问题

（引入记忆，生成专业长报告）

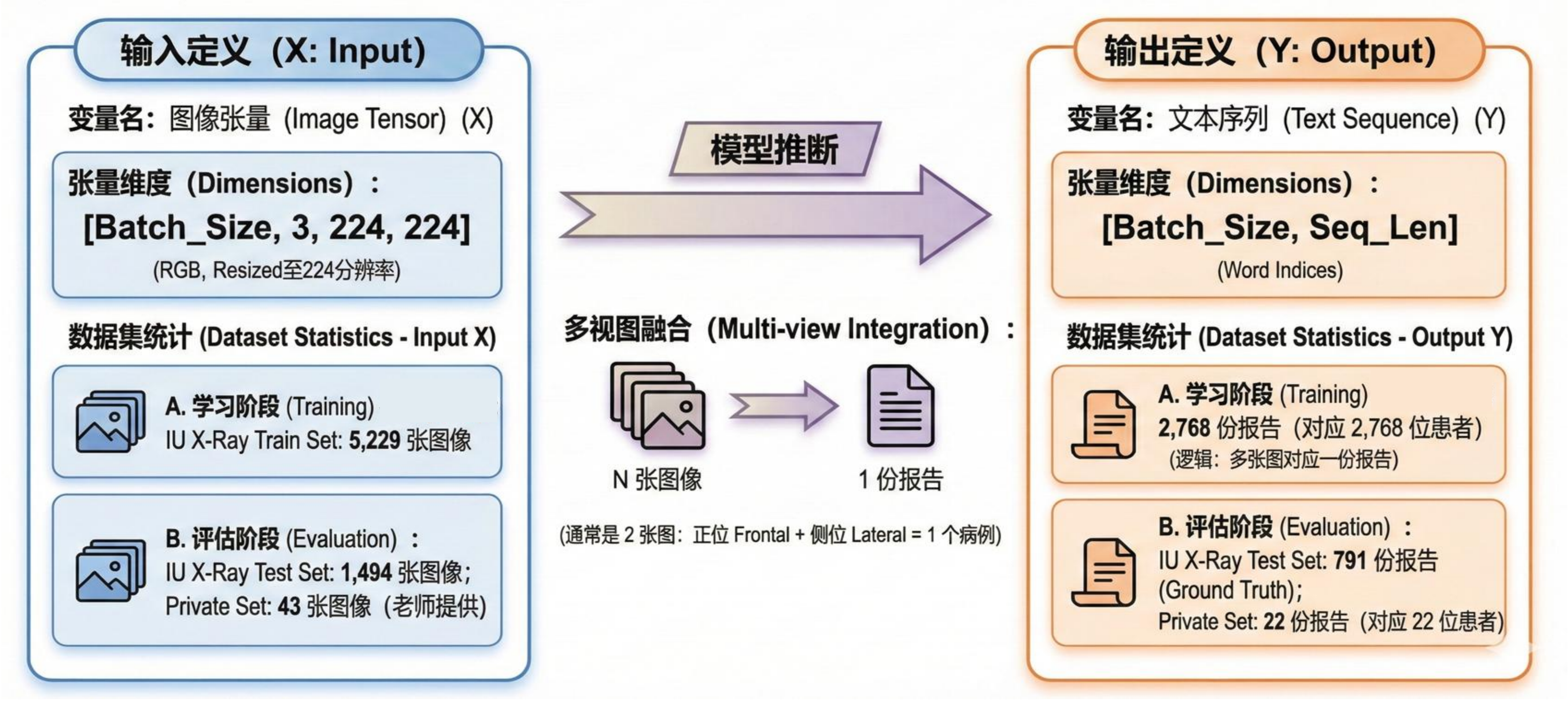
4. 大模型时代：Multimodal LLM (SOTA / Future)



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

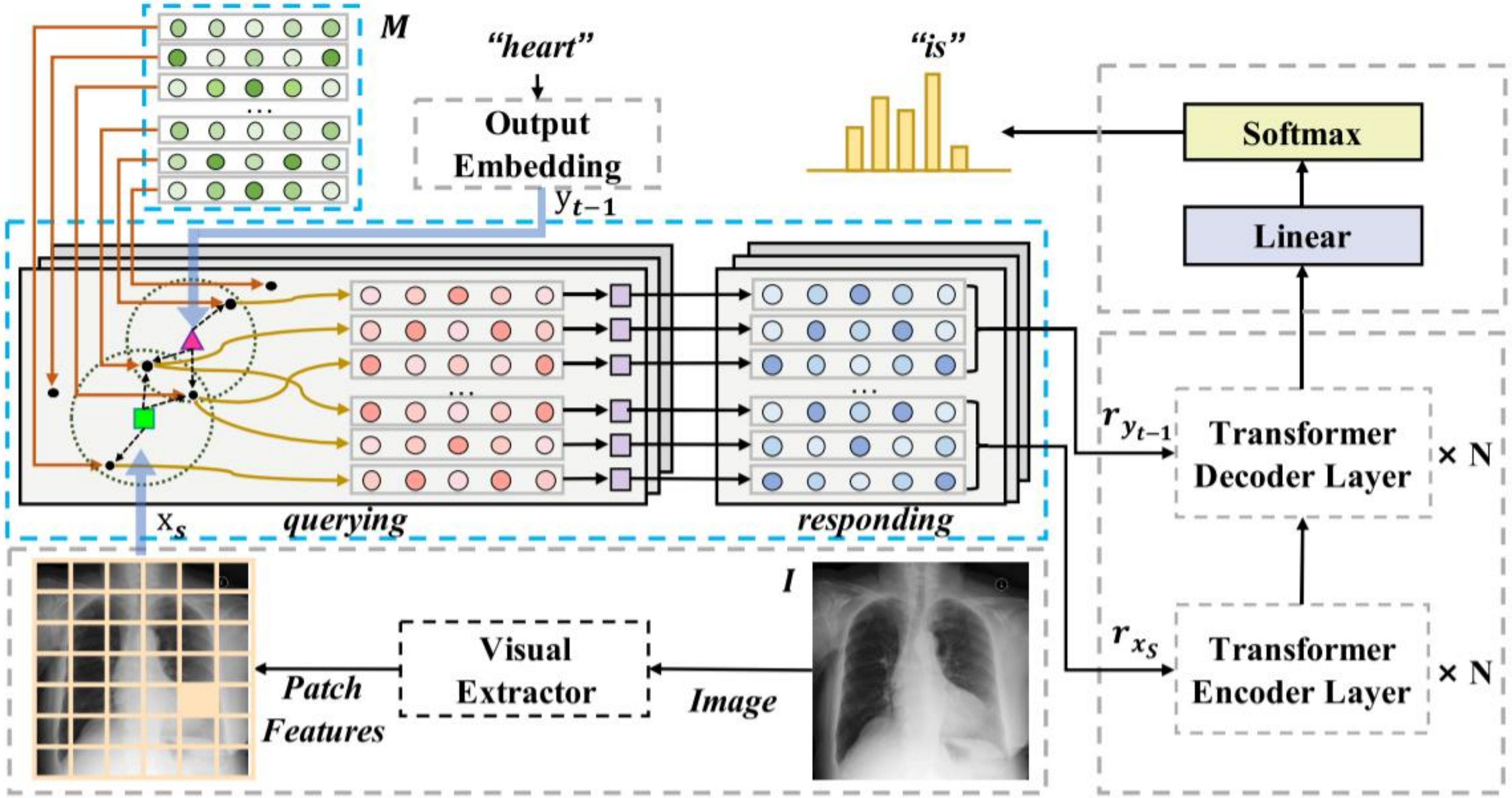
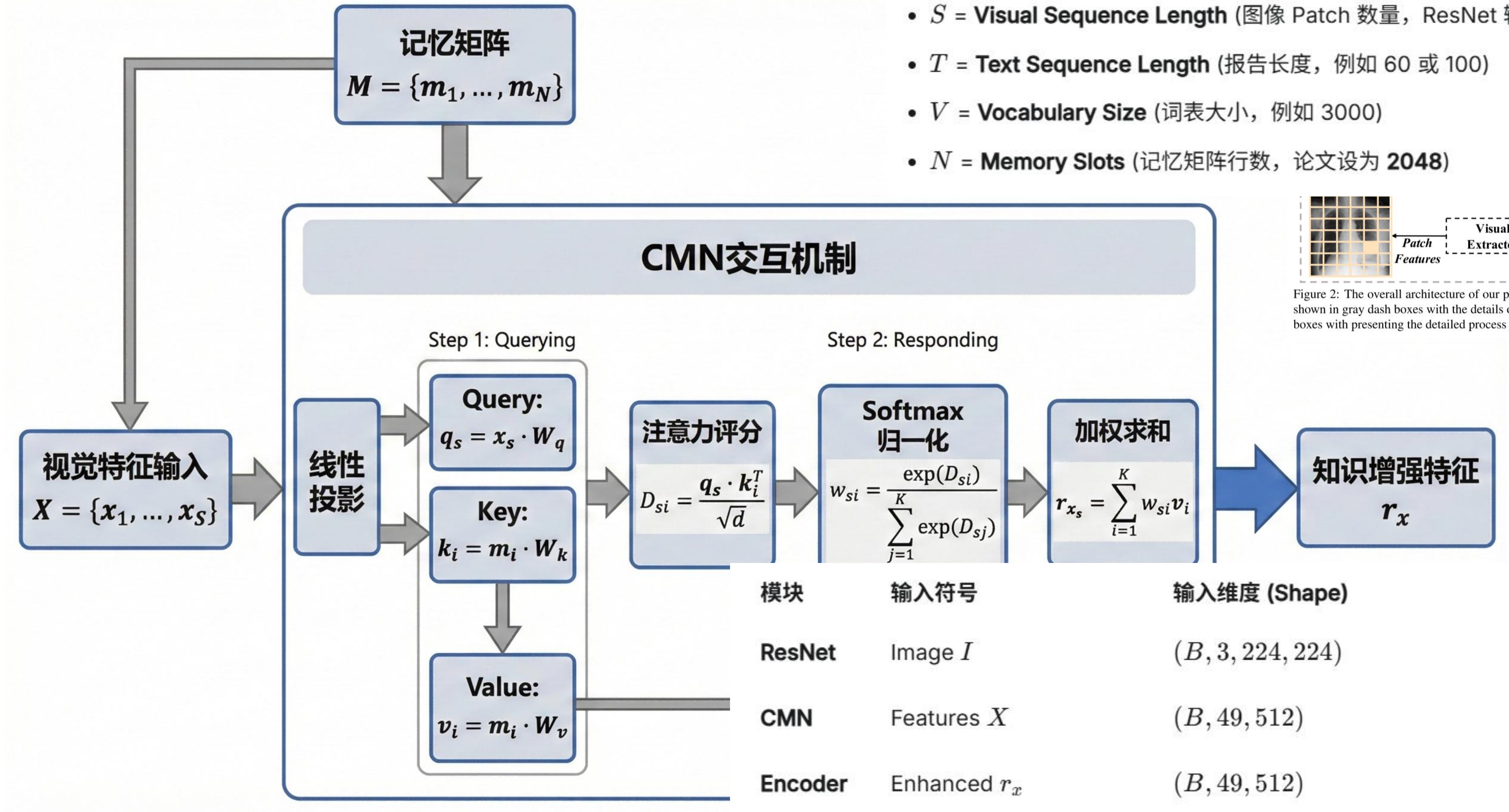


Figure 2: The overall architecture of our proposed approach, where the visual extractor, encoder and decoder are shown in gray dash boxes with the details omitted. The cross-modal memory networks are illustrated in blue dash boxes with presenting the detailed process of memory querying and responding.

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 7×7，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**)

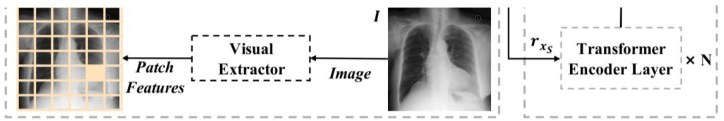
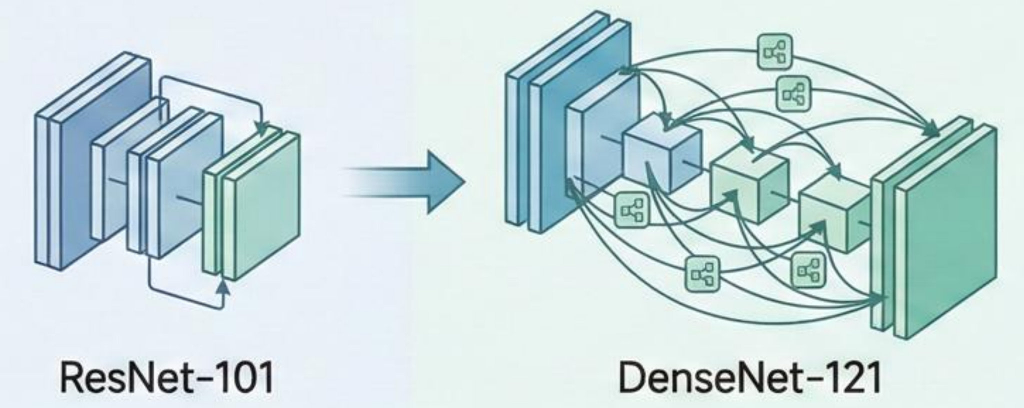


Figure 2: The overall architecture of our proposed approach, where the visual extractor, encoder and decoder are shown in gray dash boxes with the details omitted. The cross-modal memory networks are illustrated in blue dash boxes with presenting the detailed process of memory querying and responding.

模块	输入符号	输入维度 (Shape)	输出符号	输出维度 (Shape)
ResNet	Image I	$(B, 3, 224, 224)$	Features X	$(B, 49, 512)$
CMN	Features X	$(B, 49, 512)$	Enhanced r_x	$(B, 49, 512)$
Encoder	Enhanced r_x	$(B, 49, 512)$	Context Z	$(B, 49, 512)$
Decoder	Context Z Text Y	$(B, 49, 512)$ $(B, T, 512)$	Probs P	(B, T, V)

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

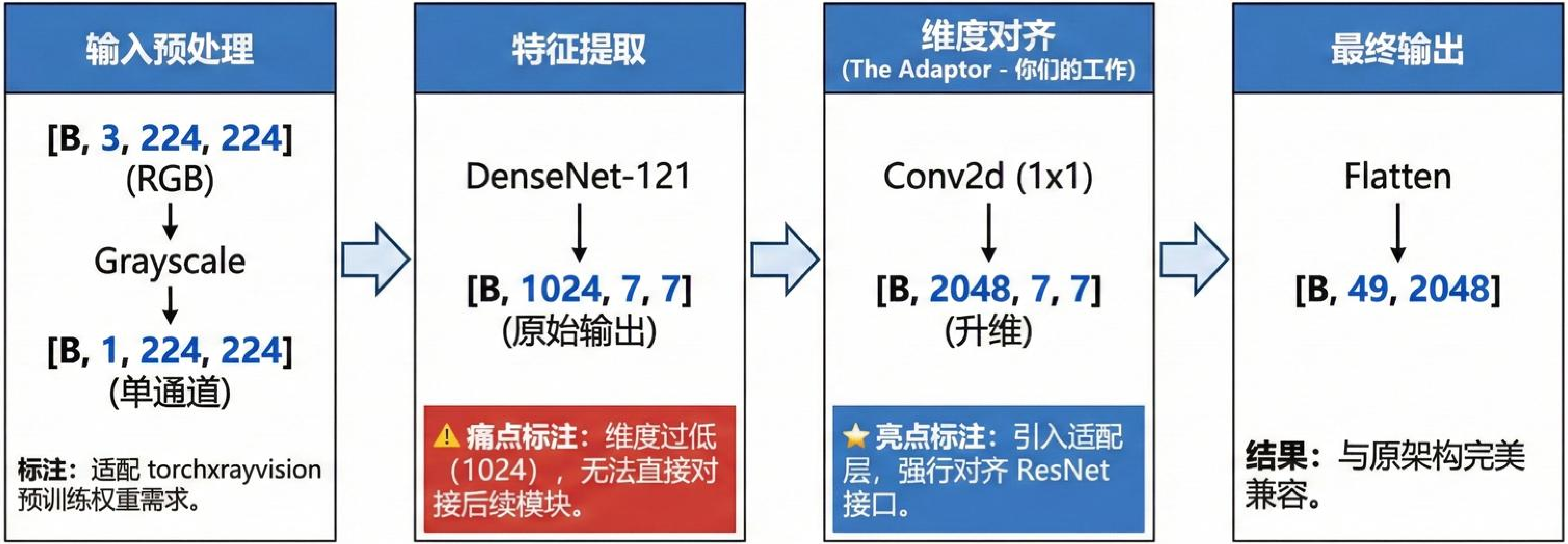
理由： DenseNet 的密集连接特性使其更擅长捕捉医学图像中细微的纹理特征（Feature Reuse），且参数利用率更高。

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



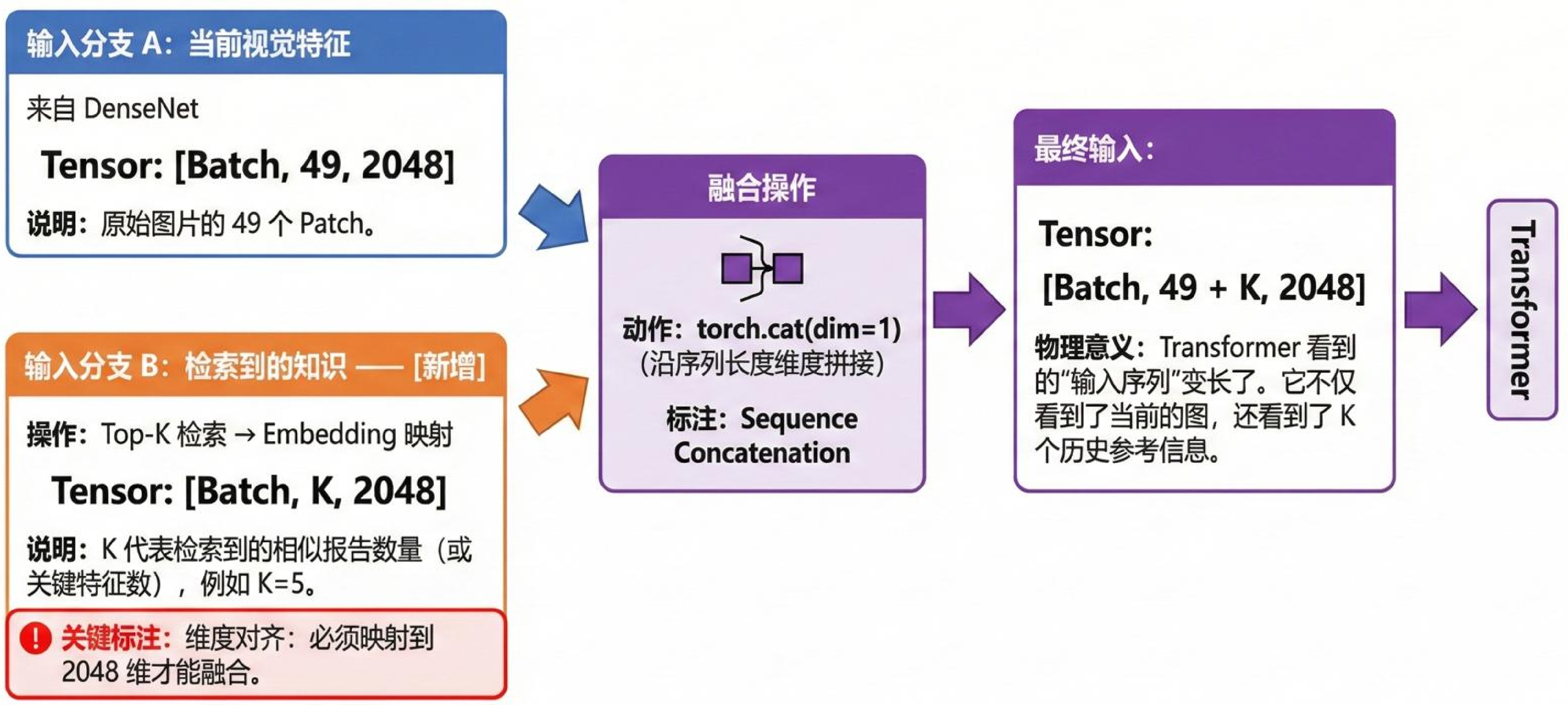
阶段二：基于相似病例检索的增强生成

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



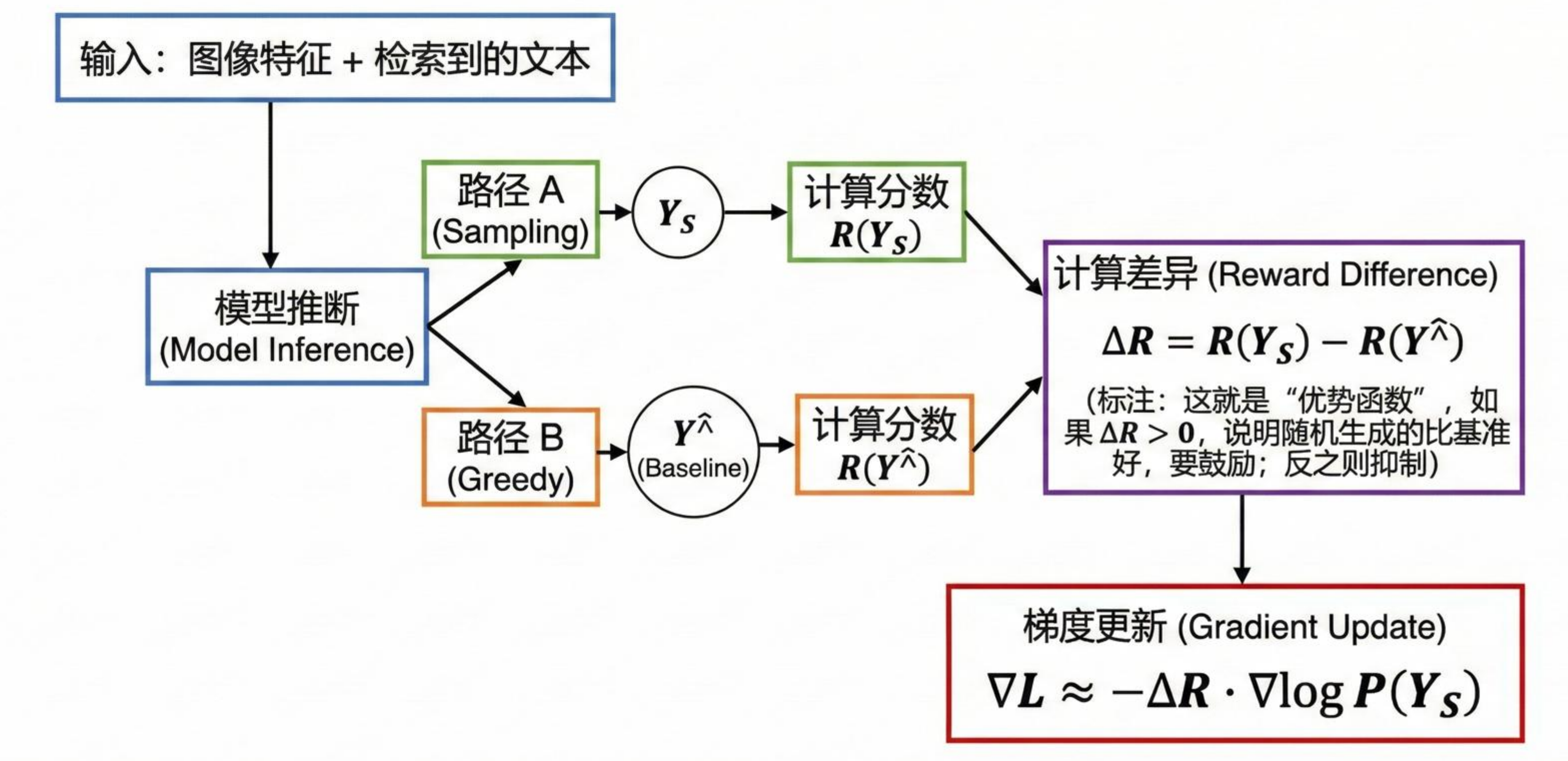
阶段三：引入强化学习优化指标

- **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

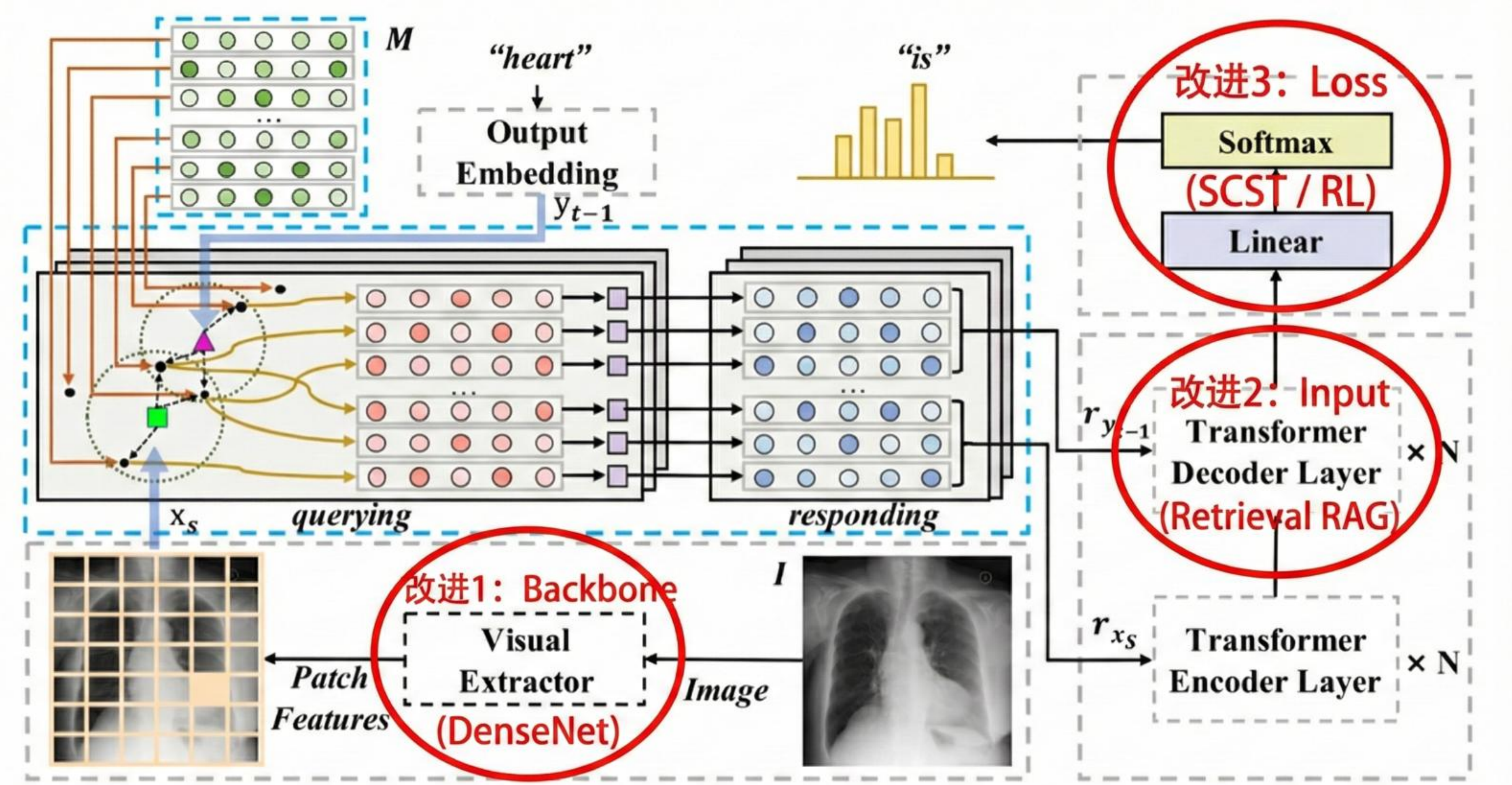


Figure 2: The overall architecture of our proposed approach, where the visual extractor, encoder and decoder are shown in gray dash boxes with the details omitted. The cross-modal memory networks are illustrated in blue dash boxes with presenting the detailed process of memory querying and responding.

运行：硬件平台（算力）

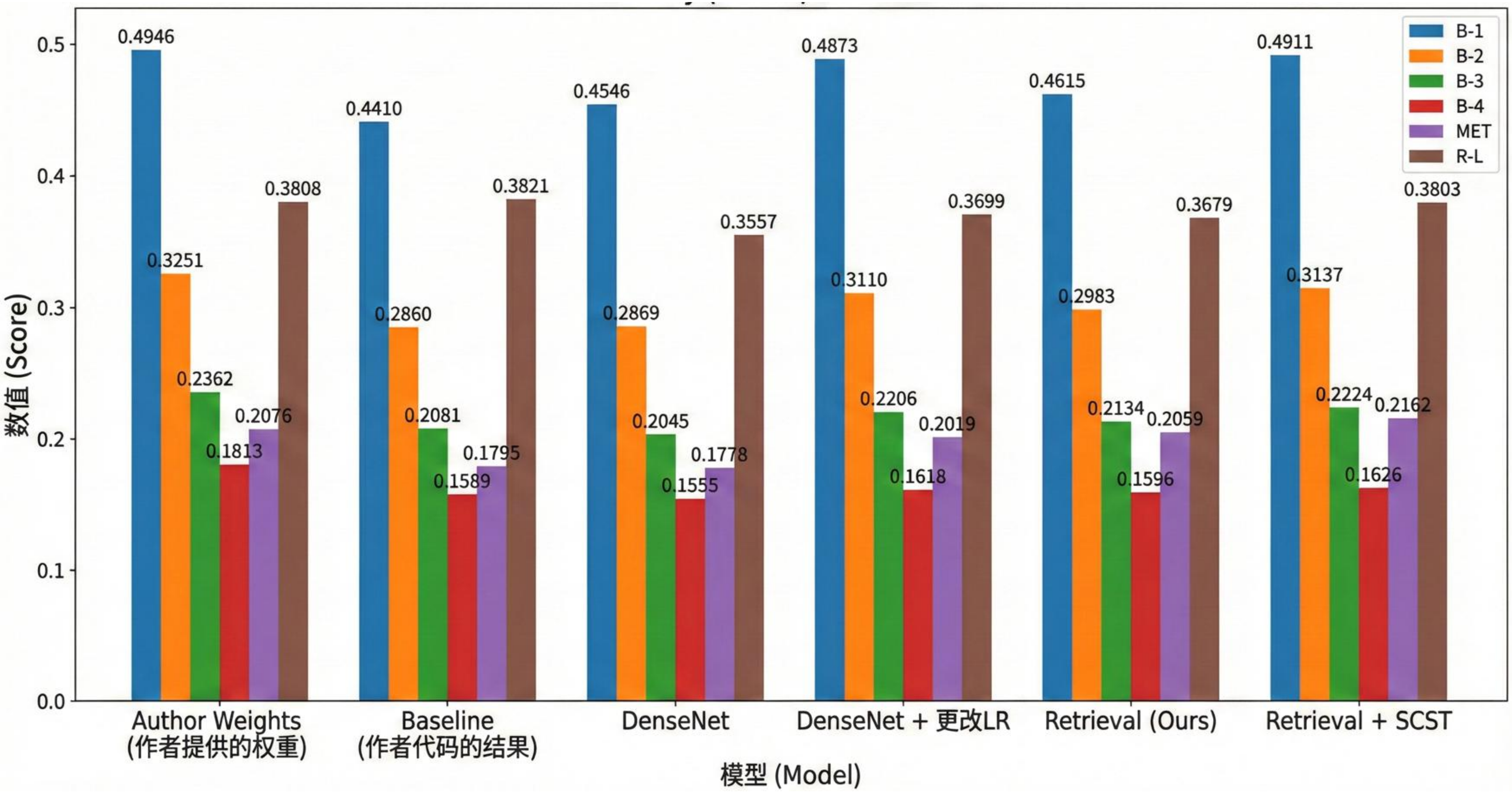
平台：AutoDL云平台 NVIDIA RTX5090

运行速度：一轮（100epoch）约1.5小时

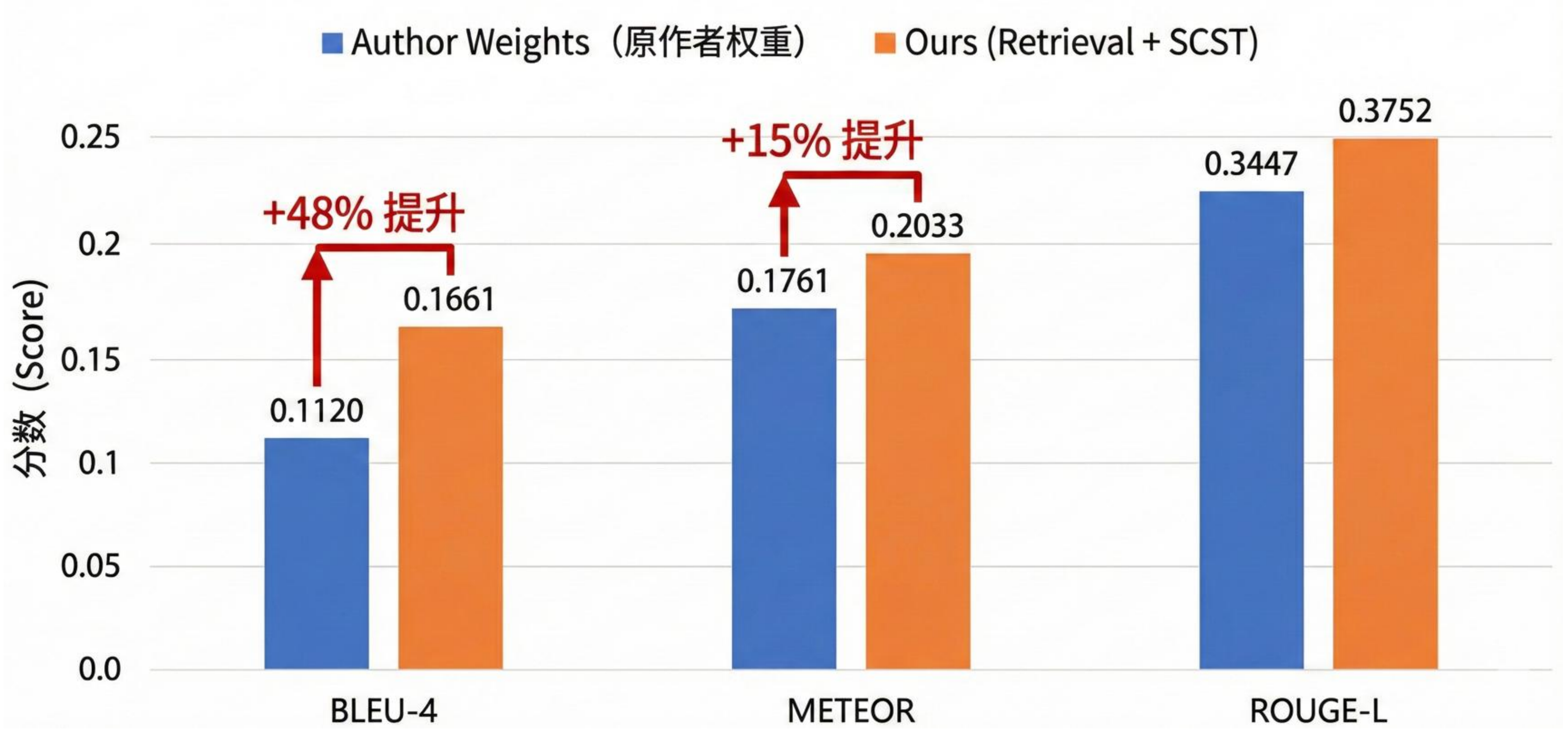
结果展示

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


Author Weights（原作者权重）
Ours (Retrieval + SCST)
分数（Score）
0.25
0.2
0.15
0.1
0.05
0.0
+48% 提升
+15% 提升
0.1120
0.1661
0.1761
0.2033
0.3447
0.3752
BLEU-4
METEOR
ROUGE-L

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	• **输入 X：** 图像张量，Shape = [B, 3, 224, 224] • **输出 Y：** 文本序列，Shape = [B, Seq_Len]
2. 模型架构	• **基座：** ResNet-101 (Encoder) + CMN (Memory) + Transformer (Decoder) • **改进：** 引入 Retrieval (检索增强) + SCST (强化学习)
3. 训练集数据	• **来源：** IU X-Ray 公开数据集 • **规模：** 5229 张图像 （对应2768份报告）
4. 测试集表现	• **来源：** IU X-Ray 测试集 (20% 切分) • **结果：** 改进后 BLEU-4 提升 45% (0.112 -> 0.163)
5. 私有集泛化	• **来源：** 老师提供的 43 张盲测图片 (对应 22 份报告) • **结果：** 成功生成流畅报告，BLEU-4 相对 Baseline 提升 48%

项目核心总结

关键问题	我们的实现与答案
6. 硬件平台	• **设备:** 单卡 NVIDIA GeForce RTX 5090 (24GB)
7. Loss 如何收敛	• **阶段一 (Warm-up):** 使用交叉熵 (CE)，Loss 快速平稳下降 • **阶段二 (Fine-tuning):** 开启 SCST，Loss (负奖励) 虽有震荡，但 CIDEr 指标持续上升并收敛
8. 对比了哪些算法	• **SOTA 参照:** 原作者提供的预训练权重 (Author Weights) • **内部基线:** 复现的原始代码 (Baseline) • **消融对象:** 替换骨干网络的变体 (DenseNet)

局限性与反思

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。

未来方向

- **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。

北京航空航天大學
BEIHANG UNIVERSITY

Thanks for listening!
请老师批评指正