Computer Science and Application 计算机科学与应用, 2026, 16(4), 378-395
Published Online April 2026 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2026.164138

Hans汉斯

面向医疗视觉问答的动态问题编码与知识对比推理模型

张澳斌

长沙理工大学数学与统计学院，湖南 长沙

收稿日期：2026年3月23日；录用日期：2026年4月21日；发布日期：2026年4月29日

摘　要

医疗视觉问答(Med-VQA)旨在基于医学图像与自然语言问题预测可信且准确答案。现有方法主要依赖医学图像特征解析，缺乏对问题语义的深入建模，未能充分考虑开放式与封闭式问题在语义理解上的差异需求。此外，面对多义性强、语境依赖性高的医疗问题，文本查询往往缺乏足够的描述性内容，仅依赖图像和文本特征的融合会导致跨模态对齐不足。针对上述问题，本文提出动态问题编码与知识对比推理(DKCR)模型。该模型利用动态问题编码模块根据问题类型动态建模，既增强开放式问题的语义表征，又避免对封闭式问题引入冗余特征。为减轻跨模态语义偏差，DKCR的知识对比学习通过知识－图像与知识－问题两条路径联合约束潜在表征空间，促进跨模态特征的一致性与互补性。同时，引入外部知识以弥补查询语义不足，丰富问题表征，并通过知识驱动的引导注意力机制，实现视觉、文本与知识模态的深度交互与细粒度对齐。在VQA-RAD与SLAKE数据集上的实验结果表明，DKCR在整体性能上优于现有方法，消融研究进一步验证了各模块的独立价值与协同增益。

关键词

医疗视觉问答，外部医学知识，知识驱动学习，对比学习，引导注意力

Dynamic Question Encoding and Knowledge Contrastive Reasoning Model for Medical Visual Question Answering

Aobin Zhang

School of Mathematics and Statistics, Changsha University of Science and Technology, Changsha Hunan

Received: March 23, 2026; accepted: April 21, 2026; published: April 29, 2026

文章引用: 张澳斌. 面向医疗视觉问答的动态问题编码与知识对比推理模型[J]. 计算机科学与应用, 2026, 16(4): 378-395. DOI: 10.12677/csa.2026.164138

Abstract

Medical Visual Question Answering (Med-VQA) is a task that aims to predict reliable and accurate answers based on medical images and natural language questions. Existing methods primarily rely on the analysis of medical image features, lacking in-depth modeling of question semantics and failing to fully consider the distinct semantic understanding requirements of open-ended and closed-ended questions. Furthermore, medical questions often exhibit strong ambiguity and high context dependence, where textual queries frequently lack sufficient descriptive content. Solely relying on the fusion of image and text features leads to insufficient cross-modal alignment. To address these issues, this paper proposes a Dynamic Question Encoding and Knowledge Contrastive Reasoning Model (DKCR) model. This model employs a dynamic question encoding module to adaptively model questions based on their types, enhancing semantic representation for open-ended questions while avoiding the introduction of redundant features for closed-ended ones. To mitigate cross-modal semantic bias, DKCR's knowledge contrastive learning constrains the latent representation space through two pathways—knowledge-image and knowledge-question—promoting consistency and complementarity of cross-modal features. Concurrently, external knowledge is incorporated to compensate for insufficient query semantics and enrich question representation. A knowledge-driven co-attention mechanism facilitates deep interaction and fine-grained alignment among visual, textual, and knowledge modalities. Experimental results on the VQA-RAD and SLAKE datasets demonstrate that DKCR outperforms existing methods in overall performance, and ablation studies further validate the individual value and synergistic benefits of each module.

Keywords

Medical Visual Question Answering, External Medical Knowledge, Knowledge Driven Learning, Contrastive Learning, Guided Attention

1. 引言

医疗视觉问答(Med-VQA)是一项融合计算机视觉与自然语言处理的任务，其目标是在给定医学图像及相关问题的条件下生成准确的答案。该任务能够辅助医生诊断、提供第二意见并提升临床效率，从而降低误诊风险[1]。

现有研究主要聚焦于医学图像特征建模，而 Med-VQA 同样依赖问题文本的深层语义解析。为此，AMAM [2]通过对齐图文注意力来定位问题中的关键字，从而在语义层面实现多模态信息融合。MAMF [3]等人则提出多级多模态语义表示，以获得单词级的细粒度特征。然而，这些方法普遍对开放式与封闭式问题采用统一编码策略，忽视了两类任务在语义建模上的差异。相较于多为判断式的封闭式问题，开放式问题更依赖于图像与文本的深层语义对齐与建模，统一的编码方式可能导致开放式问题语义表征不足或对封闭式问题引入冗余特征[4]。

与此同时，目前公开的 Med-VQA 数据集规模有限，医学图像标注过程依赖专业知识且成本高昂，这进一步制约了深度学习模型的训练与泛化能力。为缓解数据稀缺问题，一些研究尝试借助医学图像描述数据进行预训练，再在目标数据集上进行微调[5]。例如，CPCR [6]利用外部补充数据增强图像解析能

力，M2I2 [7]通过自监督学习从图像与语义模态中提取临床特征。但现有方法忽略了外部医学知识的引入，嵌入结构化医学知识不仅能丰富问题语义表征，还能提供临床先验约束，从而辅助解析图像中的潜在病理特征[8]。

事实上，Med-VQA 不仅面临数据稀缺与语义表征能力不足的挑战，其本质更是一个高度复杂的跨模态推理任务，核心难点在于实现图像与文本之间的有效语义对齐[9]。医学图像通常包含丰富的临床信息，而问题文本往往缺乏上下文细节，这种信息不对称性极易导致模态间的语义偏差，进而削弱模型的推理能力[10]。并且现有方法多依赖拼接或浅层融合机制，未能充分解决深层次模态对齐问题，最终限制模型整体性能[1]。

针对上述挑战，本文提出了一种用于医学视觉问答的动态问题编码与知识对比推理(DKCR)模型。该模型设计了一个动态问题编码模块(DQE)，可根据问题类型动态选择编码路径来增强语义表征；同时，引入外部医学知识以丰富问题表示，并通过知识－图像与知识－问题对比学习缓解模态间语义偏差。此外，构建了知识引导注意力网络，利用知识的互补作用增强跨模态交互，从而提升模型的语义对齐效果与推理能力。

本研究的贡献可以概括为三个方面：

- 提出了一种 DQE 机制，该机制根据问题类型动态调整问题的编码方式，从而增强模型的文本语义理解与表征能力；
- 结合外部医学知识，采用对比学习来缓解跨模态语义偏见，并进一步设计了一个知识引导注意力网络，该网络利用知识的互补作用来促进跨模态特征对齐和推理；
- 提出的 DKCR 模型在 VQA-RAD 和 SLAKE 数据集上取得了较好的性能，消融实验验证了每个模块的有效性。

2. 方法论

与现有的视觉问答方法一样，给定医学图像 V 和基于该图像的问题 Q，输出概率最高的答案 $\hat{a}$。描述如下：

$$\hat{a} = \arg\max_{a \in \mathcal{A}} P(a \mid V, Q, \theta), \tag{1}$$

其中 $\mathcal{A}$ 是包含所有候选答案的集合，θ 表示模型的所有参数。

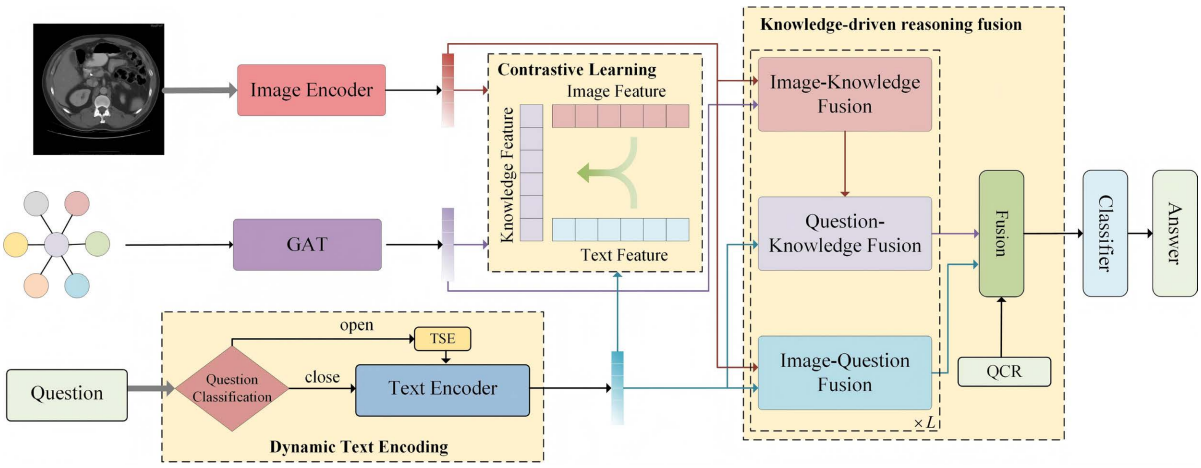

Figure 1. Overall structural framework of the DKCR model
图 1. DKCR 模型整体结构框架

本文提出的 DKCR 模型总体架构如图 1 所示，主要包含三个关键部分：(1) DQE 模块根据问题类型动态调整问题的表征方式；(2) 知识对比学习的多模态语义对齐模块，利用知识－图像和知识－问题对比学习实现特征间的语义互补与增强；(3) 知识引导注意力网络，融合外部医学知识引导模型在多模态信息之间进行更精确的推理与匹配。以下各小节将详细解释这三部分。

2.1. 动态问题编码

医学视觉问答中的问题在语义结构、答案空间及推理需求上存在显著差异。其中，封闭式问题通常对应判断式或有限候选式回答，所需语义上下文相对集中；而开放式问题往往涉及病灶属性、解剖位置、数量信息或影像征象描述，对语义理解深度和跨模态推理能力提出了更高要求。若对两类问题统一采用相同的编码与推理方式，容易导致开放式问题建模不足，或为封闭式问题引入不必要的表征冗余，从而限制整体性能。

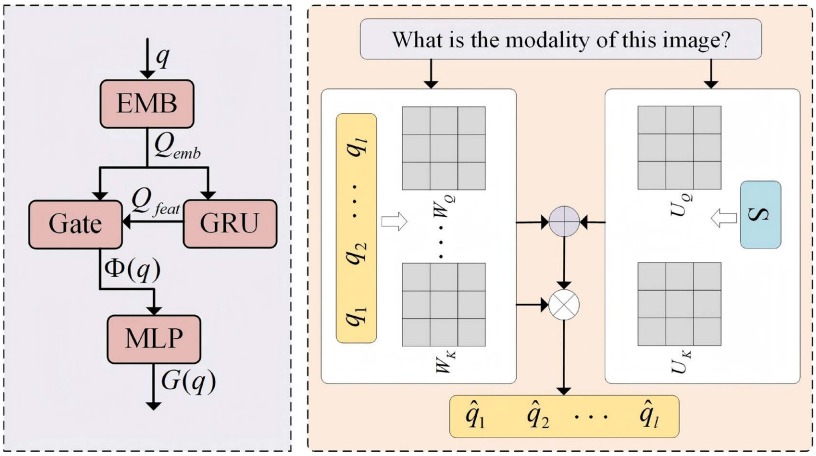

Figure 2. Dynamic question encoding module
图 2. 动态问题编码模块

基于上述考虑，本文提出 DQE 模块，以匹配其语义理解深度和推理需求。具体而言，我们首先引入一个轻量级、可学习的问题类型分类器 G，以输入问题文本为条件，自动预测其属于封闭式问题还是开放式问题，并据此选择匹配的后续编码分支。需要说明的是，该分类过程并非依赖人工设定的关键词规则，而是通过对问题语义表示的学习实现数据驱动的类别判别。尽管从表层语言形式上看，封闭式问题往往呈现判断式表达，而开放式问题更多体现为属性、数量或位置查询，但这些现象仅反映两类问题在语言分布上的统计差异，并非模型进行类别决策的硬性依据。如图 1 中的菱形模块所示，其具体结构见图 2 左侧。具体地，长度为 l 的问题字符串 q 被预训练的 Glove 映射为词嵌入序列，令 $q_i \in R^{d_q}$ 为第 i 个词的嵌入向量，则词嵌入为：

$$Q_{emb} = \text{WordEmbedding}(q) = [q_1, \cdots, q_l], \tag{2}$$

词嵌入 $Q_{emb} \in R^{d_q \times l}$ 由 GRU 处理得到问题嵌入，通过门控机制得到输出 $\mathcal{G} \in R^{d_G \times l}$：

$$Q_{feat} = \text{GRU}(Q_{emb}) = [\eta_1, \cdots, \eta_l], \tag{3}$$

$$\tilde{Q} = [Q_{emb} \| Q_{feat}], \tag{4}$$

$$Y = \tanh(W_1 \tilde{Q}), \tag{5}$$

$$\tilde{Y}=\sigma\left(W_2\tilde{Q}\right), \tag{6}$$

$$\mathcal{G}=Y\circ\tilde{Y}, \tag{7}$$

其中 $Q_{feat}\in R^{d_G\times l}$，$\eta_i$ 表示第 i 个词的嵌入，$\|$表示特征拼接，$\tilde{Q}\in R^{\left(d_q+d_G\right)\times l}$，$W_1,W_2\in R^{d_G\times\left(d_q+d_G\right)}$，$\sigma$ 和 tanh 分别表示 Sigmoid 函数和双曲正切激活函数，$\circ$为哈达玛积。

门控输出 $\mathcal{G}$ 映射为问题嵌入 $\Phi(q)$，使用多层感知机将 $\Phi(q)$ 映射到二值分类中，最终得到二元问题类型分类器 G 的输出，公式如下：

$$\alpha=\mathrm{softmax}\left(\left(W_\alpha\mathcal{G}\right)^{\mathrm{T}}\right), \tag{8}$$

$$\Phi(q)=Q_{feat}\alpha, \tag{9}$$

$$p^t=\mathrm{softmax}\left(\mathrm{MLP}\left(\Phi(q)\right)\right), \tag{10}$$

$$G(q)=\begin{cases}0, & \text{if } p_0^t>p_1^t,\\ 1 & \text{else},\end{cases} \tag{11}$$

其中 p^t 为二值分类概率，p_0^t 和 p_1^t 分别表示封闭式和开放式问题概率。

尽管词级嵌入能够捕捉问题中的局部关键词信息，但其对长距离依赖关系和整体句意的建模能力仍然有限，尤其是在开放式医学视觉问答中，问题往往包含病灶属性、解剖位置或数量描述等复合语义，仅依赖词级表示容易导致全局语义约束不足，从而削弱后续跨模态推理的准确性。为此，本文提出句词联合编码模块(SWE)，结构如图 2 右侧所示，在保留词级细粒度特征的基础上，引入句子级全局语义信息对词间关系建模进行调制，从而生成更具判别性的文本表示。考虑到开放式问题通常需要更丰富的语义补充与更深层的推理支持，SWE 专用于开放式问题分支；而对于语义模式相对简单、答案空间较小的封闭式问题，则仍采用 BERT 编码路径，以在增强复杂问题建模能力的同时避免对简单问题引入不必要的冗余表征。

对于给定问题 Q 由公式(2)得到词级表示 $q=\left[q_1,\cdots,q_l\right]\in R^{l\times d_q}$，句子表示通过使用预训练的 Sentence-BERT 提取 $S=\mathrm{BERT}(Q)\in R^{l\times d_q}$。与仅基于词级特征计算自注意力不同，本文进一步将句子级全局语义引入词–词关系建模过程，以缓解词级编码在整体语义约束上的不足。具体而言，如图 2 右侧所示，我们在注意力打分中同时考虑词级交互项与句级语义项，其计算形式为：

$$\alpha_{ij}=\frac{1}{\sqrt{2d}}\left(q_iW_Q\right)\left(q_jW_K\right)^{\mathrm{T}}+\frac{1}{\sqrt{2d}}\left(SU_Q\right)\left(SU_K\right)^{\mathrm{T}}, \tag{12}$$

$$\hat{q}_i=\sum_{j=1}^{l}\frac{\exp\left(\alpha_{ij}\right)}{\sum_{j'=1}^{l}\exp\left(\alpha_{ij'}\right)}\left(q_jW_V\right), \tag{13}$$

其中 $d=d_q$ 表示词的嵌入维度，$W_Q,W_K,W_V\in R^{d_q\times d_q}$ 表示 q 的可学习投影矩阵，$U_Q,U_K\in R^{d_q\times d_q}$ 表示 S 的可学习矩阵。式(12)中的第一项用于刻画词级表示之间的细粒度相关性，保留局部关键词和上下文依赖信息；第二项则将句子级全局语义作为补充约束注入注意力计算，使词间关系建模不再仅依赖局部共现模式，而能够受到整体句意的引导。这样设计的好处在于，模型在突出关键医学语义词的同时，也能够维持对完整问题语义结构的感知，从而减少语义丢失并提升开放式问题表示的鲁棒性。SWE 的核心作用并非简单叠加句级与词级特征，而是利用全局语义对词间关系建模进行显式引导，从而为开放式问题构建兼具局部判别性与整体一致性的文本表示。

在此基础上，问题的双重嵌入表示为：$Q_{SWE}=\left(\hat{q}_1;\cdots;\hat{q}_l\right)=\lambda*\mathrm{SWE}(Q)+S\in R^{l\times d_q}$，其中 λ 为权重系数，

用于平衡词级增强表示与句子级全局语义之间的贡献度。最终，我们得到的问题特征表示为：

$$\hat{Q}=\begin{cases}S, & G(q)=0,\\ Q_{SWE}, & G(q)=1.\end{cases} \tag{14}$$

由此，模型能够根据问题类型自适应地选择匹配的语义建模路径：对于封闭式问题，采用更紧凑的句级表示以保证编码效率；对于开放式问题，则通过句词联合编码增强其对复杂语义约束和深层推理信息的表征能力。

2.2. 知识驱动对比学习

我们采用 VIT 模型来提取图像特征，给定医学图像 V 的特征 $\hat{V}$ 可以定义为：

$$\hat{V}=\mathrm{VIT}(V)\in R^{n\times d_v}. \tag{15}$$

我们利用来自统一医学语言系统的结构化专家领域知识(UMLS)增强 Med-VQA。将文本中的实体链接到 UMLS 知识库，对于每个图像－文本对存在一个实体序列 $ES=\{x_1^e,x_2^e,\cdots,x_m^e\}$ 与 token 序列 $T=\{x_1,x_2,\cdots,x_l\}$ 对齐，其中 x_i^e 是提取出的实体， m 是实体序列长度。我们采用实体匹配矩阵 $P\in R^{l\times m}$ 来记录提取的实体位置，其中每个元素表示为：

$$P_{ij}=\begin{cases}0, & x_i\notin x_j^e,\\ 1, & x_i\in x_j^e.\end{cases} \tag{16}$$

在对所有文本进行处理后，得到包含所有 N_e 个实体的实体集 $\mathcal{E}=\{e_i\}_{i=1}^{N_e}$。若三元组的头实体和尾实体都在实体集中，则从 UMLS 知识库中提取相关的知识图谱三元组，将其表示为知识集 $EK=\{k_i=(h_i,r_i,t_i)\}_{i=1}^{N_g}$，其中 N_g 为知识图谱三元组的个数， k_i 为知识图谱三元组， h_i 、 r_i 和 t_i 分别表示头实体、关系和尾实体。将 TransE 应用于知识图谱 EK 获得实体嵌入 $\{e_i\}_{i=1}^{N_g}$，其中 $e_i\in R^{d_e}$， d_e 是实体嵌入维度。考虑到图的整体结构，我们使用 GAT 以聚合每个节点在图领域中的局部信息，并获得实体表示：

$$\hat{E}=\{\hat{e}_1,\hat{e}_2,\cdots,\hat{e}_{N_g}\}\in R^{N_g\times d_e}. \tag{17}$$

为进一步提升 Med-VQA 的多模态语义对齐能力，本文设计知识－图像与知识－问题两种对比学习机制。前者促进模型识别病理特征相似的图像表示，后者引导模型聚焦语义相近的问题嵌入，从而提升对复杂问答对中语义差异的感知能力。两者均采用三元组对比损失作为训练目标，通过最小化正样本对之间的表示相似度，最大化负样本对之间的表示相似度，有效保证跨模态表示在医学语义空间中的一致性，提升跨模态对齐质量。

在具体的训练过程中，每个样本对应的知识表示作为正样本，其余不相关样本作为负样本，图像和问题模态表示与知识表示互为对比目标，以构建双向对比学习路径。考虑到 Med-VQA 数据集中单一图像可对应多个语义差异显著的问题，我们在负样本构建中引入掩码机制，将来源于同一图像但关联不同问题的知识表示从负样本集合中剔除以缓解伪负样本干扰。知识对比学习采用三元组损失，以知识－问题对比路径为例，损失函数定义如下：

$$\mathcal{L}_{Q\to K}=\max\left(\mathrm{sim}\left(\hat{Q},\hat{E}\right)-\mathrm{sim}\left(\hat{Q},\hat{E}'\right)+\mathrm{margin},0\right), \tag{18}$$

$$\mathcal{L}_{K\to Q}=\max\left(\mathrm{sim}\left(\hat{E},\hat{Q}\right)-\mathrm{sim}\left(\hat{E},\hat{Q}'\right)+\mathrm{margin},0\right), \tag{19}$$

其中， $\hat{Q}$ 和 $\hat{E}$ 分别表示问题特征与其对应的知识特征， $\hat{Q}'$ 和 $\hat{E}'$ 为其负样本， $\mathrm{sim}(\cdot,\cdot)$ 表示余弦相似度函数，margin 为对比边界参数。最终，知识－问题对比损失表示为：

$$\mathcal{L}_Q = \mathcal{L}_{Q\to K} + \mathcal{L}_{K\to Q}, \tag{20}$$

同理，定义知识－图像对比损失 $\mathcal{L}_V$。两者加权组合构成总对比损失：

$$\mathcal{L}_{ctra} = \alpha * \mathcal{L}_Q + \beta * \mathcal{L}_V, \tag{21}$$

其中，α 和 β 为两个对比路径的权重系数，控制其对整体训练目标的贡献度。。

2.3. 知识驱动推理融合

为有效整合外部医学知识并增强跨模态语义推理，我们提出知识驱动推理融合机制(图 3)。该机制以外部医学知识为事实支撑，在图像与问题的引导下实现灵活注入，从而在多模态交互中完成语义对齐与知识增强，提升复杂语义关系建模能力。

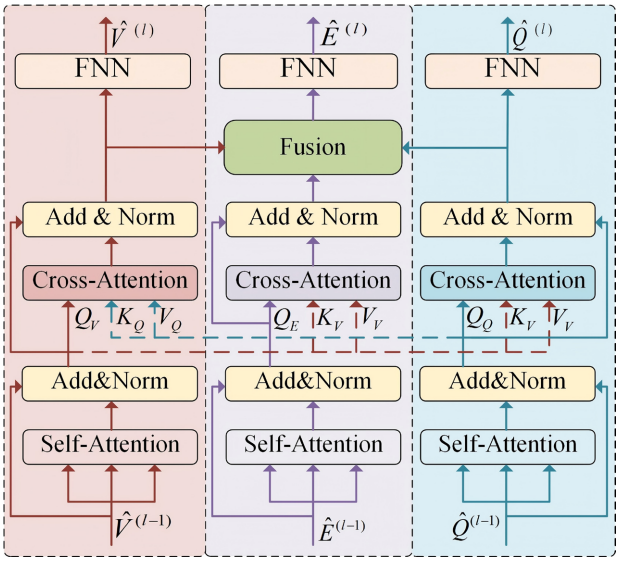

Figure 3. Knowledge driven reasoning fusion module
图 3. 知识驱动推理融合模块

单层模块在建模复杂跨模态关系时存在不足，因此我们将其堆叠为 L 层级联结构，以实现更充分的特征交互与融合。具体地，从方程(14)、(15)和(17)分别获得的问题特征 $\hat{Q}$、图像特征 $\hat{V}$ 和知识特征 $\hat{E}$ 作为融合模块的初始输入。在第 l 层，我们以第 $l-1$ 层的输出 $\hat{Q}^{(l-1)}$、$\hat{V}^{(l-1)}$ 和 $\hat{E}^{(l-1)}$ 作为输入，以实现三个特征的深层交互与对齐。

在融合阶段，我们采用自注意力与引导注意力相结合的策略。各模态特征通过多头自注意力(MSA)建模模态内关系，第 l 层的计算公式如下：

$$\begin{aligned} F_Q^{(l)} &= \mathrm{MSA}\left(\hat{Q}^{(l-1)}\right), \\ F_V^{(l)} &= \mathrm{MSA}\left(\hat{V}^{(l-1)}\right), \\ F_E^{(l)} &= \mathrm{MSA}\left(\hat{E}^{(l-1)}\right), \end{aligned} \tag{22}$$

其中 $F_Q^{(l)}$、$F_V^{(l)}$ 和 $F_E^{(l)}$ 分别为第 l 层的问题、图像和知识特征输出，再经残差连接与归一化后得到增强表示，为简洁起见，仍记作 $F_Q^{(l)}$、$F_V^{(l)}$ 和 $F_E^{(l)}$。

随后，引入多头引导注意力(MCA)实现模态间的信息融合，加强语义耦合能力，其中问题与图像特

征双向对齐，知识特征依赖图像上下文融合：

$$\begin{aligned}\hat{F}_Q^{(l)} &= \mathrm{MCA}\left(F_Q^{(l)}, F_V^{(l)}, F_V^{(l)}\right),\\ \hat{F}_V^{(l)} &= \mathrm{MCA}\left(F_V^{(l)}, F_Q^{(l)}, F_Q^{(l)}\right),\\ \hat{F}_E^{(l)} &= \mathrm{MCA}\left(F_E^{(l)}, F_V^{(l)}, F_V^{(l)}\right),\end{aligned} \tag{23}$$

其中 $\hat{F}_Q^{(l)}$， $\hat{F}_V^{(l)}$ 和 $\hat{F}_E^{(l)}$ 分别表示融合后的问题、图像与知识特征。同样采用残差连接与层归一化处理，仍记为 $\hat{F}_Q^{(l)}$， $\hat{F}_V^{(l)}$ 和 $\hat{F}_E^{(l)}$。

考虑到问题与实体之间的结构对齐关系由知识映射矩阵 P 表示，其融合过程如下：

$$\tilde{F}_Q^{(l)} = P * \hat{F}_E^{(l)} + \hat{F}_Q^{(l)}, \tag{24}$$

其中 $\tilde{F}_Q^{(l)}$ 表示融合了图像与知识信息后的问题特征，经前馈子层得到增强的多模态表示 $\hat{Q}^{(l)}$ 作为下一层的输入。同时将 $\hat{F}_V^{(l)}$ 和 $\hat{F}_E^{(l)}$ 输入到前馈网络以生成下一层图像表示 $\hat{V}^{(l)}$ 和知识表示 $\hat{E}^{(l)}$。

在获得跨模态融合表示后，本文进一步引入基于词重要性权重的语义调制机制，以保证问题中的关键医学语义在线索整合后仍能对最终决策保持足够约束。其动机在于：尽管多模态交互能够建立图像与文本之间的语义关联，但融合结果仍可能受到无关语义成分干扰，导致少量关键诊断词在联合表示中被弱化。为此，我们利用由公式(2)~(9)得到的词级重要性表示 $\Phi(q)$，通过多层感知机映射生成语义引导向量：

$$\mathrm{QCR}(q) = \mathrm{MLP}\left(\Phi(q)\right), \tag{25}$$

随后，将最终一层视觉表示与问题表示拼接，并通过哈达玛积进行逐元素语义调制：

$$Z = \mathrm{concate}\left[\hat{V}^{(L)}; \hat{Q}^{(L)}\right] \odot \mathrm{QCR}(q), \tag{26}$$

其中 Z 是最终融合特征，$\odot$ 为哈达玛积。该设计使关键问题语义能够对联合表示进行显式重标定，从而在保留多模态互补信息的同时，增强模型对核心诊断线索的关注，提高最终答案预测的针对性与鲁棒性。

综上，本文围绕医学视觉问答中的关键挑战，构建了一条由问题表示增强、知识语义约束到多源证据融合与校准的递进式建模链路。具体而言，DQE 通过问题类型感知的动态编码提升了问题表示的针对性；知识对比学习进一步增强了问题与外部医学知识之间的语义一致性，提高了知识利用的有效性；在此基础上，多模态融合与语义调制共同作用，使图像、文本与知识信息能够围绕当前问题形成更具判别性的联合表示。由此，模型不仅增强了跨模态语义对齐能力，也提升了最终答案预测的准确性与鲁棒性。

2.4. 分类器

根据先前研究，我们将 Med-VQA 建模为一个多分类任务。最终融合特征表示 Z 输入由 MLP 和 Softmax 构成的分类器中，获得每个候选答案的概率分布。最终预测答案为概率最高的类别，其计算形式如下：

$$\hat{a} = \arg\max\left(\mathrm{softmax}\left(\mathrm{MLP}(Z)\right)\right), \tag{27}$$

其中 $\hat{a}$ 表示模型输出的预测答案。我们采用交叉熵作为分类损失函数：

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\left(a_i \log\left(\hat{a}_i\right) + \left(1-a_i\right)\log\left(1-\hat{a}_i\right)\right), \tag{28}$$

其中 $\mathcal{L}_{cls}$ 为分类损失，N 为训练样本数，a_i 和 $\hat{a}_i$ 分别为第 i 个样本真实标签和预测概率。模型的总损失函

数由对比学习损失(式(21))与分类损失(式(28))共同构成，我们将分类损失的权重固定为 1，以避免多任务权重波动对主任务造成干扰，对比学习损失的权重则通过式(21)中的超参数 α 与 β 进行调节。联合优化目标可形式化表示为：

$$\mathcal{L} = \mathcal{L}_{ctra} + \mathcal{L}_{cls}. \tag{29}$$

3. 实验与结果

在 VQA-RAD 与 SLAKE 数据集上系统评估所提出的 DKCR 模型的性能表现。此外，我们还设计了消融实验，以验证各个模块的有效性及其组合的协同作用。为了进一步探索不同模块对视觉推理过程的影响，使用 Grad-CAM [11]进行视觉分析。

3.1. 数据集

本研究所采用的数据集信息汇总如表 1 所示。因 VQA-RAD 原始数据无验证集，我们复制训练集作为验证集。SLAKE 数据集是面向医学多模态问答的中英双语资源，本研究仅使用英文子集。

Table 1. Detailed information on the medical visual question answering dataset
表 1. 医学视觉问答数据集的详细信息

Dataset	Data category	Training set	Validation set	Test set
VQA-RAD	Images	315	315	315
	QA pairs	3064	3064	451
SLAKE	Images	450	96	96
	QA pairs	4919	1053	1061

3.2. 评估指标

与大多数 Med-VQA 模型一致，我们采用分类准确率(%)作为主要评估指标。设 P_i 和 Y_i 分别表示测试集中样本 i 的预测标签与真实标签，T 表示测试样本集合，则整体准确率定义如下：

$$Acc = \frac{1}{|T|}\sum_{i\in T}\mathbf{1}\left(P_i = Y_i\right), \tag{30}$$

其中，$\mathbf{1}(\cdot)$ 为指示函数，当预测标签与真实标签一致时取值为 1，否则为 0。

3.3. 实验细节

我们将图像与问题模态维度统一设定为 384，知识模态的隐藏维度设为 256，并采用 12 个注意力头(每个头的维度为 32)。问题输入的最大长度限制为 32，图像输入大小调整为 224×224。训练过程中使用 AdamW 优化器，初始学习率设置为 5e−5，批大小为 16，所有实验均在单张 NVIDIA RTX 3090 GPU 上完成。在损失函数设计上，分类损失的权重固定为 1，而对比损失通过可调超参数进行加权，以平衡不同优化目标的贡献。

3.4. 与最新技术比较

为验证所提出 DKCR 模型的整体性能，本文将其与 VQA-RAD 和 SLAKE 数据集上当前具有代表性的先进方法进行了比较，结果如表 2 所示。为了公平比较，我们在相同配置下复现了 DALNet-WSE [12]、

ARL [13]和 LaPA [14]，而其他方法的结果均来自其原始论文。

与 CR 及其扩展方法 CPCR 和 CPRD 相比，DKCR 在两个数据集上均取得了显著优势。尽管上述方法通过问题类型自适应的推理机制提升了泛化能力，但其建模完全依赖图像与文本特征，缺乏外部知识支撑。相较之下，DKCR 引入结构化知识并显式地对齐知识、视觉与问题表征，有效缓解了语义不匹配问题。与同样缺乏外部知识的 DALNet-WSE 相比，DKCR 在 VQA-RAD 与 SLAKE 上整体准确率分别提升了 2.20%与 2.21%，进一步验证了外部医学知识在弥补跨模态表征不足方面的重要作用。基于注意力模型 AMAM 与 VG-CALF 通过跨模态注意力机制增强了模态间交互，但由于潜在空间缺乏知识引导约束，其整体性能仍不及 DKCR。这表明，相比单一的注意力机制，将知识引导注意力与知识对比学习相结合，更能有效缓解跨模态语义偏差。

Table 2. Comparison and analysis of results with state-of-the-art models
表 2. 与最先进模型的结果比较和分析

Method	VQA-RAD			SLAKE		
	Open-ended	Closed-ended	Overall	Open-ended	Closed-ended	Overall
CR [4]	60.00	79.30	71.60	78.80	80.00	82.00
CPCR [6]	60.50	80.40	72.50	80.50	84.10	81.90
CPRD [15]	52.50	77.90	67.80	79.50	83.40	81.10
AMAM [2]	63.88	80.37	73.39	-	-	-
MKBN [16]	-	-	-	77.70	85.10	80.60
M2I2 [7]	61.80	81.60	73.70	74.70	**91.10**	81.20
M3AE [17]	67.23	83.46	77.01	80.31	87.82	83.25
DALNet-WSE* [12]	65.90	84.80	77.40	79.04	89.30	82.47
ARL* [13]	65.10	85.96	77.55	79.70	89.30	84.10
LaPA* [14]	66.48	85.29	77.82	79.84	86.53	82.46
PubMedCLIP [18]	60.10	80.00	72.10	78.40	82.50	80.10
VG-CALF [19]	67.00	85.50	76.10	81.40	83.80	83.30
MITER [20]	59.40	80.50	72.10	79.20	84.40	81.20
UnlCLAM [21]	59.80	82.60	73.20	81.10	85.70	83.10
MKGF [22]	61.51	83.67	72.59	69.66	82.93	76.30
DKCR	**69.83**	**86.02**	**79.60**	**82.09**	88.70	**84.68**

相较于同样引入外部知识的 MKBN 与 LaPA，DKCR 也表现出更优的性能。MKBN 未实现知识与视觉、文本嵌入对齐，LaPA 偏重答案感知提示而忽略潜在空间一致性。DKCR 通过双路径对比学习与知识引导注意力结合，实现了更全面的知识–模态对齐。进一步与使用相同外部知识源的 ARL 对比，DKCR 在开放式问题上的表现明显优于 ARL。这主要归功于 DQE 模块对问题类型的区分能力以及对比学习对

细粒度语义对齐的优化，使模型在处理开放式问题时能捕捉更丰富的语义细节，同时避免封闭式问题特征冗余。

与生成模型 M2I2 相比，尽管 M2I2 在 SLAKE 的封闭式问题上表现略优，但 DKCR 在开放式问题及总体准确率上均占优，显示出其更强的语义适应能力。同样，相较于 M3AE 等自监督预训练方法，DKCR 在两个数据集上均取得稳定提升，说明知识增强的对比学习有助于增强医学场景下的语义表示多样性。此外，尽管 MKGF 融合了检索增强策略，其性能仍低于 DKCR，说明仅靠检索增强难以保证跨模态语义结构的一致性。DKCR 通过对比学习将知识、图像与问题映射到统一的语义空间，从而实现更几何一致的表征对齐。

综上所述，DKCR 在 VQA-RAD 和 SLAKE 上均取得了较优的性能，尤其在开放式问题上提升显著。该优势源于 DQE 对问题类型动态编码、知识引导的双路径对比学习以及引导注意力机制，三者共同促成了多模态语义的细粒度与一致性对齐。

3.5. 消融实验

为系统分析所提出模型各组成模块对整体性能的贡献，并进一步探索它们之间的协同作用，我们设计并实施了以下消融实验：

(1) Baseline：在 M3AE 框架基础上构建的简化版本，不引入任何外部知识信息；

(2) DKCR (ND)：不使用 DQE，所有问题均使用统一的 BERT 编码器处理；

(3) DKCR (NG)：对开放式与封闭式问题统一采用 BERT + SWE 的联合编码方式；

(4) DKCR (NC)：移除知识对比学习，仅使用主分类任务；

(5) DKCR (NDC)：同时去除 DQE 和知识对比学习，仅保留知识驱动推理融合；

(6) DKCR (NK)：去掉外部知识模态及知识对比学习，仅输入图像与文本特征进行多模态融合；

(7) DKCR (NA)：移除密集跨模态注意力交互，将三种模态的特征直接拼接后送入分类器进行答案预测；

(8) DKCR (NQ)：删除 QCR 模块，最终融合表示仅由注意力交互后的特征构成；

(9) DKCR (Q-V)：引导注意力结构中仅保留问题引导图像特征路径；

(10) DKCR (V-Q)：引导注意力结构中仅保留图像引导问题特征路径。

表 3 给出了 DKCR 的不同配置在 VQA-RAD 和 SLAKE 数据集上的性能。

1) DQE 消融实验：表 3 显示，与 DKCR (ND)相比，DKCR 在 VQA-RAD 和 SLAKE 数据集上的整体准确率分别提升 2.99%与 1.32%。其中，开放式问题的提升较为显著，而封闭式问题准确率相近，该结果表明，开放式问题更依赖语义增强表征，BERT + SWE 的联合编码能够更有效地捕捉复杂语义信息，从而增强推理能力。

进一步对比 DKCR 与 DKCR (NG)，两者在开放式问题上差异不大，但在封闭式问题上 DKCR 分别提升 1.47%和 2.41%，这说明对于结构更固定、答案空间有限的封闭式问题，联合编码可能引入冗余特征或噪声从而削弱判别效果。总体而言，DQE 既能增强模型对开放式问题的语义建模能力，又能避免为封闭式问题引入不必要的计算开销与冗余参数，从而提升模型在不同问题类型下的适配性与鲁棒性。

值得说明的是，DQE 中的问题类型划分并非基于人工关键词规则，而是由轻量级可学习分类器自动完成。为验证其可靠性，我们进一步统计了该分类器在测试集上的判别性能，其在 VQA-RAD 和 SLAKE 上的分类准确率分别达到 99.33%和 99.81%。这表明所采用的问题类型识别策略具有较高稳定性，能够为后续动态编码路径选择提供可靠依据。少量误分类样本主要集中在句式形式与真实语义需求不完全一致的边界问题中，但由于占比较低，因此对整体模型性能影响有限。

Table 3. Study on ablation of DKCR in different states on VQA-RAD and SLAKE datasets
表 3. VQA-RAD 和 SLAKE 数据集上不同状态 DKCR 的消融研究

Method	VQA-RAD			SLAKE		
	Open-ended	Closed-ended	Overall	Open-ended	Closed-ended	Overall
Baseline	64.80	83.08	75.83	80.31	85.81	82.46
DKCR (ND)	63.68	85.29	76.71	80.46	87.98	83.41
DKCR (NG)	68.71	84.55	78.27	81.24	86.29	83.22
DKCR (NC)	68.71	84.19	78.04	80.62	**90.14**	84.35
DKCR (NDC)	64.80	81.98	75.16	81.24	87.01	83.50
DKCR (NK)	65.36	82.35	75.60	80.46	87.98	83.41
DKCR (NA)	62.01	79.41	72.50	80.93	86.29	83.03
DKCR (NQ)	68.15	85.08	78.36	81.55	88.22	84.16
DKCR (Q-V)	67.03	84.55	77.60	80.77	86.53	83.03
DKCR (V-Q)	65.92	83.82	76.71	78.75	87.98	82.37
DKCR	**69.83**	**86.02**	**79.60**	**82.09**	88.70	**84.68**

2) 知识对比学习消融实验：在 VQA-RAD 数据集上，DKCR 相较于 DKCR (NC)在开放式问题、封闭式问题及整体准确率上分别提升 3.35%、1.83%和 2.44%，这充分说明对比学习在跨模态语义对齐与判别能力增强方面发挥了重要作用，尤其在语义复杂度更高的开放式问题中带来显著增益。

相比之下，在 SLAKE 数据集上，DKCR (NC)在开放式问题上较 DKCR 下降 1.47%，整体准确率略低于 DKCR，但在封闭式问题上反而优于 DKCR，这表明封闭式问题更依赖判别性特征而非复杂知识推理，对比学习在此类问题中可能引入过强的约束，从而产生轻微干扰。因此，去除对比学习似乎有利于这种特定问题类型的表现。

进一步比较 DKCR (NC)与 DKCR (NDC)可见，后者在两个数据集上的整体性能均显著下降，说明对比学习与 DQE 具有互补性，前者通过约束潜在表征空间促进跨模态一致性，后者则通过问题类型自适应提升语义建模效果，二者协同作用确保了 DKCR 在复杂 Med-VQA 任务中的优越表现。

3) 知识推理融合消融实验：如表 3 所示，DKCR 相较于 DKCR (NK)在 VQA-RAD 和 SLAKE 数据集上的整体准确率分别提升 4.00%与 1.27%，并均优于 Baseline，这表明仅依赖视觉与语言模态难以充分弥合跨模态语义差异，外部医学知识的引入不仅为模型提供了丰富的先验信息，增强了语义理解与推理能力，而且结合对比学习机制，有效促进了跨模态语义空间对齐，从而提升模型的鲁棒性与泛化性。

在跨模态交互方面，DKCR (NA)在 VQA-RAD 上性能大幅下降，表明简单拼接难以捕捉模态间复杂依赖关系，在 SLAKE 上，虽然其整体结果优于 Baseline，但仍低于完整模型，说明在数据量较大的场景下，特征拼接虽能保持一定信息完整性，但不足以充分挖掘模态间的交互潜力。DKCR (NQ)的性能虽略低于 DKCR，但整体结果仍优于 Baseline、DKCR (NK)和 DKCR (NA)，验证了 QCR 在优化融合表示与增强判别能力方面的有效性。

进一步分析单向交叉注意力路径，结果显示 DKCR (Q-V)与 DKCR (V-Q)在 VQA-RAD 数据集上均优于 DKCR (NA)，表明单向交互在一定程度上建模跨模态依赖关系。在 SLAKE 数据集上，DKCR (Q-V)与

DKCR (NA)的整体准确率持平，而 DKCR (V-Q)低于 DKCR (NA)，说明过度依赖单向信息流可能导致信息损失与语义不足。完整模型 DKCR 在两个数据集上均优于 DKCR (Q-V)和 DKCR (V-Q)，证明了知识推理融合在实现细粒度跨模态对齐及增强推理能力方面的有效性，同时凸显了知识引导下多模态互补优势。

3.6. 参数灵敏度分析

为了彻底验证 DKCR 的设计原理，我们对其关键超参数进行了实证研究。

1) 句词权重 λ 的灵敏度分析：为系统评估 SWE 模块中超参数 λ 的作用，本文在 VQA-RAD 与 SLAKE 数据集上进行灵敏度分析，λ 取值范围设定为[0.1, 1.0]，结果如图 4 所示。在 VQA-RAD 数据集上，开放式问题的准确率随 λ 变化呈现明显波动，并在 $\lambda=0.5$ 时达到峰值；封闭式问题对 λ 的敏感性相对较低，在 $\lambda=0.8$ 取得最优性能。该现象说明 λ 虽主要作用于开放式问题的特征表示，但在多模态融合与对比学习阶段也可能引发特征分布变化，进而间接影响封闭式任务表现。VQA-RAD 在 $\lambda=0.5$ 时整体准确率最高，因此本文在该数据集上固定 $\lambda=0.5$。相比之下，SLAKE 数据集在三项指标(开放式、封闭式及整体准确率)上均表现出更强的波动性，反映其对特征权重配置更加敏感。当 $\lambda=0.9$ 时三项指标同时达到最优，验证了在该任务场景下增强句词级特征贡献有助于提升问答性能。综上，合理调节 λ 对模型性能具有显著影响，体现了在多模态表示学习中动态平衡不同粒度特征的重要性。

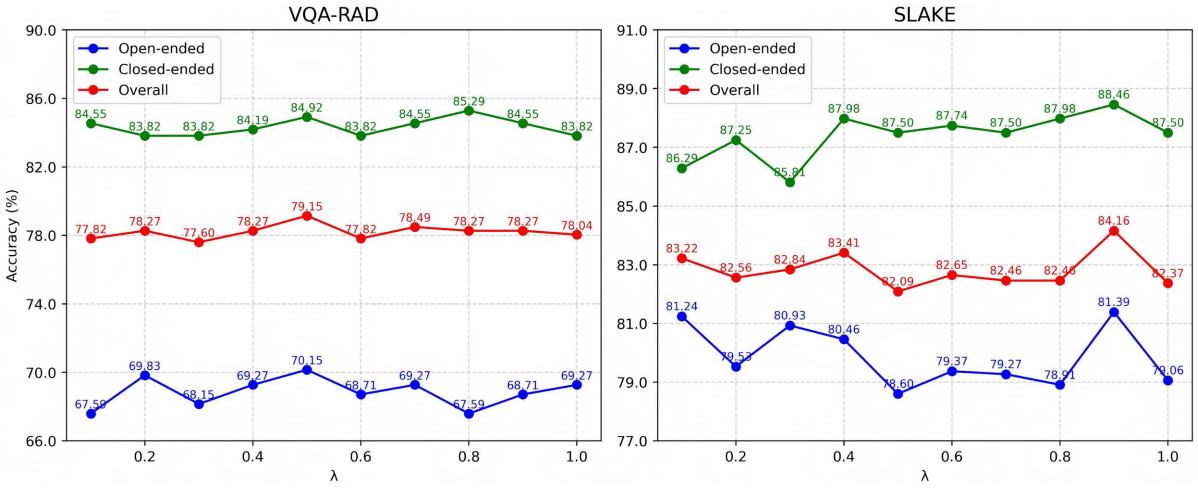

Figure 4. The impact of sentence weight λ on the overall accuracy of two datasets
图 4. 句词权重 λ 对两个数据集总体准确率的影响

2) 知识对比学习约束的影响：知识对比学习损失由知识–问题对比学习与知识–图像对比学习两部分组成，分别通过超参数 α 和 β 加权控制。为了探究两条对比路径在模型优化中的相对作用，本文在 VQA-RAD 与 SLAKE 数据集上对 α 与 β 进行了二维网格搜索，结果如图 5 所示。在 VQA-RAD 数据集上，模型整体性能对 α 与 β 的变化高度敏感，总体准确率在 77.16%到 79.15%之间波动，并于 $\alpha=0.5$，$\beta=1.0$ 时达到最高。随着 α 或 β 取值的增大，模型性能呈现下降趋势，表明该数据集对两条路径的权重配置具有不同需求，更侧重于知识–图像对比学习的作用。SLAKE 数据集的性能同样呈现显著波动，其整体准确率变化范围为 82.17%至 84.16%，并在 $\alpha=2.0$，$\beta=2.0$ 时达到最佳，表明该数据集更倾向于两条路径的均衡协同。上述实验结果凸显了不同数据集在语义复杂性和模态依赖性方面的内在差异：VQA-RAD 在约束较轻的情况下表现更优，而 SLAKE 需要更强的知识对比学习监督信号。这进一步验证了所提出的多路径对比学习框架在不同任务场景中的适应性和鲁棒性。

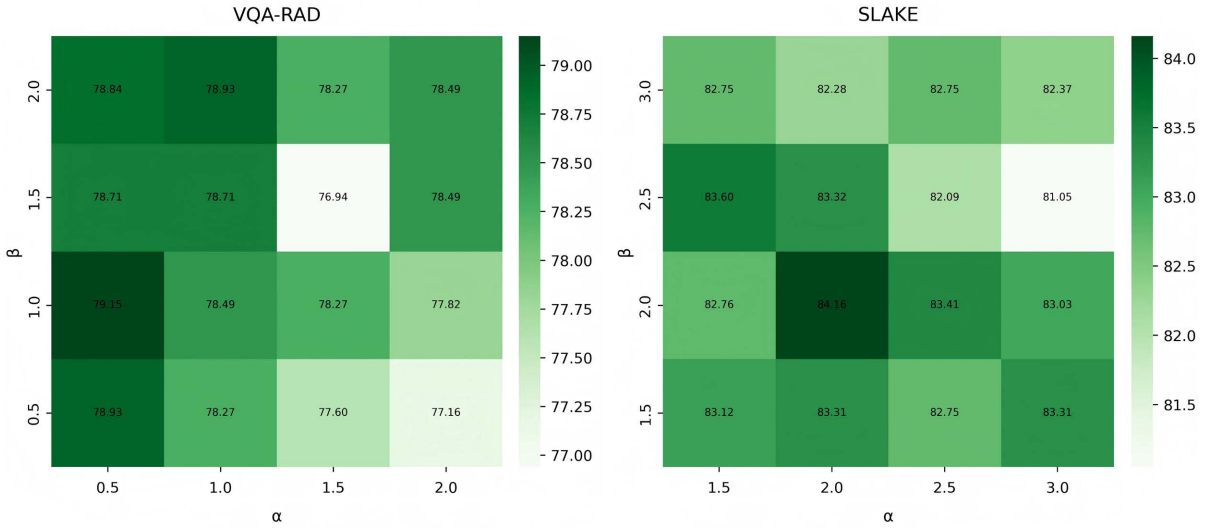

Figure 5. Impact of hyperparameters α and β in knowledge contrastive learning on the overall accuracy of the VQA-RAD and SLAKE datasets
图 5. 知识对比学习超参数 α 与 β 对 VQA-RAD 和 SLAKE 数据集总体准确率的影响

3) 跨模态交互深度的影响：在 DKCR 模型中，注意力层数 L 决定了多模态融合的深度。如表 4 所示，当 L 取 2~4 层时，模型在开放式问题上整体准确率偏低，说明此时难以充分捕捉图像与问题间复杂的语义关联，表征学习不足以支撑深层推理。随着层数增加至 6 层，模型在开放式与封闭式问题及整体准确率上均达到最优，尤其在多步推理的开放式任务中提升显著。这表明适当的深度有助于促进视觉与语言模态深层交互，逐步细化语义对齐，增强联合表示的判别能力。然而，当 L 增至 8 层时，性能未进一步提升，部分指标甚至出现轻微回落。其原因可能为过深结构引发过拟合，泛化能力下降，同时深层结构加剧梯度衰减，增加了模型优化难度，影响收敛稳定性，导致性能波动。综合而言，6 层交叉注意力结构在实验中展现了最佳平衡：既保证了足够的融合深度以捕捉复杂跨模态语义，又避免了过深网络带来的训练负担与泛化风险。

Table 4. The effect of the number of attention layers L on the accuracy of VQA-RAD and SLAKE datasets
表 4. 注意力层数 L 对 VQA-RAD 与 SLAKE 数据集准确率的影响

L	VQA-RAD			SLAKE		
	Open-ended	Closed-ended	Overall	Open-ended	Closed-ended	Overall
2	67.59	79.77	74.94	80.46	86.77	82.94
4	66.48	84.55	77.38	80.45	87.74	83.31
6	**69.13**	**85.96**	**79.28**	**82.05**	**88.70**	**84.65**
8	66.48	84.19	77.16	79.84	87.01	82.65

3.7. 定性分析

为了更直观地展示 DKCR 及其变体在不同类型问题上的表现差异，我们对 VQA-RAD 数据集的三个实例(a、b、c)及 SLAKE 数据集的三个实例(d、e、f)进行定性分析，如图 6 所示。

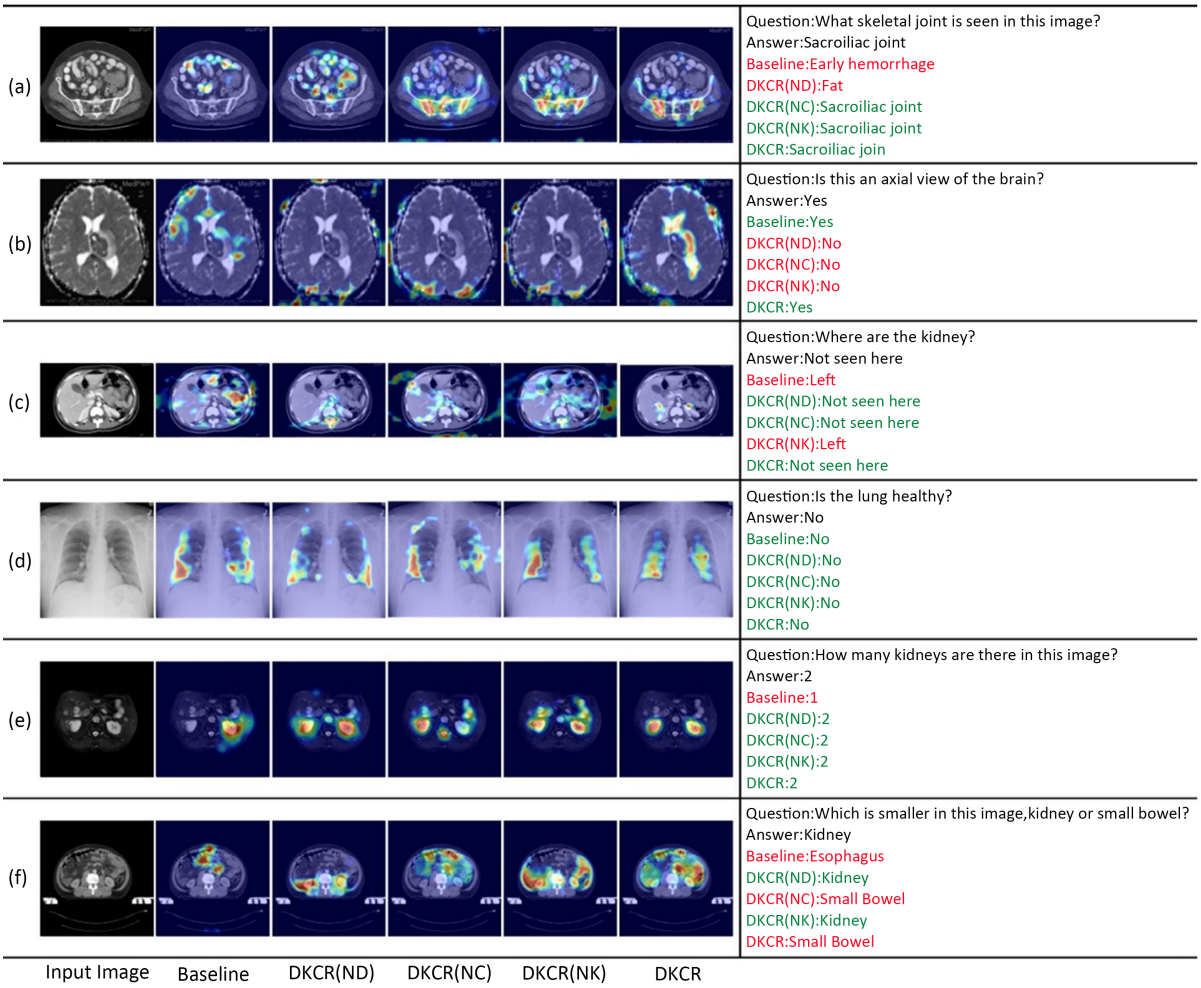

Figure 6. Visual results of DKCR, with correct and incorrect predictions highlighted in green and red
图 6. DKCR 的可视化结果，正确和错误的预测分别用绿色和红色高亮显示

案例(a)中，Baseline 与 DKCR (ND)因注意力偏向图像上半区域而无法聚焦关键解剖结构，产生错误预测。完整 DKCR 及其变体 DKCR (NC)、DKCR (NK)均能准确聚焦于骶髂关节这一关键区域，为正确预测提供了可靠视觉支撑。涉及成像模态差异的案例(b)中，DKCR 能够识别出侧脑室的对称结构，注意力分布与问题语义高度匹配；其他变体注意力更多分散于图像边缘区域，无法有效捕捉核心结构；Baseline 虽预测正确，但注意力分布较为分散，视觉依据一致性不足，决策可靠性较低。案例(c)中，该 CT 图像实际无肾脏，Baseline 与 DKCR (NK)缺乏关键语义引导，产生存在肾脏的幻觉预测；引入外部知识的 DKCR (ND)、DKCR (NC)及完整 DKCR 均能准确定位肾脏预期解剖位置，并正确判断“未可见”，表明外部知识为模型提供了关键语义线索，缓解了跨模态偏差与误判，提升了预测的准确性与合理性。

案例(d)为二元分类任务，预测主要依赖图像中的显著病理特征，对跨模态语义依赖较低，所有模型均聚焦于肺野区域，准确捕捉病理特征并给出正确答案。案例(e)中，所有模型均关注肾脏区域，但缺乏知识整合，其定位精度与覆盖范围仍受限制，Baseline 主要关注左肾，DKCR (NK)更侧重右肾，在语义调制机制作用下，DKCR (NK)可扩展注意范围至左肾区域，实现正确预测。案例(f)中 DKCR (ND)和 DKCR (NC)关注区域集中于主动脉以上结构，DKCR 同时关注小肠和肾脏，但更偏重小肠，最终导致预测失败。

3.8. 模型效率分析

为进一步评估 DKCR 的实用性，本文比较了其与主要基线模型在参数量、显存占用以及训练和推理时间上的差异，结果如图 7 所示。为保证比较的公平性，所有效率指标均在相同硬件平台和统一实验设置下统计。总体来看，DKCR 的参数量为 409M，与强基线 LaPA 的 404M 基本处于同一量级，仅有小幅增加，说明所提出的动态问题编码、知识对比学习和语义调制等模块并未带来明显的参数膨胀。从显存开销来看，DKCR 在训练阶段和测试阶段的显存占用均明显低于 LaPA，表明其在保持较强表达能力的同时具有更好的资源控制能力。

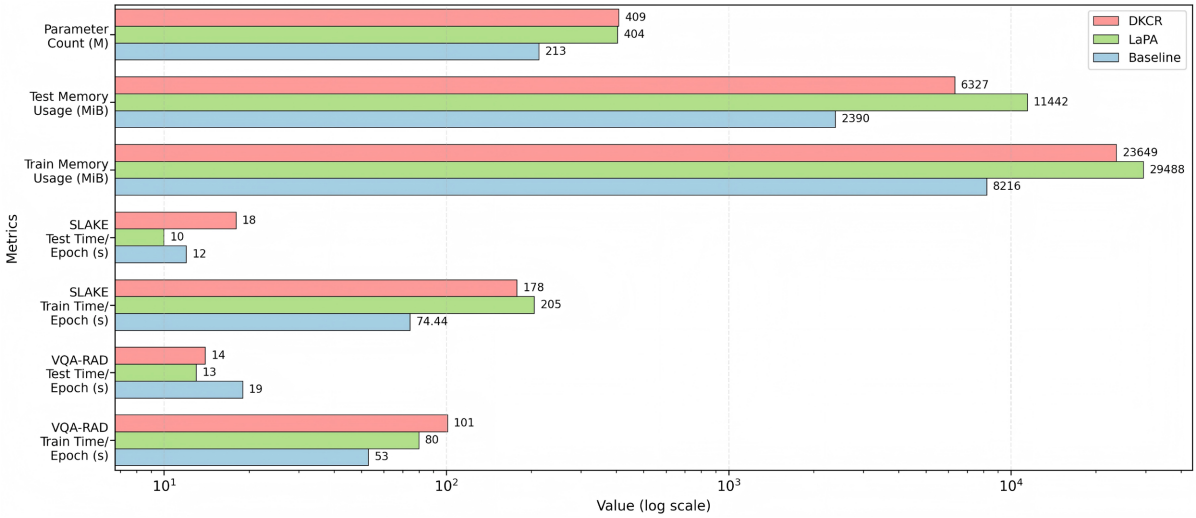

Figure 7. Time efficiency and resource consumption: DKCR vs. LaPA vs. Baseline
图 7. 时间效率与资源消耗：DKCR vs. LaPA vs.基线

从时间开销来看，DKCR 在不同数据集上的表现存在一定差异。在 VQA-RAD 上，DKCR 的训练时间高于 LaPA 和 Baseline，但其测试时间与 LaPA 基本接近，并优于 Baseline；在 SLAKE 上，DKCR 的训练时间优于 LaPA，但测试时间相对更长。这说明，DKCR 为提升问题理解、知识利用和多模态推理能力，引入了一定额外的时间开销，尤其在较大规模数据集上的推理阶段更为明显。综合而言，DKCR 虽然并非最轻量或最快的模型，但在参数规模未显著增加的情况下实现了较好的显存控制，并在性能提升与计算成本之间取得了较为合理的平衡。

3.9. 错误案例分析

为更全面地评估 DKCR 的局限性，本文对测试集中的典型失败样本进行了分析。结果表明，模型错误主要集中在三类情形：复合病理描述中的多证据整合不足、细粒度解剖定位不够精确，以及多目标比较任务中的关系推理不足。

如图 8(a)所示，在复合病理描述任务中，各模型预测均仅覆盖其中一个组成部分：Baseline 和 DKCR (ND)偏向 Pulmonary edema，DKCR (NC)、DKCR (NK)和 DKCR 则偏向 Cardiomegaly。热力图表明，模型已能够关注到部分相关胸部异常区域，但仍主要依赖单一主导证据进行判断，未能形成对复合病理的完整联合描述。

如图 8(b)所示，各模型预测多停留在较粗粒度的位置层面。热力图表明，模型已能够感知异常所在半球，但仍未稳定收敛于 temporal lobe 的局部解剖范围。考虑到脑部轴位图像通常遵循放射学显示习惯，

图像左侧实际对应患者右侧解剖位置，因此热力图的主要响应区域对应于右侧 temporal lobe 病灶。该结果说明，当前模型在细粒度解剖定位方面仍有局限。

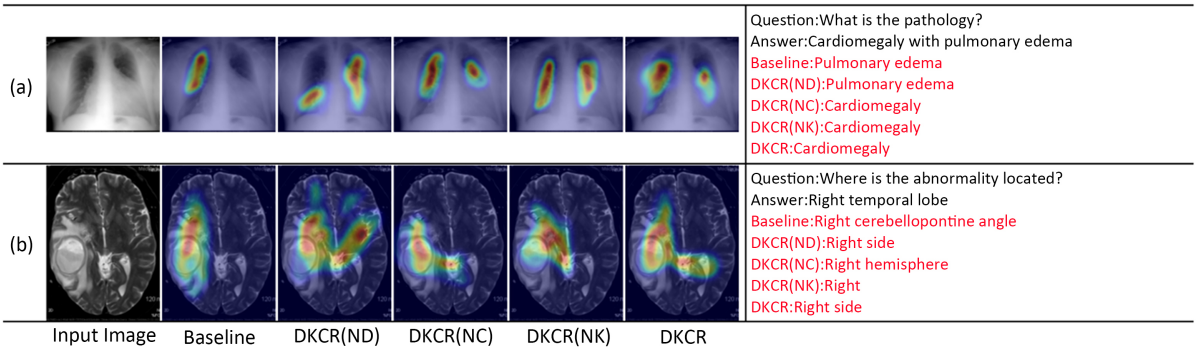

Figure 8. Failure case
图 8. 失败案例

此外，在多目标比较任务中，模型虽能同时激活多个候选区域，但仍可能因关键证据权重分配不合理而产生误判。以图 6(f)为例，模型同时关注到了 kidney 和 small bowel 区域，但最终更偏向 small bowel，导致比较判断错误。

总体而言，错误案例所揭示的局限并不否定 DKCR 在问题表示增强、知识语义约束和多模态推理方面的有效性，而是表明当前模型在复合病理语义生成、细粒度解剖定位和多目标比较推理等更复杂场景下仍存在进一步提升空间。未来可进一步结合区域级建模和更精细的推理机制，以提升模型在复杂医学视觉问答任务中的鲁棒性。

4. 结论

本文提出了 DKCR 模型，该模型集成了 DQE、知识对比学习和知识引导注意力机制，以实现更有效的跨模态语义对齐与推理。具体而言，DQE 使模型能够根据问题类型动态调整其文本编码策略，既捕捉开放式问题的复杂语义，又避免对封闭式问题产生冗余特征。知识驱动对比学习通过对齐外部医学知识与多模态表示，促进跨模态语义一致性，其双路径设计允许根据数据集灵活调整，从而增强模型泛化能力。知识引导注意力促进视觉与文本模态间的深度交互，有效缓解跨模态语义差距。在 VQA-RAD 和 SLAKE 数据集上的全面实验验证了所提出方法的有效性。未来工作将围绕区域级显式建模、更精细的推理机制以及动态利用外部知识展开，以进一步提升模型在复杂医学视觉问答任务中的鲁棒性与泛化能力。

参考文献

[1] Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., *et al*. (2023) Medical Visual Question Answering: A Survey. *Artificial Intelligence in Medicine*, **143**, Article ID: 102611. https://doi.org/10.1016/j.artmed.2023.102611

[2] Pan, H., He, S., Zhang, K., Qu, B., Chen, C. and Shi, K. (2022) AMAM: An Attention-Based Multimodal Alignment Model for Medical Visual Question Answering. *Knowledge-Based Systems*, **255**, Article ID: 109763. https://doi.org/10.1016/j.knosys.2022.109763

[3] Long, S., Yang, Z., Li, Y., Qian, X., Zeng, K. and Hao, T. (2023) MAMF: A Multi-Level Attention-Based Multimodal Fusion Model for Medical Visual Question Answering. In: Zhang, H., *et al*., Eds., *International Conference on Neural Computing for Advanced Applications*, Springer, 202-214. https://doi.org/10.1007/978-981-99-5847-4_15

[4] Zhan, L., Liu, B., Fan, L., Chen, J. and Wu, X. (2020) Medical Visual Question Answering via Conditional Reasoning. *Proceedings of the* 28*th ACM International Conference on Multimedia*, Seattle, 12-16 October 2020, 2345-2354. https://doi.org/10.1145/3394171.3413761

[5] Liu, G., He, J., Li, P., Zhao, Z. and Zhong, S. (2024) Cross-Modal Self-Supervised Vision Language Pre-Training with Multiple Objectives for Medical Visual Question Answering. *Journal of Biomedical Informatics*, **160**, Article ID: 104748. https://doi.org/10.1016/j.jbi.2024.104748

[6] Liu, B., Zhan, L., Xu, L. and Wu, X. (2023) Medical Visual Question Answering via Conditional Reasoning and Contrastive Learning. *IEEE Transactions on Medical Imaging*, **42**, 1532-1545. https://doi.org/10.1109/tmi.2022.3232411

[7] Li, P., Liu, G., Tan, L., Liao, J. and Zhong, S. (2023) Self-Supervised Vision-Language Pretraining for Medial Visual Question Answering. 2023 *IEEE* 20*th International Symposium on Biomedical Imaging* (*ISBI*), Cartagena, 18-21 April 2023, 1-5. https://doi.org/10.1109/isbi53787.2023.10230743

[8] Bhatti, U.A., Tang, H., Wu, G., Marjan, S. and Hussain, A. (2023) Deep Learning with Graph Convolutional Networks: An Overview and Latest Applications in Computational Intelligence. *International Journal of Intelligent Systems*, **2023**, Article ID: 8342104. https://doi.org/10.1155/2023/8342104

[9] Ren, F. and Zhou, Y. (2020) CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering. *IEEE Access*, **8**, 50626-50636. https://doi.org/10.1109/access.2020.2980024

[10] Wang, H. and Du, H. (2023) Knowledge-Enhanced Medical Visual Question Answering: A Survey (Invited Talk Summary). In: Yang, S. and Islam, S., Eds., *Web and Big Data. APWeb-WAIM* 2022 *International Workshops*, Springer, 3-9. https://doi.org/10.1007/978-981-99-1354-1_1

[11] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 618-626. https://doi.org/10.1109/iccv.2017.74

[12] Huang, X. and Gong, H. (2024) A Dual-Attention Learning Network with Word and Sentence Embedding for Medical Visual Question Answering. *IEEE Transactions on Medical Imaging*, **43**, 832-845. https://doi.org/10.1109/tmi.2023.3322868

[13] Chen, Z., Li, G. and Wan, X. (2022) Align, Reason and Learn: Enhancing Medical Vision-and-Language Pre-Training with Knowledge. *Proceedings of the* 30*th ACM International Conference on Multimedia*, Lisboa, 10-14 October 2022, 5152-5161. https://doi.org/10.1145/3503161.3547948

[14] Gu, T., Yang, K., Liu, D. and Cai, W. (2024) LaPA: Latent Prompt Assist Model for Medical Visual Question Answering. 2024 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Seattle, 17-18 June 2024, 4971-4980. https://doi.org/10.1109/cvprw63382.2024.00502

[15] Liu, B., Zhan, L. and Wu, X. (2021) Contrastive Pre-Training and Representation Distillation for Medical Visual Question Answering Based on Radiology Images. In: de Bruijne, M., *et al*., Eds., *Medical Image Computing and Computer Assisted Intervention—MICCAI* 2021, Springer, 210-220. https://doi.org/10.1007/978-3-030-87196-3_20

[16] Huang, J., Chen, Y., Li, Y., Yang, Z., Gong, X., Wang, F.L., *et al*. (2023) Medical Knowledge-Based Network for Patient-Oriented Visual Question Answering. *Information Processing & Management*, **60**, Article ID: 103241. https://doi.org/10.1016/j.ipm.2022.103241

[17] Chen, Z., Du, Y., Hu, J., Liu, Y., Li, G., Wan, X., *et al*. (2024) Mapping Medical Image-Text to a Joint Space via Masked Modeling. *Medical Image Analysis*, **91**, Article ID: 103018. https://doi.org/10.1016/j.media.2023.103018

[18] Eslami, S., Meinel, C. and de Melo, G. (2023) PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain? *Findings of the Association for Computational Linguistics*: *EACL* 2023, Dubrovnik, May 2023, 1181-1193. https://doi.org/10.18653/v1/2023.findings-eacl.88

[19] Lameesa, A., Silpasuwanchai, C. and Alam, M.S.B. (2025) VG-CALF: A Vision-Guided Cross-Attention and Late-Fusion Network for Radiology Images in Medical Visual Question Answering. *Neurocomputing*, **613**, Article ID: 128730. https://doi.org/10.1016/j.neucom.2024.128730

[20] Shu, C., Zhu, Y., Tang, X., Xiao, J., Chen, Y., Li, X., *et al*. (2024) MITER: Medical Image-Text Joint Adaptive Pretraining with Multi-Level Contrastive Learning. *Expert Systems with Applications*, **238**, Article ID: 121526. https://doi.org/10.1016/j.eswa.2023.121526

[21] Zhan, C., Peng, P., Wang, H., Wang, G., Lin, Y., Chen, T., *et al*. (2025) UnICLAM: Contrastive Representation Learning with Adversarial Masking for Unified and Interpretable Medical Vision Question Answering. *Medical Image Analysis*, **101**, Article ID: 103464. https://doi.org/10.1016/j.media.2025.103464

[22] Wu, Y., Lu, Y., Zhou, Y., Ding, Y., Liu, J. and Ruan, T. (2025) MKGF: A Multi-Modal Knowledge Graph Based RAG Framework to Enhance LVLMs for Medical Visual Question Answering. *Neurocomputing*, **635**, Article ID: 129999. https://doi.org/10.1016/j.neucom.2025.129999