

知识 - 视觉深度协同的 KB-VQA 应用研究

赵俊东¹ 余艳梅^{1*} 陶青川¹
ZHAO Jundong YU Yanmei TAO Qingchuan

摘要

知识的视觉问答 (KB-VQA) 任务旨在借助外部知识对复杂视觉问题进行解答, 然而现有方法仍面临诸多挑战, 如对文本检索知识的过度依赖易导致检索偏差、对图像视觉细节的捕捉能力不足, 且预训练模型的深层推理能力未得到充分挖掘。针对上述问题, 文章提出一种新型知识 - 图像融合与视觉特征集成模型 (KIVI)。该模型依托多模态预训练模型实现图文信息的深度融合, 以此提升外部知识的检索精度; 同时结合图像内容与检索知识, 生成聚焦图像关联信息的视觉多模态知识。此外, KIVI 模型整合图像的区域局部特征与全局整体特征, 保障视觉信息的完整提取; 并设计融合候选答案与推理路径的提示策略, 有效激活预训练模型的深层推理能力。实验结果表明, KIVI 模型在 OK-VQA 数据集上的准确率为 67.29%, 性能优于目前所有最优方法。

关键词

视觉问答; 知识检索; 多模态预训练模型; 深度学习

doi: 10.3969/j.issn.1672-9528.2026.02.006

0 引言

基于知识的视觉问答通过结合视觉理解与外部知识整合, 能够回答复杂的问题。这一任务需要计算机视觉技术分析图像内容、自然语言处理技术处理文本数据, 以及外部知识库提供准确知识的能力。KB-VQA 系统已广泛应用于医学图像分析、搜索引擎等领域。然而, 当前方法仍面临三大挑战: 一是知识检索易因模糊查询或算法局限引入错误 / 无关内容干扰推理 (如 “Apple” 一词的多义性导致 “Are apples red?” 与 “Is there an Apple computer?” 的检索干扰); 二是模型常忽略关键视觉信息 (如仅凭文本知识无法准确判断图中气球的颜色), 导致依赖图像的问答出错; 三是预训练模型的推理潜力未被充分挖掘, 将图像简单转成文本描述或提示会丢失视觉上下文, 限制了其处理复杂多模态任务的能力。

为解决上述挑战, 本文提出一种新型 KB-VQA 方法 KIVI:

- (1) 利用多模态预训练模型在共享嵌入空间中编码图文信息提升知识检索准确性;
- (2) 融合图文特征, 增强模型对图像相关文本的关注;
- (3) 采用多路径策略确保关键视觉细节不遗漏;
- (4) 通过双提示策略激发预训练模型生成

高质量隐式文本知识。

1 方法

本文工作主要包括五个部分: (a) 提出基于多模态预训练模型的密集检索机制, 解决同义词问题; (b) 通过文本 - 图像交互生成多模态知识特征, 降低文本噪声; (c) 设计多路径视觉模块 (融合文本描述、区域与全局特征), 全面刻画图像信息; (d) 引入候选答案与推理路径的双提示策略, 激发预训练模型潜力; (e) 最终由多模态推理模块整合信息进行答案预测, 整体框架如图 1 所示。

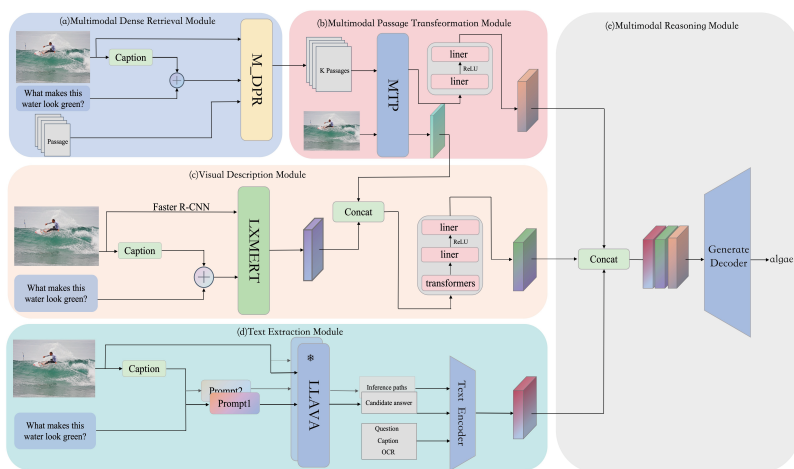


图 1 模型框架

1.1 多模态密集检索模块

现有的密集检索方法^[1]主要存在以下缺点：由于仅使用单模态语言模型以文本信息作为查询，导致无法充分利用图像信息，从而引发同义词检索错误。而多模态大模型的崛起，显著提升了图文融合理解能力，为复杂多模态任务开辟了新路径。

因此，本文提出了一种新型多模态密集检索模块以应对这些挑战。通过将图像、图像描述和问题同时输入到多模态预训练模型中，生成包含视觉信息的查询向量，从而减少同义词检索问题。该模块如图2所示，其主要工作流程如下：

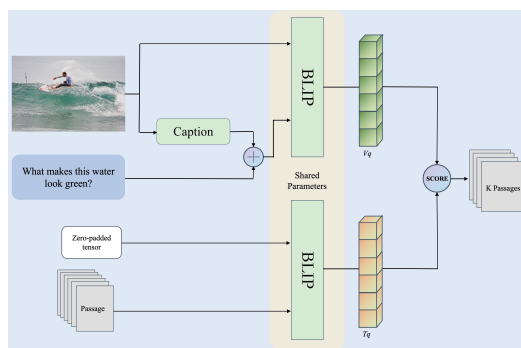


图2 多模态密集检索模块

对于查询，给定图像后，使用 ExpansionNet v2^[2] 模型生成该图像的文本描述。随后，将该描述与问题及图像一同输入至 BLIP^[3] 模型。文本信息通过 BLIP 与图像信息融合，最终查询向量通过“CLS”令牌进行表示。

对于纯文本知识文档，由于 BLIP 等多模态模型需双模态输入，采用与文献[4]相同的视觉零填充策略：将零张量作为虚拟视觉输入。共享参数的 BLIP 编码器输出“CLS”令牌作为知识向量。在检索阶段，查询向量 V_q 与所有知识向量 T_k 的内积计算相似度，并通过 FAISS 引擎加速 Top-K 检索。最终，相似度最高的 K 个段落作为显式知识输出。模型训练使用基于交叉熵的对比损失：

$$L = -\log \frac{e^{S((Q,I),P_{pos})}}{e^{S((Q,I),P_{pos})} + \sum_{P' \in P_{neg}} e^{S((Q,I),P')}} \quad (1)$$

式中： P_{pos} 是与问题-图像对 (Q, I) 正相关的段落集合； P_{neg} 是与之负相关的段落集合。

1.2 多模态知识转换模块

针对检索所得维基百科知识常含图像无关噪声，提出基于 BLIP 的多模态知识转换机制。不同于直接使用文本知识的方法^[5]，该机制利用 BLIP 强大的图文交互能力，将检索结果与图像共同编码，生成聚焦图像相关信息的多模态知识表征。

具体来说，冻结预训练最优检索模型的参数，检索 Top-K 知识条目。将该知识与图像配对输入 BLIP 编码器，

并提取其文本侧输出作为多模态交互知识表征。此表征本质上是融合了图像上下文的维基百科知识嵌入，用公式表示为：

$$W_{iki}^{(n)} = \text{BLIP}^* ((I, T_1), (I, T_2), \dots, (I, T_K)) \quad (2)$$

式中： $W_{iki}^{(n)}$ 为转换后的多模态知识；BLIP* 为多模态密集检索模型的最优模型参数初始化。

1.3 视觉特征提取模块

KB-VQA 需融合视觉感知与知识推理，但现有方法过度依赖大语言模型提供的文本信息，仅仅将图像转换为文本，导致视觉直觉缺失。为此，本文提出多特征融合机制，联合图像字幕、区域及全局特征，实现全面视觉信息提取。

为全面捕捉图像的信息，首先使用 ExpansionNet v2 生成开放式文本描述，以实现整体语义理解。同时，通过 EasyOCR 进行光学字符识别，从图像中提取可见文本。二者的协同集成显著增强模型对视觉与文本细粒度特征的感知能力。

接下来，利用 Faster R-CNN^[6] 从输入图像中检测出 36 个区域并提取其对应的 2 048 维特征向量。随后，将这些区域特征与文本输入（即问题和图像描述）共同输入 LXMERT 模型^[7] 进行跨模态融合。LXMERT 能够基于文本语义自适应地关注图像中的相关区域，从而有效增强对图文信息的联合理解与整合能力：

$$V_i = \text{LXMERT}(Q, C, \text{Faster R-CNN}(I)) \quad (3)$$

式中： V_i 表示区域特征； Q 、 C 、 I 分别表示问题、图像描述和图像。

最后，为获得更全面的图像表征，将多模态知识转换模块的视觉分支输出作为图像的全局特征，使其在向量空间中获得整体性表示，从而更准确地捕捉图像的整体语义信息：

$$V_g = \frac{1}{K} \sum_{k=1}^K \text{BLIP}^* ((I, T_1), (I, T_2), \dots, (I, T_K)) \quad (4)$$

式中： V_g 表示全局特征； $\{T_1, T_2, \dots, T_K\}$ 表示检索到的文本知识输入的数量。

1.4 文本提取模块

近期研究利用单模态语言模型生成文本知识，但其推理能力受限于无法捕捉视觉线索。多模态预训练模型虽能融合图文信息，其图文融合能力仍有提升空间。为此，提出一种提示方案，引导多模态大模型聚焦关键图文特征、深入挖掘信息关联。基于 LLaVA-1.5^[8] 模型，设计了两步优化流程：

步骤一（候选答案生成， $M_{lm,a}$ 表示）：将问题与图像说明结合，输入包含约束条件的提示（示例：“question: {question} context: {caption} Answer the question based on the image and context, using only a word or phrase.”），驱动模型综合视觉与文本上下文生成候选答案。

步骤二（推理路径生成， M_{llm_r} 表示）：输入结构化推理提示（示例：“question:{question} context:{caption}Please provide rationales for the question based on the context and image.”），引导模型输出逻辑链条清晰地逐步推理路径。该方法有效提升了复杂查询的推理精度，增强结论可解释性。

1.5 多模态推理模块

面对大量的视觉、文本及多模态知识，下一步便是预测正确答案。对于文本知识，包括来自多模态大型模型的隐式知识、图像字幕（ C_{ap} ）、光学字符识别（ O_{cr} ）以及问题（ Q ），使用预训练的 T5-Large 模型的编码器对文本进行统一编码。输入序列为：

$$Z_i = \text{enc}_{\text{txt}}(Q, O_{cr}, M_{llm_a}, C_{ap}, M_{llm_r}) \quad (5)$$

式中： Z_i 是第 i 个 token 的表示。

为整合视觉信息，沿特征维度拼接区域特征（ V_i ）和全局特征（ V_g ）。该拼接序列输入九层 Transformer，通过其注意力机制进行深度交互建模。最终融合表示由两层线性层对齐跨模态特征空间：

$$V'_i = \text{FC}_2 \left(\text{ReLU} \left(\text{FC}_1 \left(T_{\text{Transformers}}(|V_i, V_g|) \right) \right) \right) \quad (6)$$

式中： $\text{FC}()$ 是一个用于线性变换的函数； V'_i 是图像在第一维度上融合区域和全局特征后得到的总视觉特征。

为对齐特征空间，使用同一线性层将视觉交互后的维基知识嵌入序列投影至文本编码器空间：

$$W_i = \text{FC}_2 \left(\text{ReLU} \left(\text{FC}_1 \left(W_{\text{iki}}^{(n)} \right) \right) \right) \quad (7)$$

式中： W_i 表示在文本编码器的特征空间中映射后的维基知识嵌入序列； $W_{\text{iki}}^{(n)}$ 表示与视觉信息交互后的维基知识嵌入序列。

然后，将视觉嵌入序列、文本嵌入序列和维基知识嵌入序列沿第一维度进行拼接，以获得混合序列 S'_i ，用公式表示为：

$$S'_i = \text{concat}(V'_i, Z_i, W_i) \quad (8)$$

式中： V'_i 是视觉嵌入序列； Z_i 是文本嵌入序列； W_i 是维基知识嵌入序列。

最后，将该混合序列输入预训练的 T5 模型，根据预测概率生成最终答案：

$$P(a'_1), \dots, P(a'_l) = \text{softmax}(\text{dec}(S'_i)) \quad (9)$$

式中： l 表示答案的长度。

使用交叉熵损失函数为：

$$\mathcal{L} = - \sum_{l=1}^L \log P_{\theta}(a_l | a_{<l}) \quad (10)$$

式中： L 是答案文本的长度； a_l 是答案文本在位置 l 处的文本； θ 是模型参数。

2 实验

2.1 数据集与实验设计

本文研究方法主要在目前最大且最具挑战性的 KB-VQA 数据集 OK-VQA^[9] 上进行评估。该数据集包含 14 055 个问题。训练集和测试集分别包含 9 009 个和 5 046 个样本。

实验采用 NVIDIA A40（48 GB）进行训练，使用 AdamW 优化器（学习率 10^{-5} ）。基础模型训练配置：20 000 步，batch size=2，每 1 000 步评估。多模态密集检索模块训练配置：20 000 步，batch size=4，每 2 000 步评估。检索设置：Top-6 维基百科条目，最大输入长度 240 tokens；T5-Large 解码器最大输入长度 440 tokens。实验采用 VQAv2^[10] 软准确率（Soft Accuracy）作为核心评估指标。

2.2 实验结果

为了评估提出的方法的有效性，将它与四组具有影响力的最先进模型进行了比较，具体如表 1 所示。比较结果如下：

（1）与无预训练模型相比，通过 MLLM 跨模态交互增强图文关系理解，性能提升 20.61%；

（2）相较于依赖 GPT-3 进行文本提取的方法，通过融合视觉特征克服其视觉上下文局限性并降低提示敏感性，性能提升 10.69%；

（3）相比利用 GPT-3 API 生成答案的方案，在避免其高昂开销与延迟的同时，精度提升 6.19%，且效率更高；

（4）与其他预训练模型相比，凭借以图像为中心的策略和高效提示设计优化了推理与任务处理能力，性能提升 3.69%。

表 1 主实验准确率结果

#	方法	知识来源	ACC/%
无预训练模型			
1	苏振强等 ^[11]	Wikipedia	34.01
2	Xu 等 ^[12]	Wikipedia + ConceptNet	46.68
GPT-3 提取文本			
3	KAT ^[13]	GPT-3(175B)+Wikidata	53.09
4	REVIVE ^[14]	GPT-3(175B)+Wikidata	56.60
GPT-3 (API) 生成答案			
5	PromptCap ^[15]	GPT-3(175B)	60.40
6	Prophet ^[16]	GPT-3(175B)	61.10
其他预训练模型			
7	郑洪岩等 ^[17]	Google Search+GPT-3(175B)+OFA	61.30
8	GeReA ^[18]	LLaVA-1.5-7B(7B)	63.60
9	KIVI(ours)	Wikipedia(m)+LLaVA-1.5-7B(7B)	67.29

2.3 消融实验

6 个独立组件对模型性能的影响如表 2 所示。来自图像文本描述和 OCR 的上下文信息、区域特征、全局特征、通过图像交互获取的显式知识、推理路径以及候选答案。

逐步整合 6 个关键组件持续提升模型性能。具体效果如

下:

- (1) 仅使用图像文本描述时, OK-VQA 准确率为 38.20%;
- (2) 加入区域特征, 准确率提升 1.09%;
- (3) 进一步整合全局特征(增强图像细节与整体理解), 准确率再增加 1.26%;
- (4) 将显式知识与图像融合为模型提供了更多相关信息, 提升 1.48%, 有效应对高度开放式问题;
- (5) 引入推理路径引导逻辑推导(显著增强可解释性与连贯性), 带来 16.09% 的显著跃升;
- (6) 添加候选答案作为优化参考, 性能进一步提高 9.17%。结果表明, 视觉信息整合、知识-图像融合与 MLLM 的充分激活共同驱动了准确性的实质性提升。

表 2 消融实验准确率结果

Context	V_l	V_g	Exp	M_{llm_r}	M_{llm_a}	ACC/%
√						38.20
√	√					39.29
√	√	√				40.55
√	√	√	√			42.03
√	√	√	√	√		58.12
√	√	√	√	√	√	67.29

3 结果分析

3.1 MLLM 推理激活案例研究

图 3 展示了 LE 如何有效激活 MLLM 的推理能力及在该任务中提出的提示策略的具体应用。

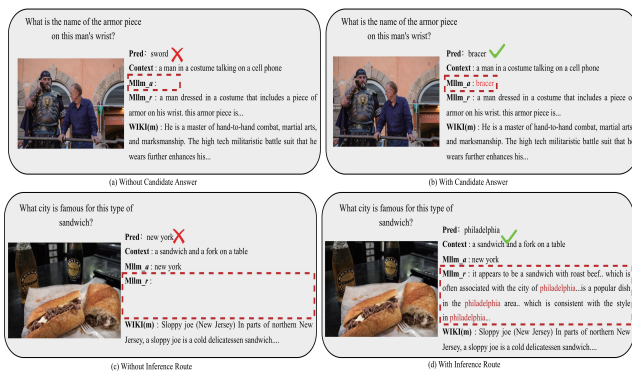


图 3 MLLM 推理激活案例研究

图 3 (a) 和 (b) 阐释了候选答案的关键作用。具体而言: 图 3 (a) 显示在候选答案缺失时, 模型因无法从上下文、推理路径及多模态知识中获取与正确答案相关的信息(存在知识盲区)而导致预测失效; 图 3 (b) 表明借助目标导向的提示策略, 模型能成功生成正确的候选答案并据此达成最终准确预测。验证了候选答案在生成可靠预测结果中的不可或缺性。

图 3(c) 和 (d) 展示了推理路径的关键作用。其中, 图 3(c)

揭示了在缺乏上下文、候选答案及相关多模态知识的支持下, 模型易于作出误导性误判。相反, 在图 3 (d) 中, 推理路径引导模型通过逐步推理探索问题与图片之间的关系。这种结构化方法提供了清晰的逻辑链, 使模型能够生成准确的答案。这验证了推理路径不仅能显著提升模型逻辑推理能力, 更能实质性地增强其在复杂情境下的正确答案识别能力。

3.2 知识与图像融合案例研究

图 4 (a) 与 (b) 展示了知识-图像融合的价值: 在应对高度开放性问题时, 候选答案、推理路径及上下文常因信息局限难以提供准确支持。如图 4 (a) 所示, 问题 “What has been planted in this field?” 中作物特征视觉不可见, 导致模型无法生成相关候选, 推理路径亦失效; 仅依赖上下文词汇 “grass” 因而产生误判。相反, 图 4 (b) 表明, 融合图像的维基百科知识提供充分支撑, 驱动模型生成准确响应。这直接证实, 通过图像交互获取的显式知识是解决此类开放性问题不可或缺的关键。

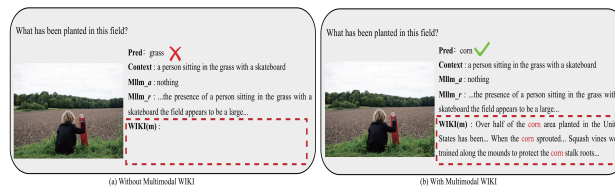


图 4 知识与图像融合案例研究

3.3 视觉描述模块案例研究

如图 5 (a) 所示, 尽管 MLLM 生成的推理路径包含正确知识, 但由于缺乏视觉描述模块, 模型最终预测的答案与 MLLM 生成的候选答案相同, 导致结果错误。这表明视觉描述模块有助于模型聚焦于与图像本身更密切相关的文本信息, 从而与文本知识更好地协作以预测正确答案。相比之下, 图 5 (b) 显示, 即使所有文本知识与问题无关, 模型仍可在视觉描述模块的帮助下提供正确答案。这两个案例均聚焦于需要深度图像理解的问题, 验证了直接从视觉信息中提取关键细节的有效性。共同证明了视觉描述模块在显著增强模型图像解析能力方面不可或缺。

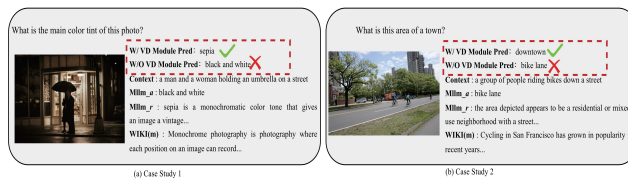


图 5 视觉描述模块案例研究

4 结语

本文提出了一种创新的多模态知识驱动型视觉问答模型, 该模型有效解决了该领域中的多个挑战。首先, 采用了一种先进的多模态预训练技术, 充分利用图像描述来引导

MLLM 模型更关注与问题密切相关的图像区域。通过引入候选答案生成和推理路径设计,成功激活了冻结的多模态模型的推理能力。此外,采用多模态密集检索策略,并通过将检索到的文本知识转换为多模态知识来减轻噪声的影响。并结合图像文本描述、区域特征和全局特征来增强视觉信息提取,确保模型不仅理解局部细节,还能深入把握整体视觉内容。实验结果表明,本文所提模型在 OK-VQA 数据集上优于当前最先进的方法。通过消融实验,进一步验证了模型各组件的有效性。

参考文献:

- [1]GAO F, PING Q, THATTAI G, et al. Transform-retrieve-generate: natural language-centric outside-knowledge visual question answering[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway:IEEE, 2022: 5067-5077.
- [2]HU J C, CAVICCHIOLI R, CAPOTONDI A. Exploiting multiple sequence lengths in fast end to end training for image captioning[EB/OL]. (2024-01-19)[2025-12-22]. <https://doi.org/10.48550/arXiv.2208.06551>.
- [3]LI J N, LI D X, XIONG C M, et al. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation[EB/OL]. (2022-02-15)[2025-12-22]. <https://doi.org/10.48550/arXiv.2201.12086>.
- [4]SALEMI A, PIZZORNO J A, ZAMANI H, et al. A symmetric dual encoding dense retrieval framework for knowledge-intensive visual question answering[C]//SIGIR '23: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2023: 110-120.
- [5]WU J L, MOONEY R J. Entity-focused dense passage retrieval for outside-knowledge visual question answering[EB/OL]. (2022-10-20)[2025-12-22]. <https://doi.org/10.48550/arXiv.2210.10176>.
- [6]REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(6):1137-1149.
- [7]TAN H, BANSAL M. LXMERT: learning cross-modality encoder representations from transformers[EB/OL]. (2019-12-03)[2025-12-22]. <https://doi.org/10.48550/arXiv.1908.07490>.
- [8]LIU H T, LI C Y, WU Q Y, et al. Visual instruction tuning[EB/OL]. (2023-12-11)[2025-12-22]. <https://doi.org/10.48550/arXiv.2304.08485>.
- [9]MARINO K, RASTEGARI M, FARHADI A, et al. OK-VQA: a visual question answering benchmark requiring external knowledge[EB/OL]. (2019-09-04)[2025-12-22]. <https://doi.org/10.48550/arXiv.1906.00067>.
- [10]AGRAWAL A, LU J S, ANTOL S, et al. VQA: visual question answering[EB/OL]. (2016-10-27)[2025-12-22]. <https://doi.org/10.48550/arXiv.1505.00468>.
- [11]苏振强, 苟刚. 联合知识和视觉信息推理的视觉问答研究[J]. 计算机工程与应用, 2024, 60 (5): 95-102.
- [12]XU N, GAO Y F, LIU A A, et al. Multi-modal validation and domain interaction learning for knowledge-based visual question answering[J]. IEEE transactions on knowledge and data engineering, 2024, 36(11):6628-6640.
- [13]GUI L K, WANG B R, HUANG Q Y, et al. KAT: a knowledge augmented transformer for vision-and-language[EB/OL]. (2022-05-05)[2025-12-22]. <https://doi.org/10.48550/arXiv.2112.08614>.
- [14]LIN L Z, XIE Y J, CHEN D D, et al. REVIVE: regional visual representation matters in knowledge-based visual question answering[C]//NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems. New York: ACM, 2022:10560-10571.
- [15]HU Y S, HUA H, YANG Z Y, et al. PromptCap: prompt-guided image captioning for VQA with GPT-3[C]//IEEE International Conference on Computer Vision. Piscataway: IEEE, 2023: 2951-2963.
- [16]SHAO Z W, YU Z, WANG M, et al. Prompting large language models with answer heuristics for knowledge-based visual question answering[C/OL]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023[2025-12-15]. <https://ieeexplore.ieee.org/document/10203158>. DOI:10.1109/CVPR52729.2023.01438.
- [17]郑洪岩, 王慧, 刘昊, 等. 基于隐式知识增强的 KB-VQA 知识检索策略研究[J]. 图学学报, 2024, 45 (6): 1231-1242.
- [18]MA Z Y, LI S T, SUN B, et al. GeReA: question-aware prompt captions for knowledge-based visual question answering[EB/OL]. (2024-02-04)[2025-12-22]. <https://doi.org/10.48550/arXiv.2402.02503>.

【作者简介】

赵俊东 (2002—), 男, 贵州毕节人, 硕士, 研究方向: 计算机视觉、大模型应用, email: zhaojd@stu.scu.edu.cn。

余艳梅 (1975—), 通信作者 (email: yuyanmei@scu.edu.cn), 女, 四川广安人, 博士, 副教授、硕士生导师, 研究方向: 图像处理。

陶青川 (1972—), 男, 四川南充人, 博士, 副教授、硕士生导师, 研究方向: 模式识别与智能系统。

(收稿日期: 2025-08-20 修回日期: 2026-02-06)