MedKInstruct: A Multimodal Knowledge Graph Based Framework for Multi-Hop and Hard-Negative Instruction Data Synthesis in MedVQA

Yinan Wu[1]*, Jihang Jin[1]*, Xuhao Bao[1]*, Hanjing Yan[1], Weiyan Zhang[1]†, Tong Ruan[1]†, Chunming Wang[2]
[1]East China University of Science and Technology, Shanghai, China,
[2]Renji Hospital Affiliated to Shanghai Jiaotong University School of Medicine, Shanghai, China
Correspondence: y21220035@mail.ecust.edu.cn, ruantong@ecust.edu.cn, weiyanzhang@ecust.edu.cn

Abstract

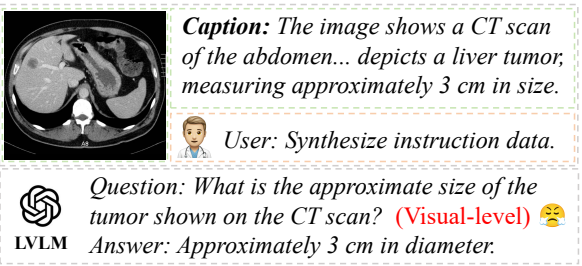
Medical visual question answering (MedVQA) requires models to provide accurate answers given a medical image and a corresponding question. Recently, instruction tuning of general large vision–language models (LVLMs) has become a dominant paradigm for this task, enabling open-ended predictions and effective integration of multimodal information. However, existing methods synthesize instruction data from image–caption pairs that primarily focus on visual attributes, rather than knowledge-level QA generation. This situation limits the model's ability to learn relevant medical knowledge during training, thereby restricting its performance on MedVQA. Hence, this paper proposes MedKInstruct, which incorporates a multimodal medical knowledge graph (MMKG) to assist LVLMs in synthesizing knowledge-intensive instruction data. Additionally, we design an MMKG path–based reward function to train a stronger MedVQA model through reinforcement learning. Experimental results on the public datasets Slake and VQA-RAD show that MedKInstruct outperforms previous methods by 4.16% and 4.50%. The source code is available at the following link: https://github.com/Sonder-hang/MedKinstruct

1 Introduction

Medical Visual Question Answering (MedVQA) aims to build a model that can answer natural language questions based on medical images. For example, given a medical image of "liver tumor" and a question "Is there an abnormality in this image?", a MedVQA model is expected to answer "Yes". Such a model effectively assists medical professionals in clinical decision-making and helps alleviate their workload (Liang et al., 2024).

The limitations of previous approaches. Traditional methods formulate MedVQA as a classification task, requiring models to select answers from a fixed set of predefined candidates. These approaches primarily focus on multimodal feature interactions (Huang and Gong, 2024; Zhang et al., 2024b) or pretraining on external medical image-caption resources (Eslami et al., 2023; Zhang et al., 2023). Such methods are restricted to closed-set responses, which limits their applicability in real-world clinical scenarios. (Li et al., 2024; Chen et al., 2024a).

With the rise of large vision–language models (LVLMs), LVLM-based generative methods have become the dominant solution for MedVQA, enabling open-set responses. These methods can be categorized into two groups. The first group focuses on in-context learning, where LVLMs are guided to perform MedVQA by few-shot cases (Moor et al., 2023), relevant medical knowledge (Wu et al., 2025), and chain-of-thought instructions (Wei et al., 2024; Jiang et al., 2025). However, LVLMs struggle to comprehend complex medical



(a) LVLMs directly generate QA pairs.

(b) LVLMs generate QA pairs with an MMKG.

Figure 1: An example of how the MMKG benefits the LVLM for generating knowledge-level QA pairs.

* Equal Contribution.
† Corresponding Authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 27935–27947

knowledge without domain-specific training, which limits the performance of these approaches (Zhang et al., 2024a).

Hence, the second group aims to build medical LVLMs. These works first employ an advanced closed-source LVLM (e.g., GPT-4o[1]) to synthesize QA pairs based on medical image captions (Li et al., 2024; Chen et al., 2024b), which are then used to instruction-tune an open-source LVLM (e.g., LLaVA (Liu et al., 2023)) for MedVQA. However, such synthetic QA pairs mainly focus on visual attributes, which limits the models' ability to learn associated medical knowledge during training. For example, given the image-caption pair shown in Figure 1 (a), it is challenging for GPT-4o to generate meaningful QA pairs related to the knowledge of "liver tumor" (e.g., its cause).

Our idea and contributions. We argue that introducing a multimodal medical knowledge graph (MMKG) provides an effective way to assist LVLMs in generating medical knowledge-intensive QA pairs. For example, after incorporating the MMKG triples shown in Figure 1(b), GPT-4o can synthesize multi-hop questions concerning the causes of liver tumors. However, incorporating the MMKG introduces two new challenges. (1) How to associate image captions with the MMKG: medical entities mentioned in image captions are often misaligned with entities in the MMKG (e.g., "lung mass" and "pulmonary mass"). (2) How to effectively utilize the knowledge-intensive QA pairs: conventional supervised fine-tuning (SFT), which focuses solely on the final answer, fails to supervise the model's use of intermediate medical knowledge required for multi-hop reasoning (Li et al., 2024; Chen et al., 2024b).

Hence, in this paper, we propose MedKInstruct, a novel framework for synthesizing medical instruction data. The core idea is that we first leverage an LVLM to generate medical QA pairs with image captions and the MMKG, and then use the synthesized QA pairs to train a stronger MedVQA model. Specifically, given an image-caption pair and an MMKG, we first design a two-stage matching strategy to ground caption entities into the MMKG. This strategy consists of exact string matching followed by embedding-based matching with LVLM verification, which effectively alleviates the entity misalignment problem. Based on the MMKG-grounded medical image–caption pairs, we further employ an LVLM to generate multi-hop and hard-negative QA pairs that require medical knowledge reasoning. Finally, we use the synthesized data and MMKG to train a MedVQA model with supervised fine-tuning (SFT) and reinforcement learning (RL) (Guo et al., 2025). SFT employs QA pairs to equip the model with fundamental medical knowledge, while RL enhances the model's complex reasoning capability through a designed MMKG path–based reward function. Our main contributions can be summarized as follows:

- We propose an instruction data synthesis framework, named MedKInstruct. To the best of our knowledge, it is the first approach to leverage an MMKG to guide LVLMs in generating multimodal medical QA pairs.
- We design a novel MMKG path-based reward function for RL. This function guides the model to think relevant medical knowledge and produce the accurate final answer during MedVQA.
- Extensive experiments are conducted on two publicly available MedVQA datasets, Slake and VQA-RAD. The proposed MedKInstruct framework achieves improvements of 4.16% and 4.50%, respectively.

2 Related works

In this section, we provide an overview of previous MedVQA and RL-based LLM reasoning methods.

MedVQA methods. Existing methods generally formulate MedVQA as a classification or generative task. The classification method primarily investigate attention-based multimodal fusion modules (Huang and Gong, 2024; Zhang et al., 2024b) or large-scale pretraining (Eslami et al., 2023; Zhang et al., 2023). However, these methods rely on predefined answer candidates, which is often unavailable in real-world scenarios. Hence, some works apply general LVLMs to perform generative MedVQA through in-context learning (Moor et al., 2023; Wu et al., 2025; Wei et al., 2024; Jiang et al., 2025). However, without domain-specific training, it is challenging for the general LVLM to comprehend complex medical knowledge (Zhang et al., 2024a). To solve this problem, recent works have explored synthesizing instruction data to train a medical LVLM. They focus on designing diverse QA scenarios (Li et al., 2024; Zhang et al., 2024a; Chen

[1] https://chat.openai.com/

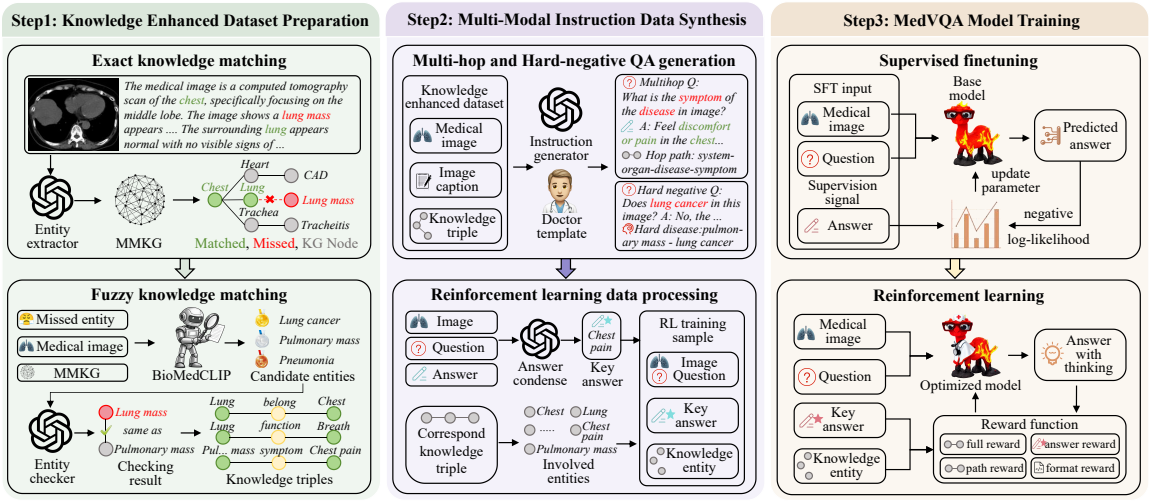


Figure 2: The framework of MedKInstruct which mainly consists of three parts: (1) knowledge enhanced dataset preparation, (2) multimodal instruction data synthesis and (3) MedVQA model training.

et al., 2024b) and incorporating thought chains (Team et al., 2025; Xu et al., 2024). However, these methods mainly synthesize visual-level QA pairs derived from the medical image–caption. In contrast, this paper introduces an MMKG to obtain knowledge-intensive QA pairs, which enabling the training of a stronger medical LVLM.

RL methods for LLM reasoning. From the perspective of training policy, existing RL methods for enhancing LLM reasoning can be divided into two groups. The first group focuses on the final answer, using answer accuracy as the sole reward signal to guide the generation of the reasoning process (Guo et al., 2025; Chen et al., 2025). Recently, this paradigm has been successfully extended to the medical domain for training medical LVLMs (Pan et al., 2025; Lai et al., 2025). However, the lack of fine-grained guidance during reasoning limits the interpretability of the model's responses. Hence, the second group is proposed, which emphasizes intermediate reasoning supervision. Existing approaches in this line primarily focus on designing self-reflection mechanisms (Shen et al., 2025; Wang et al., 2025; Yu et al., 2025) or incorporating external critic models to provide feedback (Liu et al., 2025; Wang et al., 2024). However, model-based feedback suffers from high computational cost and hallucinations issues. Hence, this paper explores the use of MMKG paths to guide the model's reasoning process.

3 Overview

In this section, we first formulate the MedVQA task and then present the overall framework.

3.1 Problem formulation

We first introduce several key concepts used in this paper and then define the MedVQA task.

MMKG. An MMKG is defined as $G = \{E, R, I\}$, where $E = \{e_1, e_2, \ldots, e_n\}$ denotes a set of medical entities, and R represents the set of relation triples between head and tail entities. Unlike textual KGs, each MMKG entity $e_i \in E$ is associated with a corresponding medical image $I_i \in I$.

Multimodal instruction data synthesis. Given a medical image–caption pair $\{P_i, C_i\}$, this process aims to leverage an LVLM to synthesize multiple instruction–response pairs $\{P_i, Q_i, A_i\}$. Formally, the data synthesis procedure can be expressed as $\{P_j, Q_j, A_j\} = \text{LVLM}(P_j, C_j)$.

MedVQA. This task requires the model M to return a correct answer A_i of the question Q_i based on the medical image P_i, which can be defined as $A_i = M(Q_i, P_i)$.

3.2 Overall Framework

As depicted in Figure 2, the proposed MedKInstruct framework mainly consists of three modules: (1) knowledge enhanced dataset preparation, (2) multimodal instruction data synthesis and (3) MedVQA model training.

Knowledge enhanced data preparation. Given a medical image–caption pair, we first extract the medical entities from the given caption and then associate these entities with the relevant knowledge triples in the MMKG.

Multimodal instruction data synthesis. After obtaining the knowledge-enhanced medical im-

age–caption data, we employ the doctor-designed template and LVLM to synthesize multi-hop and hard-negative QA pairs, which are used for Med-VQA model training.

MedVQA model training. After obtaining the synthesized instruction data, we first perform SFT on a VLM using the QA pairs. Then, we incorporate the corresponding MMKG path information and further optimize the model through RL.

4 Methodology

In this section, we detail the three main parts of the proposed MedKInstruct framework.

4.1 Knowledge enhanced dataset preparation

Given a medical image–caption dataset, this step aims to associate each image–caption pair with relevant triples in the MMKG. To this end, we design a two-stage knowledge matching strategy, which consists of string-based matching and embedding-based matching.

Exact knowledge matching. Specifically, given a medical image-caption pair $\{P_j, C_j\}$, we first feed them into an LVLM and prompt it to extract medical entities using the extraction prompt EP: "*Please extract the organ or disease entity in the given caption.*" (see Appendix A for the detailed prompt). This process can be formulated as:

$$E_j = \text{LVLM}(P_j, C_j, EP), \tag{1}$$

where E_j denotes as the set of extracted entities. Then we perform exact string matching to ground each entity in E_j to the MMKG, and obtain the string-matched entity set SE_j.

Fuzzy knowledge matching. For these unmatched entities, we further perform a fuzzy knowledge matching stage to improve coverage. Specifically, given an unmatched entity $e_j \in E_j$, we first embed e_j, the medical image P_j, and MMKG entities E using a pre-trained medical multimodal encoder (e.g., BiomedCLIP (Zhang et al., 2023)). Then, we compute the cosine similarities and retrieve the top-k most similar MMKG entities as potential matches PE_j. This process can be formulated as:

$$\begin{aligned} PE_j = \{e_p | \textit{top-}k(\textit{cos}(\text{MM-Enc}(P_j, I_p)) \\ +\textit{cos}(\text{MM-Enc}(e_j, e_p))), e_p \in E\}, \end{aligned} \tag{2}$$

where $MM\text{-}Enc$ denotes the pre-trained medical multimodal encoder, and I_p is the image of the MMKG entity e_p. After obtaining PE_j, we employ an LVLM to check the semantic consistency between e_j and each entity in PE_j (see Appendix A for the detailed prompt). The entities that fail the verification are removed. Finally, we retrieve the MMKG triples of the verified entities in PE_j and the previously string-matched entities in SE_j. The retrieved triples K_j are integrated with the original medical image-caption pair $\{P_j, C_j\}$ to construct a knowledge-enhanced dataset $\{P_j, C_j, K_j\}$.

4.2 Multimodal instruction data synthesis

After obtaining the knowledge-enhanced dataset, we first instruct the LVLM to generate multi-hop and hard negative QA pairs. Then, we simplify the answer and MMKG path to construct RL data.

Multi-hop QA generation. Specifically, given a knowledge-enhanced medical image–caption pair $\{P_j, C_j, K_j\}$, we first concatenate the knowledge triples K_j into two categories of reasoning paths: (1) short paths K_j^s, which consist of system-level and organ-level entities and (2) long paths K_j^l, which include the disease entity and its associated attributes. K_j^s is used to synthesize simple QA pairs that focus on the visual information directly observable from P_j, while K_j^l is employed to generate more complex QA pairs that require additional medical knowledge. Then, we feed each path, the image–caption pair, along with a doctor-designed template into the LVLM, and instruct it to generate QA data (see Appendix A for the detailed prompt). This process is formulated as:

$$(q_j^i, a_j^i) = \text{LVLM}\big(P_j, C_j, k_j^i\big), k_j^i \in K_j^s, K_j^l, \tag{3}$$

where k_j^i denotes a specific knowledge path, and (q_j^i, a_j^i) is a generated multi-hop QA pair.

Hard negative QA generation. Given a medical image P_j, we first use the medical multimodal encoder to retrieve the top-m disease entities from the MMKG that are highly similar to the abnormality depicted in P_j. This retrieved disease entity set DE_j is defined as:

$$\begin{aligned} DE_j = (\{e_d | \textit{top-}m(\textit{cos}(\text{MM-Enc}(P_j), \\ \text{MM-Enc}(I_d))), e_d \in E\}), \end{aligned} \tag{4}$$

where I_d is the image of the MMKG entity e_p. Then, we insert each entity $e_d \in DE_j$ into the fixed templates and construct the hard negative QA pairs (e.g., "*Q: Does {e_d} appeared in the given image, A: No*"). Finally, we attach the MMKG

path associated with DE_j to the corresponding hard negative QA pair, which is used for MedVQA model training.

RL data processing. After obtaining the generated QA pairs, we aim to transform them into the training samples for RL. Specifically, given a QA pair $\{P_j, q_j^i, a_j^i\}$ and its corresponding MMKG path k_j^i, we first employ the LVLM to condense the key part of the answer a_j^i (e.g., "chest pain" in Figure 2). This process is formulated as:

$$ka_j^i = \text{LVLM}(P_j, q_j^i, a_j^i, SP), \tag{5}$$

where $ka_j^{i'}$ is the key answer, and SP is the used summary prompt (see Appendix A for the detailed prompt). Then, we extract the set of involved entities ke_j^i from the MMKG path k_j^i. The tuple $\{P_j, q_j^i, ka_j^i, ke_j^i\}$ is treated as a training sample for subsequent RL process.

4.3 MedVQA model training

After obtaining the synthesized instruction data, we employ it to train a MedVQA model. To this end, we first introduce a VLM (e.g., LLaVA (Liu et al., 2023)) as the base model and then train it in two stages: (1) SFT and (2) RL.

SFT. Specifically, given a base model M and a training sample consisting of a medical image P_j, a question q_j^i, and the corresponding answer a_j^i, we first feed $\{P_j, q_j^i\}$ into M and obtain its predicted response. Then, we compute the negative log-likelihood of the answer a_j^i based on the model's prediction. This process is formulated as:

$$\mathcal{L}_{\text{SFT}}(M) = -\log p_M(a_j^i \mid P_j, q_j^i), \tag{6}$$

where p_M denotes the probability assigned by the model M to the answer a_j^i. Finally, we optimize the model parameters by minimizing the training loss $\mathcal{L}_{\text{SFT}}$.

RL. After obtaining the SFT model M', we employ Group Relative Policy Optimization (GRPO) (Guo et al., 2025) to strengthen its reasoning ability with relevant medical knowledge. Specifically, given a training input $\{P_j, q_j^i\}$, we first use a "think prompt" (Pan et al., 2025) to guide M' in outputting its answer and reasoning steps. Then, we employ the key answer ka_j^i and MMKG path ke_j^i to optimize its output with the designed reward function, which has four parts.

(1) Answer reward R_{ans}: if ka_j^i appears in the model's predicted answer, a reward of 1 is given. Otherwise, R_{ans} is set to 0.

Table 1: Comparison of typical multimodal medical QA data construction methods. "MH" means multi-hop and "HN" means hard negative.

Methods	Auto	Image	Think	MH	HN
VQA-RAD	✗	✓	✗	✗	✗
Slake	✗	✓	✗	✓	✗
LLava-Med	✓	✗	✗	✗	✗
PMC	✓	✗	✗	✗	✗
Huatuo	✓	✓	✗	✗	✗
Lingshu	✓	✓	✓	✗	✗
MedKInstruct	✓	✓	✓	✓	✓

(2) MMKG path reward R_{path}: If r entities in ke_j^i appear in the model's output, a reward of $\frac{r}{|ke_j^i|}$ is given, where $|ke_j^i|$ denotes the total number of entities in the MMKG path.

(3) Full reward R_{full}: If all entities in k_j^i appear in the model's output, another reward of 1 is given. This function encourages the model to make a comprehensive thinking.

(4) Format reward R_{mat}: if the model's output following the format of "<think>...</think> <ans>...</ans>", a reward of 0.5 is given. Otherwise, R_{mat} is set to 0.

5 Experiment

In this section, we first detail the experimental setup and present the results of MedKInstruct in comparison with baseline methods. Then, we conduct ablation studies and provide detailed analyses to demonstrate the importance of incorporating the MMKG and our RL policy.

5.1 Experimental Setup

Datasets. Following the previous MMKG-based work (Wu et al., 2025), we conduct experiments on two publicly available MedVQA datasets: (1) Slake (Liu et al., 2021) and (2) VQA-RAD (Lau et al., 2018). The detailed characteristics of each dataset are shown in Table 1.

Baselines. We compare MedKInstruct with other state-of-the-art multimodal medical instruction data synthesis methods, which belong to two categories: (1) caption-based methods, including LLaVA-Med-Instruct (Li et al., 2024) and PMC-Instruct (Zhang et al., 2024a), and (2) image–caption-based methods, including Huatuo-Instruct (Chen et al., 2024b) and Lingshu-Instruct (Team et al., 2025). The detailed characteristics of

Table 2: The performance of different medical instruction data synthesis methods on the Slake and VQA-RAD dataset. "Base" means the result of pre-trained LVLMs directly answering the questions. Methods with * indicate that medical images are used during instruction data synthesis.

Model	Method	Slake			VQA-RAD		
		Close	Open	Avg	Close	Open	Avg
LLava -v1.5-7B	Base	65.87	44.57	55.22	62.15	36.77	49.46
	LLaVA-Med-instruct	71.39	45.43	58.41	64.54	40.29	52.42
	PMC-instruct	69.47	50.89	60.18	64.14	42.87	53.51
	Huatuo-instruct*	70.91	53.35	62.13	67.33	43.45	55.39
	Lingshu-instruct*	71.39	54.36	62.88	64.54	45.85	55.20
	MedKInstruct(SFT)*	75.72	55.01	65.37	69.72	45.54	57.63
	MedKInstruct(RL)*	79.33	54.77	67.05	70.12	50.13	60.12
HuatuoGPT -Vision-7B	Base	76.20	51.97	64.09	73.71	51.04	62.37
	LLaVA-Med-instruct	73.56	54.79	64.17	70.92	53.38	62.15
	PMC-instruct	75.72	54.16	64.94	72.11	53.72	62.92
	Huatuo-instruct*	77.40	55.05	66.23	75.70	53.47	64.58
	Lingshu-instruct*	78.85	55.81	67.33	76.10	53.48	64.79
	MedKInstruct(SFT)*	80.77	59.36	70.06	77.69	55.15	66.42
	MedKInstruct(RL)*	81.01	62.02	71.52	80.88	57.70	69.29

each baseline method are shown in Table 1. Since the Slake and VQA-RAD datasets do not provide captions, we retrieve image-caption pairs from PubMedVision[2]. For each dataset, the image-caption pairs most similar to one of its images are collected for multimodal instruction data synthesis. For a fair comparison, all these methods generate 10 QA pairs for each image–caption pair.

Evaluation Metrics. Following the setup of previous works (Li et al., 2024; Wu et al., 2025), we evaluate the MedVQA model's performance on close-ended questions (i.e., yes/no questions) using accuracy, and assess open-ended questions by recall. Since the model is not provided with a candidate answer set, this evaluation method is more aligned with real-world scenarios and presents a more challenging setting.

Implementation details. We choose GPT-4o as the LVLM for instruction data synthesis and introduce two base models for MedVQA model training: (1) a general-domain VLM LLaVA-v1.5-7B (Liu et al., 2023) and (2) a medical-domain VLM HuatuoGPT-Vision-7B (Chen et al., 2024b). During instruction data synthesis, we set the hyperparameter k as 20 and m as 2. During MedVQA model training, we employ the LoRA method (Hu et al., 2022). We set the LoRA rank as 16, training epoch as 8, batch size as 4, and leaning rate as 1e-4. The temperature of the model inference is set to 0.0 for stability.

[2] https://huggingface.co/datasets/FreedomIntelligence/PubMedVision

Table 3: Ablation studies of MedKInstruct with the HuatuoGPT-Vision-7B model.

DATASET	**Slake**		
Methods	Close	Open	Avg
MedKInstruct	81.01	62.02	71.52
w/o R_{full}	79.81	61.27	70.54
w/o R_{path}, R_{full}	78.36	60.48	69.43
w/o MMKG	73.80	57.26	65.53
w/o MI	69.95	56.88	63.42
DATASET	**VQA-RAD**		
Methods	Close	Open	Avg
MedKInstruct	80.88	57.70	69.29
w/o R_{full}	78.88	56.48	67.68
w/o R_{path}, R_{full}	77.29	56.12	66.71
w/o MMKG	75.30	52.58	63.94
w/o MI	70.52	52.13	61.32

5.2 Main results

From the experimental results shown in Table 2, we can derive the following conclusions. (1) Overall, the models trained with synthesized instruction data outperform their pre-trained versions. This indicates that the synthetic QA pairs enable the models to learn effective medical knowledge.

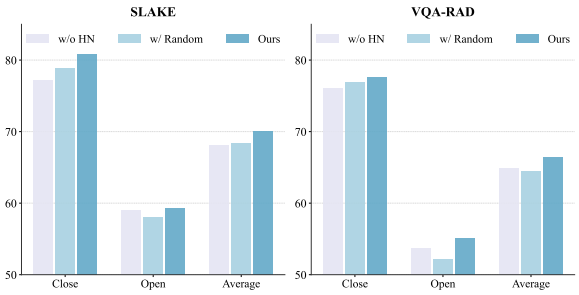


Figure 3: MedKInstruct(HuatuoGPT-Vision-7B) with different MMKG completeness on the Slake (left) and VQA-RAD (right) datasets.

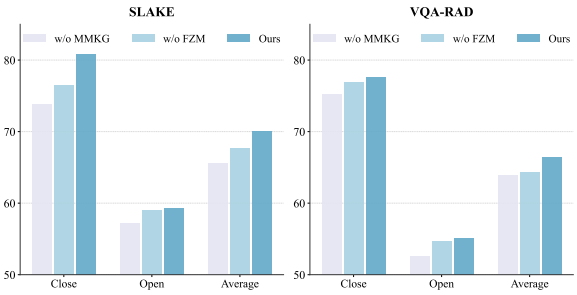


Figure 5: MedKInstruct(HuatuoGPT-Vision-7B) with different knowledge matching settings on the Slake (left) and VQA-RAD (right) datasets.

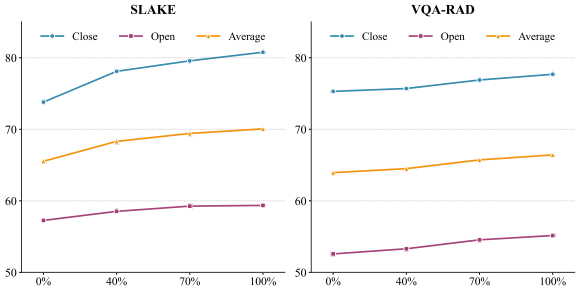


Figure 4: MedKInstruct(HuatuoGPT-Vision-7B) with different negative QA settings on the Slake (left) and VQA-RAD (right) datasets.

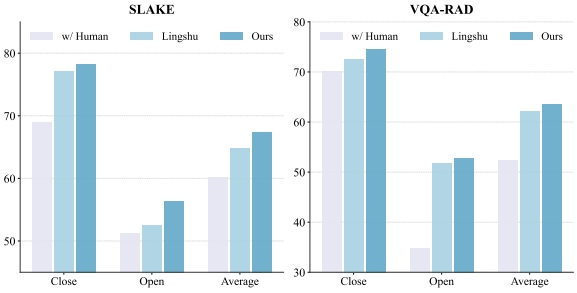


Figure 6: MedKInstruct(HuatuoGPT-Vision-7B) with different generalization settings on VQA-RAD→Slake (left) and Slake→VQA-RAD (right).

(2) The proposed MedKInstruct framework outperforms all baseline methods. More specifically, MedKInstruct(HuatuoGPT-vision-7B) achieves an improvement of 4.16% and 4.50% on two datasets. The main reason is that we introduce the MMKG to augment both the data synthesis process and model training. (3) MedKInstruct with RL achieves approximately 2.0% performance improvement compared to the SFT version. Hence, we can infer that the proposed RL policy effectively strengthens the model's ability to reason and leverage relevant medical knowledge. (4) The superior performance of HuatuoGPT-Vision-7B over LLaVA-1.5-7B under various instruction synthesis settings highlights the importance of multimodal medical pretraining. Hence, medical VLM pretraining remains an important direction for future research.

5.3 Ablation Studies

We perform an ablation study on several MedKInstruct variants: (1) w/o R_{full}: removing the full reward during RL, (2) w/o R_{path}, R_{full}: removing both the MMKG path and full rewards during RL, (3) w/o MMKG: excluding the MMKG information during instruction data synthesis, and (4) w/o MI: excluding the medical image information during instruction data synthesis.

From Table 3, we observe that: (1) the removal of any component leads to a performance drop, highlighting their importance. (2) w/o R_{path}, R_{full} shows an approximately 2.0% decrease compared to MedKInstruct. This indicates that introducing the MMKG path as a supervision signal helps the model learn relevant intermediate medical knowledge. (3) w/o R_{full} exhibits a slight performance drop of about 1.0%. We can infer that R_{full} guides the model to focus on intermediate reasoning knowledge during RL. (4) The performance of w/o MMKG and w/o MI drops more than 5.2% compared to MedKInstruct. This demonstrates that the MMKG and medical image information are essential for LVLMs to synthesize high-quality medical instruction data.

5.4 Detailed analysis

In this section, we present a detailed analysis of the components in the MedKInstruct framework. Then, we evaluate its generalization ability.

Analysis of hard negative questions. To evaluate the effectiveness of the hard negative QAs, we design an A/B test. In experiment A, hard negative QAs are removed (w/o HN). In experiment

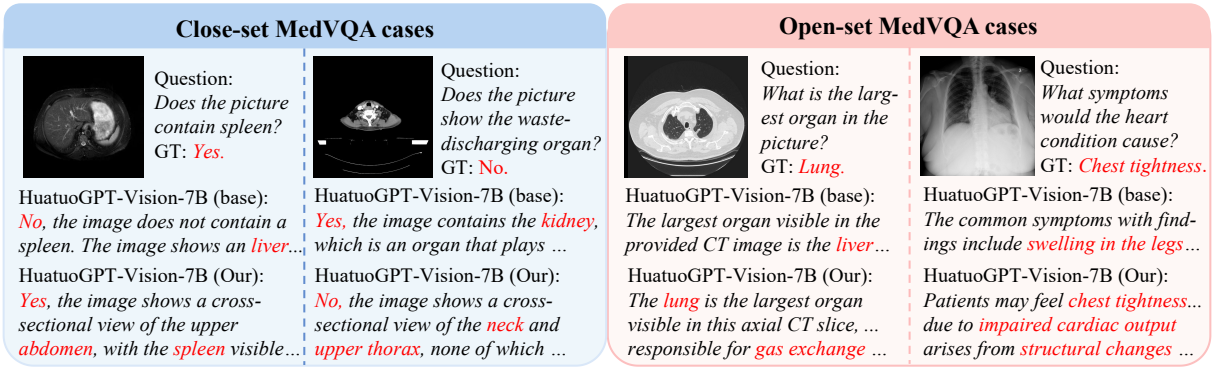


Figure 7: Cases of how the MMKG-based synthesis instruction data (MedKInstruct) contributes to training a stronger MedVQA model (HuatuoGPT-Vision-7B).

B, we consider two settings: (1) the original MedKInstruct (Ours), and (2) negative QAs generated based on randomly selecting diseases (w/ random). From Figure 3, we observe that (1) the original MedKInstruct outperforms the w/o HN setting by over 1.6%, and (2) in contrast, the w/ random setting shows no noticeable improvement over the w/o HN setting. These results highlight the necessity of retrieving highly similar diseases for hard negative QA construction.

Analysis of knowledge sparse. To further investigate the importance of MMKG, we synthesize the instruction data by incorporating knowledge triples at proportions of {0%, 40%, 70%, and 100%}. As shown in Figure 4, the performance of MedKInstruct consistently improves with the completeness of MMKG on both datasets. Hence, we can conclude that: (1) the performance gain is not due to a few specific knowledge triples, but rather to the overall knowledge of the MMKG, and (2) constructing a more complete MMKG constitutes a promising direction for future work.

To more intuitively illustrate how the incorporated MMKG assists the model in learning relevant medical knowledge, we present several representative cases in Figure 7. Compared to the base version, HuatuoGPT-Vision-7B trained with MMKG-based instruction data achieves superior performance on both close-ended and open-ended medical question answering tasks.

Analysis of the exact and fuzzy knowledge matching. To evaluate the effectiveness of our two-stage knowledge matching policy, we first remove the MMKG triples obtained via fuzzy matching from the knowledge-enhanced dataset, and then use the remaining knowledge to synthesize multimodal instruction data (denoted as w/o FZM). As shown in Figure 5, we observe that: (1) w/o FZM outperforms w/o MMKG on both datasets, demonstrating the effectiveness of exact knowledge matching; and (2) compared with the original MedKInstruct method, w/o FZM exhibits a performance drop of more than 2.1%. These results indicate that fuzzy matching further improves the recall of relevant medical knowledge.

Analysis of generalization ability. To investigate the generalization ability of the proposed MedKInstruct framework, we first employ the Slake (Liu et al., 2021) and VQA-RAD (Lau et al., 2018) based image-caption pairs to synthesize instruction data, and then evaluate the trained models on each other's test sets (e.g., training on Slake and testing on VQA-RAD). We compare MedKInstruct with Lingshu (Team et al., 2025) and the original manually annotated short QA pairs (denoted as w/ human). As shown in Figure 6, MedKInstruct achieves more than a 1.5% performance improvement over other methods, demonstrating its generalization ability.

6 Conclusions

This paper proposes MedKInstruct, a novel framework for multimodal medical instruction data synthesis. Unlike previous approaches that primarily generate QA pairs focused on visual attributes, MedKInstruct introduces an MMKG to synthesize knowledge-intensive instruction data. To leverage this data for training a stronger MedVQA model, we further design an MMKG path–based reward function that guides the model to reason over relevant medical knowledge during RL. Comprehensive evaluations on two public MedVQA benchmarks, Slake and VQA-RAD, demonstrate the effectiveness of MedKInstruct.

Limitations

Although MedKInstruct has demonstrated its effectiveness on existing MMKG-based MedVQA datasets, this work still has two main limitations. (1) Due to the limited scale of existing MMKGs, our experiments focus on only 11 major systems of the human body. As a result, we do not fully evaluate the generalization ability of MedKInstruct on more fine-grained MedVQA scenarios (e.g., gallbladder). Future work could construct larger and more comprehensive MMKGs to support knowledge-based methods on a broader range of MedVQA tasks. (2) Owing to GPU resources, our reinforcement learning experiments are conducted only on 7B-scale models. Therefore, we are unable to explore the effectiveness of MedKInstruct on larger-scale models (e.g., 32B). Future work may explore applying the MedKInstruct framework to larger open-source models to further enhance the performance and reasoning capabilities of MedVQA systems.

Ethical Statement

The experiments in this paper are conducted using publicly available MedVQA datasets, including PubMedVision, Slake and VQA-RAD. All datasets contain no offensive content and are released under the Apache License 2.0 or Creative Commons Attribution 4.0 License, which permit redistribution and re-annotation with proper attribution. The MMKG and models used in this work, including MKGF, LLaVA-v1.5-7B, and HuatuoGPT-Vision-7B, are open-source for scientific research purposes. Accordingly, this work does not raise any copyright or licensing concerns.

Acknowledgments

This work is supported by the Shanghai Natural Science Foundation Project under Grant 25ZR1402116.

References

Jiawei Chen, Dingkang Yang, Yue Jiang, Yuxuan Lei, and Lihua Zhang. 2024a. Miss: A generative pretraining and fine-tuning approach for med-vqa. In *Artificial Neural Networks and Machine Learning – ICANN 2024*, pages 299–313, Cham. Springer Nature Switzerland.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. 2024b. Towards injecting medical visual knowledge into multimodal LLMs at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7346–7370, Miami, Florida, USA. Association for Computational Linguistics.

Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. 2025. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. *Preprint*, arXiv:2506.04207.

Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. 2023. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, Dubrovnik, Croatia. Association for Computational Linguistics.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Xiaofei Huang and Hongfang Gong. 2024. A dual-attention learning network with word and sentence embedding for medical visual question answering. *IEEE Transactions on Medical Imaging*, 43(2):832–845.

Yue Jiang, Jiawei Chen, Dingkang Yang, Mingcheng Li, Shunli Wang, Tong Wu, Ke Li, and Lihua Zhang. 2025. Comt: Chain-of-medical-thought reduces hallucination in medical report generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. 2025. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *Preprint*, arXiv:2503.13939.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Xiao Liang, Di Wang, Haodi Zhong, Quan Wang, Ronghan Li, Rui Jia, and Bo Wan. 2024. Candidate-heuristic in-context learning: A new framework for enhancing medical visual question answering with llms. *Information Processing & Management*, 61(5):103805.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Qihao Liu, Luoxin Ye, Wufei Ma, Yu-Cheng Chou, and Alan Yuille. 2025. Generative adversarial reasoner: Enhancing llm reasoning with adversarial reinforcement learning. *Preprint*, arXiv:2512.16917.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 353–367. PMLR.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347. Springer.

Maohao Shen, Guangtao Zeng, Zhenting Qi, Zhang-Wei Hong, Zhenfang Chen, Wei Lu, Gregory W. Wornell, Subhro Das, David Daniel Cox, and Chuang Gan. 2025. Satori: Reinforcement learning with chain-of-action-thought enhances LLM reasoning via autoregressive search. In *Forty-second International Conference on Machine Learning*.

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *Preprint*, arXiv:2506.07044.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *Preprint*, arXiv:2504.08837.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439.

Lai Wei, Wenkai Wang, Xiaoyu Shen, Yu Xie, Zhihao Fan, Xiaojin Zhang, Zhongyu Wei, and Wei Chen. 2024. Mc-cot: A modular collaborative cot framework for zero-shot medical-vqa with llm and mllm integration. *arXiv preprint arXiv:2410.04521*.

Yinan Wu, Yuming Lu, Yan Zhou, Yifan Ding, Jingping Liu, and Tong Ruan. 2025. Mkgf: A multi-modal knowledge graph based rag framework to enhance lvlms for medical visual question answering. *Neurocomputing*, 635:129999.

Dexuan Xu, Yanyuan Chen, Jieyi Wang, Yue Huang, Hanpin Wang, Zhi Jin, Hongxing Wang, Weihua Yue, Jing He, Hang Li, and Yu Huang. 2024. MLeVLM: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4977–4997, Bangkok, Thailand. Association for Computational Linguistics.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, and 1 others. 2025. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19985–19995.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, and 1 others. 2023. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024a. Pmc-vqa: Visual instruction tuning for medical visual question answering. *Preprint*, arXiv:2305.10415.

Zhilin Zhang, Jie Wang, Zhanghao Qin, Ruiqi Zhu, and Xiaoliang Gong. 2024b. Efficient bilinear attention-based fusion for medical visual question answering. *arXiv preprint arXiv:2410.21000*.

A Prompt

We present the prompts used for knowledge-enhanced dataset preparation in Table 4 and for multi-modal instruction data synthesis in Table 5.

B Datasets

For the Slake dataset (Liu et al., 2021), we use the English version, which contains 642 radiology images and 7,033 QA pairs. We follow the data split

Knowledge Enhanced Dataset Preparation

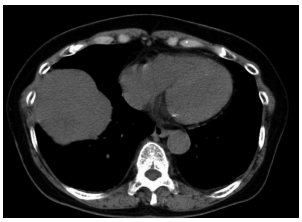
Entity Extraction (disease & organ)

Output JSON only.

```
You are an expert medical AI assistant specializing in multimodal analysis of clinical images
and text.
Your task is to extract exactly TWO pieces of information:
1. disease: The specific pathological condition (e.g., “Brain Tumour”, “Osteoporosis”,
“Retinal Detachment”). If uncertain, provide the most likely diagnosis.
2. organ: The anatomical site or organ involved (e.g., “Brain”, “Bone Marrow”, “Retina”).
Guidelines:
- Use only standardized medical terminology.
- If the description mentions multiple diseases/organs, pick the PRIMARY one.
- NEVER say “I don’t know” or leave fields blank.
- Return ONLY a valid JSON object with keys “disease” and “organ”. No other text.
Example output: {"disease":"Multiple Myeloma","organ":"Bone Marrow"}
Medical description: {caption}
```

Case (Input → Output)

Input (image + question/caption)



Caption: The medical image is a computed tomography (CT) scan of the chest, specifically focusing on the middle lobe. The image shows a homogenous, sharply defined lung mass in ..., which appears well-circumscribed and uniform in density, indicating a lack of internal heterogeneity. The surrounding lung parenchyma appears normal with no visible signs of infiltration or distortion from the mass. No significant lymphadenopathy or pleural effusion is observed...

Output

```
{"disease":"Lung mass","organ":"Lung"}
```

Entity Consistency Check

Answer true/false only.

```
You are an expert in medical terminology consistency evaluation. Determine whether the
following two medical terms refer to approximately the same {entity_type}. Answer only “true”
or “false” without explanation.
Term 1: {original entity} Term 2: {matched entity}
```

Case (Input → Output)

Input (text only)

Entity type: `Disease` Term 1: `Lung mass` Term 2: `Pulmonary mass`

Output

`True`

Table 4: The two-stage prompt engineering framework for knowledge-enhanced dataset preparation. The pipeline first extracts structured entities (Stage 1) and then verifies terminology consistency (Stage 2).

used in previous work, consisting of 4,919, 1,053, and 1,061 QA pairs for training, validation, and testing, respectively. For the VQA-RAD dataset (Lau et al., 2018), it contains 315 radiology images and 3,515 QA pairs. Following the previous split, 1,797 QA pairs are used for training, and 451 QA pairs for testing. Note that Slake is the only benchmark explicitly aligned with MMKG entities, and choosing these datasets enables a fair comparison with prior works.

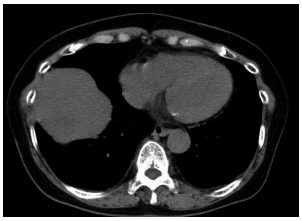
Multi-Modal Instruction Data Synthesis

Data Synthesis (QA Generation w/ KG Context) *Q starts with What/Why/How; no disease names.*

```
You are {scene_prompt[role]} and {scene_prompt[prompt]}.
Your task is to generate distinct and generalizable question-answer pairs based on the
provided medical image and supporting materials.
Reference materials include: - Relevant knowledge graph information: {kg_context}
- Image caption: {caption}
- Reference QA pairs: {reference_QA}
Question constraints: - Each question must begin with What, Why, or How.
- Disease names must not appear in any question.
Answer constraints: - Answers must be concise (2-3 sentences).
- Answers must be fact-based and grounded in standard medical knowledge.
Strict output format:  Q1: ...  A1: ...  Q2: ...  A2: ...
```

Case (Input → Output)

Input (image + materials)



Caption: The medical image is a CT scan of the chest, specifically focusing on ...

Reference QA: Q: Can you provide a detailed analysis...? A: Based on the analysis... the mass is homogenous...

KG Context: chest pain # symptom # Pulmonary Mass # location # Lung # belong to # Respiratory System

Output

```
Q: What symptom might develop if this mass enlarges significantly?
A: As part of the respiratory system, significant enlargement of the mass could compress
surrounding structures, such as ..., leading to discomfortable or pain in the chest.
```

Answer Condensation (Key Answer Summary) *Output words only; no punctuation.*

```
prompt = (
    "Based on the following medical question and model answer, "
    "give a key answer as brief as possible. "
    "Output only the summary words, no punctuation, no extra text.\n \n"
    f"Question: {question}\n"
    f"Answer: {answer}"
)
```

Case (Input → Output)

Input (text only)

Question: `What symptom might develop if this mass enlarges significantly?`

Answer: `As part of the respiratory ... , leading to discomfortable or pain in the chest.`

Output

```
Chest pain
```

Table 5: Prompts for multimodal instruction data synthesis. The pipeline synthesizes diverse QA pairs using Knowledge Graphs (Stage 1) and condenses answers for key information extraction (Stage 2).

Table 6: Analysis of knowledge matching.

Datasets	Image-caption	Exact	Fuzzy
Slake	642	72.11	27.89
VQA-RAD	315	73.46	26.54
Total	957	72.55	27.45

C Knowledge matching

Table 6 presents an analysis of knowledge matching strategies on the Slake and VQA-RAD datasets. The results indicate that approximately 73% of entities are aligned through exact string-based matching, while the remaining 27% are matched via fuzzy matching. These findings demonstrate that both strategies are essential for achieving robust and comprehensive entity alignment.