Detecting Any Instruction-to-Answer Interaction Relationship:UniversalInstruction-Vision Navigator for Med-VQA

Zhongze Wu [1] * Hongyan Xu [2] * Yitian Long [3] Shan You [4] Xiu Su [1 5] † Jun Long [1] † Yueyi Luo [1] Chang Xu [5]

*Equal contribution,†Correspondence, [1]Central South University, Changsha, Hu nan, China [2]University of New South Wales, Sydney, Australia

[3]Vanderbilt University, Nashville, Tennessee, USA [4]SenseTime,[5]University of Sydney, Sydney, Australia.

ICML International Conference On Machine Learning

Abstract

Medical Visual Question Answering (Med-VQA) interprets complex medical imagery using user instructions for precise diagnostics, yet faces challenges due to diverse, inadequately annotated images. In this paper, we introduce the Universal Instruction-Vision Navigator (Uni-Med) framework for extracting instruction-to-answer relationships, facilitating the understanding of visual evidence behind responses. Specifically, we design the Instruct-to-Answer Clues Interpreter (IAI) to generate visual explanations based on the answers and mark the core part of instructions with "real intent" labels. The IAI-Med VQA dataset, produced using IAI, is now publicly available to advance Med-VQA research. Additionally, our Token-Level Cut-Mix module dynamically aligns visual explanations with image patches, ensuring answers are traceable and learnable. We also implement intention-guided attention to minimize non-core instruction interference, sharpening focus on 'real intent'. Extensive experiments on SLAKE datasets show Uni-Med's superior accuracies (87.52 %closed, 86.12 % overall), outperforming MedVInT-PMC-VQA by 1.22 % and 0.92 %. Code and dataset are available at: https://github.com/zhongzee/Uni-Med-master.

Motivation

1.Despite the significant advancements made by multi-modal large language models (MLLMs) in biomedical applications, their practical application remains hindered by a high reliance on text and image labels. This dependency often leads to inaccuracies in medical diagnostics due to modal interference and the inability of current models to effectively interpret complex medical data.

2.Existing medical VQA systems lack efficient mechanisms for mapping instructions to accurate answers, often resulting in responses that do not align with the visual data provided. This issue is compounded by the absence of training strategies that focus on user-intent, leading to a disconnect between the questions posed and the answers generated.

3.There is a pressing need for models that can offer more granular visual explanations and align these explanations with textual instructions. The existing methods do not adequately support dynamic, instruction-specific feature enhancements, resulting in potential misalignments between learned representations and actual medical queries.

4.To address current challenges, this paper introduces the Universal Instruction-Vision Navigator (Uni-Med) framework, which integrates the Instruct-to-Answer Clues Interpreter (IAI) to enhance Med-VQA interpretability. Uni-Med refines Med-VQA by aligning visual explanations with user instructions marked for 'real intent', as shown in Figure 1. This approach not only improves the accuracy of answers but also ensures they are traceable and learnable through mechanisms like Token-Level Cut-Mix and Intention-guided Attention.

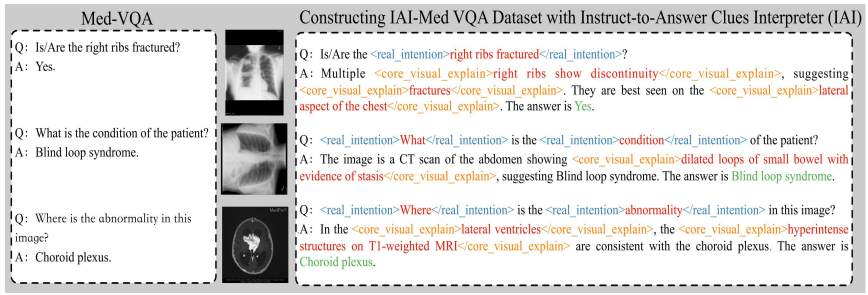

Figure 1. Details of the construction of the IAI-Med VQA dataset with the Instruct-to-Answer Clues Interpreter (IAI). We design a Universal-Navigator Prompt (UNP) to guide MLLM to articulate the reasoning behind answers based on the visual content present in medical images and the context provided by existing question-answer pairs.

Uni-Med FrameWork

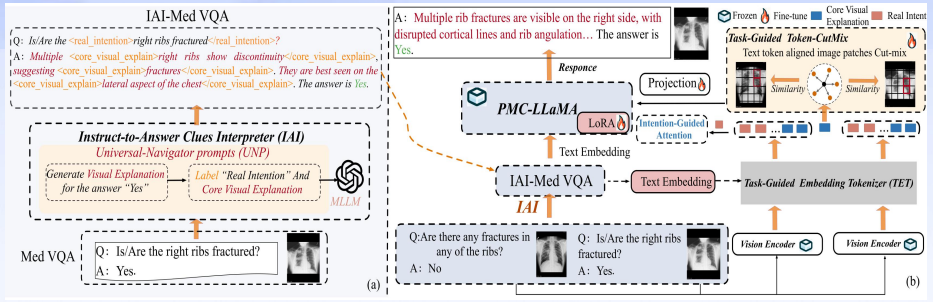

Figure 2. The Uni-Med Training Paradigm. (a): The IAI module, where UNP prompts MLLMs to identify instruction's "real intent" and generate visual explanations. The most aligned patches are selected to perform feature-level enhancement by TC-Mix. Intention-Guided Attention focuses LLM on the "real intent" to minimize modal interference.

Instruct-to-Answer Clues Interpreter

Prompting GPT-4V to generate task related marking based on COT explanation.

```
payload = {
    "model": "gpt-4-vision-preview",
    "messages": [{
        "role": "system","content": (
            """You are an AI assistant specialized in biomedical topics. Given the latest question about a medical image and its direct answer,your task is to use <task_related></task_related> tags to enclose the direct question to choose the most important emplement to reflect the user's intentions. And <task_related_visual></task_related_visual> tags to wrap the description of the visual content that supports the answer.
            1)The content should be related to the critical visual elements from the image and be as precise as possible…
            6)When marking answers, try to mark the visual content that exists in the visual content of the image and supports the answer,
            7).Identify and label the key medical terms in the question that are directly related to the diagnosis or condition being inquired about, and focus your answer on these terms…"""){},{"role": "user","content": [{"type": "text","text": ( f"The current question and answer of uploaded image is {question} and {answer}…"
            Here are some guidelines:
            1) When labeling questions, label verbs and nouns rather than predicates…
            3)To avoid wrapping invalid and redundant information…"""
            ){,{"type": "image_url","image_url": {"url": f"{base64_image}"}}]}],"max_tokens": 200}
```

Figure 3. The simplified version of prompt MLLM to generate task related explanation.

TC-Mix

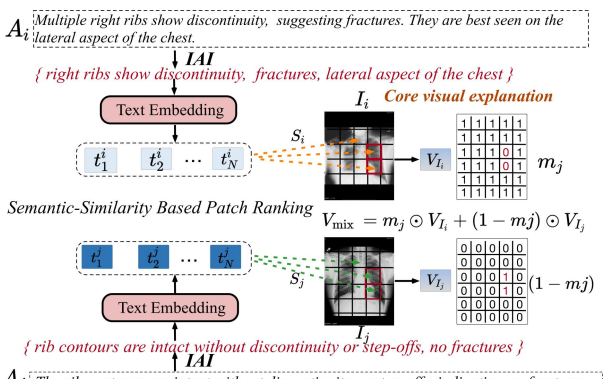

Figure 4. The details of Task-guided Token-Level Cut-Mix.

Algorithm 1 Token-Level Cut-Mix (TC-Mix)

Input: Image groups $\mathcal{G} = \{G_{o,p}\}$, Real Intent Marked answers $\mathcal{A}_{o,p} = \{A_i^{o,p}\}_{i=1}^{N_{o,p}}$
Output: Cut-Mixed feature vectors V_{mix}
for each group $G_{o,p}$ in $\mathcal{G}$ **do**
  for each answer $A_j^{o,p}$ in $\mathcal{A}_{o,p}$ **do**
    Concatenate marked answer tokens $A_j^{M_{o,p}} \leftarrow \text{Concat}(a_1^{o,p}, a_2^{o,p}, ..., a_n^{o,p})$ from $A_i^{o,p}$
    Compute concatenated text embedding $t_j^{o,p} \leftarrow \mathcal{E}_{\text{txt}}(A_j^{M_{o,p}})$
  end for
  for each image $I_i^{o,p}$ in $G_{o,p}$ **do**
    Decompose $I_i^{o,p}$ into K patches $P_{I_i}^{o,p} = \{p_{i,k}\}_{k=1}^{K}$
    Extract image features $v_{i,k} \leftarrow \mathcal{E}_{\text{img}}(p_{i,k})$, for all k
    Calculate semantic similarity scores $S'_{i,k} \leftarrow \kappa(v_{i,k}, t_j^{o,p})$, for all k
    Select top-N patches $T_S^N \leftarrow \text{top}_N(S'_{i,k})$
  end for
  for each pair (I_i, I_j) in $G_{o,p}$ **do**
    Generate binary masks m_i, m_j for T_S^N
    Perform Cut-Mix: $V_{\text{mix}} \leftarrow (m_i \odot V_{I_i}) + (1 - m_j) \odot V_{I_j}$
  end for
end for
Return: V_{mix}

Results on VQA-RAD & SLAKE Dataset

Table 1. The comparisons between Uni-Med and other SOTA methods. "Closed": closed-ended questions with a yes/no answer. "Open": open-ended questions with no fixed form answer.

Methods	Pretrain images	VQA-RAD			SLAKE		
		Closed	Open	Overall	Closed	Open	Overall
Data Enhancement							
BAN-MEVF+DAVQA (Kafle et al., 2017)	-	76.2	51.2	66.2	-	-	-
MEVF+SAN (Nguyen et al., 2019b)	-	69.7	49.2	57.1	78.4	75.3	76.5
MEVF+BAN (Nguyen et al., 2019b)	-	77.2	49.2	66.1	79.8	77.8	78.6
BAN-MEVF+CR (Zhan et al., 2020)	-	79.3	52.4	68.5	-	-	-
SEADA (Tang et al., 2020a)	-	79.6	56.6	70.4	-	-	-
HQS (Gupta et al., 2021)	-	63.4	12.9	41.1	-	-	-
CMSA-MTPT (Gong et al., 2021)	-	77.8	52.8	67.9	-	-	-
VQAMIX (Gong et al., 2022b)	-	79.6	56.6	70.4	-	-	-
Pretrain-finetuning Model							
MMBERT (Khare et al., 2021)	-	76.8	58.3	66.9	-	-	-
PUBMEDCLIP-MEVF (Eslami et al., 2021)	80K	78.1	48.6	66.5	76.2	79.9	77.6
CPRD+BAN (Liu et al., 2021a)	-	77.9	52.5	67.8	83.4	79.5	81.1
MMBERT (Khare et al., 2021)	-	76.9	55.3	66.9	83.4	79.5	81.1
MTL (Cong et al., 2022)	87K	79.8	69.8	75.8	86.1	80.2	82.5
M3AE (Chen et al., 2022)	298K	83.4	67.2	77	87.8	80.3	82.5
PTUnifier (Chen et al., 2023)	-	-	-	78.3	-	-	85.2
RAMM (Yuan et al., 2023)	700K	-	-	78.27	-	-	86.05
MUMC (Li et al., 2023b)	387K	84.2	71.5	79.2	-	-	84.9
LLaVA (7B) (Liu et al., 2023)	-	65.07	50.00	-	63.22	78.18	-
LLaVA-Med (7B) (Li et al., 2023a)	$1M$	84.19	61.52	-	85.34	83.08	-
LLaVA-Med (13B) (Li et al., 2023a)	$1M$	81.98	64.39	-	85.58	84.97	-
MedVInT-PMC-VQA (Zhang et al., 2023b)	140K	86.8	73.7	81.6	86.3	84.5	85.2
Pretrain-finetuning and Data Enhancement							
Uni-Med (7B)	140K	**87.22**	**74.21**	**82.05**	**87.52**	**85.34**	**86.12**

Conclusions

In this paper, we introduce the Uni-Med framework, an approach that significantly enhances the interpretation of complex medical images through user instructions,which make the answer 'traceable' and 'learnable'.

We design an Instruct-to-Answer Clues Interpreter (IAI) to generate the IAI-Med VQA dataset, which marks the "real intent" of instructions and generates corresponding visual explanations. To minimize errors in medical image analysis, we develop an Universal-Navigator Prompt (UNP) to enhance medical image understanding and reasoning of MLLM.

We implement a task-guided Token-level Cut-Mix (TC-Mix) strategy that leverages visual explanation aligned with user instructions, mapping them to the most relevant blocks in medical images for token-level enhancement.

References

1. Agarwal, V., Shetty, R., and Fritz, M. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.
2. Binh D. Nguyen, Thanh-Toan Do, B. X. N. T. D. E. T. Q. D. T. Overcoming data limitation in medical visual question answering. In MICCAI, 2019.
3. Cao, Y., Su, X., Tang, Q., You, S., Lu, X., and Xu, C. Searching for better spatio-temporal alignment in fewshot action recognition. Advances in Neural Information Processing Systems, 35:21429-21441, 2022.
4. Cao, Y., Tang, Q., Yang, F., Su, X., You, S., Lu, X., and Xu, C. Re-mine, learn and reason: Exploring the cross-modal semantic correlations for language-guided hoi detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 23492-23503, 2023.
5. Cao, Y., Tang, Q., Su, X., Chen, S., You, S., Lu, X., and Xu, C. Detecting any human-object interaction relationship: Universal hoi detector with spatial prompt learning on foundation models. Advances in Neural Information Processing Systems, 36, 2024.
6. Chen, J., Yang, D., Jiang, Y., Lei, Y., and Zhang, L. Miss: A generative pretraining and finetuning approach for medvqa, 2024.
7. Chen, Z., Du, Y., Hu, J., Liu, Y., Li, G., Wan, X., and Chang, T.-H. Multi-modal masked autoencoders for medical vision-and-language pre-training. In International Conference on Medical Image Computing and ComputerAssisted Intervention. Springer, 2022.
8. Chen, Z., Diao, S., Wang, B., Li, G., and Wan, X. Towards unifying medical vision-and-language pre-training via soft prompts. arXiv preprint arXiv:2302.08958, 2023.
9. Cong, F., Xu, S., Guo, L., and Tian, Y. Caption-aware medical vqa via semantic focusing and progressive crossmodality comprehension. In Proceedings of the 30th ACM International Conference on Multimedia, pp. 3569$3577,2022$.